ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

# Ανάπτυξη προγνωστικών μοντέλων κρουσμάτων γριπωδών συνδρομών με χρήση τεχνικών μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Φραγκοζίδης Γεώργιος**

**Επιβλέπουσα** : Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

**Μέλη Επιτροπής:** Ανδρέας – Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2021

# Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα της διπλωματικής μου εργασίας, καθηγήτρια κα. Κωνσταντίνα Νικήτα, χάρη στην οποία μου δόθηκε η ευκαιρία να έρθω σε επαφή με ένα θέμα εξαιρετικού ενδιαφέροντος για εμένα, αλλά και όλους τους καθηγητές που το κατέστησαν αυτό δυνατό.

Ευχαριστώ ιδιαιτέρως την υποψήφια διδάκτορα Μαρία Αθανασίου για την άμεση και πολύτιμη βοήθειά της στην συγγραφή της εργασίας μου, τόσο σε επίπεδο καθοδήγησης και συμβουλών, καθ' όλη την εκπόνησή της, όσο και σε επίπεδο στήριξης και ενθάρρυνσης.

Τέλος, θα ήθελα να εκφράσω την βαθιά μου ευγνωμοσύνη στους γονείς μου οι οποίοι είναι δίπλα μου και με στηρίζουν σε κάθε βήμα της ζωής και της εξέλιξής μου, με αγάπη και υπομονή.

Γεώργιος Φραγκοζίδης

Αθήνα, 10η Ιανουαρίου 2021

# Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η σχεδίαση, η ανάπτυξη και η αξιολόγηση υπολογιστικών μοντέλων με στόχο την πρόβλεψη του αριθμού των κρουσμάτων γριπωδών συνδρομών (Influenza Like Illness ή ILI) στον πληθυσμό της Ελλάδας, βασισμένου στη χρήση ετερογενών δεδομένων και στην εφαρμογή μεθόδων μηχανικής μάθησης. Για την δεκαετία 2010-2020 συλλέχθηκαν μετεωρολογικά δεδομένα των 13 περιφερειών της Ελλάδας, δεδομένα του συστήματος επιτήρησης γρίπης «Sentinel», που παραχωρήθηκαν από τον Εθνικό Οργανισμό Δημόσιας Υγείας (ΕΟΔΥ), και δημοσιεύσεις στο Twitter που αφορούσαν την υγεία των χρηστών. Πιο συγκεκριμένα, οι μετεωρολογικοί δείκτες που λήφθηκαν υπόψη ήταν η ημερήσια θερμοκρασία, η υγρασία, η ταχύτητα του ανέμου και η ηλιοφάνεια, και έχουν συσχετιστεί άμεσα με την εξάπλωση των ιών, που φέρουν συμπτώματα παρόμοια με αυτά της γρίπης. Για τα δεδομένα κοινωνικής δικτύωσης πραγματοποιήθηκε συλλογή δημοσιεύσεων από την πλατφόρμα του Twitter με χρήση συγκεκριμένων λέξεων κλειδιών σχετικών με τις ILI, και εφαρμόστηκαν προηγμένες τεχνικές προεπεξεργασίας όπως η φασματική ομαδοποίηση και η εποχική αποδόμηση με στόχο την εξαγωγή χαρακτηριστικών εισόδου με ποιοτικά ενισχυμένη πληροφορία. Τέλος, συνυπολογίστηκε ο εβδομαδιαίος αριθμός των κρουσμάτων ILI για την Ελλάδα.

Οι τρεις κατηγορίες δεδομένων (μετεωρολογικά, επιδημιολογικά, κοινωνικής δικτύωσης) χρησιμοποιήθηκαν για την ανάπτυξη τριών πρωταρχικών προγνωστικών μοντέλων. Κάθε πρωταρχικό μοντέλο έλαβε ως είσοδο τιμές δεδομένων προηγούμενων εβδομάδων και παρείχε εκτιμήσεις για τον αριθμό των κρουσμάτων ILI κατά την τρέχουσα εβδομάδα και τις δύο επόμενες εβδομάδες. Στη συνέχεια, τα πρωταρχικά μοντέλα αξιοποιήθηκαν στο πλαίσιο συνδυαστικών αρχιτεκτονικών με στόχο τη δημιουργία σύνθετων μοντέλων με υψηλή προγνωστική ικανότητα. Για την ανάπτυξη των προγνωστικών μοντέλων και την αποτελεσματική διαχείριση της χρονολογικής φύσης των δεδομένων εισόδου επιστρατεύτηκε η μέθοδος των Νευρωνικών Δικτύων Μακρόχρονης και Βραχύχρονης Μνήμης (Long Short Term Memory Neural Networks-LSTM). Τα πρωταρχικά και σύνθετα μοντέλα αξιολογήθηκαν ως προς την ικανότητά τους να παρέχουν ακριβείς εκτιμήσεις του αριθμού κρουσμάτων ILI για χρονικό ορίζοντα έως και τριών εβδομάδων στο μέλλον.

# Λέξεις κλειδιά

Γριπώδεις Συνδρομές, ασθένειες, καιρός, κρούσματα, Twitter, τεχνητά νευρωνικά δίκτυα, βαθιά μάθηση, ομαδοποίηση, μηχανική μάθηση, επεξεργασία φυσικής γλώσσας, παλινδρόμηση, μοντέλο πρόβλεψης

# Abstract

The present thesis aims at the design, development and evaluation of computational models towards the prediction of the number of cases of influenza syndrome (Influenza Like Illness or ILI) in the population of Greece, based on the use of heterogeneous data and the application of machine learning techniques. For the decade 2010-2020, meteorological parameters of the 13 regions of Greece along with epidemiological data of the ILI Sentinel surveillance system, provided by the National Public Health Organization, and Twitter posts concerning the health of users, were considered to compose the models' input space. In terms of meteorological parameters, the daily temperature, humidity, wind speed and sunlight were taken into account due to their association with the spread of viruses presenting flu-like symptoms. With respect to social media data, Twitter posts were collected based on specific keywords, and advanced pre-processing techniques, including spectral clustering and seasonal decomposition, were applied towards the extraction of highly informative features. Finally, information regarding the weekly number of past influenza cases was also taken into consideration.

Three primary prediction models were developed by utilizing each of the three categories of collected data (meteorological, Twitter, epidemiological). Each model received input data from previous weeks and provided estimates of the number of ILI cases for the current week and the following two weeks. Subsequently, these primary models were combined within the framework of different architectures towards the development of complex models with a high predictive power. The method of Long Short Term Memory Neural Networks (LSTM) was deployed for the primary and complex models' development, due to its efficiency in handling sequential data. The primary and complex models were assessed in terms of their ability to provide accurate estimates of ILI cases for up to three weeks in the future.



# Keywords

Influenza Like Illness, Diseases, Weather, Cases, Twitter, Artificial Neural Networks, Deep Learning, Clustering, Machine Learning, Natural Language Processing, Regression, Prediction Model

# Περιεχόμενα

# Κεφάλαιο 1

# Εισαγωγή

## *1.1 Μολυσματικές Ασθένειες*

Σαν ασθένεια, ορίζεται μια συγκεκριμένη κατάσταση η οποία παρεμποδίζει την ομαλή διεξαγωγή των λειτουργιών ενός οργανισμού ή κάποιου συστήματος αυτού του οργανισμού, όπως για παράδειγμα κάποιο σωματικό όργανο. Σε αντίθεση με έναν τραυματισμό, οι επιπλοκές από μία ασθένεια γίνονται εμφανείς σταδιακά με την πάροδο του χρόνου, σε διάστημα μερικών ημερών ή και χρόνων, όπως για παράδειγμα ισχύει με τη μαλάρια ή με την νόσο του Αλτσχάιμερ. Συνήθως, υπάρχει παρουσία από συγκεκριμένα συμπτώματα, τα οποία στους ανθρώπους μπορούν να είναι είτε σωματικά (καταρροή, βήχας, μυϊκός πόνος κλπ.) είτε ψυχολογικά (νευρογνωστικές διαταραχές, εκφυλισμός της μνήμης κλπ.) [1,3]. Οι λόγοι για τους οποίους κάποιος μπορεί να νοσήσει από μια ασθένεια μπορεί να είναι είτε εσωτερικές διεργασίες όπως μία γενετική μετάλλαξη ή μία αλλεργία, είτε εξωτερικοί παράγοντες, όπως η μόλυνση από το τσίμπημα ενός κουνουπιού. Όταν ένας οργανισμός νοσεί λόγω της εισβολής εξωτερικών παραγόντων των οποίων η δραστηριότητα είναι επιβλαβής προς αυτόν, τότε η ασθένεια χαρακτηρίζεται σαν μολυσματική [5].

Οι μολυσματικές ασθένειες προκαλούνται κατά βάση από μικροοργανισμούς, όπως οι ιοί ή κάποια βακτήρια, των οποίων η ανάπτυξη και μεταβολική δραστηριότητα οδηγεί σε ποικίλες βλάβες στο σώμα του ξενιστή[4]. Το βακτήριο C.Tetani, το οποίο ευθύνεται για τον τέτανο, για παράδειγμα, παράγει πλήθος από επικίνδυνες τοξίνες και ένζυμα τα οποία διασπούν μόρια όπως τα λιπίδια, τις πρωτεΐνες ή το κολλαγόνο, και προκαλεί σαν συμπτώματα του έντονους μυϊκούς σπασμούς και πόνους [7]. Στους περισσότερους οργανισμούς, ο κύριος μηχανισμός άμυνας απέναντι σε αυτούς τους επιβλαβείς μικροοργανισμούς είναι το ανοσοποιητικό σύστημα, το οποίο εντοπίζει τους ξενικούς παράγοντες και μέσω ενός πολυσύνθετου δικτύου βιολογικών διεργασιών και συστημάτων, τους αποβάλλει. Επιπλέον, χάρις την βελτίωση των συνθηκών διαβίωσης και υγιεινής, της ανάπτυξης εμβολίων και δημιουργίας στοχευμένων φαρμακευτικών αγωγών, αρκετές από τις ασθένειες που μάστιζαν τον ανθρώπινο πληθυσμό όπως η ευλογιά ή η πολιομυελίτιδα έχουν πλέον εξαφανιστεί [6].

Αρκετοί ιοί και βακτήρια ωστόσο, έχουν αναπτύξει ανθεκτικότητα στα συνηθισμένα φάρμακα, ενώ για πολλούς από αυτούς η περίοδος επώασης (χρονικό διάστημα από την μόλυνση μέχρι την ασθένεια) είναι μεγαλύτερη από την περίοδο αδράνειας με αποτέλεσμα ο ξενιστής να μολύνει άθελά του τα άτομα με τα οποία έρχεται σε επαφή, και να παρατηρηθεί απότομη αύξηση στα κρούσματα μιας ασθένειας.  Σε αυτή την περίπτωση, λέμε πως έχουμε

ξέσπασμα επιδημικής κρίσης. Μία επιδημική κρίση μπορεί να φέρει τεράστιο αντίκτυπο σε κάθε τομέα της κοινωνικής δραστηριότητας (οικονομικό, πολιτιστικό κλπ.). Υπολογίζεται πως σε μια τέτοια περίοδο το 5-20% του πληθυσμού θα νοσήσει, και παγκοσμίως οι κρίσεις αυτές ευθύνονται σημαντικά για το ποσοστό των απουσιών από τα σχολεία και τον εργασιακό κλάδο [8]. Το 2003 υπολογίζεται πως στις ΗΠΑ ξοδεύτηκαν πάνω από δέκα δισεκατομμύρια δολάρια σε ιατρικά έξοδα (εξοπλισμός, φάρμακα, προσωπικό) ενώ οι απώλειες στα κέρδη των επιχειρήσεων λόγω θανάτων, αδειών νοσηλείας και μειωμένης αποδοτικότητας των εργαζομένων ξεπέρασαν τα δεκαπέντε δισεκατομμύρια δολάρια [10]. Παρόμοιο κόστος παρουσιάστηκε και στη Γερμανία και τη Γαλλία την ίδια χρονιά.

## *1.2 Γριπώδεις Συνδρομές - ILI*

Ο Παγκόσμιος Οργανισμός Υγείας (World Health Organization, WHO) ορίζει τις Γριπώδεις Συνδρομές (Influenza-Like Illnesses, ILI) σαν οξείες λοιμώξεις του αναπνευστικού συστήματος οι οποίες παρουσιάζουν ως συμπτώματα τον έντονο πυρετό (>38 C°), τον πονοκέφαλο, τον ξηρό βήχα (ενδεχομένως έντονο και με διάρκεια έως και δύο εβδομάδων), τον πόνο στις αρθρώσεις και τους μύες, τον πονόλαιμο και την καταρροή [9]. Παρά το όνομά τους, τις περισσότερες φορές μια γριπώδης συνδρομή δεν προκαλείται από τον ιό της γρίπης αλλά από κάποιον άλλον ιό, όπως για παράδειγμα τους ρινοϊούς, τους αδενοϊούς και λιγότερο συχνά λόγω βακτηρίων σαν τη Λεγεωνέλλα ή το Chlamydia pneumoniae [7]. Στους περισσότερους ανθρώπους τα συμπτώματα περνούν μέσα σε μία εβδομάδα χωρίς την ανάγκη νοσηλείας. Ωστόσο, στις ευπαθείς ομάδες, ο κίνδυνος σοβαρής ασθένειας ή και θανάτου είναι πολύ πιο υψηλός.

Σε παγκόσμιο επίπεδο, εμφανίζονται τρεις έως και πέντε εκατομμύρια περιπτώσεις σοβαρής ασθένειας και 290.000 με 650.000 θάνατοι από αναπνευστικά προβλήματα, ενώ στις βιομηχανοποιημένες χώρες οι περισσότεροι θάνατοι προκύπτουν στις ηλικίες άνω των 65 ετών. Ως εκ τούτου, ένα επιδημιολογικό κύμα μπορεί να επιφέρει υψηλά επίπεδα τόσο στην αποχή από την εκπαιδευτική διαδικασία για τα παιδιά όσο και επιπτώσεις στην παραγωγική και επιχειρηματική δραστηριότητα [8-10][46-47]. Επίσης, λόγω του όγκου των κρουσμάτων, τα νοσοκομεία και οι κλινικές οδηγούνται σε κατάσταση υπερπλήρωσης με αποτέλεσμα τη δημιουργία περισσότερων προβλημάτων.

Όσον αφορά την μεταδοτικότητα της ασθένειας, οι λοιμώξεις αυτές μεταδίδονται εύκολα, με ακόμη υψηλότερους ρυθμούς σε χώρους με υψηλή συγκέντρωση ανθρώπων, όπως τα σχολεία ή τα γηροκομεία. Όταν ένας μολυσμένος άνθρωπος φτερνίζεται ή βήχει, μολυσματικά σταγονίδια διαχέονται στον αέρα σε απόσταση έως και ενός μέτρου και μολύνουν όποιον τα εισπνεύσει. Η μετάδοση πραγματοποιείται επίσης και με την σωματική επαφή, για παράδειγμα μέσω μιας χειραψίας ή αγκαλιάς. Στις χώρες με ήπιο κλίμα, οι ILI ξεσπούν κυρίως τον χειμώνα, ενώ στα τροπικά κλίματα έχουμε πιο ιδιόρρυθμα μοτίβα. Η περίοδος επώασης του ιού είναι περίπου δύο με τέσσερεις ημέρες. Υψηλά συσχετισμένοι

δείκτες με την ταχύτητα εξάπλωσης μιας επιδημίας είναι, μεταξύ άλλων, η πυκνότητα πληθυσμού και ο ρυθμός μετακίνησης από και προς τον χώρο εργασίας. Τα αποτελέσματα των προσομοιώσεών των Viboud et al. [38] φανερώνουν πως η μετάδοση των κρουσμάτων σε μια πολιτεία είναι ευθέως ανάλογη με τον πληθυσμό της και με το εύρος συνδεσιμότητας που παρέχουν τα MMM.

Η διάγνωση των συνδρομών γίνεται κυρίως κλινικά, όμως σε περιόδους χαμηλής παρουσίας του ιού ή εκτός μιας επιδημίας, η παρουσία άλλων λοιμώξεων, όπως ο ρινοϊός ή ο αδενοϊός, δυσκολεύουν την διαδικασία της διάγνωσης. Για μία έγκυρη διάγνωση, απαιτείται η συλλογή αναπνευστικών δειγμάτων και η πραγματοποίηση κλινικών δοκιμών μέσω εντοπισμού αντιγόνων, απομόνωση των ιικών στελεχών ή της ταυτοποίησης μέσω ριβονουκλεϊκού οξέος (RNA) [9].

## *1.3 Επιδημιολογική Παρακολούθηση*

Ο σκοπός της επιδημιολογικής παρακολούθησης είναι η συνεχής και συστηματική συλλογή, ανάλυση και ερμηνεία δεδομένων υγείας, που στοχεύει στον αποτελεσματικό σχεδιασμό, ανάπτυξη και αξιολόγηση των υποδομών δημόσιας υγείας. Για αυτό τον λόγο, ανά κράτος, έχουν αναπτυχθεί συστήματα παρατηρητών νοσηρότητας, τα οποία συλλέγουν και επεξεργάζονται χρήσιμες πληροφορίες γύρω από τα κρούσματα επιδημικών ασθενειών, με σκοπό την πιο αποτελεσματική διαχείριση μίας επιδημίας, τη λήψη των απαραίτητων μέτρων προστασίας των πολιτών και τον καθορισμό καλύτερου πλάνου διαχείρισης των υγειονομικών πόρων [14]. Οι νόσοι που επιτηρούνται είναι κυρίως το σοβαρό οξύ αναπνευστικό σύνδρομο ή SARS (Severe Acute Respiratory Syndrome), οι γριπώδεις συνδρομές και η γαστρεντερίτιδα.

Σύμφωνα με τον WHO, με την παροχή έγκαιρων και ποιοτικών επιδημιολογικών δεδομένων, μπορούν να διεξαχθεί ένα σύνολο αποτελεσμάτων όπως, μεταξύ άλλων, η καταγραφή της εποχικότητας μιας ασθένειας, η περιγραφή της γενετικής σύνθεσης των ιών που κυκλοφορούν ή και η αξιολόγηση της αποτελεσματικότητας των παρεμβάσεων των δημόσιων φορέων υγείας. Ποια από αυτά θα διεξαχθούν και εντέλει θα αξιοποιηθούν εξαρτάται από τις ανάγκες και τις δυνατότητες κάθε κράτους.

*Πίνακας 1.1 Στόχοι και επιλεγμένες δράσεις της επιδημιολογικής παρακολούθησης*

| ***Στόχοι της επιδημιολογικής παρακολούθησης*** | ***Προβλεπόμενες δράσεις*** |
|---|---|
| Πρόβλεψη και καταγραφή επιδημικής δραστηριότητας | Προετοιμασία ιατρικών φορέων, εφαρμογή πολιτικών εμβολιασμού, χορήγηση φαρμακευτικών αγωγών κ.α. |
| Ανίχνευση γενετικών μεταλλάξεων των ιών | Ενημέρωση ιατρών για αντιικές θεραπείες και εμβόλια |
| Περιγραφή κλινικών μοτίβων της ασθένειας | Προστασία ασθενών υψηλού ρίσκου - καθορισμός ομάδων προτεραιότητας για εμβολιασμό/θεραπεία |
| Εκτίμηση επικινδυνότητας μιας επιδημίας ή ενός ιού | Ενημέρωση των πολιτών, συνυπολογισμός του οικονομικού κόστους στη λήψη αποφάσεων |
| Αποτίμηση επιδημικού φορτίου των γριπωδών συνδρομών | Διαμοιρασμός πόρων και καθορισμός ορίου επικινδυνότητας |
| Εντοπισμός ασυνήθιστων γεγονότων | Ειδοποίηση σχετικών αρχών |
| Μέτρηση αποτελεσματικότητας παρεμβάσεων | Βέλτιστη επιλογή πολιτικής παρέμβασης |

### *1.3.1 Συλλογή Δεδομένων*

Η επίβλεψη των κρουσμάτων μια ασθένειας απαιτεί τη συλλογή δεδομένων μέσω διαγνώσεων ή κλινικών ελέγχων, από προσεκτικά επιλεγμένες εστίες παρακολούθησης σε κάθε περιοχή μιας χώρας. Τα αποτελέσματα οφείλουν να είναι αντιπροσωπευτικά του πληθυσμού της, καθώς ανάλογα με τα διαφορετικά δημογραφικά χαρακτηριστικά που

παρουσιάζονται, όπως η ηλικία ή το φύλο, αλλά και τα κοινωνικά/οικονομικά, καθώς δημιουργούνται διαφοροποιήσεις στον τρόπο μετάδοσης της νόσου.

Ο αριθμός των ασθενών που θα συμπεριληφθούν στο δείγμα επίβλεψης αποφασίζεται αφού πραγματοποιηθεί μια επισκόπηση των δυνατοτήτων και των πόρων της εκάστοτε κλινικής. Στην ιδανική περίπτωση, πραγματοποιείται κλινικός έλεγχος και συλλογή των αποτελεσμάτων για κάθε ασθενή που επισκέπτεται μια κλινική, αλλά αυτό τις περισσότερες φορές δεν είναι δυνατόν. Έτσι, καθίσταται φανερή η ανάγκη για έναν δειγματοληπτικό έλεγχο των ασθενών. Πρέπει να δοθεί ιδιαίτερη προσοχή ωστόσο, στην στρατηγική δειγματοληψίας, έτσι ώστε να αποφευχθεί η εξαγωγή εσφαλμένων συμπερασμάτων λόγω στατιστικών λαθών.

Για παράδειγμα, μια εύκολα υλοποιήσιμη και φθηνή στρατηγική, θα ήταν να επιλέγει τυχαία ο γιατρός δείγματα από ασθενείς υπό εξέταση, για κάποια από τις επιβλεπόμενες νόσους. Παρόλα αυτά, έτσι τα αποτελέσματα μπορούν να διαστρεβλωθούν, διότι είναι πιθανότερο να επιλεγούν ασθενείς που είναι ήδη άρρωστοι, μικρά παιδιά ή ηλικιωμένοι που πάσχουν από τη νόσο. Κατά αυτόν τον τρόπο, οι αναλογίες που προκύπτουν δεν είναι αντιπροσωπευτικές για το γενικότερο πληθυσμό. Μια καλύτερα προσέγγιση είναι αυτή της τυχαίας δειγματοληψίας, η οποία όμως είναι δύσκολα υλοποιήσιμη και ρεαλιστική για ερευνητικούς σκοπούς μόνο.

Έτσι, οι πιο συνηθισμένες στρατηγικές δειγματοληψίας είναι οι εξής:

- *Διαστηματική Δειγματοληψία (Interval Sampling)*

  Επαναληπτική επιλογή κάθε ν-οστού δείγματος από τον τόπο ελέγχου. Απαιτούμενη είναι η πρότερη γνώση του όγκου των επισκέψεων ανά ημέρα καθώς και ο καθημερινός εξωτερικός έλεγχος για σφάλματα. Τις περισσότερες φορές είναι ανέφικτο για μεγάλο δείγμα παρατήρησης.

- *Δειγματοληψία εναλλασσόμενης ημέρας*

  Επιλογή όλων των ασθενών που διαγιγνώσκονται με την υπό εξέταση νόσο, και έχουν εξεταστεί στην κλινική μία συγκεκριμένη μέρα (ή μέρες) της εβδομάδας. Με αυτό τον τρόπο περιορίζεται η διαδικασία συλλογής και αποθήκευσης των εργαστηριακών δειγμάτων σε μία μόνο ημέρα. Η μεροληψία που συσχετίζεται με τις διαφορές στην συμπεριφορά αναζήτησης βοήθειας των ασθενών, εξαλείφεται με την συστηματική εναλλαγή των ημερών της εβδομάδας κατά τις οποίες επιλέγονται να δείγματα. Μια εναλλακτική πρόταση είναι η ακολουθιακή δειγματοληψία, όπου η συλλογή των δειγμάτων πραγματοποιείται ανά βδομάδα και ανά διαδοχικές ημέρες (π.χ. Δευτέρα, Τρίτη και Τετάρτη).

- *Δειγματοληψία τροποποιημένης ευκολίας (Modified convenience sampling*)

  Επιλογή των πρώτων ν ασθενών προς εξέταση που πληρούν τα κριτήρια μιας ασθένειας. Το χρονικό πλαίσιο επιλογής τους, υπόκειται σε συστηματική περιστροφή για

να ληφθούν υπόψιν οι διάφοροι τρόποι και συνήθειες κατά τις οποίες κάποιο άτομο αναζητάει βοήθεια σε έναν γιατρό. Ένα παράδειγμα θα ήταν η επιλογή των πρώτων δύο υποψήφιων νοσούντων το πρωί, το μεσημέρι και το απόγευμα. [9,11,12]

### *1.3.2 Μειονεκτήματα Συμβατικών Μεθόδων Επιτήρησης*

Όλα τα συστήματα επιδημιολογικής παρακολούθησης πάσχουν από κάποιους περιορισμούς, η γνώση των οποίων είναι απαραίτητη για την λήψη εμπεριστατωμένων αποφάσεων όσον αφορά τον τρόπο συλλογής και επεξεργασίας δεδομένων, την ερμηνεία των ευρημάτων και την τοποθέτηση νέων στόχων επιτήρησης. Το κύριο μειονέκτημα αφορά την χρονική πλευρά της επίβλεψης, καθώς μέχρι να καταγραφεί η παρούσα κατάσταση μιας επιδημίας, εκείνη έχει ήδη προχωρήσει σε επόμενο στάδιο, με αποτέλεσμά οι δράσεις που ακολουθεί η πολιτεία να είναι παρωχημένες. Επιπλέον, αν δεν υπάρχουν οι απαραίτητες υποδομές επικοινωνίας και μεταφοράς, ή αν το ιατρικό προσωπικό δεν επαρκεί για την παροχή των απαραίτητων δεδομένων, για παράδειγμα σε μια κατάσταση εκτάκτου ανάγκης όπως είναι μια πανδημία, τότε οι απαραίτητες λειτουργίες των προγραμμάτων επιτήρησης δεν δύνανται να επιτευχθούν. Είναι φανερή η ανάγκη λοιπόν, για τη δημιουργία νέων μεθόδων οι οποίοι είτε θα λειτουργούν συμπληρωματικά με τις συμβατικές, είτε θα τις βελτιώσουν/αντικαταστήσουν [61].

# Κεφάλαιο 2

# Συστήματα επιδημιολογικής επιτήρησης βασισμένα στη χρήση ετερογενών δεδομένων και μεθόδων μηχανικής μάθησης

Οι συμβατικές μέθοδοι επιδημιολογικής επιτήρησης, μαζί με όλους τους συγγενείς της στόχους, υστερούν όσον αφορά την ταχύτητα υλοποίησής τους, εφόσον από το στάδιο της συλλογής και επεξεργασίας των δεδομένων μέχρι την κατάστρωση ενός σχεδίου αντιμετώπισης της επιδημίας, μεσολαβεί ένα χρονικό διάστημα (συνήθως μίας με δύο εβδομάδων) μέσα στο οποίο ο ρυθμός εξάπλωσης ενός ιού ή το πλήθος των γνωστών κρουσμάτων, καθώς ενδεχομένως και η γενετική σύστασή του, να έχουν μεταβληθεί. Έτσι, πλήθος από ερευνητές παγκοσμίως ασχολείται με την ανάπτυξη καινοτόμων μεθόδων παρακολούθησης και πρόβλεψης της εξάπλωσης μίας επιδημίας, οι οποίες βασίζονται αρχικά, όπως θα δούμε και παρακάτω, στην χρήση εναλλακτικών πηγών δεδομένων [34], δηλαδή δεδομένων που δε χρησιμοποιούνται κλασσικά στην επιδημιολογική επίβλεψη και στη συνέχεια, μέσω τεχνικών μηχανικής μάθησης ή μαθηματικής μοντελοποίησης, στην εξαγωγή χαρακτηριστικών μέσω αυτών των δεδομένων τα οποία μπορούν άμεσα να παρέχουν μια εκτίμηση για τα γνωρίσματα της εκάστοτε επιδημίας. Παρακάτω παρατίθενται οι σημαντικότερες από αυτές τις πηγές δεδομένων, σε συνδυασμό με συγγενείς εργασίες οι οποίες αξιοποιούν τη δυναμική των ετερογενών δεδομένων και τις δυνατότητες που παρέχουν οι εξελίξεις στον τομέα της μηχανικής μάθησης.

## *2.1 Δεδομένα Κοινωνικών Δικτύων*

Η πλειοψηφία από τις πλατφόρμες κοινωνικής δικτύωσης όπως το Facebook, το Twitter ή το Instagram, χρησιμοποιούνται ευρέως σαν μέσο για την δημοσίευση ειδήσεων, γεγονότων ή/και τη δημοσίευση αναρτήσεων σχετικά με την προσωπική ή συναισθηματική κατάσταση των χρηστών τους. Έτσι, έχουν διαδραματίσει σημαντικό ρόλο, στην ανάλυση γεγονότων πραγματικού χρόνου και την ταχύτερη πρόβλεψη τάσεων σε τομείς όπως η απεικόνιση της χρήσης των ΜΜΜ στις πόλεις [58], η διαχείριση κρίσεων [59] ή η λήψη επιχειρηματικών αποφάσεων [60]. Όσον αφορά τον τομέα της δημόσιας υγείας, τα κοινωνικά δίκτυα αποτελούν μια αποτελεσματική πηγή πληροφόρησης για την επιδημιολογική επίβλεψη, καθώς παρέχουν πρώιμους ενδείκτες για τις εποχιακές ασθένειες και μπορούν να λειτουργήσουν σαν ανιχνευτές ή προβλέπτες τάσεων [43]. Τα πλεονεκτήματα αυτών των μεθόδων έγκεινται στην ταχύτητα τους και στο κόστος υλοποίησής τους [36].

### *2.1.1 Μοντελοποίηση εξάπλωσης γρίπης βασισμένη στην ανίχνευση κοινωνικών επαφών με χρήση δεδομένων Twitter*

Στην εργασία τους οι Sadilek et al. [37], κατάφεραν να καταγράψουν την μετάδοση μιας μολυσματικής ασθένειας μεταξύ ατόμων στην πόλη της Νέας Υόρκης, και να μοντελοποιήσουν τη σχέση μεταξύ του τρόπου εξάπλωσής της και των συμπτωμάτων που παρουσιάζουν χρήστες του Twitter, τα οποία περιέχονται στις αναρτήσεις τους. Η συλλογή των Tweets βασίστηκε στην επιλογή δραστήριων χρηστών, δηλαδή σε αυτούς που αναρτούν σε σταθερή βάση και στη συνέχεια, με τη χρήση μηχανών διανυσμάτων υποστήριξης, εντόπισαν τις αναρτήσεις που επιβεβαιώνουν πως ένας χρήστης είναι άρρωστος. Λαμβάνοντας υπόψιν το δίκτυο φιλίας του κάθε ατόμου μοντελοποιήθηκε η εξάπλωση μιας ασθένειας, αφού έγινε μια εκτίμηση της φυσικής επαφής μεταξύ άρρωστων και υγειών ατόμων μέσω της online δραστηριότητάς τους. Τα χαρακτηριστικά που εξήχθησαν από κάθε tweet ήταν τα unigrams, bigrams, και trigrams.

### *2.1.2 Πρόβλεψη κρουσμάτων ILI με χρήση δεδομένων Instagram*

Παρόμοια με πριν, χρησιμοποιήθηκε το Instagram. Αυτή τη φορά, οι αναρτήσεις συγκεντρώθηκαν με βάση λέξεις κλειδιά («*βήχας, πυρετός, γρίπη, μυϊκός πόνος, άρρωστος, πονόλαιμος*»), που είναι τα συνηθέστερα συμπτώματα των γριπωδών συνδρομών. Επιπλέον, αναζητήθηκαν αναρτήσεις που περιείχαν εικόνες παρόμοιες με προεπιλεγμένες εικόνες αναφοράς όπως «*συσκευασίες φαρμάκων», «χάπια»* κ.α. μέσω εξαγωγής χαρακτηριστικών με συνελικτικό νευρωνικό δίκτυο και μετρικής ομοιότητας συνημίτονου. Για το τελικό μοντέλο, τα χαρακτηριστικά εισόδου ήταν το πλήθος των αναρτήσεων και ο αλγόριθμος μάθησης ήταν ο XGBoost. Το παράθυρο πρόβλεψης έφτανε μέχρι τις 3 εβδομάδες στο μέλλον [33].

### *2.1.3 Πρόβλεψη κρουσμάτων ILI με χρήση δεδομένων Sina Weibo*

Το Sina Weibo είναι μια πλατφόρμα κοινωνικής δικτύωσης στην Κίνα, παρόμοια με το Twitter. Για να παράγουν ποιοτικές προβλέψεις όσον αφορά τα κρούσματα της γρίπης στην πόλη Chongqing, η οποία χαρακτηρίζεται απο ασυνήθιστα μοτίβα επιδημικής δραστηριότητας, οι Su et al. [65], συνέλεξαν ημερήσιες δημοσιεύσεις από την πλατφόρμα Sina Weibo με βάση 63 λέξεις – κλειδιά, για το διάστημα 2012-2018. Ο αριθμός των ημερήσιων δημοσιεύσεων εισάγονταν σε ένα Self Adaptive AI μοντέλο (SAAIM) το οποίο δημιουργήθηκε μέσω του αλγόριθμου μάθησης XGBoost και του Seasonal Autoregressive Moving Average μοντέλου.

## *2.2 Μετεωρολογικά Δεδομένα*

Οι Lowen et al.[16] και οι Tamerius et al. [15] έδειξαν την επίδραση που μπορεί να έχει η υγρασία, η θερμοκρασία και η ατμοσφαιρική πίεση στα εποχιακά μοτίβα της γρίπης, ενώ οι J.J. Cannel et al. [62] και οι Joan M. et al. [18] ανέφεραν ότι υπάρχει συσχέτιση μεταξύ της συγκέντρωσης βιταμίνης D στο ανθρώπινο σώμα και της ποσότητας μολύνσεων από ILI σε έναν δεδομένο πληθυσμό. Σύμφωνα με τον WHO, η υπεριώδης ακτινοβολία είναι ικανή να διεγείρει την παραγωγή της βιταμίνης D [63], επηρεάζοντας κατά συνέπεια την εποχιακή συμπεριφορά της γρίπης.

Αρκετές έρευνες επίσης, συσχετίζουν τα χαμηλά επίπεδα υγρασίας με την επάνοδο μιας επιδημίας, καθώς ευνοείται η μετάδοση του ιού μέσω του αέρα. Οι χαμηλές θερμοκρασίες από την άλλη, ενισχύουν την τάση των ανθρώπων να συνωστίζονται σε κλειστούς χώρους, πράγμα το οποίο επιταχύνει τους ρυθμούς μετάδοσης μιας ασθένειας.

### *2.2.1 Πρόβλεψη κρουσμάτων ILI με βάση μετεωρολογικούς δείκτες στην ανατολική Κίνα*

Οι Wendong Liu et al. [64] χρησιμοποίησαν την μέθοδο των Τυχαίων Δασών (Random Forests) για να εκτιμήσουν την προβλεπτική ισχύ διάφορων μεταβλητών στο πρόβλημα της επιδημιολογικής επιτήρησης. Μεταβάλλοντας τον αριθμό των δέντρων του μοντέλου και τον αριθμό των τυχαία επιλεγμένων μεταβλητών, συσχέτισαν την μεταβολή στην προγνωστική ακρίβεια του μοντέλου με την μεταβολή στις τιμές της εκάστοτε επιλεγμένης μεταβλητής. Οι μετεωρολογικές μεταβλητές που έλαβαν υπόψιν τους ήταν η βροχόπτωση, η διάρκεια ηλιοφάνειας, η σχετική υγρασία, η ατμοσφαιρική πίεση και η ελάχιστη, μέγιστη και μέση θερμοκρασία.

### *2.2.2 Πρόβλεψη κρουσμάτων κορωνοϊού COVID-19*

Στο πλαίσιο ανάπτυξης ενός μοντέλου πρόβλεψης των κρουσμάτων του κορωνοϊού COVID-19 στην Ινδία, οι M. Mousavi *et al.* [66] επιστράτευσαν την αρχιτεκτονική των αναδρομικών νευρωνικών δικτύων (RNN). Τα σήματα εισόδου των δικτύου ήταν ο ημερήσιος αριθμός των επιβεβαιωμένων κρουσμάτων, η ημερήσια θερμοκρασία και το ημερήσιο ποσοστό υγρασίας, για τα οποία εκτελέστηκε εποχική αποδόμηση μέσω του αλγόριθμου VMD.

## *2.3 Πρόβλεψη κρουσμάτων ILI μέσω του Wikipedia*

Οι Kyle S. Hickmann *et al.* [40] σχεδίασαν ένα γραμμικό μοντέλο, δηλαδή μια εξίσωση που περιγράφει την αλλαγή μιας μεταβλητής σε σχέση με μία άλλη με γραμμικό τρόπο, για να εξάγουν προβλέψεις για τα μελλοντικά κρούσματα ILI. Τα δεδομένα που έλαβαν υπόψιν τους ήταν ο αριθμός των κρουσμάτων ILI μία βδομάδα πριν την εβδομάδα πρόβλεψης και η συχνότητα πρόσβασης σε πέντε άρθρα της Wikipedia («*Human Flu, Influenza, Influenza A virus, Influenza B virus, και Oseltamivir»*), τα οποία έκριναν πως εμφάνιζαν τη μεγαλύτερη συσχέτιση με το πλήθος κρουσμάτων. Η μετρική αξιολόγησης των αποτελεσμάτων τους ήταν η απόσταση Mahalanobis ή M-distance. Ένα προτέρημά της έναντι της MSE είναι πως υπολογίζει τόσο την ακρίβεια της μέσης πρόβλεψης του μοντέλου όσο και την ακρίβεια της διασποράς της γύρω από τη μέση τιμή.

## *2.4 Συνδυασμός ετερογενών πηγών δεδομένων*

### *2.4.1 Πρόβλεψη κρουσμάτων ILI για στρατιωτικούς πληθυσμούς*

Το πρόβλημα που κλήθηκαν να λύσουν οι Svitlana Volkova *et al.* [31], ήταν η πρόβλεψη των κρουσμάτων ILI σε συγκεκριμένες περιοχές στρατιωτικού πληθυσμού, με παράθυρο πρόβλεψης από μία μέχρι τρεις εβδομάδες χρησιμοποιώντας αναδρομικά νευρωνικά δίκτυα και δεδομένα από την πλατφόρμα του Twitter. Επιπλέον, ενσωματώθηκε σαν πληροφορία, ο αριθμός των επισκέψεων στα ιατρικά κέντρα για συμπτώματα ILI, προς τον συνολικό αριθμό επισκέψεων. Αφού εκπαιδεύτηκαν δύο αρχιτεκτονικές δικτύων μακράς-βραχέας μνήμης, ένα για κάθε τύπο δεδομένων, οι έξοδοί τους οδηγήθηκαν σε έναν πλήρως συνδεδεμένο νευρώνα εξόδου, ο οποίος παρήγαγε την τελική πρόβλεψη για κάθε περιοχή ενδιαφέροντος. Τα χαρακτηριστικά που εξήχθησαν από τις αναρτήσεις του Twitter ήταν τα εβδομαδιαία n-grams, tf-idf score, LDA Topics και text-embeddings.

### *2.4.2 Συνδυασμός δεδομένων αναζήτησης, κοινωνικών δικτύων και παραδοσιακών πηγών*

Οι Santillana *et al.* [29] αξιοποίησαν, όπως οι παραπάνω, για ένα παράθυρο πρόβλεψης μέχρι τεσσάρων εβδομάδων, τις αναρτήσεις χρηστών στο *Twitter* και επιδημιολογικά δεδομένα που τους παρείχε το *Flu Near You*, μια πλατφόρμα επιτήρησης η οποία βασίζεται στην εθελοντική συμμετοχή χρηστών που δηλώνουν σε εβδομαδιαία βάση ποια είναι η

κατάσταση της υγείας τους. Με αυτόν τον τρόπο, δημιουργείται μια γεωγραφική αναπαράσταση των κρουσμάτων ILI, η οποία είναι εξαιρετικά χρήσιμη στην αποτροπή δυνητικών πανδημιών. Σε συνδυασμό αυτές τις πηγές, ενσωματώθηκαν οι καταγραφές επισκέψεων στα νοσοκομεία της Αμερικής, μέσω του *athenahealth*, καθώς και οι αναζητήσεις που περιείχαν συγκεκριμένους όρους στην μηχανή αναζήτησης Google [42]. Οι αλγόριθμοι πρόβλεψης που χρησιμοποιήθηκαν ήταν οι μηχανές διανυσμάτων υποστήριξης, παλινδρόμηση με δέντρα απόφασης και γραμμική παλινδρόμηση, οι οποίοι στη συνέχεια συνδυάστηκαν με την προσέγγιση συνεργατικής μάθησης (ensemble learning).

# Κεφάλαιο 3

# Θεωρητικό Υπόβαθρο

Το σύνολο των τεχνικών που χρησιμοποιήθηκαν στην παρούσα εργασία, καθώς και οι δείκτες αξιολόγησης της επίδοσης των μεθόδων που εφαρμόστηκαν, προέρχονται από τον χώρο της μηχανικής μάθησης, της επεξεργασίας φυσικής γλώσσας και της στατιστικής.

## *3.1 Επεξεργασία Κειμένου*

Η προεπεξεργασία δεδομένων κειμένου αποσκοπεί στην απεικόνισή τους σε αριθμητική μορφή μέσω κατάλληλων μετασχηματισμών, οι οποίοι ανήκουν στον τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing – NLP).

### *3.1.1 Σύνολα λέξεων (Bag of Words)*

Το μοντέλο του συνόλου λέξεων, είναι μια απλουστευτική απεικόνιση δεδομένων κειμένου, η οποία χρησιμοποιείται ευρέως στους τομείς της Μηχανικής Μάθησης, της επεξεργασίας φυσικής γλώσσας και της ανάκτησης πληροφορίας (Information Retrieval). Η κατηγοριοποίηση ενός κειμένου, είναι ένα από τα προβλήματα στα οποία είναι δημοφιλής αυτή η μέθοδος.

Η κεντρική ιδέα, είναι πως ξεχωριστά κείμενα ή documents είναι όμοια αν έχουν παρόμοιο περιεχόμενο. Η νέα αναπαράσταση του κειμένου λοιπόν, περιέχει ένα λεξιλόγιο, το οποίο είναι το σύνολο των λέξεων που εμφανίζονται στο δείγμα κειμένου, και το πλήθος αυτών. Η λέξη «Bag» οφείλεται στο ότι η μοναδική πληροφορία που κρατιέται είναι ο αριθμός εμφανίσεων της κάθε λέξης, και κάθε άλλη πληροφορία, όπως η γραμματική ή συντακτική δομή απορρίπτεται. Συνοπτικά, ο αλγόριθμος «Bag-of-words» είναι ο εξής :

- ***Συλλογή δεδομένων***

Έστω ότι συλλέγονται οι εξής τρεις προτάσεις : «Ο σκύλος κάθισε», «Ο σκύλος κάθισε με τη γάτα», «Ο σκύλος σηκώθηκε κι έφυγε μακριά». Κάθε πρόταση θεωρείται ξεχωριστό δείγμα κειμένου document, και το σύνολο των documents το σώμα κειμένου (corpus) μας.

- ***Δημιουργία λεξιλογίου***

Δημιουργείται το σύνολο του λεξιλογίου το οποίο περιέχει 10 μοναδικές λέξεις, από ένα corpus 15 λέξεων.

- ***Δημιουργία διανυσμάτων κειμένου***

Κάθε document μετατρέπεται σε ένα διάνυσμα 10 στοιχείων, με κάθε στοιχείο το πλήθος της εκάστοτε λέξης στο συγκεκριμένο document, και η τελική του μορφή (για την τρίτη πρόταση) είναι [1,1,0,0,0,0,1,1,1,1].

Ωστόσο, όταν η συλλογή των κειμένων είναι πολύ μεγάλη, το λεξιλόγιο και κατ' επέκταση η διάσταση του λεξιλογίου είναι τεράστια, με αποτέλεσμα τη δημιουργία αραιών πινάκων (δηλαδή πινάκων με πολλά μηδενικά στοιχεία και ελάχιστες μονάδες), οι οποίοι είναι ακριβοί ως προς την διαχείρισή τους από πόρους όπως η υπολογιστική ισχύς και η μνήμη.

### *3.1.2 Μοντέλο N-grams*

Ένα ακόμη μειονέκτημα της παραπάνω μεθόδου είναι, όπως αναφέρθηκε, το γεγονός ότι δεν λαμβάνεται υπόψιν η σύνταξη ή η σειρά με την οποία εμφανίζονται οι λέξεις, κι έτσι χάνεται χρήσιμη πληροφορία. Για παράδειγμα, στο corpus «Ο Γιάννης απολαμβάνει να βλέπει ταινίες, αλλά και η Μαρία απολαμβάνει να βλέπει ταινίες.» η αναπαράσταση μέσω Bag-of-words δεν θα κατανοήσει πως το ρήμα «απολαμβάνω» ακολουθεί πάντα ένα όνομα. Έτσι, σαν λύση, το μοντέλο n-grams, ομαδοποιεί τις λέξεις κατά n, και στο παραπάνω παράδειγμα η εξαγόμενη πληροφορία θα είναι : [«Ο Γιάννης», «Γιάννης απολαμβάνει», «απολαμβάνει να», … , «η Μαρία», «Μαρία απολαμβάνει», «απολαμβάνει να»…].

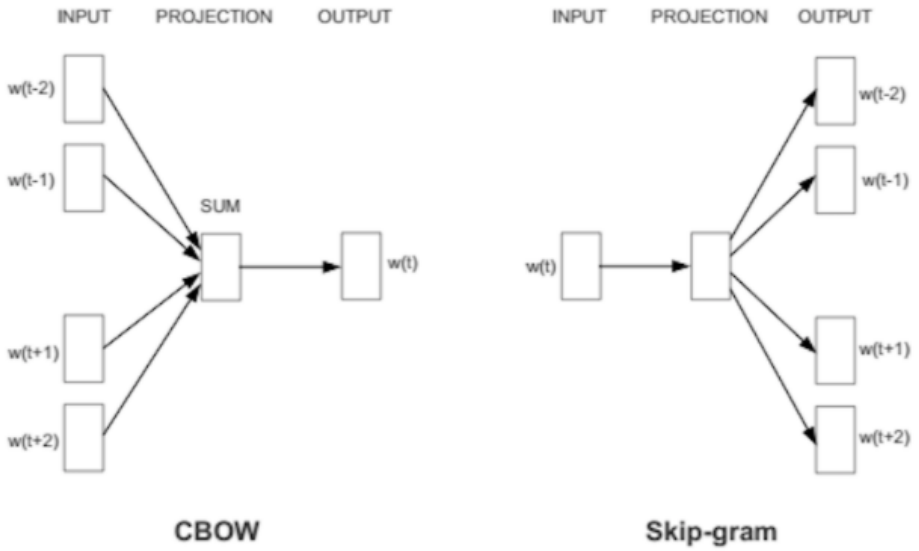
### *3.1.3 Αλγόριθμος Word-2-Vec*

Οι Mikolov *et al.* [27] παρουσίασαν έναν από τους πιο δημοφιλείς τρόπους αναπαράστασης λέξεων, μέσω του αλγορίθμου Word-2-Vec. Πρόκειται για ένα νευρωνικό δίκτυο το οποίο επεξεργάζεται ένα σώμα κειμένου (corpus) και μετατρέπει κάθε λέξη σε ένα διάνυσμα, με τέτοιο τρόπο έτσι ώστε παρόμοιες λέξεις να γειτνιάζουν στον n-διάστατο διανυσματικό χώρο. Αν υπάρχει αρκετή πληροφορία για τον τρόπο με τον οποίο χρησιμοποιείται μια λέξη ή για το συγκείμενο που την περιβάλει, ο Word2vec μπορεί να εκτιμήσει με μεγάλη ακρίβεια το νόημά της, βασιζόμενος σε προηγούμενες εμφανίσεις της στο corpus εκπαίδευσης και, εγκαθιστά μια συσχέτιση μεταξύ λέξεων (π.χ. η λέξη «άντρας» είναι για τη λέξη «αγόρι» ό,τι είναι η λέξη «γυναίκα» για το «κορίτσι»). Μια ακόμη χρησιμότητα είναι η ομαδοποίηση δειγμάτων κειμένου και η ταξινόμησή τους ανά θέμα.

Ο Word-2-vec είναι ένα νευρωνικό δίκτυο δύο επιπέδων το οποίο δημιουργεί διανύσματα, που είναι κατανεμημένες αριθμητικές αναπαραστάσεις των λεξιλογικών χαρακτηριστικών

του corpus, όπως το συγκείμενο που παρουσιάζει κάθε λέξη. Το κρυφό επίπεδο περιέχει γραμμικές συναρτήσεις ενεργοποίησης και το επίπεδο εξόδου τη συνάρτηση *softmax.* Χωρίς την παρέμβαση του προγραμματιστή, το δίκτυο εκπαιδεύεται από τις γειτονικές σχέσεις λέξεων, με τον υπολογισμό της συχνότητας συνεμφάνισής (co-occurence) τους και προσθήκη των αριθμητικών διανυσμάτων σε ένα συμπαγές λεξιλογικό διάνυσμα, το οποίο περιέχει τιμές πιθανοτήτων για ομοιότητα και συσχέτιση μεταξύ των λέξεων. Η ανάλυση ομοιότητας των τελικών διανυσμάτων πραγματοποιείται μέσω της μετρικής ομοιότητας συνημίτονου.

Οι δύο βασικές μέθοδοι δημιουργίας των αναπαραστάσεων είναι η CBOW (Continuous Bag of Words), όπου τα συμφραζόμενα προβλέπουν την λέξη στόχο, και η Skip-gram, η οποία μέσω μιας λέξης, εξάγει το ζητούμενο συγκείμενό της. Για σύνολα δεδομένων μεγάλου όγκου, η Skip-gram παρέχει μεγαλύτερη ακρίβεια.

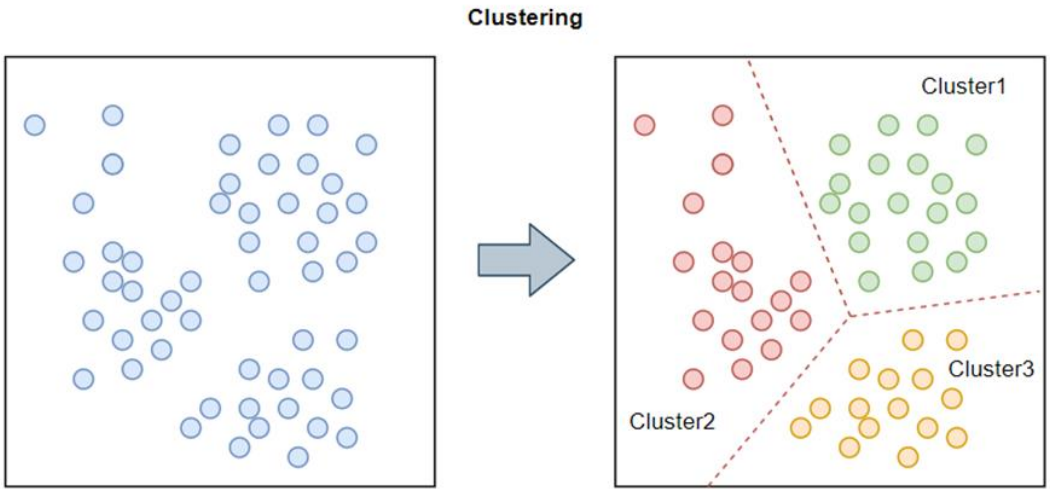



*Σχήμα 3.1 Μέθοδοι Continuous Bag of Words και Skip-gram*

## *3.2 Αλγόριθμοι ομαδοποίησης*

Ομαδοποίηση (clustering) ονομάζεται η διαίρεση του πλήθους ενός συνόλου δεδομένων σε ομάδες (ή ομάδες) έτσι ώστε τα στοιχεία μιας ομάδας να έχουν μεταξύ τους μεγαλύτερο δείκτη ομοιότητας απ’ ότι με τα στοιχεία άλλων ομάδων. Με άλλα λόγια, πραγματοποιείται ένας διαχωρισμός των δεδομένων, με βάση τα χαρακτηριστικά του κάθε δείγματος. Η ομαδοποίηση των δεδομένων μπορεί να είναι είτε αυστηρή, εννοώντας πως κάθε δείγμα ανήκει αποκλειστικά σε μία και μόνο ομάδα, είτε χαλαρή, δηλαδή αντί να πραγματοποιείται ανάθεση μιας ομάδας σε κάθε δείγμα, αντ’ αυτού προσδίδεται μια πιθανότητα να ανήκει το δείγμα για κάθε ομάδα. Η ομαδοποίηση βρίσκεται επίσης, ανάμεσα στις πιο δημοφιλείς τεχνικές μη επιβλεπόμενης μάθησης, καθώς ανακαλύπτει υποκείμενα μοτίβα στο σύνολο

των δεδομένων και τα χωρίζει σε ομάδες δίχως να χρειάζεται κάποια μορφή ετικετών ή προβλέψεων (ground truth).




*Σχήμα 3.2 Διαχωρισμός σημείων σε ομάδες*

### *3.2.1 K-means ομαδοποίηση*

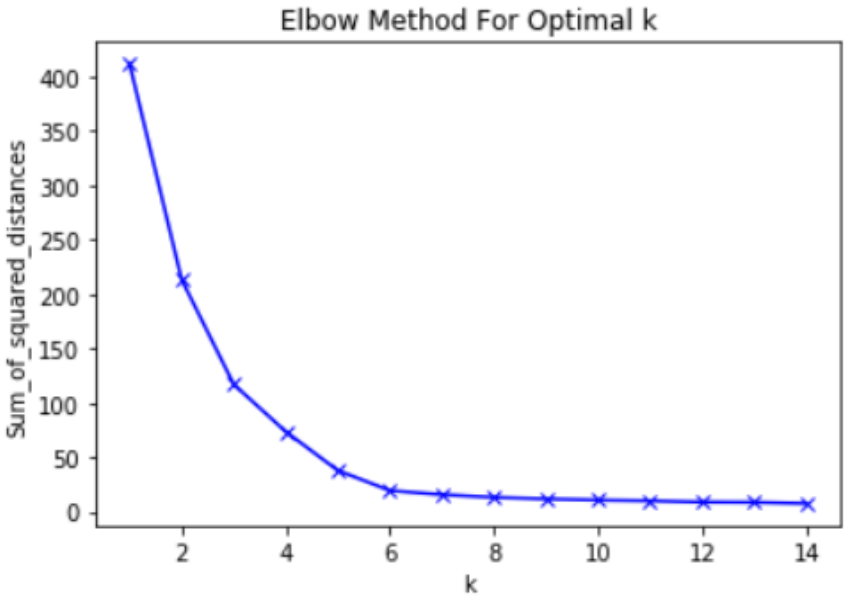
Ο αλγόριθμος αυτός βασίζεται στην ιδέα των κεντροειδών, δηλαδή των μέσων σημείων κάθε ομάδας δειγμάτων. Περιλαμβάνει τα εξής βήματα:

1. Επιλογή του αριθμού K των ομάδων τις οποίες θέλουμε να ανακαλύψουμε
2. Τυχαία παραγωγή K σημείων στον χώρο των δειγμάτων, που θα είναι τα κεντροειδή των αρχικών ομάδων.
3. Μέτρηση της απόστασης κάθε σημείου από τα κεντροειδή και ανάθεση σε κάθε σημείο το κοντινότερο κεντροειδές του και την αντίστοιχη ομάδα.
4. Επανυπολογισμός του μέσου σημείου (κεντροειδές) κάθε ομάδας.
5. Επανάληψη των βημάτων τρία και τέσσερα μέχρι να :
   α. Σταθεροποιηθούν τα κεντροειδή, δηλαδή αφού υπολογιστούν ξανά τα μέσα σημεία, να μην γίνεται εκ νέου ανάθεση ομάδας σε κανένα σημείο ή
   β. Ο προκαθορισμένος αριθμός επαναλήψεων να εξαντληθεί.

Η επιλογή του αριθμού K των ομάδων γίνεται με τέτοιο τρόπο έτσι ώστε να μειωθεί ικανοποιητικά η απόκλιση των σημείων κάθε ομάδας. Ο υπολογισμός της μπορεί να γίνει κοιτώντας το άθροισμα των τετραγωνισμένων αποστάσεων κάθε σημείου από το κεντροειδές του. Αν έχουμε μόνο μια ομάδα τότε η απόκλιση θα είναι υψηλή και, αντίστοιχα,

αν κάθε σημείο αποτελεί τη δική του ομάδα τότε θα είναι μηδενική. Έτσι, όσο αυξάνεται ο αριθμός Κ, τόσο αυτή μειώνεται.

Μια εμπειρική μέθοδος επιλογής του Κ, είναι η κατασκευή του elbow plot, απεικόνιση της απόκλισης σε συνάρτηση με το Κ. Εκεί όπου εμφανίζεται το σημείο καμπής, ή αλλιώς «αγκώνας», είναι ο βέλτιστος αριθμός ομάδων καθώς από εκεί και πέρα, περαιτέρω ομάδες δεν δικαιολογούν το επιπλέον κέρδος σε απόκλιση.




*Σχήμα 3.3 Elbow Method*

### *3.2.2 Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)*

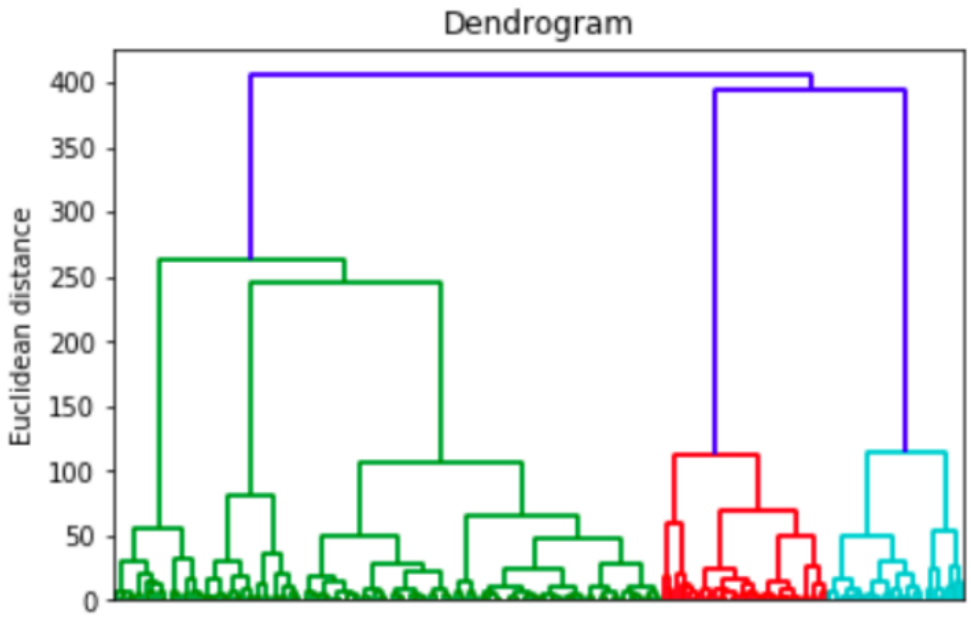
Η ανάθεση των σημείων σε ομάδες σε αυτή τη μέθοδο, γίνεται βάσει των ιεραρχικών σχέσεων μεταξύ των σημείων. Υπάρχουν δύο τύποι ιεραρχικής ομαδοποίησης, η από πάνω-προς-τα-κάτω και η από κάτω-προς-τα-πάνω. Ο δεύτερος είναι ο συνηθέστερος, και μαθηματικά πιο εύκολος να υπολογιστεί.

*Αλγόριθμος Ιεραρχικής Ομαδοποίησης*

1. Ανάθεση κάθε σημείου στην δική του ομάδα, μέχρι ο αριθμός Κ να ισούται με το πλήθος Ν των σημείων.
2. Υπολογισμός των αποστάσεων μεταξύ κάθε ομάδας.
3. Ένωση των δύο πιο όμοιων ομάδων.
4. Επανάληψη των βημάτων δύο και τρία μέχρις ότου όλα τα σημεία να βρεθούν στην ίδια ομάδα.

Ο υπολογισμός της ομοιότητας μεταξύ των ομάδων γίνεται με διάφορους τρόπους, εκ των οποίων τρεις είναι η απόλυτη σύνδεση, με χρήση της μέγιστης απόστασης μεταξύ οποιωνδήποτε δύο σημείων για κάθε υποψήφια ομάδα, η μονή σύνδεση με χρήση της ελάχιστης απόστασης μεταξύ των δύο σημείων και η μέση σύνδεση, με χρήση της μέσης απόστασης κάθε σημείου μιας ομάδας με όλα τα σημεία της άλλης ομάδας.

Η μετρική για τον υπολογισμό της απόστασης είναι συνήθως η ευκλείδεια απόσταση ενώ η επιλογή του αριθμού K μπορεί να γίνει πάλι με το elbow plot. Η απεικόνιση των ιεραρχικών σχέσεων μπορεί να γίνει με ένα δενδρόγραμμα.




*Σχήμα 3.4 Δενδρόγραμμα ιεραρχικής ομαδοποίησης*

### *3.2.3 Φασματική Ομαδοποίηση (Spectral Clustering)*

Αν υποθέσουμε ότι τα δείγματα προς ομαδοποίηση είναι κόμβοι σε ένα γράφο *G*, τότε η φασματική μέθοδος χρησιμοποιεί τη συνδεσιμότητα μεταξύ των κόμβων *V* για να βρει τα δείγματα που βρίσκονται το ένα δίπλα στο άλλο. Στη συνέχεια, οι κόμβοι απεικονίζονται σε έναν χώρο χαμηλής διαστατικότητας και μπορούν εύκολα να διαχωριστούν για να σχηματιστούν οι ομάδες τους. Για να επιτευχθεί η παραπάνω διαδικασία, είναι απαραίτητες οι ιδιοτιμές συγκεκριμένων πινάκων, που εξάγονται από τον γράφο *G*.

Οι φασματικές μέθοδοι είναι εύκολες στην υλοποίηση και σχετικά γρήγορες για αραιά σύνολα δεδομένων μέχρι μερικών χιλιάδων. Επιπλέον, δεν πραγματοποιούνται καθόλου υποθέσεις για τη μορφή ή την κατανομή που έχουν οι ομάδες, όπως για παράδειγμα ο αλγόριθμος K-means που υποθέτει πως τα σημεία μιας ομάδας είναι σφαιρικά κατανεμημένα γύρω από τα κεντροειδή τους. Μια ακόμη ενδιαφέρουσα ιδιότητα της

φασματικής μεθόδου είναι πως μπορεί να κατανέμει σωστά στην ίδια ομάδα μακρινά σημεία, μέσω της μείωσης της διαστατικότητας.

- **Γράφοι Ομοιότητας**

Δεδομένου ενός συνόλου σημείων δεδομένων $x_1, \dots, x_n$ και κάποιας έννοιας ομοιότητας $s_{ij} \geq 0$ μεταξύ όλων των ζευγών σημείων $(x_i, x_j)$, ο στόχος της ομαδοποίησης είναι να χωριστούν τα σημεία σε διάφορες ομάδες έτσι ώστε τα σημεία στην ίδια ομάδα να είναι παρόμοια και σημεία σε διαφορετικές ομάδες είναι διαφορετικά μεταξύ τους. Ένας καλός τρόπος αναπαράστασης των δεδομένων είναι με τη μορφή του γραφήματος ομοιότητας $G(V, E)$. Κάθε κορυφή σε αυτό το γράφημα αντιπροσωπεύει ένα σημείο δείγματος $x_i$. Δύο κορυφές συνδέονται εάν η ομοιότητα $s_{ij}$ μεταξύ των αντίστοιχων σημείων $(x_i, x_j)$ είναι θετική ή μεγαλύτερη από ένα συγκεκριμένο κατώφλι και η ακμή θα σταθμίζεται με βάρος $s_{ij}$. Το πρόβλημα της ομαδοποίησης μπορεί τώρα να αναδιατυπωθεί χρησιμοποιώντας το γράφημα ομοιότητας: θέλουμε να βρούμε μια διαμέριση του γραφήματος έτσι ώστε οι ακμές μεταξύ διαφορετικών ομάδων να έχουν πολύ χαμηλά βάρη (που σημαίνει ότι τα σημεία σε διαφορετικές ομάδες είναι διαφορετικά μεταξύ τους) και οι ακμές μέσα σε μια ομάδα έχουν υψηλά βάρη (που σημαίνει ότι τα σημεία μέσα στο ίδιο σύμπλεγμα είναι παρόμοια μεταξύ τους).

- **Πίνακες συγγένειας και γειτνίασης (Adjacency and Affinity Matrix)**

Ο γράφος *G* μπορεί να αναπαρασταθεί ως ένας πίνακας γειτνίασης *A* , όπου οι δείκτες των σειρών και των στηλών είναι οι κόμβοι και οι τιμές των κελιών δηλώνουν την παρουσία ή την απουσία ακμής μεταξύ δύο σημείων, με τις τιμές ένα ή μηδέν αντίστοιχα. Ο πίνακας συγγένειας αντί να δηλώνει την ύπαρξη ή μη ακμών μεταξύ σημείων, οι τιμές στα κελιά του είναι ο βαθμός ομοιότητας μεταξύ δύο σημείων (με εύρος από μηδέν έως ένα). Ουσιαστικά κρατάει τα βάρη των ακμών του γράφου.

- **Πίνακας Βαθμού Degree Matrix**

Είναι ένας διαγώνιος πίνακας *D*, με την διαγώνιο να κρατάει τον αριθμό των ακμών που καταλήγουν σε κάθε κόμβο. Μπορεί να προκύψει λαμβάνοντας το άθροισμα κάθε γραμμής στον πίνακα γειτνίασης.

- **Λαπλασιανός Πίνακας Laplacian Matrix**

Προκύπτει από την αφαίρεση του *A* από τον *D* : $L = D - A$

**Αλγόριθμος κανονικοποιημένης φασματοποίησης**

Ο αλγόριθμος έχει προταθεί από τους Ng *et al* [28]. Δέχεται ως είσοδο έναν πίνακα ομοιότητας $S \in n \times n$, και τον αριθμός *k* ομάδων προς κατασκευή και εκτελεί τα παρακάτω βήματα:

1. Κατασκευή γραφήματος ομοιότητας και έστω *W* ο σταθμισμένος πίνακας γειτνίασης
2. Υπολογισμός κανονικοποιημένου Laplacian *L*
3. Υπολογισμός των πρώτων *k* ιδιοδιανυσμάτων $u_1, \dots, u_k$ του *L*
4. Έστω $U \in n \times k$πίνακας που περιέχει τα διανύσματα $u_1, \dots, u_k$ως στήλες
5. Σχηματισμός του πίνακα $T \in n \times k$ από τον *U* κανονικοποιώντας τις σειρές βάσει του τύπου

$$t_{ij} = u_{ij} / \left(\sum_j u_{ij}^2\right)^{1/2} \qquad (3.1)$$

6. Για $i = 1, \dots, n$ έστω το διάνυσμα $y_i \in k$που αντιστοιχεί στη i-οστή σειρά του *T*
7. Ομαδοποίηση των σημείων $y_i$ με τον αλγόριθμο K-means σε ομάδες $C_1,\dots,C_k$

Έξοδος: Ομάδες $A_1, \dots, A_k$ με $A_i = j \vee y_j \in C_i$

## *3.3 Εποχική Αποδόμηση*

Κατά την περιγραφή χρονοσειρών μπορούμε να χρησιμοποιήσουμε τους παρακάτω όρους οι οποίοι αντιστοιχούν σε διαφορετικά μοτίβα συμπεριφοράς.

- **Συνιστώσα τάσης**

Σαν τάση ορίζεται μια μακροπρόθεσμη αύξηση ή μείωση στην τιμή των δεδομένων η οποία δεν είναι απαραίτητα γραμμική. Είναι πιθανό επίσης, να υπάρχει αλλαγή κατεύθυνσης στην τάση αυτή όταν μεταβαίνει από αυξανόμενη σε μια φθίνουσα και αντίστροφα. Για παράδειγμα, στο σχήμα 3.15 φαίνεται πως υπάρχει μια αυξητική τάση στον αριθμό πωλήσεων των αντιδιαβητικών φαρμάκων.

- **Εποχική συνιστώσα**

Ένα εποχιακό μοτίβο εμφανίζεται όταν μια χρονοσειρά επηρεάζεται από εποχιακούς παράγοντες, όπως η ώρα του έτους ή η ημέρα της εβδομάδας. Η εποχικότητα είναι πάντα σταθερή και γνωστή. Οι μηνιαίες πωλήσεις αντιδιαβητικών φαρμάκων παρακάτω δείχνουν εποχικότητα που προκαλείται εν μέρει από την αλλαγή στο κόστος των φαρμάκων στο τέλος του ημερολογιακού έτους.

- **Κυκλική συνιστώσα**

Ένας κύκλος συμβαίνει όταν το ύψος τιμών αυξομειώνεται σε μη σταθερή συχνότητα. Το μέσο μήκος ενός κύκλου τείνει να είναι μεγαλύτερο και πιο ευμετάβλητο από το μήκος του εποχιακού κύκλου.

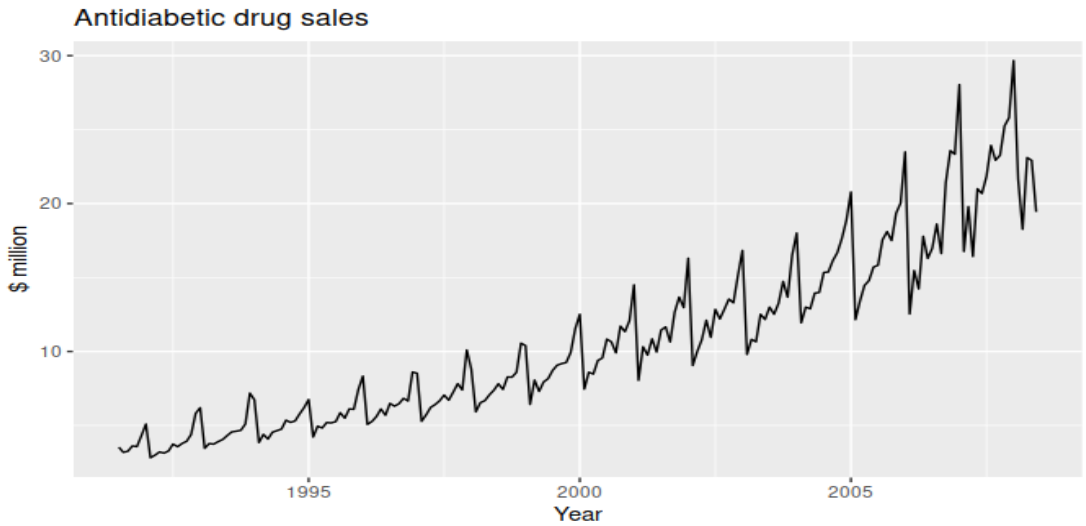
Πολλές χρονοσειρές περιλαμβάνουν τάση, κύκλους και εποχικότητα. Κατά την επιλογή μιας μεθόδου πρόβλεψης, θα πρέπει πρώτα να προσδιοριστούν τα μοτίβα χρονοσειρών στα δεδομένα και στη συνέχεια, να επιλεγεί μια μέθοδος που μπορεί να καταγράψει σωστά τα μοτίβα αυτά. Όταν αποδομείται μια χρονοσειρά στα επιμέρους κομμάτια της, η τάση και οι κύκλοι της συνήθως συνδυάζονται, κι έτσι, μπορεί η σειρά να εκφραστεί σαν ένα άθροισμα ή γινόμενο τριών συνιστωσών, *εποχής (S), τάσης-κύκλου (T)* και *θορύβου (R).*

*Αθροιστική αποδόμηση :* $y_t = S_t + T_t + R_t$ *(3.2)*

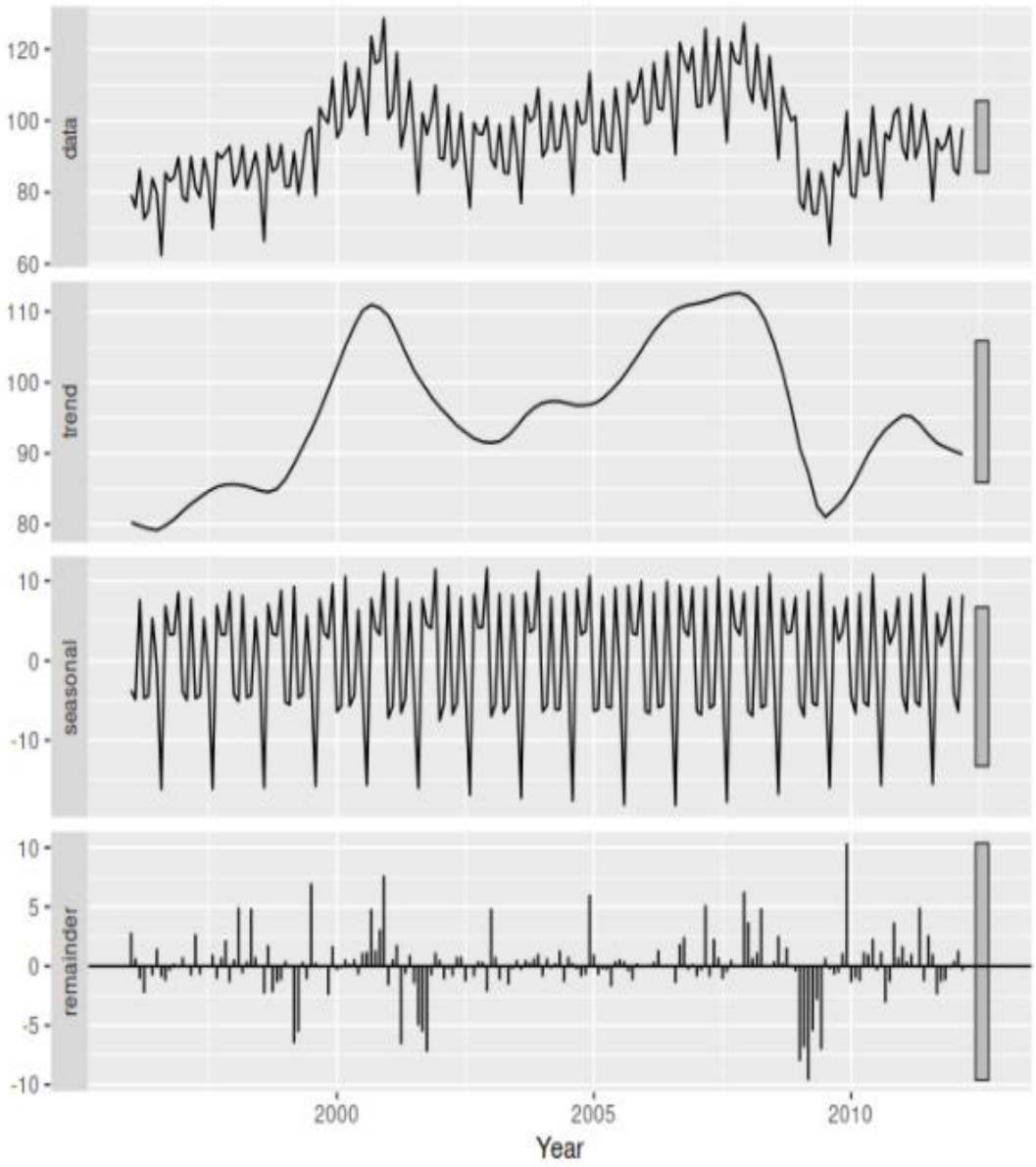
*Πολλαπλασιαστική αποδόμηση :* $y_t = S_t \times T_t \times R_t$ *(3.3α)*

*ή*

$log(y_t) = log(S_t) + log(T_t) + log(R_t)$ *(3.3b)*




*Σχήμα 3.5 Πωλήσεις αντιδιαβητικών φαρμάκων στην Αμερική*




*Σχήμα 3.6 Ανάλυση χρονοσειράς στις συνιστώσες τάσης, εποχικότητας και θορύβου*

### *3.3.1 Εκτίμηση της συνιστώσας Τάσης-Κύκλου*

Το πρώτο βήμα της κλασσικής μεθόδου εποχιακής αποδόμησης, είναι η χρήση της μεθόδου κινούμενου μέσου (Moving Average-MA), για την εκτίμηση της συνιστώσας Τάσης-Κύκλου (Trend-Cycle). Η μέθοδος κινούμενου μέσου τάξεως *m* ορίζεται ως

$$\hat{T}_t = \frac{1}{\sum_{j=-k}^{k} y_{t+j}} \qquad (3.4)$$

όπου $m = 2k + 1$.

Ένα πλεονέκτημα αυτής της μεθόδου, είναι η εξάλειψη της τυχαιότητας των δεδομένων, διότι δείγματα που είναι κοντά χρονικά, είναι πιθανό να έχουν παρόμοιες τιμές. Μεγαλύτερη τάξη *m*, επίσης, συνεπάγεται μια πιο ομαλή καμπύλη τάσης, ενώ επιλέγεται να είναι περιττή έτσι ώστε η πράξη να είναι συμμετρική. Σε περίπτωση που έχουμε άρτια τάξη *m*, μπορεί να εφαρμοστεί επιπλέον κινούμενος μέσος άρτιας τάξης για να μετατραπεί σε συμμετρικό το τελικό αποτέλεσμα. Αν η περίοδος του σήματος είναι άρτια και τάξης *m*, εφαρμόζουμε ένα 2 x m-MA, για να εκτιμηθεί η συνιστώσα τάσης, ενώ αν η περίοδος είναι περιττή, εφαρμόζουμε ένα m-MA. Κάθε όρος του $T_t$ έχει ίσο βάρος με τους άλλους, καθώς ο πρώτος και ο τελευταίος ανήκουν σε διαδοχικά παράθυρα. Για παράδειγμα, παρακάτω φαίνεται ένας 2x4-MA:

$$\begin{aligned}\hat{T}_t &= \frac{1}{2}\left[\frac{1}{4}(y_{t-2}+y_{t-1}+y_t+y_{t+1})+\frac{1}{4}(y_{t-1}+y_t+y_{t+1}+y_{t+2})\right] \\ &= \frac{1}{8}y_{t-2}+\frac{1}{4}y_{t-1}+\frac{1}{4}y_t+\frac{1}{4}y_{t+1}+\frac{1}{8}y_{t+2}\end{aligned}$$

### *3.3.2 Αλγόριθμος εποχικής αποδόμησης*

Στην κλασσική μέθοδο εποχικής αποδόμησης, η εποχιακή συνιστώσα λαμβάνεται ως σταθερή από έτος σε έτος. Ακολουθεί ο αλγόριθμος της αθροιστικής αποδόμησης, ο οποίος είναι η βάση των περισσότερων πιο εξελιγμένων μεθόδων εποχικής αποδόμησης.

1. Επιλογή περιόδου χρονοσειράς m. Αν m άρτια, εφαρμογή 2 x m-MA για συνιστώσα τάσης αλλιώς εφαρμογή m-MA
2. Υπολογισμός αποτασιοποιημένης (detrended) σειράς :
$$\hat{y} = y - T \qquad (3.5)$$
3. Για να υπολογιστεί η εποχιακή συνιστώσα για κάθε εποχή, υπολογίζεται ο μέσος όρος των τιμών της detrended σειράς για αυτή την εποχή. Για παράδειγμα, αν οι τιμές της σειράς είναι μηνιαίες, η εποχιακή τιμή της σειράς για το Μάρτιο, είναι ο μέσος όρος όλων των τιμών της σειράς για κάθε Μάρτιο. Στη συνέχεια κανονικοποιούνται οι τιμές αυτές των συνιστωσών έτσι ώστε το άθροισμά τους να είναι μηδενικό. Η τελική συνιστώσα είναι η συρραφή των τιμών για κάθε μήνα, και η επανάληψή της για κάθε χρόνο του συνόλου δεδομένων.
4. Η συνιστώσα θορύβου υπολογίζεται ως εξής :
$$R = y - T - S \qquad (3.6)$$

## *3.4 Μηχανική Μάθηση*

Το πεδίο της επιστήμης υπολογιστών στο οποίο αναπτύσσονται υπολογιστικά συστήματα που βελτιώνουν την απόδοσή *P* σε ένα πρόβλημα *T*, όσο συλλέγουν εμπειρία *E* , ονομάζεται μηχανική μάθηση [48]. Η εμπειρία εδώ χρησιμοποιείται για να περιγράψει τα δεδομένα που επεξεργάζεται ένα πρόγραμμα για να εκμεταλλευτεί την πληροφορία που περιέχουν. Πρόκειται για έναν υποκλάδο της τεχνητής νοημοσύνης, στον οποίο όμως ο σχεδιασμός του αλγορίθμου για το εκάστοτε πρόβλημα, πραγματοποιείται από τον ίδιο τον υπολογιστή, κι όχι τον προγραμματιστή [52]. Ένα πρόβλημα *T* μηχανικής μάθησης, είναι η αναγνώριση εικόνων από ένα πρόγραμμα. Εδώ, η εμπειρία *E* θεωρείται το σύνολο των επισημασμένων εικόνων, και η απόδοση *P* είναι το ποσοστό των σωστά αναγνωρισμένων εικόνων προς ένα σύνολο δοκιμής.

### *3.4.1 Είδη Μηχανικής Μάθησης*

Ανάλογα με τα δεδομένα εισόδου και εξόδου κάθε αλγορίθμου, τον τύπο του κάθε προβλήματος και την μέθοδο που ακολουθείται, υπάρχουν διάφορες κατηγορίες μάθησης, οι σημαντικότερες εκ των οποίων είναι οι εξής:

- ***Επιβλεπόμενη Μάθηση***

  Σκοπός είναι η εξαγωγή μιας συνάρτησης η οποία απεικονίζει μια είσοδο $X$ σε μία έξοδο $Y$ βάσει παραδειγμάτων ζευγαριών εισόδου-εξόδου. Έτσι, για το πρόβλημα της κατηγοριοποίησης δεδομένων σε κάποια κλάση, οι είσοδοι $X$ έχουν επισημανθεί με μία ετικέτα της αληθινής τους κλάσης $Y$, και ο αλγόριθμος μάθησης αναγνωρίζει χαρακτηριστικά της εισόδου βάσει των οποίων κατηγοριοποιεί νέα δεδομένα που δίνονται χωρίς κάποια ετικέτα.

- ***Μη Επιβλεπόμενη Μάθηση***

  Σε αυτή την κατηγορία μάθησης, τα δεδομένα εισόδου δεν συνοδεύονται από κάποια επισήμανση επιθυμητής εξόδου. Αποστολή του αλγόριθμου είναι να αναγνωρίσει πρότυπα που μπορεί να υπάρχουν στο χώρο της εισόδου και με βάσει αυτά να πραγματοποιήσει μια ομαδοποίησή τους.

- ***Ενισχυτική Μάθηση***

  Στα προβλήματα που χρησιμοποιείται η ενισχυτική μάθηση, σκοπός είναι η λήψη αποφάσεων που θα οδηγήσουν στο μέγιστο κέρδος με τις ελάχιστες δυνατές απώλειες. Για παράδειγμα, στο γνωστό πρόβλημα της εύρεσης συντομότερης διαδρομής σε έναν χώρο με εμπόδια, ο αλγόριθμος δοκιμάζει όλες τις δυνατές διαδρομές, με μία επιπλέον

επιβάρυνση κάθε φορά που συναντά κάποιο εμπόδιο, και για κάθε σωστό βήμα ή λάθος βήμα, επιβραβεύεται ή επιβαρύνεται αντίστοιχα, και συνεχίζει να μαθαίνει. Στο τέλος, ο αλγόριθμος θα επιστρέψει την στρατηγική η οποία θα αποφέρει το μεγαλύτερο κέρδος με το μικρότερο κόστος.

### *3.4.2 Τεχνητά Νευρωνικά Δίκτυα*

Ο ανθρώπινος εγκέφαλος λειτουργεί με εντελώς διαφορετικό τρόπο σε σχέση με έναν ψηφιακό υπολογιστή. Μέσω της οργάνωσης των εσωτερικών του δομών, ή αλλιώς νευρώνων, επιτυγχάνει ορισμένες δράσεις όπως την αναγνώριση διαφορετικών μοτίβων στο περιβάλλον του ή τον έλεγχο της κίνησης του σώματος, αρκετές φορές πιο γρήγορα από τον ταχύτερο υπολογιστή. Η ανάπτυξη των (τεχνητών) νευρωνικών δικτύων (Artificial Neural Networks-ANN) βασίστηκε στην μη γραμμικότητα και την παραλληλία εκτέλεσης αυτών των δράσεων, προσομοιώνοντας δύο σημεία κλειδιά αυτής της λειτουργίας. Πρώτον, η απόκτηση της γνώσης από το περιβάλλον γίνεται μέσω μιας διαδικασίας μάθησης, και δεύτερον, οι διασυνδέσεις των νευρώνων, ή αλλιώς τα συναπτικά τους βάρη, αποθηκεύουν αυτή την γνώση. Η διαδικασία κατά την οποία το νευρωνικό δίκτυο μεταβάλλει τα συναπτικά του βάρη, αποθηκεύει γνώση και αποκτά την ικανότητα να γενικεύει ονομάζεται αλγόριθμος μάθησης. Γενίκευση σε αυτή την περίπτωση θεωρείται η παραγωγή λογικών συμπερασμάτων (εξόδων) όταν το δίκτυο βρίσκεται σε μια άγνωστη κατάσταση (είσοδος). Ακολουθούν τα δομικά στοιχεία των τεχνητών νευρωνικών δικτύων.

- **Είσοδος (Input)**

Σαν είσοδοι θεωρούνται οι μετρήσεις που λαμβάνονται από το περιβάλλον, δηλαδή η πληροφορία που θα εισαχθεί στο δίκτυο, υπό μορφή ενός πίνακα ν-στοιχείων.

- **Βάρη (Synaptic Weights)**

Το σύνολο των νευρικών συνάψεων, κάθε μία από τις οποίες χαρακτηρίζεται από το δικό της μέτρο. Ο πολλαπλασιασμός τους με τα στοιχεία της εισόδου, μεταβάλλει την εκάστοτε έξοδο του δικτύου, κι έτσι ανάλογα με την τιμή τους, προσδίδεται μεγαλύτερη ή μικρότερη βαρύτητα σε κάθε στοιχείο εισόδου.

- **Αθροιστής (Summing Function)**

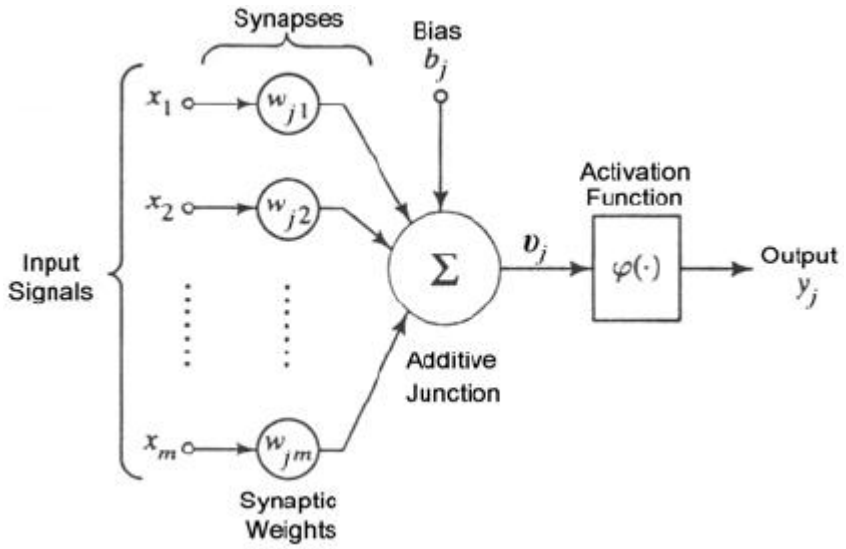
Λήψη των σταθμισμένων από τα βάρη σημάτων εισόδου, και άθροισή τους.

- **Συνάρτηση ενεργοποίησης (Activation Function)**

Λήψη του σήματος του αθροιστή και παραγωγή του σήματος εξόδου του νευρώνα με βάση μια μαθηματική συνάρτηση. Η μη-γραμμικότητα του δικτύου εισάγεται εδώ.

- **Πόλωση (Bias)**

Μία επιπλέον είσοδος στον αθροιστή, η οποία αυξάνει ή μειώνει αντίστοιχα την δικτυακή διέγερση της συνάρτησης ενεργοποίησης.




*Σχήμα 3.7 Δομικές μονάδες ενός τεχνητού νευρώνα*

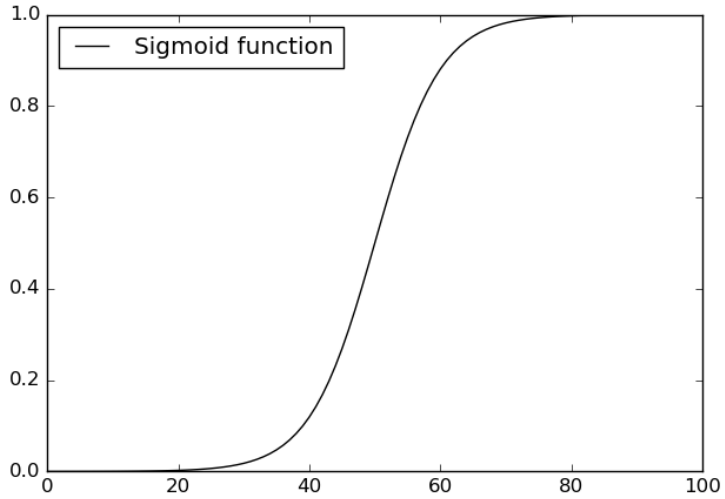
- **Τύποι συναρτήσεων ενεργοποίησης**

Στην πλειοψηφία τους πρόκεινται για μη γραμμικές συναρτήσεις, με σκοπό την αποκωδικοποίηση των δεδομένων εισόδου. Παρατίθενται δύο από τους βασικότερους τύπους συναρτήσεων ενεργοποίησης:

1. *Συνάρτηση Κατωφλίου (threshold function) :* Με βάση αυτή την συνάρτηση, η έξοδος του νευρώνα είναι ίση με 1, αν το σήμα εισόδου της συνάρτησης είναι θετικό, ειδάλλως η έξοδος ισούται με 0.

2. *Σιγμοειδής συνάρτηση* : Από τις πλέον χρησιμοποιούμενες συναρτήσεις στην κατασκευή νευρωνικών δικτύων. Παρουσιάζει αυστηρά αύξουσα συμπεριφορά, και είναι διαφορίσιμη (σε αντίθεση με την συνάρτηση κατωφλίου). Ένα παράδειγμα είναι η λογιστική συνάρτηση, ενώ αν χρειαστεί να λάβει και αρνητικές τιμές αυτή η συνάρτηση, χρησιμοποιείται η υπερβολική εφαπτομένη (tanh). Ο μαθηματικός της τύπος είναι

$$f(x) = \frac{1}{1+e^{-x}} \qquad (3.7)$$




*Σχήμα 3.8 Γραφική παράσταση της σιγμοειδούς συνάρτησης*

- **Αρχιτεκτονικές Νευρωνικών Δικτύων**

Ενός επιπέδου πρόσθιας τροφοδότησης

Η απλούστερη από τις αρχιτεκτονικές νευρωνικών δικτύων, αποτελείται από ένα επίπεδο νευρώνων εισόδου το οποίο συνδέεται άμεσα με το επίπεδο νευρώνων εξόδου, όχι όμως και αντίστροφα. Για αυτόν το λόγο, αυτού του είδους τα δίκτυα χαρακτηρίζονται πρόσθιας τροφοδότησης (feedforward).

Πολυεπίπεδα Δίκτυα Πρόσθιας Τροφοδότησης

Σε αυτή την κατηγορία δικτύων, παρουσιάζονται ένα ή περισσότερα κρυφά επίπεδα νευρώνων μεταξύ των επιπέδων εισόδου και εξόδου. Σκοπός τους είναι η ανίχνευση των χαρακτηριστικών των δεδομένων εκπαίδευσης (feature extraction) μέσω ενός μη-γραμμικού μετασχηματισμού του σήματος της εισόδου σε έναν νέο χώρο που αποκαλείται χώρος χαρακτηριστικών (feature space). Στο παράδειγμα της αναγνώρισης προτύπων, οι κλάσεις των χαρακτηριστικών μπορούν και διαχωρίζονται σε αυτόν τον χώρο από οτιδήποτε άλλο υπήρχε στον χώρο εισόδου.

Αναδρομικά Νευρωνικά Δίκτυα

Σε αντίθεση με τις παραπάνω αρχιτεκτονικές, εδώ η σύνδεση μεταξύ των επιπέδων παρουσιάζει έναν βρόχο ανάδρασης, παράλληλα με βαθμίδες μοναδιαίων καθυστερήσεων επί της εισόδου, τα οποία όλα μαζί παρέχουν την τρέχουσα καθώς και παρελθοντικές τιμές της εισόδου και της εξόδου στο δίκτυο. Η δομή αυτή προσδίδει στην αρχιτεκτονική αυτή την ικανότητα “μνήμης” των παρελθοντικών καταστάσεων στις

οποίες έχει βρεθεί, πράγμα εξαιρετικά χρήσιμο για παράδειγμα, στο πρόβλημα αναγνώρισης ακολουθιών ή πρόβλεψης χρονοσειρών.

- **Συναρτήσεις κόστους**

Αυτή η οικογένεια συναρτήσεων χρησιμεύει στην μέτρηση της απόδοσης ενός μοντέλου μηχανικής μάθησης. Η λειτουργία τους έγκειται στην απεικόνιση ενός συνόλου τιμών μιας τυχαίας μεταβλητής σε έναν αριθμό. Έτσι, μπορεί να ποσοτικοποιηθεί το σφάλμα μεταξύ της επιθυμητής και της προβλεπόμενης τιμής σε ένα πρόβλημα βελτιστοποίησης, και απώτερος σκοπός είναι η προσέγγιση του κατώτερου φράγματος αυτού του αριθμού. Στο πρόβλημα της ταξινόμησης κλάσεων, για παράδειγμα, η συνάρτηση κόστους αποδίδει την ποινή για τη λανθασμένη ταξινόμηση ενός δείγματος. Μία ευρέως χρησιμοποιούμενη συνάρτηση κόστους, είναι αυτή του μέσου τετραγωνικού σφάλματος (Mean Square Error,MSE), και είναι και αυτή που χρησιμοποιείται στην παρούσα εργασία. Εφόσον το πρόβλημα εδώ είναι η πρόβλεψη της εξέλιξης μιας χρονοσειράς, η MSE ορίζεται ως :

$$MSE = \frac{1}{n}\sum_{1}^{n} Y_i - \hat{Y}_i \qquad (3.8)$$

όπου

- $n$ είναι το πλήθος των τιμών της σειράς
- $Y_i$είναι η πρόβλεψη την στιγμή $i$
- $\hat{Y}_i$ είναι η πραγματική τιμή την στιγμή$i$

Με άλλα λόγια, η μετρική MSE είναι ο μέσος όρος του τετραγώνου του σφάλματος.

- **Αλγόριθμος Μάθησης για το Perceptron (Steepest Descent)**

Ο δημοφιλέστερος αλγόριθμος μάθησης για την επιβλεπόμενη εκπαίδευση και τη βελτιστοποίηση των (πολυεπίπεδων) νευρωνικών δικτύων, είναι ο αλγόριθμος κατάβασης κλίσης (Gradient descent). Στο επίκεντρο του αλγορίθμου βρίσκεται η ελαχιστοποίηση της μερικής παραγώγου μιας επιλεχθείσας συνάρτησης κόστους (εξ' ου και η ονομασία του) ως προς τα βάρη του δικτύου, με την επαναληπτική προσαρμογή τους ανά βήμα εκτέλεσης. Τα βήματα του αλγορίθμου είναι τα εξής:

1. *Αρχικοποίηση βαρών*

   Η επιλογή των αρχικών τιμών για τα βάρη γίνεται από μια ομοιόμορφη κατανομή με μηδενικό μέσο, και αποσκοπεί στο να μην λάβουν οι παράγωγοί τους υπερβολικά υψηλές ή χαμηλές τιμές (exploding or vanishing gradients) με αποτέλεσμα την αδυναμία σύγκλισης του αλγορίθμου ή την μη αποδοτική εκτέλεσή του.

2. *Πρόσθιος υπολογισμός σφάλματος*

Τα δείγματα εκπαίδευσης παρουσιάζονται στο δίκτυο υπό την μορφή ενός διανύσματος που περιέχει την είσοδο *x* και το διάνυσμα της επιθυμητής απόκρισης *y*. Έπειτα, για κάθε επίπεδο νευρώνων και για κάθε νευρώνα υπολογίζεται η απόκριση $d_i = \varphi(u_j)$ όπου

$$u_j = \sum_{i=0}^{m} w_{ij} y_i \qquad (3.9)$$

οι οποίες διαδοχικά οδηγούν στην έξοδο του δικτύου και στον υπολογισμό του σφάλματος *e*.

3. *Οπίσθιος υπολογισμός*

Σε αυτό το στάδιο, για κάθε νευρώνα του δικτύου , υπολογίζεται η τοπική κλίση του $\delta$ , και πραγματοποιείται η ανανέωση των βαρών μέσω του γενικευμένου κανόνα Δέλτα,

$$\delta_j = \frac{-\partial e}{\partial y_i} \frac{\partial y_j}{\partial u_j} \ , \qquad (3.10a)$$

$$\Delta w_{ij} = \eta \times \delta_j \times y_i \qquad (3.10b)$$

όπου $\eta$ ο ρυθμός μάθησης.

4. *Επανάληψη*

Επαναλαμβάνονται οι υπολογισμοί των βημάτων τρία και τέσσερα μέχρι την ικανοποίηση του κριτηρίου τερματισμού.

Η προσαρμογή των βαρών ανάλογα με τον τρόπο παρουσίασης των δειγμάτων εκπαίδευσης στο δεύτερο βήμα πραγματοποιείται με δύο τρόπους. Αν τα δείγματα παρουσιαστούν στο σύνολό τους, τότε η διαδικασία μάθησης λέγεται μαζική, ενώ εάν παρουσιάζονται δείγμα προς δείγμα, τότε μιλάμε για μάθηση σε πραγματικό χρόνο (on-line). Τις περισσότερες φορές, επιλέγεται η on-line μάθηση, λόγω του ότι απαιτεί μικρότερο χώρο αποθήκευσης από την μαζική και επειδή μειώνει τις πιθανότητες να παγιδευτεί η διαδικασία μάθησης σε κάποιο τοπικό ελάχιστο της συνάρτησης κόστους.

## *3.5 Βελτιστοποιητές*

Πρόκειται για μια οικογένεια αλγορίθμων, βάσει των οποίων αναπροσαρμόζεται ο κλασσικός αλγόριθμος κατάβασης κλίσης (Stochastic Gradient Descent-SGD) [21]. Παρά την δημοφιλία του, ο αλγόριθμος αυτός, παρουσιάζει ορισμένα μειονεκτήματα:

- Δεν είναι πάντα εύκολη η επιλογή του κατάλληλου ρυθμού μάθησης. Για μικρές τιμές, η σύγκλιση είναι εξαιρετικά αργή ενώ σε υψηλότερες δεν εγγυάται καν, καθώς η συνάρτηση κόστους ταλαντώνεται γύρω από κάποιο (τοπικό ή ολικό) ελάχιστο.

- Αλγόριθμοι επιλογής του ρυθμού μάθησης, όπως η προσομοιωμένη ανόπτηση (simulated annealing), λειτουργούν με προεπιλεγμένες τιμές κατωφλίων, οι οποίες όμως δεν είναι εφικτό να προσαρμοστούν στα ιδιαίτερα χαρακτηριστικά των δειγμάτων εκπαίδευσης.
- Ο ρυθμός μάθησης, εφαρμόζεται για κάθε αναπροσαρμοζόμενο βάρος, χωρίς να λαμβάνεται υπόψιν η συχνότητα εμφάνισης κάθε κρυφού χαρακτηριστικού του συνόλου εκπαίδευσης.
- Στον χώρο των βαρών, πολλές φορές εμφανίζονται σέλες, δηλαδή σημεία όπου μία διάσταση ανεβαίνει και μία άλλη κατεβαίνει. Στην γύρω περιοχή του σημείου λοιπόν, το σφάλμα μένει ίδιο, και είναι δύσκολο για τον SGD να ξεφύγει διότι η παράγωγος του σφάλματος είναι μηδενική σε κάθε διάσταση.

Παρουσιάζονται συνοπτικά οι κυριότεροι από τους εν χρήση βελτιστοποιητές:

*SGD με ορμή*

Σε περιοχές του χώρου των βαρών όπου κάποιες διαστάσεις έχουν μεγαλύτερη κλίση σε σχέση με άλλες, το σφάλμα ταλαντώνεται γύρω από αυτές και συγκλίνει πιο αργά προς το τοπικό ελάχιστο, το οποίο είναι σύνηθες σε αυτές τις περιοχές.

*SGD χωρίς ορμή* *SGD με ορμή*

*Σχήμα 3.9 Συμπεριφορά του SGD αλγορίθμου με και χωρίς ορμή*

Με την προσθήκη ενός ποσοστού *γ* της προηγούμενης τιμής του βάρους στην εξίσωση αυξάνεται η συνεισφορά των παραγώγων που δείχνουν προς την ίδια κατεύθυνση, μειώνοντας έτσι τις ταλαντώσεις.

*Ενημέρωση :*

$$u_t = \gamma u_{t-1} + \eta \nabla_\theta J_\theta \qquad (3.11a)$$

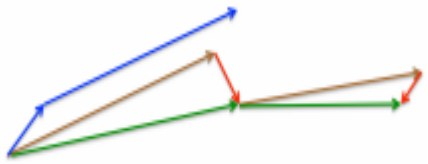
$$\theta = \theta - u_t \qquad (3.11b)$$

όπου $\vartheta$ το διάνυσμα των βαρών, *γ* ο παράγοντας ορμής και *η* ο ρυθμός μάθησης.

*Nesterov Accelarated Gradient*

Ο συγκεκριμένος βελτιστοποιητής είναι εξέλιξη του προηγούμενου (μπλε διάνυσμα). Χρησιμοποιεί για παράγωγο την προβλεπόμενη τιμή που θα πάρει το βάρος (καφέ διάνυσμα), κι έτσι εισάγει έναν διορθωτικό παράγοντα (κόκκινο διάνυσμα) στον υπολογισμό

του νέου βάρους (πράσινο διάνυσμα), έτσι ώστε να μειώσει περαιτέρω τις ταλαντώσεις. Έχει φανεί πειραματικά πως βελτιώνει την αποδοτικότητα των αναδρομικών δικτύων.



*Σχήμα 3.10 Nesterov Accelerated Gradient*

*Ενημέρωση :*

$$u_t = \gamma u_{t-1} + h\nabla_\theta J(\theta - \gamma u_{t-1}) \quad (3.12a)$$

$$\theta = \theta - u_t \quad (3.12b)$$

*Adagrad*

Η χρησιμότητα του αλγόριθμου αυτού βρίσκεται στο ότι ανάλογα με την συχνότητα ενεργοποίησης των βαρών, εκτελεί μεγαλύτερου μέτρου ενημερώσεις στα σπάνια βάρη και μικρότερου στα συχνά. Το κάνει αυτό υπολογίζοντας έναν διαγώνιο πίνακα *G* στον οποίο αποθηκεύονται όλες οι προηγούμενες τιμές ενός συγκεκριμένου βάρους και μέσω αυτού αλλάζει τον ρυθμό μάθησης. Το σαφές πλεονέκτημα εδώ, είναι πως εξαλείφεται η ανάγκη για χειροκίνητη ρύθμιση του ρυθμού μάθησης $\eta$.

*Ενημέρωση:*

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii}+\varepsilon}} g_{t,i} \quad (3.13a)$$

$$g_{t,i} = \nabla_{\theta_t} J(\theta_{t,i}) \quad (3.13b)$$

*Adam*

Ο αλγόριθμος Adaptive Moment Estimation, ο οποίος είναι και αυτός που χρησιμοποιήθηκε στην παρούσα εργασία, συνδυάζει τα πλεονεκτήματα της Adagrad και του SGD με ορμή υπολογίζοντας προσαρμοζόμενους ρυθμούς μάθησης για κάθε βάρος, καθώς και μία φθίνουσα μέση τιμή των προηγούμενων παραγώγων τους και του τετραγώνου τους (first και second moment).

Η ενημέρωση πραγματοποιείται ως εξής:

1. *Εκτίμηση First Moment (mean)* $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ *(3.14)*
2. *Εκτίμηση Second Moment (uncentered variance)* $u_t = \beta_2 u_{t-1} + (1-\beta_2) g_t^2$ *(3.15)*
3. *Διορθωτικός υπολογισμός* $\hat{m}_t = \frac{m}{1-\beta_1^t}, \hat{u}_t = \frac{u}{1-\beta_1^t}$ *(3.16)*
4. *Ενημέρωση παραμέτρων* $\theta_{t+1} = \theta_t - \frac{h}{\sqrt{\hat{u}_t}+\varepsilon} \hat{m}_t$ *(3.17)*

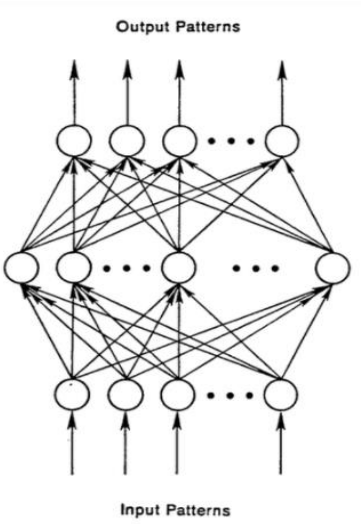
Ο διορθωτικός υπολογισμός λαμβάνει χώρα για λόγους σύγκλισης των διανυσμάτων *m* και *u*, ενώ οι προεπιλεγμένες τιμές των $\beta_1$, $\beta_2$ και $\varepsilon$ είναι 0.9 , 0.999 και $10^{-8}$ αντίστοιχα.

## *3.6 Αναδρομικά Νευρωνικά Δίκτυα*

Τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks-RNN), καθώς και τα δίκτυα μακράς-βραχείας μνήμης (LSTM) που αποτελούν μια υποκατηγορία των πρώτων, είναι ειδικά σχεδιασμένα για να αναγνωρίζουν μοτίβα σε ακολουθιακά δεδομένα, όπως οι μετρήσεις ενός αισθητήρα ή οι τιμές του χρηματιστηρίου, καθώς και σε δεδομένα κειμένου ή ακολουθιών DNA. Η κύρια διαφορά των RNN και των LSTM από τα πολυστρωματικά νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks-FNN), είναι πως λαμβάνουν υπόψιν τους την χρονική ή/και την ακολουθιακή διάσταση των δεδομένων εισόδου. Αυτό επιτυγχάνεται με την εισαγωγή μιας μονάδας μνήμης στην αρχιτεκτονική τους, η οποία προσομοιάζει αρκετά στην ανθρώπινη μνήμη.

### *3.6.1 Δίκτυα Πρόσθιας Τροφοδότησης (Feedforward Neural Network)*

Από δομικής πλευράς, η ροή της πληροφορίας μέσω ενός δικτύου πρόσθιας τροφοδότησης δεν σχηματίζει κύκλους, δηλαδή τα δεδομένα εισόδου αφού μετασχηματιστούν μέσω μαθηματικών υπολογισμών σε κάθε επίπεδο νευρώνων, προχωρούν στο επόμενο και τελικά παράγουν την έξοδο. Στα προβλήματα επιβλεπόμενης μάθησης, όπως στην αναγνώριση εικόνας για παράδειγμα, η έξοδος είναι ένα όνομα που καταδεικνύει την κλάση στην οποία ανήκει η εικόνα, μέσω αναγνώρισης υπαρχόντων μοτίβων στον χώρο της εισόδου. Έτσι, τα δίκτυα εκπαιδεύονται με κατηγοριοποιημένες εικόνες *(x,y)* μέχρι να ελαχιστοποιηθεί το σφάλμα εξόδου. Στη συνέχεια, αφού οριστικοποιηθούν οι αλλαγές στα βάρη του δικτύου, όταν παρουσιαστεί μια εικόνα στο εκπαιδευμένο μοντέλο, αυτό θα της αποδώσει μια ετικέτα.

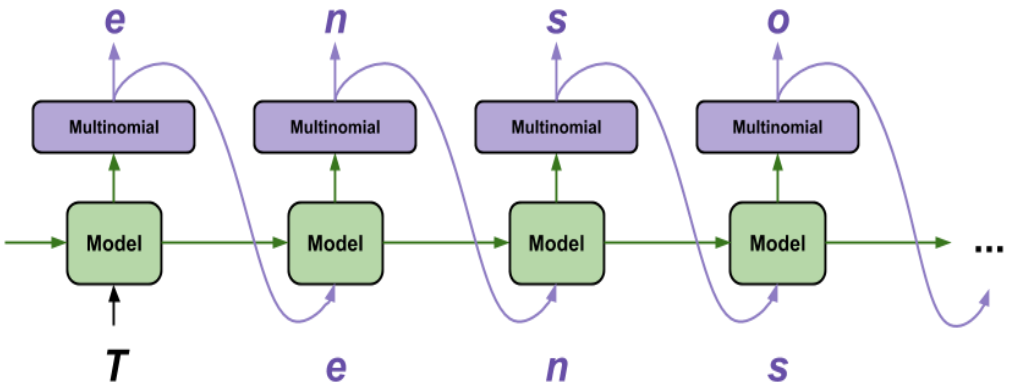



*Σχήμα 3.11 Πολυεπίπεδο Νευρωνικό Δίκτυο*

Ωστόσο, η σειρά με την οποία θα παρουσιαστούν τα δείγματα δεν επηρεάζει την έξοδο του δικτύου. Αν για παράδειγμα, η πρώτη εικόνα που δει, είναι ενός αυτοκινήτου, δε θα οδηγηθεί να κατηγοριοποιήσει την επόμενη εικόνα σαν αυτή ενός φαναριού.

### *3.6.2 Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Network)*

Τα αναδρομικά νευρωνικά δίκτυα ωστόσο, λαμβάνουν υπόψιν τους όχι μόνο την είσοδο που βλέπουν την εκάστοτε στιγμή, αλλά και όλες αυτές που έχουν προηγηθεί χρονικά. Η έξοδος του δικτύου, λοιπόν, την χρονική στιγμή *t-1* επηρεάζει την έξοδό του τη στιγμή *t*. Κατά αυτόν τον τρόπο, τα δύο αυτά είδη εισόδων συνδυάζονται για να εξάγουν την πληροφορία που περιέχεται σε καινούρια δεδομένα.




*Σχήμα 3.12 Στάδια παραγωγής εξόδου από ένα αναδρομικό δίκτυο*

Φαίνεται λοιπόν εδώ, πως αυτό που διαφοροποιεί τα RNN από τα FNN, είναι ο βρόχος ανάδρασης που συνδέει τις παρελθούσες αποφάσεις τους με τις τωρινές, στιγμή προς στιγμή. Η λειτουργία της μνήμης των RNN βρίσκεται στην κρυφή τους κατάσταση, στην οποία διατηρείται η ακολουθιακή πληροφορία των εισόδων. Μέσω της μνήμης, ανακαλύπτονται συσχετίσεις μεταξύ των καταστάσεων του δικτύου οι οποίες απέχουν χρονικά μεταξύ τους, και ονομάζονται "εξαρτήσεις μακράς διάρκειας". Αυτό κατά μία έννοια, είναι λογικό να συμβαίνει, διότι όπως και στον άνθρωπο, οι αποφάσεις που λαμβάνονται συναρτήσει μιας τρέχουσας κατάστασης, εξαρτώνται από τις καταστάσεις που έχουν προηγηθεί. Ένα παράδειγμα στο οποίο φαίνεται η χρησιμότητα της μνήμης, είναι στα προβλήματα πρόβλεψης του επόμενου χαρακτήρα σε μια ακολουθία γραμμάτων. Αν το δίκτυο δει τον χαρακτήρα «*t*», τότε αυτή η πληροφορία θα το βοηθήσει να συμπεράνει ότι ο επόμενος χαρακτήρας είναι το «*h*» κ.ο.κ. Τέτοιου είδους προβλήματα γενικότερα, είναι αρκετά δύσκολο να επιλυθούν από ένα FNN. Οι δύο βασικές εξισώσεις που περιγράφουν την λειτουργία ενός RNN είναι :

*a)* $h_t = \varphi(Wx_t + Uh_{t-1})$ *(3.18)*

όπου $h_t$ η κρυφή κατάσταση του δικτύου τη στιγμή *t*

$y_t = Vh_t$ *(3.19)*

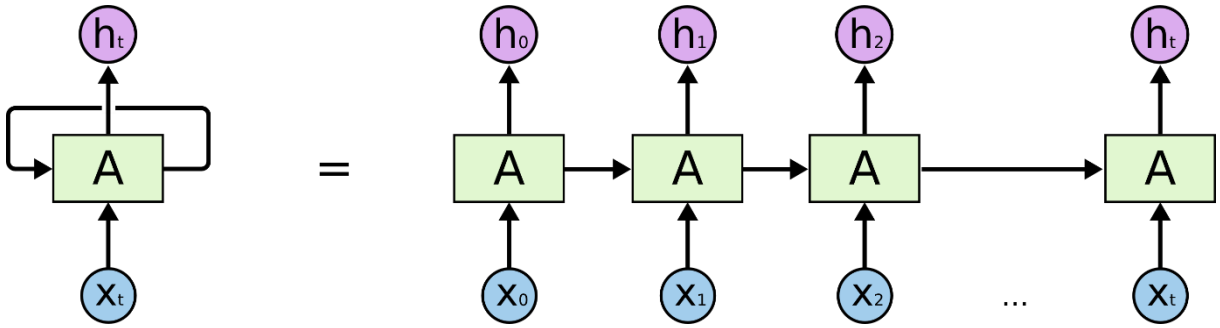
όπου $y_t$ η έξοδος του δικτύου τη στιγμή t

Η συνάρτηση ενεργοποίησης *φ* είναι συνήθως η σιγμοειδής ή η υπερβολική εφαπτομένη. Σκοπός της είναι να κάνει τις παραγώγους υπολογίσιμες για την οπισθοδιάδοση. Η *h* είναι συνάρτηση της παρούσας εισόδου πολλαπλασιασμένης με έναν πίνακα βαρών *W*, και της προηγούμενής της κατάστασης πολλαπλασιασμένης με ένα πίνακα *U*. Τα βάρη *W*, λειτουργούν σαν φίλτρα που καθορίζουν πόση σημασία πρέπει να δοθεί στην παρούσα είσοδο και στις προηγούμενες κρυφές καταστάσεις. Μέσω του αλγόριθμου backpropagation, επιστρέφει το σφάλμα που παράγουν και χρησιμοποιείται για να τροποποιηθούν τα βάρη μέχρι την ελαχιστοποίηση του σφάλματος. Ο βρόχος ανάδρασης που περιγράφεται από τις δικτυακές εξισώσεις ενεργοποιείται σε κάθε χρονικό βήμα της ακολουθίας εισόδου, κι έτσι η κατάσταση *h*, περιέχει ίχνη όχι μόνο της προηγούμενής της τιμής, αλλά και όλων αυτών που προηγήθηκαν του βήματος *t-1*, για όσο αντέξει η μνήμη.

### *3.6.3 Οπίσθια Διάδοση στο Χρόνο (Back Propagation Through Time)*

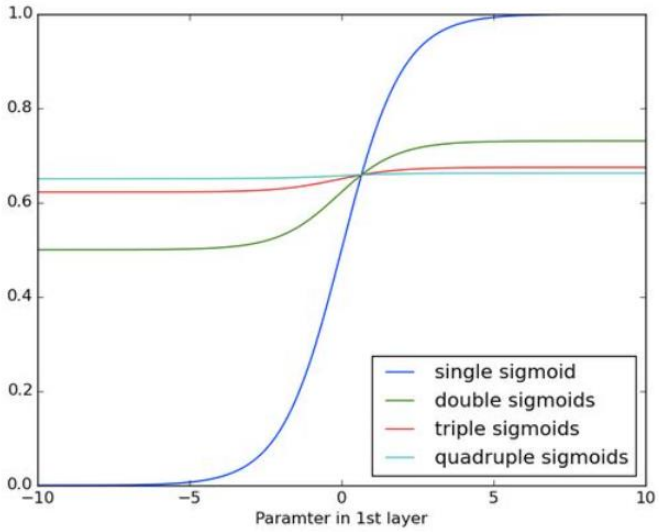
Τα νευρωνικά δίκτυα, αναδρομικά ή μη, είναι απλά εμφωλευμένες συναρτήσεις της μορφής $f\left(g\left(h(x)\right)\right)$. Όταν παρουσιάζεται το στοιχείο του χρόνου στον υπολογισμό των συναρτήσεων αυτών, επεκτείνεται η σειρά υπολογισμού των παραγώγων μέσω του κανόνα της αλυσίδας . Λόγω του βρόχου ανάδρασης των αναδρομικών δικτύων, χρησιμοποιείται μια

επέκταση του γνωστού αλγόριθμου backpropagation. Ουσιαστικά, η επέκταση αυτή πραγματοποιεί την εκδίπλωση αυτού του βρόχου (loop unrolling) παράγοντας ένα αντίγραφο του δικτύου για κάθε χρονικό βήμα, μαζί με την αντίστοιχη είσοδο και έξοδο, καθώς και της κρυφής κατάστασής του. Στη συνέχεια, πραγματοποιείται υπολογισμός των σφαλμάτων και μέσω του αλγόριθμου backpropagation βρίσκεται η παράγωγος του σφάλματος σε συνάρτηση με όλα τα δικτυακά βάρη (παράμετροι).




*Σχήμα 3.13 Αναπαράσταση του loop unrolling*

Ωστόσο, η οπίσθια διάδοση στο χρόνο έχει δύο συγκεκριμένα μειονεκτήματα που εμφανίζονται όταν έχουμε υψηλό αριθμό χρονικών βημάτων. Αφενός, το υπολογιστικό φορτίο της είναι δυσβάσταχτο και αφετέρου, σε κάθε βήμα αντιστοιχεί ο υπολογισμός μιας παραγώγου, και τόσες είναι οι παράγωγοι που χρειάζονται για μία και μόνο ενημέρωση των βαρών του δικτύου. Αφετέρου, η μακροσκελής σειρά υπολογισμού των κλίσεων σε συνάρτηση με τα βάρη, οδηγεί σε πολύ ψηλές ή χαμηλές τιμές, το οποίο σαν συνέπεια έχει τον αργό ρυθμό μάθησης και την εισαγωγή θορύβου στο μοντέλο. Όσον αφορά το πρώτο μειονέκτημα, εφαρμόζεται μια παραλλαγή του αλγόριθμου BPTT, γνωστή ως Truncated BPTT. Κατά αυτή την μέθοδο, αν υποτεθεί ότι έχουμε μια ακολουθία μήκος n, τότε μόλις περάσουν m βήματα, τρέχει ο αλγόριθμος BPTT για k βήματα, κι έτσι, αν το k είναι μικρό, τότε η ενημέρωση των παραμέτρων είναι φθηνή, ενώ ταυτόχρονα οι κρυφές καταστάσεις του δικτύου έχουν εκτεθεί σε πολλά χρονικά βήματα που περιέχουν χρήσιμη πληροφορία για το μακρύ παρελθόν, η οποία μπορεί να αξιοποιηθεί [54].




*Σχήμα 3.14 Αποτέλεσμα πολλαπλών εφαρμογών της σιγμοειδούς*

### *3.5.4 Δίκτυα Μακράς και Βραχείας Μνήμης (LSTM)*

Προς επίλυση του προβλήματος της εξαφανιζόμενης κλίσης (Vanishing Gradient) των RNN, οι ερευνητές Sepp Hochreiter και Jürgen Schmidhuber [55] πρότειναν το 1997 ένα νέο είδος αναδρομικού δικτύου, επονομαζόμενο Δίκτυο Μακράς και Βραχείας Μνήμης (Long Short-Term Memory Neural Network – LSTM). Τα LSTM διατηρούν ένα σταθερότερο σφάλμα κατά την οπίσθια διάδοση μέσω των στρωμάτων και του χρόνου, κι έτσι επιτρέπεται στο δίκτυο να συνεχίσει την διαδικασία μάθησης ακόμη κι όταν ο αριθμός των βημάτων είναι μεγάλος, συνδέοντας αίτιο κι αποτέλεσμα απομακρυσμένα μεταξύ τους. Η μνημονική επεξεργασία των δεδομένων πραγματοποιείται μέσω ενός κυττάρου μνήμης, το οποίο χάρις σε πύλες ανοίγουν και κλείνουν, εκτελεί τις εξής ενέργειες [56]:

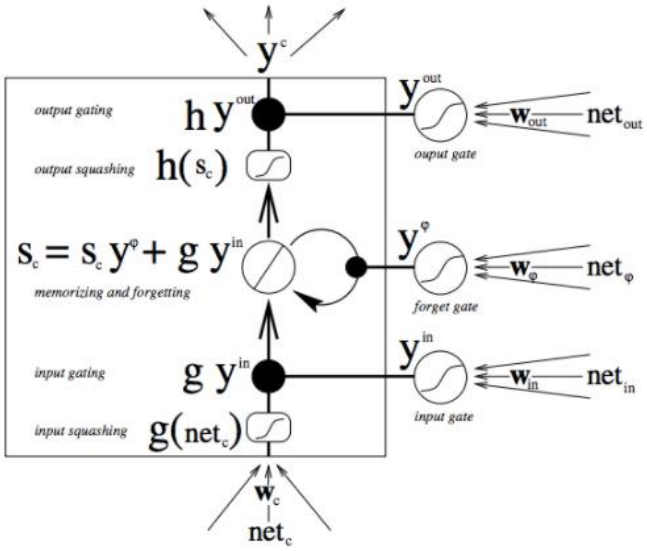
1. *Ανάκληση (Forget/Remember)*: Κάθε φορά που παρουσιάζεται μια καινούρια είσοδος στο δίκτυο, επιλέγεται ποια στοιχεία των παρελθοντικών εισόδων θα μείνουν στην κρυφή μνήμη και ποια θα διαγραφούν.
2. *Αποθήκευση (Save)* : Αναζήτηση χρήσιμων πληροφοριών στην παρούσα είσοδο, προς αποθήκευση .
3. *Εστίαση (Focus)* : Επιλογή σχετικών πληροφοριών της κρυφής κατάστασης μνήμης που θα βοηθήσουν με την παρούσα είσοδο.

Οι εξισώσεις που περιγράφουν το κύτταρο μνήμης είναι οι εξής:

1. $I_t = \sigma(X_t W_{xi} + M_{t-1} W_{hi} + C_{t-1} W_{ci} + b_i)$ *(3.20)*

2. $F_t = \sigma\left(X_t W_{xf} + M_{t-1} W_{hf} + C_{t-1} W_{cf} + b_f\right)$ *(3.21)*
3. $O_t = \sigma(X_t W_{xo} + M_{t-1} W_{ho} + C_{t-1} W_{co} + b_o)$ *(3.22)*
4. $C_t = F_t \odot C_{t-1} + I_t \odot tanh(X_t W_{xc} + M_{t-1} W_{hc} + b_c)$ *(3.23)*
5. $M_t = O_t \odot tanh(C_t)$ *(3.24)*

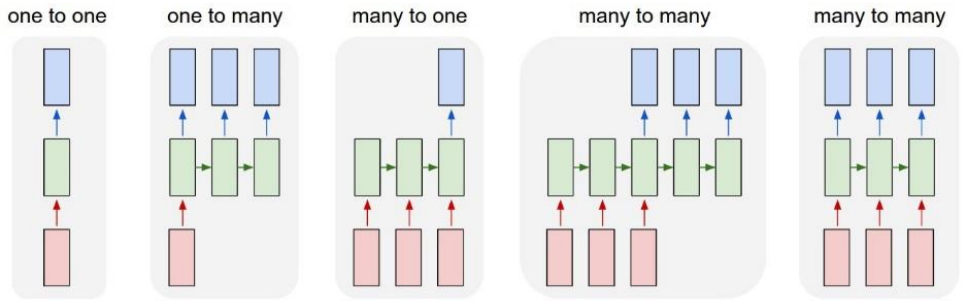
Η μνήμη μακράς διαρκείας αντιπροσωπεύεται από το διάνυσμα $C_t$ , το οποίο είναι το κύτταρο μνήμης, ενώ η κρυφή κατάσταση μνήμης που είχαμε και στα RNN βρίσκεται στο διάνυσμα $M_t$. Το διάνυσμα $F_t$ είναι ο μηχανισμός ανάκλησης και ο μηχανισμός αποθήκευσης είναι στο διάνυσμα $I_t$. Τέλος, η εστίαση πραγματοποιείται από το διάνυσμα $O_t$.




*Σχήμα 3.15 Ροή πληροφορίας στο κύτταρο μνήμης του LSTM*

## *3.6.5 Αρχιτεκτονικές Δικτύων Μακράς Βραχείας Μνήμης*

Ανάλογα με το είδος προβλήματος που έχουμε να χειριστούμε, υπάρχουν διάφορα είδη αρχιτεκτονικών LSTM. Δεδομένου πως τα δεδομένα εισόδου και εξόδου μπορούν να είναι ακολουθιακά, οι βασικές αρχιτεκτονικές είναι τέσσερις.




*Σχήμα 3.16 Αρχιτεκτονικές δικτύων LSTM*

*Ένα προς ένα (One to One)*

Βασική επεξεργασία εισόδου δίχως την χρήση κάποιου αναδρομικού δικτύου. Ένα παράδειγμα είναι η κατηγοριοποίηση μιας εικόνας

*Ένα σε πολλά (One to Many)*

Από μία μη ακολουθιακή είσοδο, π.χ. μια εικόνα, παράγεται μια ακολουθιακή έξοδος, π.χ. παραγωγή κειμένου από μία εικόνα

*Πολλά σε ένα (Many to One)*

Η είσοδος είναι μια ακολουθία που πρέπει να κατηγοριοποιηθεί

*Πολλά σε πολλά (Many to Many)*

Η είσοδος και η έξοδος είναι ακολουθίες

# Κεφάλαιο 4

# Ανάπτυξη προγνωστικών μοντέλων κρουσμάτων γριπωδών συνδρομών με χρήση τεχνικών μηχανικής μάθησης

Στόχος της παρούσας εργασίας ήταν η πρόβλεψη των κρουσμάτων ILI στον ελλαδικό χώρο, για μία έως τρεις εβδομάδες στο μέλλον. Για τον σκοπό αυτό, διερευνήθηκαν διαφορετικές αρχιτεκτονικές υπολογιστικών μοντέλων βασισμένες στη χρήση των δικτύων LSTM και την ενσωμάτωση ετερογενών πηγών δεδομένων[1]. Κίνητρο για την επιλογή της μεθόδου LSTM αποτέλεσε η ικανότητά τους να διαχειρίζονται αποτελεσματικά δεδομένα χρονοσειρών. Για το χρονικό διάστημα της δεκαετίας 2010-2019 και για κάθε μία από τις 13 περιφέρειες της Ελλάδας συλλέχθηκαν εβδομαδιαίοι μετεωρολογικοί δείκτες και οι αναρτήσεις των χρηστών του Twitter με βάση συγκεκριμένες λέξεις κλειδιά. Επιπλέον παραχωρήθηκαν από τον Εθνικό Οργανισμό Δημόσιας Υγείας (ΕΟΔΥ) δεδομένα του συστήματος επιτήρησης γρίπης «Sentinel» σχετικά με τον εβδομαδιαίο αριθμό κρουσμάτων γρίπης για όλη την Ελλάδα στο διάστημα της δεκαετίας.

Αρχικά εκπαιδεύτηκαν τρία πρωταρχικά μοντέλα (M,T,H), κάθε ένα από τα οποία εκμεταλλεύεται την προγνωστική δύναμη των μετεωρολογικών δεικτών (M), των αναρτήσεων στο Twitter (T) και των επιδημιολογικών δεδομένων (H), αντίστοιχα. Η αρχιτεκτονική των τριών πρωταρχικών μοντέλων διαμορφώθηκε με βάση τη χρήση ενός στρώματος LSTM, η έξοδος του οποίου οδηγήθηκε σε ένα πλήρως συνδεδεμένο (dense) στρώμα, με σκοπό την ανίχνευση μακροπρόθεσμων εξαρτήσεων ανάμεσα στην είσοδο και την έξοδο. Στη συνέχεια, διερευνήθηκαν δύο συνδυαστικές αρχιτεκτονικές των πρωταρχικών μοντέλων με σκοπό τον αποτελεσματικό συγκερασμό των ετερογενών δεδομένων εισόδου για τη βελτίωση των εξαγόμενων προγνώσεων.

## *4.1 Συλλογή Δεδομένων*

### *4.1.1 Δημοσιεύσεις χρηστών του Twitter*

Το επίσημο API του Twitter κρίθηκε ακατάλληλο για την συλλογή των αναρτήσεων επειδή επιτρέπει τη συλλογή αναρτήσεων μέχρι μια εβδομάδα στο παρελθόν, και έτσι, επιλέχθηκε μια βιβλιοθήκη της Python, που ονομάζεται *«GetOldTweets3»,* η οποία αναπτύχθηκε από τον *Dmitry Mottl [67]*. Περιέχει πολλές χρήσιμες λειτουργίες, όπως την καταμέτρηση του

[1] Ευχαριστούμε την NVIDIA Corporation για την υποστήριξή της μέσω της δωρεάς της Quadro P6000 GPU, που χρησιμοποιήθηκε για τη διεξαγωγή αυτής της έρευνας.

πλήθους των δημοσιεύσεων ή την γεωγραφική αναζήτησή τους. Σαν προτιμώμενη γλώσσα ορίστηκε η ελληνική. Για τη συλλογή των δεδομένων προσδιορίστηκε ένα σύνολο από λέξεις-κλειδιά σχετικές με τις ILI και τα αντίστοιχα συμπτώματά τους. Ανάμεσα στις λέξεις κλειδιά προστέθηκε, επίσης, η προσωπική αντωνυμία «*εγώ*» έτσι ώστε να αυξηθεί η πιθανότητα μία δημοσίευση να αφορά τον χρήστη που την ανάρτησε. Αλιεύθηκαν όλες οι αναρτήσεις που περιείχαν τουλάχιστον μία από τις προκαθορισμένες λέξεις κλειδιά (ή κάποια από τις μορφές τους).

Λέξεις Κλειδιά

*βήχας, γρίπη, H1N1, πόνος, ιός, ίωση, γαστρεντερίτιδα, ωτίτιδα, αντιβίωση, ασπιρίνη, άσθμα, κρύωμα, διάρροια, ζαλάδα, φάρμακο, ημικρανία, ναυτία, παυσίπονο, παναντόλ, παρακεταμόλη, χάπι, αναρρώνω, αναπνευστικό, φτάρνισμα, στομαχόπονος, σύμπτωμα, πονόλαιμος, εμβόλιο, πυρετός, άρρωστος, ντεπόν, πονοκέφαλος, ρίγος, εγώ.*


*Σχήμα 4.1: Ενδεικτική συχνότητα λέξεων στις εβδομαδιαίες αναρτήσεις της Αττικής το 2014*

Όσον αφορά την συχνότητα εμφάνισης τους, οι λέξεις κλειδιά προφανώς κυριαρχούν, ενώ οι υπόλοιπες λέξεις σχηματίζουν το συγκείμενο στο οποίο εμφανίζονται οι πρώτες.

### *4.1.2 Μετεωρολογικά Δεδομένα*

Λαμβάνοντας υπόψιν την προγνωστική δύναμη που μπορούν να παρέχουν τα μετεωρολογικά στοιχεία, εξήχθησαν αναφορές για την Ελλάδα για τη δεδομένη χρονική περίοδο από την ιστοσελίδα https://power.larc.nasa.gov/data-access-viewer, οι οποίες περιείχαν τις μετρήσεις για την βροχόπτωση, την σχετική υγρασία, την επιφανειακή πίεση,

την ελάχιστη, μέση και μέγιστη θερμοκρασία στα δύο μέτρα από το έδαφος, την μέση ταχύτητα του ανέμου και την θερμική υπέρυθρη ακτινοβολία. Το εβδομαδιαίο διάνυσμα εισόδου στα μοντέλα ήταν 1x104, δηλαδή οκτώ χαρακτηριστικά για κάθε μία από 13 περιφέρειες. Στη συνέχεια, για κάθε ξεχωριστό χαρακτηριστικό, πραγματοποιήθηκε κανονικοποίηση στο εύρος [-1,1].


*Σχήμα 4.2 Υπέρυθρη ακτινοβολία κατά τα έτη 2010-2015 στην πόλη της Θεσσαλονίκης*

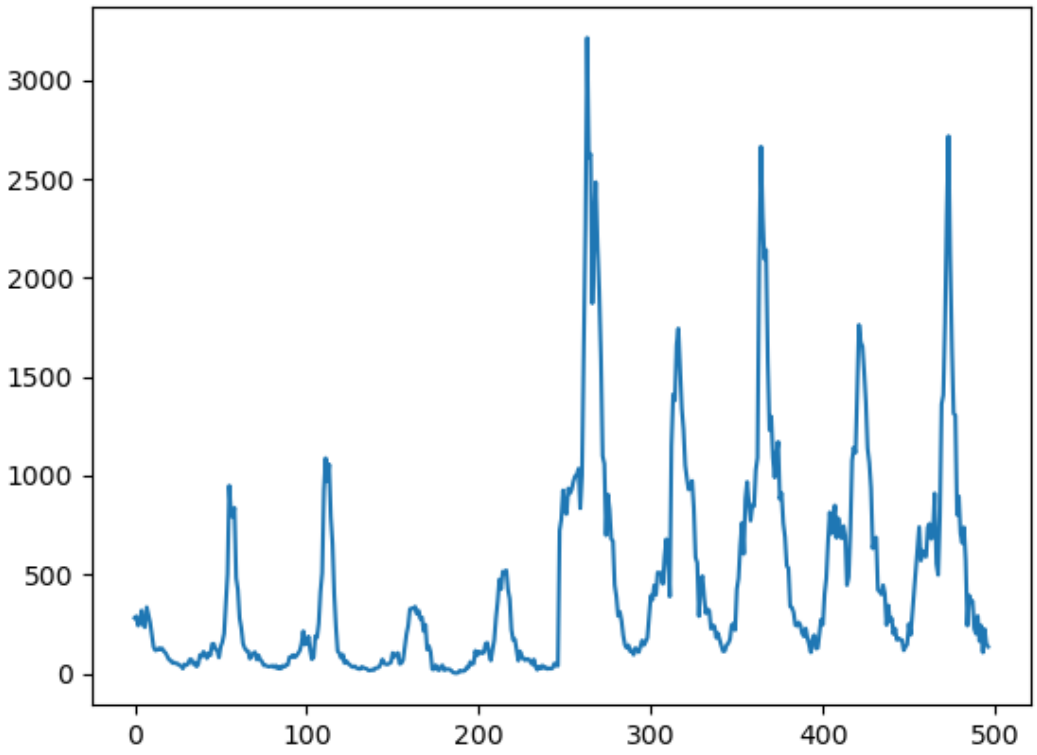

*Σχήμα 4.3 Βροχόπτωση κατά τα έτη 2010-2015 στην πόλη της Θεσσαλονίκης*


*Σχήμα 4.4 Ταχύτητα του ανέμου κατά τα έτη 2010-2015 στην πόλη της Θεσσαλονίκης*

### *4.1.3 Δεδομένα συστήματος επιτήρησης γρίπης Sentinel*

Το σύστημα παρατηρητών νοσηρότητας Sentinel είναι υπεύθυνο για την επιτήρηση νοσημάτων τα οποία παρουσιάζουν αυξημένη συχνότητα και άρα είναι υψηλής σημασίας για την δημόσια υγεία, ωστόσο λόγω της ήπιας κλινικής εικόνας τους δεν απαιτούν άμεση παρέμβαση ή νοσηλεία σε νοσοκομείο. Λειτουργεί με την εθελοντική συμμετοχή ιδιωτών γιατρών από όλη την Ελλάδα, οι οποίοι αναφέρουν και αναλύουν σε εβδομαδιαία βάση

στοιχεία για τα επιτηρούμενα νοσήματα. Μερικά από αυτά είναι οι ILI, οι λοιμώξεις του αναπνευστικού συστήματος με πυρετό και η γαστρεντερίτιδα. Για το σκοπό της εργασίας, παραχωρήθηκαν οι εβδομαδιαίες αναφορές των ILI για το διάστημα 2010-2019, οι οποίες παρουσιάζουν τον συνολικό αριθμό επισκέψεων σε ιατρικούς φορείς και το σύνολο των κρουσμάτων.




*Σχήμα 4.5 Κρούσματα γριπωδών συνδρομών στην Ελλάδα κατά τη δεκαετία 2010-2019. Στον άξονα x βρίσκεται κάθε εβδομάδα της δεκαετίας 2010-2019 και στον άξονα y το πλήθος των κρουσμάτων*

## *4.2 Προεπεξεργασία Δεδομένων Twitter*

### *4.2.1 Καθαρισμός από bots*

Μεγάλος όγκος των δημοσιεύσεων προέρχονταν από «social bots» δηλαδή λογαριασμούς χρηστών οι οποίοι ελέγχονται τουλάχιστον εν μέρει από λογισμικό, των οποίων ο σκοπός είναι να δρουν σύμφωνα με προκαθορισμένα μοτίβα. Η δράση τους αφορά μεταξύ άλλων, την απάντηση σε μια δημοσίευση ή την προκαθορισμένη ανάρτηση δημοσιεύσεων [26,50].

Για να καθαριστούν οι δημοσιεύσεις, χρησιμοποιήθηκε το “Botometer API”, ένα project που αναπτύχθηκε από το Παρατηρητήριο Κοινωνικών Δικτύων του πανεπιστημίου της Indiana [50]. Ο αλγόριθμός του API αυτού, εξάγει χαρακτηριστικά τα οποία αφορούν το προφίλ ενός χρήστη, τη δομή του κοινωνικού του κύκλου, τα χρονικά μοτίβα δραστηριότητάς του, τη γλώσσα που χρησιμοποιεί και το συναίσθημα που εκφράζει. Στη συνέχεια, παράγει ένα σκορ που καθορίζει εάν το προφίλ ανήκει σε άνθρωπο ή σε bot. Το κατώφλι τέθηκε στο 0.5, το οποίο είναι αρκετά κοντά με αυτό που χρησιμοποιήθηκε από το Pew Research Center(0.43)[51].

Αφού απομακρύνθηκαν οι περιττές δημοσιεύσεις, συμπεράναμε την τοποθεσία κάθε δημοσίευσης είτε άμεσα, μέσω του Tweepy API και του GeoPy, είτε έμμεσα, από την δηλωμένη τοποθεσία κάθε χρήστη στο προφίλ του.

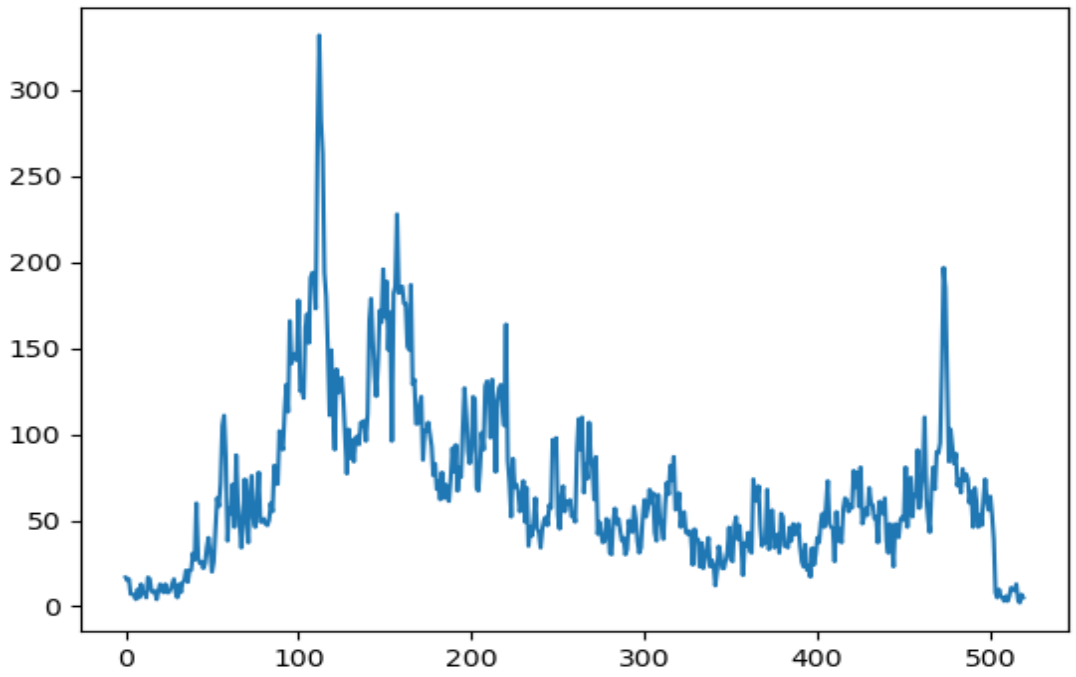
### *4.2.2 Επεξεργασία Φυσικής Γλώσσας*

Για τους σκοπούς της προεπεξεργασίας των αναρτήσεων του Twitter, χρησιμοποιήθηκε το Spacy, μια βιβλιοθήκη NLP ανοιχτού κώδικα της Python, μαζί με το pipeline της ελληνικής γλώσσας «el_core_news_md», το οποίο διαθέτει λεξιλόγιο 20.000 λέξεων.

Αρχικά, όλα τα κεφαλαία γράμματα των tweets μετατράπηκαν σε πεζά και στη συνέχεια, καταργήθηκαν όλα τα σημεία στίξης, urls, hashtag, αριθμοί, emoji, λέξεις εκτός λεξιλογίου, καθώς και οι πιο συχνές λέξεις διακοπής (π.χ. «όπως», «αντί», «αλλαγή», «διαφορετικά» κ.λπ.). Επιπλέον, καταργήθηκαν λέξεις που είχαν λιγότερες από πέντε εμφανίσεις στο σύνολο του corpus. Όλες οι λέξεις στη συνέχεια μετατράπηκαν στο αντίστοιχο διάνυσμα λέξης (1x300) χρησιμοποιώντας τον αλγόριθμο Word-2-Vec.

Από τα υπόλοιπα tweets, εξαγάγαμε n-grams και αναπαραστάσεις κειμένου TFIDF. Από τα bigrams και trigrams κρατήθηκε περίπου το 2% του συνόλου τους, επιλέγοντας τα πιο συχνά εμφανιζόμενα στο σώμα κειμένου. Στη συνέχεια, για να αξιολογηθεί ποιες λέξεις-κλειδιά είναι καλύτεροι προγνωστικοί παράγοντες και για να χαρτογραφηθεί η σημασιολογική σχέση κάθε λέξης με κάποια άλλη, πραγματοποιήθηκε φασματική ομαδοποίηση στο σύνολο του λεξιλογίου, και για κάθε n-gram (n=1,2,3), ανατέθηκε σε κάθε λέξη η αντίστοιχη ομάδα της.

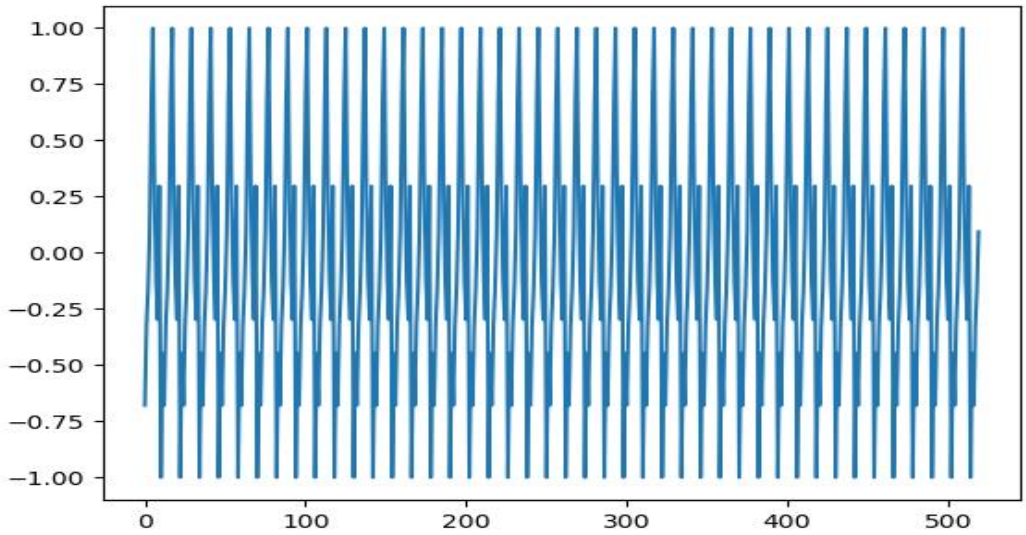
Ένα επιπλέον χαρακτηριστικό ήταν ο συνολικός αριθμός των ομάδων στις οποίες ανήκουν οι καθορισμένες λέξεις κλειδιά, που εμφανίζεται ανά εβδομάδα και ανά περιφέρεια και, τέλος, μαζί με τον εβδομαδιαίο αριθμό των δημοσιεύσεων ανά περιφέρεια, χρησιμοποιήθηκε η εβδομαδιαία μέτρηση μοναδικών χρηστών, διότι κάποιοι χρήστες παρουσίαζαν πολλαπλές αναρτήσεις ανά εβδομάδα. Όλα τα χαρακτηριστικά κανονικοποιήθηκαν στο εύρος [-1,1].



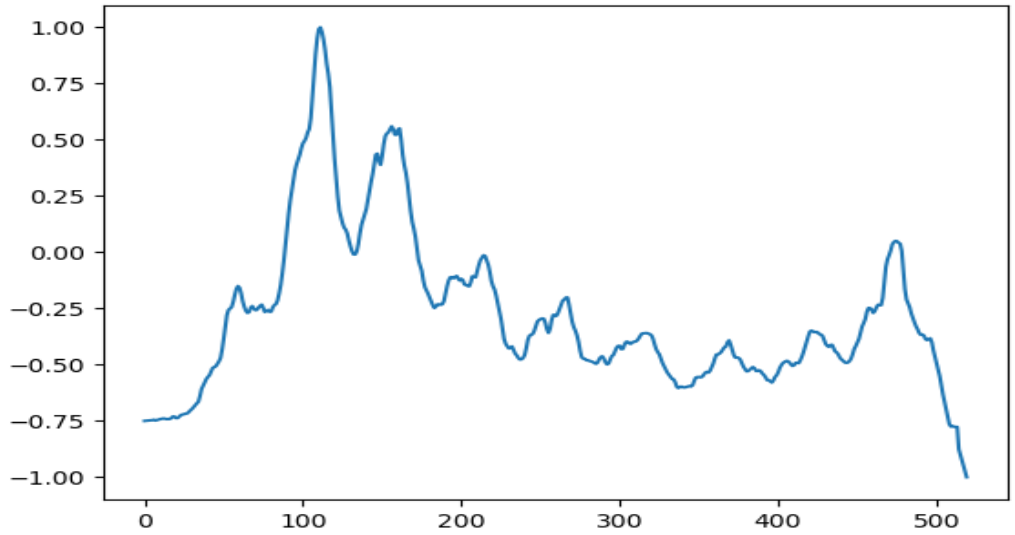

*Σχήμα 4.3 Αριθμός αναρτήσεων Twitter. Στον άξονα x βρίσκεται κάθε εβδομάδα της δεκαετίας 2010-2019 και στον άξονα y το πλήθος των αναρτήσεων*

### *4.2.3 Εποχική αποδόμηση των χαρακτηριστικών εισόδου*

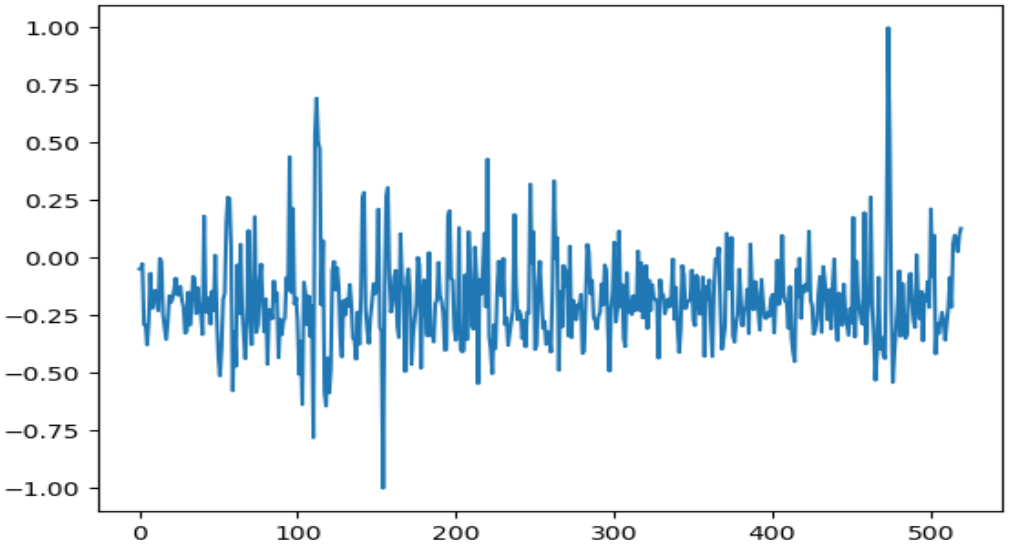
Μελετώντας τα εξαγόμενα χαρακτηριστικά από το Twitter, είναι εμφανές πως υπάρχει μια περιοδικότητα η οποία υπονοεί μια συσχέτισή τους με το πλήθος των κρουσμάτων ILI. Παρόλα αυτά, τα κοινωνικά σήματα χαρακτηρίζονται από μη-περιοδικές εναλλαγές και υψηλό θόρυβο, τα οποία οφείλονται εν μέρει στην μεταβολή του ποσοστού χρήσης του Twitter στην Ελλάδα ανά έτος. Έτσι, για να διευκολυνθούν τα μοντέλα κατά τη διαδικασία μάθησης, εκτελέστηκε εποχική αποδόμηση σε κάθε προκύπτον σήμα εκτός από τα bigrams (συχνότητα tweet, αριθμός χρηστών κλπ.), τάξης m=52 (όσες είναι δηλαδή και οι εβδομάδες σε ένα έτος) έτσι ώστε να απομονωθούν καλύτερα οι συνιστώσες που είναι καλύτεροι προγνωστικοί παράγοντες [23]. Το εβδομαδιαίο διάνυσμα εισόδου ήταν μεγέθους 1x226. Η κανονικοποίηση πράγματοποιήθηκε ανά χαρακτηριστικό στο εύρος [-1,1].




*Σχήμα 4.4 Εποχιακή συνιστώσα του πλήθους των αναρτήσεων Twitter κατά την δεκαετία 2010-2019*




*Σχήμα 4.5 Συνιστώσα τάσης του πλήθους των αναρτήσεων Twitter ανά εβδομάδα για την δεκαετία 2010-2019*




*Σχήμα 4.6 Συνιστώσα θορύβου τάσης του πλήθους των αναρτήσεων Twitter ανά εβδομάδα για την δεκαετία 2010-2019*

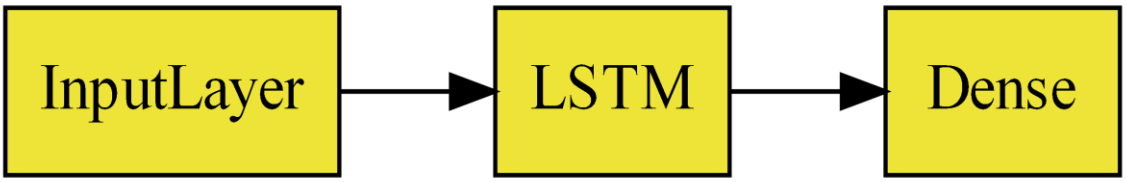
## *4.3 Πρωταρχικά Προγνωστικά Μοντέλα*

Στο πλαίσιο της παρούσας εργασίας αναπτύχθηκαν τρία πρωταρχικά μοντέλα:

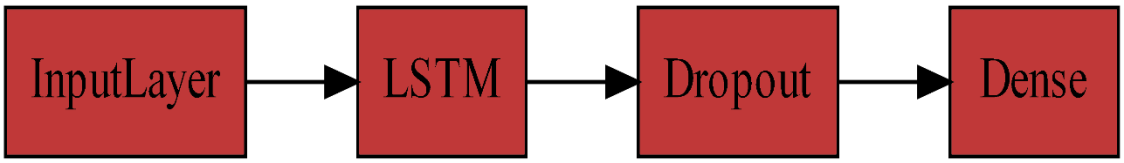
- Ένα προγνωστικό μοντέλο του αριθμού κρουσμάτων ILI βασισμένο στη χρήση δεδομένων του Twitter (T)
- Ένα προγνωστικό μοντέλο του αριθμού των κρουσμάτων ILI βασισμένο στη χρήση μετεωρολογικών δεδομένων (M)
- Ένα προγνωστικό μοντέλο του αριθμού των κρουσμάτων ILI βασισμένο στη χρήση δεδομένων του συστήματος επιτήρησης γρίπης Sentinel (H).

Στα μοντέλα M και H δόθηκαν ως είσοδος δεδομένα δύο εβδομάδων [t-2, t-1] με σκοπό την εκτίμηση των κρουσμάτων ILI κατά τις εβδομάδες t, t+1, t+2, όπου t η τρέχουσα κάθε φορά εβδομάδα. Το μοντέλο T έλαβε ως είσοδο δεδομένα τριών εβδομάδων [t-3, t-1] με σκοπό την εκτίμηση των κρουσμάτων ILI κατά τις εβδομάδες t, t+1, t+2. Για την αποτελεσματική διαχείριση των διαφορετικών δεδομένων χρονοσειρών (Twitter, μετεωρολογικά, Sentinel) και με στόχο την ακριβή πρόβλεψη των ILI κρουσμάτων, χρησιμοποιήθηκαν τα LSTM δίκτυα στο κρυφό στρώμα των μοντέλων, ακολουθούμενα από ένα πλήρως συνδεδεμένο στρώμα εξόδου (dense layer) με έναν κόμβο και γραμμική συνάρτηση ενεργοποίησης. Επιπλέον, ενσωματώθηκε στρώμα αποκοπής συνδέσεων (dropout layer) με στόχο την αποφυγή της υπερπροσαρμογής των μοντέλων στα δεδομένα εκπαίδευσης. Ανάλογα με τον τύπο δεδομένων που διαχειριζόταν κάθε μοντέλο, η θέση του στρώματος dropout μεταβλήθηκε κατάλληλα. Για το μοντέλο M το στρώμα dropout

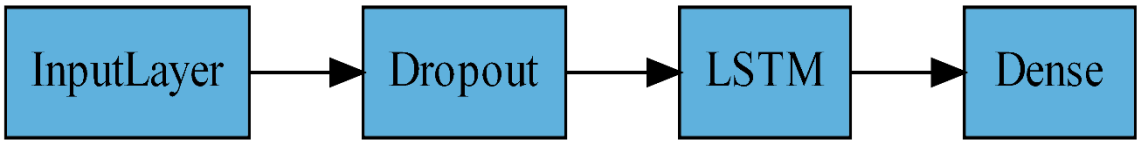
τοποθετήθηκε στην έξοδο του LSTM, για το μοντέλο T στην είσοδο του LSTM, ενώ για το H μοντέλο δε χρησιμοποιήθηκε στρώμα dropout.




*Σχήμα 4.7 Αρχιτεκτονική του μοντέλου H (δεδομένα Sentinel), αποτελούμενη από ένα στρώμα εισόδου (Input Layer), ένα κρυφό στρώμα LSTM και ένα πλήρως συνδεδεμένο στρώμα εξόδου (Dense)*




*Σχήμα 4.8 Αρχιτεκτονική του μοντέλου M (μετεωρολογικά δεδομένα), με ένα στρώμα Dropout να παρεμβάλλεται μεταξύ των στρωμάτων LSTM και Dense*




*Σχήμα 4.9 Αρχιτεκτονική του μοντέλου T (δεδομένα Twitter), όπου το στρώμα Dropout ακολουθεί την είσοδο του δικτύου*

### *4.3.1 Εκπαίδευση και ρύθμιση υπερπαραμέτρων*

Η αξιολόγηση της επίδοσης των πρωταρχικών μοντέλων βασίστηκε στη χρήση διασταυρωμένης επικύρωσης πέντε πτυχών (5-fold cross validation). Κάθε πτυχή (fold) περιελάμβανε έναν τυχαίο συνδυασμό δύο ετών μεταξύ του 2010-2019. Ο τρόπος με τον οποίο διαχωρίστηκε το σύνολο των δεδομένων ήταν ο εξής: σε κάθε επανάληψη, για το σύνολο εκπαίδευσης (training set) χρησιμοποιήθηκε το 60% του συνόλου, δηλαδή τρία ζεύγη ετών, ενώ για το σύνολο επικύρωσης (validation set) και το σύνολο αξιολόγησης (test set) χρησιμοποιήθηκε από ένα ζεύγος ετών (20% και 20% επί του συνόλου, αντίστοιχα). Για τη διερεύνηση των υπερπαραμέτρων εφαρμόστηκε άπληστη αναζήτηση μέσω διαμόρφωσης κατάλληλης συνάρτησης πλέγματος. Έχοντας λάβει υπόψιν την υπάρχουσα ερευνητική βιβλιογραφία, ο χώρος αναζήτησης των υπερπαραμέτρων για τα πρωταρχικά μοντέλα περιορίστηκε σε συγκεκριμένα εύρη τιμών για κάθε υπερπαράμετρο, μειώνοντας σημαντικά

έτσι τον χρόνο αναζήτησης. Κατά τη διαδικασία αυτή, διατηρήθηκαν οι συνδυασμοί υπερπαραμέτρων, η χρήση των οποίων οδήγησε στην εξαγωγή προβλέψεων, που παρουσίαζαν στατιστικά σημαντική συσχέτιση (p-value<0.05) με τις πραγματικές τιμές των κρουσμάτων στο σύνολο επικύρωσης. Πραγματοποιήθηκε διερεύνηση τιμών για τις εξής υπερπαραμέτρους:

- Batch size: Καθορίζει πόσα δείγματα εκπαίδευσης θα παρουσιαστούν στο δίκτυο προτού γίνει η ανανέωση των βαρών του. Διερευνήθηκαν μικρές τιμές, από 1-16 δείγματα, με οριστική επιλογή το batch size να ισούται με ένα. Αν και μεγαλύτερες τιμές οδηγούν σε γρηγορότερη εκπαίδευση, οι μικρότερες έχουν γενικότερο μικρότερο σφάλμα γενίκευσης[19-20][22].
- Optimizer: Adam [21].
- Dropout: Βοηθάει στην αποφυγή της υπερπροσαρμογής στο σύνολο εκπαίδευσης, θέτοντας τυχαία ένα ποσοστό βαρών κάποιου στρώματος στο μηδέν.
- Συναρτήσεις ενεργοποίησης: Sigmoid, Tanh
- Αρχικοποίηση Βαρών: Η επιλογή κατάλληλης μεθόδου αρχικοποίησης των τιμών που θα έχουν τα βάρη του δικτύου, βοηθάει στο να αποφευχθεί το πρόβλημα του exploding ή vanishing gradient. Επιλέχθηκε ο Xavier normal initializer [24].
- Εποχές: Ο αριθμός των επαναλήψεων της εκπαίδευσης του μοντέλου επί του συνόλου των δεδομένων εκπαίδευσης. Για όλα τα μοντέλα, ο αριθμός των εποχών ορίστηκε ίσος με 50.
- Window Size: Η παράμετρος αυτή καθορίζει πόσο «πίσω» στο χρόνο θα κοιτάξει ένα LSTM.

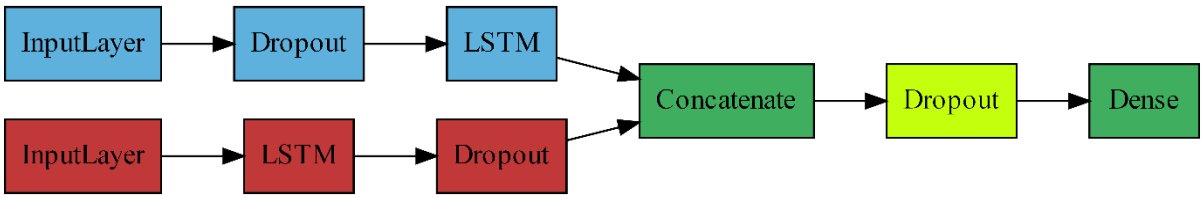
## *4.4 Ανάπτυξη σύνθετων προγνωστικών μοντέλων*

Με στόχο τον συγκερασμό των ετερογενών πηγών δεδομένων και την αποτελεσματική αξιοποίησή τους για την εξαγωγή ακριβών εκτιμήσεων των κρουσμάτων ILI, αναπτύχθηκαν συνδυαστικές αρχιτεκτονικές των πρωταρχικών μοντέλων (M, T, H). Λαμβάνοντας υπόψη πως τα δεδομένα υγείας που συλλέγονται από συστήματα επιδημιολογικής επιτήρησης όπως το Sentinel, μπορούν να είναι διαθέσιμα μετά το πέρας του ορίζοντα πρόβλεψης, διερευνήθηκε η δυνατότητα αξιοποίησης του συνδυασμού ενναλακτικών πηγών δεδομένων (Twitter, μετεωρολογικά δεδομένα, Sentinel) προς την ανάπτυξη βελτιωμένων μοντέλων για την ακριβή πρόγνωση των κρουσμάτων ILI.

- **Σύνθετο Μοντέλο Μετεωρολογικών Δεδομένων και Δεδομένων Twitter**

Για την σχεδίαση του σύνθετου μοντέλου μετεωρολογικών δεδομένων και δεδομένων Twitter **(**MT), λήφθηκαν όλα τα προεκπαιδευμένα στρώματα των μοντέλων M και T, και οι έξοδοί τους συγχωνεύθηκαν και οδηγήθηκαν σε έναν πλήρως συνδεδεμένο νευρώνα εξόδου με μια γραμμική συνάρτηση ενεργοποίησης. Καθώς η επίδοση του πρωταρχικού μοντέλου M

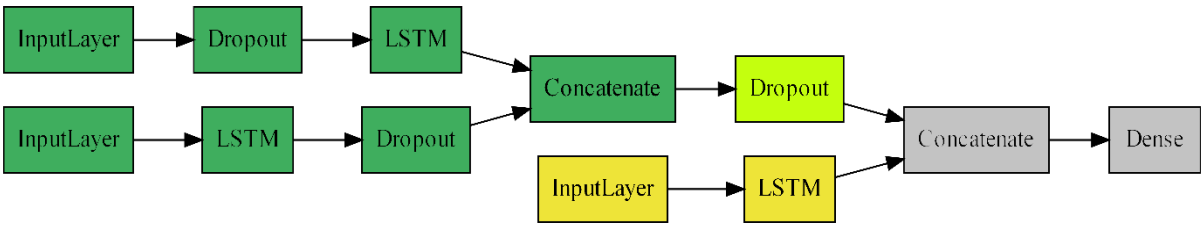
στο σύνολο επικύρωσης ήταν υψηλότερη από την επίδοση του Τ μοντέλου, επιλέχθηκε κατά την εκπαίδευση να διατηρηθούν σταθερά τα βάρη των στρωμάτων του δικτύου Μ, επιτρέποντας την επανεκπαίδευση μόνο του δικτύου Τ, με στόχο την εξαγωγή νέων χαρακτηριστικών από τα δεδομένα του Twitter, που θα συνδυάζονταν αποτελεσματικά με τα εξαγόμενα χαρακτηριστικά του Μ μοντέλου. Κατά την εκπαίδευση, οι υπερπαράμετροι των επιμέρους μοντέλων Μ και Τ διατηρήθηκαν σταθερές, ο αριθμός των εποχών τέθηκε ίσος με 50 και η τιμή του κοινού στρώματος αποκοπής συνδέσεων ίση με 0.1.




*Σχήμα 4.11 Αρχιτεκτονική του μοντέλου ΜΤ . Τα βάρη του δικτύου Τ (μπλε) διατηρούνταν σταθερά κατά την εκπαίδευση*

- **Σύνθετο Μοντέλο Μετεωρολογικών Δεδομένων, Δεδομένων Twitter και Δεδομένων Sentinel**

Το σύνθετο μοντέλο μετεωρολογικών δεδομένων, δεδομένων Twitter και δεδομένων Sentinel (ΜΤΗ) αξιοποίησε την αρχιτεκτονική του προηγούμενου βήματος, συνδυάζοντας τα μοντέλα ΜΤ και Η. Στην αρχιτεκτονική του συνδυαστικού μοντέλου χρησιμοποιήθηκαν οι επιμέρους αρχιτεκτονικές των μοντέλων ΜΤ και Η, παραλείποντας το πλήρως συνδεδεμένο (dense) στρώμα των μοντέλων ΜΤ και Η, και προσθέτοντας ένα στρώμα συγχώνευσης και έναν πλήρως συνδεδεμένο νευρώνα εξόδου. Κατά την εκπαίδευση, ο αριθμός των εποχών τέθηκε ίσος με 50 και οι υπερπαράμετροι των επιμέρους μοντέλων ΜΤ και Η διατηρήθηκαν σταθερές, με εξαίρεση την τιμή του στρώματος αποκοπής συνδέσεων (dropout) του ΜΤ μοντέλου, που τέθηκε ίση με 0.8.




*Σχήμα 4.12 Αρχιτεκτονική του μοντέλου ΜΤΗ . Ο υπογράφος με πράσινο χρώμα αντιστοιχεί στο προεκπαιδευμένο δίκτυο ΜΤ, ενώ ο υπογράφος με κίτρινο χρώμα στο προεκπαιδευμένο δίκτυο Η*

# Κεφάλαιο 5

# Αξιολόγηση επίδοσης πρωταρχικών και σύνθετων προγνωστικών μοντέλων

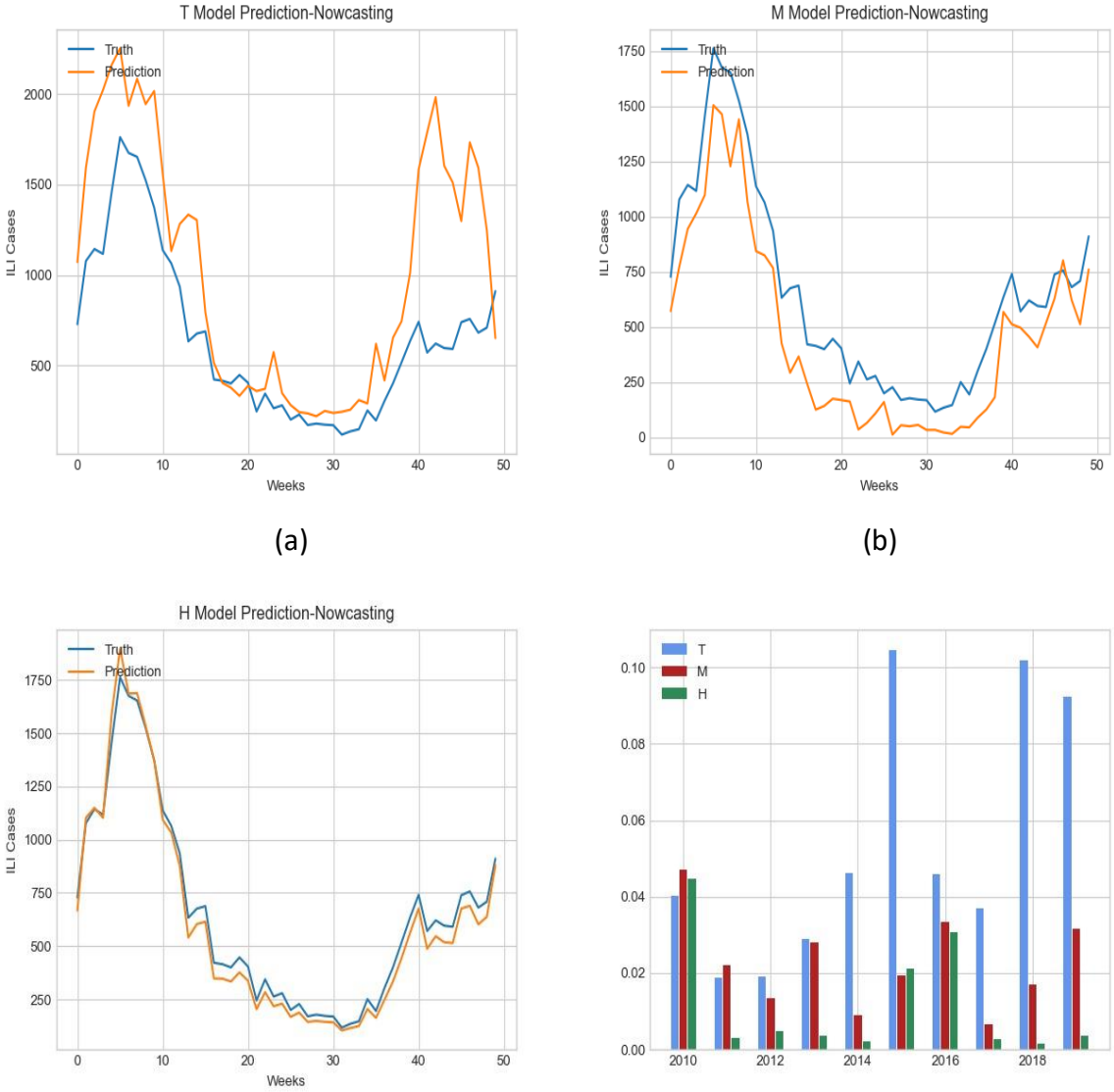
Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της αξιολόγησης των πρωταρχικών και σύνθετων προγνωστικών μοντέλων. Η διαμόρφωση του πλαισίου αξιολόγησης βασίστηκε στην εφαρμογή διασταυρωμένης επικύρωσης πέντε πτυχών (5-fold cross validation). Κάθε πτυχή (fold) περιελάμβανε έναν τυχαίο συνδυασμό δύο ετών μεταξύ του 2010-2019. Διερευνήθηκαν διαφορετικές τιμές του ορίζοντα πρόβλεψης, με στόχο την αξιολόγηση της ικανότητας των μοντέλων να προβλέπουν με ακρίβεια τον αριθμό κρουσμάτων ILI για την τρέχουσα εβδομάδα (nowcasting) και για έως δύο εβδομάδες στο μέλλον (forecasting). Ως κριτήρια αξιολόγησης χρησιμοποιήθηκαν ήταν το μέσο τετραγωνικό σφάλμα (MSE) και η συσχέτιση Pearson:

$$MSE = \frac{1}{n}\sum_{1}^{n} Y_i - \hat{Y}_i \tag{5.1}$$

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{5.2}$$

## *5.1 Αξιολόγηση πρωταρχικών μοντέλων*

Στον παρακάτω πίνακα παρουσιάζονται τα συγκριτικά αποτελέσματα, καθώς και η καλύτερη απόδοση ανά στόχο πρόβλεψης και κριτήριο. Παρατηρείται πως το μοντέλο Η που εκπαιδεύτηκε στα δεδομένα του Sentinel εμφανίζει σαφώς καλύτερη απόδοση, το Μ μοντέλο να ακολουθεί, ενώ η χαμηλότερη επίδοση καταγράφεται για το μοντέλο Τ . Τα δεδομένα από το Twitter φαίνονται ικανά να αδράξουν τη γενική τάση των κρουσμάτων, αλλά έχουν τεράστιο περιθώριο βελτίωσης σε αρκετές περιπτώσεις, ενώ οι μετεωρολογικοί δείκτες ακολουθούν την πορεία των κρουσμάτων πιο πιστά. Όσον αφορά το μοντέλο Η, η μόνη αδυναμία εντοπίζεται στην ανίχνευση του ακριβούς ύψους των κρουσμάτων σε μεμονωμένα σημεία. Ακολουθούν κάποια ενδεικτικά αποτελέσματα μαζί με τον πίνακα σύγκρισης των κριτηρίων MSE και Pearson για κάθε πρωταρχικό μοντέλο.



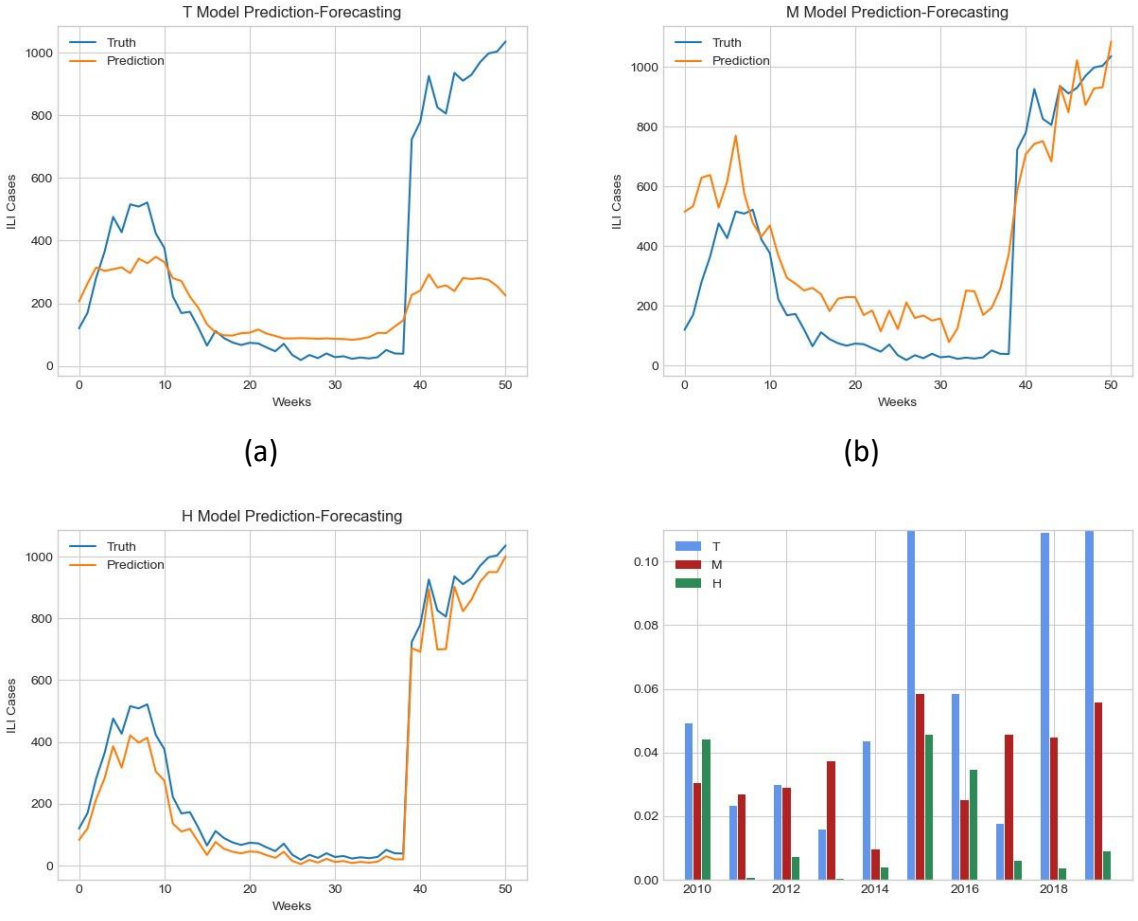

*Σχήμα 5.1*

(a) *Μοντέλο Τ : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(b) *Μοντέλο Μ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(c) *Μοντέλο Η* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(d) *MSE των μοντέλων Μ, Τ και Η για τα έτη 2010-2019 (nowcasting)*



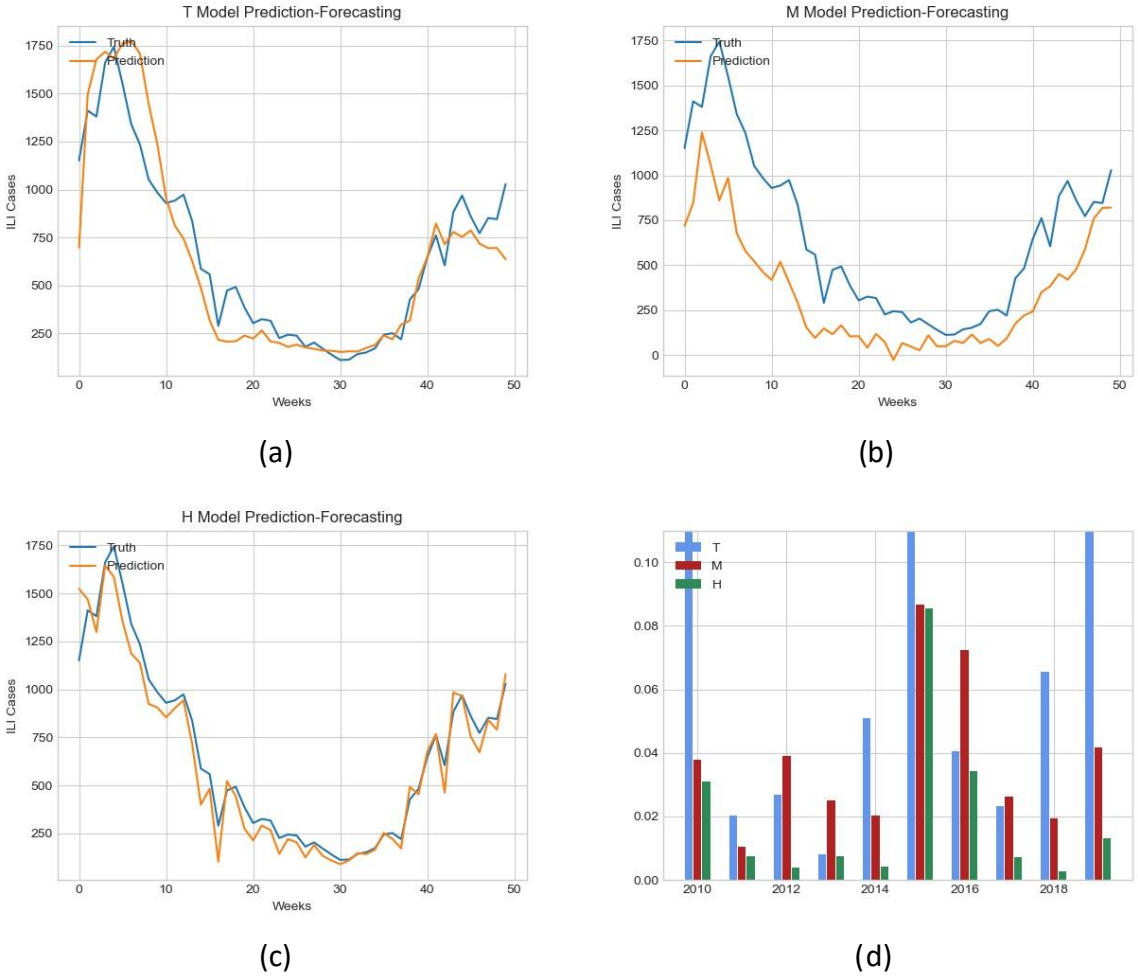

*Σχήμα 5.2*

(a) *Μοντέλο Τ : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2014*

(b) *Μοντέλο Μ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2014*

(c) *Μοντέλο Η* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2014*

(d) *MSE των μοντέλων Μ,Τ και Η για τα έτη 2010-2019 (forecasting μίας εβδομάδας)*




*Σχήμα 5.3*

(a) *Μοντέλο Τ : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2016*

(b) *Μοντέλο Μ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2016*

(c) *Μοντέλο Η* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2016*

(d) *MSE των μοντέλων Μ,Τ και Η για τα έτη 2010-2019 (forecasting δύο εβδομάδων)*

*Πίνακας 5.4 Σύγκριση ελάχιστης, μέσης και μέγιστης τιμής των κριτηρίων MSE και Pearson Correlation για κάθε μοντέλο ανά βδομάδα πρόβλεψης*

| MSE | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| T | 0.0189 | 0.0159 | 0.0082 | 0.0535 ± 0.0489 | 0.0624 ± 0.0480 | 0.0730 ± 0.0662 | 0.1045 | 0.1430 | 0.1769 |
| M | 0.0066 | 0.0095 | 0.0105 | 0.0227 ± 0.0164 | 0.0360 ± 0.0267 | 0.0379 ± 0.0215 | 0.0469 | 0.0583 | 0.0866 |
| H | **0.0014** | **0.0002** | **0.0027** | **0.0118 ± 0.0117** | **0.0154 ± 0.0134** | **0.0197 ± 0.0151** | **0.0447** | **0.0456** | **0.0854** |

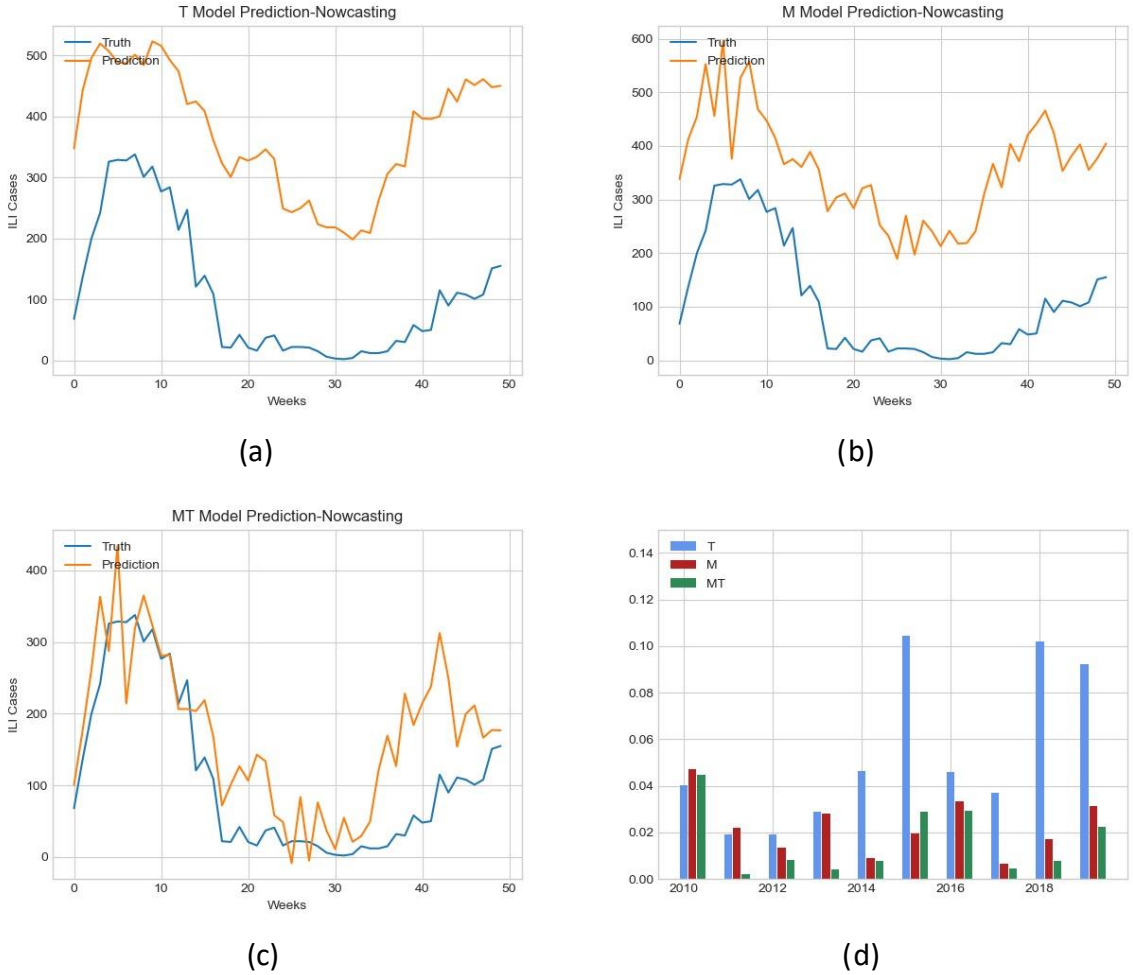
| CORR | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| T | 0.0909 | 0.2208 | 0.2335 | 0.6578 ± 0.3391 | 0.6404 ± 0.3248 | 0.5491 ± 0.3541 | 0.9380 | 0.9156 | 0.8925 |
| M | **0.4008** | **0.5451** | **0.3910** | 0.8510 ± 0.1714 | 0.8334 ± 0.0824 | 0.7732 ± 0.1271 | 0.9824 | 0.9468 | 0.8967 |
| H | 0.2663 | 0.2598 | 0.2704 | **0.8965 ± 0.2154** | **0.8885 ± 0.0120** | **0.8801 ± 0.0189** | **0.9967** | **0.9937** | **0.9885** |

## *5.2 Αξιολόγηση σύνθετων μοντέλων*

- *MT μοντέλο*

Στα επόμενα σχήματα παρουσιάζονται τα αποτελέσματα για το σύνθετο μοντέλο MT. Όπως φάνηκε στην προηγούμενη παράγραφο, το μοντέλο T υστερεί σημαντικά έναντι του M, όπως για παράδειγμα στο έτος 2013. Συμπεραίνεται ότι το MT μοντέλο είναι ικανό να

βελτιώσει την ακρίβεια της πρόβλεψης σε σύγκριση με τα δύο επιμέρους M και T μοντέλα, το οποίο φανερώνει πως με τον συνδυασμό των πρωταρχικών μοντέλων σε μια κοινή αρχιτεκτονική και την αναπροσαρμογή των βαρών του μοντέλου T επιτυγχάνεται η εξαγωγή ποιοτικότερων χαρακτηριστικών από το σύνολο δεδομένων του Twitter.



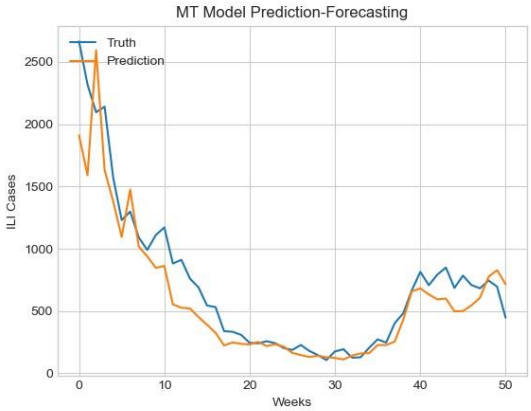

*Σχήμα 5.4*

(a) *Μοντέλο T : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

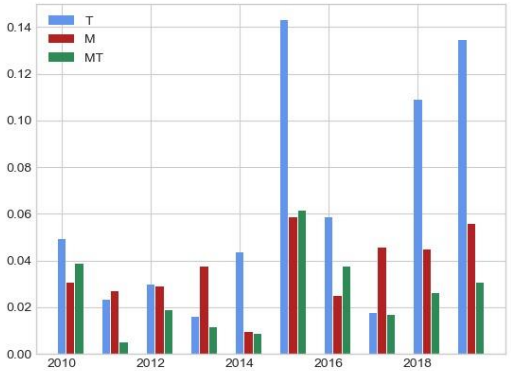
(b) *Μοντέλο M* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

(c) *Μοντέλο MT* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

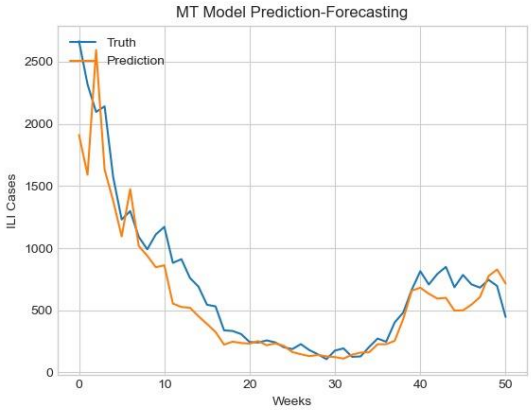
(d) *MSE των μοντέλων M,T και H για τα έτη 2010-2019 (nowcasting)*



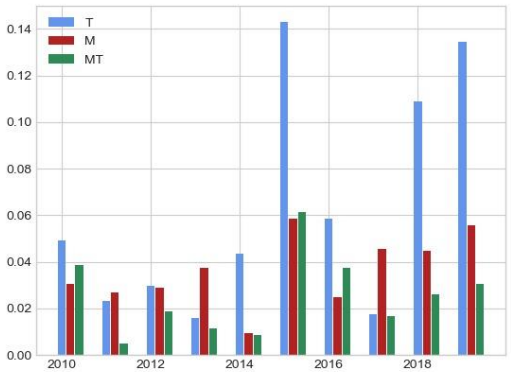

(a)




(b)




(c)




(d)

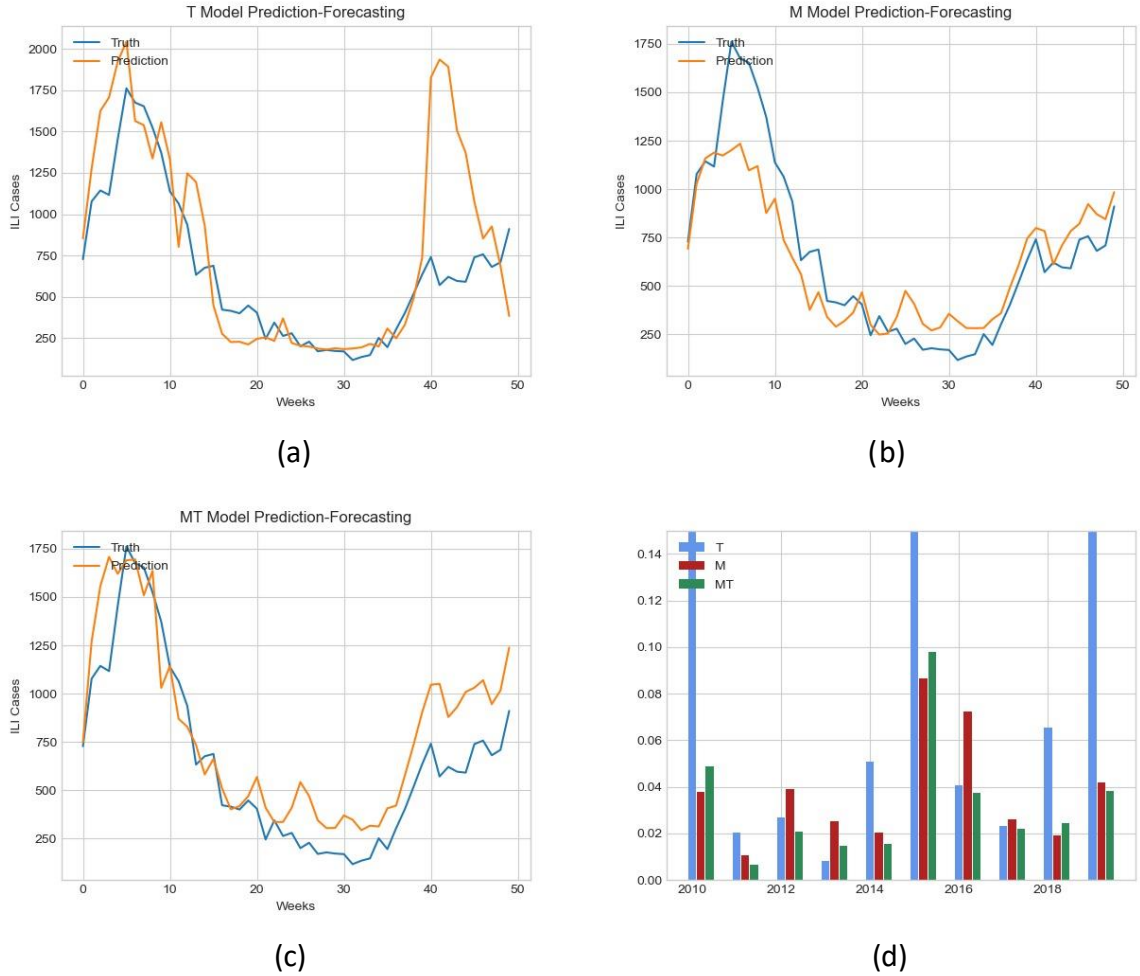
*Σχήμα 5.5*

(a) *Μοντέλο Τ : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2017*

(b) *Μοντέλο Μ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2017*

(c) *Μοντέλο ΜΤ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2017*

(d) *MSE των μοντέλων Μ,Τ και ΜΤ για τα έτη 2010-2019 (forecasting μίας εβδομάδας)*




*Σχήμα 5.6*

(a) *Μοντέλο Τ : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(b) *Μοντέλο Μ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(c) *Μοντέλο ΜΤ* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(d) *MSE των μοντέλων Μ,Τ και ΜΤ για τα έτη 2010-2019 (forecasting δύο εβδομάδων)*

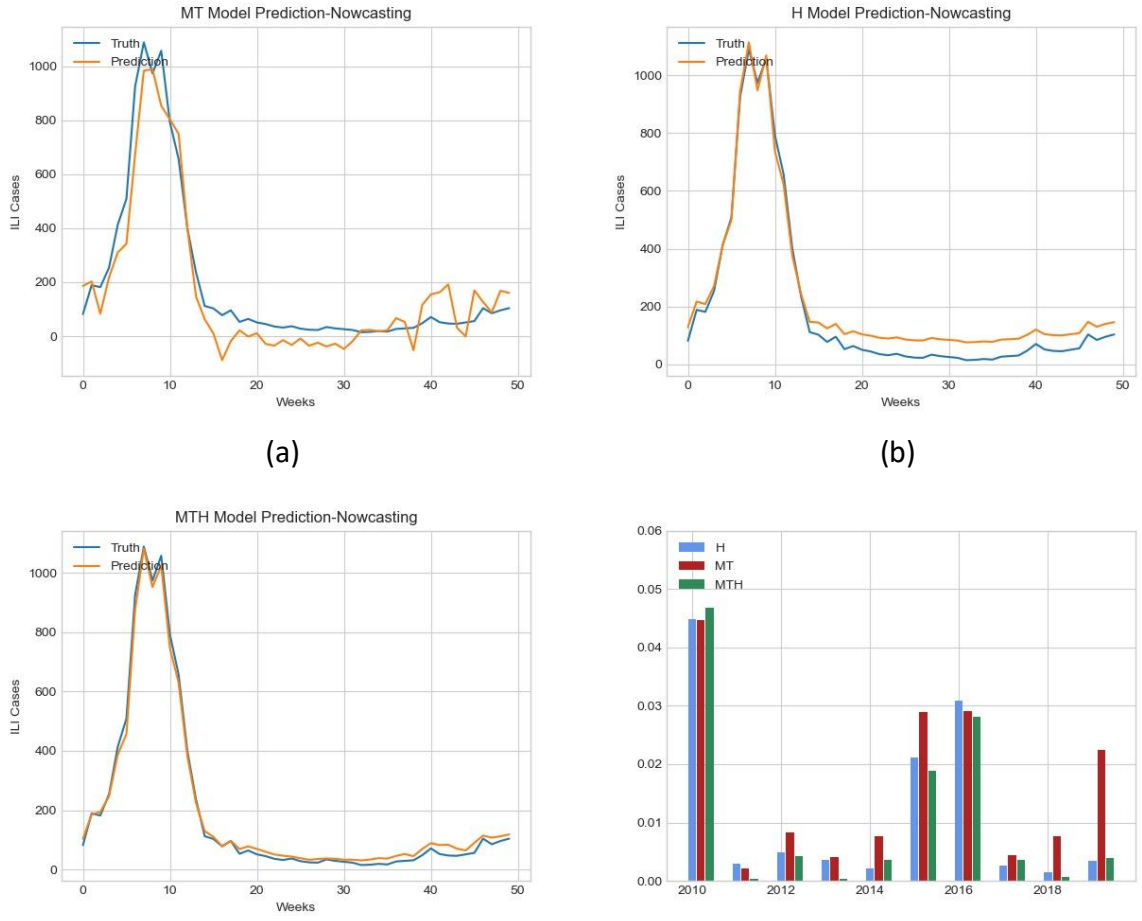
*Πίνακας 5.7 Σύγκριση ελάχιστης, μέσης και μέγιστης τιμής των κριτηρίων MSE και Pearson Correlation για τα πρωταρχικά μοντέλα M και T και το σύνθετο μοντέλο MT ανά εβδομάδα πρόβλεψης*

| MSE | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| T | 0.0189 | 0.0159 | 0.0082 | 0.0535 ± 0.0489 | 0.0624 ± 0.0480 | 0.0730 ± 0.0662 | 0.1045 | 0.1430 | 0.1769 |
| M | 0.0066 | 0.0095 | 0.0105 | 0.0227 ± 0.0164 | 0.0360 ± 0.0267 | 0.0379 ± 0.0215 | 0.0469 | **0.0583** | **0.0866** |
| MT | **0.0021** | **0.0047** | **0.0066** | **0.0159 ± 0.0142** | **0.0254 ± 0.0160** | **0.0326 ± 0.0253** | **0.0447** | 0.0614 | 0.0980 |

| COR | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| T | 0.0909 | 0.2208 | 0.2335 | 0.6578 ± 0.3391 | 0.6404 ± 0.3248 | 0.5491 ± 0.3541 | 0.9380 | 0.9156 | 0.8925 |
| M | 0.4008 | **0.5451** | **0.3910** | 0.8510 ± 0.1714 | **0.8334 ± 0.0824** | **0.7732 ± 0.1271** | 0.9824 | 0.9468 | 0.8967 |
| MT | **0.4241** | 0.4642 | 0.3591 | **0.8641 ± 0.1626** | 0.8272 ± 0.1211 | 0.7616 ± 0.1431 | **0.9833** | **0.9523** | **0.9066** |

- *MTH Μοντέλο*

Παρατηρείται πως στη μέση περίπτωση, το MTH μοντέλο υπερέχει του H κατά ένα ποσοστό, καταφέρνοντας να το βελτιώσει όσον αφορά τις μικρού πλάτους διακυμάνσεις μεταξύ των πραγματικών τιμών των κρουσμάτων και των προβλέψεων.



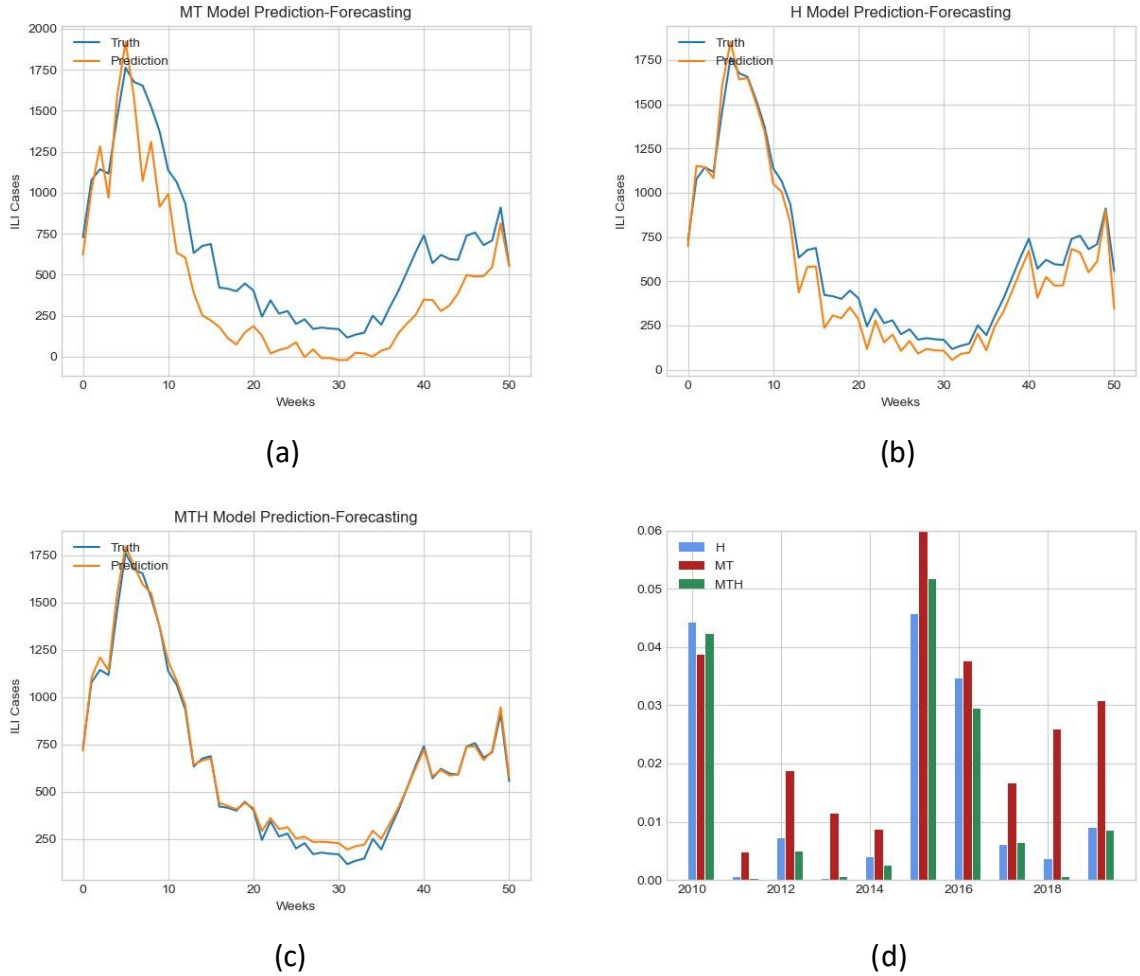

*Σχήμα 5.8*

(a) *Μοντέλο MT : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2012*

(b) *Μοντέλο H* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2012*

(c) *Μοντέλο MTH* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2012*

(d) *MSE των μοντέλων MT, H και MTH για τα έτη 2010-2019 (nowcasting)*



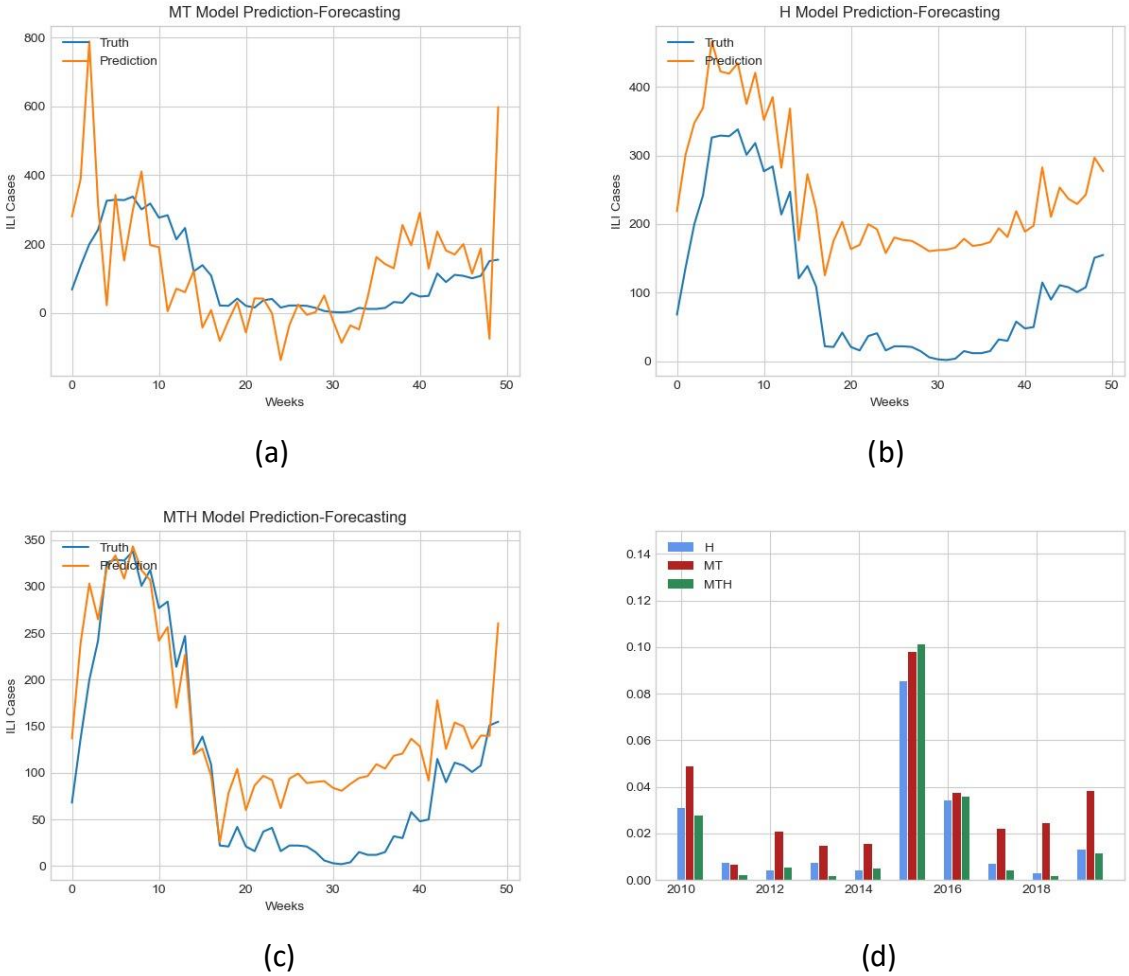

*Σχήμα 5.9*

(a) *Μοντέλο MT : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(b) *Μοντέλο H* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(c) *Μοντέλο MTH* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2018*

(d) *MSE των μοντέλων MT, H και MTH για τα έτη 2010-2019 (forecasting μίας εβδομάδας)*




*Σχήμα 5.10*

(a) *Μοντέλο MT : Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

(b) *Μοντέλο H* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

(c) *Μοντέλο MTH* : *Πραγματικός και εκτιμώμενος αριθμός κρουσμάτων για το έτος 2013*

(d) *MSE των μοντέλων MT, H και MTH για τα έτη 2010-2019 (forecasting δύο εβδομάδων)*

*Πίνακας 5.11 Σύγκριση ελάχιστης, μέσης και μέγιστης τιμής των κριτηρίων MSE και Pearson Correlation για τα μοντέλα MT, H και MTH ανά εβδομάδα πρόβλεψης*

| MSE | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| MT | 0.0021 | 0.0047 | 0.0066 | 0.0159 ± 0.0142 | 0.0254 ± 0.0160 | 0.0326 ± 0.0253 | 0.0447 | 0.0614 | 0.0980 |
| H | 0.0014 | 0.0002 | 0.0027 | 0.0118 ± 0.0117 | 0.0154 ± 0.0134 | 0.0197 ± 0.0151 | 0.0447 | 0.0456 | 0.0854 |
| MTH | **0.0002** | **0.0002** | **0.0015** | **0.0110 ± 0.0088** | **0.0146 ± 0.0122** | **0.0195 ± 0.0168** | 0.0467 | 0.0516 | 0.101 |

| COR | Min | | | Mean | | | Max | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| MT | **0.4241** | **0.4642** | 0.3591 | 0.8641 ± 0.1626 | 0.8272 ± 0.1211 | 0.7616 ± 0.1431 | 0.9833 | 0.9523 | 0.9066 |
| H | 0.2663 | 0.2598 | 0.2704 | 0.8965 ± 0.2154 | 0.8885 ± 0.0121 | 0.8801 ± 0.0189 | **0.9967** | 0.9937 | 0.9885 |
| MTH | 0.3145 | 0.3272 | **0.3936** | **0.9003 ± 0.0667** | **0.8925 ± 0.0240** | **0.8892 ± 0.0510** | 0.9962 | **0.9978** | **0.9897** |

# Κεφάλαιο 6

# Συμπεράσματα

Οι εποχιακές εξάρσεις γριπωδών συνδρομών αποτελούν μείζον πρόβλημα για την δημόσια υγεία σε όλο τον κόσμο. Είναι ικανές να προκαλέσουν από ήπια έως πολύ σοβαρή νόσηση, και τα άτομα που ανήκουν σε ομάδες υψηλού κινδύνου διατρέχουν υψηλό κίνδυνο να εμφανίσουν σοβαρές επιπλοκές εάν νοσήσουν. Επιπλέον, σε περιπτώσεις όπου τα ιικά στελέχη των ασθενειών αυτών μεταλλάσσονται, η έλλειψη ανοσίας είναι ικανή να οδηγήσει σε κατάσταση πανδημίας με εκατομμύρια θανάτους ανά τον κόσμο και τεράστιες ζημιές στον οικονομικό ή πολιτιστικό τομέα.

Ο πρώιμος εντοπισμός των μοτίβων της εξάπλωσης μιας νόσου και η γρήγορη απόκριση των κρατικών φορέων είναι καίριας σημασίας για την μείωση του κοινωνικού και οικονομικού αντίκτυπου. Οι παραδοσιακές μέθοδοι επιδημιολογικής επιτήρησης, που αξιοποιούνται από το CDC (Center for Disease Control and Prevention – ΗΠΑ) , από το EISS (European Influenza Surveillance Scheme) ή τον ΕΟΔΥ, βασίζονται σε ιολογικά και κλινικά δεδομένα, όπως ο αριθμός των ιατρικών επισκέψεων όταν κάποιος παρουσιάζει συμπτώματα ILI. Παρόλα αυτά, οι αναφορές των οργανισμών αυτών είναι έτοιμες για χρήση μετά από μία ή δύο εβδομάδες, καθιστώντας εμφανή την ανάγκη για ανάπτυξη μεθόδων οι οποίες μπορούν να εντοπίσουν ή να προβλέψουν ταχύτερα το πλήθος των κρουσμάτων της εκάστοτε νόσου.

Πλήθος ερευνητών έχει ασχοληθεί με την ενσωμάτωση ετερογενών δεδομένων και την εφαρμογή τεχνικών βαθιάς μάθησης με στόχο τη διερεύνηση της δυνατότητας αξιοποίησης των εναλλακτικών πηγών δεδομένων προς την πρόβλεψη της εξάπλωσης μεταδοτικών ασθενειών όπως οι γριπώδεις συνδρομές. Σκοπός αυτής της εργασίας ήταν η αξιοποίηση των αναρτήσεων του Twitter, των μετεωρολογικών δεδομένων και των αναφορών του συστήματος Sentinel για τη δεκαετία 2010-2019, με στόχο την πρόβλεψη των κρουσμάτων ILI με ορίζοντα πρόβλεψης έως τρεις εβδομάδες. Η διαδικασία που ακολουθήθηκε ήταν η εξής: πρώτα σχεδιάστηκαν τρία πρωταρχικά μοντέλα (Τ, Μ και Η), κάθε ένα από τα οποία λάμβανε υπόψιν έναν από τους τρεις διαφορετικούς τύπους των χρησιμοποιούμενων δεδομένων, και στη συνέχεια, διερευνήθηκε ο βέλτιστος τρόπος συγκερασμού τους, έτσι ώστε τα συνδυαστικά μοντέλα να υπερτερούν των πρωταρχικών τριών. Παρατηρήθηκε ότι ο συνδυασμός των ετερογενών δεδομένων με κατάλληλες τεχνικές μπορεί να οδηγήσει στην ανάπτυξη προγνωστικών συστημάτων του αριθμού κρουσμάτων ILI που επιτυγχάνουν ικανοποιητική ακρίβεια και θα μπορούσαν να συνεισφέρουν στη βελτιστοποίηση και την ενίσχυση των στρατηγικών επιδημιολογικής επιτήρησης.

## *6.1 Προτάσεις για μελλοντική έρευνα*

Η υψηλή συχνότητα με την οποία αναρτώνται δημοσιεύσεις στις κοινωνικές πλατφόρμες επιτρέπει την ανάλυση των διαφόρων τάσεων προς μελέτη λεπτό προς λεπτό, ενώ η πρόσβαση σε αυτές (από δημόσια προφίλ), έρχεται χωρίς κάποιο κόστος. Ωστόσο, αξιοσημείωτα εμπόδια που παρουσιάζονται κατά την αξιοποίηση δεδομένων κοινωνικών δικτύων στην επιδημιολογική επιτήρηση αφορούν την αξιοπιστία των δεδομένων αυτών καθώς και τη μορφολογική ιδιομορφία που αυτά παρουσιάζουν. Σημαντικές πληροφορίες όπως η ηλικία ενός χρήστη ή το φύλο του, πολλές φορές παραλείπονται ή είναι ψευδείς, ενώ η κατανομή των χρηστών περιλαμβάνει κυρίως νεαρότερες ηλικιακές ομάδες. Όσον αφορά το Twitter, επειδή η χρήση του είναι περιορισμένη στην Ελλάδα, θα ήταν χρήσιμη η διερεύνηση άλλων μέσων κοινωνικής δικτύωσης όπως το Facebook ή το Instagram.

Οι αναφορές του συστήματος Sentinel που χρησιμοποιήθηκαν αφορούσαν τα επιδημιολογικά δεδομένα ILI για το σύνολο της Ελλάδας, ενώ τα μετεωρολογικά δεδομένα και τα δεδομένα κοινωνικής δικτύωσης αναφέρονταν σε κάθε περιφέρεια ξεχωριστά. Για αυτό τον λόγο, τα μοντέλα έπρεπε να προσαρμοστούν έτσι ώστε να μάθουν πόσο συμβάλλει κάθε περιφέρεια στον συνολικό αριθμό κρουσμάτων. Αυτό όμως οδήγησε σε πιο αργή εκπαίδευση των μοντέλων και σε μείωση της απόδοσής τους. Έτσι, αν μελλοντικά οι αναφορές παρέχονται ανά περιφέρεια, αφενός οι προβλέψεις θα είναι πιο ακριβείς και αφετέρου, θα είναι δυνατή η εξαγωγή τους ανά γεωγραφικό διαμέρισμα.

Τέλος, η μεθοδολογία που αναπτύχθηκε για τον σχεδιασμό μοντέλων πρόβλεψης των κρουσμάτων γριπωδών συνδρομών, θα μπορούσε να αξιοποιηθεί για την ανάπτυξη προγνωστικών μοντέλων με στόχο την επιτήρηση άλλων μεταδοτικών ασθενειών ή νόσων όπως είναι αυτή του κορωνοϊού COVID-19. Η πανδημία του COVID-19 έχει εξαπλωθεί σε πάνω από 214 χώρες παγκοσμίως και έχει επηρεάσει σημαντικά κάθε πτυχή της ανθρώπινης δραστηριότητας. Δεδομένα όπως είναι οι μετεωρολογικοί δείκτες (θερμοκρασία, υγρασία) ή το περιεχόμενο των αναρτήσεων στα μέσα κοινωνικής δικτύωσης εμφανίζουν ισχυρή συσχέτιση με τον ρυθμό μετάδοσης του ιού. Έτσι, οι τεχνολογίες που εκμεταλλεύονται και συνδυάζουν την προγνωστική ισχύ ετερογενών τύπων δεδομένων είναι πιθανό να διαδραματίσουν σημαντικό ρόλο στην παγκόσμια μάχη ενάντια στις πανδημίες όπως αυτή του COVID-19.

## Βιβλιογραφία


[1] S. Munjal, S. J. Ferrando, and Z. Freyberg, “Neuropsychiatric Aspects of Infectious Diseases,” *Critical Care Clinics*, vol. 33, no. 3, pp. 681–712, Jul. 2017, doi: 10.1016/j.ccc.2017.03.007.

[2] N. I. Nii-Trebi, “Emerging and Neglected Infectious Diseases: Insights, Advances, and Challenges,” BioMed Research International, vol. 2017, pp. 1–15, 2017, doi: 10.1155/2017/5245021.

[3] “What is the Difference Between an ‘Injury’ and ‘Disease’ for Commonwealth Injury Claims? - Tindall Gask Bentley Lawyers,” Tindall Gask Bentley Lawyers, Dec. 19, 2014. https://tgb.com.au/injured-people/what-is-the-difference-between-an-%E2%80%9Cinjury%E2%80%9D-and-%E2%80%9Cdisease%E2%80%9D-for-commonwealth-injury-claims/#:~:text=With%20an%20injury%2C%20it%20is (accessed Jan. 11, 2021).

[4] A. Signore, “About inflammation and infection,” EJNMMI Research, vol. 3, no. 1, p. 8, 2013, doi: 10.1186/2191-219x-3-8.

[5] K. Todar, “Online Textbook of Bacteriology,” Textbookofbacteriology.net, 2012. http://textbookofbacteriology.net/index.html.

[6] “Standing up to infectious disease,” Nature Microbiology, vol. 4, no. 1, pp. 1–1, Dec. 2018, doi: 10.1038/s41564-018-0331-3.

[7] “Notice to Readers: Considerations for Distinguishing Influenza-Like Illness from Inhalational Anthrax,” *www.cdc.gov*. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5044a5.htm (accessed Jan. 11, 2021).

[8] L. A. Palmer, M. D. Rousculp, S. S. Johnston, P. J. Mahadevia, and K. L. Nichol, “Effect of influenza-like illness and other wintertime respiratory illnesses on worker productivity: The child and household influenza-illness and employee function (CHIEF) study,” *Vaccine*, vol. 28, no. 31, pp. 5049–5056, Jul. 2010, doi: 10.1016/j.vaccine.2010.05.011.

[9] “Global Epidemiological Surveillance Standards for Influenza.” [Online]. Available: https://www.who.int/influenza/resources/documents/WHO_Epidemiological_Influenza_Surveillance_Standards_2014.pdf.

[10] N.-A. M. Molinari *et al.*, “The annual impact of seasonal influenza in the US: Measuring disease burden and costs,” *Vaccine*, vol. 25, no. 27, pp. 5086–5096, Jun. 2007, doi: 10.1016/j.vaccine.2007.03.046.

[11] “U.S. Influenza Surveillance System: Purpose and Methods,” 2019. https://www.cdc.gov/flu/weekly/overview.htm.

[12] “European Influenza Surveillance Network (EISN),” *European Centre for Disease Prevention and Control*. https://www.ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/eisn (accessed Jan. 11, 2021).

[13] R. Ferland and S. Froda, “A statistical tool for comparing seasonal ILI surveillance data,” *Scientific Reports*, vol. 9, no. 1, Feb. 2019, doi: 10.1038/s41598-018-38292-x.

[14] B. Cakici, “Disease surveillance systems,” *kth.diva-portal.org*, 2011, Accessed: Jan. 11, 2021. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-33661.

[15] J. D. Tamerius *et al.*, “Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates,” *PLoS Pathogens*, vol. 9, no. 3, p. e1003194, Mar. 2013, doi: 10.1371/journal.ppat.1003194.

[16] A. C. Lowen and J. Steel, “Roles of Humidity and Temperature in Shaping Influenza Seasonality,” *Journal of Virology*, vol. 88, no. 14, pp. 7692–7695, Jul. 2014, doi: 10.1128/JVI.03544-13.

[17] Q. Guo *et al.*, “The effects of meteorological factors on influenza among children in Guangzhou, China,” *Influenza and Other Respiratory Viruses*, vol. 13, no. 2, pp. 166–175, Dec. 2018, doi: 10.1111/irv.12617.

[18] J. E. Moan, A. Dahlback, L. Ma, and A. Juzeniene, “Influenza, solar radiation and vitamin D,” *Dermato-Endocrinology*, vol. 1, no. 6, pp. 308–310, Nov. 2009, doi: 10.4161/derm.1.6.11357.

[19] D. Masters and C. Luschi, “Revisiting Small Batch Training for Deep Neural Networks,” arXiv.org, 20-Apr-2018. [Online]. Available: https://arxiv.org/abs/1804.07612. [Accessed: 11-Jan-2021].

[20] Y. Bengio, “Practical Recommendations for Gradient-Based Training of Deep Architectures,” *Lecture Notes in Computer Science*, pp. 437–478, 2012, doi: 10.1007/978-3-642-35289-8_26.

[21] S. Ruder, “An overview of gradient descent optimization algorithms,” Sebastian Ruder, Jan. 19, 2016. https://ruder.io/optimizing-gradient-descent/.

[22] A. Khodabakhsh, I. Ari, M. Bakır, and S. M. Alagoz, “Forecasting Multivariate Time-Series Data Using LSTM and Mini-Batches,” Data Science: From Research to Application, pp. 121–129, 2020, doi: 10.1007/978-3-030-37309-2_10.

[23] K. Bandara, C. Bergmeir, and H. Hewamalage, “LSTM-MSNet: Leveraging Forecasts on Sets of Related Time Series With Multiple Seasonal Patterns,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–14, 2020, doi: 10.1109/tnnls.2020.2985720.

[24] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks.” [Online]. Available: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf?hc_location=ufi.

[25] S. C. Kramer and J. Shaman, “Development and validation of influenza forecasting for 64 temperate and tropical countries,” *PLOS Computational Biology*, vol. 15, no. 2, p. e1006742, Feb. 2019, doi: 10.1371/journal.pcbi.1006742.

[26] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, May 2017.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv.org, 2013. https://arxiv.org/abs/1301.3781

[28] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On Spectral Clustering : Analysis and an Algorithm.”

[29] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, “Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance,” *PLOS Computational Biology*, vol. 11, no. 10, p. e1004513, Oct. 2015, doi: 10.1371/journal.pcbi.1004513.

[30] H. Woo, Y. Cho, E. Shim, J.-K. Lee, C.-G. Lee, and S. H. Kim, “Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea,” *Journal of Medical Internet Research*, vol. 18, no. 7, p. e177, Jul. 2016, doi: 10.2196/jmir.4955.

[31] S. Volkova, E. Ayton, K. Porterfield, and C. D. Corley, “Forecasting influenza-like illness dynamics for military populations using neural networks and social media,” *PLOS ONE*, vol. 12, no. 12, p. e0188941, Dec. 2017, doi: 10.1371/journal.pone.0188941.

[32] Q. Xu, Y. R. Gel, L. L. Ramirez Ramirez, K. Nezafati, Q. Zhang, and K.-L. Tsui, “Forecasting influenza in Hong Kong with Google search queries and statistical model fusion,” PLOS ONE, vol. 12, no. 5, p. e0176690, May 2017, doi: 10.1371/journal.pone.0176690.

[33] O. Gencoglu and M. Ermes, “Predicting the Flu from Instagram,” arXiv:1811.10949 [cs, stat], Nov. 2018, Accessed: Jan. 11, 2021. [Online]. Available: https://arxiv.org/abs/1811.10949.

[34] B. M. Althouse *et al.*, “Enhancing disease surveillance with novel data streams: challenges and opportunities,” *EPJ Data Science*, vol. 4, no. 1, Oct. 2015, doi: 10.1140/epjds/s13688-015-0054-0.

[35] N. A. Christakis and J. H. Fowler, “Social Network Sensors for Early Detection of Contagious Outbreaks,” PLoS ONE, vol. 5, no. 9, p. e12948, Sep. 2010, doi: 10.1371/journal.pone.0012948.

[36] A. Alessa and M. Faezipour, “A review of influenza detection and prediction through social networking sites,” Theoretical Biology and Medical Modelling, vol. 15, no. 1, Feb. 2018, doi: 10.1186/s12976-017-0074-5.

[37] A. Sadilek, H. Kautz, and V. Silenzio, “Modeling Spread of Disease from Social Interactions.”

[38] C. Viboud, O. N. Bjornstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell, “Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza,” Science, vol. 312, no. 5772, pp. 447–451, Mar. 2006, doi: 10.1126/science.1125237.

[39] T. Sugawara, Y. Ohkusa, Y. Ibuka, H. Kawanohara, K. Taniguchi, and N. Okabe, “Real-time Prescription Surveillance and its Application to Monitoring Seasonal Influenza Activity in Japan,” Journal of Medical Internet Research, vol. 14, no. 1, p. e14, Jan. 2012, doi: 10.2196/jmir.1881.

[40] K. S. Hickmann et al., “Forecasting the 2013–2014 Influenza Season Using Wikipedia,” PLOS Computational Biology, vol. 11, no. 5, p. e1004239, May 2015, doi: 10.1371/journal.pcbi.1004239.

[41] M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, and J. S. Brownstein, “Using Clinicians’ Search Query Data to Monitor Influenza Epidemics,” Clinical Infectious Diseases, vol. 59, no. 10, pp. 1446–1450, Aug. 2014, doi: 10.1093/cid/ciu647.

[42] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” Nature, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009, doi: 10.1038/nature07634.

[43] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter Improves Influenza Forecasting,” PLoS Currents, 2014, doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[44] A. Burns and S. Iliffe, “Alzheimer’s disease,” BMJ, vol. 338, no. feb05 1, pp. b158–b158, Feb. 2009, doi: 10.1136/bmj.b158.

[45] P. F. Adams and M. A. Marano, “Current estimates from the National Health Interview Survey, 1994,” Vital and Health Statistics. Series 10, Data from the National Health Survey, no. 193 Pt 1, pp. 1–260, Dec. 1995, Accessed: Jan. 11, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15789506/.

[46] M. Keech and P. Beardsworth, “The impact of influenza on working days lost: a review of the literature,” PharmacoEconomics, vol. 26, no. 11, pp. 911–24, 2008, doi: 10.2165/00019053-200826110-00004.

[47] “Economic and Social Impact of Epidemic and Pandemic Influenza,” Vaccine, vol. 24, no. 44–46, pp. 6776–6778, Nov. 2006, doi: 10.1016/j.vaccine.2006.06.072.

[48] T. M. Mitchell, *The discipline of machine learning*. Pittsburgh, Pa.: Carnegie Mellon University, School Of Computer Science, Machine Learning Dept, 2006.

[49] (*Coursera | Online Courses From Top Universities. Join for Free*, 2019) https://www.coursera.org/learn/machine-learning Andrew Ng

[50] “Botometer by OSoMe,” botometer.iuni.iu.edu. https://botometer.osome.iu.edu/faq.

[51] 1615 L. St NW, Suite 800 Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Twitter Bots: An Analysis of the Links Automated Accounts Share,” Pew Research Center: Internet, Science & Tech, Apr. 09, 2018. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/.

[52] Ethem Alpaydin, *Introduction To Machine Learning.* S.L.: Mit Press, 2020.

[53] T. Mitchell, Machine Learning, McGraw-Hill, 1997

[54] I. Sutskever, “Training Recurrent Neural Networks,” 2013.

[55] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[56] “Exploring LSTMs,” blog.echen.me. http://blog.echen.me/2017/05/30/exploring-lstms/.

[57] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[58] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, “Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network,” IEEE Transactions on Big Data, vol. 2, no. 1, pp. 85–99, Mar. 2016, doi: 10.1109/tbdata.2016.2546301.

[59] F. Abel, C. Hauff, Geert-Jan Houben, and Ke Tao, “Twitcident: Fighting fire with information from Social Web streams,” *ResearchGate*, Apr. 16, 2012. https://www.researchgate.net/publication/254008774_Twitcident_Fighting_fire_with_information_from_Social_Web_streams (accessed Jan. 11, 2021).

[60] C. Corley, D. Cook, A. Mikler, and K. Singh, “Text and Structural Data Mining of Influenza Mentions in Web and Social Media,” *International Journal of Environmental Research and Public Health*, vol. 7, no. 2, pp. 596–615, Feb. 2010, doi: 10.3390/ijerph7020596.

[61] “Introduction: Epidemiology in crises,” *conflict.lshtm.ac.uk*. http://conflict.lshtm.ac.uk/page_06.htm (accessed Jan. 11, 2021).

[62] J. J. CANNELL *et al.*, “Epidemic influenza and vitamin D,” *Epidemiology and Infection*, vol. 134, no. 6, pp. 1129–1140, Sep. 2006, doi: 10.1017/s0950268806007175.

[63] “The known health effects of UV,” *WHO*. https://www.who.int/uv/resources/FAQ/uvhealtfac/en/index1.html.

[64] W. Liu *et al.*, “Influenza activity prediction using meteorological factors in a warm temperate to subtropical transitional zone, Eastern China,” *Epidemiology and Infection*, vol. 147, 2019, doi: 10.1017/s0950268819002140.

[65] K. Su, Liang Xu, G. Li, and Y. Li, “Forecasting influenza activity using self-adaptive AI model and multi-source data in Chongqing, China,” ResearchGate, Aug. 2019.

https://www.researchgate.net/publication/335510823_Forecasting_influenza_activity_using_self-adaptive_AI_model_and_multi-source_data_in_Chongqing_China/link/5db1e8164585155e27f91479/download (accessed Jan. 14, 2021).

[66] M. Mousavi, R. Salgotra, D. Holloway, and A. H. Gandomi, "COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features," IEEE Computational Intelligence Magazine, vol. 15, no. 4, pp. 34–50, Nov. 2020, doi: 10.1109/mci.2020.3019895.

[67] "GetOldTweets3," *PyPI*, Nov. 27, 2019. https://pypi.org/project/GetOldTweets3/ (accessed Jan. 16, 2021).