Πανεπιστήμιο Δυτικής Αττικής
Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**


Αριστοτέλης Χασαπόπουλος

# Ανάλυση Συναισθήματος σε Δεδομένα από το Twitter με Χρήση Αλγορίθμων Μηχανικής Μάθησης

Επιβλέπων: Χρήστος Τρούσσας
ΠΜΣ: Επιστήμη και Τεχνολογία της Πληροφορικής και των Υπολογιστών

Αθήνα 2021

**Τριμελής Επιτροπή**

| Κλειώ Σγουροπούλου | Αθανάσιος Βουλόδημος | Χρήστος Τρούσσας |
|---|---|---|
| KLEIO SGOUROPOULOU<br>Digitally signed by KLEIO SGOUROPOULOU<br>Date: 2021.11.17 12:10:20 +02'00' | Athanasios Voulodimos<br>Digitally signed by Athanasios Voulodimos<br>Date: 2021.11.05 01:43:44 +02'00' | Christos Troussas<br>Digitally signed by Christos Troussas<br>Date: 2021.11.02 09:45:54 +02'00' |
| ..... | ..... | ..... |
| Καθηγήτρια<br>Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών<br>Πανεπιστήμιο Δυτικής Αττικής | Επίκουρος Καθηγητής<br>Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών<br>Πανεπιστήμιο Δυτικής Αττικής | Διδάκτωρ<br>Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών<br>Πανεπιστήμιο Δυτικής Αττικής |

*Στην κορούλα μου Ανδριάνα.*

**Τίτλος:** Ανάλυση Συναισθήματος σε Δεδομένα από το Twitter με Χρήση Αλγορίθμων Μηχανικής Μάθησης

**Συγγραφέας:** Αριστοτέλης Χασαπόπουλος

**Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών:**Πανεπιστήμιο Δυτικής Αττικής

**Επιβλέπων:** Χρήστος Τρούσσας, Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών
Πανεπιστήμιο Δυτικής Αττικής

**Περίληψη:** Η παρούσα διπλωματική εργασία έχει ως κύριο θεματικό άξονα την *Ανάλυση Συναισθήματος* για δεδομένα που προέρχονται από το κοινωνικό δίκτυο *Twitter*. Με τον όρο «ανάλυση συναισθήματος» αναφερόμαστε σε μια αυτοματοποιημένη διαδικασία που προσδιορίζει τη συναισθηματική πολικότητα ενός κειμένου. Ως κλάδος έχει συγκεντρώσει τα βλέμματα τόσο της επιστημονικής κοινότητας όσο και διάφορων επιχειρηματικών κλάδων, αφού η ευρεία ενσωμάτωση των διαδικτυακών κοινωνικών δικτύων στην καθημερινότητα, έχει δημιουργήσει άφθονα και ευκόλως προσβάσιμα δεδομένα προς επεξεργασία και εξαγωγή συμπερασμάτων. Ο κλάδος της Ανάλυσης Συναισθήματος περιέχει μεθόδους που ανήκουν στους επιστημονικούς τομείς της *Τεχνητής Νοημοσύνης*, της *Μηχανικής Μάθησης* και της *Βαθιάς Μάθησης*.

Το Κεφάλαιο 1 αποτελεί μια σύντομη εισαγωγή στα κοινωνικά δίκτυα και ειδικότερα στο Twitter. Συγκεκριμένα, περιγράφεται η δομή, η λειτουργία του καθώς και οι τομείς στους οποίους μπορεί να διενεργηθεί περαιτέρω ανάλυσή του.

Στη συνέχεια, παρουσιάζονται οι δύο κύριες προσεγγίσεις του κλάδου της Ανάλυσης Συναισθήματος, αυτή της Μηχανικής Μάθησης και αυτή του σημασιολογικού προσανατολισμού. Επιπλέον, παρουσιάζονται τρόποι αναπαράστασης χαρακτηριστικών των δεδομένων ενός κειμένου, καθώς και μέθοδοι προεπεξεργασίας κειμένου. Το Κεφάλαιο 2 ολοκληρώνεται με την περιγραφή της διαδικασίας αξιολόγησης των μοντέλων ανάλυσης συναισθήματος.

Έπειτα, στο Κεφάλαιο 3, γίνεται παρουσίαση των πιο διαδεδομένων μοντέλων Μηχανικής Μάθησης που χρησιμοποιούνται στον κλάδο της ανάλυσης συναισθήματος. Τέτοια μοντέλα είναι η οικογένεια ταξινομητών Bayes, οι μηχανές διανυσματικής υποστήριξης, ο αλγόριθμος k-κοντινότερων γειτόνων και τα συνελικτικά νευρωνικά δίκτυα.

Στο τελευταίο κεφάλαιο γίνεται εφαρμογή 3 αλγορίθμων Μηχανικής Μάθησης και σύγκριση της απόδοσής τους ως προς την κατηγοριοποίηση συναισθήματος. Μετά την παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε στα πειράματα, πραγματοποιείται η περιγραφή των μοντέλων που χρησιμοποιήθηκαν και ο σχολιασμός των αποτελεσμάτων. Τα μοντέλα που υλοποιήθηκαν, αναπτύχθηκαν στην έκδοση 3.8 της Python.

**Λέξεις κλειδιά:** Μηχανική Μάθηση, Εξόρυξη δεδομένων, Ανάλυση συναισθήματος, Κείμενο

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

# 1. Κοινωνικά Δίκτυα και Παγκόσμιος Ιστός

## 1.1 Εξόρυξη και Ανάλυση Δεδομένων

Στη σύγχρονη εποχή του διαδικτύου, η οποία έχει καταστήσει τους ίδιους τους χρήστες του πρωταγωνιστικούς παράγοντες παραγωγής δεδομένων, είναι ιδιαιτέρως χρήσιμη η κατασκευή και αξιοποίηση εργαλείων που επιτρέπουν τη διαχείριση του τεράστιου όγκου δεδομένων (Σχήμα 1.1) με σκοπό τον εντοπισμό χρήσιμων μοτίβων [1] και κατ' επέκταση τη δημιουργία αξιοποιήσιμων πληροφοριών.


Σχήμα 1.1: Παραγωγή δεδομένων ανά λεπτό (2019)

Οι κλάδοι που χρειάζονται την προκύπτουσα γνώση από την εξόρυξη και ανάλυση των δεδομένων είναι ποικίλοι. Ενδεικτικά αναφέρονται ο κλάδος της προώθησης προϊόντων - για τη μελέτη της γνώμης των καταναλωτών σχετικά με κάποιο προϊόν ή υπηρεσία, ο κλάδος των δημοσίων σχέσεων - για τη μελέτη της στάσης των πολιτών σχετικά με τις εφαρμοζόμενες πολιτικές. Τα εξαγόμενα πρότυπα και οι τάσεις απασχολούν τις επιχειρήσεις, οι οποίες τα αξιοποιούν προκειμένου να σχεδιάσουν την επιχειρησιακή τους στρατηγική κατά το βέλτιστο δυνατό τρόπο [2]. Η κάλυψη των διαφορετικών αναγκών κάθε κλάδου ωθεί τους επιστήμονες της πληροφορικής στην παροχή ολοένα και πιο εξειδικευμένων εργαλείων που να εξορύττουν όσο το δυνατό πιο «χρήσιμα» δεδομένα.

Η Εξόρυξη Δεδομένων αντλεί μεθοδολογίες από διάφορους τομείς της Επιστήμης και της Τεχνολογίας, όπως η Μηχανική μάθηση, οι Βάσεις Δεδομένων και η Στατιστική, προκειμένου να ανακαλύψει πληροφορία μέσα από μεγάλους όγκους δεδομένων. Ως αποτέλεσμα, δεν υπάρχει ένας ορισμός για την εξόρυξη δεδομένων που να είναι κοινώς αποδεκτός γιατί οι διάφοροι ορισμοί που έχουν δοθεί, αντιπροσωπεύουν την οπτική γωνία των συγγραφέων τους. Στη συνέχεια παρουσιάζουμε τους ορισμούς σύμφωνα με τους Witten, Frank Han, Kamber [3, 4]

**Ορισμός 1.1.** *Η Εξόρυξη Δεδομένων αξιοποιείται προς την ανακάλυψη γνώσης μέσα σε μεγάλους όγκους δεδομένων.*

**Ορισμός 1.2.** *Η Εξόρυξη Δεδομένων δύναται να οριστεί ως η διαδικασία εντοπισμού προτύπων μέσα από σύνολα δεδομένων, αποδίδοντας έτσι, έμφαση στη διάσταση της Μηχανικής μάθησης.*

Κατά την εξόρυξη δεδομένων ο χρήστης δύναται να επιλέξει ανάμεσα σε δύο είδη ανάλυσης:

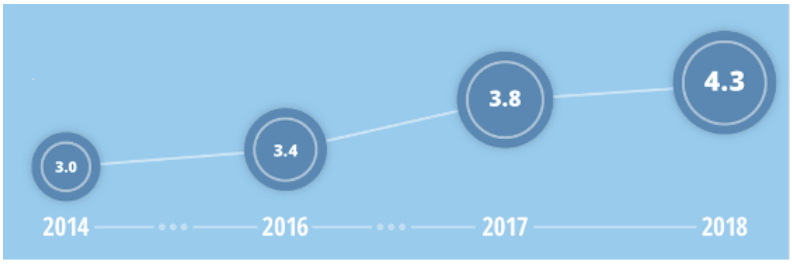
1. Την περιγραφική ανάλυση όπου γίνεται ομαδοποίηση των δεδομένων και παρουσίαση των ιδιοτήτων τους.

2. Την προγνωστική ανάλυση όπου μέσω της κατάρτισης κάποιου μοντέλου επιδιώκεται η διατύπωση προβλέψεων.

Η Ανάλυση Δεδομένων είναι μια διαδικασία κατά την οποία τα δεδομένα που έχουν συλλεχθεί, υφίσταντο κάποια μορφή προεπεξεργασίας όπως ο καθαρισμός ή/και η μετατροπή και στη συνέχεια μοντελοποιούνται με σκοπό την παραγωγή χρήσιμης πληροφορίας.

## 1.2 Κοινωνικά Δίκτυα

Η εμφάνιση των μέσων κοινωνικής δικτύωσης προκάλεσε αξιοσημείωτη αλλαγή στον τρόπο δομής και ανάπτυξης του διαδικτύου. Πλέον τα παραδοσιακά μέσα ενημέρωσης και επικοινωνίας έχουν παραγκωνιστεί από τους ίδιους τους χρήστες αφού τα μέσα κοινωνικής δικτύωσης παρέχουν διαδραστικότητα και άμεση αλληλεπίδραση μεταξύ των χρηστών. Επίσης η άνθιση της κινητής τηλεφωνίας, επιτρέπει σε καθημερινή βάση, τη συμμετοχή των χρηστών σε μια πληθώρα μέσων κοινωνικής δικτύωσης (Σχήμα 1.2). Η ευκολία με την οποία οποιοσδήποτε χρήστης μπορεί ν᾽ αναρτήσει οποιουδήποτε είδους πληροφορία (κείμενο, πολυμέσα, υπερσυνδέσμους), τα καθιστά μια σημαντική πηγή αλληλεπίδρασης και έχει δημιουργήσει την ανάγκη αξιολόγησης και αξιοποίησης αυτής της πληροφορίας από επιστημονικούς και επιχειρηματικούς κλάδους.




Σχήμα 1.2: Παγκόσμιος πληθυσμός Διαδικτύου (σε δισεκατομμύρια)

Οι ιστοσελίδες κοινωνικής δικτύωσης, κοινώς γνωστές για τη διάδοση πληροφοριών, είναι εικονικές κοινότητες όπου οι εγγεγραμμένοι χρήστες μπορούν να δημιουργήσουν τα εικονικά τους προφίλ και ν΄ αναπτύξουν ένα δίκτυο επαφών. Επιπλέον παρέχονται στους χρήστες εργαλεία ώστε να μπορούν να επικαιροποιούν το προφίλ τους με αναρτήσεις κάθε είδους, καθώς και να σχολιάζουν αναρτήσεις άλλων χρηστών. Είναι φανερό πως η δημοφιλία και το περιεχόμενο της εκάστοτε πλατφόρμας κοινωνικής δικτύωσης, εξαρτάται αποκλειστικά από τους χρήστες του.

Σε αυτό το σημείο θα μπορούσαμε να περιγράψουμε ένα μέσον κοινωνικής δικτύωσης ως μια δομή η οποία συντελείται από κόμβους. Με τη σειρά τους οι κόμβοι συνδέονται μεταξύ τους με διαφορετικούς τύπους αλληλεξάρτησης, έναν ή περισσότερους (Σχήμα 1.3).

Σχήμα 1.3: Σχηματική απεικόνιση κοινωνικού δικτύου.

Σύμφωνα με τους Kaplan και Haenlein [5]: «*Τα μέσα κοινωνικής δικτύωσης ορίζονται σαν ένα σύνολο από διαδικτυακές εφαρμογές που βασίζονται στα ιδεολογικά και τεχνολογικά θεμέλια του Web 2.0 και επιτρέπουν τη δημιουργία και την ανταλλαγή περιεχομένου από τους χρήστες.*»

Τα μέσα κοινωνικής δικτύωσης έχουν πέντε βασικές συνιστώσες [6]:

1. *Συμμετοχή (Participation)*
   Δίδεται στα εμπλεκόμενα μέρη (ιδιώτες, επιχειρήσεις, οργανισμούς) η ευκαιρία να αλληλεπιδράσουν μεταξύ τους μέσω της ατομικής τους συμβολής στην κατάρτιση του περιεχομένου των αναρτήσεών τους.

2. *Διαφάνεια (Openness)*
   Η διαθεσιμότητα μηχανισμών για τη σύνθεση και το διαμοιρασμό περιεχομένου, δίνει τη δυνατότητα για ανατροφοδότηση και συμμετοχή των χρηστών κάτω από ελάχιστους περιορισμούς ως προς την πρόσβαση και τη χρήση του περιεχομένου.

3. *Συνομιλία (Conversation)*
   Σε αντίθεση με το παραδοσιακά μέσα που απευθύνονται σε παθητικούς χρήστες, στα μέσα κοινωνικής δικτύωσης επιτρέπεται η επικοινωνία διπλής κατεύθυνσης.

4. *Κοινότητα (Community)*
   Είναι δυνατή, με ιδιαίτερη ευκολία, η δημιουργία κοινοτήτων χρηστών που έχουν κοινά ενδιαφέροντα. Ωστόσο, ο σκοπός των μέσων κοινωνικής δικτύωσης είναι η επικοινωνία μεταξύ των χρηστών που εμφανίζουν κοινά ενδιαφέροντα σε μια συγκεκριμένη περίοδο.

5. *Συνεκτικότητα (Connectedness)*
   Μπορεί να θεωρηθεί ως ένα μέτρο εγγύτητας και ταύτισης μεταξύ των χρηστών ενός μέσου κοινωνικής δικτύωσης. Η συνεκτικότητα ενισχύεται κατά τη χρήση συνδέσεων με άλλες ιστοσελίδες, πόρους και χρήστες.

Τα διαδικτυακά κοινωνικά δίκτυα μπορούν να αντιμετωπιστούν ως μια σύγχρονη διαδικτυακή κοινωνία, μέσω της οποίας εξάγονται ενδιαφέροντα συμπεράσματα για την ανθρώπινη αλληλεπίδραση. Αυτό που τα καθιστά όμως τόσο σημαντικό ερευνητικό πεδίο είναι η διάδοσή τους σε παγκόσμια κλίμακα, το γεγονός ότι όλα συμβαίνουν σε πραγματικό χρόνο και το χαμηλό κόστος πρόσβασης στα δεδομένα τους.

### 1.2.1 Εξόρυξη Δεδομένων Στα Κοινωνικά Δίκτυα

Η Εξόρυξη Δεδομένων από τα κοινωνικά δίκτυα είναι ιδιαιτέρως σημαντική αφού τα ψηφιακά μέσα μπορούν να μελετήσουν τις ανθρώπινες σχέσεις και τα συναισθήματα δίχως την πραγματοποίηση ενδελεχών ερευνών [7]. Η εξόρυξη που λαμβάνει χώρα στα κοινωνικά δίκτυα, δύναται να οριστεί ως η σειρά κατάλληλων ενεργειών που οδηγεί στην εξαγωγή προτύπων από δεδομένα, προφανώς μέσω κοινωνικής δικτύωσης, και αξιοποιεί τα κατάλληλα εργαλεία ανάλυσης και μοντελοποίησης για τη διερεύνηση μεγάλων σε όγκο δεδομένων.

Είναι κοινώς αποδεκτό ότι τα μέσα κοινωνικής δικτύωσης επηρεάζουν το μάρκετινγκ, αναδεικνύουν τις τάσεις και αποτελούν την καταλληλότερη πηγή έρευνας προκειμένου να μελετηθούν διεξωδικά οι μηχανισμοί επιρροής του κοινού, ώστε να αξιοποιούνται με το βέλτιστο δυνατό τρόπο [8]. Αξιοποιώντας το γεγονός αυτό, οι επιχειρήσεις διατηρούν λογαριασμούς στα μέσα κοινωνικής δικτύωσης και έτσι, αποκτούν άμεση πρόσβαση με τεράστιο όγκο δεδομένων. Η ανάλυση αυτών των δεδομένων παρέχει τη δυνατότητα να βελτιώσουν τις επιδόσεις τους σε διαφορετικούς τομείς και φυσικά, να διατηρούνται ανταγωνιστικές. Από την άλλη, όμως, η πρόσβαση σε τόσο χρήσιμη πληροφορία αποτελεί μια σημαντική και ταυτόχρονα πολύ δύσκολη διαδικασία, καθώς χρειάζεται να λάβουν υπόψιν τους μια σειρά από παραμέτρους:

1. Το μοντέλο που θα δημιουργήσουν να μπορεί να επεξεργάζεται τεράστιο όγκο δεδομένων σε ένα εύλογο χρονικό διάστημα. Οι συχνές αλλαγές σε σύντομο χρονικό διάστημα, που παρατηρούνται στη φύση των κοινωνικών δικτύων, διαδραματίζει σημαντικό ρόλο και, σίγουρα, δεν πρέπει να υποτιμηθεί ή να μη ληφθεί σοβαρά υπόψιν.

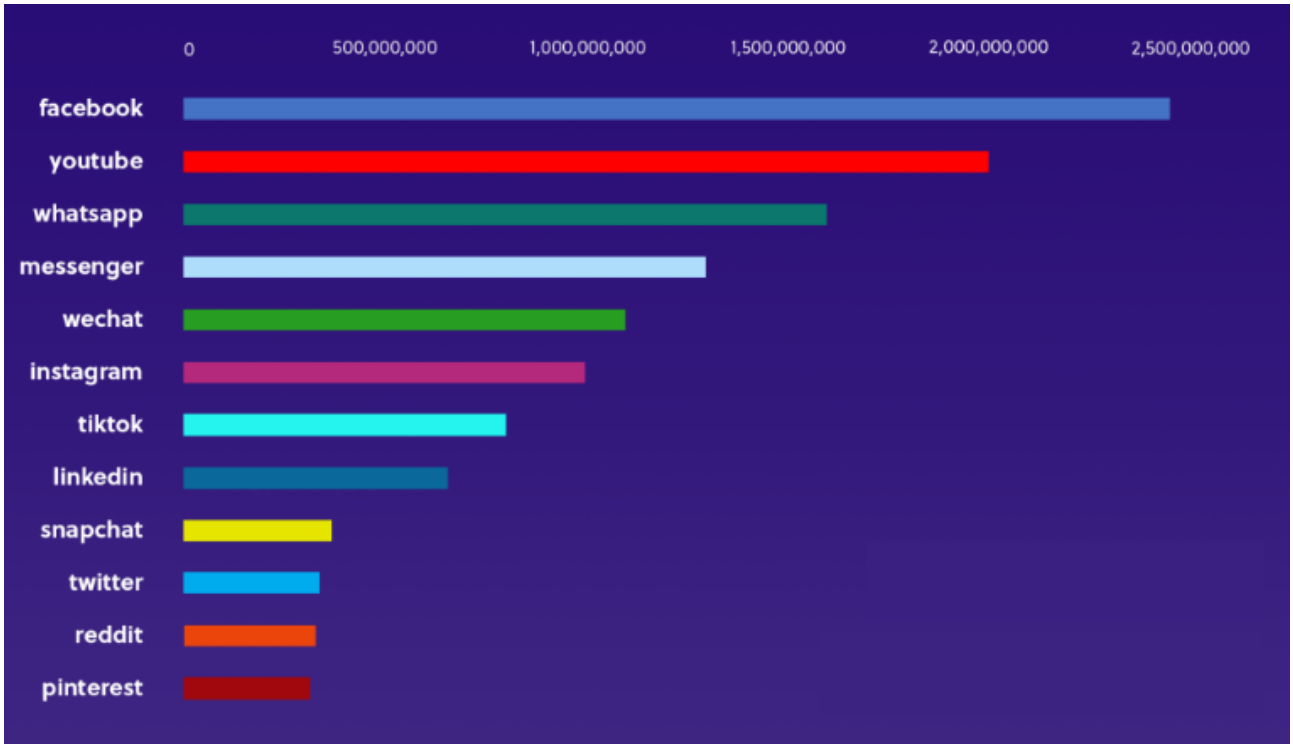
2. Να υπάρχει κάποια μορφή προεπεξεργασίας των δεδομένων ή ανοχή του μοντέλου στον θόρυβο που πιθανώς να υπάρχει στα δεδομένα. Στα δεδομένα, κυρίως στη μορφή κειμένου, συναντώνται συχνά σημαντικά προβλήματα που σχετίζονται με το λεκτικό περιεχόμενο καθώς και το συντακτικό μέρος του κειμένου. Οι παραπάνω παράμετροι δύναται να συντελέσουν καθοριστικό ρόλο στην ανάλυση και σε ορισμένες περιπτώσεις, μάλιστα, να επηρεάσουν αρνητικά το αποτέλεσμα.

Συμπερασματικά, η εξόρυξη γνώσης μέσω των ποικίλων μέσων κοινωνικής δικτύωσης, δύναται να βοηθήσει επιχειρήσεις και οργανισμούς να εφαρμόσουν εξειδικευμένο και προσωποποιημένο μάρκετινγκ, να μελετήσουν τη συμπεριφορά των χρηστών, να αποκτήσουν ενδείξεις σχετικά με την κοινωνική δομή, να ανιχνεύουν και να προλαμβάνουν ανεπιθύμητα γεγονότα [9].

## 1.3 Το Twitter

Ένα αναπόσπαστο κομμάτι της ανθρώπινης καθημερινότητας είναι η επικοινωνία και η αλληλεπίδραση με άλλους ανθρώπους. Ο άνθρωπος όντας ένα κοινωνικό ον, θέλει να ακούσει, να ακουστεί, να εκφράσει απόψεις και να συμμετέχει σε διαλόγους αναφορικά με τα ενδιαφέροντά του. Το Twitter όπως και άλλα διαδικτυακά μέσα κοινωνικής δικτύωσης (Σχήμα 1.4) αποτελεί ένα ευρέως διαδεδομένο, εύχρηστο μέσο όπου η τεχνολογία μπορεί να ικανοποιήσει στο έπακρο την ανάγκη των ανθρώπων για επικοινωνία. Δεν πρέπει όμως να παρομοιαστεί με μια δωρεάν, παγκόσμιας εμβέλειας, πλατφόρμα ανταλλαγής μηνυμάτων.

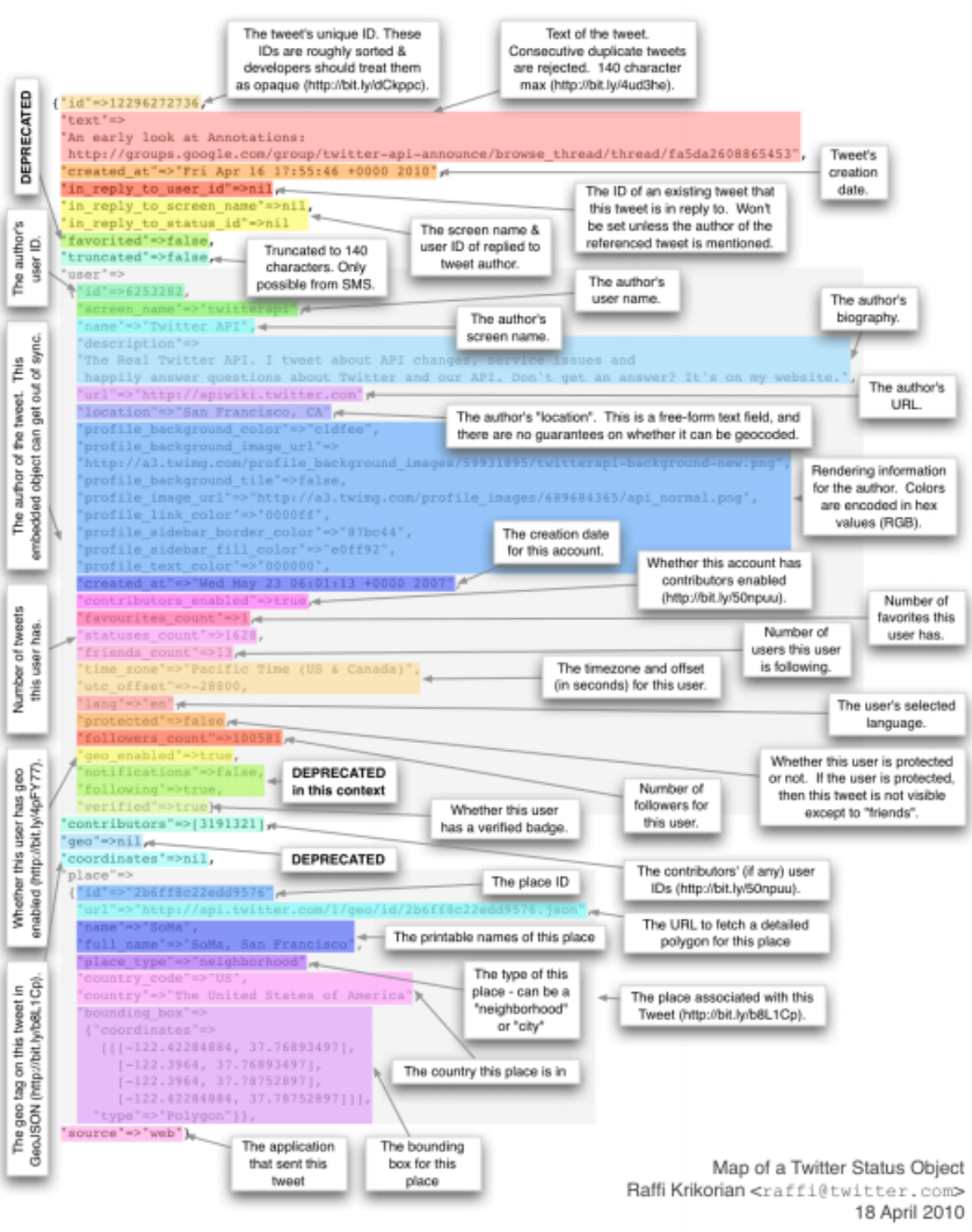
Το Twitter είναι μια ιστοσελίδα κοινωνικής δικτύωσης που δημιουργήθηκε το Μάρτιο του 2006 από τον Jack Dorsey και έγινε σύντομα δημοφιλής απαριθμώντας εκατοντάδες εκατομμύρια ενεργούς χρήστες. Το βασικό γνώρισμα του Twitter είναι η ανάρτηση tweets από τους χρήστες. Με τον όρο tweet δηλώνουμε σύντομα μηνύματα, μέχρι 140 χαρακτήρες, τα οποία μπορούν να αναγνωστούν και από μη εγγεγραμμένους στην ιστοσελίδα. Ένα tweet μπορεί να περιλαμβάνει απλό κείμενο, εικόνες, URLs, αναφορές σε άλλους χρήστες και hashtags. Η αναφορά σε άλλο χρήστη καλείται mention και έχει τη μορφή **@**ακολουθούμενο από το όνομα χρήστη (πχ. **@**a_hasap). Τα hashtags είναι λέξεις που ο χρήστης τους βάζει το πρόθεμα # (πχ. #coronavirus), συνήθως χρησιμοποιούνται για τα επίκαιρα θέματα. Μεγάλη σημασία σε ένα tweet, πέρα από το κείμενο που έχει αναρτηθεί, έχουν οι δύο ακόμα συνιστώσες του, οι οντότητες και οι θέσεις. Με τον όρο *οντότητα* (entity) εννοούμε αναφορές άλλων χρηστών, hashtags και πολυμέσα, ενώ με τον όρο *θέσεις* αναφερόμαστε σε πραγματικές τοποθεσίες που αναφέρονται σε ένα tweet ή την τοποθεσία στην οποία αναρτήθηκε το tweet. Η δομή ενός tweet αναπαρίσταται στο Σχήμα 1.5




Σχήμα 1.4: Ενεργοί χρήστες στα μέσα κοινωνικής δικτύωσης, παγκοσμίως (2020)

Ένας χρήστης στο Twitter μπορεί να ακολουθεί ή να ακολουθείται από κάποιον άλλο χωρίς να απαιτείται η αμφίδρομη σχέση, να απαντήσει σε tweet άλλου χρήστη, να κοινοποιήσει το tweet κάποιου άλλου (retweet) ή να επισημάνει κάποιο tweet ως αγαπημένο. Τέλος πρέπει να αναφέρουμε τα χρονοδιαγράμματα (timelines) που είναι χρονολογικά ταξινομημένες συλλογές από tweets. Τα πιο σημαντικά χρονοδιαγράμματα είναι τα (1) **home timeline:** περιέχει όλα τα tweets των χρηστών που ακολουθεί κάποιος χρήστης (2) **user timeline:** περιέχει όλα τα tweets που έχει αναρτήσει ένας χρήστης.

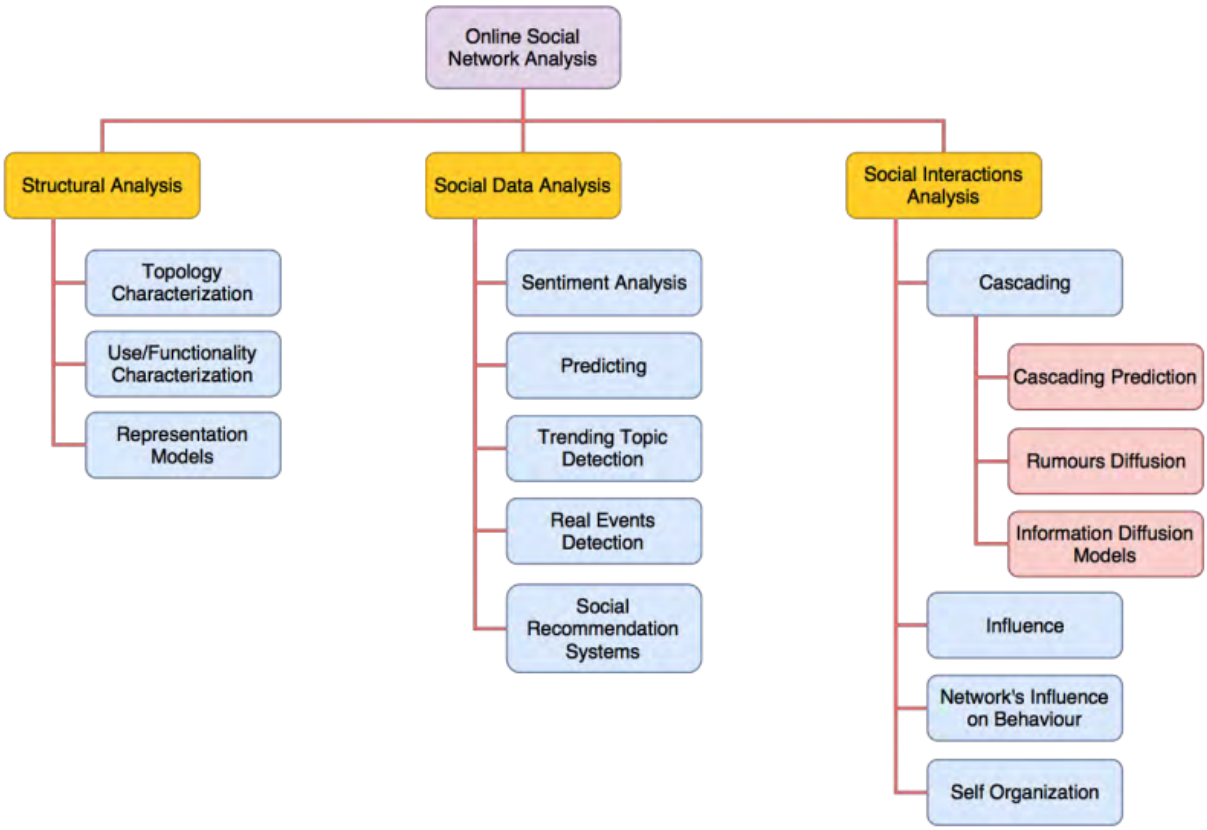
Θα μπορούσαμε να πούμε ότι το Twitter έχει δομηθεί ώστε να «εκμεταλεύεται» την ανθρώπινη περιέργεια. Κατά το πλαίσιο λειτουργίας του, κάθε χρήστης μπορεί να παρακολουθεί τις αναρτήσεις οποιουδήποτε άλλου έχει επιλέξει να ακολουθεί, δεδομένης όμως της επιλογής αποκλεισμού του ακολουθούμενο προς τον ακόλουθο. Δημιουργούνται έτσι γραφήματα ενδιαφέροντος που παρέχουν ευελιξία στο πεδίο της εξόρυξης δεδομένων, προκειμένου να χρησιμοποιηθούν εφαρμογές μηχανικής μάθησης (machine learning) για διάφορους σκοπούς.




Σχήμα 1.5: Η ανατομία ενός tweet

## 1.4 Ανάλυση Κοινωνικών Δικτύων

Το Twitter είναι η εικονική προσωμοίωση ενός κοινωνικού δικτύου και παρέχει ένα μεγάλο φάσμα τομέων προς έρευνα και ανάλυση. Οι Kurka, Godoy και Von Zuben [10] πρότειναν την κατηγοριοποίηση των τομέων έρευνας (Σχήμα 1.6) την κατηγοριοποίηση των τομέων έρευνας στους τρεις βασικούς άξονες που ακολουθούν.




Σχήμα 1.6: Τομείς έρευνας κοινωνικών δικτύων

1. **Δομική ανάλυση**
   Αντικείμενο της είναι οι έρευνες που σχετίζονται με τη δομή και το καθεστός λειτουργίας των μέσων κοινωνικής δικτύωσης.

2. **Ανάλυση κοινωνικών δεδομένων**
   Σε αυτόν τον κλάδο οι ερευνητές στρέφονται προς το περιεχόμενο που παράγεται μέσα σε ένα κοινωνικό δίκτυο. Ειδικότερα, γίνεται επεξεργασία του αδόμητου περιεχομένου που παράγεται από τους χρήστες προκειμένου να υλοποιηθεί μια πληθώρα εφαρμογών. Στα πλαίσια αυτής της εργασίας θα ασχοληθούμε με την ανάλυση συναισθήματος (sentiment analysis)

3. **Ανάλυση της κοινωνικής αλληλεπίδρασης**
   Είναι ίσως ο πιο δύσκολος από τους τρεις κλάδους διότι απαιτεί την επεξεργασία και την αξιοποίηση τεράστιου όγκου πληροφορίας προκειμένου να ερμηνευθούν οι συνδέσεις μεταξύ των χρηστών του κοινωνικού δικτύου. Στόχος των ερευνητών είναι ν᾽ ανιχνεύσουν πως τόσο οι αλληλεπιδράσεις μεταξύ των χρηστών, όσο και η ατομική συμπεριφορά κάθε χρήστη, μπορούν να χρησιμοποιηθούν ώστε να εξαχθούν χρήσιμα συμπεράσματα ως προς τις τάσεις και τις επιρροές σε μια κοινωνία.

Τα κριτήρια που χρησιμοποιεί κάποιος (φυσικό πρόσωπο ή εταιρία) για τη λήψη αποφάσεων επηρεάζονται από τις απόψεις των ανθρώπων στο στενό περιβάλλον και στο ευρύτερο διαδικτυακό. Στη δεύτερη περίπτωση ο διαθέσιμος όγκος πληροφοριών είναι τεράστιος, γεγονός που έχει στρέψει την επιστημονική κοινότητα στην υλοποίηση τεχνικών που να «ανακαλύπτει» γνώση μέσα από αυτή την ακατέργαστη πληροφορία.

### 1.4.1 Ανάλυση Δεδομένων Twitter

Το Twitter ανήκει στην κατηγορία των microblogs [11]. Η ευκολία με την οποία δημιουργούνται και διαμοιράζονται τα δεδομένα μέσω του Twitter, σε συνδυασμό με την εξάπλωσή του ανά την υφήλιο, το έχει καταστήσει μία από τις δημοφιλέστερες πλατφόρμες κοινωνικής δικτύωσης και συνεπώς, μια ανεξάντλητη πηγή πληροφοριών κατάλληλη για ανάλυση και εξαγωγή χρήσιμων συμπερασμάτων. Τα προαναφερθέντα πλεονεκτήματα του Twitter δεν είναι

από μόνα τους ικανά να το καταστήσουν μια δεξαμενή άντλησης πληροφορίας διότι σχεδόν όλα τα μέσα κοινωνικής δικτύωσης εκμεταλλεύονται τα παραπάνω προτερήματα. Στο σημείο αυτό, θα παραθέσουμε τους παράγοντες που το έχουν ξεχωρίσει και επιπλέον, το καθιστούν ιδιαίτερα πρόσφορο για ανάλυση:

(1) η άμεση και ελεύθερη πρόσβαση στην πηγή μέσω της διεπαφής και του λεγόμενου streaming API, το οποίο είναι το μέσο εκείνο που επιτρέπει την αλληλεπίδραση μεταξύ προγραμμάτων ηλεκτρονικών υπολογιστών και υπηρεσιών διαδικτύου. Πέρα από τις πολλές δυνατότητες που παρέχει στους προγραμματιστές, τους επιτρέπει να έχουν πρόσβαση σε ολόκληρο το γράφο του δικτύου. Επιπλέον η πρόσβαση στα δεδομένα επιτυγχάνεται και με τη χρήση κάποιου προγράμματος ανίχνευσης ιστού (web crawler). Αν και το Twitter επιβάλλει περιορισμούς στον αριθμό των tweets που δύναται κάποιος να συλλέξει ανά ημέρα, είναι δυνατή η συλλογή ενός επαρκούς αριθμού δεδομένων.

(2) τα παραγόμενα tweets έχουν σαν χαρακτηριστικό μια χρονοσφραγίδα (timestamp) που επιτρέπει τη διάκριση των γεγονότων σε χρονολογική σειρά.

(3) ο περιορισμός των χαρακτήρων (140) ανά ανάρτηση δημιουργεί δεδομένα μικρά σε όγκο, που είναι ευκολότερο για μια εφαρμογή να τα επεξεργαστεί.

Από την άλλη μεριά, η ενασχόληση με το Twitter επιβάλλει στους αναλυτές να υπερκεράσουν κάποιες δυσκολίες που συνοψίζονται στις παρακάτω κατηγορίες [12].

> **Λεξιλόγιο.** Το κείμενο των tweets περιλαμβάνει αργκό, συντομογραφίες, αρκτικόλεξα αλλά και παραλλαγές λέξεων με σκοπό να δοθεί έμφαση. Επίσης, προκειμένου να εξοικονομηθεί πλήθος χαρακτήρων, υπάρχει περίπτωση ν᾽ αλλοιώνονται κάποιες λέξεις.
>
> **Θόρυβος.** Η δημιουργία γραπτού λόγου είναι συνυφασμένη με την παρουσία ορθογραφικών λαθών και τη χρήση ακατανόητων εκφράσεων. Τα παραπάνω συνιστούν θόρυβο.
>
> **Πολυγλωσσικό περιεχόμενο.** Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης παγκόσμιας εμβέλειας με αποτέλεσμα το περιεχόμενο των αναρτήσεων να μην είναι μόνο σε μια γλώσσα. Έτσι τεχνικές που λειτουργούν αποκλειστικά με συγκεκριμένες γλώσσες, καθιστώνται μη εφαρμόσιμες.

Η σημασία του Twitter για την άντληση πληροφοριών είναι καταλυτική. Αυτή αποδεικνύεται από το γεγονός ότι πολλές και μεγάλες εταιρίες επενδύουν κεφάλαιο και ανθρώπινο δυναμικό στην εξόρυξη γνώσης από αυτό το μέσον. Έτσι, δύνανται να βελτιώσουν τα προσφερόμενα προς τους καταναλωτές μέσω της ανάλυσης της γνώμης του αγοραστικού κοινού που χρησιμοποιούν το Twitter.

# 2. Ανάλυση Συναισθήματος

## 2.1 Εισαγωγή

Η ευκολία και η αμεσότητα με την οποία οι χρήστες δημιουργούν περιεχόμενο στο διαδίκτυο, τους επιτρέπει να το χρησιμοποιήσουν προκειμένου να εκφράσουν τη γνώμη τους, ντα συναισθήματά τους και να μοιραστούν τις εμπειρίες τους. Έχει, έτσι, δημιουργηθεί ο επιστημονικός κλάδος της **Επεξεργασίας Φυσικής Γλώσσας** (Natural Language Processing - NLP) που πραγματεύεται την ανάλυση απόψεων ντων χρηστών σε σχέση με άλλους χρήστες, προϊόντα - υπηρεσίες και γεγονότα [13]. Είναι ένας τομέας όπου συναντιόνται τα πεδία της επιστήμης Η/Υ, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και σχετίζεται με την ικανότητα των υπολογιστών να κατανοήσουν τη φυσική γλώσσα και να μπορούν μέσω κάποιας μορφής επεξεργασίας, να εξάγουν συμπεράσματα από εισόδους τέτοιας μορφής. Οι σύγχρονοι NLP αλγόριθμοι που βασίζονται στη στατιστική Μηχανική μάθηση, σε αντίθεση με τους αλγορίθμους της πρώτης εποχής όπου παρήγαγαν συστήματα αποτελούμενα μόνο από κανόνες της μορφής «if-then», μπορούν να εκφράσουν σχετική βεβαιότητα πολλών πιθανών διαφορετικών απαντήσεων αντί μόνο μίας, παράγοντας έτσι αξιόπιστα αποτελέσματα [14].


Σχήμα 2.1: Συνιστώσες της Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος, η οποία συναντάται σε μεγάλο βαθμό στην ταξινόμηση κειμένων, αφορά την ανάλυση της πολικότητας του συναισθήματος που περιέχεται σε κάποιο κείμενο με χρήση μόνο γλωσσολογικών χαρακτηριστικών του περιεχομένου του κειμένου. Δηλαδή, μετά την ανάλυση του κειμένου, αποδίδεται μια συναισθηματική κλίση είτε σε μια κλίμακα θετικού - αρνητικού, είτε ως κάποιο συγκεκριμένο συναίσθημα. Αξίζει να σημειωθεί πως η Ανάλυση Συναισθήματος είναι εφαρμόσιμη και σε δεδομένα διαφορετικού τύπου από οτι το κείμενο, όπως για παράδειγμα στη μουσική. Ο συγκεκριμένος τομέας έχει μεγάλη απήχηση στις επιχειρήσεις και την πολιτική, γεγονός που αντικατοπτρίζεται στη χρήση εργαλείων ανάλυσης συναισθήματος στις ιστοσελίδες των επιχειρήσεων καθώς και στα ιστολόγια πολιτικών προσώπων ή παρατάξεων. Η Ανάλυση Συναισθήματος μπορεί να πραγματοποιηθεί υπό τρεις οπτικές [15]:

1. **Επίπεδο εγγράφου.** Υπό αυτή την οπτική θεωρούμε ολόκληρο το κείμενο σαν μία οντότητα και προσπαθούμε να αναγνωρίσουμε το γενικό συναίσθημα που απορρέει από αυτό. Αυτή η μορφή ανάλυσης είναι εύκολο να υλοποιηθεί, όμως χάνονται σημαντικές πληροφορίες.

2. **Επίπεδο φράσης.** Σε κάθε πρόταση ενός εγγράφου προσδίδεται ένα συναίσθημα. Η συγκεκριμένη περίπτωση είναι αρκετά πιο δύσκολη από την πρώτη αφού τα δεδομένα είναι λιγότερα, όμως μπορούν να αξιοποιηθούν περισσότερες πληροφορίες. Αυτού του είδους ανάλυσης, συναντάται αρκετά συχνά στο πρόβλημα της αναγνώρισης υποκειμενικότητας (subjectivity detection).

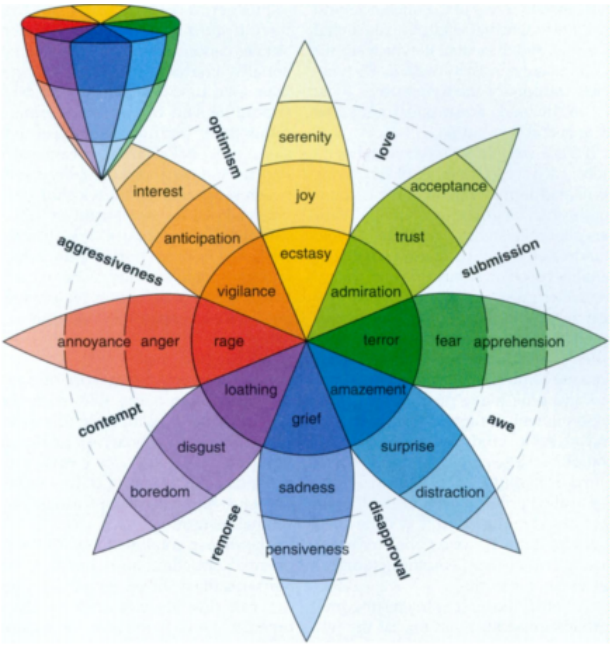
3. **Επίπεδο πτυχής.** Είναι το πιο δύσκολο επίπεδο ανάλυσης αφού η πρόβλεψη για το συναίσθημα προϋποθέτει τον προσδιορισμό των πτυχών. Η δυσκολία έγκειται στο γεγονός ότι μπορεί μία πτυχή να αναφέρεται

έμμεσα στο κείμενο ή ορισμένες πτυχές να αναφέρονται με παραπλήσιο τρόπο καθιστώντας δύσκολο το να αναγνωριστεί ότι αποτελούν την ίδια πτυχή.

Αν και ο διαχωρισμός δεν είναι πολύ εμφανής, ο τομέας της Συναισθηματικής Ανάλυσης χωρίζεται σε δύο κατηγορίες ως προς τον τρόπο αξιολόγησης του συναισθήματος [16]. Στην κατηγορία της **Εξόρυξη Γνώμης** ανήκουν όλες οι μέθοδοι και τεχνικές που για δεδομένο κείμενο αποσκοπούν στο να καταλάβουν τις απόψεις του γράφοντος και συνήθως τις ομαδοποιούν σε τρεις κλάσεις συναισθήματος (θετικό, ουδέτερο και αρνητικό). Σε αυτή την κατηγορία, η ετικέτα που αποδίδεται στο κείμενο είναι μοναδική. Στη δεύτερη κατηγορία, αυτή της **Ανάλυσης Συναισθήματος** (Affective Analysis), οι μέθοδοι που χρησιμοποιούνται, προσπαθούν να καθορίσουν το ακριβές συναίσθημα που προκύπτει από το δεδομένο κείμενο. Οι περισσότερες μέθοδοι αυτής της κατηγορίας μπορούν να ανακαλύψουν τα παρακάτω 6 βασικά συναισθήματα [17]:

| | | |
|---|---|---|
| Χαρά | Λύπη | Φόβος |
| Θυμός | Αγάπη | Μίσος |

Πρέπει να τονιστεί πως υπάρχουν και τεχνικές που έχουν τη δυνατότητα να χρησιμοποιήσουν παραπάνω από τα προαναφερθέντα συναισθήματα. Σε αντίθεση με την πρώτη κατηγορία, εδώ δεν επιλέγεται μονοσήμαντα ένα από τα εμπλεκόμενα συναισθήματα, αλλά δημιουργείται για κάθε κείμενο ένα διάνυσμα με δεκαδικούς αριθμούς, έναν για κάθε συναίσθημα. Προφανώς ο μεγαλύτερος αριθμός αντιστοιχεί στο συναίσθημα που επικρατεί. Στο Σχήμα 2.2 παρουσιάζεται ένα διάγραμμα που αναπαριστά την ύπαρξη αλληλεπίδρασης και αλληλοεξάρτησης μεταξύ των συναισθημάτων (Circumplex του Plutchik [18]). Εύκολα παρατηρεί κανείς ότι κινούμενοι από το κέντρο προς τα έξω, τα συναισθήματα που συναντούμε γίνονται όλο και πιο ηπιότερα. Οι εξαρτήσεις που δημιουργούνται μεταξύ των συναισθημάτων, μας οδηγούν στο συμπέρασμα πως δεν μπορούμε να ταυτίσουμε μονοσήμαντα μια πρόταση με ένα συναίσθημα.

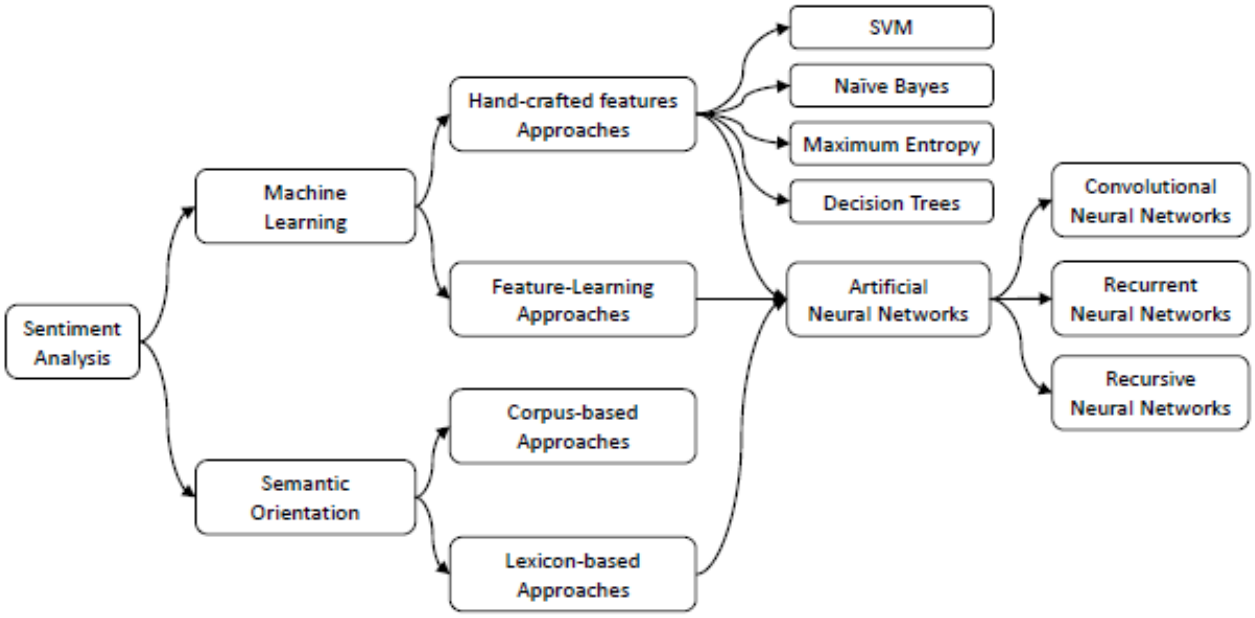



Σχήμα 2.2: Αλληλεπιδράσεις συναισθημάτων

Η επιλογή μιας τεχνικής για την ανάλυση συναισθήματος ενός γραπτού κειμένου επηρεάζεται από πολλούς παράγοντες όπως η ύπαρξη ή όχι ετικετοποιημένων δεδομένων, η γλώσσα του κειμένου και η πλατφόρμα στην οποία εφαρμόζεται. Έτσι έχουν συσταθεί τρεις διακριτές κύριες κατηγορίες για την ανάλυση συναισθήματος (επιβλεπόμενη, μη επιβλεπόμενη, με λεξικό). Προφανώς κάθε τεχνική έχει υλοποιηθεί για να ανταπεξέρχεται επιτυχώς σε συγκεκριμένες συνθήκες ανά πρόβλημα και ως εκ τούτου δεν υπάρχει κάποια που να αποτελεί πανάκεια. Αξίζει να σημειωθεί πως υπάρχουν μέθοδοι που αποτελούν προϊόν συγκερασμού τεχνικών από διαφορετικές κατηγορίες. Στη συνέχεια αναλύονται οι προσεγγίσεις που χρησιμοποιούνται στον τομέα της ανάλυσης συναισθήματος.

## 2.2 Προσεγγίσεις

Οι τεχνικές που συναντάμε στην Ανάλυση Συναισθήματος χωρίζονται σε δύο κατηγορίες όπως αυτές φαίνονται στο Σχήμα 2.3. Στην πρώτη κατηγορία περιέχονται τεχνικές που κάνουν χρήση Μηχανικής Μάθησης, ενώ στη δεύτερη τεχνικές που εστιάζουν στο Σημασιολογικό Προσανατολισμό. Η κατηγοριοποίηση αυτή γίνεται βάσει του τρόπου με τον οποίο, κάθε τεχνική, διαχειρίζεται τα χαρακτηριστικά για την μοντελοποίηση της φυσικής γλώσσας.




Σχήμα 2.3: Κατηγοριοποίηση εφαρμόσιμων τεχνικών ανά προσέγγιση.

### 2.2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση, είναι ένας σχετικά νέος και ραγδαία αναπτυσσόμενος κλάδος που ενισχύθηκε σημαντικά χάριν στην εξέλιξη των υπολογιστικών πόρων που διατίθενται πλέον στο μέσο χρήστη. Επιπλέον, αποτελεί ένα μεγάλο υποσύνολο του τομέα της Τεχνητής Νοημοσύνης, η οποία ασχολείται με τη μελέτη και κατασκευή αλγορίθμων για την εξαγωγή πολύτιμων συμπερασμάτων από τα δεδομένα. Η δημιουργία και εγκαθίδρυση αυτού του τομέα είναι αποτέλεσμα της προσπάθειας των ερευνητών να δημιουργήσουν πολύπλοκα υπολογιστικά συστήματα που θα είναι σε θέση να μάθουν και κατ΄ επέκταση να μιμηθούν τον ίδιο τον άνθρωπο και τους τρόπους με τους οποίους μαθαίνει. Ακόμα, σημαντική παράμετρος είναι η αξιοποίηση της παραγόμενης γνώσης για την εξαγωγή περαιτέρω χρήσιμων συμπεράσματων [19]. Ένας γενικός ορισμός της Μηχανικής Μάθησης δίνεται από τον [20]:

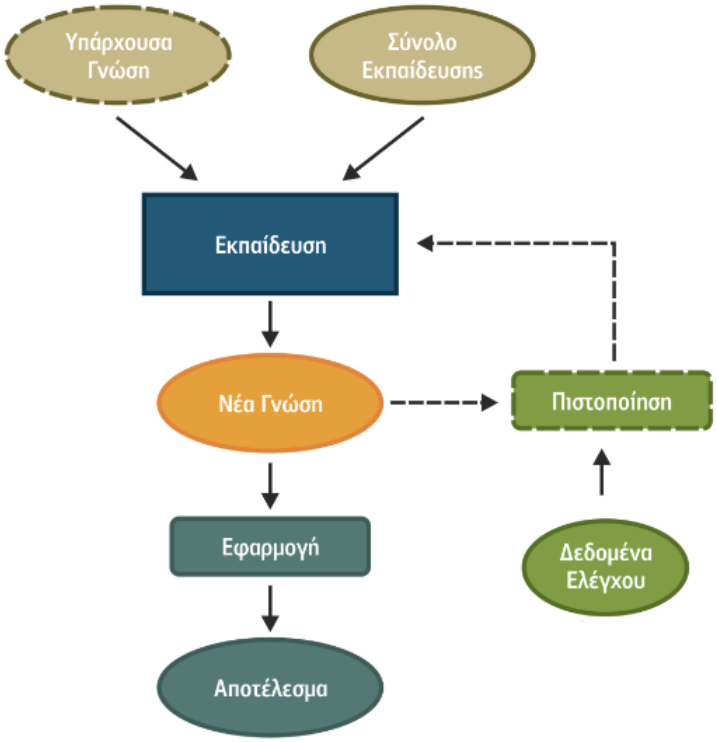
**Ορισμός 2.1.** *«Ένα πρόγραμμα Η/Υ δύναται να μαθαίνει από την εμπειρία, έστω $E$, αναφορικά με μια κλάση εργασιών, έστω $T$, και ένα μέτρο απόδοσης, έστω $P$, αν η απόδοσή του σχετικά με το $T$, όπως μετριέται από το $P$, βελτιώνεται μέσω της εμπειρίας $E$.»*

Έχει συγκεντρώσει μεγάλο μέρος της ερευνητικής δραστηριότητας διότι εκτός από την ταχύτητα αρκετών μεθόδων που περιλαμβάνει, οι μέθοδοι, μέσω των μοντέλων που δημιουργούνται, είναι ικανές να βρίσκουν και ν΄ αξιοποιούν κρυμμένα μοτίβα και πολύπλοκες αλληλουχίες που βρίσκονται στα δεδομένα.

Ως κλάδος της τεχνητής νοημοσύνης, η Μηχανική Μάθηση καθιστά εφικτή την κατασκευή προσαρμόσιμων προγραμμάτων υπολογιστών, ικανά να βελτιώνουν τη συμπεριφορά τους, ως προς την εργασία που επιτελούν, αξιοποιώντας την εμπειρία τους. Με βάση τα όσα αναφέραμε, δίνουμε τον παρακάτω εναλλακτικό ορισμό για το τί εστί Μηχανική Μάθηση [21]:

**Ορισμός 2.2.** *«Μηχανική Μάθηση ορίζεται ως η ικανότητα ενός Η/Υ να παράγει μοντέλα ή πρότυπα στηριζόμενος σε σύνολα δεδομένων.»*

Ο κύκλος ζωής ενός μοντέλου Μηχανικής Μάθησης παρουσιάζεται στο Σχήμα 2.4.




Σχήμα 2.4: Φάσεις Μηχανικής Μάθησης

Η Μηχανική Μάθηση δύναται να διαχειριστεί πολύπλοκες εργασίες που είναι αδύνατο να αντιμετωπιστούν μόνο από ένα και μοναδικό πρόγραμμα υπολογιστή, όπως οι κύριες δραστηριότητες του ανθρώπου (κατανόηση εικόνας, αναγνώριση φωνής) ή ακόμη και εργασίες που δεν μπορούν να εκτελεστούν από τον άνθρωπο λόγω της ύπαρξης δεδομένων μεγάλου όγκου και διάστασης. Κλείνοντας, αξίζει να αναφερθεί οτι για κάθε πρόβλημα στο χώρο της Μηχανικής Μάθησης, συναντάται ένας ξεχωριστός, ίσως και καταλληλότερος, τρόπος μάθησης και για κάθε τέτοιο τρόπο μπορεί να σχεδιαστεί ή υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να ανταπεξέλθει στην αντιμετώπιση του προβλήματος. Ο παραπάνω ισχυρισμός προκύπτει από το γεγονός ότι στα περισσότερα προβλήματα Μηχανικής Μάθησης χρησιμοποιείται κάποιος αλγόριθμος βελτιστοποίησης και με βάση το [22] δεν υπάρχει κάποιος που να υπερέχει όλων των υπολοίπων.

Ένα σύστημα Μηχανικής Μάθησης αποτελείται από δύο τμήματα [23]:

1. **Καθορισμός χρήσιμων χαρακτηριστικών.** Αναδεικνύονται ή/και δημιουργούνται τα πιο αντιπροσωπευτικά και χρήσιμα χαρακτηριστικά για τη μοντελοποίηση των δεδομένων. Η παραπάνω διαδικασία απαιτεί τα παρακάτω βήματα:
   - Εξαγωγή των αρχικών χαρακτηριστικών
   - Επιλογή των σημαντικότερων χαρακτηριστικών
   - Παραγωγή νέων από τα αρχικά χαρακτηριστικά
   - Απόδοση βαρών στα χαρακτηριστικά, βάσει της σημαντικότητάς τους
2. **Εφαρμογή αλγορίθμου Μηχανικής Μάθησης.** Αξιοποιώντας τα χαρακτηριστικά που εξίχθησαν στο προηγούμενο τμήμα, εφαρμόζεται ένα στατιστικό μοντέλο που αποτελείται από έναν ή συνδυασμό πολλών αλγορίθμων Μηχανικής Μάθησης.

Οι τεχνικές Μηχανικής Μάθησης διακρίνονται σε δύο βασικές υποκατηγορίες, οι οποίες περιγράφονται στη συνέχεια.

| Επιβλεπόμενης μάθηση (supervised learning) | Μη επιβλεπόμενης μάθηση (unsupervised learning) |
|---|---|

### Αλγόριθμοι επιβλεπόμενης μάθησης

Σε αυτή την κατηγορία υπάρχει ένα σύνολο δεδομένων που το κείμενο κάθε εγγραφής αντιστοιχεί σε μια προκαθορισμένη κλάση, συνήθως οι κλάσεις είναι θετικά, αρνητικά και ουδέτερα συναισθήματα. Το διαθέσιμο σύνολο δεδομένων χρησιμοποιείται ώστε να εκπαιδευτούν τα χρησιμοποιούμενα μοντέλα, μέσω κάποιων αλγορίθμων ταξινόμησης. Μετά το πέρας της εκπαίδευσης, τα μοντέλα που έχουν δημιουργηθεί, θα μπορούν να προβλέπουν την κλάση οποιασδήποτε νέας εγγραφής.

Ειδικότερα, έστω ότι έχουμε ένα σύνολο $D = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ όπου το $x_i$ είναι το κείμενο κάθε εγγραφής και το $y_i$ δηλώνει την κατηγορία που ανήκει το αντίστοιχο κείμενο, αναζητούμε μια απεικόνιση $f : X \to Y$ όπου κάθε στοιχείο του συνόλου $D$ θα εκχωρείται στη σωστή κατηγορία. Το βασικό πλεονέκτημα των μεθόδων επιβλεπόμενης μάθησης είναι η δημιουργία προσαρμόσιμων μοντέλων που ανταπεξέρχονται επιτυχώς σε συγκεκριμένους σκοπούς και πλαίσια. Από την άλλη μεριά, η αναγκαιότητα ύπαρξης ετικετών στα σύνολα δεδομένων, που σε κάποιες περιπτώσεις είναι δαπανηρή ή ακόμα και αδύνατη, αποτελεί ανασταλτικό παράγοντα στην επιλογή αυτών των μεθόδων [24].

Επίλυση προβλήματος επιβλεπόμενης μάθησης

1. Προσδιορισμός του συνόλου δεδομένων που θα χρησιμοποιηθεί, όπως έχει συλλεχθεί είτε από ειδικούς είτε από μετρήσεις.
2. Καθορισμός του αλγορίθμου που θα χρησιμοποιηθεί. Συνήθως αυτή η επιλογή γίνεται μετά από επάλληλες δοκιμές.
3. Καθορισμός του συνόλου εκπαίδευσης και ελέγχου του αλγορίθμου.
4. Καθορισμός παραμέτρων ελέγχου για τερματισμό της εκπαίδευσης.
5. Εκπαίδευση του μοντέλου.
6. Αξιολόγηση του μοντέλου με βάση το σύνολο ελέγχου.

Πρέπει να επισημανθεί πως από κοινού τα σύνολα εκπαίδευσης και ελέγχου είναι υποσύνολα του συνόλου δεδομένων του προβλήματος και ξένα μεταξύ τους. Στη συνέχεια, παρουσιάζεται μια βιβλιογραφική ανασκόπηση των προσεγγίσεων επιβλεπόμενης μάθησης για την ανάλυση συναισθήματος.

Το Twitter αποτελεί μια ανεξάντλητη πηγή πληροφοριών για ερευνητές με επιχειρηματικούς και ακαδημαϊκούς προσανατολισμούς. Στην εργασία [25] προτάθηκε μια προσέγγιση μείωσης χαρακτηριστικών, συνδυάζοντας την αναπαράσταση n-gram (Σελ. 21) και τη στατιστική ανάλυση για την ανάπτυξη ενός λεξικού ανάλυσης συναισθημάτων, αποκλειστικά για τη χρήση του σε εργασίες που επεξεργάζονται δεδομένα από το Twitter. Προκειμένου να ενισχυθεί η ισχύς του μικρού αρχικού λεξικού, περιέχει 187 χαρακτηριστικά, εισάγονται όροι από tweets αναφορικά με το θέμα για το οποίο θα εφαρμοστεί ανάλυση συναισθήματος. Αρχικά συγκρίθηκαν μοντέλα SVM [26] με τη διαφοροποίηση ως προς το λεξικό που χρησιμοποίησαν και διαπιστώθηκε υπεροχή των μοντέλων με χρήση μειωμένου λεξικού έναντι αυτών που χρησιμοποιούσαν παραδοσιακό λεξικό συναισθημάτων. Στη συνέχεια, χρησιμοποιώντας το μειωμένο λεξικό, σύγκριναν μοντέλα SVM με το δυναμικό τεχνητό νευρωνικό δίκτυο DAN2 [27] και απέδειξαν ότι το δεύτερο παράγει πιο ακριβή αποτελέσματα ταξινόμησης συναισθημάτων από το πρώτο.

Η ραγδαία αύξηση των διαδικτυακών αγορών έχει αυξήσει τον αριθμό των πελατών σε ιστότοπους ηλεκτρονικού εμπορίου, δημιουργώντας ένα μεγάλο όγκο αξιοποιήσιμων δεδομένων που προέκυψε από τις αξιολογήσεις που αφήνουν οι χρήστες για κάθε προϊόν ή υπηρεσία. Προκειμένου να αξιοποιήσουν τα διαθέσιμα δεδομένα, οι συγγραφείς στο [28] πρότειναν έναν αλγόριθμο που κατατάσσει τις κριτικές, ανάλογα με το συναίσθημα που εκφράζουν, σε θετικές ή αρνητικές. Αρχικά μέσα γίνεται εξαγωγή χαρακτηριστικών ώστε να προσδιοριστούν τα πιο σημαντικά χαρακτηριστικά ενός προϊόντος και στη συνέχεια, χρησιμοποιώντας έναν ταξινομητή πολικότητας, καθορίζεται το συναίσθημα των κριτικών σε σχέση με τα χαρακτηριστικά του προϊόντος που επισημάνθηκαν ως σημαντικά. Τα πειραματικά αποτελέσματα και η αξιολόγηση έδειξαν ότι ο ταξινομητής είναι αποδοτικός, ανταγωνιστικός και επιτυγχάνει ακρίβεια της τάξεως του 79.67%.

### Αλγόριθμοι μη επιβλεπόμενης μάθησης

Η μη επιβλεπόμενη μάθηση, στενά συνδεδεμένη με το πρόβλημα της εκτίμησης πυκνότητας στην στατιστική, είναι μια διεργασία Μηχανικής Μάθησης κατά την οποία εξάγεται μια συνάρτηση για να περιγράψει κρυφές δομές από δεδομένα χωρίς ετικέτες. Αφού το διαθέσιμο σύνολο δεδομένων αποτελείται από πρότυπα άγνωστης κατηγορίας, δεν υπάρχει σφάλμα και ως εκ τούτου, ούτε κάποιο σήμα αξιολόγησης ώστε να γίνει εκτίμηση της πιθανής λύσης.

Η συσταδοποίηση είναι μια από τις τεχνικές που χρησιμοποιούνται ώστε να αναδειχτούν εξαρτήσεις μεταξύ των χαρακτηριστικών του συνόλου δεδομένων του εκάστοτε προβλήματος. Σύμφωνα με αυτή την τεχνική, από ένα μη ετικετοποιημένο σύνολο, δημιουργούνται συστάδες (clusters) των δεδομένων εισόδου, δηλαδή ομάδες που έχουν

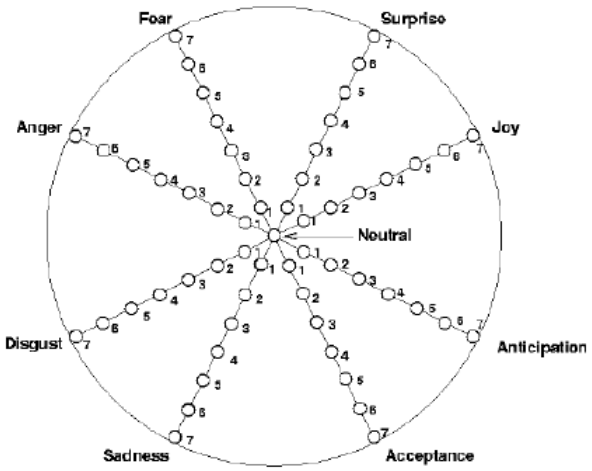
κάποιες κοινές ιδιότητες ή γνωρίσματα. Η συσταδοποίηση γίνεται βασισμένη σε ένα σύνολο από κανόνες που δίνει ο χρήστης ή υπάρχουν προεπιλεγμένοι στην εκάστοτε τεχνική. Οι αλγόριθμοι συσταδοποίησης δημιουργούν ομάδες εκμεταλλευόμενοι την ομοιότητα που υπάρχει μεταξύ των προτύπων του συνόλου δεδομένων. Στόχος τους είναι η δημιουργία ομάδων που τα πρότυπα εντός της ίδιας ομάδας να έχουν τη μεγαλύτερη δυνατή ομοιότητα μεταξύ τους και να διαφέρουν από τα πρότυπα άλλων ομάδων.

Μετά τον καταρτισμό των ομάδων, ο αναλυτής μπορεί να τις θεωρήσει ως τις κατηγορίες του προβλήματος και να ερμηνεύσει τους κανόνες που τις περιγράφουν. Οι παράγοντες [29] που επηρεάζουν τη συσταδοποίηση είναι το μέτρο ομοιότητας, ο αριθμός προτύπων ανά συστάδα και η μορφή των συστάδων. Εύκολα συμπεραίνει κανείς πως δεν υπάρχει μοναδική λύση. Στη συνέχεια παρουσιάζεται μια βιβλιογραφική ανασκόπηση των προσεγγίσεων μη επιβλεπόμενης μάθησης για την ανάλυση συναισθήματος.

Ο Paragraph Vector [30] είναι ένας αλγόριθμος μηχανικής μάθησης που δεχόμενος σαν είσοδο κείμενα μεταβλητού μήκους (προτάσεις, παραγράφους, έγγραφα), εκπαιδεύεται μέσω της μεθόδου στοχαστικής κλίσης (stochastic gradient descent) ώστε να μαθαίνει αναπαραστάσεις χαρακτηριστικών σταθερού μήκους. Ο αλγόριθμος αποτελείται από 3 διακριτά μέρη: (1) την εκπαίδευση για τη λήψη διανυσμάτων λέξεων W, (2) το στάδιο συμπερασμάτων για τη λήψη διανυσμάτων παραγράφου D. (3) τη χρήση του D σε συνεργασία με έναν ταξινομητή προκειμένου να προβλεφθούν οι κλάσεις. Αξίζει να σημειωθούν τα προτερήματα του προτεινόμενου αλγορίθμου. Πέρα από το γεγονός ότι μπορεί να χρησιμοποιηθεί για μη ετικετοποιημένα δεδομένα, λαμβάνει υπόψιν τη σημασιολογία των λέξεων και διατηρεί πολλές πληροφορίες της παραγράφου, συμπεριλαμβανομένης της σειράς λέξεων.

Στην προσπάθεια τους να υπερκεράσουν τρία σημαντικά ζητήματα που δε λήφθηκαν υπόψιν στην προηγούμενη εργασία, οι συγγραφείς του [31], χρησιμοποιώντας αναδρομικούς αυτοκωδικοποιητές (autoencoders) πρότειναν ένα μοντέλο πρόβλεψης συναισθημάτων για μικρές φράσεις. Το παραγόμενο μοντέλο δέχεται σαν είσοδο διανύσματα λέξεων και προβλέπει πολύπλοκα, αλληλοσυνδεόμενα συναισθήματα. Μέσα από πειράματα απέδειξαν πως η προσέγγισή τους αποδίδει καλύτερα από τις άλλες συγκρινόμενες παρόλο που δεν χρησιμοποιήθηκαν λεξικά προκαθορισμένων συναισθημάτων ή κανόνες αλλαγής πολικότητας. Αξίζει να σημειωθεί πως η εν λόγω προσέγγιση μπορεί να λειτουργήσει με ετικετοποιημένα και μη, δεδομένα. Τέλος, αξίζει να αναφερθεί ότι η μέθοδος δεν επιτυγχάνει σημαντικά υψηλά ποσοστά ακρίβειας και εκτός αυτού, υστερούσε έναντι των άλλων, όταν στο σύνολο εκπαίδευσης εμφανιζόταν άρνηση.

Αντί να κατατάξουν ένα κείμενο σε μια κλίμακα συναισθήματος (αρνητικό - θετικό), οι συγγραφείς του [32] χρησιμοποίησαν υπολογιστικά γλωσσικά εργαλεία (LIWC [33], LSA [34], (HAL [35]) προκειμένου να προσδιορίσουν το ακριβές συναίσθημα. Σαν σύνολο εκπαίδευσης χρησιμοποίησαν κείμενα ιστολογίων των 50 και 200 λέξεων και υιοθέτησαν ένα μοντέλο από λίστες λέξεων που αντιπροσωπεύουν τα οκτώ πρωτογενή συναισθήματα του Σχήματος 2.5 (όσο μεγαλύτερη είναι η απόσταση από το κέντρο, τόσο μεγαλύτερη είναι η ισχύς του συναισθήματος). Μέσα από τα πειράματα, διαπιστώθηκε ότι ορισμένα συναισθήματα όπως ο θυμός και η χαρά είναι πιο διακριτά από άλλα.




Σχήμα 2.5: Τροχός συναισθημάτων

### 2.2.2 Σημασιολογικός Προσανατολισμός

Ο σημασιολογικός προσανατολισμός ενός κειμένου, μπορεί να ταξινομηθεί σε τρεις κλάσεις (θετικό, αρνητικό ή ουδέτερο) και προκύπτει συμψηφίζοντας τον προσανατολισμό/πολικότητα όλων των επιμέρους λέξεων που το

συγκροτούν. Ο τρόπος υπολογισμού του Σημασιολογικού Προσανατολισμού κάθε λέξης ενός κειμένου έχει κατηγοριοποιήσει της τεχνικές της συγκεκριμένης κατηγορίας σε δύο υποκατηγορίες:

| Κειμενοσκοπική Ανάλυση | Λεξικογραφική Ανάλυση |
|---|---|
| (Corpus-based Analysis) | (Lexicon-based Analysis) |

Πρέπει να τονισθεί πως οι τεχνικές που ανήκουν στην κατηγορία του Σημασιολογικού Προσανατολισμού, εμφανίζουν την τάση να μεροληπτούν υπέρ του θετικού προσανατολισμού. Η αιτία είναι πως οι άνθρωποι χρησιμοποιούν στα κείμενα τους πολύ πιο συχνά λέξεις με «θετικό» προσανατολισμό έναντι αυτό με «αρνητικό» προσανατολισμό, έτσι ο προκύπτον προσανατολισμός του κειμένου δεν αντιστοιχεί στον πραγματικό.

### Κειμενοσκοπική Ανάλυση

Στις μεθόδους αυτής της υποκατηγορίας διερευνάται η σχέση μεταξύ των λέξεων που υπάρχουν σε κάποιο κείμενο και κάποιας άλλης λέξης του κειμένου που είναι συναισθηματικά προσημασμένη. Αρχικά δημιουργούνται δύο σύνολα λέξεων, ένα με λέξεις που έχουν θετικό προσανατολισμό και ένα με αρνητικό. Η μέθοδος διατρέχει το κείμενο και μετρά τον αριθμό των εμφανίσεων των μη συναισθηματικά προσημασμένων όρων του κειμένου, που είναι σε μικρή απόσταση (αν, για παράδειγμα, βρίσκονται σε απόσταση το πολύ μέχρι 10 λέξεις μεταξύ τους) από τους προσημασμένους. Με βάση τα στοιχεία που συλλέχθηκαν, προσδίδει θετικό προσανατολισμό σε μια λέξη αν η συχνότητα εμφάνισης της κοντά στις θετικά προσανατολισμένες λέξεις, είναι μεγαλύτερη απ΄ οτι για τις αρνητικά προσανατολισμένες λέξεις.

Οι κειμενοσκοπικές προσεγγίσεις αξιοποιούν το συντακτικό και τα συμφραζόμενα και απαιτούν για το σωστό προσδιορισμό του προσανατολισμού κάθε λέξης, μεγάλη συλλογή κειμένων. Επιπλέον, υιοθετούν την παραδοχή πως ο προσανατολισμός μίας λέξης είναι σταθερός ανεξαρτήτως πλαισίου στο οποίο αναφέρεται η λέξη, γεγονός που δεν υφίσταται στην πραγματικότητα, αφού ο προσανατολισμός των λέξεων επηρεάζεται από το πλαίσιο και το πεδίο μέσα στο οποίο αναφέρονται.

### Λεξικογραφική Ανάλυση

Στη συγκεκριμένη κατηγορία, με τον όρο λεξικό εννοείται μια λίστα από λέξεις, συνήθως τα θέματά τους, που έχουν βαθμονομηθεί για την κλίμακα αρνητικό - θετικό (π.χ. SentiWordNet [36]) ή σχετικά με το πόσο ταιριάζουν σε κάποιο από τα βασικά συναισθήματα. Λαμβάνοντας υπόψη τη βαθμολογία κάθε λέξης μιας εγγραφής, προκύπτει η συνολική πολικότητά της εγγραφής ή η εγγραφή αντιστοιχίζεται μονοσήμαντα σ' ένα συναίσθημα. Στα λεξικά οι προσεγγίσεις που διαχειρίζονται ενδεχόμενη άρνηση στο κείμενο, συνήθως αλλάζουν το πρόσημο της συναισθηματικής βαθμολογίας της εμπλεκόμενης λέξης, διατηρώντας έτσι την ένταση αλλά αλλάζοντας την πολικότητα του συναισθήματος. Τα λεξικά που χρησιμοποιούνται στην εκάστοτε μέθοδο είτε κατασκευάζονται απευθείας [37] από τον προγραμματιστή ή/και το χρήστη, είτε μέσω μιας αυτοματοποιημένης διαδικασίας [38] που χρησιμοποιούνται λέξεις ((seed)) ώστε να αναβαθμιστεί το λεξικό. Η αποδοτικότητα των τεχνικών της Λεξικογραφικής Ανάλυσης βασίζεται σε μεγάλο βαθμό στην πληρότητα και την ευστοχία των διαθέσιμων λεξικών. Ωστόσο, ο καταρτισμός ενός μόνο λεξικού, ικανού να ανταπεξέρχεται σε διαφορετικά πλαίσια λειτουργίας, είναι εξαιρετικά δύσκολη.

Οι μέθοδοι της Μηχανικής Μάθησης που σχετίζονται με την Ανάλυση συναισθήματος δεν παρουσιάζουν πολύ υψηλά ποσοστά επιτυχίας [16] και χρειάζονται χρόνο μέχρι να εκπαιδεύσουν ένα μοντέλο. Από την άλλη μεριά, οι Λεξικογραφικές τεχνικές που χρησιμοποιούν ένα εγκεκριμένο, από την επιστημονική κοινότητα, λεξικό για την πρόβλεψη συναισθημάτων σε άγνωστα κείμενα, δεν απαιτούν κανενός είδους εκπαίδευση. Άρα οι Λεξικογραφικές τεχνικές είναι θεωρητικά πιο γρήγορες από αυτές της Μηχανικής Μάθησης και δεν κινδυνεύουν από τις προβλέψεις ενός λάθος εκπαιδευμένου μοντέλου. Πρέπει όμως να επισημάνουμε πως οι Λεξικογραφικές Τεχνικές δεν έχουν την ικανότητα να ανακαλύπτουν τις αλληλουχίες που υπάρχουν στα δεδομένα, γεγονός ιδιαίτερης σημασίας, αφού οι αλληλουχίες που υποβόσκουν σε ένα κείμενο και επηρεάζουν την πρόβλεψη συναισθήματος, είναι εξαιρετικά πολλές. Στον Πίνακα 2.1 παρουσιάζεται συνοπτικά η σύγκριση μεταξύ τεχνικών της Μηχανικής Μάθησης και της Λεξικογραφικής Ανάλυσης. Τέλος, αξίζει να σχολιάσουμε πως οι Λεξικογραφικές Τεχνικές (lexicon - based) συνήθως συνδυάζονται με τεχνικές μη επιβλεπόμενης μάθησης, βελτιώνοντας την ταξινόμηση των συναισθημάτων [39].

## 2.3 Προεπεξεργασία Δεδομένων Κειμένου

Η προεπεξεργασία των δεδομένων, στη συγκεκριμένη εργασία κειμένων, αποτελεί την πιο σημαντική πτυχή στην κατάρτιση ενός μοντέλου Ανάλυσης Συναισθήματος. Περιλαμβάνει τον καθαρισμό και το φιλτράρισμα του κειμένου ώστε να περιέχει τους πιο ουσιώδεις όρους. Λόγω της συρρίκνωσης του λεξιλογίου από την απομάκρυνση μη

| Χαρακτηριστικό | Μηχανική Μάθηση | Λεξικογραφική Ανάλυση |
|---|---|---|
| Χρόνος εκτέλεσης | Μέτριος | Μικρός |
| Ανακάλυψη αλληλουχιών | Ναι | Όχι |
| Ακρίβεια προβλέψεων | Ανάλογα το μοντέλο | Μεγάλη |

∗ Στον όρο Μηχανική Μάθηση συμπεριλαμβάνονται τεχνικές με επιβλεπόμενη και μη επιβλεπόμενη μάθηση.

Πίνακας 2.1: Σύγκριση τομέων Συναισθηματικής Ανάλυσης

χρήσιμων όρων, προκύπτει ένα σύνολο δεδομένων με αυξημένες δυνατότητες ως προς την ανάκτηση πληροφορίας, σχετικά με το αρχικό.

Τα ηλεκτρονικά κείμενα και ειδικά αυτά του Twitter, εξαιτίας της ιδιαιτερότητας των tweets (αναφερόμαστε στη φυσική τους σύσταση) είναι γεμάτα θόρυβο (email, υπερσυνδέσμοι, λέξεις με ορθογραφικά λάθη), καθιστώντας απαραίτητη την προεπεξεργασία τους. Εκτός από τον θόρυβο, αφαιρούνται και πολλές λέξεις που δεν επηρεάζουν το εννοιολογικό πλαίσιο του κειμένου. Η απόρριψη τέτοιων όρων είναι ιδιαιτέρως σημαντική αφού μειώνεται η διάσταση του προβλήματος και αυξάνεται η απόδοση του μοντέλου. Στη συνέχεια παρουσιάζονται τα βήματα που ακολουθούνται κατά τη διαδικασία προεπεξεργασίας των κειμένων.

1. **Λημματοποίηση (Tokenization).** Ένα κείμενο μπορεί να περιέχει όρους όπως λέξεις, αριθμούς, σύμβολα ή σημεία στίξης. Από αυτούς τους όρους, κρατάμε στο σύνολο δεδομένων αυτούς που προσφέρουν πληροφορία ως προς το εννοιολογικό περιεχόμενο του κειμένου, δηλαδή το κείμενο από ακολουθία χαρακτήρων μετατρέπεται σε ακολουθία μόνο από λέξεις. Το βήμα αυτό είναι καθοριστικό για την απόδοση του μοντέλου αφού τα όποια λάθη προκύψουν σε αυτό το βήμα, θα προωθηθούν και στα επόμενα.

2. **Αφαίρεση των Stopwords.** Αφαιρούνται από το κείμενο τετριμμένες λέξεις που έχουν μεγάλη συχνότητα εμφάνισης μέσα στο κείμενο αλλά δεν προσθέτουν ιδιαίτερη σημασία στο νόημα του κειμένου. Για παράδειγμα, στην ελληνική γλώσσα, στους τετριμένους όρους συγκαταλέγονται τα άρθρα, επίθετα όπως τα κάποιος, άλλος, ... και αντωνυμίες όπως το κάθε.

3. **Αποκατάληξη (Stemming).** Προκειμένου να μειωθεί ακόμα περισσότερο η διάσταση του προβλήματος πρέπει λέξεις που έχουν το ίδιο θέμα να αντιστοιχηθούν στο ίδιο γνώρισμα. Για να επιτευχθεί κάτι τέτοιο γίνεται περικοπή των καταλήξεων καθώς και των προθεμάτων των λέξεων.

## 2.4 Αναπαράσταση Χαρακτηριστικών σε Δεδομένα Κειμένου

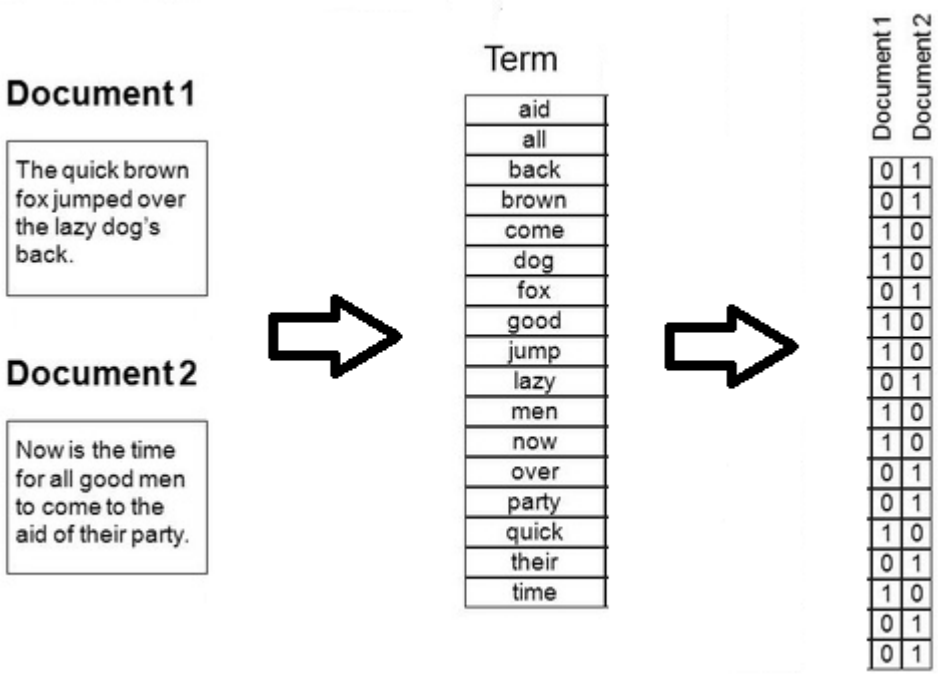
Το κείμενο ως δεδομένο δεν έχει εκ των προτέρων ορισμένη δομή και δε μπορεί να χρησιμοποιηθεί, με αυτή την ακατέργαστη δομή, από τις τεχνικές μηχανικής μάθησης για την εξαγωγή συμπερασμάτων στον τομέα της Ανάλυσης Συναισθήματος. Η διαδικασία κωδικοποίησης των δεδομένων από τα κείμενα μπορεί να θεωρηθεί ως μία πολύ απλή διαδικασία ως και μία εξίσου σύνθετη. Η ιδιαιτερότητα αυτή προκύπτει από το γεγονός πως πρέπει να επιλεχθεί η καταλληλότερη μορφή αναπαράστασης του κειμένου σύμφωνα με τις ιδιαιτερότητες του προβλήματος που θέλουμε να αντιμετωπίσουμε. Για τη χρησιμοποίηση δεδομένων μορφής κειμένου ως είσοδο σε κάποιο μοντέλο μηχανικής μάθησης, πρέπει να αναπαραστήσουμε το κείμενο σε κάποια αριθμητική μορφή (συνήθως διανύσματα χαρακτηριστικών), η οποία είτε αξιοποιείται ως έχει ως το τελικό διάνυσμα χαρακτηριστικών, είτε αξιοποιείται για τη δημιουργία νέων σύνθετων χαρακτηριστικών. Στη συνέχεια, περιγράφονται διάφορες τεχνικές αναπαράστασης δεδομένων υπό μορφή αριθμητικών διανυσμάτων.

### 2.4.1 Σύνολα Λέξεων (Bag-of-words)

Θεωρείται ως η πιο απλή μέθοδος αναπαράστασης ενός κειμένου με αριθμητικά διανύσματα. Αφού έχει συγκεντρωθεί το σύνολο των κειμένων προς επεξεργασία, δημιουργείται μία λίστα από τις λέξεις που εμφανίζονται στο σύνολο των κειμένων, εξού και η ονομασία της μεθόδου. Οι λέξεις που περιέχονται στη λίστα ονομάζονται χαρακτηριστικά του κειμένου. Έτσι δημιουργείται ένα αραιό διάνυσμα για κάθε κείμενο που εξαρτάται από τον τρόπο με τον οποίο θα ορίσουμε τις τιμές του. Οι πιο ευρέως χρησιμοποιούμενοι είναι οι:

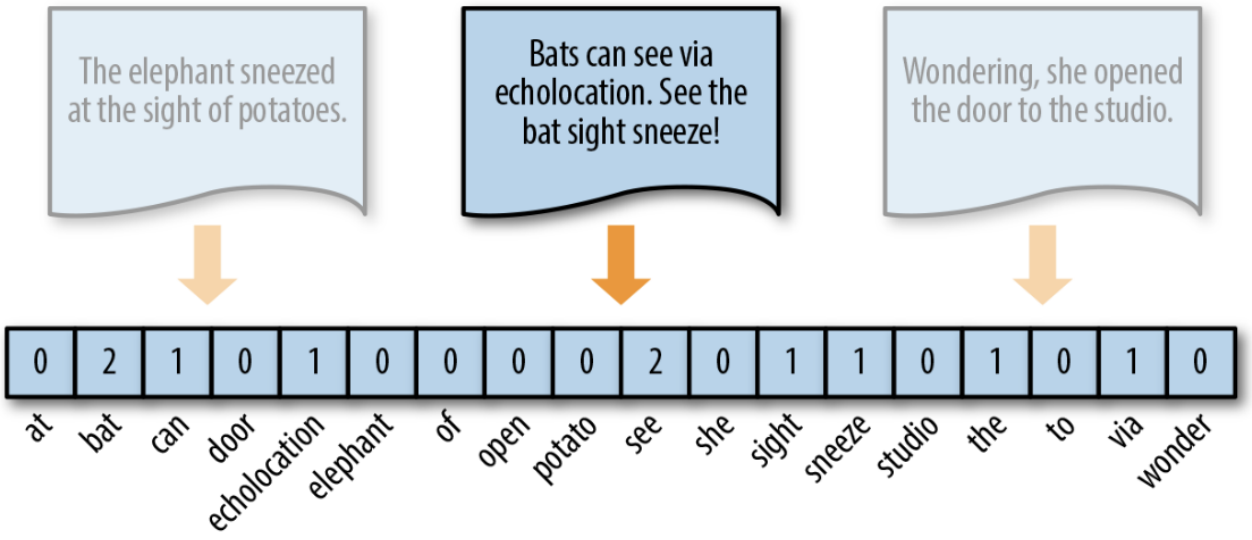
1. Δυαδικές τιμές: Στη θέση του διανύσματος που αντιστοιχεί η εκάστοτε λέξη, δίνεται η τιμή 1 αν η λέξη υπάρχει στο κείμενο, διαφορετικά το 0 (Σχήμα 2.6).

   Στις προτάσεις που αποτελούνται από μία μόνο λέξη θα έχουμε διανύσματα με 1 σε μια θέση και 0 σε όλα τις υπόλοιπες. Αν οι λέξεις είναι ίδιες, τότε η Ευκλείδεια απόστασή (Σχέση 3.20) τους θα είναι 0 ενώ σε οποιοδήποτε συνδυασμό διαφορετικών, θα είναι $\sqrt{2}$. Δηλαδή ο παραγόμενος διανυσματικός χώρος δε δίνει πληροφορίες αναφορικά με τις σημασιολογικές σχέσεις μεταξύ των λέξεων.




Σχήμα 2.6: Παράδειγμα αναπαράστασης Bag-of-words για δυαδικές τιμές

2. Αριθμός εμφανίσεων: Στη θέση του διανύσματος που αντιστοιχεί η εκάστοτε λέξη, η τιμή είναι ίση με το πλήθος των εμφανίσεων της λέξης στο κείμενο (Σχήμα 2.7).




Σχήμα 2.7: Παράδειγμα αναπαράστασης Bag-of-words για αριθμό εμφανίσεων

3. Συχνότητα εμφανίσεων: Στη θέση του διανύσματος που αντιστοιχεί η εκάστοτε λέξη, η τιμή είναι ίση με τη συχνότητα των εμφανίσεων της λέξης στο κείμενο.

4. TF-IDF: Στη θέση του διανύσματος που αντιστοιχεί η εκάστοτε λέξη, δίνεται η TF-IDF τιμή της λέξης. Ο όρος TF αντιστοιχεί στη συχνότητα κάποιας λέξης, ενώ ο όρος IDF αντιστοιχεί με κάποιο βάρος που σχετίζεται με τη συχνότητα μιας λέξης του κείμενου, σε σχέση με το σύνολο των κειμένων. Εύκολα παρατηρεί κανείς πως με αυτή την αναπαράσταση, οι σπάνιοι όροι έχουν υψηλό IDF, ενώ οι συχνοί επιβαρύνονται με

χαμηλότερο IDF. Ο αριθμός TF-IDF για τη λέξη $i$ στο κείμενο $k$ δίνεται από τον τύπο

$$TF - IDF_{ik} = f_{ik} \cdot \log\left(\frac{N}{n_i}\right)$$

όπου $f_{ik}$ η συχνότητα της λέξης $i$ στο κείμενο $k$, $N$ ο συνολικός αριθμός κειμένων και $n_i$ ο αριθμός των κειμένων που περιέχουν τουλάχιστον μία φορά τη λέξη $i$. Μια παραλλαγή της μεθόδου TF-IDF για κείμενα διαφορετικού μήκους είναι η TFC. Η διαφορά με την TF-IDF είναι πως περιλαμβάνει κανονικοποίηση του κειμένου για να αποδώσει την τιμή για τη λέξη $i$ στο κείμενο $k$ και δίνεται από τον τύπο

$$TFC_{ik} = \frac{f_{ik} \cdot \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[f_{ik} \cdot \log\left(\frac{N}{n_i}\right)\right]^2}}$$

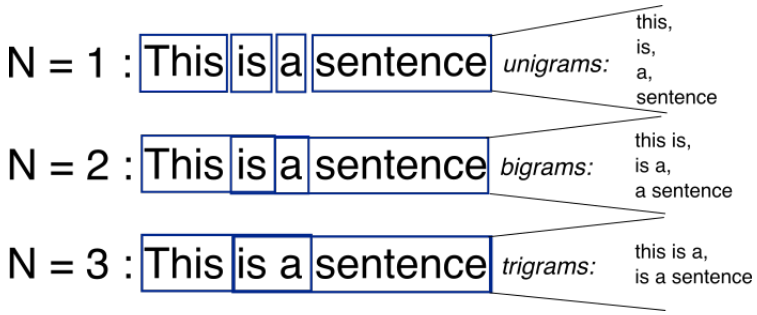
όπου $M$ ο αριθμός των λέξεων στη λίστα.

Εύκολα διαπιστώνει κανείς πως η συγκεκριμένη μέθοδος δεν αξιοποιεί καμία πληροφορία από τη σειρά εμφάνισης των λέξεων και πως κείμενα που περιέχουν τις ίδιες λέξεις θα έχουν σίγουρα την ίδια αναπαράσταση, ανεξαρτήτως της σημασιολογικής τους υπόστασης. Στα θετικά συγκαταλέγεται η δημιουργία αραιών διανυσμάτων μεγάλης διάστασης, για την αναπαράσταση των κειμένων,που τις περισσότερες φορές είναι γραμμικώς διαχωρίσιμα. Τέλος πρέπει να επισημάνουμε πως πρέπει να προηγηθεί επεξεργασία των κειμένων ώστε να αφαιρεθούν όροι που δε δίνουν πληροφορία καθώς και να γίνει word stemming (Σελ. 19). Επιπλέον, επειδή το μέγεθος ενός λεξικού μπορεί να γίνει δυσθεώρητα μεγάλο, υπάρχει δυνατότητα ο χρήστης να επιλέξει τον αριθμό των χαρακτηριστικών της αναπαράστασης (συνήθως 1000 - 5000 πιο συχνοί όροι) ή να εφαρμόσει κάποια αυτόματη μέθοδο εξαγωγής χαρακτηριστικών.

### 2.4.2 N-Gram

Στη συγκεκριμένη υποενότητα ξεκινάμε παραθέτωντας τον ορισμό ενός n-gram [16]:

**Ορισμός 2.3.** *«Με τον όρο* n-gram*, θεωρείται οποιαδήποτε αναπαράσταση μιας συλλογής $n$ το πλήθος λέξεων ενός tweet (πρότασης), από τις επιμέρους λέξεις της. Ένα n-gram μίας λέξης, καλείται 1-gram ή unigram, ένα n-gram δύο λέξεων 2-gram κ.ο.κ.»*




Σχήμα 2.8: Παραδείγματα για n-gram

Δημιουργούνται δηλαδή ακολουθίες διαδοχικών λέξεων οι οποίες αν υπάρχουν αυτούσιες στο κείμενο, αναπαρίστανται με τρόπο πανομοιότυπο με την περίπτωση των απλών λέξεων. Το μεγάλο μειονέκτημα των ν-γραμς είναι πως αυξάνοντας το μέγεθός του ν, αυξάνεται σημαντικά το πλήθος των όρων του λεξιλογίου και ο θόρυβος στο μοντέλο αφού οι περισσότεροι όροι δεν περιέχουν αρκετή πληροφορία.

Για παράδειγμα, η πρόταση «A sentence is a group of words.» έχει τις ακόλουθες αναπαραστάσεις αν χρησιμοποιηθούν τα n-gram του Σχήματος 2.8:

$n = 1$ (0,1,1,1)
$n = 2$ (0,1,0)
$n = 3$ (0,0)

### 2.4.3 Τεχνικές Μείωσης Διαστάσεων

Στο πρόβλημα της Ανάλυσης Συναισθήματος από δεδομένα του Twitter, χρειαζόμαστε ένα τρόπο να απεικονίζουμε τα δεδομένα υπο μορφή κειμένου σε ένα διανυσματικό χώρο λίγων διαστάσεων. Με αυτό τον τρόπο επιτυγχάνεται κοντινά σημεία στο χώρο να αντιπροσωπεύουν παρόμοιο νόημα. Ερχόμαστε έτσι αντιμέτωποι με το πρόβλημα διαχείρισης δεδομένων που αποτελούνται από μεγάλο αριθμό χαρακτηριστικών διότι τα περισσότερα μοντέλα αδυνατούν να διαχειριστούν πολύ μεγάλο όγκο δεδομένων ή αργούν πάρα πολύ να εξάγουν αποτελέσματα ή υπερπροσαρμόζονται στα δεδομένα. Η λύση είναι η μείωση της διάστασης των δεδομένων, κρατώντας τα χαρακτηριστικά εκείνα που προσφέρουν τη μέγιστη δυνατή πληροφορία.

Η *μείωση διαστάσεων* (dimensionality reduction) περιλαμβάνει μεθόδους που μετατρέπουν ένα σύνολο δεδομένων από το χώρο $R^m$ στο χώρο $R^n$ όπου $m \gg n$, χάνοντας μικρό ποσοστό πληροφορίας του αρχικού χώρου. Είναι απαραίτητη για το πρόβλημα της Ανάλυσης Συναισθήματος αφενός επειδή μειώνεται ο χρόνος επεξεργασίας των δεδομένων αφετέρου επειδή αποκαλύπτεται πιο εύκολα η κρυφή δομή στον ελαττομένο χώρο. Συνήθως για την εκτίμηση της απόδοσης μιας μεθόδου μείωσης διάστασης, χρησιμοποιείται ο δείκτης stress [40] που εκτιμά τη διατήρηση των αποστάσεων ανά ζεύγη σημείων στον ελαττομένο χώρο και υπολογίζεται σύμφωνα με την παρακάτω εξισωση.

$$stress = \sqrt{\frac{\sum_{i=1}^{d}\sum_{j=1}^{d}(d_{ij}^{(n)^2} - d_{ij}^{(m)^2})}{\sum_{i=1}^{d}\sum_{j=1}^{d} d_{ij}^{(n)^2}}}$$

όπου $d_{ij}^{(n)}$, $d_{ij}^{(m)}$ η Ευκλείδεια απόσταση ανάμεσα στα σημεία $i$ και $j$ στον $n$-διάστατο και $m$-διάστατο χώρο.

#### Λανθάνουσα Σημασιολογική Ανάλυση

Η Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis) [34] αποτελεί μια στατιστική προσέγγιση για γραμμική μείωση των διαστάσεων ενός συνόλου δεδομένων. Η μέθοδος αυτή δύναται να ανακαλύπτει τις αλληλοεξαρτήσεις μεταξύ των όρων ενός συνόλου κειμένων και να μετατρέπει την αναπαράσταση των κειμένων από το χώρο των όρων στο χώρο των εννοιών. Η διαδικασία που ακολουθείται κατά την εκτέλεση της εν λόγω μεθόδου, είναι η εξής:

1. Μέσα από τη μέθοδο Bag-of-words δημιουργούμε το λεξιλόγιο και το διάνυσμα αναπαράστασης κάθε κειμένου.
2. Κανονικοποιούμε τα διανύσματα βάσει της μεθόδου TF-IDF.
3. Εφαρμόζουμε μία τεχνική μείωσης διαστάσεων (π.χ. SVD).

**Ορισμός 2.4.** *Παραγοντοποίηση Ιδιαζουσών Τιμών (SVD)*
*Η παραγοντοποίηση ενός μητρώου A διάστασης $n \times m$ ως*

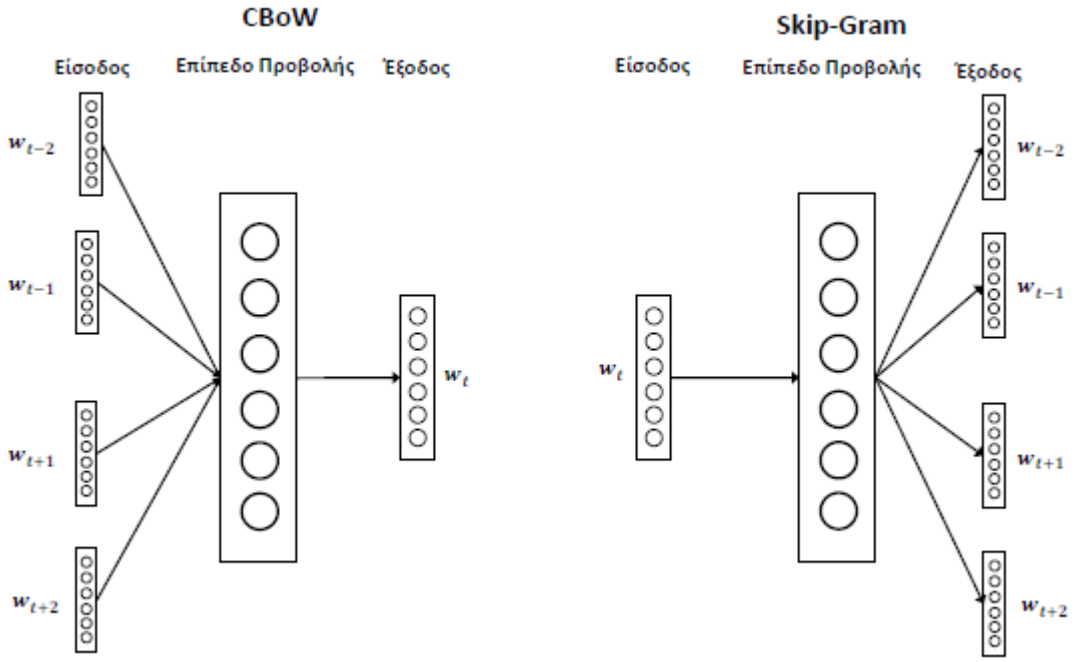
$$A = U \cdot D \cdot V^T$$

*όπου το U είναι ένα ορθογώνιο μητρώο διάστασης $n \times n$, το V είναι ένα ορθογώνιο μητρώο διάστασης $m \times m$ και το D είναι ένα ορθογώνιο μητρώο διάστασης $n \times m$ με μη αρνητικές καταχωρήσεις, ονομάζεται παραγοντοποίηση ιδιαζουσών τιμών [41].*

#### word2vec

Το μοντέλο word2vec [42] αποτελείται από ένα τεχνητό νευρωνικό δίκτυο με ένα κρυφό επίπεδο, που εκπαιδεύεται χωρίς επίβλεψη σε μία αρκετά μεγάλη συλλογή από κείμενα ώστε να είναι σε θέση να αντιστοιχίσει κάθε ξεχωριστή λέξη σε διάνυσμα μικρής διάστασης. Πρακτικά αυτό που εκπαιδεύεται να κάνει το τεχνητό νευρωνικό δίκτυο, είναι να αποδίδει πόσο πιθανό είναι να δημιουργηθεί μία ακολουθία λέξεων. Υπάρχουν οι εξής δύο διαφορετικές υλοποιήσεις του μοντέλου word2vec:

1. **Continuous Bag-of-Words (CBOW):** Για δεδομένο αριθμό λέξεων (μέγεθος παραθύρου), γίνεται πρόβλεψη της κεντρικής λέξης. Π.χ. αν το μέγεθος παραθύρου είναι 3, η είσοδος του δικτύου είναι οι λέξεις $w_{i-3}$, $w_{i-2}$, $w_{i-1}$, $w_{i+1}$, $w_{i+2}$, $w_{i+3}$ και $w_i$ η έξοδος.
2. **Skip-Gram (SG):** Επιτελεί την αντίστροφη διαδικασία από το CBOW. Δηλαδή, δίνοντας ως είσοδο στο δίκτυο μια λέξη, προσπαθεί να προβλέψει πόσο πιθανό είναι άλλες λέξεις να βρεθούν στο πλαίσιο της δεδομένης της λέξης. Εδώ εκτός από το μέγεθος παραθύρου, παράμετρος είναι και πόσες λέξεις θα παραλειφθούν πριν και μετά τη δεδομένη λέξη.




Σχήμα 2.9: Γραφική αναπαράσταση Continuous Bag-of-Words και Skip-Gram

## 2.5 Αξιολόγηση Μοντέλων Ανάλυσης Συναισθημάτων

Το πιο σημαντικό στάδιο της ανάδειξης ενός μοντέλου ως επιτυχημένου ή μη είναι η αξιολόγηση του. Κατά την αξιολόγηση του μοντέλου, το τροφοδοτούμε με νέα δεδομένα και εξετάζουμε αν η έξοδός του ταυτίζεται με την επιθυμητή έξοδο. Παραδείγματος χάριν, αν είχαμε ένα πρόβλημα ταξινόμησης όπως αυτό στο σύνολο δεδομένων Iris[1], θα δημιουργούσαμε δύο ξένα σύνολα (εκπαίδευσης και ελέγχου) και μετά την εκπαίδευση, θα τροφοδοτούσαμε το μοντέλο μας με το σύνολο ελέγχου. Σε αυτή την περίπτωση μπορούμε εύκολα να διαπιστώσουμε αν η έξοδος του μοντέλου ταυτίζεται με την επιθυμητή έξοδο, σύμφωνα με τα δεδομένα εισόδου. Στην περίπτωση όμως της Ανάλυσης Συναισθήματος τα πράγματα είναι πιο δύσκολα γιατί η αξιολόγηση ενός τέτοιου συστήματος, έγκειται βασικά στο πόσο καλά συμφωνεί με την ανθρώπινη κρίση.

| | Προβλεπόμενα Θετικά Στοιχεία | Προβλεπόμενα Αρνητικά Στοιχεία |
|---|---|---|
| Πραγματικά Θετικά Στοιχεία | Αριθμός Πραγματικά Θετικών Στοιχείων TP | Αριθμός Ψευδώς Αρνητικών Στοιχείων FN |
| Πραγματικά Αρνητικά Στοιχεία | Αριθμός Ψευδώς Θετικών Στοιχείων FP | Αριθμός Πραγματικά Αρνητικών Στοιχείων TN |

Πίνακας 2.2: Μορφή πίνακα σύγχυσης

Η ακρίβεια και η αξιοπιστία ενός μοντέλου Ανάλυσης Συναισθημάτων καθορίζεται κυρίως μέσα από κάποιες μετρήσεις και εκτιμήσεις εμπειρογνωμόνων. Η χρήση των δεικτών: Accuracy, Precision, Recall και F-score είναι χρήσιμη για την εκτίμηση της απόδοσης μοντέλων ανάλυσης συναισθήματος. Ο υπολογισμός τους είναι αρκετά εύκολος όταν είναι διαθέσιμος ο πίνακας σύγχυσης του εκάστοτε προβλήματος (Σχήμα 2.2). Ο τύπος για τον υπολογισμό της τιμής κάθε δείκτη παρουσιάζεται στις εξισώσεις 2.1 έως 2.4

[1] $http://archive.ics.uci.edu/ml/datasets/Iris/$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

Σύμφωνα με το [43] όμως, οι άνθρωποι που ορίζονται ως εκτιμητές σε προβλήματα ανάλυσης συναισθημάτων, συνήθως συμφωνούν μόνο στο 80% των περιπτώσεων. Ως εκ τούτου, θα μπορούσε εύκολα κάποιος να ισχυριστεί ότι μια μέθοδος με ακρίβεια πάνω από 70%, αποδίδει σχεδόν το ίδιο καλά με τους ανθρώπους, αφού και απολύτως σωστή να ήταν η μέθοδος, οι άνθρωποι και πάλι θα διαφωνούσαν με αυτό κατά 20%.

Ερευνητικές και επιχειρηματικές ομάδες παγκοσμίως ασχολούνται με την ανάλυση συναισθημάτων. Οι επιχειρηματικές ομάδες επικεντρώνονται στην εξόρυξη γνώμης από κριτικές (reviews) και από τα μέσα κοινωνικής δικτύωσης για ένα μεγάλο πλήθος εφαρμογών που εκτείνεται από τη διαφήμιση μέχρι την εξυπηρέτηση πελατών. Από την άλλη μεριά, ο επιστημονικός κλάδος επικεντρώνεται κυρίως στην κατανόηση της δυναμικότητας του συναισθήματος στις e-κοινωνίες μέσω των τεχνικών της ανάλυσης συναισθημάτων. Το έργο και των δύο δυσχεραίνεται κυρίως από πολιτισμικούς παράγοντες και γλωσσολογικές αποχρώσεις, και σε συνδυασμό με το γεγονός ότι ακόμα και οι άνθρωποι συχνά διαφωνούν μεταξύ τους για το συναίσθημα των κειμένων, αποδεικνύεται τη δυσκολία που αντιμετωπίζει ένας υπολογιστής για την μετατροπή ενός τμήματος γραπτού κειμένου σε κάποιο συναίσθημα. Στον Πίνακα 2.3 παρουσιάζεται η σύγκριση της απόδοσης μεθόδων ανάλυσης συναισθημάτων για διάφορους τομείς και προσεγγίσεις.

| | Μέθοδος | Σύνολο Δεδομένων | Ακρίβεια |
|---|---|---|---|
| Μηχανική Μάθηση | SVM | Movie reviews | 86.4% |
| | CoTraining SVM | Twitter | 82.52% |
| | Deep learning | Stanford Sentiment Teebank | 80.7% |
| Λεξικογραφική Ανάλυση | Corpus | Product reviews | 74% |
| | Dictionary | Amazon's Mechanical Turk | |
| Διαγλωσσική (Cross - lingual) | Ensemble | Amazon | 81% |
| | Co-Train | Amazon, IT168 | 81.30% |
| | EWGA | IMDB movie review | $> 90\%$ |
| | CLMM | MPQA, NTCIR, ISI | 83.02% |
| Διατομεακή (Cross - domain) | Active learning | Book, DVD, Electronics, Kitchen | 80% (M.O.) |
| | Thesaurus | | |
| | SFA | | |

Πίνακας 2.3: Σύγκριση απόδοσης μεθόδων ανάλυσης συναισθημάτων

# 3. Μοντέλα Μηχανικής Μάθησης

## 3.1 Ταξινομητές Bayes

Εμπνευσμένοι από τη θεωρία αποφάσεων του Bayes, έχει δημιουργηθεί μια οικογένεια ταξινομητών με ευρεία χρήση σε προβλήματα που τα δεδομένα εισόδου είναι μεγάλης τάξης. Οι ταξινομητές Bayes [44] προβλέπουν την κλάση $y_i$, $i = 1, 2, \ldots, k$ στην οποία ανήκει ένα διάνυσμα χαρακτηριστικών $X = (x_1, x_2, \ldots, x_n)$. Ουσιαστικά, ταξινομούν το διάνυσμα στην κλάση που εμφανίζει τη μεγαλύτερη δεσμευμένη πιθανότητα $P(y_i|X) = P(y_i|x_1, x_2, \ldots, x_n)$. Η πιθανότητα αυτή υπολογίζεται σύμφωνα με τον τύπο

$$P(y_i|X) = \frac{P(y_i) \cdot P(X|y_i)}{P(X)} \tag{3.1}$$

όπου $P(A|B)$ η πιθανότητα να συμβεί το ενδεχόμενο $A$ δεδομένου ότι συμβαίνει το $B$ και $P(A)$ η πιθανότητα να συμβεί το $A$. Άρα, η έξοδος $out$ του ταξινομητή για το εκάστοτε διάνυσμα εισόδου $X$ είναι

$$out = \underset{i\in\{1,2,\ldots,k\}}{argmax} \frac{P(y_i) \cdot P(X|y_i)}{P(X)} \tag{3.2}$$

Επειδή ο παρονομαστής θα είναι κοινός για όλες τις κλάσεις (ανεξάρτητος του $i$), μπορεί να απαλειφθεί και η έξοδος το ταξινομητή να έχει τη μορφή

$$out = \underset{i\in\{1,2,\ldots,k\}}{argmax} P(y_i) \cdot P(X|y_i) \tag{3.3}$$

Στην εξίσωση 3.3 ο υπολογισμός των συναρτήσεων πυκνότητας πιθανότητας $P(X|y_i)$ έχει μεγάλο υπολογιστικό φόρτο και προκειμένου να παρακαμφθεί, γίνεται η υπόθεση πως τα χαρακτηριστικά $x_i$ του διανύσματος εισόδου $X$, είναι στατιστικά ανεξάρτητα. Άρα η πιθανότητα που υπολογίζεται για την ταξινόμηση του εκάστοτε διανύσματος δίνεται από τον τύπο

$$\begin{aligned}
\frac{P(y_i) \cdot P(X|y_i)}{P(X)} &= P(x_1, x_2, \ldots, x_n, y_i) \\
&= \underbrace{P(x_1|x_2, \ldots, x_n, y_i) \cdot P(x_2|x_3, \ldots, x_n, y_i) \cdot \cdots \cdot P(x_{n-1}|x_n, y_i) \cdot P(x_n|y_i)}_{P(x_j|x_{j+1},\ldots,x_n,y_i) = P(x_j|y_i)} \\
&\cdot P(y_i) \\
&= P(x_1|y_i) \cdot P(x_2|y_i) \cdot \cdots \cdot P(x_n|y_i) \cdot P(y_i) \\
&= \underset{i\in\{1,2,\ldots,k\}}{argmax} P(y_i) \cdot \prod_{j=1}^{n} P(x_j|y_i)
\end{aligned} \tag{3.4}$$

Από την εξίσωση 3.4 γίνεται εύκολα αντιληπτό ότι η ποσότητα $P(x_j|y_i)$ δείχνει την επίδραση του όρου $x_j$ ως προς τη σωστή ταξινόμηση της κλάσης $y_i$.

Πρέπει να αναφέρουμε πως παρά την υπόθεση ανεξαρτησίας σχετικά με τα χαρακτηριστικά του χώρου εισόδου, ο απλοϊκός ταξινομητής Bayes παρουσιάζει καλή επίδοση σε προβλήματα του πραγματικού κόσμου. Επιπλέον έχει χαμηλές απαιτήσεις υπολογιστικών πόρων και κατ' επέκταση απαιτεί μικρό χρόνο εκπαίδευσης. Από την άλλη μεριά κάποιες φορές υπερεκτιμά τις πιθανότητες εξόδου, κάνοντας έτσι λάθος προβλέψεις στα νέα δεδομένα.

Τέλος, δύο άλλες παραλλαγές του απλοϊκού ταξινομητή Bayes είναι ο ταξινομητής:

1. **Gaussian Naive Bayes** που υιοθετεί την παραδοχή ότι τα δεδομένα ακολουθούν κανονική κατανομή, δηλαδή η συνάρτηση πυκνότητας πιθανότητας έχει τη μορφή

$$P(x_j|y_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}^2}} e^{-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2}}$$

όπου τα $\mu_{ji}$ και $\sigma_{ji}^2$ υπολογίζονται με τη μέθοδο μέγιστης πιθανοφάνειας.

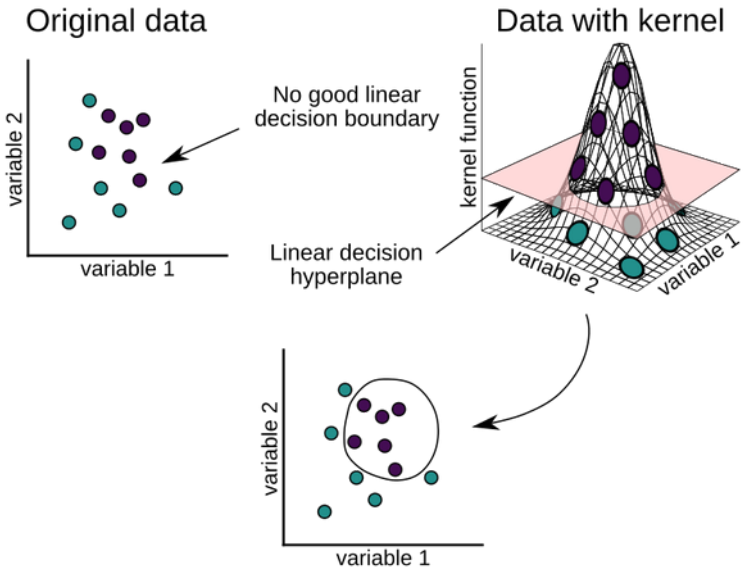
2. **Multinomial Naive Bayes** που εφαρμόζεται κυρίως σε προβλήματα ταξινόμησης κειμένου και η συνάρτηση πυκνότητας πιθανότητας του, είναι

$$P(x_j|y_i) = \frac{(\sum_{j=1}^{n} x_j)!}{\prod_{j=1}^{n} x_j!} \prod_{j=1}^{n} p_{ij}^{x_j}$$

όπου $p_{ij}^{x_j}$ είναι παράμετροι προς εκτίμηση και τα χαρακτηριστικά του διανύσματος $X$ αντιστοιχούν στη συχνότητα εμφάνισης κάθε λέξης.

## 3.2 Μηχανές Διανυσματικής Υποστήριξης

Η Μηχανή Διανυσματικής Υποστήριξης (Support Vector Machine-SVM) αποτελεί ένα ταξινομητή επιβλεπόμενης μάθησης για δεδομένα που ανήκουν σε δύο κλάσεις. Προκειμένου να μπορούν να ανταπεξέλθουν και σε προβλήματα με περισσότερες κλάσεις, χρησιμοποιούν το συνδυασμό των αποτελεσμάτων που απορρέουν από την εφαρμογή ενός συνόλου δυαδικών SVM. Οι δύο κλάσεις διαχωρίζονται μεταξύ τους μέσω ενός υπερεπιπέδου που έχει την ιδιότητα να απέχει τη μεγαλύτερη δυνατή απόσταση από τα κοντινότερα πρότυπα και των δύο κλάσεων. Για να επιτευχθεί η επιτυχής ταξινόμηση των προτύπων κάθε κλάσης, οι SVMs χρησιμοποιούν μια γραμμική ή μη γραμμική απεικόνιση των δεδομένων εισόδου σε ένα χώρο μεγαλύτερης διάστασης [45]. Η απεικόνιση αυτή εγυυάται πως τα δεδομένα που μπορεί να μην είναι διαχωρίσιμα στον αρχικό χώρο, θα είναι σε αυτόν της μεγαλύτερης διάστασης (Σχήμα 3.1).




Σχήμα 3.1: Μετασχηματισμός δεδομένων

Οι SVMs παρουσιάζουν καλή απόδοση στο πρόβλημα ταξινόμησης ως προς το συναίσθημα κάποιου κειμένου. Αυτό επιτυγχάνεται διότι μπορούν να λειτουργήσουν αποδοτικά ακόμα και όταν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από αυτόν των προτύπων του συνόλου δεδομένων. Επιπλέον έχουν μικρές απαιτήσεις ως προς τη μνήμη αφού για τον καταρτισμό του συνόρου απόφασης και κατ᾽ επέκταση της διαχωριστικής επιφάνειας, χρησιμοποιούν ένα μικρό υποσύνολο των προτύπων του προβλήματος.

Τα κύρια συστατικά μιας SVM είναι [46]:

1. Οι *συναρτήσεις πυρήνα* (kernel functions) μετασχηματίζουν τα δεδομένα από τον αρχικό χώρο εισόδου σε ένα χώρο χαρακτηριστικών μεγαλύτερης διάστασης. Η συνάρτηση πυρήνα

$$k(x, x^{'}) = \phi^{T}(x)\phi(x^{'}) =< \phi(x)\phi(x^{'}) >$$

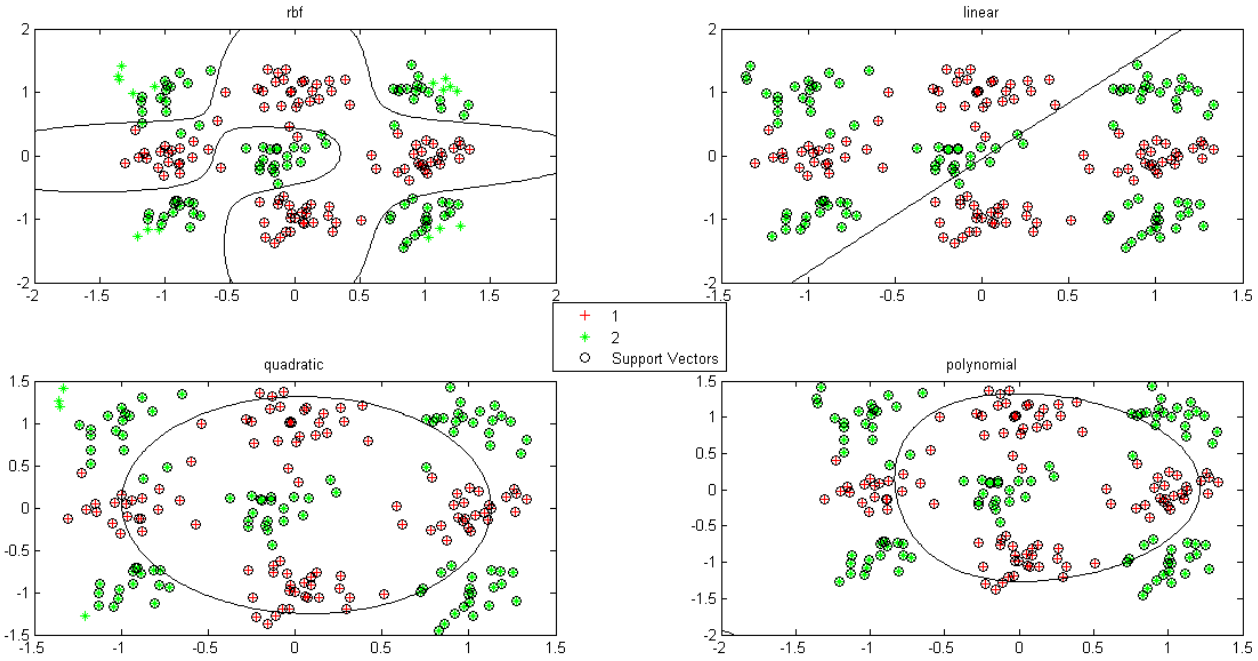
είναι το εσωτερικό γινόμενο των διανυσμάτων $\phi(x)$ και $\phi(x^{'})$ όπου $\phi(x)$ μια μη γραμμική συνάρτηση. Μερικές μορφές των συναρτήσεων πυρήνα παρουσιάζονται στις εξισώσεις 3.5 ως 3.7 και στο Σχήμα 3.2 αποτυπώνεται

η μορφή του διαχωριστικού επιπέδου που δημιουργούν.

$$k(x_i, x_j) = (x_i x_j + 1)^h \qquad \text{Πολυωνυμικός πυρήνας βαθμού } h \tag{3.5}$$

$$k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad \text{Πυρήνας συνάρτησης ακτινικής βάσης Gauss} \tag{3.6}$$

$$k(x_i, x_j) = \tanh(K x_i x_j - \delta) \qquad \text{Σιγμοειδής πυρήνας} \tag{3.7}$$




Σχήμα 3.2: Σύνορο απόφασης για διαφορετικά είδη συναρτήσεων πυρήνα.

2. *Το περιθώριο διαχωρισμού*: είναι το σημείο του υπερεπιπέδου απόφασης με τη μικρότερη απόσταση από τα δεδομένα εισόδου.

3. *Το βέλτιστο υπερεπίπεδο διαχωρισμού*: διαχωρίζει τις δύο κλάσεις (Σχήμα 3.3)

4. *Τα διανύσματα υποστήριξης*: τα πλησιέστερα σημεία στο υπερεπίπεδο διαχωρισμού (Σχήμα 3.3)

### 3.2.1 Γραμμικώς Διαχωρίσιμο Πρόβλημα

Έστω ότι έχουμε το γραμμικώς διαχωρίσιμο σύνολο εκπαίδευσης $\{(x_i, d_i)\}_{i=1}^N$ όπου $x_i$ η είσοδος και $d_i$ το αναμενόμενο αποτέλεσμα. Η εξίσωση του υπερεπιπέδου απόφασης δίνεται μέσω της σχέσης [47]:

$$w^T x + b = 0 \tag{3.8}$$

όπου $w$ το διάνυσμα βάρους και $b$ η μεροληψία. Αν $w_0, b_0$ οι βέλτιστες τιμές του διανύσματος βάρους και της μεροληψίας και $x^{(s)}$ ένα διάνυσμα υποστήριξης για το οποίο η κλάση είναι το 1 ($d^{(s)} = 1$), θα έχουμε ότι:

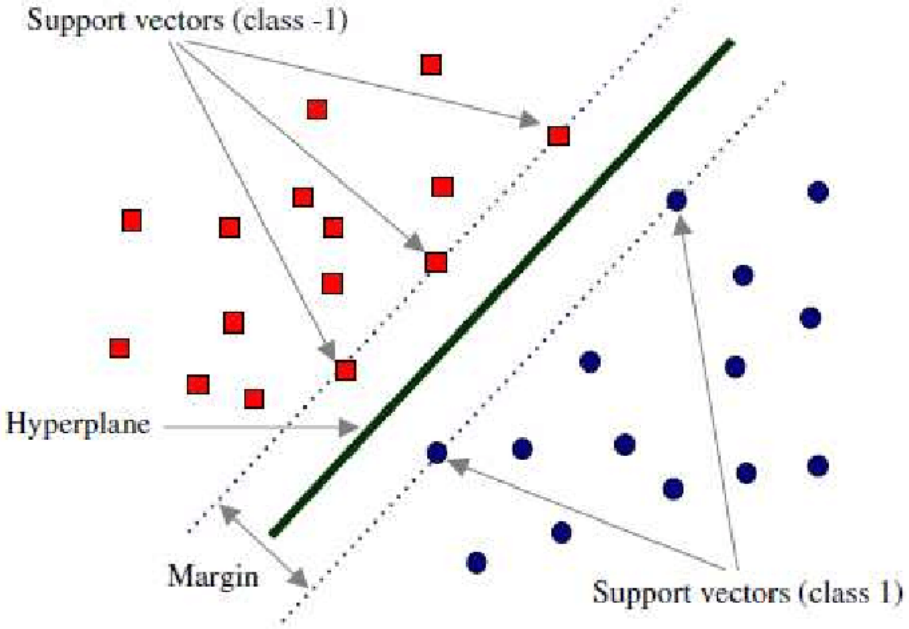
$$g(x^{(s)}) = w_0^T x^{(s)} \pm b_0 = \pm 1, \quad \text{για} \quad d^{(s)} = \pm 1 \tag{3.9}$$

Η απόσταση του διανύσματος $x^{(s)}$ από το βέλτιστο υπερεπίπεδο είναι [48]:

$$r = \frac{g(x^{(s)})}{\|w_0\|} = \begin{cases} \frac{1}{\|w_0\|}, & \text{αν} \quad d^{(s)} = 1 \\ -\frac{1}{\|w_0\|}, & \text{αν} \quad d^{(s)} = -1 \end{cases}$$

όπου το πρόσημο δείχνει σε ποια μεριά του υπερεπιπέδου βρίσκεται το $x^{(s)}$. Η βέλτιστη τιμή για το περιθώριο διαχωρισμού δίνεται από τον τύπο:

$$\rho = 2r = \frac{2}{\|w_0\|} \tag{3.10}$$




Σχήμα 3.3: Διαχωρισμός κλάσεων με SVM

Ο στόχος είναι να βρούμε τις βέλτιστες τιμές ενός διανύσματος βάρους $w$ και τη βέλτιστη τιμή της μεροληψίας $b$ ώστε για δεδομένο σύνολο εκπαίδευσης $\{(x_i, d_i)\}_{i=1}^{N}$ να ικανοποιείται ο περιορισμός [49]:

$$d_i(w^T x_i + b) \geq 1 \quad \text{για} \quad i = 1, 2, \ldots, N \tag{3.11}$$

και το διάνυσμα βάρους $w$ να ελαχιστοποιεί την κυρτή συνάρτηση κόστους

$$E(w) = \frac{1}{2} w^T w \tag{3.12}$$

Για να λύσουμε αυτό το πρόβλημα θα χρησιμοποιήσουμε τους πολλαπλασιαστές Lagrange [50]. Η συνάρτηση Lagrange δίνεται μέσω του τύπου

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i [d_i(w^T x_i + b) - 1] \tag{3.13}$$

όπου $\alpha_i$ μη αρνητικές σταθερές που ονομάζονται πολλαπλασιαστές Lagrange. Χρησιμοποιώντας τις μερικές παραγώγους ως προς $w$ και ως προς $b$ καταλήγουμε στις παρακάτω σχέσεις

$$\frac{\partial J(w, b, \alpha)}{\partial w} = 0 \tag{3.14}$$

και

$$\frac{\partial J(w, b, \alpha)}{\partial b} = 0 \tag{3.15}$$

Από τη (Σχέση 3.14) έχουμε ότι

$$w = \sum_{i=1}^{N} \alpha_i d_i x_i \tag{3.16}$$

και από τη (Σχέση 3.15) έχουμε ότι

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \tag{3.17}$$

Κατασκευάζοντας το δυικό του παραπάνω προβλήματος καταλήγουμε ότι το βέλτιστο βάρος $w_0$ υπολογίζεται ως [48]:

$$w_0 = \sum_{i=1}^{N} \alpha_{oi} d_i x_i \tag{3.18}$$

και η βέλτιστη διασπορά ως

$$b_0 = 1 - w_0^T x^{(s)}, \qquad d^{(s)} = 1 \tag{3.19}$$

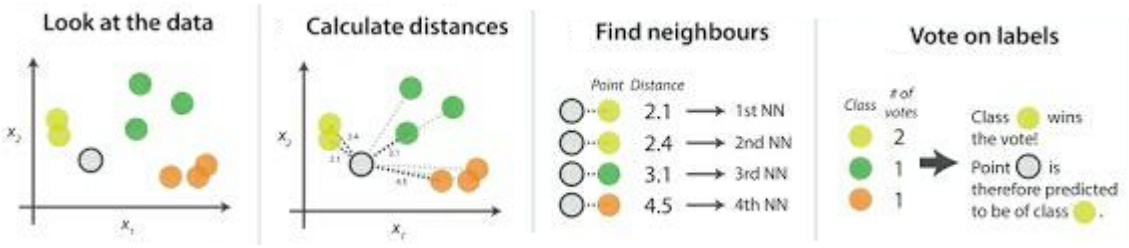
όπου $\alpha_{oi}$ οι βέλτιστοι πολλαπλασιαστές Lagrange και $x^{(s)}$ ένα θετικό διάνυσμα υποστήριξης. Η πλήρης και αναλυτική κατασκευή και περιγραφή του δυικού του προβλήματος ελαχιστοποίησης της συνάρτησης κόστους παρουσιάζεται στο [48].

## 3.3 Αλγόριθμος $k$-Κοντινότερων Γειτόνων

Ο αλγόριθμος των $k$-Κοντινότερων Γειτόνων (k-Nearest Neighbors - k-NN) θεωρείται ως ένας από τους γνωστότερους και ευρέως χρησιμοποιούμενους αλγόριθμους μηχανικής μάθησης για ταξινόμηση προτύπων σε ομάδες [51]. Βασίζεται στην παραδοχή ότι κάθε πρότυπο του συνόλου δεδομένων είναι ένα διάνυσμα άρα και ένα σημείο στον αντίστοιχο χώρο. Χρησιμοποιεί μετρικές απόστασης για να εκτιμηθεί η ομοιότητα μεταξύ των προτύπων και να επιτευχθεί σωστή ταξινόμηση για κάθε στοιχείο. Η εφαρμογή του προϋποθέτει την ύπαρξη κλάσης για κάθε πρότυπο ενός αρχικού συνόλου εισόδου αφού για να ταξινομήσει κάθε νέο πρότυπο, χρησιμοποιεί την κλάση που είναι η πιο κοινή μεταξύ των $k$ πλησιέστερων προτύπων.

Στη συνέχεια παραθέτονται τα βήματα του αλγόριθμου KNN

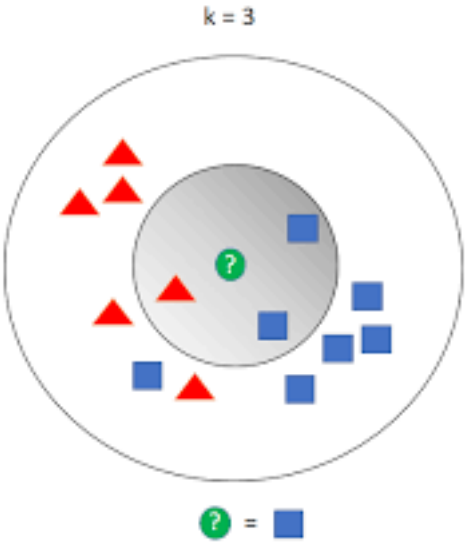
1. Συλλογή αρχικού κατάλληλου συνόλου δεδομένων $S$
2. Καθορισμός της παραμέτρου $k$
3. Καθορισμός της μετρικής απόστασης
4. Για κάθε νέο πρότυπο $X_i$
   (α′) Δημιουργία συνόλου $S_x \subset S$ με τους $k$-κοντινότερους γείτονες
   (β′) Εύρεση της κλάσης που πλειοψηφεί στο σύνολο $S_x$
5. Ταξινόμηση του προτύπου στην αντίστοιχη κλάση




Σχήμα 3.4: Διαδικασία ταξινόμησης νέου στοιχείου με 4NN

Είναι προφανές πως KNN έχει μόνο δύο παραμέτρους που πρέπει να οριστούν:

1. Η παράμετρος $k$
   Δεν υπάρχει κάποια διαδικασία που να καθορίζει τον αριθμό $k$ ανάλογα με το πρόβλημα. Συνήθως ακολουθείται η τεχνική trial and error, δηλαδή γίνονται πολλές επαναλήψεις με διαφορετικές αποτιμήσεις, μέχρι να επιτευχθεί ο επιθυμητός στόχος. Ωστόσο υπάρχουν κάποιες παρατηρήσεις - κατευθυντήριες [52]:
   (α′) επιλογή πολύ μικρής τιμής για το $k$ οδηγεί σε ευαισθησία στα σημεία θορύβου
   (β′) επιλογή πολύ μεγάλης τιμής για το $k$ ίσως δημιουργήσει γειτονιές με ισάριθμα στοιχεία από διαφορετικές κλάσεις
   (γ′) καλή προσέγγιση είναι $k = \sqrt{n}$, όπου $n$ το μέγεθος του αρχικού διαθέσιμου συνόλου
   (δ′) σε πολλά λογισμικά έχει καθοριστεί ως προεπιλογή το $k = 10$
2. Η μετρική της απόστασης
   Η επιλογή της εξαρτάται από το χώρο προτύπων του προβλήματος και υπάρχουν πολλές εναλλακτικές επιλογές [53].




Σχήμα 3.5: Παράδειγμα ταξινόμησης για τον 3NN

(α′) **Ευκλείδεια απόσταση.** Είναι η πιο συχνά χρησιμοποιούμενη μετρική και το μέτρο της εξαρτάται από την κλίμακα μέτρησης κάθε χαρακτηριστικού των προτύπων. Επειδή οι διαφορές στις κλίμακες έχουν μεγάλη επίδραση στον υπολογισμό της απόστασης, συνήθως κάθε χαρακτηριστικό κανονικοποιείται στο διάστημα $[0, 1]$.

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.20}$$

(β′) **Manhattan απόσταση.** Προέκυψε από την απόσταση που διανύει ένα αυτοκίνητο σε μια πόλη (Νέα Υόρκη) που ορίζεται από οικοδομικά τετράγωνα. Είναι πιο ανθεκτική από την Ευκλείδεια στην περίπτωση ύπαρξης outliers αφού οι διαφορές εξομαλύνονται με χρήση της απόλυτης τιμής σε σχέση με την ύψωση στο τετράγωνο.

$$d(X, Y) = \sum_{i=1}^{n} \|x_i - y_i\| \tag{3.21}$$

(γ′) **Minkowski απόσταση.**

$$d(X, Y) = \Big(\sum_{i=1}^{n}(x_i - y_i)^p\Big)^{\frac{1}{p}} \tag{3.22}$$

(δ′) **Chebychev απόσταση.** Σε αντίθεση με τις προαναφερθείσες αποστάσεις, εδώ δε χρησιμοποιούνται όλα τα χαρακτηριστικά του προτύπου αλλά μόνο αυτό με τη μεγαλύτερη απόκλιση. Χρησιμοποιείται κυρίως για την αναγνώριση διαφορών μεταξύ προτύπων σε μια μεταβλητή και εξαρτάται πάρα πολύ από την κλίμακα μέτρησης της αντίστοιχης μεταβλητής.

$$d(X, Y) = \max\{|x_i - y_i|, \quad i = 1, \ldots, n \tag{3.23}$$

(ε′) **Mahalanobis απόσταση.** Λαμβάνει υπόψη τις όποιες διαφορές στην κλίμακα των μεταβλητών όπως επίσης και τις διαφορές στις διακυμάνσεις τους.

$$d(X, Y) = \sqrt{(\mu_X - \mu_Y)^T \cdot S^{-1} \cdot (\mu_X - \mu_Y)} \tag{3.24}$$

όπου $S$ το μητρώο διακυμάνσεων και $\mu_X$ το διάνυσμα με τη μέση τιμή για κάθε χαρακτηριστικό $X$.

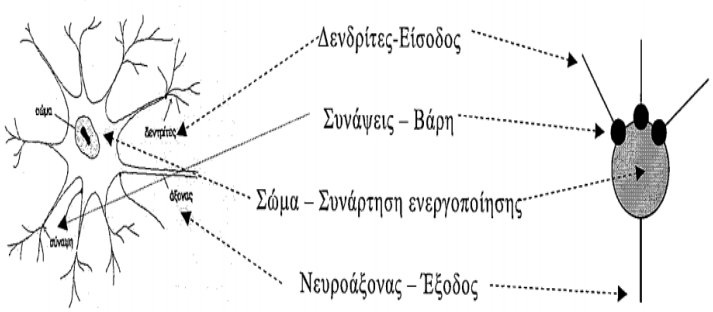
Κλείνοντας πρέπει να αναφέρουμε πως ο αλγόριθμος knn είναι εύκολα υλοποιήσιμος όμως έχει υψηλό υπολογιστικό κόστος για μεγάλα σύνολα γιατί απαιτείται ο υπολογισμός των αποστάσεων μεταξύ του εκάστοτε νέου προτύπου με όλα τα ήδη υπάρχοντα ταξινομημένα πρότυπα. Το γεγονός αυτό αντισταθμίζεται με την ευκολία παραλληλοποίησης του αλγορίθμου. Τέλος η απόδοσή του επηρεάζεται στην περίπτωση που υπάρχουν κλάσεις με

μεγάλη διαφορά ως προς το πλήθος των στοιχείων, αφού η πολυπληθέστερη είναι πιθανό να υπερισχύει κατά την ταξινόμηση των νέων προτύπων.

## 3.4 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα-ΤΝΔ (Artificial Neural Networks - ANNs) αποτελούν υπολογιστικά μοντέλα που προέκυψαν κατά την προσπάθεια προσομοίωσης της λειτουργίας του ανθρώπινου νευρωνικού δικτύου (Σχήμα 3.6). Έχουν σχεδιαστεί έτσι, ώστε να συνδυάζουν τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου με τον αφηρημένο μαθηματικό τρόπο σκέψης. Κατ΄ επέκταση ένα ΤΝΔ μπορεί να μαθαίνει, να θυμάται αλλά και να χρησιμοποιεί περίπλοκες μαθηματικές συναρτήσεις. Οι βασικές αρχές που τα διέπουν είναι η μη γραμμικότητα, η παραλληλία και το μεγάλο πλήθος νευρώνων. Ένα ΤΝΔ ορίζεται ως εξής [54]:

**Ορισμός 3.1.** *«Τεχνητό νευρωνικό δίκτυο είναι μια αρχιτεκτονική δομή που απαρτίζεται από ένα σύνολο διασυνδεδεμένων μονάδων επεξεργασίας. Κάθε τέτοια μονάδα χαρακτηρίζεται από εισόδους και εξόδους. Πραγματοποιεί τοπικά έναν υπολογισμό με βάση τις εισόδους που δέχεται και μεταδίδει το αποτέλεσμα σε άλλες μονάδες επεξεργασίας με τις οποίες συνδέεται. Οι τιμές των βαρών που εντοπιζονται στις συνδέσεις αποτελούν τη γνώση που είναι αποθηκευμένη στο τεχνητό νευρωνικό δίκτυο και καθορίζουν τη λειτουργικότητά του. Συχνά το δίκτυο αναπτύσσει μια συνολική λειτουργικότητα μέσω μιας μορφής εκπαίδευσης.»*




Σχήμα 3.6: Αντιστοιχία βιολογικού-τεχνητού νευρώνα

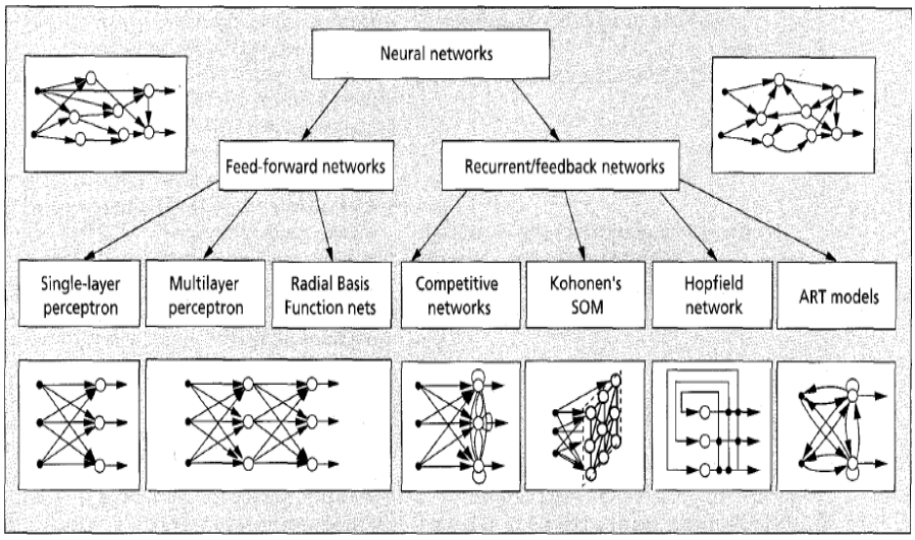
Οι μονάδες επεξεργασίας ονομάζονται νευρώνες και καθένας τους μπορεί να έχει πολλές εισόδους όμως μόνο μία έξοδο. Οι είσοδοί τους είναι είτε οι έξοδοι άλλων νευρώνων, είτε το πρωταρχικό σήμα εισόδου του δικτύου. Ως μία έξοδο εννοούμε τη μοναδικότητα της τιμής και όχι το πλήθος των συνδέσεων του νευρώνα. Η λειτουργία του νευρώνα θα αναλυθεί στην υποενότητα 3.4.1. Οι συνδέσεις μεταξύ των νευρώνων ονομάζονται συνάξεις και ανάλογα με το πόσο σημαντικές είναι, σταθμίζονται με ένα βάρος $w_{ij}$ όπου $i$ ο νευρώνας πομπός και $j$ ο νευρώνας δέκτης. Τα βάρη $w_{ij}$ ονομάζονται συναπτικά βάρη και είναι οι παράμετροι του μοντέλου που ρυθμίζονται μέσω μιας διαδικασίας μάθησης. Εύκολα συμπεραίνουμε πως η απόδοση των ΤΝΔ εξαρτάται πρωτίστως από την αρχιτεκτονική τους, δηλαδή τη διάταξη των συνδέσεων των νευρώνων καθώς και τον αριθμό αλλά και τον τύπο των νευρώνων. Στο Σχήμα 3.7 παρουσιάζονται οι κυριότερες αρχιτεκτονικές ΤΝΔ

### 3.4.1 Το μοντέλο Perceptron

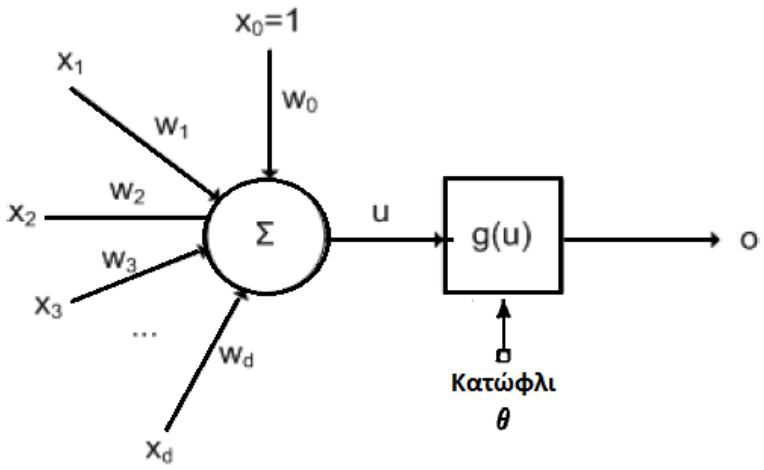
Το πρώτο μοντέλο τεχνητού νευρώνα προτάθηκε από τον Rosenblatt [55], είναι το πιο απλό νευρωνικό δίκτυο, και αποτελεί ένα γραμμικό ταξινομητή δύο κλάσεων. Το μοντέλο αυτό μετά την εκπαίδευσή του θα είναι σε θέση να να αντιστοιχεί δηλαδή κάθε σετ εισόδων που δέχεται στη σωστή κλάση, ορίζοντας ένα υπερεπίπεδο της μορφής

$$u = \sum_{i=1}^{d} w_i x_i + b = 0 \tag{3.25}$$

όπου $b$ ένας όρος πόλωσης. Σύμφωνα με το υπόδειγμα του νευρώνα που παρουσιάζεται στο Σχήμα 3.8, παρατηρούμε ότι υπάρχουν $d$ συνδέσεις εισόδου με τα αντίστοιχα σήματα $x_i$, που το καθένα χαρακτηρίζεται από μια τιμή βάρους $w_i$. Όταν το $w_i$ είναι μεγάλο (μικρό), τότε η συνεισφορά του σήματος είναι μεγάλη (μικρή). Το σήμα εισόδου $x_0$ είναι πάντα 1 και το αντίστοιχο βάρος $w_0$ αντιστοιχεί στη *μεροληψία* (bias) του νευρώνα.




Σχήμα 3.7: Τοπολογίες νευρωνικών δικτύων




Σχήμα 3.8: Το μοντέλο του τεχνητού νευρώνα

Η λειτουργία ενός νευρώνα συνοψίζεται σε δύο κύρια τμήματα με πλήρη εξάρτηση μεταξύ τους. Αρχικά, μέσω ενός αθροιστή, υπολογίζεται η ποσότητα $u(x, w) = \sum_{i=1}^{d} w_i x_i + w_0$ που αντιστοιχεί στο εσωτερικό γινόμενο του διανύσματος εισόδου με το αντίστοιχο των συναπτικών βαρών επαυξημένο με τη μεροληψία του νευρώνα. Αν η ποσότητα $u(x, w)$ ξεπεράσει μια προκαθορισμένη τιμή κατωφλιού, ο νευρώνας λέμε ότι ενεργοποιείται και προχωράμε στο δεύτερο τμήμα όπου γίνεται ο υπολογισμός της εξόδου $o(x)$ του νευρώνα. Η τιμή εξόδου διαμορφώνεται από την τιμή $u(x, w)$ και μια συνάρτηση ενεργοποίησης $g$.

### Μάθηση στο Perceptron

Προκειμένου το μοντέλο Perceptron να είναι σε θέση να ταξινομεί σωστά τα εισερχόμενα πρότυπα εισόδου, πρέπει να εκπαιδευτεί ώστε να μάθει το χώρο του προβλήματος.

**Ορισμός 3.2.** *«Μάθηση (learning) είναι μια διαδικασία με την οποία προσαρμόζονται οι ελεύθερες παράμετροι ενός νευρωνικού δικτύου μέσω μιας συνεχούς διαδικασίας διέγερσης από το περιβάλλον στο οποίο βρίσκεται το δίκτυο.»[56]*

Στο μοντέλο Perceptron χρησιμοποιείται επιβλεπόμενη μάθηση για τη ρύθμιση των συναπτικών βαρών, δηλαδή το δίκτυο τροφοδοτείται από ένα σύνολο της μορφής $(x_i, t_i)$, όπου $x_i$ είναι η είσοδος και $t_i$ η επιθυμητή έξοδος.

Μετά το πέρας της εκπαίδευσης, οι τιμές για τα συναπτικά βάρη και τις μεροληψίες του δικτύου θα πρέπει να έχουν οριστεί έτσι ώστε η έξοδος του δικτύου ($o_i$) να συγκλίνει στην επιθυμητή έξοδο $t_i$. Στη συνέχεια παρατίθεται η διαδικασία που ακολουθείται κατά την εκπαίδευση του Perceptron [57] για ένα πρόβλημα ταξινόμησης δύο κλάσεων ($c_1$ και $c_2$).

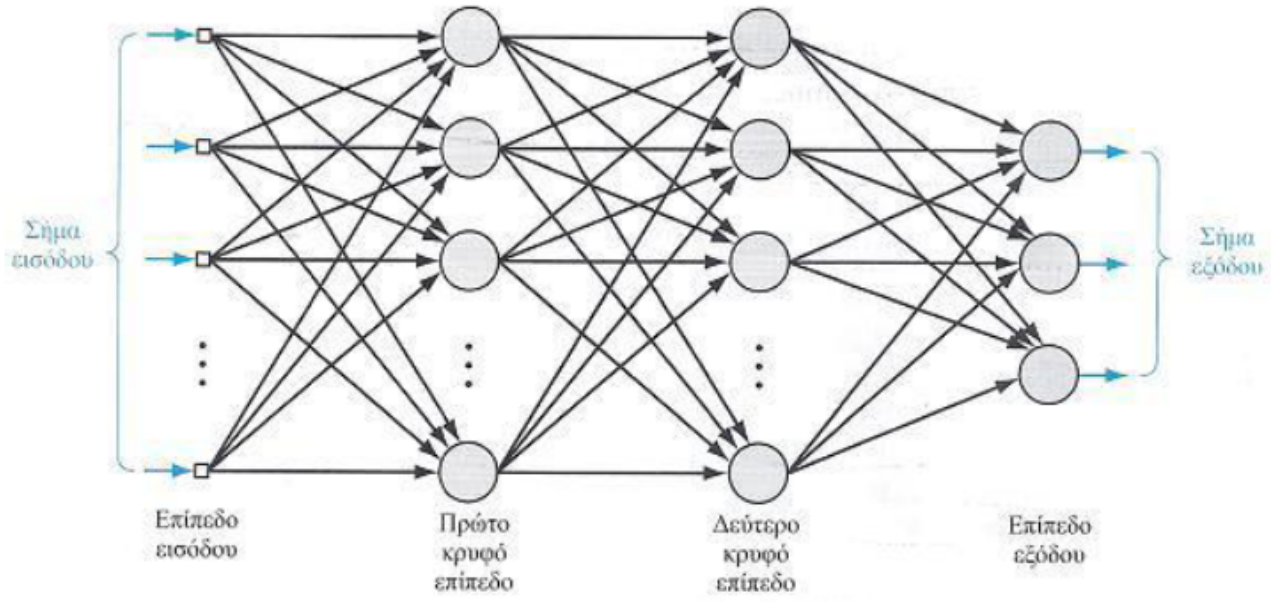
1. Αρχικοποίηση του $w(n) = (w_1, w_2, \ldots, w_n)$ με τυχαίες τιμές στο διάστημα [0,1]
2. Τροφοδότηση του μοντέλου με τα πρότυπα $x(n) = (x_1, x_2, \ldots, x_n)$.
3. Για κάθε πρότυπο η έξοδος δίνεται ως
$$o(n) = w(n)^T x(n)$$
για την οποία υπάρχει ένα διάνυσμα βαρών ώστε
$$w^T x \geq 0 \quad \forall x \in c_1 \quad \text{ή} \quad w^T x < 0 \quad \forall x \in c_2$$
4. Έλεγχος διόρθωσης των βαρών:
   (α′) Δε γίνεται αλλαγή στα βάρη.
   α) αν $o(n) \geq 0$ και $x(n) \in c_1 \Rightarrow w(n+1) = w(n)$
   β) αν $o(n) < 0$ και $x(n) \in c_2 \Rightarrow w(n+1) = w(n)$
   (β′) Γίνεται αλλαγή στα βάρη.
   α) αν $o(n) \geq 0$ και $x(n) \in c_2 \Rightarrow w(n+1) = w(n) - \eta(n)x(n)$ ·
   β) αν $o(n) < 0$ και $x(n) \in c_1 \Rightarrow w(n+1) = w(n) + \eta(n)x(n)$
   Το $\eta(n)$ αντιστοιχεί στο ρυθμό μάθησης.
5. Ο αλγόριθμος τερματίζεται όταν δε γίνεται καμία διόρθωση στα βάρη.

### 3.4.2 Δίκτυα Πολλαπλών Επιπέδων

Το δίκτυο Perceptron πολλών επιπέδων (Multilayer Perceptron-MLP) είναι μια γενίκευση του Perceptron και ανήκει στην κατηγορία δικτύων πρόσθιας τροφοδότησης. Δηλαδή το σήμα μέσα στο δίκτυο κινείται μόνο με κατεύθυνση από την είσοδο προς την έξοδο χωρίς να υπάρχει κάποιος νευρώνας ανατροφοδότησης. Η διαδικασία περιγράφεται εν περιλήψει ως εξής: οι νευρώνες του δικτύου τροφοδοτούν αποκλειστικά τους νευρώνες του στρώματος ακολουθεί και τροφοδοτούνται αποκλειστικά από τους νευρώνες του προηγούμενου. Σε αντίθεση με το Perceptron που αποτελείται από το επίπεδο εξόδου και ένα νευρώνα στο επίπεδο εξόδου, εδώ μπορούμε να έχουμε παραπάνω από ένα νευρώνα στο επίπεδο εξόδου καθώς και τουλάχιστον ένα ακόμα επίπεδο ανάμεσα σε αυτά της εισόδου και της εξόδου [58]. Τα ενδιάμεσα επίπεδα ονομάζονται *κρυφά επίπεδα*. Στο Σχήμα 3.9 παρουσιάζεται ένα πολυεπίπεδο δίκτυο (Perceptron) με δύο κρυφά επίπεδα.

Στο επίπεδο εισόδου δεν πραγματοποιείται κάποιος υπολογισμός, οι νευρώνες λειτουργούν ως υποδοχείς του εισερχόμενου σήματος, το οποίο μεταβιβάζουν στους νευρώνες του πρώτου κρυφού επιπέδου. Ο αριθμός των νευρώνων στο επίπεδο εισόδου είναι ίσος με τον αριθμό των χαρακτηριστικών κάθε προτύπου του προβλήματος. Δηλαδή αν τα πρότυπα του προβλήματός μας αναπαριστώνται ως διάνυσμα 15 χαρακτηριστικών, θα έχουμε 15 νευρώνες στο επίπεδο εισόδου. Στα κρυφά επίπεδα η επιλογή του πλήθους των νευρώνων παραμένει ανοιχτό πρόβλημα και αποτελεί επιλογή του χρήστη του δικτύου, συνήθως μετά από μια διαδικασία trial and error. Οι μη γραμμικοί νευρώνες που χρησιμοποιούνται στα κρυφά επίπεδα, επιτρέπουν στο δίκτυο να «μαθαίνει» και να εκτελεί περίπλοκες εργασίες, εξάγοντας προοδευτικά τα χαρακτηριστικά εκείνα των προτύπων εισόδου που έχουν τη μεγαλύτερη σημασία για τη σωστή απόκριση του δικτύου. Στο επίπεδο εξόδου ο αριθμός των νευρώνων συχνά καθορίζεται να είναι ίσος με τον αριθμό των κλάσεων για τα προβλήματα ταξινόμησης. Το είδος της συνάρτησης ενεργοποίησης ποικίλει ανάλογα με το πρόβλημα. Παραδείγματος χάρη, σε πρόβλημα προσέγγισης συνάρτησης επιλέγονται συνήθως γραμμικοί νευρώνες, ενώ σε προβλήματα ταξινόμησης επιλέγονται κατά κύριο λόγο μη γραμμικοί νευρώνες.

Η μετάβαση από κάθε επίπεδο στο επόμενο, αντιστοιχεί στο μετασχηματισμό του προτύπου σε ένα χώρο διάστασης ίσης με τον αριθμό των νευρώνων του επιπέδου όπου μεταφέρεται το σήμα. Οι αναπαραστάσεις των προτύπων σε διαφορετικούς χώρους βοηθάει στην ανακάλυψη κρυφών δομών στα δεδομένα που δεν είναι εμφανή στον αρχικό χώρο, για την αντιμετώπιση ενός προβλήματος με τη μεγαλύτερη δυνατή επιτυχία. Εύκολα διαπιστώνει κανείς ότι, όσο μεγαλύτερος είναι ο αριθμός των κρυφών επιπέδων και όσο περισσότεροι οι νευρώνες σε καθένα από αυτά, τόσο πιο σύνθετες μπορούν να γίνουν και οι αναπαραστάσεις των αρχικών δεδομένων και συνεπώς, τόσο μεγαλύτερη η δύναμη του δικτύου στο να ταξινομεί δεδομένα. Αν και έχει αποδειχθεί ότι ένα δίκτυο πρόσθιας τροφοδότησης με ένα κρυφό επίπεδο μπορεί να προσεγγίσει κάθε συνεχή συνάρτηση με οποιαδήποτε επιθυμητή ακρίβεια [59], το πρόβλημα της Ανάλυσης Συναισθήματος δε μπορεί να αντιμετωπιστεί με δίκτυα λίγων κρυφών επιπέδων, έτσι η επιστημονική κοινότητα έχει στραφεί στη δημιουργία βαθιών νευρωνικών δικτύων.




Σχήμα 3.9: Πολυεπίπεδο δίκτυο

### 3.4.3 Εκπαίδευση Δικτύων Πολλαπλών Επιπέδων

Η εκπαίδευση των MLP βασίζεται σε μια απλή ιδέα. Το δίκτυο αρχικοποιείται με τυχαίους αριθμούς στο διάστημα $[0, 1]$ για κάθε συναπτικό βάρος του. Στη συνέχεια τροφοδοτείται με τα πρότυπα εισόδου $x_i$ και δημιουργείται σφάλμα ανάμεσα στην επιθυμητή έξοδο $y_i$ και την έξοδο του δικτύου $o_i$. Τα βάρη διορθώνονται με σκοπό να ελαχιστοποιηθεί το σφάλμα μέχρι να φτάσει την ακρίβεια που έχει τεθεί από το χρήστη [60]. Η τροφοδότηση όλων των προτύπων μαζί με τις διορθώσεις που προκύπτουν, ονομάζεται εποχή εκπαίδευσης και επαναλαμβάνεται αρκετές φορές μέχρι να επιτευχθεί η επιθυμητή ακρίβεια για το σφάλμα του δικτύου.

Προκύπτουν έτσι δύο ερωτήματα που αφορούν αφενός την επιλογή μιας κατάλληλης μετρικής αξιολόγησης της εξόδου του δικτύου και αφετέρου την επιλογή ενός αλγορίθμου εκπαίδευσης που θα αναπροσαρμόζει τα συναπτικά βάρη. Η διαφορά ανάμεσα στην επιθυμητή έξοδο και αυτή του νευρωνικού δικτύου, είναι η ποσότητα που θέλουμε να ελαχιστοποιήσουμε. Σαν συνάρτηση σφάλματος μπορούμε να χρησιμοποιήσουμε το μέσο τετραγωνικό σφάλμα (mean squared error)

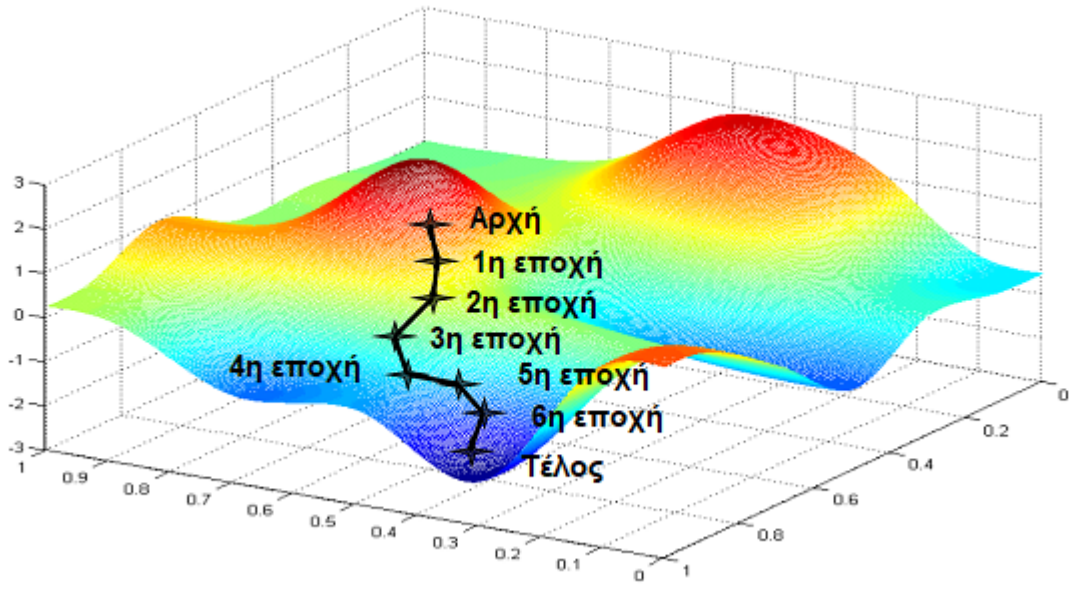
$$J = \frac{1}{2N} \sum_{n=1}^{N} \sum_{i=1}^{m} [o_i^{(n)} - y_i^{(n)}]^2 \tag{3.26}$$

ή το σφάλμα διεντροπίας (cross-entropy error)

$$J = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{m} y_{in} \log(o_{in}) \tag{3.27}$$

όπου $N$ το μέγιστο πλήθος των εποχών εκπαίδευσης και $m$ το πλήθος των νευρώνων στο επίπεδο εξόδου. Η σχέση 3.27 υπολογίζει την απόσταση δύο κατανομών πιθανοτήτων. Διαισθητικά η σχέση 3.26 μπορεί να αναπαρασταθεί ως μια πολυδιάστατη επιφάνεια με συντεταγμένες τα συναπτικά βάρη του δικτύου. Για να βελτιώνεται η απόδοση του δικτύου πρέπει το σημείο, που αντιστοιχεί στο διάνυσμα των συναπτικών βαρών, να κινείται προς το ελάχιστο της επιφάνειας (Σχήμα 3.10).

Αφού υπολογιστεί η απόκλιση του δικτύου από τα πρότυπα εισόδου, επιστρατεύεται ένας αλγόριθμος εκπαίδευσης για την αναπροσαρμογή των συναπτικών βαρών. Το πρόβλημα της εκπαίδευσης ενός ΤΝΔ ανάγεται στην ουσία στο πρόβλημα ελαχιστοποίησης της συνάρτησης σφάλματος. Δεν υπάρχει μοναδικός αλγόριθμος ελαχιστοποίησης που η χρήση του να έχει αποδειχθεί πανάκεια για όλες τις περιπτώσεις. Αντίθετα, υπάρχει ένα σύνολο αλγορίθμων που χρησιμοποιούνται κατά περίπτωση ανάλογα, με τα πλεονεκτήματα και τα μειονεκτήματα που παρουσιάζουν κατά την εφαρμογή τους σε ένα πρόβλημα. Η μεγαλύτερη διαφορά μεταξύ των αλγορίθμων είναι ο τρόπος με τον οποίο προσαρμόζουν τα βάρη $w_{ij}$ των συνάψεων των νευρώνων του δικτύου. Στην υποενότητα 3.4.3 θα περιγράψουμε τον πιο συχνά χρησιμοποιούμενο αλγόριθμο εκπαίδευσης πολυεπίπεδων δικτύων, Backpropagation.




Σχήμα 3.10: Το διάνυσμα συναπτικών βαρών στην επιφάνεια σφάλματος

Για την εκπαίδευση του δικτύου υπάρχουν οι παρακάτω τρόποι:

1. *Τρόπος προτύπων* (Pattern mode): Οι προσαρμογές στα βάρη εκτελούνται μετά την παρουσίαση κάθε διανύσματος εκπαίδευσης [61].
2. *Μαζική μάθηση* (Batch mode): Οι προσαρμογές στα βάρη επιτελούνται μετά από την παρουσίαση όλων των παραδειγμάτων εκπαίδευσης [62].

Τα επικρατέστερα κριτήρια τερματισμού της διαδικασίας εκπαίδευσης είναι η επιλογή της τιμής κατωφλίου για τη συνάρτηση σφάλματος ή ο καθορισμός του μέγιστου αριθμού εποχών. Επιπλέον, υπάρχει και η ικανότητα να χρησιμοποιηθεί ένα σύνολο επικύρωσης ώστε να ελέγχουμε την απόδοση του δικτύου σε ξένα πρότυπα κατά τη διάρκεια της εκπαίδευσης. Η εκπαίδευση, σε αυτή την περίπτωση, σταματά αν η απόδοση στο σύνολο επικύρωσης δεν βελτιώνεται για συνεχόμενες εποχές εκπαίδευσης.

### Αλγόριθμος Backpropagation

Κατά τον αλγόριθμο Backpropagation παρουσιάζονται στο δίκτυο ένα σύνολο παραδειγμάτων εκπαίδευσης $D = (x^n, t^n), n = 1, 2, \ldots, d$ όπου $x^n = (x_1^n, \ldots, x_d^n) \in \mathbb{R}^d$ η είσοδος του δικτύου και $t^n = (t_1^n, \ldots, t_p^n) \in \mathbb{R}^p$ η επιθυμητή έξοδος. Το δίκτυο με βάση τα διανύσματα εισόδου υπολογίζει την έξοδο $o(x^n, w)$ όπου $w = (w_1, \ldots, w_L)^T$ το διάνυσμα στο οποίο συγκεντρώνουμε όλα τα βάρη και τις πολώσεις του δικτύου.

Η διαδικασία της οπίσθιας διάδοσης του σφάλματος αποτελείται από δύο περάσματα διαμέσου των διαφορετικών επιπέδων του δικτύου, ένα προς τα εμπρός και ένα προς τα πίσω πέρασμα [63]:

1. Έστω λοιπόν ένα ΤΝΔ με $d$ εισόδους, $p$ εξόδους και $H$ κρυμμένα επίπεδα. Το σήμα στο προς τα εμπρός πέρασμα μεταδίδεται στα διάφορα επίπεδα ακολουθώντας τα παρακάτω βήματα:

   (α′) Ένα διάνυσμα εισόδου εφαρμόζεται στους νευρώνες εισόδου του δικτύου.

   $$y_i^{(0)}(n) = x_i(n), \quad y_0^{(0)} = x_0 = 1 \tag{3.28}$$

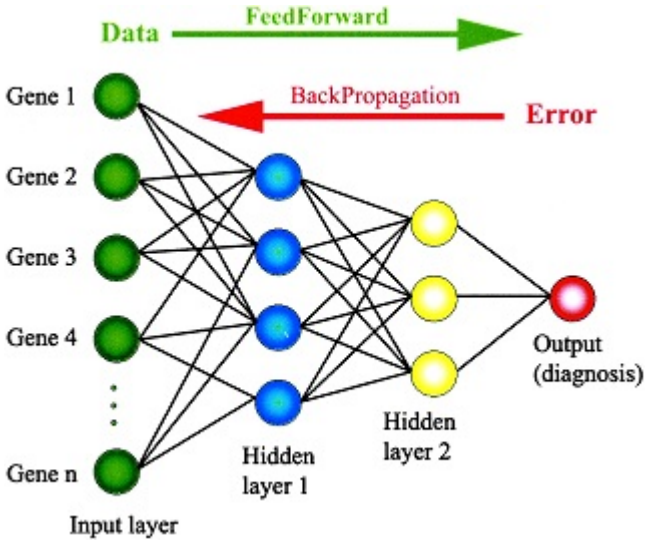
   (β′) Η επίδρασή του διαδίδεται μέσα στο δίκτυο από επίπεδο σε επίπεδο.

   $$u_i^{(h)}(n) = \sum_{j=0}^{d_h} w_i^h(n) y_j^{(h-1)}(n), \quad h = 1, \ldots, H+1, \quad i = 1, \ldots, d_h \tag{3.29}$$

   $$y_i^{(h)} = g_h(u_i^{(h)}(n)), \quad h = 1, \ldots, H+1, \quad i = 1, \ldots, d_h, \quad y_0^{(h)} = 1 \tag{3.30}$$

   (γ′) Ένα σύνολο από εξόδους παράγεται ως η πραγματική απόκριση του δικτύου.

   $$o_i(n) = y_i^{(H+1)}(n), \quad i = 1, \ldots, p \tag{3.31}$$




Σχήμα 3.11: Οπισθοδιάδοση σφάλματος

2. Στο πίσω πέρασμα το σήμα σφάλματος μεταδίδεται κατά την αντίθετη φορά και υπολογίζονται τα επιμέρους σήματα σφάλματος που αντιστοιχούν στους κρυμμένους νευρώνες:

   (α′) Η πραγματική απόκριση του δικτύου αφαιρείται από την επιθυμητή απόκριση για την παραγωγή ενός σήματος λάθους.

   $$\delta_i^{(H+1)}(n) = (t_i(n) - o_i(n))o_i(n)(1 - o_i(n)) \tag{3.32}$$

   (β′) Το σήμα λάθους διαδίδεται προς τα πίσω στο δίκτυο.

   $$\delta_i^{(h)}(n) = y_i^{(h)}(n)(1 - y_i^{(h)}(n)) \sum_k \delta_k^{(h+1)}(n) w_{ki}^{(h+1)}(n), \tag{3.33}$$
   $$h = H, \ldots, 1, \quad i = 1, \ldots, d_h$$

   (γ′) Τα βάρη προσαρμόζονται σε συμφωνία με τον κανόνα διόρθωσης λάθους.

   $$w_{ij}^{(h)}(n+1) = w_{ij}^{(h)}(n) + \alpha[w_{ij}^{(h)}(n) - w_{ij}^{(h)}(n-1)] + \eta\delta_i^{(h)}(n) y_j^{(h-1)}(n) \tag{3.34}$$

Οι συμβολισμοί που χρησιμοποιήθηκαν έχουν ως εξής:

$w^{(l)}$ το διάνυσμα βαρών των συνάψεων ενός νευρώνα στο επίπεδο $l$

$u^{(l)}$ το διάνυσμα εσωτερικής δραστηριότητας των νευρώνων στο επίπεδο $l$

$y^{(l)}$ το διάνυσμα σήματος εξόδων των νευρώνων του επιπέδου $l$

$\delta^{(l)}$ το διάνυσμα τοπικών κλίσεων των νευρώνων στο επίπεδο $l$

Οι παραπάνω υπολογισμοί θα επαναληφθούν μέχρις ότου οι ελεύθερες παράμετροι του δικτύου σταθεροποιηθούν και η συνάρτηση σφάλματος αποκτήσει μια αποδεκτή τιμή. Η σειρά που θα τροφοδοτούνται τα πρότυπα στο δίκτυο σε κάθε εποχή θα πρέπει να μεταβάλλεται τυχαία. Στη συνέχεια παρατίθεται ένα κριτήριο σύγκλισης του αλγόριθμου Backpropagation.

**Κριτήριο 3.1.** *Ο αλγόριθμος Backpropagation θεωρείται ότι συγκλίνει όταν η Ευκλείδεια νόρμα του διανύσματος κλίσης φτάσει ένα επαρκώς μικρό κατώφλι:*

$$\| \bigtriangledown E(w_k) \| \leq \epsilon$$

### 3.4.4 Συναρτήσεις Ενεργοποίησης Νευρώνα

Η *συνάρτηση ενεργοποίησης* (activation function) είναι μια γραμμική ή μη γραμμική συνεχής συνάρτηση η οποία επιλέγεται, έτσι ώστε να περιορίσει την έξοδο του συστήματος σε ένα επιθυμητό διάστημα. Κατά το πλείστον χρησιμοποιούμε συναρτήσεις που είναι συνεχείς και παραγωγίσιμες γιατί οι περισσότεροι αλγόριθμοι ελαχιστοποίησης

που χρησιμοποιούνται κατά τη διάρκεια της εκπαίδευσης αξιοποιούν την πληροφορία των παραγώγων. Στον Πίνακα 3.1 παρατίθενται μερικές από τις πιο ευρέως χρησιμοποιούμενες συναρτήσεις ενεργοποίησης. Τέλος δεν πρέπει να παραληφθεί η συνάρτηση Softmax που χρησιμοποιείται στους νευρώνες του επιπέδου εξόδου, κυρίως σε προβλήματα ταξινόμησης. Το ιδιαίτερο χαρακτηριστικό της είναι ότι απεικονίζει κάθε πρότυπο εισόδου σε ένα διάνυσμα που τα χαρακτηριστικά του είναι όσα και ο αριθμός των κλάσεων ταξινόμησης και οι τιμές κάθε χαρακτηριστικού κυμαίνονται στο διάστημα $[0, 1]$. Οι τιμές των χαρακτηριστικών μπορούν να θεωρηθούν ως πιθανότητες ταξινόμησης του εκάστοτε προτύπου στην αντίστοιχη κλάση. Ο τύπος της δίνεται ως

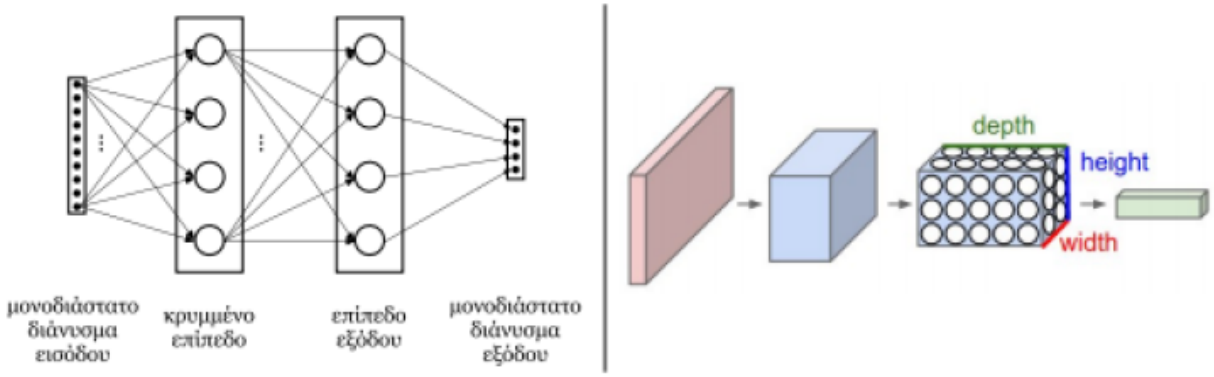
$$\phi_i(s) = \frac{e^{s_i}}{\sum_{j=1}^{k} e^{s_j}}$$

όπου $i$ ο εκάστοτε νευρώνας εξόδου και $k$ το πλήθος των νευρώνων στο επίπεδο εξόδου.

## 3.5 Βαθιά Νευρωνικά Δίκτυα

Η Βαθιά Μάθηση αποκρυπτογραφεί τις πληροφορίες μέσω της επισήμανσης και της εκχώρησης των αντικειμένων σε διάφορες κατηγορίες, διαδικασία που ακολουθεί και ο ανθρώπινος εγκέφαλος. Η διαφορά των δικτύων Βαθιάς μάθησης με τα απλά τεχνητά νευρωνικά δίκτυα είναι ότι τα πρώτα περιέχουν πιο σύνθετες συνδέσεις μεταξύ των νευρώνων αλλά κυρίως έχουν πολύ μεγαλύτερο αριθμό κρυφών επιπέδων σε σχέση με τα δεύτερα, γεγονός που τους επιτρέπει να αναγνωρίζουν την πολύπλοκη δομή σε μεγάλα, μη δομημένα, σύνολα δεδομένων. Τα βαθιά νευρωνικά δίκτυα [64] βασιζόμενα στο ότι πολλά φυσικά σήματα είναι ιεραρχίες σύνθεσης, μετατρέποντάς τα χαρακτηριστικά χαμηλότερου επιπέδου σε μια κατάλληλη και ευκόλως επεξεργάσιμη εσωτερική δομή, αναδεικνύουν χαρακτηριστικά υψηλότερου επιπέδου. Έχει αποδειχθεί ότι έχουν μεγαλύτερη ακρίβεια από τα τεχνητά νευρωνικά δίκτυα, απαιτούν αρκετό χρόνο εκπαίδευσης και χρειάζονται τεράστιο όγκο δεδομένων εκπαίδευσης, μαζί με ακριβό υλικό. Όμως με τη βελτίωση της τεχνολογίας και τις τεχνικές επιτάχυνσης μέσω χρήσης GPU, τα δίκτυα αυτής της κατηγορίας εκπαιδεύονται σε αποδεκτούς χρόνους.

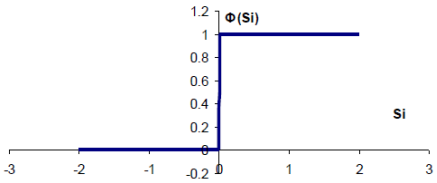
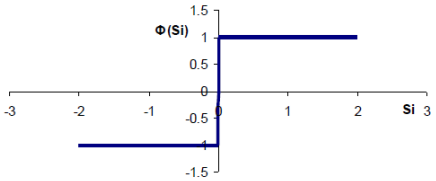
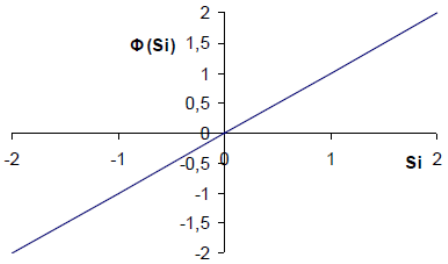
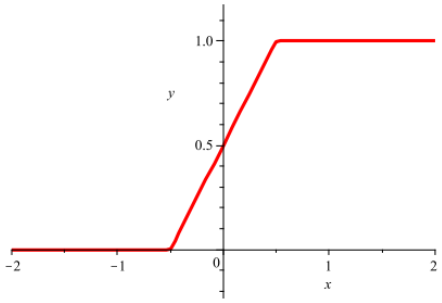
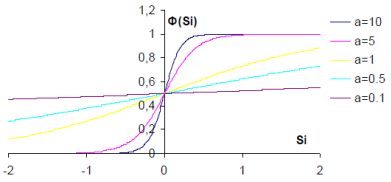
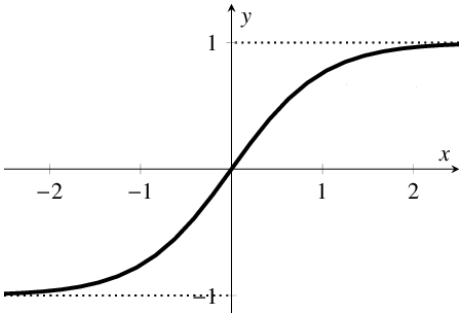
### 3.5.1 Συνελικτικά Νευρωνικά Δίκτυα




Σχήμα 3.12: Αρχιτεκτονική ΤΝΔ vs ΣΝΔ

Τα Συνελικτικά Νευρωνικά Δίκτυα - ΣΝΔ (Convolutional Neural Networks - CNN) αποτελούν αρχιτεκτονικές εμπνευσμένες από τον τρόπο με τον οποίο λειτουργεί το οπτικό σύστημα του ανθρώπινου εγκεφάλου, δηλαδή οι νευρώνες τους είναι οργανωμένοι κατά τέτοιο τρόπο ώστε να αποκρίνονται σε επικαλυπτόμενες περιοχές του οπτικού πεδίου. Έχουν ευρεία και ιδιαιτέρως αποδοτική χρήση σε προβλήματα υπολογιστικής όρασης όπως η αναγνώριση προσώπων σε στατική εικόνα ή βίντεο. Η αρχιτεκτονική τους αποτελείται από τρία είδη επιπέδων, το επίπεδο συνέλιξης (convolutional layer), το επίπεδο συγκέντρωσης (pooling layer) και από ένα σύνολο από πλήρως συνδεδεμένα επίπεδα όπως αυτά που συναντάμε στα πολυεπίπεδα ΤΝΔ πρόσθιας τροφοδότησης.

Τα ΣΝΔ διαφέρουν από τα ΤΝΔ ως προς τις συνδέσεις μεταξύ των νευρώνων και την πράξη που γίνεται στους νευρώνες. Εδώ κάθε νευρώνας συνδέεται με ένα υποσύνολο των νευρώνων του προηγούμενου επιπέδου αντί για το εσωτερικό γινόμενο του σήματος με το διάνυσμα των συναπτικών βαρών, εφαρμόζεται η συνέλιξη που θα αναλυθεί στη συνέχεια. Τέλος, συμπληρώνουμε πως η εκπαίδευση τέτοιου είδους δικτύων είναι αρκετά απλή και μπορούν να χρησιμοποιηθούν αλγόριθμοι που χρησιμοποιούνται και στα ΤΝΔ. Ακολουθεί περιγραφή των επιπέδων ενός ΣΝΔ.

| Όνομα | Τύπος | Σχήμα |
|---|---|---|
| Βηματική<br><br>(Hard-Limit) | $\phi(s) = \begin{cases} 1, & \text{αν} \quad s > 0 \\ 0, & \text{αν} \quad s \leq 0 \end{cases}$ |  |
| Προσήμου<br><br>(Symmetric Hard-Limit) | $\phi(s) = \begin{cases} 1, & \text{αν} \quad s > 0 \\ -1, & \text{αν} \quad s \leq 0 \end{cases}$ |  |
| Γραμμική<br><br>(Linear) | $\phi(s) = \lambda s$ |  |
| Τμηματικά γραμμική<br><br>(Piecewise Linear) | $\phi(s) = \begin{cases} 1, & \text{αν} \quad s \geq \frac{1}{2} \\ s, & \text{αν} \quad -\frac{1}{2} < s < \frac{1}{2} \\ 0, & \text{αν} \quad s \leq -\frac{1}{2} \end{cases}$ |  |
| Σιγμοειδής<br><br>(Sigmoid) | $\phi(s) = \frac{1}{1+e^{-\alpha s}}$ |  |
| Υπερβολική εφαπτομένη<br><br>(Hyperbolic tangent) | $\phi(s) = \tanh(s) = \frac{1-e^{-s}}{1+e^{-s}}$ |  |

Πίνακας 3.1: Συναρτήσεις ενεργοποίησης

1. **Επίπεδο εισόδου.** Η είσοδος στα ΤΝΔ είναι ένα διάνυσμα χαρακτηριστικών ενώ στα ΣΝΔ η είσοδος είναι οργανωμένη σε ένα δισδιάστατο ή τρισδιάστατο μητρώο (Σχήμα 3.12). Η λειτουργία του επιπέδου εισόδου είναι απλά να μεταβιβάζει το σύνολο εισόδου που έχει διάσταση $W \times H \times D$, στο επόμενο επίπεδο. Το $W$ αντιστοιχεί στις στήλες του μητρώου εισόδου, το $H$ αντιστοιχεί στις γραμμές του μητρώου εισόδου και το $D$ στο βάθος. Αν το διάνυσμα εισόδου είναι μια εικόνα σε grayscale έχουμε ότι $D = 1$, ενώ σε RGB έχουμε

ότι $D = 3$.

2. **Επίπεδο συνέλιξης.** Ο σκοπός των συνελικτικών επιπέδων είναι η εξαγωγή χαρακτηριστικών και η σχέση μεταξύ των θέσεων τους. Σε κάθε επίπεδο συνέλιξης εφαρμόζονται $K$ φίλτρα πάνω στο εισερχόμενο μητρώο ώστε να δημιουργηθούν οι αντίστοιχοι χάρτες χαρακτηριστικών (feature maps) [65]. Σε κάθε επίπεδο εφαρμόζεται διαφορετικό φίλτρο που το καθένα αντιστοιχεί σ᾽ ένα χαρακτηριστικό. Τα φίλτρα είναι συνήθως συμμετρικά, μεγέθους $F$, διάστασης $F \times F$ και αποτελούνται από διαφορετικές παραμέτρους, οι οποίες ρυθμίζονται μέσω της εκπαίδευσης. Αν το φίλτρο δεν είναι συμμετρικό, πρέπει να γίνει κατοπτρισμός ως προς το κεντρικό στοιχείο. Η εφαρμογή του φίλτρου σε κάθε τμήμα του μητρώου δημιουργεί μία μόνο τιμή στο χάρτη χαρακτηριστικών. Αν το μητρώο εισόδου έχει διάσταση $W \times H \times D$, με $D = 1$, θα έχουμε $K$ χάρτες χαρακτηριστικών (ένα χάρτη χαρακτηριστικών για κάθε φίλτρο). Αν $D > 1$, για κάθε φίλτρο θα έχουμε $D$ χάρτες χαρακτηριστικών, οι οποίοι αθροίζονται κατά στοιχείο για να προκύψει ένας χάρτης χαρακτηριστικών ανά φίλτρο. Ο υπολογισμός που επιτελείται για τον καθορισμό της τιμής στο χάρτη χαρακτηριστικών περιλαμβάνει τον πολλαπλασιασμό στοιχείο προς στοιχείο μεταξύ του φίλτρου και του αντίστοιχου σημείου του μητρώου και μετά την άθροιση των παραπάνω γινομένων (Σχέση 3.35).

$$y(i,j) = (x * h)(i,j) = \sum_{m,n} x(m,n)h(i-m, j-n) \tag{3.35}$$

όπου $x$ το μητρώο και $h$ το φίλτρο.

Σε κάθε στοιχείο του χάρτη χαρακτηριστικών εφαρμόζεται μια πόλωση και έπειτα κάθε στοιχείο περνά από μια μη γραμμική συνάρτηση. Οι επικρατέστερες είναι οι:

| | |
|---|---|
| ReLU | $f(x) = \max(0, x)$ |
| Συνάρτηση υπερβολικής εφαπτομένης | $f(x = \tanh(x)$ |
| Λογιστική Συνάρτηση | $f(x) = \dfrac{1}{1+e^{-x}}$ |

Συνοψίζοντας, έστω ότι έχουμε

(α′) είσοδο διάστασης $W \times H \times D$

(β′) $K$ φίλτρα διάστασης $F \times F$

(γ′) $w_{ijk}$ τα βάρη των φίλτρων με $i = 1, 2, \ldots F$, $j = 1, 2, \ldots F$, $k == 1, 2, \ldots D$

(δ′) $S$ το βήμα (stride) που καθορίζει τον αριθμό των στοιχείων που μετακινούνται οριζόντια και κάθετα

(ε′) $P$ το γέμισμα (padding) με προσθήκη μηδενικών που χρησιμοποιείται για τη διατήρηση των χωρικών διαστάσεων εισόδου ώστε να εφαρμοστεί το φίλτρο μεγέθους $F$

προκύπτει ο χάρτης χαρακτηριστικών[66] $W' \times H' \times D$ όπου

$$W' = \frac{W - F + 2P}{S} + 1 \tag{3.36}$$

$$H' = \frac{H - F + 2P}{S} + 1 \tag{3.37}$$

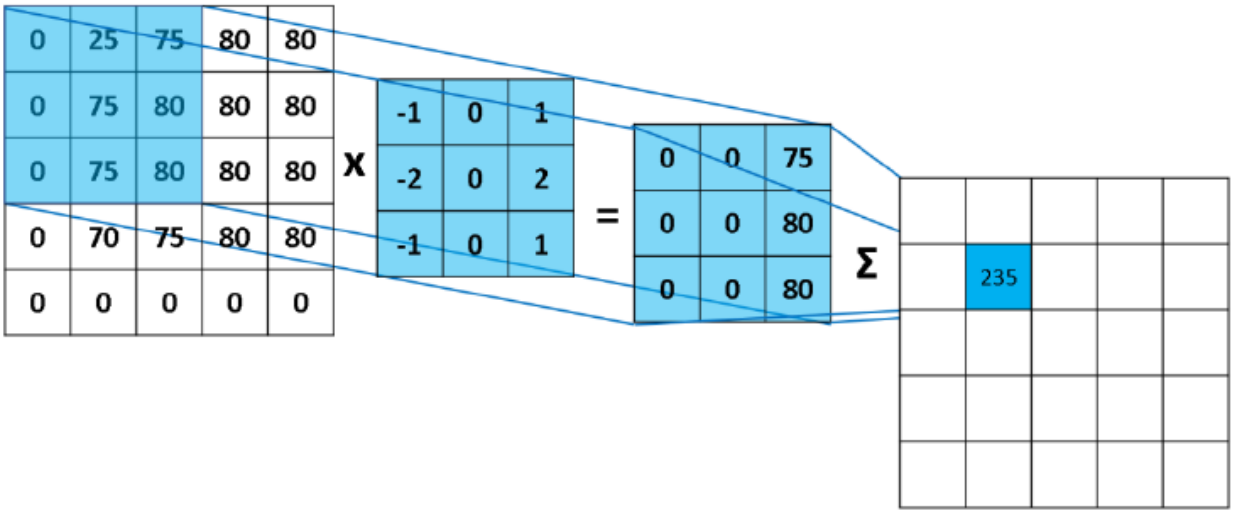
που μετά την άθροιση κατά στοιχείο για το $D$ (αν $D > 1$) , θα έχει τη μορφή $W' \times H'$. Είναι εύκολα αντιληπτό (Σχήμα 3.13) πως ο χάρτης χαρακτηριστικών θα έχει μικρότερη διάσταση από αυτή του αρχικού μητρώου.

3. **Επίπεδο συγκέντρωσης.** Το επίπεδο συγκέντρωσης συνήθως βρίσκεται ενδιάμεσα σε δύο συνελικτικά επίπεδα. Στα επίπεδα συγκέντρωσης γίνεται μείωση των διαστάσεων του μητρώου, διατηρώντας όμως τα σημαντικά χαρακτηριστικά. Δηλαδή, έχοντας ως είσοδο ένα χάρτη χαρακτηριστικών, τον χωρίζει σε ορθογώνιες μη επικαλυπτόμενες υποπεριοχές και για κάθε τέτοια περιοχή εφαρμόζεται μια πράξη η οποία επιτρέπει στο δίκτυο να αναγνωρίζει την ύπαρξη ή όχι ενός χαρακτηριστικού. Οι συνηθέστερες πράξεις είναι η εύρεση μεγίστου, μέσου όρου ή αθροίσματος (Πίνακας 3.2). Με αυτό τον τρόπο μειώνεται προοδευτικά τόσο το μέγεθος της αναπαράστασης όσο και ο αριθμός των προσαρμόσιμων παραμέτρων και κατ᾽ επέκταση ο αριθμός των υπολογισμών.

Συνοψίζοντας, έστω ότι έχουμε

(α′) είσοδο διάστασης $W' \times H' \times D'$

(β′) $D'$ το πλήθος των φίλτρων στο προηγούμενο επίπεδο συνέλιξης




Σχήμα 3.13: Η διαδικασία της συνέλιξης

(γ′) $F$ το μέγεθος παραθύρου

(δ′) το βήμα $S = F$

προκύπτει ο χάρτης χαρακτηριστικών[66] $W'' \times H'' \times D'$ όπου

$$W'' = \frac{W' - F}{S} + 1 \tag{3.38}$$

$$H'' = \frac{H' - F}{S} + 1 \tag{3.39}$$

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | ↗ | 20 30<br>112 37 | Μέγιστο |
| 12 | 20 | 30 | 0 | | | |
| 8 | 12 | 2 | 0 | → | 13 8<br>79 20 | Μέσος όρος |
| 34 | 70 | 37 | 4 | | | |
| 112 | 100 | 25 | 12 | | | |
| | | | | ↘ | 52 32<br>316 78 | Άθροισμα |

Πίνακας 3.2: Η διαδικασία της συγκέντρωσης

4. **Πλήρως συνδεδεμένα επίπεδα.** Τα πλήρως συνδεδεμένα επίπεδα, όπως αυτά που χρησιμοποιούνται στο πολυεπίπεδο Perceptron, τοποθετούνται μετά από αρκετές διαδοχικές επανεμφανίσεις του μοτίβου επίπεδο συνέλιξης - επίπεδο συγκέντρωσης. Η έξοδος του προηγούμενου επιπέδου, που συνήθως είναι επίπεδο συγκέντρωσης, συνδέεται πλήρως με κάθε νευρώνα του πρώτου πλήρως συνδεδεμένου επιπέδου. Η τοποθέτηση των πλήρως συνδεδεμένων επιπέδων στο τέλος του συνελικτικού δικτύου αποσκοπεί στην επιτυχή υλοποίηση της κύριας λειτουργίας του δικτύου, που τις περισσότερες φορές είναι η ταξινόμηση.

Προκειμένου να χρησιμοποιηθούν τα ΣΝΔ για το πρόβλημα της Ανάλυσης Συναισθήματος πρέπει να αναπαραστήσουμε το κείμενο ως ένα διδιάστατο μητρώο. Θα έχουμε $M$ γραμμές όπου $M$ το πλήθος των διαφορετικών λέξεων του κειμένου και $N$ στήλες όπου $N$ οι διαστάσεις των διανυσμάτων των λέξεων. Οι διαστάσεις του εκάστοτε φίλτρου ορίζονται ως $W \times N$ όπου $W$ ο αριθμός των λέξεων που θέτει ο χρήστης (κυμαίνεται μεταξύ 2 και 5). Ο αναγνώστης μπορεί να μελετήσει τεχνικές βαθειάς μάθησης που σχετίζονται με την πρόβλεψη της απόδοσης μαθητών στην ακόλουθη έρευνα [67].

## 3.6 Εφαρμογές και σύγχρονες προσεγγίσεις

Στην ενότητα αυτή θα παρουσιάσουμε σύγχρονες εφαρμογές που άπτονται των μεθοδολογιών που παρουσιάστηκαν παραπάνω. Ειδικότερα, ο αναγνώστης θα συναντήσει τεχνικές που άπτονται του Εξελικτικού Υπολογισμού και καθιερωμένων τεχνικών της Μηχανικής Μάθησης στο χώρο της εκπαίδευσης. Επιπλέον, οι συγκεκριμένες αναφορές σχετίζονται με τις διδακτικές επιδόσεις την αξιολόγηση και τη συμπεριφορά των μαθητών, μεταξύ άλλων.

Ο τρόπος με τον οποίο μαθαίνει ένα παιδί είναι σημαντικός στη διαδικασία μάθησης, καθώς ο τρόπος που συλλαμβάνει καλύτερα την πληροφορία ένας εκπαιδευόμενος αξίζει να μελετηθεί. Τα τελευταία χρόνια, με την ραγδαία αύξηση της τεχνολογίας και εξαιτίας συνθηκών που προκλήθηκαν λόγω της πανδημίας του Covid-19 οι ηλεκτρονικές πλατφόρμες μάθησης έγιναν μέρος της καθημερινότητας ενός εκπαιδευόμενου. Ακόμα, ηλεκτρονικές κονσόλες, έξυπνα κινητά και γενικά, έξυπνες συσκευές έχουν κατακλύσει την αγορά και την καθημερινότητά μας. Έτσι, Η μάθηση μέσω παιχνιδιού μέσω κινητού τηλεφώνου ή άλλων ηλεκτρονικών βιντεοπαιχνιδιών αποτελεί ένα σύχρονο στη σχετική επιστημονική βιβλιογραφία [68] ζήτημα, καθώς προάγει τη μάθηση με έναν διασκεδαστικό τρόπο και ενισχύει τα κίνητρα των μαθητών να αυξήσουν τη συμμετοχή τους στην εκπαιδευτική διαδικασία. Ως εκ τούτου, μπορεί να ενισχύσει τη μαθησιακή διαδικασία και να βελτιώσει τη συμμετοχή των μαθητών. Προς αυτή την κατεύθυνση, έχει διερευνηθεί [68] πώς μπορεί να χρησιμοποιηθεί η μάθηση με βάση τα παιχνίδια σε περιβάλλοντα τριτοβάθμιας εκπαίδευσης και μάλιστα, δόθηκε μια ανάλυση για την παιδαγωγική δυνατότητα της υιοθέτησής τους. Ως δοκιμαστικό πεδίο για την έρευνα αυτή, σχεδιάστηκε και εφαρμόστηκε ένα Quiz ερωτήσεων. Ειδικότερα, η συγκεκριμένη εφαρμογή στηρίχθηκε στη γλώσσα προγραμματισμού C# και είναι μια έξυπνη εφαρμογή εκμάθησης που βασίζεται σε παιχνίδια για κινητά για την αξιολόγηση και την προώθηση των γνώσεων των μαθητών. Ακόμα, χρησιμοποιήθηκε σε ίδρυμα τριτοβάθμιας εκπαίδευσης για ένα ακαδημαϊκό εξάμηνο και αξιολογήθηκε από φοιτητές και ειδικούς στην επιστήμη των υπολογιστών χρησιμοποιώντας ένα καθιερωμένο στατιστικό πλαίσιο ελέγχου. Όσον αφορά τα αποτελέσματα, οι ειδικοί «επικύρωσαν» την παιδαγωγική επάρκεια της εφαρμογής και οι μαθητές τόνισαν το θετικό αντίκτυπό της στη μάθηση και τη χρησιμότητά της.

Αντίστοιχες έρευνες [69] έχουν πραγματοποιηθεί, ώστε να εξαχθούν χρήσιμα συμπεράσματα για τον τρόπο που ο εγκέφαλος ενός εκπαιδευόμενου αντιδρά στην πληροφορία. Η μάθηση βασισμένη στον εγκέφαλο είναι η κατανόηση των λειτουργιών του ανθρώπινου εγκεφάλου και η εφαρμογή της σε εκπαιδευτικά περιβάλλοντα για ουσιαστική μάθηση. Η μάθηση βασισμένη στον εγκέφαλο προσαρμόζει τη διαδικασία μάθησης με βάση τη λειτουργία του ανθρώπινου εγκεφάλου, παρέχοντας ένα περιβάλλον διδασκαλίας με επίκεντρο τους μαθητές (ή τουλάχιστον εκεί στοχεύει). Προς αυτή την κατεύθυνση, έχει αναπτυχθεί [69] ένα εξατομικευμένο παιχνίδι ερωτήσεων με βάση τον εγκέφαλο που εφαρμόζει τις αρχές της εγκεφαλικής μάθησης και της ταξινόμησης Marzano για την προώθηση ουσιαστικής μάθησης και τη βελτίωση των γνωστικών λειτουργιών ανώτερης τάξης των μαθητών. Έτσι, το παραπάνω σύστημα προσαρμόζει το περιεχόμενο των ερωτήσεων με βάση το επίπεδο γνώσης των μαθητών, τη συναισθηματική κατάσταση και τον καθορισμένο μαθησιακό στόχο. Τα αποτελέσματα αποκαλύπτουν ότι αυτή η προσέγγιση έχει θετική επίδραση στις επιδόσεις των μαθητών, ξεπερνώντας τα παραδοσιακά συστήματα ηλεκτρονικής αξιολόγησης.

Η εξατομικευμένη μάθηση μέσω κινητού μπορεί να επιτευχθεί με την αναγνώριση των διαφορετικών μορφών μάθησης των μαθητών. Ωστόσο, αυτό συμβαίνει με τη συμπλήρωση μεγάλων ερωτηματολογίων. Αυτή η διαδικασία έχει θεωρηθεί ως κουραστική και χρονοβόρα, προκαλώντας τυχαία επιλογή των επιλογών των ερωτηματολογίων και, συνεπώς, εσφαλμένη προσαρμογή στις ανάγκες των μαθητών, θέτοντας σε κίνδυνο την απόκτηση γνώσης. Επιπλέον, τα κινητά περιβάλλοντα καθιστούν την επιλογή στα ερωτηματολόγια μη πρακτική λόγω περιορισμένων διεπαφών. Λαμβάνοντας υπόψη τα παραπάνω, η έρευνα [70] παρουσιάζει, ένα πλήρως ανεπτυγμένο σύστημα εκμάθησης γλωσσών για κινητά που ενσωματώνει αυτόματη αναγνώριση των μαθησιακών στυλ των μαθητών σύμφωνα με το μοντέλο Felder-Silverman (FSLSM) χρησιμοποιώντας ταξινόμηση που στηρίζεται σε ομάδες ταξινομητών (ensemble learners). Συγκεκριμένα, τρεις ταξινομητές, οι SVM, NB και KNN, συνδυάζονται με βάση τον κανόνα της πλειοψηφικής ψηφοφορίας. Επιπλέον, ενσωματώνει προσαρμοσμένες εκπαιδευτικές ρουτίνες για να δημιουργήσει ένα εξατομικευμένο περιβάλλον μάθησης με βάση τις μαθησιακές προτιμήσεις των μαθητών, όπως καθορίζονται από το στυλ τους.

Η μάθηση μέσω παιχνιδιών μέσω smartphones αξιοποιεί ένα διασκεδαστικό περιβάλλον για την παροχή ψηφιακής εκπαίδευσης. Μια τέτοια προσέγγιση περιλαμβάνει την ομαδοποίηση μαθητών για να παίξουν μαζί προς την κατεύθυνση της προώθησης των γνώσεών τους. Ωστόσο, η δημιουργία κατάλληλων ομάδων έχει σημαντικές παιδαγωγικές επιπτώσεις, καθώς η σύσταση των κατάλληλων συνεργατών θα μπορούσε να ενισχύσει περαιτέρω τις γνωστικές ικανότητες των μαθητών. Προς αυτή την κατεύθυνση, μια πολύ ενδιαφέρουσα έρευνα δημοσιεύθηκε πρόσφατα [71]. Στην εφαρμογή που αναπτύχθηκε, το σύστημα συνιστά σε κάθε μαθητή τέσσερις συνομηλίκους του προκειμένου να παίξουν ως ανταγωνιστές χρησιμοποιώντας έναν γενετικό αλγόριθμο. Ο γενετικός αλγόριθμος βρίσκει τους πιο κατάλληλους συνομηλίκους για κάθε μαθητή λαμβάνοντας υπόψη τον τρόπο εκμάθησης των εκπαιδευομένων, την προηγούμενη γνώση, την τρέχουσα γνώση και τις λανθασμένες αντιλήψεις. Ως εκ τούτου, ο μαθητής μπορεί να επιλέξει από τη λίστα ένα άτομο από τα προτεινόμενα, τα οποία έχουν κοινά χαρακτηριστικά. Οι δύο κύριοι λόγοι για τους οποίους επιλέγονται να δημιουργηθούν ομοιογενείς ομάδες είναι η προώθηση του θεμιτού

ανταγωνισμού και η παροχή προσαρμοστικού περιεχομένου παιχνιδιών με βάση τα χαρακτηριστικά των παικτών για τη βελτίωση των μαθησιακών τους αποτελεσμάτων.

Μπορεί οι συνθήκες που βιώσαμε κατά τα έτη 2019-2021 (και συνεχίζουμε να βιώνουμε παγκοσμίως) να ήταν πρωτόγνωρες λόγω της πανδημίας, όμως οι διεθνείς επιχειρήσεις, τα εκπαιδευτικά ιδρύματα, οι κρατικές υπηρεσία κ.ά κατάφεραν να αναπροσαρμόσουν την τεχνογνωσία ή τα μέσα που αξιοποιούσαν μέχρι πρώτινος. Αυτό επετεύχθη κυρίως χάριν των εξ αποστάσεων μέσων που αξιοποιήθηκαν. Στις μέρες μας, όσο ποτέ άλλοτε, τονίζεται η χρησιμότητα ή/και αναγκαιότητα των εξ αποστάσεως περιβαλλόντων. Ας δούμε και πάλι, στη συνέχεια, πώς όλα αυτά σκετίζονται με την εκπαιδευτική διαδικασία. Η κοινωνική δικτύωση έχει εκσυγχρονίσει την ψηφιακή εκπαίδευση μέσω της παροχής νέων λειτουργιών, όπως η αντίδραση, ο σχολιασμός, το κίνητρο ή η δημιουργία ομάδας. Υπό το φως των νέων εξελίξεων, ερευνητικές προσπάθειες όπως η [72] παρουσιάζει ένα λογισμικό ηλεκτρονικής μάθησης που ενσωματώνει κοινωνικά χαρακτηριστικά για τη διδασκαλία προγραμματισμού σε Η/Υ. Ωστόσο, η διερεύνηση του αντίκτυπου του λογισμικού ηλεκτρονικής μάθησης που κατέχει κοινωνικά χαρακτηριστικά είναι ακόμη υπό ζωηρή μελέτη. Για το σκοπό αυτό, πραγματοποήθηκε μια εκτεταμένη διερεύνηση [72], η οποία εξέτασε διαφορετικούς παράγοντες που επηρεάζουν τη μάθηση που βασίζεται στην κοινωνική δικτύωση. Ο πληθυσμός αυτής της μελέτης περιελάμβανε 200 προπτυχιακούς φοιτητές της επιστήμης των υπολογιστών. Για την ανάλυση των δεδομένων, χρησιμοποιήθηκε η λεγόμενη μοντελοποίηση δομικών εξισώσεων. Ολοκληρώνοντας, η μελέτη επιβεβαίωσε ότι το μοντέλο εξήγησε επαρκώς τις αιτιώδεις σχέσεις μεταξύ των μεταβλητών και παρουσίασε άμεσες και έμμεσες σημαντικές επιπτώσεις αυτών που μπορούν να προωθήσουν την καλύτερη ακαδημαϊκή απόδοση και την απόκτηση γνώσης των μαθητών.

Πέραν των έξυπνων τεχνολογιών και ιδιαίτερα των smartphones, tablets, PCs κλ.π. αξίζει να σταθούμε στο χώρο της Επαυξημένης Πραγματικότητας (ΕΠ), μια τεχνολογική καινοτομία που συνεισφέρει σε πλήθος διαφορετικών κλάδων, όπως η Ιατρική, η Μηχανική, η Αεροναυπηγική κ.ά. Η ενσωμάτωση της Επαυξημένης Πραγματικότητας στην εκπαίδευση των μηχανικών θεωρείται ότι αυξάνει την αποδοτικότητα, την ασφάλεια και το κέρδος χρόνου στις εργασίες, μειώνοντας τα αναλώσιμα και το κόστος υποδομών. Το κίνητρο συγχρονων ερευνών [73] είναι η συνεχής χρήση τεχνικών της ΕΠ στην κατάρτιση στη βιομηχανία και η καινοτομία τους είναι η ανάλυση των πιο σημαντικών παραγόντων που επηρεάζουν την πραγματική χρήση συστημάτων ΕΠ. Τα αποτελέσματα βοηθούν τους προγραμματιστές της ΕΠ να βελτιώσουν την ποιότητα των συστημάτων εκπαίδευσης προσομοίωσης ΕΠ για να βελτιώσουν την εμπειρία των χρηστών και να τις καταστήσουν ρεαλιστικά χρησιμοποιήσιμες. Περισσότερες πληροφορίες για λογισμικά και εφαρμογές που έχουν αναπτυχθεί για το χώρο της βιομηχανίας, ο ανγνώστης μπορεί να ανατρέξει στο βιβλίο [74]. Ακόμα, περισσότερες πληροφορίες για τις έρευνες που σχετίζονται με τη διδασκαλία και την ΕΠ μπορεί να μελετήσει στις ακόλουθες αναφορές [75, 76, 77].

Η διάδοση του Διαδικτύου εισήγαγε νέες τεχνολογίες στην ψηφιακή εκπαίδευση. Μια από αυτές είναι τα λεγόμενα Μαζικά Ανοικτά Εξ αποστάσεως Μαθήματα, γνωστά και ως MOOCs (Massive Open Online Courses). Τα MOOCs είναι διαδικτυακά περιβάλλοντα μάθησης που προσφέρουν δωρεάν εκπαιδευτικά προγράμματα σε μεγάλο αριθμό γεωγραφικά διασκορπισμένων μαθητών. Η ταχεία ανάπτυξη των MOOCs οδηγεί στη διερεύνηση της παρεχόμενης ποιότητας μάθησής τους, που αποτελείται από ένα συνδυασμό διαφορετικών και ποικίλων παραγόντων, τα οποία δε θα παρουσιάσουμε στα πλαίσια αυτής της εργασίας. Λαμβάνοντας υπόψη τα παραπάνω, έχουν αναπτυχθεί μοντέλα, όπως το [78], όπου γίνεται αξιολόγηση της ποιότητας στην ηλεκτρονική μάθηση γενικότερα και ειδικότερα στα MOOCs. Τα αποτελέσματα που έχουν λάβει οι ερευνητές είναι άκρως υποσχόμενα και υποδεικνύουν ένα πρόσφορο έδαφος για την προώθηση της ποιότητας στην ηλεκτρονική μάθηση μέσω των MOOCs.

Οι προσεγγίσεις που υποστηρίζονται από Η/Υ έχουν χρησιμοποιηθεί ευρέως για τον εμπλουτισμό της μαθησιακής διαδικασίας, όπως φαίνεται και από το πλήθος των παραπάνω αναφορών που παραθέσαμε και παρουσιάσαμε εν συντομία. Η τεχνολογική πρόοδος οδήγησε τα συστήματα διδασκαλίας να ενσωματώσουν τη νοημοσύνη στις λειτουργίες τους. Ωστόσο, μέχρι στιγμής, δεν καταφέρνουν να ενσωματώσουν επαρκώς τη νοημοσύνη και την προσαρμοστικότητα στους διαγνωστικούς και συλλογιστικούς μηχανισμούς τους. Για το σκοπό αυτό έχουν αναπτυχθεί εφαρμογές, όπως η [79], όπου παρουσιάζει ένα νέο σύστημα εμπειρογνωμόνων για τη διδασκαλία της γλώσσας προγραμματισμού Java. Μια πολυστρωματική μηχανή συμπερασμάτων αναπτύχθηκε και χρησιμοποιήθηκε σε αυτό το σύστημα για να παρέχει εξατομικευμένες οδηγίες στους μαθητές ανάλογα με τις ανάγκες και τις προτιμήσεις τους. Η πολυστρωματική μηχανή συμπερασμάτων ενσωματώνει ένα σύνολο αλγοριθμικών μεθόδων σε διαφορετικά επίπεδα προωθώντας την εξατομίκευση στις στρατηγικές διδασκαλίας. Συγκεκριμένα, ένα τεχνητό νευρωνικό δίκτυο και η ανάλυση πολλαπλών αποφάσεων χρησιμοποιούνται σε ένα στρώμα για την προσαρμογή των μαθησιακών μονάδων σύμφωνα με το στυλ εκμάθησης των μαθητών. Από την άλλη, ένα μοντέλο ασαφούς λογικής εφαρμόζεται για τον καθορισμό της ευκρίνειας των μαθησιακών μονάδων σύμφωνα με χαρακτηριστικά του προφίλ των μαθητών, όπως το στυλ μάθησης, το επίπεδο γνώσης και οι λανθασμένες αντιλήψεις.

Παρά την πρόοδο που έχει σημειωθεί, δεν θα πρέπει να παραβλέπουμε ότι τα λογισμικά ηλεκτρονικής μάθησης προσανατολίζονται σε μια ετερογενή ομάδα μαθητών. Έτσι, τέτοια συστήματα πρέπει να παρέχουν εξατομίκευση στις ανάγκες και τις προτιμήσεις των μαθητών, έτσι ώστε η απόκτηση γνώσεων να γίνει πιο αποτελεσματική. Ένας μηχανισμός εξατομίκευσης είναι η προσαρμογή στις μαθησιακές μορφές των μαθητών. Ωστόσο, αυτή η διαδικασία

απαιτεί πολύ χρόνο όταν γίνεται χειροκίνητα και είναι επιρρεπής σε σφάλματα. Συνεπώς, η ερευνητική κοινότητα οδηγήθηκε σε μια νέα τεχνική για την ανίχνευση τρόπων εκμάθησης. Συγκεκριμένα, χρησιμοποιήθηκε [80] το γνωστό μοντέλο Honey-Mumford, το οποίο κατατάσσει τους μαθητές σε τέσσερις κατηγορίες. Επιπλέον, η αυτόματη ανίχνευση χρησιμοποιεί την τεχνική της ασαφούς λογικής [81] λαμβάνοντας ως είσοδο την αλληλεπίδραση των μαθητών με το μαθησιακό περιβάλλον, όπως τα σχόλια των μαθητών σχετικά με τις μαθησιακές διαδικασίες και τη συμμετοχή τους σε συζητήσεις. Καταλήγοντας, πέραν των ηλεκτρονικών μέσων ή της προόδου που έχει σημειωθεί μέσω των τεχνικών και αλγορίθμων Μηχανικής Μάθησης θα πρέπει να δοθεί ιδιαίτερη σημασία σε ορισμένα ακόμα χαρακτηριστικά, όπως η συλλογιστική των γνωστικών καταστάσεων των μαθητών, ο επανασχεδιασμός στρατηγικών διδασκαλίας και η εκπαιδευτική προσαρμοστικότητα των διαθέσιμων εργαλείων κ.ά. [82, 83, 84, 85, 86]. Κλείνοντας αυτή την ενότητα και πριν περάσουμε στην πειραματική μελέτη της εργασίας, αξίζει να αναφέρουμε πως το Twitter είναι ένα μόνο από τα μέσα κοινωνικής δικτύωσης που έχουν προσεγγίσει το ενδιαφέρον της επιστημονικής κοινότητας. Από τα πρώτα και ευρέως χρησιμοποιούμενα μέσα, αυτό του Facebook, έχει επίσης συγκερτρώσει το ενδιαφέρον για ανάλογες με τις προαναφερθείσες μελέτες. Ενδεικτικά, ο αναγνώστης μπορεί να ανατρέξει στις εργασίες [87, 88].

# 4. Σύγκριση Μοντέλων Ανάλυσης Συναισθήματος

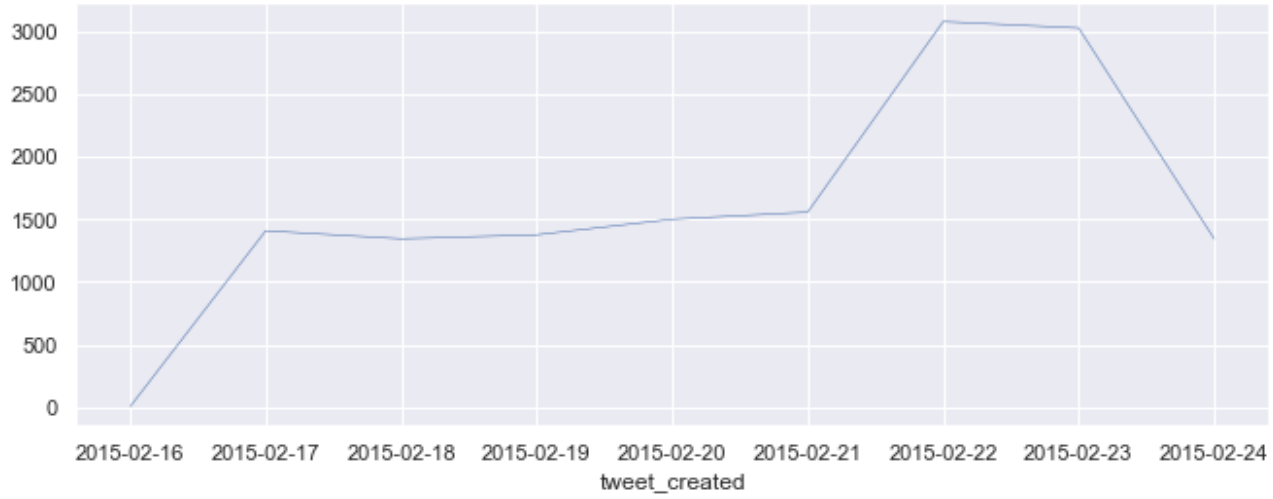
Στο κεφάλαιο αυτό θα συγκρίνουμε την απόδοση τριών αλγορίθμων μηχανικής μάθησης για το πρόβλημα πρόβλεψης συναισθήματος από δεδομένα κειμένου.

## 4.1 Παρουσίαση Συνόλου Δεδομένων

Για να συγκρίνουμε την απόδοση τριών μοντέλων ανάλυσης συναισθήματος με επιβλεπόμενη μάθηση, χρησιμοποιήσαμε το σύνολο δεδομένων Twitter US Airline Sentiment που είναι διαθέσιμο μέσω του υπερσυνδέσμου *https : //www.kaggle.com/crowdflower/twitter − airline − sentiment*. Το εν λόγω σύνολο αποτελείται από 14640 εγγραφές με 15 χαρακτηριστικά η κάθε μία (Πίνακας 4.1). Στα πλαίσια της ανάλυσής μας λήφθηκαν υπόψιν μόνο τα χαρακτηριστικά του Πίνακα 4.1 με έντονη γραφή, όπου το χαρακτηριστικό text αποτελεί το κείμενο που τροφοδότησε την είσοδο των μοντέλων και το χαρακτηριστικό airline_ sentiment αντιπροσωπεύει την επιθυμητή έξοδο των μοντέλων ως προς την πρόβλεψη συναισθήματος, για κάθε κείμενο.

| | | |
|---|---|---|
| tweet_id | **airline_sentiment** | airline_sentiment_confidence |
| negativereason | negativereason_confidence | airline |
| airline_sentiment_gold | name | negativereason_gold |
| retweet_count | **text** | tweet_coord |
| tweet_created | tweet_location | user_timezone |

Πίνακας 4.1: Τα χαρακτηριστικά του συνόλου δεδομένων.

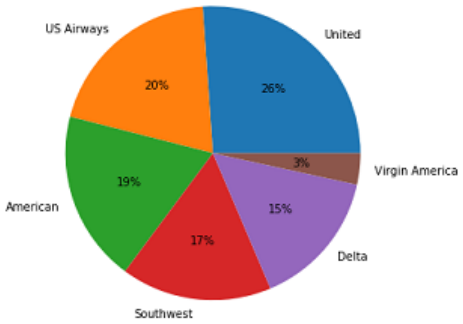



Σχήμα 4.1: Σχόλια ανά ημέρα.

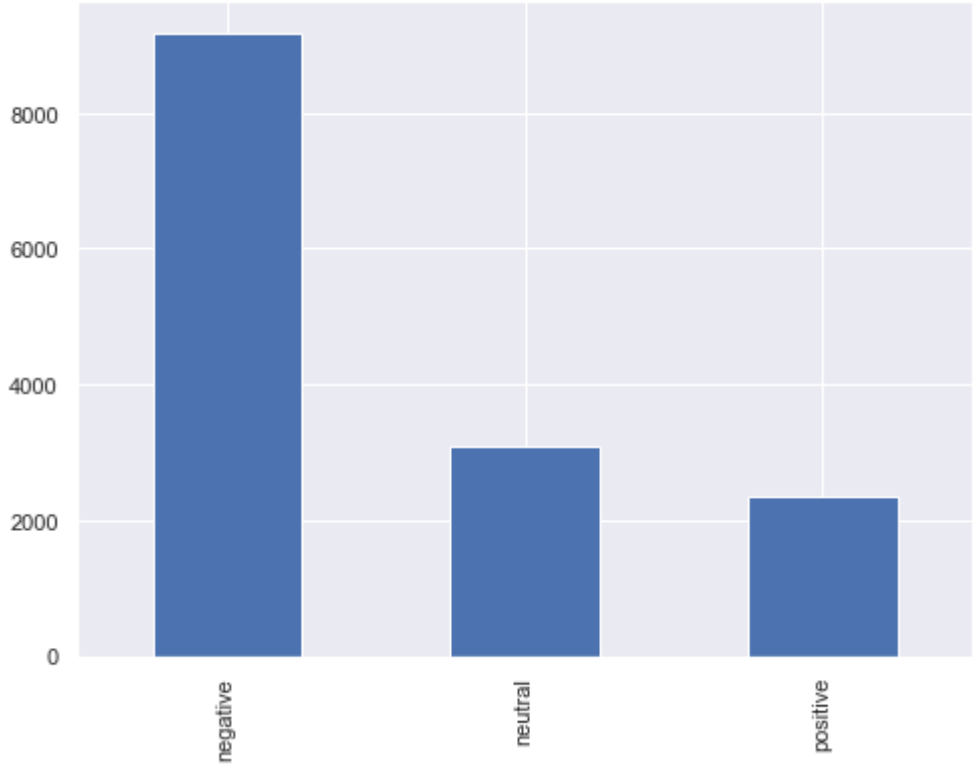
Τα tweets που απαρτίζουν το σύνολο δεδομένων αφορούν τα σχόλια των επιβατών για 6 αμερικάνικες εταιρίες αερομεταφορών (Σχήμα 4.2) για το μήνα Φεβρουάριο του 2015 (Σχήμα 4.1). Το κείμενο κάθε tweet έχει ταξηνομηθεί, ως προς το συναίσθημα, στις ακόλουθες τρεις κλάσεις negative, neutral και positive (Σχήμα 4.3). Παρατηρούμε ότι τα αρνητικά σχόλια υπερτερούν κατά πολύ τόσο των θετικών όσο και των ουδέτερων.

## 4.2 Περιγραφή Πειραμάτων Και Αποτελεσμάτων

Στο προγραμματιστικό περιβάλλον της έκδοσης 3.8 της Python, χρησιμοποιώντας κατά κύριο λόγο τα modules nltk και scikit-learn αναπτύχθηκαν τα μοντέλα ταξινομητών Multinomial Bayes, Lineal Svm και Knn σε υπολογιστικό σύστημα του οποίου τα χαρακτηριστικά αναφέρονται στον Πίνακα 4.2.




Σχήμα 4.2: Συχνότητα σχόλιων ανά εταιρία.




Σχήμα 4.3: Αριθμός σχόλιων ανά κλάση.

| | |
|---|---|
| Λογισμικό | macOS Big Sur |
| Επεξεργαστής | 1.4 GHz Τετραπύρηνος Intel Core i5 |
| Μνήμη RAM | 8 GB 2133 MHz LPDDR3 |
| Γραφικά | Intel Iris Plus Graphics 645 1536 MB |

Πίνακας 4.2: Χαρακτηριστικά υπολογιστικού συστήματος

Τα δεδομένα χωρίστηκαν σε δύο μέρη, το σύνολο εκπαίδευσης και το σύνολο ελέγχου με αναλογίες επί του αρχικού 80% και 20% αντίστοιχα. Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε 100 φορές με τη μέση απόδοση για το κάθε ένα να παρουσιάζεται στον Πίνακα 4.3.

Προκειμένου να ελέγξουμε σε επίπεδο σημαντικότητας 95%, αν οι διαφορές στο μέσο ποσοστό επιτυχίας είναι στατιστικά σημαντικές, κάναμε τον παρακάτω έλεγχο υποθέσεων:

- $H_0 : \mu_1 = \mu_2 = \mu_3$
  Όλα τα μοντέλα έχουν περίπου ίδια μέση τιμή.

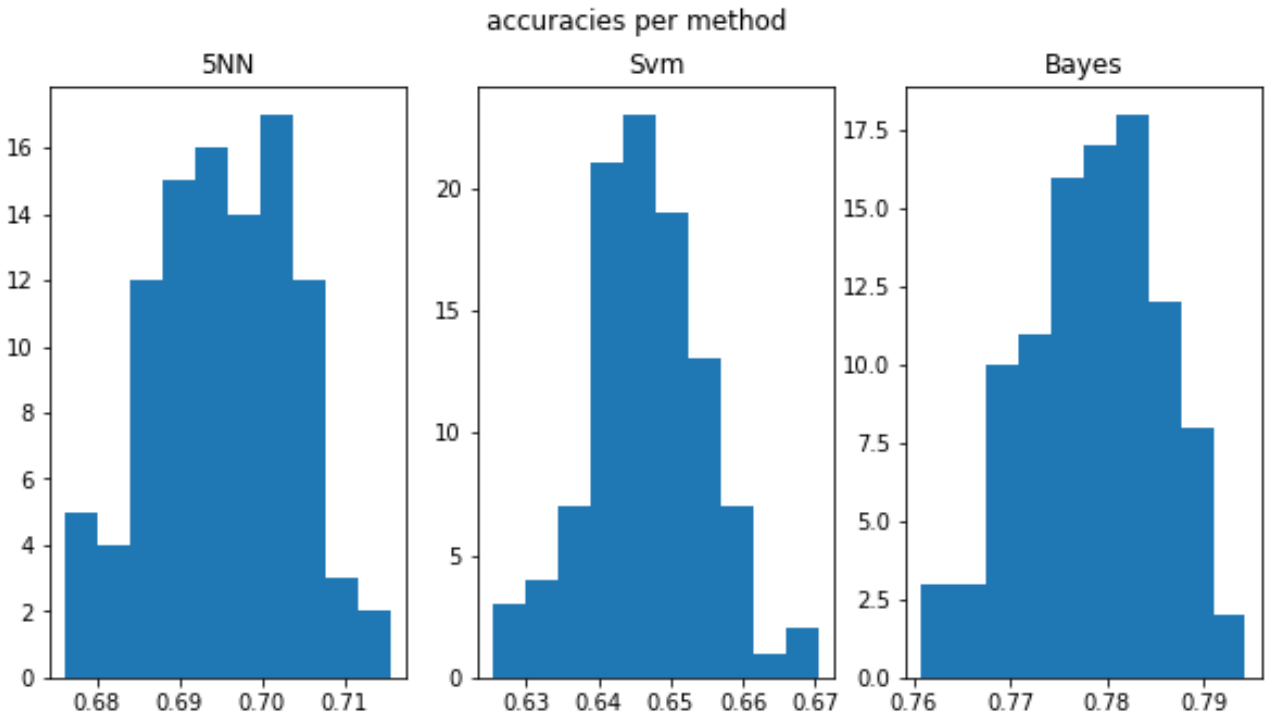
| Μέθοδος | **5-nn** | **Svm** | **Bayes** |
|---|---|---|---|
| Μέσο ποσοστό επιτυχίας (%) | 69.52390710382512 | 64.63893683120469 | 77.87193648602003 |
| Μικρότερο ποσοστό επιτυχίας (%) | 67.62295081967213 | 62.54746289264757 | 76. 07870210562651 |
| Μεγαλύτερο ποσοστό επιτυχίας (%) | 71.55054644808743 | 67.03486365205384 | 79.4269934414912 |

Πίνακας 4.3: Ποσοστά επιτυχίας συγκρινόμενων μοντέλων

- $H_1 : \exists\, \mu_i \neq \mu_j$
  Υπάρχει κάποιο μοντέλο με αρκετά διαφορετική μέση τιμή από τα άλλα.

όπου $\mu_i$ η μέση τιμή για κάθε μοντέλο.

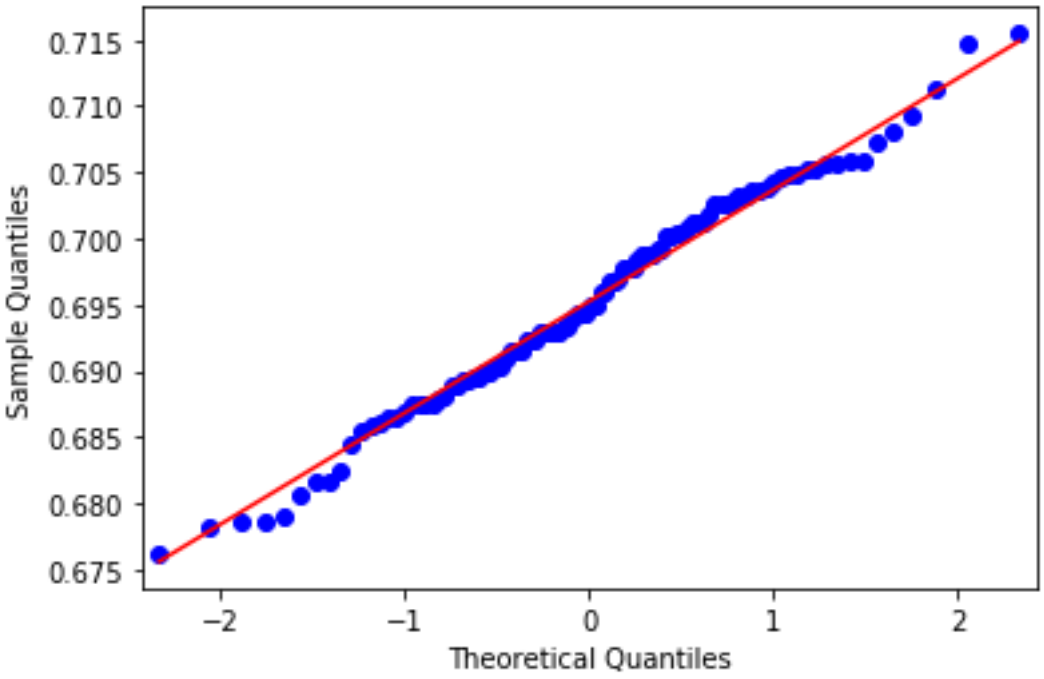
Χρησιμοποιώντας το module scipy κάναμε τους ελέγχους Shapiro Wilk και D' Agostino για να ελέγξουμε αν τα σύνολα δεδομένων από τα ποσοστά επιτυχίας ακολουθούν την κανονική κατανομή ώστε να μπορέσουμε να εφαρμόσουμε One Way Anova. Η κανονικότητα ως προς την κατανομή μπορεί να εκτιμηθεί μέσω των Σχημάτων 4.4 έως 4.7.



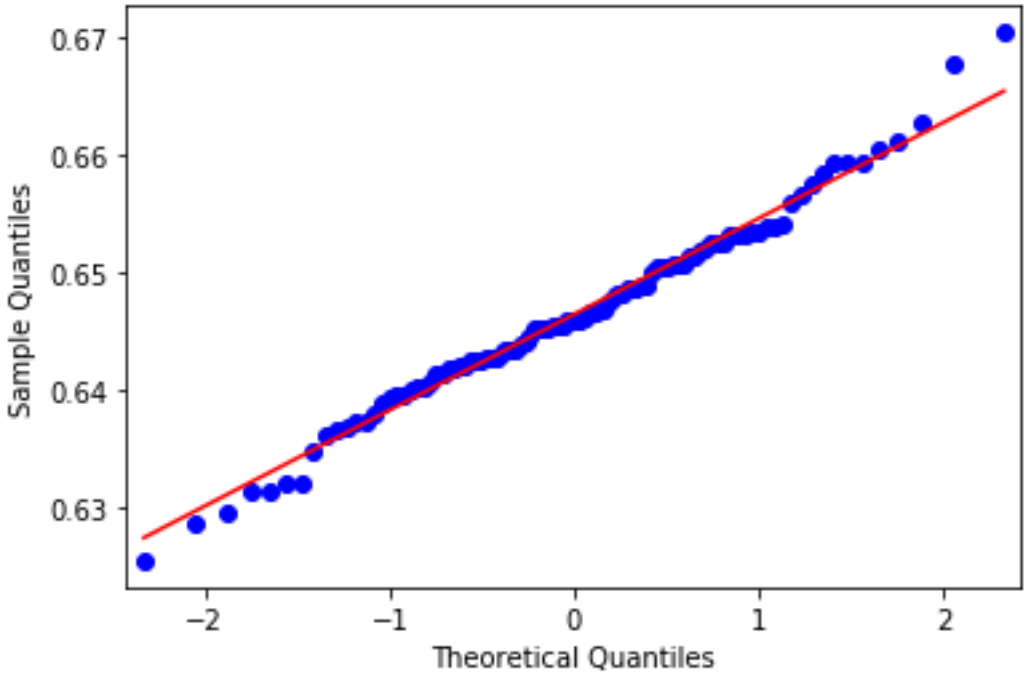

Σχήμα 4.4: Ιστόγραμμα συχνοτήτων

Η άλλη προϋπόθεση για το One Way Anova είναι τα σύνολα, να είναι ξένα μεταξύ τους, γεγονός που ισχύει αφού αποτελούν τα ποσοστά επιτυχίας διαφορετικών μοντέλων. Με την εφαρμογή του προαναφερθέντος ελέγχου (Πίνακες 4.4 και 4.5) προκύπτει ότι το p-value είναι $1.3120771932613894e^{-251} < 0.05$ άρα απορρίπτεται η μηδενική υπόθεση και συμπεραίνουμε ότι τα τρία μοντέλα δεν είναι ισοδύναμα.

Προκειμένου να αξιολογηθεί η διαφορά στις επιδόσεις μεταξύ των μοντέλων, εφαρμόστηκε ο post hoc έλεγχος Tukey HSD που κατέδειξε τη μεγαλύτερη διαφορά μεταξύ των μοντέλων Bayes και 5nn. Επιπλέον όλες οι συγκρίσεις ανά ζεύγος έδειξαν ότι δεν υπάρχει κάποιο ζευγάρι μοντέλων που να χει παρόμοια απόδοση και άρα συμπεραίνουμε πως το μοντέλο Bayes είναι καλύτερο των άλλων δύο για το συγκεκριμένο σύνολο δεδομένων.



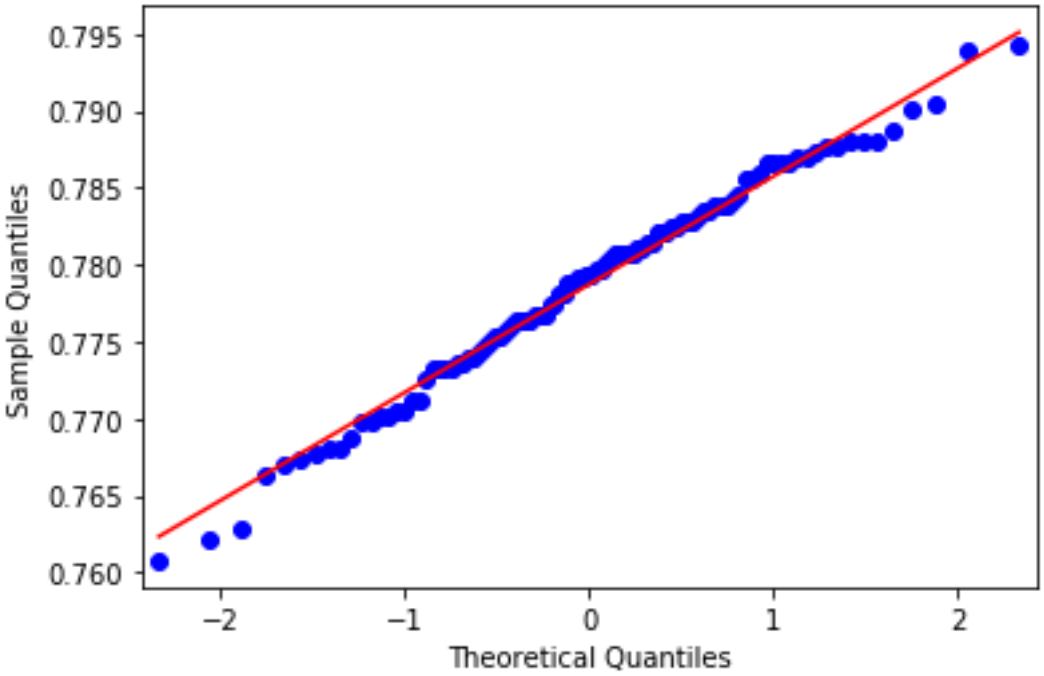

Σχήμα 4.5: QQ-plot 5nn




Σχήμα 4.6: QQ-plot Svm

| | **5-nn** | **Svm** | **Bayes** | Total |
|---|---|---|---|---|
| N | 100 | 100 | 100 | 300 |
| $\sum X$ | 69.5239 | 64.6389 | 77.8719 | 212.0348 |
| Mean | 0.6952 | 0.6464 | 0.7787 | 0.707 |
| $\sum X^2$ | 48.3429 | 41.7886 | 60.6453 | 150.7767 |
| Std.Dev. | 0.0085 | 0.0082 | 0.0071 | 0.0553 |

Πίνακας 4.4: Περίληψη Δεδομένων




Σχήμα 4.7: QQ-plot Bayes

| Source | **SS** | **df** | **MS** | |
|---|---|---|---|---|
| Between-treatments | 0.8955 | 2 | 0.4478 | F = 7115.36979 |
| Within-treatments | 0.0187 | 297 | 0.0001 | |
| Total | 0.9142 | 299 | | |

Πίνακας 4.5: One Way Anova

# Bibliography


[1] Umman Tugba Gursoy, Diren Bulut, and Cemil Yigit. Social media mining and sentiment analysis for brand management. *Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology*, 3(1):497–551, 2017.

[2] B Umadevi and P Surya. A review on various data mining techniques in social media. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(4):8082–8086, 2017.

[3] Jiawe Han and Micheline Kamber. Data mining concepts and techniques san francisco moraga kaufman. 2001.

[4] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[5] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[6] Antony Mayfield. What is social media. 2008.

[7] Hady Lauw, John C Shafer, Rakesh Agrawal, and Alexandros Ntoulas. Homophily in the digital world: A livejournal case study. *IEEE Internet Computing*, 14(2):15–23, 2010.

[8] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66, 2001.

[9] Irwin King, Jiexing Li, and Kam Tong Chan. A brief survey of computational approaches in social computing. In *2009 International Joint Conference on Neural Networks*, pages 1625–1632. IEEE, 2009.

[10] David Burth Kurka, Alan Godoy, and Fernando J Von Zuben. Online social network analysis: A survey of research applications in computer science. *arXiv preprint arXiv:1504.05655*, 2015.

[11] Marc Cheong and Sid Ray. A literature review of recent microblogging developments. *Victoria, Australia: Clayton School of Information Technology, Monash University*, page 7, 2011.

[12] Kalina Bontcheva and Dominic Rout. Making sense of social media streams through semantics: a survey. *Semantic Web*, 5(5):373–403, 2014.

[13] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[14] Κωνσταντίνος Καρποδίνης. *Ανάλυση συναισθημάτων σε δεδομένα από το twitter χρησιμοποιώντας εργαλεία της R και μοντέλα μηχανικής μάθησης*. PhD thesis, 2016.

[15] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[16] Ελευθέριος Ιωσήφ Παρασκευάς. *Μια προσέγγιση για την ανάλυση συναισθήματος στον εντοπισμό γεγονότων*. Technical report, Aristotle University of Thessaloniki, 2016.

[17] Despoina Chatzakou and Athena Vakali. Harvesting opinions and emotions from social media textual resources. *IEEE Internet Computing*, 19(4):46–50, 2015.

[18] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[19] Andreas C Müller, Sarah Guido, et al. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[20] Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[21] Αικατερίνη Γεωργούλη. *Τεχνητή νοημοσύνη*. 2015.

[22] Stavros P Adam, Stamatios-Aggelos N Alexandropoulos, Panos M Pardalos, and Michael N Vrahatis. No free lunch theorem: a review. In *Approximation and Optimization*, pages 57–82. Springer, 2019.

[23] Χρήστος Μπαζιώτης. Ανάλυση συναισθήματος στο twitter με βαθιά νευρωνικά δίκτυα. Master's thesis, Πανεπιστήμιο Πειραιώς, 2017.

[24] Γεωργία Χαρίση. Ανίχνευση συναισθήματος σε δεδομένα κοινωνικών δικτύων μέσω εξόρυξης και ανάλυσης. 2018.

[25] Manoochehr Ghiassi, James Skinner, and David Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[26] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[27] Manoochehr Ghiassi and H Saidane. A dynamic architecture for artificial neural networks. *Neurocomputing*, 63:397–413, 2005.

[28] Siddharth Aravindan and Asif Ekbal. Feature extraction and opinion mining in online product reviews. In *2014 International Conference on Information Technology*, pages 94–99. IEEE, 2014.

[29] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[30] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[31] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161, 2011.

[32] Alastair J Gill, Robert M French, Darren Gergle, and Jon Oberlander. Identifying emotional characteristics from short blog texts. In *30th Annual Conference of the Cognitive Science Society*, pages 2237–2242. Cognitive Science Society Washington, DC, 2008.

[33] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[34] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[35] Kevin Lund. Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995*, 1995.

[36] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[37] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[38] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[39] SM Vohra and JB Teraiya. A comparative study of sentiment analysis techniques.

[40] Panagis Magdalinos, Christos Doulkeridis, and Michalis Vazirgiannis. Fedra: A fast and efficient dimensionality reduction algorithm. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 509–520. SIAM, 2009.

[41] James E Gentle. Matrix algebra. *Springer texts in statistics, Springer, New York, NY, doi*, 10:978–0, 2007.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[43] Maria Ogneva. How companies can use sentiment analysis to improve their business. *Retrieved August*, 30, 2010.

[44] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[45] Tristan Fletcher. Support vector machines explained. *Tutorial paper., Mar*, page 28, 2009.

[46] Alex M Andrew. An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor, cambridge university press, cambridge, 2000, xiii+ 189 pp., isbn 0-521-78019-5 (hbk,£ 27.50)., 2000.

[47] Yu Hen Hu and Jenq-Neng Hwang. *Handbook of neural network signal processing*. CRC press, 2001.

[48] Simon Haykin and Neural Network. A comprehensive foundation. *Neural Networks*, 2(2004):41, 2004.

[49] Δανάη Π Γιαννούλη and Danai P Giannouli. Εφαρμογές των μηχανών διανυσματικής υποστήριξης σε προβλήματα ταξινόμησης και παλινδρόμησης. Master's thesis, 2014.

[50] Vojislav Kecman. Learning and soft computing: Support vector machines, neural networks, and fuzzy logic models (complex adaptive systems), 2001.

[51] Laszlo Kozma. k nearest neighbors algorithm (knn). *Helsinki University of Technology*, 2008.

[52] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5133–5142, 2018.

[53] Γεώργιος Μαστραπάς. Ανάλυση συναισθήματος σε δεδομένα του κοινωνικού δικτύου twitter με μεθόδους μηχανικής μάθησης. 2016.

[54] Igor Aleksander and Helen Morton. *An introduction to neural computing*, volume 240. Chapman and Hall London, 1990.

[55] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[56] JM Mendel and RW McLaren. 8 reinforcement-learning control and pattern recognition systems. *Mathematics in Science and Engineering*, 66:287–318, 1970.

[57] Richard Lippmann. An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2):4–22, 1987.

[58] Paul D McNelis. *Neural networks in finance: gaining predictive edge in the market*. Academic Press, 2005.

[59] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[60] Kevin Gurney. *An introduction to neural networks*. CRC press, 1997.

[61] Nadeem Ahmed Syed, Syed Huan, Liu Kah, and Kay Sung. Incremental learning with support vector machines. 1999.

[62] Tom Heskes and Wim Wiegerinck. A theoretical comparison of batch-mode, on-line, cyclic, and almost-cyclic learning. *IEEE transactions on neural networks*, 7(4):919–925, 1996.

[63] Μιχαήλ Επιτροπάκης. *Εκπαίδευση τεχνητών νευρωνικών δικτύων με την χρήση εξελικτικών αλγορίθμων, σε σειριακά και κατανεμημένα συστήματα*. PhD thesis, 2008.

[64] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[65] Νικόλαος Μπούας. Βαθιά νευρωνικά δίκτυα για την ανάλυση ακτινογραφιών και αξονικών τομογραφιών ασθενών με covid-19. 2020.

[66] Εμμανουήλ Παπαδάκης. Ανάλυση συναισθήματος από κείμενο με τεχνικές μηχανικής μάθησης και χρήση λεξικού. 2016.

[67] F Giannakas, C Troussas, I Voyiatzis, and C Sgouropoulou. A deep learning classification framework for early prediction of team-based academic performance. *Applied Soft Computing*, 106:107355, 2021.

[68] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education. *Computers & Education*, 144:103698, 2020.

[69] Akrivi Krouska, Christos Troussas, and Cleo Sgouropoulou. A personalized brain-based quiz game for improving students' cognitive functions. In *International Conference on Brain Function Assessment in Learning*, pages 102–106. Springer, 2020.

[70] Christos Troussas, Akrivi Krouska, Cleo Sgouropoulou, and Ioannis Voyiatzis. Ensemble learning using fuzzy weights to improve learning style identification for adapted instructional routines. *Entropy*, 22(7):735, 2020.

[71] Akrivi Krouska, Christos Troussas, and Cleo Sgouropoulou. Applying genetic algorithms for recommending adequate competitors in mobile game-based learning environments. In *International Conference on Intelligent Tutoring Systems*, pages 196–204. Springer, 2020.

[72] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Impact of social networking for advancing learners' knowledge in e-learning environments. *Education and Information Technologies*, pages 1–21, 2021.

[73] Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. User acceptance of augmented reality welding simulator in engineering training. *Education and Information Technologies*, pages 1–27, 2021.

[74] Christos Troussas and Cleo Sgouropoulou. *Innovative trends in personalized software engineering and information systems: the case of intelligent and adaptive e-learning systems*, volume 324. IOS Press, 2020.

[75] Hui-Chin Yeh, Sheng-Shiang Tseng, and Leechin Heng. Enhancing efl students' intracultural learning through virtual reality. *Interactive Learning Environments*, pages 1–10, 2020.

[76] Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Exploration of augmented reality in spatial abilities training: A systematic literature review for the last decade. *Informatics in Education*, 20(1):107–130, 2021.

[77] Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Measuring user experience, usability and interactivity of a personalized mobile augmented reality training system. *Sensors*, 21(11):3888, 2021.

[78] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Towards a reference model to ensure the quality of massive open online courses and e-learning. In *International Conference on Brain Function Assessment in Learning*, pages 169–175. Springer, 2020.

[79] Christos Troussas, Akrivi Krouska, and Maria Virvou. A multilayer inference engine for individualized tutoring model: adapting learning material and its granularity. *Neural Computing and Applications*, pages 1–15, 2021.

[80] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. Dynamic detection of learning modalities using fuzzy logic in students' interaction activities. In *International conference on intelligent tutoring systems*, pages 205–213. Springer, 2020.

[81] Akrivi Krouska, Christos Troussas, and Cleo Sgouropoulou. Fuzzy logic for refining the evaluation of learners' performance in online engineering education. *European Journal of Engineering and Technology Research*, 4(6):50–56, 2019.

[82] Akrivi Krouska, Christos Troussas, and Cleo Sgouropoulou. Usability and educational affordance of web 2.0 tools from teachers' perspectives. In *24th Pan-Hellenic Conference on Informatics*, pages 107–110, 2020.

[83] Christos Troussas, Akrivi Krouska, Filippos Giannakas, Cleo Sgouropoulou, and Ioannis Voyiatzis. Automated reasoning of learners' cognitive states using classification analysis. In *24th Pan-Hellenic Conference on Informatics*, pages 103–106, 2020.

[84] Christos Troussas, Akrivi Krouska, Filippos Giannakas, Cleo Sgouropoulou, and Ioannis Voyiatzis. Redesigning teaching strategies through an information filtering system. In *24th Pan-Hellenic Conference on Informatics*, pages 111–114, 2020.

[85] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. A novel teaching strategy through adaptive learning activities for computer programming. *IEEE Transactions on Education*, 64(2):103–109, 2020.

[86] Christos Troussas, Akrivi Krouska, Efthimios Alepis, and Maria Virvou. Intelligent and adaptive tutoring through a social network for higher education. *New Review of Hypermedia and Multimedia*, pages 1–30, 2021.

[87] Christos Troussas, Maria Virvou, Jaime Caro, and Kurt Junshean Espinosa. Language learning assisted by group profiling in social networks. *International Journal of Emerging Technologies in Learning (IJET)*, 8(3):35–38, 2013.

[88] Christos Troussas, Maria Virvou, and Kurt Junshean Espinosa. Using visualization algorithms for discovering patterns in groups of users for tutoring multiple languages through social networking. *J. Networks*, 10(12):668–674, 2015.