ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και
Συστημάτων Πληροφορικής

Τεχνικές Μηχανικής Μάθησης για Εκτίμηση Τηλεθέασης με Δεδομένα από Μέσα Κοινωνικής Δικτύωσης και Μηχανές Αναζήτησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Κωνσταντίνας-Μαρίας Θ. Γιαννακοπούλου

Επιβλέπουσα: Ιωάννα Ρουσσάκη
Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Αύγουστος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και
Συστημάτων Πληροφορικής

Τεχνικές Μηχανικής Μάθησης για Εκτίμηση Τηλεθέασης με Δεδομένα από Μέσα Κοινωνικής Δικτύωσης και Μηχανές Αναζήτησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Κωνσταντίνας-Μαρίας Θ. Γιαννακοπούλου

Επιβλέπουσα: Ιωάννα Ρουσσάκη
Επίκουρη Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Αυγούστου 2020.

.....	.....	.....
Ιωάννα Ρουσσάκη	Μιλτιάδης Αναγνώστου	Συμεών Παπαβασιλείου
Επ. Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Αύγουστος 2020

.....
Κωνσταντίνα-Μαρία Θ. Γιαννακοπούλου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ραγδαία εξάπλωση των μέσων κοινωνικής δικτύωσης και γενικότερα του διαδικτύου και κυρίως των μηχανών αναζήτησης σε αυτό έχει συντελέσει στην παραγωγή ενός τεράστιου όγκου δεδομένων, ο οποίος μπορεί να αξιοποιηθεί αποδοτικά με τη βοήθεια του ταχέως αναπτυσσόμενου πεδίου της Μηχανικής Μάθησης για την εξαγωγή χρήσιμων συμπερασμάτων και προβλέψεων. Η προβλεπτική ικανότητα αυτών των δεδομένων έχει πλέον αναγνωριστεί μέσα από μελέτες σε διάφορους τομείς του επιστητού, όπως στην οικονομία, την υγεία και την πολιτική μεταξύ άλλων. Στην παρούσα διπλωματική εργασία αντιμετωπίζουμε το πρόβλημα της εκτίμησης της τηλεθέασης προγραμμάτων με δεδομένα που συλλέγονται από το μέσο κοινωνικής δικτύωσης Twitter και από την πλατφόρμα Google Trends, η οποία παρέχει στατιστικά στοιχεία για τις αναζητήσεις στη μηχανή αναζήτησης της Google.

Από αυτά τα δεδομένα εξάγονται σε διάφορα χρονικά παράθυρα χαρακτηριστικά, τα οποία είτε αντιστοιχούν σε ποσοτικούς δείκτες, όπως ο όγκος των δημοσιεύσεων στο Twitter και ο όγκος των αναζητήσεων στη μηχανή αναζήτησης της Google είτε προκύπτουν από ανάλυση συναισθήματος στο κειμενικό περιεχόμενο των tweets. Με αυτά ή με κάποια από αυτά, αφού έχει προηγηθεί μείωση της διαστατικότητας, εκπαιδεύονται διάφορα μοντέλα παλινδρόμησης με πληθώρα διαφορετικών μεθόδων και αλγορίθμων Μηχανικής Μάθησης. Ενδεικτικά, χρησιμοποιούνται απλές τεχνικές γραμμικής και πολυωνυμικής παλινδρόμησης υλοποιημένες με τη μέθοδο των ελαχίστων τετραγώνων, τεχνικές κανονικοποίησης, όπως ridge, LASSO και elastic net, πιθανοτικά μοντέλα με γκαουσιανές διεργασίες, δέντρα αποφάσεων, μέθοδοι συλλογικής μάθησης, όπως τυχαία δάση και gradient boosting μηχανές, νευρωνικά δίκτυα, όπως πολυεπίπεδα perceptron και μηχανές διανυσμάτων υποστήριξης.

Η αξιολόγηση των μοντέλων πραγματοποιείται με διάφορες μετρικές και τα αποτελέσματα συγκρίνονται με αυτά προηγούμενων εργασιών. Τόσο η διαδικασία της εκπαίδευσης όσο και η διαδικασία του ελέγχου βασίζονται συγκεκριμένα στα δεδομένα τηλεθέασης της ιταλικής σατιρικής εκπομπής Le Iene για δύο ημερολογιακά έτη. Τέλος, τα συμπεράσματα που εξάγονται ενισχύουν την αρχική υπόθεση ότι ο συνδυασμός δεδομένων από τις πλατφόρμες Twitter και Google Trends μπορεί να αποδειχθεί ικανός για την εκτίμηση της τηλεθέασης και η προσθήκη του τελευταίου να αποτελέσει καταλύτη για τη βελτίωση της απόδοσης πιο διαδεδομένων μοντέλων που χρησιμοποιούν μόνο δεδομένα από το Twitter.

Λέξεις κλειδιά:

Επιστήμη Δεδομένων, Κοινωνικά δίκτυα, Μηχανές Αναζήτησης, Twitter, Google Trends, Εξόρυξη Δεδομένων, Εξαγωγή Γνώσης, Μηχανική Μάθηση, Παλινδρόμηση, Ανάλυση Συναισθήματος, Τηλεθέαση

Abstract

The rapid spread of the Internet in general and search engines and social media in particular has led to the production of a huge amount of data, which can be used efficiently with the help of the rapidly growing field of Machine Learning to draw useful conclusions and predictions. The predictive power of these data has been recognized through several studies in various fields, such as economics, health and politics among others. In this dissertation we address the problem of estimating TV viewership with data collected from Twitter microblogging service and Google Trends platform, which provides statistics on searches in the Google search engine.

Various features are extracted from these data in several time windows. Some of them correspond to quantitative indicators, such as the volume of tweets and the volume of searches in Google search engine, while others have emerged from sentiment analysis of the textual content of tweets. Various regression models are trained with these or some of these features, after implementing dimensionality reduction, as input and by deploying a variety of different methods and Machine Learning algorithms. Indicatively, simple linear and polynomial regression techniques are implemented with the ordinary least squares method, regularization techniques are used, such as ridge, LASSO and elastic net, as well as probabilistic models with gaussian processes, decision trees, ensemble learning methods, such as random forests and gradient boosting machines, neural networks such as multilayer perceptron and support vector machines.

The evaluation of the models is performed on the basis of various metrics and the results are compared with those of previous work. Both training and testing process is based specifically on the TV ratings of the Italian satirical show Le Iene for two years. Finally, the conclusions support the initial hypothesis that the combination of data from Twitter and Google Trends platforms may prove capable of estimating TV viewership and the addition of the latter could be a catalyst for improving the performance of more popular models which make use only of data collected from Twitter.

Keywords:

Data Science, Social Media, Search Engines, Twitter, Google Trends, Data Mining, Knowledge Extraction, Machine Learning, Regression, Sentiment Analysis, TV Viewership

Ευχαριστίες

Καταρχάς, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια Ιωάννα Ρουσσάκη για την ευκαιρία που μου έδωσε να ενασχοληθώ με ένα τόσο ενδιαφέρον και επίκαιρο θέμα στα πλαίσια της διπλωματικής μου εργασίας. Η πολύτιμη βοήθεια και η καθοδήγηση που μου προσέφερε ήταν καθοριστικής σημασίας για την ολοκλήρωση της εργασίας αλλά και τη θεμελίωση της μετέπειτα ερευνητικής μου πορείας. Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Χρήστο Ματσούκα για τις χρήσιμες συμβουλές καθόλη τη διάρκεια εκπόνησης της εργασίας, ιδιαίτερα κάτω από τις ιδιάζουσες συνθήκες που εισήγαγε στην καθημερινότητά μας η εμφάνιση της πανδημίας COVID-19. Επιπλέον, θα ήθελα να ευχαριστήσω τους καθηγητές Μιλτιάδη Αναγνώστου και Συμεών Παπαβασιλείου που παρευρέθησαν στην παρουσίαση της παρούσας διπλωματικής εργασίας και συνέθεσαν την τριμελή επιτροπή.

Φυσικά, δε θα μπορούσα να μην ευχαριστήσω τους φίλους μου, που ήταν δίπλα μου στις πιο χαρούμενες αλλά και στις πιο δύσκολες στιγμές και δημιουργήσαμε μαζί μοναδικές αναμνήσεις από τα φοιτητικά μας χρόνια. Τέλος, θα ήθελα να αφιερώσω την εργασία αυτή στην οικογένεια μου -τους γονείς μου, Μαίρη και Θωμά, και την αδερφή μου, Μελίνα- τους οποίους υπερευχαριστώ για την αμέριστη αγάπη, στήριξη και υπομονή που μου έχουν δείξει όλα αυτά τα χρόνια. Η πίστη τους στις δυνατότητές μου και η ενθάρρυνσή τους να ακολουθήσω τα όνειρά μου αποδείχτηκαν τα πιο πολύτιμα εφόδια στη μέχρι τώρα πορεία μου και θα μου δίνουν ώθηση και στη συνέχεια σε κάθε νέα μου προσπάθεια.

Κωνσταντίνα-Μαρία Γιαννακοπούλου
Αθήνα, 31 Αυγούστου 2020

Περιεχόμενα

Περίληψη **5**

Abstract **7**

Ευχαριστίες **9**

1 Εισαγωγή **19**
1.1 Μέσα κοινωνικής δικτύωσης . . . 19
1.1.1 Twitter . . . 20
1.2 Μηχανές αναζήτησης . . . 21
1.2.1 Google Trends . . . 22
1.3 Τηλεοπτική βιομηχανία και σημασία μετρήσεων τηλεθέασης . . . 23
1.4 Κίνητρο και συμβολή εργασίας . . . 24
1.5 Διάρθρωση εργασίας . . . 25

2 Συναφής βιβλιογραφία **27**
2.1 Προβλέψεις σε διάφορους τομείς . . . 27
2.1.1 Με δεδομένα από μέσα κοινωνικής δικτύωσης . . . 27
2.1.2 Με δεδομένα από μηχανές αναζήτησης . . . 28
2.2 Προβλέψεις στη βιομηχανία της τηλεόρασης και του κινηματογράφου . . . 29
2.2.1 Με δεδομένα από μέσα κοινωνικής δικτύωσης . . . 29
2.2.2 Με δεδομένα από μηχανές αναζήτησης . . . 31
2.2.3 Με συνδυασμό δεδομένων από μέσα κοινωνικής δικτύωσης και μηχανές αναζήτησης . . . 31

3 Θεωρητικό υπόβαθρο **33**
3.1 Μηχανική μάθηση . . . 33
3.1.1 Επιβλεπόμενη μάθηση . . . 33
3.1.1.1 Ταξινόμηση . . . 34
3.1.1.2 Παλινδρόμηση . . . 34
3.1.2 Μη επιβλεπόμενη μάθηση . . . 34
3.1.3 Ημι-επιβλεπόμενη μάθηση . . . 35
3.1.4 Ενισχυτική μάθηση . . . 35
3.2 Παλινδρόμηση . . . 36
3.2.1 Γραμμική παλινδρόμηση με ελάχιστα τετράγωνα . . . 36
3.2.2 Πολυωνυμική παλινδρόμηση με ελάχιστα τετράγωνα . . . 37
3.2.3 Μη γραμμική παλινδρόμηση με ελάχιστα τετραγώνα . . . 38
3.2.4 Λογιστική παλινδρόμηση . . . 38
3.2.5 Βηματική παλινδρόμηση . . . 39
3.2.6 Παλινδρόμηση κορυφογραμμής . . . 39
3.2.7 Παλινδρόμηση LASSO . . . 40
3.2.8 Παλινδρόμηση elastic net . . . 40

3.2.9 Παλινδρόμηση διανυσμάτων υποστήριξης . . . 40
3.2.10 Παλινδρόμηση με πολυεπίπεδο perceptron . . . 43
3.2.11 Παλινδρόμηση με δέντρα αποφάσεων . . . 46
3.2.12 Παλινδρόμηση με τυχαία δάση . . . 47
3.2.13 Παλινδρόμηση με gradient boosting μηχανές . . . 48
3.2.14 Μπεϋζιανή παλινδρόμηση . . . 49
3.2.15 Παλινδρόμηση με γκαουσιανές διεργασίες . . . 50
3.3 Αξιολόγηση μοντέλων παλινδρόμησης . . . 51
3.3.1 Διαδικασίες αξιολόγησης . . . 51
3.3.1.1 Έλεγχος . . . 51
3.3.1.2 Επικύρωση . . . 51
3.3.1.3 Διασταυρούμενη επικύρωση . . . 52
3.3.2 Μετρικές αξιολόγησης . . . 52
3.3.2.1 Μέσο απόλυτο σφάλμα και μέσο απόλυτο ποσοστιαίο σφάλμα . 53
3.3.2.2 Μέσο τετραγωνικό σφάλμα και ριζικό μέσο τετραγωνικό σφάλμα 53
3.3.2.3 Συντελεστής προσδιορισμού . . . 53
3.3.2.4 Σκορ δικαιολογημένης διακύμανσης . . . 54
3.3.2.5 Ακρίβεια . . . 54
3.3.3 Μετρικές συσχέτισης . . . 54
3.3.3.1 Συντελεστής συσχέτισης Pearson . . . 54
3.3.3.2 Συντελεστής συσχέτισης Spearman . . . 55
3.4 Τεχνικές προεπεξεργασίας δεδομένων . . . 55
3.4.1 Μείωση διαστατικότητας . . . 55
3.4.1.1 Επιλογή χαρακτηριστικών . . . 56
3.4.1.2 Ανάλυση σε κύριες συνιστώσες . . . 56
3.4.2 Κανονικοποίηση χαρακτηριστικών . . . 56
3.4.2.1 Κλιμάκωση μεγίστου-ελαχίστου . . . 56
3.4.2.2 Τυποποίηση z-score . . . 57
3.4.3 Άλλα είδη προεπεξεργασίας δεδομένων . . . 57
3.5 Επεξεργασία φυσικής γλώσσας . . . 58
3.5.1 Ανάλυση συναισθήματος . . . 59

4 Τεχνικό υπόβαθρο **61**
4.1 Γλώσσα προγραμματισμού Python . . . 61
4.1.1 Βιβλιοθήκες που χρησιμοποιήθηκαν στα πειράματα . . . 61
4.1.1.1 pandas . . . 61
4.1.1.2 NumPy . . . 62
4.1.1.3 SciPy . . . 62
4.1.1.4 scikit-learn . . . 62
4.1.1.5 Matplotlib . . . 62
4.1.1.6 seaborn . . . 63
4.1.1.7 math . . . 63
4.1.1.8 spaCy . . . 63
4.1.1.9 datetime, dateutil, pytz, time, calendar . . . 63
4.1.1.10 Άλλες βιβλιοθήκες . . . 64
4.2 Εργαλεία συλλογής δεδομένων . . . 64
4.2.1 Twitter API . . . 64
4.2.2 Get Old Tweets script . . . 65
4.2.3 Google Trends API . . . 66
4.2.4 pytrends . . . 67

5 Προετοιμασία πειραμάτων **69**
5.1 Επισκόπηση δεδομένων αναφοράς . . . 69
5.2 Συλλογή δεδομένων . . . 70
5.2.1 Συλλογή δεδομένων από το Twitter . . . 71
5.2.2 Συλλογή δεδομένων από το Google Trends . . . 71
5.3 Εξαγωγή χαρακτηριστικών . . . 73
5.3.1 Εξαγωγή χαρακτηριστικών από το Twitter . . . 73
5.3.1.1 Χαρακτηριστικά ποσοτικών δεικτών . . . 74
5.3.1.2 Χαρακτηριστικά από ανάλυση συναισθήματος . . . 76
5.3.1.3 Καθορισμός χρονικών παραθύρων . . . 78
5.3.2 Εξαγωγή χαρακτηριστικών από το Google Trends . . . 78
5.3.2.1 Καθορισμός χρονικών παραθύρων . . . 78
5.3.3 Τελικό σύνολο δεδομένων και σχόλια . . . 79
5.4 Διερευνητική ανάλυση δεδομένων . . . 80
5.4.1 Συσχετίσεις μεταξύ χαρακτηριστικών και εξαρτημένων μεταβλητών . . . 80
5.4.2 Οπτικοποίηση δεδομένων . . . 86
5.5 Προεπεξεργασία χαρακτηριστικών και άλλες σχεδιαστικές επιλογές . . . 88
5.5.1 Σχεδιαστικές επιλογές για τα χρονικά παράθυρα . . . 88
5.5.2 Καθορισμός και επιλογή τελικών χαρακτηριστικών . . . 89
5.5.3 Τυποποίηση χαρακτηριστικών . . . 90
5.5.4 Διασταυρούμενη επικύρωση . . . 90
5.5.5 Διασωλήνωση . . . 91

6 Διεξαγωγή πειραμάτων και αξιολόγηση αποτελεσμάτων **93**
6.1 Μοντέλα αναφοράς . . . 94
6.2 Μοντέλα πολλαπλής γραμμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων . . . 96
6.3 Μοντέλα πολυωνυμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων 97
6.4 Μοντέλα παλινδρόμησης κορυφογραμμής . . . 99
6.5 Μοντέλα παλινδρόμησης LASSO . . . 101
6.6 Μοντέλα παλινδρόμησης elastic net . . . 103
6.7 Μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες . . . 104
6.8 Μοντέλα παλινδρόμησης με δέντρα απόφασης . . . 106
6.9 Μοντέλα παλινδρόμησης με τυχαία δάση . . . 108
6.10 Μοντέλα παλινδρόμησης με gradient boosting μηχανές . . . 110
6.11 Μοντέλα παλινδρόμησης με πολυεπίπεδα perceptron . . . 112
6.12 Μοντέλα παλινδρόμησης με διανύσματα υποστήριξης . . . 114
6.12.1 Με γραμμικό πυρήνα . . . 114
6.12.2 Με πυρήνα ακτινικής συνάρτησης βάσης . . . 116
6.12.3 Με πολυωνυμικό πυρήνα . . . 117
6.12.4 Με σιγμοειδή πυρήνα . . . 118
6.13 Σύγκριση μεταξύ διαφορετικών μοντέλων . . . 120
6.14 Σύγκριση αποτελεσμάτων με προηγούμενες εργασίες . . . 122

7 Επίλογος **125**
7.1 Σύνοψη και συμπεράσματα . . . 125
7.2 Μελλοντικές επεκτάσεις . . . 127

Βιβλιογραφία **129**

Κατάλογος Πινάκων

5.1 Μετρικές συσχέτισης χαρακτηριστικών ποσοτικών δεικτών με SHR% 82
5.2 Μετρικές συσχέτισης χαρακτηριστικών από ανάλυση συναισθήματος με SHR% . 83
5.3 Μετρικές συσχέτισης χαρακτηριστικών ποσοτικών δεικτών με AMR 84
5.4 Μετρικές συσχέτισης χαρακτηριστικών από ανάλυση συναισθήματος με AMR . 85

6.1 Αξιολόγηση του μοντέλου αναφοράς για την εκτίμηση του SHR% 94
6.2 Αξιολόγηση του μοντέλου αναφοράς για την εκτίμηση του AMR 95
6.3 Αξιολόγηση των μοντέλων πολλαπλής γραμμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του SHR% 96
6.4 Αξιολόγηση των μοντέλων πολλαπλής γραμμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του AMR 96
6.5 Αξιολόγηση των μοντέλων πολυωνυμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του SHR% 98
6.6 Αξιολόγηση των μοντέλων πολυωνυμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του AMR 98
6.7 Αξιολόγηση των μοντέλων παλινδρόμησης κορυφογραμμής για την εκτίμηση του SHR% 100
6.8 Αξιολόγηση των μοντέλων παλινδρόμησης κορυφογραμμής για την εκτίμηση του AMR 100
6.9 Αξιολόγηση των μοντέλων παλινδρόμησης LASSO για την εκτίμηση του SHR% 102
6.10 Αξιολόγηση των μοντέλων παλινδρόμησης LASSO για την εκτίμηση του AMR . 102
6.11 Αξιολόγηση των μοντέλων παλινδρόμησης elastic net για την εκτίμηση του SHR% 103
6.12 Αξιολόγηση των μοντέλων παλινδρόμησης elastic net για την εκτίμηση του AMR 103
6.13 Αξιολόγηση των μοντέλων παλινδρόμησης με γκαουσιανές διεργασίες για την εκτίμηση του SHR% 105
6.14 Αξιολόγηση των μοντέλων παλινδρόμησης με γκαουσιανές διεργασίες για την εκτίμηση του AMR 105
6.15 Αξιολόγηση των μοντέλων παλινδρόμησης με δέντρα απόφασης για την εκτίμηση του SHR% 107
6.16 Αξιολόγηση των μοντέλων παλινδρόμησης με δέντρα απόφασης για την εκτίμηση του AMR 107
6.17 Αξιολόγηση των μοντέλων παλινδρόμησης με τυχαία δάση για την εκτίμηση του SHR% 109
6.18 Αξιολόγηση των μοντέλων παλινδρόμησης με τυχαία δάση για την εκτίμηση του AMR 109
6.19 Αξιολόγηση των μοντέλων παλινδρόμησης με gradient boosting μηχανές για την εκτίμηση του SHR% 111
6.20 Αξιολόγηση των μοντέλων παλινδρόμησης με gradient boosting μηχανές για την εκτίμηση του AMR 111
6.21 Αξιολόγηση των μοντέλων παλινδρόμησης με πολυεπίπεδο perceptron για την εκτίμηση του SHR% 112

6.22 Αξιολόγηση των μοντέλων παλινδρόμησης με πολυεπίπεδο perceptron για την εκτίμηση του AMR 113
6.23 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα για την εκτίμηση του SHR% 114
6.24 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα για την εκτίμηση του AMR 115
6.25 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα για την εκτίμηση του SHR% 116
6.26 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα για την εκτίμηση του AMR 117
6.27 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα για την εκτίμηση του SHR% 118
6.28 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα για την εκτίμηση του AMR 118
6.29 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα για την εκτίμηση του SHR% 119
6.30 Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα για την εκτίμηση του AMR 119
6.31 Σύγκριση των μοντέλων που κατασκευάσαμε με τα προϋπάρχοντα όσον αφορά στη μετρική της ακρίβειας 123

Κατάλογος Σχημάτων

1.1 Παράδειγμα ενός tweet . . . 20
1.2 Μερίδιο στην παγκόσμια αγορά των πιο δημοφιλών μηχανών αναζήτησης στο διαδίκτυο από τον Ιανουάριο του 2010 έως τον Απρίλιο του 2020 . . . 21
1.3 Παράδειγμα αποτελεσμάτων από την πλατφόρμα Google Trends . . . 22
1.4 Εκτίμηση εσόδων από την παραδοσιακή τηλεοπτική βιομηχανία παγκοσμίως για την περίοδο 2018-2022 . . . 24

3.1 Παράδειγμα συμβιβασμού μεταξύ απόκλισης και διακύμανσης σε εργασία παλινδρόμησης (πηγή: towardsdatascience) . . . 34
3.2 Παράδειγμα γραμμικής και μη γραμμικής σχέσης των δεδομένων (πηγή: Laerd Statistics) . . . 38
3.3 Διαφορά γραμμικής και λογιστικής παλινδρόμησης (πηγή: DataCamp) . . . 39
3.4 Παράδειγμα γραμμικής παλινδρόμησης με μηχανή διανυσμάτων υποστήριξης (πηγή: Saed Sayad) . . . 41
3.5 Παράδειγμα μετασχηματισμού μη γραμμικής παλινδρόμησης σε γραμμική (πηγή: Saed Sayad) . . . 43
3.6 Παράδειγμα Multilayer Perceptron με 3 κρυφά επίπεδα (πηγή: [34]) . . . 44
3.7 Παράδειγμα δέντρου απόφασης σε πρόβλημα παλινδρόμησης με δύο χαρακτηριστικά (πηγή: SuperDataScience) . . . 46
3.8 Παράδειγμα παλινδρόμησης με δέντρα απόφασης διαφορετικού μέγιστου βάθους (πηγή: Data Science Stack Exchange) . . . 47
3.9 Παράδειγμα παλινδρόμησης με τυχαίο δάσος 600 δέντρων απόφασης - βάσης (πηγή: gitconnected) . . . 48
3.10 Παράδειγμα παλινδρόμησης με γκαουσιανή διεργασία (πηγή: scikit-learn) . . . 50
3.11 Διασταυρούμενη επικύρωση 5 τμημάτων (πηγή: github) . . . 52
3.12 Παράδειγμα ανάλυσης σε κύριες συνιστώσες (πηγή: TIBCO) . . . 57

5.1 Περιγραφικά στατιστικά στοιχεία για τα δεδομένα αναφοράς . . . 70
5.2 Ενδιαφέρον με την πάροδο του χρόνου για την εκπομπή . . . 72
5.3 Ενδιαφέρον με την πάροδο του χρόνου για την εκπομπή ανά εξάμηνο . . . 73
5.4 Όγκος των δημοσιεύσεων ανά ημέρα . . . 74
5.5 Πλήθος των διακριτών χρηστών που δημοσίευσαν ανά ημέρα . . . 74
5.6 Όγκος των αναδημοσιεύσεων ανά ημέρα . . . 75
5.7 Όγκος των επισημάνσεων «Μου αρέσει» ανά ημέρα . . . 75
5.8 Κανονικοποιημένα αποτελέσματα από το Google Trends ανά ημέρα . . . 79
5.9 Οπτικοποίηση όλων των χαρακτηριστικών από ποσοτικούς δείκτες . . . 87
5.10 Οπτικοποίηση όλων των χαρακτηριστικών από ανάλυση συναισθήματος . . . 88

6.1 Πραγματικές και προβλεπόμενες τιμές από τα μοντέλα αναφοράς . . . 95
6.2 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα πολλαπλής γραμμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων . . . 97

6.3 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα πολυωνυμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων 99
6.4 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης κορυφογραμμής 101
6.5 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης LASSO 102
6.6 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης elastic net 104
6.7 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες 106
6.8 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με δέντρα απόφασης 108
6.9 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με τυχαία δάση 109
6.10 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με gradient boosting μηχανές 111
6.11 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με πολυεπίπεδο perceptron 113
6.12 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα 115
6.13 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα 117
6.14 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα 119
6.15 Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα 120
6.16 Σύγκριση διαφορετικών μοντέλων για την εκτίμηση του SHR% ως προς το MAPE 121
6.17 Σύγκριση διαφορετικών μοντέλων για την εκτίμηση του AMR ως προς το MAPE 121

Κεφάλαιο 1

Εισαγωγή

1.1 Μέσα κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης (social media) έχουν πλέον εδραιωθεί στην καθημερινότητα δισεκατομμυρίων ανθρώπων, κερδίζοντας ολοένα και περισσότερους χρήστες. Σύμφωνα με δημοσιευμένη στατιστική μελέτη υπολογίζεται ότι πλέον 3.96 δισεκατομμύρια άνθρωποι παγκοσμίως χρησιμοποιούν μέσα κοινωνικής δικτύωσης ενώ όσο περνάει ο χρόνος αναμένεται να αυξάνεται το πλήθος τους [16].

Αυτή η μεγάλη δημοτικότητά τους οφείλεται στις πολλαπλές δυνατότητες ψυχαγωγίας, ενημέρωσης και κοινωνικοποίησης που προσφέρουν. Διευκολύνουν το διαμοιρασμό πληροφοριών, ιδεών, ειδήσεων και άλλων μορφών περιεχομένου μέσω εικονικών κοινοτήτων και δικτύων. Ενδεικτικά, στα περισσότερα μέσα κοινωνικής δικτύωσης ένας χρήστης μπορεί να διατηρεί ένα προσωπικό προφίλ και μέσω αυτού να επικοινωνεί άμεσα και να ανταλλάσσει πολυμεσικό περιεχόμενο με άλλους χρήστες, να αναρτά προσωπικές του απόψεις ή να κοινοποιεί δημοσιεύσεις τρίτων που τον εκφράζουν και να παρακολουθεί δημοσιευμένο περιεχόμενο που εμπίπτει στα ενδιαφέροντά του.

Τα μέσα κοινωνικής δικτύωσης ποικίλλουν ως προς τη μορφή και το σκοπό τους και χωρίζονται σε διάφορες κατηγορίες, μερικές από τις οποίες παραθέτουμε παρακάτω. Καταρχάς, οι ιστοσελίδες κοινωνικής δικτύωσης (social networking sites), όπως το Facebook και το Linkedin. Σε αυτές τις πλατφόρμες ο χρήστης κυρίως διαχειρίζεται και αναπτύσσει τον κοινωνικό του κύκλο, αλληλεπιδρώντας με όσους τον απαρτίζουν. Δύο άλλες κατηγορίες είναι τα blogs και τα microblogs, όπως το Twitter. Σε αυτές τις πλατφόρμες δημοσιεύεται κυρίως κειμενικό αλλά και γενικότερα πολυμεσικό περιεχόμενο γύρω από ένα συγκεκριμένο θέμα ενδιαφέροντος κάθε φορά και ενθαρρύνεται η συζήτηση γύρω από αυτό. Άλλες σημαντικές κατηγορίες είναι οι ιστότοποι διαμοιρασμού βίντεο, εικόνων και ηχητικών κλιπ (media sharing sites), όπως το YouTube, το Instagram και το TikTok και οι υπηρεσίες τηλεφωνίας μέσω διαδικτύου και ανταλλαγής μηνυμάτων (VoIP and messaging services) όπως το WhatsApp, το Facebook Messenger και το Viber μεταξύ άλλων. [2, 57]

Πέρα από τα άμεσα οφέλη που προσφέρουν στους τελικούς χρήστες, τα μέσα κοινωνικής δικτύωσης αποτελούν πηγή ενός τεράστιου όγκου δεδομένων. Αυτά μπορούν να χρησιμοποιηθούν σε διάφορες έρευνες για την ανάδειξη κοινωνικών φαινομένων και συμπεριφορών, σε οικονομικές μελέτες για εμπορικούς σκοπούς αλλά και προς εξυπηρέτηση πολιτικών σκοπιμοτήτων. Ενδεικτικά παραδείγματα αποτελούν η πρόβλεψη αποτελεσμάτων πολιτικών εκλογών, οι εξατομικευμένες διαφημίσεις, η διαμόρφωση της κοινής γνώμης και η ανίχνευση και η διαχείριση κρίσεων μεταξύ άλλων. Μάλιστα, δεν είναι λίγες οι φορές που τα τελευταία χρόνια έγινε κατάχρηση αυτών των δεδομένων, ακόμα και προσωπικών στοιχείων χωρίς τη γνώση και τη συγκατάθεση των φυσικών προσώπων, καταδεικνύοντας την ανάγκη για θεσμοθέτηση νόμων με σκοπό την προστασία των προσωπικών δεδομένων[1].

[1] `https://gdpr-info.eu/`

Χωρίς την καταστρατήγηση αυτών των νόμων, οι πληροφορίες που μοιράζονται εν γνώσει τους οι χρήστες με δημόσιες αναρτήσεις είναι υπεραρκετές για την εξαγωγή χρήσιμων στατιστικών συμπερασμάτων σε πολλές εκφάνσεις του κοινωνικού και επαγγελματικού βίου. Στην παρούσα διπλωματική εργασία, στοχεύουμε να εκμεταλλευτούμε αυτήν την πλούσια πηγή γνώσης προς όφελος όσων δραστηριοποιούνται στη βιομηχανία της τηλεόρασης και όσων συνεργάζονται ή γενικότερα σχετίζονται με αυτήν.

1.1.1 Twitter

Το Twitter[2] αποτελεί το πιο δημοφιλές μέσο κοινωνικής δικτύωσης στην κατηγορία του microblogging. Ιδρύθηκε το Μάρτιο του 2006 από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams και κυκλοφόρησε για πρώτη φορά τον Ιούλιο του ίδιου χρόνου. Πλέον έχει 326 εκατομμύρια ενεργούς χρήστες το μήνα παγκοσμίως και κατατάσσεται στα 15 πιο δημοφιλή μέσα κοινωνικά δικτύωσης ανεξαρτήτως κατηγορίας [15].

Ένα micro-blog διαφέρει από ένα παραδοσιακό blog όσον αφορά στο μέγεθος των δημοσιεύσεων που φιλοξενεί. Όπως άλλωστε προδιαθέτει το όνομά του, κάθε ανάρτηση έχει πολύ μικρότερο περιεχόμενο σε ένα micro-blog από ό,τι σε ένα κλασικό blog. Το Twitter δεν αποτελεί εξαίρεση σε αυτόν τον κανόνα και κάθε δημοσίευση (tweet) στην πλατφόρμα οφείλει να τηρεί τον περιορισμό των 280 χαρακτήρων. Οι μη συνδεδεμένοι χρήστες μπορούν να έχουν πρόσβαση στα αναρτημένα tweets με δικαίωμα ανάγνωσης μόνο. Οι συνδεδεμένοι χρήστες πέρα από την ανάγνωση έχουν το δικαίωμα να δημοσιεύουν περιεχόμενο στην πλατφόρμα, να κάνουν like σε tweets της αρεσκείας τους, να αναδημοσιεύουν (retweet) αναρτήσεις, να απαντούν (reply) σε tweets άλλων χρηστών και να ανταλλάσσουν απευθείας ιδιωτικά μηνύματα μεταξύ τους (direct messages). Επίσης, μπορούν να ακολουθούν (follow) όσους χρήστες επιθυμούν με σκοπό να ενημερώνονται για τα tweets που αυτοί δημοσιεύουν και αντίστοιχα να ακολουθούνται από άλλους χρήστες (followers).

Σχήμα 1.1: Παράδειγμα ενός tweet

Το περιεχόμενο ενός tweet μπορεί να είναι απλά κειμενικό αλλά και να περιλαμβάνει φωτογραφίες, βίντεο και συνδέσμους. Επίσης, συνήθως περιέχει τουλάχιστον ένα hashtag, δηλαδή το σύμβολο # ακολουθούμενο από μια λέξη ή φράση χωρίς κενά, το οποίο χρησιμεύει για την ομαδοποίηση των tweets και την κατάταξή τους σε θεματικές ενότητες. Ακόμα, ένα tweet είναι δυνατό να περιέχει mentions, δηλαδή ονόματα χρηστών έπειτα από το σύμβολο @, με σκοπό την αναφορά σε άλλους χρήστες. Όλα αυτά τα στοιχεία ενός tweet φαίνονται στο σχήμα 1.1. Επιπλέον, μπορούμε να διακρίνουμε αρκετά μεταδεδομένα, όπως το μοναδικό όνομα χρήστη αλλά

[2] https://twitter.com/

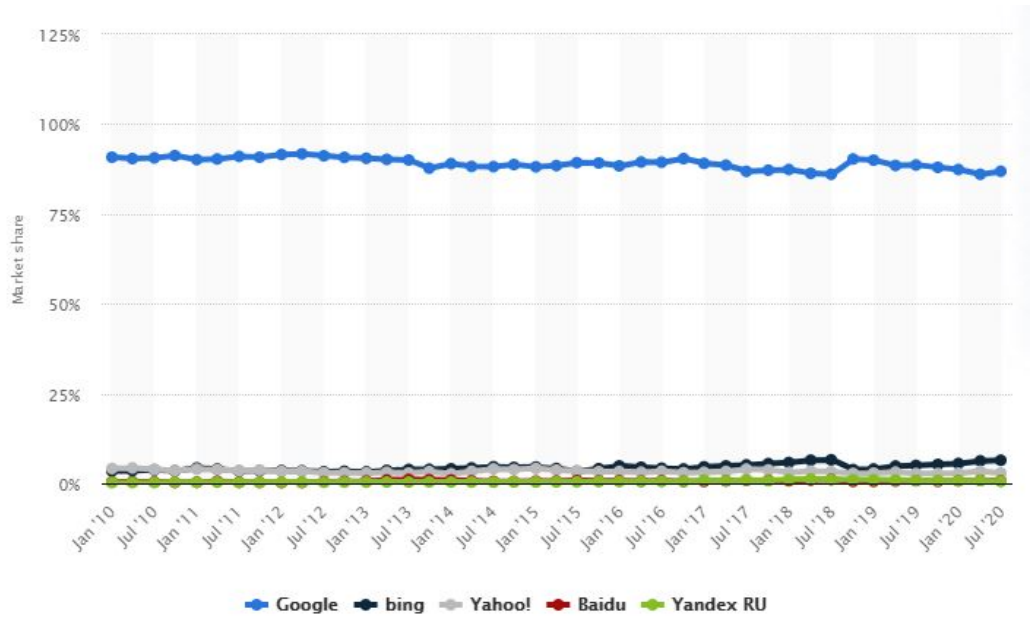
και το ονοματεπώνυμο εκείνου που το ανήρτησε, την ημερομηνία και την ώρα που δημοσιεύτηκε, το γεωγραφικό στίγμα του χρήστη τη στιγμή της δημοσίευσης (αν υπάρχει), το πλήθος των retweets και των favorites. Τέλος, όπως είναι λογικό, είναι ορατό το permalink, δηλαδή η διεύθυνση που βρίσκεται διαθέσιμο το tweet και κατ΄ επέκταση το αναγνωριστικό του (id), το οποίο φαίνεται στο τέλος του μόνιμου συνδέσμου.

Σε κάθε περίπτωση, οι αναρτήσεις στο Twitter σχετίζονται με επίκαιρα γεγονότα, με κοινωνικοπολιτικά τεκταινόμενα και ψυχαγωγικές εκδηλώσεις και περιστρέφονται γύρω από ειδήσεις ευρείας θεματολογίας και τις προσωπικές απόψεις των χρηστών πάνω σε αυτές. Ο προσανατολισμός του είναι σαφής και διαφέρει από αυτόν άλλων δημοφιλών μέσων κοινωνικής δικτύωσης, όπως π.χ. το Facebook και το Instagram, όπου οι χρήστες συνηθίζουν να αναρτούν φωτογραφίες και βίντεο από την προσωπική τους ζωή, κοινοποιούν την παρουσία τους σε διάφορα μέρη κλπ. Γίνεται, λοιπόν, φανερό ότι το Twitter αποτελεί μια πλούσια πηγή δεδομένων από την οποία μπορεί να εξαχθεί χρήσιμη γνώση σχετικά με τα ενδιαφέροντα και τις απόψεις των χρηστών επί παντός επιστητού.

1.2 Μηχανές αναζήτησης

Οι μηχανές αναζήτησης στο διαδίκτυο (web search engines ή Internet search engines) είναι συστήματα λογισμικού σχεδιασμένα για αναζήτηση με συστηματικό τρόπο στον Παγκόσμιο Ιστό (World Wide Web). Μια τέτοια αναζήτηση πραγματοποιείται με βάση κάποιο ερώτημα (query) που υποβάλλεται σε κειμενική μορφή. Τα αποτελέσματα παρατίθενται σε φθίνουσα σειρά σχετικότητας σύμφωνα με τους αλγορίθμους βελτιστοποίησης της εκάστοτε μηχανής αναζήτησης (search engine optimization) και εμφανίζονται με τη μορφή ενός καταλόγου ιστοσελίδων, εικόνων, βίντεο, άρθρων και άλλων τύπων αρχείων [81].

Υπάρχουν διάφορες διαθέσιμες μηχανές αναζήτησης, οι δημοφιλέστερες από τις οποίες είναι οι εξής: Google, bing, Yahoo!, Baidu, Yandex RU και DuckDuckGo. Με συντριπτική διαφορά η μηχανή αναζήτησης της Google κατέχει το μεγαλύτερο μερίδιο της αγοράς παγκοσμίως τα τελευταία 10 χρόνια, όπως φαίνεται και στο σχήμα 1.2 [17].



Σχήμα 1.2: Μερίδιο στην παγκόσμια αγορά των πιο δημοφιλών μηχανών αναζήτησης στο διαδίκτυο από τον Ιανουάριο του 2010 έως τον Απρίλιο του 2020

Σε μια μηχανή αναζήτησης μπορεί κανείς να βρει πληροφορίες για ένα σύγχρονο ή ιστορικό γεγονός πάνω σε οποιοδήποτε θέμα. Επίσης, είναι λογικό πως στις αναζητήσεις των χρηστών

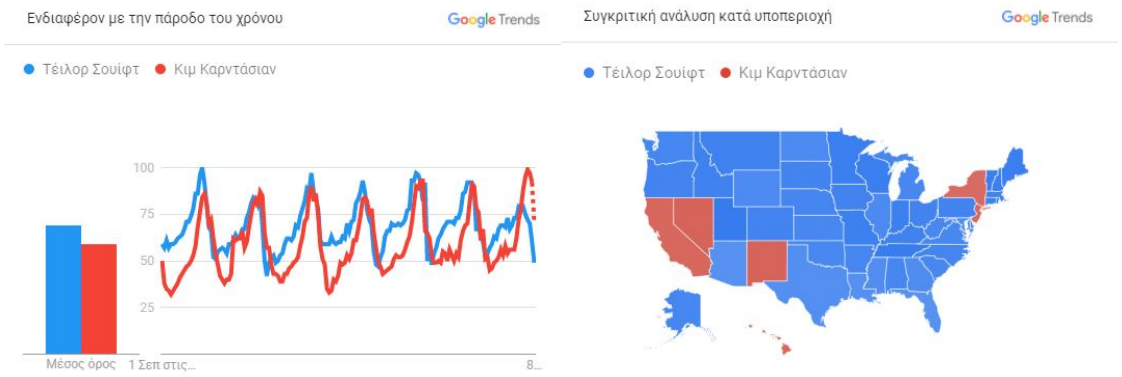
στον Παγκόσμιο Ιστό αντανακλώνται τα ενδιαφέροντα τους και οι προτιμήσεις τους, οπότε η γνώση των στατιστικών στοιχείων των αναζητήσεων σε μια μηχανή αναζήτησης γύρω από ένα θέμα μπορεί να αποδειχθεί ιδιαίτερα διαφωτιστική για τη δημοφιλία του συγκεκριμένου θέματος.

1.2.1 Google Trends

Η πλατφόρμα Google Trends[3] παρέχει πρόσβαση σε ένα δείγμα πραγματικών αιτημάτων που υποβάλλονται στη μηχανή αναζήτησης της Google. Κυκλοφόρησε για πρώτη φορά το Μάιο του 2006 και συγχωνεύτηκε με την Google Insights for Search το Σεπτέμβριο του 2012 [48]. Όπως φαίνεται από το σχήμα 1.2, η μηχανή αναζήτησης της Google συγκεντρώνει με μεγάλη διαφορά περισσότερους χρήστες σε σχέση με τις άλλες μηχανές αναζήτησης στο διαδίκτυο [17], οπότε και στην παρούσα εργασία θα ασχοληθούμε μόνο με αυτή.

Ένας επισκέπτης στην ιστοσελίδα Google Trends έχει τη δυνατότητα να πληροφορηθεί για τη δημοφιλία όρων αναζήτησης ορίζοντας ένα χρονικό διάστημα της επιλογής του και συγκεκριμένο γεωγραφικό περιορισμό σε επίπεδο χωρών. Μπορεί να μελετηθεί η δημοφιλία ενός μόνο όρου αναζήτησης αλλά και η σχετική δημοφιλία περισσότερων συγκρινόμενων μεταξύ τους στο συγκεκριμένο χρονικό διάστημα και τοπικό επίπεδο. Πέρα από απλούς όρους αναζήτησης που ορίζονται από σύνολα λέξεων μπορούν να χρησιμοποιηθούν και θέματα αναζήτησης. Η διαφορά έγκειται στο ότι με τους όρους αναζήτησης στα αποτελέσματα περιλαμβάνονται αναζητήσεις που περιέχουν τις λέξεις που ορίστηκαν σε αυτούς ενώ με τα θέματα αναζήτησης στα αποτελέσματα περιλαμβάνονται αναζητήσεις με βάση όρους σημασιολογικά όμοιους με αυτά και μεταξύ τους αλλά πιθανόν αρκετά διαφορετικούς λεξιλογικά. Τέλος, μπορεί να επιλεγεί κάποια κατηγορία στην οποία είναι επιθυμητό να εμπίπτουν οι αναζητήσεις σχετικά με αυτόν τον όρο αλλά και να οριστεί ο χώρος αναζήτησης (αναζήτηση στον Ιστό, αναζήτηση εικόνων, αναζήτηση ειδήσεων, αγορές Google και αναζήτηση YouTube).

Υπάρχουν δύο μορφές με τις οποίες επιστρέφονται τα αποτελέσματα. Η πρώτη είναι οι χρονοσειρές, όπου μπορούν να φανούν οι διακυμάνσεις στη σχετική δημοφιλία των όρων συναρτήσει του χρόνου. Αν το χρονικό διάστημα μελέτης είναι η τελευταία ώρα τα αποτελέσματα δίνονται ανά λεπτό, αν είναι η τελευταία εβδομάδα ανά ώρα ενώ όταν πρόκειται για ιστορικά δεδομένα χρονικού διαστήματος έως και 270 ημερών δίνονται σε ημερήσια βάση. Για περιόδους μεγαλύτερης διάρκειας των 9 μηνών και μικρότερης των 5 ετών επιστρέφονται αποτελέσματα σε εβδομαδιαία βάση ενώ για ακόμα μεγαλύτερης διάρκειας σε μηνιαία. Η δεύτερη μορφή με την οποία επιστρέφονται τα αποτελέσματα είναι αυτή των χρωματισμένων χαρτών. Σε αυτήν την περίπτωση, φαίνονται οι διακυμάνσεις της σχετικής δημοφιλίας των όρων στις διάφορες γεωγραφικές υποενότητες της ευρύτερης περιοχής που ορίστηκε αρχικά.



(α΄) με χρονοσειρά (β΄) με χρωματισμένο χάρτη

Σχήμα 1.3: Παράδειγμα αποτελεσμάτων από την πλατφόρμα Google Trends

[3] https://trends.google.com/trends/?geo

Αξίζει να σημειωθεί ότι η επιστροφή του απόλυτου πλήθους των αναζητήσεων θα ήταν χρονικά ασύμφορη, καθώς ο όγκος των δεδομένων είναι υπερβολικά μεγάλος. Ενδεικτικά, αναφέρουμε ότι το 2019 κατά μέσο όρο πραγματοποιούνταν 3.5 δισεκατομμύρια αναζητήσεις τη μέρα ενώ το 2020 αυτός ο αριθμός είναι πιθανό να υπερβεί τα 7 δισεκατομμύρια [30]. Για αυτόν το λόγο, επιστρέφονται δειγματοληπτημένα αποτελέσματα ώστε να έχουμε γρήγορα μια αντιπροσωπευτική εικόνα των αναζητήσεων. Επιπλέον, τα αποτελέσματα είναι κανονικοποιημένα ως προς το συνολικό πλήθος των αναζητήσεων στην υπό εξέταση χρονική περίοδο και περιοχή. Τελικά, οι τιμές που επιστρέφονται βρίσκονται στο εύρος 0-100, με το 100 να αντιστοιχεί στο μέγιστο και το 0 στο ελάχιστο πλήθος αναζητήσεων πάντα στον συγκεκριμένο χωροχρονικό περιορισμό που έχει τεθεί. Αυτό το σημείο χρήζει ιδιαίτερης προσοχής, καθώς καθιστά αδύνατη τη σύγκριση τιμών σε διαφορετικές χρονικές περιόδους και χώρες. Τέλος, το Google Trends προσφέρεται για παροχή στατιστικών στοιχείων μόνο για δημοφιλείς όρους. Όλοι οι όροι που συγκεντρώνουν μικρό πλήθος αναζητήσεων αντιστοιχούνται στο 0 και δεν μπορούμε να εξάγουμε πληροφορίες για τη δημοφιλία τους σε διάφορες χρονικές στιγμές και υποπεριοχές. [24]

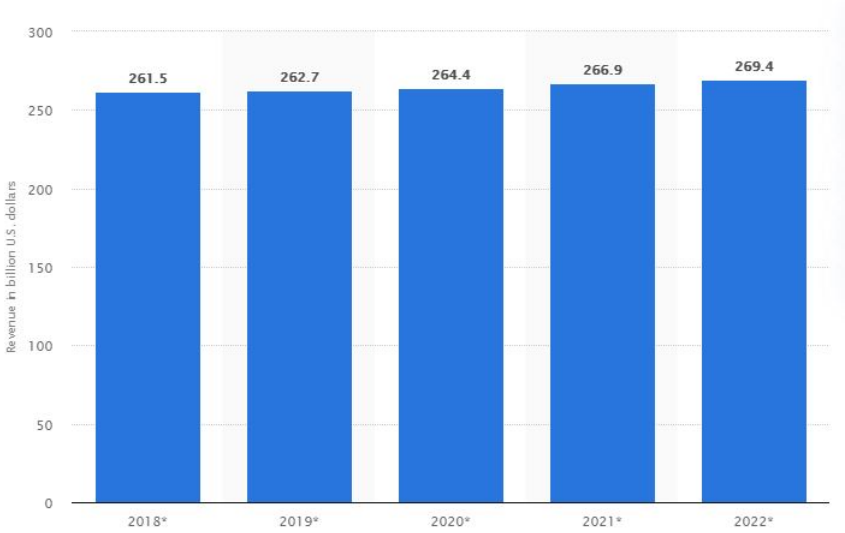
Συνοψίζοντας, οι περιορισμοί που εισάγονται λόγω δειγματοληψίας, πιθανού φιλτραρίσματος των αναζητήσεων και πιο αραιών χρονικά ιστορικών δεδομένων σε μεγαλύτερες περιόδους δεν είναι ικανοί να επισκιάσουν τις δυνατότητες που προσφέρει η πλατφόρμα. Ακόμα και αν τα δεδομένα δεν είναι απολύτως ακριβή, δεν παύει να πρόκειται για μια πολύτιμη πηγή γνώσης που μπορεί να αποκαλύψει ενδιαφέρουσες πτυχές στις προτιμήσεις των χρηστών και κατ΄ επέκταση στην κοινωνία και την οικονομία.

1.3 Τηλεοπτική βιομηχανία και σημασία μετρήσεων τηλεθέασης

Η βιομηχανία της τηλεόρασης αποτελεί μια μεγάλη αγορά με δισεκατομμύρια δολάρια ως ετήσια κέρδη. Αν και έχουν περάσει κάποιες δεκαετίες από τότε που πρωτοεμφανίστηκε, ακόμα το ενδιαφέρον των χρηστών παραμένει αμείωτο για αυτήν. Φαινομενικά, μπορεί να ισχυριστεί κανείς ότι η επίδρασή της έχει αρχίσει να φθίνει τα τελευταία χρόνια, καθώς όλο και περισσότεροι άνθρωποι αναζητούν εναλλακτικές μορφές και υπηρεσίες για την παρακολούθηση εκπομπών, σειρών και ταινιών. Αυτό όμως δεν ισχύει, αφού στην πραγματικότητα δεν πρόκειται για πτώση του ενδιαφέροντος αλλά για μεταβατική φάση προς μία νέα εποχή παρακολούθησης τηλεοπτικών προγραμμάτων με τη βοήθεια της διαρκούς τεχνολογικής εξέλιξης.

Πιο αναλυτικά, οι νέες δυνατότητες που παρέχονται αφορούν αφενός τον τρόπο μετάδοσης, π.χ. καλωδιακή τηλεόραση (cable tv), δορυφορική τηλεόραση (sattelite tv), ψηφιακή επίγεια τηλεόραση (DTT), τηλεόραση μέσω πρωτοκόλλου διαδικτύου (IPTV)) και αφετέρου τις τελικές υπηρεσίες που παρέχονται στον τηλεθεατή, π.χ. over-the-top (OTT), video-on-demand (VoD), Digital Video Recorder - DVR viewing. Όλες οι τελευταίες σηματοδοτούν το τέλος της παθητικής θέασης και την αρχή μιας νέας εποχής, όπου ο τηλεθεατής μπορεί να επιλέξει τι επιθυμεί να δει και πότε, ενισχύοντας την αλληλεπίδρασή του με την τηλεόραση. Όπως γίνεται φανερό λοιπόν, όλες αυτές οι νέες δυνατότητες εντάσσονται στην τηλεοπτική βιομηχανία και εκτιμάται πως η εισαγωγή τους θα οδηγήσει σε ακόμα μεγαλύτερη ανάπτυξη της τελευταίας [93, 11]. Μάλιστα, σύμφωνα με μελέτη που διεξήχθη το 2018, εκτιμάται ότι τα έσοδα από τη βιομηχανία μόνο της παραδοσιακής τηλεόρασης τα επόμενα χρόνια θα γνωρίζουν σταδιακή μικρή αύξηση, ξεπερνώντας τα 269 δισεκατομμύρια δολάρια παγκοσμίως το 2022 [94].

Ο μεγάλος τζίρος στη βιομηχανία της τηλεόρασης οφείλεται ως επί το πλείστον πρώτα στις διαφημίσεις και έπειτα σε συνδρομητικές υπηρεσίες [11]. Και στις δύο περιπτώσεις, είναι προφανές ότι όσο υψηλότερα νούμερα τηλεθέασης συγκεντρώνει ένα πρόγραμμα, δηλαδή όσο μεγαλύτερη απήχηση έχει στο κοινό, τόσο πιο πολλούς διαφημιστές και χορηγούς θα προσελκύει αλλά και τόσο περισσότερους συνδρομητές, με αποτέλεσμα την αύξηση των εσόδων. Συνεπώς, η γνώση της τηλεθέασης των προγραμμάτων, και δη η πρότερη, είναι βαρύνουσας σημασίας για μια εταιρία παραγωγής, έναν τηλεοπτικό σταθμό, ένα συνδρομητικό κανάλι και οποιονδήποτε δραστηριο-



Σχήμα 1.4: Εκτίμηση εσόδων από την παραδοσιακή τηλεοπτική βιομηχανία παγκοσμίως για την περίοδο 2018-2022

ποιείται σε αυτόν το χώρο. Αυτά τα δεδομένα δύνανται να υποδείξουν μια πιθανή αλλαγή στο πρόγραμμα ενός σταθμού ή και μιας μεμονωμένης εκπομπής, η οποία θα αποβεί κερδοφόρα στη συνέχεια. Πέρα από την προφανή χρησιμότητα στους τηλεοπτικούς σταθμούς και οι διαφημιστικές εταιρίες αλλά και γενικότερα οι εταιρίες που σκοπεύουν να προάγουν ένα προϊόν στην αγορά, με βάση αυτά μπορούν να κρίνουν πού και πότε είναι η κατάλληλη στιγμή να το πράξουν ώστε να απευθυνθούν σε μεγαλύτερο μέρος του καταναλωτικού κοινού που τους ενδιαφέρει. Σε κάθε περίπτωση, τα νούμερα τηλεθέασης αποτελούν βασικούς και αξιόπιστους δείκτες για την κερδοφορία γύρω από ένα τηλεοπτικό πρόγραμμα.

Σε μία εποχή που οι σύγχρονες τεχνολογίες προσφέρουν ολοένα και περισσότερες δυνατότητες θέασης, οι παραδοσιακοί τρόποι μετρήσεων της τηλεθέασης (π.χ. με ειδικές εγκατεστημένες συσκευές) αρχίζουν να αδυνατούν να ανιχνεύουν όλο το φάσμα των διαφορετικών ειδών τηλεθεατών. Είναι αδήριτη η ανάγκη, λοιπόν, να ανακαλυφθεί ένας νέος τρόπος εκτίμησής της, σε αυτό το μεταβαλλόμενο πεδίο, όχι για να αντικαταστήσει τους ήδη υπάρχοντες αλλά για να τους συμπληρώσει και να τους υποβοηθήσει. Αυτό μπορεί να επιτευχθεί συλλέγοντας δεδομένα από μέσα κοινωνικής δικτύωσης και από μηχανές αναζήτησης στο διαδίκτυο, τα οποία, όπως θα δούμε και αναλυτικότερα στη συνέχεια, έχουν υψηλή προβλεπτική ικανότητα και σε αυτόν τον τομέα ανάμεσα σε πολλούς άλλους.

1.4 Κίνητρο και συμβολή εργασίας

Όπως γίνεται φανερό από τα παραπάνω, η εκτίμηση της τηλεθέασης προγραμμάτων με μεθόδους πέρα από τις κλασικές που χρησιμοποιούν εγκατεστημένα μηχανήματα σε δειγματοληπτημένα νοικοκυριά, είναι ένα ανοιχτό πεδίο έρευνας με μεγάλα οικονομικά οφέλη για την τηλεοπτική βιομηχανία και τους κλάδους που σχετίζονται άμεσα με αυτήν. Λύση σε αυτό το πρόβλημα μπορεί να δώσει η αξιοποίηση του μεγάλου όγκου δεδομένων που προκύπτει από τη χρήση μέσων κοινωνικής δικτύωσης και μηχανών αναζήτησης στο διαδίκτυο. Στην παρούσα διπλωματική εργασία, θα ασχοληθούμε με αυτό το πρόβλημα και με τη βοήθεια αλγορίθμων Μηχανικής Μάθησης θα εκμεταλλευτούμε αυτά τα δεδομένα για την υλοποίηση μοντέλων παλινδρόμησης, τα οποία θα είναι σε θέση να κάνουν έγκυρες προβλέψεις για την τηλεθέαση στο μέλλον. Όπως θα δούμε και στη συνέχεια, έχουν ήδη διεξαχθεί αρκετές έρευνες για την επίτευξη αυτού του σκοπού, εκ των οποίων οι περισσότερες βασίζονται μόνο σε δεδομένα που έχουν συλλεχθεί από το Twitter και λιγότερες μόνο σε δεδομένα που έχουν συλλεχθεί από το Facebook. Τα στατιστικά δεδο-

μένα από τη χρήση μηχανών αναζήτησης, όπως αυτή της Google συνήθως χρησιμοποιούνται για την αντιμετώπιση του παρεμφερούς προβλήματος της εκτίμησης της απήχησης κινηματογραφικών ταινιών τόσο από άποψη εσόδων όσο και άποψη κριτικών και αξιολογήσεων. Από όσο γνωρίζουμε, δεν έχει γίνει κάποια απόπειρα συνδυασμού των δεδομένων από τις πλατφόρμες Twitter και Google Trends για την εκτίμηση της τηλεθέασης πέρα από την αναφορά στη μελέτη [38], που συνοδεύτηκε από τη διπλωματική εργασία του Μάριου Παρασκευόπουλου [61], στην οποία θα στηριχθούμε για να προβούμε σε προεκτάσεις.

Στην παρούσα διπλωματική εργασία, θα επιχειρήσουμε να υποστηρίξουμε την υπόθεση ότι ο συνυπολογισμός των στατιστικών δεδομένων από τη μηχανή αναζήτησης της Google μπορεί να αποτελέσει καταλύτη για τη βελτίωση της απόδοσης ήδη προτεινόμενων μοντέλων που χρησιμοποιούν μόνο δεδομένα από το Twitter. Αυτό θα γίνει με ένα μεγαλύτερο σύνολο δεδομένων από αυτό που χρησιμοποιήθηκε στην προηγούμενη προσπάθεια ούτως ώστε να διασφαλίσουμε τη στιβαρότητα των μοντέλων μας και να αυξήσουμε την πιθανότητα γενίκευσης των αποτελεσμάτων μας. Πιο συγκεκριμένα, ως παρατηρήσεις θα χρησιμοποιηθούν όλα τα επεισόδια από τα έτη 2016-2017 του ιταλικού τηλεοπτικού προγράμματος Le Iene και θα εξαχθούν διάφορα χαρακτηριστικά που αφορούν ποσοτικούς δείκτες ή προκύπτουν από ανάλυση συναισθήματος στο περιεχόμενο των tweets. Όλα θα οριστούν σε διάφορα χρονικά παράθυρα γύρω από την προβολή του κάθε επεισοδίου, όπως έχει ήδη προταθεί από συγγενείς εργασίες. Πέρα από την καινοτομία που εισάγεται με την προσθήκη των δεδομένων από την πλατφόρμα Google Trends, σημαντικό είναι και το γεγονός ότι θα εξεταστούν πολλοί αλγόριθμοι Μηχανικής Μάθησης και αντίστοιχα θα κατασκευαστεί μια πληθώρα μοντέλων παλινδρόμησης. Από όσο γνωρίζουμε είναι η πρώτη φορά που θα δοκιμαστούν τόσες πολλές μέθοδοι πάνω στο ίδιο σύνολο δεδομένων για την αντιμετώπιση του συγκεκριμένου προβλήματος, καθιστώντας δυνατή στη συνέχεια τη σύγκριση της ποιότητας των προβλέψεων που παράγουν όλα τα μοντέλα για να αποφανθούμε ποιο υπερτερεί. Αξίζει να σημειώσουμε ότι η μελέτη μας περιλαμβάνει την εκτίμηση τόσο του μεριδίου τηλεθέασης, δηλαδή του ποσοστού τηλεθέασης επί των ανοιχτών δεκτών όσο και του απόλυτου πλήθους τηλεθεατών που παρακολούθησαν ένα επεισόδιο έστω και για ένα λεπτό. Τέλος, θα συγκρίνουμε ένα υποσύνολο των αποτελεσμάτων με αυτά προηγούμενης διπλωματικής εργασίας [65], που βασίστηκε στο ίδιο σύνολο δεδομένων αναφοράς, για να δούμε αν πράγματι τελικά η εξαγωγή περισσότερων χαρακτηριστικών τόσο από την ίδια πηγή (Twitter) όσο και από την πλατφόρμα Google Trends σε περισσότερα χρονικά παράθυρα και η προεπεξεργασία αυτών οδήγησε σε αύξηση της απόδοσης των εκτιμητών.

1.5 Διάρθρωση εργασίας

Στο συγκεκριμένο εισαγωγικό κεφάλαιο αναπτύχθηκε το αντικείμενο της παρούσας εργασίας, το κίνητρο που μας οδήγησε στη διεξαγωγή της, οι συνθήκες που καθιστούν εφικτή την τελευταία αλλά και η σημασία των αποτελεσμάτων που θα προκύψουν. Πέρα από αυτό, στη συνέχεια ακολουθούν άλλα 6 κεφάλαια, τα οποία δομούνται ως εξής:

Στο κεφάλαιο 2, γίνεται μια εκτενής αναφορά στη συναφή βιβλιογραφία, που μελετήθηκε. Πρώτα, δίνεται έμφαση γενικά στην προβλεπτική ικανότητα των δεδομένων που συλλέγονται από μέσα κοινωνικής δικτύωσης και μηχανές αναζήτησης στο διαδίκτυο, μέσα από έρευνες για την αντιμετώπιση ποικίλων προβλημάτων και την εξαγωγή εκτιμήσεων σε διάφορους τομείς του επιστητού. Έπειτα, επικεντρωνόμαστε σε σχετικές μελέτες, οι οποίες πραγματεύονται το συγκεκριμένο πρόβλημα, που καλούμαστε να αντιμετωπίσουμε στην παρούσα εργασία, δηλαδή την εκτίμηση της τηλεθέασης εκπομπών αλλά και το παρεμφερές πρόβλημα της εκτίμησης της απήχησης κινηματογραφικών ταινιών.

Στο κεφάλαιο 3, αναπτύσσεται το θεωρητικό υπόβαθρο της εργασίας. Αρχικά, αναλύεται γενικότερα ο κλάδος της Μηχανικής Μάθησης, οι υποκατηγορίες του και οι εφαρμογές του. Στη συνέχεια, παρατίθενται διάφορες τεχνικές και αλγόριθμοι με τους οποίους μπορεί να υλοποιηθεί ένα μοντέλο παλινδρόμησης καθώς και κάποιες διαδικασίες και μετρικές αξιολόγησής

του. Στο τέλος, αναφέρονται μερικές τεχνικές προεπεξεργασίας των δεδομένων, οι οποίες τα μετασχηματίζουν στην κατάλληλη μορφή για την αξιοποίησή τους από τους αλγορίθμους αλλά και παρουσιάζεται συνοπτικά το πεδίο της Επεξεργασίας Φυσικής Γλώσσας, και ειδικότερα της ανάλυσης συναισθήματος για την εξαγωγή επιπλέον χαρακτηριστικών.

Στο κεφάλαιο 4, αναλύεται το τεχνικό υπόβαθρο της εργασίας. Πιο συγκεκριμένα, ορίζεται η γλώσσα προγραμματισμού και οι διάφορες βιβλιοθήκες που θα χρησιμοποιηθούν αλλά και τα όποια εργαλεία και προγραμματιστικές διεπαφές θα επιστρατευτούν για τη συλλογή των δεδομένων.

Στο κεφάλαιο 5, αρχίζει πραγματικά η υλοποίηση της μελέτης μας, καθώς πραγματοποιείται η προετοιμασία των πειραμάτων. Πρώτα, αναλύονται το είδος και η κατανομή των εξαρτημένων μεταβλητών, για να έχουμε μια καλύτερη εικόνα του προβλήματος. Έπειτα, γίνεται γνωστός ο τρόπος που συλλέξαμε τα δεδομένα, ποια χαρακτηριστικά και σε ποια χρονικά παράθυρα τελικά εξήχθησαν καθώς και ποιες τεχνικές προεπεξεργασίας εφαρμόστηκαν σε αυτά. Τέλος, πραγματοποιείται διερευνητική ανάλυση των δεδομένων με σκοπό την ανάδειξη εκείνων των χαρακτηριστικών που φαίνονται καλύτερα για την εκτίμηση της τηλεθέασης εκπομπών.

Στο κεφάλαιο 6, διεξάγονται όλα τα πειράματα και αξιολογούνται τα αποτελέσματά τους. Πιο συγκεκριμένα, κατασκευάζονται μοντέλα παλινδρόμησης με διάφορες τεχνικές και συνδυασμούς χαρακτηριστικών ως εισόδους και ελέγχεται η ποιότητα των προβλέψεών τους για όλες τις εξαρτημένες μεταβλητές, που καλούμαστε να εκτιμήσουμε. Τέλος, συγκρίνεται η απόδοση όλων των μοντέλων τόσο μεταξύ τους όσο και με αποτελέσματα προϋπαρχουσών εργασιών.

Καταληκτικά, στο κεφάλαιο 7, ανακεφαλαιώνουμε τη δουλειά μας, εξάγουμε τα τελικά συμπεράσματα και προτείνουμε παραπάνω επεκτάσεις που θα μπορούσαν να πραγματοποιηθούν στο μέλλον.

Κεφάλαιο 2

Συναφής βιβλιογραφία

2.1 Προβλέψεις σε διάφορους τομείς

2.1.1 Με δεδομένα από μέσα κοινωνικής δικτύωσης

Την τελευταία δεκαετία έχει αναγνωριστεί η προβλεπτική ικανότητα των δεδομένων που εξάγονται από μέσα κοινωνικής δικτύωσης και έχει διερευνηθεί η χρήση τους για την εξαγωγή συμπερασμάτων σε ένα ευρύ φάσμα πεδίων μελέτης, όπως η οικονομία, η υγεία, η πολιτική, το περιβάλλον και πολλά άλλα. Σε αυτήν την ενότητα θα αναφερθούμε σε έρευνες που αξιοποίησαν δεδομένα από την πλατφόρμα του Twitter, μιας και αυτό είναι το μέσο κοινωνικής δικτύωσης που θα εκμεταλλευτούμε στην παρούσα εργασία, για την εκτίμηση μεγεθών και την ανίχνευση φαινομένων σε διάφορους τομείς.

Ένας από τους κύριους τομείς ενδιαφέροντος είναι αυτός της οικονομίας, στα πλαίσια του οποίου, έχουν πραγματοποιηθεί διάφορες μελέτες για την πρόβλεψη των χρηματιστηριακών δεικτών. Οι Zhang, Fuehres και Gloor (2011) [98] πραγματοποίησαν ανάλυση συναισθήματος σε ένα σύνολο αναρτήσεων από το Twitter και παρατήρησαν υψηλή αρνητική τις περισσότερες φορές αλλά και θετική σε μία περίπτωση συσχέτιση μεταξύ των συναισθηματικών μετρικών και των διάφορων χρηματιστηριακών δεικτών. Προς την ίδια κατεύθυνση κινήθηκαν οι Bollen, Mao και Zeng (2011) [9], οι οποίοι εξήγαγαν μετρήσεις της κοινής γνώμης και διάθεσης και έλεγξαν κατά πόσο αυτή σχετίζεται και μπορεί συμπληρωματικά να χρησιμοποιηθεί για την πρόβλεψη των αλλαγών σε ένα χρηματιστηριακό δείκτη, αυξάνοντας την πιστότητα ενός αυτο-οργανούμενου ασαφούς νευρωνικού δικτύου στο 87.6%. Επίσης, έχει μελετηθεί η πρόβλεψη της αξίας κρυπτονομισμάτων, με βάση δεδομένα από το Twitter, όπως στην περίπτωση της εργασίας των Shen, Urquhart και Wang (2019) [77], όπου παρατηρήθηκε ότι ο όγκος των σχετικών δημοσιεύσεων στο Twitter μπορεί να αποβεί προβλεπτικός για την τιμή του bitcoin την επόμενη μέρα.

Άλλη μια κλασική περίπτωση μελέτης, συνιστά η πρόβλεψη των αποτελεσμάτων πολιτικών εκλογών βάσει δεδομένων από μέσα κοινωνικής δικτύωσης και ειδικά το Twitter. Οι Tumasjan, Sprenger, Sandner και Welpe (2010) [85] προβλέπουν τα αποτελέσματα των γερμανικών ομοσπονδιακών εκλογών με βάση τον όγκο των δημοσιεύσεων στο Twitter για πολιτικά κόμματα και πρόσωπα που συμμετέχουν σε αυτές, πετυχαίνοντας MAE 1.65. Επίσης, οι Wang, Can, Kazemzadeh, Bar και Narayanan (2012) [90] πέρα από τον όγκο των tweets, εκμεταλλεύτηκαν και τα συναισθηματικά πρόσημα με τα οποία αυτά ήταν φορτισμένα για την εκτίμηση των αλλαγών της κοινής γνώμης σε πραγματικό χρόνο για τους υποψήφιους στις προεδρικές εκλογές των ΗΠΑ.

Οι Aramaki, Maskawa και Morita (2011)[5] ανίχνευσαν επιδημιολογικά κύματα της γρίπης χρησιμοποιώντας αναρτήσεις στο Twitter που αναφέρονται σε αυτήν. Κατασκεύασαν έναν ταξινομητή με τη βοήθεια των μηχανών διανυσμάτων υποστήριξης και αξιολόγησαν τις προβλέψεις του, υπολογίζοντας το συντελεστή συσχέτισης του Pearson μεταξύ αυτών και των πραγματικών αποτελεσμάτων, ο οποίος τελικά προέκυψε πολύ υψηλός (0.89) και άγγιξε ακόμα και το 0.97 στα αρχικά στάδια της επιδημίας. Παρόμοια προβλήματα υγειονομικού ενδιαφέροντος μελέτη-

σαν οι Achrekar, Gandhe, Lazarus, Yu και Liu το 2011 στο [1] αλλά και οι Paul, Dredze και Broniatowski το 2014 στο [62], καθώς και πολλοί άλλοι ερευνητές έκτοτε.

Άλλες περιπτώσεις αξιοποίησης δεδομένων από τα μέσα κοινωνικής δικτύωσης αποτελούν η ανίχνευση ακραίων φυσικών και καιρικών φαινομένων, π.χ. ένας σεισμός και μια δασική πυρκαγιά, όπως έχουν αναλύσει οι Sakaki, Okazaki, Matsuo το 2010 στο [73] και οι Slavkovikj, Verstockt, Van Hoecke και Van de Walle το 2014 στο [78]. Επίσης, τα δεδομένα από το Twitter μπορούν να συνδράμουν στην αυτόματη ανίχνευση εγκλημάτων και την πρόβλεψη της εκδήλωσης παραβατικής συμπεριφοράς, όπως απέδειξαν οι Wang, Gerber και Brown το 2012 [91] αναλύοντας με τεχνικές επεξεργασίας φυσικής γλώσσας το κειμενικό περιεχόμενο των tweets και χρησιμοποιώντας ένα γραμμικό μοντέλο πρόβλεψης. Τέλος, ένα παρεμφερές παράδειγμα με το αντικείμενο της παρούσας εργασίας αποτελεί η εκτίμηση της απήχησης των διαφημίσεων, η οποία μελετήθηκε από τους Oh, Sasser και Almahmoud το 2015 [59] στην περίπτωση του μεγάλου αθλητικού γεγονότος Super Bowl με τη βοήθεια μοντέλων πολλαπλής γραμμικής παλινδρόμησης με δεδομένα από το Twitter και άλλα στοιχεία.

2.1.2 Με δεδομένα από μηχανές αναζήτησης

Αντίστοιχα, τα στατιστικά δεδομένα για αναζητήσεις που πραγματοποιούνται σε μηχανές αναζήτησης στο διαδίκτυο έχουν αναδειχθεί ως ιδιαίτερα χρήσιμα για την εξαγωγή συμπερασμάτων και προβλέψεων σε πολλούς τομείς. Μερικά παραδείγματα μελετών που συνέβαλαν στην καθιέρωσή τους ως χαρακτηριστικά εισόδου για μοντέλα πρόβλεψης είναι τα ακόλουθα:

Καταρχάς, η πλατφόρμα Google Trends έχει χρησιμοποιηθεί ευρέως για την εξαγωγή δεδομένων για αναζητήσεις υγειονομικού ενδιαφέροντος πρωτοστατώντας ως εργαλείο για έρευνες στον τομέα της υγείας, όπως παρατήρησαν και οι Nuti, Wayda, Ranasinghe, Wang, Dreyer, Chen και Murugiah (2014) [56]. Μάλιστα, οι Lin, Liu και Chiu (2020) [43] το χρησιμοποίησαν για την ανάδειξη της συσχέτισης της εξάπλωσης της πανδημίας του COVID-19 με τον όγκο των αναζητήσεων των χρηστών για προστατευτικά μέτρα στο αμέσως προηγούμενο χρονικό διάστημα, η οποία βρέθηκε έντονα αρνητική ($r = -0.70$), υπογραμμίζοντας τη σημασία της εφαρμογής τους για τον περιορισμό της εξάπλωσης της πανδημίας. Επίσης, πέρα από τη μηχανή αναζήτησης της Google, και άλλες, όπως η Yahoo! και το Baidu, έχουν χρησιμοποιηθεί για την πρόβλεψη της εξάπλωσης της γρίπης [66, 97].

Επιπλέον, ήδη από το 2012, οι Choi και Varian ανέδειξαν τις δυνατότητες της πλατφόρμας Google Trends για την πραγματοποίηση προβλέψεων σε διάφορες περιπτώσεις χρήσης και κατάφεραν με απλά αυτοπαλινδρομούμενα μοντέλα να υπερβούν την απόδοση παραδοσιακών τεχνικών πρόβλεψης [14]. Πάλι με αυτοπαλινδρομούμενα μοντέλα, οι Dimpfl, Jank (2016) [22] έδειξαν ότι ο αυξημένος αριθμός αναζητήσεων σχετικών με το χρηματιστήριο μια μέρα οδηγεί σε αυξημένη μεταβλητότητα σε αυτό την επόμενη. Επίσης, οι Wu, Brynjolfsson (2015) [95] εκτιμούν το πώς θα κινηθεί η αγορά ακινήτων με βάση τον όγκο των αναζητήσεων στη μηχανή αναζήτησης της Google το αμέσως προηγούμενο χρονικό διάστημα και κάποια άλλα στοιχεία, τα οποία εισάγουν σε μοντέλα πολλαπλής γραμμικής παλινδρόμησης. Η αξιολόγηση των αποτελεσμάτων είναι κατά 23.6% καλύτερη από αυτή εμπειρογνωμόνων που βασίστηκαν σε συμβατικά μοντέλα πρόβλεψης και η μελέτη μπορεί να επεκταθεί και σε άλλες αγορές.

Επίσης, ο όγκος των αναζητήσεων από τη μηχανή της Google ή και άλλες μηχανές αναζήτησης για μια θέση εργασίας μπορεί να αξιοποιηθεί ως δείκτης της ανεργίας. Αυτό έχει μελετηθεί σε διάφορες έρευνες, μία εκ των οποίων αφορά τις ΗΠΑ και διεξήχθη από τους D'Amuri, Marcucci το 2017 [20] χρησιμοποιώντας απλά αυτοπαλινδρομούμενα μοντέλα. Τέλος, ο όγκος των αναζητήσεων για αεροπορικά εισιτήρια μπορεί να συνδράμει στην πρόβλεψη της κίνησης στην τουριστική περίοδο για συγκεκριμένους προορισμούς. Ένα παράδειγμα μιας τέτοιας μελέτης αποτελεί αυτή των Yang, Pan, Evans και Lv (2015) [96] για τον τουρισμό στην Κίνα που πραγματοποιήθηκε με δεδομένα που συλλέχθηκαν από το Baidu.

2.2 Προβλέψεις στη βιομηχανία της τηλεόρασης και του κινηματογράφου

2.2.1 Με δεδομένα από μέσα κοινωνικής δικτύωσης

Το Twitter έχει χρησιμοποιηθεί αρκετές φορές στο παρελθόν ως πηγή γνώσης για την εξυπηρέτηση του σκοπού της παρούσας διπλωματικής εργασίας, που δεν είναι άλλος από την εκτίμηση της τηλεθέασης προγραμμάτων. Ενδεικτικά, παρακάτω παραθέτουμε μερικές από αυτές τις δημοσιευμένες εργασίες και τα αποτελέσματά τους καθώς και πιθανές προεκτάσεις τους.

Από τις πρώτες απόπειρες για σύνδεση δεδομένων από την πλατφόρμα Twitter με την τηλεθέαση αποτελούν οι εργασίες των Wakamiya, Lee, Sumiya το 2011 [89, 88]. Σε αυτές μελετήθηκε η συσχέτιση μεταξύ των αναρτήσεων των χρηστών στο Twitter και της απήχησης τηλεοπτικών προγραμμάτων στην Ιαπωνία και υπογραμμίστηκε η ανάγκη για πρόταση νέων μεθόδων εκτίμησης της τηλεθέασης στο μεταβαλλόμενο πεδίο της τηλεοπτικής βιομηχανίας, χωρίς ωστόσο να προτείνεται κάποιο μοντέλο πρόβλεψης.

Το πρώτο στατιστικό μοντέλο πρόβλεψης της τηλεθέασης εκπομπών με βάση τη δραστηριότητα των χρηστών στο Twitter μάλλον προτάθηκε από τον Giglietto το 2013 [28]. Αυτή η έρευνα βασίστηκε σε 11 ιταλικές τηλεοπτικές εκπομπές με συζητήσεις πάνω σε πολιτικά θέματα και στα πλαίσιά της εκπαιδεύτηκε ένα μοντέλο πολλαπλής παλινδρόμησης με είσοδο το ρυθμό σχετικών δημοσιεύσεων στο Twitter και το πλήθος των χρηστών που τις πραγματοποίησε κατά τη διάρκεια προβολής του κάθε επεισοδίου, επιτυγχάνοντας συντελεστή προσδιορισμού R^2 έως και 0.96 κατά την εκπαίδευση.

Στη συνέχεια, οι Sommerdijk, Sanders και van den Bosch (2016) [80] μελέτησαν τη συσχέτιση μεταξύ του πλήθους των tweets και των τηλεθεατών για τα 11 δημοφιλέστερα τηλεοπτικά προγράμματα στην Ολλανδία. Χρησιμοποιήθηκαν τρία χρονικά παράθυρα (κατά τη διάρκεια προβολής, μισή ώρα πριν και μισή ώρα μετά από το εκάστοτε επεισόδιο) και στα αντίστοιχα πειράματα προέκυψε συντελεστής συσχέτισης του Pearson 0.57-0.82. Όμως, η συσχέτιση μειώνεται κατά πολύ αν αφαιρεθούν από το σύνολο δεδομένων τα πιο δημοφιλή προγράμματα. Ταυτόχρονα, αν και για την ίδια εκπομπή μεταξύ διαδοχικών επεισοδίων παρατηρήθηκε μεγάλη συσχέτιση στην τηλεθεασή τους, αυτό δεν αντικατοπτρίστηκε στο πλήθος των tweets. Τέλος, δεν προτάθηκε κάποιο μοντέλο πρόβλεψης και δε χρησιμοποιήθηκαν περισσότερα και άλλου είδους χαρακτηριστικά π.χ. από συναισθηματική ανάλυση.

Έπειτα, οι Molteni και De Leon (2016) [53] εκτίμησαν την τηλεθέαση δημοφιλών αμερικάνικων prime time σειρών. Μια καινοτομία που εισήγαγαν ήταν η διάσπαση των σειρών του συνόλου δεδομένων σε 3 ομάδες ανάλογα με το μέσο πλήθος και τη διασπορά των τηλεθεατών τους και η πρόταση ενός ξεχωριστού μοντέλου πρόβλεψης για την κάθε ομάδα. Πρόκειται για μοντέλα γραμμικής παλινδρόμησης υλοποιημένα με τη μέθοδο των ελάχιστων τετραγώνων, τα οποία παρά την απλότητα τους παρουσίασαν πολύ καλή απόδοση (R 0.86-0.97, συντελεστή προσδιορισμού R^2 0.73-0.94 και τυπικό σφάλμα 0.43-0.72). Τα χαρακτηριστικά που χρησιμοποιήθηκαν ήταν ο όγκος των συνολικών tweets και των θετικών, αρνητικών, ουδέτερων και συναισθηματικά φορτισμένων (θετικών/αρνητικών) tweets καθώς και τα ποσοστά αυτών επί του συνόλου για πέντε διαφορετικά χρονικά παράθυρα κατά τη διάρκεια και πριν την προβολή του εκάστοτε επεισοδίου. Θα ήταν ενδιαφέρον η συγκεκριμένη μελέτη να επεκταθεί σε μεγαλύτερο σύνολο δεδομένων τόσο από άποψη σειρών όσο και από άποψη επεισοδίων της κάθε σειράς και να εξετασθούν παραπάνω τεχνικές παλινδρόμησης.

Επίσης, οι Crisci, Grasso, Nesi, Pantaleo, Paoli και Zara (2018) [19] εκτίμησαν την τηλεθέαση τριών reality shows που προβλήθηκαν στην Ιταλία τις σεζόν 2015-16. Πιο συγκεκριμένα, προβλέφθηκε η τηλεθέαση λίγων τελευταίων επεισοδίων σύμφωνα με μοντέλα που κατασκευάστηκαν με βάση τα υπόλοιπα επεισόδια. Προτάθηκε ένα μοντέλο γραμμικής παλινδρόμησης με είσοδο χαρακτηριστικά ποσοτικών δεικτών (πλήθος των tweets, των retweets και των διακριτών χρηστών που τα δημοσίευσαν) αλλά και πιο σύνθετα μοντέλα παλινδρόμησης κορυφογραμμής, LASSO και elastic net που πέρα από χαρακτηριστικά ποσοτικών δεικτών ως είσοδο έχουν και

χαρακτηριστικά που προέκυψαν από επεξεργασία φυσικής γλώσσας και ανάλυση συναισθήματος στα tweets. Πραγματοποιήθηκε ανάλυση κύριων συνιστωσών και κρατήθηκαν οι σημαντικότερες, καθώς το σύνολο δεδομένων δεν ήταν αρκετά μεγάλο για να μπορούν να χρησιμοποιηθούν όλα τα χαρακτηριστικά που παρήχθησαν ταυτόχρονα. Τελικά, στο σύνολο αυτών των μοντέλων επιτεύχθηκε MAPE 0.05-0.30 αλλά δε δοκιμάστηκαν πιο σύνθετα μοντέλα, όπως μηχανές διανυσμάτων υποστήριξης.

Τέλος, στη διπλωματική εργασία του Χρήστου Πιερράκου [65] προτάθηκαν διάφορα μοντέλα παλινδρόμησης, όπως γραμμική, πολυωνυμική, κορυφογραμμής, LASSO, elastic net, με δέντρα αποφάσεων, τυχαία δάση και XGBoost για την εκτίμηση του μεριδίου τηλεθέασης και του απόλυτου αριθμού των τηλεθεατών μιας ιταλικής σατιρικής εκπομπής. Τα μοντέλα εκπαιδεύτηκαν με βάση δεδομένα από το Twitter που σχετίζονται είτε με ποσοτικούς δείκτες είτε με ανάλυση συναισθήματος στο περιεχόμενο των tweets, πετυχαίνοντας ακρίβεια $92.09\% - 96.31\%$ στο πρώτο πρόβλημα εκτίμησης και $90.86\% - 94.57\%$ στο δεύτερο. Επίσης, χρησιμοποιήθηκαν τεχνικές χρονοσειρών για την εκτίμηση των χαρακτηριστικών μελλοντικών επεισοδίων και την πρόβλεψη της τηλεθέασής τους με είσοδο αυτά στο καλύτερο μοντέλο παλινδρόμησης που κατασκευάστηκε προηγουμένως.

Ένα άλλο δημοφιλές μέσο κοινωνικής δικτύωσης το όποιο έχει χρησιμοποιηθεί ως πηγή γνώσης για επίλυση προβλημάτων σε αυτόν τον τομέα είναι το Facebook. Πιο αναλυτικά, οι Hsieh, Chou, Cheng, Wu (2013) [35] και οι Cheng, Wu, Chen (2016) [13] πρότειναν μοντέλα για την εκτίμηση του ποσοστού τηλεθέασης τηλεοπτικών προγραμμάτων που προβλήθηκαν στην Ταϊβάν βάσει δεδομένων που συλλέχθηκαν από fan pages που διατηρούνταν για αυτά στο Facebook. Στην πρώτη εργασία χρησιμοποιήθηκε ένα νευρωνικό δίκτυο 3 επιπέδων, το οποίο εκπαιδεύτηκε με τον αλγόριθμο back-propagation με τη μέθοδο κατάβασης κλίσης και επιτεύχθηκαν MAPE 5.73%-23.71% και MAE 0.18-0.51. Στη δεύτερη εργασία παρήχθησαν μοντέλα παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων και με διανύσματα υποστήριξης και προέκυψαν MAPE 10%-20% και MSE 0.01-0.42.

Επίσης, οι Sereday και Cui [76] πρότειναν ένα νέο σύστημα για την πρόβλεψη της τηλεθέασης προγραμμάτων. Χρησιμοποίησαν διάφορα δεδομένα από τη Nielsen, όπως το είδος, ο χρόνος προβολής, οι δαπάνες προώθησης και οι ιστορικές μετρήσεις τηλεθέασης για κάποια τηλεοπτικά προγράμματα καθώς και δεδομένα από μέσα κοινωνικής δικτύωσης. Το μοντέλο που επικράτησε ως καλύτερο ανάμεσα σε πολλά που δοκιμάστηκαν βασίστηκε σε gradient boosting μηχανές και εμφάνισε μέση βελτίωση 16% στο WAPE και 41% στο συντελεστή προσδιορισμού R^2 σε σχέση με το προϋπάρχον μοντέλο.

Συμπληρωματικά, ένα συναφές πεδίο έρευνας αποτελεί η εκτίμηση της απήχησης κινηματογραφικών ταινιών. Πρώτα, οι Oghina, Breuss, Tsagkias και De Rijke (2012) [58] εξόρυξαν δεδομένα από το YouTube και το Twitter και τα εκμεταλλεύτηκαν για να εκτιμήσουν τις βαθμολογίες ταινιών που είχαν καταχωρηθεί στο IMDb. Κατασκεύασαν μοντέλα γραμμικής παλινδρόμησης που τροφοδοτούνταν μόνο από χαρακτηριστικά ποσοτικών δεικτών από τις δύο πλατφόρμες, μόνο από χαρακτηριστικά που προέκυψαν από ανάλυση περιεχομένου και από συνδυασμό αυτών των δύο. Χρησιμοποιήθηκε διασταυρούμενη επικύρωση για την αξιολόγηση και τελικά επιτεύχθηκαν έως και $\rho = 0.85$, $\text{MAE} = 0.42$ και $\text{MAPE} = 0.52$.

Προς την ίδια κατεύθυνση κινήθηκαν και οι Schmit και Wubben (2015) [74], οι οποίοι με χαρακτηριστικά που προέκυψαν από ανάλυση περιεχομένου σε tweets εκτιμούν τις ακριβείς βαθμολογίες ταινιών στο IMDb ή γενικότερα την κλάση τους (χαμηλή έως πολύ υψηλή). Για την πρώτη προσέγγιση κατασκευάστηκαν μοντέλα γραμμικής παλινδρόμησης με ελάχιστα τετράγωνα αλλά και με διανύσματα υποστήριξης και σημειώθηκε MSE 0.53-0.57. Για τη δεύτερη προσέγγιση, κατασκευάστηκαν ταξινομητές με διανύσματα υποστήριξης και με στοχαστική βαθμωτή κατάβαση και από την αξιολόγησή τους προέκυψε F1-score περίπου 0.53. Και στις δύο περιπτώσεις η αξιολόγηση πραγματοποιήθηκε με διασταυρούμενη επικύρωση.

Πέρα από την εκτίμηση της βαθμολογίας στο IMDb, ένα παρεμφερές πρόβλημα στο χώρο του κινηματογράφου είναι η εκτίμηση των εισπράξεων από τις πρώτες προβολές μιας ταινίας. Οι

συγγραφείς στα [6, 68, 82, 92] ασχολήθηκαν με αυτό το θέμα ακριβώς. Εξήγαγαν δεδομένα από το Twitter που προέκυπταν είτε από ποσοτικούς δείκτες είτε έπειτα από ανάλυση περιεχομένου με τεχνικές επεξεργασίας φυσικής γλώσσας και τα συνδύασαν με γνωστά χαρακτηριστικά των ταινιών, όπως η δημοφιλία των πρωταγωνιστών, το είδος, ο αριθμός των προβολών, των κινηματογραφικών αιθουσών και η μέση τιμή εισιτηρίου για την εκτίμηση του ζητουμένου. Τα μοντέλα που προτάθηκαν ήταν είτε πολλαπλής γραμμικής είτε ασαφούς παλινδρόμησης και απέδωσαν αρκετά ικανοποιητικά.

Συμπληρωματικά, οι συγγραφείς στα [46] και [8] αντιμετώπισαν το ίδιο πρόβλημα συνδυάζοντας δεδομένα από περισσότερα μέσα κοινωνικής δικτύωσης, όπως το Twitter, το YouTube και το Facebook. Όπως και προηγουμένως, αντλήθηκαν επιπλέον κάποιες γενικές πληροφορίες για τις ταινίες από το IMDb ή το Rotten Tomatoes. Στην πρώτη εκ των δύο εργασιών αναπτύχθηκαν μοντέλα παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων, με διανύσματα υποστήριξης και με πολυεπίπεδο perceptron εκπαιδευμένο με τον αλγόριθμο back-propagation. Η αξιολόγηση με διασταυρούμενη επικύρωση έδωσε προσαρμοσμένο συντελεστή προσδιορισμού R^2 0.58-0.75. Στη δεύτερη εργασία, κατασκευάστηκαν απλούστερα μοντέλα γραμμικής παλινδρόμησης και δόθηκε μεγαλύτερη έμφαση στην παρατήρηση ισχυρής συσχέτισης μεταξύ των δεδομένων από τα μέσα κοινωνικής δικτύωσης και των εισπράξεων από τις προβολές των ταινιών.

2.2.2 Με δεδομένα από μηχανές αναζήτησης

Από όσο γνωρίζουμε δεν έχει διεξαχθεί κάποια έρευνα για την εκτίμηση της τηλεθέασης εκπομπών με βάση μόνο στατιστικά δεδομένα χρήσης μηχανών αναζήτησης. Ωστόσο έχουν δημοσιευθεί κάποιες μελέτες, τις οποίες θα παραθέσουμε στη συνέχεια, σχετικές με το παρεμφερές πρόβλημα της εκτίμησης της απήχησης κινηματογραφικών ταινιών, όσον αφορά την αξιολόγησή τους και τις κριτικές που λαμβάνουν αλλά και τα έσοδα που συγκεντρώνουν από τις πρώτες προβολές.

Καταρχάς, οι Goel, Hofman, Lahaie, Pennock και Watts (2010) [29] παρουσίασαν μία από τις πρώτες απόπειρες για συσχέτιση του όγκου των αναζητήσεων στη μηχανή αναζήτησης της Yahoo με τις εισπράξεις μιας ταινίας από τις πρώτες μέρες κυκλοφορίας της. Το γραμμικό μοντέλο με γκαουσιανό σφάλμα που αναπτύχθηκε απέδωσε αρκετά καλά αλλά όχι καλύτερα από τα ήδη υπάρχοντα παραδοσιακά μοντέλα πρόβλεψης.

Παρόμοια κινήθηκαν οι Miao και Ma (2015) [50] οι οποίοι εκτίμησαν τα έσοδα από τις προβολές κινεζικών ταινιών στους κινηματογράφους με βάση τον όγκο των σχετικών αναζητήσεων στη μηχανή αναζήτησης Baidu. Ανέπτυξαν μοντέλα γραμμικής παλινδρόμησης για διάφορα χρονικά παράθυρα και πέτυχαν συντελεστή προσδιορισμού R^2 έως και 0.85.

Τέλος, οι Demir, Kapralova και Lai (2012) [21] συνέλεξαν δεδομένα από το IMDb για ταινίες και σε συνδυασμό με τη συχνότητα αναζητήσεων για αυτές στη μηχανή αναζήτησης της Google ταξινόμησαν τις ταινίες σε 2 κατηγορίες (χαμηλή-υψηλή αξιολόγηση). Χρησιμοποιήθηκαν μοντέλο λογιστικής παλινδρόμησης, μηχανή διανυσμάτων υποστήριξης και πολυεπίπεδο perceptron, τα οποία ταξινόμησαν σωστά το 64%, το 72% και το 68% των ταινιών, αντίστοιχα.

2.2.3 Με συνδυασμό δεδομένων από μέσα κοινωνικής δικτύωσης και μηχανές αναζήτησης

Πρώτοι οι Huang, Yen, Ku, Lin, Hsieh και Ku (2014) [36] συνδύασαν δεδομένα από τις πλατφόρμες Facebook και Google Trends με σκοπό να εκτιμήσουν το ποσοστά τηλεθέασης διάφορων σειρών. Ανέπτυξαν ένα μοντέλο παλινδρόμησης γκαουσιανής διεργασίας με πυρήνα διαμοιραζομένων βαρών και πέτυχαν αρκετά μικρό MAPE 8.91%-11.18%. Όσον αφορά το συνδυασμό δεδομένων από τις πλατφόρμες Twitter και Google Trends, από όσο γνωρίζουμε δεν έχει υπάρξει κάποια άλλη προσπάθεια για την εκτίμηση της τηλεθέασης, πέρα από τη προγενέστερη δουλειά των συναδέλφων και τη διπλωματική εργασία του Μάριου Παρασκευόπουλου.

Πιο συγκεκριμένα, οι Kalatzis, Roussaki, Matsoukas, Paraskevopoulos, Papavassiliou και Tonoli (2018) [38] πρότειναν μια γενική μεθοδολογία για την αντιμετώπιση του προβλήματος και έδειξαν ότι υπάρχει ισχυρή συσχέτιση (συντελεστής συσχέτισης του Pearson 0.816-0.893) μεταξύ του όγκου των αναζητήσεων στη μηχανή αναζήτησης της Google και του όγκου των δημοσιεύσεων στο Twitter σχετικά με ένα ιταλικό talent show ενώ δεν προτάθηκε κάποιο συγκεκριμένο μοντέλο πρόβλεψης.

Για το ίδιο σύνολο δεδομένων, στη διπλωματική εργασία [61] εξετάστηκαν μοντέλα γραμμικής, πολυωνυμικής και μη γραμμικής παλινδρόμησης, τα οποία εκπαιδεύτηκαν με χαρακτηριστικά ποσοτικών δεικτών, όπως το πλήθος των tweets και το πλήθος των διακριτών χρηστών που τα δημοσίευσαν εντός της μέρας προβολής του εκάστοτε επεισοδίου, καθώς και το κανονικοποιημένο πλήθος των σχετικών αναζητήσεων από την πλατφόρμα Google Trends για όλη τη βδομάδα έως και την προβολή κάθε επεισοδίου. Η αξιολόγηση των αποτελεσμάτων πραγματοποιήθηκε με διασταυρούμενη επικύρωση, λόγω του μικρού μεγέθους του συνόλου δεδομένων και τελικά επιτεύχθηκε MAPE 6.1%-9.6% στο σύνολο των μοντέλων που κατασκευάστηκαν τόσο για την εκτίμηση του μεριδίου τηλεθέασης όσο και για την εκτίμηση του απόλυτου πλήθους τηλεθεατών. Τέλος, επιστρατεύτηκαν γενετικοί αλγόριθμοι για την επαλήθευση των αποτελεσμάτων, με τους οποίους βρέθηκε MAPE 6.1%-11%.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

3.1 Μηχανική μάθηση

Η Μηχανική Μάθηση (Machine Learning - ML) συνιστά έναν υποκλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI), ο οποίος ασχολείται με τους αλγορίθμους που καθιστούν μια υπολογιστική μηχανή ικανή να βελτιώνει την απόδοση της αυτόματα με την απόκτηση εμπειρίας, δηλαδή να «μαθαίνει». Ουσιαστικά, μια υπολογιστική μηχανή σε αυτόν τον τομέα δύναται να παίρνει αποφάσεις και να πραγματοποιεί προβλέψεις αυτόματα, χωρίς να έχει προγραμματιστεί ρητά για αυτό εξ αρχής. Πιο συγκεκριμένα, ο Mitcell το 1997 έγραψε για το ερευνητικό πεδίο της μηχανικής μάθησης: «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετράται από το P, βελτιώνεται μέσω της εμπειρίας E.» [51].

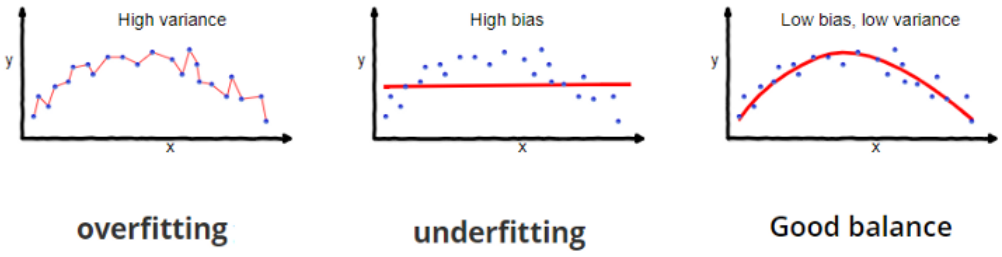
Η μηχανική μάθηση παρουσιάζει πολλές και διάφορες εφαρμογές στην καθημερινή μας ζωή. Κάποιες από αυτές είναι οι εικονικοί προσωπικοί βοηθοί, η πρόβλεψη της κίνησης στους δρόμους, η προσωποποίηση των news feed στα social media, τα συστήματα συστάσεων, το φιλτράρισμα για spam στα email και για malware και πολλά άλλα. Υπάρχουν τέσσερις βασικές κατηγορίες μάθησης: η Επιβλεπόμενη Μάθηση (Supervised Learning), η Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), η Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning) και η Ενισχυτική Μάθηση (Reinforcement Learning). Τα βασικά χαρακτηριστικά αυτών θα αναλυθούν παρακάτω.

Σε όλες τις κατηγορίες οι αλγόριθμοι μηχανικής μάθησης δέχονται ως είσοδο κάποια χαρακτηριστικά (features), τα οποία περιγράφουν τα δεδομένα. Συνήθως πρόκειται για αριθμητικές τιμές, διαφορετικά απαιτείται κατάλληλη προεπεξεργασία τους ώστε να μετατραπούν σε μορφή αξιοποιήσιμη από τον εκάστοτε αλγόριθμο. Με βάση τα δοσμένα χαρακτηριστικά οι αλγόριθμοι μηχανικής μάθησης κατασκευάζουν μοντέλα και πραγματοποιούν την επιθυμητή εργασία, αναλόγως την κατηγορία. Στο τέλος, τα αποτελέσματα που παράγονται αξιολογούνται για την ποιότητα τους σύμφωνα με κάποιες μετρικές, οι οποίες θα αναλυθούν παρακάτω στα πλαίσια του παρόντος προβλήματος. Όλα όσα αναφέρουμε σε αυτό το κεφάλαιο, έχουν βασιστεί στα βιβλία [32, 3, 52, 31], στα οποία αναπτύσσεται η θεωρητική θεμελίωση του πεδίου της Μηχανικής Μάθησης.

3.1.1 Επιβλεπόμενη μάθηση

Στην επιβλεπόμενη μάθηση το σύνολο δεδομένων εκτός από τα χαρακτηριστικά περιλαμβάνει και την επιθυμητή έξοδο για κάθε παρατήρηση - δείγμα εισόδου (labelled data). Ουσιαστικά, ο στόχος είναι η εύρεση ενός γενικού κανόνα αντιστοίχισης των δεδομένων στα αποτελέσματα που να ελαχιστοποιεί μια συνάρτηση κόστους. Αυτός ο γενικός κανόνας συμπεραίνεται μέσα από τη διαδικασία εκπαίδευσης, στην οποία ο αλγόριθμος «μαθαίνει» με βάση τα δεδομένα εκπαίδευσης (training data), ένα υποσύνολο των παρατηρήσεων. Έπειτα, ο αλγόριθμος καλείται να κάνει προβλέψεις σύμφωνα με όσα «έμαθε» και αξιολογείται η απόδοσή του με βάση τις εναπομείνασες παρατηρήσεις, τα δεδομένα ελέγχου (testing data).

Για να προκύψει ένα ικανοποιητικό αποτέλεσμα, αφενός φυσικά το μοντέλο πρέπει να έχει «μάθει» από τα δεδομένα, δηλαδή να έχει αφομοιώσει τη μορφολογική δομή τους ώστε να είναι σε θέση να κάνει μελλοντικές προβλέψεις. Με άλλα λόγια πρέπει να εμφανίζει χαμηλή απόκλιση (bias) ώστε να μην έχει συμβεί υποεκπαίδευση (underfitting). Αφετέρου πρέπει να μπορεί να γενικεύει, δηλαδή να κάνει σωστές προβλέψεις για δεδομένα που δεν έχει ξανασυναντήσει, καθώς δεν περιέχονταν στο σύνολο εκπαίδευσης. Όταν ένα μοντέλο «μαθαίνει» από πολλές λεπτομέρειες και από το θόρυβο των δεδομένων εκπαίδευσης, θα εμφανίσει υψηλή διακύμανση (variance) και θα πέσει στην παγίδα της υπερεκπαίδευσης (overfitting) καθιστώντας αδύνατη την επίτευξη γενίκευσης (generalization). Συνεπώς, είναι επιθυμητό να επιτευχθεί μια καλή ισορροπία μεταξύ απόκλισης και διακύμανσης, όπως απεικονίζεται στο σχήμα 3.1.



Σχήμα 3.1: Παράδειγμα συμβιβασμού μεταξύ απόκλισης και διακύμανσης σε εργασία παλινδρόμησης (πηγή: towardsdatascience)

Υπάρχουν δύο χαρακτηριστικές περιπτώσεις επιβλεπόμενης μάθησης: οι εργασίες ταξινόμησης (classification) και οι εργασίες παλινδρόμησης (regression).

3.1.1.1 Ταξινόμηση

Σε αυτήν την περίπτωση, οι επιθυμητές έξοδοι ή ετικέτες (labels) αποτελούνται από ένα προκαθορισμένο πεπερασμένο σύνολο διακριτών τιμών. Αυτές οι διακριτές τιμές αντιστοιχούν στις διάφορες κλάσεις που υπάρχουν, δηλαδή το σύνολο τιμών της εξόδου είναι $C = [c_1, c_2, ..., c_m]$, όπου m το πλήθος των διαφορετικών κλάσεων. Στόχος είναι να προβλεφθεί η κλάση στην οποία ανήκει ένα νέο παράδειγμα εισόδου. Υπάρχει η δυαδική ταξινόμηση (binary classification), όπου το πλήθος των κλάσεων είναι μόνο δύο και ενδείκνυται για προβλήματα απόφασης που μπορούν να απαντηθούν με ΝΑΙ ή ΟΧΙ. Αλλά υπάρχει και η περίπτωση ταξινόμησης σε περισσότερες εκ των δύο κλάσεων (multi class classification).

3.1.1.2 Παλινδρόμηση

Σε αυτήν την περίπτωση η επιθυμητή έξοδος ανήκει σε ένα εύρος συνεχών τιμών και όχι διακριτών. Με άλλα λόγια, το σύνολο τιμών της εξόδου μπορεί να οριστεί ως $C = [a, b]$, όπου $a, b \in \mathbf{R}$. Στόχος είναι για κάθε νέο δείγμα εισόδου να προβλέπεται ως έξοδος μία τιμή που είναι όσο το δυνατόν πιο κοντά στην πραγματική. Το πρόβλημα της εκτίμησης της τηλεθέασης εκπομπών, τόσο σε επίπεδο ποσοστού τηλεθέασης όσο και σε επίπεδο απόλυτου πλήθους τηλεθεατών, που καλούμαστε να αντιμετωπίσουμε στην παρούσα εργασία ανήκει σε αυτήν την κατηγορία. Συνεπώς, η παλινδρόμηση και οι διάφοροι τρόποι με τους οποίους αυτή μπορεί να υλοποιηθεί θα αναλυθούν εκτενώς στη συνέχεια.

3.1.2 Μη επιβλεπόμενη μάθηση

Στη μη επιβλεπόμενη μάθηση το σύνολο των δεδομένων δεν περιλαμβάνει τις επιθυμητές εξόδους για τις παρατηρήσεις (unlabelled data). Συνεπώς, δεν υπάρχει εξωτερικός εκπαιδευτής ή κριτής που να επιβλέπει τη διαδικασία μάθησης, όπως προηγουμένως. Σε αυτήν την περίπτωση,

ο αλγόριθμος έχοντας στη διάθεση του μόνο μη χαρακτηρισμένα δεδομένα καλείται να μάθει τη δομή τους με αυτο-οργανούμενο τρόπο. Η μη επιβλεπόμενη μάθηση χρησιμοποιείται για την αναγνώριση μοτίβων που μπορεί να κρύβονται στη δομή των δεδομένων και την ανακάλυψη συσχετίσεων σε αυτά χωρίς να είναι γνωστό εκ των προτέρων εάν υπάρχουν, πόσες και ποιες είναι αυτές. Συχνά, για την υλοποίησή της μπορεί να επιστρατευτεί ένας κανόνας ανταγωνιστικής μάθησης (competitive learning).

Συνήθως απαντάται σε προβλήματα συσταδοποίησης ή ομαδοποίησης (clustering), όπου τα δεδομένα σχηματίζουν ομάδες ανάλογα με τις ομοιότητες και τις διαφορές τους και αναγνώρισης ανωμαλιών (anomaly detection), όπου εντοπίζονται αποκλίνουσες συμπεριφορές σε σχέση με την πλειοψηφία των δεδομένων. Επίσης, με μη επιβλεπόμενη μάθηση αντιμετωπίζονται τα προβλήματα αυτοσυσχέτισης (autoassociation), όπου με είσοδο μια ημιτελή ή παραμορφωμένη (με θόρυβο) περιγραφή πρέπει να ανακτηθεί το πρωτότυπο αλλά και η εξαγωγή κανόνων συσχέτισης (association rules), όπου ανακαλύπτονται συσχετίσεις μεταξύ των δεδομένων, π.χ. ένας καταναλωτής που αγοράζει το προϊόν X είναι πολύ πιθανό να αγοράσει και το προϊόν Y.

3.1.3 Ημι-επιβλεπόμενη μάθηση

Στην ημι-επιβλεπόμενη μάθηση το σύνολο εκπαίδευσης αποτελείται τόσο από χαρακτηρισμένα όσο και από μη χαρακτηρισμένα παραδείγματα. Συνήθως τα χαρακτηρισμένα παραδείγματα είναι πολύ λιγότερα από τα μη χαρακτηρισμένα αλλά η ύπαρξη των τελευταίων μπορεί να οδηγήσει σε σημαντική αύξηση της ακριβείας (accuracy) των προβλέψεων. Η ημι-επιβλεπόμενη μάθηση είναι πολύ χρήσιμη σε περιπτώσεις όπου το κόστος προσθήκης ετικετών είναι ιδιαίτερα υψηλό αλλά τα μη χαρακτηρισμένα παραδείγματα αποκτώνται εύκολα, όπως συμβαίνει πολύ συχνά στον πραγματικό κόσμο. Με αυτόν τον τρόπο αντισταθμίζεται το συγκεκριμένο μειονέκτημα της επιβλεπόμενης μάθησης. Συγχρόνως, αμβλύνεται και το περιορισμένο εύρος εφαρμογής που παρουσιάζει η μη επιβλεπόμενη μάθηση. Συνήθως, τα μη χαρακτηρισμένα παραδείγματα θα αποκτήσουν ετικέτες αυτόματα με βάση τα χαρακτηρισμένα αφού πρώτα ομαδοποιηθούν με αυτά με τεχνικές μη επιβλεπόμενης μάθησης. Η ημι-επιβλεπόμενη μάθηση χρησιμοποιείται αρκετά συχνά σε προβλήματα ανάλυσης ομιλίας (speech analysis) και σε προβλήματα κατηγοριοποίησης κειμένου (text classification). Τέλος, χάρη σε αυτή μπορεί να σχεδιαστεί ένα σύστημα μάθησης με καλή κλιμάκωση που επιτρέπει την ταξινόμηση προτύπων σε προβλήματα μεγάλης κλίμακας.

3.1.4 Ενισχυτική μάθηση

Η ενισχυτική μάθηση βρίσκεται κάπου ενδιάμεσα στην επιβλεπόμενη και στη μη επιβλεπόμενη μάθηση. Σε αυτήν την περίπτωση η αντιστοίχιση εισόδου-εξόδου υλοποιείται μέσω της συνεχούς αλληλεπίδρασης ενός συστήματος μάθησης ή πράκτορα (agent) με το περιβάλλον του, με τέτοιο τρόπο ώστε να ελαχιστοποιείται ένας βαθμωτός δείκτης απόδοσης. Πιο συγκεκριμένα, το σύστημα μάθησης εκτελεί μια ενέργεια και «μαθαίνει» από την απόκριση του περιβάλλοντός του ως προς αυτήν την ενέργεια. Η διαδικασία επαναλαμβάνεται έως ότου εξερευνήσει όλες τις πιθανές καταστάσεις που μπορεί να βρεθεί. Ο πράκτορας έχει σκοπό να επιτύχει κάποιο στόχο και κάθε φορά ενεργεί χωρίς να γνωρίζει αν αυτή η κίνηση ήταν προς τη σωστή κατεύθυνση ή όχι παρά μόνο ότι ήταν η βέλτιστη που μπορούσε να κάνει σύμφωνα με αυτά που έχει ήδη «μάθει». Η στρατηγική ενεργειών που ακολουθεί βασίζεται είτε σε μια συνάρτηση ανταμοιβής (reward function) που πρέπει να μεγιστοποιήσει είτε σε μια συνάρτηση κινδύνου (risk function) που πρέπει να ελαχιστοποιήσει. Πρακτικά, ένας κριτής, που μπορεί να είναι ενσωματωμένος στο μηχανισμό μάθησης, έχει το ρόλο του εκπαιδευτή που υπήρχε στην επιβλεπόμενη μάθηση. Χαρακτηριστικά παραδείγματα προβλημάτων τα οποία αντιμετωπίζονται με ενισχυτική μάθηση είναι ο κινηματικός έλεγχος ρομποτικών χειριστών (kinematic control of robotic manipulators), η βελτιστοποίηση εργασιών σε εργοστασιακά περιβάλλοντα, η αυτοοδήγηση οχημάτων (self-driving cars) αλλά και η χάραξη στρατηγικής σε παίγνια δύο αντιπάλων (adversary/two-person games), όπως επιτραπέζια παιχνίδια με αντίπαλο μια υπολογιστική μηχανή (π.χ. σκάκι, Go).

3.2 Παλινδρόμηση

Η παλινδρόμηση είναι μια στατιστική μέθοδος μοντελοποίησης με την οποία μπορεί να προσδιορισθεί αν υπάρχει συσχέτιση μεταξύ μιας εξαρτημένης μεταβλητής ή μεταβλητής εξόδου $\mathbf{y}$ (dependent/outcome/response variable) και μίας ή περισσότερων ανεξάρτητων μεταβλητών ή χαρακτηριστικών $\mathbf{X}$ (independent/explanatory variables / predictors/features). Ορίζεται ως στατιστική μέθοδος ανεξάρτητα του πεδίου της Μηχανική Μάθησης αλλά όπως αναφέραμε και προηγουμένως, χρησιμοποιείται στα πλαίσια της επιβλεπόμενης μάθησης για προβλήματα όπου η επιθυμητή έξοδος ανήκει σε ένα εύρος συνεχών τιμών. Πρώτα, γίνεται η υπόθεση ότι τα δεδομένα ταιριάζουν με κάποιο γνωστό είδος συνάρτησης και έπειτα στη φάση της εκπαίδευσης, καθορίζεται επακριβώς η συνάρτηση αυτού του είδους που μοντελοποιεί καλύτερα τα δεδομένα. Το αποτέλεσμα, λοιπόν, της παλινδρόμησης είναι αυτή η συνάρτηση ή μοντέλο, το οποίο μπορεί να χρησιμοποιηθεί στη συνέχεια για να προβλέψει τις τιμές εξόδου νέων δεδομένων.

Ένα μοντέλο παλινδρόμησης πέρα από τη μεταβλητή εξόδου $\mathbf{y}$ και τις ανεξάρτητες μεταβλητές $\mathbf{X}$ περιλαμβάνει και τις άγνωστες παραμέτρους $\boldsymbol{\beta}$ (coefficients). Αξίζει να σημειωθεί ότι οι ανεξάρτητες μεταβλητές και οι άγνωστες παράμετροι δίνονται σε διανύσματα στη γενική περίπτωση, ωστόσο μπορεί να είναι και βαθμωτά μεγέθη. Τελικά, σύμφωνα με το μοντέλο παλινδρόμησης για ένα παράδειγμα εισόδου i η πραγματική έξοδος y_i θα δίνεται ως συνάρτηση των ανεξάρτητων μεταβλητών $\mathbf{x}_i$ και των συντελεστών β συνοδευόμενη από έναν όρο σφάλματος e_i, ο οποίος εκφράζει το πόσο απέχει η εκτιμώμενη από το μοντέλο τιμή εξόδου για το συγκεκριμένο παράδειγμα από την πραγματική τιμή εξόδου του.

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + e_i \tag{3.1}$$

Υπάρχουν πολλά παραδείγματα προβλημάτων εκτίμησης στην καθημερινότητα όπου ενδείκνυται η χρήση παλινδρόμησης. Μερικά από αυτά είναι η εκτίμηση των τιμών πώλησης ακινήτων συναρτήσει της θέσης τους, του έτους κατασκευής τους και άλλων χαρακτηριστικών, η πρόβλεψη της ζήτησης ενός νέου προϊόντος με βάση τις διαφημιστικές δαπάνες, η εκτίμηση της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση κ.ά. Επίσης, όπως αναφέραμε και παραπάνω, το πρόβλημα που μελετάμε σε αυτήν την εργασία, η εκτίμηση της τηλεθέασης εκπομπών είναι άλλο ένα παράδειγμα προβλήματος παλινδρόμησης.

Τέλος, υπάρχουν διάφοροι τύποι παλινδρόμησης ανάλογα με το είδος της συνάρτησης που έχει υποτεθεί ότι μοντελοποιεί καλύτερα τα δεδομένα (γραμμική, πολυωνυμική, μη γραμμική). Επίσης, η κατηγοριοποίηση μπορεί να γίνει και με βάση το πλήθος των ανεξάρτητων μεταβλητών (απλή, πολλαπλή) και τον τύπο της εξαρτημένης μεταβλητής (π.χ. λογιστική). Στη συνέχεια, αναλύουμε τα διάφορα είδη παλινδρόμησης μαζί με τους βασικότερους αλγορίθμους και τις πιο διαδεδομένες τεχνικές με τις οποίες μπορούν να υλοποιηθούν. Οι πληροφορίες για αυτήν την ενότητα πέρα από τα παραπάνω βιβλία αντλήθηκαν και από τα [23, 71, 54].

3.2.1 Γραμμική παλινδρόμηση με ελάχιστα τετράγωνα

Η γραμμική παλινδρόμηση (linear regression) είναι η πιο απλή και διαδεδομένη μορφή παλινδρόμησης. Σε αυτήν, η εξαρτημένη μεταβλητή $\mathbf{y}$ ορίζεται ως γραμμικός συνδυασμός των ανεξαρτήτων μεταβλητών $\mathbf{x}_j$. Αν υπάρχει μόνο μία ανεξάρτητη μεταβλητή x τότε πρόκειται για απλή γραμμική παλινδρόμηση (simple linear regression) και η εξίσωση που συνδέει το μοναδικό χαρακτηριστικό με τη μεταβλητή εξόδου για την παρατήρηση i είναι η εξής:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{3.2}$$

Αν πάλι υπάρχουν περισσότερες ανεξάρτητες μεταβλητές τότε πρόκειται για πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) και η εξίσωση που συνδέει τη μεταβλητή εξόδου με τις ανεξάρτητες μεταβλητές για την παρατήρηση i είναι η εξής:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + e_i \tag{3.3}$$

όπου m το πλήθος των χαρακτηριστικών. Όπως είναι λογικό, κατά την εκπαίδευση είναι επιθυμητό να βρεθεί η ευθεία στην περίπτωση ενός χαρακτηριστικού, το επίπεδο στην περίπτωση δύο χαρακτηριστικών ή το υπερεπίπεδο στην περίπτωση περισσότερων εκ των δύο χαρακτηριστικών που ταιριάζει περισσότερο στα δεδομένα. Το ποιο τελικά μοντελοποιεί καλύτερα τα δεδομένα μπορεί να αποφασιστεί με βάση διάφορες μετρικές, που θα αναλυθούν εκτενώς παρακάτω, και να βρεθεί με διάφορους αλγορίθμους.

Η πιο απλή και δημοφιλής μέθοδος είναι αυτή των ελάχιστων τετραγώνων (ordinary least squares - OLS). Σύμφωνα με αυτήν, καλύτερο γραμμικό μοντέλο είναι εκείνο που ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολοίπων (residual sum of squares - RSS), όπου ως υπόλοιπο λογίζεται το σφάλμα μεταξύ της προβλεπόμενης από το μοντέλο τιμής εξόδου $\hat{\mathbf{y}}$ και της πραγματικής μεταβλητής εξόδου $\mathbf{y}$. Τελικά, τα βάρη β_j που καθορίζουν το μοντέλο προκύπτουν από την επίλυση του προβλήματος ελαχιστοποίησης τετραγωνικής συνάρτησης (quadratic minimization problem):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \boldsymbol{S}(\boldsymbol{\beta}) \tag{3.4}$$

όπου $\boldsymbol{S}$ είναι η τετραγωνική αντικειμενική συνάρτηση προς ελαχιστοποίηση και ορίζεται ως εξής:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} x_{ij}\beta_j \right|^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{3.5}$$

Στον παραπάνω τύπο, για λόγους συντομίας έχουμε θεωρήσει ότι στα m χαρακτηριστικά περιλαμβάνεται ένα επιπλέον που έχει ως τιμή τη μονάδα και πολλαπλασιάζεται με το σταθερό όρο του γραμμικού μοντέλου. Αυτό πέρα από την παρουσίαση δεν επηρεάζει καθόλου περαιτέρω την ανάλυση μας. Μάλιστα αν οι m στήλες του πίνακα $\mathbf{X}$ είναι γραμμικώς ανεξάρτητες τότε το πρόβλημα έχει γνωστή μοναδική λύση που δίνεται από τον ακόλουθο τύπο:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.6}$$

3.2.2 Πολυωνυμική παλινδρόμηση με ελάχιστα τετράγωνα

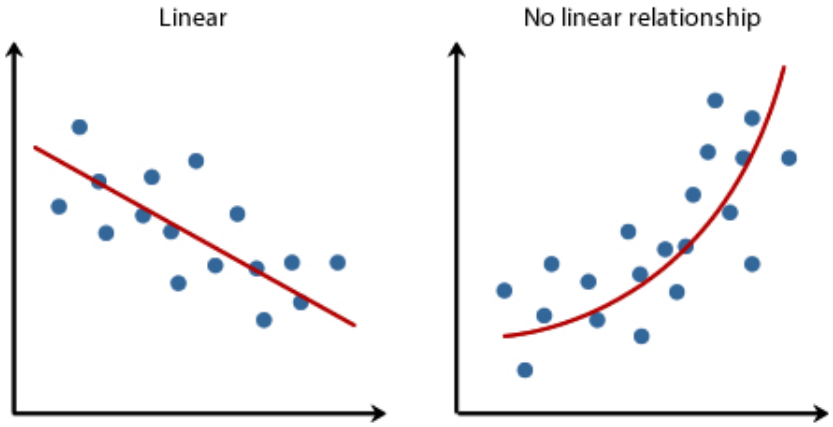
Στην πολυωνυμική παλινδρόμηση (polynomial regression) η μεταβλητή εξόδου $\mathbf{y}$ συσχετίζεται με κάποιο πολυώνυμο των ανεξάρτητων μεταβλητών $\mathbf{x}_j$. Αξίζει να σημειωθεί πως και σε αυτήν την περίπτωση το μοντέλο παραμένει γραμμικό ως προς τους συντελεστές β_j, που είναι οι παράμετροι οι οποίες θα καθοριστούν από την εκπαίδευση του μοντέλου. Για μία ανεξάρτητη μεταβλητή, η εξίσωση που τη συνδέει με τη μεταβλητή εξόδου για την παρατήρηση i γράφεται ως εξής:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + ... + \beta_l {x_i}^l + e_i \tag{3.7}$$

όπου l ο βαθμός του πολυωνύμου που έχει επιλεγεί για τη μοντελοποίηση. Είναι προφανές ότι αυτός ο τύπος παλινδρόμησης προτιμάται όταν μια καμπύλη φαίνεται να ταιριάζει περισσότερο στα δεδομένα από μια ευθεία. Για περισσότερες ανεξάρτητες μεταβλητές η εξίσωση θα περιλαμβάνει όλα τα δυνατά πεπλεγμένα πολυώνυμα έως και βαθμού l. Όλα τα πολυώνυμα χαρακτηριστικών που προκύπτουν, πεπλεγμένα ή μη, μπορούν να θεωρηθούν ως νέα χαρακτηριστικά $\mathbf{X}'$ οπότε και η εξίσωση για την παρατήρηση i γράφεται:

$$y_i = \beta'_0 + \beta'_1 x'_{i1} + \beta'_2 x'_{i2} + ... + \beta'_k x'_{ik} + e_i \tag{3.8}$$

όπου k το πλήθος των νέων χαρακτηριστικών. Εφόσον, λοιπόν, το μοντέλο παραμένει γραμμικό ως προς τα βάρη $\boldsymbol{\beta}$, μας δίνεται η δυνατότητα θεωρώντας νέα χαρακτηριστικά να ανάγουμε την πολυωνυμική παλινδρόμηση στην προηγούμενη περίπτωση της γραμμικής. Συνεπώς, κατά τα γνωστά μπορεί να εφαρμοστεί η μέθοδος των ελαχίστων τετραγώνων. Τέλος, πρέπει να σημειώσουμε ότι επιλέγοντας κάποιο μεγαλύτερο βαθμό πολυωνύμου είναι δυνατό να επιτευχθεί μεγαλύτερη ακρίβεια αλλά ταυτόχρονα είναι επικίνδυνο να οδηγηθούμε σε υπερεκπαίδευση. Όπως αναφέραμε και σε προηγούμενη παράγραφο, είναι σημαντικό να τηρείται ισορροπία μεταξύ απόκλισης και διακύμανσης.



Σχήμα 3.2: Παράδειγμα γραμμικής και μη γραμμικής σχέσης των δεδομένων (πηγή: Laerd Statistics)

3.2.3 Μη γραμμική παλινδρόμηση με ελάχιστα τετραγώνα

Η μη γραμμική παλινδρόμηση (nonlinear regression) είναι η πιο γενική μορφή παλινδρόμησης όπου η εξαρτημένη μεταβλητή $\mathbf{y}$ συσχετίζεται με τις ανεξάρτητες μεταβλητές $\mathbf{x}_j$ μέσω μιας μη γραμμικής συνάρτησης. Για την παρατήρηση i, δηλαδή, ισχύει

$$y_i = \mathbf{f}(\mathbf{x}_i, \boldsymbol{\beta}) + e_i \tag{3.9}$$

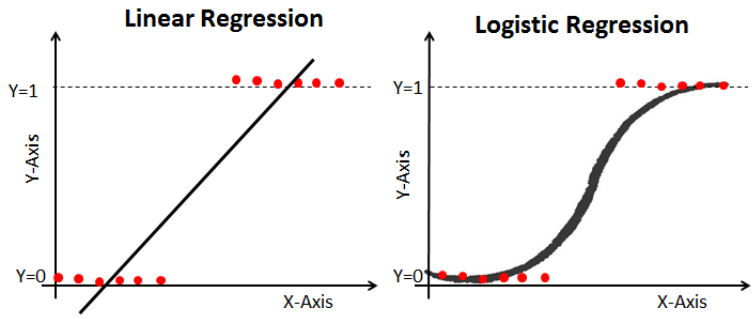
Σε αντίθεση με προηγουμένως, τώρα η συνάρτηση $\mathbf{f}$ είναι μη γραμμική και ως προς τις παραμέτρους β_j, π.χ. μπορεί να είναι της μορφής $f(x, \boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x}$. Άλλες συνήθεις μορφές μη γραμμικών συναρτήσεων είναι οι λογαριθμικές, οι εκθετικές, οι τριγωνομετρικές κ.ά. Μερικές από αυτές είναι δυνατό να γραμμικοποιηθούν και με κάποιο μετασχηματισμό να μετατραπούν στη μορφή

$$y' \sim \mathbf{X}'\boldsymbol{\beta}' \tag{3.10}$$

οπότε και ανάγονται στην περίπτωση της γραμμικής παλινδρόμησης που μπορεί να αντιμετωπιστεί με τη μέθοδο των ελάχιστων τετραγώνων. Αυτές οι συναρτήσεις ονομάζονται εγγενώς γραμμικές (intrinsically linear). Για κάποιες άλλες μη γραμμικές συναρτήσεις προσφέρεται ο εναλλακτικός τρόπος γραμμικοποίησης της τμηματοποίησης. Σύμφωνα με αυτόν, μια μη γραμμική συνάρτηση μπορεί να τεμαχιστεί σε επιμέρους γραμμικά τμήματα και καθένα από αυτά να μοντελοποιηθεί με τη μέθοδο των ελάχιστων τετραγώνων, όπως ακριβώς στην περίπτωση της γραμμικής παλινδρόμησης. Διαφορετικά, αν δηλαδή δεν μπορεί να πραγματοποιηθεί γραμμικοποίηση, η μοντελοποίηση γίνεται με διαδοχικές προσεγγίσεις και επιστρατεύονται αλγόριθμοι αριθμητικής βελτιστοποίησης. Συνεπώς, στη γενική περίπτωση δεν υπάρχει κλειστή μορφή λύσης για το καλύτερο ταιριαστό μοντέλο. [39]

3.2.4 Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση (logistic regression) χρησιμοποιείται όταν θέλουμε να προβλέψουμε την παρουσία ή απουσία ενός χαρακτηριστικού ή συμβάντος. Συνεπώς, απαιτείται η εξαρτημένη μεταβλητή y να είναι δυαδική, δηλαδή να λαμβάνει τιμές 0 και 1, οι οποίες αντιστοιχούν σε νίκη/ήττα, υγιής/ασθενής κλπ. Αξίζει να σημειώσουμε ότι είναι δυνατό να επεκταθεί και για μεγαλύτερο πλήθος κλάσεων (multinomial logistic regression). Το λογιστικό μοντέλο βασίζεται στη λογιστική συνάρτηση, η οποία ανήκει στην οικογένεια των σιγμοειδών καμπυλών που έχουν αποκτήσει αυτή την ονομασία λόγω της μορφής τους, και ουσιαστικά μοντελοποιεί την πιθανότητα εμφάνισης καθεμιάς εκ των δύο κλάσεων. [18]



Σχήμα 3.3: Διαφορά γραμμικής και λογιστικής παλινδρόμησης (πηγή: DataCamp)

3.2.5 Βηματική παλινδρόμηση

Στη βηματική παλινδρόμηση (stepwise regression) πραγματοποιείται με κάποια αυτόματη διαδικασία η επιλογή ενός «καλού» υποσυνόλου χαρακτηριστικών σύμφωνα με ένα προκαθορισμένο κριτήριο. Συνήθως, αυτό λαμβάνει τη μορφή μιας σειράς από F-tests ή t-tests ή και άλλα κριτήρια όπως ο προσαρμοσμένος συντελεστής προσδιορισμού R^2, το Akaike κριτήριο πληροφορίας (AIC) κ.ά. Όποιο κριτήριο και αν χρησιμοποιηθεί, ο στόχος είναι να μεγιστοποιηθεί η προβλεπτική ικανότητα του μοντέλου με το μικρότερο δυνατό πλήθος ανεξάρτητων μεταβλητών. Τα παραπάνω μπορούν να υλοποιηθούν με τρεις προσεγγίσεις. Πρώτον, η προς τα εμπρός επιλογή (forward selection) ξεκινάει τη μοντελοποίηση με καμία ανεξάρτητη μεταβλητή και προοδευτικά προσθέτει σε κάθε βήμα τη μεταβλητή, η συμπερίληψη της οποίας οδηγεί στη στατιστικά σημαντικότερη αύξηση της απόδοσης, αν φυσικά υπάρχει τέτοια μεταβλητή. Δεύτερον, η προς τα πίσω εξάλειψη (backward elimination) ξεκινάει τη μοντελοποίηση λαμβάνοντας υπόψη όλες τις ανεξάρτητες μεταβλητές και σταδιακά αφαιρεί εκείνες που η συμπερίληψή τους δεν οδηγούσε σε στατιστικά σημαντική αύξηση στην απόδοση του μοντέλου. Τέλος, μπορεί να γίνει συνδυασμός των δύο προηγούμενων προσεγγίσεων με εξάλειψη δύο κατευθύνσεων (bidirectional elimination).

3.2.6 Παλινδρόμηση κορυφογραμμής

Η παλινδρόμηση κορυφογραμμής (ridge regression) χρησιμοποιείται όταν υπάρχει υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών που θα χρησιμοποιηθούν ως είσοδοι στο μοντέλο. Σε αυτό το σημείο, οφείλουμε να σημειώσουμε ότι δεν πρέπει να μας παραπλανά το όνομα τους, καθώς οι ανεξάρτητες μεταβλητές δεν είναι απαραίτητα ανεξάρτητες μεταξύ τους και δεν είναι σπάνιο να παρουσιάζεται το φαινόμενο της συγγραμμικότητας (collinearity). Σε αυτήν την περίπτωση, αν εφαρμόσουμε κλασική παλινδρόμηση με τη μέθοδο των ελάχιστων τετραγώνων, το σφάλμα λόγω διακύμανσης είναι μεγάλο. Με την παλινδρόμηση κορυφογραμμής, όμως, εισάγεται σε έναν βαθμό απόκλιση (bias) στο μοντέλο και με αυτόν τον τρόπο τα σφάλματα λόγω συγγραμμικότητας των χαρακτηριστικών μειώνονται. Πιο συγκεκριμένα, αυτό επιτυγχάνεται με την παράμετρο συρρίκνωσης λ στον παρακάτω τύπο:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{Loss} + \lambda \underbrace{\|\boldsymbol{\beta}\|_2^2}_{Penalty} \tag{3.11}$$

Όπως υποδεικνύεται και από τον τύπο, πρόκειται για μέθοδο $L2$ κανονικοποίησης, γνωστή και ως Tikhonov regularization, που ονομάστηκε έτσι από τον εμπνευστή της Andrey Tikhonov. Τέλος, αξίζει να σημειωθεί ότι μπορεί οι συντελεστές των μεταβλητών εισόδου να μικραίνουν αλλά δε φτάνουν ποτέ το 0. Άρα, δεν πραγματοποιείται επιλογή χαρακτηριστικών. [84]

3.2.7 Παλινδρόμηση LASSO

Η μέθοδος παλινδρόμησης LASSO (Least Absolute Shrinkage and Selection Operator), όπως και η παραπάνω, μικραίνει τους συντελεστές του μοντέλου για να επιτύχει μικρότερο σφάλμα. Η κύρια διαφορά με την παλινδρόμηση κορυφογραμμής έγκειται στο γεγονός ότι πρόκειται για μέθοδο $L1$ κανονικοποίησης, όπως φαίνεται και στον παρακάτω τύπο:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{Loss} + \lambda \underbrace{\|\boldsymbol{\beta}\|_1}_{Penalty} \tag{3.12}$$

Αυτό έχει ως συνέπεια τη δυνατότητα να μηδενιστούν κάποιοι συντελεστές στο μοντέλο, επιτρέποντας συνεπώς την επιλογή χαρακτηριστικών. Ειδικότερα, αν υπάρχει μια ομάδα χαρακτηριστικών τα οποία είναι υψηλά συσχετισμένα μεταξύ τους, η μέθοδος LASSO επιλέγει να κρατήσει μόνο ένα εξ αυτών μηδενίζοντας τους συντελεστές όλων των υπολοίπων στην ομάδα. [83]

3.2.8 Παλινδρόμηση elastic net

Μία ακόμα μέθοδος κανονικοποίησης στην παλινδρόμηση είναι το Elastic Net. Πρόκειται για υβριδική μέθοδο των δύο παραπάνω, κορυφογραμμής και LASSO, καθώς όπως φαίνεται στον παρακάτω τύπο οι όροι που αφορούν το πέναλτι περιέχουν $L1$ και $L2$ νόρμες.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \tag{3.13}$$

Συνεπώς, η παλινδρόμηση Elastic Net μπορεί να συνδυάζει τα πλεονεκτήματα και των δύο παραπάνω μεθόδων κανονικοποίησης. [99]

3.2.9 Παλινδρόμηση διανυσμάτων υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) αποτελούν μοντέλα επιβλεπόμενης μάθησης που αρχικά επινοήθηκαν για προβλήματα ταξινόμησης. Στην περίπτωση δυαδικής ταξινόμησης, στοχεύουν στην κατασκευή ενός υπερεπιπέδου που να διαχωρίζει σωστά σε δύο κλάσεις τα δεδομένα εκπαίδευσης ορίζοντας παράλληλα το μεγαλύτερο δυνατό περιθώριο μεταξύ τους. Με άλλα λόγια, στον χώρο χαρακτηριστικών, το υπερεπίπεδο που διαχωρίζει τα δεδομένα εκπαίδευσης θα έχει τη μέγιστη δυνατή απόσταση από τα πιο κοντινά σημεία των δύο κλάσεων, ώστε να επιτευχθεί τελικά χαμηλότερο σφάλμα γενίκευσης. Κατ΄ αντιστοιχία, αν υπάρχουν περισσότερες από δύο κλάσεις, έστω p, με τη μηχανή διανυσμάτων υποστήριξης θα κατασκευαστούν $p-1$ υπερεπίπεδα που θα διαχωρίζουν επιτυχώς τα δεδομένα εκπαίδευσης στις κλάσεις ορίζοντας συγχρόνως τα μεγαλύτερα δυνατά περιθώρια μεταξύ τους.

Αργότερα χρησιμοποιήθηκε η ίδια λογική σε προβλήματα παλινδρόμησης και το σχετικό μοντέλο ονομάστηκε μοντέλο Παλινδρόμησης Διανυσμάτων Υποστήριξης (Support Vector Regression - SVR) [7, 79]. Θεωρούμε $\mathbf{x}_i$ το διάνυσμα χαρακτηριστικών και y_i την πραγματική τιμή εξόδου που καλούμαστε να εκτιμήσουμε της i-οστής παρατήρησης. Ο στόχος είναι να βρεθεί μια συνάρτηση $\mathbf{f}(\mathbf{X})$ που για κάθε δείγμα εκπαίδευσης i το $\mathbf{f}(\mathbf{x_i})$ να μη διαφέρει από το y_i παραπάνω από ϵ ενώ ταυτόχρονα να είναι όσο πιο απλή-επίπεδη γίνεται ώστε να γενικεύει καλύτερα. Με άλλα λόγια, ορίζεται ένα περιθώριο εύρους ϵ γύρω από την καμπύλη παλινδρόμησης που χρησιμοποιείται για την πρόβλεψη νέων τιμών, το οποίο οριοθετείται από δύο παράλληλες οριακές καμπύλες. Τα σημεία που αντιστοιχούν σε δεδομένα εκπαίδευσης και απέχουν το λιγότερο από τις οριακές καμπύλες ονομάζονται διανύσματα υποστήριξης (support vectors).

Όλα αυτά συνοψίζονται στο πρόβλημα βελτιστοποίησης που ορίζεται στις παρακάτω εξισώσεις. Σημειώνουμε ότι όλες οι εξισώσεις σε αυτήν την υποενότητα αφορούν, για λόγους απλότητας, την ειδική περίπτωση της γραμμικής παλινδρόμησης και στο τέλος θα αναφέρουμε πώς μπορούν να μεταβληθούν για άλλα είδη παλινδρόμησης. Η αντικειμενική συνάρτηση προς

ελαχιστοποίηση, δεδομένου ότι η μεγιστοποίηση του περιθωρίου ϵ ισοδυναμεί με την ελαχιστοποίηση της ευκλείδειας νόρμας των βαρών $\boldsymbol{w}$ είναι η εξής:

$$\boldsymbol{\Phi}(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{3.14}$$

με περιορισμούς για όλα τα δείγματα εκπαίδευσης i:

$$-\epsilon \leq y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \leq \epsilon \tag{3.15}$$

Αξίζει να σημειώσουμε πως μπορεί να μην υπάρχει συνάρτηση που να ικανοποιεί αυτούς τους περιορισμούς για όλα τα δείγματα εκπαίδευσης, οπότε εισάγονται μεταβλητές χαλάρωσης (slack variables) ξ_i που επιτρέπουν σφάλμα παλινδρόμησης μέχρι ενός σημείου. Με άλλα λόγια, πρόκειται για περιθώριο ανοχής εύρους ϵ γύρω από την καμπύλη παλινδρόμησης, στο οποίο είναι επιθυμητό να περιέχονται όσο το δυνατόν περισσότερα δείγματα εκπαίδευσης, διατηρώντας ταυτόχρονα χαμηλό σφάλμα γενίκευσης. Η διαφορά με την απλή παλινδρόμηση που βασίζεται στη μέθοδο των ελάχιστων τετραγώνων έγκειται στο ότι εδώ περιορίζουμε το σφάλμα κάτω από ένα συγκεκριμένο κατώφλι, λαμβάνοντας υπόψη μόνο τα σημεία που βρίσκονται εντός των οριακών καμπυλών, δηλαδή αυτών που παρουσιάζουν μικρότερο σφάλμα.

Τώρα, αν N είναι το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης, η αντικειμενική συνάρτηση προς ελαχιστοποίηση γράφεται:

$$\boldsymbol{\Phi}(\boldsymbol{w}, \xi) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{3.16}$$

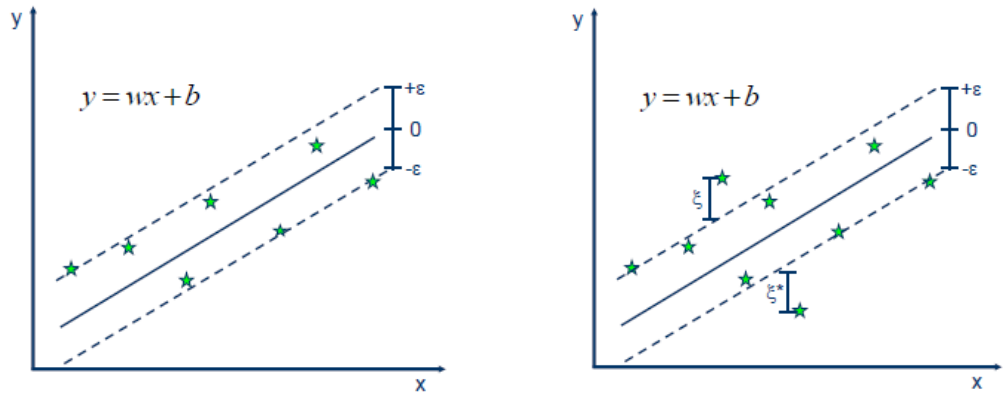
με περιορισμούς για όλα τα δείγματα εκπαίδευσης i:

$$y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \leq \epsilon + \xi_i \tag{3.17}$$

$$\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \leq \epsilon + \xi_i^* \tag{3.18}$$

$$\xi_i, \xi_i^* \geq 0 \tag{3.19}$$

Η σταθερά C εφαρμόζεται ως τιμωρία στις παρατηρήσεις που βρίσκονται εκτός των οριακών καμπυλών και είναι απαραίτητη για την αποφυγή της υπερεκπαίδευσης. Γενικά, υπάρχει συμβιβασμός μεταξύ της γενίκευσης και της ανοχής στο πόσα σημεία μπορούν να βρίσκονται εκτός του περιθωρίου ανοχής.



(α′) χωρίς μεταβλητές χαλάρωσης (β′) με μεταβλητές χαλάρωσης

Σχήμα 3.4: Παράδειγμα γραμμικής παλινδρόμησης με μηχανή διανυσμάτων υποστήριξης (πηγή: Saed Sayad)

Είναι υπολογιστικά αποδοτικότερο να επιλύσουμε το δυϊκό πρόβλημα με πολλαπλασιαστές Langrange από το να επιλύσουμε το ανώτερο πρόβλημα βελτιστοποίησης. Γνωρίζουμε ότι το

δυικό πρόβλημα θα έχει την ίδια λύση με το αρχικό, εφόσον πρόκειται για ελαχιστοποίηση κυρτής συνάρτησης με γραμμικούς περιορισμούς. Η Λανγκρατσιανή συνάρτηση που προκύπτει είναι η εξής:

$$\boldsymbol{L}(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\boldsymbol{x}_i^T\boldsymbol{x}_j + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \tag{3.20}$$

με τους περιορισμούς:

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \tag{3.21}$$

$$0 \leq \alpha_i \leq C, \forall i \tag{3.22}$$

$$0 \leq \alpha_i^* \leq C, \forall i \tag{3.23}$$

Τελικά, τα βάρη $\boldsymbol{w}$ μπορούν να περιγραφούν ως γραμμικός συνδυασμός των παρατηρήσεων εκπαίδευσης:

$$\mathbf{w} = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{3.24}$$

και η συνάρτηση παλινδρόμησης προκύπτει:

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\mathbf{x}_i^T\mathbf{x}) + b \tag{3.25}$$

Τέλος, είναι πολύ σημαντικό ότι το μοντέλο παλινδρόμησης εξαρτάται μόνο από τα διανύσματα στήριξης, καθώς οι πολλαπλασιαστές Langrange a_i, a_i^* είναι μη μηδενικοί μόνο για αυτά, σύμφωνα με τις συνθήκες Karush-Kuhn-Tucker (KKT).

Ένας από τους λόγους που καθιστούν τις μηχανές διανυσμάτων υποστήριξης τόσο δημοφιλείς είναι ότι με αυτές πέρα από γραμμική παλινδρόμηση μπορούν να υλοποιηθούν και άλλα είδη παλινδρόμησης, όπως πολυωνυμική και μη γραμμική. Αυτό είναι εφικτό αν στη συνάρτηση παλινδρόμησης 3.25 το εσωτερικό γινόμενο μεταξύ των διανυσμάτων χαρακτηριστικών αντικατασταθεί από μια συνάρτηση πυρήνα που ορίζεται ως το εσωτερικό γινόμενο των μετασχηματισμένων διανυσμάτων χαρακτηριστικών σε έναν χώρο υψηλότερης διαστατικότητας. Έστω Φ ο μετασχηματισμός αυτός, τότε ως συνάρτηση πυρήνα ορίζεται ως εξής:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle = \mathbf{\Phi}^T(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_j) \tag{3.26}$$

Στα προβλήματα ταξινόμησης, με αυτόν τον τρόπο αρχικά μη γραμμικά διαχωρίσιμα δεδομένα μπορούν να ταξινομηθούν με γραμμικό όριο απόφασης στον χώρο υψηλότερης διαστατικότητας. Αντίστοιχα, στα προβλήματα παλινδρόμησης μη γραμμικές συναρτήσεις στο χώρο των χαρακτηριστικών αρχικά μπορεί να μετασχηματιστούν σε γραμμικές στον χώρο υψηλότερης διαστατικότητας των μετασχηματισμένων χαρακτηριστικών. Στη συνέχεια, παραθέτουμε κάποιους τύπους πυρήνων πέρα από τον τετριμμένο γραμμικό:

- πολυωνυμικός πυρήνας (polynomial kernel):

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + 1)^p \tag{3.27}$$

- πυρήνας ακτινικής συνάρτησης βάσης (radial basis function - RBF kernel):

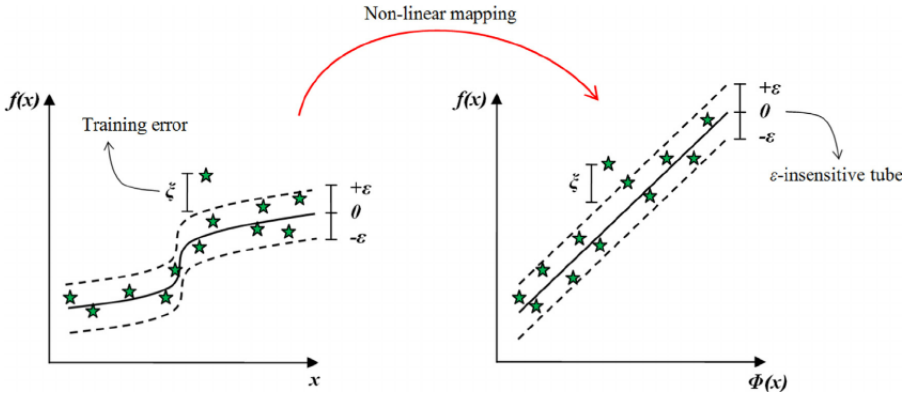
$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} \tag{3.28}$$

- γκαουσιανός πυρήνας (Gaussian kernel) - ειδική περίπτωση RBF:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2} \tag{3.29}$$

- σιγμοειδής πυρήνας (sigmoid kernel):

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(k\mathbf{x}_i^T\mathbf{x}_j + c) \tag{3.30}$$



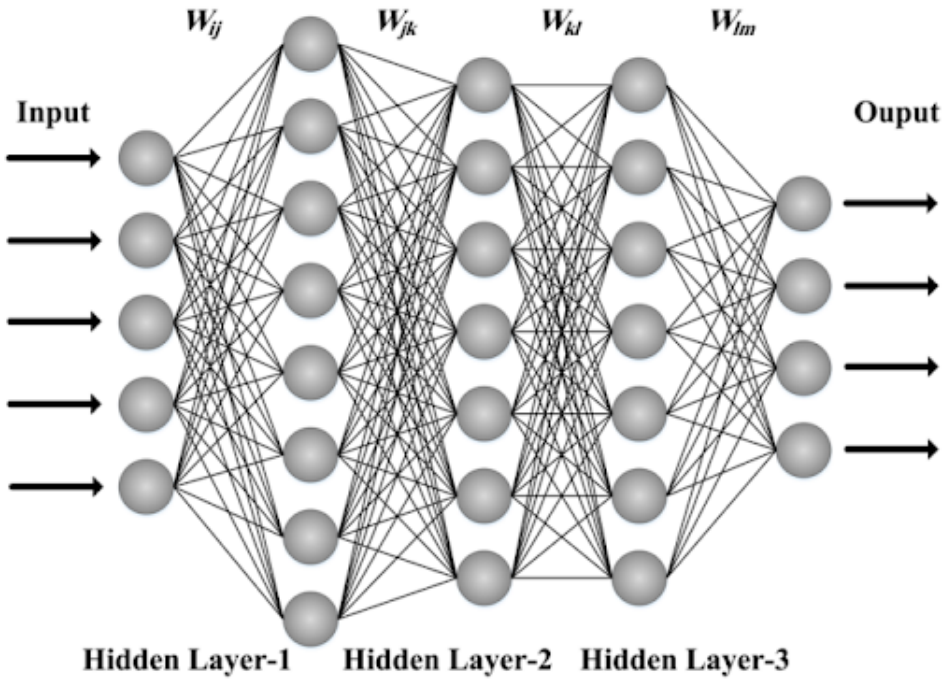
Σχήμα 3.5: Παράδειγμα μετασχηματισμού μη γραμμικής παλινδρόμησης σε γραμμική (πηγή: Saed Sayad)

3.2.10 Παλινδρόμηση με πολυεπίπεδο perceptron

Τα πολυεπίπεδα Perceptron (Multilayer Perceptron - MLP) είναι τεχνητά νευρωνικά δίκτυα που αποτελούνται από πολλές στοιχειώδεις υπολογιστές μονάδες, τους νευρώνες, οργανωμένους σε διάφορα επίπεδα ή στρώματα (layers). Πρόκειται για δίκτυα πρόσθιας τροφοδότησης (feedforward networks), καθώς δεν υπάρχει ανατροφοδότηση της εξόδου κάποιου νευρώνα στην είσοδο του ίδιου ή κάποιου άλλου νευρώνα σε προηγούμενο επίπεδο (κύκλος). Συνήθως, κάθε νευρώνας συνδέεται μόνο με νευρώνες του προηγούμενου και του επόμενου επιπέδου από το επίπεδο που βρίσκεται ο ίδιος. Στην ειδική περίπτωση που κάθε νευρώνας συνδέεται με όλους τους νευρώνες των γειτονικών του επιπέδων, το δίκτυο ονομάζεται πλήρως συνδεδεμένο (fully connected).

Ένα MLP αποτελείται υποχρεωτικά από ένα επίπεδο εισόδου (input layer), τουλάχιστον ένα κρυφό επίπεδο (hidden layer) και ένα επίπεδο εξόδου (output layer). Με το MLP υλοποιείται μια απεικόνιση από το χώρο εισόδου στο χώρο εξόδου και καθίσταται δυνατή η επίλυση προβλημάτων επιβλεπόμενης μάθησης. Ο χώρος εισόδου αντιστοιχεί στα χαρακτηριστικά ενός δείγματος στο πρόβλημα που εξετάζουμε. Ενώ στο επίπεδο εξόδου κάθε νευρώνας αντιστοιχεί σε μια κλάση σε προβλήματα ταξινόμησης ή σε ένα συνεχές μέγεθος που καλούμαστε να εκτιμήσουμε σε προβλήματα παλινδρόμησης. Τα κρυφά επίπεδα με τη σειρά τους εκτελούν ένα μετασχηματισμό από ένα χώρο σε έναν άλλον και ουσιαστικά χρησιμεύουν ως ανιχνευτές χαρακτηριστικών, καθώς μέσα από τη διαδικασία μάθησης αρχίζουν να ανακαλύπτουν τα εξέχοντα χαρακτηριστικά. Αυτό είναι το διαφοροποιητικό στοιχείο μεταξύ του πολυεπίπεδου perceptron και του απλού perceptron του Rosenblatt. Ως βάθος του δικτύου ορίζεται το πλήθος των διαφορετικών επιπέδων που διαθέτει ενώ ως πλάτος το πλήθος των υπολογιστικών μονάδων που υπάρχουν σε κάθε επίπεδο. Τα λεγόμενα βαθιά νευρωνικά δίκτυα (deep neural networks) αναφέρονται σε δίκτυα με πολλά επίπεδα.

Ο μετασχηματισμός από έναν χώρο σε άλλον που πραγματοποιείται σε κάθε νευρώνα εξαρτάται από τα συναπτικά βάρη με τα οποία σταθμίζονται οι είσοδοί του σε αυτόν και τη συνάρτηση ενεργοποίησής του (activation function). Αυτή επιφορτίζεται με τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Ως είσοδο παίρνει το τοπικό πεδίο ή δυναμικό ενεργοποίησης



Σχήμα 3.6: Παράδειγμα Multilayer Perceptron με 3 κρυφά επίπεδα (πηγή: [34])

(activation potential) v του νευρώνα, δηλαδή το σταθμισμένο με τα βάρη άθροισμα εισόδων και το αντιστοιχεί σε μια τιμή από ένα συγκεκριμένο εύρος, τυπικά $[0, 1]$ ή $[-1, 1]$. Η συνάρτηση ενεργοποίησης μπορεί να λαμβάνει διάφορες μορφές αλλά υποχρεωτικά πρέπει να είναι διαφορίσιμη ώστε να καθίσταται δυνατή η εφαρμογή του αλγορίθμου μάθησης, όπως θα δούμε παρακάτω. Σε αυτό το σημείο παραθέτουμε τις πιο διαδεδομένες συναρτήσεις ενεργοποίησης:

- Γραμμική συνάρτηση ταυτότητας (identity):

$$\phi(x) = x \tag{3.31}$$

  Ουσιαστικά είναι σαν να μην υπήρχε συνάρτηση ενεργοποίησης καθώς ισοδυναμεί με το να μεταφέρεται κατευθείαν στην έξοδο το δυναμικό ενεργοποίησης ως έχει, χωρίς κάποιο φιλτράρισμα.

- Σιγμοειδής (Λογιστική) συνάρτηση (sigmoid):

$$\phi(x) = \frac{1}{1 + \mathrm{e}^{-\alpha x}} \tag{3.32}$$

  όπου α η παράμετρος κλίσης. Η συγκεκριμένη συνάρτηση έχει λάβει το όνομά της από τη χαρακτηριστική μορφή της, που θυμίζει το γράμμα S. Είναι αυστηρά αύξουσα και λαμβάνει τιμές από 0 έως 1. Το βασικό μειονέκτημά της είναι ότι δεν παρουσιάζει αισθητή διαφορά στην έξοδο για πολύ μεγάλες ή πολύ μικρές τιμές εισόδου, δημιουργώντας το πρόβλημα της εξαφανιζόμενης κλίσης (vanishing gradient).

- Συνάρτηση υπερβολικής εφαπτομένης (hyperbolic tangent):

$$\phi(x) = \tanh(x) \tag{3.33}$$

  Οι ιδιότητες αυτής της συνάρτησης είναι ίδιες με της σιγμοειδούς με μόνη διαφορά ότι το πεδίο τιμών τώρα είναι $[-1, 1]$ και η συνάρτηση περιττή. Αν και δε φαίνεται μεγάλη διαφορά, διευκολύνει στη διάκριση τριών στάθμεων (αρνητικό, ουδέτερο, θετικό) σε αντίθεση με προηγουμένως που μπορούν να διακριθούν μόνο δύο.

- Ανορθωμένη γραμμική μονάδα ή συνάρτηση ράμπας (rectified linear unit - ReLU):

$$\phi(x) = x^+ = \max(0, x) \tag{3.34}$$

Αυτή η συνάρτηση ταυτίζεται με τη γραμμική συνάρτηση για θετικές τιμές εισόδου ενώ αποκόπτει όλες τις αρνητικές τιμές εισόδου. Το πρόβλημα που εισάγεται σε αυτήν την περίπτωση είναι η «θανάτωση» όλων των νευρώνων που κάποια στιγμή το δυναμικό ενεργοποίησής τους έλαβε αρνητικό πρόσημο, δηλαδή ο αποκλεισμός τους από τη διαδικασία μάθησης (the dying ReLU problem). Το φαινόμενο αυτό αντιμετωπίζεται με μια παραλλαγή της, την Leaky ReLU , η οποία επιτρέπει και τις αρνητικές τιμές να περνούν με μια πολύ μικρή κλίση.

Σύμφωνα με το θεώρημα της καθολικής προσέγγισης (universal approximation theorem), είναι δυνατό με την κατάλληλη αρχιτεκτονική, επιλογή παραμέτρων και συναρτήσεων ενεργοποίησης να προσεγγιστεί οποιαδήποτε συνάρτηση με οσοδήποτε μικρό σφάλμα.

Ένα πολυεπίπεδο perceptron μπορεί να εκπαιδευτεί με τον αλγόριθμο οπισθοδιάδοσης σφάλματος (back propagation) [70], ο οποίος αποτελεί γενίκευση του αλγορίθμου ελαχίστων τετραγώνων. Υποχρεωτικά πρέπει να χρησιμοποιηθεί κάποιος επαναληπτικός αλγόριθμος, γιατί δεν υπάρχει κλειστή μορφή της βέλτιστης λύσης. Στον αλγόριθμο back propagation η εκπαίδευση χωρίζεται σε δύο φάσεις. Στην πρώτη φάση, έχουμε ροή της πληροφορίας προς τα εμπρός. Τα συναπτικά βάρη παραμένουν αμετάβλητα και καθώς προχωράμε από το επίπεδο εισόδου στο επίπεδο εξόδου υπολογίζονται τα νέα δυναμικά ενεργοποίησης και οι έξοδοι των νευρώνων. Στο τέλος της πρώτης φάσης υπολογίζεται η συνολική στιγμιαία ενέργεια σφάλματος όλου του δικτύου από τα σφάλματα που μετρώνται στους νευρώνες εξόδου, η οποία δίνεται συναρτήσει του πλήθους των επαναλήψεων n από τον τύπο:

$$E(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n) \tag{3.35}$$

όπου C το πλήθος των νευρώνων στο επίπεδο εξόδου. Στη δεύτερη φάση, η ροή της πληροφορίας γίνεται προς τα πίσω. Διαδίδεται από το επίπεδο της εξόδου προς το επίπεδο της εισόδου ένα σήμα σφάλματος και υπολογίζεται αναδρομικά σε κάθε βήμα η συνεισφορά κάθε νευρώνα σε αυτό με βάση την κλίση της καμπύλης σφάλματος ως προς τα αντίστοιχα συναπτικά βάρη. Με άλλα λόγια, σε αυτή τη φάση λαμβάνει χώρα η ανάθεση εμπιστοσύνης/υπαιτιότητας (credit assignment). Τα βάρη ενημερώνονται ανάλογα με αυτήν και τον ρυθμό μάθησης η, σύμφωνα με τη μέθοδο της στοχαστικής βαθμωτής κατάβασης (stochastic gradient descent). Η βαθμωτή κατάβαση κλίσης στο χώρο των βαρών σημαίνει αναζήτηση μιας κατεύθυνσης για τη μεταβολή των βαρών η οποία θα μειώνει τη συνολική στιγμιαία ενέργεια σφάλματος, που συνιστά τη συνάρτηση κόστους προς ελαχιστοποίηση. Η διόρθωση των βαρών στην επανάληψη n μεταξύ των νευρώνων i και j δίνεται από τη σχέση:

$$\Delta w_{ij}(n) = -\eta \frac{\partial E(n)}{\partial w_{ij}(n)} = -\eta \frac{\partial E(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ij}(n)} = -\eta \frac{\partial E(n)}{\partial v_j(n)} y_i(n) \tag{3.36}$$

όπου $y_i(n)$ η έξοδος του i νευρώνα στην επανάληψη n. Τώρα, αν ο νευρώνας j ανήκει στο επίπεδο εξόδου:

$$-\frac{\partial E(n)}{\partial v_j(n)} = e_j(n) \phi'(v_j(n)) \tag{3.37}$$

διαφορετικά, αν ανήκει σε κρυφό επίπεδο:

$$-\frac{\partial E(n)}{\partial v_j(n)} = \phi'(v_j(n)) \sum_k -\frac{\partial E(n)}{\partial v_k(n)} w_{kj}(n) \tag{3.38}$$

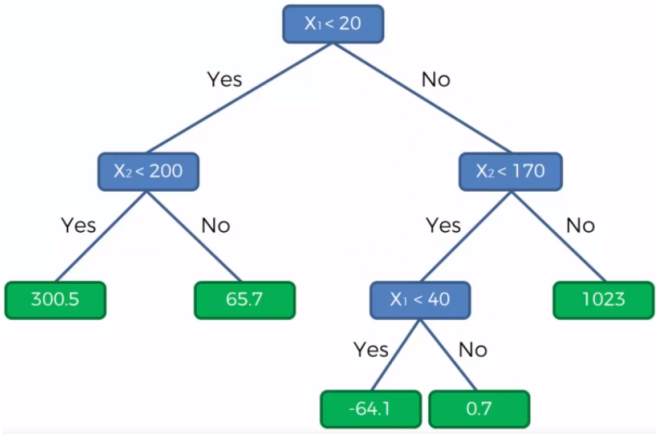
Εφόσον οι ενημερώσεις λαμβάνουν χώρα παράδειγμα προς παράδειγμα, η συγκεκριμένη μέθοδος ανήκει στην κατηγορία της on-line μάθησης. Επιπλέον, ο αλγόριθμος back propagation

είναι απλό να υπολογιστεί σε τοπικό επίπεδο, επιτρέποντας τη χρήση παράλληλων αρχιτεκτονικών. Επίσης, έχει γραμμική πολυπλοκότητα ως προς το πλήθος των παραμέτρων που καλείται να βελτιστοποιήσει, δηλαδή ως προς το σύνολο των συναπτικών βαρών. Ωστόσο δεν έχουμε καμία διαβεβαίωση ότι ο αλγόριθμος θα συγκλίνει και ότι θα εντοπίσει το σύνολο των τιμών των συναπτικών βαρών που οδηγούν στο ολικό ελάχιστο της καμπύλης σφάλματος. Μάλιστα, λόγω της τοπικότητας των υπολογισμών ελλοχεύει ο κίνδυνος να παγιδευτεί σε ένα τοπικό ελάχιστο, αναλόγως και με την αρχικοποίηση των βαρών που έχει προηγηθεί. Συμπληρωματικά, λόγω της στοχαστικής φύσης του αλγορίθμου, η σύγκλιση, όπως και αν οριστεί, τείνει να είναι αργή. Ένα ακόμα μειονέκτημα είναι η κακή κλιμάκωση ως προς το πλήθος των χαρακτηριστικών και την πολυπλοκότητα της υπολογιστικής εργασίας.

Τέλος, αξίζει να σημειωθεί πως στη θέση της στοχαστικής βαθμωτής κατάβασης μπορεί να χρησιμοποιηθεί ο αλγόριθμος βελτιστοποίησης Adam (Adaptive Moment Estimation) [40]. Η βασική διαφοροποίησή του είναι ότι χρησιμοποιείται διαφορετικός προσαρμόσιμος ρυθμός μάθησης για κάθε παράμετρο που πρόκειται να βελτιστοποιηθεί. Επίσης, εκμεταλλεύεται τόσο την πρώτη όσο και τη δεύτερη βαθμίδα της κλίσης. Εναλλακτικά, ο αλγόριθμος L-BFGS που προσεγγίζει τον αλγόριθμο των Broyden-Fletcher-Goldfarb-Shanno κάνοντας χρήση περιορισμένης μνήμης είναι κατάλληλος για την εκπαίδευση ενός MLP [45].

3.2.11 Παλινδρόμηση με δέντρα αποφάσεων

Τα Δέντρα Αποφάσεων (Decision Trees) [4, 69] χρησιμοποιούνται για την κατασκευή μοντέλων ταξινόμησης και παλινδρόμησης με βάση μια δενδρική δομή, όπως άλλωστε μας προϊδεάζει το όνομά τους. Ένα δέντρο απόφασης δημιουργείται επαυξητικά, ξεκινώντας από τον ανώτερο κόμβο-ρίζα και διαιρώντας σε κάθε βήμα το σύνολο των δεδομένων σε όλο και μικρότερα τμήματα. Τελικά, το δέντρο απαρτίζεται από ενδιάμεσους κόμβους-απόφασης και από τερματικούς κόμβους-φύλλα. Κάθε κόμβος απόφασης αφορά ένα χαρακτηριστικό και οι διακλαδώσεις που προκύπτουν από αυτόν σχετίζονται με τις διάφορες τιμές που αυτό μπορεί να λάβει. Τα φύλλα αντιστοιχούν στις προβλέψεις (αριθμητικές στην περίπτωση της παλινδρόμησης, κλάσεις στην περίπτωση της ταξινόμησης).

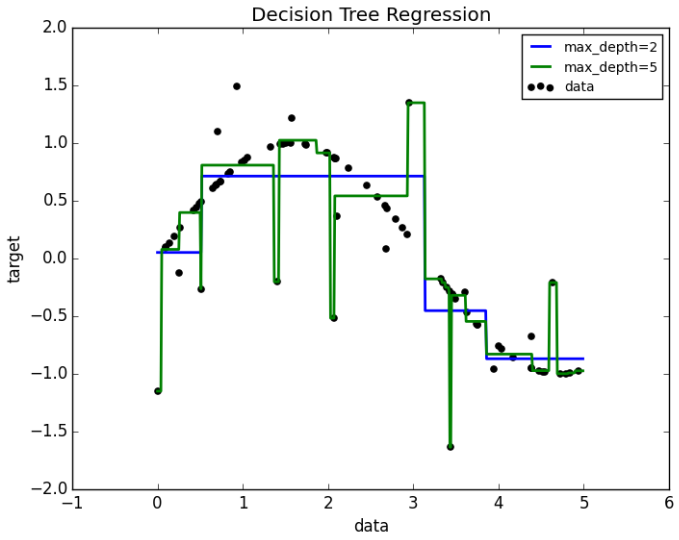


Σχήμα 3.7: Παράδειγμα δέντρου απόφασης σε πρόβλημα παλινδρόμησης με δύο χαρακτηριστικά (πηγή: SuperDataScience)

Υπάρχουν διάφοροι αλγόριθμοι για την εκπαίδευση δέντρων απόφασης, ανάμεσα στους οποίους ο CART, ο ID3 και ο ID5 [87]. Ο ID3 είναι άπληστος αλγόριθμος, καθώς εκτελεί αναζήτηση στο δέντρο ξεκινώντας από την ρίζα και κάθε φορά αποφασίζει προς τα πού θα κατευθυνθεί σύμφωνα με κάποιο άπληστο κριτήριο χωρίς τη δυνατότητα οπισθοχώρησης. Άρα, υπάρχει ο κίνδυνος να μη βρεθεί η βέλτιστη λύση του προβλήματος και ο αλγόριθμος να εγκλωβιστεί σε κάποιο τοπικό ακρότατο. Το άπληστο κριτήριο που χρησιμοποιείται σε προβλήματα παλινδρόμησης είναι η μείωση της τυπικής απόκλισης ή η μείωση του μέσου τετραγωνικού σφάλματος. Ο στόχος

είναι η τελική διάσπαση του συνόλου δεδομένων σε όσο το δυνατόν πιο ομοιόμορφα υποσύνολα, που θα έχουν μικρή τυπική απόκλιση. Ως κριτήριο τερματισμού μπορεί να θεωρηθεί η επίτευξη χαμηλού συντελεστή μεταβλητότητας κάτω από ένα κατώφλι ή ο μικρός πληθικός αριθμός των υποσυνόλων που έχουν δημιουργηθεί. Κάθε κόμβος απόφασης σε ένα μονοπάτι του δέντρου αντιστοιχεί στο χαρακτηριστικό που οδηγεί στη μεγαλύτερη μείωση της τυπικής απόκλισης (άπληστο κριτήριο) σε εκείνο το σημείο του μονοπατιού. Αφού τελειώσει η εκπαίδευση, κάθε τελικό υποσύνολο αντιστοιχεί σε ένα φύλλο, το οποίο με τη σειρά του αντιστοιχεί στην πρόβλεψη μιας τιμής ίσης με τον μέσο όρο των τιμών εξόδου των στιγμιοτύπων αυτού του υποσυνόλου. Η συνάρτηση παλινδρόμησης που παράγεται θα αποτελείται το πολύ από τόσα οριζόντια επίπεδα όσα φύλλα διαθέτει το δέντρο απόφασης.

Τελικά, τα δέντρα απόφασης είναι κατάλληλα για προσέγγιση μη γραμμικών, σύνθετων σχέσεων μεταξύ δεδομένων εισόδου και εξόδου. Ωστόσο, συχνά υπερεκπαιδεύονται και είναι ευαίσθητα στον θόρυβο. Αυτό μπορεί να αποφευχθεί ορίζοντας το κατάλληλο επιτρεπτό μέγιστο βάθος στο δέντρο. Επίσης, μια μικρή αλλαγή στο σύνολο δεδομένων εκπαίδευσης μπορεί να οδηγήσει στην κατασκευή ενός πολύ διαφορετικού δέντρου και κατ΄ επέκταση στην πρόβλεψη πολύ διαφορετικών αποτελεσμάτων. Τέλος, πρέπει να αρκεστούμε σε άπληστους αλγορίθμους για την εκπαίδευση των δέντρων απόφασης, καθώς η εκπαίδευση ενός βέλτιστου δέντρου αποτελεί NP πρόβλημα.

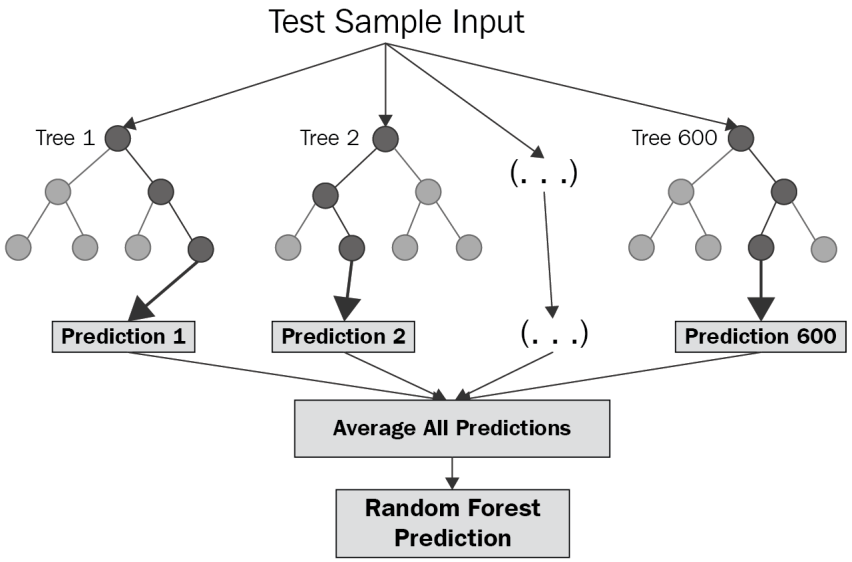


Σχήμα 3.8: Παράδειγμα παλινδρόμησης με δέντρα απόφασης διαφορετικού μέγιστου βάθους (πηγή: Data Science Stack Exchange)

3.2.12 Παλινδρόμηση με τυχαία δάση

Τα Τυχαία Δάση (Random Forests) [42] αποτελούν παράδειγμα συνδυαστικής μεθόδου μάθησης (ensemble learning) και πιο συγκεκριμένα bootstrap aggregating (bagging). Κατά την εκπαίδευσή τους, χτίζονται πολλά δέντρα απόφασης ανεξάρτητα το ένα από το άλλο, τα οποία είναι αρκετά διαφορετικά μεταξύ τους. Αυτό επιτυγχάνεται γιατί κάθε δέντρο χρησιμοποιεί ένα ισάριθμο τυχαία επιλεγμένο με δειγματοληψία με αντικατάσταση (bootstrap) υποσύνολο παρατηρήσεων. Προς την ίδια κατεύθυνση, ο αλγόριθμος εκπαίδευσης κάθε δέντρου είναι ελαφρώς τροποποιημένος ούτως ώστε να χρησιμοποιεί για κάθε δέντρο ένα διαφορετικό υποσύνολο χαρακτηριστικών (feature bagging). Η διαδικασία αυτή λόγω της ανεξαρτησίας των δέντρων και της παντελούς έλλειψης αλληλεπίδρασης μεταξύ τους μπορεί να παραλληλοποιηθεί. Τελικά, η πρόβλεψη του μοντέλου ενός τυχαίου δάσους σε προβλήματα παλινδρόμησης προκύπτει ως ο μέσος όρος των προβλέψεων των επιμέρους ανεξάρτητων βασικών μοντέλων-δέντρων απόφασης. Όλα τα επιμέρους βασικά μοντέλα συνεισφέρουν το ίδιο στο αποτέλεσμα.

Λογικά, όσο περισσότερα δέντρα χρησιμοποιούνται τόσο πιο ακριβή αποτελέσματα θα λαμβάνονται. Επιπλέον, χάρη στην τυχαιότητα που χαρακτηρίζει ένα μοντέλο τυχαίου δάσους, αυτό γενικεύει καλύτερα και κάνει πιο ακριβείς προβλέψεις με μικρότερη διακύμανση από τα ατομικά δέντρα απόφασης. Επίσης, είναι λιγότερο ευαίσθητο σε outliers και η μείωση της διακύμανσης δε συνοδεύεται από σημαντική αύξηση στην απόκλιση. Τέλος, με τα τυχαία δάση δύναται να προσδιορισθούν ποιες μεταβλητές είναι πιο σημαντικές για την παλινδρόμηση.



Σχήμα 3.9: Παράδειγμα παλινδρόμησης με τυχαίο δάσος 600 δέντρων απόφασης - βάσης (πηγή: gitconnected)

3.2.13 Παλινδρόμηση με gradient boosting μηχανές

Μια άλλη κατηγορία μεθόδων συνδυαστικής μάθησης (ensemble learning) είναι η boosting, παράδειγμα της οποίας αποτελούν οι Gradient Boosting Machines [25, 55]. Σε αυτήν την περίπτωση, συνδυάζονται πολλά αδύναμα μοντέλα, που έχουν λίγο καλύτερη απόδοση από την τυχαία επιλογή, με τρόπο ώστε να μετατρέπονται σε ισχυρά. Ως αδύναμα μοντέλα μπορούν να χρησιμοποιηθούν τα δέντρα απόφασης. Η μετατροπή τους σε δυνατά, γίνεται σταδιακά-ακολουθιακά. Κάθε νέο δέντρο που προστίθεται μαθαίνει από τα σφάλματα του προηγούμενου και έτσι γίνεται καλύτερο. Σε αυτό το προσθετικό μοντέλο, τα ήδη υπάρχοντα δέντρα απόφασης παραμένουν αμετάβλητα σε κάθε νέο στάδιο.

Αρχικά, αυτή η ιδέα χρησιμοποιήθηκε σε προβλήματα ταξινόμησης με τον αλγόριθμο Adaptive Boosting (AdaBoost), ο οποίος κάθε φορά απέδιδε μεγαλύτερα βάρη στις παρατηρήσεις που ήταν δύσκολο να ταξινομηθούν και μικρότερα σε εκείνες που ήταν εύκολο. Τελικά, οι προβλέψεις του συνολικού μοντέλου ήταν το σταθμισμένο άθροισμα των προβλέψεων των προηγούμενων μοντέλων. Αργότερα, χρησιμοποιήθηκαν οι Gradient Boosting μηχανές και για προβλήματα παλινδρόμησης. Η βασική διαφοροποίησή τους είναι ότι αντί να αλλάζουν τα βάρη των παρατηρήσεων, χρησιμοποιούν την τεχνική της βαθμωτής κατάβασης κλίσης στη συνάρτηση σφάλματος. Ουσιαστικά, τα βασικά υπομοντέλα προβλέπουν τα σφάλματα και όχι τη μεταβλητή-στόχο. Οι παράμετροι κάθε νέου δέντρου απόφασης αλλάζουν προς την κατεύθυνση που έχουμε μείωση του σφάλματος. Επίσης, τα δέντρα που χρησιμοποιούνται στο Gradient Boosting έχουν μεγαλύτερο βάθος από αυτά στο Adaptive Boosting. Τελικά, οι παρατηρήσεις με μεγαλύτερο σφάλμα εμφανίζονται με μεγαλύτερη πιθανότητα στα νέα βασικά μοντέλα σε αντίθεση με τις bagging μεθόδους μάθησης όπου όλες οι παρατηρήσεις είχαν ίση πιθανότητα να εμφανιστούν σε κάθε μοντέλο βάσης.

Επίσης, αξίζει να σημειώσουμε πως είναι σημαντικό να σταματήσουμε τη διαδικασία της εκπαίδευσης πρόωρα ούτως ώστε να μην πέσουμε στην παγίδα της υπερεκπαίδευσης. Αυτή η κατάλληλη στιγμή είναι όταν τα σφάλματα πλέον κατανέμονται τυχαία και δεν μπορούν να μοντελοποιηθούν περαιτέρω.

3.2.14 Μπεϋζιανή παλινδρόμηση

Στην κλασική γραμμική παλινδρόμηση, για την κατασκευή του μοντέλου χρησιμοποιείται συνήθως η μέθοδος των ελάχιστων τετραγώνων. Σύμφωνα με αυτήν κατασκευάζεται το βέλτιστο υπερεπίπεδο που ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης, ελαχιστοποιώντας το άθροισμα των τετραγώνων των σφαλμάτων. Αποδίδεται, δηλαδή, μία πραγματική τιμή-εκτίμηση σε κάθε συντελεστή β_j της εξίσωσης παλινδρόμησης:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + e_i \tag{3.39}$$

και για κάθε δείγμα εκπαίδευσης i προκύπτει το αντίστοιχο σφάλμα e_i. Κατ΄ επέκταση, για την πρόβλεψη μιας νέας τιμής $\hat{y}_i$ με βάση το διάνυσμα χαρακτηριστικών x_i προκύπτει σύμφωνα με το μοντέλο μία και μόνο πραγματική τιμή.

Το μοντέλο που προκύπτει από την Μπεϋζιανή παλινδρόμηση (Bayesian regression) [27] ακολουθεί την ίδια γραμμική εξίσωση που παραθέσαμε παραπάνω με την υπόθεση ότι τα σφάλματα είναι ανεξάρτητα και ισοκατανεμημένα ακολουθώντας κανονική κατανομή με μηδενική μέση τιμή και διασπορά σ^2, δηλαδή

$$e_i \sim N(0, \sigma^2) \tag{3.40}$$

Κατ΄ επέκταση, σε αυτήν την περίπτωση ως αποτέλεσμα προκύπτει μια κατανομή για τη μεταβλητή εξόδου και όχι απλά μια εκτίμηση ενός σημείου:

$$y_i \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2) \tag{3.41}$$

Κατά την εκπαίδευση, λοιπόν, δε βρίσκονται απλά οι καλύτερες τιμές για τις παραμέτρους του μοντέλου αλλά προσδιορίζεται η a-posteriori κατανομή τους, δεδομένων των δειγμάτων εκπαίδευσης (διανύσματα εισόδου και έξοδοι). Η a-posteriori κατανομή τους είναι ανάλογη του γινομένου της a-priori κατανομής τους και της πιθανότητας εξόδου δεδομένης της εισόδου και των συντελεστών του μοντέλου, όπως ορίζει ο νόμος του Bayes.

$$P(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \frac{P(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{X})P(\boldsymbol{\beta}|\mathbf{X})}{P(\boldsymbol{y}|\mathbf{X})} \tag{3.42}$$

Με αυτόν τον τρόπο, μπορεί να ενσωματωθεί στο μοντέλο και οποιαδήποτε πρότερη γνώση έχουμε για αυτό πέρα από την κλασική διαμόρφωσή του από τα δεδομένα εκπαίδευσης. Αυτό είναι βολικό στην περίπτωση που έχουμε λίγες παρατηρήσεις, καθώς η έλλειψη δεδομένων καλύπτεται από την a-priori κατανομή, δηλαδή την όποια πρότερη γνώση διαθέτουμε. Όσο περισσότερα δεδομένα διαθέτει το σύνολο εκπαίδευσης τόσο περισσότερο αυτά θα επηρεάζουν το μοντέλο, το οποίο θα τείνει να συγκλίνει όλο και περισσότερο στη λύση των ελάχιστων τετραγώνων. Με άλλα λόγια, ξεκινώντας από μία αρχική εκτίμηση για το μοντέλο, τα δεδομένα τη συμπληρώνουν μειώνοντας την αβεβαιότητα για αυτό, είτε ενισχύοντας την αρχική εκτίμηση είτε αποκλίνοντας από αυτήν. Για αυτόν το λόγο, μπορεί να χρησιμοποιηθεί ως τεχνική αυξητικής μάθησης, φορτώνοντας λίγα δεδομένα κάθε φορά στο μοντέλο και αξιοποιώντας αυτά που ήδη έχουν περάσει μέσω της a-priori κατανομής, ενισχύοντας την κλιμακωσιμότητα του μοντέλου.

Επίσης, καθίσταται δυνατή η ποσοτικοποίηση της αβεβαιότητας μας για το μοντέλο, καθώς μπορούμε να λάβουμε εκτίμηση ενός σημείου από τη μέση τιμή της κατανομής και την αβεβαιότητα που αυτή ενέχει μέσω της διασποράς της κατανομής. Προφανώς, η διασπορά, ως μέτρο αβεβαιότητας, μικραίνει όσο περισσότερα δείγματα εκπαίδευσης είναι διαθέσιμα.

Τέλος, στην πράξη η αξιολόγηση των a-posteriori κατανομών των παραμέτρων του μοντέλου είναι υπολογιστικά δύσκολη για συνεχείς μεταβλητές, οπότε επιστρατεύονται οι δειγματοληπτικές

μέθοδοι Monte Carlo για την προσέγγισή τους. Ο πιο συχνός αλγόριθμος για αυτόν τον σκοπό βασίζεται σε Μαρκοβιανές αλυσίδες. Φυσικά, αξίζει να σημειωθεί ότι η Μπεϋζιανή παλινδρόμηση μπορεί να επεκταθεί και για μη γραμμικά μοντέλα.

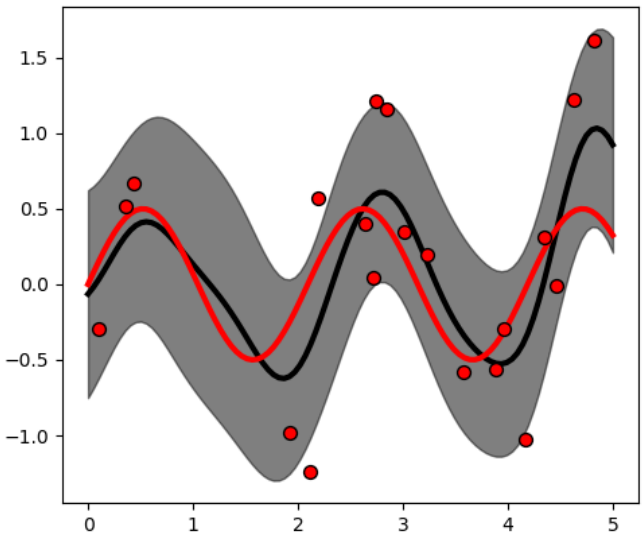
3.2.15 Παλινδρόμηση με γκαουσιανές διεργασίες

Η Παλινδρόμηση με Γκαουσιανές Διεργασίες (Gaussian Process Regression-GPR) [67] είναι μια μη παραμετρική μπεϋζιανή προσέγγιση στην παλινδρόμηση. Προς την ίδια κατεύθυνση με την Μπεϋζιανή παλινδρόμηση, εμφανίζει πιθανοτικό χαρακτήρα αλλά αυτή τη φορά στο χώρο των συναρτήσεων. Πιο αναλυτικά, αντί το μοντέλο να περιγράφεται με μία συνάρτηση, της οποίας το διάνυσμα βαρών έχει βελτιστοποιηθεί, περιγράφεται από μία κατανομή όλων των πιθανών συναρτήσεων που μπορούν να προκύψουν με βάση τα δεδομένα εισόδου. Για αυτό είναι και μη παραμετρική, αφού δεν υπολογίζει κατανομές για τους συντελεστές μιας συνάρτησης αλλά την κατανομή όλων των πιθανών συναρτήσεων του μοντέλου.

Επίσης, πάλι ορίζεται μια a-priori κατανομή για τις συναρτήσεις και με βάση αυτή και τα δεδομένα εκπαίδευσης υπολογίζεται η a-posteriori κατανομή των συναρτήσεων. Με αυτόν τον τρόπο, όπως εξηγήσαμε και παραπάνω, ενσωματώνεται πρότερη γνώση, η οποία μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη όταν έχουμε μικρό πλήθος παρατηρήσεων. Ακόμα, η τεχνική αυτή επιτρέπει την ποσοτικοποίηση της αβεβαιότητας των προβλέψεων. Ουσιαστικά, κατά την εκπαίδευση ρυθμίζεται η μέση συνάρτηση και η συνάρτηση πυρήνα-συνδιακύμανσης, οι οποίες αρκούν για να ορίσουν την κατανομή των συναρτήσεων παλινδρόμησης.

$$\mathbf{f}(\mathbf{x}) \sim \mathbf{GP}\left(\mathbf{m}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}')\right) \tag{3.43}$$

Στους πυρήνες κωδικοποιούνται οι υποθέσεις που γίνονται για το μοντέλο που μαθαίνεται, καθορίζοντας την ομοιότητα δύο σημείων. Όπως είναι λογικό, δύο όμοια σημεία εισόδου πρέπει να έχουν όμοιες εξόδους. Συχνά συναντώνται στατικοί πυρήνες, οι οποίοι βασίζονται μόνο στη χωρική απόσταση δύο σημείων. Γενικότερα, κάποια παραδείγματα αυτών είναι: ο σταθερός, ο γραμμικός, ο τετραγωνικός εκθετικός, η συνάρτηση ακτινικής βάσης, εσωτερικού γινομένου, Matern κ.ά. Επίσης, επιτρέπεται η σύνθεση πυρήνων.



Σχήμα 3.10: Παράδειγμα παλινδρόμησης με γκαουσιανή διεργασία (πηγή: scikit-learn)

Τέλος, γίνεται η υπόθεση ότι όλα τα χαρακτηριστικά ακολουθούν από κοινού γκαουσιανή κατανομή. Γκαουσιανές είναι και η a-priori και η a-posteriori κατανομές. Αυτό μπορεί να είναι κάπως περιοριστικό αλλά στην πραγματικότητα αυτή η τεχνική προσφέρει πολλές δυνατότητες. Όμως, το αρνητικό είναι ότι είναι αρκετά υπολογιστικά ακριβή. Στο παρακάτω σχήμα φαίνεται ένα

παράδειγμα παλινδρόμησης με γκαουσιανή διεργασία, όπου με κόκκινο χρώμα έχει σχεδιαστεί η πραγματική συνάρτηση που στοχεύουμε να εκτιμήσουμε, με μαύρο χρώμα είναι η συνάρτηση παλινδρόμησης, δηλαδή η συνάρτηση μέση τιμής βάσει της οποίας πραγματοποιούνται οι προβλέψεις και η γκριζαρισμένη περιοχή αντιστοιχεί στην αβεβαιότητα αυτής της συνάρτησης και οριοθετεί ένα διάστημα εμπιστοσύνης γύρω από αυτή.

3.3 Αξιολόγηση μοντέλων παλινδρόμησης

Στην προηγούμενη ενότητα μελετήσαμε διάφορες μεθόδους και αλγορίθμους με τους οποίους μπορεί να αντιμετωπιστεί ένα πρόβλημα παλινδρόμησης εκπαιδεύοντας το κατάλληλο μοντέλο. Σε αυτήν, θα ασχοληθούμε με την αξιολόγησή τους και τι θα πρέπει να κάνουμε για να είμαστε σε θέση να τις συγκρίνουμε και να μπορέσουμε να αποφανθούμε ποια επιστρέφει τα καλύτερα αποτελέσματα σε ένα συγκεκριμένο πρόβλημα.

3.3.1 Διαδικασίες αξιολόγησης

Η αξιολόγηση των μοντέλων παλινδρόμησης είναι ο ακρογωνιαίος λίθος στη φάση ελέγχου του μοντέλου, όπου πραγματικά ελέγχεται η απόδοσή του στα δεδομένα ελέγχου. Όμως, συχνά χρειάζεται να αξιολογήσουμε το μοντέλο αρκετά νωρίτερα στη φάση εκπαίδευσης και μάλιστα επαναλαμβανόμενα ώστε να αποφανθούμε ποια παραμετροποίηση του αλγορίθμου οδηγεί στα καλύτερα αποτελέσματα και πρέπει να επιλεγεί για να πραγματοποιήσει τις τελικές προβλέψεις.

3.3.1.1 Έλεγχος

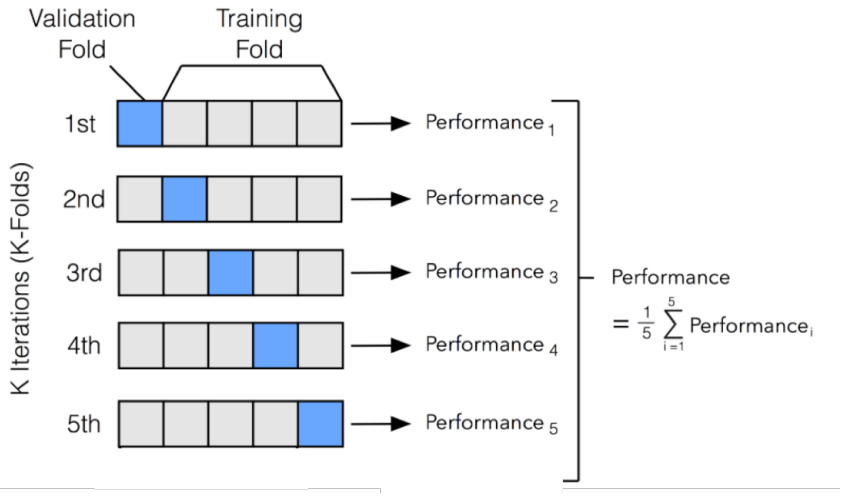
Η φάση ελέγχου (testing) είναι απαραίτητη ακόμα και στην πιο απλή περίπτωση. Κατά τη διάρκειά της, το μοντέλο λαμβάνει ως είσοδο νέα δεδομένα, που δεν έχει ξανασυναντήσει και καλείται να παράξει εκτιμήσεις για αυτά. Τα νέα δεδομένα εξάγονται από το σύνολο δεδομένων ελέγχου, που είναι το ένα από τα δύο ξένα υποσύνολα που προέκυψαν από την αρχική διάσπαση του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Το σύνολο εκπαίδευσης έχει ήδη χρησιμοποιηθεί στη φάση της εκπαίδευσης, στην οποία απαγορεύεται ρητά να εκμεταλλευτεί οποιαδήποτε πληροφορία από το σύνολο ελέγχου για να μην έχουμε διαρροή δεδομένων (data leakage) και συνεπώς να είμαστε βέβαιοι ότι το μοντέλο μας έχει προβλεπτικές ικανότητες.

3.3.1.2 Επικύρωση

Στην περίπτωση που για να χτιστεί το μοντέλο στη φάση της εκπαίδευσης χρησιμοποιείται κάποιος αλγόριθμος με παραμέτρους ή υπερπαραμέτρους, οι τιμές των οποίων πρέπει να βελτιστοποιηθούν, χρειάζεται να αξιολογούμε το μοντέλο και στη φάση της εκπαίδευσης. Όπως εξηγήσαμε προηγουμένως, δε γίνεται αυτή η αξιολόγηση να γίνει με βάση τα δεδομένα του συνόλου ελέγχου, καθώς θα είχαμε διαρροή δεδομένων αλλά ούτε και με βάση τα δεδομένα του συνόλου εκπαίδευσης ή υποσύνολο αυτών, καθώς τότε δε θα αξιολογούσαμε το μοντέλο μας με άγνωστα μέχρι στιγμής δεδομένα. Η πιο απλοϊκή λύση είναι να διασπάσουμε στην αρχή το σύνολο δεδομένων μας σε τρία ξένα υποσύνολα, το σύνολο εκπαίδευσης (training set), το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου (testing set). Τα σύνολα εκπαίδευσης και ελέγχου έχουν ακριβώς την ίδια χρησιμότητα που αναφέραμε παραπάνω ενώ το σύνολο επικύρωσης χρησιμοποιείται για την αξιολόγηση των μοντέλων στη φάση της εκπαίδευσης ώστε να επιλεγεί το καλύτερο για να κάνει τις προβλέψεις. Αυτή η αντιμετώπιση, όμως, έχει ένα πολύ βασικό μειονέκτημα. Θυσιάζεται ένας αξιοσημείωτος όγκος δεδομένων που θα μπορούσε να αξιοποιηθεί στην εκπαίδευση και να βοηθήσει στην ανακάλυψη της δομής των δεδομένων επηρεάζοντας θετικά τη διαμόρφωση του τελικού μοντέλου.

3.3.1.3 Διασταυρούμενη επικύρωση

Η διασταυρούμενη επικύρωση (cross validation) [64] είναι μια τεχνική η οποία επιλύει το τελευταίο πρόβλημα. Με αυτή δε χρειάζεται εξ αρχής να δεσμεύσουμε ένα υποσύνολο των δεδομένων για επικύρωση. Αντ΄ αυτού εκμεταλλευόμαστε επαναληπτικά ένα μικρό ποσοστό των δεδομένων εκπαίδευσης διαφορετικό για κάθε γύρο. Πιο αναλυτικά, τα δεδομένα εκπαίδευσης χωρίζονται σε n τμήματα (folds), τυπικά 5 ή 10 (5-fold ή 10-fold, αντίστοιχα). Όσα τμήματα ορίζονται τόσες επαναλήψεις πραγματοποιούνται και σε καθεμία από αυτές αξιολογείται το ίδιο μοντέλο. Σε κάθε γύρο, χρησιμοποιείται 1 από το n τμήματα για επικύρωση-αξιολόγηση και τα υπόλοιπα $n-1$ για εκπαίδευση. Στο τέλος των n γύρων, αποφαίνεται για την ποιότητα του μοντέλου με βάση κάποιες μετρικές, σύμφωνα με το μέσο όρο των τιμών τους από τους n γύρους. Με αυτόν τον τρόπο, δε θυσιάζουμε πολύτιμα δεδομένα ενώ ταυτόχρονα πετυχαίνουμε να αξιολογήσουμε το μοντέλο μας πάνω σε άγνωστα για αυτό δεδομένα, αποφεύγοντας τη διαρροή δεδομένων.



Σχήμα 3.11: Διασταυρούμενη επικύρωση 5 τμημάτων (πηγή: github)

3.3.2 Μετρικές αξιολόγησης

Για να είναι εφικτή οποιαδήποτε από τις παραπάνω μορφές αξιολόγησης, θα πρέπει πρώτα να ορίσουμε ορισμένες μετρικές αξιολόγησης [49], οι οποίες μετρούν ακριβώς αυτό, την ποιότητα των εκτιμήσεων του μοντέλου.

Σε εργασίες ταξινόμησης, οι μετρικές αξιολόγησης σχετίζονται με το πόσες παρατηρήσεις ταξινομούνται σωστά και πόσες λάθος. Η απλούστερη εξ αυτών είναι η πιστότητα (accuracy) και εκφράζει το πλήθος των σωστά ταξινομημένων δειγμάτων προς το συνολικό πλήθος των προβλέψεων. Άλλες πιο σύνθετες αλλά πάλι προσανατολισμένες στη σχέση σωστά και λανθασμένα ταξινομημένων δειγμάτων είναι η ακρίβεια (precision), η ανάκληση (recall) και το f1 score, ο αρμονικός μέσος όρος ακρίβειας και ανάκλησης. Ιδανικά θέλουμε όλα να είναι υψηλά (κοντά στο 1) όμως η ακρίβεια και η ανάκληση είναι αντικρουόμενες οπότε πρέπει να υπάρξει ένας συμβιβασμός μεταξύ τους. Στη παρούσα εργασία, δεν ασχολούμαστε με πρόβλημα ταξινόμησης οπότε και δε θα επεκταθούμε περαιτέρω.

Σε εργασίες παλινδρόμησης, οι μετρικές αξιολόγησης σχετίζονται με το πόσο καλά η καμπύλη που παρήχθη μοντελοποιεί και προσεγγίζει τα δεδομένα. Παρακάτω αναλύουμε τις βασικότερες μετρικές αξιολόγησης στην περίπτωση της παλινδρόμησης.

3.3.2.1 Μέσο απόλυτο σφάλμα και μέσο απόλυτο ποσοστιαίο σφάλμα

Το μέσο απόλυτο σφάλμα (mean absolute error - MAE) είναι η πιο απλή μετρική που μπορεί να οριστεί και αντιστοιχεί στο μέσο όρο των σφαλμάτων στις προβλέψεις. Ορίζεται ως το άθροισμα των σφαλμάτων ή υπολοίπων (residuals), δηλαδή των απόλυτων διαφορών μεταξύ πραγματικής και εκτιμώμενης από το μοντέλο τιμής εξόδου, διαιρεμένο με το πλήθος των προβλέψεων, πρακτικά το πλήθος των παρατηρήσεων που χρησιμοποιούνται στην αξιολόγηση.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.44}$$

Το μέσο απόλυτο ποσοστιαίο σφάλμα (mean absolute percentage error - MAPE) είναι η κανονικοποιημένη εκδοχή του παραπάνω. Τώρα, οι απόλυτες διαφορές μεταξύ των πραγματικών τιμών εξόδου και των εκτιμήσεών τους εκφράζονται ως ποσοστό των πραγματικών τιμών και ο τύπος παίρνει τη μορφή:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}\cdot 100\% \tag{3.45}$$

3.3.2.2 Μέσο τετραγωνικό σφάλμα και ριζικό μέσο τετραγωνικό σφάλμα

Το μέσο τετραγωνικό σφάλμα (mean squared error - MSE) διαφοροποιείται από το μέσο απόλυτο σφάλμα στο ότι τα υπόλοιπα είναι υψωμένα στο τετράγωνο.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.46}$$

Ο τετραγωνισμός των σφαλμάτων συνεπάγεται την απόδοση πολύ μεγαλύτερης βαρύτητας στα μεγαλύτερα σφάλματα από ό,τι στα μικρότερα. Στη στατιστική, ο δείκτης αυτός ταυτίζεται με τη διασπορά μιας αμερόληπτης εκτιμήτριας.

Το μειονέκτημα του μέσου τετραγωνικού σφάλματος είναι ότι δε μετράται στις μονάδες που βρίσκονται τα αποτελέσματα. Το ριζικό μέσο τετραγωνικό σφάλμα (root mean squared error - RMSE) διορθώνει αυτήν τη διαφορά στις μονάδες, βάζοντας κάτω από το υπόριζο την ποσότητα του μέσου τετραγωνικού σφάλματος.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.47}$$

Στη στατιστική, ο δείκτης αυτός ταυτίζεται με την τυπική απόκλιση μιας αμερόληπτης εκτιμήτριας.

3.3.2.3 Συντελεστής προσδιορισμού

Ο συντελεστής προσδιορισμού R^2 (coefficient of determination) μετρά το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το σύνολο των ανεξάρτητων μεταβλητών. Δεν έχει μονάδες μέτρησης και λαμβάνει τιμές στο εύρος $[0, 1]$. Σε αντίθεση με όλες τις προηγούμενες μετρικές αξιολόγησης που έχουμε παραθέσει, ο συντελεστής προσδιορισμού είναι επιθυμητό να έχει όσο μεγαλύτερη τιμή γίνεται. Εκφράζει το πόσο καλά έχει προσαρμοστεί η καμπύλη παλινδρόμησης στα πραγματικά δεδομένα, δηλαδή το πόσο ταιριάζουν οι εκτιμήσεις του μοντέλου με τις πραγματικές τιμές και αποτελεί μέτρο της συσχέτισης δύο τυχαίων μεταβλητών.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.48}$$

όπου $\bar{y}$ είναι ο μέσος όρος των πραγματικών τιμών εξόδου

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.49}$$

Ιδανικά, ο συντελεστής προσδιορισμού λαμβάνει την τιμή 1 όταν οι εκτιμήσεις που παράγονται από το μοντέλο ταυτίζονται με τις πραγματικές τιμές εξόδου ενώ στη χειρότερη περίπτωση θεωρητικά θα είναι 0, καθώς θα προβλέπεται πάντα η μέση τιμή ανεξαρτήτως των τιμών των χαρακτηριστικών. Πρακτικά, όπως θα δούμε και στα πειράματα, αν το μοντέλο που θα κατασκευαστεί είναι χειρότερο από αυτό το αφελές μοντέλο βάσης, ο συντελεστής προσδιορισμού μπορεί να λάβει και αρνητικές τιμές.

3.3.2.4 Σκορ δικαιολογημένης διακύμανσης

Το σκορ δικαιολογημένης διακύμανσης (explained variance score) εκφράζει το ποσοστό της διακύμανσης των παρατηρήσεων που μπορεί να δικαιολογήσει το μοντέλο. Η μόνη διαφορά με το συντελεστή προσδιορισμού είναι ότι λαμβάνει υπόψη του την κατανομή των σφαλμάτων και πιο συγκεκριμένα συνυπολογίζει τη μέση τιμή τους. Κατ' επέκταση είναι προτιμότερο να χρησιμοποιηθεί σε περιπτώσεις όπου η μέση τιμή των σφαλμάτων δεν είναι μηδενική. Διαφορετικά, αυτές οι δύο μετρικές ταυτίζονται. Ο τύπος της δικαιολογημένης διακύμανσης είναι:

$$\text{explained} \quad \text{variance} = 1 - \frac{\text{Var}\{\mathbf{y} - \hat{\mathbf{y}}\}}{\text{Var}\{\mathbf{y}\}} \tag{3.50}$$

όπου Var η διακύμανση.

3.3.2.5 Ακρίβεια

Τέλος, η ακρίβεια (accuracy) μπορεί να οριστεί ως το αντίθετο του MAPE και να επεκταθεί και σε προβλήματα παλινδρόμησης, πέρα από τα προβλήματα ταξινόμησης όπου συνηθίζεται. Ο τύπος της είναι ο ακόλουθος:

$$\text{accuracy} = (100 - \text{MAPE})\% \tag{3.51}$$

3.3.3 Μετρικές συσχέτισης

Προτού κατασκευάσουμε το μοντέλο παλινδρόμησης, μπορούμε να χρησιμοποιήσουμε διάφορες μετρικές για να εκτιμήσουμε τις συσχετίσεις των δεδομένων και να αξιολογήσουμε κατά πόσο ένα μέγεθος θα μπορούσε να προβλεφθεί ικανοποιητικά από ένα σύνολο χαρακτηριστικών. Με άλλα λόγια, υπάρχουν διάφορες μετρικές συσχέτισης [75], οι οποίες υποδεικνύουν τις σχέσεις μεταξύ των χαρακτηριστικών και των εξαρτημένων μεταβλητών αλλά και των χαρακτηριστικών μεταξύ τους. Αναλόγως τις τιμές αυτών των μετρικών, μπορούμε να κρίνουμε αν έχει νόημα να προσπαθήσουμε να κατασκευάσουμε ένα μοντέλο παλινδρόμησης και ποιες ανεξάρτητες μεταβλητές μπορούν να βοηθήσουν στην κατασκευή του, συνεισφέροντας στις πληροφορίες που αξιοποιεί.

3.3.3.1 Συντελεστής συσχέτισης Pearson

Η μετρική συσχέτισης του Pearson (Pearson's correlation coefficient) ποσοτικοποιεί τη γραμμική συσχέτιση μεταξύ δύο τυχαίων μεταβλητών, είτε αυτές είναι μία ανεξάρτητη και μία εξαρτημένη είτε είναι δύο ανεξάρτητες ή δύο εξαρτημένες. Ορίζεται από τον τύπο:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.52}$$

όπου με cov συμβολίζεται η συνδιακύμανση και με σ η τυπική απόκλιση. Οι τιμές της κυμαίνονται στο εύρος $[-1, 1]$ και όπως γίνεται φανερό, πέρα από την ισχύ της γραμμικής συσχέτισης μπορεί να προσδιορισθεί και η κατεύθυνσή της (θετική ή αρνητική). Συνεπώς, η τιμή -1 λαμβάνεται για έντονη αρνητική γραμμική συσχέτιση, η τιμή 1 για έντονη θετική γραμμική συσχέτιση ενώ η τιμή 0 για πλήρη απουσία συσχέτισης.

3.3.3.2 Συντελεστής συσχέτισης Spearman

Η μετρική συσχέτισης του Spearman (Spearman's correlation coefficient) ελέγχει αν υπάρχει μονότονη σχέση μεταξύ δύο τυχαίων μεταβλητών, είτε αυτές είναι μία ανεξάρτητη και μία εξαρτημένη είτε είναι δύο ανεξάρτητες ή δύο εξαρτημένες. Ορίζεται από τον τύπο:

$$\rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}} \tag{3.53}$$

όπου με cov συμβολίζεται η συνδιακύμανση, με σ η τυπική απόκλιση και οι rg_X, rg_Y είναι οι μεταβλητές κατάταξης (ordinal variables) των αντίστοιχων τυχαίων μεταβλητών X και Y. Όπως μπορούμε να δούμε, ο συντελεστής συσχέτισης του Spearman ταυτίζεται με το συντελεστή συσχέτισης του Pearson για τις μεταβλητές τάξεως, που αφορούν τη διάταξη των τιμών τους και όχι τις ίδιες τις τιμές αυτές καθ' αυτές.

3.4 Τεχνικές προεπεξεργασίας δεδομένων

Παραπάνω αναλύσαμε μεθόδους για την εκπαίδευση και τον έλεγχο μοντέλων παλινδρόμησης. Όμως, οφείλουμε να αναφέρουμε ότι ενδέχεται προτού αρχίσουμε να χτίζουμε ένα μοντέλο να προβούμε σε κάποια προεπεξεργασία των δεδομένων, ώστε αυτά να είναι σε μορφή αξιοποιήσιμη από τον αλγόριθμο εκπαίδευσης [26, 41]. Στη συνέχεια, αναφέρουμε κάποιες από τις πιο συχνές περιπτώσεις όπου τα δεδομένα δεν μπορούν να χρησιμοποιηθούν ως είσοδος στον αλγόριθμο παλινδρόμησης ως έχουν, καθώς και τρόπους με τους οποίους γίνεται να υπερκεραστεί αυτό το εμπόδιο.

3.4.1 Μείωση διαστατικότητας

Η κατάλληλη διαστατικότητα των δεδομένων, δηλαδή το ιδανικό πλήθος χαρακτηριστικών που θα χρησιμοποιηθούν ως είσοδοι στον αλγόριθμο εκπαίδευσης είναι άρρηκτα συνδεδεμένη με το διαθέσιμο αριθμό δειγμάτων. Έστω ότι έχουμε στη διάθεσή μας ένα σύνολο δεδομένων με $n \gg 1$ παρατηρήσεις. Αν ξεκινήσουμε τη μοντελοποίηση με ένα χαρακτηριστικό και σταδιακά αυξάνουμε το πλήθος τους, θα παρατηρήσουμε ότι όσο αυτό παραμένει σημαντικά μικρότερο του n ενώ αυξάνεται, η προβλεπτική ικανότητα του μοντέλου θα βελτιώνεται. Μόλις πάψει να είναι τάξεις μικρότερο ή πολλαπλάσια μικρότερο από το n και συνεχίσει να αυξάνει, η απόδοση του μοντέλου παλινδρόμησης θα αρχίσει να φθίνει. Συμπεραίνουμε ότι ενώ περισσότερα χαρακτηριστικά μπορούν να φανούν χρήσιμα και να αποκαλύψουν νέες πτυχές για τη δομή των δεδομένων και τη συσχέτισή τους, το πλήθος τους πρέπει να είναι σημαντικά μικρότερο των διαθέσιμων παρατηρήσεων για να μπορέσουν να αξιοποιηθούν. Διαφορετικά σύμφωνα με την «κατάρα της διαστατικότητας» (curse of dimensionality), το μοντέλο θα αδυνατεί να γενικεύσει και θα μαθαίνει από θόρυβο.

Για να το αποφύγουμε αυτό, μπορούμε να προβούμε στη μείωση της διαστατικότητας της εισόδου (dimensionality reduction) με διάφορες τεχνικές είτε επιλογής χαρακτηριστικών (feature selection) είτε εξαγωγής χαρακτηριστικών (feature extraction). Στην πρώτη περίπτωση, όπως εξηγείται από το όνομα της τεχνικής, απορρίπτονται κάποια χαρακτηριστικά με βάση ένα κριτήριο ενώ στη δεύτερη περίπτωση μετασχηματίζονται οι τιμές των υπαρχόντων χαρακτηριστικών σε νέες σε ένα χώρο λιγότερων διαστάσεων, πρακτικά εξάγοντας νέα χαρακτηριστικά σε αυτόν το χώρο.

3.4.1.1 Επιλογή χαρακτηριστικών

Ο τρόπος μείωσης της διαστατικότητας σε αυτήν την περίπτωση είναι εμφανής. Επιλέγονται να χρησιμοποιηθούν ως είσοδοι στον αλγόριθμο εκπαίδευσης μόνο κάποια χαρακτηριστικά ενώ τα υπόλοιπα απορρίπτονται σύμφωνα με κάποιο κριτήριο. Το κριτήριο αυτό μπορεί ενδεικτικά να αφορά τη διακύμανση ενός χαρακτηριστικού ή τη συσχέτισή του με την εξαρτημένη μεταβλητή. Πιο αναλυτικά, χαρακτηριστικά τα οποία παρουσιάζουν μηδενική έως και πολύ μικρή διακύμανση στις τιμές τους απορρίπτονται, καθώς εφόσον οι τιμές τους παραμένουν σχεδόν σταθερές ανεξαρτήτως των μεταβολών της τιμής εξόδου που πρόκειται να προβλεφθεί αδυνατούν να επηρεάσουν με κάποιον τρόπο το μοντέλο. Αν τις κρατούσαμε θα υπήρχε ο κίνδυνος της υπερεκπαίδευσης αλλά και η καθυστέρηση και παρακώλυση της διαδικασίας εκπαίδευσης. Επιπλέον, μπορούν να επιλεγούν να χρησιμοποιηθούν τα k καλύτερα χαρακτηριστικά σύμφωνα με κάποια μετρική συσχέτισης, π.χ. τη μετρική συσχέτισης του Pearson που αναλύσαμε παραπάνω, με τη λογική ότι τα χαρακτηριστικά που εμφανίζουν μεγαλύτερη συσχέτιση με τη μεταβλητή εξόδου θα διαδραματίζουν σημαντικότερο ρόλο στην κατασκευή του μοντέλου.

3.4.1.2 Ανάλυση σε κύριες συνιστώσες

Η ανάλυση σε κύριες συνιστώσες (principal components analysis - PCA) αποτελεί χαρακτηριστικό παράδειγμα τεχνικής εξαγωγής νέων χαρακτηριστικών για την επίτευξη της μείωσης της διαστατικότητας. Ουσιαστικά, αυτό πραγματοποιείται με τον επαναπροσδιορισμό των αξόνων αναπαράστασης του συνόλου δεδομένων. Αυτοί οι νέοι άξονες είναι ορθογώνιοι και ορίζουν τη βάση του χώρου χαμηλότερης διαστατικότητας που προκύπτει. Πρόκειται για γραμμικώς ασυσχέτιστες μεταβλητές που προέκυψαν ως γραμμικός συνδυασμός των αρχικών χαρακτηριστικών και ονομάζονται κύριες συνιστώσες (principal components) . Υπολογίζονται και διατηρούνται σε φθίνουσα σειρά διασποράς. Πιο αναλυτικά, πρώτα υπολογίζεται ο πίνακας συνδιακύμανσης των αρχικών χαρακτηριστικών μαζί με τις ιδιοτιμές και τα ιδιοδιανύσματα του. Το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή καθορίζει και τη διεύθυνση στην οποία υπάρχει η μεγαλύτερη διασπορά και συνεπώς η περισσότερη πληροφορία. Τελικά, κρατώντας τα ιδιοδιανύσματα, τους άξονες, που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές, συγκεντρώνουμε το μεγαλύτερο ποσοστό της πληροφορίας. Με αυτόν τον τρόπο είναι δυνατό να επιτευχθεί συμπίεση δεδομένων, καθώς η ίδια ή σχεδόν η ίδια πληροφορία που εμπεριέχεται στο σύνολο των αρχικών χαρακτηριστικών, θα περιέχεται στο σαφώς μικρότερο σύνολο αυτών των αξόνων.

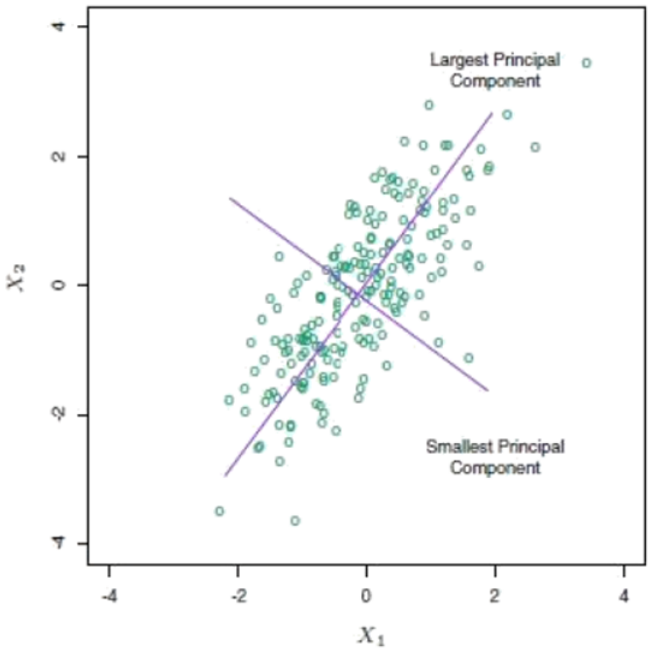
3.4.2 Κανονικοποίηση χαρακτηριστικών

Έχει παρατηρηθεί ότι χαρακτηριστικά με πολύ μεγάλες διαφορές στις απόλυτες τιμές τους μπορούν να προκαλέσουν προβλήματα στην εκπαίδευση και να μην επιτρέψουν στους αλγόριθμους να αποδώσουν βέλτιστα. Για παράδειγμα, ένα χαρακτηριστικό με πολύ μεγάλες τιμές μπορεί να έχει μεγαλύτερη επίδραση στην κατασκευή του μοντέλου από ό,τι ένα με μικρές τιμές, χωρίς αυτό να σημαίνει απαραίτητα ότι είναι περισσότερο καθοριστικό για τη διεργασία παλινδρόμησης που μελετάμε. Με την κανονικοποίηση καθίσταται εφικτή η άμβλυνση αυτών των διαφορών, έπειτα από το μετασχηματισμό των τιμών των χαρακτηριστικών. Παρακάτω παραθέτουμε δύο είδη μετασχηματισμών με αυτόν τον σκοπό.

3.4.2.1 Κλιμάκωση μεγίστου-ελαχίστου

Στην κλιμάκωση μεγίστου-ελαχίστου των χαρακτηριστικών (min-max feature scaling) , οι τιμές των χαρακτηριστικών περιορίζονται στο εύρος $[0, 1]$. Αυτό επιτυγχάνεται με τη διαίρεση των αρχικών τιμών τους με τη διαφορά μέγιστης-ελάχιστης τιμής, όπως υποδεικνύει ο ακόλουθος τύπος:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.54}$$



Σχήμα 3.12: Παράδειγμα ανάλυσης σε κύριες συνιστώσες (πηγή: TIBCO)

Η κλιμάκωση σε $[0, 1]$ είναι λιγότερο ευαίσθητη σε πολύ μικρές αποκλίσεις και η ύπαρξη ενός outlier αρκεί για να συμπιέσει τις υπόλοιπες τιμές στο άλλο άκρο του διαστήματος. Επίσης, στην περίπτωση των αραιών (sparse) διανυσμάτων χαρακτηριστικών, δηλαδή εκείνων που έχουν πολλές μηδενικές τιμές, αυτές οι μηδενικές τιμές διατηρούνται επηρεάζοντας την ταχύτητα της εκπαίδευσης.

3.4.2.2 Τυποποίηση z-score

Στην τυποποίηση z-score χαρακτηριστικών (z-score feature standardization) , οι αρχικές τιμές των χαρακτηριστικών μετατρέπονται σε z-score ή standard score. Πλέον τα χαρακτηριστικά περιγράφονται από το πόσες τυπικές αποκλίσεις απέχουν από τη μέση τιμή τους και δίνονται από τον τύπο:

$$z = \frac{X - \mu}{\sigma} \tag{3.55}$$

Με αυτόν τον τρόπο τα νέα χαρακτηριστικό έχουν μέση τιμή 0 και διακύμανση 1, σαν την κανονική κατανομή. Αξίζει να σημειώσουμε ότι δεν υπάρχει πρόβλημα αν η πραγματική κατανομή των δεδομένων δεν είναι η κανονική. Τέλος, αυτή η τεχνική κανονικοποίησης αντιμετωπίζει πιο αποτελεσματικά τις ακραίες τιμές, καθώς δεν είναι απαραίτητο να συμπιεστούν υπερβολικά όλες οι υπόλοιπες.

3.4.3 Άλλα είδη προεπεξεργασίας δεδομένων

Υπάρχουν και άλλες περιπτώσεις όπου απαιτείται προεπεξεργασία δεδομένων για να μπορέσουν οι αλγόριθμοι εκπαίδευσης να εκτελεστούν αποδοτικά. Για παράδειγμα, κάποιες φορές το σύνολο δεδομένων που διαθέτουμε δεν είναι πλήρες και για κάποιες παρατηρήσεις απουσιάζουν οι τιμές ορισμένων χαρακτηριστικών. Αυτά τα δεδομένα δεν μπορούν να δοθούν ως είσοδος στον αλγόριθμο εκπαίδευσης ως έχουν. Πρέπει να απορριφθούν εξ ολοκλήρου οι αντίστοιχες παρατηρήσεις ή τα αντίστοιχα χαρακτηριστικά. Διαφορετικά, για να μη χαθούν χρήσιμες πληροφορίες, οι απουσιάζουσες τιμές μπορούν να εκτιμηθούν από στατιστικά μεγέθη των αντίστοιχων χαρακτηριστικών, όπως η μέση τιμή ή η πιο συχνή.

Επίσης, αν το σύνολο δεδομένων περιέχει κατηγορικά χαρακτηριστικά, αυτά πρέπει να μετατραπούν κατάλληλα ώστε τελικά στον αλγόριθμο εκπαίδευσης να δοθούν στην είσοδο μόνο χαρακτηριστικά με αριθμητικές τιμές. Τέλος, σε προβλήματα κατηγοριοποίησης έχει παρατηρηθεί ότι οι ταξινομητές αποδίδουν καλύτερα όταν ο πληθικός αριθμός των διαφορετικών κλάσεων είναι περίπου ο ίδιος. Στην πράξη όμως, συχνά συναντάμε σύνολα δεδομένων όπου κάποιες κλάσεις εμφανίζονται πολύ πιο σπάνια από κάποιες άλλες, οπότε και το σύνολο δεδομένων είναι μη ισορροπημένο. Τότε είναι απαραίτητο να προβούμε σε εξισορρόπηση του συνόλου δεδομένων πραγματοποιώντας είτε υποδειγματοληψία στα δείγματα των πιο συχνών κλάσεων είτε υπερδειγματοληψία, επαναλαμβάνοντας παρατηρήσεις των πιο σπάνιων κλάσεων.

3.5 Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας (natural language processing - NLP) [37, 47] είναι ένας διεπιστημονικός κλάδος που βρίσκεται στην τομή της υπολογιστικής γλωσσολογίας, της τεχνητής νοημοσύνης και της τεχνολογίας πληροφορίας. Συνδέεται άμεσα με την αλληλεπίδραση ανθρώπου-υπολογιστή (human-computer interaction) στα πλαίσια της κατανόησης και της παραγωγής φυσικής (ανθρώπινης) γλώσσας από τους υπολογιστές. Τα πεδία έρευνας με τα οποία ασχολείται μπορούν να χωριστούν σε υποκατηγορίες, με διάφορες εργασίες επικαλυπτόμενες μερικές φορές μεταξύ τους.

Η πρώτη από αυτές αφορά τη μορφολογία και τη σύνταξη (morphosyntax) και περιλαμβάνει γραμματική και συντακτική ανάλυση (parsing και grammar induction), λημματοποίηση και αποκατάληξη (lemmatization και stemming), όπου στην πρώτη προσδιορίζεται η ρίζα μιας λέξης ενώ στη δεύτερη αφαιρείται η κατάληξη της σε μια προσπάθεια προσέγγισης της ρίζας της, επισήμανση μερών του λόγου, π.χ. ουσιαστικό, ρήμα, επίθετο κλπ. (part-of-speech tagging), διάσπαση σε προτάσεις, λέξεις και μορφήματα.

Η δεύτερη κατηγορία αφορά τη σημασιολογία (semantics) και σχετίζεται με την εξαγωγή πληροφορίας από κείμενα. Πιο συγκεκριμένα, μελετάται η αναγνώριση ονοματισμένων οντοτήτων, π.χ. οργανισμών, κύριων ονομάτων ανθρώπων, χωρών κ.ά. (named entity regognition - NER) και εξαγωγή σχέσεων μεταξύ τους (relationship extraction), η ανάλυση συναισθήματος (sentiment analysis) και η εξαγωγή ορολογιών (terminology extraction) μεταξύ άλλων. Φυσικά η σημασιολογία συνδέεται άμεσα με τη νόηση (cognition), δηλαδή τη διαδικασία απόκτησης γνώσης και κατανόησης μέσω σκέψης, εμπειρίας και ερεθισμάτων. Ένα παράδειγμα αυτής αποτελεί η κατανόηση μεταφορών, η κατανόηση δηλαδή μιας έννοιας σε ένα διαφορετικό πεδίο από αυτό που χρησιμοποιείται συνήθως (word sense disambiguation).

Επίσης, η επεξεργασία φυσικής γλώσσας ασχολείται με εργασίες λόγου (discourse), οι οποίες πάλι αφορούν τη σημασιολογία αλλά σε υψηλότερο επίπεδο, σε επίπεδο κειμένου. Αυτή η κατηγορία περιλαμβάνει την επίλυση σχέσεων συναναφοράς (coreference resolution), όπου εντοπίζεται ποιες λέξεις αναφέρονται στις ίδιες οντότητες σε ένα κομμάτι κειμένου, την αυτόματη εξαγωγή θέματος και την κατάτμηση ενός κειμένου σε θεματικές ενότητες (topic extraction and segmentation) καθώς και την ανάλυση λόγου (discourse analysis), η οποία αναφέρεται στην αναγνώριση της δομής του λόγου, όπως η φύση των σχέσεων του λόγου μεταξύ δύο προτάσεων π.χ. επεξήγηση, αντίθεση κλπ.

Τέλος, ορίζονται εργασίες ακόμα υψηλότερου επιπέδου, όπως η αυτόματη μετάφραση ενός κειμένου από μία φυσική γλώσσα σε άλλη (machine translation), η παραγωγή φυσικής γλώσσας (natural language generation), η αυτόματη ερωταπόκριση (question answering), η αυτόματη παραγωγή περιλήψεων (automatic summarization) και η αυτόματη παραγωγή ολόκληρων μυθιστορημάτων.

Αξίζει να σημειωθεί πως ο κλάδος της επεξεργασίας φυσικής γλώσσας ασχολείται και με τον προφορικό λόγο, την ομιλία (speech), πέρα από τον γραπτό που αναλύθηκε παραπάνω. Πιο συγκεκριμένα, περιλαμβάνει την αναγνώριση ομιλίας από ηχητικό κλιπ (speech recognition), την κατάτμηση ομιλίας πάλι από ηχητικό κλιπ (speech segmentation) και τη μετατροπή κειμένου σε

ομιλία. Επίσης, είναι δυνατή η οπτική αναγνώριση χαρακτήρων (optical character recognition) από κειμενική πληροφορία ενσωματωμένη σε εικόνες.

3.5.1 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος ή εξόρυξη γνώμης (sentiment analysis or opinion mining) [44] συνιστά την αυτοματοποιημένη διαδικασία εξαγωγής πληροφοριών για την συναισθηματική πολικότητα ενός σώματος κειμένου. Όπως αναφέραμε και προηγουμένως, εντάσσεται στη σημασιολογική ενότητα της επεξεργασίας φυσικής γλώσσας και αποτελεί ειδική περίπτωση της κατηγοριοποίησης κειμένων (text categorization/classification), με τις κλάσεις αυτή τη φορά να εκφράζουν συναίσθημα. Πιο αναλυτικά, στόχος είναι ο προσδιορισμός της συναισθηματικής κατάστασης ενός ατόμου κατά τη συγγραφή του υπό μελέτη κειμένου ή του συναισθήματος που σκόπιμα επιχειρεί να μεταδώσει στον αναγνώστη ή την άποψή του πάνω σε ένα θέμα που πραγματεύεται στο εν λόγω κείμενο. Οι κατηγορίες, λοιπόν, μπορούν να αντιστοιχούν σε ένα συγκεκριμένο συναίσθημα, όπως χαρά, λύπη, θυμός ακόμα και ειρωνεία ή γενικότερα σε θετικό, αρνητικό και ουδέτερο πρόσημο. Μάλιστα, μπορεί να οριστούν επιπλέον επίπεδα πολικότητας (polarity), όπως μάλλον θετικό, μάλλον αρνητικό κλπ.

Επίσης, το συναίσθημα ή η πολικότητα είναι δυνατό να μελετηθεί σε διάφορες εκτάσεις του κειμένου που εξετάζεται. Η συναισθηματική ανάλυση, λοιπόν, μπορεί να πραγματοποιηθεί σε επίπεδο κειμένου-εγγράφου (document level), σε επίπεδο πρότασης (sentence level) και σε επίπεδο χαρακτηριστικών (feature level), όπου εξετάζονται μεμονωμένες φράσεις οι οποίες αναφέρονται σε χαρακτηριστικά μιας οντότητας ή ενός θέματος.

Σε όποιο επίπεδο και αν μελετάται υπάρχουν δύο βασικές προσεγγίσεις για την επίτευξη της εξόρυξης γνώμης των χρηστών. Η πρώτη βασίζεται σε προκαθορισμένα λεξικά (lexicon based), στα οπαία περιέχονται διάφορες λέξεις μιας γλώσσας και η συναισθηματική αποτίμηση αυτών. Σε αυτήν την περίπτωση, αποδίδονται στις λέξεις του κειμένου που μελετάται τα συναισθηματικά τους σκορ σύμφωνα με ένα λεξικό και η τελική αποτίμηση του κειμένου προκύπτει από το άθροισμα ή τον μέσο όρο αυτών. Αυτή η προσέγγιση είναι πολύ εύκολο να εφαρμοστεί ωστόσο συνήθως εισάγει τον περιορισμό του ότι οι λέξεις έχουν μία στατική προκαθορισμένη αποτίμηση ανεξαρτήτως του γενικότερου πλαισίου στο οποίο γράφονται. Η δεύτερη προσέγγιση είναι η εκπαίδευση ενός ταξινομητή με μεθόδους μηχανικής μάθησης. Αν και με αυτόν τον τρόπο μπορούν να ληφθούν πιο αξιόλογα αποτελέσματα, για να είναι δυνατή η εφαρμογή του απαιτείται σημαντική προετοιμασία. Πρώτον, πρέπει να έχει προηγηθεί η επισήμανση των παρατηρήσεων (είτε κειμένων είτε φράσεων) με την πολικότητα ή το συναίσθημα με το οποίο είναι επιφορτισμένα. Στον πραγματικό κόσμο, συναντώνται πολύ λίγα χαρακτηρισμένα σύνολα δεδομένων, συνεπώς τις περισσότερες φορές η διαδικασία αυτή χρειάζεται να γίνει χειροκίνητα από τους ερευνητές και είναι ιδιαίτερα χρονοβόρα. Επίσης, στους αλγορίθμους εκπαίδευσης δεν μπορούν να δοθούν ως είσοδος τα κείμενα ως έχουν και είναι απαραίτητη η εξαγωγή χαρακτηριστικών από αυτά σε ένα διανυσματικό χώρο.

Κεφάλαιο 4

Τεχνικό υπόβαθρο

4.1 Γλώσσα προγραμματισμού Python

Η Python[1] είναι μία διερμηνευμένη (interpreted), γενικού σκοπού (general-purpose) και υψηλού επιπέδου (high-level) γλώσσα προγραμματισμού. Δημιουργήθηκε από τον Guido van Rossum και κυκλοφόρησε για πρώτη φορά το 1991. Αναπτύσσεται ως ανοιχτό λογισμικό (open source) και διαχειρίζεται από το μη κερδοσκοπικό οργανισμό Python Software Foundation. Το όνομα της προέρχεται από τους Monty Python (ομάδα Άγγλων κωμικών) και όχι από το φίδι-πύθωνα, όπως προδιαθέτει το λογότυπό της. Σήμερα χρησιμοποιείται κυρίως η Python 3 η οποία δεν είναι πλήρως συμβατή προς τα πίσω. Συχνά, λοιπόν, χρησιμοποιείται και η Python 2, η οποία υποστηρίζεται από την κοινότητα εξίσου με την Python 3.

Υποστηρίζει πολλά προγραμματιστικά υποδείγματα (programming paradigms), όπως προστακτικό (imperative), διαδικαστικό (procedural), αντικειμενοστραφές (object-oriented) και συναρτησιακό (functional). Είναι δυναμική γλώσσα προγραμματισμού (dynamically typed) και υποστηρίζει συλλογή απορριμμάτων (garbage collection). Τα κύρια χαρακτηριστικά της είναι η αναγνωσιμότητα του κώδικα και η ευκολία χρήσης της. Το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ό,τι σε γλώσσες όπως η C++ και η Java. Αυτό οφείλεται σε ένα βαθμό στην εκμετάλλευση της στοίχισης και των κενών χαρακτήρων για το διαχωρισμό των συντακτικών δομών του προγράμματος. Στις περισσότερες γλώσσες προγραμματισμού αυτός ο διαχωρισμός πραγματοποιείται με ειδικούς χαρακτήρες, όπως αγκύλες και φαίνεται πιο σύνθετος στο ανθρώπινο μάτι. Επίσης, οι διάφορες λειτουργίες που παρέχει ορίζονται από τον προγραμματιστή συνήθως με ολόκληρες αγγλικές λέξεις ή συντομογραφίες αυτών και σε ελάχιστες περιπτώσεις απαιτούνται ειδικά σύμβολα. Για αυτούς τους λόγους χαρακτηρίζεται ως εύκολη στην εκμάθηση και καθιστά τον κώδικα που γράφεται σε αυτήν ιδιαίτερα ευανάγνωστο. Τέλος, διαθέτει πολλές βιβλιοθήκες που διευκολύνουν αρκετές συνηθισμένες εργασίες [60]. Μερικές από αυτές χρησιμοποιήθηκαν στην παρούσα μελέτη και αναλύονται πιο ενδελεχώς παρακάτω.

4.1.1 Βιβλιοθήκες που χρησιμοποιήθηκαν στα πειράματα

4.1.1.1 pandas

Η βιβλιοθήκη ανοιχτού λογισμικού pandas[2] χρησιμοποιείται για τη διαχείριση και την ανάλυση δεδομένων στη γλώσσα προγραμματισμού Python. Το όνομά της προέρχεται από τον όρο "panel data", που χρησιμοποιείται στην οικονομετρία για να περιγράψει σύνολα δεδομένων με παρατηρήσεις σε διάφορα χρονικά διαστήματα. Δημιουργήθηκε από το Wes McKinney και κυκλοφόρησε πρώτη φορά το 2008.

[1] https://www.python.org/
[2] https://pandas.pydata.org/

Η συγκεκριμένη βιβλιοθήκη επιτρέπει την εισαγωγή δεδομένων από διάφορες μορφές αρχείων, όπως csv (comma-separated values), JSON, υπολογιστικά φύλλα Excel και βάσεις δεδομένων SQL και προσφέρει δομές δεδομένων για την αποθήκευσή τους, με κυριότερη αυτή των data frames, οι οποίες συνοδεύονται από διάφορες λειτουργίες για τη διαχείριση τους. Τα data frames μοιάζουν με δισδιάστατους πίνακες των οποίων οι στήλες είναι πιθανό να περιέχουν διαφορετικούς τύπους δεδομένων. Στην πράξη, είθισται κάθε γραμμή να αντιστοιχεί σε μια παρατήρηση του συνόλου δεδομένων και κάθε στήλη σε ένα χαρακτηριστικό. Διατρέχουν ιδιαίτερα αποδοτικά τα δεδομένα που είναι αποθηκευμένα σε αυτά και κατ' επέκταση είναι βολικά για εφαρμογές μεγάλης κλίμακας και προβλήματα μεγάλων δεδομένων (big data). Επίσης, υλοποιούν διάφορες συνήθεις λειτουργίες, όπως συγχώνευση, καθάρισμα δεδομένων, αλλαγή διαστάσεων κ.ά.

4.1.1.2 NumPy

Η NumPy[3] είναι μία βιβλιοθήκη ανοιχτού λογισμικού για την προγραμματιστική γλώσσα Python. Υποστηρίζει μεγάλους πολυδιάστατους πίνακες και μαθηματικές συναρτήσεις υψηλού επιπέδου που μπορούν να εφαρμοστούν σε αυτούς. Δημιουργήθηκε από τον Travis Oliphant το 2005 από τη συνένωση δύο προϋπαρχόντων ανταγωνιστικών βιβλιοθηκών για αυτόν τον σκοπό.

4.1.1.3 SciPy

Η SciPy[4] είναι μία βιβλιοθήκη ανοιχτού λογισμικού που χρησιμοποιείται για επιστημονικούς υπολογισμούς. Δημιουργήθηκε από τους Travis Oliphant, Pearu Peterson και Eric Jones και κυκλοφόρησε για πρώτη φορά το 2001. Ως βασική δομή δεδομένων χρησιμοποιεί τους πολυδιάστατους πίνακες της NumPy και επεκτείνει τις συναρτήσεις που η τελευταία παρέχει για αυτούς. Μπορεί να χρησιμοποιηθεί για διάφορες συνήθεις εργασίες στα πεδία της επιστήμης και της μηχανικής, όπως η βελτιστοποίηση, η γραμμική άλγεβρα, ο ολοκληρωτικός λογισμός, η παρεμβολή, ο μετασχηματισμός Fourier, η επεξεργασία σημάτων και εικόνας και η επίλυση διαφορικών εξισώσεων. Τέλος, με τον όρο SciPy μπορούμε να αναφερθούμε σε ένα ευρύτερο οικοσύστημα ανοιχτού λογισμικού με τον ίδιο προσανατολισμό, το οποίο πέρα από τις βιβλιοθήκες NumPy και SciPy περιλαμβάνει και άλλα πακέτα, όπως pandas, Matplotlib, SymPy και IPython.

4.1.1.4 scikit-learn

Η scikit-learn[5] είναι μια βιβλιοθήκη ανοιχτού λογισμικού για τη γλώσσα προγραμματισμού Python. Αποτελεί μία από τις πιο δημοφιλείς βιβλιοθήκες για μηχανική μάθηση και στατιστική μοντελοποίηση, καθώς προσφέρει μια πλούσια φαρέτρα εργαλείων για εργασίες ταξινόμησης, παλινδρόμησης, συσταδοποίησης και μείωσης της διαστατικότητας [63]. Ενδεικτικά, προσφέρει πληθώρα υλοποιημένων αλγορίθμων μηχανικής μάθησης, όπως μηχανές διανυσμάτων υποστήριξης, δέντρα απόφασης, τυχαία δάση, κοντινότεροι γείτονες, ομαδοποίηση k μέσων και πολλούς άλλους. Δημιουργήθηκε από τον David Cournapeau το 2007, αρχικά στα πλαίσια του Google Summer of Code και έκτοτε χρησιμοποιείται ευρέως για τους παραπάνω σκοπούς. Σε αυτό συνέβαλε και το γεγονός ότι εμφανίζει διαλειτουργικότητα με τις πολύ διαδεδομένες βιβλιοθήκες NumPy και SciPy.

4.1.1.5 Matplotlib

Η Matplotlib[6] είναι μια βιβλιοθήκη ανοιχτού λογισμικού για τη γλώσσα προγραμματισμού Python, η οποία ασχολείται με την οπτικοποίηση δεδομένων και την παραγωγή γραφικών παραστάσεων. Δημιουργήθηκε αρχικά από τον John D. Hunter το 2003 και πλέον αποτελεί μία

[3] https://numpy.org/
[4] https://www.scipy.org/
[5] https://scikit-learn.org/stable/
[6] https://matplotlib.org/

από τις πιο δημοφιλείς βιβλιοθήκες σχεδίασης. Παρέχει διάφορων ειδών διαγράμματα, όπως line plots, histograms, scatter plots, 3D plots, image plots, contour plots, polar plots κ.ά. Επίσης, περιλαμβάνει τη συλλογή συναρτήσεων pyplot, η οποία προσφέρει μια διεπαφή όμοια με αυτή του Matlab.

4.1.1.6 seaborn

Η seaborn[7] αποτελεί επίσης μία βιβλιοθήκη οπτικοποίησης δεδομένων για τη γλώσσα προγραμματισμού Python, η οποία μάλιστα βασίζεται στη matplotlib που αναλύσαμε παραπάνω. Τα επιπλέον βασικά στοιχεία που έχει να προσφέρει σε σχέση με τη matplotlib είναι η προγραμματιστική διεπαφή με υψηλού επιπέδου συναρτήσεις για συνήθεις τύπους διαγραμμάτων, το ευκολότερο συντακτικό, η δυνατότητα παραγωγής των ίδιων γραφικών παραστάσεων με λιγότερες γραμμές κώδικα και η ενσωμάτωση της λειτουργικότητας των δομών δεδομένων της βιβλιοθήκης pandas και ειδικότερα των data frames. Αν και η ίδια η βιβλιοθήκη matplotlib έχει κάνει προσπάθειες προς αυτές τις κατευθύνσεις τα τελευταία χρόνια, η χρησιμότητα της seaborn παραμένει αδιαφιλονίκητης αξίας.

4.1.1.7 math

Η βιβλιοθήκη math[8] της Python καθιστά προσβάσιμες συνήθεις μαθηματικές συναρτήσεις και σταθερές. Δίνει τη δυνατότητα άμεσης χρήσης και ενσωμάτωσης αυτών σε πιο σύνθετους μαθηματικούς υπολογισμούς σε κώδικα γραμμένο στη γλώσσα προγραμματισμού Python. Οι συναρτήσεις παρέχονται όπως ακριβώς έχουν οριστεί από τα πρότυπα της C. Τέλος, αξίζει να σημειώσουμε ότι η βιβλιοθήκη math αποτελεί ενσωματωμένη μονάδα της Python και συνεπώς δεν απαιτείται κάποια εγκατάσταση προτού χρησιμοποιηθεί.

4.1.1.8 spaCy

Η spaCy[9] είναι μια βιβλιοθήκη ανοιχτού λογισμικού που υλοποιεί και διευκολύνει εργασίες επεξεργασίας φυσικής γλώσσας στη γλώσσα προγραμματισμού Python. Αρχικός συγγραφέας υπήρξε ο Matthew Honnibal, συνιδρυτής της εταιρείας λογισμικού Explosion AI και η πρώτη κυκλοφορία της πραγματοποιήθηκε το 2015. Διαφοροποιείται από την ευρέως χρησιμοποιούμενη στην εκπαίδευση και την έρευνα NLTK και είναι εμφανώς στοχευμένη στον τομέα της παραγωγής.

Παρέχει προεκπαιδευμένα μοντέλα για εργασίες επισήμανσης μερών του λόγου, κατηγοριοποίησης κειμένων, αναγνώρισης ονοματισμένων οντοτήτων και γραμματικής ανάλυσης σε 15 διαφορετικές γλώσσες. Ενδεικτικά, στις υποστηριζόμενες γλώσσες συγκαταλέγονται τα Αγγλικά, Γαλλικά, Γερμανικά, Ιταλικά, Κινέζικα, Ισπανικά και Ελληνικά. Επίσης, είναι διαθέσιμο ένα πολυγλωσσικό μοντέλο. Η κατάτμηση ενός κειμένου σε προτάσεις, λέξεις και μορφήματα υποστηρίζεται σε ακόμα περισσότερες γλώσσες και φυσικά πάντα ο χρήστης μπορεί να εκπαιδεύσει ένα δικό του μοντέλο. Τέλος, η spaCy είναι ικανή να διαχειριστεί μεγάλο όγκο δεδομένων και να προετοιμάσει κείμενα για εφαρμογές βαθιάς μάθησης (deep learning). Εμφανίζει διαλειτουργικότητα με πολλές βιβλιοθήκες βαθιάς μάθησης, όπως οι TensorFlow, PyTorch, scikit-learn, Gensim αλλά και με το υπόλοιπο οικοσύστημα τεχνητής νοημοσύνης της Python.

4.1.1.9 datetime, dateutil, pytz, time, calendar

Στους αρχέγονους τύπους της Python δεν περιλαμβάνεται κάποιος τύπος για την περιγραφή και την αποθήκευση ημερομηνίας και ώρας. Σε ένα πρόγραμμα όμως συχνά καλούμαστε

[7] `https://seaborn.pydata.org/`
[8] `https://docs.python.org/3.0/library/math.html`
[9] `https://spacy.io/`

να διαχειριστούμε ημερολογιακά και ωρολογιακά δεδομένα. Οι μονάδες datetime[10], time[11] και calendar[12] καλύπτουν ακριβώς αυτό το κενό. Παρέχουν αποδοτικές αναπαραστάσεις χρονικών στιγμιοτύπων και στοιχειώδεις λειτουργίες πάνω σε αυτά. Επίσης, η dateutil[13] επεκτείνει τις λειτουργικότητες της datetime και η pytz[14] εισάγει πληροφορίες για τις διάφορες ζώνες ώρας και επιτρέπει μετατροπές μεταξύ αυτών.

4.1.1.10 Άλλες βιβλιοθήκες

Σε αυτό το σημείο συνοψίζουμε κάποιες βιβλιοθήκες που χρησιμοποιήθηκαν για την επίλυση τεχνικών θεμάτων. Αρχικά, η βιβλιοθήκη csv[15], όπως μας προϊδεάζει το όνομά της, παρέχει τη δυνατότητα ανάγνωσης και αποθήκευσης δεδομένων από ή σε, αντίστοιχα, αρχεία τύπου csv (comma-separated values). Η βιβλιοθήκη re[16] αφορά τις κανονικές εκφράσεις (regular expressions - regex) και τη διαχείρισή τους. Ως γνωστόν με μια κανονική έκφραση ορίζεται ένα σύνολο συμβολοσειρών που ικανοποιούν κάποιους κανόνες/περιορισμούς. Η συγκεκριμένη βιβλιοθήκη, λοιπόν, χρησιμεύει για την κατασκευή κανονικών εκφράσεων και τον έλεγχο ταιριάσματος μιας συμβολοσειράς με κάποια από αυτές.

Επίσης, η βιβλιοθήκη functools[17] χρησιμοποιείται για την κατασκευή και το χειρισμό υψηλότερης τάξης συναρτήσεων, δηλαδή συναρτήσεων που επιστρέφουν ή δρουν πάνω σε άλλες συναρτήσεις ή γενικότερα καλέσιμα αντικείμενα. Δίνει τη δυνατότητα στις συναρτήσεις υψηλότερης τάξης να χρησιμοποιήσουν ή να επεκτείνουν συναρτήσεις χαμηλότερης τάξης, χωρίς οι τελευταίες να ξαναγραφούν εξ ολοκλήρου. Τέλος, η βιβλιοθήκη itertools[18] χρησιμοποιείται για το συνδυασμό απλών επαναληπτών και τη δημιουργία πιο σύνθετων, οι οποίοι εκτελούν επαναληπτικές διαδικασίες γρήγορα και με αποδοτικό όσον αφορά τη μνήμη τρόπο.

4.2 Εργαλεία συλλογής δεδομένων

Σε αυτήν την ενότητα παραθέτουμε κάποιους τρόπους με τους οποίους μπορούν να συλλεχθούν δεδομένα από τις πλατφόρμες που θα χρησιμοποιήσουμε στην παρούσα εργασία, το μέσο κοινωνικής δικτύωσης Twitter και τη μηχανή αναζήτησης της Google.

4.2.1 Twitter API

Καταρχάς, το Twitter, όπως και τα περισσότερα μέσα κοινωνικής δικτύωσης, προσφέρει μια διεπαφή μέσα από την οποία ένας προγραμματιστής μπορεί να αξιοποιήσει μέρος των υπηρεσιών της πλατφόρμας και να ενσωματώσει δεδομένα δημόσια κοινοποιημένα σε αυτή στην εφαρμογή του [86]. Εν ολίγοις, αυτή η διεπαφή προγραμματισμού εφαρμογών (application programming interface - API) δίνει στον προγραμματιστή τις εξής δυνατότητες μέσω αντίστοιχων endpoints:

- Search API: Αναζήτηση και ανάκτηση tweets που δημοσιεύθηκαν την τελευταία εβδομάδα με βάση κάποιο ερώτημα (query), στο οποίο μπορεί να καθορίζεται ο χρήστης που τα δημοσίευσε, τα hashtags που περιέχουν, η χωροχρονική στάμπα της δημοσίευσης τους, το μέγιστο πλήθος αποτελεσμάτων που είναι επιθυμητό να επιστραφούν, η γλώσσα στην οποία έχουν γραφεί κ.ά. Τα αποτελέσματα επιστρέφονται σε JSON και πέρα από το περιεχόμενο των tweets συμπεριλαμβάνουν και μεταδεδομένα, όπως στοιχεία του χρήστη που πραγματοποίησε τη δημοσίευση, η χρονική στιγμή και το γεωγραφικό στίγμα της δημοσίευσης και

[10] https://docs.python.org/3/library/datetime.html
[11] https://docs.python.org/3/library/time.html
[12] https://docs.python.org/3/library/calendar.html
[13] https://pypi.org/project/python-dateutil/
[14] https://pypi.org/project/pytz/
[15] https://docs.python.org/3/library/csv.html
[16] https://docs.python.org/3/library/re.html
[17] https://docs.python.org/3/library/functools.html
[18] https://docs.python.org/3/library/itertools.html

το πλήθος των αναδημοσιεύσεων και των favorites που συγκέντρωσε μεταξύ άλλων. Αυτά τα δεδομένα μπορούν να αναλυθούν περαιτέρω με στατιστικές μεθόδους και να δώσουν διαφωτιστικά αποτελέσματα στα οποία θα στηριχτούν οι επιχειρήσεις στις διαδικασίες λήψης αποφάσεων. Η συγκεκριμένη λειτουργία, λοιπόν, είναι αυτή που μας ενδιαφέρει και έχουμε σκοπό να εκμεταλλευτούμε για την παρούσα εργασία. Ωστόσο, ο περιορισμός των 7 ημερών δε μας επιτρέπει να ανακτήσουμε τα δεδομένα που θέλουμε, όπως θα δούμε στη συνέχεια. Επίσης, υπάρχει άνω όριο στα αιτήματα (requests) που μπορούμε να υποβάλουμε (180 αιτήματα ανά χρήστη και 450 αιτήματα ανά εφαρμογή για κάθε χρονικό παράθυρο μηδενισμού, που διαρκεί 15 λεπτά). Φυσικά, διατίθενται συνδρομητικές επιλογές για επιχειρήσεις, όπου αναλόγως το συνδρομητικό πακέτο που έχει επιλεγεί τα παραπάνω όρια είναι πιο χαλαρά και μπορούν να ανακτηθούν δημοσιεύσεις έως και το πρώτο πρώτο tweet που αναρτήθηκε το 2006.

- Ads API: Δημιουργία και διαχείριση μιας διαφημιστικής καμπάνιας μιας εταιρείας στο Twitter. Συμπληρωματικά, παρέχεται η δυνατότητα στοχευμένης διαφήμισης και ανάλυσης της απήχησής της.

- Engagement API: Άντληση πληροφοριών για τη συμμετοχή/δέσμευση/εμπλοκή (engagement) των χρηστών σε συγκεκριμένες δημοσιεύσεις της επιλογής του προγραμματιστή. Υπάρχουν διάφορες μετρικές δέσμευσης, όπως ο όγκος των favorites, ο όγκος των retweets κ.ά. Επίσης, είναι πορφανές ότι μια επιχείρηση μπορεί με ποικίλους τρόπους να αυξήσει την εμπλοκή των χρηστών στην πλατφόρμα του Twitter κυρίως δημοσιεύοντας το κατάλληλο περιεχόμενο.

- Direct Message API: Προγραμματισμένη αποστολή προσωπικών μηνυμάτων σε συγκεκριμένους παραλήπτες που ορίζει ο προγραμματιστής. Με αυτόν τον τρόπο μπορεί να ενισχυθεί το engagement των χρηστών με τις εταιρείες, που αναφέραμε προηγουμένως, καθώς διευκολύνεται η εξυπηρέτηση πελατών και το μάρκετινγκ μέσω διαλόγων είτε με φυσικά πρόσωπα είτε με διαλογικούς πράκτορες (chatbots).

- Account Activity API: Λήψη αυτοματοποιημένων μηνυμάτων μέσω webhook. Αυτά τα μηνύματα πυροδοτούνται από συγκεκριμένα γεγονότα, π.χ. την εγγραφή ενός νέου χρήστη σε μια υπηρεσία και αποστέλλονται σε πραγματικό χρόνο. Συνεπώς, δίνουν τη δυνατότητα σε έναν προγραμματιστή να παρακολουθεί τις ενέργειες διάφορων χρηστών σε σχέση με τις υπηρεσίες της επιχείρησης με την οποία συνεργάζεται χωρίς καθυστερήσεις.

- Ενσωμάτωση περιεχομένου από το Twitter σε ιστοσελίδες και εφαρμογές iOS και Android. Σε αυτήν την περίπτωση, πάλι μπορεί να ενισχυθεί το engagement αλλά τώρα άμεσα μέσω των ιστοσελίδων και των εφαρμογών που διαθέτει η κάθε επιχείρηση.

4.2.2 Get Old Tweets script

Όσον αφορά τη λειτουργία της ανάκτησης tweets, το επίσημο API του Twitter, όπως είδαμε προηγουμένως, εισάγει κάποιους περιορισμούς στο χρονικό εύρος της αναζήτησης και στο ρυθμό των αιτημάτων που μπορούν να υποβληθούν. Το "Get Old Tweets" είναι ένα script γραμμένο στη γλώσσα προγραμματισμού Python που σχεδιάστηκε για τον υπερκερασμό αυτών των εμποδίων, χωρίς να είναι υποχρεωτική η καταβολή κάποιου χρηματικού ποσού. Η λειτουργία του μιμείται αυτήν της επίσημης αναζήτησης του Twitter και εκμεταλλεύεται την κύλιση προς τα κάτω (scrolling) για να είναι δυνατή η ανάκτηση οσοδήποτε παλιών δημοσιεύσεων στην πλατφόρμα.

Η πρώτη έκδοση[19] του γράφτηκε σε Python 2 και διατηρείται σε ένα αποθετήριο στο GitHub από τον Jefferson Henrique. Ο χρήστης μπορεί να χρησιμοποιήσει αυτό το script για να ανακτήσει οσαδήποτε tweets έχοντας θέσει κάποια κριτήρια αναζήτησης. Αυτά τα κριτήρια μπορεί

[19] https://github.com/Jefferson-Henrique/GetOldTweets-python

να αφορούν το χρήστη που πραγματοποίησε τις αναρτήσεις, το χρονικό διάστημα και το γεωγραφικό εύρος στο οποίο είναι επιθυμητό να λάβει χώρα η αναζήτηση, κάποια συμβολοσειρά που θέλουμε να περιέχουν τα tweets, αν επιθυμούμε να μας επιστραφούν μόνο top tweets ή όχι και πόσα το πολύ να είναι αυτά. Τα αποτελέσματα που επιστρέφονται αντιστοιχούν σε tweets και αποτελούν στιγμιότυπα της κλάσης Tweet η οποία περιλαμβάνει τα εξής πεδία: το αναγνωριστικό της δημοσίευσης (id), το μόνιμο σύνδεσμο (permalink) όπου βρίσκεται η ανάρτηση, το όνομα χρήστη (username) που την έκανε, το περιεχόμενό της (text), την ημερομηνία και την ώρα που πραγματοποιήθηκε (date), το πλήθος των αναδημοσιεύσεων (retweets) και των επισημάνσεων «Μου αρέσει» (favorites) που συγκέντρωσε, τις αναφορές σε άλλους χρήστες (mentions) και τις ετικέτες (hashtags) που περιέχει και το γεωγραφικό στίγμα (geo) του χρήστη τη στιγμή της δημοσίευσης εφόσον ο ίδος έχει ενεργοποιήσει την αντίστοιχη επιλογή. Τέλος, τα αποτελέσματα είναι δυνατό να αποθηκευτούν σε αρχείο της μορφής csv για μελλοντική χρήση και ανάλυση.

Υπάρχει και μια μεταγενέστερη έκδοση[20] του script γραμμένη σε Python 3 που διατηρείται σε αποθετήριο στο GitHub από τον Victor Irekponor. Κατά βάση η αναζήτηση πραγματοποιείται με τον ίδιο τρόπο και τα αποτελέσματα επιστρέφονται με την ίδια μορφή, όπως περιγράφηκε παραπάνω. Δύο μικρές προσθήκες είναι ότι τώρα ο χρήστης μπορεί να ορίσει και τη γλώσσα ως κριτήριο αναζήτησης και στα πεδία της κλάσης Tweet συμπεριλαμβάνονται επιπλέον το πλήθος των απαντήσεων (replies) και στην περίπτωση που η εν λόγω δημοσίευση είναι απάντηση σε κάποια άλλη, το όνομα χρήστη εκείνου που ανήρτησε την τελευταία. Μια τελευταία διαφορά είναι ότι το αρχικό script επιστρέφει τα δεδομένα στην τοπική ώρα του συστήματός μας (Ελλάδας) ενώ το μεταγενέστερο τα επιστρέφει στη Συντονισμένη Παγκόσμια Ώρα (UTC).

4.2.3 Google Trends API

Όπως αναφέραμε και παραπάνω, η πλατφόρμα Google Trends[21] παρέχει πρόσβαση σε ένα δείγμα πραγματικών αιτημάτων που υποβάλλονται στη μηχανή αναζήτησης της Google. Όπως μας προδιαθέτει το όνομά της, μπορεί να χρησιμοποιηθεί για την ανακάλυψη τάσεων, των πιο πρόσφατων ανερχόμενων αναζητήσεων, την εύρεση της δημοφιλίας όρων αναζήτησης και σχετικών θεμάτων και αναζητήσεων με αυτούς. Υπενθυμίζουμε ότι δίνει τη δυνατότητα σε ένα χρήστη να πληροφορηθεί για τη δημοφιλία όρων αναζήτησης είτε για τον καθένα χωριστά (απόλυτη) είτε σε σύγκριση μεταξύ τους (σχετική) σε ένα συγκεκριμένο χρονικό διάστημα και γεωγραφικό εύρος σε επίπεδο χωρών. Η αναζήτηση μπορεί να πραγματοποιήθει είτε με όρους είτε με θέματα. Η διαφορά έγκειται στο ότι με τους όρους αναζήτησης στα αποτελέσματα περιλαμβάνονται αναζητήσεις που περιέχουν τις λέξεις που ορίστηκαν σε αυτούς ενώ με τα θέματα αναζήτησης στα αποτελέσματα περιλαμβάνονται αναζητήσεις με βάση όρους σημασιολογικά όμοιους με αυτά και μεταξύ τους αλλά πιθανόν αρκετά διαφορετικούς λεξιλογικά. Τέλος, μπορεί να επιλεγεί κάποια κατηγορία στην οποία είναι επιθυμητό να εμπίπτουν οι αναζητήσεις αλλά και να οριστεί ο χώρος αναζήτησης (αναζήτηση στον Ιστό, αναζήτηση εικόνων, αναζήτηση ειδήσεων, αγορές Google και αναζήτηση YouTube).

Υπάρχουν δύο μορφές με τις οποίες επιστρέφονται τα αποτελέσματα και είναι διαθέσιμες για λήψη σε αρχείο τύπου csv, για ενσωμάτωση σε σελίδες HTML σε υπολογιστή ή κινητό και για κοινοποίηση σε διάφορα μέσα κοινωνικής δικτύωσης. Η πρώτη μορφή είναι οι χρονοσειρές, όπου μπορούν να φανούν οι διακυμάνσεις στη σχετική δημοφιλία των όρων συναρτήσει του χρόνου. Αν το χρονικό διάστημα μελέτης είναι η τελευταία ώρα τα αποτελέσματα δίνονται ανά λεπτό, αν είναι η τελευταία εβδομάδα ανά ώρα ενώ όταν πρόκειται για ιστορικά δεδομένα χρονικού διαστήματος έως και 270 ημερών δίνονται σε ημερήσια βάση. Για περιόδους μεγαλύτερης διάρκειας των 9 μηνών και μικρότερης των 5 ετών επιστρέφονται αποτελέσματα σε εβδομαδιαία βάση ενώ για ακόμα μεγαλύτερης διάρκειας σε μηνιαία. Η δεύτερη μορφή με την οποία επιστρέφονται τα αποτελέσματα είναι αυτή των χρωματισμένων χαρτών. Σε αυτήν την περίπτωση, φαίνονται οι διακυμάνσεις της σχετικής δημοφιλίας των όρων στις διάφορες γεωγραφικές υποενότητες της ευρύτερης περιοχής

[20] https://github.com/marquisvictor/Optimized-Modified-GetOldTweets3-OMGOT
[21] https://trends.google.com/trends/?geo

που ορίστηκε αρχικά. Επίσης, αξίζει να σημειώσουμε πάλι ότι τα αποτελέσματα είναι δειγματοληπτημένα και κανονικοποιημένα ως προς το συνολικό πλήθος των αναζητήσεων στην υπό εξέταση χρονική περίοδο και περιοχή στο εύρος [0-100]. Οι δύο τελευταίες παρατηρήσεις είναι ιδιαίτερα περιοριστικές για την ανάλυση που επιχειρούμε να κάνουμε στην παρούσα εργασία, όπως θα δούμε και παρακάτω, καθώς δε μας επιτρέπουν να συλλέξουμε ημερήσια δεδομένα για μεγαλύτερα χρονικά διαστήματα και αν διασπάσουμε ένα μεγάλο χρονικό διάστημα αναζήτησης σε μικρότερα αυτά είναι μη συγκρίσιμα μεταξύ τους λόγω κανονικοποίησης. Τέλος, οι μη δημοφιλείς όροι που συγκεντρώνουν μικρό πλήθος αναζητήσεων έχουν πάντα μηδενική τιμή και δεν μπορούμε να εξάγουμε πληροφορίες για τη δημοφιλία τους σε διάφορες χρονικές στιγμές και υποπεριοχές.

4.2.4 pytrends

Πρόκειται για ένα ανεπίσημο API[22] για την πλατφόρμα Google Trends, που επιτρέπει την αυτοματοποίηση της μεταφόρτωσης δεδομένων από την τελευταία. Συντηρείται σε αποθετήριο στο GitHub από την εταιρεία General Mills και πρέπει να ενημερώνεται αρκετά συχνά ακολουθώντας τις αλλαγές της Google. Ως επί το πλείστον, διαθέτει την ίδια λειτουργικότητα με το επίσημο API, που αναλύσαμε παραπάνω και κάποιες επιπλέον δυνατότητες που το καθιστούν ισχυρό εργαλείο. Πιο συγκεκριμένα, υποβάλει αιτήματα στην πλατφόρμα συνοδευόμενα από διάφορες μεθόδους αναλόγως τα δεδομένο που είναι επιθυμητό να συλλεχθούν. Οι μέθοδοι που διαθέτει είναι οι ακόλουθες:

- Interest Over Time: επιστρέφει ιστορικά δεδομένα από τη πλατφόρμα συναρτήσει του χρόνου, όπως ακριβώς συμβαίνει και στο επίσημο API.
- Historical Hourly Interest: επιστρέφει ιστορικά δεδομένα από την πλατφόρμα σε ωριαία βάση, κάτι που δεν είναι δυνατό μέσω του επίσημου API. Το πετυχαίνει στέλνοντας επαναλαμβανόμενα αιτήματα ανά εβδομάδα και αποκτώντας ωριαία δεδομένα για καθεμία από αυτές και έπειτα ανακατασκευάζοντάς τα κατάλληλα.
- Interest by Region: επιστρέφει δεδομένα σε σχέση με τις διάφορες υποπεριοχές που ορίζονται στην περιοχή ενδιαφέροντος κατά αντιστοιχία με τους χρωματισμένους χάρτες του επίσημου API.
- Related Queries: επιστρέφει τα σχετικά ερωτήματα που υπέβαλαν οι χρήστες στη μηχανή αναζήτησης της Google σε σχέση με το ερώτημα τη δημοφιλία του οποίου αναζητήσαμε, κατά αντιστοιχία με το τμήμα «Σχετικά ερωτήματα» του επίσημου API.
- Related Topics: επιστρέφει τα σχετικά θέματα με τον όρο αναζήτησης τη δημοφιλία του οποίου εξετάζουμε, κατά αντιστοιχία με το τμήμα «Σχετικά Θέματα» του επίσημου API.
- Trending Searches: επιστρέφει δεδομένα για τις τελευταίες τάσεις, ανερχόμενες αναζητήσεις στη μηχανή αναζήτησης της Google. Αυτή η λειτουργία υπάρχει και στο επίσημο API.
- Top Charts: επιστρέφει τα δεδομένα σχετικά με ένα συγκεκριμένο θέμα, όπως ακριβώς συμβαίνει και στο αντίστοιχο τμήμα του επίσημου API.
- Suggestions: επιστρέφει μια λίστα με προτάσεις για όρους αναζήτησης, όπως ακριβώς όταν πληκτρολογούμε έναν όρο στο πεδίο αναζήτησης του επίσημου API. Αυτό το στοιχείο μάς διευκολύνει να επιλέξουμε τον καταλληλότερο όρο αναζήτησης για το θέμα που μας ενδιαφέρει να μελετήσουμε.

[22] `https://pypi.org/project/pytrends/`

Ως παράμετροι στις παραπάνω μεθόδους συνήθως δίνονται στοιχεία ανάλογα με αυτά που απαιτούνται για την υποβολή ενός ερωτήματος στο επίσημο Google Trends API. Τέτοιες παράμετροι μπορεί να είναι μια λίστα όρων ή/και θεμάτων αναζήτησης, η κατηγορία στην οποία αυτά εμπίπτουν, η γεωγραφική περιοχή αναζήτησης, η ζώνη ώρας σύμφωνα με την οποία θα επιστραφούν τα αποτελέσματα, το χρονικό διάστημα αναζήτησης και η ιδιότητα της Google που προσδιορίζει το χώρο αναζήτησης.

Επίσης, από την ίδια ομάδα προγραμματιστών που ανέπτυξαν και συντηρούν τη βιβλιοθήκη pytrends, έχει προταθεί μία συνάρτηση `get_daily_data`, η οποία ανακατασκευάζει ιστορικά ημερήσια δεδομένα για χρονικές περιόδους μεγαλύτερες των 5 ετών. Υπενθυμίζουμε ότι για τόσο μεγάλα χρονικά διαστήματα μπορούμε να ανακτήσουμε τα δεδομένα μόνο σε μηνιαία βάση ενώ για μικρότερα διαστήματα εύρους ενός μήνα μπορούμε να τα ανακτήσουμε σε ημερήσια. Η συγκεκριμένη συνάρτηση βασίζεται σε αυτό το γεγονός και συλλέγει τα δεδομένα για το επιθυμητό διάστημα που είναι μεγαλύτερο των 5 ετών σε δύο φάσεις. Στην πρώτη φάση υποβάλλεται ένα ερώτημα και ανακτώνται τα μηνιαία δεδομένα ενώ στη δεύτερη φάση υποβάλλονται επαναλαμβανόμενα ερωτήματα, ένα για καθέναν από τους παραπάνω μήνες, με αποτέλεσμα να ανακτώνται τα ημερήσια δεδομένα για κάθε μήνα. Τα τελευταία δεν μπορούν να συγκριθούν μεταξύ διαφορετικών μηνών, λόγω της κανονικοποίησης που υφίστανται τα δεδομένα. Για να γίνουν συγκρίσιμα μεταξύ τους και να ληφθούν τελικά τα επιθυμητά ημερήσια δεδομένα για το συνολικό διάστημα ενδιαφέροντος, εκμεταλλεύονται τα μηνιαία δεδομένα τα οποία χρησιμοποιούνται ως συντελεστές για την επανακλιμάκωση των προηγούμενων. Δηλαδή κάθε ημερήσια τιμή πολλαπλασιάζεται με το συντελεστή που προκύπτει από την αντίστοιχη μηνιαία τιμή. Με αυτόν τον τρόπο, τελικά, προκύπτει μια προσέγγιση των ημερήσιων αναζητήσεων και για μεγαλύτερα χρονικά διαστήματα των 5 ετών.

Κεφάλαιο 5

Προετοιμασία πειραμάτων

5.1 Επισκόπηση δεδομένων αναφοράς

Όπως έχουμε αναφέρει και προηγουμένως, στην παρούσα εργασία θα ασχοληθούμε με την εκτίμηση της τηλεθέασης εκπομπών. Πιο συγκεκριμένα, ως δεδομένα αναφοράς (ground truth data) έχουμε στη διάθεσή μας τα νούμερα τηλεθέασης του ιταλικού ψυχαγωγικού τηλεοπτικού show "Le Iene"[1] για τις χρονιές 2016-2017. Το συγκεκριμένο πρόγραμμα προβάλλεται στη γείτονα χώρα από το κανάλι Italia 1 από το 1997. Η θεματολογία του κινείται γύρω από κοινωνικοπολιτικά τεκταινόμενα τόσο της ιταλικής όσο και της παγκόσμιας επικαιρότητας, τα οποία αναδεικνύονται μέσω ρεπορτάζ και σατιρικών σκετς. Επίσης, περιλαμβάνει συνεντεύξεις διάσημων προσώπων και έχει αναφορές στην ταινία "Reservoir Dogs" του Quentin Tarantino. Όλα αυτά τα χρόνια προβολής του, πολλοί ηθοποιοί και κωμικοί έχουν περάσει από το τιμόνι της παρουσίασής του, με πρώτη τη Simona Ventura.

Το σύνολο των δεδομένων αναφοράς αποτελείται από 107 επεισόδια τεσσάρων τηλεοπτικών σεζόν ($37^{η}$ έως και $40^{η}$ σεζόν) που αντιστοιχούν σε δύο ημερολογιακά έτη (2016-2017). Για καθένα από αυτά τα επεισόδια έχουμε στη διάθεσή μας τις χρονικές στιγμές έναρξης και λήξης του, τη διάρκειά του, τα ποσοστά τηλεθέασης επί του συνόλου των ανοιχτών δεκτών και τον απόλυτο αριθμό τηλεθεατών που συγκέντρωσε. Τα δύο τελευταία μεγέθη, που είναι και τα ζητούμενα προς εκτίμηση, υπάρχουν τόσο ανεξαρτήτου φύλου και ηλικίας όσο και για τρεις ηλικιακές ομάδες (15-34, 35-54, 55+) και ανά φύλο χωριστά. Η εξαγωγή δημογραφικών χαρακτηριστικών είναι δύσκολη από τα μέσα κοινωνικής δικτύωσης έως και αδύνατη από τις μηχανές αναζήτησης, οπότε θα περιορίσουμε την ανάλυσή μας στην εκτίμηση των συνολικών ποσοστών τηλεθέασης και πλήθους τηλεθεατών ανεξαρτήτου φύλου και ηλικίας. Ωστόσο και από τα υπόλοιπα αντλήσαμε την πληροφορία ότι καταγράφονται υψηλότερα ποσοστά τηλεθέασης στις νεότερες ηλικιακές ομάδες (15-54 ετών). Αν κάνουμε την απλοϊκή υπόθεση ότι οι άνθρωποι που ανήκουν σε αυτές τείνουν να είναι πιο ενεργοί στα μέσα κοινωνικής δικτύωσης, και συγκεκριμένα στο Twitter και να χρησιμοποιούν συχνότερα μηχανές αναζήτησης στο διαδίκτυο, μπορούμε να συμπεράνουμε ότι η συγκεκριμένη εκπομπή ενδείκνυται για τη μελέτη που επιχειρούμε να πραγματοποιήσουμε.

Η εκπομπή προβάλλεται δύο μέρες τη βδομάδα, η μία εκ των οποίων είναι σταθερά η Κυριακή ενώ η άλλη κάποιες φορές είναι η Τρίτη, κάποιες η Τετάρτη και κάποιες άλλες η Πέμπτη. Επίσης, η προβολή της διακόπτεται μόνο την περίοδο του καλοκαιριού (μέσα Μαΐου ή Ιουνίου - Σεπτέμβριος), στις γιορτές των Χριστουγέννων (τέλη Δεκεμβρίου - τέλη Ιανουαρίου) και μια βδομάδα λόγω Πάσχα μέσα στον Απρίλιο του 2017. Συνήθως, η διάρκεια των επεισοδίων κυμαίνεται μεταξύ 2 και 3 ωρών και σε λίγες περιπτώσεις φτάνει έως και τις 3.5 ώρες. Η ώρα έναρξης είναι λίγο μετά τις 21.00 και η ώρα λήξης λίγο πριν ή λίγο μετά τα μεσάνυχτα. Οφείλουμε να σημειώσουμε ότι οι ώρες αναφέρονται σε χειμερινή ή εαρινή, αναλόγως την περίοδο, τοπική ώρα Ιταλίας. Αυτές οι πληροφορίες θα επηρεάσουν την απόφασή μας για το ποιο χρονικό παράθυρο είναι το καλύτερο για την ομαδοποίηση των χαρακτηριστικών, όπως θα δούμε αργότερα.

[1] `https://www.iene.mediaset.it/`

Κάθε παρατήρηση του συνόλου δεδομένων (dataset), που θα χρησιμοποιήσουμε, αντιστοιχεί σε ένα επεισόδιο από τα 107 συνολικά. Αρχικά, το σύνολο δεδομένων περιλαμβάνει 5 μεταβλητές, την ημερομηνία προβολής του κάθε επεισοδίου, τη χρονική στιγμή έναρξης και λήξης του, το συνολικό ποσοστό τηλεθέασης επί των ανοικτών δεκτών και το απόλυτο πλήθος τηλεθεατών που συγκέντρωσε. Οι δύο τελευταίες αποτελούν τις εξαρτημένες μεταβλητές ή μεταβλητές εξόδου. Στη συνέχεια, θα προσθέτουμε σταδιακά παραπάνω μεταβλητές που θα αντιστοιχούν σε χαρακτηριστικά τα οποία πιθανώς να χρησιμοποιηθούν ως είσοδοι στους αλγορίθμους εκπαίδευσης που θα εκτελέσουμε. Επίσης, αξίζει να αναφέρουμε ότι τα χρονικά δεδομένα μετατράπηκαν από τοπική ζώνη ώρας Ιταλίας στην Παγκόσμια Συντονισμένη Ώρα (UTC). Αυτό ήταν αναγκαίο για να υπάρχει συνέπεια των δεδομένων μεταξύ διαφορετικών πηγών. Όπως θα δούμε αργότερα, κατά τη συλλογή των δεδομένων από την πλατφόρμα Google Trends με τη βιβλιοθήκη pytrends τα δεδομένα επιστρέφονται σε UTC και όχι σε τοπική ώρα Ιταλίας και μάλλον δεν είναι δυνατό να το αλλάξουμε αυτό. Επιπλέον, η μετατροπή σε UTC έχει το πλεονέκτημα ότι με εξαίρεση ενός επεισοδίου όλα τα υπόλοιπα προβάλλονται ολόκληρα εντός της ίδιας μέρας που ξεκινούν. Έτσι θα μπορούμε να αναφερθούμε καθαρά σε μέρα προβολής και επόμενη αυτής. Στη συνέχεια, τόσο τα δεδομένα που συλλέγονται από το Google Trends όσο και αυτά από το Twitter αν δεν είναι ήδη σε UTC θα πρέπει να μετατρέπονται.

Τέλος, προβήκαμε στην εξαγωγή κάποιων στατιστικών στοιχείων για τα δεδομένα αναφοράς, ώστε να έχουμε μια καλύτερη εικόνα για τα μεγέθη που σκοπεύουμε να εκτιμήσουμε. Φυσικά, μεριμνήσαμε για την αποφυγή της διαρροής δεδομένων και κατ΄ επέκταση της υπερεκπαίδευσης, εκμεταλλευόμενοι παρατηρήσεις μόνο από το σύνολο δεδομένων εκπαίδευσης για τη στατιστική μας ανάλυση. Τα συμπεράσματα ήταν πως το συνολικό ποσοστό τηλεθέασης επί των ανοιχτών δεκτών κυμαίνεται μεταξύ 6.5% και 15.1%, με μέση τιμή 10.11% και τυπική απόκλιση 1.36. Αντίστοιχα, το απόλυτο πλήθος των τηλεθεατών κυμαίνεται μεταξύ 1.29 και 2.86 εκατομμυρίων, με μέση τιμή 1.99 εκατομμύρια και τυπική απόκλιση 0.26 εκατομμύρια. Τα αποτελέσματα παρουσιάζονται αναλυτικά στην εικόνα που ακολουθεί.

	SHR %	AMR
count	96.000000	9.600000e+01
mean	10.113542	1.994179e+06
std	1.355816	2.629714e+05
min	6.500000	1.292489e+06
25%	9.200000	1.837154e+06
50%	10.050000	1.974116e+06
75%	10.825000	2.167976e+06
max	15.100000	2.863314e+06

Σχήμα 5.1: Περιγραφικά στατιστικά στοιχεία για τα δεδομένα αναφοράς

5.2 Συλλογή δεδομένων

Στην παρούσα εργασία, όπως έχουμε προαναφέρει, θα συλλέξουμε δεδομένα από δύο πηγές, το μέσο κοινωνικής δικτύωσης Twitter και την πλατφόρμα Google Trends που παρέχει στατιστικά στοιχεία για τη μηχανή αναζήτησης της Google. Στο τεχνικό υπόβαθρο περιγράψαμε διάφορους τρόπους με τους οποίους μπορούν αυτά να συλλεχθούν και σε αυτό το σημείο θα αναλυθούν οι μέθοδοι που τελικά επιλέχτηκαν.

5.2.1 Συλλογή δεδομένων από το Twitter

Τα δεδομένα από το Twitter για το τηλεοπτικό show που εξετάζουμε συλλέχθηκαν με βάση το επίσημο hashtag της εκπομπής (#leiene) για το χρονικό διάστημα των 2 ετών που μελετάμε. Για τη συλλογή τους χρησιμοποιήθηκε κατάλληλος crawler, και πιο συγκεκριμένα η δεύτερη έκδοση σε Python 3 του Get Old Tweets script. Το επίσημο API του Twitter απορρίφθηκε, καθώς δε μας επέτρεπε να ανακτήσουμε τόσο παλιά tweets χωρίς συνδρομή. Μάλιστα, επειδή ο συνολικός όγκος των δεδομένων ήταν πολύ μεγάλος (190947 tweets), τα tweets συλλέχθηκαν σε μηνιαία βάση για αυτά τα 2 έτη και αποθηκεύτηκαν σε ξεχωριστά αρχεία. Τα δεδομένα επιστράφηκαν σε UTC, οπότε και βρίσκονται σε συνέπεια όσον αφορά τη ζώνη ώρας με τα δεδομένα αναφοράς έπειτα από τη μετατροπή. Επίσης, στο χρονικό παράθυρο που χρησιμοποιήθηκε για τη συλλογή συμπεριλήφθηκε και η πρώτη μέρα του επόμενου μήνα, καθώς τα δεδομένα ανακτώνται με βάση το ανοιχτό δεξιά διάστημα που ορίζει ο χρήστης.

Τελικά, για κάθε tweet συγκεντρώσαμε τις εξής πληροφορίες: το username του χρήστη που το δημοσίευσε, την ημερομηνία και την ώρα που δημοσιεύθηκε, το πλήθος των retweets, των favorites και των replies που συγκέντρωσε, το περιεχόμενό του, τα mentions σε άλλους χρήστες και τα hashtags που περιέχει, το αναγνωριστικό του και το μόνιμο σύνδεσμο όπου μπορεί κάποιος να το επισκεφθεί καθώς και στην περίπτωση που είναι απάντηση σε κάποια άλλη ανάρτηση το όνομα χρήστη εκείνου που έκανε την αρχική ανάρτηση. Τα περισσότερα από αυτά θα τα εκμεταλλευτούμε για την εξαγωγή χαρακτηριστικών σε διάφορα χρονικά παράθυρα, όπως θα δούμε και σε επόμενη ενότητα.

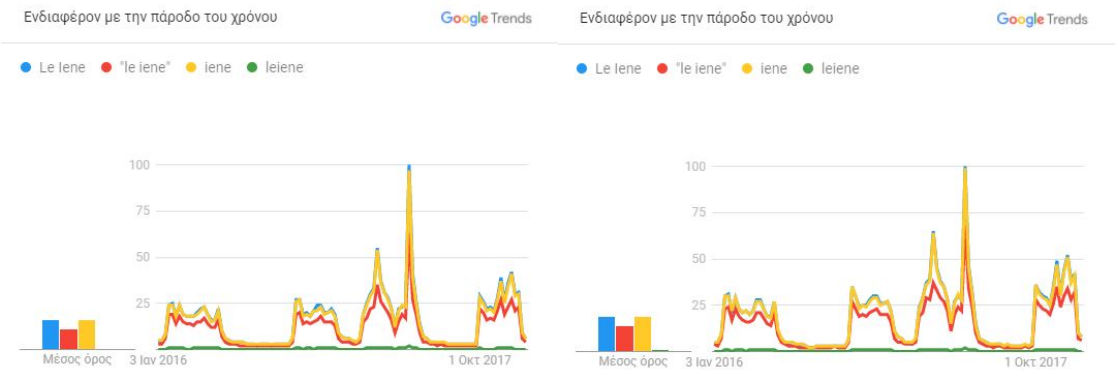
5.2.2 Συλλογή δεδομένων από το Google Trends

Αρχικά, εξετάζουμε τις δυνατότητες που παρέχει το επίσημο API της πλατφόρμας Google Trends. Μέσω αυτού για το χρονικό εύρος των 2 ετών που μας ενδιαφέρει, μπορούμε να λάβουμε δεδομένα σε εβδομαδιαία βάση μόνο, από την Κυριακή έως και το Σάββατο. Πιο συγκεκριμένα, τα αποτελέσματα που επιστρέφονται αφορούν το διάστημα 03/01/2016 - 06/01/2018, όμως αυτή η μικρή παρέκκλιση από το ζητούμενο διάστημα δε μας επηρεάζει, καθώς την εορταστική περίοδο γύρω από την αλλαγή του χρόνου η υπό μελέτη εκπομπή δεν προβάλλεται.

Αρχικά, η αναζήτηση γίνεται με 4 όρους που θεωρούμε ότι καλύπτουν το μεγαλύτερο μέρος των αναζητήσεων των χρηστών για τη συγκεκριμένη εκπομπή. Πιο αναλυτικά, χρησιμοποιείται ένα θέμα (Le Iene) και 3 όροι αναζήτησης (”le iene”, iene, leiene). Η χρήση του θέματος, όπως εξηγήσαμε και παραπάνω, συνεπάγεται την αναζήτηση με βάση μια ομάδα όρων, οι οποίοι συνδέονται σημασιολογικά μεταξύ τους και με το συγκεκριμένο τηλεοπτικό πρόγραμμα ακόμα και αν οι λέξεις που περιέχουν διαφέρουν αρκετά μεταξύ τους. Θεωρητικά, αυτό το θέμα παρέχει πιο αξιόπιστα και ολοκληρωμένα αποτελέσματα. Για λόγους πληρότητας, προσθέτουμε στην αναζήτησή μας και τους άλλους 3 όρους, οι οποίοι σχετίζονται με αναζητήσεις που περιλάμβαναν αυστηρά αυτές τις λέξεις. Μάλιστα, ο όρος ”le iene” επειδή αποτελείται από δύο λέξεις, περικλείεται σε εισαγωγικά ώστε να αναφέρεται μόνο σε αναζητήσεις που εμφανίζουν αυτές τις δύο λέξεις με αυτή τη σειρά και χωρίς κάποια άλλη να παρεμβάλλεται μεταξύ τους. Φυσικά, πέρα από τον χρονικό περιορισμό, εφαρμόστηκε και γεωγραφικός περιορισμός που δεν είναι άλλος από την Ιταλία, τη χώρα στην οποία προβάλλεται το υπό μελέτη show.

Σε πρώτη φάση, λαμβάνουμε τις αναζητήσεις στον Ιστό για όλες τις θεματικές κατηγορίες. Στη συνέχεια, για να βεβαιωθούμε ότι κάποιοι όροι δεν αφορούν άσχετες κατηγορίες (π.χ. IENE: Ινστιτούτο Ενέργειας Νοτιοανατολικής Ευρώπης, η ύαινα ως ζώο), επιλέξαμε να λαμβάνονται οι αναζητήσεις στον Ιστό που ανήκουν στην κατηγορία «Τέχνες και Ψυχαγωγία». Τα αποτελέσματα φαίνονται στο σχήμα 5.2. Παρατηρούμε ότι οι διαφορές των αποτελεσμάτων στις δύο περιπτώσεις είναι πολύ μικρές. Επίσης, και στις δύο περιπτώσεις οι χρονοσειρές που αφορούν το θέμα Le Iene και τους δύο όρους αναζήτησης ”le iene” και iene έχουν σχεδόν πανομοιότυπη μορφή και φαίνεται να είναι έντονα συσχετισμένες μεταξύ τους, όπως άλλωστε αναμενόταν. Από την άλλη, η χρονοσειρά που αφορά τον όρο leiene (η συμβολοσειρά που αντιστοιχεί στο επίσημο hashtag

της εκπομπής) είναι σχεδόν σταθερά μηδενική, γιατί αυτός ο όρος περιλαμβάνεται σε συντριπτικά μικρότερο αριθμό αναζητήσεων σε σχέση με τους υπόλοιπους στο ίδιο χρονικό διάστημα. Θα ήταν ασφαλές, λοιπόν, να κρατήσουμε μόνο το θέμα Le Iene που όπως προείπαμε είναι πιο αξιόπιστο και με αυτόν τον τρόπο να μην έχουμε πλεονάζουσα πληροφορία αλλά και να μη μας προβληματίζει ποια κατηγορία θα επιλέξουμε.



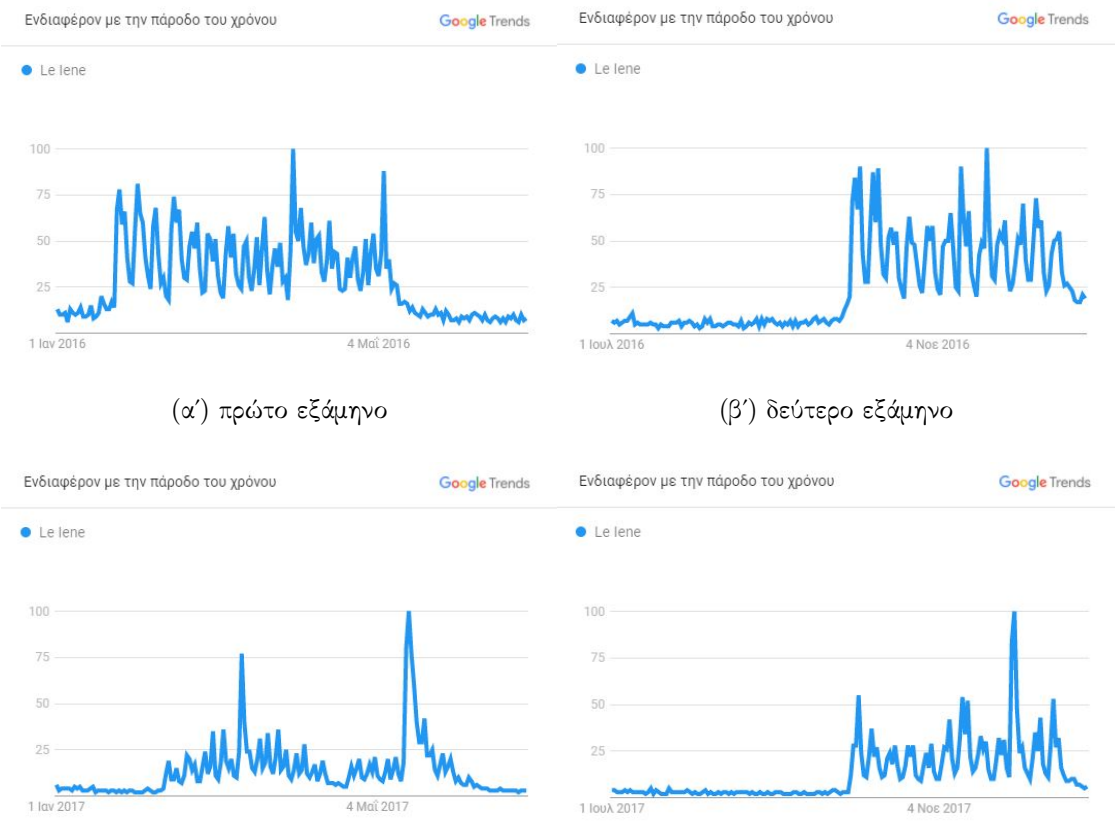
(α′) ανεξαρτήτως κατηγορίας (β′) για την κατηγορία Τέχνες και Ψυχαγωγία

Σχήμα 5.2: Ενδιαφέρον με την πάροδο του χρόνου για την εκπομπή

Είναι αρκετά ενθαρρυντικό το γεγονός ότι το πλήθος των αναζητήσεων στον Ιστό είναι εμφανώς μειωμένο σε περιόδους που το show δεν προβάλλεται στην τηλεόραση, όπως στην καλοκαιρινή περίοδο και στις διακοπές των Χριστουγέννων. Εν αντιθέσει, σε περιόδους που προβάλλεται, το ενδιαφέρον των χρηστών για αυτό είναι ιδιαίτερα αυξημένο, ενισχύοντας την υπόθεσή μας ότι τα δεδομένα από μηχανές αναζήτησης μπορούν να χρησιμοποιηθούν ως αξιόπιστοι εκτιμητές της τηλεθέασης. Βέβαια, εμείς επιθυμούμε να εκτιμήσουμε την τηλεθέαση κάθε επεισοδίου ξεχωριστά και δεδομένου ότι κάθε βδομάδα προβάλλονται δύο επεισόδια, δεν μας αρκούν τα εβδομαδιαία δεδομένα. Για να λάβουμε τα δεδομένα σε ημερήσια βάση, χρειάζεται να σπάσουμε το χρονικό διάστημα μελέτης σε μικρότερα τμήματα, 4 εξάμηνα εν προκειμένω. Ανακύπτει όμως το εξής πρόβλημα. Τα ημερήσια δεδομένα των 4 διαφορετικών εξαμήνων δεν μπορούν να συγκριθούν μεταξύ τους καθώς έχει γίνει διαφορετική κανονικοποίηση στο καθένα. Σε κάθε ένα από αυτά η τιμή 100 αντιστοιχήθηκε στη μέρα με το μεγαλύτερο πλήθος αναζητήσεων σε αυτό το διάστημα και η τιμή 0 στη μέρα με το μικρότερο, αντίστοιχα. Μπορεί όμως το πλήθος των αναζητήσεων που αντιστοιχεί στην τιμή x ενός εξαμήνου να είναι πολύ διαφορετικό από εκείνο που αντιστοιχεί στην ίδια τιμή x κάποιου άλλου εξαμήνου. Οπότε αν θέλουμε να παραγάγουμε ένα και μόνο μοντέλο για την εκτίμηση της τηλεθέασης στο συνολικό διάστημα των 2 ετών θα πρέπει να προηγηθεί μια επανακλιμάκωση αυτών των δεδομένων ούτως ώστε να είναι συγκρίσιμα μεταξύ τους.

Για να το πετύχουμε αυτό, τελικά θα συλλέξουμε τα δεδομένα με τη βοήθεια του ανεπίσημου API που προσφέρει η pytrends. Πιο συγκεκριμένα, προσαρμόζουμε στις ανάγκες του προβλήματός μας το script που προτείνουν οι δημιουργοί της βιβλιοθήκης για την επανακλιμάκωση των δεδομένων σε διαστήματα μεγαλύτερα των 5 ετών. Σε πλήρη αντιστοιχία με αυτό, συλλέγονται δεδομένα σε δύο φάσεις από την πλατφόρμα. Πρώτα για το συνολικό διάστημα ενδιαφέροντος των 2 ετών, οπότε και τα αποτελέσματα δίνονται σε εβδομαδιαία βάση και έπειτα για καθεμία από αυτές τις βδομάδες χωριστά, οπότε και τα αποτελέσματα δίνονται σε ημερήσια βάση. Στο τέλος, πραγματοποιείται επανακλιμάκωση των ημερήσιων δεδομένων συλλεγμένων ανά εβδομάδα με έναν συντελεστή που καθορίζεται από την εβδομαδιαία τιμή της αντίστοιχης εβδομάδας. Στη συνάρτηση που υποβάλλει τα ερωτήματα, δίνουμε ως όρισμα το αναγνωριστικό του θέματος "Le Iene", ώστε να ληφθεί ως θέμα και όχι ως απλός όρος αναζήτησης, το οποίο βρέθηκε με τη βοήθεια της μεθόδου suggestions του pytrends. Το API της Google φαίνεται να αγνοεί την παράμετρο που καθορίζει τη ζώνη ώρας και επιστρέφει τα δεδομένα σε UTC, οπότε τα αποτελέσματα είναι συνεπή με όλα τα υπόλοιπα όσον αφορά τη ζώνη ώρας. Επιπλέον, αξίζει να σημειώσουμε πως

ένας εναλλακτικός τρόπος να υπερπηδήσουμε αυτό το εμπόδιο είναι να λάβουμε επικαλυπτόμενα χρονικά διαστήματα και όχι ανεξάρτητα ανά εβδομάδα, όπως προηγουμένως, ούτως ώστε να επιβάλουμε κατά κάποιον τρόπο κοινή κλιμάκωση. Η μέθοδος αυτή δεν εξετάζεται στην παρούσα εργασία, οπότε δε θα την αναλύσουμε περισσότερο. Τελικά, καταφέραμε να ανακατασκευάσουμε προσεγγιστικά τα ημερήσια δεδομένα, τα οποία θα χρησιμοποιηθούν στη συνέχεια για την εξαγωγή χαρακτηριστικών σε διάφορα χρονικά παράθυρα.



(α′) πρώτο εξάμηνο (β′) δεύτερο εξάμηνο

(γ′) τρίτο εξάμηνο (δ′) τέταρτο εξάμηνο

Σχήμα 5.3: Ενδιαφέρον με την πάροδο του χρόνου για την εκπομπή ανά εξάμηνο

5.3 Εξαγωγή χαρακτηριστικών

Με τα δεδομένα που συλλέξαμε μπορούμε να εξάγουμε ποικίλα χαρακτηριστικά σε διάφορα χρονικά παράθυρα. Τα περισσότερα από αυτά χρησιμοποιούνταν ευρέως και στις εργασίες της συναφούς βιβλιογραφίας που μελετήσαμε. Αναλόγως την πηγή από την οποία προέρχονται τα χαρακτηριστικά, Twitter ή Google Trends, μπορούν να ομαδοποιηθούν σε διαφορετικά χρονικά παράθυρα, όπως θα δούμε στη συνέχεια, ώστε να εμπεριέχουν όση περισσότερη χρήσιμη πληροφορία γίνεται για την κατασκευή των προβλεπτικών μοντέλων στη συνέχεια.

5.3.1 Εξαγωγή χαρακτηριστικών από το Twitter

Με τα δεδομένα που συλλέξαμε από το Twitter, μπορούμε να εξάγουμε δύο ειδών χαρακτηριστικά. Αφενός χαρακτηριστικά που προκύπτουν από ποσοτικούς δείκτες και αφορούν τον όγκο κάποιων ενεργειών και αφετέρου χαρακτηριστικά που προκύπτουν από ανάλυση συναισθήματος πάνω στο κειμενικό περιεχόμενο των tweets και συνεπώς συνδέονται με ποιοτικούς δείκτες. Ο Oghina και λοιποί στο [58] διαχωρίζουν με ανάλογο τρόπο τα χαρακτηριστικά σε δύο κατηγο-

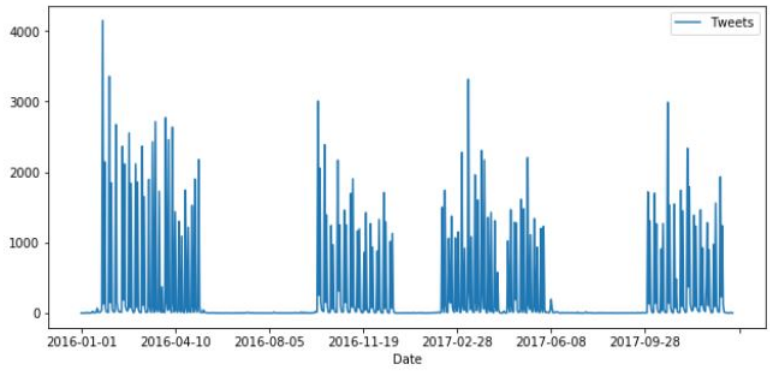
ρίες και αναφέρονται σε αυτές με τους όρους «χαρακτηριστικά επιφάνειας» (surface features) και «κειμενικά χαρακτηριστικά» (textual features), αντίστοιχα.

5.3.1.1 Χαρακτηριστικά ποσοτικών δεικτών

Όπως μας προϊδεάζει και ο τίτλος, τα χαρακτηριστικά που θα αναλύσουμε σε αυτήν την παράγραφο αντιστοιχούν στον όγκο συγκεκριμένων ενεργειών, το πλήθος προσώπων που σχετίζονται με αυτές αλλά και γραμμικούς ή μη συνδυασμούς των προαναφερθέντων. Αυτές οι ενέργειες θεωρούνται διαισθητικά και με γνώμονα συγγενείς προϋπάρχουσες εργασίες ότι σχετίζονται με τις εξαρτημένες μεταβλητές που σκοπεύουμε να εκτιμήσουμε. Δεν εξετάζεται η αιτία που οδήγησε σε αυτές, το περιεχόμενό τους και ο αντίκτυπος που μπορεί να έχουν, παρά μόνο η ποσότητά τους. Παρακάτω, αναφέρουμε συγκεκριμένα ποιες θα αξιοποιηθούν στη συγκεκριμένη εργασία.

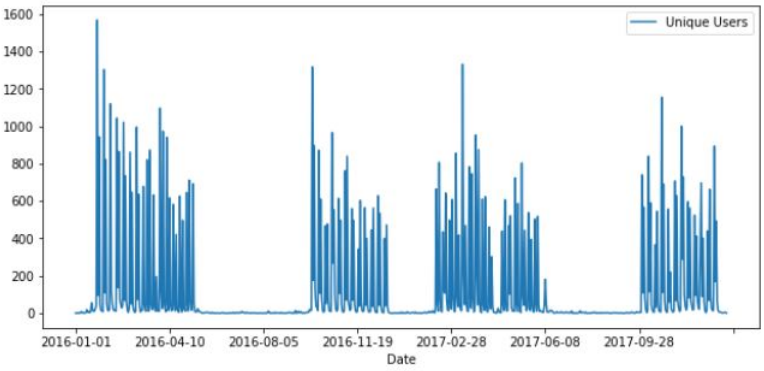
- **Πλήθος αναρτήσεων**

  Πρόκειται για το πιο απλό χαρακτηριστικό (Tweets) που μπορεί να εξαχθεί και αντιστοιχεί στο πλήθος των διακριτών δημοσιεύσεων που έχουν αναρτηθεί στο εκάστοτε χρονικό παράθυρο που μελετάται. Στο παρακάτω διάγραμμα φαίνεται το πλήθος των tweets που συγκεντρώθηκαν για την εκπομπή Le Iene ανά ημέρα.



Σχήμα 5.4: Όγκος των δημοσιεύσεων ανά ημέρα

- **Πλήθος διακριτών χρηστών**

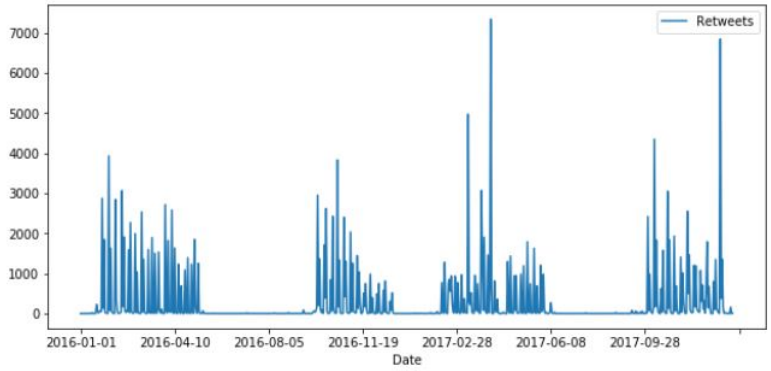


Σχήμα 5.5: Πλήθος των διακριτών χρηστών που δημοσίευσαν ανά ημέρα

Ένα ακόμα προφανές χαρακτηριστικό είναι το πλήθος των διακριτών χρηστών (UnqUsers) που πραγματοποίησαν κάποια δημοσίευση στο Twitter από αυτές που συγκεντρώσαμε εντός του χρονικού διαστήματος που μελετάται. Στο σχήμα 5.5 παρουσιάζεται η μεταβολή αυτού του χαρακτηριστικού για το διάστημα των 2 ετών συναρτήσει του χρόνου ανά ημέρα.

- **Πλήθος αναδημοσιεύσεων**

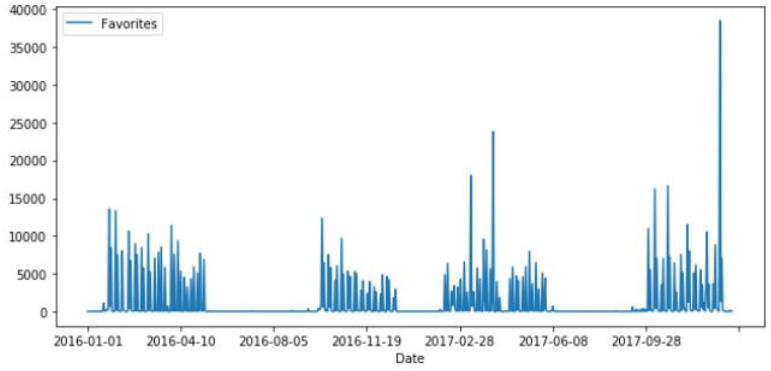
  Το πλήθος των αναδημοσιεύσεων (Retweets) αναφέρεται στο σύνολο των αναδημοσιεύσεων που συγκέντρωσαν όλες οι αναρτήσεις μαζί, οι οποίες συμπεριλήφθηκαν στο πρώτο χαρακτηριστικό. Στο σχήμα 5.6 διαφαίνεται η μεταβολή του σε ημερήσια βάση για το συνολικό διάστημα των 2 ετών που μελετάμε.



Σχήμα 5.6: Όγκος των αναδημοσιεύσεων ανά ημέρα

- **Πλήθος επισημάνσεων «Μου Αρέσει»**

  Το πλήθος των επισημάνσεων «Μου αρέσει» (Favorites) που συγκέντρωσαν όλα τα tweets μαζί, που αναρτήθηκαν στο υπό μελέτη χρονικό παράθυρο. Το σχήμα 5.7 αποδίδει τη χρονική μεταβολή αυτού του χαρακτηριστικού σε ημερήσια βάση.



Σχήμα 5.7: Όγκος των επισημάνσεων «Μου αρέσει» ανά ημέρα

- **Πλήθος δημοσιεύσεων και αναδημοσιεύσεων**

  Τώρα, το πλήθος των δημοσιεύσεων στο επιθυμητό χρονικό παράθυρο προστίθεται με το πλήθος των αναδημοσιεύσεων που αυτές συγκέντρωσαν (Tweets&Retweets).

- **Λόγος επισημάνσεων «Μου Αρέσει» προς αναρτήσεις**

  Επίσης, μπορούμε να εξάγουμε χαρακτηριστικά από το μη γραμμικό συνδυασμό άλλων χαρακτηριστικών. Σε αυτήν την περίπτωση ανήκει ο λόγος του πλήθους των επισημάνσεων «Μου αρέσει» που συγκέντρωσαν τα tweets τα οποία αναρτήθηκαν στο υπό μελέτη χρονικό διάστημα προς το πλήθος των τελευταίων (FavoritesRatio).

- **Λόγος αναδημοσιεύσεων προς δημοσιεύσεις**

  Ένα άλλο παράδειγμα, της παραπάνω κατηγορίας είναι ο λόγος του πλήθους των αναδημοσιεύσεων που συγκέντρωσαν τα tweets τα οποία αναρτήθηκαν στο υπό μελέτη χρονικό παράθυρο προς το πλήθος των τελευταίων (RetweetsRatio).

- **Λόγος δημοσιεύσεων και αναδημοσιεύσεων προς δημοσιεύσεις**

  Το συγκεκριμένο χαρακτηριστικό (Tweets&RetweetsRatio) είναι γραμμικώς εξαρτημένο με το προηγούμενο, καθώς

$$\text{Tweets\&RetweetsRatio} = \frac{\text{Tweets} + \text{Retweets}}{\text{Tweets}} = 1 + \frac{\text{Retweets}}{\text{Tweets}} \Leftrightarrow \tag{5.1}$$

$$\text{Tweets\&RetweetsRatio} = 1 + \text{RetweetsRatio} \tag{5.2}$$

  Συνεπώς, δεν περιέχεται κάποια επιπλέον πληροφορία σε αυτό, οπότε και δε θα χρησιμοποιηθεί στη συνέχεια ως είσοδος στους αλγορίθμους εκπαίδευσης.

- **Λόγος αναρτήσεων προς διακριτούς χρήστες**

  Το τελευταίο χαρακτηριστικό αυτής της κατηγορίας που θα εξεταστεί είναι ο λόγος του όγκου των αναρτήσεων σε ένα προκαθορισμένο χρονικό παράθυρο προς το πλήθος των διακριτών χρηστών που τις πραγματοποίησαν (UnqUsersRatio).

5.3.1.2 Χαρακτηριστικά από ανάλυση συναισθήματος

Τα χαρακτηριστικά που θα αναλύσουμε σε αυτήν την παράγραφο προκύπτουν από την ανάλυση συναισθήματος που εφαρμόστηκε στο σύνολο των κειμενικών περιεχομένων των tweets. Η ανάλυση συναισθήματος πραγματοποιήθηκε σε επίπεδο ανάρτησης και υλοποιήθηκε με τη βοήθεια της βιβλιοθήκης spaCy. Το αποτέλεσμα ήταν να βρεθεί η πολικότητα κάθε tweet, δηλαδή ένας πραγματικός αριθμός στο διάστημα $[-1, 1]$, που αντιστοιχεί στο πόσο θετικό ή αρνητικό χαρακτηρίστηκε το εκάστοτε tweet όσον αφορά τη γνώμη ή το συναίσθημα που μετέδιδε. Όσο πιο κοντά στο -1 τόσο πιο πολύ είναι φορτισμένο με αρνητικά συναισθήματα ενώ όσο πιο κοντά στο 1 τόσο πιο πολύ είναι φορτισμένο με θετικά συναισθήματα. Τιμές κοντά στο 0 σημαίνουν πως το εν λόγω tweet είναι σχετικά ουδέτερο ενώ το απόλυτο 0 συναντάται όταν όλες οι λέξεις που ανήκουν σε αυτό δεν έχουν κάποιο πρόσημο θετικό ή αρνητικό. Η τελευταία φράση προδίδει τη μέθοδο με την οποία υλοποιήθηκε η ανάλυση συναισθήματος στην εργασία μας.

Πιο αναλυτικά, η βιβλιοθήκη spaCy διαθέτει τρία προεκπαιδευμένα μοντέλα στην ιταλική γλώσσα. Όλα έχουν εκπαιδευτεί σε κείμενα της Wikipedia και η διαφορά τους έγκειται στο μέγεθός τους ή πιο συγκεκριμένα στα διανύσματα που χρησιμοποιούνται ως χαρακτηριστικά. Είναι αναμενόμενο τα μεγαλύτερα μοντέλα να έχουν πιο ακριβή αποτελέσματα, θυσιάζοντας περισσότερη μνήμη και χρόνο. Σημειώνεται ότι γενικά, οι εργασίες επεξεργασίας φυσικής γλώσσας εξαρτώνται από το πλαίσιο στο οποίο είναι γραμμένα τα εξεταζόμενα κείμενα. Συνεπώς, είναι πιθανό τα μοντέλα που έχουν εκπαιδευτεί με κείμενα της Wikipedia να μην αποδίδουν τόσο καλά με κείμενα από μέσα κοινωνικής δικτύωσης, όπως αυτά που θα μελετήσουμε. Ωστόσο, στην παρούσα εργασία θα χρειαστούμε μόνο τις βασικές λειτουργίες της διάσπασης σε προτάσεις και λέξεις και της λημματοποίησης, οπότε λογικά δε θα συναντήσουμε κάποιο πρόβλημα. Πιο συγκεκριμένα, στο pipeline της spaCy περιέχονται επισημαντής μερών του λόγου (POS tagger), συντακτικός/γραμματικός αναλυτής εξαρτήσεων (parser) και αναγνωριστής ονοματισμένων οντοτήτων (ner) αλλά όχι αναλυτής συναισθήματος.

Επομένως, θα χρειαστεί να κατασκευάσουμε το δικό μας αναλυτή συναισθήματος. Όπως αναφέραμε και παραπάνω, υπάρχουν δύο τρόποι για να επιτευχθεί αυτό. Ο πρώτος είναι με εκπαίδευση ενός ταξινομητή με μεθόδους μηχανικής μάθησης και ο δεύτερος είναι με ένα προκαθορισμένο λεξικό συναισθήματος. Στην εργασία μας, θα ακολουθήσουμε το δεύτερο τρόπο και θα εκμεταλλευτούμε ένα λεξικό συναισθήματος[2] στα Ιταλικά που βρήκαμε στο kaggle. Αυτό αποτελείται από δύο αρχεία. Το πρώτο περιέχει ιταλικές λέξεις με «θετικό» πρόσημο ενώ το δεύτερο ιταλικές λέξεις με «αρνητικό» πρόσημο. Το τελικό λεξικό, που θα χρησιμοποιήσουμε, δημιουργείται ως στιγμιότυπο της κλάσης Lexicon του οποίου όλες οι καταχωρήσεις είναι τα λήμματα των παραπάνω λέξεων με την αποτίμηση 1 αν ανήκουν στο πρώτο αρχείο και την αποτίμηση -1

[2] https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages

αν ανήκουν στο δεύτερο. Τελικά, για να βρούμε την πολικότητα ενός tweet το διασπάμε πρώτα σε προτάσεις και έπειτα την κάθε πρόταση σε λέξεις. Βρίσκουμε το λήμμα της κάθε λέξης και αναζητούμε αν αυτό είναι καταχωρημένο στο λεξικό. Αν ναι προσθέτουμε την αποτίμησή του στη συνολική αποτίμηση της πρότασης. Λαμβάνεται ειδική μέριμνα στην περίπτωση που η πρόταση περιέχει άρνηση, οπότε και αντιστρέφεται το πρόσημό της συνολικής της αποτίμησης. Τελικά, ως πολικότητα μιας δημοσίευσης προκύπτει το άθροισμα των αποτιμήσεων των επιμέρους προτάσεων προς το συνολικό πλήθος των συναισθηματικά φορτισμένων λέξεων που αυτές περιείχαν και ανήκει στο εύρος $[-1, 1]$.

Έπειτα, από αυτήν τη διαδικασία, είμαστε σε θέση να εξάγουμε τα παρακάτω χαρακτηριστικά, τα οποία σε αντίθεση με τα προηγούμενα αναφέρονται σε ποιοτικούς δείκτες και έχουν άμεση σχέση με το περιεχόμενο των ενεργειών, την αιτία τους και τον αντίκτυπό τους.

- **Συνολική αποτίμηση συναισθήματος**

  Αυτό το χαρακτηριστικό (SentScore) αντιστοιχεί στο άθροισμα των πολικοτήτων των αναρτήσεων που πραγματοποιήθηκαν στο υπό μελέτη χρονικό διάστημα.

- **Πλήθος θετικών αναρτήσεων**

  Αυτό το χαρακτηριστικό (PosTweets) προκύπτει από την καταμέτρηση των tweets που έχουν πολικότητα με θετικό πρόσημο και δημοσιεύθηκαν στο υπό μελέτη χρονικό διάστημα.

- **Πλήθος αρνητικών αναρτήσεων**

  Αντίστοιχα με το προηγούμενο χαρακτηριστικό, το πλήθος των αρνητικών αναρτήσεων (NegTweets) προκύπτει από την καταμέτρηση των αναρτήσεων με αρνητική πολικότητα που πραγματοποιήθηκαν στο χρονικό παράθυρο που μελετάται.

- **Πλήθος ουδέτερων αναρτήσεων**

  Το συγκεκριμένο χαρακτηριστικό (NeutTweets) αντιστοιχεί στο πλήθος των αναρτήσεων με ακριβώς μηδενική πολικότητα που πραγματοποιήθηκαν στο χρονικό παράθυρο που μελετάται.

- **Πλήθος θετικών και αρνητικών αναρτήσεων**

  Σε αντίθεση με το προηγούμενο χαρακτηριστικό, αυτό (Pos&NegTweets) αντιστοιχεί στο πλήθος των αναρτήσεων με μη μηδενική πολικότητα, εκείνων δηλαδή που είναι συναισθηματικά φορτισμένες είτε θετικά είτε αρνητικά και πραγματοποιήθηκαν στο χρονικό παράθυρο που μελετάται.

- **Μέση αποτίμηση συναισθήματος**

  Το παρόν χαρακτηριστικό (AvgSentScore) ορίζεται ως ο λόγος της συνολικής αποτίμησης συναισθήματος, του πρώτου χαρακτηριστικού που αναφέραμε σε αυτήν την κατηγορία, προς το πλήθος των αναρτήσεων που πραγματοποιήθηκαν στο εν λόγω χρονικό παράθυρο.

- **Λόγος θετικών αναρτήσεων προς συνολικές αναρτήσεις**

  Σε αυτή την περίπτωση συνδυάζονται με μη γραμμικό τρόπο δύο παραπάνω χαρακτηριστικά και ορίζεται ο λόγος του πλήθους των θετικών προς το πλήθος των συνολικών αναρτήσεων (PosTweetsRatio) σε ένα συγκεκριμένο χρονικό παράθυρο.

- **Λόγος αρνητικών αναρτήσεων προς συνολικές αναρτήσεις**

  Ομοίως με το προηγούμενο, ορίζεται και το εν λόγω χαρακτηριστικό αλλά αυτή τη φορά για το πλήθος των αρνητικών δημοσιεύσεων (NegTweetsRatio)

- **Λόγος θετικών και αρνητικών αναρτήσεων προς συνολικές αναρτήσεις**

  Αντίστοιχα, ορίζεται ο λόγος με αριθμητή το πλήθος των συναισθηματικά φορτισμένων αναρτήσεων, είτε θετικά είτε αρνητικά (Pos&NegTweetsRatio).

- **Λόγος ουδέτερων αναρτήσεων προς συνολικές αναρτήσεις**

  Τέλος, ορίζεται ο λόγος του πλήθους των ουδέτερων προς τις συνολικές αναρτήσεις που πραγματοποιήθηκαν στο χρονικό διάστημα που μελετάται (NeutTweetsRatio). Το συγκεκριμένο χαρακτηριστικό είναι γραμμικώς εξαρτημένο με το προηγούμενο, καθώς:

$$\text{NeutTweetsRatio} = \frac{\text{NeutTweets}}{\text{Tweets}} \Leftrightarrow \tag{5.3}$$

$$\text{NeutTweetsRatio} = \frac{\text{Tweets} - \text{Pos\&NegTweets}}{\text{Tweets}} = 1 - \frac{\text{Pos\&NegTweets}}{\text{Tweets}} \Leftrightarrow \tag{5.4}$$

$$\text{NeutTweetsRatio} = 1 - \text{Pos\&NegTweetsRatio} \tag{5.5}$$

  Συνεπώς, δεν περιέχει κάποια παραπάνω χρήσιμη πληροφορία για την κατασκευή των μοντέλων πρόβλεψης στη συνέχεια, οπότε και δε θα ασχοληθούμε περεταίρω με αυτό.

5.3.1.3 Καθορισμός χρονικών παραθύρων

Όπως είδαμε, για τα δεδομένα που συλλέχθηκαν από το Twitter, έχουμε στη διάθεσή μας την ακριβή ώρα δημοσίευσης κάθε tweet. Αυτό μας δίνει τη δυνατότητα να εξάγουμε χαρακτηριστικά από αυτά σε οποιοδήποτε χρονικό παράθυρο, σε επίπεδο ωρών, ημερών κλπ. χωρίς ιδιαίτερους περιορισμούς. Όσον αφορά τα ωριαία δεδομένα, παρατηρούμε ότι ο όγκος των δημοσιεύσεων είναι ιδιαίτερα αυξημένος κατά τη διάρκεια προβολής της εκπομπής, καθώς και λίγο νωρίτερα και λίγο αργότερα από αυτήν. Κατά συνέπεια, ένα χρονικό παράθυρο που θα επιλεγεί είναι από μισή ώρα πριν έως και μισή ώρα μετά την προβολή του κάθε επεισοδίου. Αυτό το χρονικό παράθυρο είχε μελετηθεί ως καταλληλότερο και στη διπλωματική εργασία του Χρήστου Πιερράκου [65], που βασιζόταν στο ίδιο σύνολο δεδομένων αναφοράς αλλά συναντήθηκε και στη συναφή βιβλιογραφία.

Επιπλέον, ομαδοποιούμε τις δημοσιεύσεις σε ημερήσια βάση, για να μπορέσουμε να αποφασίσουμε πιο χρονικό παράθυρο είναι καταλληλότερο σε επίπεδο ημερών. Από τις γραφικές παραστάσεις που παρουσιάσαμε παραπάνω σε ημερήσια βάση κάποιων χαρακτηριστικών, παρατηρούμε ότι ανήμερα της προβολής του κάθε επεισοδίου αλλά και την επόμενη και ελαφρώς την προηγούμενη μέρα από αυτήν, υπάρχει μεγάλη αύξηση σε αυτά. Συνεπώς, επιλέγονται επιπλέον τα ακόλουθα χρονικά παράθυρα: μόνο η μέρα προβολής του εκάστοτε επεισοδίου, η μέρα προβολής και η επόμενη αυτής και τέλος, η μέρα προβολής, η προηγούμενη και η επόμενή της. Επίσης, θα δοκιμαστούν και ευρύτερα παράθυρα, όπως από τη δεύτερη επόμενη μέρα της προβολής του προηγούμενου επεισοδίου έως και την επόμενη μέρα της προβολής του υπό εξέταση επεισοδίου αλλά και από τη μέρα προβολής του υπό εξέταση επεισοδίου έως και την προηγούμενη από το επόμενο επεισόδιο.

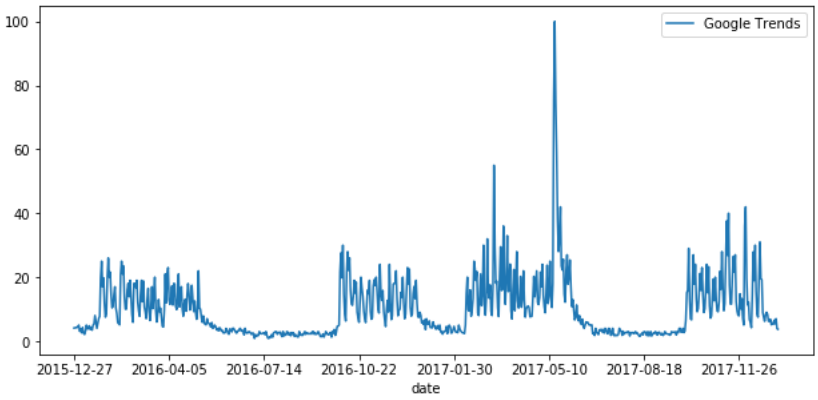
Σε επόμενη ενότητα, θα συγκρίνουμε αυτά τα 6 χρονικά παράθυρα μεταξύ τους και θα αποφανθούμε ποιο φαίνεται καταλληλότερο για την εκτίμηση της τηλεθέασης του ιταλικού προγράμματος "Le Iene".

5.3.2 Εξαγωγή χαρακτηριστικών από το Google Trends

Από την πλατφόρμα Google Trends μπορούμε να εξάγουμε μόνο ένα χαρακτηριστικό, που δεν είναι άλλο από το κανονικοποιημένο πλήθος αναζητήσεων στη μηχανή αναζήτησης της Google (GTrends) σε διάφορα χρονικά παράθυρα. Παρακάτω, φαίνεται η μεταβολή του ανά ημέρα, έπειτα από την επανακλιμάκωση που εφαρμόστηκε, για το σύνολο των 2 ετών που μελετάμε.

5.3.2.1 Καθορισμός χρονικών παραθύρων

Επειδή, η πλατφόρμα δε μας δίνει τη δυνατότητα να λάβουμε παλιά δεδομένα σε επίπεδο ωρών, τα χρονικά παράθυρα δε γίνεται να είναι μικρότερα από μια ολόκληρη μέρα. Από τα δεδομένα που συλλέξαμε και την οπτικοποίησή τους, παρατηρούμε ότι υπάρχει μεγάλη αύξηση στις αναζητήσεις τη μέρα της προβολής κάθε επεισοδίου και την επόμενη αυτής. Στη συντριπτική πλειοψηφία των



Σχήμα 5.8: Κανονικοποιημένα αποτελέσματα από το Google Trends ανά ημέρα

περιπτώσεων μάλιστα, την επόμενη μέρα από την προβολή του εκάστοτε επεισοδίου υπήρχαν οι περισσότερες αναζητήσεις σχετικά με το show. Με βάση αυτές τις παρατηρήσεις θα προσδιορίσουμε χρονικά παράθυρα για τον ορισμό του συγκεκριμένου χαρακτηριστικού.

Πιο αναλυτικά, επιλέγονται τα εξής χρονικά παράθυρα: μόνο η μέρα προβολής, η μέρα προβολής και η επόμενη, μόνο η επόμενη μέρα από την προβολή, η μέρα προβολής, η προηγούμενη και η επόμενη αυτής, από τη δεύτερη επόμενη μέρα του προηγούμενου επεισοδίου μέχρι και την επόμενη μέρα προβολής του υπό εξέταση, από τη μέρα προβολής του υπό εξέταση επεισοδίου μέχρι και την προηγούμενη μέρα από το επόμενο επεισόδιο. Σε επόμενη ενότητα, θα συγκρίνουμε αυτά τα 6 χρονικά παράθυρα μεταξύ τους και θα αποφανθούμε ποιο φαίνεται καταλληλότερο για την εκτίμηση των εξαρτημένων μεταβλητών στο πρόβλημά μας.

5.3.3 Τελικό σύνολο δεδομένων και σχόλια

Όλα τα παραπάνω χαρακτηριστικά, 20 στο σύνολο τους, θα προστεθούν στο σύνολο δεδομένων, 6 φορές το καθένα, μία για το κάθε χρονικό παράθυρο. Τελικά, δηλαδή, στο σύνολο δεδομένων, πέρα από τις 5 αρχικές μεταβλητές, που εισάγαμε σε προηγούμενη ενότητα, προστίθενται άλλες 120 ανεξάρτητες μεταβλητές. Είναι σημαντικό να σημειώσουμε πως οι μεταβλητές εισόδου ονομάζονται έτσι χωρίς να σημαίνει απαραίτητα ότι είναι πραγματικά ανεξάρτητες μεταξύ τους. Μάλιστα, όπως είδαμε παραπάνω, υπάρχουν δύο ζεύγη γραμμικά εξαρτημένων χαρακτηριστικών. Από καθένα εξ αυτών θα κρατήσουμε μόνο το ένα χαρακτηριστικό, καθώς η πλεονάζουσα πληροφορία μόνο ζημιογόνα μπορεί να αποβεί για την κατασκευή μοντέλων πρόβλεψης στη συνέχεια. Άρα, τελικά, απομένουν 18 ανεξάρτητες μεταβλητές για κάθε χρονικό παράθυρο, ή 108 συνολικά. Οι τιμές και των 125 μεταβλητών αποθηκεύονται ως στήλες σε ένα αρχείο της μορφής csv. Τα δεδομένα ανακτώνται πολύ εύκολα και αποθηκεύονται προσωρινά στη δομή δεδομένων data frame που παρέχει η βιβλιοθήκη pandas, κάθε φορά που τα χρειαζόμαστε.

Επίσης, παρατηρήσαμε ότι τα ευρύτερα χρονικά παράθυρα από το προηγούμενο επεισόδιο ή μέχρι το επόμενο είναι προβληματικά λόγω της συχνότητας προβολής της εκπομπής που μελετάμε. Πιο αναλυτικά, όπως αναφέραμε και προηγουμένως, το πρόγραμμα "Le Iene" προβάλλεται στην τηλεόραση δύο φορές τη βδομάδα. Η μία μέρα προβολής είναι σταθερά η Κυριακή ενώ η δεύτερη αλλάζει ανά διαστήματα από Τρίτη σε Τετάρτη και Πέμπτη. Αυτό έχει ως συνέπεια να μη μεσολαβεί χρονικό διάστημα ίδιας διάρκειας κάθε φορά μεταξύ της προβολής δύο διαδοχικών επεισοδίων. Η ανομοιότητα στο εύρος των χρονικών διαστημάτων από επεισόδιο σε επεισόδιο συνεπάγεται, όπως είναι λογικό, αυξημένες τιμές ειδικά στα χαρακτηριστικά από ποσοτικούς δείκτες όταν μεσολαβούν περισσότερες μέρες μεταξύ δύο επεισοδίων (π.χ. Τρίτη - Κυριακή). Επιπλέον, πρέπει να ληφθεί ειδική μέριμνα για περιπτώσεις όπου επαναπροβάλλεται το show μετά από ένα διάστημα απουσίας, όπως στην καλοκαιρινή περίοδο.

Η απλούστερη λύση που σκεφτήκαμε για να καταπολεμήσουμε αυτό το φαινόμενο ήταν να κανονικοποιήσουμε τα χαρακτηριστικά που ορίζονται σε αυτά τα δύο χρονικά παράθυρα. Με άλλα λόγια, διαιρέσαμε την αρχική τιμή του κάθε χαρακτηριστικού με το πλήθος των ημερών του αντίστοιχου χρονικού διαστήματος. Όμως και πάλι το πρόβλημα δε λύθηκε. Μάλιστα, αντιστράφηκε, καθώς προέκυψαν πλασματικά μικρότερες τιμές για επεισόδια που προβάλλονταν μετά από μεγαλύτερο χρονικό διάστημα. Και στις δύο περιπτώσεις, λοιπόν, τα δεδομένα που συγκεντρώσαμε είναι παραπλανητικά και η χρήση τους στη συνέχεια θα αποφευχθεί όπου είναι δυνατόν. Το πρόβλημα θα μπορούσε να λυθεί πιθανώς, λαμβάνοντας ένα σταθμισμένο μέσο όρο των ημερήσιων δεδομένων, όπου οι μέρες που απέχουν περισσότερο από μέρα προβολής θα κλιμακώνονταν με μικρότερο συντελεστή από τις υπόλοιπες. Κάτι τέτοιο όμως δε θεωρήθηκε απαραίτητο στην παρούσα ανάλυση.

Τέλος, από εδώ και πέρα, σε όλη την υπόλοιπη εργασία το σύνολο δεδομένων θα διασπασθεί σε σύνολο δεδομένων εκπαίδευσης και σύνολο δεδομένων ελέγχου. Αυτή η διάσπαση υλοποιείται με τη βοήθεια της βιβλιοθήκης scikit-learn και πιο συγκεκριμένα της συνάρτησης train_test_split, στην οποία δίνονται ως ορίσματα οι στήλες των χαρακτηριστικών X, οι στήλες των εξαρτημένων μεταβλητών y, το μέγεθος του συνόλου δεδομένων ελέγχου ως ποσόστωση του συνολικού (0.1) και ένας τυχαίος ακέραιος (10), ο οποίος θα χρησιμοποιηθεί από τη γεννήτρια τυχαίων αριθμών ώστε να παράγεται πάντα η ίδια τυχαία διάσπαση. Ως αποτέλεσμα προκύπτουν δύο ξένα συμπληρωματικά σύνολα παρατηρήσεων με τυχαίο τρόπο, το σύνολο εκπαίδευσης με 96 παρατηρήσεις ($\approx 90\%$) και το σύνολο ελέγχου με 11 παρατηρήσεις ($\approx 10\%$). Οι συνήθεις αναλογίες που επιλέγονται είναι από $75\% - 25\%$ έως και $90\% - 10\%$. Στην εργασία μας, προτιμήθηκε αυτή η αναλογία, καθώς το σύνολο των παρατηρήσεων που έχουμε στη διάθεσή μας είναι 107 και όχι ιδιαίτερα μεγάλο, οπότε και θα εκμεταλλευτούμε όσες περισσότερες από αυτές γίνεται για εκπαίδευση ώστε να παράγουμε στιβαρά μοντέλα με περισσότερη «εμπειρία». Είναι πολύ σημαντικό από εδώ και στο εξής τα δείγματα που ανήκουν στο σύνολο ελέγχου να αντιμετωπίζονται ως τελείως άγνωστα και να μη ληφθούν υπόψη σε οποιαδήποτε επιλογή κάνουμε για τη βελτίωση της απόδοσης των μοντέλων πρόβλεψης. Με αυτόν τον τρόπο, θα αποφύγουμε την υπερεκπαίδευση και θα είμαστε πιο βέβαιοι για την ικανότητα των μοντέλων μας να γενικεύσουν.

5.4 Διερευνητική ανάλυση δεδομένων

Σε αυτό το σημείο, αφού έχουμε εξάγει όλα τα χαρακτηριστικά που θα χρησιμοποιήσουμε και έχουμε διασπάσει το σύνολο των δεδομένων μας σε εκπαίδευσης και ελέγχου, προτού προχωρήσουμε στην κατασκευή μοντέλων για την εκτίμηση της τηλεθέασης, θα προβούμε σε διερευνητική ανάλυση των δεδομένων (exploratory data analysis - EDA). Σε αυτή τη φάση χρησιμοποιείται ένα σύνολο στατιστικών και γραφικών μεθόδων, οι οποίες μας βοηθούν να αποκτήσουμε μια εικόνα για την κατανομή των τιμών των μεταβλητών, τη δομή τους, τις συσχετίσεις που μπορεί να υπάρχουν μεταξύ τους αλλά και για την ανίχνευση ανωμαλιών, την αναγνώριση προτύπων και γενικότερα τον έλεγχο υποθέσεων και παραδοχών.

5.4.1 Συσχετίσεις μεταξύ χαρακτηριστικών και εξαρτημένων μεταβλητών

Στην περίπτωσή μας, θα ελέγξουμε τις συσχετίσεις σε πρώτη φάση μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών για διάφορα χρονικά παράθυρα και σε δεύτερη φάση και μεταξύ των ανεξάρτητων μεταβλητών. Αυτές θα μετρηθούν με τους συντελεστές συσχέτισης Pearson και Spearman. Όπως αναλύσαμε παραπάνω, η πρώτη ποσοτικοποιεί τη γραμμική συσχέτιση ενώ η δεύτερη τη μονότονη σχέση μεταξύ δύο μεταβλητών. Και οι δύο πέρα από την απόλυτη τιμή που προσδιορίζει την ισχύ της συσχέτισης συνοδεύονται από το πρόσημο που προσδιορίζει την κατεύθυνσή της. Αυτά είναι τα δύο είδη συσχέτισης που θα μας απασχολήσουν στην παρούσα μελέτη, καθώς φαίνεται εύλογο τα νούμερα τηλεθέασης να αυξάνονται όσο περισσότερες αναζη-

τήσεις πραγματοποιούνται στη μηχανή αναζήτησης της Google και όσο μεγαλύτερο ενδιαφέρον δείχνουν οι χρήστες στο Twitter για ένα επεισόδιο του show.

Πιο αναλυτικά, αυτή η διαδικασία ισοδυναμεί με το να κάνουμε αρχικά την υπόθεση ότι υπάρχει συσχέτιση και έπειτα να ελέγχουμε αν αυτή ισχύει με ένα στατιστικό τεστ, του οποίου τα αποτελέσματα, δηλαδή οι τιμές των συντελεστών συσχέτισης που θα έχουν προκύψει, θα συνοδεύονται από την αντίστοιχη τιμή σημαντικότητας (p-value) που υποδεικνύει πόσο στατιστικά σημαντικά είναι αυτά. Με άλλα λόγια, η p-value εκφράζει την πιθανότητα να προκύψουν τουλάχιστον τόσο ακραία αποτελέσματα όταν η μηδενική υπόθεση (null hypothesis) είναι αληθής. Είναι επιθυμητό, λοιπόν, η p-value να παίρνει τιμές όσο πιο κοντά στο 0 γίνεται, για να μπορούμε να απορρίψουμε τη μηδενική υπόθεση, σύμφωνα με την οποία δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών που εξετάζουμε. Ορίζουμε εκ των προτέρων ως άνω όριο για αυτήν alpha $= 0.05$, κάτω από το οποίο θεωρούμε ότι έχει επιτευχθεί στατιστική σημαντικότητα.

Στους πίνακες 5.1, 5.2, 5.3 και 5.4 φαίνονται οι τιμές των συντελεστών συσχέτισης που προέκυψαν μεταξύ των εξαρτημένων μεταβλητών και όλων των χαρακτηριστικών ορισμένων στα 6 διαφορετικά χρονικά παράθυρα που προσδιορίσαμε παραπάνω, μαζί με τις αντίστοιχες τιμές σημαντικότητας. Συμβολίζουμε με r το συντελεστή συσχέτισης του Pearson και με ρ το συντελεστή συσχέτισης του Spearman. Επίσης, παρακάτω παραθέτουμε την αντιστοιχία αρίθμησης - χρονικών παραθύρων για τα χαρακτηριστικά που προέκυψαν από τις δύο πηγές δεδομένων.
Πρώτα, για το πλήθος των αναζητήσεων στη μηχανή αναζήτησης της Google:

1. μόνο η μέρα προβολής κάθε επεισοδίου
2. η μέρα προβολής κάθε επεισοδίου και η επόμενή της
3. μόνο η επόμενη μέρα από την προβολή του κάθε επεισοδίου
4. η μέρα προβολής κάθε επεισοδίου, η προηγούμενη και η επόμενη της
5. από τη δεύτερη επόμενη μέρα του προηγούμενου επεισοδίου μέχρι και την επόμενη μέρα από την προβολή του υπό εξέταση επεισοδίου
6. από τη μέρα προβολής του υπό εξέταση επεισοδίου μέχρι και την προηγούμενη μέρα από το επόμενο επεισόδιο

Έπειτα, για όλα τα χαρακτηριστικά που προέκυψαν από τα δεδομένα που αντλήθηκαν από το Twitter:

1. μόνο η μέρα προβολής κάθε επεισοδίου
2. η μέρα προβολής κάθε επεισοδίου και η επόμενή της
3. μισή ώρα πριν έως και μισή ώρα μετά την προβολή κάθε επεισοδίου
4. η μέρα προβολής κάθε επεισοδίου, η προηγούμενη και η επόμενη της
5. από τη δεύτερη επόμενη μέρα του προηγούμενου επεισοδίου μέχρι και την επόμενη μέρα από την προβολή του υπό εξέταση επεισοδίου
6. από τη μέρα προβολής του υπό εξέταση επεισοδίου μέχρι και την προηγούμενη μέρα από το επόμενο επεισόδιο

Συγκρίνοντας τα αποτελέσματα των συσχετίσεων για διάφορα χρονικά παράθυρα για κάθε χαρακτηριστικό, μπορούμε να αποφανθούμε ποιο από αυτά φαίνεται καταλληλότερο για την εκτίμηση των εξαρτημένων μεταβλητών. Στους πίνακες, επισημαίνουμε με έντονη γραμματοσειρά το χρονικό παράθυρο στο οποίο επιτεύχθηκε η υψηλότερη συσχέτιση για κάθε χαρακτηριστικό και με πράσινο χρώμα αν το αποτέλεσμα θεωρείται στατιστικά σημαντικό και η μηδενική υπόθεση μπορεί να απορριφθεί, διαφορετικά με κόκκινο. Επίσης, μπορούμε να συγκρίνουμε τις καλύτερες

τιμές των χαρακτηριστικών μεταξύ τους, για να αποφανθούμε ποια θα διαδραματίσουν περισσότερο ευεργετικό ρόλο στα μοντέλα μας και ποια πιθανώς θα ήταν καλύτερο να μη χρησιμοποιηθούν λόγω πενιχρής συσχέτισης με τις εξαρτημένες μεταβλητές. Στη συνέχεια, συνοψίζουμε τις παρατηρήσεις μας σχετικά με τα αποτελέσματα που προέκυψαν.

Χαρακτηριστικά		**Χρονικά παράθυρα**					
		1	**2**	**3**	**4**	**5**	**6**
GTrends	**r**	0.03	0.17	**0.25**	0.11	0.09	0.16
	p-value	0.80	0.10	**0.01**	0.31	0.41	0.12
	ρ	0.15	0.30	**0.36**	0.23	0.22	0.31
	p-value	0.15	0.00	**0.00**	0.02	0.03	0.00
Tweets	**r**	0.14	0.16	0.14	**0.17**	0.15	0.17
	p-value	0.16	0.12	0.18	**0.10**	0.15	0.10
	ρ	0.17	0.18	0.16	0.16	0.17	**0.19**
	p-value	0.10	0.09	0.11	0.11	0.10	**0.07**
UnqUsers	**r**	0.15	0.18	0.15	**0.20**	0.15	0.20
	p-value	0.14	0.08	0.14	**0.05**	0.16	0.06
	ρ	0.16	0.17	0.16	**0.17**	0.15	0.17
	p-value	0.12	0.11	0.13	**0.09**	0.14	0.10
Retweets	**r**	0.22	0.23	0.16	0.23	0.22	**0.24**
	p-value	0.03	0.02	0.11	0.02	0.03	**0.02**
	ρ	0.07	0.06	0.07	0.07	0.06	**0.08**
	p-value	0.53	0.56	0.49	0.47	0.59	**0.44**
Favorites	**r**	0.40	0.41	0.34	0.41	0.39	**0.42**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
	ρ	0.24	0.24	0.25	0.25	0.24	**0.27**
	p-value	0.02	0.02	0.01	0.01	0.02	**0.01**
Tweets&Retweets	**r**	0.22	0.23	0.17	0.24	0.22	**0.24**
	p-value	0.03	0.02	0.10	0.02	0.03	**0.02**
	ρ	0.11	0.10	**0.12**	0.10	0.10	0.11
	p-value	0.30	0.34	**0.26**	0.31	0.35	0.27
FavoritesRatio	**r**	0.37	0.36	0.33	0.36	0.35	**0.38**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
	ρ	0.27	0.27	0.29	0.26	0.25	**0.29**
	p-value	0.01	0.01	0.00	0.01	0.01	**0.00**
RetweetsRatio	**r**	0.12	0.11	0.07	0.12	0.11	**0.13**
	p-value	0.23	0.27	0.47	0.24	0.30	**0.20**
	ρ	**0.04**	0.01	0.04	0.03	0.00	0.03
	p-value	**0.68**	0.96	0.70	0.81	1.00	0.75
UnqUsersRatio	**r**	**0.11**	0.09	0.06	0.07	0.11	0.09
	p-value	**0.28**	0.41	0.55	0.51	0.30	0.45
	ρ	0.11	0.09	0.05	0.08	**0.12**	0.10
	p-value	0.28	0.38	0.61	0.46	**0.26**	0.32

Πίνακας 5.1: Μετρικές συσχέτισης χαρακτηριστικών ποσοτικών δεικτών με SHR%

Κατ΄ αρχάς, όσον αφορά το κανονικοποιημένο πλήθος αναζητήσεων στη μηχανή αναζήτησης της Google, το χρονικό παράθυρο 3, που αναφέρεται μόνο στην επόμενη μέρα από την προβολή του εκάστοτε επεισοδίου φαίνεται ξεκάθαρα καλύτερο από τα υπόλοιπα για την εκτίμηση του ποσοστού τηλεθέασης επί των ανοικτών δεκτών. Και στις δύο μετρικές συσχέτισης επιτυγχάνονται ικανοποιητικά ψηλές τιμές (0.25 και 0.36), με στατιστική σημαντικότητα, οπότε και η μηδενική υπόθεση μπορεί να απορριφθεί με ασφάλεια. Για τη συσχέτιση αυτού του χαρακτηριστικού με το

απόλυτο πλήθος τηλεθεατών, μόνο η μετρική συσχέτισης του Spearman εμφανίζει ικανοποιητικά υψηλή τιμή (0.24) με στατιστική σημαντικότητα για το χρονικό παράθυρο 6, που αντιστοιχεί στο διάστημα από τη μέρα προβολής τους εκάστοτε επεισοδίου έως και την προηγούμενη μέρα από την προβολή του επόμενου. Αμέσως μετά ακολουθεί το χρονικό παράθυρο 3 για αυτόν τον συντελεστή, το οποίο μάλιστα είναι το καλύτερο για το συντελεστή συσχέτισης Pearson, χωρίς όμως να είναι δυνατή η απόρριψη της μηδενικής υπόθεσης.

Χαρακτηριστικά		**Χρονικά παράθυρα**					
		1	**2**	**3**	**4**	**5**	**6**
SentScore	**r**	-0.30	-0.31	-0.26	**-0.34**	-0.32	-0.32
	p-value	0.00	0.00	0.01	**0.00**	0.00	0.00
	ρ	-0.27	-0.28	-0.22	-0.29	**-0.29**	-0.29
	p-value	0.01	0.01	0.03	0.00	**0.00**	0.00
AvgSentScore	**r**	-0.32	-0.33	-0.29	-0.33	**-0.34**	-0.32
	p-value	0.00	0.00	0.00	0.00	**0.00**	0.00
	ρ	-0.27	-0.26	-0.25	-0.25	**-0.28**	-0.26
	p-value	0.01	0.01	0.01	0.01	**0.01**	0.01
PosTweets	**r**	0.12	0.13	0.12	0.13	0.12	**0.14**
	p-value	0.24	0.20	0.24	0.22	0.26	**0.18**
	ρ	0.15	0.14	**0.16**	0.14	0.12	0.15
	p-value	0.15	0.18	**0.12**	0.18	0.25	0.14
NegTweets	**r**	0.22	0.24	0.21	**0.26**	0.23	0.25
	p-value	0.03	0.02	0.04	**0.01**	0.02	0.01
	ρ	0.22	0.23	0.21	**0.25**	0.23	0.24
	p-value	0.03	0.02	0.04	**0.01**	0.02	0.02
NeutTweets	**r**	0.09	0.11	0.08	**0.12**	0.10	0.12
	p-value	0.39	0.28	0.44	**0.25**	0.34	0.25
	ρ	0.13	0.15	0.12	0.14	0.14	**0.15**
	p-value	0.22	0.16	0.24	0.16	0.18	**0.16**
Pos&NegTweets	**r**	0.18	0.20	0.18	**0.21**	0.19	0.21
	p-value	0.07	0.05	0.08	**0.04**	0.07	0.04
	ρ	0.20	0.20	0.20	0.20	0.20	**0.21**
	p-value	0.06	0.05	0.05	0.05	0.05	**0.04**
PosTweetsRatio	**r**	-0.08	-0.11	-0.05	**-0.14**	-0.13	-0.12
	p-value	0.44	0.27	0.66	**0.18**	0.22	0.25
	ρ	-0.10	-0.11	-0.08	-0.10	**-0.13**	-0.11
	p-value	0.34	0.27	0.42	0.32	**0.20**	0.29
NegTweetsRatio	**r**	0.43	0.43	0.42	0.43	**0.44**	0.43
	p-value	0.00	0.00	0.00	0.00	**0.00**	0.00
	ρ	0.34	0.35	0.34	0.35	**0.36**	0.34
	p-value	0.00	0.00	0.00	0.00	**0.00**	0.00
Pos&NegTweetsRatio	**r**	**0.40**	0.37	0.12	0.34	0.36	0.36
	p-value	**0.00**	0.00	0.24	0.00	0.00	0.00
	ρ	**0.29**	0.27	0.16	0.25	0.25	0.28
	p-value	**0.00**	0.01	0.12	0.01	0.01	0.01

Πίνακας 5.2: Μετρικές συσχέτισης χαρακτηριστικών από ανάλυση συναισθήματος με SHR%

Τέλος, επειδή όπως αναφέραμε παραπάνω θα προσπαθήσουμε να αποφεύγουμε τα χρονικά παράθυρα 5 και 6 όπου είναι δυνατόν, θα κρατήσουμε το χρονικό παράθυρο 3 ως καταλληλότερο για την εκτίμηση των εξαρτημένων μεταβλητών SHR% και AMR με το χαρακτηριστικό GTrends. Αξίζει να σημειώσουμε ότι και στις δύο περιπτώσεις φαίνονται πιο χρήσιμα τα χρονικά διαστήματα

μετά από την προβολή του υπό εξέταση επεισοδίου και όχι πριν. Αυτό μπορεί να ερμηνευθεί διαισθητικά από το γεγονός ότι πριν την προβολή ενός επεισοδίου οι δημοσιεύσεις έχουν διαφημιστικό κυρίως χαρακτήρα ενώ μετά σχολιαστικό. Είναι λογικό, λοιπόν, ο σχολιασμός που ακολουθεί ένα επεισόδιο να κεντρίζει περισσότερο το ενδιαφέρον του κοινού, το οποίο τον αποζητά με σχετικές αναζητήσεις στη μηχανή αναζήτησης της Google.

Χαρακτηριστικά		**Χρονικά παράθυρα**					
		1	**2**	**3**	**4**	**5**	**6**
GTrends	**r**	-0.04	0.04	**0.10**	-0.03	-0.03	0.08
	p-value	0.67	0.69	**0.34**	0.77	0.74	0.43
	ρ	0.09	0.15	0.15	0.03	0.02	**0.24**
	p-value	0.38	0.15	0.15	0.76	0.86	**0.02**
Tweets	**r**	0.41	0.42	0.39	**0.43**	0.39	0.41
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	**0.45**	0.45	0.43	0.45	0.43	0.44
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
UnqUsers	**r**	0.43	0.43	0.41	**0.44**	0.36	0.41
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	**0.47**	0.46	0.45	0.46	0.43	0.43
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
Retweets	**r**	0.35	0.35	0.31	0.35	0.33	**0.36**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
	ρ	**0.32**	0.31	0.28	0.32	0.30	0.31
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
Favorites	**r**	0.48	0.48	0.45	0.48	0.46	**0.49**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
	ρ	0.46	0.46	0.44	0.46	0.46	**0.47**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
Tweets&Retweets	**r**	0.42	0.42	0.38	0.42	0.39	**0.43**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
	ρ	**0.39**	0.37	0.37	0.37	0.37	0.37
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
FavoritesRatio	**r**	0.31	0.30	0.26	0.30	0.29	**0.32**
	p-value	0.00	0.00	0.01	0.00	0.00	**0.00**
	ρ	0.25	**0.26**	0.23	0.25	0.23	0.25
	p-value	0.01	**0.01**	0.03	0.01	0.02	0.01
RetweetsRatio	**r**	0.13	0.12	0.07	0.12	0.10	**0.13**
	p-value	0.22	0.26	0.47	0.26	0.31	**0.21**
	ρ	**0.13**	0.10	0.09	0.10	0.08	0.11
	p-value	**0.21**	0.32	0.36	0.33	0.42	0.27
UnqUsersRatio	**r**	0.43	0.43	0.41	**0.44**	0.36	0.41
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	**0.47**	0.46	0.45	0.46	0.43	0.43
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00

Πίνακας 5.3: Μετρικές συσχέτισης χαρακτηριστικών ποσοτικών δεικτών με AMR

Όσον αφορά τα χαρακτηριστικά που εξάγαμε από το Twitter, τα πράγματα είναι λιγότερο ξεκάθαρα. Σε γενικές γραμμές τα αποτελέσματα για τα διάφορα χρονικά παράθυρα δεν εμφανίζουν αξιοσημείωτες διαφορές μεταξύ τους, οπότε και δε γίνεται εμφανές ποιο από αυτά πρέπει να θεωρηθεί το καταλληλότερο για τη μελέτη που πρόκειται πραγματοποιήσουμε. Στο σύνολο των αποτελεσμάτων, τα παράθυρα που εμφάνισαν τις μεγαλύτερες συσχετίσεις χαρακτηριστικού-

εξαρτημένης μεταβλητής ήταν το 6 και το 4. Το 6 αφορά το χρονικό διάστημα από τη μέρα προβολής του υπό εξέταση επεισοδίου έως και την προηγούμενη μέρα από το επόμενο επεισόδιο ενώ το 4 το τριήμερο από την προηγούμενη έως και την επόμενη μέρα της προβολής του εκάστοτε επεισοδίου. Έπειτα, με διαφορά ακολουθεί το απλούστερο παράθυρο 1, που αντιστοιχεί μόνο στη μέρα προβολής του εκάστοτε επεισοδίου. Στα χρονικά παράθυρα 2,3 και 5 εμφανίστηκαν πολύ μικρότερες συσχετίσεις, οδηγώντας μας στο συμπέρασμα ότι είναι πιο ακατάλληλα για την παραγωγή μοντέλων που θα ακολουθήσει.

Χαρακτηριστικά		**Χρονικά παράθυρα**					
		1	**2**	**3**	**4**	**5**	**6**
SentScore	**r**	-0.34	-0.35	-0.30	**-0.37**	-0.35	-0.36
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	-0.32	-0.33	-0.31	-0.34	-0.34	**-0.36**
	p-value	0.00	0.00	0.00	0.00	0.00	**0.00**
AvgSentScore	**r**	-0.17	-0.19	-0.14	-0.21	-0.20	**-0.21**
	p-value	0.09	0.07	0.18	0.04	0.06	**0.04**
	ρ	-0.12	-0.12	-0.11	-0.11	**-0.14**	-0.14
	p-value	0.26	0.24	0.28	0.28	**0.16**	0.17
PosTweets	**r**	0.41	**0.41**	0.40	0.40	0.38	0.40
	p-value	0.00	**0.00**	0.00	0.00	0.00	0.00
	ρ	**0.45**	0.44	0.43	0.43	0.41	0.42
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
NegTweets	**r**	0.45	0.45	0.42	**0.47**	0.43	0.45
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	0.47	0.47	0.45	**0.49**	0.46	0.47
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
NeutTweets	**r**	0.37	0.38	0.34	**0.39**	0.35	0.37
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	0.43	0.43	0.41	**0.43**	0.41	0.42
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
Pos&NegTweets	**r**	0.44	0.44	0.42	**0.45**	0.42	0.44
	p-value	0.00	0.00	0.00	**0.00**	0.00	0.00
	ρ	**0.47**	0.47	0.46	0.47	0.46	0.46
	p-value	**0.00**	0.00	0.00	0.00	0.00	0.00
PosTweetsRatio	**r**	-0.01	-0.05	0.03	-0.08	-0.06	**-0.08**
	p-value	0.90	0.66	0.76	0.44	0.58	**0.44**
	ρ	-0.01	-0.02	0.01	-0.00	-0.04	**-0.06**
	p-value	0.95	0.86	0.93	0.96	0.67	**0.55**
NegTweetsRatio	**r**	0.26	0.27	0.25	**0.28**	0.27	0.28
	p-value	0.01	0.01	0.01	**0.01**	0.01	0.01
	ρ	0.16	0.19	0.17	0.18	**0.19**	0.19
	p-value	0.11	0.07	0.09	0.07	**0.06**	0.06
Pos&NegTweetsRatio	**r**	0.27	0.25	**0.40**	0.23	0.24	0.23
	p-value	0.01	0.01	**0.00**	0.02	0.02	0.02
	ρ	0.17	0.16	**0.44**	0.16	0.14	0.14
	p-value	0.10	0.13	**0.00**	0.12	0.18	0.17

Πίνακας 5.4: Μετρικές συσχέτισης χαρακτηριστικών από ανάλυση συναισθήματος με AMR

Τα χαρακτηριστικά που εμφάνισαν υψηλότερη συσχέτιση με τη μεταβλητή εξόδου SHR% είναι τα εξής: NegTweetsRatio, Pos&NegTweetsRatio, Favorites, SentScore, AvgSentScore, FavoritesRatio και NegTweets. Όλα αυτά παρουσίασαν συντελεστές συσχέτισης κατ΄ απόλυτη

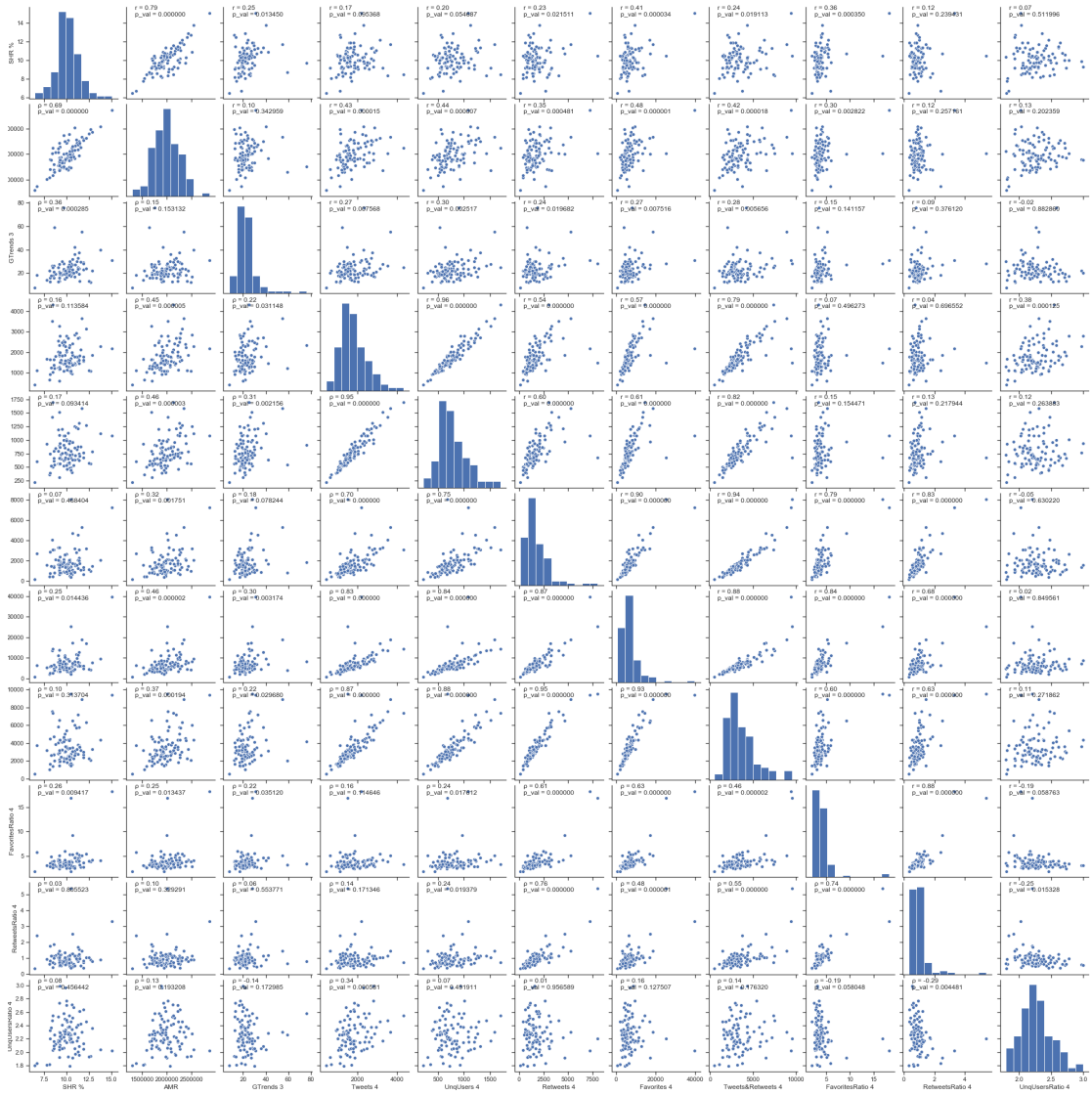
τιμή 0.25 - 0.44 συνοδευόμενους από πολύ μικρές τιμές p-value, οπότε και η μηδενική υπόθεση μπορεί να απορριφθεί με ασφάλεια. Σε αυτό το φάσμα των, φαινομενικά τουλάχιστον, καταλληλότερων χαρακτηριστικών ανήκει και το GTrends. Ακολουθούν τα χαρακτηριστικά UnqUsers, Retweets, Tweets&Retweets και Pos&NegTweets, τα οποία παρουσίασαν συσχέτιση σύμφωνα με το συντελεστή Pearson με την εξαρτημένη μεταβλητή SHR% 0.20 - 0.24 και τιμή σημαντικότητας τέτοια ώστε να μπορεί να απορριφθεί η μηδενική υπόθεση. Ωστόσο με την ίδια εξαρτημένη μεταβλητή εμφάνισαν συσχέτιση σύμφωνα με τη μετρική Spearman 0.08 - 0.21, μην τηρώντας το όριο της στατιστικής σημαντικότητας τις περισσότερες φορές. Τέλος, τα χαρακτηριστικά Tweets, RetweetsRatio, UnqUsersRatio, PosTweets, NeutTweets και PosTweetsRatio εμφάνισαν τιμές συσχέτισης -0.14 - 0.19 και η μηδενική υπόθεση δεν μπορεί να απορριφθεί σε καμία εξ αυτών των περιπτώσεων.

Όσον αφορά τη μεταβλητή εξόδου AMR, τα χαρακτηριστικά που εμφάνισαν την υψηλότερη συσχέτιση με αυτήν είναι τα εξής: Tweets, UnqUsers, Retweets, Favorites, Tweets&Retweets, FavoritesRatio, UnqUsersRatio, SentScore, PosTweets, NegTweets, NeutTweets, Pos&Neg Tweets και Pos&NegTweetsRatio. Όλα αυτά παρουσίασαν συντελεστές συσχέτισης κατ΄ απόλυτη τιμή 0.26 - 0.49, συνοδευόμενους από σχεδόν μηδενικές τιμές p-value, οπότε και η μηδενική υπόθεση μπορεί να απορριφθεί με ασφάλεια. Ακολουθούν τα χαρακτηριστικά RetweetsRatio, AvgSentScore, PosTweetsRatio και NegTweetsRatio, τα οποία παρουσίασαν τιμές συντελεστών συσχέτισης κατ΄ απόλυτη τιμή 0.06 - 0.28 και τιμή σημαντικότητας τέτοια ώστε να μην μπορεί να απορριφθεί η μηδενική υπόθεση. Η συσχέτιση που εμφανίζει το χαρακτηριστικό GTrends με τη μεταβλητή εξόδου AMR ανήκει σε αυτό το εύρος ενώ η μηδενική υπόθεση για αυτό μπορεί να απορριφθεί μόνο στην περίπτωση του συντελεστή συσχέτισης Spearman.

Καταλήγοντας, για την εκτίμηση της εξαρτημένης μεταβλητής SHR% φαίνονται πιο χρήσιμα τα χαρακτηριστικά που προέκυψαν από συναισθηματική ανάλυση στα tweets και ιδιαίτερα αυτά που σχετίζονται με αρνητικό συναίσθημα ή γενικότερα με συναισθηματική φόρτιση ανεξαρτήτως προσήμου και τα σχετικά σκορ. Από χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες μόνο το πλήθος των αναζητήσεων από το Google Trends και το πλήθος και ο λόγος των επισημάνσεων «Μου αρέσει» συσχετίζονται σε αξιοσημείωτο βαθμό με την προς εκτίμηση μεταβλητή. Σε αντίθεση, στην περίπτωση της εκτίμησης του απόλυτου πλήθους των τηλεθεατών AMR τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες από το Twitter φαίνονται πιο κατάλληλα εμφανίζοντας υψηλές συσχετίσεις με το προς εκτίμηση μέγεθος, με μοναδική εξαίρεση το λόγο των αναδημοσιεύσεων. Επίσης, τα χαρακτηριστικά που προκύπτουν από συναισθηματική ανάλυση στα tweets εμφανίζουν παρόμοια συμπεριφορά με βασική εξαίρεση το λόγο των θετικών δημοσιεύσεων. Τέλος, το πλήθος των αναζητήσεων από την πλατφόρμα Google Trends εμφανίζει μικρότερη συσχέτιση με το AMR από ό,τι με το SHR%. Παρατηρούμε, λοιπόν, ότι παρόλο που τα δύο μεγέθη εμφανίζουν μεγάλη γραμμική συσχέτιση μεταξύ τους (0.79), όπως φαίνεται και στο δεύτερο διάγραμμα της πρώτης στήλης ή γραμμής του σχήματος 5.9 ή 5.10, διαφορετικά χαρακτηριστικά φαίνεται να περιέχουν περισσότερη χρήσιμη πληροφορία για το καθένα και συνεπώς διαφορετικά χαρακτηριστικά είναι πιο συνετό να χρησιμοποιηθούν για την εκτίμηση καθενός εκ των δύο. Άρα τελικά, κατά την εκπαίδευση θα αντιμετωπιστούν ως δύο τελείως διαφορετικά μεταξύ τους προβλήματα.

5.4.2 Οπτικοποίηση δεδομένων

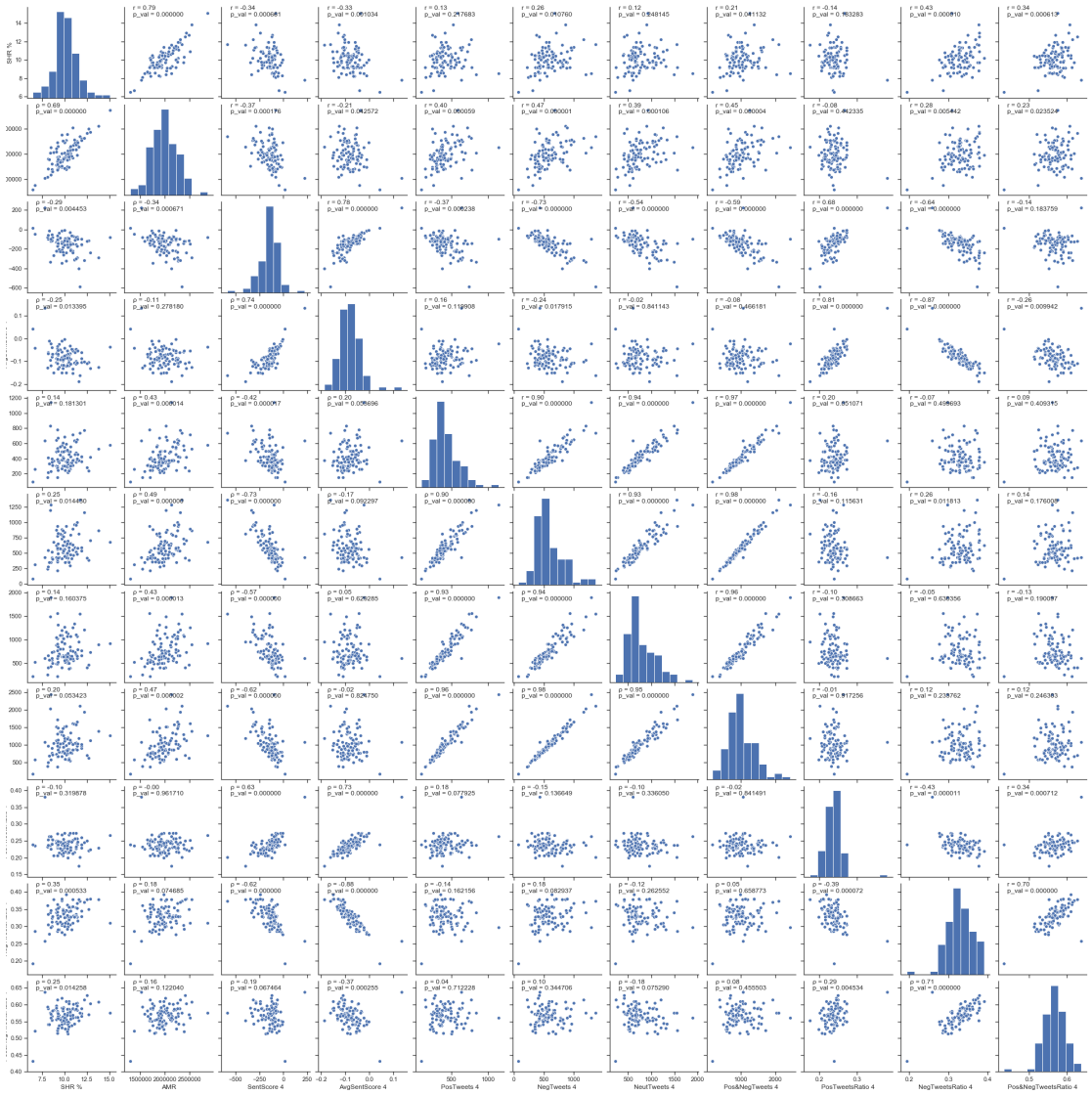
Τέλος, με τη βοήθεια της βιβλιοθήκης seaborn σχεδιάζουμε διάφορα scatter plots (από κοινού κατανομές) που απεικονίζουν τις σχέσεις ανεξάρτητη μεταβλητή i - ανεξάρτητη μεταβλητή j, ανεξάρτητη μεταβλητή - εξαρτημένη μεταβλητή και εξαρτημένη μεταβλητή i - εξαρτημένη μεταβλητή j και τα παρουσιάζουμε με τη μορφή πίνακα. Στα διαγράμματα άνω της διαγωνίου προστίθεται ο συντελεστής συσχέτισης του Pearson ενώ στα διαγράμματα κάτω της διαγωνίου προστίθεται ο συντελεστής συσχέτισης του Spearman. Επίσης, στη διαγώνιο παρουσιάζονται οι κατανομές της καθεμίας μεταβλητής, είτε ανεξάρτητης είτε εξαρτημένης. Για λόγους οικονομίας χώρου παρακάτω παρουσιάζονται τα διαγράμματα μόνο για ένα χρονικό παράθυρο του κάθε χαρακτηριστικού.



Σχήμα 5.9: Οπτικοποίηση όλων των χαρακτηριστικών από ποσοτικούς δείκτες

Πιο συγκεκριμένα, για τα δεδομένα από το Google Trends επιλέχθηκε το χρονικό παράθυρο 3, που αντιστοιχεί μόνο στην επόμενη μέρα από την προβολή του υπό εξέταση επεισοδίου και για τα δεδομένα από το Twitter επιλέχθηκε το χρονικό παράθυρο 4, που αντιστοιχεί στο τριήμερο από την προηγούμενη έως και την επόμενη μέρα από την προβολή του υπό εξέταση επεισοδίου. Οι πρώτες δύο γραμμές και οι πρώτες δύο στήλες του πίνακα διαγραμμάτων αντιστοιχούν στις εξαρτημένες μεταβλητές, SHR% και AMR ενώ οι υπόλοιπες γραμμές και στήλες αντιστοιχούν στις ανεξάρτητες μεταβλητές με την ίδια σειρά που παρουσιάστηκαν στους παραπάνω πίνακες.

Παρατηρούμε ότι το οπτικό αποτέλεσμα στις δύο πρώτες στήλες και γραμμές επιβεβαιώνει τις συσχετίσεις που παρατέθηκαν στους παραπάνω πίνακες μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών. Συμπληρωματικά, γίνεται φανερό ότι υπάρχει υψηλή γραμμική συσχέτιση μεταξύ των χαρακτηριστικών ποσοτικών δεικτών από το Twitter που δε συμπεριλαμβάνονται στους λόγους, δηλαδή μεταξύ των Tweets, UnqUsers, Retweets, Favorites και Tweets&Retweets. Αντίστοιχα, παρατηρείται υψηλή γραμμική συσχέτιση μεταξύ των χαρακτηριστικών που προέκυψαν από ανάλυση συναισθήματος στο κειμενικό περιεχόμενο των tweets και αφορούν απόλυτα μεγέθη, δηλαδή μεταξύ των PosTweets, NegTweets, NeutTweets και Pos&NegTweets. Αυτό και στις δύο περιπτώσεις ίσως να ήταν αναμενόμενο αν αναλογιστούμε τη φυσική υπόσταση των συγκεκριμένων χαρακτηριστικών.



Σχήμα 5.10: Οπτικοποίηση όλων των χαρακτηριστικών από ανάλυση συναισθήματος

5.5 Προεπεξεργασία χαρακτηριστικών και άλλες σχεδιαστικές επιλογές

Πριν προχωρήσουμε στην παραγωγή μοντέλων εκτίμησης στη φάση διεξαγωγής πειραμάτων, θα αναφερθούμε σε ορισμένες παραδοχές και επιλογές που κάναμε αλλά και σε όλες τις μορφές προεπεξεργασίας που απαιτήθηκαν για να μετατραπούν τα χαρακτηριστικά στην τελική μορφή τους που θα χρησιμοποιηθεί ως είσοδος στους αλγορίθμους.

5.5.1 Σχεδιαστικές επιλογές για τα χρονικά παράθυρα

Κατ΄ αρχάς, κάθε φορά θα χρησιμοποιείται μόνο ένα χρονικό παράθυρο για κάθε χαρακτηριστικό. Αυτό είναι απαραίτητο, καθώς για το ίδιο χαρακτηριστικό τα μεταξύ τους χρονικά παράθυρα περιέχουν κοινή πληροφορία σε μεγάλο βαθμό, δηλαδή είναι υψηλά συσχετισμένα μεταξύ τους, οπότε η συνύπαρξή τους μόνο αρνητικά θα μπορούσε να επιδράσει στην κατασκευή των μοντέλων, εισάγοντας καθυστερήσεις στο χρόνο εκπαίδευσης και οδηγώντας σε σφάλματα λόγω διακύμανσης. Συμπληρωματικά, αν τα κρατούσαμε όλα και έπειτα πραγματοποιούσαμε επιλογή χαρακτηριστικών με βάση κάποια μετρική συσχέτισης, κατά πάσα πιθανότητα θα επιλέγονταν όλα τα χρονικά παράθυρα των καλύτερων χαρακτηριστικών, καθώς όπως είδαμε και στους παραπάνω

πίνακες για το Twitter δε διαφέρουν ιδιαίτερα μεταξύ τους και τα υπόλοιπα θα απορρίπτονταν, χάνοντας με αυτόν τον τρόπο χρήσιμες πληροφορίες. Επίσης, για όλα τα χαρακτηριστικά που έχουν εξαχθεί με δεδομένα από το Twitter θα χρησιμοποιείται το ίδιο επιλεγμένο χρονικό παράθυρο κάθε φορά.

Τέλος, κάθε πείραμα που θα ακολουθήσει διεξάγεται σε δύο διαφορετικές εκτελέσεις. Η πρώτη εκτέλεση περιλαμβάνει τη δοκιμή όλων, και των 6, χρονικών παραθύρων που ορίστηκαν για τα χαρακτηριστικά. Σε κάθε επανάληψη ελέγχεται ένα παράθυρο για το πλήθος αναζητήσεων από το Google Trends και ένα παράθυρο για τα χαρακτηριστικά από το Twitter, δηλαδή συνολικά ελέγχονται 36 συνδυασμοί και κατασκευάζονται 36 μοντέλα. Τελικά, επιλέγεται να πραγματοποιήσει εκτιμήσεις στο σύνολο ελέγχου εκείνο το μοντέλο που εμφανίζει την καλύτερη απόδοση, και πιο συγκεκριμένα το ελάχιστο μέσο απόλυτο σφάλμα, στο σύνολο εκπαίδευσης έπειτα από διασταυρούμενη επικύρωση. Η δεύτερη εκτέλεση περιλαμβάνει την εξαγωγή μοντέλου εκτίμησης με βάση μόνο τα καλύτερα χρονικά παράθυρα του κάθε χαρακτηριστικού. Για το χαρακτηριστικό που προέκυψε από τα δεδομένα της πλατφόρμας Google Trends επιλέχθηκε το χρονικό παράθυρο 3, δηλαδή μόνο η επόμενη μέρα από την προβολή του εκάστοτε επεισοδίου, καθώς σε αυτό παρατηρήθηκαν οι υψηλότερες συσχετίσεις με τις εξαρτημένες μεταβλητές. Από την άλλη, για τα χαρακτηριστικά που προέκυψαν από δεδομένα της πλατφόρμας Twitter επιλέχθηκε το χρονικό παράθυρο 4, που αντιστοιχεί στο τριήμερο από την προηγούμενη μέρα έως και την επόμενη από την προβολή του κάθε επεισοδίου, καθώς ήταν ένα από τα δύο χρονικά παράθυρα που εμφάνισαν τις υψηλότερες συσχετίσεις. Και στις δύο περιπτώσεις το εναλλακτικό χρονικό παράθυρο ήταν το 6, το οποίο αποφεύγεται για λόγους που έχουν περιγραφεί εκτενώς παραπάνω.

5.5.2 Καθορισμός και επιλογή τελικών χαρακτηριστικών

Για κάθε αλγόριθμο παλινδρόμησης που θα εξεταστεί υπάρχουν 4 δυνατότητες για το ποιες και πόσες ανεξάρτητες μεταβλητές θα χρησιμοποιηθούν ως είσοδοι σε αυτόν. Και οι 4 δυνατότητες, που παραθέτουμε παρακάτω, εφαρμόζονται σε δύο διαφορετικές εκτελέσεις. Στην πρώτη εκτέλεση περιλαμβάνονται μόνο τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες, τα οποία καταγράφονται στους πίνακες 5.1 και 5.3. Στη δεύτερη εκτέλεση περιλαμβάνονται όλα τα χαρακτηριστικά που έχουν εξαχθεί, τόσο αυτά που σχετίζονται με ποσοτικούς δείκτες όσο και αυτά που προέκυψαν από συναισθηματική ανάλυση στα tweets. Οι 4 δυνατότητες εισαγωγής χαρακτηριστικών στους αλγορίθμους εκπαίδευσης είναι οι εξής:

- Καμία προεπεξεργασία

  Πρόκειται για την απλούστερη περίπτωση, καθώς ως είσοδοι στον εκάστοτε αλγόριθμο δίνονται όλα τα χαρακτηριστικά ανεξαιρέτως σε ένα επιλεγμένο χρονικό παράθυρο το καθένα κάθε φορά.

- Επιλογή χαρακτηριστικών

  Σε αυτήν την περίπτωση επιλέγονται να χρησιμοποιηθούν ως είσοδοι τα καλύτερα k χαρακτηριστικά σύμφωνα με το συντελεστή συσχέτισης του Pearson. Αυτό πραγματοποιείται με τη βοήθεια της κλάσης SelectKBest και τη συνάρτηση αξιολόγησης f_regression της βιβλιοθήκης scikit-learn. Στη συντριπτική πλειοψηφία των πειραμάτων, το εύρος των τιμών του k που δοκιμάστηκαν ήταν από 1 έως και το πλήθος όλων των διαθέσιμων χαρακτηριστικών. Η καλύτερη τιμή του k, δηλαδή αυτή που οδηγούσε σε μοντέλο με μικρότερο μέσο απόλυτο σφάλμα στο σύνολο δεδομένων εκπαίδευσης βρέθηκε με διασταυρούμενη επικύρωση.

- Ανάλυση σε κύριες συνιστώσες

  Σε αυτήν την περίπτωση εξάγονται νέα χαρακτηριστικά από τα ήδη υπάρχοντα σε έναν νέο χώρο με διανύσματα βάσης τις κύριες συνιστώσες. Η λογική είναι ότι μπορούμε να

κρατήσουμε την ίδια πληροφορία με λιγότερες μεταβλητές, διατηρώντας μόνο τις κύριες συνιστώσες που συνεισφέρουν περισσότερο στη διασπορά των δειγμάτων. Το εύρος τιμών που δοκιμάζεται για το πλήθος των κύριων συνιστωσών που θα διατηρηθούν (n_components) είναι συνήθως από 1 έως και το συνολικό πλήθος των αρχικών ανεξάρτητων μεταβλητών. Η όλη διαδικασία επιτυγχάνεται με την κλάση PCA της βιβλιοθήκης scikit-learn και τελικά επιλέγεται η καλύτερη τιμή για το πλήθος των κύριων συνιστωσών με διασταυρούμενη επικύρωση.

- Όλοι οι δυνατοί συνδυασμοί

  Τέλος, δοκιμάζονται να δοθούν όλοι οι δυνατοί συνδυασμοί χαρακτηριστικών ως είσοδοι στους αλγορίθμους. Η συγκεκριμένη διαδικασία ξεφεύγει από τα όρια της «έξυπνης» υλοποίησης, καθώς πρόκειται για εξαντλητική αναζήτηση και είναι εξαιρετικά χρονοβόρα. Για αυτόν το λόγο εφαρμόστηκε μόνο στους πιο απλούς αλγορίθμους που θα εξετάσουμε. Ωστόσο έστω και η εφαρμογή του σε πολύ λίγες περιπτώσεις ανέδειξε χρήσιμες πληροφορίες. Η εύρεση όλων των δυνατών συνδυασμών πραγματοποιήθηκε με τη βοήθεια της συνάρτησης combinations της βιβλιοθήκης itertools. Τελικά, επιλέγεται ως μοντέλο που θα πραγματοποιήσει τις προβλέψεις στο σύνολο ελέγχου εκείνο που έχει εκπαιδευτεί με είσοδο έναν από αυτούς τους συνδυασμούς και παρουσίασε το μικρότερο μέσο απόλυτο σφάλμα στο σύνολο εκπαίδευσης.

5.5.3 Τυποποίηση χαρακτηριστικών

Πολλοί αλγόριθμοι παλινδρόμησης από αυτούς που θα εξετάσουμε στη συνέχεια βασίζονται στη μέθοδο κατάβασης κλίσης, η οποία επιταχύνεται με την κανονικοποίηση των χαρακτηριστικών. Επίσης, σε περίπτωση που εφαρμοστεί ανάλυση σε κύριες συνιστώσες πρέπει να έχει προηγηθεί η κανονικοποίησή τους ώστε να μη συνεισφέρουν στη διασπορά περισσότερο ανεξάρτητες μεταβλητές που λαμβάνουν μεγαλύτερες τιμές. Αυτοί είναι οι κυριότεροι λόγοι που μας οδήγησαν να κανονικοποιήσουμε τα χαρακτηριστικά πριν αυτά εισαχθούν ως είσοδοι στους αλγορίθμους εκπαίδευσης. Ως τεχνική κανονικοποίησης επιλέχθηκε η τυποποίηση σε z-score, η οποία μετατρέπει τα χαρακτηριστικά με τέτοιο τρόπο ώστε να ακολουθούν κανονική κατανομή. Η τυποποίηση σε z-score προτιμήθηκε από την κλιμάκωση μεγίστου-ελαχίστου, καθώς όπως έχει αναφερθεί και προηγουμένως χειρίζεται καλύτερα αποκλίνουσες τιμές (outliers). Για την υλοποίησή της χρησιμοποιήθηκε η κλάση StandardScaler της βιβλιοθήκης scikit-learn. Αξίζει να σημειωθεί ότι στις τεχνικές παλινδρόμησης που βασίζονται σε δέντρα αποφάσεων, δηλαδή στην παλινδρόμηση με δέντρα αποφάσεων, με τυχαία δάση και με gradient boosting μηχανές, δεν πραγματοποιήθηκε κανονικοποίηση των χαρακτηριστικών, καθώς στους αλγορίθμους για την εκπαίδευση αυτών των μοντέλων είναι σημαντικές οι πραγματικές τιμές των χαρακτηριστικών και μπορούν να τις διαχειριστούν με επιτυχία.

5.5.4 Διασταυρούμενη επικύρωση

Οι αλγόριθμοι που θα επιστρατευτούμε στη συνέχεια για την κατασκευή μοντέλων παλινδρόμησης αλλά και διάφορες τεχνικές προεπεξεργασίας, όπως είδαμε π.χ. η επιλογή χαρακτηριστικών και η ανάλυση σε κύριες συνιστώσες δέχονται πλήθος υπερπαραμέτρων. Οι τιμές αυτών των υπερπαραμέτρων μπορεί να είναι καθοριστικές τόσο για το χρόνο εκπαίδευσης όσο και για την απόδοση του τελικού μοντέλου και την ικανότητά του να γενικεύει. Συνεπώς, η βελτιστοποίηση αυτών έχει βαρύνουσα σημασία και πρέπει να γίνει με τρόπο ώστε να μην υπάρχει διαρροή δεδομένων και να αποφευχθεί η υπερεκπαίδευση. Αυτό σε όλα τα πειράματα επιτεύχθηκε με τη διασταυρούμενη επικύρωση (cross-validation), η οποία υλοποιήθηκε με τη βοήθεια της κλάσης GridSearchCV της βιβλιοθήκης scikit-learn. Η GridSearchCV[3] μεταξύ άλλων δέχεται ως πα-

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

ραμέτρους έναν εκτιμητή, ένα εύρος τιμών για κάθε υπερπαράμετρο που χρησιμοποιεί αυτός ο εκτιμητής, τη συνάρτηση αξιολόγησης (scoring) σύμφωνα με την οποία κρίνεται καλύτερη ή όχι η απόδοση ενός μοντέλου και το πλήθος των τμημάτων-folds στα οποία θα κατατμηθεί το σύνολο δεδομένων εκπαίδευσης (cv). Στην περίπτωσή μας, σε όλα τα πειράματα χρησιμοποιήθηκαν 5 τμήματα και η συνάρτηση αξιολόγησης βασίζεται στο μέσο απόλυτο σφάλμα. Σε κάθε επανάληψη, δηλαδή, χρησιμοποιούνται τα $\frac{4}{5}$ του συνόλου εκπαίδευσης για να εκπαιδευτεί ένα μοντέλο και το $\frac{1}{5}$ για να ελεγχθεί η απόδοσή του. Πραγματοποιείται εξαντλητική αναζήτηση στο χώρο των υπερπαραμέτρων και κάθε φορά δοκιμάζεται ένας συνδυασμός τιμών υπερπαραμέτρων του εκτιμητή.

5.5.5 Διασωλήνωση

Όπως είδαμε προηγουμένως, η GridSearchCV δέχεται ως όρισμα έναν εκτιμητή για να ρυθμίσει βέλτιστα τις υπερπαραμέτρους του. Ωστόσο πέρα από τις υπερπαραμέτρους του τελικό εκτιμητή θέλουμε να συνυπολογιστούν και οι υπερπαράμετροι τυχόν μετασχηματιστών που εφαρμόζονται στα χαρακτηριστικά στο στάδιο της προεπεξεργασίας. Μία λύση είναι να οριστεί ένας συνολικός εκτιμητής, ο οποίος έχει ενσωματωμένους τους μετασχηματιστές με τη σειρά που εφαρμόζονται και στο τέλος τον εκτιμητή. Αυτό μπορεί να γίνει πράξη με τη διασωλήνωση (pipeline), η οποία υλοποιείται με την ομώνυμη κλάση Pipeline[4] που διαθέτει η βιβλιοθήκη scikit-learn. Στην περίπτωση που δεν εφαρμόζεται κάποια τεχνική μείωσης της διαστατικότητας και ο εκτιμητής είναι δέντρο απόφασης, τυχαίο δάσος ή gradient boosting μηχανή, τετριμμένα η διασωλήνωση περιλαμβάνει μόνο τον τελικό εκτιμητή. Για όλες τις υπόλοιπες τεχνικές παλινδρόμησης αν δεν εφαρμόζεται κάποια τεχνική μείωσης της διαστατικότητας η διασωλήνωση περιέχει τον κλιμακωτή (StandardScaler) και τον τελικό εκτιμητή. Αν εφαρμόζεται επιλογή χαρακτηριστικών, τότε αυτή προστίθεται στην αρχή της διασωλήνωσης και στις δύο παραπάνω περιπτώσεις. Ενώ αν εφαρμόζεται ανάλυση σε κύριες συνιστώσες, αυτή προστίθεται ακριβώς μετά τον κλιμακωτή, εάν φυσικά υπάρχει και ακριβώς πριν τον τελικό εκτιμητή. Τελικά, στην GridSearchCV δίνεται στο όρισμα του εκτιμητή ο συνολικός εκτιμητής που προκύπτει από τη διασωλήνωση και στο πλέγμα παραμέτρων δίνεται το εύρος των τιμών για κάθε υπερπαράμετρο που χρησιμοποιείται σε οποιοδήποτε στάδιο της διασωλήνωσης.

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html`

Κεφάλαιο 6

Διεξαγωγή πειραμάτων και αξιολόγηση αποτελεσμάτων

Σε αυτήν την ενότητα θα παρουσιάσουμε όλα τα μοντέλα παλινδρόμησης, τα οποία αναπτύχθηκαν στα πλαίσια της συγκεκριμένης διπλωματικής για την εκτίμηση τόσο του ποσοστού τηλεθέασης επί των ανοικτών δεκτών (SHR%) όσο και του απόλυτου πλήθους των τηλεθεατών που συγκέντρωσε (AMR) το ιταλικό τηλεοπτικό πρόγραμμα Le Iene. Όπως αναφέρθηκε και προηγουμένως, σε όλες τις εκτελέσεις χρησιμοποιήθηκε το ίδιο σύνολο 96 δειγμάτων για εκπαίδευση και το ίδιο σύνολο 11 δειγμάτων για αξιολόγηση. Αυτό διασφαλίστηκε δίνοντας στη γεννήτρια τυχαίων αριθμών ως όρισμα ένα σταθερό αριθμό στη συνάρτηση train_test_split, η οποία υλοποιεί τη διάσπαση του συνόλου δεδομένων σε σύνολο εκπαίδευσης και ελέγχου. Κατά την εκπαίδευση του κάθε μοντέλου, ανεξαρτήτως της μεθόδου που ακολουθήθηκε, χρησιμοποιήθηκε διασταυρούμενη επικύρωση (cross-validation) για τη βελτιστοποίηση των υπερπαραμέτρων του.

Επίσης, πραγματοποιήθηκαν διάφορες εκτελέσεις με βάση τα χρονικά παράθυρα, τα χαρακτηριστικά που θα αξιοποιηθούν και την προεπεξεργασία των τελευταίων. Παρακάτω, παραθέτουμε την αρίθμηση, που αντιστοιχεί σε αυτές, η οποία θα χρησιμοποιηθεί για την παράθεση των αποτελεσμάτων σε πίνακες στη συνέχεια. Πρώτα, όσον αφορά την επιλογή και την προεπεξεργασία των χαρακτηριστικών που θα χρησιμοποιηθούν ως είσοδοι στους αλγορίθμους:

1. όλα τα χαρακτηριστικά χωρίς προεπεξεργασία
2. τα επιλεγμένα k καλύτερα χαρακτηριστικά με βάση τη μετρική συσχέτισης του Pearson
3. οι n κύριες συνιστώσες στις οποίες οφείλεται το μεγαλύτερο μέρος της διασποράς
4. ο καλύτερος συνδυασμός χαρακτηριστικών από όλους τους δυνατούς συνδυασμούς

Έπειτα, όσον αφορά το είδος των χαρακτηριστικών και τα χρονικά παράθυρα στα οποία αυτά ορίζονται:

α′) μόνο τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες, ορισμένα σε όλα, και τα 6, διαφορετικά χρονικά παράθυρα

β′) μόνο τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων που προηγήθηκε, δηλαδή το 3 για το GTrends και το 4 για τα χαρακτηριστικά από το Twitter

γ′) όλα τα χαρακτηριστικά που προκύπτουν είτε από ποσοτικούς δείκτες είτε από ανάλυση συναισθήματος στα tweets, ορισμένα σε όλα, και τα 6, διαφορετικά χρονικά παράθυρα

δ′) όλα τα χαρακτηριστικά που προκύπτουν είτε από ποσοτικούς δείκτες είτε από ανάλυση συναισθήματος στα tweets, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων που προηγήθηκε, δηλαδή το 3 για το GTrends και το 4 για τα χαρακτηριστικά από το Twitter

Οφείλουμε να υπενθυμίσουμε ότι το χρονικό παράθυρο 3 για τις αναζητήσεις από την πλατφόρμα Google Trends αναφέρεται μόνο στην επόμενη μέρα από την προβολή του κάθε επεισοδίου ενώ το χρονικό παράθυρο 4 για τα δεδομένα από το Twitter αναφέρεται στο τριήμερο από την προηγούμενη έως και την επόμενη μέρα από την προβολή του κάθε επεισοδίου.

Αφού ολοκληρωθεί η διαδικασία της εκπαίδευσης, επιλέγεται το μοντέλο που είχε την καλύτερη απόδοση στο σύνολο των δεδομένων εκπαίδευσης, σύμφωνα με τη μετρική του μέσου απολύτου σφάλματος (MAE), για να πραγματοποιήσει τις τελικές εκτιμήσεις. Γίνεται γνωστό ποια χαρακτηριστικά και σε ποια χρονικά παράθυρα χρησιμοποιήθηκαν ως είσοδοι για την εκπαίδευσή του και ποιες είναι οι βέλτιστες τιμές υπερπαραμέτρων που προέκυψαν για αυτό από τη διαδικασία της διασταυρούμενης επικύρωσης. Επίσης, επιστρέφεται το σκορ σύμφωνα με το οποίο αποφασίστηκε ότι αυτό είναι το καλύτερο μοντέλο και στην περίπτωση της ανάλυσης με κύριες συνιστώσες το ποσοστό της διακύμανσης που δικαιολογείται από τις κύριες συνιστώσες που κρατήθηκαν, τόσο για τη κάθε μία ξεχωριστά όσο και αθροιστικά.

Τέλος, η αξιολόγηση των αποτελεσμάτων πραγματοποιήθηκε πάνω στα 11 δείγματα που περιλαμβάνει το σύνολο δεδομένων ελέγχου για τα μεγέθη SHR% και AMR. Μέσω αυτής δίνεται μια εικόνα για την ποιότητα των προβλέψεων, οι οποίες παρήχθησαν από το καλύτερο μοντέλο, που προέκυψε κατά τη διαδικασία της εκπαίδευσης. Η αξιολόγηση βασίστηκε σε όλες τις μετρικές που παρουσιάστηκαν στην ενότητα 3.3.2, οι οποίες είναι το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE), το μέσο απόλυτο σφάλμα (MAE), το μέσο τετραγωνικό σφάλμα (MSE), ο συντελεστής προσδιορισμού (R^2), το ποσοστό της δικαιολογημένης διακύμανσης (explained variance score) και η ακρίβεια (accuracy).

Στη συνέχεια, για κάθε τεχνική παλινδρόμησης, παρατίθενται όλα τα στοιχεία των μοντέλων που κατασκευάστηκαν στη φάση της εκπαίδευσης, δίνονται σε πίνακες οι τιμές των μετρικών αξιολόγησης που προέκυψαν από τη φάση του ελέγχου και παρουσιάζονται σε γραφικές απεικονίσεις οι προβλεπόμενες τιμές των εξαρτημένων μεταβλητών σε σχέση με τις πραγματικές για τα 11 δείγματα που ανήκουν στο σύνολο ελέγχου.

6.1 Μοντέλα αναφοράς

Κατασκευάστηκαν δύο μοντέλα αναφοράς (baseline models), ένα για κάθε μεταβλητή προς εκτίμηση. Και τα δύο ανεξαρτήτως των χαρακτηριστικών, προβλέπουν πάντα την ίδια σταθερή τιμή, που δεν είναι άλλη από το μέσο όρο των τιμών της αντίστοιχης εξαρτημένης μεταβλητής στην κάθε περίπτωση στο σύνολο εκπαίδευσης. Αυτός βρίσκεται με τη βοήθεια της συνάρτησης mean της βιβλιοθήκης numpy και στην περίπτωση του SHR% είναι περίπου 10.11% ενώ στην περίπτωση του AMR είναι λίγο λιγότερο από 2 εκατομμύρια (1994179). Στους πίνακες 6.1 και 6.2 φαίνονται ποσοτικοποιημένα τα αποτελέσματα της αξιολόγησης των εκτιμήσεων ενώ στο σχήμα 6.1 φαίνονται αντίστοιχα γραφικά οι διαφορές μεταξύ πραγματικών και προβλεπόμενων τιμών.

MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
11.15%	1.03	1.68	-0.27	0.00	88.85%

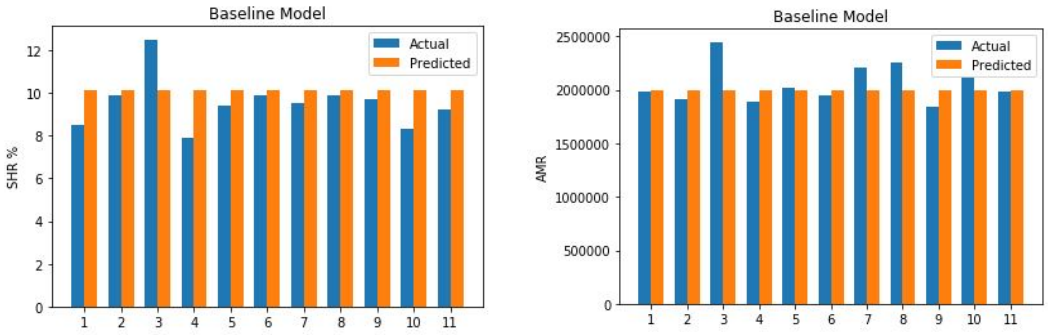
Πίνακας 6.1: Αξιολόγηση του μοντέλου αναφοράς για την εκτίμηση του SHR%

Προφανώς, και στις δύο περιπτώσεις πρόκειται για υπεραπλουστευτικά μοντέλα που δεν αξιοποιούν καμία πληροφορία για να εξάγουν προβλέψεις. Ο μόνος λόγος για τον οποίο κατασκευάστηκαν είναι για να αποτελέσουν σημείο αναφοράς και να διευκολύνουν την αξιολόγηση των υπόλοιπων μοντέλων που θα εξαχθούν στη συνέχεια. Με τα αποτελέσματα που λαμβάνονται από

MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
6.18%	133940	34199143098	-0.12	-0.00	93.82%

Πίνακας 6.2: Αξιολόγηση του μοντέλου αναφοράς για την εκτίμηση του AMR

αυτά, μπορούμε να αποκτήσουμε μια πιο καθαρή εικόνα για το τι να περιμένουμε από τα μοντέλα που θα εκπαιδεύσουμε και να θέσουμε πιο σαφείς στόχους για την απόδοσή τους. Σίγουρα το επιθυμητό είναι να πέσουν οι τιμές των MAPE, MAE και MSE όσο το δυνατόν περισσότερο, να γίνουν θετικές οι τιμές των R^2 και explained variane και όσο μεγαλύτερες γίνεται κατ΄ απόλυτη τιμή και η τιμή της accuracy να πλησιάσει πιο κοντά στο 100%.



(α΄) για το SHR%

(β΄) για το AMR

Σχήμα 6.1: Πραγματικές και προβλεπόμενες τιμές από τα μοντέλα αναφοράς

Παρατηρούμε ότι η σταθερή πρόβλεψη του μέσου όρου από τις παρατηρήσεις εκπαίδευσης φαίνεται να ταιριάζει καλύτερα στις πραγματικές τιμές του AMR από ό,τι του SHR%. Αυτό γίνεται φανερό από το χαμηλότερο MAPE και την υψηλότερη accuracy στην περίπτωση του AMR και σημαίνει ότι δεν υπάρχει μεγάλη διακύμανση στις τιμές των παρατηρήσεων του συνόλου ελέγχου για αυτήν την εξαρτημένη μεταβλητή. Αν και, λοιπόν, αυτό το πρόβλημα εκτίμησης εμφανίζει πολύ μικρό σφάλμα ήδη από το πιο απλό μοντέλο, θα αποδειχθεί πολύ πιο δύσκολο πρόβλημα παλινδρόμησης στη συνέχεια, ακριβώς λόγω των μικρών διαφορών στις προς εκτίμηση τιμές. Επιπλέον, όπως είναι φυσικό τα MAE και MSE είναι ασύγκριτα μεγαλύτερα στην περίπτωση του AMR, καθώς πρόκειται για την εκτίμηση ενός απόλυτου μεγέθους που λαμβάνει τιμές της τάξης των εκατομμυρίων σε αντίθεση με την εκτίμηση του SHR%, το οποίο είναι ένα ποσοστό και οι τιμές του περιορίζονται σε ένα σαφώς μικρότερο εύρος. Πιο συγκεκριμένα, σε αντιστοιχία με τα ληφθέντα αποτελέσματα, θέτουμε τους εξής στόχους για το καθένα πρόβλημα:

- Το MAPE να πέσει κάτω από 8% στην περίπτωση του SHR% και να πλησιάσει το 5% στην περίπτωση του AMR.
- Το MAE να πέσει κάτω από το 0.8 για το SHR% και να πλησιάσει τις 100000 για το AMR.
- Αντίστοιχα, το MSE να πέσει κάτω από το 1 για το SHR% και να πλησιάσει τα 20 εκατομμύρια για το AMR.
- Ο συντελεστής προσδιορισμού R^2 να υπερβεί το 0.50 στην περίπτωση του SHR% και το 0.30 στην περίπτωση του AMR.
- Το ποσοστό δικαιολογημένης διακύμανσης να φτάσει το 0.60 στην περίπτωση του SHR% και το 0.40 στην περίπτωση του AMR.
- Τέλος, η ακρίβεια να υπερβεί το 92% για το SHR% και να φτάσει το 95% για το AMR.

6.2 Μοντέλα πολλαπλής γραμμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων

Η πολλαπλή γραμμική παλινδρόμηση σε αυτήν την ενότητα στηρίζεται στη μέθοδο των ελάχιστων τετραγώνων και υλοποιήθηκε με την κλάση LinearRegression[1] της βιβλιοθήκης scikit-learn. Συνολικά, παρήχθησαν 14 μοντέλα για την εκτίμηση καθενός εκ των δύο εξαρτημένων μεταβλητών. Καθένα από αυτά βασίστηκε στο συνδυασμό μιας επιλογής 1-4 για την προεπεξεργασία των χαρακτηριστικών και μιας επιλογής α-δ για το είδος τους και το χρονικό παράθυρο στο οποίο αυτά ορίστηκαν. Τα αποτελέσματα που προέκυψαν και η αξιολόγησή τους φαίνονται στους πίνακες 6.3, 6.4 και στο σχήμα 6.2.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**5.45%**	**0.48**	**0.53**	**0.60**	**0.62**	**94.55%**
	β	6.76%	0.62	0.86	0.35	0.53	93.24%
	γ	10.96%	1.02	1.43	-0.08	-0.05	89.04%
	δ	9.31%	0.86	1.34	-0.01	0.08	90.69%
2	α	5.45%	0.48	0.53	0.60	0.62	94.55%
	β	6.76%	0.62	0.86	0.35	0.53	93.24%
	γ	9.93%	0.95	1.26	0.05	0.07	90.07%
	δ	9.31%	0.86	1.34	-0.01	0.08	90.69%
3	α	5.45%	0.48	0.53	0.60	0.62	94.55%
	β	6.82%	0.62	0.83	0.38	0.56	93.18%
	γ	11.02%	1.08	1.69	-0.27	-0.25	88.98%
	δ	8.67%	0.80	1.06	0.20	0.32	91.33%
4	α	6.61%	0.59	0.77	0.42	0.52	93.39%
	β	7.19%	0.65	0.90	0.32	0.56	92.81%

Πίνακας 6.3: Αξιολόγηση των μοντέλων πολλαπλής γραμμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του SHR%

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	6.81%	140289	41204055995	-0.35	-0.01	93.19%
	β	6.81%	140783	34722430557	-0.14	-0.05	93.19%
	γ	9.84%	204235	64258843947	-1.11	-0.76	90.16%
	δ	9.21%	189509	55684762746	-0.82	-0.65	90.79%
2	α	6.23%	127992	38200027372	-0.25	0.05	93.76%
	β	**4.96%**	**102882**	**23232737622**	**0.24**	**0.36**	**95.04%**
	γ	9.84%	204235	64258843947	-1.11	-0.76	90.16%
	δ	8.50%	174806	49580341370	-0.62	-0.40	91.50%
3	α	6.81%	140289	41204055995	-0.35	-0.01	93.19%
	β	6.80%	140783	34722430557	-0.14	-0.05	93.19%
	γ	9.66%	200020	64424821290	-1.11	-0.85	90.33%
	δ	8.37%	171062	52152255206	-0.71	-0.61	91.63%
4	α	6.34%	133417	30485587593	0.00	0.26	93.66%
	β	5.44%	112531	25921141825	0.15	0.29	94.56%

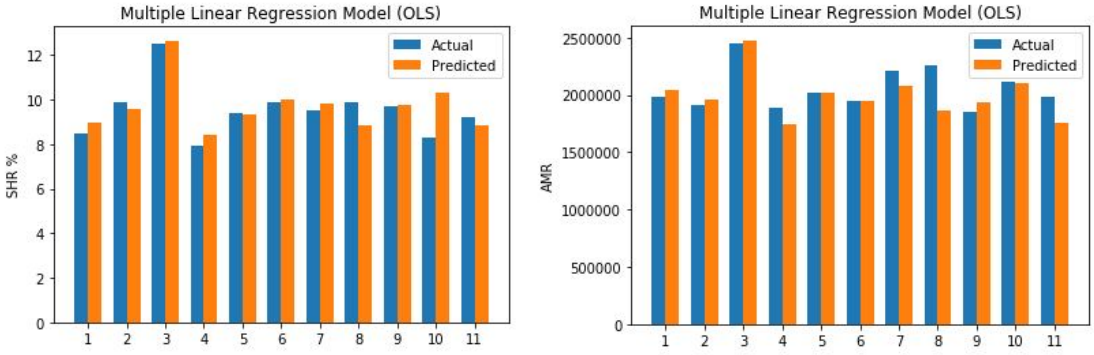
Πίνακας 6.4: Αξιολόγηση των μοντέλων πολλαπλής γραμμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του AMR

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Παρατηρούμε ότι απλώς και μόνο με ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης βασιζόμενης στη μέθοδο των ελάχιστων τετραγώνων, το οποίο είναι το πιο απλό μοντέλο που μπορούμε να κατασκευάσουμε, επιτυγχάνονται ήδη πολύ ικανοποιητικά αποτελέσματα. Μάλιστα, στην περίπτωση της εκτίμησης του ποσοστού τηλεθέασης επί των ανοικτών δεκτών (SHR%), όλα τα μοντέλα που παρήχθησαν εμφάνισαν καλύτερες τιμές όσον αφορά τις μετρικές MAPE και Accuracy από ό,τι το μοντέλο αναφοράς ενώ μόλις ένα μοντέλο παρουσίασε ελαφρώς χειρότερη απόδοση από αυτό όσον αφορά τις μετρικές MAE, MSE και ποσοστό δικαιολογημένης διακύμανσης. Από την άλλη, πολλά μοντέλα από αυτά που κατασκευάστηκαν για την εκτίμηση του AMR παρουσίασαν χειρότερη απόδοση από αυτήν του αντίστοιχου μοντέλου αναφοράς. Όμως υπήρχαν και δύο μοντέλα που ήταν εμφανώς καλύτερα από το τελευταίο.

Το μοντέλο που παρήγαγε τις καλύτερες εκτιμήσεις για το SHR% εκπαιδεύτηκε με όλα τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες, ορισμένα στο χρονικό παράθυρο 3 και για τις δύο πηγές δεδομένων, το οποίο αντιστοιχεί μόνο στη μέρα προβολής του εκάστοτε επεισοδίου για το GTrends και στο χρονικό διάστημα από μισή ώρα πριν έως και μισή ώρα μετά την προβολή του εκάστοτε επεισοδίου για τα χαρακτηριστικά από το Twitter. Χρησιμοποιώντας αυτό στο σύνολο ελέγχου, όπως μπορούμε να δούμε και στον πίνακα 6.3, επιτεύχθηκε MAPE = 5.45%, MAE = 0.48, MSE = 0.53, $R^2 = 0.60$, explained_variance = 0.62 και accuracy = 94.55%.

Το μοντέλο που παρήγαγε τις καλύτερες εκτιμήσεις για το AMR εκπαιδεύτηκε με τα 7 καλύτερα σύμφωνα με το συντελεστή συσχέτισης του Pearson χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για GTrends και 4 για τα χαρακτηριστικά από το Twitter). Πιο συγκεκριμένα, τα χαρακτηριστικά που επικράτησαν ήταν τα εξής: Tweets, UnqUsers, Retweets, Favorites, Tweets&Retweets, FavoritesRatio και UnqUsersRatio. Οι εκτιμήσεις που πραγματοποίησε για τις παρατηρήσεις στο σύνολο ελέγχου οδήγησαν στα εξής αποτελέσματα: MAPE = 4.96%, MAE = 102882, MSE = 23232737622, $R^2 = 0.24$, explained_variance = 0.36 και accuracy = 95.04%.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.2: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα πολλαπλής γραμμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων

6.3 Μοντέλα πολυωνυμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων

Είναι πιθανό τα μοντέλα γραμμικής παλινδρόμησης που κατασκευάσαμε προηγουμένως να ήταν υπερβολικά απλουστευτικά για τα δεδομένα μας και κατ΄ επέκταση, να παρουσίαζαν υψηλή απόκλιση. Για να αποφευχθεί αυτό το φαινόμενο της υποεκπαίδευσης, θα αυξήσουμε την πολυπλοκότητα των μοντέλων μας για να δούμε αν θα λάβουμε καλύτερα αποτελέσματα. Αυτό μπορεί να επιτευχθεί με μοντέλα πολυωνυμικής παλινδρόμησης. Ουσιαστικά, πρόκειται πάλι

για γραμμικά μοντέλα ως προς τα βάρη, θεωρώντας περισσότερα χαρακτηριστικά αυτή τη φορά. Πιο συγκεκριμένα, τα νέα χαρακτηριστικά προκύπτουν από όλα τα πεπλεγμένα πολυώνυμα των αρχικών χαρακτηριστικών βαθμού μικρότερου ή ίσου με αυτόν που ορίζουμε. Εδώ, δοκιμάσαμε βαθμούς 2 και 3 και τα νέα χαρακτηριστικά που δημιουργήθηκαν από τα πεπλεγμένα πολυώνυμα των αρχικών παρήχθησαν με τη βοήθεια της κλάσης PolynomialFeatures[2] της βιβλιοθήκης scikit-learn. Το μοντέλο κατασκευάστηκε πάλι με τη βοήθεια της κλάσης LinearRegression της βιβλιοθήκης scikit-learn και στηρίχθηκε στη μέθοδο των ελάχιστων τετραγώνων. Τέλος, δε δοκιμάζονται πολυώνυμα μεγαλύτερου βαθμού ώστε να μην κατασκευαστεί ένα υπερβολικά περίπλοκο μοντέλο, που θα μαθαίνει από θόρυβο και θα έχει υψηλή διακύμανση και συνεπώς μικρή ικανότητα γενίκευσης, έχοντας πέσει στην παγίδα της υπερεκπαίδευσης.

Συνολικά, παρήχθησαν 9 μοντέλα για την εκτίμηση καθενός εκ των δύο εξαρτημένων μεταβλητών. Καθένα από αυτά βασίστηκε στο συνδυασμό μιας επιλογής 1-3 για την προεπεξεργασία των χαρακτηριστικών και μιας επιλογής α-δ για το είδος τους και το χρονικό παράθυρο στο οποίο αυτά ορίστηκαν. Από εδώ και πέρα, δε θα ελέγχουμε όλους τους δυνατούς συνδυασμούς χαρακτηριστικών ως είσοδο (επιλογή 4), καθώς τα μοντέλα που θα αναπτύξουμε είναι σημαντικά πιο σύνθετα και αυτή η διαδικασία γίνεται υπερβολικά χρονοβόρα. Τα αποτελέσματα που προέκυψαν και η αξιολόγησή τους φαίνονται στους πίνακες 6.5, 6.6 και στο σχήμα 6.3.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	21.43%	2.30	22.80	-16.2	-15.65	78.57%
2	**α**	7.83%	0.73	1.07	0.20	0.51	92.17%
	β	8.92%	0.79	1.13	0.15	0.38	91.08%
	γ	12.57%	1.15	2.23	-0.67	-0.33	87.42%
	δ	12.77%	1.19	2.27	-0.71	-0.55	87.23%
3	**α**	7.14%	0.64	0.76	0.43	0.59	92.86%
	β	**6.70%**	**0.61**	**0.63**	**0.52**	**0.63**	**93.30%**
	γ	9.21%	0.87	0.98	0.26	0.34	90.79%
	δ	11.05%	1.04	1.48	-0.11	0.04	88.95%

Πίνακας 6.5: Αξιολόγηση των μοντέλων πολυωνυμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του SHR%

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	26.46%	618211	2487147935230	-80	-70	73.54%
2	**α**	9.51%	209786	115020037741	-2.77	-1.80	90.49%
	β	5.81%	126780	30723407362	-0.01	0.23	94.19%
	γ	7.86%	166224	43796815575	-0.44	-0.29	92.14%
	δ	6.89%	148947	38609760484	-0.27	-0.10	93.11%
3	**α**	7.24%	150190	36653048165	-0.20	-0.02	92.76%
	β	**5.08%**	**106343**	**21028791469**	**0.31**	**0.42**	**94.92%**
	γ	7.94%	168848	48727414866	-0.60	-0.50	92.06%
	δ	7.12%	151329	40096476809	-0.31	-0.26	92.88%

Πίνακας 6.6: Αξιολόγηση των μοντέλων πολυωνυμικής παλινδρόμησης που υλοποιήθηκε με τη μέθοδο των ελάχιστων τετραγώνων για την εκτίμηση του AMR

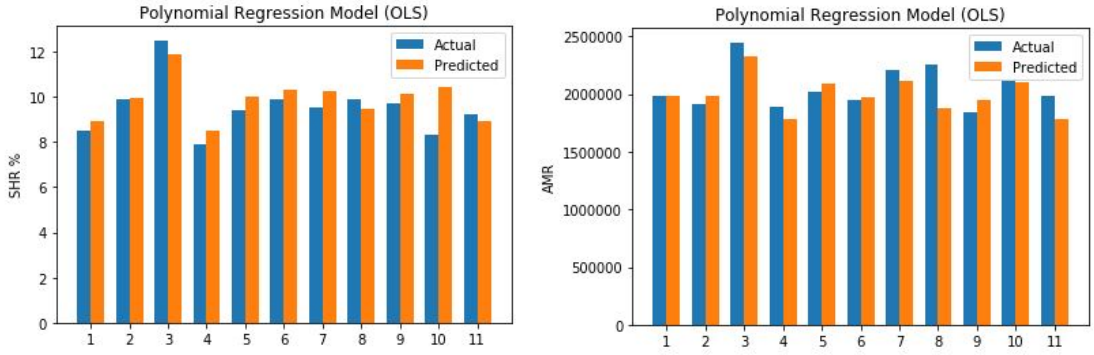
Όπως μπορούμε να δούμε από τους πίνακες, όταν χρησιμοποιούνται όλα τα χαρακτηριστικά ως είσοδος στους αλγορίθμους εκπαίδευσης, χωρίς να έχει προηγηθεί κάποια τεχνική μείωσης της διαστατικότητας, η απόδοση του μοντέλου πέφτει κατακόρυφα. Αυτό ήταν αναμενόμενο να

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

συμβεί, καθώς το πλήθος των νέων εξαγόμενων χαρακτηριστικών γίνεται συγκρίσιμο ή ακόμα και υπερβαίνει κάποιες φορές το πλήθος των παρατηρήσεων, οπότε και οδηγούμαστε σε υπερεκπαίδευση. Ακόμα και στην πιο απλή περίπτωση, που χρησιμοποιούνται μόνο χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και πολυώνυμο δευτέρου βαθμού, προκύπτουν 55 νέα από τα 9 αρχικά χαρακτηριστικά.

Όσον αφορά την εκτίμηση του SHR%, τα μοντέλα που εκπαιδεύτηκαν με χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες, με εξαίρεση φυσικά αυτό που δεν προέκυψε έπειτα από κάποια τεχνική μείωσης της διαστατικότητας, εμφάνισαν πολύ ικανοποιητικά αποτελέσματα και σημαντικά καλύτερα από αυτά του μοντέλου αναφοράς. Από την άλλη, τα μοντέλα που εκπαιδεύτηκαν με χαρακτηριστικά που προέκυψαν τόσο από ποσοτικούς δείκτες όσο και από ανάλυση συναισθήματος στα tweets εμφάνισαν χαμηλότερα απόδοση, με μόνο το ένα από αυτά να παρουσιάζει αποτελέσματα πολύ καλύτερα από το μοντέλο αναφοράς, ένα ελαφρώς βελτιωμένα και τα άλλα δύο ελαφρώς χειρότερα. Όσον αφορά την εκτίμηση του AMR, μόλις δύο μοντέλα από αυτά που κατασκευάστηκαν ξεπέρασαν το μοντέλο αναφοράς σε απόδοση. Επίσης, σε αυτήν την περίπτωση δεν παρατηρήθηκε κάποια μεγάλη διαφορά μεταξύ των μοντέλων που χρησιμοποιούσαν μόνο χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και των μοντέλων που χρησιμοποιούσαν πέρα από αυτά και χαρακτηριστικά από ανάλυση συναισθήματος.

Και στις δύο περιπτώσεις, τα μοντέλα που πραγματοποίησαν τις καλύτερες προβλέψεις ήταν εκείνα στα οποία χρησιμοποιήθηκε η ανάλυση σε κύριες συνιστώσες ως τεχνική μείωσης της διαστατικότητας πάνω σε όλα τα χαρακτηριστικά που προέκυψαν από ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα, σύμφωνα με τα συμπεράσματα από τη διερευνητική ανάλυση δεδομένων (3 για το πλήθος των αναζητήσεων από την πλατφόρμα Google Trends και 4 για τα δεδομένα από το Twitter). Επίσης, και στις δύο περιπτώσεις επιλέχθηκε έπειτα από τα αποτελέσματα της διασταυρούμενης επικύρωσης να χρησιμοποιηθεί πολυώνυμο δευτέρου βαθμού και να κρατηθούν 10 κύριες συνιστώσες, στις οποίες αποδίδεται περίπου το 99.8% της διακύμανσης των παρατηρήσεων. Τελικά, για την εκτίμηση του SHR% στο σύνολο ελέγχου επιτεύχθηκε: MAPE $= 6.70\%$, MAE $= 0.61$, MSE $= 0.63$, $R^2 = 0.52$, explained_variance $= 0.63$ και accuracy $= 93.30\%$ και για την εκτίμηση του AMR στο σύνολο ελέγχου επιτεύχθηκε: MAPE $= 5.08\%$, MAE $= 106343$, MSE $= 21028791469$, $R^2 = 0.31$, explained_variance $= 0.42$ και accuracy $= 94.92\%$.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.3: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα πολυωνυμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων

6.4 Μοντέλα παλινδρόμησης κορυφογραμμής

Η παλινδρόμηση κορυφογραμμής ανήκει στις τεχνικές παλινδρόμησης κανονικοποίησης και μπορεί να αντιμετωπίσει την ύπαρξη συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Αυτό το

πετυχαίνει ομαλοποιώντας την καμπύλη της παλινδρόμησης και συνεπώς μικραίνοντας τους συντελεστές στη συνάρτηση αυτής, με αποτέλεσμα να μειώνεται η διακύμανση και να αυξάνεται η απόκλιση. Για αυτούς τους λόγους και επειδή ακριβώς είναι σε θέση να δίνει λιγότερη ή περισσότερη βαρύτητα σε όποια χαρακτηριστικά χρειάζεται, μικραίνοντας ή μεγαλώνοντας αντίστοιχα τις τιμές των συντελεστών τους, κρίθηκε σκόπιμο να μην εφαρμοστούν τεχνικές μείωσης της διαστατικότητας. Κατασκευάστηκαν, λοιπόν, 4 μοντέλα, τα οποία όλα αντιστοιχούν στην επιλογή 1 για τα διάφορα είδη και χρονικά παράθυρα χαρακτηριστικών που μπορούν να χρησιμοποιηθούν (α-δ). Η υλοποίηση έλαβε χώρα με την κλάση Ridge[3] της βιβλιοθήκης scikit-learn. Δοκιμάστηκαν πολλές τιμές για την υπερπαράμετρο alpha που αυτή φέρει, η οποία εκφράζει την ισχύ της κανονικοποίησης. Όσο πιο κοντά στο 0 είναι τόσο μικρότερη είναι η κανονικοποίηση που εφαρμόζεται και τόσο πιο πολύ πλησιάζει η καμπύλη παλινδρόμησης εκείνη που προέκυψε από τη λύση των ελάχιστων τετραγώνων. Δοκιμάσαμε από πολύ μικρές τιμές στο εύρος $[10^{-10}, 10^{-2}]$, με πολλαπλασιαστικό βήμα 100, μετά μεσαίες στο εύρος $[0.1, 0.4]$ με αθροιστικό βήμα 0.1 έως και πιο μεγάλες στο εύρος $[0.5, 10.0]$ με αθροιστικό βήμα 0.5. Τα αποτελέσματα που λάβαμε φαίνονται στους πίνακες 6.7 και 6.8 και στο σχήμα 6.4.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**5.45%**	**0.48**	**0.53**	**0.60**	0.62	**94.55%**
	β	6.54%	0.59	0.57	0.57	**0.64**	93.46%
	γ	9.76%	0.95	1.41	-0.06	-0.04	90.24%
	δ	9.79%	0.91	1.21	0.09	0.23	90.21%

Πίνακας 6.7: Αξιολόγηση των μοντέλων παλινδρόμησης κορυφογραμμής για την εκτίμηση του SHR%

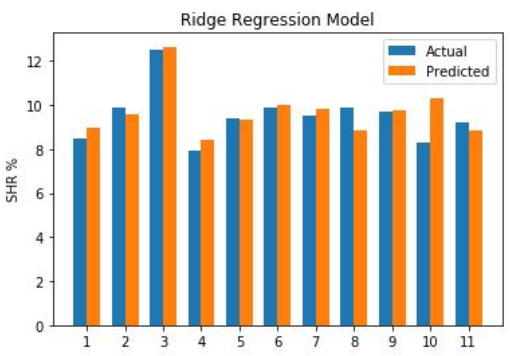
		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**6.81%**	**140298**	41204055988	-0.35	**-0.01**	**93.19%**
	β	6.81%	140783	**34722430550**	**-0.14**	-0.05	93.19%
	γ	9.59%	198648	62015013702	-1.03	-0.76	90.41%
	δ	8.57%	176273	51166837030	-0.68	-0.61	91.43%

Πίνακας 6.8: Αξιολόγηση των μοντέλων παλινδρόμησης κορυφογραμμής για την εκτίμηση του AMR

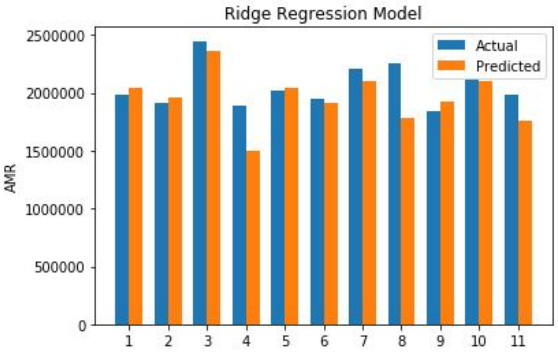
Παρατηρούμε ότι στην περίπτωση της εκτίμησης του SHR% όλα τα μοντέλα παλινδρόμησης κορυφογραμμής απέδωσαν καλύτερα από το μοντέλο αναφοράς. Μάλιστα, εκείνα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες εμφάνισαν πολύ πιο ικανοποιητικές τιμές στις μετρικές αξιολόγησης από εκείνα που εκπαιδεύτηκαν με βάση όλα τα χαρακτηριστικά που προέκυψαν είτε από ποσοτικούς δείκτες είτε από ανάλυση συναισθήματος στα tweets. Αντιθέτως, στην περίπτωση της εκτίμησης του AMR, κανένα μοντέλο παλινδρόμησης κορυφογραμμής δεν απέδωσε καλύτερα από το αντίστοιχο μοντέλο αναφοράς. Αλλά πάλι τα μοντέλα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που σχετίζονται μόνο με ποσοτικούς δείκτες παρουσίασαν καλύτερη συμπεριφορά από τα υπόλοιπα, με αποτελέσματα συγκρίσιμα με εκείνα του μοντέλου αναφοράς.

Τελικά, το καλύτερο μοντέλο κορυφογραμμής για την εκτίμηση του SHR% είναι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο, ορισμένα στο χρονικό παράθυρο 3 (που αντιστοιχεί μόνο στην επόμενη μέρα από την προβολή για το GTrends και στο διάστημα από μισή ώρα πριν έως και μισή ώρα μετά την προβολή του κάθε επεισοδίου για τα χαρακτηριστικά από το Twitter). Για την υπερπαράμετρο alpha χρησιμοποιήθηκε η τιμή 10^{-10}, καθώς αναδείχθηκε ως η καταλληλότερη από τη διαδικασία της διασταυρούμενης επικύρωσης.

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`



(α′) για το SHR%



(β′) για το AMR

Σχήμα 6.4: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης κορυφογραμμής

Τα αποτελέσματα που πέτυχε στο σύνολο ελέγχου είναι πανομοιότυπα με εκείνα της πολλαπλής γραμμικής παλινδρόμησης που στηρίχθηκε στη μέθοδο των ελάχιστων τετραγώνων (MAPE = 5.45%, MAE = 0.48, MSE = 0.53, R^2 = 0.60, explained_variance = 0.62 και accuracy = 94.55%). Αυτό είναι λογικό, καθώς η κανονικοποίηση που εφαρμόζεται έχει πολύ μικρή ισχύ λόγω της πολύ μικρής τιμής της υπερπαραμέτρου alpha.

Ως καλύτερο μοντέλο για την εκτίμηση του AMR προέκυψε πάλι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο. Αυτή τη φορά όμως το πλήθος των αναζητήσεων από τη πλατφόρμα Google Trends είναι ορισμένο στο χρονικό παράθυρο 6, που αντιστοιχεί στο διάστημα από τη μέρα προβολής του υπό εξέταση επεισοδίου έως και την προηγούμενη μέρα από το επόμενο επεισόδιο και τα χαρακτηριστικά από το Twitter είναι ορισμένα στο χρονικό παράθυρο 1 που αντιστοιχεί μόνο στη μέρα προβολής κάθε επεισοδίου. Επίσης, η τιμή της υπερπαραμέτρου alpha είναι και σε αυτήν την περίπτωση η μικρότερη δυνατή (10^{-10}), οδηγώντας σε μια σχεδόν ανεπαίσθητη κανονικοποίηση. Αυτό επιβεβαιώνεται και από τις σχεδόν ταυτόσημες τιμές των μετρικών αξιολόγησης στο σύνολο ελέγχου με την περίπτωση του αντίστοιχου μοντέλου πολλαπλής γραμμικής παλινδρόμησης που κατασκευάστηκε σύμφωνα με τη μέθοδο των ελάχιστων τετραγώνων (MAPE = 6.81%, MAE = 140298, MSE = 41204055988, R^2 = −0.35, explained_variance = −0.01 και accuracy = 93.19%).

6.5 Μοντέλα παλινδρόμησης LASSO

Ομοίως, η παλινδρόμηση LASSO ανήκει στις τεχνικές παλινδρόμησης κανονικοποίησης και μπορεί να αντιμετωπίσει την ύπαρξη συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Μάλιστα είναι δυνατόν ακόμα και να μηδενιστούν οι συντελεστές κάποιων χαρακτηριστικών, πραγματοποιώντας κατ΄ επέκταση επιλογή χαρακτηριστικών. Για αυτό, πάλι κρίθηκε σκόπιμο να μην εφαρμοστούν τεχνικές μείωσης της διαστατικότητας. Όπως και προηγουμένως, κατασκευάστηκαν 4 μοντέλα, τα οποία όλα αντιστοιχούν στην επιλογή 1 για τα διάφορα είδη και χρονικά παράθυρα χαρακτηριστικών που μπορούν να χρησιμοποιηθούν (α-δ). Η υλοποίησή τους πραγματοποιήθηκε με τη βοήθεια της ομώνυμης κλάσης Lasso[4] της βιβλιοθήκης scikit-learn. Δοκιμάστηκαν οι ίδιες τιμές για την υπερπαράμετρο alpha, η οποία εκφράζει την ισχύ της κανονικοποίησης, με την προηγούμενη περίπτωση της παλινδρόμησης κορυφογραμμής. Τα αποτελέσματα που λάβαμε φαίνονται στους πίνακες 6.9 και 6.10 και στο σχήμα 6.5.

Παρατηρούμε ότι στην περίπτωση της εκτίμησης του SHR% τα μοντέλα παλινδρόμησης που κατασκευάστηκαν, με εξαίρεση ένα, απέδωσαν καλύτερα από το μοντέλο αναφοράς. Μάλιστα, εκείνα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες εμ-

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

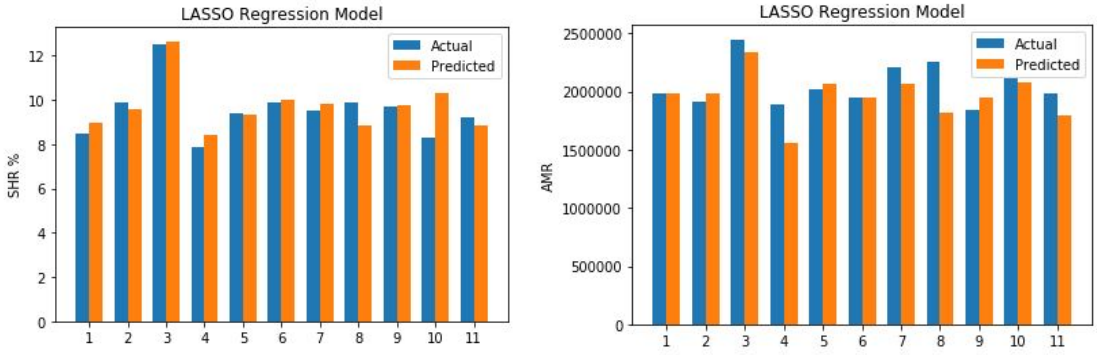
		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	α	**5.45%**	**0.49**	**0.53**	**0.60**	0.62	**94.55%**
	β	6.53%	0.59	0.57	0.57	**0.64**	93.47%
	γ	9.90%	0.95	1.32	0.01	0.05	90.10%
	δ	11.40%	1.06	1.59	-0.19	0.10	88.60%

Πίνακας 6.9: Αξιολόγηση των μοντέλων παλινδρόμησης LASSO για την εκτίμηση του SHR%

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	α	6.81%	140244	41120671278	-0.35	-0.01	93.19%
	β	**6.47%**	**134021**	**35317197738**	**-0.16**	**0.13**	**93.53%**
	γ	9.64%	199574	63722833344	-1.09	-0.83	90.36%
	δ	8.81%	180215	56296936329	-0.86	-0.79	91.19%

Πίνακας 6.10: Αξιολόγηση των μοντέλων παλινδρόμησης LASSO για την εκτίμηση του AMR

φάνισαν πολύ πιο ικανοποιητικές τιμές στις μετρικές αξιολόγησης από εκείνα που εκπαιδεύτηκαν με βάση όλα τα χαρακτηριστικά που προέκυψαν είτε από ποσοτικούς δείκτες είτε από ανάλυση συναισθήματος στα tweets. Αντιθέτως, στην περίπτωση της εκτίμησης του AMR, κανένα μοντέλο δεν απέδωσε καλύτερα από το αντίστοιχο μοντέλο αναφοράς. Σε πλήρη αντιστοιχία με την περίπτωση του SHR% όμως, τα μοντέλα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που σχετίζονται μόνο με ποσοτικούς δείκτες παρουσίασαν καλύτερη συμπεριφορά από τα υπόλοιπα, με αποτελέσματα συγκρίσιμα με εκείνα του μοντέλου αναφοράς.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.5: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης LASSO

Τελικά, το καλύτερο μοντέλο για την εκτίμηση του SHR% είναι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο, ορισμένα στο χρονικό παράθυρο 3 (που αντιστοιχεί μόνο στη επόμενη μέρα από την προβολή για το GTrends και στο διάστημα από μισή ώρα πριν έως και μισή ώρα μετά την προβολή του κάθε επεισοδίου για τα χαρακτηριστικά από το Twitter). Πάλι ξεχώρισε μικρή τιμή (10^{-6}) για την υπερπαράμετρο alpha, από τη διαδικασία της διασταυρούμενης επικύρωσης, γεγονός που οδήγησε σε αποτελέσματα στο σύνολο ελέγχου σχεδόν ίδια με εκείνα της πολλαπλής γραμμικής παλινδρόμησης που στηρίχθηκε στη μέθοδο των ελάχιστων τετραγώνων και της παλινδρόμησης κορυφογραμμής (MAPE = 5.45%, MAE = 0.49, MSE = 0.53, $R^2 = 0.60$, explained_variance = 0.62 και accuracy = 94.55%).

Ως καλύτερο μοντέλο παλινδρόμησης LASSO για την εκτίμηση του AMR προέκυψε πάλι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο. Αυτή τη φορά όμως τα χαρακτηριστικά ήταν ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τα αποτελέσματα της διερευνητικής ανάλυσης δεδομένων (3 για το GTrends και 4 για τα χαρακτη-

ριστικά από το Twitter). Επίσης, η τιμή της υπερπαραμέτρου alpha είναι (10^{-10}), οδηγώντας σε μια σχεδόν ανεπαίσθητη κανονικοποίηση. Τα αποτελέσματα της αξιολόγησης στο σύνολο ελέγχου είναι ελαφρώς καλύτερα από αυτά του αντίστοιχου γραμμικού μοντέλου που κατασκευάστηκε σύμφωνα με τη μέθοδο των ελάχιστων τετραγώνων (MAPE = 6.47%, MAE = 134021, MSE = 35317197738, $R^2 = -0.16$, explained_variance = 0.13 και accuracy = 93.53%).

6.6 Μοντέλα παλινδρόμησης elastic net

Η παλινδρόμηση elastic net αποτελεί άλλη μία περίπτωση κανονικοποίησης στην παλινδρόμηση, η οποία μάλιστα συνδυάζει τις δύο προηγούμενες τεχνικές (κορυφογραμμής και LASSO). Όπως και προηγουμένως, κατασκευάστηκαν 4 μοντέλα, τα οποία όλα αντιστοιχούν στην επιλογή 1 για τα διάφορα είδη και χρονικά παράθυρα χαρακτηριστικών που μπορούν να χρησιμοποιηθούν (α-δ). Η υλοποίηση τους πραγματοποιήθηκε με τη βοήθεια της ομώνυμης κλάσης ElasticNet[5] που διαθέτει η βιβλιοθήκη scikit-learn. Τώρα, πέρα από τη γνωστή μας υπερπαράμετρο alpha χρησιμοποιείται και η υπερπαράμετρος l1_ratio, η οποία ρυθμίζει σε τι ποσοστό πραγματοποιείται L_2 κανονικοποίηση, όπως στην περίπτωση της παλινδρόμησης κορυφογραμμής και σε τι ποσοστό πραγματοποιείται L_1 κανονικοποίηση, όπως στην περίπτωση της παλινδρόμησης LASSO. Η τιμή 0 ισοδυναμεί με το να χρησιμοποιείται μόνο L_2 κανονικοποίηση και η τιμή 1 με το να χρησιμοποιείται μόνο L_1 κανονικοποίηση. Για την υπερπαράμετρο alpha δοκιμάστηκαν οι ίδιες τιμές με τις δύο προηγούμενες περιπτώσεις ενώ για την υπερπαράμετρο l1_ratio δοκιμάστηκαν τιμές από 0 έως 1, καθώς μόνο αυτές έχουν νόημα για ένα ποσοστό, με βήμα 0.1. Τα αποτελέσματα που λάβαμε φαίνονται στους πίνακες 6.11 και 6.12 και αντίστοιχα στο σχήμα 6.6.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**5.45%**	**0.49**	**0.53**	**0.60**	0.62	**94.55%**
	β	6.53%	0.59	0.57	0.57	**0.64**	93.47%
	γ	10.08%	0.98	1.45	-0.09	-0.07	89.92%
	δ	10.19%	0.95	1.23	0.07	0.21	89.81%

Πίνακας 6.11: Αξιολόγηση των μοντέλων παλινδρόμησης elastic net για την εκτίμηση του SHR%

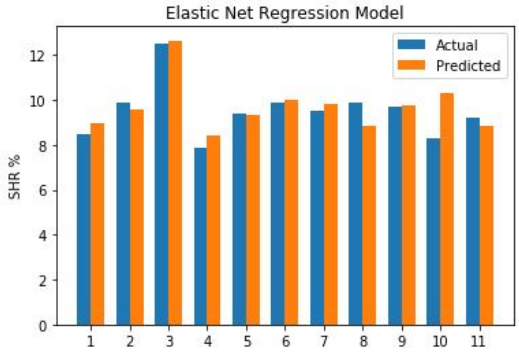
		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	6.81%	140244	41120671278	-0.35	-0.01	93.19%
	β	**6.47%**	**134021**	**35317197738**	**-0.16**	**0.13**	**93.53%**
	γ	9.41%	195684	59602666428	-0.95	-0.68	90.59%
	δ	8.57%	176164	51349897590	-0.68	-0.62	91.43%

Πίνακας 6.12: Αξιολόγηση των μοντέλων παλινδρόμησης elastic net για την εκτίμηση του AMR

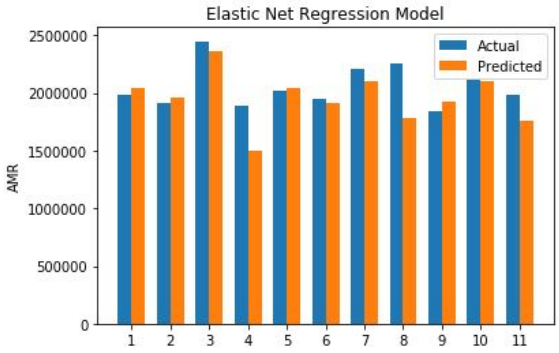
Παρατηρούμε ότι στην περίπτωση της εκτίμησης του SHR% τα μοντέλα παλινδρόμησης που κατασκευάστηκαν απέδωσαν καλύτερα από το μοντέλο αναφοράς. Όπως και προηγουμένως, εκείνα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που σχετίζονται με ποσοτικούς δείκτες εμφάνισαν πολύ πιο ικανοποιητικές τιμές στις μετρικές αξιολόγησης από εκείνα που εκπαιδεύτηκαν με βάση όλα τα χαρακτηριστικά που προέκυψαν είτε από ποσοτικούς δείκτες είτε από ανάλυση συναισθήματος στα tweets. Αντιθέτως, στην περίπτωση της εκτίμησης του AMR, κανένα μοντέλο δεν απέδωσε καλύτερα από το αντίστοιχο μοντέλο αναφοράς. Σε πλήρη αντιστοιχία με την περίπτωση του SHR% όμως, τα μοντέλα που εκπαιδεύτηκαν με όλα τα χαρακτηριστικά που

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

σχετίζονται μόνο με ποσοτικούς δείκτες παρουσίασαν καλύτερη συμπεριφορά από τα υπόλοιπα, με αποτελέσματα συγκρίσιμα με εκείνα του μοντέλου αναφοράς.



(α′) για το SHR%



(β′) για το AMR

Σχήμα 6.6: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης elastic net

Τελικά, το καλύτερο μοντέλο για την εκτίμηση του SHR% είναι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο, ορισμένα στο χρονικό παράθυρο 3 (που αντιστοιχεί μόνο στην επόμενη μέρα από την προβολή για το GTrends και στο διάστημα από μισή ώρα πριν έως και μισή ώρα μετά την προβολή του κάθε επεισοδίου για τα χαρακτηριστικά από το Twitter). Προέκυψε πάλι μικρή τιμή (10^{-6}) για την υπερπαράμετρο alpha από τη διαδικασία της διασταυρούμενης επικύρωσης. Το καλύτερο μοντέλο για την εκτίμηση του AMR προέκυψε πάλι εκείνο που εκπαιδεύτηκε με όλα τα χαρακτηριστικά ποσοτικών δεικτών ως είσοδο. Αυτή τη φορά όμως τα χαρακτηριστικά ήταν ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τα αποτελέσματα της διερευνητικής ανάλυσης δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Επίσης, η τιμή της υπερπαραμέτρου alpha είναι (10^{-10}), οδηγώντας σε μια σχεδόν ανεπαίσθητη κανονικοποίηση. Και στις δύο περιπτώσεις επιλέχθηκε η τιμή 1 για την υπερπαράμετρο l1_ratio, πράγμα που σημαίνει ότι τα καλύτερα μοντέλα παλινδρόμησης elastic net ταυτίζονται με τα καλύτερα μοντέλα παλινδρόμησης LASSO. Συνεπώς, οι αντίστοιχες τιμές αξιολόγησης στο σύνολο ελέγχου για την εκτίμηση και των δύο εξαρτημένων μεταβλητών είναι ίδιες με προηγουμένως.

6.7 Μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες

Επιπλέον, κατασκευάστηκαν 6 μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες για την εκτίμηση κάθε εξαρτημένης μεταβλητής. Αυτά χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με το συντελεστή συσχέτισης Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α ή β, οι οποίες και οι δύο αφορούν χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και διαφοροποιούνται ως προς τα χρονικά παράθυρα στα οποία αυτά ορίζονται. Αξίζει να σημειώσουμε πως από εδώ και πέρα τις περισσότερες φορές θα χρησιμοποιούμε μόνο χαρακτηριστικά ποσοτικών δεικτών ως εισόδους στα μοντέλα, αφενός γιατί τα μοντέλα αρχίζουν να γίνονται πιο σύνθετα και όσο πιο πολλές ανεξάρτητες μεταβλητές χρησιμοποιούνται τόσο πιο πολύ διαρκεί η εκπαίδευση και αφετέρου γιατί η μέχρι τώρα εμπειρία μας μάς δείχνει ότι τα χαρακτηριστικά από ανάλυση συναισθήματος δεν οδηγούν σε βελτιωμένα αποτελέσματα.

Η υλοποίηση όλων των μοντέλων παλινδρόμησης με γκαουσιανή διεργασία πραγματοποιήθηκε με τη βοήθεια της κλάσης GaussianProcessRegressor[6], που διαθέτει η βιβλιοθήκη scikit-learn.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.`

Οι υπερπαράμετροι που χρησιμοποιήθηκαν είναι οι εξής: kernel, η οποία αντιστοιχεί στον πυρήνα που καθορίζει τη συνάρτηση συνδιακύμανσης μιας γκαουσιανής διεργασίας, alpha, η οποία αντιστοιχεί σε μια τιμή που προστίθεται στη διαγώνιο του πίνακα πυρήνα και όσο μεγαλύτερη είναι τόσο υψηλότερο επίπεδο θορύβου υπάρχει στις παρατηρήσεις και η n_restarts_optimizer, η οποία καθορίζει το πλήθος των επανεκκινήσεων που θα κάνει ο αλγόριθμος βελτιστοποίησης για να βρει τις παραμέτρους του πυρήνα που μεγιστοποιούν την οριακή συνάρτηση πιθανοφάνειας. Στα πειράματά μας, ως πυρήνες δοκιμάστηκαν εσωτερικού γινομένου, ακτινικής συνάρτησης βάσης, τετραγωνικός ρητός και Matern, πάντα συνδυασμένοι με πυρήνα λευκού θορύβου ([DotProduct() + WhiteKernel(),RBF() + WhiteKernel(),RationalQuadratic() + WhiteKernel(), Matern() + WhiteKernel()]). Οι τιμές που δοκιμάστηκαν για την υπερπαράμετρο alpha είναι $[10^{-10}, 10^{-8}, 10^{-4}]$ ενώ για την n_restarts_optimizer δοκιμάστηκαν οι τιμές $[0, 5, 10]$. Τα αποτελέσματα που παράχθηκαν παρουσιάζονται στους πίνακες 6.13 και 6.14 και στο αντίστοιχο σχήμα 6.7.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**5.39%**	**0.49**	**0.52**	**0.61**	0.65	**94.61%**
	β	6.04%	0.53	0.60	0.55	**0.69**	93.96%
2	**α**	6.01%	0.55	0.57	0.57	0.62	93.99%
	β	6.04%	0.53	0.60	0.55	0.69	93.96%
3	**α**	6.00%	0.57	0.60	0.55	0.57	94.00%
	β	6.72%	0.62	0.63	0.53	0.61	93.28%

Πίνακας 6.13: Αξιολόγηση των μοντέλων παλινδρόμησης με γκαουσιανές διεργασίες για την εκτίμηση του SHR%

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	4.70%	100609	20194604574	0.34	0.51	95.30%
	β	**4.70%**	**100609**	**20194604574**	**0.34**	**0.51**	**95.30%**
2	**α**	4.70%	100609	20194604574	0.34	0.51	95.30%
	β	4.70%	100609	20194604574	0.34	0.51	95.30%
3	**α**	4.70%	100609	20194604574	0.34	0.51	95.30%
	β	4.70%	100609	20194604574	0.34	0.51	95.30%

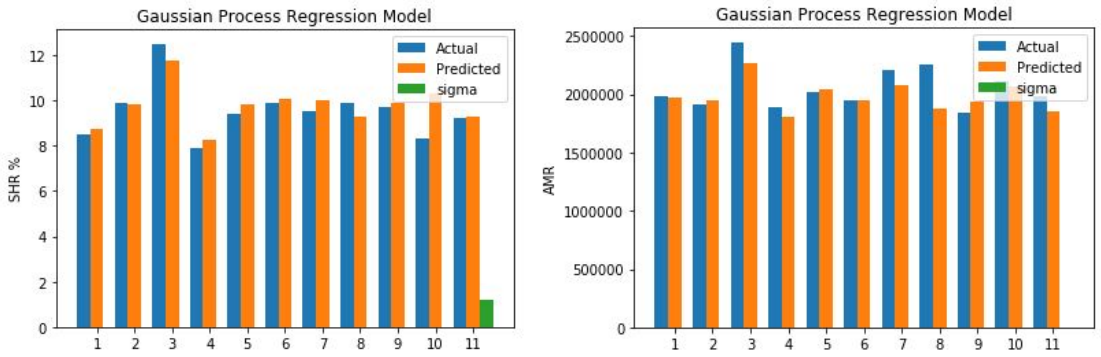
Πίνακας 6.14: Αξιολόγηση των μοντέλων παλινδρόμησης με γκαουσιανές διεργασίες για την εκτίμηση του AMR

Όπως γίνεται φανερό, στην περίπτωση της εκτίμησης του μεγέθους SHR% όλα τα μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες που κατασκευάστηκαν παρουσίασαν καλύτερα αποτελέσματα από το μοντέλο αναφοράς. Εκείνο που ξεχώρισε περισσότερο είναι αυτό που χρησιμοποίησε όλα τα χαρακτηριστικά ποσοτικών δεικτών, εκ των οποίων το GTrends είναι ορισμένο στο χρονικό παράθυρο 3, που αντιστοιχεί στην επόμενη μέρα από την προβολή κάθε επεισοδίου και τα χαρακτηριστικά από το Twitter είναι ορισμένα στο χρονικό παράθυρο 1, που αντιστοιχεί μόνο στη μέρα προβολής του κάθε επεισοδίου. Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχθηκε ο πυρήνας γινομένου συνδυασμένος με λευκό θόρυβο, η τιμή 10^{-8} για την υπερπαράμετρο alpha και ως πλήθος επανεκκινήσεων του αλγορίθμου βελτιστοποίησης 10. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.39%, MAE = 0.49, MSE = 0.52, R^2 = 0.61, explained_variance = 0.65 και accuracy = 94.61%. Αυτές οι μετρήσεις προκύπτουν από τις προβλέψεις που έκανε τελικά το μοντέλο στο σύνολο ελέγχου χρησιμοποιώντας ως καμπύλη παλινδρόμησης αυτή που αντιστοιχεί στη μέση τιμή. Κάθε προβλεπόμενη τιμή συνοδεύτηκε από μια τιμή τυπικής απόκλισης, η οποία ήταν αρκετά μικρή, σχεδόν μηδενική τις περισσότερες φορές,

GaussianProcessRegressor.html

όπως φαίνεται στο σχήμα 6.7.

Όσον αφορά την εκτίμηση του μεγέθους AMR, όλα τα μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες που κατασκευάστηκαν οδήγησαν στα ίδια αποτελέσματα, τα οποία είναι εμφανώς καλύτερα από εκείνα του μοντέλου αναφοράς. Για την εκπαίδευσή τους, χρησιμοποιήθηκαν όλα τα χαρακτηριστικά ποσοτικών δεικτών, τα οποία ορίζονται στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχθηκε ο πυρήνας γινομένου συνδυασμένος με λευκό θόρυβο, η τιμή 10^{-10} για την υπερπαράμετρο alpha και ως πλήθος επανεκκινήσεων του αλγορίθμους βελτιστοποίησης 0. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE $= 4.70\%$, MAE $= 100609$, MSE $= 20194604574$, $R^2 = 0.34$, explained_variance $= 0.51$ και accuracy $= 95.30\%$. Πάλι, η κάθε πρόβλεψη συνοδεύτηκε από μια τιμή τυπικής απόκλισης, τάξεις μεγέθους μικρότερη από την εκτιμώμενη τιμή (~ 300) για όλες τις παρατηρήσεις στο σύνολο ελέγχου, οπότε και δεν είναι καν ορατή στο σχήμα 6.7.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.7: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με γκαουσιανές διεργασίες

6.8 Μοντέλα παλινδρόμησης με δέντρα απόφασης

Επίσης, κατασκευάστηκαν 6 μοντέλα παλινδρόμησης με δέντρα απόφασης για την εκτίμηση της καθεμίας εξαρτημένης μεταβλητής. Αυτά χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με το συντελεστή συσχέτισης Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α ή β, οι οποίες και οι δύο αφορούν χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και διαφοροποιούνται ως προς τα χρονικά παράθυρα στα οποία αυτά ορίζονται.

Η υλοποίησή όλων αυτών των μοντέλων πραγματοποιήθηκε με τη βοήθεια της κλάσης DecisionTreeRegressor[7], που διαθέτει η βιβλιοθήκη scikit-learn. Σε όλους τους κατασκευαστές των μοντέλων δόθηκε ο ίδιος ακέραιος αριθμός ως τιμή στην υπερπαράμετρο random_state, ώστε να διασφαλίζεται η αναπαραξιμότητα των αποτελεσμάτων και να καθίσταται δυνατή η σύγκριση μεταξύ των διαφορετικών μοντέλων. Πέρα από αυτήν, τα μοντέλα παλινδρόμησης με δέντρα απόφασης διαθέτουν ένα μεγάλο πλήθος υπερπαραμέτρων. Κατ' αρχάς, το criterion καθορίζει τη συνάρτηση σύμφωνα με την οποία θα μετρηθεί η ποιότητα της διάσπασης ενός κόμβου-πατέρα σε περισσότερους κόμβους-παιδιά. Οι συναρτήσεις που δοκιμάστηκαν είναι η 'mse', η οποία ισοδυναμεί με τη μείωση της διακύμανσης, ελαχιστοποιώντας κάθε φορά την L_2 νόρμα των απωλειών, η 'friedman_mse', που ταυτίζεται με το mse, προσθέτοντας τη βελτίωση που πρότεινε ο Friedman

[7] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

για το ποια διάσπαση πιθανοτικά είναι καλύτερη και η 'mae', η οποία ισοδυναμεί με την ελαχιστοποίηση της L_1 νόρμας των απωλειών. Μια ακόμη υπερπαράμετρος είναι ο splitter, ο οποίος καθορίζει τη στρατηγική της διάσπασης και μπορεί να λάβει τις τιμές είτε 'best' για την επιλογή της καλύτερης δυνατής διάσπασης κάθε φορά είτε 'random' για την επιλογή της καλύτερης τυχαίας διάσπασης κάθε φορά.

Επίσης, η υπερπαράμετρος max_depth καθορίζει το μέγιστο επιτρεπτό βάθος στο οποίο μπορεί να επεκταθεί το υπό κατασκευή δέντρο απόφασης. Οι τιμές που δοκιμάστηκαν για αυτήν ανήκουν στο εύρος από 10 έως 100 με βήμα 10, καθώς και η τιμή 'None', η οποία ισοδυναμεί με την απουσία άνω ορίου στο βάθος που μπορεί να επεκταθεί. Η υπερπαράμετρος min_samples_split ορίζει το ελάχιστο πλήθος παρατηρήσεων που πρέπει να αντιστοιχούν σε έναν κόμβο ώστε αυτός να μπορεί να διασπαστεί σε περισσότερους. Οι τιμές που δοκιμάστηκαν για αυτήν είναι όλοι οι ακέραιοι στο διάστημα $[2, 5]$. Προς την ίδια κατεύθυνση, η υπερπαράμετρος min_samples_leaf καθορίζει το πλήθος των παρατηρήσεων που πρέπει να έχει τουλάχιστον ένας κόμβος για να θεωρηθεί φύλλο. Στα πειράματά μας δοκιμάστηκαν οι τιμές 1,2 και 3. Τέλος, η υπερπαράμετρος max_leaf_nodes καθορίζει αν υπάρχει μέγιστο επιτρεπτό πλήθος κόμβων-φύλλων και ποιο είναι αυτό. Οι τιμές που δοκιμάστηκαν για αυτήν την υπερπαράμετρο είναι οι εξής: $[10, 20, 50, 80, 100]$. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.15 και 6.16 και στο αντίστοιχο σχήμα 6.8.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	11.05%	1.01	1.68	-0.26	-0.10	88.95%
	β	10.41%	0.95	1.56	-0.18	-0.10	89.59%
2	**α**	8.44%	0.75	0.92	0.31	0.50	91.56%
	β	9.16%	0.84	1.15	0.13	0.32	90.84%
3	**α**	8.82%	0.78	0.92	0.31	0.60	91.18%
	β	**6.62%**	**0.59**	**0.62**	**0.54**	**0.74**	**93.38%**

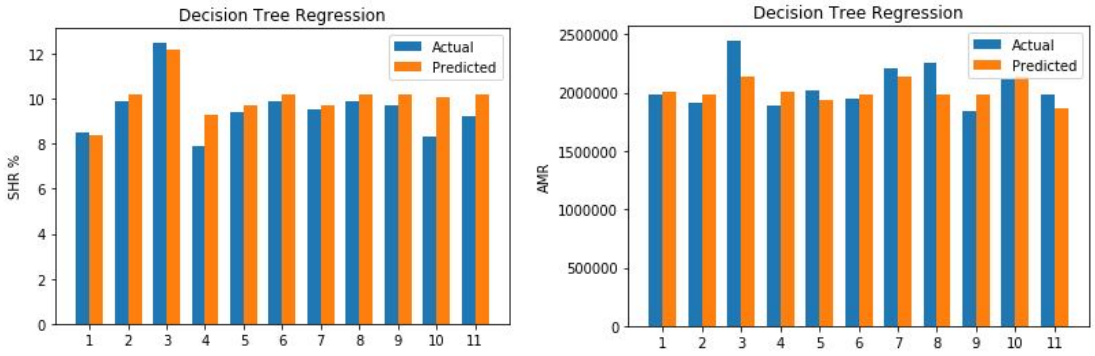
Πίνακας 6.15: Αξιολόγηση των μοντέλων παλινδρόμησης με δέντρα απόφασης για την εκτίμηση του SHR%

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	6.86%	148193	42589479082	-0.40	-0.20	93.14%
	β	8.28%	171750	41757459727	-0.37	-0.37	91.72%
2	**α**	8.32%	178028	57765387887	-0.89	-0.77	91.68%
	β	5.84%	124493	29938669438	0.02	0.09	94.16%
3	**α**	7.29%	152262	40325220217	-0.32	-0.07	92.71%
	β	**5.40%**	**114676**	**21505302309**	**0.30**	**0.35**	**94.60%**

Πίνακας 6.16: Αξιολόγηση των μοντέλων παλινδρόμησης με δέντρα απόφασης για την εκτίμηση του AMR

Όπως μπορούμε να δούμε, στην περίπτωση της εκτίμησης του μεγέθους SHR% σχεδόν όλα τα μοντέλα παλινδρόμησης με δέντρα απόφασης που παράχθηκαν παρουσίασαν καλύτερη απόδοση στο σύνολο ελέγχου σε σχέση με το μοντέλο αναφοράς, με εξαίρεση ένα που παρουσίασε συγκρίσιμα αποτελέσματα με αυτά του μοντέλου αναφοράς. Το καλύτερο, όμως, που ξεχώρισε είναι εκείνο που χρησιμοποίησε χαρακτηριστικά ποσοτικών δεικτών, έπειτα από ανάλυση σε κύριες συνιστώσες, τα οποία ορίζονταν στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως κριτήριο το 'mae', ως στρατηγική διάσπασης 'random', ως μέγιστο επιτρεπτό βάθος 10, ως min_samples_split 2, ως min_samples_leaf 1 και ως max_leaf_nodes 20 ενώ κρατήθηκαν μόνο 3 κύριες συνιστώσες. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 6.62%, MAE = 0.59, MSE = 0.62, R^2 = 0.54, explained_variance = 0.74 και accuracy = 93.38%.

Όσον αφορά την εκτίμηση του μεγέθους AMR μόλις δύο μοντέλα παλινδρόμησης με δέντρα απόφασης που παράχθηκαν παρουσίασαν καλύτερη απόδοση στο σύνολο ελέγχου σε σχέση με το μοντέλο αναφοράς ενώ όλα τα υπόλοιπα παρουσίασαν συγκρίσιμα ή και χειρότερα αποτελέσματα από αυτό. Το καλύτερο, όμως, που ξεχώρισε, είναι εκείνο που χρησιμοποίησε χαρακτηριστικά ποσοτικών δεικτών, έπειτα από ανάλυση σε κύριες συνιστώσες, τα οποία ορίζονταν στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως κριτήριο το 'mse', ως στρατηγική διάσπασης 'random', ως μέγιστο επιτρεπτό βάθος 10, ως min_samples_split 3, ως min_samples_leaf 2 και ως max_leaf_nodes 10 ενώ κρατήθηκαν 6 κύριες συνιστώσες. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.40%, MAE = 114676, MSE = 21505302309, $R^2 = 0.30$, explained_variance = 0.35 και accuracy = 94.60%.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.8: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με δέντρα απόφασης

6.9 Μοντέλα παλινδρόμησης με τυχαία δάση

Επίσης, κατασκευάστηκαν 4 μοντέλα παλινδρόμησης με τυχαία δάση για την εκτίμηση της καθεμίας εξαρτημένης μεταβλητής, τα οποία συνδυάζουν πολλούς μεμονωμένους ανεξάρτητους εκτιμητές δέντρων απόφασης. Αυτά χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με τη μετρική Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α ή β, οι οποίες και οι δύο αφορούν χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και διαφοροποιούνται ως προς τα χρονικά παράθυρα στα οποία αυτά ορίζονται.

Η υλοποίησή όλων αυτών των μοντέλων πραγματοποιήθηκε με τη βοήθεια της κλάσης RandomForestRegressor[8], που διαθέτει η βιβλιοθήκη scikit-learn. Χρησιμοποιήθηκαν και πάλι όλες οι υπερπαράμετροι που ορίστηκαν στην προηγούμενη περίπτωση των δέντρων απόφασης, εκτός από το splitter, με ακριβώς τα ίδια εύρη τιμών να δοκιμάζονται. Επιπλέον, ορίστηκε η υπερπαράμετρος n_estimators, η οποία αντιστοιχεί στο πλήθος των ανεξάρτητων δέντρων απόφασης, που θα συνδυαστούν σε έναν εκτιμητή μέσω του τυχαίου δάσους. Οι τιμές που δοκιμάστηκαν για αυτό είναι $[10, 30, 50, 70, 100]$. Επίσης, χρησιμοποιήθηκε και η υπερπαράμετρος bootstrap, η οποία παίρνει τιμές True ή False αναλόγως με το αν χρησιμοποιείται bootsrap δειγματοληψία ώστε κάθε δέντρο απόφασης να εκπαιδεύεται με ένα μόνο μέρος των δεδομένων εκπαίδευσης ή όχι ώστε κάθε δέντρο απόφασης να εκπαιδεύεται με το σύνολο των δεδομένων εκπαίδευσης. Τα

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

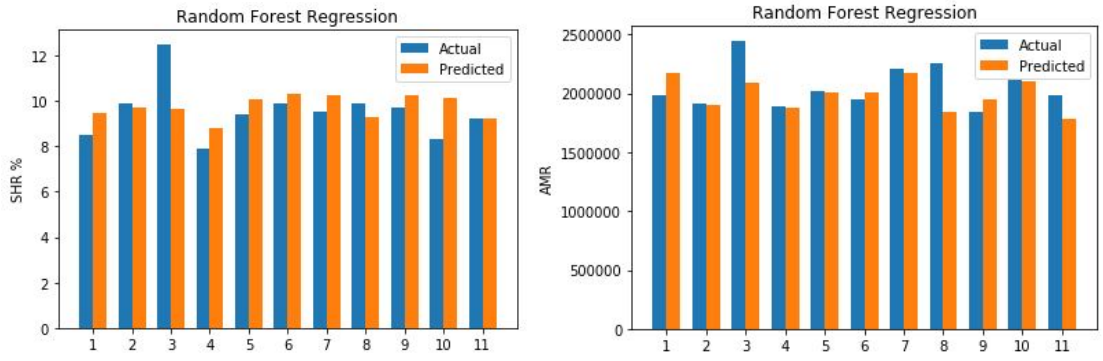
αποτελέσματα που προέκυψαν παρουσιάζονται στους πίνακες 6.17 και 6.18 και στο αντίστοιχο σχήμα 6.9.

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	9.94%	0.90	1.47	-0.11	-0.03	90.06%
	β	10.89%	1.05	2.19	-0.65	-0.62	89.11%
2	**β**	11.05%	1.06	2.49	-0.88	-0.84	88.95%
3	**β**	**9.12%**	**0.88**	**1.36**	**-0.03**	**0.01**	**90.88%**

Πίνακας 6.17: Αξιολόγηση των μοντέλων παλινδρόμησης με τυχαία δάση για την εκτίμηση του SHR%

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	7.72%	162945	41062206568	-0.35	-0.26	92.28%
	β	**5.92%**	**127828**	**35918457874**	**-0.18**	**-0.04**	**94.08%**
2	**β**	6.08%	131107	36521504825	-0.20	-0.09	93.92%
3	**β**	8.14%	171500	49055529890	-0.61	-0.40	91.86%

Πίνακας 6.18: Αξιολόγηση των μοντέλων παλινδρόμησης με τυχαία δάση για την εκτίμηση του AMR



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.9: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με τυχαία δάση

Όπως γίνεται φανερό, στην περίπτωση της εκτίμησης του μεγέθους SHR% τα μισά μοντέλα παλινδρόμησης με τυχαία δάση παρουσίασαν καλύτερη απόδοση στο σύνολο ελέγχου σε σχέση με το μοντέλο αναφοράς, ενώ τα άλλα μισά κυμάνθηκαν στο ίδιο επίπεδο αποτελεσμάτων. Κανένα όμως δεν εμφάνισε εξαιρετικά καλή απόδοση. Εκείνο που ξεχώρισε περισσότερο είναι αυτό που χρησιμοποίησε χαρακτηριστικά ποσοτικών δεικτών, έπειτα από ανάλυση σε κύριες συνιστώσες, τα οποία ορίζονταν στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως πλήθος επιμέρους δέντρων απόφασης 50, ως κριτήριο το 'mse', ως μέγιστο επιτρεπτό βάθος 20, ως min_samples_split 2, ως min_samples_leaf 1, ως max_leaf_nodes 20 και τέθηκε bootstrap = True ενώ κρατήθηκαν 9 κύριες συνιστώσες. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 9.12%, MAE = 0.88, MSE = 1.36, $R^2 = -0.03$, explained_variance = 0.01 και accuracy = 90.88%.

Όσον αφορά την εκτίμηση του μεγέθους AMR μόνο ένα μοντέλο από αυτά που κατασκευάστηκαν παρουσίασε σαφώς καλύτερη απόδοση στο σύνολο ελέγχου σε σχέση με το μοντέλο

αναφοράς, άλλο ένα παρουσίασε σχεδόν την ίδια με ελαφρώς βελτιωμένη απόδοση σε κάποιες μετρικές ενώ τα υπόλοιπα δύο παρουσίασαν χειρότερη. Εκείνο που ξεχώρισε είναι αυτό που χρησιμοποίησε όλα τα χαρακτηριστικά ποσοτικών δεικτών, τα οποία ορίζονται στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως πλήθος επιμέρους δέντρων απόφασης 70, ως κριτήριο το 'mae', ως μέγιστο επιτρεπτό βάθος 10, ως min_samples_split 2, ως min_samples_leaf 2, ως max_leaf_nodes 20 και τέθηκε bootstrap = True. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.92%, MAE = 127828, MSE = 35918457874, $R^2 = -0.18$, explained_variance = -0.04 και accuracy = 94.08%.

6.10 Μοντέλα παλινδρόμησης με gradient boosting μηχανές

Άλλη μία συνδυαστική μέθοδος παλινδρόμησης είναι αυτή των gradient boosting μηχανών, οι οποίες συνδυάζουν σειριακά δέντρα απόφασης σε έναν εκτιμητή με τέτοιο τρόπο, ώστε το επόμενο να βελτιώνει την απόδοση του προηγούμενου. Συνολικά, κατασκευάστηκαν 10 μοντέλα για την εκτίμηση και των δύο εξαρτημένων μεταβλητών. Αυτά χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με τη μετρική Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α ή β, οι οποίες και οι δύο αφορούν χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και διαφοροποιούνται ως προς τα χρονικά παράθυρα στα οποία αυτά ορίζονται.

Η υλοποίηση όλων αυτών των μοντέλων πραγματοποιήθηκε με τη βοήθεια της κλάσης GradientBoostingRegressor[9], που διαθέτει η βιβλιοθήκη scikit-learn. Χρησιμοποιήθηκαν και πάλι όλες οι υπερπαράμετροι n_estimators, criterion, min_samples_split, min_samples_leaf, max_leaf_nodes και max_depth. Για τις 4 πρώτες υπερπαραμέτρους, που αναφέρθηκαν, δοκιμάστηκαν οι ίδιες τιμές με προηγουμένως. Τώρα όμως, για να μην υπάρχει ο κίνδυνος της υπερεκπαίδευσης, θα ορίσουμε πολύ μικρότερο εύρος τιμών για το μέγιστο επιτρεπτό βάθος και κατ΄ επέκταση για το μέγιστο επιτρεπτό πλήθος φύλλων. Πιο συγκεκριμένα, για το μέγιστο βάθος δοκιμάζονται οι τιμές $[3, 5, 8]$ ενώ για το άνω όριο στο πλήθος των φύλλων δοκιμάζονται οι τιμές $[10, 20, 50]$.

Επίσης, χρησιμοποιήθηκε η υπερπαράμετρος loss, η οποία αντιστοιχεί στη συνάρτηση κόστους που πρόκειται να ελαχιστοποιηθεί και μπορεί να λάβει τις τιμές 'ls','lad' και 'huber'. Η πρώτη αντιστοιχεί στα ελάχιστα τετράγωνα, η δεύτερη στην ελάχιστη απόλυτη απόκλιση και η τρίτη στο συνδυασμό αυτών των δύο. Μια ακόμη υπερπαράμετρος, που χρησιμοποιήθηκε, είναι το learning_rate, ο ρυθμός εκπαίδευσης, ο οποίος μεταβάλλει το πόσο συνεισφέρει κάθε επιπλέον δέντρο απόφασης που προστίθεται στο τελικό αποτέλεσμα. Οι τιμές που δοκιμάστηκαν για αυτό είναι $[0.05, 0.1, 0.2]$. Τέλος, η υπερπαράμετρος subsample καθορίζει το ποσοστό των δειγμάτων που χρησιμοποιούνται για την εκπαίδευση των επιμέρους εκτιμητών - δέντρων απόφασης. Όσο μικρότερο είναι αυτό το ποσοστό τόσο μειώνεται η διακύμανση και αυξάνεται η απόκλιση στο τελικό μοντέλο. Στα πειράματά μάς, δοκιμάστηκαν οι τιμές $[0.8, 0.9, 1.0]$. Τα αποτελέσματα που παράχθηκαν παρουσιάζονται στους πίνακες 6.19 και 6.20 και στο αντίστοιχο σχήμα 6.10.

Όπως γίνεται φανερό, στην περίπτωση της εκτίμησης του μεγέθους SHR% όλα τα μοντέλα παλινδρόμησης με gradient boosting μηχανές που κατασκευάστηκαν, με εξαίρεση ένα, παρουσίασαν καλύτερα αποτελέσματα από το μοντέλο αναφοράς. Κανένα όμως δεν εμφάνισε εξαιρετικά καλή απόδοση. Εκείνο που ξεχώρισε περισσότερο είναι αυτό που χρησιμοποίησε όλα τα χαρακτηριστικά ποσοτικών δεικτών, εκ των οποίων το GTrends είναι ορισμένο στο χρονικό παράθυρο 2, που αντιστοιχεί στη μέρα προβολής και την επόμενη κάθε επεισοδίου και τα χαρακτηριστικά από το Twitter είναι ορισμένα στο χρονικό παράθυρο 5, που αντιστοιχεί στο διάστημα από

[9] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

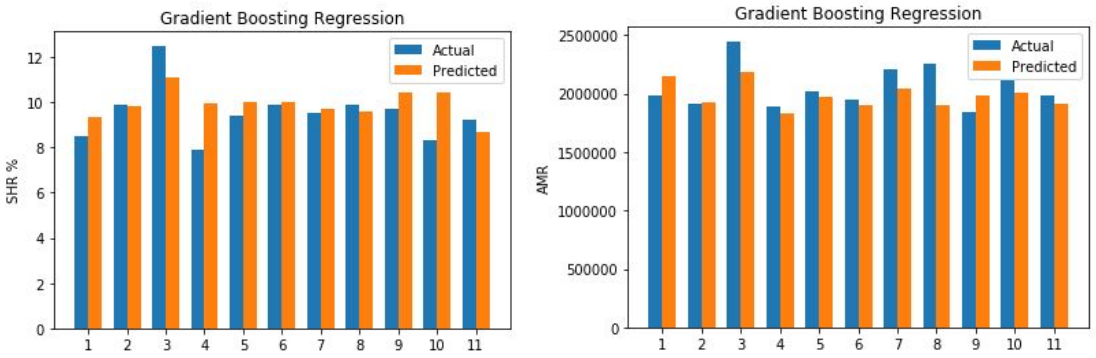
		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	**9.04%**	**0.82**	1.16	0.13	**0.24**	**90.96%**
	β	11.97%	1.15	2.66	-1.00	-0.99	88.02%
2	**α**	10.26%	0.98	1.60	-0.21	-0.20	89.74%
	β	10.72%	1.00	1.67	-0.25	-0.15	89.28%
3	**α**	10.06%	0.92	1.32	0.00	0.22	89.94%
	β	9.27%	0.88	**1.12**	**0.16**	0.16	90.73%

Πίνακας 6.19: Αξιολόγηση των μοντέλων παλινδρόμησης με gradient boosting μηχανές για την εκτίμηση του SHR%

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	7.59%	160999	53004928277	-0.74	-0.32	92.41%
	β	7.43%	154497	50943039914	-0.67	-0.49	92.56%
2	β	8.29%	180509	72676085435	-1.38	-1.15	91.71%
3	β	**6.08%**	**129706**	**26168817442**	**0.14**	**0.31**	**93.92%**

Πίνακας 6.20: Αξιολόγηση των μοντέλων παλινδρόμησης με gradient boosting μηχανές για την εκτίμηση του AMR

τη δεύτερη επόμενη μέρα του προηγούμενου επεισοδίου έως και την επόμενη από την προβολή του υπό εξέταση επεισοδίου. Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως πλήθος επιμέρους δέντρων απόφασης 100, ως κριτήριο το 'friedman_mse', ως μέγιστο επιτρεπτό βάθος 5, ως min_samples_split 2, ως min_samples_leaf 1, ως max_leaf_nodes 20 και ως συνάρτηση κόστους 'lad'. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 9.04%, MAE = 0.82, MSE = 1.16, $R^2 = 0.13$, explained_variance = 0.24 και accuracy = 90.96%.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.10: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με gradient boosting μηχανές

Όσον αφορά την εκτίμηση του μεγέθους AMR μόνο ένα μοντέλο από αυτά που κατασκευάστηκαν παρουσίασε ελαφρώς καλύτερη απόδοση στο σύνολο ελέγχου από το μοντέλο αναφοράς ενώ όλα τα άλλα παρουσίασαν χειρότερη. Το καλύτερο μοντέλο είναι αυτό που χρησιμοποίησε χαρακτηριστικά ποσοτικών δεικτών, έπειτα από την ανάλυσή τους σε κύριες συνιστώσες, τα οποία ορίζονται στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, επιλέχτηκαν ως πλήθος επιμέρους δέντρων απόφασης 10, ως κριτήριο το 'mae', ως μέγιστο επιτρεπτό βάθος 3, ως min_samples_split 3, ως min_samples_leaf 2, ως max_leaf_nodes 10 και ως loss 'ls' ενώ επιλέχθηκαν 7 κύριες συνιστώσες στις οποίες αποδίδεται

το 100% της διακύμανσης των παρατηρήσεων. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 6.08%, MAE = 129706, MSE = 26168817442, $R^2 = 0.14$, explained_variance = 0.31 και accuracy = 93.92%.

6.11 Μοντέλα παλινδρόμησης με πολυεπίπεδα perceptron

Επίσης, κατασκευάστηκαν 6 μοντέλα παλινδρόμησης με πολυεπίπεδα perceptron για την εκτίμηση της καθεμίας εξαρτημένης μεταβλητής. Αυτά χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με τη μετρική Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α ή β, οι οποίες και οι δύο αφορούν χαρακτηριστικά σχετιζόμενα με ποσοτικούς δείκτες και διαφοροποιούνται ως προς τα χρονικά παράθυρα στα οποία αυτά ορίζονται.

Η υλοποίησή όλων αυτών των μοντέλων πραγματοποιήθηκε με τη βοήθεια της κλάσης MLPRegressor[10], που διαθέτει η βιβλιοθήκη scikit-learn. Σε όλα χρησιμοποιήθηκε ο αλγόριθμος 'lbfgs' για τη βελτιστοποίηση των βαρών, ο οποίος δόθηκε ως τιμή στην υπερπαράμετρο solver, καθώς όπως αναφέρεται στην τεκμηρίωση της κλάσης ο συγκεκριμένος αλγόριθμος συγκλίνει πιο γρήγορα και αποδίδει καλύτερα σε σχετικά μικρά σύνολα δεδομένων, όπως το δικό μας. Επίσης, κάθε φορά δινόταν ο ίδιος ακέραιος αριθμός ως τιμή στην υπερπαράμετρο random_state, ώστε να διασφαλίζεται η αναπαραξιμότητα των αποτελεσμάτων και να καθίσταται δυνατή η σύγκριση μεταξύ των διαφορετικών μοντέλων.

Άλλες υπερπαράμετροι που χρησιμοποιήθηκαν είναι οι εξής: hidden_layer_sizes, activation και alpha. Η πρώτη καθορίζει το πλήθος των κρυφών επιπέδων και το πλήθος των νευρώνων σε καθένα από αυτά. Επειδή το πρόβλημα που καλούμαστε να αντιμετωπίσουμε δεν είναι ιδιαίτερα σύνθετο και τα δείγματα δε φαίνεται να ακολουθούν κάποιο ιδιαίτερα περίπλοκο πρότυπο, επιλέξαμε να χρησιμοποιήσουμε σχετικά «ρηχά» νευρωνικά δίκτυα με ένα ή δύο κρυφά επίπεδα και από 2 έως 20 νευρώνες το καθένα. Πιο συγκεκριμένα οι τιμές που δόθηκαν στην υπερπαράμετρο hidden_layer_sizes είναι οι εξής: $(2,), (5,), (10,), (20,), (2,2), (5,5), (5,2), (10,10), (10,5), (20,10)$. Η δεύτερη υπερπαράμετρος (activation) αντιστοιχεί στη συνάρτηση ενεργοποίησης και για αυτή δοκιμάστηκαν οι τιμές 'identity', 'logistic', 'tanh', 'relu', οι οποίες αντιστοιχούν όπως προδίδουν τα ονόματά τους στις συναρτήσεις γραμμική ταυτότητας, λογιστική, υπερβολικής εφαπτομένης και ράμπας. Η τελευταία υπερπαράμετρος alpha είναι μια παράμετρος L_2 κανονικοποίησης - τιμωρίας. Οι τιμές που δοκιμάστηκαν για αυτήν είναι από 0.00001 έως και 0.001 με πολλαπλασιαστικό βήμα 10. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.21 και 6.22 και στο αντίστοιχο σχήμα 6.11.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	6.90%	0.63	**0.66**	**0.51**	**0.56**	93.10%
	β	**6.73%**	**0.62**	0.85	0.36	0.54	**93.27%**
2	**α**	10.83%	1.04	2.01	-0.51	-0.42	89.17%
	β	6.73%	0.62	0.85	0.36	0.54	93.27%
3	**α**	8.40%	0.80	1.10	0.17	0.17	91.60%
	β	10.79%	1.01	1.66	-0.25	-0.10	89.21%

Πίνακας 6.21: Αξιολόγηση των μοντέλων παλινδρόμησης με πολυεπίπεδο perceptron για την εκτίμηση του SHR%

Όπως μπορούμε να δούμε, στην περίπτωση της εκτίμησης του μεγέθους SHR% τα δύο μοντέλα που ξεχώρισαν όσον αφορά την απόδοσή τους στο σύνολο των δεδομένων ελέγχου είναι εκείνα που χρησιμοποίησαν όλα τα χαρακτηριστικά ποσοτικών δεικτών, εκ των οποίων το

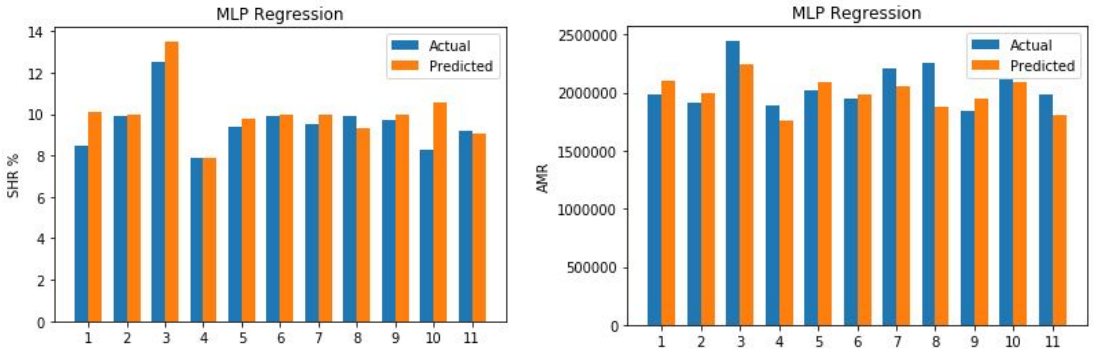
[10] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	8.66%	177093	51757548710	-0.70	-0.48	91.34%
	β	6.74%	139302	33951802957	-0.11	-0.02	93.26%
2	**α**	7.32%	150695	36982108667	-0.21	0.05	92.68%
	β	**6.36%**	**133724**	**26529425505**	**0.13**	**0.25**	**93.64%**
3	**α**	8.56%	175474	47548497741	-0.56	-0.22	91.43%
	β	8.13%	167161	46059915805	-0.51	-0.46	91.87%

Πίνακας 6.22: Αξιολόγηση των μοντέλων παλινδρόμησης με πολυεπίπεδο perceptron για την εκτίμηση του AMR

GTrends ήταν ορισμένο στο χρονικό παράθυρο 3 και τα εξαγόμενα χαρακτηριστικά από το Twitter ήταν ορισμένα είτε στο χρονικό παράθυρο 3 είτε στο χρονικό παράθυρο 4. Στην πρώτη περίπτωση, από τη διασταυρούμενη επικύρωση, τελικά προέκυψαν 1 κρυφό επίπεδο με 2 νευρώνες, συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης και alpha = 0.01. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 6.90%, MAE = 0.63, MSE = 0.66, R^2 = 0.51, explained_variance = 0.56 και accuracy = 93.10%. Στη δεύτερη περίπτωση, από τη διασταυρούμενη επικύρωση τελικά επιλέχτηκαν πάλι 1 κρυφό επίπεδο με 2 νευρώνες, η ράμπα ως συνάρτηση ενεργοποίησης και η τιμή 10^{-5} ως παράμετρος κανονικοποίησης. Τα αποτελέσματα που επιτεύχθηκαν είναι εξίσου ικανοποιητικά (MAPE = 6.73%, MAE = 0.62, MSE = 0.85, R^2 = 0.36, explained_variance = 0.54 και accuracy = 93.27%).

Όσον αφορά την εκτίμηση του μεγέθους AMR, κανένα μοντέλο δεν παρουσιάσε καλύτερη συμπεριφορά από το μοντέλο αναφοράς. Το καλύτερο μοντέλο είναι εκείνο που χρησιμοποίησε επιλεγμένα χαρακτηριστικά ποσοτικών δεικτών, ορισμένα στα καλύτερα χρονικά παράθυρα, σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά επιλέχθηκαν 2 κρυμμένα επίπεδα με 2 νευρώνες το καθένα, η ράμπα ως συνάρτηση ενεργοποίησης και η τιμή 0.001 για την υπερπαράμετρο κανονικοποίησης ενώ κρατήθηκαν τα 5 καλύτερα χαρακτηριστικά, σύμφωνα με το συντελεστή συσχέτισης του Pearson. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 6.36%, MAE = 133724, MSE = 26529425505, R^2 = 0.13, explained_variance = 0.25 και accuracy = 93.64%.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.11: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης με πολυεπίπεδο perceptron

6.12 Μοντέλα παλινδρόμησης με διανύσματα υποστήριξης

Σε αυτό το σημείο, κατασκευάστηκαν μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με τη βοήθεια της κλάσης SVR[11], που διαθέτει η βιβλιοθήκη scikit-learn. Για τα μοντέλα χρησιμοποιήθηκαν 4 είδη πυρήνων, γραμμικός, ακτινικής συνάρτησης βάσης, πολυωνυμικός και σιγμοειδής, οι οποίοι κάθε φορά δίνονταν στο όρισμα kernel, που παίρνει ο κονστράκτορας της κλάσης SVR, ως 'linear', 'rbf', 'poly' και 'sigmoid', αντίστοιχα. Για κάθε είδος πυρήνα, παράχθηκαν από 6 έως 12 μοντέλα, τα οποία χτίστηκαν τόσο με βάση κάποια τεχνική μείωσης της διαστατικότητας, όπως επιλογή των k καλύτερων χαρακτηριστικών σύμφωνα με τη μετρική Pearson και ανάλυση σε κύριες συνιστώσες (επιλογές 2 και 3, αντίστοιχα), όσο και χρησιμοποιώντας όλα τα δοσμένα χαρακτηριστικά (επιλογή 1). Καθένα από αυτά συνδυάστηκε με μία από τις επιλογές α-δ για το είδος των χαρακτηριστικών και τα χρονικά παράθυρα στα οποία αυτά ορίζονται.

Επίσης, για όλα τα παραγόμενα μοντέλα πέρα από την υπερπαράμετρο που καθορίζει τον πυρήνα, χρησιμοποιούνται και οι υπερπαράμετροι epsilon και C. Αυτές, όπως μας προδιαθέτει το όνομά τους, αντιστοιχούν στις μεταβλητές ϵ και C, που είδαμε στις θεωρητικές εξισώσεις ενός μοντέλου παλινδρόμησης διανυσμάτων υποστήριξης. Η πρώτη ορίζει το εύρος του περιθωρίου ανοχής, εντός του οποίου δεν εφαρμόζεται κάποια τιμωρία στις παρατηρήσεις. Η δεύτερη αποτελεί παράμετρο κανονικοποίησης και πιο συγκεκριμένα είναι αντιστρόφως ανάλογη με την ισχύ της L_2 κανονικοποίησης - τιμωρίας που εφαρμόζεται στις παρατηρήσεις εκτός του παραπάνω εύρους. Οι τιμές που δοκιμάστηκαν για την υπερπαράμετρο C ανήκουν στο εύρος $[0.001, 1000]$ και για την υπερπαράμετρο epsilon στο εύρος $[0.0001, 10]$, με πολλαπλασιαστικό βήμα 10 και στις δύο περιπτώσεις.

6.12.1 Με γραμμικό πυρήνα

Όταν χρησιμοποιείται γραμμικός πυρήνας, δε χρειάζεται να ορίσουμε άλλες υπερπαραμέτρους για τα μοντέλα. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.23 και 6.24 και στο αντίστοιχο σχήμα 6.12.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	5.67%	0.54	0.61	0.54	0.54	94.33%
	β	5.67%	0.55	0.61	0.54	0.55	94.33%
	γ	11.56%	1.12	1.83	-0.38	-0.32	88.44%
	δ	11.55%	1.10	1.57	-0.19	-0.05	88.84%
2	**α**	5.67%	0.54	0.61	0.54	0.54	94.33%
	β	5.37%	0.51	0.53	0.60	0.61	94.63%
	γ	11.44%	1.11	1.85	-0.39	-0.36	88.56%
	δ	10.88%	1.04	1.48	-0.12	-0.04	89.12%
3	**α**	5.41%	0.51	0.54	0.59	0.60	94.59%
	β	**5.08%**	**0.46**	**0.46**	**0.65**	**0.68**	**94.92%**
	γ	10.32%	1.00	1.42	-0.07	-0.03	89.68%
	δ	8.64%	0.85	1.07	0.20	0.20	91.36%

Πίνακας 6.23: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα για την εκτίμηση του SHR%

Όπως βλέπουμε, στην περίπτωση της εκτίμησης του μεγέθους SHR%, τα μοντέλα που εκπαιδεύτηκαν μόνο με χαρακτηριστικά ποσοτικών δεικτών οδήγησαν σε πολύ καλύτερα αποτελέσματα από τα υπόλοιπα, τα οποία παρουσίασαν αποτελέσματα συγκρίσιμα με εκείνα του μοντέλου αναφοράς. Ως μοντέλο με την καλύτερη απόδοση στο σύνολο των δεδομένων ελέγχου αποδείχθηκε εκείνο που χρησιμοποίησε ανάλυση σε κύριες συνιστώσες και χαρακτηριστικά που σχετίζονται

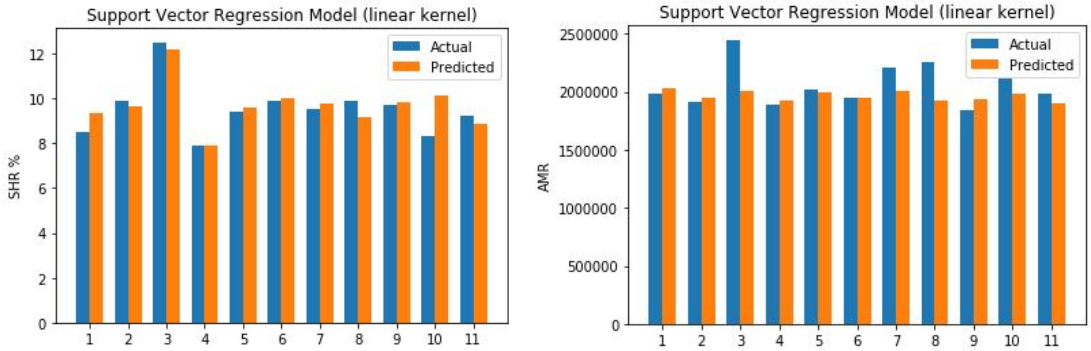
[11] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

μόνο με ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά, κρατήθηκαν 7 κύριες συνιστώσες στις οποίες οφείλεται περίπου το 99.97% της διασποράς των δειγμάτων και για τις υπερπαραμέτρους epsilon και C επιλέχθηκαν οι τιμές 0.001 και 10, αντίστοιχα, ως καταλληλότερες. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.08%, MAE = 0.46, MSE = 0.46, R^2 = 0.65, explained_variance = 0.68 και accuracy = 94.92%.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	**5.86%**	**129002**	34326951654	-0.13	0.14	**94.14%**
	β	5.87%	129331	34734221279	-0.14	0.14	94.13%
	γ	6.75%	149672	51001419375	-0.67	-0.21	93.25%
	δ	6.62%	146862	48483175867	-0.59	-0.18	93.37%
2	**α**	5.93%	130490	**34144718633**	**-0.12**	**0.15**	94.06%
	β	5.94%	130605	34303106164	-0.12	0.14	94.06%
	γ	6.20%	137069	40655770612	-0.33	-0.02	93.80%
	δ	5.94%	132301	40721131727	-0.33	0.04	94.06%
3	**α**	6.20%	136953	40553904045	-0.33	-0.01	93.80%
	β	5.99%	132425	38191649953	-0.25	0.07	94.01%
	γ	6.28%	139151	42786953978	-0.40	-0.06	93.72%
	δ	6.28%	139151	42786953978	-0.40	-0.06	93.72%

Πίνακας 6.24: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα για την εκτίμηση του AMR

Όσον αφορά την εκτίμηση του μεγέθους AMR, τα μοντέλα που εκπαιδεύτηκαν μόνο με χαρακτηριστικά ποσοτικών δεικτών παρουσίασαν ελαφρώς καλύτερα αποτελέσματα από τα υπόλοιπα και το μοντέλο αναφοράς. Ως μοντέλο με την καλύτερη απόδοση στο σύνολο των δεδομένων ελέγχου αναδείχθηκε εκείνο που χρησιμοποίησε όλα τα χαρακτηριστικά που σχετίζονται μόνο με ποσοτικούς δείκτες, ορισμένα στα χρονικά παράθυρα 6 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter. Από τη διασταυρούμενη επικύρωση για τις υπερπαραμέτρους epsilon και C επιλέχθηκαν οι τιμές 0.0001 και 1000, αντίστοιχα, ως καταλληλότερες. Τελικά, επιτεύχθηκαν τα εξής αποτελέσματα: MAPE = 5.86%, MAE = 129002, MSE = 34326951654, $R^2 = -0.13$, explained_variance = 0.14 και accuracy = 94.14%. Αξίζει να σημειωθεί ότι



(α′) για το SHR% (β′) για το AMR

Σχήμα 6.12: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με γραμμικό πυρήνα

παρόμοια αποτελέσματα παρουσιάστηκαν και στην περίπτωση που χρησιμοποιήθηκε επιλογή χαρακτηριστικών που σχετίζονται με ποσοτικούς δείκτες, ορισμένα στα χρονικά παράθυρα 1 και

4, αντίστοιχα. Πάλι, επιλέχθηκαν οι τιμές 0.0001 και 1000 για τις υπερπαραμέτρους epsilon και C ενώ επιλέχθηκαν τα 6 καλύτερα χαρακτηριστικά, σύμφωνα με το συντελεστή συσχέτισης του Pearson. Τελικά, τα αποτελέσματα ήταν τα εξής: MAPE = 5.93%, MAE = 130490, MSE = 34144718633, $R^2 = -0.12$, explained_variance = 0.15 και accuracy = 94.06%.

6.12.2 Με πυρήνα ακτινικής συνάρτησης βάσης

Όταν χρησιμοποιείται πυρήνας ακτινικής συνάρτησης βάσης (rbf), πέρα από τις υπερπαραμέτρους epsilon και C, ορίζεται και η υπερπαράμετρος gamma. Για τις δύο πρώτες δοκιμάστηκαν οι ίδιες τιμές με προηγουμένως, με μόνη διαφορά ότι οι τιμές της C ξεκινούν από το 0.01 σε αυτήν την περίπτωση. Η υπερπαράμετρος gamma, που προστέθηκε τώρα, είναι απλά ένας συντελεστής στη συνάρτηση πυρήνα, όπως μπορούμε να δούμε και από τη θεωρητική εξίσωση 3.28. Το εύρος τιμών της κυμαίνεται από 0.00001 έως και 1.0 με πολλαπλασιαστικό βήμα 10 και επιπλέον περιλαμβάνει τις επιλογές 'scale' και 'auto'. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.25 και 6.26 και στο αντίστοιχο σχήμα 6.13.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	6.08%	0.59	0.74	0.45	0.45	93.92%
	β	**5.99%**	**0.57**	**0.58**	**0.56**	**0.58**	**94.01%**
	γ	8.65%	0.86	1.21	0.09	0.09	91.35%
	δ	8.78%	0.86	1.10	0.18	0.19	91.22%
2	**α**	7.02%	0.66	0.76	0.43	0.43	92.98%
	β	5.99%	0.57	0.58	0.56	0.58	94.01%
	γ	10.29%	0.98	1.40	-0.06	0.03	89.70%
	δ	19.74%	1.83	8.19	-5.16	-4.13	80.26%
3	**α**	6.05%	0.59	0.74	0.44	0.44	93.95%
	β	6.88%	0.65	0.69	0.48	0.54	93.12%
	γ	10.59%	1.02	1.51	-0.14	-0.08	89.41%
	δ	11.57%	1.07	1.58	-0.20	0.06	88.42%

Πίνακας 6.25: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα για την εκτίμηση του SHR%

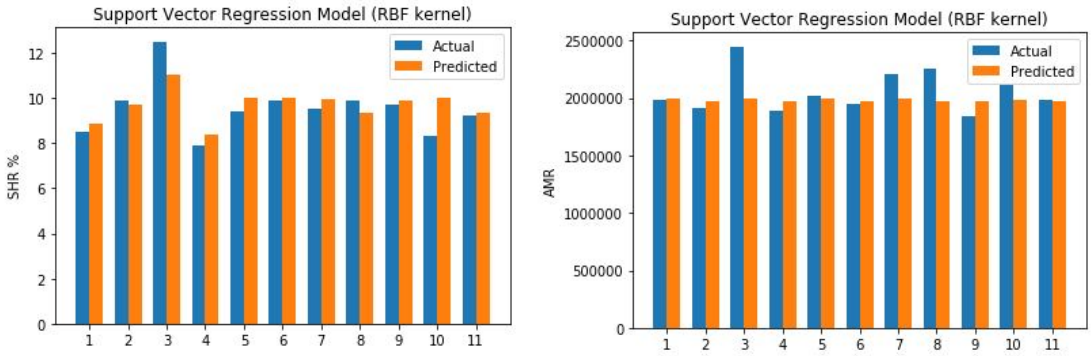
Όπως βλέπουμε, στην περίπτωση της εκτίμησης του μεγέθους SHR%, πάλι τα μοντέλα που εκπαιδεύτηκαν μόνο με χαρακτηριστικά ποσοτικών δεικτών οδήγησαν σε πολύ καλύτερα αποτελέσματα από τα υπόλοιπα. Ως μοντέλο με την καλύτερη απόδοση στο σύνολο των δεδομένων ελέγχου προέκυψε εκείνο που χρησιμοποίησε όλα τα χαρακτηριστικά που σχετίζονται μόνο με ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά επιλέχθηκαν οι τιμές 0.01, 1000 και 0.0001 για τις υπερπαραμέτρους epsilon, C και gamma, αντίστοιχα. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.99%, MAE = 0.57, MSE = 0.58, $R^2 = 0.56$, explained_variance = 0.58 και accuracy = 94.01%.

Όσον αφορά την εκτίμηση του μεγέθους AMR, είναι από τις λίγες φορές που τα μοντέλα, τα οποία εκπαιδεύτηκαν με βάση μόνο χαρακτηριστικά ποσοτικών δεικτών οδήγησαν σε εξίσου καλά αποτελέσματα με τα υπόλοιπα. Το καλύτερο μοντέλο είναι εκείνο που χρησιμοποίησε επιλεγμένα χαρακτηριστικά ποσοτικών δεικτών, ορισμένα στα καλύτερα χρονικά παράθυρα, σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά επιλέχθηκαν οι τιμές 0.0001, 1000 και 'scale' για τις υπερπαραμέτρους epsilon, C και gamma ενώ κρατήθηκαν τα 3 μόνο καλύτερα χαρακτηριστικά, σύμφωνα με το συντελεστή συσχέτισης του Pearson. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.95%, MAE = 130219, MSE = 34567593582, $R^2 =$

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	6.05%	132461	35851657126	-0.17	0.02	93.95%
	β	6.04%	132236	35735686293	-0.17	0.02	93.96%
	γ	6.10%	133426	36111023465	-0.18	0.01	93.90%
	δ	6.09%	133273	36016649456	-0.18	0.01	93.91%
2	**α**	6.00%	131171	34637183680	-0.13	0.04	94.00%
	β	**5.95%**	**130219**	**34567593582**	**-0.13**	**0.04**	**94.05%**
	γ	6.03%	131706	34853884536	-0.14	0.04	93.97%
	δ	6.01%	131291	34756027866	-0.14	0.04	93.99%
3	**α**	6.06%	132462	35533197879	-0.16	0.02	93.94%
	β	6.05%	132391	35358590082	-0.16	0.03	93.95%
	γ	6.00%	131394	35467987404	-0.16	0.03	94.00%
	δ	6.00%	131394	35467987404	-0.16	0.03	94.00%

Πίνακας 6.26: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα για την εκτίμηση του AMR

-0.13, explained_variance $= 0.04$ και accuracy $= 94.05\%$.



(α′) για το SHR%

(β′) για το AMR

Σχήμα 6.13: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα

6.12.3 Με πολυωνυμικό πυρήνα

Όταν χρησιμοποιείται πολυωνυμικός πυρήνας, πέρα από τις υπερπαραμέτρους epsilon, C και gamma ορίζονται και οι υπερπαράμετροι degree και coef0. Όπως μας προδιαθέτουν και τα ονόματα των νέων υπερπαραμέτρων, η πρώτη αντιστοιχεί στο βαθμό του πολυωνύμου, που θα χρησιμοποιηθεί ως συνάρτηση βάσης και η δεύτερη στο σταθερό - ανεξάρτητο όρο αυτού. Για τις τρεις πρώτες υπερπαραμέτρους δοκιμάστηκαν ακριβώς οι ίδιες τιμές με προηγουμένως. Για το βαθμό του πολυωνύμου δοκιμάστηκαν οι τιμές 2 και 3 ενώ για το σταθερό όρο το εύρος των τιμών που δοκιμάστηκαν είναι από 0.001 έως και 100 με πολλαπλασιαστικό βήμα 10 και επιπλέον η τιμή 0 που ισοδυναμεί με απουσία σταθερού όρου. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.27 και 6.28 και στο αντίστοιχο σχήμα 6.14.

Όπως γίνεται φανερό, στην περίπτωση της εκτίμησης του μεγέθους SHR%, για ακόμη μία φορά τα μοντέλα που εκπαιδεύτηκαν μόνο με χαρακτηριστικά ποσοτικών δεικτών οδήγησαν σε καλύτερα αποτελέσματα από τα υπόλοιπα. Ως μοντέλο με την καλύτερη απόδοση στο σύνολο των δεδομένων ελέγχου προέκυψε εκείνο που χρησιμοποίησε όλα τα χαρακτηριστικά που σχετίζονται μόνο με ποσοτικούς δείκτες, ορισμένα στα καλύτερα χρονικά παράθυρα σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από

		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	6.71%	0.63	0.65	0.51	0.51	93.29%
	β	**5.05%**	**0.47**	**0.44**	**0.67**	0.68	**94.95%**
	γ	12.21%	1.18	2.20	-0.66	-0.64	87.78%
	δ	8.57%	0.84	1.05	0.21	0.22	91.43%
2	**α**	6.20%	0.56	0.59	0.57	**0.70**	93.80%
	β	5.05%	0.47	0.44	0.67	0.68	94.95%
	γ	11.06%	1.06	1.57	-0.19	-0.11	88.94%
	δ	8.57%	0.84	1.05	0.21	0.22	91.43%
3	**α**	6.98%	0.65	0.65	0.51	0.51	93.02%
	β	5.20%	0.49	0.45	0.66	0.67	94.80%

Πίνακας 6.27: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα για την εκτίμηση του SHR%

το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά προτιμήθηκε ο συνδυασμός τιμών 0.1, 1, 0.0001, 3 και 100 για τις υπερπαραμέτρους epsilon, C, gamma, degree και coef0, αντίστοιχα. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.05%, MAE = 0.47, MSE = 0.44, $R^2 = 0.67$, explained_variance = 0.68 και accuracy = 94.95%.

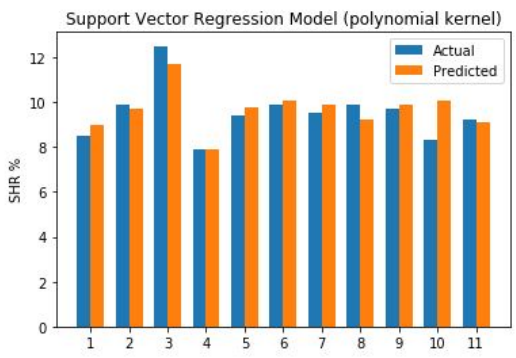
		MAPE	MAE	MSE	R^2	Expl_variance	Accuracy
1	**α**	6.97%	146434	34642948883	-0.14	-0.04	93.03%
	β	6.53%	142156	39809225466	-0.30	0.06	93.47%
	γ	9.01%	193139	66730486054	-1.18	-0.59	90.99%
	δ	7.15%	151037	37522472927	-0.23	0.01	92.85%
2	**α**	6.97%	146434	34642948883	-0.14	-0.04	93.03%
	β	**4.17%**	**87349**	**16834142301**	**0.45**	**0.52**	**95.83%**
	γ	6.04%	125407	26720604895	0.12	0.19	93.96%
	δ	7.26%	155471	43233752350	-0.42	0.07	92.74%
3	**α**	6.92%	145502	34722778724	-0.14	-0.03	93.08%
	β	6.56%	144547	45213111683	-0.48	-0.07	93.44%

Πίνακας 6.28: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα για την εκτίμηση του AMR

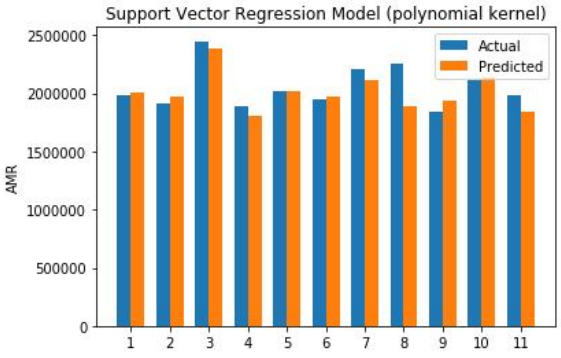
Όσον αφορά την εκτίμηση του μεγέθους AMR, τις περισσότερες φορές τα μοντέλα, τα οποία εκπαιδεύτηκαν με βάση μόνο χαρακτηριστικά ποσοτικών δεικτών οδήγησαν σε καλύτερα αποτελέσματα από τα υπόλοιπα. Τελικά, το καλύτερο μοντέλο είναι εκείνο που χρησιμοποίησε επιλεγμένα χαρακτηριστικά ποσοτικών δεικτών, ορισμένα στα καλύτερα χρονικά παράθυρα, σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμενα χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά επιλέχθηκαν οι τιμές 10, 100, 0.1, 3 και 100 για τις υπερπαραμέτρους epsilon, C, gamma, degree και coef0, αντίστοιχα, ενώ κρατήθηκαν τα 5 καλύτερα χαρακτηριστικά, σύμφωνα με το συντελεστή συσχέτισης του Pearson. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 4.17%, MAE = 87349, MSE = 16834142301, $R^2 = 0.45$, explained_variance = 0.52 και accuracy = 95.83%.

6.12.4 Με σιγμοειδή πυρήνα

Όταν χρησιμοποιείται σιγμοειδής πυρήνας, χρειάζονται οι ίδιες υπερπαράμετροι με την προηγούμενη περίπτωση του πολυωνυμικού πυρήνα, εκτός από εκείνη που αντιστοιχεί στο βαθμό του πολυωνύμου (degree). Για όλες τις υπερπαραμέτρους δοκιμάζονται οι ίδιες ακριβώς τιμές με προηγουμένως. Τα ληφθέντα αποτελέσματα παρουσιάζονται στους πίνακες 6.29 και 6.30 και στο



(α′) για το SHR%



(β′) για το AMR

Σχήμα 6.14: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με πολυωνυμικό πυρήνα

αντίστοιχο σχήμα 6.15.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	5.49%	0.54	0.60	0.55	0.55	94.51%
	β	5.37%	0.51	0.52	0.61	0.62	94.63%
2	**α**	5.49%	0.54	0.60	0.55	0.55	94.51%
	β	5.37%	0.51	0.52	0.61	0.62	94.63%
3	**α**	**4.81%**	**0.46**	**0.47**	**0.65**	**0.65**	**95.19%**
	β	5.35%	0.51	0.52	0.61	0.62	94.65%

Πίνακας 6.29: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα για την εκτίμηση του SHR%

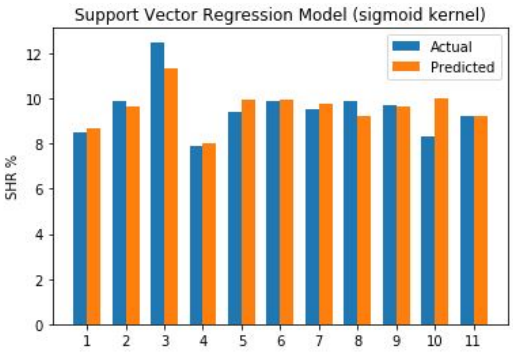
Όπως μπορούμε να δούμε, στην περίπτωση της εκτίμησης του μεγέθους SHR% το μοντέλο που εμφάνισε την καλύτερη απόδοση στο σύνολο των δεδομένων ελέγχου είναι εκείνο που χρησιμοποίησε χαρακτηριστικά ποσοτικών δεικτών έπειτα από την ανάλυσή τους σε κύριες συνιστώσες, τα οποία ήταν ορισμένα στα χρονικά παράθυρα 3 για το GTrends και 6 για τα εξαγόμενα χαρακτηριστικά από το Twitter. Από τη διασταυρούμενη επικύρωση, τελικά προτιμήθηκε ο συνδυασμός τιμών 0.001, 1000, 0.001 και 0.1 για τις υπερπαραμέτρους epsilon, C, gamma και coef0, αντίστοιχα ενώ κρατήθηκαν 7 κύριες συνιστώσες, στις οποίες αποδίδεται το 99.96% της διασποράς των παρατηρήσεων. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 4.81%, MAE = 0.46, MSE = 0.47, $R^2 = 0.65$, explained_variance = 0.65 και accuracy = 95.19%.

		MAPE	**MAE**	**MSE**	**R^2**	**Expl_variance**	**Accuracy**
1	**α**	5.73%	125981	34360643268	-0.13	0.07	94.27%
	β	5.59%	122580	31671425915	-0.04	**0.14**	94.41%
2	**α**	5.65%	124157	33369356392	-0.09	0.09	94.35%
	β	**5.56%**	**122130**	**31583461044**	**-0.04**	0.13	**94.44%**
3	**α**	5.73%	125981	34360643268	-0.13	0.07	94.27%
	β	5.64%	123876	32691178525	-0.07	0.11	94.36%

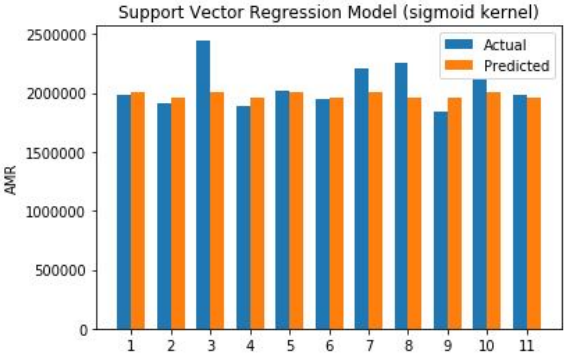
Πίνακας 6.30: Αξιολόγηση των μοντέλων παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα για την εκτίμηση του AMR

Όσον αφορά την εκτίμηση του μεγέθους AMR, το καλύτερο μοντέλο είναι εκείνο που χρησιμοποίησε επιλεγμένα χαρακτηριστικά ποσοτικών δεικτών, ορισμένα στα καλύτερα χρονικά παράθυρα, σύμφωνα με τη διερευνητική ανάλυση δεδομένων (3 για το GTrends και 4 για τα εξαγόμε-

να χαρακτηριστικά από το Twitter). Από τη διασταυρούμενη επικύρωση, τελικά επιλέχθηκαν οι τιμές 10, 1000, 1 και 0 για τις υπερπαραμέτρους epsilon, C, gamma, degree και coef0, αντίστοιχα, ενώ κρατήθηκαν τα 8 καλύτερα χαρακτηριστικά, σύμφωνα με το συντελεστή συσχέτισης του Pearson. Τα αποτελέσματα που επιτεύχθηκαν είναι τα εξής: MAPE = 5.56%, MAE = 122130, MSE = 31583461044, $R^2 = -0.04$, explained_variance = 0.13 και accuracy = 94.44%.



(α′) για το SHR%



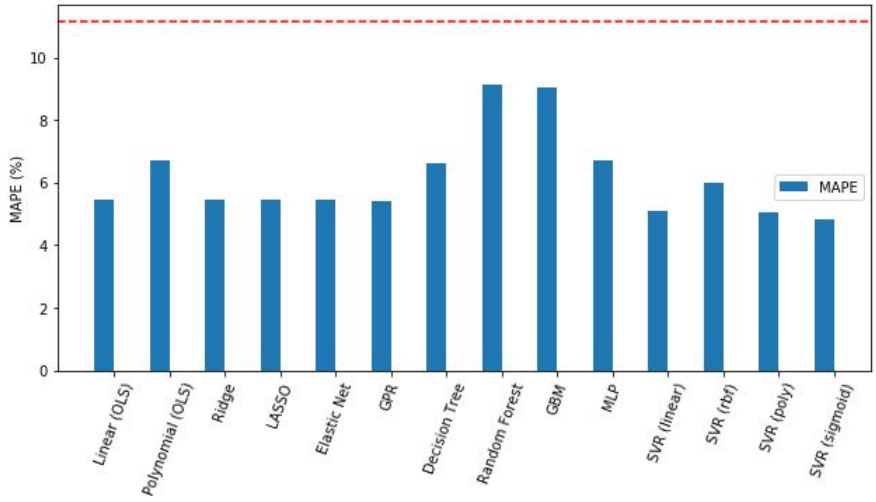
(β′) για το AMR

Σχήμα 6.15: Πραγματικές και προβλεπόμενες τιμές για τα καλύτερα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα

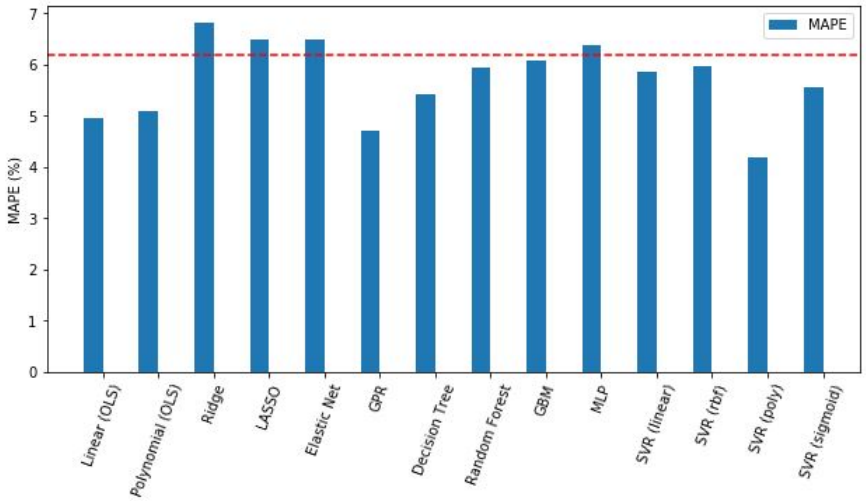
6.13 Σύγκριση μεταξύ διαφορετικών μοντέλων

Σε αυτό το σημείο έχουμε ολοκληρώσει τα πειράματά μας και θα προβούμε στη σύγκριση των αποτελεσμάτων που παράχθηκαν. Μία από τις σημαντικότερες μετρικές που χρησιμοποιήθηκε στην αξιολόγηση και θα ληφθεί υπ' όψη σε αυτήν την ενότητα είναι το μέσο απόλυτο ποσοστιαίο σφάλμα MAPE. Από κάθε μέθοδο παλινδρόμησης που υλοποιήθηκε παραπάνω, κρατάμε μόνο το μοντέλο που παρουσίασε το μικρότερο MAPE στη φάση ελέγχου (αυτό που επισημαίνεται με έντονη γραμματοσειρά στους παραπάνω πίνακες). Για τα δύο προβλήματα παλινδρόμησης που αντιμετωπίσαμε, την εκτίμηση του SHR% και την εκτίμηση του AMR, παραθέτουμε δύο γραφήματα (6.16 και 6.17), ένα για το καθένα, στα οποία φαίνεται το MAPE που πέτυχε το καλύτερο μοντέλο από κάθε κατηγορία παλινδρόμησης που εξετάστηκε. Επίσης, σε αυτά έχει σημειωθεί με κόκκινη διακεκομμένη οριζόντια γραμμή, η τιμή του MAPE που πέτυχε το αντίστοιχο μοντέλο αναφοράς.

Αρχικά, παρατηρούμε ότι η εκτίμηση του ποσοστού τηλεθέασης επί των ανοικτών δεκτών αποδείχθηκε ευκολότερο πρόβλημα από εκείνο της εκτίμησης του απόλυτου πλήθους των τηλεθεατών που συγκέντρωσε ένα επεισόδιο της υπό εξέτασης εκπομπής. Αν και το μοντέλο αναφοράς που κατασκευάστηκε για την εκτίμηση του AMR οδηγούσε ήδη σε μικρή τιμή MAPE, αυτή ήταν πολύ δύσκολο να βελτιωθεί στη συνέχεια. Αντιθέτως, όπως βλέπουμε από το σχήμα 6.16, όλα τα μοντέλα που κατασκευάσαμε για την εκτίμηση του SHR% οδήγησαν σε σημαντικά μικρότερο MAPE. Αυτή η διαφορά μπορεί να αποδοθεί στη διαφορετική διακύμανση των τιμών των δύο εξαρτημένων μεταβλητών στο σύνολο δεδομένων ελέγχου. Όπως μπορούμε να δούμε από το δεύτερο διάγραμμα του σχήματος 6.1, με εξαίρεση τρεις παρατηρήσεις που αποκλίνουν περισσότερο από τη μέση τιμή του συνόλου δεδομένων εκπαίδευσης, όλες οι υπόλοιπες έχουν πολύ μικρές διαφορές μεταξύ τους αλλά και με τη σταθερή προβλεπόμενη τιμή. Αυτό έχει ως συνέπεια αφενός να λαμβάνουμε πολύ μικρά MAPE, MAE και MSE ήδη από το μοντέλο αναφοράς, χωρίς να έχει γίνει πετυχημένη προσαρμογή του μοντέλου στα δεδομένα, γεγονός που φαίνεται από την αρνητική τιμή του R^2 και τη μηδενική του explained_variance και αφετέρου να καθίσταται πολύ δύσκολη η μοντελοποίηση τόσο παρόμοιων δεδομένων. Για αυτόν το λόγο, και τα υπόλοιπα μοντέλα που κατασκευάσαμε για την εκτίμηση του AMR εξακολούθησαν να εμφανίζουν μικρές



Σχήμα 6.16: Σύγκριση διαφορετικών μοντέλων για την εκτίμηση του SHR% ως προς το MAPE



Σχήμα 6.17: Σύγκριση διαφορετικών μοντέλων για την εκτίμηση του AMR ως προς το MAPE

τιμές σφάλματος στην αξιολόγηση ακόμα και όταν οι μετρικές R^2 και explained_variance είχαν πολύ μικρές τιμές, ακόμα και αρνητικές κάποιες φορές. Αντιθέτως, στα μοντέλα που κατασκευάσαμε για την εκτίμηση του SHR% η σημαντική μείωση στα σφάλματα συνοδευόταν πάντα από αντίστοιχα σημαντική αύξηση στις μετρικές που εκφράζουν το πόσο καλά ταιριάζει το εκάστοτε μοντέλο στα δεδομένα.

Μάλιστα στην περίπτωση του SHR% μπορούμε να δούμε ότι ήδη πολύ απλά μοντέλα, όπως αυτό της πολλαπλής γραμμικής παλινδρόμησης που βασίστηκε στη μέθοδο των ελάχιστων τετραγώνων εμφανίζουν πάρα πολύ ικανοποιητική απόδοση, σημειώνοντας MAPE 5.45%. Εξίσου καλή απόδοση εμφανίζουν τα μοντέλα κανονικοποίησης ridge, LASSO και elastic net, γεγονός αναμενόμενο καθώς η λύση τους πλησιάζει πολύ αυτή των ελάχιστων τετραγώνων δεδομένου ότι η ισχύς της κανονικοποίησης που εφαρμόζεται και στις τρεις περιπτώσεις είναι πολύ μικρή λόγω της μικρής τιμής που επιλέγεται από τη διασταυρούμενη επικύρωση για την αντίστοιχη υπερπαράμετρο. Ελαφρώς καλύτερα αποτελέσματα (MAPE = 5.39%) επιτυγχάνονται και με το μοντέλο παλινδρόμησης με γκαουσιανή διεργασία, του οποίου οι προβλέψεις συνοδεύονται από ένα μικρό διάστημα αβεβαιότητας. Ακόμα καλύτερη απόδοση παρουσιάζουν τα μοντέλα παλινδρόμησης που κατασκευάστηκαν με διανύσματα υποστήριξης και με πυρήνα γραμμικό, πολυωνυμικό

ή σιγμοειδή. Μάλιστα, το μοντέλο παλινδρόμησης διανυσμάτων υποστήριξης με σιγμοειδή πυρήνα και υπερπαραμέτρους epsilon = 0.001, $C = 1000$, gamma = 0.001 και $coef0 = 0.1$, που εκπαιδεύτηκε με 7 κύριες συνιστώσες, οι οποίες προέκυψαν από χαρακτηριστικά ποσοτικών δεικτών, εμφάνισε MAPE μικρότερο του 5% (4.81%). Από την άλλη, το μοντέλο παλινδρόμησης διανυσμάτων υποστήριξης με rbf πυρήνα εμφάνισε ελαφρώς χειρότερη απόδοση, με MAPE σχεδόν 6% (5.99%). Ακολούθησαν τα μοντέλα πολυωνυμικής παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων, παλινδρόμησης με πολυεπίπεδο perceptron και με δέντρο απόφασης, παρουσιάζοντας MAPE 6-7%. Τέλος, τα μοντέλα συνδυαστικής μάθησης (τυχαίο δάσος και gradient boosting μηχανή) παραδόξως παρουσίασαν αρκετά χειρότερη απόδοση από εκείνο του μεμονωμένου εκτιμητή - δέντρου απόφασης. Αυτό μπορεί να οφείλεται στο ότι καταφέραμε να βρούμε ένα δέντρο απόφασης που να ταιριάζει πολύ καλά στο πρόβλημά μας ωστόσο πολύ μικρές αλλαγές στα δέντρα απόφασης μπορούν να οδηγήσουν σε εξαιρετικά διαφορετικά αποτελέσματα. Συνεπώς, ελαφρώς παραλλαγμένα δέντρα απόφασης μπορεί να είχαν αρκετά χειρότερη απόδοση και ο συνδυασμός τους τελικά να οδήγησε σε εξίσου χειρότερα αποτελέσματα. Και πάλι όμως, με αυτά τα μοντέλα επιτυγχάνεται MAPE περίπου 9%, διατηρώντας μια μεγάλη απόσταση ασφαλείας από το αποτέλεσμα του αντίστοιχου μοντέλου αναφοράς (MAPE = 11.15%). Συνεπώς, σε όλες τις περιπτώσεις επιτεύχθηκε ο σκοπός της μάθησης και της εξαγωγής γνώσης από τα μέσα κοινωνικής δικτύωσης και τις μηχανές αναζήτησης στο διαδίκτυο.

Όσον αφορά το πρόβλημα της εκτίμησης του AMR τα πράγματα ήταν πιο δύσκολα, όπως αναφέρθηκε και παραπάνω. Είναι αξιοσημείωτο πως πάλι τα απλούστερα μοντέλα που βασίστηκαν στη μέθοδο των ελάχιστων τετραγώνων, γραμμικό ή πολυωνυμικό, παρουσίασαν πολύ ικανοποιητικά αποτελέσματα, με MAPE περίπου 5%, διατηρώντας μια σημαντική απόσταση από τα αποτελέσματα του μοντέλου αναφοράς. Επίσης, πάλι ένα μοντέλο παλινδρόμησης διανυσμάτων υποστήριξης αποδείχθηκε το καλύτερο, σημειώνοντας το χαμηλότερο MAPE (4.17%). Αυτή τη φορά εκπαιδεύεται πάνω στα 5 καλύτερα σύμφωνα με το συντελεστή συσχέτισης Pearson χαρακτηριστικά, τα οποία σχετίζονται με ποσοτικούς δείκτες και χρησιμοποιεί πολυωνυμικό πυρήνα ενώ οι υπόλοιπες υπερπαράμετροι του λαμβάνουν τις εξής τιμές: epsilon = 10, $C = 100$, gamma = 0.1, degree = 3 και coef0 = 100. Τα υπόλοιπα μοντέλα παλινδρόμησης διανυσμάτων υποστήριξης, με γραμμικό, rbf και σιγμοειδή πυρήνα πετυχαίνουν ελαφρώς καλύτερη απόδοση από το μοντέλο αναφοράς με MAPE 5.5-6%. Στο ίδιο εύρος τιμών κυμαίνονται και τα μοντέλα που χρησιμοποιούν δέντρα απόφασης (ο μεμονωμένος εκτιμητής - δέντρο απόφασης, το τυχαίο δάσος και η gradient boosting μηχανή). Επίσης, το μοντέλο παλινδρόμησης γκαουσιανής διεργασίας οδήγησε σε πολύ ικανοποιητικά αποτελέσματα (MAPE = 4.70%) και οι προβλέψεις του συνοδεύονταν από ένα πολύ μικρό διάστημα αβεβαιότητας. Τέλος, το μοντέλο παλινδρόμησης με πολυεπίπεδο perceptron αλλά και τα μοντέλα παλινδρόμησης με τεχνικές κανονικοποίησης (ridge, LASSO και elastic net) παρουσίασαν ελαφρώς υψηλότερο MAPE από το μοντέλο αναφοράς, πάλι μικρό αντικειμενικά βέβαια, περίπου 6.3-6.8%.

6.14 Σύγκριση αποτελεσμάτων με προηγούμενες εργασίες

Όπως έχουμε αναφέρει και προηγουμένως, η παρούσα διπλωματική εργασία βασίστηκε σε προϋπάρχουσες διπλωματικές εργασίες, οι οποίες αντιμετώπιζαν το ίδιο πρόβλημα. Καταρχάς, η διπλωματική εργασία του Μάριου Παρασκευόπουλου [61] συνδύασε δεδομένα από τις πλατφόρμες Google Trends και Twitter, όπως ακριβώς πράξαμε και εμείς, αλλά βασίστηκε σε ένα άλλο πολύ μικρότερο σύνολο δεδομένων, το οποίο δεν επέτρεπε τον έλεγχο της καταλληλότητας των μοντέλων σε ένα ξεχωριστό υποσύνολο δεδομένων παρά μόνο με τη μέθοδο της διασταυρούμενης επικύρωσης. Εφόσον, λοιπόν, πρόκειται για τελείως διαφορετικό σύνολο δεδομένων, δεν μπορούμε να προβούμε σε σύγκριση των αποτελεσμάτων μας. Ωστόσο αξίζει να σημειώσουμε πως τα αποτελέσματα που προέκυψαν στη μελέτη του ήταν θετικά με MAPE 6.1%-9.6% σε όλα τα μοντέλα για την εκτίμηση και των δύο μεγεθών. Τα δικά μας αποτελέσματα είναι εξίσου καλά με MAPE 4.17%-9.12%, ενισχύοντας την αρχική υπόθεση ότι μπορεί να εξαχθεί ιδιαίτερα χρήσιμη

πληροφορία από τις πλατφόρμες Google Trends και Twitter για την εκτίμηση της τηλεθέασης εκπομπών.

Επίσης, στη διπλωματική εργασία του Χρήστου Πιερράκου [65] αντιμετωπίζεται το ίδιο πρόβλημα με δεδομένα που έχουν συλλεχθεί μόνο από το Twitter αλλά το σύνολο δεδομένων αναφοράς είναι το ίδιο με αυτό που χρησιμοποιήσαμε στην παρούσα εργασία για την ιταλική εκπομπή Le Iene και τα έτη 2016-2017. Συνεπώς, είμαστε σε θέση να προβούμε σε σύγκριση των αποτελεσμάτων οπότε και στον πίνακα που ακολουθεί παραθέτουμε την ακρίβεια που επιτεύχθηκε για όλες τις κοινές μεθόδους παλινδρόμησης που αναπτύχθηκαν.

Μοντέλα	**Προηγούμενη εργασία**		**Παρούσα εργασία**	
	SHR%	**AMR**	**SHR%**	**AMR**
Linear (OLS)	92.09%	90.86%	94.55%	95.04%
Polynomial (OLS)	93.83%	93.59%	93.30%	94.92%
Ridge	92.08%	-	94.55%	93.19%
LASSO	92.04%	-	94.55%	93.53%
Elastic Net	92.04%	-	94.55%	93.53%
Decision Tree	94.33%	93.13%	93.38%	94.60%
Random Forest	94.46%	92.03%	90.88%	94.08%

Πίνακας 6.31: Σύγκριση των μοντέλων που κατασκευάσαμε με τα προϋπάρχοντα όσον αφορά στη μετρική της ακρίβειας

Είναι ιδιαίτερα ενθαρρυντικό το γεγονός ότι στην πλειοψηφία των περιπτώσεων η προσθήκη του χαρακτηριστικού από το Google Trends σε συνδυασμό με τις τεχνικές προεπεξεργασίας που δοκιμάστηκαν οδήγησαν σε βελτίωση των αποτελεσμάτων. Εξαίρεση αποτέλεσαν τα μοντέλα παλινδρόμησης που βασίστηκαν σε δέντρα απόφασης, όπως αυτό του μεμονωμένου εκτιμητή - δέντρου απόφασης και του τυχαίου δάσους μόνο όσον αφορά στο πρόβλημα της εκτίμησης του ποσοστού τηλεθέασης SHR%. Επίσης, για πολύ λίγο υπερίσχυσε το προϋπάρχον μοντέλο πολυωνυμικής παλινδρόμησης που υλοποιήθηκε με βάση τη μέθοδο των ελάχιστων τετραγώνων έναντι του δικού μας, πάλι μόνο στην περίπτωση της εκτίμησης του SHR%. Σε όλες τις άλλες περιπτώσεις τα νέα μοντέλα αποδείχθηκαν καλύτερα από τα προηγούμενα.

Κεφάλαιο 7

Επίλογος

7.1 Σύνοψη και συμπεράσματα

Στις μέρες μας, δισεκατομμύρια άνθρωποι χρησιμοποιούν σε καθημερινή βάση μέσα κοινωνικής δικτύωσης για να κοινωνικοποιηθούν, να ψυχαγωγηθούν και να ενημερωθούν. Προς την ίδια κατεύθυνση, με την ευρεία ανάπτυξη των μηχανών αναζήτησης στο διαδίκτυο ολοένα και περισσότεροι άνθρωποι προστρέχουν σε αυτές για να ανακτήσουν οποιαδήποτε πληροφορία επιθυμούν. Τόσο τα μέσα κοινωνικής δικτύωσης όσο και οι μηχανές αναζήτησης στο διαδίκτυο πέρα από τις κύριες, άμεσες λειτουργίες και υπηρεσίες που παρέχουν στους χρήστες τους, συνιστούν πλούσια πηγή ενός τεράστιου όγκου δεδομένων. Ο επιστημονικός και ο επιχειρησιακός κλάδος τα τελευταία χρόνια έχουν δείξει μεγάλο ενδιαφέρον για αυτά τα δεδομένα, λόγω της προβλεπτικής ικανότητας που φαίνεται ότι διαθέτουν, καθώς προσφέρουν διαφωτιστικές πληροφορίες, χρήσιμες για την επίλυση προβλημάτων και την πρόβλεψη φαινομένων, που με τις παραδοσιακές μεθόδους θα ήταν πολύ δύσκολο έως και αδύνατο να επιτευχθούν.

Στην παρούσα διπλωματική εργασία, ασχοληθήκαμε με το πρόβλημα της εκτίμησης της τηλεθέασης εκπομπών με βάση δεδομένα από μέσα κοινωνικής δικτύωσης και μηχανές αναζήτησης στο διαδίκτυο. Πιο συγκεκριμένα, εξήγαμε δεδομένα από τις πλατφόρμες Twitter και Google Trends και με βάση αυτά προβήκαμε στην εκτίμηση τόσο του ποσοστού τηλεθέασης (SHR%) όσο και του απόλυτου πλήθους των τηλεθεατών (AMR) που συγκέντρωσε κάθε επεισόδιο της ιταλικής εκπομπής Le Iene. Το χρονικό εύρος μελέτης αποτελούταν από τα δύο ημερολογιακά έτη 2016-2017, όπου το εν λόγω τηλεοπτικό πρόγραμμα προβαλλόταν δύο φορές την εβδομάδα.

Όπως είδαμε, έχει ήδη μελετηθεί σε αρκετές εργασίες η προβλεπτική ικανότητα των δεδομένων που συλλέγονται από το Twitter για την εκτίμηση της τηλεθέασης εκπομπών, αλλά και από άλλα μέσα κοινωνικής δικτύωσης, όπως το Facebook. Επιπλέον, λίγες έρευνες έχουν επιστρατευτεί και στατιστικά δεδομένα που εξάγονται από διάφορες μηχανές αναζήτησης στο διαδίκτυο, προς την επίτευξη αυτού του σκοπού. Η καινοτομία που εισάγει η παρούσα διπλωματική εργασία έγκειται στο συνδυασμό των δεδομένων από τις δύο πλατφόρμες για την εξαγωγή διαφόρων ειδών χαρακτηριστικών και την αξιοποίηση τους χρησιμοποιώντας ποικίλες τεχνικές Μηχανικής Μάθησης, πάνω σε ένα ικανοποιητικά μεγάλο και στιβαρό σύνολο δεδομένων αναφοράς.

Πιο αναλυτικά, εξάγονται χαρακτηριστικά που σχετίζονται τόσο με ποσοτικούς δείκτες από τα δεδομένα στις δύο πλατφόρμες όσο και με ανάλυση συναισθήματος στο κειμενικό περιεχόμενο των δημοσιεύσεων στο Twitter, σε διάφορα χρονικά παράθυρα γύρω από την προβολή του εκάστοτε επεισοδίου. Αυτά στη συνέχεια, αφού περάσουν πιθανώς από διάφορες τεχνικές προεπεξεργασίας, όπως κανονικοποίηση, μείωση της διαστατικότητας κλπ., εισάγονται ως είσοδοι στους αλγορίθμους εκπαίδευσης για την κατασκευή των μοντέλων και τέλος στα ίδια τα μοντέλα για την τελική παραγωγή των εκτιμήσεων. Μελετήθηκαν διάφορες τεχνικές παλινδρόμησης, όπως γραμμική και πολυωνυμική με τη μέθοδο των ελάχιστων τετραγώνων, με τεχνικές κανονικοποίησης (π.χ. κορυφογραμμής, LASSO και elastic net), πιθανοτικές (π.χ. με γκαουσιανή διεργασία), με δέντρο απόφασης ως μεμονωμένο εκτιμητή αλλά και ως επιμέρους εκτιμητή σε τεχνικές συλ-

λογικής μάθησης (π.χ. τυχαίο δάσος και gradient boosting μηχανή), με νευρωνικά δίκτυα (π.χ. πολυεπίπεδο perceptron) και με μηχανές διανυσμάτων υποστήριξης.

Τα αποτελέσματα που λάβαμε ενισχύουν την υπόθεση ότι ο συνδυασμός δεδομένων από το Twitter και το Google Trends μπορεί να αποβεί ιδιαίτερα ωφέλιμος για την αντιμετώπιση του συγκεκριμένου προβλήματος, που δεν είναι άλλο από την εκτίμηση της τηλεθέασης εκπομπών. Επίσης, είναι αξιοσημείωτο πως στο σύνολο των τεχνικών παλινδρόμησης που μελετήθηκαν, σε καμία δεν προέκυψε ως καλύτερο μοντέλο κάποιο που εκπαιδεύτηκε χρησιμοποιώντας και χαρακτηριστικά που προέκυψαν από ανάλυση συναισθήματος μεταξύ άλλων. Αυτό έρχεται σε αντίθεση με το γεγονός ότι πολλά χαρακτηριστικά που συνδέονται με το συναίσθημα φαίνονταν να περιέχουν χρήσιμη πληροφορία, εμφανίζοντας υψηλές τιμές συσχέτισης με τις εξαρτημένες μεταβλητές, σύμφωνα με τη διερευνητική ανάλυση δεδομένων. Ωστόσο η ανάδειξη των χρονικών παραθύρων 3 για το χαρακτηριστικό από το Google Trends, που αντιστοιχεί στην επόμενη μέρα από την προβολή κάθε επεισοδίου και 4 για τα χαρακτηριστικά από το Twitter, που αντιστοιχεί στο τριήμερο από την προηγούμενη έως και την επόμενη μέρα από την προβολή κάθε επεισοδίου, ως καλύτερα από τη διερευνητική ανάλυση δεδομένων επιβεβαιώθηκε, καθώς στη συντριπτική πλειοψηφία των περιπτώσεων των καλύτερων μοντέλων χρησιμοποιήθηκαν αυτά για τα χαρακτηριστικά με τα οποία εκπαιδεύτηκαν.

Και στις δύο περιπτώσεις εξαρτημένων μεταβλητών (SHR% και AMR), ως καλύτερα μοντέλα από άποψη μικρότερου μέσου απόλυτου σφάλματος (MAPE), αναδείχθηκαν εκείνα που υλοποιήθηκαν με διανύσματα υποστήριξης, με σιγμοειδή πυρήνα στην πρώτη περίπτωση και με πολυωνυμικό στη δεύτερη. Όσον αφορά στην εκτίμηση του SHR% παρατηρήθηκαν πολύ ικανοποιητικά αποτελέσματα με MAPE $= 4.81\% - 9.12\%$, MAE $= 0.46 - 0.88$, MSE $= 0.44 - 1.36$, $R^2 = -0.03 - 0.67$, explained_variance $= 0.01 - 0.74$ και accuracy $= 90.88\% - 95.19\%$. Μάλιστα, αν εξαιρέσουμε τα δύο χειρότερα μοντέλα, που αντιστοιχούν στο τυχαίο δάσος και στην gradient boosting μηχανή, οι τιμές των μετρικών αξιολόγησης βελτιώνονται σε μεγάλο βαθμό ως εξής: MAPE $= 4.81\% - 6.73\%$, MAE $= 0.46 - 0.62$, MSE $= 0.44 - 0.66$, $R^2 = 0.51 - 0.67$, explained_variance $= 0.56 - 0.74$ και accuracy $= 93.27\% - 95.19\%$, οπότε και επιτυγχάνονται όλοι στόχοι που είχαμε θέσει εξ αρχής στην ενότητα 6.1.

Από την άλλη, τα μοντέλα που κατασκευάστηκαν για την εκτίμηση του AMR ενώ οδήγησαν σε μικρά σφάλματα δεν κατάφεραν να ταιριάξουν τόσο καλά στα δεδομένα εξηγώντας τις διάφορες διακυμάνσεις, καθώς οι διακυμάνσεις αυτές δεν ήταν τόσο έντονες. Πιο συγκεκριμένα, τα αποτελέσματα της αξιολόγησής τους ήταν τα εξής: MAPE $= 4.17\% - 6.81\%$, MAE $= 87349 - 140298$, MSE $= 16834142301 - 35918457874$, $R^2 = -0.18 - 0.45$, explained_variance $= -0.04 - 0.52$ και accuracy $= 93.19\% - 95.83\%$. Όπως γίνεται φανερό, η ποιότητα της μοντελοποίησης αρκετές φορές είναι αμφιλεγόμενη, καθώς παρουσιάζονται αρνητικές τιμές στα μεγέθη R^2 και explained_variance, οπότε και οι ικανοποιητικές τιμές στα σφάλματα και στην ακρίβεια αποδίδονται στην εγγενή χαμηλή διακύμανση των παρατηρήσεων. Παρόλα αυτά, κάποια μοντέλα κατάφεραν να «μάθουν» και απομονώνοντας μόνο αυτά επιτεύχθηκαν: MAPE $= 4.17\% - 5.40\%$, MAE $= 87349 - 114676$, MSE $= 16834142301 - 23232737622$, $R^2 = 0.24 - 0.45$, explained_variance $= 0.35 - 0.52$ και accuracy $= 94.60\% - 95.83\%$.

Επιπρόσθετα, για την αξιολόγηση των ληφθέντων αποτελεσμάτων, οφείλουμε να λάβουμε υπόψη μας ότι το ιταλικό τηλεοπτικό πρόγραμμα Le Iene συνιστά μια δύσκολη περίπτωση συνόλου δεδομένων αναφοράς λόγω δύο βασικών ιδιαιτεροτήτων. Αφενός η εκπομπή προβάλλεται δύο φορές την εβδομάδα, εκ των οποίων η μία είναι καθημερινή ενώ η άλλη Κυριακή, γεγονός που σίγουρα επηρεάζει την τηλεθέασή της. Αφετέρου, το είδος της και τα θέματα που πραγματεύεται είναι τέτοια ώστε να εισάγονται κάποιες δυσκολίες, ειδικά στην ανάλυση συναισθήματος, με αποτέλεσμα να μην αποτελεί ένα σύνηθες παράδειγμα εκπομπής για μελέτες σαν και τη δική μας. Πιο συγκεκριμένα, λόγω του σατιρικού χαρακτήρα της και της σκωπτικής διάθεσης πάνω σε φλέγοντα κοινωνικοπολιτικά ζητήματα, οι σχετικές αναρτήσεις των χρηστών στο Twitter εμπεριέχουν ειρωνικά σχόλια πολλές φορές, γεγονός που είναι δύσκολο να ανιχνευτεί με τεχνικές επεξεργασίας φυσικής γλώσσας και δυσχεραίνει τη διαδικασία ανάλυσης συναισθήματος στην

οποία προβήκαμε. Αυτό ίσως μπορεί να δικαιολογήσει τη συστηματική μη επιλογή των σχετικών χαρακτηριστικών στα καλύτερα μοντέλα.

Καταλήγοντας, είναι ιδιαίτερα σημαντικό πως παρά τις δυσκολίες που αναφέραμε πετύχαμε να μοντελοποιήσουμε σε πολύ ικανοποιητικό βαθμό τα δεδομένα, ειδικά στην περίπτωση του μεριδίου της τηλεθέασης, και να πραγματοποιήσουμε εκτιμήσεις για τελείως άγνωστα επεισόδια. Καθώς, όπως παρατηρήθηκε προηγουμένως τις περισσότερες φορές χρησιμοποιήθηκαν δεδομένα, ορισμένα σε χρονικά παράθυρα που περιλάμβαναν και μέρες έπειτα από τη την προβολή του εκάστοτε επεισοδίου, δεν πρόκειται για προβλέψεις με την έννοια της προγενέστερης χρονικά γνώσης. Ωστόσο η σπουδαιότητα των αποτελεσμάτων εξακολουθεί να είναι μεγάλη, αφού μπορούν να χρησιμοποιηθούν ως συμπλήρωμα στον παραδοσιακό τρόπο μετρήσεων της τηλεθέασης, πράγμα που καθίσταται όλο και πιο αναγκαίο με τις νέες δυνατότητες παρακολούθησης που προσφέρει η τεχνολογία. Συνεπώς, η παρούσα εργασία συντέλεσε ως ένα βαθμό στην ανάδειξη της χρησιμότητας τόσο των δεδομένων από το μέσο κοινωνικής δικτύωσης Twitter όσο και των στατιστικών δεδομένων που παρέχει η πλατφόρμα Google Trends για τη μηχανή αναζήτησης της Google για το πρόβλημα της εκτίμησης της τηλεθέασης προγραμμάτων.

7.2 Μελλοντικές επεκτάσεις

Είναι δυνατόν να πραγματοποιηθούν διάφορες προεκτάσεις τόσο για τη βελτίωση και την επιβεβαίωση της στατιστικής σημαντικότητας των αποτελεσμάτων που προέκυψαν όσο και για την εξαγωγή νέων συμπερασμάτων. Καταρχάς, όσον αφορά την παρούσα προσέγγιση και το συγκεκριμένο σύνολο δεδομένων αναφοράς, όπως αναφέρθηκε και παραπάνω υπάρχουν κάποια σημεία τα οποία χρήζουν περισσότερης προσοχής και απαιτούν ειδικούς χειρισμούς. Πρώτα από όλα, το γεγονός ότι η εκπομπή, της οποίας η τηλεθέαση εκτιμάται, προβάλλεται δύο μέρες την εβδομάδα εκ των οποίων η μία είναι Κυριακή, μη εργάσιμη μέρα, και η άλλη καθημερινή, έχει αντίκτυπο στην τηλεθέαση των επεισοδίων της. Αυτό δε λήφθηκε υπόψη στην παραπάνω ανάλυση και θα μπορούσε να εισαχθεί ως πληροφορία στους αλγορίθμους εκπαίδευσης για τη μοντελοποίηση, με τη μορφή μιας ανεξάρτητης μεταβλητής, η οποία θα λαμβάνει τις τιμές 1 και 0 (True και False), αναλόγως με το αν πρόκειται για καθημερινή ή όχι ημέρα προβολής.

Επιπλέον, λόγω της ιδιαιτερότητας του είδους της εκπομπής ίσως θα ήταν πιο χρήσιμο να ανιχνεύεται το συναίσθημα και όχι το πρόσημο (θετικό, αρνητικό, ουδέτερο) του κάθε tweet. Με αυτόν τον τρόπο, οι αναρτήσεις που περιέχουν χιούμορ και ειρωνεία, που όπως ήταν αναμενόμενο, είναι πάρα πολύ συχνές για μια σατιρική εκπομπή, θα μπορούσαν να κατηγοριοποιηθούν με μεγαλύτερη ακρίβεια. Επίσης, στη διαδικασία ανάλυσης συναισθήματος είναι πολύ σημαντικό να ληφθούν υπόψη τα emoji, να αντιστοιχιστούν σε κανονικές λέξεις οποιαδήποτε αρκτικόλεξα χρησιμοποιούνται και να αποκωδικοποιηθούν εκφράσεις αργκό, όλα από τα οποία είναι πολύ συχνά φαινόμενα σε μέσα κοινωνικής δικτύωσης, όπως το Twitter. Πάλι στα πλαίσια της ανάλυσης συναισθήματος, έναντι της απλουστευτικής στατικής προσέγγισης των προκαθορισμένων λεξικών, θα μπορούσε η αποτίμηση κάθε λέξης να επηρεάζεται από τα συμφραζόμενά της και το πλαίσιο στο οποίο εμφανίζεται, όπως υποδεικνύεται στην έρευνα [72].

Όσον αφορά το μέγεθος του συνόλου δεδομένων, αυτό μπορεί να επεκταθεί ούτως ώστε να περιέχει περισσότερες παρατηρήσεις και περισσότερα χαρακτηριστικά. Το πλήθος των παρατηρήσεων του συνόλου δεδομένων που χρησιμοποιήθηκε στην παρούσα μελέτη ήταν επαρκώς μεγάλο ώστε να μας επιτρέπει να το διασπάσουμε σε σύνολο εκπαίδευσης και ελέγχου. Ωστόσο το ιδανικό θα ήταν να συντίθεται από πολλές εκατοντάδες ή και χιλιάδες δείγματα για να είμαστε βέβαιοι για τη στατιστική σημαντικότητα των αποτελεσμάτων έπειτα από τον έλεγχο σε πολύ μεγαλύτερο πλήθος παρατηρήσεων και τη στιβαρότητα των κατασκευασμένων μοντέλων, τα οποία θα είναι πιο «έμπειρα» αφού θα έχουν εκπαιδευτεί με περισσότερα δείγματα. Αυτό μπορεί να επιτευχθεί για το ίδιο τηλεοπτικό πρόγραμμα λαμβάνοντας υπόψη περισσότερες σεζόν αλλά το καλύτερο θα ήταν να χτιστούν και περισσότερα μοντέλα για διάφορα είδη τηλεοπτικών εκπομπών ούτως ώστε τα συμπεράσματα να μπορούν να γενικευτούν για περισσότερες εκπομπές. Επίσης,

θα μπορούσαμε να εξάγουμε επιπλέον χαρακτηριστικά από την πλατφόρμα του Twitter, όπως το πλήθος των ακολούθων των χρηστών που έκαναν κάποια δημοσίευση, ώστε να λαμβάνεται υπόψη και η επιδραστικότητά τους, το πλήθος των διακριτών χρηστών που πραγματοποίησαν κάποια αναδημοσίευση και άλλα.

Συμπληρωματικά θα μπορούσαν να εξαχθούν δημογραφικά στοιχεία, όπως το φύλο, η ηλικία και το γεωγραφικό στίγμα. Τα δύο πρώτα είναι προσβάσιμα μόνο από την πλατφόρμα του Twitter και μόνο για όσους χρήστες έχουν επιλέξει να είναι ορατά δημόσια αυτά τα προσωπικά τους στοιχεία. Το γεωγραφικό στίγμα μπορεί να εξαχθεί από την πλατφόρμα Google Trends και τους χρωματισμένους χάρτες, που αυτή διαθέτει, σε επίπεδο γεωγραφικού διαμερίσματος της χώρας, την οποία μελετάμε. Αυτό μπορεί να αποβεί διαφωτιστικό για την απήχηση μιας εκπομπής στις πόλεις ή στην ύπαιθρο και σε συγκεκριμένες γεωγραφικές ενότητες, ελκύοντας διαφημίσεις από εταιρίες της τοπικής κοινωνίας. Από την άλλη, η γνώση του φύλου και της ηλικίας θα μπορούσε να χρησιμοποιηθεί για τη διάσπαση του συνόλου των χαρακτηριστικών σε υποσύνολα σύμφωνα με αυτά, με τέτοιο τρόπο ώστε να εκπαιδευτούν επιμέρους μοντέλα παλινδρόμησης για την εκτίμηση της τηλεθέασης μιας εκπομπής ανά ηλικιακές ομάδες και ανά φύλο. Για την εκπομπή Le Iene έχουμε στη διάθεση μας αυτά τα δεδομένα τηλεθέασης, τα οποία θα μπορούσαν να αξιοποιηθούν με μια τέτοια προσέγγιση.

Μια ακόμη πρόταση είναι να εξεταστούν και άλλοι αλγόριθμοι Μηχανικής Μάθησης και τεχνικές παλινδρόμησης. Πιο συγκεκριμένα, θα μπορούσε να δοθεί περισσότερη έμφαση στα πιθανοτικά μοντέλα και να εφαρμοστεί μπεϋζιανή παλινδρόμηση. Επίσης, αντί των gradient boosting μηχανών που υλοποιήθηκαν με τη βοήθεια της βιβλιοθήκης scikit-learn, θα ήταν καλή ιδέα να χρησιμοποιηθεί η βιβλιοθήκη XGBoost για την εφαρμογή αυτής της τεχνικής συλλογικής μάθησης, καθώς έχει μικρότερη τάση προς την υπερεκπαίδευση από ό,τι οι gradient boosting μηχανές [12]. Επιπλέον, σε όλες τις εκτελέσεις που πραγματοποιήθηκαν, η διάσπαση σε σύνολο δεδομένων εκπαίδευσης και σύνολο δεδομένων ελέγχου ήταν τυχαία. Μια εναλλακτική λύση είναι να θεωρήσουμε τις παρατηρήσεις που αντιστοιχούν στα πρώτα επεισόδια ως σύνολο εκπαίδευσης και αυτές που αντιστοιχούν στα τελευταία ως σύνολο ελέγχου, δηλαδή να τα αντιμετωπίσουμε σαν χρονοσειρά. Σε αυτήν την προσέγγιση είναι δυνατόν να αξιοποιηθούν συμπληρωματικά ιστορικές μετρήσεις τηλεθέασης, από προηγούμενα επεισόδια και οι προβλέψεις να παραχθούν από ένα επαναληπτικό νευρωνικό δίκτυο (recurrent neural network - RNN), το οποίο λόγω της δομής του και των αναδράσεων που διαθέτει είναι σε θέση να αντιλαμβάνεται το χρόνο, παρέχοντας λειτουργία μνήμης [33].

Τέλος, η προσέγγιση που παρουσιάσαμε στη συγκεκριμένη διπλωματική εργασία δεν πραγματοποιεί πραγματικά μελλοντικές προβλέψεις, καθώς τα δεδομένα που χρησιμοποιεί για την εκτίμηση της τηλεθέασης ενός επεισοδίου συνήθως παράγονται κατά τη διάρκεια προβολής του ή και μετά από αυτήν. Συνεπώς, για να καταστεί δυνατή η πρόβλεψη της τηλεθέασης ενός μελλοντικού επεισοδίου προτού αυτό προβληθεί στην τηλεόραση θα πρέπει πρώτα να εκτιμηθούν τα χαρακτηριστικά αυτού με μεθόδους ανάλυσης και πρόβλεψης χρονοσειρών. Ενδεικτικά, μπορούν να χρησιμοποιηθούν οι τεχνικές απλού μέσου (simple average), κινητού μέσου (moving average), εκθετικής εξομάλυνσης (exponential smoothing) και ολοκληρωμένων αυτοπαλινδρομικών μοντέλων κινητού μέσου (ARIMA) μεταξύ άλλων [10]. Έπειτα, οι εκτιμώμενες τιμές των χαρακτηριστικών θα χρησιμοποιηθούν ως είσοδοι στα ήδη κατασκευασμένα μοντέλα παλινδρόμησης για να προκύψουν τελικά οι προβλέψεις για το ποσοστό τηλεθέασης επί των ανοικτών δεκτών και το απόλυτο πλήθος τηλεθεατών που θα συγκεντρώσει το υπό μελέτη επερχόμενο επεισόδιο.

Βιβλιογραφία

[1] Harshavardhan Achrekar et al. "Predicting flu trends using twitter data". In: *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*. IEEE. 2011, pp. 702–707.

[2] Thomas Aichner and Frank Jacob. "Measuring the degree of corporate social media use". In: *International Journal of market research* 57.2 (2015), pp. 257–276.

[3] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[4] Chidanand Apté and Sholom Weiss. "Data mining with decision trees and decision rules". In: *Future generation computer systems* 13.2-3 (1997), pp. 197–210.

[5] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter". In: *Proceedings of the 2011 Conference on empirical methods in natural language processing*. 2011, pp. 1568–1576.

[6] Sitaram Asur and Bernardo A Huberman. "Predicting the future with social media". In: *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*. Vol. 1. IEEE. 2010, pp. 492–499.

[7] Mariette Awad and Rahul Khanna. "Support vector regression". In: *Efficient learning machines*. Springer, 2015, pp. 67–80.

[8] Biplab Bhattacharjee, Amulyashree Sridhar, and Anirban Dutta. "Identifying the causal relationship between social media content of a Bollywood movie and its box-office success-a text mining approach". In: *International Journal of Business Information Systems* 24.3 (2017), pp. 344–368.

[9] Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market". In: *Journal of computational science* 2.1 (2011), pp. 1–8.

[10] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[11] *Broadcasting & Cable TV Market Size, Share & Trends Analysis Report By Technology (Cable TV, Satellite TV, IPTV, DTT), By Revenue Channel (Advertising, Subscription), By Region, And Segment Forecasts, 2020 - 2027*. GRAND VIEW RESEARCH. May 2020. URL: https://www.grandviewresearch.com/industry-analysis/broadcasting-and-cable-tv-market.

[12] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[13] Mei-Hua Cheng, Yi-Chen Wu, Ming-Chih Chen, et al. "Television meets facebook: The correlation between tv ratings and social media". In: *American Journal of Industrial and Business Management* 6.03 (2016), p. 282.

[14] Hyunyoung Choi and Hal Varian. "Predicting the present with Google Trends". In: *Economic record* 88 (2012), pp. 2–9.

[15] J Clement, ed. *Most popular social networks worldwide as of July 2020, ranked by number of active users*. statista. Aug. 2020. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[16] J Clement. *Social media - Statistics & Facts*. statista. May 2020. URL: https://www.statista.com/topics/1164/social-networks/#dossierSummary__chapter1.

[17] J Clement. *Worldwide desktop market share of leading search engines from January 2010 to July 2020*. statista. Sept. 2020. URL: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/.

[18] Jan Salomon Cramer. "The origins of logistic regression". In: (2002).

[19] Alfonso Crisci et al. "Predicting TV programme audience by using twitter based metrics". In: *Multimedia Tools and Applications* 77.10 (2018), pp. 12203–12232.

[20] Francesco D'Amuri and Juri Marcucci. "The predictive power of Google searches in forecasting US unemployment". In: *International Journal of Forecasting* 33.4 (2017), pp. 801–816.

[21] Deniz Demir, Olga Kapralova, and Hongze Lai. "Predicting IMDB movie ratings using Google Trends". In: 2012.

[22] Thomas Dimpfl and Stephan Jank. "Can internet search queries help to predict stock market volatility?". In: *European Financial Management* 22.2 (2016), pp. 171–192.

[23] Norman R Draper and Harry Smith. *Applied regression analysis*. Vol. 326. John Wiley & Sons, 1998.

[24] *FAQ about Google Trends data*. Google. URL: https://support.google.com/trends/answer/4365533.

[25] Yoav Freund, Robert Schapire, and Naoki Abe. "A short introduction to boosting". In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780 (1999), p. 1612.

[26] Salvador Garcia, Julian Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.

[27] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[28] Fabio Giglietto. "Exploring correlations between TV viewership and Twitter conversations in Italian political talk shows". In: *Available at SSRN 2306512* (2013).

[29] Sharad Goel et al. "Predicting consumer behavior with Web search". In: *Proceedings of the National academy of sciences* 107.41 (2010), pp. 17486–17490.

[30] *Google Search Statistics*. Internet Live Stats. URL: https://www.internetlivestats.com/google-search-statistics/#trend.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[32] Simon Haykin. *Neural Networks and Learning Machines, 3/E*. Pearson Education India, 2010.

[33] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[34] Seyed Hamid Hosseini and Mahdi Samanipour. "Prediction of final concentrate grade using artificial neural networks from Gol-E-Gohar iron ore plant". In: *American Journal of Mining and Metallurgy* 3.3 (2015), pp. 58–62.

[35] Wen-Tai Hsieh et al. "Predicting tv audience rating with social media". In: *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*. 2013, pp. 1–5.

[36] Yu-Yang Huang et al. “A Weight-Sharing Gaussian Process Model Using Web-Based Information for Audience Rating Prediction”. In: *International Conference on Technologies and Applications of Artificial Intelligence*. Springer. 2014, pp. 198–208.

[37] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*. Vol. 2. CRC Press, 2010.

[38] Nikos Kalatzis et al. “Social Media and Google Trends in Support of Audience Analytics: Methodology and Architecture”. In: *DATA ANALYTICS 2018* (2018), p. 49.

[39] Autar Kaw and Luke Snyder. “Nonlinear regression”. In: (1989).

[40] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[41] SB Kotsiantis, Dimitris Kanellopoulos, and PE Pintelas. “Data preprocessing for supervised leaning”. In: *International Journal of Computer Science* 1.2 (2006), pp. 111–117.

[42] Andy Liaw, Matthew Wiener, et al. “Classification and regression by randomForest”. In: *R news* 2.3 (2002), pp. 18–22.

[43] Yu-Hsuan Lin, Chun-Hao Liu, and Yu-Chuan Chiu. “Google searches for the keywords of “wash hands” predict the speed of national spread of COVID-19 outbreak among 21 countries”. In: *Brain, Behavior, and Immunity* (2020).

[44] Bing Liu. “Sentiment analysis and opinion mining”. In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167.

[45] Dong C Liu and Jorge Nocedal. “On the limited memory BFGS method for large scale optimization”. In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[46] Yafeng Lu et al. “Integrating predictive analytics and social media”. In: *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2014, pp. 193–202.

[47] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[48] Yossi Matias. *Insights into what the world is searching for – the new Google Trends*. Google Search blog. Sept. 2012. URL: https://search.googleblog.com/2012/09/insights-into-what-world-is-searching.html.

[49] *Metrics and scoring: quantifying the quality of predictions*. scikit-learn. URL: https://scikit-learn.org/stable/modules/model_evaluation.html.

[50] Rui Miao and Yueyue Ma. “The Dynamic Impact of Web Search Volume on Product Sales—An Empirical Study Based on Box Office Revenues”. In: *Wuhan International Conference on e-Business (WHICEB 2015)*. 2015.

[51] Tom M Mitchell et al. “Machine learning. 1997”. In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

[52] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[53] Luca Molteni and J Ponce De Leon. “Forecasting with twitter data: an application to Usa Tv series audience”. In: *International Journal of Design & Nature and Ecodynamics* 11.3 (2016), pp. 220–229.

[54] Raymond H Myers and Raymond H Myers. *Classical and modern regression with applications*. Vol. 2. Duxbury press Belmont, CA, 1990.

[55] Alexey Natekin and Alois Knoll. “Gradient boosting machines, a tutorial”. In: *Frontiers in neurorobotics* 7 (2013), p. 21.

[56] Sudhakar V Nuti et al. “The use of google trends in health care research: a systematic review”. In: *PloS one* 9.10 (2014), e109583.

[57] Jonathan A Obar and Steven S Wildman. “Social media definition and the governance challenge-an introduction to the special issue”. In: *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39.9 (2015), pp. 745–750.

[58] Andrei Oghina et al. “Predicting imdb movie ratings using social media”. In: *European Conference on Information Retrieval*. Springer. 2012, pp. 503–507.

[59] Chong Oh, Sheila Sasser, and Soliman Almahmoud. “Social media analytics framework: the case of Twitter and Super Bowl ads”. In: *Journal of Information Technology Management* 26.1 (2015), pp. 1–18.

[60] Travis E Oliphant. “Python for scientific computing”. In: *Computing in Science & Engineering* 9.3 (2007), pp. 10–20.

[61] Marios Paraskevopoulos. “Machine learning for TV ratings estimation based on social media data”. Bachelor’s thesis. National Technical University of Athens, 2018.

[62] Michael J Paul, Mark Dredze, and David Broniatowski. “Twitter improves influenza forecasting”. In: *PLoS currents* 6 (2014).

[63] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[64] Richard R Picard and R Dennis Cook. “Cross-validation of regression models”. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 575–583.

[65] Christos Pierrakos. “Machine learning on social media data for TV ratings estimation”. Bachelor’s thesis. National Technical University of Athens, 2019.

[66] Philip M Polgreen et al. “Using internet searches for influenza surveillance”. In: *Clinical infectious diseases* 47.11 (2008), pp. 1443–1448.

[67] Carl Edward Rasmussen. “Gaussian processes in machine learning”. In: *Summer School on Machine Learning*. Springer. 2003, pp. 63–71.

[68] Ajay Siva Santosh Reddy, Pratik Kasat, and Abhiyash Jain. “Box-office opening prediction of movies based on hype analysis through data mining”. In: *International Journal of Computer Applications* 56.1 (2012), pp. 1–5.

[69] Lior Rokach and Oded Z Maimon. *Data mining with decision trees: theory and applications*. Vol. 69. World scientific, 2008.

[70] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[71] Thomas P Ryan. *Modern regression methods*. Vol. 655. John Wiley & Sons, 2008.

[72] Hassan Saif et al. “Contextual semantics for sentiment analysis of Twitter”. In: *Information Processing & Management* 52.1 (2016), pp. 5–19.

[73] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. “Earthquake shakes Twitter users: real-time event detection by social sensors”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 851–860.

[74] Wernard Schmit and Sander Wubben. “Predicting ratings for new movie releases from twitter content”. In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2015, pp. 122–126.

[75] Patrick Schober, Christa Boer, and Lothar A Schwarte. “Correlation coefficients: appropriate use and interpretation”. In: *Anesthesia & Analgesia* 126.5 (2018), pp. 1763–1768.

[76] Scott Sereday and Jingsong Cui. “Using machine learning to predict future tv ratings”. In: *Data Science, Nielsen* 1.3 (2017), pp. 3–12.

[77] Dehua Shen, Andrew Urquhart, and Pengfei Wang. “Does twitter predict Bitcoin?”. In: *Economics Letters* 174 (2019), pp. 118–122.

[78] Viktor Slavkovikj et al. “Review of wildfire detection using social media”. In: *Fire safety journal* 68 (2014), pp. 109–118.

[79] Alex J Smola and Bernhard Schölkopf. “A tutorial on support vector regression”. In: *Statistics and computing* 14.3 (2004), pp. 199–222.

[80] Bridget Sommerdijk, Eric Sanders, and Antal van den Bosch. “Can Tweets Predict TV Ratings?”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2965–2970.

[81] Danny Sullivan. “How search engines work”. In: *SEARCH ENGINE WATCH (on file with the New York University Journal of Legislation and Public Policy)* (2002).

[82] Sameer Thigale et al. “Prediction of box office success of movies using hype analysis of Twitter data”. In: *International Journal of Innovative Engineering and Science* 3.1 (2014), pp. 1–6.

[83] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[84] Andrei Nikolaevich Tikhonov et al. *Numerical methods for the solution of ill-posed problems*. Vol. 328. Springer Science & Business Media, 2013.

[85] Andranik Tumasjan et al. “Predicting elections with twitter: What 140 characters reveal about political sentiment”. In: *Fourth international AAAI conference on weblogs and social media*. Citeseer. 2010.

[86] *Twitter API*. Twitter. URL: `https://developer.twitter.com/en`.

[87] Paul E Utgoff. “Incremental induction of decision trees”. In: *Machine learning* 4.2 (1989), pp. 161–186.

[88] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. “Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs”. In: *International Conference on Database Systems for Advanced Applications*. Springer. 2011, pp. 390–401.

[89] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. “Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating”. In: *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. 2011, pp. 1–10.

[90] Hao Wang et al. “A system for real-time twitter sentiment analysis of 2012 us presidential election cycle”. In: *Proceedings of the ACL 2012 system demonstrations*. 2012, pp. 115–120.

[91] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. “Automatic crime prediction using events extracted from twitter posts”. In: *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer. 2012, pp. 231–238.

[92] Xiaojun Wang et al. “Using Twitter data to predict the performance of Bollywood movies”. In: *Industrial Management & Data Systems* (2015).

[93] Amy Watson, ed. *Television Industry - Statistics & Facts*. statista. Oct. 2019. URL: `https://www.statista.com/topics/977/television/`.

[94] Amy Watson, ed. *Traditional TV industry revenue worldwide from 2018 to 2022(in billion U.S. dollars)*. statista. Nov. 2018. URL: `https://www.statista.com/statistics/265983/global-tv-industry-revenue/`.

[95] Lynn Wu and Erik Brynjolfsson. "The future of prediction: How Google searches foreshadow housing prices and sales". In: *Economic analysis of the digital economy*. University of Chicago Press, 2015, pp. 89–118.

[96] Xin Yang et al. "Forecasting Chinese tourist volume with search engine data". In: *Tourism Management* 46 (2015), pp. 386–397.

[97] Qingyu Yuan et al. "Monitoring influenza epidemics in china with search query from baidu". In: *PloS one* 8.5 (2013), e64323.

[98] Xue Zhang, Hauke Fuehres, and Peter A Gloor. "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"". In: *Procedia-Social and Behavioral Sciences* 26 (2011), pp. 55–62.

[99] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.