ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# Εκτίμηση της εμπιστοσύνης των πολιτών προς δημόσιους οργανισμούς υγείας, με χρήση τεχνικών εξόρυξης δεδομένων από μέσα κοινωνικής δικτύωσης και αλγορίθμων βαθιάς μηχανικής μάθησης για επεξεργασία φυσικού λόγου

Πέτρος-Φλώριος Μπάκαλος

Φλώριος Γλέζος

Επιβλέπων: Δουλάμης Νικόλαος
Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Οκτώβριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# Εκτίμηση της εμπιστοσύνης των πολιτών προς δημόσιους οργανισμούς υγείας, με χρήση τεχνικών εξόρυξης δεδομένων από μέσα κοινωνικής δικτύωσης και αλγορίθμων βαθιάς μηχανικής μάθησης για επεξεργασία φυσικού λόγου

Πέτρος-Φλώριος Μπάκαλος

Φλώριος Γλέζος

Υπεύθυνος: Δουλάμης Νικόλαος
Αναπληρωτής Καθηγητής ΕΜΠ

| Δουλάμης Νικόλαος | Δουλάμης Αναστάσιος | Πρωτονοτάριος Εμμανουήλ |
|---|---|---|
| Αναπληρωτής Καθηγητής ΕΜΠ | Αναπληρωτής Καθηγητής ΕΜΠ | Ομότιμος Καθηγητής ΕΜΠ |

Αθήνα, Οκτώβριος 2021

## Περίληψη

Το περιεχόμενο της παρούσας διπλωματικής εργασίας είναι η έρευνα για την ανάπτυξη ενός εργαλείου που θα είναι χρήσιμο, σε περιόδους που υπάρχει μία μεγάλη υγειονομική κρίση, όπως μια πανδημία. Τους τελευταίους μήνες, ο πλανήτης ολόκληρος βρίσκεται μπροστά σε μία πρωτόγνωρη κατάσταση. Η εξάπλωση του SARS-Cov-2 στοίχησε πολλές ανθρώπινες ζωές αλλά είχε και αντίκτυπο σε κάθε είδους ανθρώπινη δραστηριότητα. Το οικονομικό και κοινωνικό πλήγμα που συνεχίζει να αφήνει η πανδημία είναι τεράστιο. Κυβερνήσεις, φορείς υγείας κι άλλοι δημόσιοι φορείς αναγκάστηκαν πολλές φορές να λαμβάνουν αποφάσεις σε πολύ μικρό χρονικό διάστημα, χωρίς να είναι δυνατό πάντα να υπολογίζονται όλες οι συνέπειες. Το διάστημα αυτό, στο οποίο εφαρμόστηκε το μέτρο της καραντίνας σε όλες τις γωνιές του πλανήτη και τα ταξίδια περιορίστηκαν στο ελάχιστο, η παρουσία του κόσμου στα μέσα κοινωνικής δικτύωσης ήταν ακόμη μεγαλύτερη. Στόχος της εργασίας αυτής, είναι να αξιοποιήσει την κίνηση αυτή στα μέσα κοινωνικής δικτύωσης, ώστε να βοηθήσει σε παρόμοιες υγειονομικές κρίσεις. Μέσω της έρευνας που έγινε, δίνεται η δυνατότητα για ανάλυση συναισθήματος των πολιτών (sentiment analysis) μέσω διάφορων κοινωνικών δικτύων, ώστε να εκτιμηθεί η εμπιστοσύνη τους προς τους δημόσιους φορείς, σχετικά με μέτρα και αποφάσεις που λαμβάνονται σε περίοδο πανδημίας.

Λέξεις κλειδιά: Μηχανική Μάθηση, Επεξεργασία Φυσικού Λόγου, Ανάλυση Συναισθήματος, Υγειονομική Κρίση, Δημόσιοι Φορείς και Φορείς Υγείας

## Abstract

The content of this dissertation is research to develop a tool that will be useful in times of great health crisis, such as a pandemic. In recent months, the entire planet is facing an unprecedented situation. The spread of SARS-Cov-2 cost many lives but also had an impact on all kinds of human activity. The economic and social blow that the pandemic continues to inflict is enormous. Governments, health agencies and other public bodies have often been forced to make decisions in a very short time, without it being always possible to calculate all the consequences. During this period, during which the quarantine measure was applied in all corners of the globe and travel was reduced to a minimum, the presence of people on social media was even greater. The aim of this work is to take advantage of this move on social media to help in similar health crises. Through the research that was done, it is possible to analyze citizens' feelings (sentiment analysis) through various social networks, in order to assess their trust in public bodies, regarding measures and decisions taken during a pandemic.

Keywords: Machine Learning, Natural Language Processing, Sentiment Analysis, Health Crisis, Public and Health Bodies

# Περιεχόμενα

# Εικόνες

Κεφάλαιο 1

# Εισαγωγή

## 1.1 Κίνητρα

Στην καθημερινή ζωή, οι άνθρωποι λαμβάνουν αποφάσεις και αναπτύσσουν συμπεριφορές με βάση τις σκέψεις και τις απόψεις των άλλων. Με την εκρηκτική ανάπτυξη πηγών πλούσιων σε γνώμες, όπως ιστολόγια, διαδικτυακά φόρουμ, κοινωνικά μέσα και micro-blogging ιστοσελίδες, προκύπτουν νέες ευκαιρίες και προκλήσεις καθώς οι άνθρωποι μπορούν τώρα να χρησιμοποιούν ενεργά τεχνολογίες πληροφοριών για να αναζητήσουν και να αξιολογήσουν τις απόψεις άλλων. Τέτοιες πηγές περιέχουν δεδομένα απόψεων για οποιαδήποτε υπηρεσία, προϊόν, θέμα, συμβάν και ιδέα. Έτσι, οι προαναφερθείσες πηγές μπορούν να χρησιμοποιηθούν αποτελεσματικά για την εξαγωγή πολύτιμων πληροφοριών από αυτές. Οι επιχειρήσεις και οι οργανισμοί δεν χρειάζεται πλέον να διεξάγουν έρευνες ή δημοσκοπήσεις για τη συλλογή των απόψεων των καταναλωτών ή της κοινής γνώμης, καθώς η αφθονία των διαθέσιμων διαδικτυακών δεδομένων καθιστά πολύ πιο εύκολη τη συλλογή όλων των πληροφοριών που απαιτούνται για την κατανόηση του κοινωνικού συναισθήματος περί μιας επωνυμίας, είτε πρόκειται για υπηρεσία ή προϊόν.

Το Twitter είναι μια δημοφιλής υπηρεσία micro-blogging και κοινωνικής δικτύωσης που, την τελευταία δεκαετία, έχει γίνει μια από τις πλουσιότερες πηγές δεδομένων απόψεων. Οι χρήστες του Twitter δημιουργούν διαδικτυακές δημοσιεύσεις, που ονομάζονται "tweets", προκειμένου να μοιράζονται απόψεις, σκέψεις και συναισθήματα για οποιοδήποτε θέμα τους ενδιαφέρει. Τα Tweets είναι σύντομα μηνύματα, που περιορίζονται σε έναν σχετικά μικρό αριθμό χαρακτήρων, ώστε να μπορούν να θεωρηθούν ως πρόταση ή τίτλος παρά ως έγγραφο. Η γλώσσα που χρησιμοποιείται στα tweets είναι αρκετά ανεπίσημη και περιέχει μια ποικιλία από “αλλιώτικες” λέξεις, ορθογραφικά λάθη, emoticon, σημεία στίξης, διευθύνσεις URL, δημιουργική ορθογραφία και ειδική ορολογία ή συντομογραφίες, όπως "re-tweet" (RT) και hashtag (#). Μια άλλη πτυχή των tweets είναι ότι περιλαμβάνουν δομημένες πληροφορίες σχετικά με τους χρήστες που εμπλέκονται στην επικοινωνία, όπως ποιος ακολουθεί ποιον.

Η εξόρυξη απόψεων και συναισθημάτων από αυτές τις διαδικτυακές πηγές είναι πολύ δύσκολη λόγω του τεράστιου όγκου πληροφοριών που δημιουργούνται καθημερινά. Κάθε πηγή περιέχει συνήθως έναν τεράστιο όγκο δεδομένων κειμένου που είναι σχεδόν αδύνατο για τον μέσο άνθρωπο να αποκρυπτογραφήσει και να συνοψίσει τις απόψεις σε αυτά. Για το λόγο αυτό, οι ερευνητές έχουν αρχίσει να ερευνούν και να αναπτύσσουν συστήματα που μπορούν να

ανιχνεύσουν ή να προβλέψουν αυτόματα το συναίσθημα σε κείμενο και να εξορύξουν αποτελεσματικά τις γνώμες, ακόμη και από μια τεράστια ποσότητα δεδομένων. Το πεδίο της έρευνας στην εξόρυξη κειμένου που σκοπεύει να προσδιορίσει τις γνώμες, τις σκέψεις και τις οπτικές των ανθρώπων που εκφράζονται σε ένα κομμάτι κειμένου ονομάζεται "ανάλυση συναισθημάτων". Η ανάλυση συναισθημάτων (ή η εξόρυξη γνώμης) εφαρμόζει ένα συνδυασμό στατιστικών στοιχείων, επεξεργασίας φυσικής γλώσσας (NLP) και μηχανικής μάθησης (ML) για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών από αρχεία κειμένου. Συγκεκριμένα, το Twitter Sentiment Analysis (TSA) είναι μια τεχνική που αντιμετωπίζει το πρόβλημα της ανάλυσης των tweets ως προς τα συναισθήματα που εκφράζουν. Με άλλα λόγια, η TSA περιλαμβάνει τον διαχωρισμό των tweets και του περιεχομένου αυτών των εκφράσεων.

Σε αυτήν τη διπλωματική εργασία, εστιάζουμε στην εκτίμηση και την εμπιστοσύνη των χρηστών του Twitter και του Reddit προς τους δημόσιους οργανισμούς υγείας. Η πρωτοφανής υγειονομική κρίση των τελευταίων χρόνων έδειξε σε όλους ότι ο πλανήτης δεν ήταν κατάλληλα προετοιμασμένος. Οι κυβερνήσεις όλων των χωρών κλήθηκαν να πάρουν αποφάσεις σε πολύ μικρό χρονικό διάστημα, οι οποίες θα είχαν τεράστιο αντίκτυπο στη ζωή εκατομμυρίων ανθρώπων. Η ανάλυση της κίνησης και των δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης θα μπορούσε να φανεί πολύ χρήσιμη, ώστε να λαμβάνονται ορθότερες αποφάσεις και να διορθώνονται άλλες σε σχεδόν πραγματικό χρόνο.

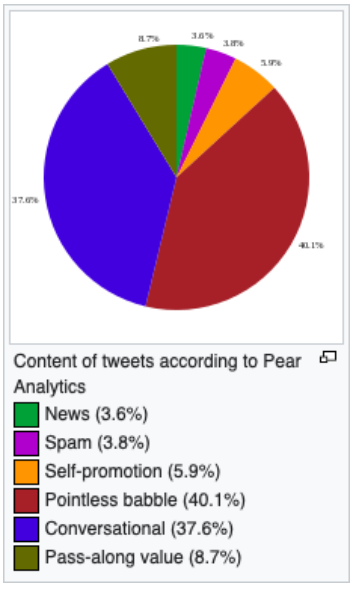
## 1.2 Twitter

Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης στο οποίο οι χρήστες δημοσιεύουν και αλληλεπιδρούν με μηνύματα γνωστά ως "tweets". Οι εγγεγραμμένοι χρήστες μπορούν να δημοσιεύουν, να κάνουν like και να αναδημοσιεύουν tweets. Μη εγγεγραμμένοι χρήστες έχουν πρόσβαση στο Twitter αλλά μπορούν να διαβάζουν μόνο αυτά που είναι δημόσια. Το Twitter διατίθεται σαν web και mobile εφαρμογή ενώ παρέχεται και η δυνατότητα μέσω προγραμματισμού μέσω των API του. Τα tweets έχουν περιορισμένο αριθμό χαρακτήρων, αρχικά ήταν 140 χαρακτήρες αλλά αργότερα το όριο διπλασιάστηκε σε 280. Επίσης, παρέχεται η δυανότητα προσθήκης ήχου και βίντεο περιορισμένης διάρκειας. Είναι ένα από τα μέσα κοινωνικής δικτύωσης με τη μεγαλύτερη κίνηση. Λόγω του περιορισμένου αριθμού χαρακτήρων που επιτρέπονται σε κάθε δημοσίευση, χρησιμοποιείται συχνά από διάσημα

πρόσωπα και πολιτικούς αλλά και οργανισμούς και φορείς για ανακοινώσεις. Επίσης ένα άλλο χαρακτηριστικό είναι ότι πολλές φορές αναπτύσσονται συζητήσεις με τη μορφή των tweets.

## 1.3 Reddit

Το Reddit είναι ένα μέσο κοινωνικής δικτύωσης όπου τα εγγεγραμμένα μέλη υποβάλλουν περιεχόμενο στον ιστότοπο, όπως συνδέσμους, αναρτήσεις κειμένου, εικόνες και βίντεο, τα οποία στη συνέχεια ψηφίζονται από τα άλλα μέλη. Οι αναρτήσεις οργανώνονται ανά θέμα σε πίνακες που δημιουργούνται από τους χρήστες που ονομάζονται "κοινότητες" ή "subreddits", οι οποίοι καλύπτουν μια ποικιλία θεμάτων όπως ειδήσεις, πολιτική, θρησκεία, επιστήμη, ταινίες, βιντεοπαιχνίδια, μουσική, βιβλία, αθλήματα, γυμναστήριο, μαγείρεμα, κατοικίδια ζώα και κοινή χρήση εικόνων. Οι αναρτήσεις με τις περισσότερες ψήφους εμφανίζονται στο επάνω μέρος της δημοσίευσης. Αν και υπάρχουν αυστηροί κανόνες που απαγορεύουν την παρενόχληση, εξακολουθεί να συμβαίνει και οι διαχειριστές του Reddit προσπαθούν να ηρεμούν τις κοινότητες, τις κλείνουν ή τις περιορίζουν κατά καιρούς.




*Εικόνα 1 Ανάλυση του περιεχομένου των tweets*

## 1.4 Research Objectives & Contributions

Οι περισσότερες από τις προσεγγίσεις State-Of-The-Art (SOTA) σχετικά με την αυτόματη ανάλυση συναισθημάτων κειμένων που συλλέγονται από κοινωνικά δίκτυα και υπηρεσίες micto-blogging επικεντρώνονται κυρίως στην ταξινόμηση δυαδικών ή τριμερών

συναισθημάτων. Με άλλα λόγια, ταξινομούν τα κείμενα σε "θετικά" και "αρνητικά" ή σε "θετικά", "αρνητικά" και "ουδέτερα". Σε αυτήν τη διπλωματική εργασία, στοχεύουμε να προχωρήσουμε βαθύτερα στην ταξινόμηση των κειμένων που συλλέγονται από το Twitter και το Reddit και να τα ταξινομήσουμε σε εξαιρετικά αρνητικό, αρνητικό, ουδέτερο, θετικό και πολύ θετικό, με βάση το συναίσθημα προς τους δημόσιους οργανισμούς υγείας. Για το σκοπό αυτό, ερευνούμε μεθόδους που μπορούν να βοηθήσουν τους αλγόριθμους να αντιμετωπίσουν αποτελεσματικά το θέμα της ανάλυσης συναισθημάτων σε πολλαπλές κατηγορίες.

Πιο συγκεκριμένα, χρησιμοποιούμε εποπτευόμενα μοντέλα βαθιάς μάθησης, τα οποία εκπαιδεύονται να επισημαίνουν δημοσιεύσεις με βάση το εκφραστικό τους συναίσθημα. Στη συνέχεια, τα μοντέλα δοκιμάζονται και αξιολογούνται για την απόδοσή τους σε διάφορες μετρήσεις αξιολόγησης. Για να εκπαιδεύσουμε και να αξιολογήσουμε τα μοντέλα, χρησιμοποιούμε ένα μη ισορροπημένα σχολιασμένο σύνολο δημοσιεύσεων που σχετίζονται με δημόσιους οργανισμούς υγείας. Προκειμένου να βρούμε τη γνώμη των ανθρώπων σχετικά με τους δημόσιους οργανισμούς υγείας, συλλαμβάνουμε και ταξινομούμε posts που σχετίζονται με δημόσιους οργανισμούς υγεία. Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα της ανάλυσης συναισθημάτων που πραγματοποιήθηκε και τα τρέχοντα επίπεδα εκτίμησης και εμπιστοσύνης προς τους δημόσιους οργανισμούς υγείας.

## 1.5 Δομή διπλωματικής εργασίας

Στο κεφάλαιο 2, συζητάμε την προηγούμενη έρευνα στον τομέα της ανάλυσης συναισθημάτων, και συγκεκριμένα την TSA. Επιπλέον, παρουσιάζουμε μελέτες που επικεντρώνονται στην αντίληψη του κοινού για δημόσιους οργανισμούς υγείας.

Στο κεφάλαιο 3, παρουσιάζουμε τον τομέα της μηχανικής μάθησης και παρέχουμε τις γνώσεις που απαιτούνται για τα επόμενα κεφάλαια. Ειδικότερα, μετά την παρουσίαση των τύπων μάθησης σε αυτόν τον τομέα, εστιάζουμε στα προβλήματα ταξινόμησης και τα βασικά βήματα που απαιτούνται για την επίλυση και την αξιολόγηση αυτών των ειδών προβλημάτων.

Στο κεφάλαιο 4, παρουσιάζουμε στον αναγνώστη τα τεχνητά νευρωνικά δίκτυα και γενικά τη βαθιά μάθηση. Πιο συγκεκριμένα, παρουσιάζουμε τους τύπους μοντέλων βαθιάς μάθησης που χρησιμοποιούνται κυρίως στη διπλωματική εργασία, δηλαδή επαναλαμβανόμενα νευρωνικά δίκτυα, και εξηγούμε τις έννοιες των μηχανισμών προσοχής και της μεταφοράς μάθησης.

Στο κεφάλαιο 5, παρουσιάζουμε το υπόβαθρο επεξεργασίας φυσικής γλώσσας που απαιτείται για την κατανόηση αυτού του αντικειμένου. Πρώτον, παρουσιάζουμε εν συντομία το έργο της

ανάλυσης συναισθημάτων. Στη συνέχεια, αφού εξερευνήσουμε δημοφιλείς τεχνικές προεπεξεργασίας κειμένου, παρουσιάζουμε τη μοντελοποίηση γλωσσών, αρχικά με τη μορφή μοντέλου γλώσσας n-gram και στη συνέχεια ως μοντέλο νευρωνικής γλώσσας. Τέλος, εξηγούμε τις μεθόδους ενσωμάτωσης και μεταφοράς μεθόδων μάθησης που χρησιμοποιούνται σήμερα για την εκπαίδευση μοντέλων επεξεργασίας φυσικής γλώσσας.

Στο κεφάλαιο 6, ο κύριος στόχος μας είναι να παρουσιάσουμε τη δουλειά μας σχετικά με την ανάλυση συναισθημάτων στο Twitter σε πολλαπλές κατηγορίες. Ξεκινάμε με μια περιγραφή των σετ δεδομένων πάνω στα οποία εκπαιδεύουμε τα μοντέλα μας, καθώς και μια επεξηγηματική ανάλυση δεδομένων. Στη συνέχεια, περιγράφουμε τα μοντέλα που χρησιμοποιήσαμε στα πειράματά μας και τα αντίστοιχα αποτελέσματα. Επιπλέον, διεξάγουμε ανάλυση συναισθημάτων πολλαπλών κατηγοριών σε tweets που δεν έχουν επισημανθεί σχετικά με δημόσιους οργανισμούς υγείας.

Τέλος, το κεφάλαιο 7 περιέχει το συμπέρασμα όπου συνοψίζουμε τα ευρήματά μας και συζητάμε για μελλοντικές κατευθύνσεις.

Κεφάλαιο 2

# Σχετική δουλειά

## 2.1 Ανάλυση συναισθημάτων Twitter

Με την ανάπτυξη των υπηρεσιών κοινωνικής δικτύωσης και micro-blogging, οι άνθρωποι άρχισαν να συζητούν ανοιχτά τις απόψεις και τις σκέψεις τους στο διαδίκτυο. Η πλατφόρμα Twitter αποτέλεσε αντικείμενο πολύ πρόσφατης έρευνας ανάλυσης συναισθημάτων, καθώς τα tweets εκφράζουν συχνά τη γνώμη ενός χρήστη για ένα θέμα ενδιαφέροντος. Οι ερευνητές έχουν χρησιμοποιήσει πληροφορίες που προέρχονται από την TSA για να εξηγήσουν και να προβλέψουν τις πωλήσεις προϊόντων (1), τις κινήσεις του χρηματιστηρίου (2) και τα αποτελέσματα των πολιτικών εκλογών (3), (4), (5), (6), (7). Επίσης, η TSA έχει χρησιμοποιηθεί για την κατανόηση της γνώμης των χρηστών για διάφορα επιχειρηματικά και κοινωνικά ζητήματα, όπως μια επωνυμία προϊόντος (8), η παραγωγή πυρηνικής ενέργειας (9) και οι υποψηφιότητες για προεδρικές εκλογές (10).

Μερικές από τις έρευνες σχετικά με τα tweets επικεντρώνονται στα χαρακτηριστικά τους που καθιστούν την TSA προκλητική, όπως η συντομία των tweets και η προκύπτουσα συμπαγής, νέα γλώσσα με στοιχεία επικοινωνίας ειδικά για το Twitter (5), (11), μια ισχυρή ανισορροπία κατηγορίας συναισθημάτων (12), (13) και δημιουργία tweet βάσει ροής (14), (15). Επιπλέον, ορισμένοι ερευνητές έχουν ασχοληθεί με τη μορφή των δεδομένων, τη χρήση της αργκό και πώς αυτά εξελίσσονται με την πάροδο του χρόνου, τη χρήση emoticon και τη φύση των ίδιων των tweets (16), (17).

Το πρόβλημα ταξινόμησης της πολικότητας των συναισθημάτων μοντελοποιείται συχνά ως ταξινόμηση κειμένου αμφίδρομης (δυαδικής) ή τριών κατευθύνσεων (τριμερής). Λίγες έρευνες έχουν διεξαχθεί για το έργο ταξινόμησης συναισθημάτων πολλαπλών τάξεων. Σημειώστε ότι η ταξινόμηση πολλαπλών κλάσεων αναφέρεται συμβατικά στην απόδοση ενός από τα πολλά δυνατά συναισθήματα σε ένα κείμενο ή ένα tweet. Οι περισσότερες από τις μελέτες επικεντρώθηκαν στην αξιολόγηση της δύναμης του συναισθήματος σε διαφορετικά επίπεδα ισχύος συναισθημάτων (π.χ. «πολύ αρνητικό», «αρνητικό», «ουδέτερο», «θετικό» και «πολύ θετικό») ή απλώς δίνουν βαθμολογίες που κυμαίνονται από −1 έως 1 στα κείμενα, δείχνοντας ταυτόχρονα την πολικότητα και τη δύναμη του συναισθήματος (18), (19). Ορισμένοι ερευνητές έχουν αξιολογήσει μοντέλα συναισθημάτων πέντε κατηγοριών σε TSA που σχετίζονται με το εμπορικό σήμα για να στοχεύσουν ισχυρά και ήπια θετικά και αρνητικά

συναισθήματα που παρέχουν πιο ενεργή νοημοσύνη στους επαγγελματίες της διαχείρισης επωνυμίας (8), (11), (20).

Υπάρχουν δύο κοινά χρησιμοποιούμενες προσεγγίσεις για την ανάλυση του αισθήματος των κειμένων: η προσέγγιση με βάση το λεξικό και οι μέθοδοι ML. Η πρώτη προσέγγιση περιλαμβάνει τη χρήση ενός λεξικού όρων που σχετίζονται με τη γνώμη με μια μέθοδο βαθμολόγησης για την αξιολόγηση του αισθήματος σε μια μη εποπτευόμενη εφαρμογή (21), (22) Ωστόσο, οι επιδόσεις αυτών των μεθόδων είναι περιορισμένες, καθώς δεν είναι σε θέση να λάβουν υπόψη τις πληροφορίες με βάση τα συμφραζόμενα, τους αποχρωματισμένους δείκτες της έκφρασης συναισθημάτων ή το νέο λεξιλόγιο. Η δεύτερη προσέγγιση ποσοτικοποιεί το κείμενο με βάση μια αναπαράσταση χαρακτηριστικών και εφαρμόζει έναν αλγόριθμο ML για να αντλήσει τη σχέση μεταξύ συναισθημάτων και τιμών χαρακτηριστικών χρησιμοποιώντας εποπτευόμενη μάθηση (23), (24). Τα μοντέλα που βασίζονται σε αυτόν τον τύπο μάθησης απαιτούν ένα μεγάλο σετ εκπαίδευσης με ετικέτες κατηγορίας συναισθημάτων για τη βαθμονόμηση των παραμέτρων του μοντέλου.

Οι Gao et al (25) πρότειναν μια προσέγγιση που εστιάζει στη συχνότητα των τάξεων συναισθημάτων στο σύνολο δεδομένων που αναλύουν. Οι συγγραφείς κατέληξαν στο συμπέρασμα ότι, σε αντίθεση με τους κανονικούς αλγόριθμους που βασίζονται στην ταξινόμηση, η χρήση ενός αλγόριθμου ειδικής ποσοτικοποίησης παρουσιάζει μια καλύτερη εκτίμηση συχνότητας. Επιπλέον, με την ευρεία υιοθέτηση του DL ως τεχνολογίας αιχμής, τα πιο πρόσφατα έργα πήγαν "βαθύτερα" και δημιουργήθηκαν νέα μοντέλα. Για παράδειγμα, το έργο της ανάλυσης συναισθημάτων πολλαπλών τάξεων έχει αντιμετωπιστεί επίσης σε έργα όπως αυτό των Araque et al. (26).

Κεφάλαιο 3

# Μηχανική Μάθηση – Machine Learning

## 3.1 Ορισμός Μηχανικής Μάθησης

Η μηχανική μάθηση μπορεί να θεωρηθεί ως μια συλλογή από μεθόδους που μπορούν αυτόματα να αναγνωρίσουν διάφορα μοτίβα στα δεδομένα και στη συνέχεια με βάση αυτά τα μοτίβα να προβλέψουν μελλοντικά αποτελέσματα ή να πάρουν αποφάσεις κάτω από συγκεκριμένες καταστάσεις. Όλα αυτά έχουν τη δυνατότητα να αξιοποιηθούν χρησιμοποιώντας διάφορους αλγορίθμους που επιτρέπουν στις μηχανές να καταλαβαίνουν διάφορες καταστάσεις και βασισμένες σε αυτές να παίρνονται οι αποφάσεις. Η μηχανική μάθηση είναι χρήσιμη σε όλους τους τομείς αφού παρέχει διάφορα πλεονεκτήματα, όπως:

- Γρήγορη απόφαση: Η μηχανική μάθηση παρέχει γρήγορα τα καλύτερα αποτελέσματα.
- Ικανότητα προσαρμογής: Έχει τη δυνατότητα να προσαρμόζεται γρήγορα στο νέο περιβάλλον, το οποίο μεταβάλλεται συνεχώς, αφού τα δεδομένα ενημερώνονται συνεχώς.
- Καινοτομία: Με τη χρησιμοποίηση προηγμένων αλγορίθμων βελτιώνεται η ικανότητα λήψης αποφάσεων, βοηθώντας έτσι στην ανάπτυξη καινοτόμων επιχειρηματικών υπηρεσιών και μοντέλων.
- Διορατικότητα: Γίνεται η κατανόηση μοναδικών προτύπων δεδομένων και με βάση αυτών βασίζονται στις ενέργειες που μπορούν να παρθούν.
- Επιχειρηματική ανάπτυξη: Η συνολική επιχειρηματική διαδικασία και η ροή εργασίας είναι ταχύτερες, βοηθώντας έτσι στην επιχειρηματική ανάπτυξη.
- Καλό αποτέλεσμα: Το αποτέλεσμα θα βελτιώνεται σε αντίθεση με τη πιθανότητα σφάλματος, η οποία θα μειώνεται.

## 3.2 Τύποι Μηχανικής Μάθησης

Υπάρχουν τρεις τύποι μηχανικής μάθησης, η επιβλεπόμενη μάθηση (Supervised Learning), η μη επιβλεπόμενη μάθηση (Unsupervised Learning) και η ενισχυτική μάθηση (Reinforcement Learning). Η επιβλεπόμενη μάθηση έχει σαν κύριες μεθόδους την κατηγοριοποίηση και την παλινδρόμηση, ενώ η μη επιβλεπόμενη μάθηση το μετασχηματισμό και τη συσταδοποίηση, έννοιες οι οποίες αναλύονται στη συνέχεια. Και στις δύο περιπτώσεις, τα δεδομένα εισόδου

πρέπει να έχουν σωστή αναπαράσταση για να μπορεί να τα καταλάβει ένας υπολογιστής. Η ενισχυτική μάθηση ασχολείται κυρίως με διάφορες οντότητες που ονομάζονται πράκτορες, οι οποίοι παίρνουν τις αποφάσεις τους από το περιβάλλον, με σκοπό να εκτελέσουν κάποια ενέργεια.

### 3.2.1 Επιβλεπόμενη μάθηση - Supervised Learning

Η επιβλεπόμενη μάθηση είναι μια από τις πιο κοινές και επιτυχημένες μεθόδους που χρησιμοποιούνται στη μηχανική μάθηση, η οποία χρησιμοποιείται όταν θέλουμε να προβλέψουμε ένα σίγουρο αποτέλεσμα από μία δεδομένη είσοδο. Ονομάζεται επιβλεπόμενη μάθηση γιατί το μοντέλο μας μαθαίνει από ένα σύνολο εκπαίδευσης δημιουργώντας έτσι ένα άλλο μοντέλο, όπου με βάση αυτό εφαρμόζεται στο νέο σύνολο δεδομένων για να προβλέψει τα αποτελέσματα. Υπάρχουν δύο υποκατηγορίες επιβλεπόμενης μάθησης, η κατηγοριοποίηση και η παλινδρόμηση.

Η κατηγοριοποίηση (Classification) είναι μια από τις βασικότερες εργασίες της μηχανικής μάθησης. Κύριος στόχος της κατηγοριοποίησης είναι η πρόβλεψη μιας κατηγοριοποιημένης ετικέτας κατηγοριών νέων περιπτώσεων βασισμένη σε προηγούμενες παρατηρήσεις. Ως επί το πλείστον, η κατηγοριοποίηση είναι χωρισμένη σε δυαδική κατηγοριοποίηση, όπου ο αλγόριθμος μαθαίνει μια σειρά από κανόνες για τη διάκριση των ετικετών μεταξύ δύο πιθανών κατηγοριών, και σε πολλαπλή κατηγοριοποίηση (Multiclass Classification), η οποία κατηγοριοποιεί τα δεδομένα σε περισσότερες από δύο κατηγορίες. Η κατηγοριοποίηση μπορεί να περιγραφεί ως μια διαδικασία με δύο στάδια, την εκμάθηση και την κατηγοριοποίηση / εκπαίδευση.

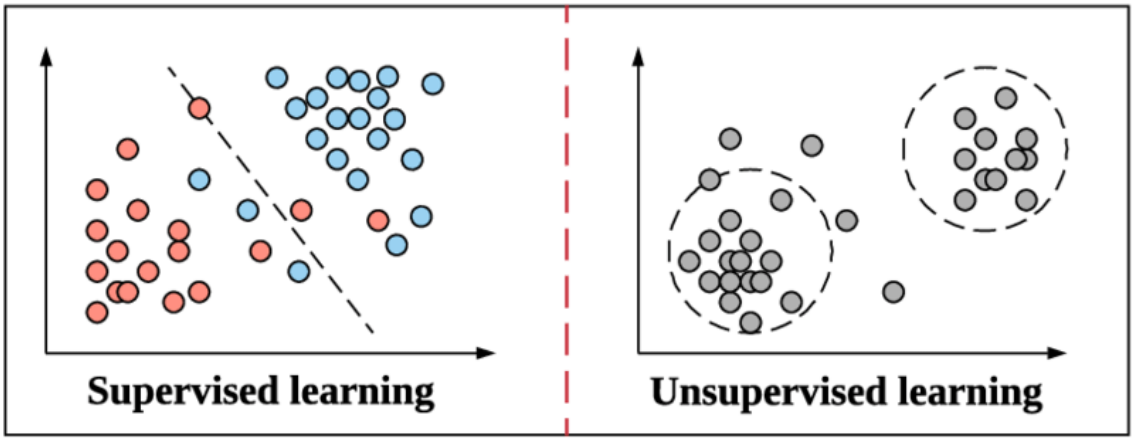
Δεύτερος τύπος της επιβλεπόμενης μάθησης είναι η μέθοδος της παλινδρόμησης (Regression), η οποία χρησιμοποιείται για την πρόβλεψη συνεχών αποτελεσμάτων. Στόχος αυτής της μεθόδου είναι να προβλεφθεί ένας συνεχής αριθμός, αφού αρχικά δοθεί ένας αριθμός από τον παράγοντα της μεταβλητής καθώς και μια συνεχής μεταβλητή αποτελέσματος. Με βάση τα πιο πάνω προσπαθούμε να βρούμε μια σχέση μεταξύ των μεταβλητών, η οποία θα μας επιτρέψει να προβλέψουμε το αποτέλεσμα.

### 3.2.2 Μη επιβλεπόμενη μάθηση - Unsupervised Learning

Δεύτερος τύπος μηχανικής μάθησης είναι η μη επιβλεπόμενη μάθηση. Σε αντίθεση με την επιβλεπόμενη μάθηση, όπου γνωρίζαμε εξαρχής τη σωστή απάντηση, στη μη επιβλεπόμενη μάθηση ερχόμαστε αντιμέτωποι με μη κατηγοριοποιημένα δεδομένα ή με δεδομένα με άγνωστη δομή. Χρησιμοποιώντας αυτή την τεχνική, μπορούμε να εξερευνήσουμε τη δομή των δεδομένων και να εξάγουμε σημαντική πληροφορία χωρίς την καθοδήγηση ενός γνωστού αποτελέσματος. Στη μη επιβλεπόμενη μάθηση, ο αλγόριθμος δέχεται απλά τα δεδομένα και ζητείται η εξαγωγή γνώσης από τα δεδομένα αυτά. Όπως και στην επιβλεπόμενη μάθηση, έτσι και στη μη επιβλεπόμενη μάθηση υπάρχουν δύο υποκατηγορίες, οι οποίες ονομάζονται μετασχηματισμός δεδομένων και συσταδοποίηση.

Η χρησιμοποίηση αλγορίθμων για τη δημιουργία νέων αναπαραστάσεων του συνόλου δεδομένων μας ονομάζεται μετασχηματισμός (Transformation) δεδομένων. Οι αναπαραστάσεις αυτές είναι ευκολότερες στην κατανόηση από τους ανθρώπους ή από τους αλγορίθμους μηχανικής μάθησης, έτσι ώστε να είναι εφικτή η καλύτερη σύγκριση των δεδομένων από την αρχική.

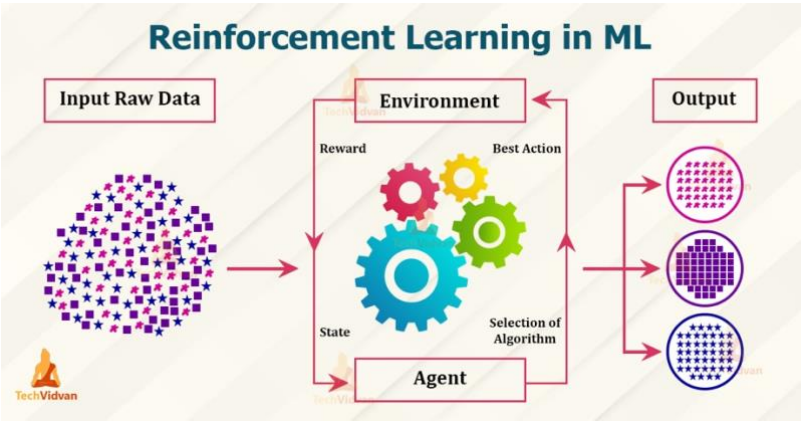
Δεύτερος τύπος μη επιβλεπόμενης μάθησης είναι η μέθοδος συσταδοποίησης (Clustering), μιας διερευνητικής τεχνικής ανάλυσης δεδομένων, που μας επιτρέπει να οργανώσουμε μια στοίβα από πληροφορίες σε σημαντικές υποομάδες (clusters), χωρίς όμως να υπάρχει κάποια προηγούμενη γνώση των μελών της ομάδας τους. Κάθε υποομάδα (cluster) η οποία μπορεί να προκύψει κατά τη διάρκεια της ανάλυσης, ορίζεται από μια ομάδα αντικειμένων που έχουν κάποια κοινά χαρακτηριστικά, αλλά είναι και περισσότερο ανόμοια με αντικείμενα από τις άλλες υποομάδες (clusters).




*Εικόνα 2 Unsupervised και Supervised Learning (27)*

### 3.2.3 Ενισχυτική μάθηση - Reinforcement Learning

Εκτός από την επιβλεπόμενη και μη-επιβλεπόμενη μάθηση, οι οποίες είναι τα σημαντικότερα είδη μάθησης, άλλο ένα είδος μάθησης είναι η ενισχυτική μάθηση (Reinforcement Learning). Σκοπός του συστήματος είναι η μεγιστοποίηση της αριθμητικής τιμής της ανταμοιβής. Η μέθοδος αυτή χρησιμοποιεί τρεις μεθόδους (Components), τον πράκτορα (Agent), το περιβάλλον (Environment) και την ενέργεια (Action). Ο πράκτορας είναι η οντότητα η οποία μαθαίνει και παίρνει διάφορες αποφάσεις, ενώ οτιδήποτε άλλο είναι το περιβάλλον. Υπάρχει μια συνεχείς αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος, όπου ο πράκτορας επιλέγει διάφορες ενέργειες για να πραγματοποιήσει το περιβάλλον, παρουσιάζοντας του έτσι καινούριες καταστάσεις. Το περιβάλλον δίνει στον πράκτορα διάφορες ανταμοιβές, όπου με βάση και το σκοπό του συστήματος προσπαθεί να μεγιστοποιήσει μακροπρόθεσμα.




*Εικόνα 3 Reinforcement Learning (28)*

## 3.3 Feature Engineering

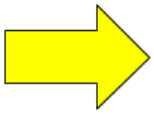
Οι εργασίες μηχανικής εκμάθησης, όπως η ταξινόμηση ή η πρόβλεψη, απαιτούν συχνά είσοδο που είναι μαθηματικά και υπολογιστικά βολική στην επεξεργασία. Ωστόσο, δεδομένα πραγματικού κόσμου όπως εικόνες, βίντεο και μέτρηση αισθητήρων είναι συνήθως περίπλοκα, περιττά και υψηλών διαστάσεων, καθιστώντας δυσκολότερο να βρεθούν μοτίβα και ανωμαλίες. Επομένως, είναι απαραίτητο να ανακαλυφθούν χρήσιμα χαρακτηριστικά ή αναπαραστάσεις από μη επεξεργασμένα δεδομένα.

Η διαδικασία εξαγωγής χαρακτηριστικών από ανεπεξέργαστα δεδομένα με χρήση γνώσης τομέα ονομάζεται Feature Engineering. Σκοπός της είναι να μετατρέψει τα δεδομένα σε μορφές συμβατές με τον αλγόριθμο ML και να βελτιώσει την απόδοσή του. Στη συνέχεια παρουσιάζονται μερικές δημοφιλείς τεχνικές που μπορούν να είναι χρήσιμες σε διαφορετικά προβλήματα.

### 3.3.1 Κωδικοποίηση One-Hot

Οι κατηγοριοποιημένες μεταβλητές αντιπροσωπεύουν τύπους δεδομένων που μπορούν να χωριστούν σε περιορισμένο αριθμό ομάδων. Η αναπαράσταση κατηγοριοποιημένων δεδομένων ως δυαδικών διανυσμάτων (αριθμητική μορφή) ονομάζεται κωδικοποίηση one-hot. Είναι μια από τις πιο κοινές μεθόδους κωδικοποίησης στο ML. Δημιουργεί ένα διάνυσμα με μήκος ίσο με τον αριθμό των κατηγοριών στο σύνολο δεδομένων. Εάν ένα δείγμα ανήκει στην κατηγορία i, τότε στο i-οστό στοιχείο του διανύσματος εκχωρείται η τιμή 1 και σε όλα τα άλλα εκχωρείται η τιμή 0.

| Color |
|---|
| Red |
| Red |
| Yellow |
| Green |
| Yellow |



| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

*Εικόνα 4 Παράδειγμα (29)*

## 3.4 Feature Learning

Η μη αυτόματη μηχανική χαρακτηριστικών απαιτεί συχνά δαπανηρή ανθρώπινη εργασία, βασίζεται σε ειδικές γνώσεις και συνήθως δεν γενικεύεται καλά. Αυτό παρακινεί το σχεδιασμό αποτελεσματικών τεχνικών εκμάθησης χαρακτηριστικών, για αυτοματοποίηση και γενίκευση αυτών. Η εκμάθηση χαρακτηριστικών, που ονομάζεται επίσης εκμάθηση αναπαράστασης, είναι ένα σύνολο τεχνικών που μαθαίνουν αναπαραστάσεις εισόδου μη επεξεργασμένων δεδομένων, συνήθως μετατρέποντάς τις ή εξάγοντας χαρακτηριστικά από αυτές. Αυτό καθιστά ευκολότερο για μια μηχανή να μάθει τόσο τα χαρακτηριστικά όσο και να τα χρησιμοποιήσει για την εκτέλεση μιας εργασίας μηχανικής μάθησης. Αυτές οι τεχνικές επιτρέπουν την ανακατασκευή των εισόδων που προέρχονται από την άγνωστη κατανομή που δημιουργεί δεδομένα, ενώ δεν είναι απαραίτητα πιστές στις διαμορφώσεις που είναι αβάσιμες κάτω από αυτήν την κατανομή.
Η εκμάθηση χαρακτηριστικών μπορεί να χωριστεί σε δύο κατηγορίες: εποπτευόμενη και μη εποπτευόμενη εκμάθηση χαρακτηριστικών, ανάλογη με αυτές των κατηγοριών της μηχανικής μάθησης. Στην εποπτευόμενη εκμάθηση χαρακτηριστικών, τα χαρακτηριστικά μαθαίνονται χρησιμοποιώντας δεδομένα εισόδου με ετικέτα. Στη μη εποπτευόμενη εκμάθηση χαρακτηριστικών, οι δυνατότητες μαθαίνονται με δεδομένα εισόδου χωρίς ετικέτα.

### 3.4.1 Ανάλυση κυρίων συνιστωσών

Η μέθοδος ανάλυσης κυρίων συνιστωσών (Principal Components Analysis - PCA) (30) είναι μια τεχνική γραμμικού μετασχηματισμού, η οποία χρησιμοποιείται για μείωση διαστάσεων του συνόλου δεδομένων κατά τη φάση της προεπεξεργασίας, με σκοπό ο αλγόριθμος κατηγοριοποίησης να γίνει πιο αποτελεσματικός. Είναι επίσης μια δημοφιλής μέθοδος για την εξαγωγή των σημαντικών χαρακτηριστικών από τα δεδομένα εκπαίδευσης που χρησιμοποιούνται για την εκμάθηση ενός μοντέλου μηχανικής μάθησης. Αν θεωρήσουμε ότι ένα σύνολο δεδομένων έχει σύνολο αριθμό γραμμών N και αριθμό στηλών M, τότε με τη μέθοδο PCA βρίσκουμε ένα σύστημα K κάθετων διανυσμάτων, το οποίο είναι μικρότερο από το συνολικό αριθμό των στηλών του αρχικού συνόλου δεδομένων. Στη συνέχεια προβάλλουμε τα δεδομένα στο νέο μας σύστημα K, δημιουργώντας έτσι γραμμικούς συνδυασμούς των

αρχικών μεταβλητών, οι οποίοι είναι ασυσχέτιστοι μεταξύ τους και περιέχουν το μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

### 3.4.2 Ενσωμάτωση στοχαστικών γειτονικών t-κατανεμημένων

Η (t-SNE) (31) είναι μια μη εποπτευόμενη μη γραμμική τεχνική για μείωση των διαστάσεων που βοηθά στον εντοπισμό σχετικών προτύπων. Είναι ιδιαίτερα κατάλληλη για οπτικοποίηση συνόλων δεδομένων υψηλής διάστασης.

Συγκεκριμένα, μοντελοποιεί κάθε αντικείμενο υψηλών διαστάσεων από ένα σημείο χαμηλότερης διάστασης με τέτοιο τρόπο ώστε παρόμοια αντικείμενα να διαμορφώνονται από κοντινά σημεία και διαφορετικά αντικείμενα να διαμορφώνονται από μακρινά σημεία με μεγάλη πιθανότητα. Εφαρμόζεται στην επεξεργασία εικόνων, στην επεξεργασία φυσικής γλώσσας, στα γονιδιωματικά δεδομένα και στην επεξεργασία ομιλίας.

Με απλούς όρους, το t-SNE ελαχιστοποιεί την απόκλιση μεταξύ μιας κατανομής που μετρά τις ομοιότητες των αντικειμένων εισαγωγής κατά ζεύγη και μιας κατανομής που μετρά τις ομοιότητες κατά ζεύγη των αντίστοιχων σημείων χαμηλής διάστασης στην ενσωμάτωση. Με αυτόν τον τρόπο, το t-SNE χαρτογραφεί τα πολυδιάστατα δεδομένα σε ένα δισδιάστατο ή τρισδιάστατο χώρο και επιχειρεί να βρει μοτίβα στα δεδομένα προσδιορίζοντας τις παρατηρούμενες συστάδες με βάση την ομοιότητα των σημείων δεδομένων με πολλά χαρακτηριστικά.

## 3.5 Classification Problems

Όπως αναφέρεται στην Ενότητα 3.2.1, τα προβλήματα ταξινόμησης είναι ένας τύπος εποπτευόμενης μηχανικής μάθησης, όπου τα δεδομένα εκπαίδευσης έχουν ήδη επισημανθεί. Πρόκειται για προβλήματα προσδιορισμού σε ποια κατηγορία εμπίπτει μια συγκεκριμένη νέα παρατήρηση. Πιο συγκεκριμένα, η ταξινόμηση προγνωστικής μοντελοποίησης είναι η εργασία της προσέγγισης της λειτουργίας χαρτογράφησης (f) από μεταβλητές εισόδου (X) σε διακριτές μεταβλητές εξόδου (Y), προκειμένου να προσδιοριστεί σε ποια τάξη / ετικέτα / κατηγορία πιθανότατα θα ανήκουν τα νέα δεδομένα. Ο αλγόριθμος που εφαρμόζει την ταξινόμηση είναι γνωστός ως ταξινομητής. Τα δεδομένα με ετικέτα χρησιμοποιούνται για την εκπαίδευση ενός ταξινομητή, έτσι ώστε ο αλγόριθμος να έχει καλή απόδοση σε δεδομένα που δεν έχουν ακόμη

επισημανθεί. Η επανάληψη αυτής της διαδικασίας εκπαίδευσης ενός ταξινομητή σε ήδη επισημασμένα δεδομένα είναι γνωστή ως «μάθηση». Στους εποπτευόμενους αλγόριθμους μηχανικής μάθησης, θέλουμε να ελαχιστοποιήσουμε το σφάλμα για κάθε παράδειγμα εκπαίδευσης κατά τη διάρκεια της μαθησιακής διαδικασίας. Αυτό γίνεται με τη χρήση ορισμένων στρατηγικών βελτιστοποίησης όπως η καθοδική κλίση και αυτό το σφάλμα προέρχεται από μια συνάρτηση απώλειας.

## 3.6 Συνάρτηση Απώλειας

Στη μαθηματική βελτιστοποίηση και θεωρία αποφάσεων, μια συνάρτηση απώλειας είναι μια συνάρτηση που χαρτογραφεί ένα συμβάν ή τιμές μιας ή περισσότερων μεταβλητών σε έναν πραγματικό αριθμό διαισθητικά αντιπροσωπεύοντας κάποιο "κόστος" που σχετίζεται με το συμβάν. Στα στατιστικά στοιχεία, συνήθως χρησιμοποιείται μια συνάρτηση απώλειας για την εκτίμηση παραμέτρων και το εν λόγω συμβάν είναι κάποια συνάρτηση της διαφοράς μεταξύ εκτιμώμενων και αληθών τιμών για μια σειρά δεδομένων. Οι συναρτήσεις απώλειας παίζουν σημαντικό ρόλο σε οποιοδήποτε στατιστικό μοντέλο, καθώς καθορίζουν έναν στόχο βάσει του οποίου αξιολογείται η απόδοση του μοντέλου. Οι παράμετροι που μαθαίνει το μοντέλο καθορίζονται ελαχιστοποιώντας μια επιλεγμένη συνάρτηση απώλειας.
Στη μηχανική εκμάθηση, οι συναρτήσεις απώλειας (ή σφάλματος) για την ταξινόμηση είναι υπολογιστικά εφικτές συναρτήσεις απώλειας που αναπαριστούν την ποινή για ανακρίβεια των προβλέψεων σε προβλήματα ταξινόμησης. Είναι μια μέθοδος αξιολόγησης του πόσο καλά ένας συγκεκριμένος αλγόριθμος μοντελοποιεί τα δοσμένα δεδομένα. Τυπικά, μια συνάρτηση απώλειας L (ˆy, y) εκχωρεί έναν πραγματικό αριθμό σε μια προβλεπόμενη έξοδο y ˆ δεδομένης της πραγματικής αναμενόμενης εξόδου y. Εάν η πρόβλεψη αποκλίνει πάρα πολύ από το πραγματικό αποτέλεσμα, η συνάρτηση απώλειας θα δώσει στην έξοδο έναν μεγάλο αριθμό. Σταδιακά, με τη βοήθεια κάποιας τεχνικής βελτιστοποίησης, η συνάρτηση απώλειας μαθαίνει να μειώνει το σφάλμα στην πρόβλεψη. Η συνάρτηση απώλειας πρέπει να περιοριστεί από χαμηλά, με το ελάχιστο να επιτυγχάνεται μόνο για περιπτώσεις όπου η πρόβλεψη είναι σωστή. Υπάρχουν πολλοί τρόποι προσδιορισμού της απώλειας. Δεν ταιριάζουν όλοι στο σύνολο των πιθανών εργασιών. Η επιλογή μιας συνάρτησης απώλειας για ένα συγκεκριμένο πρόβλημα περιλαμβάνει διάφορους παράγοντες, όπως τον τύπο αλγορίθμου μηχανικής μάθησης που επιλέχθηκε, την ευκολία υπολογισμού των παραγώγων και σε κάποιο βαθμό το ποσοστό των ακραίων τιμών στο σετ δεδομένων. Σε γενικές γραμμές, οι συναρτήσεις απώλειας μπορούν να

ταξινομηθούν σε δύο μεγάλες κατηγορίες ανάλογα με τον τύπο της εργασίας εκμάθησης που αντιμετωπίζουμε: συναρτήσεις ταξινόμησης και παλινδρόμησης.

### 3.6.1 Binary Cross-Entropy Loss

Γενικά, ο όρος «εντροπία» αναφέρεται στη διαταραχή ή την αβεβαιότητα στα δεδομένα. Όσο μεγαλύτερη είναι η τιμή της εντροπίας, τόσο υψηλότερο είναι το επίπεδο αβεβαιότητας.
Για μια τυχαία διακριτή μεταβλητή $y$ με κατανομή πιθανότητας $q(y)$, μετριέται ως:

$$H(q) = -\sum_{c=1}^{C} q(y_c)\log(q(y_c))$$

όπου $C$ είναι το πλήθος των πιθανών αποτελεσμάτων. Η εντροπία είναι το αρνητικό του αθροίσματος του προϊόντος της πιθανότητας εμφάνισης ενός γεγονότος με το $log$ του σε όλα τα πιθανά αποτελέσματα. Έτσι, εάν είναι γνωστή η πραγματική κατανομή μιας τυχαίας μεταβλητής, η εντροπία της μπορεί να υπολογιστεί. Ωστόσο, η πραγματική κατανομή των δεδομένων στις περισσότερες περιπτώσεις είναι άγνωστη και προσπαθούμε να την προσεγγίσουμε με κάποια άλλη κατανομή $p(y)$. Η cross-entropy μετρά την εντροπία μεταξύ δύο κατανομών πιθανότητας. Είναι ο μέσος αριθμός των bit που απαιτούνται για την επικοινωνία ενός συμβάντος από μια διανομή σε μία άλλη:

$$H_p(q) = -\sum_{c=1}^{C} q(y_c)\log(p(y_c))$$

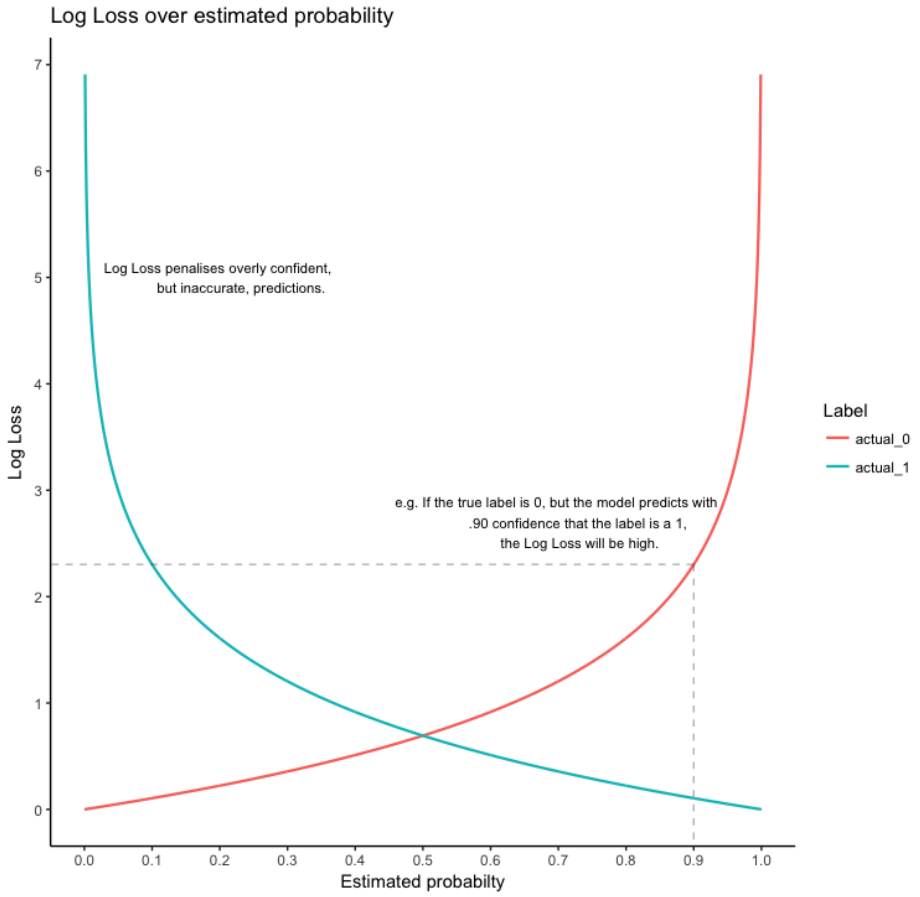
Εάν οι κατανομές $p(y)$ και $q(y)$ ταιριάζουν απόλυτα, οι υπολογισμένες τιμές τόσο για την cross-entropy όσο και για την εντροπία ταιριάζουν επίσης. Διαφορετικά, η cross-entropy θα έχει μεγαλύτερη αξία από την εντροπία που υπολογίζεται στην πραγματική κατανομή. Το καλύτερο δυνατό $p(y)$ είναι αυτό που ελαχιστοποιεί την cross-entropy εντροπία.
Η cross-entropy ως έννοια εφαρμόζεται στον τομέα της μηχανικής μάθησης όταν οι αλγόριθμοι είναι κατασκευασμένοι για να προβλέπουν από το μοντέλο. Η δημιουργία του μοντέλου βασίζεται στη σύγκριση των πραγματικών αποτελεσμάτων με τα προβλεπόμενα αποτελέσματα. Η cross-entropy σημαίνει μείωση της εντροπίας ή αβεβαιότητας για την

κατηγορία που προβλέπεται, και είναι κατάλληλη ως συνάρτηση απώλειας επειδή ο στόχος είναι να ελαχιστοποιηθεί η αξία της. Σε ένα πρόβλημα δυαδικής ταξινόμησης, το $C = 2$ και η ετικέτα εξόδου $y$ μπορούν να λάβουν τιμές 0 και 1. Σε τέτοια περίπτωση, η δυαδική διασταυρούμενη εντροπία ($BCE$) ανά δείγμα ορίζεται ως:

$$BCE = -(y_i \log(p(y_c)) + (1 - y_i) \log(1 - p(y_i))$$

όπου $p(y_i)$ είναι η προβλεπόμενη πιθανότητα p ότι το δείγμα $i$ ανήκει στην κλάση και $1 - p(y_i)$ είναι η προβλεπόμενη πιθανότητα ότι το δείγμα $i$ ανήκει στην κλάση 0.

Συνοψίζοντας, η δυαδική cross-entropy απώλεια, που ονομάζεται επίσης Log Loss, μετρά την απόδοση ενός μοντέλου ταξινόμησης του οποίου η προβλεπόμενη απόδοση είναι τιμή πιθανότητας μεταξύ 0 και 1. Το Σχήμα 3.3 δείχνει το εύρος των πιθανών τιμών απώλειας με μια πραγματική παρατήρηση. Καθώς η προβλεπόμενη πιθανότητα πλησιάζει το 1, η απώλεια ημερολογίου μειώνεται αργά, ενώ όσο μειώνεται η προβλεπόμενη πιθανότητα, η απώλεια ημερολογίου αυξάνεται γρήγορα.
Δεδομένου ότι η πιθανότητα απαιτεί μια τιμή μεταξύ 0 και 1, η συνάρτηση Sigmoid μπορεί να χρησιμοποιηθεί για τον υπολογισμό του p. Η σιγμοειδής συνάρτηση είναι επίσης γνωστή ως συνάρτηση συμπίεσης επειδή μπορεί να στριμώξει οποιαδήποτε πραγματική τιμή στο εύρος [0, 1], γεγονός που την καθιστά κατάλληλη για ένα μοντέλο όπου έχουμε πιθανότητα ως έξοδο.




*Εικόνα 5 Log Loss (32)*

### 3.6.2 Categorical Cross-Entropy Loss

Η Categorical Cross-Entropy Loss είναι μια συνάρτηση απώλειας που χρησιμοποιείται σε εργασίες ταξινόμησης πολλαπλών κατηγοριών. Πρόκειται για εργασίες όπου μια παρατήρηση μπορεί να ανήκει μόνο σε μία από τρεις ή περισσότερες πιθανές κατηγορίες με πιθανότητα 1 και σε άλλες κατηγορίες με πιθανότητα 0. Το μοντέλο πρέπει να προβλέπει μία πιθανή έξοδο κλάσης κάθε φορά.

Η Categorical Cross-Entropy (CCE) είναι ουσιαστικά η Binary Cross-Entropy που επεκτείνεται σε πολλαπλές κατηγορίες. Μαθηματικά, δίνεται ως η ακόλουθη εξίσωση:

$$CCE = -\sum_{c=1}^{C} y_{i,c} \log(p(y_{i,c}))$$

όπου $C$ είναι το πλήθος των κλάσεων, $y_{i,c}$ είναι ένας δυαδικός δείκτης (0 ή 1) που υποδεικνύει εάν το $c$ είναι η σωστή κλάση για το δείγμα $i$ και το $p(y_{i,c})$ δηλώνει την πιθανότητα παρατήρησης $i$ για την κλάση $c$.

## 3.7 Συνάρτηση Κόστους

Η συνάρτηση κόστους είναι μια συνάρτηση που μετρά την απόδοση ενός μοντέλου μηχανικής εκμάθησης για δοσμένα δεδομένα. Επιστρέφει το κόστος μεταξύ των προβλεπόμενων αποτελεσμάτων σε σύγκριση με τα πραγματικά αποτελέσματα και το παρουσιάζει με τη μορφή ενός πραγματικού αριθμού. Όταν η διαφορά μεταξύ της εξόδου και της σωστής απάντησης είναι μικρή, το σφάλμα είναι χαμηλό. Ανάλογα με το πρόβλημα, η συνάρτηση κόστους μπορεί να διαμορφωθεί με πολλούς διαφορετικούς τρόπους. Παρόλο που είναι δυνατό να οριστεί μια συνάρτηση κόστους ad hoc, συχνά η επιλογή καθορίζεται από τις επιθυμητές ιδιότητες της συνάρτησης ή επειδή προκύπτει από το μοντέλο.
Παρόλο που οι συνάρτηση κόστους και απώλειας είναι συνώνυμες και χρησιμοποιούνται εναλλακτικά, είναι διαφορετικές. Η συνάρτηση απώλειας χρησιμοποιείται για ένα μοναδικό παράδειγμα εκπαίδευσης. Μια συνάρτηση κόστους, από την άλλη πλευρά, είναι η μέση απώλεια ολόκληρου του συνόλου δεδομένων εκπαίδευσης. Ο στόχος της εποπτευόμενης μηχανικής μάθησης είναι η ελαχιστοποίηση του συνολικού κόστους, βελτιστοποιώντας έτσι τη συσχέτιση του μοντέλου με το σύστημα που προσπαθεί να εκπροσωπήσει. Δεδομένου ότι η πιθανότητα κάθε σημείου είναι 1/N, και y ∈ {0, 1}, η συνάρτηση κόστους BCE σε ολόκληρο το σύνολο δειγμάτων ορίζεται ως:

$$\Xi J_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(y_i)\big) + (1 - y_i)\log\big(1 - p(y_i)\big)$$

## 3.8 Υπερ-παράμετρος

Στη μηχανική εκμάθηση, ο όρος υπερ-παράμετρος χρησιμοποιείται για να ξεχωρίσει από τις τυπικές παραμέτρους μοντέλου που μπορούν να εξαχθούν άμεσα από την κανονική διαδικασία εκπαίδευσης για να ταιριάζουν στα δεδομένα. Η υπερ-παράμετρος είναι μια σταθερή παράμετρος που ορίζεται πριν από την έναρξη της εκπαιδευτικής διαδικασίας και της οποίας

η τιμή χρησιμοποιείται για τον έλεγχο της εκπαιδευτικής διαδικασίας. Οι υπερ-παράμετροι μπορούν να ταξινομηθούν ως μοντέλα υπερ-παραμέτρων, οι οποίες δεν μπορούν να εξαχθούν κατά την προσαρμογή της μηχανής στο σετ εκπαίδευσης επειδή αναφέρονται στην εργασία επιλογής μοντέλου, ή στον αλγόριθμο υπερ-παραμέτρων, που κατ 'αρχήν δεν επηρεάζουν την απόδοση του μοντέλου αλλά επηρεάζουν την ταχύτητα και την ποιότητα της εκπαιδευτικής διαδικασίας.

## 3.9 Στρατηγικές βελτιστοποίησης

Ο κύριος στόχος της μηχανικής μάθησης είναι να δημιουργήσει ένα μοντέλο που αποδίδει καλά και δίνει ακριβείς προβλέψεις σε ένα συγκεκριμένο σύνολο περιπτώσεων. Για να το επιτύχουμε αυτό, χρειαζόμαστε βελτιστοποίηση μηχανικής μάθησης. Η βελτιστοποίηση είναι το πρόβλημα της προσαρμογής των υπερ-παραμέτρων προκειμένου να ελαχιστοποιηθεί η συνάρτηση καθορισμένης απώλειας / συνάρτησης κόστους κατά τη χρήση κάποιας τεχνικής βελτιστοποίησης.
Υπάρχουν ίσως εκατοντάδες δημοφιλείς αλγόριθμοι βελτιστοποίησης, κάτι που σημαίνει πως είναι μια πρόκληση η επιλογή των αλγορίθμων που πρέπει να ληφθούν υπόψη για ένα συγκεκριμένο πρόβλημα βελτιστοποίησης. Η επιλογή, ωστόσο, του αλγορίθμου βελτιστοποίησης μπορεί να κάνει τη διαφορά μεταξύ της καλής ακρίβειας σε ώρες ή ημέρες. Οι βασικές μέθοδοι βελτιστοποίησης κατηγοριοποιούνται συνήθως σε μεθόδους βελτιστοποίησης πρώτης τάξης, υψηλής τάξης και μηδενικής παραγώγου. Συνήθως συναντάμε μεθόδους που εμπίπτουν στην κατηγορία της βελτιστοποίησης πρώτης τάξης, όπως η κάθοδος με βάση την κλίση και οι παραλλαγές της, οι οποίες χρησιμοποιούν την πρώτη παράγωγο (κλίση) για να επιλέξουν την κατεύθυνση που θα κινηθεί στο χώρο αναζήτησης.

### 3.9.1 Κάθοδος με βάση την κλίση

Είναι η πιο δημοφιλής μέθοδος βελτιστοποίησης που χρησιμοποιείται κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Είναι ένας αλγόριθμος επαναληπτικής βελτιστοποίησης πρώτης τάξεως για την εύρεση ενός τοπικού ελάχιστου από μια διαφοροποιημένη αντικειμενική συνάρτηση. Η ιδέα είναι να ενημερώνονται επαναληπτικά οι παράμετροι στην αντίθετη κατεύθυνση της διαβάθμισης (ή κατά προσέγγιση κλίσης) της αντικειμενικής

συνάρτησης, επειδή αυτή είναι η κατεύθυνση της απότομης καθόδου. Με κάθε ενημέρωση, αυτή η μέθοδος καθοδηγεί το μοντέλο για να βρει το στόχο και σταδιακά συγκλίνει στη βέλτιστη τιμή της αντικειμενικής λειτουργίας.

Υπάρχουν τρεις παραλλαγές του GD που διαφέρουν στον συνολικό αριθμό δειγμάτων από ένα σύνολο δεδομένων (παρτίδα) που χρησιμοποιείται για τον υπολογισμό της διαβάθμισης της αντικειμενικής συνάρτησης. Ανάλογα με την ποσότητα των δεδομένων, υπάρχει μια αντιστάθμιση μεταξύ της ακρίβειας της ενημέρωσης παραμέτρων και του χρόνου που απαιτείται για την εκτέλεση μιας ενημέρωσης.

Προκειμένου να εκτελεστεί Batch Gradient Descent (BGD), πρέπει να επαναληφθεί ολόκληρο το σύνολο των δεδομένων εκμάθησης ενώ αναπροσαρμόζεται το μοντέλο. Η παρτίδα σε αυτήν την περίπτωση θεωρείται ότι είναι ολόκληρο το σύνολο δεδομένων. Πιο συγκεκριμένα, το BGD υπολογίζει το σφάλμα για κάθε παράδειγμα στο σύνολο δεδομένων εκμάθησης, αλλά μόνο αφού αξιολογηθούν όλα τα παραδείγματα εκμάθησης, το μοντέλο ενημερώνεται. Αυτή η όλη διαδικασία μοιάζει με έναν κύκλο και ονομάζεται εποχική εκμάθηση.

### 3.9.2 Στοχαστική κάθοδος

Η στοχαστική κάθοδος με βάση την κλίση (SGD) μπορεί να θεωρηθεί ως στοχαστική προσέγγιση της GD βελτιστοποίησης, καθώς αντικαθιστά τον πραγματικό υπολογισμό της κλίσης από το συνολικό σετ δεδομένων από μια εκτιμώμενη κλίση που υπολογίζεται σε ένα τυχαία επιλεγμένο υποσύνολο των δεδομένων, το οποίο μπορεί να είναι τόσο μικρό όσο ένα μεμονωμένο δείγμα εκμάθησης.

### 3.9.3 Ποσοστό εκμάθησης

Όπως έχει ήδη αναφερθεί, το ποσοστό εκμάθησης είναι μια παράμετρος συντονισμού σε έναν αλγόριθμο βελτιστοποίησης που καθορίζει το μέγεθος του διορθωτικού βήματος σε κάθε επανάληψη ενώ κινείται προς ένα ελάχιστο μιας συνάρτησης απώλειας. Ελέγχει πόσο πρέπει να αλλάξει το μοντέλο για να ανταποκριθεί στο εκτιμώμενο σφάλμα κάθε φορά που ενημερώνονται οι παράμετροι του μοντέλου. Αναπαριστά μεταφορικά την ταχύτητα με την οποία «μαθαίνει» ένα μοντέλο μηχανικής μάθησης.

Κατά τον καθορισμό ενός ποσοστού εκμάθησης, υπάρχει μια αντιστάθμιση μεταξύ του ποσοστού σύγκλισης και της υπέρβασης. Ενώ η κατεύθυνση της καθόδου καθορίζεται

συνήθως από την κλίση της συνάρτησης απώλειας, ο ρυθμός εκμάθησης καθορίζει πόσο μεγάλο είναι ένα βήμα προς αυτήν την κατεύθυνση. Μια πολύ μικρή τιμή μπορεί να οδηγήσει σε μια μακρά διαδικασία εκμάθησης που θα μπορούσε να κολλήσει σε ένα ανεπιθύμητο τοπικό ελάχιστο και μια τιμή πολύ μεγάλη μπορεί να οδηγήσει στην εκμάθηση ενός υπο-βέλτιστου συνόλου παραμέτρων πολύ γρήγορα ή σε μια ασταθή διαδικασίας εκμάθησης.

Κεφάλαιο 4

# Βαθιά Μάθηση - Deep Learning

## 4.1 Τεχνητά Νευρωνικά Δίκτυα

### 4.1.1 Ορισμός

Η πηγή έμπνευσης των νευρωνικών δικτύων, προέρχεται από τη βιολογία. Οι ζωντανοί οργανισμοί, από τους πιο απλούς μέχρι τον άνθρωπο, έχουν ένα νευρικό σύστημα το οποίο είναι υπεύθυνο για μια πλειάδα διεργασιών, όπως η επαφή με τον εξωτερικό κόσμο, η μάθηση, η μνήμη και πολλές άλλες. Η κεντρική μονάδα του νευρικού συστήματος είναι, οπωσδήποτε ο εγκέφαλος, ο οποίος επίσης αποτελείται από νευρωνικά δίκτυα. Κάθε νευρωνικό δίκτυο αποτελείται από ένα μεγάλο αριθμό μονάδων, που λέγονται νευρώνες ή νευρώνια (neurons).
Ο νευρώνας είναι η πιο μικρή ανεξάρτητη μονάδα του δικτύου, όπως λ.χ. το άτομο είναι η μικρότερη μονάδα της ύλης. Οι νευρώνες ασταμάτητα επεξεργάζονται πληροφορίες, λαμβάνοντας και στέλνοντας ηλεκτρικά σήματα σε άλλους νευρώνες. Ο τρόπος που ο ανθρώπινος εγκέφαλος λειτουργεί, έχει οδηγήσει την επιστήμη στον πειραματισμό δημιουργίας μοντέλων παρόμοιων του νευρωνικού συστήματος του ανθρώπου, εμπεριέχοντας όλα τα χαρακτηριστικά του. Τα δίκτυα αυτά ονομάζονται τεχνητά νευρωνικά δίκτυα (Artificial Neural Nets - ANN). Τα δίκτυα αυτά, παρόμοια με τα βιολογικά δίκτυα παίρνουν γνώσεις (μαθαίνουν) με την εξάσκηση και την εμπειρία, όπως ακριβώς και οι άνθρωποι, και μαθαίνουν να εκτελούν κάποια συγκεκριμένη διαδικασία. Έτσι ένα τεχνητό νευρωνικό δίκτυο (ΤΝΔ) μπορεί να ορισθεί επίσης, ως ένα μαθηματικό μοντέλο επεξεργασίας πληροφορίας που μιμείται τον ανθρώπινο εγκέφαλο και μπορεί να λύσει ορισμένης φύσεως προβλήματα σε πολλά πεδία της επιστήμης.
Τα ΤΝΔ μαθαίνουν από παραδείγματα, όπως και οι άνθρωποι, και ρυθμίζονται προκειμένου να μπορούν να χρησιμοποιηθούν σε συγκεκριμένες εφαρμογές, όπως είναι η αναγνώριση προτύπων ή η ταξινόμηση δεδομένων, μέσα από τη διαδικασία της εκπαίδευσης. Η διαδικασία της εκπαίδευσης στα βιολογικά συστήματα ορίζεται ως οι αναπροσαρμογές των συναπτικών συνδέσεων που υπάρχουν μεταξύ των νευρώνων. Το ίδιο ισχύει και για τα ΤΝΔ, όπου μέσω της διαδικασίας της εκπαίδευσης ρυθμίζονται τα βάρη των διασυνδέσεων μεταξύ των νευρώνων, οι οποίες ονομάζονται συναπτικά βάρη

### 4.1.2 Ανάλυση του νευρώνα

#### *4.1.2.1 Αντιστοίχιση βιολογικού συστήματος με τα ΤΝΔ*

Τα τεχνητά νευρωνικά δίκτυα αποτελούν μία προσπάθεια μοντελοποίησης των δυνατοτήτων επεξεργασίας του νευρωνικού συστήματος του ανθρώπου. Οι σημαντικές ιδιότητες των βιολογικών συστημάτων, όπως η προσαρμοστικότητα, η ικανότητα αναγνώρισης, η ανοχή στα λάθη, η μεγάλη χωρητικότητα μνήμης και η ικανότητα επεξεργασίας βιολογικών πληροφοριών σε πραγματικό χρόνο, μας κατευθύνουν στη μελέτη και την προσπάθεια προσομοίωσης αυτών των βιολογικών αρχιτεκτονικών.

Σήμερα, ο μηχανισμός της παραγωγής και της μεταφοράς των σημάτων μεταξύ των νευρώνων έχει κατανοηθεί πλήρως, όμως ο τρόπος που συνεργάζονται για να σχηματίσουν ένα πολύπλοκο και παράλληλο σύστημα, ικανό να επεξεργάζεται με απίστευτο τρόπο την πληροφορία δεν έχει ακόμα πλήρως διευκρινιστεί. Στο τομέα των θετικών επιστημών, μαθηματικοί, φυσικοί και μηχανικοί υπολογιστών μπορούν να συνεισφέρουν στη μελέτη των πολύπλοκων αυτών συστημάτων. Άλλωστε, δεν είναι τυχαίο ότι η μελέτη του εγκεφάλου έχει γίνει το πιο διεπιστημονικό πεδίο έρευνας τα τελευταία χρόνια.

Στον ανθρώπινο εγκέφαλο, ένας τυπικός νευρώνας συλλέγει σήματα από άλλους νευρώνες μέσα από μία σειρά από εκλεπτυσμένες δομές που ονομάζονται δενδρίτες. Ένας νευρώνας δεν είναι τίποτε άλλο παρά ένας διακόπτης ο οποίος δέχεται και εξάγει πληροφορία αναλόγως την κατάσταση την οποία βρίσκεται. Η διακοπτική κατάσταση εξαρτάται πλήρως από τη ποσότητα της πληροφορίας που δέχεται μέσω του συμπλέγματος νευρώνων που αλληλεπιδρούν στην είσοδο. Ο νευρώνας αποτελείται από ένα μακρύ και λεπτό άξονα, ο οποίος χωρίζεται σε χιλιάδες διακλαδώσεις. Στο τέλος κάθε διακλάδωσης, μία δομή η οποία ονομάζεται σύναψη, μετατρέπει την ενέργεια από τον άξονα σε ηλεκτρικούς παλμούς που είτε αναστέλλουν, είτε διεγείρουν τους διασυνδεδεμένους νευρώνες. Συνεπώς, όταν ένας νευρώνας δέχεται διεγερτικές εισόδους, τότε στέλνει ένα παλμό ηλεκτρικής ενέργειας κατά μήκος του άξονά του, μεταδίδοντας τη διέγερση σε άλλους διασυνδεδεμένους νευρώνες. Η μάθηση προκύπτει με την αλλαγή των συνάψεων, που έχει σαν αποτέλεσμα την αλλαγή της κατάστασης του νευρώνα αναλόγως την επιρροή που δέχεται.

#### *4.1.2.2 Δομή του τεχνητού νευρώνα*

Τα ΤΝΔ αρχικά προτάθηκαν ως ένα μαθηματικό μοντέλο προσομοίωσης της λειτουργίας του ανθρώπινου εγκεφάλου. Η δομή του εγκεφάλου είναι τέτοια ώστε να επιτρέπει την παράλληλη επεξεργασία δεδομένων και τη δυνατότητα συνεχούς μάθησης μέσω της αλληλεπίδρασης με το περιβάλλον. Τα δύο αυτά βασικά χαρακτηριστικά συμβάλλουν στην ικανότητα, αφενός, να εκτελεί δύσκολα καθήκοντα, όπως ταχύτατη αναγνώριση μορφών, αφετέρου, να εξελίσσεται συνεχώς, μαθαίνοντας από το περιβάλλον του κατά την αλληλεπίδρασή του με αυτό. Ένα πολύ σημαντικό στοιχείο είναι ότι η επίδραση ενός νευρώνα στους γειτονικούς του, μπορεί να είναι είτε διεγερτική, είτε ανασταλτική. Στη πλήρη αντιστοιχία με το απλοποιημένο μοντέλο του βιολογικού νευρώνα, αναπτύχθηκε το μοντέλο του τεχνητού νευρώνα.

Κάθε σήμα που μεταδίδεται από ένα νευρώνα σε ένα άλλο μέσα στον νευρωνικό δίκτυο συνδέεται με την τιμή βάρους, w, η οποία υποδηλώνει πόσο στενά είναι συνδεδεμένοι οι δύο νευρώνες μεταξύ τους. Όταν ένα νευρωνικό δίκτυο αλληλεπιδρά με το περιβάλλον και μαθαίνει από αυτό, τα συναπτικά βάρη μεταβάλλονται συνεχώς, ενδυναμώνοντας ή αποδυναμώνοντας την ισχύ του κάθε δεσμού. Η τιμή του κάθε συναπτικού βάρους συνήθως κυμαίνεται σε ένα συγκεκριμένο διάστημα, λ.χ. στο διάστημα από –1 ως 1, όμως αυτό είναι αυθαίρετο και εξαρτάται από το πρόβλημα που προσπαθούμε να λύσουμε. Η έννοια του συναπτικού βάρους μπορεί και να αποδοθεί ως ο χημικός δεσμός ανάμεσα σε δύο άτομα που απαρτίζουν ένα μόριο. Ο δεσμός μας δείχνει πόσο δυνατά είναι συνδεδεμένα τα δύο άτομα του μορίου. Έτσι και ένα βάρος αντιπροσωπεύει πόσο σημαντική είναι η συνεισφορά του συγκεκριμένου σήματος στην διαμόρφωση της δομής του δικτύου για τους δύο νευρώνες τους οποίους συνδέει. Όταν το βάρος (w) είναι μεγάλο/μικρό, τότε η συνεισφορά του σήματος είναι μεγάλη/μικρή.

Το ΤΝΔ έχει τη δυνατότητα να γενικεύει, δηλαδή να εξάγει τα βασικά χαρακτηριστικά ενός συστήματος, ακόμα και όταν αυτά είναι κρυμμένα σε θορυβώδη δεδομένα. Όλη η εμπειρική γνώση που αποκτά το νευρωνικό δίκτυο από το περιβάλλον, κωδικοποιείται στα συναπτικά βάρη. Αυτά αποτελούν το χαρακτηριστικό εκείνο που δίνει στο δίκτυο την ικανότητα για εξέλιξη και προσαρμογή στο περιβάλλον.

Υπάρχουν διάφορα είδη νευρώνων. Το είδος που θα επιλεχθεί για να δομηθεί ένα συγκεκριμένο ΤΝΔ, εξαρτάται από τη φύση του εκάστοτε προβλήματος που εξετάζουμε. Σε πολλές περιπτώσεις χρησιμοποιείται συνδυασμός διαφορετικών ειδών νευρώνων.

### 4.1.3 Μάθηση

Το ΤΝΔ είναι ένα πολύπλοκο προσαρμοστικό σύστημα που μπορεί να αλλάξει την εσωτερική του δομή με βάση τις πληροφορίες που διέρχονται από αυτό και την αλλαγή του περιβάλλοντος. Όντας ένα πολύπλοκο σύστημα προσαρμογής, η εκμάθηση στα Νευρωνικά Δίκτυα συνεπάγεται την προσαρμογή του δικτύου για την καλύτερη διαχείριση μιας εργασίας λαμβάνοντας υπόψη δείγματα παρατηρήσεων. Η μάθηση περιλαμβάνει την προσαρμογή των παραμέτρων του δικτύου (βάρη, προκαταλήψεις και προαιρετικά όρια) προκειμένου να αλλάξει η συμπεριφορά εισόδου/εξόδου και να βελτιωθεί η ακρίβεια του αποτελέσματος. Αυτό γίνεται ελαχιστοποιώντας τα παρατηρούμενα σφάλματα.
Η εκμάθηση ολοκληρώνεται όταν η εξέταση πρόσθετων παρατηρήσεων δεν μειώνει χρήσιμα το ποσοστό σφάλματος. Ακόμα και μετά την εκμάθηση, το ποσοστό σφάλματος συνήθως δεν φτάνει το 0 και, σε περίπτωση που είναι πολύ υψηλό, το δίκτυο πρέπει να επανασχεδιαστεί. Πρακτικά αυτό γίνεται καθορίζοντας μια συνάρτηση κόστους που αξιολογείται περιοδικά κατά τη διάρκεια της μάθησης. Όπως αναφέρεται στην Ενότητα 3.9, χρησιμοποιείται ένας αλγόριθμος βελτιστοποίησης προκειμένου να ελαχιστοποιηθεί η συνάρτηση καθορισμένου κόστους προσαρμόζοντας τις υπερ-παραμέτρους του δικτύου. Παραδείγματα υπερπαραμέτρων περιλαμβάνουν το ρυθμό εκμάθησης, τον αριθμό των εποχών και τον αριθμό των κρυφών στρωμάτων.

### 4.1.4 Οπισθοδιάδοση

Πρόκειται για την πιο διαδεδομένη μέθοδο για την εκπαίδευση ενός νευρωνικού δικτύου, είναι η μέθοδος της οπισθοδιάδοσης του λάθους. Η μέθοδος αυτή βασίζεται στην μαθηματική μέθοδο της ελαχιστοποίησης του σφάλματος με την τεχνική της πλέον απότομης καθόδου (steepest descent) στην επιφάνεια του σφάλματος, ένα πρόβλημα που ανήκει στη γενικότερη κατηγορία προβλημάτων επικλινούς καθόδου (gradient descent). Αυτό που επιτελεί είναι να ελαχιστοποιεί το τετράγωνο της διαφοράς μεταξύ του σήματος που λαμβάνεται στην έξοδο και της επιθυμητής τιμής (στόχος), για όλους τους νευρώνες εξόδου και για όλα τα πρότυπα. Αυτό σημαίνει ότι η παράγωγος του σφάλματος ως προς κάθε βάρος wij είναι ανάλογος ως προς την μεταβολή της τιμής του βάρους, όπως δίνεται από τον κανόνα Δέλτα.

### 4.1.5 Συναρτήσεις ενεργοποίησης τεχνητού νευρώνα

#### *4.1.5.1 Γραμμική συνάρτηση ενεργοποίησης*

Στην πιο απλή περίπτωση η έξοδος του νευρώνα αποτελεί μόνο το βεβαρημένο άθροισμα (weighted sum) των εισόδων στον νευρώνα συν τον όρο bias. Τέτοιου είδους γραμμικοί νευρώνες εξάγουν γραμμικό μετασχηματισμό του διανύσματος εισόδου. Αυτή η συνάρτηση μεταφοράς είναι πολύ χρήσιμη συνήθως στα πρώιμα στάδια ενός πολύ-επίπεδου νευρωνικού δικτύου, ενώ πιο συχνά τη συναντάμε όταν έχουμε να αντιμετωπίσουμε προβλήματα γραμμικής παλινδρόμησης (linear regression) και λιγότερο σε προβλήματα αναγνώρισης προτύπων. Ένα τεχνητό νευρωνικό δίκτυο που αποτελείται από τέτοιους νευρώνες θα είναι γραμμικό.

Το μαθηματικό μοντέλο που αντιπροσωπεύει τη γραμμική συνάρτηση μεταφοράς, αποτελεί την απλούστερη συνάρτηση μεταφοράς στην οποία η μεταβλητή m αντιπροσωπεύει το σύνολο των δειγμάτων εκπαίδευσης, η μεταβλητή των βαρών wi αντιπροσωπεύει την κλίση της ευθείας ή του πολύ-επιπέδου, ενώ η μεταβλητή b αντιπροσωπεύει τη μετατόπιση της ευθείας ως προς την αρχή των αξόνων. Στη vectorized μορφή, η τιμή του bias αναπαρίσταται από το x0 και πολλαπλασιάζεται με το ανάλογο βάρος στην κορυφή του διανύσματος βαρών. Ύστερα από την εύρεση της βέλτιστης γραμμής ή του πολύ-επιπέδου που τέμνει τα δεδομένα, οι τιμές των βαρών μπορούν να χρησιμοποιηθούν σαν αρχικές τιμές σε προβλήματα αναγνώρισης προτύπων.

#### *4.1.5.2 Βηματική συνάρτηση ενεργοποίησης*

Στην περίπτωση του νευρώνα δυαδικής κατάστασης, ο νευρώνας μπορεί να βρεθεί σε μία από τις δύο παρακάτω δυνατές καταστάσεις: να είναι ενεργός ή να είναι αδρανής (Heaviside function). Όταν ένας νευρώνας δέχεται διάφορα σήματα, τότε υπολογίζει μία ποσότητα x από όλα τα δεδομένα που έχει και συγκρίνει την τιμή της ποσότητας αυτής με μια τιμή κατωφλίου, θ, η οποία είναι χαρακτηριστική (σταθερή) και ορισμένη από την αρχή για τον νευρώνα αυτόν. Αν η τιμή της ποσότητας είναι μεγαλύτερη από την τιμή κατωφλίου, τότε λέμε ότι ο νευρώνας ενεργοποιείται. Αν όμως είναι μικρότερη, τότε ο νευρώνας παραμένει αδρανής. Επειδή ο νευρώνας εδώ δρα ως δυαδικό στοιχείο, η έξοδος της f(x), θα είναι 1 όταν είναι ενεργοποιημένος και 0 όταν είναι αδρανής.

Το παραπάνω μοντέλο αναφέρεται συχνά ως μοντέλο McCulloch-Pitts προς τιμή αυτών που το πρότειναν. Αργότερα, η εξέλιξη στο θεωρητικό υπόβαθρο των τεχνητών νευρωνικών δικτύων φανέρωσε ότι η παράγωγος της συνάρτησης ενεργοποίησης μπορεί να δώσει χρήσιμες πληροφορίες για το νευρωνικό δίκτυο και να χρησιμοποιηθεί στην εκπαίδευσή του, γεγονός που υποδεικνύει ότι είναι προτιμότερο να χρησιμοποιηθεί μία παραγωγίσιμη συνάρτηση και όχι η βηματική συνάρτηση, που είναι προφανώς μη παραγωγίσιμη.

#### *4.1.5.3 Σιγμοειδής συνάρτηση ενεργοποίησης*

Σήμερα, στα περισσότερα μοντέλα η συνάρτηση ενεργοποίησης είναι μία σιγμοειδής συνάρτηση. Αυτή είναι γενικά μία πραγματική, συνεχής και φραγμένη συνάρτηση, της οποίας η παράγωγος είναι θετική. Το πεδίο ορισμού της μπορεί θεωρητικά να είναι όλο το σύνολο των πραγματικών αριθμών, αλλά στην πράξη μπορεί να περιοριστεί, θέτοντας όρια στις τιμές των συναπτικών βαρών. Το σύνολο των τιμών ορίζεται στο διάστημα [0,1] παίρνοντας μόνο θετικές τιμές στην έξοδο. Ένα από τα πιο γνωστά παραδείγματα σιγμοειδούς συνάρτησης που χρησιμοποιείται ως συνάρτηση ενεργοποίησης είναι η λογιστική συνάρτηση (logistic function ή Fermi function), ενώ παρόμοια συνάρτηση μεταφοράς είναι αυτή της υπερβολικής εφαπτομένης (hyperbolic tangent) της οποίας η έξοδος μπορεί να πάρει τις τιμές [-1,1].
Με την εισαγωγή της συνάρτησης ενεργοποίησης, ο νευρώνας γίνεται μη γραμμικός. Αντίστοιχα, ένα ΤΝΔ που αποτελείται από τέτοιου είδους νευρώνες θα είναι μη γραμμικό. Αυτή η εγγενής μη γραμμικότητα των νευρωνικών δικτύων είναι ένα πλεονέκτημα έναντι άλλων γνωστών μεθόδων αντιμετώπισης πολλών προβλημάτων. Για παράδειγμα, όταν σε ένα πρόβλημα πρόβλεψης το σύστημα που μελετάμε είναι μη γραμμικό και ιδιαίτερα όταν παρουσιάζει χαοτική συμπεριφορά, τα γνωστά γραμμικά μοντέλα πρόβλεψης αδυνατούν να δώσουν σωστά αποτελέσματα. Σε αυτές τις περιπτώσεις, τα μη γραμμικά ΤΝΔ είναι προτιμότερα. Όσο μικρότερη είναι η παράμετρος του αθροίσματος, τόσο πιο πολύ συρρικνώνεται η συνάρτηση στον x άξονα. Η παραπάνω συνάρτηση μπορεί εύκολα να προσεγγιστεί συνδυάζοντας τη γραμμική συνάρτηση και τη συνάρτηση δυαδικής κατάστασης που σαν αποτέλεσμα έχει τη λογιστική συνάρτηση.

#### *4.1.5.4 Συνάρτηση υπερβολικής εφαπτομένης*

Ως λειτουργία ενεργοποίησης, η tanh χρησιμοποιείται ως επί το πλείστον για τις εισόδους μοντέλων που έχουν έντονα αρνητικές και θετικές τιμές όπως είναι το μηδενικό κέντρο. Οι έξοδοί της κυμαίνονται μεταξύ -1 και 1, διευκολύνοντας τις εισόδους μοντέλων που έχουν έντονα αρνητικές, ουδέτερες και έντονα θετικές τιμές. Οι τιμές με το μηδενικό κέντρο βοηθούν τον επόμενο νευρώνα κατά τη διάρκεια της διάδοσης. Η tanh είναι συνεχής και διαφορίσιμη σε όλα τα σημεία και ακολουθεί την:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

#### *4.1.5.5 Rectified Linear Unit - ReLU*

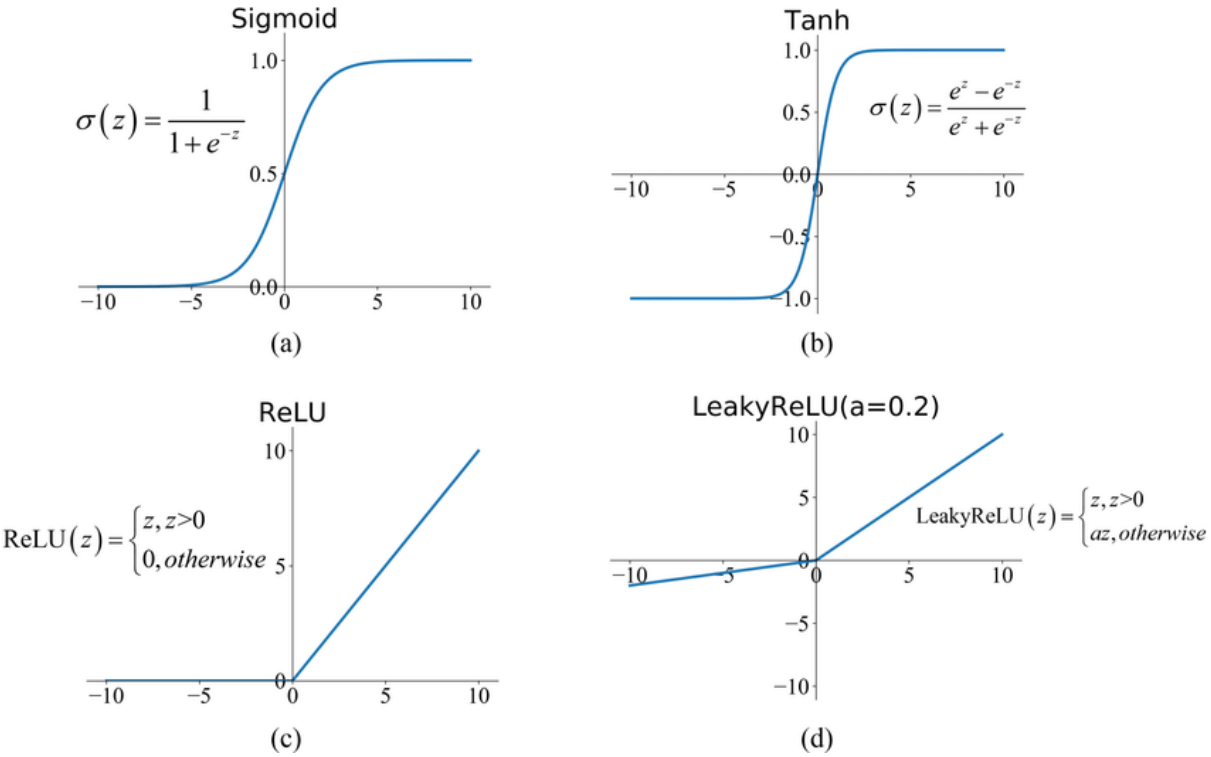
Η Rectified Linear Unit (Relu), επίσης γνωστή ως λειτουργία ράμπας, είναι μια μη γραμμική συνάρτηση και μία από τις πιο ευρέως χρησιμοποιούμενες. Είναι ένας απλός υπολογισμός που επιστρέφει την τιμή της εισόδου X, ή 0 εάν η τιμή εισόδου είναι αρνητική. Η ReLU, που απεικονίζεται στην Εικόνα 6, υπολογίζεται από την:

$$\mathrm{f}(x) = x^+ = \max(0, x)$$

όπου το x είναι η είσοδος ενός νευρώνα. Η ReLU είναι λιγότερο ακριβής υπολογιστικά από την tanh και τη σιγμοειδή και επιτρέπει στο δίκτυο να συγκλίνει πολύ γρήγορα. Το κύριο πλεονέκτημα της χρήσης της ReLU, ωστόσο, είναι ότι δεν ενεργοποιεί ταυτόχρονα όλους τους νευρώνες. Αυτό σημαίνει ότι οι νευρώνες θα απενεργοποιηθούν μόνο εάν η έξοδος του γραμμικού μετασχηματισμού είναι μικρότερη από 0.

Το εύρος της είναι [0, inf), που σημαίνει ότι δεν είναι δεσμευμένο. Επίσης, δεν είναι zero-centered. Το κύριο μειονέκτημα, ωστόσο, της ReLU είναι ένα πρόβλημα που ονομάζεται "dying ReLU problem", το οποίο είναι μια μορφή του προβλήματος βαθμίδων εξαφάνισης.

Όταν οι εισροές προσεγγίζουν το μηδέν, ή είναι αρνητικές, η κλίση της λειτουργίας καθίσταται μηδέν. Έτσι, κατά τη διάρκεια της διαδικασίας backprop, τα βάρη και οι προκαταλήψεις για ορισμένους νευρώνες δεν ενημερώνονται. Αυτό μπορεί να δημιουργήσει "νεκρούς" νευρώνες που δεν ενεργοποιούνται ποτέ, καθιστώντας ένα σημαντικό μέρος του δικτύου παθητικό. Αυτό το πρόβλημα εμφανίζεται συνήθως όταν ο ρυθμός εκμάθησης είναι πολύ υψηλός. Το πρόβλημα αυτό μπορεί να λυθεί από την Leaky ReLU, η οποία απλώς εκχωρεί μια μικρή θετική κλίση για το x < 0, ωστόσο η απόδοση μειώνεται.




*Εικόνα 6 Παραδείγματα συναρτήσεων (33)*

### *4.1.5.6 Softmax*

Η συνάρτηση softmax, γνωστή και ως συνάρτηση softargmax, χρησιμοποιείται συχνά ως η τελευταία συνάρτηση ενεργοποίησης ενός νευρωνικού δικτύου για την ομαλοποίηση της εξόδου ενός δικτύου, σε κατανομή πιθανότητας σε πολλαπλές κλάσεις εξόδου. Η συνάρτηση μετατρέπει ένα διάνυσμα $K$ πραγματικών τιμών σε ένα διάνυσμα $K$ πραγματικών τιμών που έχουν άθροισμα 1. Οι τιμές εισόδου μπορεί να είναι θετικές, αρνητικές, μηδενικές ή μεγαλύτερες από 1, αλλά το softmax τις μετατρέπει σε τιμές μεταξύ 0 και 1, ώστε να μπορούν να ερμηνευτούν ως πιθανότητες. Βασικά, ομαλοποιεί τις εξόδους για κάθε κλάση μεταξύ 0 και

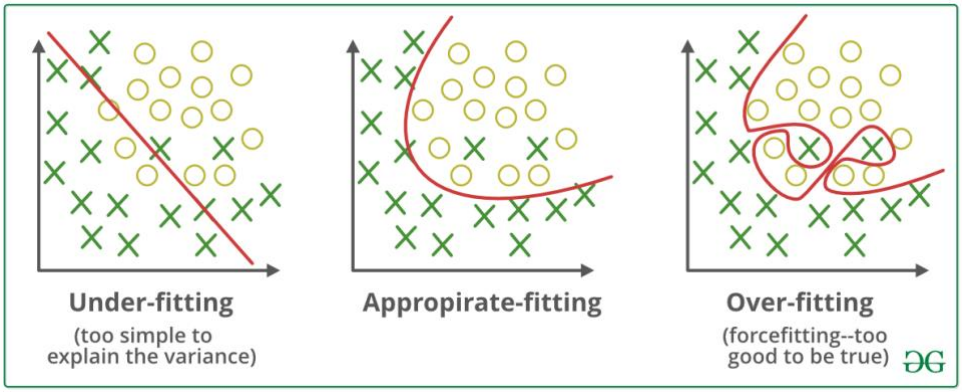
1 και διαιρείται με το άθροισμά τους, δίνοντας την πιθανότητα η τιμή εισόδου να βρίσκεται σε μια συγκεκριμένη κλάση. Η τυπική συνάρτηση softmax ορίζεται από τον τύπο:

$$f(x)_i = \frac{e^{x_i}}{\sum_{n=1}^{K} e^{x_n}}$$

### 4.1.6 Κανονικοποίηση – Regularization

Κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, και ιδιαίτερα ενός νευρωνικού δικτύου, υπάρχουν δύο κύρια είδη προβλημάτων που μπορεί να προκύψουν: προβλήματα μεροληψίας και διακύμανσης. Ο όρος μεροληψία αναφέρεται στις απλουστευτικές παραδοχές που γίνονται από ένα μοντέλο για να διευκολυνθεί η εκμάθηση του στόχου του δικτύου. Διακύμανση είναι το ποσό που θα αλλάξει η εκτίμηση της συνάρτησης στόχου εάν χρησιμοποιηθούν διαφορετικά δεδομένα εκπαίδευσης. Ο στόχος κάθε εποπτευόμενου αλγορίθμου ML είναι να επιτευχθεί χαμηλή μεροληψία και χαμηλή διακύμανση.

Τα προβλήματα μεγάλης μεροληψίας και χαμηλής διακύμανσης ονομάζονται υπό-προσαρμογή του μοντέλου και τα προβλήματα μεγάλης διακύμανσης και χαμηλής προκατάληψης ονομάζονται υπερβολική προσαρμογή του μοντέλου. Σε περίπτωση υπολειτουργίας, το δίκτυο δεν εκπαιδεύεται σωστά και οδηγεί σε υψηλά σφάλματα εκπαίδευσης και δοκιμών. Σε περίπτωση υπερβολικής προσαρμογής, το δίκτυο εκπαιδεύεται με τέτοιο τρόπο ώστε να έχει προσαρμοστεί για τις τιμές κατά τη διάρκεια της εκπαίδευσης, αλλά όταν δίνονται αόρατα δεδομένα, το μοντέλο αποδίδει άσχημα και δίνει λάθη.




*Εικόνα 7 Παραδείγματα προσαρμογής (34)*

Η συλλογή τεχνικών που χρησιμοποιούνται για την καταπολέμηση της υπερβολικής προσαρμογής και τη μείωση του σφάλματος στο σετ δοκιμών είναι γνωστή ως Κανονικοποίηση. Η κανονικοποίηση μπορεί να οριστεί ως οποιαδήποτε τροποποίηση σε έναν αλγόριθμο εκμάθησης που αποσκοπεί στη μείωση του σφάλματος γενίκευσης και, ιδανικά, δεν μειώνει το σφάλμα κατάρτισης. Αυτό επιτυγχάνεται συχνά με την προσθήκη ενός επιπλέον όρου ποινής στη συνάρτηση σφάλματος. Μερικές από τις πιο συχνά χρησιμοποιούμενες τεχνικές κανονικοποίησης παρουσιάζονται παρακάτω.

#### *4.1.6.1 Κανονικοποίηση L1*

Η κανονικοποίηση L1 προσθέτει μια ποινή L1 ίση με την απόλυτη τιμή του μεγέθους των συντελεστών. Με άλλα λόγια, περιορίζει το μέγεθος των συντελεστών. Η κανονικοποίηση αυτή μπορεί να αποδώσει αραιά μοντέλα, δηλαδή μοντέλα με λίγους συντελεστές. Τείνει να συρρικνώσει τους συντελεστές στο μηδέν και λόγω αυτού μπορεί να μηδενιστεί και να εξαλειφθεί. Η L1 είναι επομένως χρήσιμη για την επιλογή χαρακτηριστικών, καθώς όλες οι μεταβλητές που σχετίζονται με συντελεστές που πηγαίνουν στο μηδέν μπορούν να απορριφθούν.

#### *4.1.6.2 Κανονικοποίηση L2*

Η κανονικοποίηση L2 προσθέτει στη συνάρτηση απώλειας μια ποινή L2 ίση με το τετράγωνο του μεγέθους των συντελεστών. Δεν θα αποδώσει αραιά μοντέλα και όλοι οι συντελεστές συρρικνώνονται από τον ίδιο παράγοντα (κανένας δεν εξαλείφεται). Το L2 είναι χρήσιμο όταν τα χαρακτηριστικά εξαρτώνται από κώδικα. Αυτή η τεχνική είναι επίσης γνωστή ως αποσύνθεση βάρους αφού μειώνει τα μεγέθη των βαρών των νευρωνικών δικτύων κατά τη διάρκεια της εκπαίδευσης, με αποτέλεσμα ένα νέο βελτιωμένο μοντέλο που δεν είναι τόσο πιθανό να είναι υπερβολικά προσαρμοσμένο.

#### *4.1.6.3 Εγκατάλειψη - Dropout*

Η Εγκατάλειψη - Dropout (35) είναι μια τεχνική κανονικοποίησης για τη μείωση της υπερβολικής προσαρμογής σε νευρωνικά δίκτυα με την πρόληψη σύνθετων συν-προσαρμογών σε δεδομένα εκπαίδευσης. Η ιδέα είναι ότι επιλέγει τυχαία έναν αριθμό νευρώνων που

"εγκαταλείπονται" ή αγνοούνται μαζί με τις εισερχόμενες και εξερχόμενες συνδέσεις τους κατά τη διάρκεια της εκπαίδευσης. Αυτό σημαίνει ότι, σε κάθε κύκλο ενημέρωσης βάρους, η συμβολή των επιλεγμένων αγνοημένων κόμβων στην ενεργοποίηση των νευρώνων αφαιρείται προσωρινά στο εμπρόσθιο πέρασμα και οποιεσδήποτε ενημερώσεις βάρους δεν εφαρμόζονται στον νευρώνα στο οπίσθιο πέρασμα.

Το Dropout μπορεί εύκολα να εφαρμοστεί επιλέγοντας τυχαία κόμβους που θα εγκαταλειφθούν με μια δεδομένη πιθανότητα (ορίζεται από μια υπερ-παράμετρο εγκατάλειψης) σε κάθε επανάληψη. Με αυτόν τον τρόπο, κάθε επανάληψη αποτελείται από ένα διαφορετικό σύνολο κόμβων που παράγουν διαφορετικές εξόδους. Χρησιμοποιείται μόνο κατά την εκπαίδευση του μοντέλου και δεν χρησιμοποιείται κατά την αξιολόγησή του.

#### *4.1.6.4 Επαύξηση Δεδομένων - Data Augmentation*

Η επαύξηση δεδομένων είναι μια ενδιαφέρουσα τεχνική κανονικοποίησης που χρησιμοποιείται κυρίως όταν το διαθέσιμο σύνολο δεδομένων είναι μικρό. Αυτή η τεχνική παράγει νέα (ψεύτικα) δεδομένα εκπαίδευσης από το αρχικό σύνολο δεδομένων, αυξάνοντας τον αριθμό των παρατηρήσεων και επομένως ελέγχοντας την υπερβολική προσαρμογή. Πιο συγκεκριμένα, αυξάνει τον όγκο των δεδομένων προσθέτοντας ελαφρώς τροποποιημένα αντίγραφα ήδη υπαρχόντων δεδομένων ή νεοδημιουργηθέντα συνθετικά δεδομένα από τα υπάρχοντα. Είναι ένας φθηνός και εύκολος τρόπος για να αυξήθεί ο αριθμός των δεδομένων εκπαίδευσης.

#### *4.1.6.5 Πρόωρη Διακοπή – Early Stopping*

Ένα από τα μεγαλύτερα προβλήματα στην εκπαίδευση των νευρωνικών δικτύων, ειδικά στην αντιμετώπιση μεγάλων συνόλων δεδομένων, είναι το πόσο χρόνο εκπαιδεύεται το μοντέλο. Ένα μικρό χρονικό διάστημα θα οδηγήσει σε υπο-προσαρμογή του μοντέλου και ένα μεγάλο θα οδηγήσει σε υπερβολική προσαρμογή, με αποτέλεσμα να δώσει κακά αποτελέσματα όταν δοκιμαστεί. Έτσι, η πρόκληση είναι να εκπαιδεύσουμε το δίκτυο αρκετό διάστημα ώστε να είναι σε θέση να μάθει τη χαρτογράφηση από εισόδους σε εξόδους, αλλά όχι να εκπαιδευτεί τόσο πολύ ώστε να ταιριάζει υπερβολικά στα δεδομένα εκπαίδευσης.

Μια τεχνική για την αποφυγή υπερβολικής προσαρμογής είναι η χρήση σφάλματος επικύρωσης για να αποφασιστεί πότε θα σταματήσει η εκπαίδευση. Το σφάλμα επικύρωσης

αξιολογείται σε ένα σύνολο δεδομένων αναμονής, που ονομάζεται σύνολο επικύρωσης, μετά από κάθε κύκλο κατά τη δημιουργία του μοντέλου. Εάν η ακρίβεια του μοντέλου στο σύνολο επικύρωσης αρχίσει να υποβαθμίζεται (η απώλεια αρχίζει να αυξάνεται ή η ακρίβεια αρχίζει να μειώνεται), τότε η διαδικασία εκπαίδευσης διακόπτεται. Αυτή η τεχνική ονομάζεται Early Stopping. Προκειμένου να δηλωθεί ο αριθμός των κύκλων μετά τις οποίες η εκπαίδευση θα σταματήσει εάν δεν υπάρξει περαιτέρω βελτίωση, ορίζεται μια υπερ-παράμετρος που ονομάζεται υπομονή-patience.

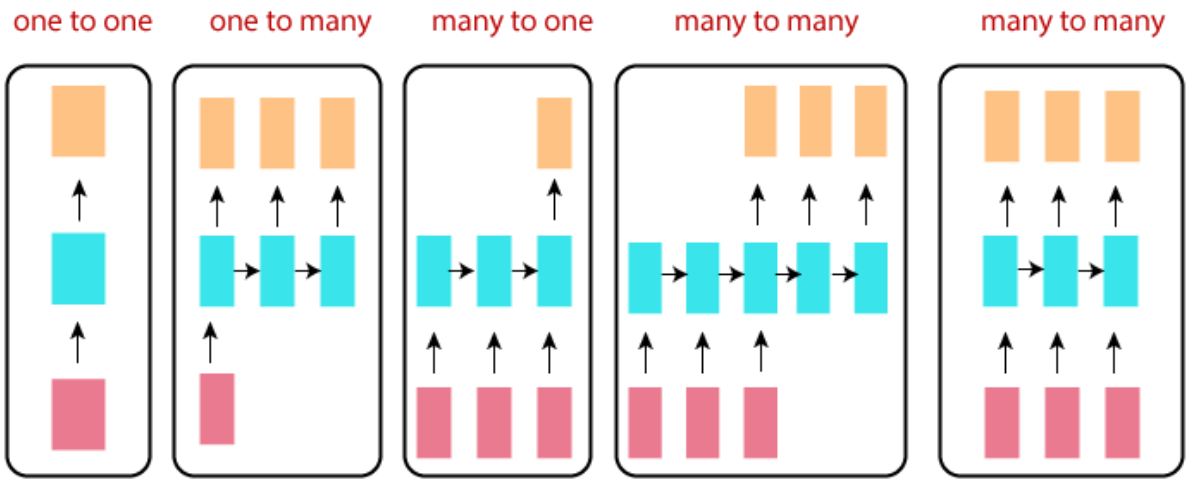
## 4.2 Ορισμός της Βαθιάς Μάθησης – Deep Learning

Το Deep Learning (DL), γνωστό και ως βαθιά δομημένη μάθηση, είναι μέρος ενός ευρύτερου τομέα της Μηχανική Μάθησης που βασίζεται σε τεχνητά νευρωνικά δίκτυα με μάθηση αναπαράστασης. Σύμφωνα με τον Jeff Dean (36), οι αλγόριθμοι Βαθιάς Μάθησης χρησιμοποιούν πολύ βαθιά νευρωνικά δίκτυα, όπου το «βαθύ» αναφέρεται στον αριθμό των στρωμάτων στο νευρωνικό δίκτυο ή στις επαναλήψεις μεταξύ εισόδου και εξόδου. Η Βαθιά Μάθηση μπορεί να είναι υπό επίβλεψη, ημι-επίβλεψη ή χωρίς επίβλεψη. Οι αρχιτεκτονικές Βαθιάς Μάθησης έχουν εφαρμοστεί σε πολυάριθμα πεδία σπουδών, όπως η όραση υπολογιστή, η αναγνώριση ομιλίας, η επεξεργασία φυσικής γλώσσας, η αναγνώριση ήχου, η αυτόματη μετάφραση, η πρόβλεψη χρονοσειρών και η βιοπληροφορική.

## 4.3 Αναδρομικά Νευρωνικά Δίκτυα – Recurrent Neural Networks (RNN)

Ένα Αναδρομικό Νευρωνικό Δίκτυο (RNN) είναι μια γενίκευση νευρωνικών δικτύων που έχουν εσωτερική "μνήμη". Είναι μια κατηγορία τεχνητών νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ των κόμβων σχηματίζουν ένα κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής ακολουθίας, το οποίο του επιτρέπει να επιδεικνύει χρονική δυναμική συμπεριφορά. Ενώ τα παραδοσιακά ουδέτερα δίκτυα υποθέτουν ότι οι είσοδοι και οι έξοδοι είναι ανεξάρτητες μεταξύ τους, τα RNN λαμβάνουν υπόψη την τρέχουσα είσοδο μαζί με την έξοδο από το προηγούμενο βήμα (προηγούμενη είσοδος) για τη λήψη απόφασης. Με άλλα λόγια, είναι δίκτυα με βρόχους ανάδρασης που επιτρέπουν τη διατήρηση των πληροφοριών - ένα χαρακτηριστικό που είναι ανάλογο με τη βραχυπρόθεσμη μνήμη. Σε αντίθεση με ένα δίκτυο προώθησης, ανάλογα με τις προηγούμενες εισόδους, η ίδια είσοδος μπορεί να παράγει διαφορετικές εξόδους.

Ο βασικός λόγος που τα RNN είναι σημαντικά είναι ότι, λόγω της εσωτερικής τους κατάστασης (μνήμης), επιτρέπουν τη λειτουργία σε αλληλουχίες διανυσμάτων: ακολουθίες στην είσοδο, την έξοδο ή στην πιο γενική περίπτωση και τα δύο. Είναι ιδανικά για προβλήματα μηχανικής μάθησης που περιλαμβάνουν κείμενο, ήχο, βίντεο και χρονοσειρές.




*Εικόνα 8 Διαφορετικά είδη αναδρομικών νευρωνικών δικτύων (37)*

### 4.3.1 Μονοκατευθυντικά ENN – Uni-Direxional RNN

Όπως τα feed-forward νευρωνικά δίκτυα, τα επαναλαμβανόμενα νευρωνικά δίκτυα χρησιμοποιούν δεδομένα εκπαίδευσης για τη μάθηση. Διακρίνονται από τη «μνήμη» τους, η οποία διατηρείται στο διάνυσμα κρυφής κατάστασης του επαναλαμβανόμενου δικτύου και αντιπροσωπεύει το πλαίσιο με βάση τις προηγούμενες εισόδους και εξόδους. Ενώ τα μελλοντικά γεγονότα θα ήταν επίσης χρήσιμα για τον προσδιορισμό της εξόδου μιας δεδομένης ακολουθίας, τα μονόδρομα επαναλαμβανόμενα νευρωνικά δίκτυα δεν μπορούν να συμπεριλάβουν αυτά τα γεγονότα στις προβλέψεις τους.

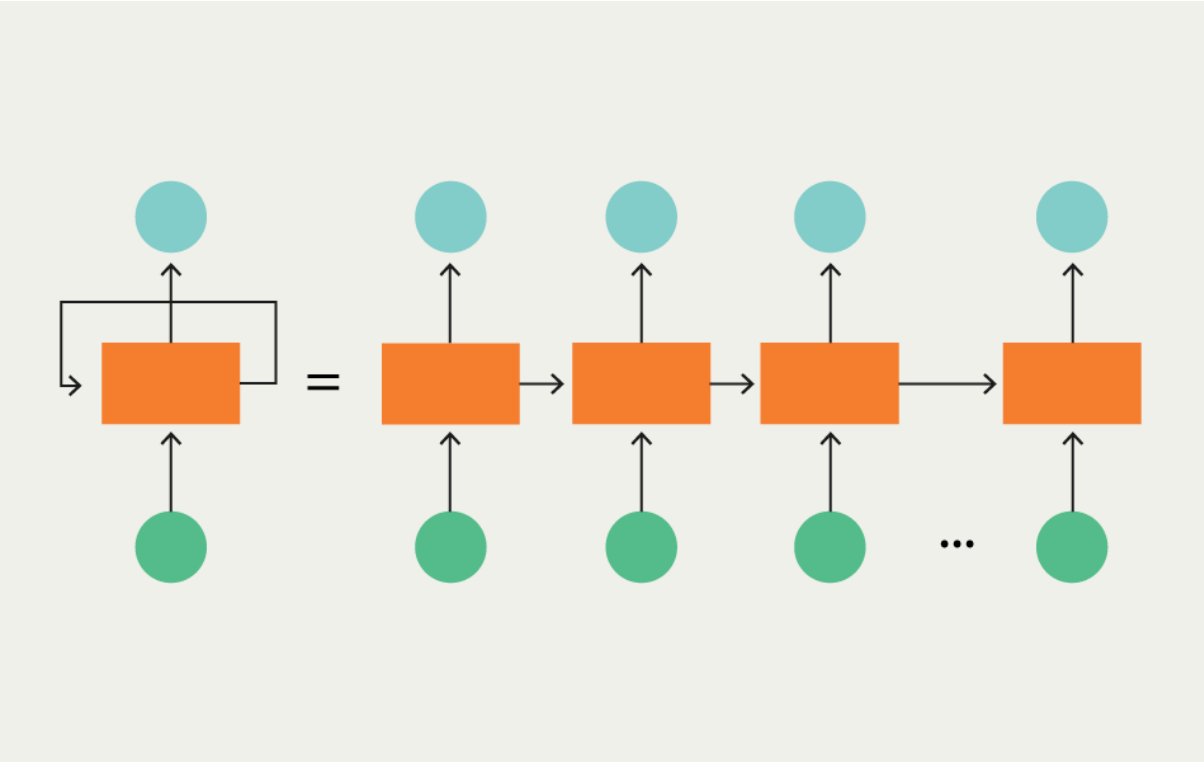
Ένα one-to-many RNN φαίνεται στο παρακάτω σχήμα. Πρώτον, κατά τη διάδοση προς τα εμπρός, το RNN παίρνει το $x_1$ από την ακολουθία εισόδου και παράγει μια έξοδο $h_1$ (κρυφή κατάσταση). Σημειώστε ότι η παραγωγή της κρυφής κατάστασης είναι απλώς το γινόμενο της εισόδου και της κρυφής κατάστασης με έναν πίνακα με βάρη W. Στο επόμενο χρονικό βήμα, η κρυφή κατάσταση $h_1$ μαζί με την επόμενη είσοδο $x_2$ δίνεται ως είσοδος. Κατά συνέπεια, το $h_1$ μαζί με το $x_2$ είναι η είσοδος για το επόμενο βήμα και ούτω καθεξής. Έτσι, ο μηχανισμός "μνήμης" των RNN εφαρμόζεται χρησιμοποιώντας ένα εσωτερικό κρυφό στρώμα που παράγει

μια κρυφή κατάσταση, η οποία θυμάται όλες τις πληροφορίες σχετικά με τα αυτά που έχουν υπολογιστεί νωρίτερα. Τυπικά, σε κάθε χρονικό βήμα $t$, η κρυφή κατάσταση $h_t$ και η έξοδος $y_t$ υπολογίζονται ως εξής:

$$h_t = f_h(W_{hh}h_{t-1} + W_{hx}x_t + b_h)$$

$$y_t = f_y(W_{yh}h_t + b_y)$$

όπου $x_t$ είναι το διάνυσμα εισόδου, $b_h$ είναι η μεροληψία για το $h$, $b_y$είναι η μεροληψία για το $y$ και $f_h$, $f_y$ είναι οι συναρτήσεις ενεργοποίησης για $h$ και $y$ αντίστοιχα. Επίσης, το $W_{hh}$ αντιπροσωπεύει το βάρος στην προηγούμενη κρυφή κατάσταση, το $W_{hx}$ το βάρος στην τρέχουσα κατάσταση εισόδου και $W_{yh}$ το βάρος στην κατάσταση εξόδου.



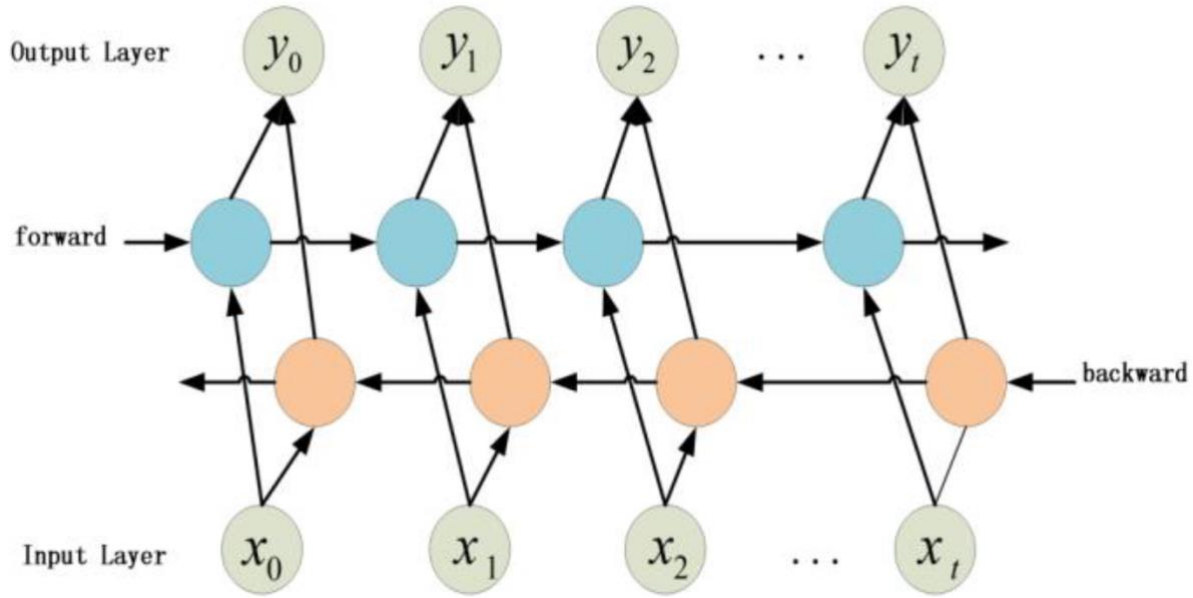
*Εικόνα 9 Ένα παράδειγμα Αναδρομικού Νευρωνικού Δικτύου με ανάλυση στον χρόνο (38)*

Τα RNN αξιοποιούν τον αλγόριθμο Back-Propagation Through Time (BPTT) για να καθορίσουν τις κλίσεις. Αυτός ο αλγόριθμος διαφέρει από το παραδοσιακό backprop που περιγράφεται στην Ενότητα 4.1.4. Οι αρχές του BPTT είναι ίδιες με την παραδοσιακή

οπισθοδρόμηση, όπου το μοντέλο εκπαιδεύεται υπολογίζοντας σφάλματα από την εξόδο στην είσοδο προκειμένου να προσαρμόσει τις παραμέτρους του μοντέλου κατάλληλα. Ωστόσο, το BPTT διαφέρει επειδή αθροίζει σφάλματα σε κάθε χρονικό βήμα, ενώ τα feed-forward δίκτυα δεν αθροίζουν τα σφάλματα καθώς δεν μοιράζονται παραμέτρους μεταξύ των επιπέδων.
Μέσω αυτής της διαδικασίας, τα RNN τείνουν να αντιμετωπίζουν δύο μεγάλα προβλήματα, γνωστά ως διαβαθμίσεις εξαφάνισης (vanishing gradient) και διαβαθμίσεις έκρηξης (exploding gradient). Αυτά τα ζητήματα καθορίζονται από το μέγεθος της κλίσης (η κλίση της συνάρτησης απώλειας κατά μήκος της καμπύλης σφάλματος). Όταν η κλίση είναι μικρή, συνεχίζει να γίνεται μικρότερη, ενημερώνοντας τις παραμέτρους βάρους μέχρι να γίνουν ασήμαντες. Αυτό καθιστά δύσκολη την εκμάθηση μεγάλων ακολουθιών δεδομένων. Αντίθετα, οι διαβαθμίσεις έκρηξης συμβαίνουν όταν η κλίση τείνει να αυξάνεται εκθετικά αντί να φθείρεται, δημιουργώντας ένα ασταθές μοντέλο. Σε αυτή την περίπτωση, οι παράμετροι βάρους θα αυξηθούν υπερβολικά (πλησιάζοντας το άπειρο) κατά τη διάρκεια της διαδικασίας εκπαίδευσης με αποτέλεσμα μία πολύ κακή απόδοση. Μια λύση σε αυτά τα ζητήματα είναι η μείωση του αριθμού των κρυφών στρωμάτων εντός του RNN, εξαλείφοντας μέρος της πολυπλοκότητάς του.

### 4.3.2 Αμφίδρομα ENN – Βι-Direxτional RNN

Το αμφίδρομο επαναλαμβανόμενο νευρωνικό δίκτυο (BRNN) (39) είναι μια παραλλαγή αρχιτεκτονικής δικτύου RNN και εισήχθη με σκοπό να αυξήσει τον όγκο των διαθέσιμων πληροφοριών εισόδου στο δίκτυο. Ενώ τα μονοκατευθυντικά RNN μπορούν να αντλήσουν πληροφορίες από προηγούμενες εισόδους (προς τα πίσω) για να κάνουν προβλέψεις σχετικά με την τρέχουσα κατάσταση, τα BRNN εισάγουν ταυτόχρονα μελλοντικά δεδομένα για να βελτιώσουν την ακρίβειά τους.
Η δομή ενός τυπικού BRNN απεικονίζεται στο παρακάτω σχήμα. Τα BRNN χωρίζουν τους νευρώνες ενός κανονικού RNN σε δύο κατευθύνσεις, μία για τη θετική χρονική κατεύθυνση (προς τα εμπρός) και μια άλλη για την αρνητική χρονική κατεύθυνση (προς τα πίσω). Η έξοδος αυτών των δύο καταστάσεων δεν συνδέεται με εισόδους των καταστάσεων αντίθετης κατεύθυνσης. Για να βρούμε την κρυφή κατάσταση για κάθε χρονικό βήμα, τα BRNN συνδυάζουν τα δύο κρυμμένα επίπεδα αντίθετων κατευθύνσεων στην ίδια έξοδο.




*Εικόνα 10 Παράδειγμα Αμφίδρομου Αναδρομικού Νευρωνικού Δικτύου (40)*

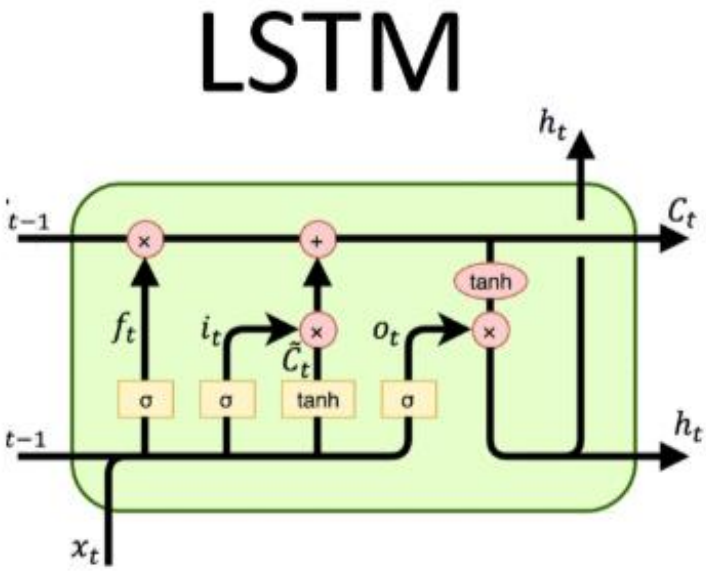
Τα BRNN μπορούν να εκπαιδευτούν χρησιμοποιώντας παρόμοιους αλγόριθμους με τους RNN, επειδή οι δύο νευρώνες δεν έχουν καμία αλληλεπίδραση. Ωστόσο, όταν εφαρμόζεται το BPTT, απαιτούνται πρόσθετες διαδικασίες επειδή η ενημέρωση των επιπέδων εισόδου και εξόδου δεν μπορεί να γίνει ταυτόχρονα. Γενικά, για τη διέλευση προς τα εμπρός, περνούν πρώτα οι forward και backward καταστάσεις και στη συνέχεια περνούν οι νευρώνες εξόδου. Για το πέρασμα προς τα πίσω, οι νευρώνες εξόδου περνούν πρώτα και στη συνέχεια περνούν οι forward και backward καταστάσεις. Αφού γίνουν πάσες εμπρός και πίσω, τα βάρη ενημερώνονται.

### 4.3.3 Long Short-Term Memory Networks

Τα Long Short-Term Memory (LSTM) (41) είναι μια δημοφιλής αρχιτεκτονική RNN που χρησιμοποιείται στον τομέα της βαθιάς μάθησης. Εισήχθη ως λύση στο πρόβλημα της διαβάθμισης που εξαφανίζεται στην οπίσθια διάδοση, γεγονός που προκαλεί τα RNN να είναι εγγενώς ανεπαρκή στη διατήρηση πληροφοριών για μεγάλα χρονικά διαστήματα. Πιο συγκεκριμένα, εάν η προηγούμενη κατάσταση που επηρεάζει την τρέχουσα πρόβλεψη δεν είναι στο πρόσφατο παρελθόν, το μοντέλο RNN μπορεί να μην είναι σε θέση να προβλέψει με ακρίβεια την τρέχουσα κατάσταση. Τα LSTM μπορούν να ξεπεράσουν αυτό το πρόβλημα διατηρώντας τις βασικές πληροφορίες, να τις διατηρήσουν για μεγάλα χρονικά διαστήματα και

στη συνέχεια να χρησιμοποιήσουν αυτές τις πληροφορίες όταν είναι απαραίτητο πολύ αργότερα στην ακολουθία.

Η δομή ενός LSTM απεικονίζεται στην Εικόνα 11. Αποτελείται από "κελιά" στα κρυμμένα στρώματα, τα οποία έχουν τρεις πύλες: μια forget πύλη, μια πύλη εισόδου και μια πύλη εξόδου. Ένα κελί θυμάται τιμές σε αυθαίρετα χρονικά διαστήματα και οι τρεις πύλες ρυθμίζουν τη ροή πληροφοριών μέσα και έξω από το κελί. Σε κάθε χρονικό βήμα $t$, κάθε κελί του LSTM ενημερώνει το διάνυσμα εσωτερικής κατάστασης $c_t$ και δημιουργεί ένα κρυφό διάνυσμα εξόδου $h_t$ με βάση την κατάσταση του κελιού. Στη συνέχεια περιγράφονται οι εξισώσεις για το εμπρός πέρασμα μιας μονάδας LSTM με forget πύλη.




*Εικόνα 11 LSTM (42)*

- Forget πύλη: Αποφασίζει ποιες μη σημαντικές πληροφορίες θα διαγραφθούν από την κατάσταση του κελιού. Η τρέχουσα είσοδος $x_t$ και η προηγούμενη κρυφή κατάσταση $h_{t-1}$ περνούν μέσω της συνάρτησης sigmoid, η οποία εξάγει έναν αριθμό μεταξύ 0 (πρέπει να ξεχαστεί) και 1 (πρέπει να διατηρηθεί) για κάθε bit στην κατάσταση κελιού $c_{t-1}$. Το διάνυσμα ενεργοποίησης της forget πύλης υπολογίζεται ως εξής:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$

- Πύλη εισόδου/ενημέρωσης: Ελέγχει ποιες πληροφορίες πρέπει να αποθηκεύονται στην κατάσταση κελιού. Αρχικά, η σιγμοειδής συνάρτηση (που ονομάζεται στρώμα πύλης εισόδου) αποφασίζει ποιες τιμές θα περάσουν, στριμώχνοντάς τις μεταξύ 0 και 1 ως εξής:

$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i)$$

όπου το $i$ είναι το διάνυσμα ενεργοποίησης της πύλης εισόδου. Στη συνέχεια, η προηγούμενη κρυφή κατάσταση και η τρέχουσα είσοδος περνούν επίσης στη συνάρτηση $tanh$, η οποία αναγκάζει τις τιμές να είναι μεταξύ $-1$ και 1, αποφασίζοντας το επίπεδο σημασίας τους. Μαθηματικά, το διάνυσμα των νέων υποψήφιων τιμών που θα μπορούσαν να προστεθούν στην κατάσταση, που ονομάζεται διάνυσμα ενεργοποίησης εισόδου κελιού, ορίζεται ως:

$$\widetilde{c_t} = tanh(W_c x_t + U_c h_{t-1} + b_c)$$

Πολλαπλασιάζοντας την έξοδο $tanh$ με την σιγμοειδή έξοδο, φιλτράρονται οι σημαντικές πληροφορίες της τρέχουσας εισόδου.

- Κατάσταση κελιού: Η παλιά κατάσταση $c_{t-1}$ πολλαπλασιάζεται με το διάνυσμα λήψης για να ξεχαστούν οι πληροφορίες από τα προηγούμενα βήματα. Η έξοδος αυτού του πολλαπλασιασμού προστίθεται στη συνέχεια στο διάνυσμα εξόδου από την πύλη εισόδου (it ◦ c̃t) προκειμένου να ενημερωθεί η κατάσταση. Η κατάσταση του κελιού $c_t$ περιέχει τώρα τις νέες τιμές που το νευρωνικό δίκτυο θεωρεί σχετικές.

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c_t}$$

- Πύλη εξόδου: Τέλος, αυτή η πύλη αποφασίζει την έξοδο με βάση την είσοδο και τη μνήμη του μπλοκ. Η σιγμοειδής συνάρτηση περιορίζει για άλλη μια φορά την προηγούμενη κρυφή κατάσταση και την τρέχουσα είσοδο μεταξύ 0 και 1 για να αποφασίσει ποιες τιμές θα περάσουν. Έτσι, το διάνυσμα ενεργοποίησης της πύλης εξόδου υπολογίζεται ως εξής:

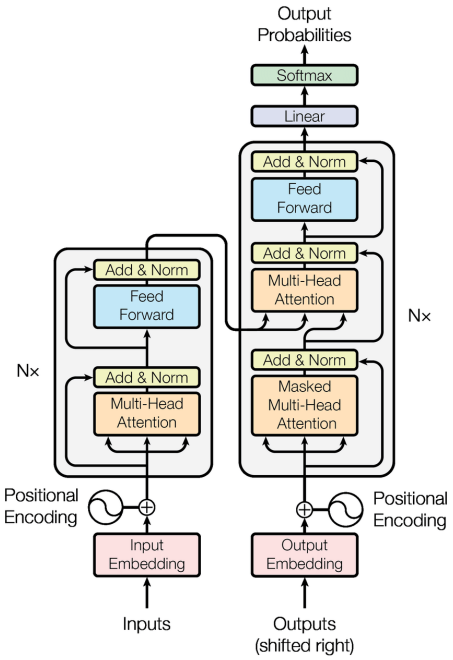
$$c_t = o_t \circ \tanh(c_t)$$

Σε αυτό το σημείο αξίζει να σημειωθεί ότι $W$, $U$ και $b$ είναι οι πίνακες βάρους και οι παράμετροι για τις οποίες πρέπει να γίνει η εκπαίδευση. Επίσης, οι αρχικές τιμές είναι $c_0 = 0$ και $h_0 = 0$ και ο τελεστής ◦ δηλώνει το προϊόν Hadamard.

## 4.4 Μηχανισμοί Προσοχής – Attention Mechanism

Το Attention (προσοχή) (43) είναι ένας μηχανισμός που αναπτύχθηκε ως λύση στους περιορισμούς του μοντέλου Κωδικοποιητή - Αποκωδικοποιητή για τα δίκτυα που πραγματοποιούσαν αυτόματη μετάφραση (44) (45), όπου οι ακολουθίες εισόδου διαφέρουν κατά μήκος από τις ακολουθίες εξόδου. Συγκεκριμένα, βοηθά στην απομνημόνευση μεγάλων προτάσεων σε μορφή αναγνωρίσιμη από το νευρωνικό δίκτυο.

Το Attention είναι μια τεχνική που μιμείται τη γνωστική προσοχή και είναι μια από τις σημαντικότερες ανακαλύψεις στην έρευνα του DL την τελευταία δεκαετία. Βοηθά τα μοντέλα να κατευθύνουν την εστίασή τους και να αφιερώνουν περισσότερη υπολογιστική ισχύ στα σημαντικά μέρη των δεδομένων εισόδου και όχι σε άλλα λιγότερο σημαντικά. Ποιο μέρος των δεδομένων είναι πιο σημαντικό εξαρτάται από το πλαίσιο και μαθαίνεται μέσω των δεδομένων εκπαίδευσης. Επομένως, το Attention στη βαθιά μάθηση μπορεί να ερμηνευτεί ευρέως ως ένα διάνυσμα βαρών σπουδαιότητας. Οι μηχανισμοί προσοχής χρησιμοποιούνται εκτενώς στα δίκτυα μετασχηματιστών (46).




*Εικόνα 12 Ένας Μετασχηματιστής (46)*

Ο μηχανισμός προσοχής που εισήχθη στη νευρωνική μηχανική μετάφραση (43) μπορεί να οριστεί ως εξής. Έστω μια ακολουθία πηγής $x$ μήκους $n$, $x = [x_1, x_2, \dots, x_n]$ και μια ακολουθία εξόδου $y$ μήκους $n$, $y = [y_1, y_2, \dots, y_n]$. Επίσης, ας υποθέσουμε ότι ο κωδικοποιητής είναι ένα BRNN με κρυφή κατάσταση προς τα εμπρός και προς τα πίσω. Μια απλή συνένωση των δύο αντιπροσωπεύει την κατάσταση κωδικοποιητή $h_i$, $i = 1, \dots, n$. Τόσο οι προηγούμενες όσο και οι επόμενες λέξεις πρέπει να περιλαμβάνονται στον σχολιασμό μίας λέξης. Για τη λέξη εξόδου στη θέση $t$, $t = 1, \dots, m$, το δίκτυο αποκωδικοποιητή έχει μια κρυφή κατάσταση:

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

όπου το $c_t$ είναι ένα σταθμισμένο άθροισμα των βαθμών ευθυγράμμισης των κρυφών καταστάσεων της ακολουθίας εισόδου:

$$c_t = \sum_{i-1}^{n} a_{t,i}\, h_i$$

$$a_{t,i} = align(y_t, x_i) = \frac{e^{score(s_{t-1}, h_i)}}{\sum_{i'=1}^{n} e^{score(s_{t-1}, h_{i-1})}}$$

Το μοντέλο ευθυγράμμισης αποδίδει μια βαθμολογία $a_{t,i}$ για το ζεύγος $(y_t, x_i)$, με βάση το πόσο καλά ταιριάζουν. Το σύνολο των $\{a_{t,i}\}$ είναι βάρη και καθορίζουν το πόσο η είσοδος από κάθε κρυφή κατάσταση πρέπει να λαμβάνεται υπόψη για κάθε έξοδο.

Μερικοί από τους πιο συνηθισμένους μηχανισμούς προσοχής που χρησιμοποιούνται είναι το additive (47), το dot-product (48) και το content-based attention (49). Επίσης, οι μηχανισμοί προσοχής μπορούν να κατηγοριοποιηθούν σε ευρύτερες κατηγορίες, όπως ο αυτό-έλεγχος (self-attention) (50). Το τελευταίο, επίσης γνωστό ως intra-attention, είναι ένας μηχανισμός που συνδέει διαφορετικές θέσεις μιας μεμονωμένης πρότασης προκειμένου να υπολογίσει μια αναπαράσταση της ίδιας πρότασης.

## 4.5 Μεταφορική Μάθηση – Transfer Learning

Η Μεταφορική Μάθηση (TL) είναι μια τεχνική μηχανικής μάθησης όπου ένα μοντέλο εκπαιδευμένο σε ένα έργο επαναχρησιμοποιείται ως σημείο εκκίνησης για ένα μοντέλο σε ένα άλλο έργο. Είναι μια δημοφιλής προσέγγιση ειδικά στη βαθιά μάθηση λόγω των τεράστιων υπολογιστικών και χρονικών πόρων που απαιτούνται για την ανάπτυξη νευρωνικών δικτύων για εργασίες όπως η όραση υπολογιστή και η επεξεργασία φυσικής γλώσσας. Επίσης, επιτρέπει την ανάπτυξη μοντέλων βαθιάς νευρωνικής δικτύωσης με συγκριτικά λίγα δεδομένα. Αυτό είναι πραγματικά χρήσιμο καθώς τα εποπτευόμενα μοντέλα που επιλύουν σύνθετα προβλήματα απαιτούν τεράστιες ποσότητες επισημασμένων δεδομένων που συνήθως δεν μπορούν να ληφθούν λόγω του χρόνου και της προσπάθειας που απαιτείται. Επιπλέον, τα περισσότερα μοντέλα που είναι εξειδικευμένα σε μια συγκεκριμένη εργασία έχουν μειωμένη απόδοση όταν χρησιμοποιούνται σε νέες εργασίες, ακόμα κι αν είναι παρόμοιες με αυτές στις οποίες εκπαιδεύτηκαν. Ο στόχος του TL είναι να βελτιώσει την απόδοση του νέου έργου αξιοποιώντας τη γνώση από το προηγούμενο.

Κεφάλαιο 5

# Επεξεργασία Φυσικής Γλώσσας

## 5.1 Ορισμός

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι ένας κλάδος της Τεχνητής Νοημοσύνης, της επιστήμης των υπολογιστών και της υπολογιστικής γλωσσολογίας (βασισμένη σε κανόνες μοντελοποίησης της ανθρώπινης γλώσσας) που αφορά κυρίως τις αλληλεπιδράσεις μεταξύ υπολογιστών και φυσικών (ανθρώπινων) γλωσσών. Με απλά λόγια, το NLP αντιπροσωπεύει την αυτόματη επεξεργασία φυσικής γλώσσας, όπως ομιλία (φωνητικά δεδομένα) ή κείμενο, προκειμένου να κατανοηθεί, να ερμηνευθεί και να επεξεργαστεί. Χρησιμοποιώντας το NLP, οι μηχανές μπορούν να εκτελέσουν εργασίες όπως αυτόματη σύνοψη κειμένου, ανάλυση συναισθημάτων, αναγνώριση ομιλίας, μετάφραση, εξαγωγή σχέσεων, τμηματοποίηση θεμάτων και άλλα.

Η κύρια πρόκληση που αντιμετωπίζει το NLP είναι ότι η ανθρώπινη γλώσσα είναι γεμάτη με ασάφειες. Εκτός από την ύπαρξη πολλών διαφορετικών γλωσσών και διαλέκτων, κάθε γλώσσα περιλαμβάνει παρατυπίες στη γραμματική και τη σύνταξη, ιδιώματα, σαρκασμούς, μεταφορές, ομώνυμα και ομόφωνα. Έτσι, η κατανόηση και ο χειρισμός της γλώσσας είναι εξαιρετικά περίπλοκες διαδικασίες, γεγονός που καθιστά δύσκολη τη δημιουργία λογισμικού που να μπορεί να εξάγει το νόημα ενός κειμένου. Για το λόγο αυτό, είναι συνηθισμένο να χρησιμοποιούνται αλγόριθμοι όπως οι tokenization, stop words removal, stemming, lemmatization, topic modeling, κ.λπ..

## 5.2 Sentiment Analysis

Η ανάλυση συναισθημάτων, που ονομάζεται επίσης εξόρυξη απόψεων ή τεχνητή νοημοσύνη συναισθημάτων, είναι ένας τύπος έρευνας κειμένου ή εξόρυξης κειμένου. Εφαρμόζει ένα μείγμα NLP, ανάλυσης κειμένου, υπολογιστικής γλωσσολογίας και βιομετρίας για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών (απόψεις ανθρώπων, συναισθήματα, αξιολογήσεις κ.λπ.) για οντότητες όπως προϊόντα, εκδηλώσεις, θέματα και υπηρεσίες. Περιλαμβάνει την ταξινόμηση των συναισθημάτων στο κείμενο σε κατηγορίες όπως "θετικά", "αρνητικά" ή "ουδέτερα". Κατά συνέπεια, η ανάλυση συναισθημάτων μπορεί να θεωρηθεί ως πρόβλημα ταξινόμησης κειμένου που στοχεύει στην κατηγοριοποίηση ενός κειμένου με βάση

την αίσθηση που περιέχει το κείμενο. Μπορεί να εφαρμοστεί στα ακόλουθα διαφορετικά επίπεδα:

- Ανάλυση συναισθημάτων σε επίπεδο εγγράφου: Το συναίσθημα εξάγεται από ολόκληρο το έγγραφο και μια ολόκληρη γνώμη ταξινομείται με βάση το γενικό συναίσθημα του κατόχου της γνώμης. Ο στόχος είναι να ταξινομηθεί το έγγραφο σε θετικό, αρνητικό ή ουδέτερο συναίσθημα.
- Ανάλυση συναισθημάτων σε επίπεδο πρότασης: Συνδέεται με μια φράση ή πρόταση. Καθορίζει αν κάθε πρόταση εκφράζει θετική, αρνητική ή ουδέτερη γνώμη για ένα προϊόν ή μια υπηρεσία.
- Ανάλυση συναισθημάτων βασισμένη σε όψη: Είναι η ανάλυση γνώμης βάσει των χαρακτηριστικών ενός προϊόντος σε μια αξιολόγηση. Ο στόχος είναι να εντοπιστούν και να εξαχθούν τα χαρακτηριστικά του προϊόντος και να προσδιοριστεί εάν η γνώμη είναι θετική, αρνητική ή ουδέτερη.

## 5.3 Text pro-processing

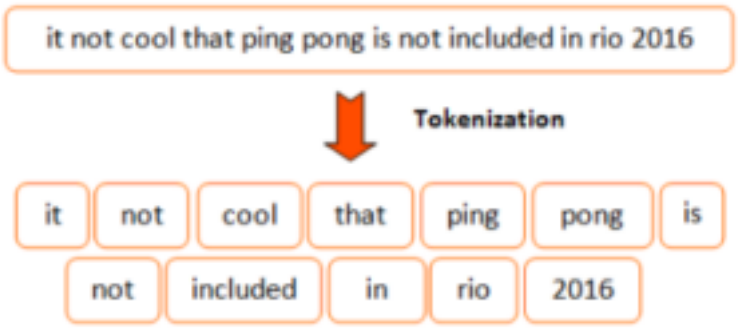
Για την ανάπτυξη μιας εφαρμογής NLP, συνήθως απαιτείται ένας τεράστιος όγκος δεδομένων κειμένου. Μια μεγάλη και δομημένη συλλογή από μηχανικά αναγνώσιμα κείμενα που έχουν παραχθεί από γραπτό ή προφορικό υλικό φυσικής γλώσσας ονομάζεται corpus (πληθυντικός αριθμός). Ως εκ τούτου, είναι ζωτικής σημασίας η προετοιμασία και η μετατροπή των δεδομένων κειμένου σε μια τυπική μορφή που μπορεί εύκολα να επεξεργαστεί με αλγόριθμους.

### 5.3.1 Toketization

Το Tokenization είναι μια μέθοδος κατά την οποία ένα κομμάτι κειμένου διαχωρίζεται σε μικρότερες μονάδες ή "tokens". Τα tokens είναι τα δομικά στοιχεία της φυσικής γλώσσας και μπορούν να είναι χαρακτήρες, λέξεις ή υπολέξεις (ένα πλήθος από n χαρακτήρες). Ως εκ τούτου, μπορούμε να διακρίνουμε τρεις τύπους:

- tokenization χαρακτήρα,
- tokenization λέξης
- tokenization υπολέξης (χαρακτήρες n-gram).

Η διακριτοποίηση μπορεί επίσης να απορρίψει ορισμένους περιττούς χαρακτήρες, όπως τα σημεία στίξης, διευκολύνοντας τη διαδρομή προς μια σωστή κατάτμηση λέξεων αλλά και ενεργοποιώντας πιθανές συμβιβασμούς. Για παράδειγμα, όταν ασχολούμαστε με βιοϊατρικούς τομείς κειμένου που περιέχουν πολλά σημεία στίξης, η διακριτοποίηση μπορεί να είναι ιδιαίτερα προβληματική.




*Εικόνα 13 Μια απλή διαδικασία Toketization (51)*

Ο πιο συνηθισμένος τρόπος επεξεργασίας ακατέργαστου κειμένου συμβαίνει σε επίπεδο των tokens, οπότε το πρώτο βήμα είναι η δημιουργία αυτών των tokens. Σε μια εργασία NLP, το tokenization πραγματοποιείται στο corpus για τη λήψη tokens, τα οποία στη συνέχεια χρησιμοποιούνται για την προετοιμασία ενός λεξιλογίου. Το λεξιλόγιο αναφέρεται στο σύνολο των μοναδικών μαρκών που χρησιμοποιούνται στο σώμα κειμένου και μπορεί να κατασκευαστεί λαμβάνοντας υπόψη είτε κάθε μοναδικό διακριτικό στο σώμα είτε τις κορυφαίες k που εμφανίζονται συχνά.

### 5.3.2 Lowercasing

Η μείωση του αριθμού όλων των δεδομένων κειμένου ή των διακριτικών λέξεων (tokens) είναι μια από τις απλούστερες και πιο αποτελεσματικές μορφές προ-επεξεργασίας κειμένου. Ισχύει για τις περισσότερες εργασίες εξόρυξης κειμένου και NLP καθώς βοηθά στη συνέπεια της αναμενόμενης παραγωγής. Για παράδειγμα, οι λέξεις "NLP", "nlp" και "Nlp" δεν αντιμετωπίζονται ως διαφορετικές λέξεις, αφού και οι τρεις αντιπροσωπεύουν την ίδια έννοια, ανεξάρτητα αν έχουν πεζά ή κεφαλαία γράμματα. Συνηθίζεται, να μετατρέπονται όλοι οι χαρακτήρες ενός κειμένου σε πεζά γράμματα για να υπάρχει ομοιομορφία ανάμεσα στα δεδομένα και τα tokens.

### 5.3.3 Stop Word Removal

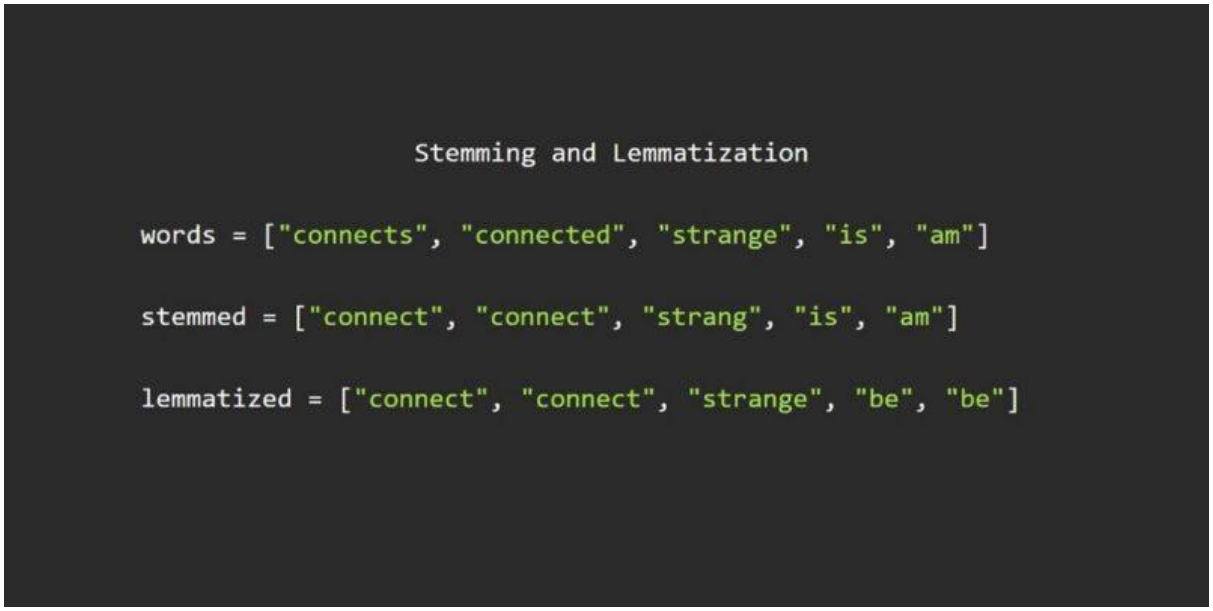
Οι stop words είναι οι πιο συνηθισμένες λέξεις σε οποιαδήποτε φυσική γλώσσα. Αυτές περιλαμβάνουν τα άρθρα, τις αντωνυμίες, τις προθέσεις, τους συνδέσμους και άλλα. Γενικά φιλτράρονται και αποκλείονται πριν από την επεξεργασία του κειμένου καθώς δεν προσθέτουν πολλές πληροφορίες σε αυτό. Παραδείγματα τέτοιων λέξεων είναι τα "και", "ο", "η", "το", "έτσι", "and", "the", "to", "a". Οι λέξεις αυτές μπορούν να αφαιρεθούν με ασφάλεια πραγματοποιώντας μια αναζήτηση με τη βοήθεια μιας προκαθορισμένης λίστας λέξεων-κλειδιών, μειώνοντας με αυτόν τον τρόπο το μέγεθος του συνολικού κειμένου και βελτιώνοντας τον χρόνο επεξεργασίας.
Η κατάργηση των παραπάνω λέξεων δεν είναι πάντα καλή ιδέα και εξαρτάται σε μεγάλο βαθμό από το έργο NLP. Γενικά, σε εργασίες όπως η ταξινόμηση κειμένου, οι stop words δεν χρειάζονται, καθώς οι άλλες λέξεις που υπάρχουν στο κείμενο είναι πιο σημαντικές. Ωστόσο, ενδέχεται οι εργασίες ανάλυσης συναισθημάτων να μην μπορούν να πραγματοποιηθούν σωστά μετά την αφαίρεση ορισμένων λέξεων.

### 5.3.4 Stemming

Το Stemming είναι η διαδικασία μείωσης των παράγωγων των λέξεων στη ρίζα της λέξης τους μέσω της απόρριψης περιττών χαρακτήρων, συνήθως προθεμάτων ή επιθεμάτων. Τα προθέματα είναι οι χαρακτήρες που είναι προσαρτημένοι στην αρχή της λέξης και τα επιθέματα τα αντίστοιχα που είναι προσαρτημένα στο τέλος της λέξης. Οι σχετικές λέξεις συνήθως αντιστοιχίζονται στην ίδια ρίζα και επομένως αντιμετωπίζονται ως συνώνυμα από τους αλγόριθμους. Ωστόσο, η "ρίζα" σε αυτή την περίπτωση μπορεί να μην είναι πανομοιότυπη με τη γλωσσολογική ρίζα των λέξεων. Το Stemming χρησιμοποιεί μια ακατέργαστη διαδικασία που αφαιρεί τα άκρα των λέξεων που χοροπηδούν για να τις μετατρέψει σωστά στις βασικές τους μορφές. Για παράδειγμα, οι λέξεις "compute", "computer" και "computing" θα μετατραπούν στη λέξη "comput" αντί για τη λέξη "compute". Ένα λογισμικό που πραγματοποιεί αυτήν τη διαδικασία ονομάζεται stemmer. Αυτήν τη μέθοδο τη χρησιμοποιούμε σε περιπτώσεις που η ταχύτητα και η απόδοση είναι σημαντικά για το έργο, επειδή είναι απλά στη χρήση και εκτελούν απλές λειτουργίες.

### 5.3.5 Lemmatization

Το Lemmatization είναι η διαδικασία ομαδοποίησης των παράγωγων των λέξεων, έτσι ώστε να μπορούν να αναλυθούν ως ένα μόνο στοιχείο, που προσδιορίζεται από το lemma (λήμμα) της λέξης. Για παράδειγμα, το ρήμα "περπατάω", το οποίο μπορεί να εμφανίζεται ως "περπατάω", "περπατώ", "περπατάει" ή "περπατάω", θα μειωνόταν στην κανονική του μορφή "περπατώ", που ονομάζεται και λήμμα της λέξης. Το lemmatization είναι παρόμοιο με το stemming αλλά έχουν μία κύρια διαφορά. Στο stemming δεν γίνεται έλεγχος αν η λέξη υπάρχει, γίνεται μόνο προσπάθεια να απλουστευτούν οι όροι αναζήτησης, ενώ στο lemmatization γίνεται προσπάθεια αναγνώρισης της λέξης και της έννοιας που έχει σε μία πρόταση και της αποδίδεται η πραγματική ρίζα της (lemma). Δεδομένου ότι το lemmatization απαιτεί περισσότερη γνώση για τη γλωσσική δομή από το Stemming, απαιτεί επίσης περισσότερη υπολογιστική ισχύ



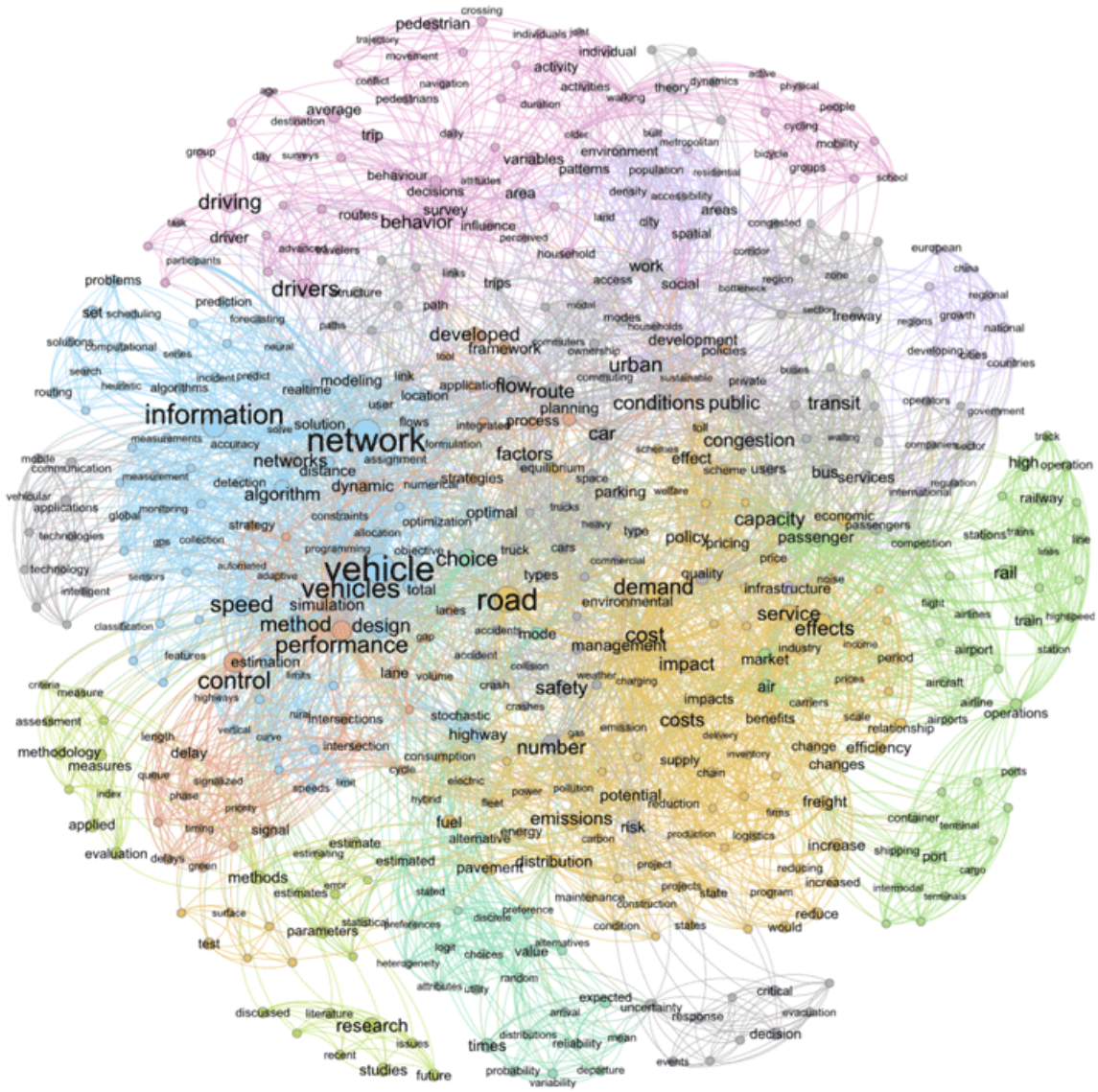

*Εικόνα 14 Διαφορά Stemming και Lemmatization (52)*

### 5.3.6 Topic Modeling

Η μοντελοποίηση ανά θέμα είναι ένας τύπος στατιστικού μοντέλου για την ανακάλυψη των κρυφών δομών ή αφηρημένων "θεμάτων" που εμφανίζονται σε ένα κείμενο. Ουσιαστικά, είναι μια μέθοδος unsupervised learning που σαρώνει ένα σύνολο εγγράφων, ανιχνεύει μοτίβα

λέξεων και φράσεων μέσα σε αυτές και δημιουργεί ομάδες λέξεων και παρόμοιων εκφράσεων που χαρακτηρίζουν καλύτερα αυτό το σύνολο.

Αυτή η τεχνική βασίζεται στην παραδοχή ότι κάθε έγγραφο είναι ένα μείγμα πολλαπλών θεμάτων σε διαφορετικές αναλογίες και ότι η παρουσία κάθε λέξης αποδίδεται σε ένα από τα θέματα του εγγράφου. Η μοντελοποίηση βάσει θεμάτων είναι εξαιρετικά χρήσιμη για την ταξινόμηση κειμένων και τον εντοπισμό τάσεων σε διαδικτυακές δημοσιεύσεις. Ορισμένα δημοφιλή μοντέλα θεμάτων περιλαμβάνουν την Latent Dirichlet Allocation (LDA), την Latent Semantic Analysis (LSA) και την Term Frequency-Inverse Document Frequency (tf-idf).




*Εικόνα 15 Μοντελοποίηση ανά θέμα (53)*

#### *5.3.6.1 Latent Dirichlet Allocation – Λανθάσουσα Κατανομή Dirichlet*

Η Latent Dirichlet Allocation (LDA) είναι η πιο συχνά χρησιμοποιούμενη τεχνική μοντελοποίησης ανά θέμα. Η λέξη «Λανθάνουσα» υποδηλώνει ότι το μοντέλο προσδιορίζει κρυμμένα θέματα μέσα στα έγγραφα. Το «Dirichlet» δηλώνει την υπόθεση ότι η κατανομή των θεμάτων στο έγγραφο και η κατανομή των λέξεων στα θέματα είναι και οι δύο Dirichlet.
Το LDA δέχεται έγγραφα ως θέματα εισόδου και εξόδου. Υποθέτει ότι κάθε έγγραφο παράγεται από ένα μείγμα θεμάτων και κάθε θέμα από ένα μείγμα λέξεων. Επίσης, υποθέτει ότι κάθε κομμάτι κειμένου που εισάγεται στον αλγόριθμο θα περιέχει λέξεις που σχετίζονται μεταξύ τους. Συγκεκριμένα, προκειμένου ο αλγόριθμος να βρει ομάδες σχετικών λέξεων, καθορίζεται πρώτα ο αριθμός των θεμάτων που επιθυμούνται να αποκαλυφθούν. Το LDA θα αναθέσει όλα τα έγγραφα σε θέματα με τέτοιον τρόπο που οι λέξεις κάθε εγγράφου να περιγράφονται από αυτά τα θέματα. Στη συνέχεια, εκχωρεί επαναληπτικά κάθε λέξη σε ένα θέμα λαμβάνοντας υπόψη την πιθανότητα να ανήκει σε αυτό και την πιθανότητα ότι το έγγραφο έχει δημιουργηθεί από αυτό. Αυτές οι πιθανότητες υπολογίζονται πολλές φορές, έως ότου ο αλγόριθμος συγκλίνει και κάθε έγγραφο εκχωρηθεί σε μία ομάδα θεμάτων.

## 5.4 Language Modeling

Η Γλωσσική Μοντελοποίηση (Language Modeling) είναι η χρήση διαφόρων τεχνικών για την πρόβλεψη της πιθανότητας εμφάνισης μιας ακολουθίας λέξεων σε μια πρόταση. Γενικά, δεδομένης μίας ακολουθίας $T$ λέξεων, $w_1, \dots, w_t$, ένα γλωσσικό μοντέλο εκχωρεί στην ακολουθία την πιθανότητα:

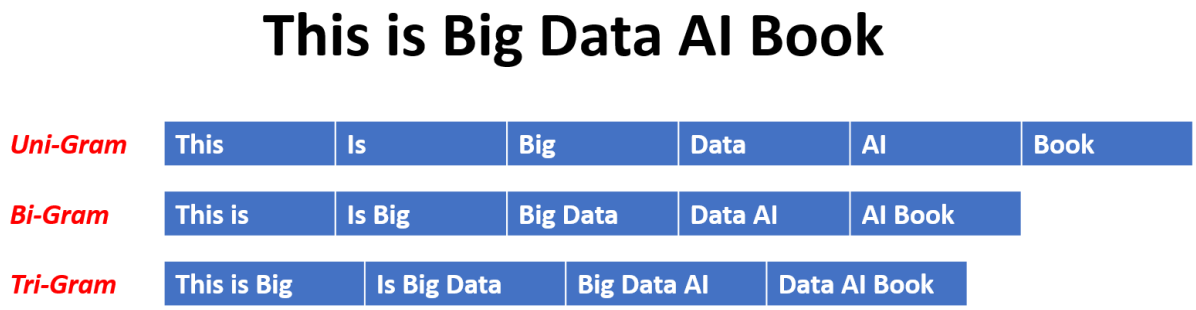
$$P(w_1, \dots, w_t) = \prod_{t=1}^{T} P(w_t | w_1, \dots, w_t)$$

Ο στόχος της LM είναι να παρέχει επαρκείς πληροφορίες, έτσι ώστε οι πιθανές ακολουθίες λέξεων να είναι πιθανότερες. Τα μοντέλα γλώσσας είναι χρήσιμα σε πολλές NLP εφαρμογές, ειδικά σε αυτές που έχουν κείμενο ως έξοδο.
Υπάρχουν κυρίως δύο κατηγορίες γλωσσικών μοντέλων:

- count based. Χρησιμοποιούν παραδοσιακές στατιστικές τεχνικές όπως n-gramm, Hidden Markov Models (HMM) και ορισμένους γλωσσικούς κανόνες για την ανάπτυξη μοντέλων που μπορούν να προβλέψουν την επόμενη λέξη ή χαρακτήρα σε ένα έγγραφο , δεδομένης μιας ακολουθίας λέξεων που προηγούνται.
- continuous-space based. Χρησιμοποιούν διαφορετικά νευρωνικά δίκτυα και συχνά θεωρούνται προηγμένα για την χρήση σε NLP εργασίες.

### 5.4.1 Μοντέλο N-Gram

Τα μοντέλα N-gram είναι μία από τις απλούστερες προσεγγίσεις των Γλωσσικών Μοντέλων. Δημιουργούν μια κατανομή πιθανοτήτων για μια ακολουθία $n$, όπου ο αριθμός $n$ καθορίζει το μέγεθος του "gram" (μία σειρά λέξεων). Για παράδειγμα, αν $n = 4$, ένα gram μπορεί να είναι: "μπορείς να με βοηθήσεις". Υπάρχουν διαφορετικοί τύποι όπως τα unigrams ($n = 1$), τα bigrams ($n = 2$) και τα trigrams ($n = 3$).




*Εικόνα 16 Παραδείγματα μοντέλων με n = 1, 2, 3 (54)*

Προκειμένου να απλοποιηθεί το πρόβλημα της εκτίμησης του γλωσσικού μοντέλου από τα δεδομένα, το μοντέλο n-gram υποθέτει ότι κάθε λέξη εξαρτάται μόνο από τις τελευταίες $n - 1$ λέξεις αντί για τις προηγούμενες λέξεις $t - 1$. Έτσι, η πιθανότητα παρατήρησης της $t$-οστής λέξης μπορεί να προσεγγιστεί με την πιθανότητα παρατήρησής της στο συντομευμένο ιστορικό περιβάλλον των προηγούμενων $n - 1$ λέξεων. Επομένως, η πιθανότητα υπολογίζεται ως εξής:

$$P(w_1, \dots, w_t) \approx \prod_{t=1}^{T} P\left(w_t | w_{t-(n-1)}, \dots, w_{t-1}\right)$$

όπου

$$P\left(w_t | w_{t-(n-1)}, \dots, w_{t-1}\right) = \frac{count\left(w_{t-(n-1)}, \dots, w_{t-1}, w_t\right)}{count\left(w_{t-(n-1)}, \dots, w_{t-1}\right)}$$

Ωστόσο, για μεγάλο $n$ υπάρχει πρόβλημα αραιότητας των δεδομένων και το μοντέλο της γλώσσας δεν είναι ακριβές. Η αραιότητα των δεδομένων (data sparsity) είναι ένας όρος που περιγράφει το φαινόμενο πολλών πιθανών ακολουθιών λέξεων που έχουν πολύ λίγες εμφανίσεις σε ένα σώμα κειμένου (corpus), που σημαίνει δηλαδή ότι δεν παρατηρούνται αρκετά κατά τη διάρκεια της εκπαίδευσης.

Υπάρχουν διάφοροι τρόποι αξιολόγησης ενός γλωσσικού μοντέλου. Το perplexity (PP) είναι το αντίστροφο της πιθανότητας που εκχωρεί το μοντέλο στο κανονικοποιημένο από τον αριθμό των λέξεων κείμενο (corpus). Η βαθμολογία PP ενός συνόλου δοκιμών $W = w_1, w_2 \dots, w_N$ σε ένα μοντέλο n-gram υπολογίζεται ως εξής:

$$PP(W) = \left(P(w_1, w_2 \dots, w_N)\right)^{-1/N} = \left(\prod_{t=1}^{N} P\left(w_t | w_{t-(n-1)}, \dots, w_{t-1}\right)\right)^{-1/N}$$

Όσο καλύτερο είναι ένα μοντέλο τόσο χαμηλότερο είναι το PP, καθώς υποδεικνύει ότι η κατανομή πιθανότητας ή το μοντέλο πιθανότητας είναι αποτελεσμάτικά στην πρόβλεψη του δείγματος. Το PP μπορεί επίσης να ερμηνευθεί ως ο σταθμισμένος μέσος συντελεστής διακλάδωσης μιας γλώσσας στην πρόβλεψη της επόμενης λέξης. Σημειώστε ότι ο παράγοντας διακλάδωσης μιας γλώσσας (branching factor) είναι ο αριθμός των πιθανών λέξεων που μπορούν να ακολουθήσουν οποιαδήποτε λέξη.

### 5.4.2 Neural Language Model

Τα μοντέλα νευρικής γλώσσας χρησιμοποιούν την τεχνική του Word Embeddings (ενσωμάτωση λέξεων), που περιγράφεται παρακάτω, για να κάνουν τις προβλέψεις τους. Καθώς τα μοντέλα εκπαιδεύονται σε μεγαλύτερα και μεγαλύτερα σώματα, το μέγεθος του λεξιλογίου (ο αριθμός των μοναδικών λέξεων) και κατά συνέπεια ο αριθμός των πιθανών

ακολουθιών αυξάνεται, έχοντας ως αποτέλεσμα το πρόβλημα της αραιότητας δεδομένων. Πιο συγκεκριμένα, τα μοντέλα Neural Language αποφεύγουν αυτό το πρόβλημα χρησιμοποιώντας μη γραμμικά νευρωνικά δίκτυα, όπως τα FFNN ή τα RNN, που έχουν τη δυνατότητα να αναπαριστούν λέξεις με κατανεμημένο τρόπο, ως μη γραμμικούς συνδυασμούς βαρών σε ένα δίκτυο.

Όπως και στα μοντέλα n-gram, τα περισσότερα πιθανολογικά γλωσσικά μοντέλα προσεγγίζουν την πιθανότητα μιας ακολουθίας λέξεων $P(w_t|w_{t-(n-1)}, \dots, w_{t-1})$ χρησιμοποιώντας ένα σταθερό πλαίσιο μεγέθους $n-1$. Το νευρωνικό πιθανολογικό γλωσσικό μοντέλο που εισήχθη από τους Bengio et al. (55), χρησιμοποιεί ένα FFNN τριών στρωμάτων για να υπολογίσει αυτήν την πιθανολογική πρόβλεψη. Κατ' αρχάς, κάθε λέξη στο αντιστοιχίζεται σε ένα διάνυσμα d-διαστάσεων $C_{w_{t-1}}$, το οποίο είναι η στήλη $w_{t-1}$ του πίνακα παραμέτρων $C$. Τα διανύσματα χαρακτηριστικών (στήλες του $C$) μαθαίνονται ταυτόχρονα με τις παραμέτρους του NN. Η συνένωση αυτών των $n-1$ διανυσμάτων, που συμβολίζονται με το διάνυσμα $x$, είναι η είσοδος στο FFNN:

$$x = (C_{w_{t-(n-1)},1}, \dots, C_{w_{t-(n-1)},d}, C_{w_{t-(n-2)},1}, \dots, C_{w_{t-(n-2)},d}, C_{w_{t-1},1}, \dots, C_{w_{t-1},d}$$

Το διάνυσμα $x$ τροφοδοτείται σε ένα κρυφό στρώμα το οποίο στη συνέχεια τροφοδοτείται σε ένα στρώμα softmax για να εκτιμήσει την πιθανότητα της επόμενης λέξης $k$:

$$P(w_t = k|w_{t-(n-1)}, \dots, w_{t-1}) = \frac{e^{a_k}}{\sum_{l=1}^{N} e^{a_l}}$$

όπου

$$a_k = b_k + \sum_{i=1}^{h} W_{ki} \tanh\left(c_i + \sum_{j=1}^{(n-1)d} V_{ij} x_j\right)$$

όπου τα διανύσματα $b$, $c$ και οι πίνακες $W$, $V$ είναι παράμετροι του μοντέλου και $h$ είναι ο αριθμός των κρυφών μονάδων. Το NN εκπαιδεύεται χρησιμοποιώντας έναν gradient-based αλγόριθμο βελτιστοποίησης για να μεγιστοποιήσει την πιθανότητα καταγραφής του συνόλου εκπαίδευσης:

$$L(\theta) = \sum_{t} \log P\left(w_t \middle| w_{t-(n-1)}, \dots, w_{t-1}\right)$$

όπου το $\theta$ υποδηλώνει τη συνένωση όλων των παραμέτρων.

## 5.5 Word Embeddings

Στο NLP, ένα word embedding (ενσωμάτωση λέξης) είναι μια αριθμητική αναπαράσταση λέξεων για ανάλυση κειμένου που μπορεί να ληφθεί χρησιμοποιώντας ένα σύνολο τεχνικών εκμάθησης χαρακτηριστικών και μοντελοποίησης γλώσσας. Τυπικά, έχει τη μορφή ενός διανύσματος πραγματικής αξίας σε έναν προκαθορισμένο διανυσματικό χώρο που κωδικοποιεί τη σημασία της λέξης έτσι ώστε οι λέξεις που έχουν το ίδιο νόημα να έχουν παρόμοια αναπαράσταση. Οι ενσωματώσεις λέξεων προσφέρουν έναν αποτελεσματικό τρόπο αναπαράστασης συμβολοσειρών και απλού κειμένου ως διανυσμάτα πραγματικών αριθμών, τα οποία μπορούν να επεξεργαστούν από τους περισσότερους ML αλγόριθμους.
Οι διαφορετικές τεχνικές ενσωμάτωσης λέξεων μπορούν να ταξινομηθούν ευρέως σε δύο κατηγορίες:

- τεχνικές ενσωμάτωσης με βάση τη συχνότητα, όπως το διάνυσμα καταμέτρησης, η διανυσματικότητα TF-IDF και ο πίνακας συμπτωμάτων, διανύουν το κείμενο ανάλογα με τη συχνότητα εμφάνισης των λέξεων στο σώμα
- τεχνικές ενσωμάτωσης με βάση την πρόβλεψη, περιλαμβάνουν μεθόδους όπως το Word2Vec (56) και το GloVe (57).

Τέλος, τόσο το PCA όσο και το t-SNE μπορούν να χρησιμοποιηθούν προκειμένου να μειωθεί η διάσταση των διανυσμάτων και να απεικονιστούν ενσωματώσεις και συστάδες λέξεων.

### 5.5.1 Term Frequency-Inverse Document Frequency Vector

Η Term Frequency – Inverse Document Frequency είναι ένα term-weighting σχήμα που αντικατοπτρίζει πόσο σχετική είναι μια λέξη με ένα έγγραφο σε μια συλλογή εγγράφων. Η τιμή της tf -idf αυξάνεται αναλογικά με τον αριθμό εμφανίσεων μιας λέξης στο έγγραφο και

αντισταθμίζεται από τον αριθμό των εγγράφων στο σώμα που περιέχουν τη λέξη. Με αυτόν τον τρόπο, οι λέξεις που είναι κοινές σε κάθε έγγραφο, όπως «το», «εάν» και «είναι», κατατάσσονται χαμηλότερα από τις λέξεις που εμφανίζονται μόνο σε ορισμένα έγγραφα, αφού οι πρώτες δεν περιέχουν σημαντικές πληροφορίες και οι δεύτερες μπορεί να είναι σχετικές. Η βαθμολογία tf -idf για τη λέξη $t$ στο έγγραφο $d$ από το σύνολο εγγράφων $D$ είναι το προϊόν των ακόλουθων δύο μετρήσεων:

- Συχνότητας Όρου($tf$): Αναφέρεται στη συχνότητα μιας λέξης, δηλαδή πόσο συχνά μια δεδομένη λέξη εμφανίζεται μέσα σε κάθε έγγραφο. Υπάρχουν διάφοροι τρόποι υπολογισμού αυτής της μέτρησης. Το πιο απλό από αυτά είναι η χρήση της ακατέργαστης μέτρησης $f_{t,d}$ από τις περιπτώσεις που ο όρος $t$ εμφανίζεται στο έγγραφο $d$:

  $$tf(t,d) = f_{t,d}$$

  Στη συνέχεια, υπάρχουν διάφοροι τρόποι για να το προσαρμόσουμε. Για παράδειγμα, το $tf$ θα μπορούσε να προσαρμοστεί με το μήκος του εγγράφου:

  $$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

  ή με την επαυξημένη συχνότητα, προκειμένου να αποφευχθεί όποια προκατάληψη προς μεγαλύτερα έγγραφα:

  $$tf(t,d) = 0{,}5 + 0{,}5\frac{f_{t,d}}{max_{\{t' \in d\}} f_{t',d}}$$

  Το tf θα μπορούσε επίσης να οριστεί ως η λογαριθμικά κλιμακούμενη συχνότητα:

  $$tf(t,d) = \log(1 + f_{t,d})$$

- Inverse Document Frequency (idf): Το idf μιας λέξης σε ένα σύνολο εγγράφων είναι το πλήθος των εγγράφων που εμφανίζεται η λέξη. Το idf μετρά πόσο σημαντική είναι η λέξη σε ολόκληρο το σώμα των κειμένων. Αυτή η μέτρηση μπορεί να υπολογιστεί

διαιρώντας τον συνολικό αριθμό εγγράφων στο σώμα, με τον αριθμό των εγγράφων που περιέχουν τη συγκεκριμένη λέξη και στη συνέχεια υπολογίζοντας το λογάριθμο του αποτελέσματος. Εάν η λέξη είναι κοινή σε πολλά έγγραφα, αυτός ο αριθμός θα πλησιάσει το 0. Διαφορετικά, αν η λέξη είναι σπάνια, θα πλησιάσει το 1. Έτσι έχουμε:

$$idf(t, D) = \log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right)$$

Συνεπώς, για το συνολικό tf-idf έχουμε τον παρακάτω υπολογισμό:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

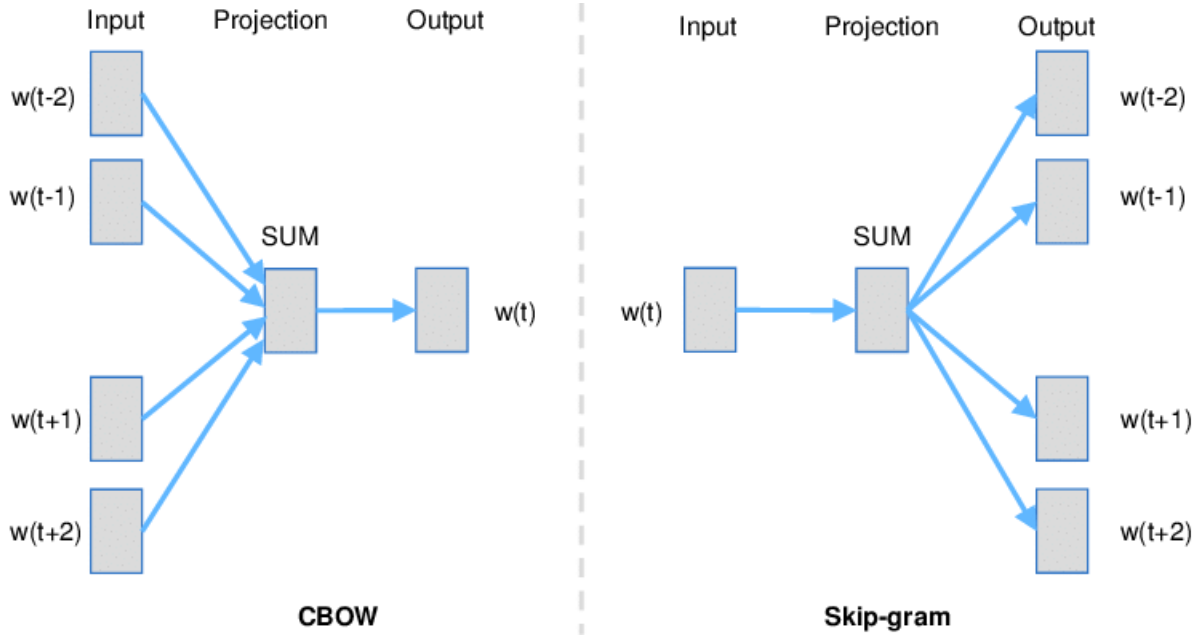
Ένα υψηλό tf στο δεδομένο έγγραφο και μια χαμηλή συχνότητα εγγράφων της λέξης σε ολόκληρο το σώμα προκαλεί την επίτευξη ενός μεγάλου βάρους στο tf -idf. Έτσι, τα βάρη τείνουν να φιλτράρουν τους κοινούς όρους. Συμπερασματικά, το tf -idf παρέχει έναν τρόπο χαρτογράφησης εγγράφων σε διανύσματα λέξεων συνδέοντας κάθε λέξη σε ένα έγγραφο με έναν αριθμό που αντιπροσωπεύει πόσο σχετική είναι κάθε λέξη σε αυτό το έγγραφο. Τα έγγραφα με παρόμοιες, σχετικές λέξεις θα έχουν παρόμοια διανύσματα.

### 5.5.2 Word2Vec

Το Word2Vec (56) είναι μια οικογένεια αρχιτεκτονικών μοντέλων και μοντέλων βελτιστοποιήσης που μπορούν να χρησιμοποιηθούν για την παραγωγή word-embeddings. Αυτά τα μοντέλα είναι νευρωνικά δίκτυα δύο επιπέδων που μετατρέπουν το κείμενο σε μια αριθμητική μορφή που τα νευρωνικά δίκτυα περισσότερων επιπέδων μπορούν να κατανοήσουν. Σε αντίθεση με τεχνικές όπως οι συνήθεις Bag-of-Words (BOW) και το tf – idf, τα μοντέλα του Word2Vec μπορούν να αποτυπώσουν το νόημα ή τη σχέση των λέξεων μόλις εκπαιδευτούν. Το Word2Vec λαμβάνει ως είσοδο ένα σώμα κειμένου και εξάγει ένα διανυσματικό χώρο, συνήθως με εκατοντάδες διαστάσεις. Σε αυτό το διάστημα, σε κάθε μοναδική λέξη στο σώμα αποδίδεται ένα αντίστοιχο διάνυσμα. Τα διανύσματα επιλέγονται με τέτοιο τρόπο ώστε μια απλή μαθηματική συνάρτηση, να δείχνει το επίπεδο σημασιολογικής ομοιότητας μεταξύ των λέξεων που αντιπροσωπεύονται από αυτά τα διανύσματα. Ο στόχος

είναι να υπάρχουν λέξεις, οι οποίες μοιράζονται κοινό περιεχόμενο στο κείμενο, να βρίσκονται το ένα κοντά στο άλλο στο διανυσματικό χώρο.

Το Word2Vec έχει δύο μορφές, το μοντέλο Skip-Gram και το μοντέλο Continuous Bag-of-Words (CBOW). Με τον πρώτο τρόπο, μια δεδομένη λέξη χρησιμοποιείται για την πρόβλεψη ενός πλαισίου ενώ στο δεύτερο, το πλαίσιο χρησιμοποιείται για την πρόβλεψη μιας λέξης.




*Εικόνα 17 CBOW και Skip-Gram (58)*

### *5.5.2.1 Skip-Gram Model*

Το μοντέλο Skip-Gram προβλέπει το πλαίσιο από τη λέξη. Το επίπεδο εισόδου του μοντέλου έχει μέγεθος (1 × V), όπου V είναι ο αριθμός των λέξεων στο λεξιλόγιο. Η είσοδος στο δίκτυο είναι η μοναδική αναπαράσταση της λέξης-στόχου. Αυτό το διάνυσμα εισόδου μετασχηματίζεται από τη μήτρα βάρους W, μεγέθους (V × E), και περνά μέσα από ένα κρυφό στρώμα μεγέθους (1 × E), όπου το E είναι η επιθυμητή διάσταση ενσωμάτωσης. Όσο υψηλότερη είναι η τιμή αυτής της υπερ-παραμέτρου, τόσο περισσότερες πληροφορίες θα συλλάβουν οι ενσωματώσεις, αλλά τόσο πιο δύσκολο θα είναι να γίνει. Τέλος, ο πίνακας βάρους W′ μεγέθους (E × V) μετατρέπει το κρυφό στρώμα στο επίπεδο εξόδου μεγέθους (1 × V), αφού οι προβλέψεις θα είναι λέξεις κωδικοποιημένες με την one-hot τεχνική.

Το μοντέλο skip-gram θα μάθει με την εκπαίδευση να προβλέπει το πλαίσιο που δίνεται σε μια λέξη-στόχο. Μόλις ολοκληρωθεί η εκπαίδευση σε ολόκληρο το λεξιλόγιο, πρόκειται να

παραχθεί ένας πίνακας βάρους $W_{V\times E}$ που συνδέει την είσοδο με το κρυφό στρώμα και μπορούμε να πάρουμε τα embeddings. Αυτή η αναπαράσταση, ιδανικά, περικλείει τη σημασιολογία και παρόμοιες λέξεις είναι κοντά η μία στην άλλη στον διανυσματικό χώρο. Το Skip-gram αποδίδει καλύτερα με μικρή ποσότητα δεδομένων και διαπιστώνεται ότι αντιπροσωπεύει καλά τις σπάνιες λέξεις.

#### *5.5.2.2 Continuous Bag-of-Words Model*

Το μοντέλο Continuous Bag-of-Words (CBOW) μοιάζει με την αντίθετη διαδικασία του μοντέλου skip-gram, αφού προβλέπει την τρέχουσα λέξη με βάση τις γύρω λέξεις. Το πλαίσιο αποτελείται από ένα παράθυρο λέξεων γύρω από την τρέχουσα (μεσαία) λέξη, δηλαδή μερικές λέξεις πριν και μετά την κεντρική λέξη. Το στρώμα εισόδου μεγέθους (1 × V) αποτελείται από τις λέξεις του πλαισίου με κωδικοποίηση one-hot και για κάθε τέτοια λέξη ο πίνακας βάρους $W_{V\times E}$ καταλήγει στο κρυφό επίπεδο. Στη συνέχεια υπολογίζονται κατά μέσο όρο σε ένα κρυφό στρώμα, το οποίο μεταφέρεται στην έξοδο.
Το μοντέλο CBOW μαθαίνει ουσιαστικά τις λέξεις προς ενσωμάτωση, εκπαιδεύοντας ένα μοντέλο να προβλέπει μια λέξη δεδομένου του πλαισίου που βρίσκεται. Και πάλι, ο πίνακας βάρους $W_{V\times E}$ χρησιμοποιείται για τη δημιουργία της λέξης από τις one-hot κωδικοποιήσεις μόλις ολοκληρωθεί η εκπαίδευση. Το CBOW μπορεί να αντιπροσωπεύει καλύτερα τις πιο συχνές λέξεις και είναι γρηγορότερο από το skip-gram. Σημειώστε ότι αυτή η αρχιτεκτονική ονομάζεται έτσι επειδή η σειρά των λέξεων περιβάλλοντος δεν επηρεάζει την πρόβλεψη.

### 5.5.3 Global Vectors for Word Representation

Τα Global Vectors (GloVe) (57) είναι ένας αλγόριθμος μάθησης χωρίς επίβλεψη για τη λήψη διανυσματικών αναπαραστάσεων για λέξεις. Όπως και οι περισσότεροι αλγόριθμοι χωρίς επίβλεψη, βασίζεται σε μέτρα όπως η συχνότητα μιας λέξης και το πλήθος της συνύπαρξης της με άλλες λέξεις. Το μοντέλο GloVe εκπαιδεύεται με μετρήσεις συνύπαρξης λέξεων, οι οποίες υποδεικνύουν πόσο συχνά κάθε ζεύγος λέξεων χρησιμοποιείται στο δεδομένο σώμα και ελαχιστοποιεί το σφάλμα ελάχιστων τετραγώνων. Ως αποτέλεσμα, παράγει ένα διανυσματικό χώρο λέξεων, όπου η απόσταση μεταξύ των λέξεων σχετίζεται με τη σημασιολογική ομοιότητα. Σημειώστε ότι το GloVe συνδυάζει τα χαρακτηριστικά δύο βασικών οικογενειών

εκμάθησης λέξης-διανύσματος, συγκεκριμένα την καθολική παραγοντοποίηση μήτρας (π.χ. LSA) και μεθόδους όπως η skip-gram.

Τυπικά, το GloVe κατασκευάζει έναν πίνακα $X$ χρησιμοποιώντας στατιστικά στοιχεία σε ολόκληρο το σώμα. Κάθε κελί $X_{ij}$ αντιπροσωπεύει πόσο συχνά εμφανίζεται η λέξη $i$ στο πλαίσιο της λέξης $j$. Συνήθως, για κάθε λέξη στο σώμα, οι όροι περιβάλλοντος αναζητούνται σε κάποια περιοχή που ορίζεται από ένα πλήθος λέξεων πριν και από ένα πλήθος λέξεων μετά τον όρο. Επίσης, σε πιο μακρινές λέξεις δίνεται μικρότερο βάρος. Οι περιορισμοί ορίζονται για κάθε ζεύγος λέξεων:

$$w_i^t w_j + b_i + b_j = \log(X_{ij})$$

όπου $w_i$, $b_i$ είναι το διάνυσμα και η κλιμακωτή μεροληψία για την κύρια λέξη και $w_j$, $b_j$ είναι η διανυσματική και κλιμακωτή μεροληψία για τις λέξεις πλαισίου. Το GloVe χρησιμοποιεί έναν σταθμισμένο στόχο ελάχιστων τετραγώνων που ελαχιστοποιεί τη διαφορά μεταξύ του τελικού προϊόντος των διανυσμάτων δύο λέξεων και του λογάριθμου του πλήθους των συνυπάρξεών τους:

$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^t w_j + b_i + b_j - \log(X_{ij})\right)^2$$

όπου $f$ είναι μια συνάρτηση στάθμισης που αποδίδει χαμηλότερα βάρη σε σπάνιες και συχνές συνυπάρξεις. Μια κατηγορία συναρτήσεων στάθμισης που διαπιστώνεται ότι λειτουργεί καλά μπορεί να είναι:

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{X_{max}}\right)^a & , if\ X_{ij} < X_{max} \\ 1 & , if\ X_{ij} \geq X_{max} \end{cases}$$

## 5.6 Μεταφορική Μάθηση (Transfer Learning, TL) στην Επεξεργασία Φυσικής Γλώσσας

Οι περισσότερες από τις εργασίες στο NLP είναι ακολουθίες εργασιών μοντελοποίησης. Σε αντίθεση με τα παραδοσιακά NN, τα επαναλαμβανόμενα νευρωνικά δίκτυα όπως τα RNN και

τα LSTM μπορούν να μοντελοποιήσουν διαδοχικές πληροφορίες που υπάρχουν στο κείμενο, επομένως χρησιμοποιούνται συνήθως για τέτοιου είδους εργασίες. Ωστόσο, αυτά τα δίκτυα έχουν κάποια σημαντικά ζητήματα που καθιστούν προφανή την ανάγκη για μεταφορά μάθησης στο NLP. Ένα από αυτά είναι ότι, στην περίπτωση μιας ακολουθίας κειμένου, τα RNN ή τα LSTM δεν μπορούν να παραλληλιστούν, καθώς λαμβάνουν κάθε φορά ένα διακριτικό ως είσοδο. Ως εκ τούτου, η εκπαίδευση ενός τέτοιου μοντέλου σε ένα μεγάλο σύνολο δεδομένων θα πάρει πολύ χρόνο.

Η μεταφορική μάθηση, είναι μια τεχνική όπου ένα μοντέλο βαθιάς εκμάθησης εκπαιδευμένο σε ένα μεγάλο σύνολο δεδομένων, που ονομάζεται προ-εκπαιδευμένο (pre-trained) μοντέλο, χρησιμοποιείται ως αφετηρία για την επίλυση ενός προβλήματος σε ένα άλλο, συνήθως μικρότερο, σύνολο δεδομένων. Η μορφή TL που οδήγησε στις μεγαλύτερες βελτιώσεις μέχρι τώρα στο NLP είναι η διαδοχική TL. Περιλαμβάνει τη μεταφορά γνώσης με μια ακολουθία βημάτων, όπου η εργασία προέλευσης και στόχου δεν είναι απαραίτητα παρόμοια. Η γενική πρακτική είναι h προ-εκπαιδεύση ενός μεγάλου κείμενου χρησιμοποιώντας μια μέθοδο όπως το ELMo (59) ή το BERT (60), και στη συνέχεια να προσαρμοστεί σε ένα supervised task.

### 5.6.1 Ενσωματώσεις από Γλωσσικά Μοντέλα

Οι Ενσωματώσεις από Γλωσσικά Μοντέλα (ELMo) (59) είναι ένας τύπος βαθιάς αναπαράστασης λέξεων, που μοντελοποιεί τόσο τα σύνθετα χαρακτηριστικά της χρήσης λέξεων, όπως η σύνταξη, όσο και το πώς αυτές οι χρήσεις ποικίλλουν σε γλωσσικά πλαίσια. Σε αντίθεση με τις προηγούμενες διανυσματικές προσεγγίσεις όπως το Word2Vec και το GloVe, οι ενσωματώσεις ELMo είναι ευαίσθητες στο περιβάλλον. Με άλλα λόγια, είναι σε θέση να συλλάβουν το πλαίσιο της λέξης που χρησιμοποιείται στην πρόταση και μπορούν να δημιουργήσουν διαφορετικές αναπαραστάσεις για την ίδια λέξη που χρησιμοποιείται με άλληέννοια σε διαφορετικές προτάσεις. Το ELMo εκπαιδεύεται στην πρόβλεψη της επόμενης λέξης σε μια ακολουθία λέξεων, η οποία είναι μια εργασία που ονομάζεται Μοντελοποίηση Γλώσσας. Αυτό το μοντέλο εκπαιδεύεται σε ένα τεράστιο σύνολο δεδομένων και στη συνέχεια μπορεί να χρησιμοποιηθεί ως συστατικό σε άλλες αρχιτεκτονικές για την εκτέλεση συγκεκριμένων γλωσσικών εργασιών.

Το ELMo χρησιμοποιεί ένα συγκεκριμένο τύπο γλωσσικού μοντέλου που ονομάζεται Bi-Direction Language Model (biLM), το οποίο είναι ένας συνδυασμός ενός forward κι ενός backward γλωσσικού μοντέλου. Τα διανύσματα λέξεων είναι λειτουργίες των εσωτερικών

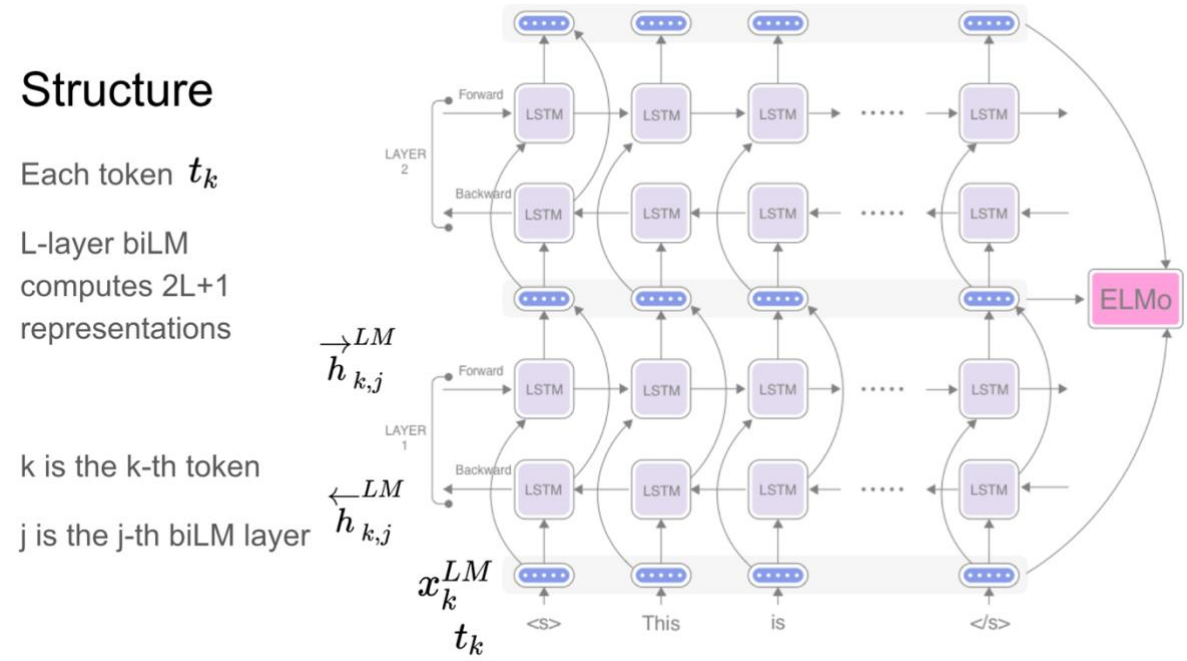
καταστάσεων αυτού του biLM, το οποίο είναι προ-εκπαιδευμένο σε ένα μεγάλο σώμα κειμένου. Πιο συγκεκριμένα, σύμφωνα με τους Peters et al. (59), το ELMo είναι ένας συνδυασμός του ενδιάμεσου επιπέδου στο biLM. Δεδομένου ότι για κάθε $t_k$, το $x_k^{LM}$ είναι μια ανεξάρτητης από το περιβάλλον αναπαράσταση token, ένα αμφίδρομο LSTM L επιπέδων υπολογίζει ένα σύνολο 2L + 1 αναπαραστάσεων:

$$R_k = \{x_k^{LM}, \vec{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM} | j = 1, \dots, L\} = \{h_{k,j}^{LM} | j = 0, \dots, L\}$$

όπου το $h_{k,0}^{LM}$ είναι το επίπεδο token και $h_{k,j}^{LM} = [\vec{h}_{k,j}^{LM}; \overleftarrow{h}_{k,j}^{LM}]$ για κάθε αμφίδρομο LSTM στρώμα. Οι ενσωματώσεις ELMo υπολογίζονται από μια στάθμιση όλων των επιπέδων biLM για κάθε εργασία:

$$ELMo_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} h_{k,j}^{LM}$$

όπου το $s_j^{task}$ είναι τα κανονικοποιημένα βάρη (softmax) και η παράμετρος $\gamma^{task}$ επιτρέπει στο μοντέλο να κλιμακώσει το διάνυσμα ELMo.

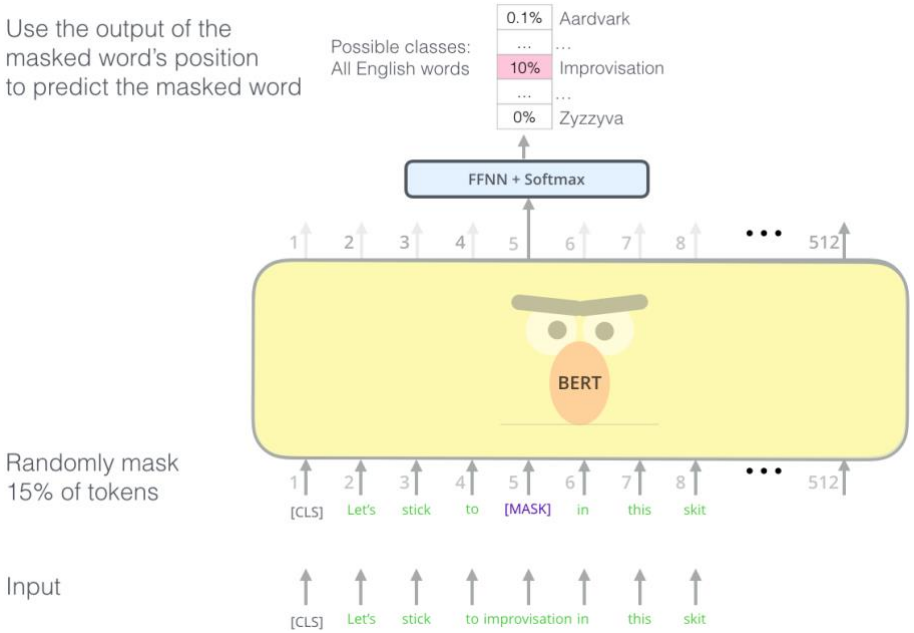



*Εικόνα 18 ELMo (61)*

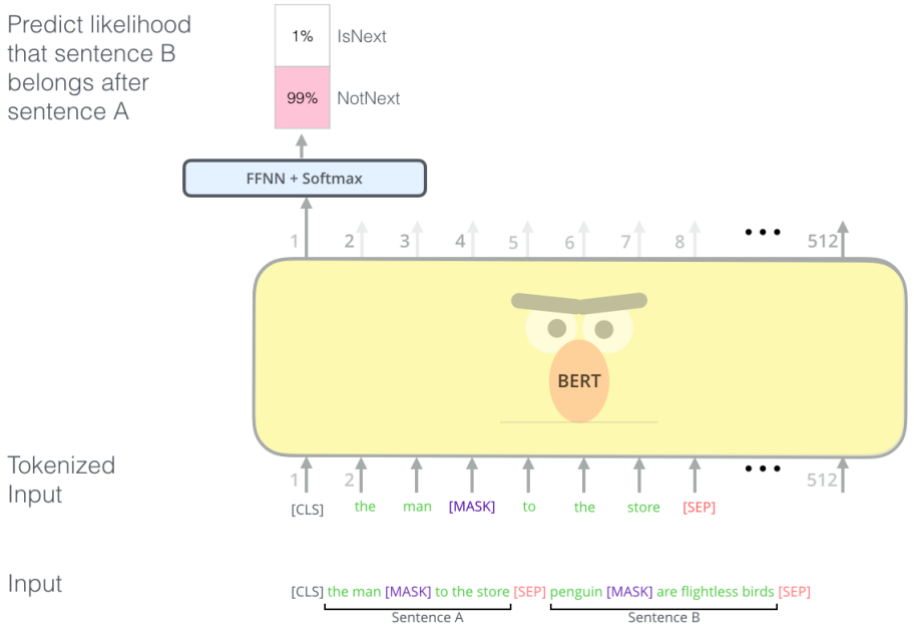
### 5.6.2 Bidirectional Encoder Representations from Transformers

Οι Αμφίδρομες Αναπαραστάσεις Κωδικοποιητή από Μετασχηματιστές (BERT) (60) είναι μια τεχνική μηχανικής μάθησης που βασίζεται στον μετασχηματιστή για την προ-εκπαίδευση NLP. Βελτιώνεται με τους τυπικούς μετασχηματιστές, χρησιμοποιώντας ένα μοντέλο γλωσσικής μάσκας (MLM) που επιτρέπει εκπαίδευση διπλής κατεύθυνσης σε μοντέλα που ήταν προηγουμένως αδύνατο. Πιο συγκεκριμένα, το BERT είναι προ-εκπαιδευμένο με δύο στόχους:

- Μοντελοποίηση Μάσκας Γλώσσας (MLM). Λαμβάνοντας μια πρόταση, το MLM (Εικόνα 19) καλύπτει τυχαία το 15% των λέξεων στην είσοδο και στη συνέχεια προσπαθεί να τις προβλέψει. Η απόκρυψη σημαίνει ότι το μοντέλο κοιτάζει και προς τις δύο κατευθύνσεις και χρησιμοποιεί το πλήρες πλαίσιο της πρότασης, τόσο στο αριστερό όσο και στο δεξί περιβάλλον, για να προβλέψει τη λέξη που καλύπτεται. Σε αντίθεση με τα προηγούμενα μοντέλα γλώσσας από αριστερά προς τα δεξιά, λαμβάνει υπόψη τόσο τα προηγούμενα όσο και τα επόμενα tokens ταυτόχρονα, γεγονός που επιτρέπει την προ-εκπαίδευση ενός βαθύ αμφίδρομου μετασχηματιστή.
- Πρόβλεψη Επόμενης Πρότασης (NSP). Το NSP προ-εκπαιδεύει από κοινού αναπαραστάσεις ζεύγους κειμένου. Πιο συγκεκριμένα, το μοντέλο συνδυάζει δύο καλυμμένες προτάσεις ως εισόδους κατά την προ-εκπαίδευση και στη συνέχεια πρέπει να προβλέψει εάν οι δύο προτάσεις ακολουθούσαν ή όχι η μία την άλλη.




*Εικόνα 19 MLM (62)*




*Εικόνα 20 NSP (62)*

Καλό θα ήταν να σημειώσουμε εδώ, ότι τα ειδικά διακριτικά [SEP] και [CLS] που εμφανίζονται στα παραπάνω σχήματα είναι ειδικά διακριτικά που απαιτούνται από το BERT. Το ειδικό διακριτικό [SEP] χρησιμοποιείται για τη διαφοροποίηση μεταξύ των διαφορετικών προτάσεων εισαγωγής. Το διακριτικό [CLS] εμφανίζεται πάντα στην αρχή του κειμένου και είναι συγκεκριμένο για εργασίες ταξινόμησης.

Το BERT αποτελείται από δύο βήματα: προ-εκπαίδευση και τελειοποίηση. Κατά το στάδιο της προ-εκαπίδευσης, το μοντέλο BERT εκπαιδεύεται σε δεδομένα κειμένου χωρίς ετικέτα, σε διαφορετικές εργασίες προ-εκπαίδευσης. Για το βήμα της τελειοποίησης, το μοντέλο αρχικοποιείται με τις προ-εκπαιδευμένες παραμέτρους και, στη συνέχεια, όλες οι παράμετροι ρυθμίζονται με τη χρήση επισημασμένων δεδομένων από τις μεταγενέστερες εργασίες. Κάθε μεταγενέστερη εργασία, παρόλο που έχει αρχικοποιηθεί με τις ίδιες προ-εκπαιδευμένες παραμέτρους, έχει ξεχωριστά λεπτομερή μοντέλα.

Το πρωτότυπο αγγλικό BERT έχει δύο εκ των προτέρων εκπαιδευμένες εκδόσεις ανάλογα με την κλίμακα της αρχιτεκτονικής του μοντέλου:

- BERT-base: 12 επίπεδα κωδικοποιητή (μπλοκ μετασχηματιστή), 12 attention heads και 110 εκατομμύρια παράμετρους.
- BERT-large: 24 επίπεδα κωδικοποιητή, 16 attention heads και 340 εκατομμύρια παράμετρους.

Και τα δύο μοντέλα, έχουν εκπαιδευτεί εκ των προτέρων από δεδομένα χωρίς ετικέτα που εξήχθησαν από την Αγγλική Βικιπαίδεια και το BooksCorpus με αυτοελεγχόμενο τρόπο. Με άλλα λόγια, εκπαιδεύτηκαν μόνο στα ακατέργαστα κείμενα, με αυτόματη διαδικασία δημιουργίας εισόδων και ετικετών από αυτά τα κείμενα.

Κεφάλαιο 6

# Ανάλυση Ιστού και Σελίδων Κοινωνικής Δικτύωσης

## 6.1 Γενικά

Διάφορα εργαλεία παρακολούθησης των κοινωνικών μέσων είναι διαθέσιμα για παρακολούθηση σε πραγματικό χρόνο της επιδημίας των ασθενειών. Ωστόσο, ορισμένα από αυτά δεν χρησιμοποιούν δεδομένα κοινωνικών μέσων (π.χ. HealthMap) ή χρησιμοποιούν μόνο δεδομένα Twitter και επικεντρώνονται στην ανίχνευση σήματος (π.χ., εργαλείο ECDC epitwittr) ή είναι ειδικά για μία ασθένεια (π.χ. WHO EARS ταμπλό) και συνήθως παρέχουν χαμηλό βαθμό προσαρμογής. Υπάρχουν διαθέσιμα και πιο ευέλικτα εργαλεία, αλλά είναι συνήθως εμπορικές και ακριβές λύσεις (π.χ. ClaraBridge). Θα συγκεντρώσουμε δεδομένα από τουλάχιστον δύο πλατφόρμες κοινωνικών μέσων (Twitter και Reddit), δεν θα εξειδικεύεται στη πανδημία και, το πιο σημαντικό, η δομή του θα επιτρέπει στον τελικό χρήστη να καθορίζει περισσότερα από ένα έργα παρακολούθησης και να δημιουργεί ad-hoc αναλύσεις, με βάση συγκεκριμένες ανάγκες του τελικού χρήστη. Το εργαλείο θα μπορεί να συν-σχεδιαστεί με τελικούς χρήστες από διαφορετικούς σχετικούς τομείς και θα αναπτυχθεί ακολουθώντας μια κοινωνικο-τεχνική προσέγγιση για να διασφαλιστεί η συμμόρφωση με τις αρχές δεοντολογίας, απορρήτου και διαφάνειας.

## 6.2 Παράδειγμα χρήσης

Οι αστυνομικές αρχές μπορούν να πραγματοποιήσουν το προτεινόμενο εργαλείο κατά τη διάρκεια της διαχείρισης μιας μολυσματικής ασθένειας ώστε να αυξήσουν την επίγνωση της κατάστασης και να κατανοήσουν το συναίσθημα του κοινού σχετικά με τα μέτρα ελέγχου που εφαρμόστηκαν. Οι αρχές μπορούν να δημιουργήσουν ένα έργο στην οθόνη διαχειριστή έργου και να εισαγάγουν τις λέξεις-κλειδιά που θέλουν να παρακολουθήσουν (όροι που σχετίζονται με την εκδήλωση της νόσου, με τη γεωγραφική περιοχή στην οποία λειτουργούν και με τα μέτρα που εφαρμόστηκαν) και τον αριθμό των ημερών που θέλουν η ανάλυση να τρέξει. Στο τέλος της περιόδου ανάλυσης, το εργαλείο δημιουργεί μια έκθεση όπου η αστυνομία μπορεί να παρατηρήσει την εξέλιξη της βαθμολογίας συναισθημάτων με την πάροδο του χρόνου για αναρτήσεις που αναφέρουν διαφορετικούς συνδυασμούς παρακολουθούμενων λέξεων-κλειδιών, όπως όρους που σχετίζονται με ασθένειες και όρους που σχετίζονται με μέτρα που

έχουν ληφθεί. Για tweets που παρέχουν γεωγραφικές πληροφορίες, η ανάλυση μπορεί να γίνεται ανά τοποθεσία. Επιπλέον, δημοσιεύσεις που ταιριάζουν με τα κριτήρια των παραμέτρων του έργου, συνοψίζονται από το εργαλείο με wordclouds και / ή barplot των πιο συχνά χρησιμοποιούμενων hashtags. Αυτά τα αναλυτικά στοιχεία μπορούν να ρίξουν φως στη στάση των χρηστών των μέσων κοινωνικής δικτύωσης έναντι των αρχών. Είναι γνωστό ότι ο σκεπτικισμός και η χαμηλή εμπιστοσύνη προς τα θεσμικά όργανα οδηγούν σε χαμηλότερη συμμόρφωση με τις πολιτικές αποφάσεις. Έτσι, η αστυνομία μπορεί να βελτιώσει την επικοινωνιακή της στρατηγική και να καθησυχάσει τους πολίτες σχετικά με τα θέματα (που συνάγονται από λέξεις-κλειδιά) που φαίνεται να σχετίζονται με πιο αρνητικά συναισθήματα.

## 6.3 Data ingestion

Αυτό το στοιχείο είναι υπεύθυνο για την επεξεργασία των παραμέτρων των έργων σε ερωτήματα που θα τροφοδοτούνται στα προγράμματα ανίχνευσης Twitter και Reddit. Αυτό περιλαμβάνει τη σύνδεση των όρων για παρακολούθηση με λογικούς τελεστές και τη μετατροπή τους στη σύνταξη που απαιτείται από τα API. Για αναζήτηση γεωγραφικών αναρτήσεων, το Twitter απαιτεί τις συντεταγμένες της τοποθεσίας ενδιαφέροντος. Αυτά λαμβάνονται μέσω ενός δωρεάν API υπηρεσίας web geocoding (Nominatim, πρόσβαση μέσω βιβλιοθήκης GeoPy Python). Τα δεδομένα υφίστανται βασική μορφοποίηση κειμένου και στη συνέχεια εισάγονται σε συλλογές βάσεων δεδομένων ανεπεξέργαστων δημοσιεύσεων Twitter και Reddit, μαζί με σχετικά μεταδεδομένα.

## 6.4 Insight Analytics

Αυτό το μέρος ανακτά τα μη επεξεργασμένα δεδομένα και εκτελεί διάφορες εργασίες ανάλυσης. Αυτές περιλαμβάνουν:

- συγκέντρωση δεδομένων εντός του χρόνου ενδιαφέροντος για τον υπολογισμό μετρήσεων βάσει του Natural Language Processing (NLP), όπως bigrams και δίκτυα λέξεων,
- δημιουργία διαφορετικών χρονοσειρών των μετρήσεων αναφορών των όρων προς αναζήτηση,
- εκτέλεση του αλγορίθμου ανίχνευσης σήματος για την αναγνώριση προειδοποιήσεων στις χρονοσειρές,

- αξιολόγηση της γεωγραφικής κατανομής των tweets. Για περιεχόμενο που περιλαμβάνει τοποθεσία, μαζί με τα γενικά αναλυτικά στοιχεία, δημιουργούνται και ανά τοποθεσία.

Η πλειονότητα των εργασιών ανάλυσης εκτελείται ξεχωριστά για το Twitter και το Reddit. Η διαφορετική δομή των δεδομένων που ανακτώνται από τα δύο API και ο διαφορετικός τύπος περιεχομένου και τρόπος κοινής χρήσης στις δύο πλατφόρμες, απαιτούν συγκεκριμένο χειρισμό και υποκινούν την ανάγκη εξαγωγής συγκεκριμένων πληροφοριών για την πλατφόρμα. Παρ 'όλα αυτά, ο συνδυασμός στατιστικών από τις δύο πηγές δεδομένων μπορεί να έχει ενδιαφέρον και, όταν είναι δυνατόν, πραγματοποιείται. Τα στοιχεία του Insight Analytics είναι επίσης υπεύθυνα για την επεξεργασία και την εναρμόνιση των αναλυτικών στοιχείων συναισθημάτων που παρέχονται από ένα εξωτερικό εργαλείο. Τέλος, συγκεντρώνει όλα τα αναλυτικά στοιχεία σε μια αναφορά, ενημερώνεται με δεδομένη συχνότητα, που αποθηκεύεται στη βάση δεδομένων και ανακτάται για την παραγωγή οπτικοποίησης στον πίνακα ελέγχου και για τροφοδοσία σε άλλα εργαλεία.

## 6.5 Scheduler

Ο προγραμματιστής είναι υπεύθυνος για την εκτέλεση των ερωτημάτων σύμφωνα με τις καθορισμένες προδιαγραφές χρόνου (συχνότητα ενημέρωσης και διάρκεια της ανάλυσης) και για τη διαχείριση των διαφόρων περιορισμών των APIσ. Ο προγραμματιστής ξεκινά εργασίες ανίχνευσης κάθε T ώρες (όπου το T είναι η επιλεγμένη περίοδος, που δεν έχει καθοριστεί ακόμη) για να αναζητήσει περιεχόμενο κοινωνικών μέσων που ταιριάζει με τα κριτήρια αναζήτησης τις τελευταίες T ώρες. Μόνο η πρώτη εργασία που εκτελείται από τον Scheduler είναι διαφορετική καθώς κοιτάζει πίσω το περιεχόμενο των προηγούμενων N ημερών, προκειμένου να δημιουργήσει ιστορικές γραμμές βάσης για το μοντέλο ανίχνευσης. Σημειώστε ότι το N μπορεί να διαφέρει ανάμεσα στο Twitter και στο Reddit. Το N ορίζεται σε 7 ημέρες για το Twitter (η ανίχνευση παλαιότερων δεδομένων δεν επιτρέπεται από το API), ενώ για το Reddit είναι μια παράμετρος για συντονισμό. Αυτό το στοιχείο ενημερώνει επίσης το Insight Analytics σχετικά με το χρόνο συγκέντρωσης στατιστικών στοιχείων και ενημερώνει τη δημιουργία αναφορών.

## 6.6 Data Layer

Αυτό το στοιχείο, που εφαρμόζεται στη MongoDB, φιλοξενεί όλα τα δεδομένα που χρησιμοποιούνται και παράγονται από το σύστημα. Η «Pandemic and social media knowledge-base» περιέχει γλωσσάρια για όρους που σχετίζονται με την πανδημία, μία λίστα αξιόπιστων και αναξιόπιστων διευθύνσεων URL για ανίχνευση αναξιόπιστων ειδήσεων, διανύσματα λέξεων και άλλα βοηθητικά προγράμματα. Οι «Social media content collections» περιλαμβάνουν ανεπεξέργαστες αναρτήσεις από το Twitter και το Reddit. Τα επεξεργασμένα δεδομένα, μετά την εκτέλεση όλων των αγωγών ανάλυσης, συγκεντρώνονται και αποθηκεύονται στη συλλογή "Analytics Reports".

Η διαδικασία δεδομένων που λαμβάνει χώρα στη μονάδα βασίζεται σε μία ontology-based ανάλυση και έχει ως αποτέλεσμα τον υπολογισμό της βαθμολογίας πολικότητας, μαζί με τη συχνότητα, τη συναίνεση, τη συνοχή και τη συνάφεια με συγκεκριμένους όρους (και τάσεις). Η βαθμολογία πολικότητας (αναφέρεται επίσης ως βαθμολογία συναισθημάτων ή συναισθημάτων) είναι μια συνεχής τιμή μεταξύ -1 και +1 που θα παρέχεται στην αρχική της μορφή και ως παράγοντας, που προκύπτει από την κατηγοριοποίηση της τιμής, υποδεικνύοντας διαφορετικά επίπεδα πολικότητας (π.χ. , σχετικά αρνητικά, έντονα αρνητικά, ουδέτερα, σχετικά θετικά, έντονα θετικά). Η οντολογία (που βασίζεται σε RDF / XML) που δημιουργείται είναι συγκεκριμένη περίπτωση χρήσης, ενώ η υποστηριζόμενη διαδικασία δεδομένων βασίζεται σε Linear Support Vector Machine (SVM) ή Naïve Bayes Network. Για περιεχόμενο με αναφορά σε τοποθεσία, θα συσχετισθεί το συναίσθημα κάθε ανάρτησης με τη χώρα στην οποία δημιουργήθηκε η ανάρτηση για τον εντοπισμό πιθανών πολιτισμικών προκαταλήψεων και θα επιτρέψει τη σύγκριση συναισθημάτων μεταξύ χωρών.

Οι είσοδοι είναι μια συλλογή εγγράφων για συγκεκριμένες περιπτώσεις (π.χ., αρχεία κειμένου / html), καθώς και συγκεκριμένοι όροι (π.χ. hashtags, εθνικότητα, χώρα, ορολογία / στόχοι που σχετίζονται με πανδημία, κάτοχοι γνώμης, προβολές, αριθμοί, ημερομηνίες, διευθύνσεις URL που χρησιμοποιούνται στην ανάλυση. Τέλος, περιλαμβάνεται ένας ενισχυμένος μηχανισμός μάθησης που επιτρέπει την εκπαίδευση των εφαρμοσμένων μοντέλων με βάση τα σχόλια που δημιουργούνται από τους χρήστες.

Τα κύρια τμήματα του εργαλείου ανάλυσης συναισθήματος μαζί με τη ροή δεδομένων παρουσιάζονται στην παρακάτω εικόνα. Αυτά περιλαμβάνουν:

- Data Connector: περιλαμβάνει το API για επικοινωνία με τα υπόλοιπα στοιχεία του εργαλείου, καθώς και το χειρισμό πτυχών διαχείρισης πρόσβασης. Αναπτύχθηκε ένα RESTful API. Μια σύντομη περιγραφή του API βρίσκεται στον ακόλουθο πίνακα (Πίνακας 2).
- Data Normalizer: αυτή η ενότητα εκτελεί δραστηριότητες προετοιμασίας δεδομένων (απενεργοποίηση θορύβου, επιμέλεια κ.λπ.) πριν πραγματοποιηθεί οποιαδήποτε διαδικασία.
- Αποθήκη δεδομένων: αποθηκεύει τα δεδομένα που συλλέχθηκαν πρόσφατα, καθώς και τα αποτελέσματα της ανάλυσης δεδομένων.
- Ontology Generator: είναι υπεύθυνη για τη δημιουργία της οντολογίας βάσει της οποίας θα πραγματοποιηθεί η ανάλυση συναισθημάτων.
- Data Extractor: εξάγει όρους που θα χρησιμοποιηθούν στην ανάλυση συναισθημάτων. Μπορεί να θεωρηθεί ως το προκαταρκτικό βήμα της ανάλυσης συναισθημάτων.
- Sentiment Analyzer: είναι υπεύθυνος για την εκτέλεση της ανάλυσης συναισθημάτων. Περιλαμβάνει τα μοντέλα (με βάση το Python) για αυτήν τη διαδικασία. Τα μοντέλα επανεκπαιδεύονται με τα δεδομένα της ίδιας μορφής και τύπου εκείνων που απορροφήθηκαν και υποβλήθηκαν.
- Validator: εφαρμόζει τον ενισχυμένο μηχανισμό εκμάθησης για να βελτιώνει συνεχώς τον αναλυτή και την ακρίβεια των μοντέλων AI.

## 6.7 Εφαμογή

Εφαρμόζουμε όλα όσα αναφέρονται στα προηγούμενα κεφάλαια για να κάνουμε μερικά πειράματα με δεδομένα που θα πάρουμε από το Twitter, ώστε να μπορέσουμε να κάνουμε την ανάλυση συναισθήματος που επιθυμούμε. Για τον σκοπό αυτό, απαιτείται μία διαδικασία ταυτοποίησης με την πλατφόρμα του Twitter, για να μπορούμε να έχουμε πρόσβαση στα tweets που μας ενδιαφέρουν.

### 6.7.1 Εξαγωγή tweets

Στην αναζήτησή μας ανάμεσα στα tweets βάζουμε ως όρο αναζήτησης τη λέξη *covid* και όριο τα 300 tweets, όπως φαίνεται και στο παρακάτω κομμάτι κώδικα:

```
#@title Twitter Search API Inputs
#@markdown ### Enter Search Query:
searchQuery = 'covid' #@param {type:"string"}
#@markdown ### Enter Max Tweets To Scrape:
#@markdown #### The Twitter API Rate Limit (currently) is 45,000
tweets every 15 minutes.
maxTweets = 300 #@param {type:"slider", min:0, max:45000, step:100}
Filter_Retweets = True #@param {type:"boolean"}

tweetsPerQry = 100  # this is the max the API permits
tweet_lst = []

if Filter_Retweets:
  searchQuery = searchQuery + ' -filter:retweets'  # to exclude
retweets

# If results from a specific ID onwards are reqd, set since_id to
that ID.
# else default to no lower limit, go as far back as API allows
sinceId = None

# If results only below a specific ID are, set max_id to that ID.
# else default to no upper limit, start from the most recent tweet
matching the search query.
max_id = -10000000000

tweetCount = 0
print("Downloading max {0} tweets".format(maxTweets))
while tweetCount < maxTweets:
    try:
        if (max_id <= 0):
            if (not sinceId):
                new_tweets = api.search(q=searchQuery,
count=tweetsPerQry, lang="en")
            else:
                new_tweets = api.search(q=searchQuery,
count=tweetsPerQry,
                                        lang="en", since_id=sinceId)
        else:
            if (not sinceId):
                new_tweets = api.search(q=searchQuery,
count=tweetsPerQry,
                                        lang="en", max_id=str(max_id
- 1))
            else:
                new_tweets = api.search(q=searchQuery,
count=tweetsPerQry,
```

```
                                                lang="en", max_id=str(max_id
- 1),
                                                since_id=sinceId)
            if not new_tweets:
                print("No more tweets found")
                break
            for tweet in new_tweets:
              if hasattr(tweet, 'reply_count'):
                reply_count = tweet.reply_count
              else:
                reply_count = 0
              if hasattr(tweet, 'retweeted'):
                retweeted = tweet.retweeted
              else:
                retweeted = "NA"

              # fixup search query to get topic
              topic = searchQuery[:searchQuery.find('-
')].capitalize().strip()

              # fixup date
              tweetDate = tweet.created_at.date()

              tweet_lst.append([tweetDate, topic,
                            tweet.id, tweet.user.screen_name,
tweet.user.name, tweet.text, tweet.favorite_count,
                            reply_count, tweet.retweet_count, retweeted])

            tweetCount += len(new_tweets)
            print("Downloaded {0} tweets".format(tweetCount))
            max_id = new_tweets[-1].id
        except tweepy.TweepError as e:
            # Just exit if any error
            print("some error : " + str(e))
            break

clear_output()
print("Downloaded {0} tweets".format(tweetCount))
```

Παίρνοντας τα tweets μπορούμε να τα αποθηκεύσουμε και να τα επεξεργαστούμε όπως επιθυμούμε.

```
pd.set_option('display.max_colwidth', -1)

# load it into a pandas dataframe
tweet_df = pd.DataFrame(tweet_lst, columns=['tweet_dt', 'topic',
'id', 'username', 'name', 'tweet', 'like_count', 'reply_count',
'retweet_count', 'retweeted'])
tweet_df.to_csv('tweets.csv')
tweet_df.head()
```

Τρέχοντας και το παραπάνω κομμάτι κώδικα, έχουμε σας αποτέλεσμα έναν πίνακα από τα tweets, όπως ενδεικτικά φαίνεται στην Εικόνα 21.

| | tweet_dt | topic | id | username | name | tweet | like_count | reply_count | retweet_count | retweeted |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2021-10-01 | Covid | 1443924233499910190 | SmallBigWord | Irina Smalley | @AugustaLees @dgurdasani1 @BBCNews @CMO_England @IndependentSage : This is a crime against our children, refusing t… https://t.co/BicBO21A8O | 0 | 0 | 0 | False |
| 1 | 2021-10-01 | Covid | 1443924232145154051 | NewPatchMomma | divided line | @ssolyom @ou812_Kess @prairiecentrist @Scribulatora Canadian Media is not owned by Canadians therefore it is not Ca… https://t.co/WhxU34GqfX | 0 | 0 | 0 | False |
| 2 | 2021-10-01 | Covid | 1443924231213879298 | Iblinkzz | Iblinkzz | @daiseskat @OldCriedPants @Reuters These are meant to be taken once you get covid, and likely it may be the case yo… https://t.co/IdNr8KFu33 | 0 | 0 | 0 | False |
| 3 | 2021-10-01 | Covid | 1443924228844179463 | CFNU | Canada's Nurses | This Sunday, CFNU Sec-Treas @PaulineWorsfold will join Ian Hanomansing on CBC's Cross Country Checkup to discuss AI… https://t.co/pZSHmkhAhz | 0 | 0 | 0 | False |
| 4 | 2021-10-01 | Covid | 1443924226570964995 | dragonflyeye | DragonFlyEye | Oh? Did it force itself on you? Wait to Squee hears about this. \n\nJustice Brett Kavanaugh tests positive for Covid-… https://t.co/9cISvbKw4Q | 0 | 0 | 0 | False |

*Εικόνα 21 Παράδειγμα αποτελεσμάτων ανεύρεσης tweets*

Δυστυχώς, το Twitter δεν επιτρέπει την αναζήτηση βάσει ημερομηνίας αλλά μπορούμε να φιλτράρουμε εμείς τα αποτελέσματα εκ των υστέρων.

### 6.7.2 NER και Ανάλυση Συναισθήματος

Στη συνέχεια θα προχωρήσουμε στην Ανάλυση συναισθήματος. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε τη βιβλιοθήκη Flair (63).

```
# predict NER
nerlst = []

for index, row in tqdm(tweet_df.iterrows(), total=tweet_df.shape[0]):
  cleanedTweet = row['tweet'].replace("#", "")
  sentence = Sentence(cleanedTweet, use_tokenizer=True)

  # predict NER tags
  tagger.predict(sentence)

  # get ner
  ners = sentence.to_dict(tag_type='ner')['entities']
```

```
  # predict sentiment
  classifier.predict(sentence)

  label = sentence.labels[0]
  response = {'result': label.value, 'polarity':label.score}

  # get hashtags
  hashtags = re.findall(r'#\w+', row['tweet'])
  if len(hashtags) >= 1:
    for hashtag in hashtags:
      ners.append({ 'type': 'Hashtag', 'text': hashtag })

  for ner in ners:
    adj_polarity = response['polarity']
    if response['result'] == 'NEGATIVE':
      adj_polarity = response['polarity'] * -1
    try:
      ner['type']
    except:
      ner['type'] = ''
    nerlst.append([ row['tweet_dt'], row['topic'], row['id'],
row['username'],
                    row['name'], row['tweet'], ner['type'],
ner['text'], response['result'],
                    response['polarity'], adj_polarity,
row['like_count'], row['reply_count'],
                   row['retweet_count'] ])

clear_output()

for index, row in tqdm(tweet_df.iterrows(), total=tweet_df.shape[0]):
  cleanedTweet = row['tweet'].replace("#", "")
  sentence = Sentence(cleanedTweet, use_tokenizer=True)

  # predict NER tags
  tagger.predict(sentence)

  # get ner
  ners = sentence.to_dict(tag_type='ner')['entities']

  # predict sentiment
  classifier.predict(sentence)

  label = sentence.labels[0]
  response = {'result': label.value, 'polarity':label.score}

  # get hashtags
  hashtags = re.findall(r'#\w+', row['tweet'])
```

```
  if len(hashtags) >= 1:
    for hashtag in hashtags:
      ners.append({ 'type': 'Hashtag', 'text': hashtag })

  for ner in ners:
    adj_polarity = response['polarity']
    if response['result'] == 'NEGATIVE':
      adj_polarity = response['polarity'] * -1
    try:
      ner['type']
    except:
      ner['type'] = ''
    neglst.append([ row['tweet'], ner['type'], ner['text'],
response['result'],
                   response['polarity'], adj_polarity ])

df_neg = pd.DataFrame(neglst, columns=['text', 'type',
'ner','sentiment', 'polarity', 'adj_polarity'])
df_neg.head()
```

Χρησιμοποιώντας τον κώδικα, που φαίνεται παραπάνω, δημιουργούμε ένα νέο tag και παίρνουμε την πόλωση (Polarity) και την ενσωματώνουμε στα δεδομένα μας (Εικόνα 22).

| | text | type | ner | sentiment | polarity | adj_polarity |
|---|---|---|---|---|---|---|
| 0 | @AugustaLees @dgurdasani1 @BBCNews @CMO_England @IndependentSage : This is a crime against our children, refusing t… https://t.co/BicBO21A8O | | England | NEGATIVE | 0.988741 | -0.988741 |
| 1 | @ssolyom @ou812_Kess @prairiecentrist @Scribulatora Canadian Media is not owned by Canadians therefore it is not Ca… https://t.co/WhxU34GqfX | | Scribulatora Canadian Media | NEGATIVE | 0.877705 | -0.877705 |
| 2 | @ssolyom @ou812_Kess @prairiecentrist @Scribulatora Canadian Media is not owned by Canadians therefore it is not Ca… https://t.co/WhxU34GqfX | | Canadians | NEGATIVE | 0.877705 | -0.877705 |
| 3 | @daiseskat @OldCriedPants @Reuters These are meant to be taken once you get covid, and likely it may be the case yo… https://t.co/IdNr8KFu33 | | Reuters | POSITIVE | 0.840035 | 0.840035 |
| 4 | This Sunday, CFNU Sec-Treas @PaulineWorsfold will join Ian Hanomansing on CBC's Cross Country Checkup to discuss Al… https://t.co/pZSHmkhAhz | | CFNU Sec-Treas | NEGATIVE | 0.744585 | -0.744585 |

*Εικόνα 22 Παράδειγμα αποτελεσμάτων όπου φαίνεται και το Polarity του κάθε tweet*

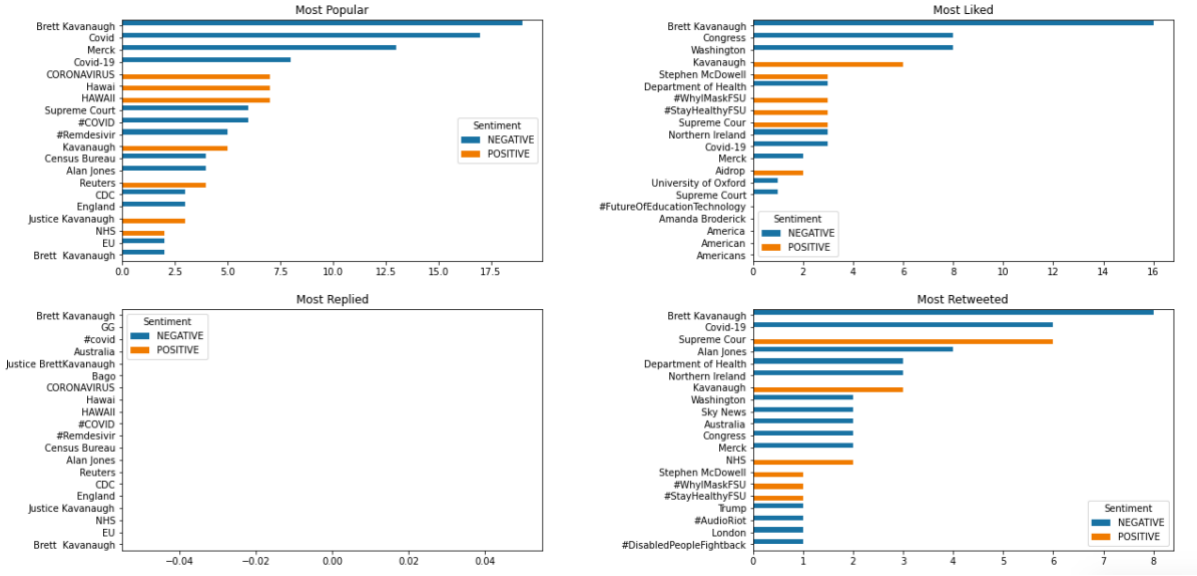
Μπορούμε να ομαδοποιήσουμε τα δεδομένα με βάση το tag και λαμβάνοντας υπόψιν τη μέση πόλωση (Average Polarity) να βγάλουμε συμπέρασμα για το συναίσθημα (Εικόνα 23).

| | index | tag | tag_type | Frequency | Avg_Polarity | Total_Likes | Total_Replies | Total_Retweets | Sentiment |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Brett Kavanaugh | | 19 | -0.327235 | 16 | 0 | 8 | NEGATIVE |
| 1 | 1 | Covid | | 17 | -0.088444 | 0 | 0 | 0 | NEGATIVE |
| 2 | 2 | Merck | | 13 | -0.504667 | 2 | 0 | 2 | NEGATIVE |
| 3 | 3 | Covid-19 | | 8 | -0.045827 | 3 | 0 | 6 | NEGATIVE |
| 4 | 4 | CORONAVIRUS | | 7 | 0.794719 | 0 | 0 | 0 | POSITIVE |

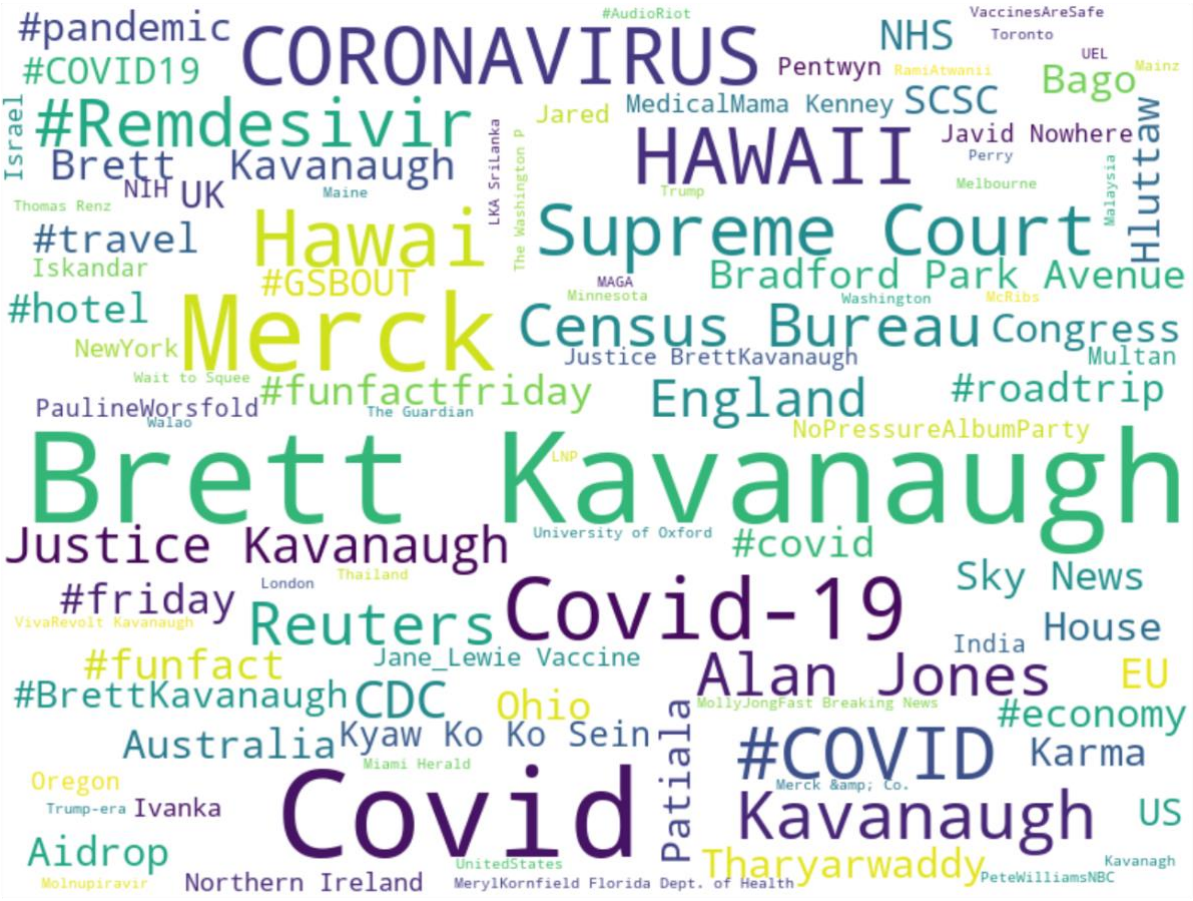
*Εικόνα 23 Συχνότητα εμφάνισης κάποιου tag και μέσο συναίσθημα*

### 6.7.3 Οπτικοποίηση αποτελεσμάτων

Στην Εικόνα 24 φαίνονται 4 διαγράμματα που δείχνουν ανά κατηγορία tweet (most popular, most liked κλπ) το συναίσθημα βάσει του tag και στην Εικόνα 25 το αντίστοιχο Σύννεφο Λέξεων (Word Cloud).




*Εικόνα 24 Διαγράμματα SA ανά tag, χωρισμένα σε 4 κατηγορίες*




*Εικόνα 25 Σύννεφο Λέξεων (Word Cloud)*

## 6.8 Μελλοντικές επεκτάσεις

Το αντικείμενο της παρούσας διπλωματικής εργασίας έχει πολύ χώρο ανάπτυξης και εξέλιξης.

- Δέσμευση τελικών χρηστών για την οριστικοποίηση της λίστας λειτουργιών που θα συμπεριληφθούν στο εργαλείο
- Οριστικοποίηση της οντολογίας που χρησιμοποιείται στην ανάλυση συναισθημάτων
- Εφαρμογή αγωγών Twitter και αναλυτικών στοιχείων Twitter και Reddit
- Καθορισμός προδιαγραφών για επικοινωνία με άλλα εργαλεία
- Βελτίωση των μηχανισμών ανίχνευσης και των κριτηρίων αναζήτησης για την ανάκτηση σχετικού περιεχομένου και την ελαχιστοποίηση των θορυβωδών σημάτων
- Εργασία για δείκτες ερμηνείας και αξιοπιστίας για τα παρεχόμενα αναλυτικά στοιχεία.
- Δοκιμή σε πραγματικά και προσομοιωμένα δεδομένα, ιδίως σε σχέση με τον αλγόριθμο ανίχνευσης σήματος / ανωμαλιών
- Δημιουργία βάσης γνώσεων για τον εμπλουτισμό πληροφοριών που δημιουργούνται από δεδομένα
- Αύξηση της ευελιξίας και της λειτουργικότητας του εργαλείου
- Βελτίωση της αποτελεσματικότητας της ανίχνευσης και αποθήκευσης δεδομένων

Όσον αφορά τη διαχείριση δεδομένων και την ενοποίηση-επικοινωνία, τα δεδομένα του Twitter συλλέγονται από το Twitter μέσω ενός δωρεάν Twitter API με το οποίο αλληλεπιδρούμε μέσω της βιβλιοθήκης Tweepy wrapper για την Python. Για να χρησιμοποιήθεί το Twitter API απαιτείται ένα σύνολο κλειδιών που σχετίζονται με έναν λογαριασμό προγραμματιστή Twitter, ο οποίος συνδέεται με έναν τυπικό λογαριασμό Twitter. Μέσω του δωρεάν τελικού σημείου API, μπορούν να ληφθούν τα χρονοδιαγράμματα των tweets ή τα tweets που έγιναν τις προηγούμενες 7 ημέρες μπορούν να ανακτηθούν και να φιλτραριστούν ανά λέξη-κλειδί, τοποθεσία και γλώσσα. Είναι επίσης δυνατή η ροή tweets σε πραγματικό χρόνο με τα ίδια φίλτρα. Σύμφωνα με την πολιτική δεδομένων του Twitter, δεν δημοσιεύουμε ούτε προβάλλουμε πληροφορίες που μπορούν να χρησιμοποιηθούν για την εκτέλεση αντιστοίχισης εκτός Twitter, δηλαδή: συσχέτιση ενός λογαριασμού Twitter με οποιονδήποτε χρήστη, άτομο, πρακτορείο με οποιοδήποτε αναγνωριστικό όχι στο Twitter. Επιπλέον, διασφαλίζουμε ότι προβάλλουμε μόνο στον χρήστη την πιο πρόσφατη έκδοση οποιουδήποτε κειμένου tweet και δεν εμφανίζουμε tweet από λογαριασμούς που έχουν

περιορίσει το κοινό τους από ενεργοποιημένα φίλτρα απορρήτου (γνωστά ως \ προστατευμένα tweets »). Τα δεδομένα Tweet αποθηκεύονται σε μια βάση δεδομένων πριν ανακτηθούν, συγκεντρωθούν και εμφανίζονται στον τελικό χρήστη. Αυτή η διαδικασία συγκέντρωσης ανωνυμοποιεί πλήρως τα δεδομένα. Όπου προβάλλουμε το πλήρες περιεχόμενο των tweets, αυτό ανωνυμοποιείται αφαιρώντας τη λαβή χρήστη από τα δεδομένα και το πλήρες κείμενο παρουσιάζεται αναλλοίωτο (απαγορεύοντας τη μορφοποίηση για παρουσίαση). Δεν συλλέγουμε ούτε αποθηκεύουμε πληροφορίες που θα μπορούσαν να χρησιμοποιηθούν για την ταυτοποίηση μεμονωμένων χρηστών Twitter, ούτε χρησιμοποιούμε κανένα από τα δεδομένα που συλλέχθηκαν για να συμπεράνουμε τυχόν χαρακτηριστικά των χρηστών Twitter, όπως ηλικία, φύλο, εθνικότητα κ.λπ. Χρησιμοποιούμε τα API σε -χτισμένα φίλτρα για την επιλογή δεδομένων που σχετίζονται με το ερώτημα του τελικού χρήστη, αλλά δεν φιλτράρουμε tweet βάσει οποιασδήποτε ιδιότητας εκτός από την παρουσία λέξεων-κλειδιών, δεδομένων τοποθεσίας, γλώσσας που εντοπίστηκε και την ημερομηνία δημιουργίας του tweet.

Για το Reddit, δεν έχουμε άμεση πρόσβαση σε δεδομένα από το Reddit. Χρησιμοποιούμε τον πόρο PushShift, που είναι ένα ψηφιακό αρχείο περιεχομένου Reddit. Το PushShift διαθέτει ένα API για αναζήτηση και ανάκτηση αρχειοθετημένου κειμένου. Το αρχείο PushShift ενημερώνεται σε πραγματικό χρόνο, με το περιεχόμενο να προστίθεται και να αφαιρείται από το αρχείο για να διατηρείται η ισοτιμία με το ίδιο το Reddit. Το PushShift δεν είναι επίσημο προϊόν ή υπηρεσία Reddit, αν και τα δεδομένα στα οποία γίνεται πρόσβαση μέσω του PushShift API υπόκεινται στους ίδιους όρους με τα δεδομένα που έχουν πρόσβαση απευθείας μέσω του Reddit API, το οποίο περιγράφει παρόμοιες σκέψεις με αυτές που αναφέρονται παραπάνω στο Twitter. Ως εκ τούτου, συγκεντρώνουμε και ανώνυμα δεδομένα για παρουσίαση στον τελικό χρήστη. Τα δεδομένα που επεξεργαζόμαστε δεν χρησιμοποιούνται για το προφίλ ή την ταυτοποίηση των χρηστών του Reddit και όπου παρουσιάζουμε περιεχόμενο πλήρους κειμένου, συλλέγουμε το πιο ενημερωμένο αντίγραφο των δεδομένων και δεν τα επεξεργαζόμαστε ή τα τροποποιούμε εκτός από τη μορφοποίηση για αναγνωσιμότητα.

Όσον αφορά τα σχετικά πρότυπα, η μορφή RDF / XML χρησιμοποιείται για τις οντολογίες συναισθήματος. Για την επικοινωνία χρησιμοποιείται ένα RESTful API και για την ανάπτυξη χρησιμοποιούνται οι τεχνολογίες FastAPI και React.

Κεφάλαιο 7

# Συμπεράσματα

Σε αυτήν την εργασία, συζητήσαμε και συγκρίναμε μεθόδους για τη χρήση του προβλήματος της ανάλυσης συναισθήματος. Το εφαρμόσουμε σε δημοσιεύσεις από τα κοινωνικά δίκτυα, με σκοπό να εξάγουμε ένα συμπέρασμα, τον βαθμό εμπιστοσύνης των πολιτών σε δημόσιους φορείς και φορείς υγείας, σε έκτακτες υγειονομικές κρίσεις. Είδαμε με εύκολο τρόπο, πως να ανακτούμε δημοσιεύσεις από το διαδίκτυο (tweets) και εφαρμόζοντας τις τεχνολογίες που αναλύονται στα πλαίσια της εργασίας, να εξάγουμε χρήσιμα συμπεράσματα.

Λόγω έλλειψης πόρων, δεν μπορούσαμε να αντιμετωπίσουμε μεγαλύτερα μοντέλα βαθιάς μάθησης. Η έρευνα δείχνει ότι τα μοντέλα με πολύ περισσότερες παραμέτρους από αυτές που χρησιμοποιήσαμε, επαυξημένες με μεθόδους βασισμένες στη γνώση, μπορούν να επιτύχουν επαρκώς το σημασιολογικό πλαίσιο για ακρίβεια 70-80% στην ανάλυση συναισθημάτων. Έτσι, μια ιδέα για μελλοντική εργασία είναι να γίνουν πειράματα με μοντέλα SOTA. Επιπλέον, μπορούμε να πειραματιστούμε με περισσότερα σύνολα δεδομένων και να πραγματοποιήσουμε μια βελτιστοποίηση των υπερ-παραμέτρων.

# Βιβλιογραφία

1. *Whose and What Chatter Matters? The Impact of Tweets on Movie Sales.* **H. Rui, Y. Liu, and A. Whinston.** s.l. : Decision Support Systems, 2011, Vol. 55.
2. *Twitter Mood Predicts the Stock Market.* **J. Bollen, H. Mao, and X.-J. Zeng.** s.l. : Journal of Computational Science, 2010, Vol. 2.
3. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* **A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe.** 2010, Vol. 2.
4. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* **B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith.** 2010, Vol. 11.
5. *Classifying sentiment in microblogs: Is brevity an advantage.* **Smeaton, A. Bermingham and A.** 2010.
6. *Can Collective Sentiment Expressed on Twitter Predict Political Elections?* **Mustafaraj, J. Chung and E.** 2011, Vol. 11.
7. *A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data.* **Gayo-Avello, D.** s.l. : Social Science Computer Review, 2012, Vol. 31.
8. *Twitter Power: Tweets as Electronic Word of Mouth.* **J. Jansen, M. Zhang, K. Sobel, and A. Chowdury.** s.l. : JASIST, 2009, Vol. 60.
9. *Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power.* **Kim, K. Dong Sung and J.** 2014.
10. *Characterizing debate performance via aggregated twitter sentiment.* **Shamm, N. Diakopoulos and D.** 2010, Vol. 2.
11. *Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network.* **M. Ghiassi, J. Skinner, and D. Zimbra.** s.l. : Expert Systems with Applications, 2013, Vol. 40.
12. *Webis:AnEnsemblefor Twitter Sentiment Detection.* **M.Hagen, M.Potthast,M.Buĺchner,andB.Stein.** 2015.
13. *Sentiment Classification from Multi- class Imbalanced Twitter Data Using Binarization.* **B. Krawczyk, B. Mcinnes, and A. Cano.** 2017.
14. *A context-based model for Sentiment Analysis in Twitte.* **A. Vanzo, D. Croce, and R. Basili.** 2014.
15. *Sentiment Estimation on Twitter.* **G. Amati, M. Bianchi, and G. Marcone.** 2014.
16. *Short text classification in twitter to improve information filtering.* **B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas.** 2010.
17. *A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets.* **M. Boia, B. Faltings, C. Musat, and P. Pu.** 2013.
18. *Sentiment analysis: Measuring sentiment strength of call centre conversations.* **Y. Priyadarshana, K. Gunathunga, K. N. N. Perera, L. Ranathunga, P. Karunaratne, and T. Thanthriwatta.** 2015.
19. *Quantifying modified opinion strength: A fuzzy inference system for Sentiment Analysis.* **Bhatia, R. Srivastava and M. : S.** 2013.
20. *Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks.* **M. Ghiassi, D. Zimbra, and S. Lee.** s.l. : Journal of Management Information Systems, 2016, Vol. 33.
21. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* **Turney, P.** s.l. : Computing Research Repository - CORR, 2002.
22. *Determining the sentiment of opinions.* **Hovy, S.-M. Kim and E.** 2004.
23. *Thumbs up? Sentiment Classification Using Machine Learning Techniques.* **B. Pang, L. Lee, and S. Vaithyanathan.** s.l. : EMNLP, 2002, Vol. 10.

24. *Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis.* **Gamon, M.** 2014.
25. *From Classification to Quantification in Tweet Sentiment Analysis.* **Sebastiani, W. Gao and F.** s.l. : Social Network Analysis and Mining, 2016, Vol. 6.
26. *Enhanc- ing Deep Learning Sentiment Analysis with Ensemble Techniques in So- cial Applications.* **O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. Iglesias.** s.l. : Expert Systems with Applications, 2017, Vol. 77.
27. **[Online] https://www.researchgate.net/figure/Examples-of-Supervised-Learning-Linear-Regression-and-Unsupervised-Learning_fig3_336642133.**
**28. [Online] https://techvidvan.com/tutorials/reinforcement-learning/.**
**29. [Online] https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding.**
**30. *LIII. On lines and planes of closest fit to systems of points in space.* F.R.S., K. P. 11, s.l. : The London, Edinburgh, and Dublin Philosophical Magazine and Jour- nal of Science, Vol. 2.**
**31. *Viualizing data using t-SNE.* Hinton, L. van der Maaten and G. s.l. : Journal of Machine Learning Research, 2008, Vol. 9.**
**32. Mc., Conor. [Online] https://conrmcdonald.medium.com/log-loss-a-short-note-b50e42af0713.**
**33. *Reconstruction of porous media from extremely limited information using conditional generative adversarial networks* . Junxi Feng, Xiaohai He, Qizhi Teng, Chao Ren. 2019.**
**34. [Online] https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/.**
**35. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting.* N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdi- nov. s.l. : Journal of Machine Learning Research, 2014, Vol. 15.**
**36. Dean, J. *Large-Scale Deep Learning for Intelligent Computer Systems.* [Online] Google TechTalk, 2016. https://www.youtube.com/watch?v=QSaZGT4-6EY.**
**37. Karpathy, Andrej. [Online] http://karpathy.github.io/2015/05/21/rnn-effectiveness/.**
**38. [Online] https://www.telusinternational.com/articles/difference-between-cnn-and-rnn.**
**39. *Bidirectional recurrent neural networks.* Paliwal, M. Schuster and K. s.l. : Signal Processing, IEEE Transactions on, 1997, Vol. 45.**
**40. *A Traffic Flow Prediction Method Based on Road Crossing Vector Coding and a Bidirectional Recursive Neural Network.* Shuanfeng Zhao, Qingqing Zhao, Yunrui Bai and Shijun Li. s.l. : Electronics, 2019.**
**41. *Long Short-term Memory.* Schmidhuber, S. Hochreiter and J. s.l. : Neural com-putation, 1997, Vol. 9.**
**42. [Online] https://www.tutorialexample.com/understand-the-effect-of-lstm-input-gate-forget-gate-and-output-gate-lstm-network-tutorial/.**
**43. D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. 2016.**
**44. I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. 2014.**
**45. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder- Decoder for Statistical Machine Translation. 2014.**
**46. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. 2017.**

**47. D. Bahdanau, K. Cho, and Y. Bengio.** ***Neural Machine Translation by Jointly Learning to Align and Translate.*** **2016.**
**48. M.-T. Luong, H. Pham, and C. D. Manning.** ***Effective Approaches to Attention- based Neural Machine Translation.*** **2015.**
**49. A. Graves, G. Wayne, and I. Danihelka.** ***Neural Turing Machines.*** **2014.**
**50. J. Cheng, L. Dong, and M. Lapata.** ***Long Short-Term Memory-Networks for Machine Reading.*** **2016.**
**51. Sharma, Palash. [Online] https://machinelearningknowledge.ai/keras-tokenizer-tutorial-with-examples-for-fit_on_texts-texts_to_sequences-texts_to_matrix-sequences_to_matrix/.**
**52. Mehta, Deep. [Online] https://byteiota.com/stemming-and-lemmatization/.**
**53. Altaweel, Mark. GIS and Topic Modeling. [Online] https://www.gislounge.com/gis-topic-modeling/.**
**54. Mehmood, Arshad. Generate Unigrams Bigrams Trigrams Ngrams Etc In Python. [Online] https://arshadmehmood.com/development/generate-unigrams-bigrams-trigrams-ngrams-etc-in-python/.**
**55.** ***A Neural Probabilistic Language Model.*** **Y. Bengio, R. Ducharme, and P. Vincent. 2000, Vol. 3.**
**56. T. Mikolov, K. Chen, G. S. Corrado, and J. Dean.** ***Efficient Estimation of Word Representations in Vector Space.*** **[Online] 2013. http://arxiv.org/abs/1301.3781.**
**57.** ***Glove: Global Vectors for Word Representation.*** **J. Pennington, R. Socher, and C. Manning. 2014, Vol. 14.**
**58.** ***Extending Thesauri Using Word Embeddings and the Intersection Method.*** **Jörg Landthaler, Bernhard Waltl, Dominik Huth, Daniel Braun and Florian Matthes.**
**59. M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer.** ***Deep contextualized word representations.*** **2018.**
**60. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova.** ***BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*** **2019.**
**61. ELMo in Practice. [Online] https://ireneli.eu/2018/12/17/elmo-in-practice/.**
**62. Alammar, J. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning. [Online] http://jalammar.github.io/illustrated-bert/.**
**63. [Online] https://github.com/zalandoresearch/flair.**