Πανεπιστήμιο Πειραιώς,
Σχολή Τεχνολογιών Πληροφορικής και Επικοινωνιών,
Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών,
"Προηγμένα Συστήματα Πληροφορικής"

Ανάλυση Συναισθήματος στο Twitter με Βαθιά Νευρωνικά Δίκτυα

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Χρήστος Μπαζιώτης

Επιβλέποντες: Γιάννης Θεοδωρίδης
Νίκος Πελέκης
Χρήστος Δουλκερίδης

Πειραιάς,
Σεπτέμβριος 2017

Στους γονείς μου

Περίληψη

Η εργασία ασχολείται με το πρόβλημα της πρόβλεψης του συναισθηματικού προσανατολισμού, σε μηνύματα του κοινωνικού δικτύου Twitter. Είναι ένα πρόβλημα Επεξεργασίας Φυσικής Γλώσσας (ΕΦΓ), το οποίο στα πλαίσια της εργασίας, προσεγγίζεται με την χρήση Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) αξιοποιώντας κατανεμημένες αναπαραστάσεις λέξεων (word embeddings).

Αρχικά, γίνεται αναδρομή της προόδου του επιστημονικού πεδίου της Ανάλυσης Συναισθήματος (Sentiment Analysis). Στη συνέχεια καταγράφονται και συγκρίνονται οι σημαντικότερες προσεγγίσεις, οι οποίες έχουν προταθεί για την επίλυση του προβλήματος. Ιδιαίτερη σημασία δίνεται στην αναζωπύρωση της έρευνας στα ΤΝΔ και παρουσιάζονται οι σημαντικότερες αρχιτεκτονικές ΤΝΔ, οι οποίες έχουν εφαρμοστεί στην Ανάλυσης Συναισθήματος. Επιπλέον, γίνεται σύγκριση των ΤΝΔ με τις παραδοσιακές τεχνικές μηχανικής μάθησης και παρουσιάζονται επιχειρήματα υπέρ της καταλληλότητάς τους σε προβλήματα ΕΦΓ.

Ακόμη, για την καλύτερη προετοιμασία των μηνυμάτων του Twitter ως είσοδο στα μοντέλα μηχανικής μάθησης, αναπτύχθηκε ένα εργαλείο προ-επεξεργασίας κειμένων. Το εργαλείο αυτό είναι ικανό να αναγνωρίσει και να επεξεργαστεί κείμενα από κοινωνικά δίκτυα, στα οποία υπάρχουν αρκετά ορθογραφικά, συντακτικά και γραμματικά λάθη, καθώς και γενικότερα "δημιουργική" γραφή. Μερικές από τις δυνατότητες του εργαλείου είναι, λεκτική ανάλυση, ορθογραφική διόρθωση και κανονικοποίηση λέξεων και φράσεων.

Τέλος, στα πλαίσια της έρευνας για την εργασία, συμμετείχαμε στον διεθνή διαγωνισμό σημασιολογικής αξιολόγησης Semeval-2017. Τα μοντέλα του διαγωνισμού είναι το ουσιαστικό αποτέλεσμα της έρευνάς μου. Γίνεται αναλυτική παρουσίαση των σχετικών μοντέλων και δίνονται θεωρητικά επιχειρήματα για την καταλληλότητα της τελικής προσέγγισης. Τα μοντέλα που αναπτύχθηκαν ήταν ιδιαίτερα ανταγωνιστικά, πετυχαίνοντας την πρώτη θέση στο Task 4:"Sentiment Analysis in Twitter" και τη δεύτερη θέση στο Task 6: "#HashtagWars: Learning a Sense of Humor" του Semeval-2017.

Λέξεις-κλειδιά: ανάλυση συναισθήματος, εξόρυξη γνώμης, επεξεργασία φυσικής γλώσσας, κατηγοριοποίηση κειμένων, μηχανική μάθηση, τεχνητά νευρωνικά δίκτυα, βαθιά νευρωνικά δίκτυα

Abstract

Sentiment analysis is an area in Natural Language Processing (NLP), studying the identification and quantification of the sentiment expressed in text. The thesis addresses the problem of predicting the sentiment of messages from Twitter micro-blogging service. The task is approached using Artificial Neural Networks (ANN), utilizing distributed text representations (word embeddings).

Firstly, a survey of the field of sentiment analysis is performed. Next, the most important approaches for addressing the problem are reviewed. Special focus is given to the resurgence of research in ANNs. The most common ANN architectures, which have been applied to sentiment analysis, are presented. Furthermore, a comparison is being made between ANNs with the more traditional machine learning approaches, providing theoretical justifications for choosing ANNs for modeling natural language.

Moreover, a text pre-processing tool was developed, for preparing the Twitter messages before passing them as inputs to the machine learning models. The tool is geared towards texts from social networks, which are very challenging to deal with, because of their informal and "creative" writing style, with improper use of grammar, figurative language, misspellings and slang. The text processing tool, is able to utilize most of the information in text, performing sentiment-aware tokenization, spell correction, word normalization, word segmentation (for splitting hashtags) and word annotation.

Finally, in the context of my research, we participated in Semeval-2017, which is an international competition for semantic evaluation of computational semantic analysis systems. The models that were developed for the participation in Semeval, were essentially the result of my research. A thorough analysis of these models is given, along with the rationale behind each design decision. The models were very competitive, achieving the first place in Task 4:"Sentiment Analysis in Twitter" and the second place in Task 6: "#HashtagWars: Learning a Sense of Humor".

Keywords: sentiment analysis, opinion mining, natural language processing, text classification, machine learning, artificial neural networks, deep neural networks

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές μου, Γιάννη Θεοδωρίδη, Νίκο Πελέκη και Χρήστο Δουλκερίδη, για την ευκαιρία που μου έδωσαν να ασχοληθώ με αυτό το αντικείμενο και την καθοδήγηση που μου παρείχαν καθ΄ όλη τη διάρκεια της έρευνάς μου. Με στήριξαν σε όλη την προσπάθεια μου και με ενθάρρυναν από την πρώτη στιγμή. Ήταν πάντα διαθέσιμοι, πρόθυμοι να συζητήσουν μαζί μου και να με κατευθύνουν. Είμαι ευγνώμων που είχα την ευκαιρία να συνεργαστώ μαζί τους.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Ακρωνύμια

BoW Bag of Words.

CC Classify & Count.

CNN Convolutional Neural Network.

DNN Deep Neural Networks.

FFNN Feed-Forward Neural Network.

GD Gradient Descent.

LDA Latent Dirichlet Allocation.

LSA Latent Semantic Analysis.

MLP Multi-Layer Perceptron.

PCA Principal Component Analysis.

PCC Probabilistic Classify & Count.

RNN Recurrent Neural Network.

SGD Stochastic Gradient Descent.

SVD Singular Value Decomposition.

SVM Support Vector Machine.

TreeNN Recursive Neural Network.

ΑΣ Ανάλυση Συναισθήματος.

ΕΦΓ Επεξεργασία Φυσικής Γλώσσας.

ΤΝΔ Τεχνητό Νευρωνικό Δίκτυο.

Κεφάλαιο 1

Εισαγωγή

Η γνώμη των άλλων είναι καθοριστική όταν πρέπει να παρθεί μία απόφαση. Όταν ένας άνθρωπος θέλει να αγοράσει ένα προϊόν συνήθως ρωτάει την γνώμη άλλων ανθρώπων ή αναζητά σχετικές κριτικές στο διαδίκτυο. Αντίστοιχα, όταν ένας οργανισμός, μία εταιρία ή ένας πολιτικός σχηματισμός χρειάζεται να πάρει μία απόφαση, λαμβάνει υπόψη του την γνώμη του κόσμου.

Τα τελευταία χρόνια, η ανάπτυξη του διαδικτύου και η άνθηση των υπηρεσιών που βασίζονται σε αυτό, έχει ως συνέπεια την δημιουργία ενός τεράστιου όγκου δεδομένων. Καθημερινά, σε υπηρεσίες κοινωνικής δικτύωσης (Facebook, Twitter), forums, blogs και ιστοσελίδες με ειδησιογραφικό περιεχόμενο, δημοσιεύονται και ανταλλάσσονται πάρα πολλές απόψεις. Τα δεδομένα αυτά είναι ένα χρυσωρυχείο πληροφοριών, στα οποία κρύβονται οι γνώμες πολλών ανθρώπων για διάφορα θέματα. Η εξαγωγή τους, μπορεί να βοηθήσει άτομα ή οργανισμούς στη διαδικασία λήψης αποφάσεων. Όμως, η ανακάλυψη και παρακολούθηση των συναισθημάτων και των απόψεων στο διαδίκτυο είναι μία σύνθετη διαδικασία. Είναι εξαιρετικά δύσκολο για έναν άνθρωπο να συλλέξει όλες τις χρήσιμες πληροφορίες από ένα τόσο μεγάλο όγκο δεδομένων. Συνεπώς υπάρχει η ανάγκη αυτοματισμού της διαδικασίας.

Η *Ανάλυση Συναισθήματος (ΑΣ)* ή αλλιώς *Εξόρυξη Γνώμης* έχει σαν αντικείμενο την ανάλυση κειμένων, για την αναγνώριση του συναισθηματικού προσανατολισμού (θετικού ή αρνητικού) ενός ανθρώπου προς μία οντότητα. Η οντότητα αυτή μπορεί να είναι ένα προϊόν, μία υπηρεσία, μία εταιρία, ένα πολιτικό κόμμα ή μία πεποίθηση. Η ΑΣ είναι ένα πρόβλημα με πολλές τεχνικές δυσκολίες. Ο βασική αιτία είναι ότι η φυσική γλώσσα είναι αδόμητη και δεν είναι εύκολο να μοντελοποιηθεί. Συνεπώς, το νόημα που φέρει μία πρόταση είναι δύσκολο να γίνει κατανοητό από ένα μηχάνημα.

Τα τελευταία χρόνια η ΑΣ έχει συγκεντρώσει μεγάλο ερευνητικό ενδιαφέρον. Η κινητήρια δύναμη πίσω από την ανάπτυξη του πεδίου είναι η διαθεσιμότητα μεγάλου όγκου δεδομένων, κυρίως από τα μέσα κοινωνικής δικτύωσης. Η αξιοποίηση αυτών των δεδομένων επιτρέπει την δημιουργία στατιστικών μοντέλων, τα οποία προσπαθούν να αναγνωρίσουν τα συντακτικά και σημασιολογικά χαρακτηριστικά της γλώσσας.

1.1 Επιστημονικές Προσεγγίσεις

Η γλώσσα είναι ένα από τα σημαντικότερα χαρακτηριστικά που κάνουν τον άνθρωπο να ξεχωρίζει από τα υπόλοιπα ζώα. Η δημιουργία αλγορίθμων ή μοντέλων που να μπορούν να κατανοήσουν φυσική γλώσσα, είναι ένα από τα πιο καθοριστικά βήματα για την δημιουργία ευφυών μηχανών (Τεστ Τούρινγκ (Turing 1950)). Η ΑΣ γεννήθηκε μέσα από το πεδίο της *Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ)*, η οποία με τη σειρά της είναι ένα από τα βασικά ερευνητικά αντικείμενα της Τεχνητής Νοημοσύνης.

Για την σχεδίαση ενός συστήματος ΕΦΓ, υπάρχουν δύο προϋποθέσεις. Αρχικά απαιτείται ένα θεωρητικό πλαίσιο. Αυτό το πλαίσιο αφορά την άντληση ιδεών από θεωρίες άλλων επιστημών, όπως τη Ψυχολογία, τη Γλωσσολογία, τη Γνωσιακή Επιστήμη και την Νευροβιολογία. Αυτές οι θεωρίες προτείνουν μοντέλα για την δομή και τη λειτουργία της γλώσσας, όπως τη σύνταξη και τη σημασιολογία, αλλά ακόμη και τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος μαθαίνει και χρησιμοποιεί τη γλώσσα. Με αυτό τον τρόπο, συγκεντρώνουμε ένα σύνολο από κανόνες ή αρχές, βάση των οποίων αποφασίζουμε το πως θα προσεγγίσουμε το εκάστοτε πρόβλημα.

Στη συνέχεια, αφού έχει επιλεχθεί το θεωρητικό πλαίσιο, θα πρέπει να μοντελοποιηθεί με μαθηματικό τρόπο ώστε να υλοποιηθεί το αντίστοιχο λογισμικό. Αυτό είναι και το αντικείμενο μελέτης αυτής της εργασίας, όπου γίνεται προσπάθεια για την καλύτερη δυνατή μοντελοποίηση της φυσικής γλώσσας, με στόχο την ΑΣ. Ο πιο επιτυχημένος τρόπος για την μοντελοποίηση προβλημάτων ΕΦΓ είναι με τη χρήση τεχνικών *Μηχανικής Μάθησης* (Machine Learning). Η μηχανική μάθηση αφορά τη δημιουργία στατιστικών μοντέλων μόνο από δεδομένα, χωρίς ρητό προγραμματισμό από τον άνθρωπο, με σκοπό την επίλυση ενός συγκεκριμένου προβλήματος.

Τα τελευταία χρόνια έχει αναζωπυρωθεί το ενδιαφέρον για ένα συγκεκριμένο είδος τεχνικών μηχανικών μάθησης, τα *Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ)*. Τα ΤΝΔ είναι εμπνευσμένα από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου και βασίζονται σε ένα πολύ απλοϊκό μοντέλο του ανθρώπινου νευρώνα. Είναι σημαντικό να σημειωθεί ότι τα ΤΝΔ δεν αποτελούν πιστό μοντέλο των Βιολογικών Νευρωνικών Δικτύων, αλλά αντλούν έμπνευση από αυτά για τον σχεδιασμό νέων στατιστικών μοντέλων.

Τα ΤΝΔ αν και έχουν μεγάλες απαιτήσεις σε υπολογιστικούς πόρους, λόγω της μεγάλης πολυπλοκότητας τους, έχουν την ιδιότητα να μοντελοποιούν πιο αφηρημένα χαρακτηριστικά χωρίς τον εκ των προτέρων ορισμό τους. Σε αυτό το πολύ σημαντικό πλεονέκτημα, έναντι των υπόλοιπων στατιστικών τεχνικών, οφείλεται η μεγάλη απήχησή τους. Η χρήση ΤΝΔ είναι ένας από τους βασικούς λόγους για την μεγάλη πρόοδο που έχει σημειωθεί πρόσφατα σε τομείς όπως η Υπολογιστική Όραση και η Επεξεργασία Φυσικής Γλώσσας.

1.2 Twitter

Ο βασικός λόγος για την ανάπτυξη του πεδίου της ΑΣ ήταν η διαθεσιμότητα δεδομένων στο διαδίκτυο. Η ύπαρξη ποιοτικών και μεγάλων συνόλων δεδομένων, είναι απαραίτητη για την δημιουργία καλών στατιστικών μοντέλων. Με την ανάπτυξη του διαδικτύου έκαναν την εμφάνισή τους ιστοσελίδες στις οποίες οι χρήστες μπορούσαν να γράψουν τη γνώμη τους για προϊόντα και υπηρεσίες όπως, Rotten-

Tomatoes[1], IMDb[2], Amazon[3], Epinions[4]. Απόψεις στο διαδίκτυο εκφράζονται όχι μόνο σε ειδικές ιστοσελίδες αλλά και σε άλλες όπως blogs και forums. Οι πρώτες εργασίες στην ΤΝΔ, αφορούσαν την δημιουργία μοντέλων για την εύρεση του συναισθηματικού προσανατολισμού κειμένων και φράσεων σε αξιολογήσεις ταινιών (Pang κ.ά. 2002; Turney 2002; Pang κ.ά. 2004) και προϊόντων (Blitzer κ.ά. 2007).

Όμως, τα τελευταία χρόνια ιδιαίτερη έμφαση έχει δοθεί στην αξιοποίηση των δεδομένων που παράγονται στα κοινωνικά δίκτυα. Τα κοινωνικά δίκτυα προσφέρουν την δυνατότητα στους χρήστες τους, να μοιραστούν άμεσα οποιαδήποτε πληροφορία επιθυμούν, όπως απόψεις, εικόνες και βίντεο. Αποτελούν πλέον ένα βασικό κομμάτι της κοινωνικής ζωής πάρα πολλών ανθρώπων. Ένα ακόμη σημαντικό χαρακτηριστικό των κοινωνικών δικτύων είναι, ότι παράγουν μία συνεχή ροή πληροφοριών. Συχνά πολλές ειδήσεις αναρτώνται και αναμεταδίδονται πρώτα μέσα από μηνύματα σε κοινωνικά δίκτυα, πριν ακόμα δημοσιευθούν από τα παραδοσιακά μέσα μαζικής ενημέρωσης. Αυτό έχει επιφέρει μεγάλες αλλαγές στον τρόπο με τον οποίο ενημερώνονται και επικοινωνούν οι άνθρωποι σήμερα.

Το μεγαλύτερο ενδιαφέρον το έχει συγκεντρώσει το Twitter, το οποίο είναι ένα *microblogging* κοινωνικό δίκτυο. Ένα *blog* (σύμπτυξη της έκφρασης *web-log*) είναι ένας ιστοχώρος στο διαδίκτυο σε μορφή ημερολογίου, όπου ο συγγραφέας εκφράζει απόψεις και ιδέες για θέματα που τον απασχολούν. Το microblogging διαφέρει από το παραδοσιακό blogging, ως προς την έκταση των κειμένων, τα οποία αποτελούνται συνήθως από μικρές προτάσεις. Το Twitter είναι το πιο δημοφιλές μέσο *microblogging* με περισσότερους από 300 εκατομμύρια χρήστες. Η σημασία του είναι τέτοια που αποτελεί αντικείμενο μελέτης από διάφορες επιστήμες.

Ο λόγος για την τόσο μεγάλη απήχηση αυτού του μέσου, είναι η αμεσότητα και η απλότητα των μηνυμάτων. Το σύνθημα του Twitter είναι "Τί συμβαίνει?" ("What is happening?") και έχει ως στόχο να ωθήσει τους χρήστες να σχολιάσουν τι συμβαίνει γύρω τους. Το κύριο χαρακτηριστικό του είναι ότι η έκτασης των μηνυμάτων περιορίζεται στους 140 χαρακτήρες, το οποίο σημαίνει ότι ένα μήνυμα αποτελείται πρακτικά από μία ή δύο προτάσεις το πολύ. Αυτό ωθεί τους χρήστες να γράφουν πιο περιεκτικά μηνύματα. Επίσης είναι δυνατό να γίνουν συζητήσεις ανάμεσα σε πολλούς χρήστες, οι οποίες είναι ορατές σε όσους τους ακολουθούν.

Ένα άλλο σημαντικό χαρακτηριστικό αφορά την εύκολη αναμετάδοση των μηνυμάτων (retweet). Εάν ένας χρήστης διαβάσει ένα μήνυμα το οποίο θέλει να διαδώσει, μπορεί να το κάνει πολύ εύκολα αναμεταδίδοντας το σε όσους τον ακολουθούν. Ομοίως, μερικοί από τους ακολούθους του αρχικού χρήστη, ενδεχομένως να αποφασίσουν να το αναμεταδώσουν στους δικούς τους ακολούθους. Όπως είναι προφανές, αυτό μπορεί να οδηγήσει σε μία "αλυσιδωτή αντίδραση", με αποτέλεσμα οι σημαντικές ειδήσεις ή σχόλια να διαδίδονται πάρα πολύ γρήγορα σε μεγάλη μάζα χρηστών.

Πλέον, η διείσδυση του μέσου είναι πολύ μεγάλη στην κοινωνία, ειδικά στις αναπτυγμένες χώρες. Εταιρίες και πολιτικοί σχηματισμοί, δραστηριοποιούνται στο Twitter με σκοπό να προσεγγίσουν μεγάλο κομμάτι του πληθυσμού και να βελτιώσουν την εικόνα τους. Επίσης αυτό προσφέρει την ευκαιρία σε πολλούς ανθρώπους να αλληλεπιδράσουν και να εκφράσουν τη γνώμη τους, σε εταιρίες, κόμματα ή ανθρώπους που υπό άλλες συνθήκες θα ήταν δύσκολο ή και ακόμη αδύνατο να το

[1] www.rottentomatoes.com
[2] www.imdb.com
[3] www.amazon.com
[4] www.epinions.com

κάνουν. Ένα αντιπροσωπευτικό πρόσφατο παράδειγμα είναι οι προεδρικές εκλογές στις Η.Π.Α. για το 2016, όπου το Twitter ήταν η βασική πηγή νέων ειδήσεων, με την δημοσίευση 40 εκατομμυρίων μηνυμάτων σχετικά με τις εκλογές (Isaac κ.ά. 2016).

1.3 Εφαρμογές

Η ΑΣ έχει πολλές εφαρμογές στο μάρκετινγκ, τις επιχειρήσεις, την πολιτική ακόμα και το χρηματιστήριο. Μία ομάδα εφαρμογών αφορά την πρόβλεψη της συμπεριφοράς των ανθρώπων. Με την εφαρμογή τεχνικών ΑΣ μπορούν να εξαχθούν συμπεράσματα σχετικά με την καταναλωτική συμπεριφορά του κόσμου, όπως ποιες ταινίες είναι πιθανό να θέλουν να δουν ή ποια κινητά να αγοράσουν, αλλά και την συμπεριφορά του εκλογικού σώματος, προβλέποντας ποιους υποψηφίους είναι πιθανότερο να ψηφίσουν. Επιπρόσθετα, η συναισθηματική κατάσταση του κόσμου (ανησυχία, εμπιστοσύνη κλπ.) είναι σημαντικός δείκτης για πολλά οικονομικά μοντέλα.

Επιπλέον, όπως έχει δειχθεί από τους (Bollen κ.ά. 2011b), αναλύοντας μηνύματα στο Twitter, παρατηρείται ότι κοινωνικοοικονομικές αναταραχές επηρεάζουν τη συναισθηματική κατάσταση των ανθρώπων σε 6 συναισθηματικούς άξονες (όπως θυμό, σύγχυση, κόπωση κλπ.). Αυτό το συμπέρασμα αποτυπώνεται στα μηνύματα των χρηστών στα κοινωνικά δίκτυα, αναδεικνύοντας την αμφίδρομη σχέση ανάμεσα στη συναισθηματική κατάσταση των ανθρώπων μίας κοινωνίας και στα γεγονότα που συμβαίνουν σε αυτή. Έτσι, ένα άλλο είδος εφαρμογής της ΑΣ αφορά την πρόβλεψη ορισμένων γεγονότων ή έλεγχο της εξέλιξης τους, παρατηρώντας τις αυξομοιώσεις σε ορισμένους συναισθηματικούς δείκτες.

1.3.1 Επιχειρήσεις

Για μία επιχείρηση είναι σημαντικό να γνωρίζει την γνώμη των πελατών της για τα προϊόντα ή τις υπηρεσίες της ώστε να μπορεί να τα βελτιώσει και να σχεδιάσει τη στρατηγική της. Επίσης είναι σημαντικό να γνωρίζει την γνώμη του κόσμου για τους ανταγωνιστές της και να ανακαλύψει ποια είναι τα συγκριτικά της πλεονεκτήματα και σε ποιους τομείς υστερεί, ώστε να βελτιώσει την ανταγωνιστικότητά της. Για την επίτευξη αυτών των στόχων είναι εξαιρετικά χρήσιμη η αξιοποίηση των αξιολογήσεων (reviews) σε μεγάλα ηλεκτρονικά καταστήματα (όπως Amazon) αλλά και σε μηνύματα σε κοινωνικά δίκτυα (Twitter, Facebook). Τα δεδομένα αυτά είναι σχετικά εύκολο να συγκεντρωθούν, όμως είναι δύσκολο να εξαχθούν χρήσιμες πληροφορίες από αυτά.

Οι πρώτες εφαρμογές συστημάτων ΑΣ αφορούσαν την εξαγωγή γνώμης από κριτικές ταινιών (Pang κ.ά. 2002; Turney 2002; Pang κ.ά. 2005), ηλεκτρονικών συσκευών (M. Hu κ.ά. 2004; Popescu κ.ά. 2005), βιβλίων (Blitzer κ.ά. 2007) κλπ. Η απλή αναγνώριση του συναισθηματικού προσανατολισμού των αξιολογήσεων πολλές φορές δεν είναι αρκετή. Μεγαλύτερη αξία έχει η αναγνώριση του προσανατολισμού ως προς τις ιδιότητες του προϊόντος αλλά και η παραγωγή περιλήψεων (M. Hu κ.ά. 2004; Pang κ.ά. 2004; Zhuang κ.ά. 2006), όπου θα συνοψίζονται τα κυριότερα σημεία και οι παρατηρήσεις των αγοραστών. Με αυτό τον τρόπο οι υπεύθυνοι λήψης αποφάσεων θα μπορούν να έχουν γρήγορα την γενική εικόνα.

Εκτός από την εξαγωγή πληροφοριών σχετικά με τις προτιμήσεις του αγοραστικού κοινού, το επόμενο βήμα είναι η πρόβλεψη της συμπεριφοράς του. Στόχος είναι η πρόβλεψη των πωλήσεων ενός προϊόντος και ο σχεδιασμός προϊόντων τα οποία θα έχουν την μεγαλύτερη δυνατή απήχηση στην αγορά. Μία από τις πρώτες έρευνες (Basuroy κ.ά. 2003) ανακάλυψαν συσχέτιση ανάμεσα στις αξιολογήσεις των κριτικών κινηματογράφου και στις πωλήσεις εισιτηρίων των ταινιών. Ακόλουθες έρευνες ανακάλυψαν ότι οι κριτικές των καταναλωτών μπορούν να βοηθήσουν στην εκτίμηση των πωλήσεων σε βιβλία (Chevalier κ.ά. 2006), ταινίες (Yong Liu 2006; Mishne κ.ά. 2006; Yang Liu κ.ά. 2007), ακόμα και σε βιντεοπαιχνίδια (Zhu κ.ά. 2010). Όπως είναι αναμενόμενο, οι πρόσφατες έρευνες αξιοποιούν το συναίσθημα και τις γνώμες στα κοινωνικά δίκτυα (Jansen κ.ά. 2009; Asur κ.ά. 2010; Rui κ.ά. 2013; Arias κ.ά. 2014) με στόχο να ανακαλύψουν περισσότερα σχετικά με την συμπεριφορά των καταναλωτών (Chamlertwat κ.ά. 2012).

1.3.2 Χρηματιστήριο

Η διάθεση των ανθρώπων επηρεάζει την συμπεριφορά τους και κατά συνέπεια την οικονομία μίας κοινωνίας. Όπως έχει δειχθεί από τους (Lemmon κ.ά. 2006; Han 2008) υπάρχει συσχέτιση ανάμεσα στο συναίσθημα των επενδυτών και στις διακυμάνσεις των αγορών. Επιπλέον, ο (Tetlock 2007) μελετώντας την επίδραση των μέσων ενημέρωσης στο χρηματιστήριο, δείχνει ότι ασκούν επιρροή με εντονότερη αυτή των αρνητικών νέων. Ακόμη, οι (Gilbert κ.ά. 2010) αναλύοντας τις γνώμες και τη διάθεση εκατομμυρίων χρηστών σε μία μεγάλη διαδικτυακή κοινότητα (LiveJournal), ανακαλύπτουν συσχετισμούς με τις τιμές των μετοχών στο χρηματιστήριο (S&P 500). Αυτές οι παρατηρήσεις έχουν ωθήσει στην ανάπτυξη συστημάτων ΑΣ τα οποία εξάγουν απόψεις και συναισθηματικά χαρακτηριστικά από τίτλους ειδήσεων, άρθρα (Dougal κ.ά. 2012) και μηνύματα σε κοινωνικά δίκτυα, επιδιώκοντας να προβλέψουν (αυτόνομα ή ως μέρος μεγαλύτερων συστημάτων) τις αγορές.

Το μεγαλύτερο ερευνητικό ενδιαφέρον το έχει συγκεντρώσει η δημιουργία στατιστικών μοντέλων τα οποία βασίζονται σε μηνύματα στο Twitter. Η πρώτη έρευνα στην οποία σχεδιάζεται ένα τέτοιο μοντέλο είναι η (Bollen κ.ά. 2011a), στην οποία οι ερευνητές παρατηρούν θετική συσχέτιση ανάμεσα στη συνολική συναισθηματική κατάσταση στα μηνύματα στο Twitter και σε χρηματιστηριακούς δείκτες (DJIA) σημειώνοντας ότι η εισαγωγή των σχετικών χαρακτηριστικών σε μοντέλα προβλέψεων βελτιώνει την ακρίβειά τους. Έκτοτε έχουν πραγματοποιηθεί αρκετές εργασίες (Oh κ.ά. 2011; Xue Zhang κ.ά. 2011; Makrehchi κ.ά. 2013; Si κ.ά. 2013; Smailović κ.ά. 2013, 2014; Sprenger κ.ά. 2014) και στην (Khadjeh Nassirtoussi κ.ά. 2014) γίνεται μία σύνοψη της έρευνας στο πεδίο. Τέλος, στη (Yu κ.ά. 2013) γίνεται σύγκριση της επιρροής των μέσων κοινωνικής δικτύωσης και των "παραδοσιακών" μέσων, καταλήγοντας στο ότι και τα δύο μέσα παίζουν ρόλο, αλλά τα κοινωνικά δίκτυα έχουν πιο ισχυρή επιρροή.

1.3.3 Πολιτική

Τα μέσα κοινωνικής δικτύωσης προσφέρουν την δυνατότητα στους ανθρώπους να συμμετάσχουν σε πολιτικές συζητήσεις σχετικά με θέματα που επηρεάζουν τη ζωή τους. Αυτό έχει ιδιαίτερη αξία στις αναπτυσσόμενες χώρες, με ίσως πιο αντιπροσωπευτικό παράδειγμα, τα γεγονότα της *Αραβικής Άνοιξης*, όπου κοινό χα-

ρακτηριστικό των κινητοποιήσεων, ήταν ότι οργανώθηκαν κατά βάση μέσα από υπηρεσίες όπως το Facebook και το Twitter.

Στην πολιτική, η κοινή γνώμη είναι καθοριστικής σημασίας για κόμματα ή υποψηφίους, καθώς επηρεάζει σε μεγάλο βαθμό την χάραξη της στρατηγικής τους. Η ΑΣ έχει δύο βασικές εφαρμογές στην πολιτική, 1) την σφυγμομέτρηση σε πραγματικό χρόνο (O'Connor κ.ά. 2010; Maynard κ.ά. 2011; Stieglitz κ.ά. 2012; Wang κ.ά. 2012; Zhou κ.ά. 2013; Ceron κ.ά. 2014) και κατά συνέπεια 2) την πρόγνωση των εκλογικών αποτελεσμάτων (το οποίο όμως όπως εξηγείται στη συνέχεια είναι αμφιλεγόμενο) (Ceron κ.ά. 2014).

Η πραγματοποίηση δημοσκοπήσεων είναι ακριβή και χρονοβόρα διαδικασία. Απαιτείται ο προσεκτικός σχεδιασμός ερωτηματολογίων, καθώς είναι πιθανό να παραληφθούν χρήσιμες ερωτήσεις, ή ορισμένες από αυτές να διατυπωθούν λάθος. Επιπλέον, ένα βασικό πρόβλημα είναι η χρονική διάρκεια των δημοσκοπήσεων. Για παράδειγμα, έστω ότι προκύπτει ένα σημαντικό ζήτημα (φυσική καταστροφή, πολιτική κρίση, πολιτικό σκάνδαλο κ.λ.π.) σε μία χώρα και μία εταιρία αναλαμβάνει να εκτελέσει μία δημοσκόπηση. Στόχος της δημοσκόπησης είναι η ανάλυση της γνώμης των πολιτών για τον τρόπο διαχείρισης της κατάστασης από κάθε πολιτικό φορέα. Η δημοσκόπηση διαρκεί λίγες μέρες, όμως στο διάστημα αυτό είναι πιθανό να έχουν προκύψει εξελίξεις, οι οποίες δεν θα έχουν αποτυπωθεί στις απαντήσεις όσων ερωτήθηκαν στις αρχές της δημοσκόπησης. Συνεπώς, τα αποτελέσματα δεν θα είναι τα πλέον αντιπροσωπευτικά.

Αυτά τα προβλήματα μπορούν να αντιμετωπισθούν με την πραγματοποίηση σφυγμομετρήσεων της κοινής γνώμης, σε πραγματικό χρόνο. Αυτό επιτυγχάνεται με την χρήση συστημάτων ΑΣ, αναλύοντας τις γνώμες που διατυπώνονται στα μέσα κοινωνικής δικτύωσης. Αρχικά, αυτού του είδους η σφυγμομέτρηση είναι παθητική και από τη στιγμή που συγκεντρωθούν τα δεδομένα (μηνύματα, σχόλια, άρθρα κλπ.) δεν υπάρχει περιορισμός στο πλήθος και το είδος των ερωτημάτων που θα εκτελεστούν. Έτσι, αν κριθεί σε δευτερεύοντα χρόνο ότι θα ήταν χρήσιμο να εκτελεστεί ένα νέο ερώτημα στα δεδομένα, για το οποίο δεν είχε προκύψει η ανάγκη μέχρι πρότινος, αυτό είναι εφικτό. Επιπλέον, είναι δυνατό οι ενδιαφερόμενοι να έχουν άμεσα τα αποτελέσματα στη διάθεση τους. Αυτό σημαίνει ότι ένα κόμμα ή ένας υποψήφιος μπορεί να είναι διαρκώς ενήμερος για τις τάσεις της κοινής γνώμης και κατά συνέπεια να λαμβάνει πιο ενημερωμένες αποφάσεις.

Σχετικά με την ικανότητα της ΑΣ να συνεισφέρει στην πρόβλεψη των εκλογικών αποτελεσμάτων, υπάρχει διαφωνία στην επιστημονική κοινότητα. Η πρώτη έρευνα η οποία αποτέλεσε την αρχή της συζήτησης είναι η (Tumasjan κ.ά. 2010), στην οποία οι ερευνητές ανέλυσαν μηνύματα στο Twitter σχετικά με τις ομοσπονδιακές εκλογές της Γερμανίας του 2009 και κατέληξαν στο συμπέρασμα ότι το μοντέλο τους προσεγγίζει σε μεγάλο βαθμό τα αποτελέσματα των κανονικών δημοσκοπήσεων. Η αντίθετη άποψη παρουσιάζεται στην δημοσίευση (Jungherr κ.ά. 2012), όπου οι συγγραφείς εξηγούν τους λόγους για τους οποίους αμφισβητούν την προβλεπτική ικανότητα του προηγούμενου μοντέλου αλλά και στην (Metaxas κ.ά. 2011) όπου οι συγγραφείς περιγράφουν τρεις αναγκαίες συνθήκες που θα πρέπει να ικανοποιεί ένα μοντέλο, ώστε να μπορεί να προσεγγίσει τα αποτελέσματα των παραδοσιακών δημοσκοπήσεων. Από την άλλη μεριά, σε δύο σχετικά πρόσφατες εργασίες (Zhou κ.ά. 2013; Ceron κ.ά. 2014) παρουσιάζονται θετικά αποτελέσματα. Τέλος, στο (Jungherr 2016) γίνεται μία αναλυτική σύγκριση 127 ερευνών.

1.3.4 Λοιπές Εφαρμογές

Εκτός από τις εφαρμογές που ήδη αναφέρθηκαν, στην βιβλιογραφία έχουν προταθεί και πολλές άλλες. Ένα παράδειγμα είναι η πρόγνωση μελλοντικών συμβάντων. Η συνήθης προσέγγιση σε αυτό το πρόβλημα είναι η παρακολούθηση των θεμάτων που συζητούνται στο διαδίκτυο (Sakaki κ.ά. 2010; Aramaki κ.ά. 2011; Weng κ.ά. 2011) ή των ερωτημάτων σε μηχανές αναζήτησης (Culotta 2010). Όμως, για ορισμένα συμβάντα αυτό δεν είναι αρκετό. Η υπόθεση είναι ότι έντονες αυξομειώσεις στο συναίσθημα που φέρουν τα μηνύματα των χρηστών σε κοινωνικά δίκτυα όπως στο Twitter, καθώς και των πεποιθήσεων που εκφράζονται σε αυτά, είναι ένδειξη ενός συμβάντος (Thelwall κ.ά. 2011; Kavanaugh κ.ά. 2012; Y. Hu κ.ά. 2013). Με τον τρόπο αυτό δίνεται η ευκαιρία της πιο άμεσης διαχείρισής τους. Ακόμη, έχει δειχθεί ότι με αυτό τον τρόπο είναι εφικτό να εκτιμηθεί η εξέλιξη ήδη γνωστών γεγονότων όπως επιδημιών (Ritterman κ.ά. 2009).

1.4 Στόχοι της εργασίας

Ο στόχος της εργασίας είναι η έρευνα στο πεδίο της ΑΣ και η ανάπτυξη ενός μοντέλου, για την εύρεση του συναισθηματικού προσανατολισμού σε μηνύματα από το Twitter. Η ΑΣ είναι ένα ερευνητικό πεδίο με μεγάλο ενδιαφέρον, τόσο από την επιστημονική κοινότητα, όσο και από την βιομηχανία, λόγω της πληθώρας των εφαρμογών. Η ΑΣ στο Twitter είναι ακόμη δυσκολότερο πρόβλημα, εξαιτίας της ιδιαιτερότητας του μέσου. Για την ανάπτυξη του αντίστοιχου μοντέλου (ή μοντέλων), αρχικά απαιτείται μία κατανόηση της ΑΣ σε θεωρητικό επίπεδο και στη συνέχεια των πιο σημαντικών προσεγγίσεων, οι οποίες έχουν δοκιμαστεί από την επιστημονική κοινότητα.

Πιο αναλυτικά, οι βασικοί στόχοι της εργασίας είναι:

- Σύνοψη της ΑΣ. Αυτό αφορά την παρουσίαση του θεωρητικού πλαισίου σχετικά με την ΑΣ (υπό το πρίσμα της υπολογιστικής γλωσσολογίας) και μία ιστορική αναδρομή αναφορικά με τις τεχνικές που έχουν χρησιμοποιηθεί για την υλοποίηση μοντέλων ΑΣ.

- Έρευνα στις τεχνικές μηχανικής μάθησης για την ανάπτυξη μοντέλων ΕΦΓ. Αυτό αφορά την έρευνα στους διάφορους τρόπους για την αναπαράσταση κειμένων (εξαγωγή χαρακτηριστικών) και στους αλγορίθμους μηχανικής μάθησης για την δημιουργία μοντέλων, τα οποία θα αξιοποιούν αυτές τις αναπαραστάσεις. Η μεγαλύτερη βαρύτητα θα δοθεί στην έρευνα σε τεχνικές βαθιών τεχνητών νευρωνικών δικτύων, καθώς τα τελευταία χρόνια έχουν κυριαρχήσει στο πεδίο της ΕΦΓ.

- Ανάπτυξη μοντέλου ΑΣ για μηνύματα στο Twitter. Αυτός είναι ο απώτερος στόχος της εργασίας και της έρευνας. Για την καλύτερη σύγκριση των μοντέλων των οποίων θα αναπτυχθούν στα πλαίσια της εργασίας, ο ιδανικότερος τρόπος είναι η συμμετοχή στον διαγωνισμό "Sentiment Analysis in Twitter" του SemEval. Έτσι, συγκρίνοντας τις επιδόσεις των μοντέλων μας, με αυτά από ομάδες από όλο των κόσμο, θα έχουμε καλύτερη εικόνα για την πραγματική ποιότητα και χρησιμότητάς τους.

1.5 Δομή της εργασίας

Το πρώτο τμήμα της εργασίας ασχολείται με την σύνοψη του πεδίου της ΑΣ (Κεφ. 2). Αρχικά δίνεται ένας πιο αυστηρός ορισμών των σχετικών εννοιών (όπως, τι είναι η άποψη στο πλαίσιο της ΑΣ). Στην συνέχεια καταγράφονται οι σημαντικότερες προσεγγίσεις για την δημιουργία συστημάτων και μοντέλων για την εφαρμογή ΑΣ σε κείμενα. Αυτό το κεφάλαιο είναι ιδανικό για κάποιον, ο οποίος θέλει να αποκτήσει γρήγορα μία εικόνα για το πεδίο της ΑΣ, χωρίς τις τεχνικές λεπτομέρειες.

Στο επόμενο κεφάλαιο (Κεφ. 3), καταγράφονται οι κυρίαρχες τεχνικές για την μοντελοποίηση κειμένων. Αυτό αφορά την αναπαράσταση ενός κειμένου με μαθηματική μορφή, δηλαδή ως ένα διάνυσμα χαρακτηριστικών. Στη συνέχεια, αυτές οι αναπαραστάσεις μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων μηχανικής μάθησης ή εξαγωγή γνώσης από κείμενα. Για κάθε μία προσέγγιση, αναλύονται οι τεχνικές με τις οποίες μπορεί να εξάγει κανείς τα αντίστοιχα χαρακτηριστικά. Ακόμη, παρουσιάζονται και ορισμένες τεχνικές για την προ-επεξεργασία των κειμένων, με στόχο την προετοιμασία τους πριν την μοντελοποίησή τους. Το κεφάλαιο αυτό, απευθύνεται σε όσους έχουν ενδιαφέρον για την ΕΦΓ, και θα ήθελαν μία σύνοψη των διάφορων τεχνικών για την εξαγωγή αναπαραστάσεων από κείμενα.

Συνεχίζοντας στο κεφάλαιο 4, γίνεται καταγραφή των δημοφιλέστερων προσεγγίσεων μηχανικής μάθησης για προβλήματα ΕΦΓ. Οι τεχνικές που παρουσιάζονται αφορούν κυρίως την κατηγοριοποίηση κειμένων, καθώς κατά βάση έτσι προσεγγίζεται η ΑΣ (αν και στο Κεφ. 5, παρουσιάζονται και άλλα μοντέλα). Γίνεται σύγκριση των παραδοσιακών τεχνικών, όπως Naive Bayes και SVM, με τις νέες αρχιτεκτονικές Βαθιών Νευρωνικών Δικτύων (Deep Neural Networks - DNNs, η πιο απλά Deep Learning), όπως RNNs, CNNs. Ιδιαίτερη έμφαση δίνεται στις τεχνικές deep learning, και παρουσιάζονται τα θεωρητικά και πρακτικά πλεονεκτήματα που έχουν, σε σχέση με τις πιο παραδοσιακές προσεγγίσεις. Το κεφάλαιο αυτό απευθύνεται σε αυτούς που έχουν μια εμπειρία από προβλήματα ΕΦΓ, πιθανότατα με την χρήση τεχνικών όπως Naive Bayes και θα ήθελαν να μάθουν τι νέο έχουν να προσφέρουν οι τεχνικές deep learning.

Τέλος, στο κεφάλαιο 5 παρουσιάζονται ορισμένα μοντέλα, τα οποία υλοποιήθηκαν για την εφαρμογή ΑΣ σε μηνύματα από το Twitter. Αυτό το κεφάλαιο είναι το ουσιαστικό αποτέλεσμα της έρευνας μου και πρακτικά αφορά τα μοντέλα τα οποία αναπτύχθηκαν για την συμμετοχή στον διαγωνισμό SemEval-2017, Task 4 “Sentiment Analysis in Twitter”. Γίνεται αναλυτική παρουσίαση των αντίστοιχων υποκατηγοριών του διαγωνισμού και των μοντέλων, των οποίων σχεδιάστηκαν για κάθε υποκατηγορία. Επίσης, παρουσιάζεται σύγκριση διάφορων μοντέλων (SVMs, SVMs + Feature Selection, RNNs, RNNs + Attention), εκτός αυτών που στάλθηκαν στον διαγωνισμό. Το κεφάλαιο αυτό απευθύνεται σε αυτούς που έχουν γνώση των τεχνικών deep learning και θα ήθελαν να δουν παραδείγματα εφαρμογής τους σε ένα πρόβλημα ΑΣ.

Κεφάλαιο 2

Σύνοψη της Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος (ΑΣ) αποτελεί στην ουσία έναν όρο "ομπρέλα" για μία ομάδα προβλημάτων, τα οποία έχουν πολλά κοινά μεταξύ τους όπως, αναγνώριση υποκειμενικότητας (subjectivity detection), εξόρυξη γνώμης (opinion mining), εύρεση συναισθηματικού προσανατολισμού (sentiment polarity) κλπ. Είναι εύκολο να γίνει σύγχυση των εννοιών καθώς στη βιβλιογραφία συχνά οι όροι αυτοί αναφέρονται αδιακρίτως. Όμως υπάρχουν διαφορές μεταξύ τους, οι οποίες πρέπει να επισημανθούν.

Σε αυτό το κεφάλαιο θα δοθεί ένας πιο αυστηρός ορισμός των εννοιών που σχετίζονται με τη ΑΣ, όπως συναίσθημα ή γνώμη. Επίσης θα γίνει καταγραφή των κύριων προσεγγίσεων, αναφορικά με το κομμάτι της μοντελοποίησης, αναλύοντας τα βασικά τους χαρακτηριστικά. Θα δοθεί ιδιαίτερη έμφαση στις τεχνικές μηχανικής μάθησης. Έτσι, θα σχηματιστεί η "γενική εικόνα", η οποία θα περιγράφει το ερευνητικό πεδίο της ΑΣ και η οποία θα είναι σημείο αναφοράς, για τα επόμενα κεφάλαια.

2.1 Ορισμοί εννοιών

Αρχικά, θα ορίσουμε τις βασικές έννοιες της ΑΣ. Ως *γνώμη* θα ορίσουμε την στάση και την αξιολόγηση ενός ανθρώπου, για μία *οντότητα*. Η *οντότητα* αποτελεί μία αφηρημένη έννοια, η οποία μπορεί να αντιστοιχεί σε έναν άνθρωπο, ένα προϊόν, μία υπηρεσία ή μία πεποίθηση.

Μία οντότητα μπορεί να αποτελείται από ορισμένες πτυχές (aspects), οι οποίες αντιστοιχούν σε ιδιότητες ή χαρακτηριστικά τα οποία την προσδιορίζουν. Για παράδειγμα, ένα κινητό τηλέφωνο είναι μία *οντότητα*, για την οποία ένας άνθρωπος μπορεί να έχει μία *γνώμη*. Το κινητό αποτελείται από ένα σύνολο από χαρακτηριστικά, όπως μπαταρία ή οθόνη, τα οποία είναι μερικές από τις *πτυχές* του.

Μία γνώμη δεν αφορά απαραίτητα την οντότητα μόνο ως σύνολο, αλλά συχνά και τις επιμέρους πτυχές της. Κάθε γνώμη φέρει ένα *συναίσθημα* (sentiment) με ένα *συναισθηματικό προσανατολισμό* (sentiment polarity), ο οποίος μπορεί να είναι θετικός, αρνητικός ή ουδέτερος. Συνεχίζοντας το προηγούμενο παράδειγμα, μπορεί η γνώμη κάποιου για την μπαταρία του κινητού να είναι θετική, ενώ η γνώμη του για την οθόνη του να είναι αρνητική.

2.1.1 Ορισμός γνώμης

Η γνώμη είναι μία πολυδιάστατη έννοια και δεν υπάρχει ένας ξεκάθαρος, κοινά αποδεκτός ορισμός στην βιβλιογραφία. Επιλέγουμε τον απλοποιημένο ορισμό, όπως δίνεται στο (B. Liu 2015). Ως γνώμη ορίζεται η τετράδα της μορφής:

$$(h, p, a, s, t) \tag{2.1}$$

Όπου:

- h είναι ο φορέας της γνώμης, δηλαδή ο άνθρωπος ή ο οργανισμός ο οποίος εκφράζει τη γνώμη.
- p είναι το αντικείμενο (ή στόχος) της γνώμης.
- s είναι το συναίσθημα που φέρει η γνώμη.
- a είναι μία πτυχή της γνώμης. Αυτή η ιδιότητα είναι προαιρετική και χρησιμοποιείται όταν η ΑΣ γίνεται σε επίπεδο χαρακτηριστικών.
- t είναι η χρονική στιγμή την οποία εκφράστηκε μία γνώμη. Αυτή είναι ακόμα μία προαιρετική ιδιότητα, η οποία παρόλο που έχει σημασία, καθώς μία γνώμη μπορεί να αλλάξει στο χρόνο, δεν χρησιμοποιείται συχνά στην πράξη.

2.1.2 Φορέας της γνώμης

Ο *φορέας της γνώμης* είναι συχνά ο συγγραφέας του κειμένου στο οποίο εκφράζεται μία γνώμη. Όμως αυτό δεν ισχύει πάντα, καθώς είναι πιθανό ο συγγραφέας να παραθέτει τη γνώμη ενός τρίτου προσώπου. Επιπλέον, ο φορέας της γνώμης μπορεί να μην είναι ένα φυσικό πρόσωπο, αλλά ένας οργανισμός ή μία ομάδα ανθρώπων. Για παράδειγμα μπορεί να είναι μία εφημερίδα, ή ένα πολιτικό κόμμα, το οποίο εκφράζει τη γνώμη του για ένα δημόσιο συμβάν, όπως το αποτέλεσμα ενός δημοψηφίσματος.

2.1.3 Αντικείμενο της γνώμης

Το αντικείμενο της γνώμης είναι η οντότητα, ή μία πτυχή της οντότητας, για την οποία εκφράζεται η γνώμη. Μπορεί επίσης να αναπαρασταθεί και ως μία ιεραρχία, στην οποία μία πτυχή είναι και αυτή με τη σειρά της πτυχή-χαρακτηριστικό μίας άλλης πτυχής. Για παράδειγμα, μία οντότητα θα μπορούσε να είναι ένα αυτοκίνητο, για το οποίο να εκφράζεται γνώμη για μία πτυχή του, την "αξιοπιστία", η οποία ομοίως να διακρίνεται σε άλλες πτυχές όπως "αντοχή κινητήρα" ή "ανθεκτικότητα πλαστικών".

2.1.4 Συναίσθημα της γνώμης

Το συναίσθημα είναι η πιο σημαντική ιδιότητα που προσδιορίζει μία γνώμη και αυτή η οποία έχει μελετηθεί περισσότερο από όλες τις άλλες, στο πλαίσιο της ΑΣ. Αφορά το σύνολο των συναισθημάτων, των ιδεών και των στάσεων που απορρέουν από μία γνώμη ή ένα κείμενο.

Το συναίσθημα μπορεί να αναλυθεί στα εξής χαρακτηριστικά:

Προσανατολισμός. Ο συναισθηματικός προσανατολισμός αφορά το κατά πόσο μία γνώμη είναι θετική ή αρνητική. Μία γνώμη μπορεί να έχει και ουδέτερο προσανατολισμό. Στην ουσία αφορά το συναισθηματικό "πρόσημο" μίας γνώμης.
Ένταση. Προσδιορίζει το πόσο δυνατό είναι το συναίσθημα που εκφράζεται. Η βαρύτητα ενός συναισθήματος μπορεί να δοθεί είτε με λέξεις, όπως "άριστος" αντί του "καλός", ή με ποσοτικά επιρρήματα, όπως "πολύ", ή παραθετικά επιρρημάτων, όπως "ελάχιστα".
Βαθμός. Ο βαθμός έχει αξία σε εφαρμογές ΑΣ, στις οποίες είναι χρήσιμο να ποσοτικοποιηθεί το συναίσθημα μίας γνώμης. Έτσι, συχνά απαιτείται ο καθορισμός ενός βαθμού ή σκορ για το συναίσθημα μίας γνώμης, συνήθως σε κλίμακες (1-3) ή (1-5).

2.2 Τα επίπεδα της ανάλυσης

Η έρευνα στην ΑΣ κατά κύριο λόγο εκτελείται σε τρία επίπεδα διακριτότητας: σε επίπεδο εγγράφου (document-level), σε επίπεδο πρότασης (sentence-level) και σε επίπεδο χαρακτηριστικών/πτυχών (aspect-/feature-level).

2.2.1 Επίπεδο Εγγράφου

Όταν εκτελούμε ανάλυση σε επίπεδο εγγράφου αντιμετωπίζομε ολόκληρο το έγγραφο σαν μία οντότητα. Αυτό σημαίνει ότι κάνουμε την παραδοχή ότι το κείμενο εκφράζει μία μόνο γνώμη και φέρει ένα μοναδικό συναίσθημα. Στόχος είναι η αναγνώριση ή και ποσοτικοποίηση του γενικού συναισθήματος που απορρέει από το κείμενο. Το πιο χαρακτηριστικό παράδειγμα είναι η εφαρμογή ΑΣ σε κριτικές προϊόντων ή ταινιών, ένα από τα πρώτα προβλήματα που μελετήθηκαν (Pang κ.ά. 2002; Turney 2002), όπου στόχος είναι η εύρεση του συναισθηματικού προσανατολισμού της κριτικής, δηλαδή αν η κριτική είναι θετική (thumbs-up) ή αρνητική (thumbs-down).

Όπως είναι προφανές, μία τέτοια ανάλυση έχει πολλούς περιορισμούς, καθώς χάνονται σημαντικές πληροφορίες. Σε ένα κείμενο, όπως σε μία αξιολόγηση, εκφράζονται πολλές γνώμες. Για παράδειγμα σε μία κριτική μίας κινηματογραφικής ταινίας, μπορεί ο κριτικός να θεωρεί ότι οι ερμηνείες των ηθοποιών ήταν καλές, αλλά η σκηνοθεσία και το σενάριο της ταινίας απογοητευτικά, με αποτέλεσμα να δίνει κακή βαθμολογία στη ταινία. Όμως, με αυτό τον τρόπο δεν αξιοποιούνται οι θετικές γνώμες για ορισμένες πτυχές της ταινίας, ούτε διακρίνονται οι λόγοι, για τους οποίους η ταινία κατέληξε να έχει αρνητική βαθμολογία.

Από την άλλη μεριά, αυτή η αφελής ανάλυση είναι πολύ πιο εύκολο να υλοποιηθεί. Ο λόγος είναι πως υπάρχουν λιγότερες ιδιότητες προς μοντελοποίηση. Παρατηρώντας τον ορισμό της γνώμης που δώσαμε παραπάνω (2.1.1), σε αυτού του είδους την ανάλυση, ο φορέας της γνώμης είναι ένας (ο συγγραφέας), το αντικείμενο της γνώμης είναι ένα (η επίμαχη ταινία) και δεν διακρίνουμε τις διάφορες πτυχές του αντικειμένου (της ταινίας).

2.2.2 Επίπεδο Πρότασης (Sentence-Level)

Σε αυτό το επίπεδο ανάλυσης ο στόχος είναι αντίστοιχος με αυτόν της ανάλυσης σε επίπεδο εγγράφου, μόνο που εδώ εξετάζεται το συναίσθημα που εκφράζεται σε

μία πρόταση. Και πάλι κάνουμε μία απλοϊκή παραδοχή, ότι δηλαδή κάθε πρόταση εκφράζει μία (το πολύ) γνώμη, για μία οντότητα. Όμως, αντίθετα με πριν έχουμε ένα βαθύτερο επίπεδο ανάλυσης, το οποίο προσφέρει την ευκαιρία για αξιοποίηση αρκετά περισσότερων πληροφοριών.

Συνεχίζοντας το προηγούμενο παράδειγμα της κριτικής μία κινηματογραφικής ταινίας, σε αυτή την περίπτωση μπορούμε να διακρίνουμε για κάθε πρόταση ξεχωριστά ποιο είναι το συναίσθημα που εκφράζεται. Επίσης, αν συνδυαστεί η πληροφορία αυτή, με την αναγνώριση της οντότητας (Topic/Target-based) για την οποία εκφράζεται το συναίσθημα, τότε αξιοποιείται ένα πολύ μεγάλο ποσοστό της διαθέσιμης πληροφορίας.

Αυτό το επίπεδο ανάλυσης, σχετίζεται αρκετά με το πρόβλημα της *αναγνώρισης υποκειμενικότητας* (subjectivity detection). Στόχος αυτής της εργασίας είναι ελεγχθεί αν μία πρόταση είναι αντικειμενική, δηλαδή αν καταγράφει απλά ένα γεγονός ("η θάλασσα είναι μπλε.") ή αν περιέχει μία γνώμη ("Το νέο iPhone είναι όμορφο."). Ο λόγος που συνδέεται η *αναγνώριση υποκειμενικότητας* με τη ΑΣ είναι ότι, μία πρόταση με μη ουδέτερο προσανατολισμό είναι εξορισμού υποκειμενική. Έτσι, σαν ένα πρώτο βήμα όταν εκτελείται ΑΣ εγγράφων σε επίπεδο πρότασης, αναγνωρίζονται οι υποκειμενικές προτάσεις και στη συνέχεια εφαρμόζεται ΑΣ μόνο σε αυτές.

2.2.3 Επίπεδο Χαρακτηριστικών (Aspect-Based)

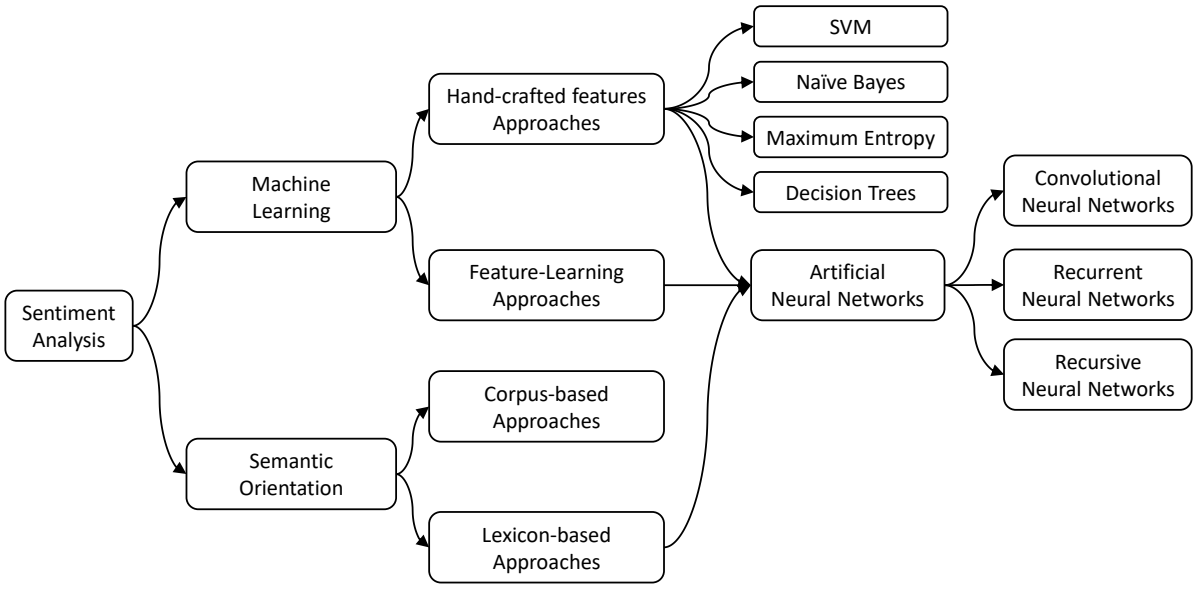
Αυτό το επίπεδο ανάλυσης είναι το πιο ενδιαφέρον και το πιο δύσκολο. Στα δύο προηγούμενα επίπεδα ανάλυσης, κάνουμε ορισμένες παραδοχές, όπως το πόσες γνώμες περιέχονται σε ένα κείμενο, με στόχο να απλοποιήσουμε το πρόβλημα. Όμως, στην ανάλυση σε επίπεδο χαρακτηριστικών σε ένα κείμενο, στόχος είναι η αναγνώριση του συναισθήματος για όλες τις οντότητες ή και τις πτυχές τους.

Το πρώτο βήμα είναι η αναγνώριση των οντοτήτων. Στη συνέχεια, αναλόγως του βαθμού της ανάλυσης, γίνεται εξαγωγή των χαρακτηριστικών (aspects) για κάθε οντότητα. Αυτό είναι ακόμα πιο δύσκολο πρόβλημα, καθώς αρκετές φορές μία πτυχή μπορεί να αναφέρεται έμμεσα στο κείμενο. Επιπλέον, ένα άλλο πρόβλημα είναι ότι ορισμένες πτυχές αναφέρονται με παραπλήσιο τρόπο και κατά συνέπεια είναι δύσκολο να αναγνωριστεί ότι αποτελούν την ίδια πτυχή.

2.3 Προσεγγίσεις

Σε αυτό το σημείο θα γίνει μία επισκόπηση των βασικών προσεγγίσεων σε εφαρμογές ΑΣ. Οι προσεγγίσεις θα διαχωριστούν βάση των μεθοδολογιών που χρησιμοποιούν. Ένα βασικό κριτήριο για την ομαδοποίηση των προσεγγίσεων, αφορά τον τρόπο με τον οποίο διαχειρίζονται τα χαρακτηριστικά (features) για την μοντελοποίηση της φυσικής γλώσσας. Η Εικόνα 2.1 παρουσιάζει μία ένα διάγραμμα στο οποίο αποτυπώνεται η ιεραρχία των διαφόρων τεχνικών.

Οι τεχνικές μπορούν να διαχωριστούν σε δύο βασικές κατηγορίες, αυτές που κάνουν χρήση Μηχανικής Μάθησης (με επιτήρηση και χωρίς επιτήρηση) και σε αυτές που χρησιμοποιούν τεχνικές Σημασιολογικού Προσανατολισμού (χωρίς επιτήρηση).



Σχήμα 2.1: Υψηλού επιπέδου σύγκριση των διαφόρων προσεγγίσεων.

2.3.1 Μηχανική Μάθηση

Οι τεχνικές Μηχανικής Μάθησης (Machine Learning) έχουν ως στόχο, την δημιουργία αλγορίθμων που θα επιτρέψουν σε ένα μηχάνημα *να μάθει* να εκτελεί μία εργασία, μέσα από δεδομένα. Δηλαδή, *χωρίς* να έχει ρητά προγραμματιστεί εκ των προτέρων. Αυτό γίνεται με την δημιουργία στατιστικών μοντέλων, όπου βασίζονται σε ένα σύνολο από χαρακτηριστικά (features) τα οποία εξάγονται από τα δεδομένα. Σημαντική προϋπόθεση είναι η χρήση χαρακτηριστικών τα οποία θα περιγράφουν όσο το δυνατό καλύτερα τα δεδομένα.

Στο πλαίσιο της μηχανικής μάθησης, ένα *χαρακτηριστικό* (feature) είναι μία μετρήσιμη ιδιότητα μίας οντότητας. Σε ένα σύνολο δεδομένων περιέχονται αρκετά παραδείγματα, από τα οποία "μαθαίνει" ένας αλγόριθμος μηχανικής μάθησης. Στόχος είναι η επιλογή ή η εξαγωγή ενός συνόλου από αντιπροσωπευτικά χαρακτηριστικά, τα οποία θα χρησιμοποιήσει ο αλγόριθμος για να μάθει να αναγνωρίζει τι περιγράφει κάθε ένα από τα παραδείγματα. Έστω ότι στόχος είναι ο σχεδιασμός ενός αλγορίθμου μηχανικής μάθησης, ο οποίος θα υπολογίζει την πιθανότητα εμφάνισης καρδιακών παθήσεων για έναν άνθρωπο. Για την "εκπαίδευση" του αλγορίθμου παρέχεται ένα σύνολο με το ιατρικό ιστορικό διαφόρων ασθενών. Για κάθε ασθενή είναι γνωστό το αποτέλεσμα, δηλαδή αν εμφάνισε μία καρδιακή πάθηση η όχι. Για την δημιουργία του στατιστικού μοντέλου, μερικά χαρακτηριστικά θα μπορούσαν να είναι:

(Φύλο, Ηλικία, Βάρος, Πίεση αίματος, Σφυγμοί)

Με αυτό τον τρόπο για κάθε ένα παράδειγμα σχηματίζεται ένα διάνυσμα χαρακτηριστικών (feature vector). Τα χαρακτηριστικά καθορίζουν τον τρόπο με τον οποίο μοντελοποιείται κάθε ένα από τα παραδείγματα του συνόλου δεδομένων. Όπως είναι προφανές, όσο πιο αντιπροσωπευτικά είναι τα χαρακτηριστικά, τόσο καλύτερο θα είναι και το αντίστοιχο μοντέλο.

Οι τεχνικές μηχανικής μάθησης διακρίνονται σε δύο βασικές υποκατηγορίες, στις τεχνικές *επιτηρούμενης μάθησης* (supervised learning) και στις τεχνικές *μη επιτηρούμενης μάθησης* (unsupervised learning). Στις τεχνικές επιτηρούμενης μάθησης, για κάθε παράδειγμα στα δεδομένα είναι γνωστό το αποτέλεσμα ή η κλάση του (π.χ. ποιος αριθμός απεικονίζεται στην εικόνα, αν είναι spam το email ή όχι κλπ.) και αξιοποιώντας αυτή την πληροφορία, σε συνδυασμό με τα χαρακτηριστικά κάθε παραδείγματος, χτίζεται το αντίστοιχο στατιστικό μοντέλο. Αντίθετα, στις τεχνικές μη επιτηρούμενης μάθησης, στόχος είναι η ανακάλυψη δομών και μοτίβων στα δεδομένα (π.χ. συσταδοποίηση), χωρίς είναι γνωστό πόσα είναι ή αν υπάρχουν. Στο πλαίσιο της ΑΣ έχουν εφαρμοστεί και οι δύο τεχνικές, με κάθε μία να προσανατολίζεται σε διαφορετικά υποπροβλήματα.

Για παράδειγμα, σε προβλήματα ταξινόμησης κειμένων ως προς τον συναισθηματικό προσανατολισμό τους (θετικός, αρνητικός ή ουδέτερος), εφαρμόζονται συνήθως τεχνικές επιτηρούμενης μάθησης. Στα προβλήματα αυτά χρησιμοποιούνται σύνολα από έγγραφα ή προτάσεις, για τις οποίες είναι γνωστός ο προσανατολισμός τους και το μοντέλο μηχανικής μάθησης μαθαίνει από τα δεδομένα, να αναγνωρίζει το προσανατολισμό νέων κειμένων. Από την άλλη μεριά, σε προβλήματα όπως την παραγωγή περιλήψεων για τις γνώμες που εκφράζονται σε ένα κείμενο, συχνά χρησιμοποιούνται τεχνικές μη επιτηρούμενης μάθησης (M. Hu κ.ά. 2004; Pang κ.ά. 2004; Zhuang κ.ά. 2006) (ως προς τις περιλήψεις). Ο λόγος είναι ότι σε τέτοιου είδους προβλήματα είναι δύσκολο να παραχθούν σημειωμένα δεδομένα.

Οι πιο δημοφιλής προσεγγίσεις, είναι αυτές με επιτηρούμενη μηχανική μάθηση και σε αυτές θα δοθεί έμφαση στην παρούσα εργασία. Ένα σύστημα μηχανικής μάθησης αποτελείται από δύο τμήματα:

1. **Χαρακτηριστικά (features)**. Σε αυτό το βήμα δημιουργούνται τα χαρακτηριστικά για τη μοντελοποίηση των παραδειγμάτων. Στη ΑΣ μερικά συνηθισμένα χαρακτηριστικά είναι από απλά σύνολα λέξεων (bag-of-words), μέχρι πιο εξεζητημένα όπως σημασιολογικά, συντακτικά και γνωστικά χαρακτηριστικά. Αυτό το βήμα αποτελείται συνήθως από διάφορα υποβήματα όπως:

   - Εξαγωγή των αρχικών χαρακτηριστικών
   - Επιλογή των σημαντικότερων χαρακτηριστικών
   - Παραγωγή νέων από τα αρχικά χαρακτηριστικά
   - Απόδοση βαρών στα χαρακτηριστικά βάση της σημαντικότητας τους (σύμφωνα με κάποιο κριτήριο)

2. **Αλγόριθμος Μηχανικής Μάθησης**. Αυτό το βήμα αφορά την εφαρμογή ενός (ή συνδυασμού πολλών) αλγορίθμου μηχανικής μάθησης, ο οποίος χτίζει το ένα στατιστικό μοντέλο, χρησιμοποιώντας τα χαρακτηριστικά του προηγούμενου βήματος.

Τα παραπάνω βήματα θα αναλυθούν στο επόμενο κεφάλαιο.

2.3.1.1 Πλεονεκτήματα - Μειονεκτήματα

Οι τεχνικές μηχανικής μάθησης έχουν εφαρμοστεί με επιτυχία στα περισσότερα προβλήματα επεξεργασίας φυσικής γλώσσας, όπως Μηχανική Μετάφραση (Machine Translation), Συστήματα Διαλόγου (Dialog Systems), Ερωταπαντήσεων

(Question-Answering) και φυσικά Ταξινόμηση Κειμένων. Η ΑΣ προσεγγίζεται κυρίως ως πρόβλημα ταξινόμησης κειμένων, καθώς στις περισσότερες εφαρμογές στόχος είναι η κατηγοριοποίηση κειμένων ανάλογα με τον συναισθηματικό-σημασιολογικό τους προσανατολισμό.

Οι τεχνικές αυτές επιτρέπουν τη δημιουργία αρκετά εξεζητημένων μοντέλων, αρκετά μεγάλης ακρίβειας. Έχουν την ικανότητα να αξιοποιούν σύνθετα σύνολα χαρακτηριστικών και να ανακαλύπτουν μοτίβα και δομές κρυμμένες σε αυτά. Στηρίζονται σε ισχυρά μαθηματικά θεμέλια.

2.3.2 Σημασιολογικός Προσανατολισμός (Semantic Orientation)

Οι τεχνικές αυτές βασίζονται στην παραδοχή, ότι ο σημασιολογικός προσανατολισμός (θετικός, αρνητικός ή ουδέτερος) ενός κειμένου, καθορίζεται από τον συμψηφισμό του προσανατολισμού των επιμέρους όρων (λέξεων ή φράσεων) που το αποτελούν.

Γενικά, αφορούν την διαδικασία υπολογισμού του σημασιολογικού προσανατολισμού κάθε όρου ενός κειμένου. Αφού έχει υπολογιστεί ο προσανατολισμός των όρων, τότε γίνεται συμψηφισμός των τιμών τους και υπολογίζεται ο συνολικός προσανατολισμός του κειμένου (έγγραφο, πρόταση ή φράση). Ανάλογα με τον τρόπο που γίνεται αυτός ο υπολογισμός, διακρίνονται σε δύο υποκατηγορίες: σημασιολογικός προσανατολισμός (1) *βασισμένος σε κείμενα* (corpus-based) και (2) *βασισμένος σε λεξικά* (lexicon/dictionary-based).

2.3.2.1 Βασισμένος σε κείμενα (corpus-based)

Η βασισμένη-σε-κείμενα προσέγγιση στηρίζεται στην παραδοχή, ότι όταν μία λέξη *συν-εμφανίζεται*[1] πιο συχνά με λέξεις με θετικό προσανατολισμό (π.χ. «excellent»), η τιμή του προσανατολισμού της τείνει να είναι θετική και αντίστοιχα όταν συν-εμφανίζεται πιο συχνά με λέξεις με αρνητικό προσανατολισμό (π.χ. «poor»), η τιμή του προσανατολισμού της τείνει να είναι αρνητική. Αρχικά, επιλέγεται ένα σύνολο από όρους με ήδη γνωστό προσανατολισμό, οι οποίοι ανήκουν σε μία από δύο κλάσεις (“θετικό” και “αρνητικό” προσανατολισμό). Στη συνέχεια, συγκεντρώνονται στατιστικά για την συν-εμφάνιση κάθε λέξης με λέξεις των δύο κλάσεων. Έτσι, αν μία λέξη εμφανίζεται πιο συχνά κοντά σε “αρνητικές” λέξεις θεωρούμε ότι αυτή ή λέξη είναι “αρνητική” και το αντίστροφο. Μία βασική δυσκολία είναι, ότι απαιτείται μεγάλη συλλογή κειμένων για να τον αξιόπιστο προσδιορισμό του προσανατολισμού κάθε λέξης. Αυτό σημαίνει ότι θα πρέπει να συγκεντρωθούν αρκετά κείμενα, ώστε το δείγμα να είναι αντιπροσωπευτικό του συνόλου (σε μία προσέγγιση) και τα στατιστικά να είναι αξιόπιστα.

Ένα άλλο πρόβλημα με αυτή την προσέγγιση είναι η παραδοχή ότι ο προσανατολισμός μίας λέξης είναι μονοδιάστατος. Αυτό σημαίνει ότι ο προσανατολισμός που ορίζεται για μία λέξη είναι σταθερός ανεξαρτήτως πλαισίου στο οποίο αναφέρεται η λέξη. Αυτό όμως δεν ανταποκρίνεται στην πραγματικότητα, καθώς ο προσανατολισμός των λέξεων αλλάζει ανάλογα με το εννοιολογικό πλαίσιο (context) και το πεδίο (domain) μέσα στο οποίο αναφέρεται. Για παράδειγμα, η λέξη “cheap”

[1] Ο όρος *συν-εμφάνιση* (co-occurrence) αναφέρεται στην εμφάνιση ορισμένων όρων (λέξεων ή φράσεων) σε μικρή απόσταση μεταξύ τους, σε ένα κείμενο. Το μέγεθος της απόστασης είναι μία παράμετρος. Έτσι για παράδειγμα, μπορεί να θεωρούμε ότι δύο λέξεις συν-εμφανίζονται, αν απέχουν το πολύ μέχρι 10 λέξεις μεταξύ τους ή αν βρίσκονται στην ίδια πρόταση.

μπορεί να είναι θετική όταν αναφέρεται σε ένα κινητό τηλέφωνο, αλλά αρνητική, όταν αναφέρεται σε μία ταινία.

2.3.2.2 Βασισμένος σε λεξικά (lexicon-based)

Σε αυτού του είδους τις τεχνικές χρησιμοποιούνται λεξικά, στα οποία για κάθε λέξη δίνεται και η αντίστοιχη τιμή του σημασιολογικού προσανατολισμού της. Δημοφιλή τέτοια λεξικά είναι τα WordNet (Miller 1995), SenticNet (Cambria κ.ά. 2014, 2016), SentiWordNet (Esuli κ.ά. 2007; Baccianella κ.ά. 2010) αλλά και πολλά άλλα. Συνήθως τα λεξικά αυτά για κάθε μία λέξη σημειώνουν το αν έχει θετικό ή αρνητικό προσανατολισμό, καθώς και ορισμένες φορές ακόμη και το πόσο έντονος είναι αυτός ο προσανατολισμός. Το γεγονός ότι χρησιμοποιείται μία πηγή η οποία παρέχει έτοιμες τις τιμές για κάθε λέξη, σημαίνει ότι δεν υπάρχει η ανάγκη ενός μεγάλου συνόλου κειμένων.

Η δυσκολία σε αυτή την προσέγγιση είναι η δημιουργία ενός αρκετά μεγάλου λεξικού το οποίο να καλύπτει αρκετά πεδία (κριτικές ταινιών, ηλεκτρονικών προϊόντων, βιβλίων κλπ). Συνήθως αυτό επιτυγχάνεται μέσω του συνδυασμού επιμέρους λεξικών. Ακόμη, κάτι το οποίο συχνά παραβλέπεται είναι η σημαντικότητα των διαφόρων λέξεων-χαρακτηριστικών. Αυτό μπορεί να αντιμετωπιστεί με την εφαρμογή βαρών στα χαρακτηριστικά, μέσω κάποιου μηχανισμού ο οποίος αποτιμά την σημαντικότητά τους.

2.3.2.3 Πλεονεκτήματα - Μειονεκτήματα

Το πλεονέκτημα αυτών των τεχνικών είναι ότι δεν απαιτούν την ύπαρξη σημειωμένων δεδομένων, όπως στις τεχνικές μη επιτηρούμενης μάθησης. Επίσης, είναι αρκετά εύκολο να υλοποιηθούν μοντέλα, τα οποία να αξιοποιούν αυτές τις τεχνικές. Αυτός είναι ένας από τους βασικούς λόγους για την απήχησή τους, ειδικά στα πρώτα χρόνια της έρευνα στην ΑΣ.

Από την άλλη μεριά, οι μεθοδολογίες σημασιολογικού προσανατολισμού έχει παρατηρηθεί ότι μπορεί να μεροληπτήσουν υπέρ του "θετικού" προσανατολισμού. Ο λόγος είναι ότι οι άνθρωποι τείνουν να χρησιμοποιούν πιο συχνά "θετικές" λέξεις όταν εκφράζονται. Αυτό σημαίνει ότι σε ένα κείμενο, οι "θετικές" λέξεις συνήθως είναι περισσότερες σε πλήθος από τις "αρνητικές" και κατά συνέπεια ο προσανατολισμός ενός κειμένου, δεν είναι αντιστοιχεί στον πραγματικό.

Κεφάλαιο 3

Εξαγωγή Χαρακτηριστικών από Κείμενα

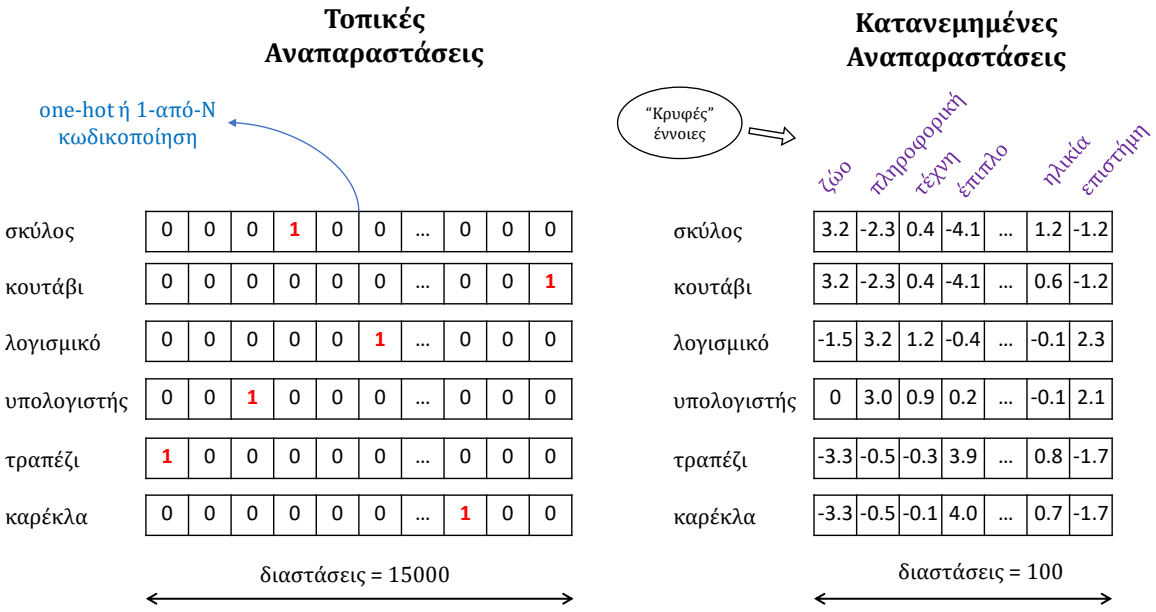
Για την εφαρμογή ενός αλγορίθμου μηχανικής μάθησης, πρέπει αρχικά να κωδικοποιήσουμε τα δεδομένα (παρατηρήσεις) σε μία αριθμητική μορφή (διανύσματα χαρακτηριστικών), η οποία στη συνέχεια θα χρησιμοποιηθεί σαν είσοδος στον αλγόριθμο. Για τον λόγο αυτό θα πρέπει να κατασκευάσουμε μία αναπαράσταση των παρατηρήσεων, η οποία μπορεί είτε να χρησιμοποιηθεί αυτούσια ως το τελικό διάνυσμα χαρακτηριστικών, είτε για την εξαγωγή / παραγωγή νέων σύνθετων χαρακτηριστικών. Ο τρόπος με τον οποίο αποφασίζουμε να μοντελοποιήσουμε κάθε μία παρατήρηση, αντιστοιχεί στα χαρακτηριστικά που θα σχηματίσουμε. Δηλαδή "σύνολο χαρακτηριστικών" = "αναπαράσταση παρατήρησης".

Η ομιλία (και ο γραπτός λόγος κατά συνέπεια) είναι ίσως το πιο ξεχωριστό χαρακτηριστικό του ανθρώπου, σε σχέση με τα υπόλοιπα ζώα και συνδέεται άμεσα με την ανθρώπινη συνείδηση. Στην γνωσιακή επιστήμη (cognitive science) υπάρχουν δύο βασικές θεωρίες σε ότι αφορά τον τρόπο με τον οποίο χαρτογραφούνται οι έννοιες στον εγκέφαλο (Roy 2012).

- τοπική-εντοπιστική αναπαράσταση (local-localist representation): κάθε νευρώνας του εγκεφάλου αντιστοιχεί σε μία και μόνο έννοια.
- κατανεμημένη αναπαράσταση (distributed representations): μία έννοια αναπαρίσταται από ένα μοτίβο δραστηριότητας πολλών νευρώνων. Δηλαδή μία έννοια χρειάζεται πολλούς νευρώνες για να αναπαρασταθεί στον εγκέφαλο και ένας νευρώνας συμμετέχει στην αναπαράσταση πολλών εννοιών.

Οι παραπάνω θεωρίες αποτελούν ένα χρήσιμο νοητικό μοντέλο για τις διάφορες τεχνικές αναπαράστασης της φυσικής γλώσσας, οι οποίες θα παρουσιαστούν στη συνέχεια. Στις μη-κατανεμημένες αναπαραστάσεις, μία παρατήρηση αναπαρίσταται ως το σύνολο πολλών μονοσήμαντων χαρακτηριστικών, ενώ στις κατανεμημένες αναπαραστάσεις μία παρατήρηση αναπαρίσταται ως σύνολο συσχετίσεων ανάμεσα σε χαρακτηριστικά.

Το πιο εύκολο παράδειγμα για να γίνει κατανοητή η διαφορά είναι η αναπαράσταση λέξεων. Έστω ότι έχουμε N διαφορετικές λέξεις τις οποίες θέλουμε να αναπαραστήσουμε μαθητικά. Σε μία μη-κατανεμημένη αναπαράσταση κάθε λέξη χαρτογραφείται σε ένα διάνυσμα N διαστάσεων, και κάθε στοιχείο του διανύσματος αντιστοιχεί *σε μία* από τις λέξεις. Αντίθετα στις κατανεμημένες αναπαραστάσεις, κάθε λέξη αναπαρίσταται ως μία *κατανομή βαρών* σε ορισμένες λανθάνουσες



Σχήμα 3.1: Στην τοπική αναπαράσταση (one-hot) οι λέξεις "σκύλος" και "κουτάβι" χαρτογραφούνται σε διαφορετικά διανύσματα. Στην κατανεμημένη αναπαράσταση όμως τα διανύσματά τους είναι σχεδόν ίσα, δηλαδή βρίσκονται πολύ κοντά στον αντίστοιχο διανυσματικό χώρο. Το βασικό πρόβλημα με τις μη κατανεμημένες αναπαραστάσεις είναι ότι δεν μπορούμε να κάνουμε καμία ουσιαστική σύγκριση μεταξύ των λέξεων παρά μόνο έλεγχο ισότητας.

(latent) μεταβλητές, οι οποίες αντιστοιχούν σε μία αφηρημένη έννοια. Με αυτό τον τρόπο, η έννοια την οποία αντιπροσωπεύει η λέξη είναι κατανεμημένη *σε όλα* τα στοιχεία του διανύσματος, και ομοίως *κάθε* στοιχείο του διανύσματος (λανθάνουσα μεταβλητή) συμμετέχει στον ορισμό της λέξης. Ο τρόπος παραγωγής κάθε μίας αναπαράστασης και οι σχετικές ερμηνείες θα παρουσιαστούν στις ακόλουθες ενότητες του κεφαλαίου.

Στην Εικόνα 3.1, φαίνεται ένα παράδειγμα με τις διαφορές στην αναπαράσταση των λέξεων. Συνοψίζοντας, κάθε λέξη αναπαριστάται μαθηματικά ως ένα διάνυσμα και ισχύει:

- *Τοπικές* αναπαραστάσεις: κάθε διάσταση του διανύσματος αντιστοιχεί σε *μία και μόνο* λέξη.

- *Κατανεμημένες* αναπαραστάσεις: κάθε διάσταση του διανύσματος αντιστοιχεί σε μία *έννοια*. Μία λέξη προσδιορίζεται ως μία *κατανομή* σε αυτές τις έννοιες.

3.1 Μονοδιάστατες Αναπαραστάσεις (One-Hot / Sparse)

Σε αυτού του είδους τις αναπαραστάσεις, αναπαριστούμε ένα κείμενο σαν το σύνολο των μερών του. Αρχικά, *προκαθορίζουμε* ένα σύνολο χαρακτηριστικών (λέξεις, φράσεις, μέρη του λόγου κλπ.) τα οποία μπορούν να περιγράψουν μία παρα-

τήρηση. Σε μία απλή περίπτωση, επιλέγονται οι N (συνήθως είναι σε τάξη μεγέθους των δεκάδων χιλιάδων) πιο συχνές λέξεις από ένα σύνολο κειμένων.

Στη συνέχεια, κατά την διαδικασία εξαγωγής του διανύσματος χαρακτηριστικών από μία παρατήρηση, ελέγχουμε τι τιμές παίρνουν τα προεπιλεγμένα χαρακτηριστικά στην τρέχουσα παρατήρηση, όπως ποιες από τις προκαθορισμένες λέξεις εμφανίζονται σε ένα κείμενο και πόσες φορές. Ως επακόλουθο, οι αναπαραστάσεις αυτές, παράγουν αραιά διανύσματα χαρακτηριστικών (sparse feature vectors), διότι κάθε παρατήρηση έχει τιμές διάφορες του μηδέν, σε λίγες μόνο διαστάσεις του διανύσματος, καθώς ένα κείμενο περιέχει ένα πολύ μικρό ποσοστό του συνόλου των προκαθορισμένων λέξεων.

3.1.1 Σύνολα Λέξεων (Bag-of-words)

Είναι ο πιο απλός τρόπος αναπαράστασης ενός κειμένου. Το κείμενο αναπαρίσταται ως το σύνολο των λέξεων που το αποτελούν. Αρχικά συγκεντρώνονται όλες οι λέξεις που εμφανίζονται στο σύνολο των κειμένων προς εκπαίδευση και σχηματίζεται ένα λεξιλόγιο (vocabulary). Οι λέξεις στο λεξιλόγιο αποτελούν τα χαρακτηριστικά (features). Σε κάθε μία λέξη ανατίθεται ένα μοναδικό αναγνωριστικό (id). Έτσι σχηματίζεται ένας πίνακας, όπου το αναγνωριστικό που έχει δοθεί σε μία λέξη, ορίζει την θέση που θα έχει στο διάνυσμα χαρακτηριστικών. Στη διαδικασία εξαγωγής των χαρακτηριστικών από ένα κείμενο, αρχικά μετράμε το πλήθος των εμφανίσεων κάθε λέξης στο κείμενο. Στη συνέχεια σχηματίζουμε το διάνυσμα χαρακτηριστικών του κειμένου, θέτοντας την τιμή σε κάθε διάσταση του διανύσματος ίση με το πλήθος των εμφανίσεων της αντίστοιχης λέξης στο κείμενο. Για παράδειγμα, έστω ότι το σύνολο δεδομένων περιέχει τα εξής κείμενα:

```
1: ”The cat sat on the hat”
2: ”The dog ate the cat and the hat”
Vocabulary: {the, cat, sat, on, hat, dog, ate, and }
```

Έχοντας το λεξιλόγιο, μπορούμε πλέον να εξάγουμε τα διανύσματα χαρακτηριστικών. Έτσι για τα δύο κείμενα του παραδείγματος έχουμε:

```
V= {the, cat, sat, on, hat, dog, ate, and }
1: (2, 1, 1, 1, 1, 0, 0, 0)
2: (3, 1, 0, 0, 1, 1, 1, 1)
```

Οι πιο δημοφιλείς τρόποι για τον ορισμό των τιμών σε κάθε χαρακτηριστικό είναι:

- Δυαδικές τιμές: δίνεται η τιμή 1 αν ο όρος (λέξη) υπάρχει στο κείμενο ή 0 αν δεν υπάρχει.
- Αριθμός εμφανίσεων: η τιμή είναι ίση με το πλήθος των εμφανίσεων του όρου στην παρατήρηση.
- Συχνότητα εμφανίσεων: η τιμή είναι ίση με τη συχνότητα (ποσοστό) των εμφανίσεων του όρου στην παρατήρηση.
- TF-IDF (Salton κ.ά. 1983; Manning κ.ά. 2008): δίνεται η TF-IDF τιμή του όρου. Το TF-IDF είναι ένα μετρικό, το οποίο χρησιμοποιείται συχνά στην Ανάκτηση Πληροφοριών (Information Retrieval). Το βάρος αυτό είναι ένα μέτρο της σημαντικότητας κάθε όρου (λέξης) σε ένα κείμενο, το οποίο αποτελεί

τμήμα μίας συλλογής κειμένων. Με τον τρόπο αυτό δίνεται μεγαλύτερη βαρύτητα σε πιο "σημαντικές" λέξεις.

Σαν επόμενο βήμα μπορεί να εφαρμοστεί κάποια τεχνική Επιλογής Χαρακτηριστικών (Feature Selection), όπως χ^2, Mutual Information - Information Gain (Forman 2003), όπου με αυτό τον τρόπο χρησιμοποιείται ένα υποσύνολο των χαρακτηριστικών. Στόχος αυτών των τεχνικών είναι η επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών.

Το Bag of Words (BoW) αν και είναι σχετικά αφελής προσέγγιση, έχει αποδειχτεί ότι πετυχαίνει αρκετά καλά αποτελέσματα σε πολλά προβλήματα ταξινόμησης κειμένων. Ένα πλεονέκτημα των αναπαραστάσεων αυτών είναι πως, επειδή παράγουν αραιά διανύσματα και μεγάλων διαστάσεων, τις περισσότερες φορές είναι γραμμικώς διαχωρίσιμα και συνεπώς για την ταξινόμησή τους μπορεί να χρησιμοποιηθούν με επιτυχία αποδοτικοί αλγόριθμοι όπως Support Vector Machine (SVM) με γραμμικό πυρήνα (linear kernel).

Μερικά από τα μειονεκτήματα αυτής της τεχνικής αναπαράστασης είναι:

- Χάνεται η πληροφορίας της σειράς εμφάνισης των λέξεων.
- Δεν περιέχεται συντακτική ή σημασιολογική πληροφορία.
- Κείμενα ή προτάσεις με τις ίδιες λέξεις έχουν την ίδια αναπαράσταση. Για παράδειγμα, οι προτάσεις "the story was good, but the acting was bad." και "the story was bad, but the acting was good.", θα έχουν την ίδια αναπαράσταση.

3.1.2 N-Grams

Το N-Gram είναι μία συνεχής ακολουθία N λέξεων από ένα κείμενο. Αποτελεί γενίκευση της BoW αναπαράστασης. Έστω η πρόταση:

```
"dog that barks does not bite"
```

Τότε εξάγονται τα ακόλουθα n-grams:

```
unigrams (n=1): dog, that, barks, does, not, bite
bigrams  (n=2): dog that, that barks, barks does, does not, not
   bite
trigrams (n=3): dog that barks, that barks does, barks does not,
   does not bite
```

Με την χρήση n-grams έχουμε την δυνατότητα να αναγνωρίσουμε εκφράσεις που αποτελούνται από πολλές λέξεις. Όμως αυξάνοντας το μέγεθος τον n-grams αυξάνουμε σημαντικά το λεξιλόγιο, δηλαδή το μέγεθος των διανυσμάτων χαρακτηριστικών (feature vectors), και εισάγουμε θόρυβο στο μοντέλο, διότι οι περισσότερες εκφράσεις (όροι) που εξάγονται δεν περιέχουν αρκετή πληροφορία.

3.1.3 Skip-grams

Είναι μία γενίκευση του N-Gram, με την διαφορά ότι παραλείπονται λέξεις, ανάμεσα στις λέξεις της ακολουθίας. Έστω η πρόταση:

```
the rain in Spain falls mainly on the plain
```

Τα 1-skip-2-grams (ακολουθία 2 λέξεων με την παράληψη μίας λέξης) που εξάγονται είναι τα εξής:

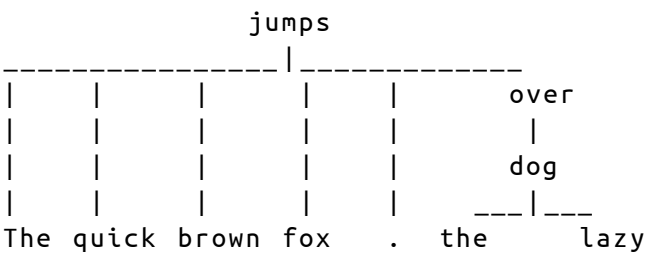
```
["the in", "rain Spain", "in falls", "Spain mainly", "falls on", "
   mainly the", "on plain"]
```

3.1.4 Συντακτικά Χαρακτηριστικά

Ένας άλλος τρόπος αναπαράστασης ενός κειμένου, είναι ως το σύνολο των συντακτικών σχέσεων εξάρτησης, μεταξύ των λέξεων του κειμένου (syntactic dependencies / dependency relations) (De Marneffe κ.ά. 2008). Για την εξαγωγή των σχέσεων θα πρέπει να προηγηθεί διαδικασία συντακτικής ανάλυσης εξαρτήσεων (dependency parsing). Αφού εκτελεστεί η συντακτική ανάλυση, μπορούμε να σχηματίσουμε ένα δέντρο εξαρτήσεων. Έστω η πρόταση:

```
"The quick brown fox jumps over the lazy dog."
```

Ένα δέντρο εξαρτήσεων της πρότασης είναι:



Είναι σημαντικό να σημειωθεί ότι υπάρχουν πολλές διαφορετικές τεχνικές (Choi κ.ά. 2015) για την συντακτική ανάλυση ενός κειμένου, με συνέπεια την παραγωγή διαφορετικών αποτελεσμάτων (δέντρων, εξαρτήσεων κλπ.). Από το δέντρο του παραδείγματος μπορούμε να εξάγουμε τριπλέτες με τις σχέσεις / εξαρτήσεις της μορφής (head, relation, modifier). Για την πρόταση του προηγούμενου παραδείγματος εξάγουμε τις εξής σχέσεις:

```
('jumps', 'det', 'The'),
('jumps', 'amod', 'quick'),
('jumps', 'amod', 'brown'),
('jumps', 'nsubj', 'fox'),
('jumps', 'prep', 'over'),
('jumps', 'punct', '.'),
('over', 'pobj', 'dog'),
('dog', 'det', 'the'),
('dog', 'amod', 'lazy')('jumps', 'det', 'The'),
('jumps', 'amod', 'quick'),
('jumps', 'amod', 'brown'),
('jumps', 'nsubj', 'fox'),
('jumps', 'prep', 'over'),
('jumps', 'punct', '.'),
('over', 'pobj', 'dog'),
('dog', 'det', 'the'),
('dog', 'amod', 'lazy')
```

Οι σχέσεις αυτές περιέχουν πλούσια συντακτική και σημασιολογική πληροφορία, η οποία δεν μπορεί να αποτυπωθεί από το Bag of Words (BoW) μοντέλο. Εξάγονται σχέσεις όπως:

```
'jumps' - 'quick'
```

```
'jumps' - 'fox'
'dog' - 'lazy'
```

3.1.4.1 Χαρακτηριστικά Εξαρτήσεων

Οι σχέσεις εξαρτήσεων μπορούν να χρησιμοποιηθούν με διάφορες παραλλαγές για την αναπαράσταση ενός κειμένου. Επιπλέον σε ορισμένες προσεγγίσεις χρησιμοποιούνται σε συνδυασμό με Οντολογίες ή Βάσεις Γνώσης. Οι συνήθεις τεχνικές είναι οι εξής:

- Σύνολο σχέσεων: ακολουθείται αντίστοιχη διαδικασία με αυτή του BoW όπου αντί για λέξεις χρησιμοποιούνται οι σχέσεις.
- Σύνολο σχέσεων με οπισθοδρόμηση: σε αυτή την τεχνική μία από τις δύο λέξεις αντικαθίσταται από το μέρος του λόγου (Part-of-Speech Tag) της λέξης (Joshi κ.ά. 2009). Η λογική πίσω από αυτή την παραλλαγή είναι, ότι με αυτό τον τρόπο σε προβλήματα όπως στην ΑΣ, παράγονται πιο γενικές σχέσεις. Για παράδειγμα η σχέση: ('dog', 'amod', 'lazy') γίνεται ('NN', 'amod', 'lazy') και συνεπώς η σχέση αυτή θα ταυτιστεί με παρόμοιες σχέσεις όπως ('student', 'amod', 'lazy'), όπου σε στην απλή περίπτωση ο αλγόριθμος Μηχανικής Μάθησης θα τις αντιμετώπιζε ως διαφορετικά χαρακτηριστικά.
- Υποδέντρα του Δέντρου Εξαρτήσεων: Σε αυτή την τεχνική χρησιμοποιούνται υποδέντρα του δέντρου εξαρτήσεων ως όροι (Özgür κ.ά. 2010; Sidorov κ.ά. 2012), όπως οι λέξεις στο BoW μοντέλο.

Τα μειονεκτήματα των χαρακτηριστικών που βασίζονται σε χαρακτηριστικά συντακτικών εξαρτήσεων είναι τα εξής:

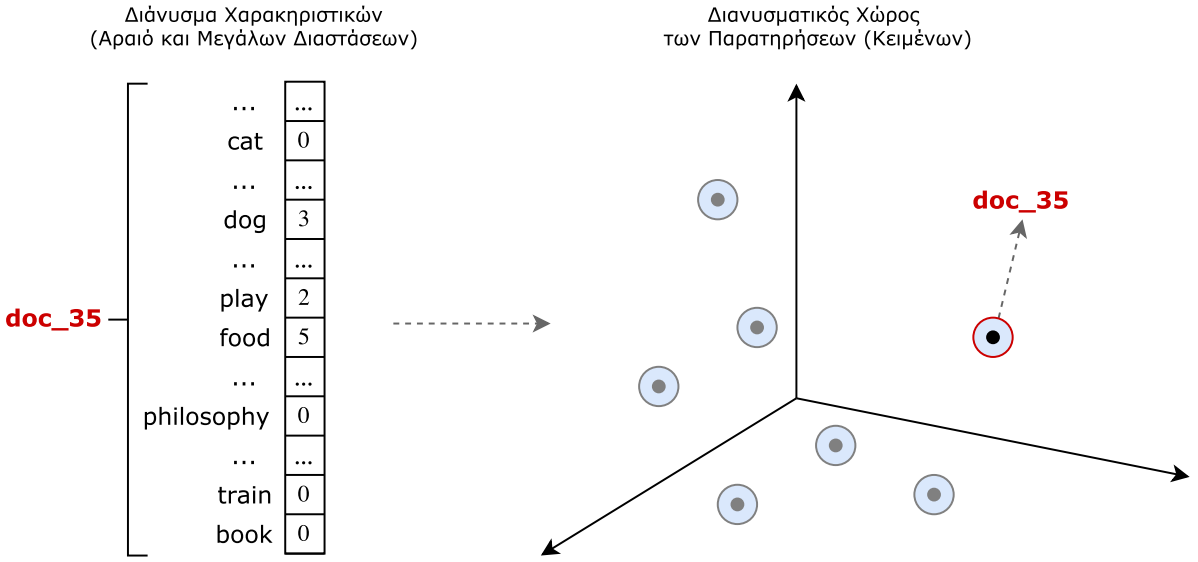
- Η διαδικασία του της συντακτικής ανάλυσης είναι μία αρκετά χρονοβόρα διαδικασία
- Η επιλογή της τεχνικής συντακτικής ανάλυσης είναι πολύ σημαντική, καθώς κάθε τεχνική παράγει σημαντικά διαφορετικές εξαρτήσεις.
- Τα διανύσματα χαρακτηριστικών που παράγονται είναι πολύ μεγάλων διαστάσεων.

3.1.5 Ανακεφαλαίωση

Ανακεφαλαιώνοντας στην Εικόνα 3.2 φαίνεται σχηματικά το πως αναπαρίσταται ένα κείμενο με μία *τοπική αναπαράσταση*. Αφού χαρτογραφηθούν όλες οι παρατηρήσεις (κείμενα) στον χώρο των χαρακτηριστικών, στην συνέχεια δίνονται ως είσοδο σε έναν αλγόριθμο Μηχανικής Μάθησης (SVM, Logistic Regression) για ταξινόμηση.

3.2 Κατανεμημένες Αναπαραστάσεις

Σε αυτού του είδους τις αναπαραστάσεις, οντότητες όπως ολόκληρα κείμενα, τμήματα ενός κειμένου (παράγραφοι, προτάσεις) ή λέξεις, χαρτογραφούνται σε *πυκνά* διανύσματα χαρακτηριστικών. Τα διανύσματα αυτά βρίσκονται στον χώρο R^N,



Σχήμα 3.2: Bag-of-Words αναπαράσταση ενός κειμένου. Η διαστάσεις των χαρακτηριστικών είναι συνήθως αρκετά μεγάλων διαστάσεων (δεκάδες χιλιάδες) και τα διανύσματα είναι αρκετά αραιά, δηλαδή λίγες διαστάσεις έχουν τιμή διάφορη του μηδέν για κάθε παρατήρηση.

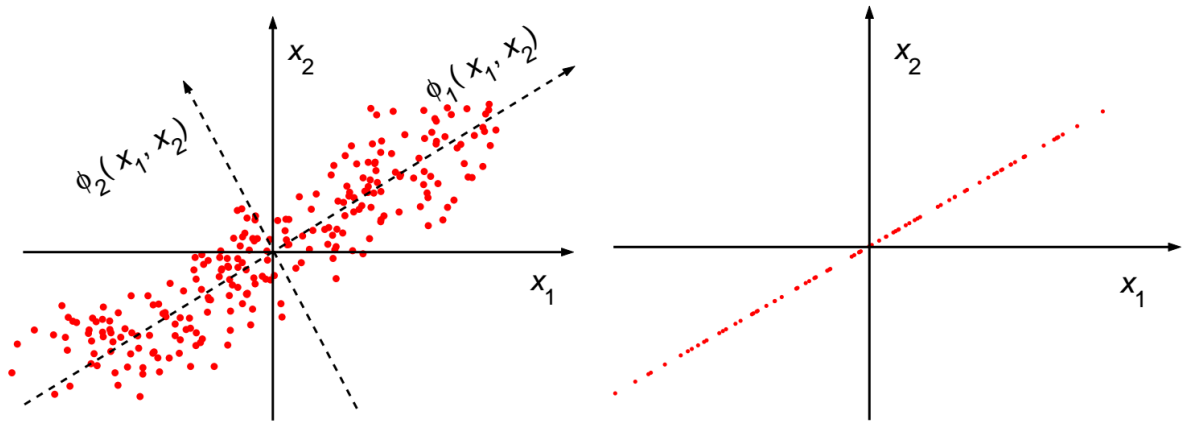
όπου κάθε μία από τις N διαστάσεις δεν αντιστοιχεί σε μία συγκεκριμένη λέξη η φράση, όπως στις τοπικές αναπαραστάσεις, αλλά σε μία λανθάνουσα μεταβλητή (latent variable). Θα μπορούσε κανείς να ερμηνεύσει μία λανθάνουσα μεταβλητή ως μία αφηρημένη έννοια, η οποία είναι συνδυασμός επιμέρους απλών χαρακτηριστικών. Η έννοια αυτή *δεν καθορίζεται από εμάς*, αλλά προκύπτει μέσα από την διαδικασία σχηματισμού των αναπαραστάσεων. Η τιμή που παίρνει η οντότητα (κείμενο, λέξη κλπ.) σε αυτή τη διάσταση αντιστοιχεί στον τρόπο με τον οποίο σχετίζεται με την αντίστοιχη "έννοια". Το παράδειγμα στην Εικόνα 3.1 είναι χαρακτηριστικό. Η παραγωγή κατανεμημένων αναπαραστάσεων γίνεται με διάφορους τρόπους. Ακολουθεί μία περιγραφή των πιο βασικών τεχνικών.

3.2.1 Τεχνικές Μείωσης Διαστάσεων

Στα μαθηματικά *μείωση διαστάσεων* (dimensionality reduction) είναι η διαδικασία κατά την οποία ένα σύνολο παρατηρήσεων στον χώρο R^M προβάλλεται σε ένα χώρο χαμηλότερων διαστάσεων R^N, όπου $M > N$. Στόχος αυτής της διαδικασίας είναι η μείωση των διαστάσεων διατηρώντας την μέγιστη δυνατή πληροφορία. Συνήθεις τεχνικές είναι οι Singular Value Decomposition (SVD) και Principal Component Analysis (PCA) οι οποίες εκτελούν χαρτογράφηση των παρατηρήσεων στον νέο χώρο.

Οι τεχνικές αυτές "συμπιέζουν" διαστάσεις με μικρή διακύμανση, διατηρώντας αυτές με μεγάλη διακύμανση, δηλαδή με μεγάλη πληροφορία. Κατά την διαδικασία της προβολής στον χώρο των χαμηλότερων διαστάσεων υπάρχει απώλεια πληροφορίας και συνεπώς δεν μπορούμε να ξαναγυρίσουμε στον αρχικό χώρο. Όσο λιγότερες οι τελικές διαστάσεις, τόσο περισσότερη είναι η απώλεια πληροφορίας, αλλά και πιο "περιεκτικές" οι νέες αναπαραστάσεις. Η απώλεια πληροφορίας δεν γίνεται γραμμικά καθώς μειώνονται οι διαστάσεις. Δηλαδή αν από τις 1000 γίνει προβολή στις 500 διαστάσεις δεν χάνεται το 50% της πληροφορίας. Συνήθως

μπορούμε να προβάλουμε τις παρατηρήσεις σε πολύ χαμηλότερο χώρο (100-200 διαστάσεις) διατηρώντας 80%-90% της αρχικής πληροφορίας.



(a) PCA basis (b) PCA reduction to 1D

Σχήμα 3.3: Προβολή παρατηρήσεων από τον χώρο R^2 στον χώρο R^1. Πηγή (Shakhnarovich κ.ά. 2011)

Ένα χαρακτηριστικό παράδειγμα χρήσης της τεχνικής αυτής σε ένα πρόβλημα ΕΦΓ είναι η μετατροπή αραιών διανυσμάτων μεγάλων διαστάσεων τα οποία έχουν παραχθεί με BoW τεχνική, σε πυκνά διανύσματα. Μία ερμηνεία του νέου χώρου χαρακτηριστικών είναι, ότι τα νέα χαρακτηριστικά αποτυπώνουν τις αλληλεπιδράσεις μεταξύ των αρχικών χαρακτηριστικών των παρατηρήσεων. Κάθε διάσταση δεν αντιστοιχεί πλέον σε μία *ιδιότητα*, αλλά σε μία *σχέση*.

3.2.2 Τεχνικές δημιουργίας Κατανεμημένων Αναπαραστάσεων για κείμενα

Οι πρώτες τεχνικές παραγωγής κατανεμημένων αναπαραστάσεων, αναπτύχθηκαν στα πλαίσια επίλυση προβλημάτων Ανάκτησης Πληροφορίας και ΕΦΓ, όπως τη σύγκριση κειμένων (ομοιότητα). Ένα χαρακτηριστικό παράδειγμα είναι το *Topic Modeling* στο οποίο θέλουμε να ομαδοποιήσουμε ένα σύνολο κειμένων (corpus) βάση του περιεχομένου τους. Η ιδέα είναι ότι κάθε κείμενο περιέχει ένα ή πολλά θέματα (topics), όπως πολιτική, επιστήμη, αθλητισμός κλπ., και κείμενα με κοινά θέματα θα πρέπει να ανήκουν στην ίδια ομάδα.

3.2.2.1 Τεχνικές γραμμικής μείωσης διαστάσεων

Η πιο απλή τεχνική είναι η Latent Semantic Analysis (LSA), η οποία δημιουργεί μια κατανεμημένη αναπαράσταση για ένα κείμενο. Η διαδικασία είναι εξαιρετικά απλή:

1. Χρησιμοποιούμε το BoW μοντέλο. Δηλαδή:

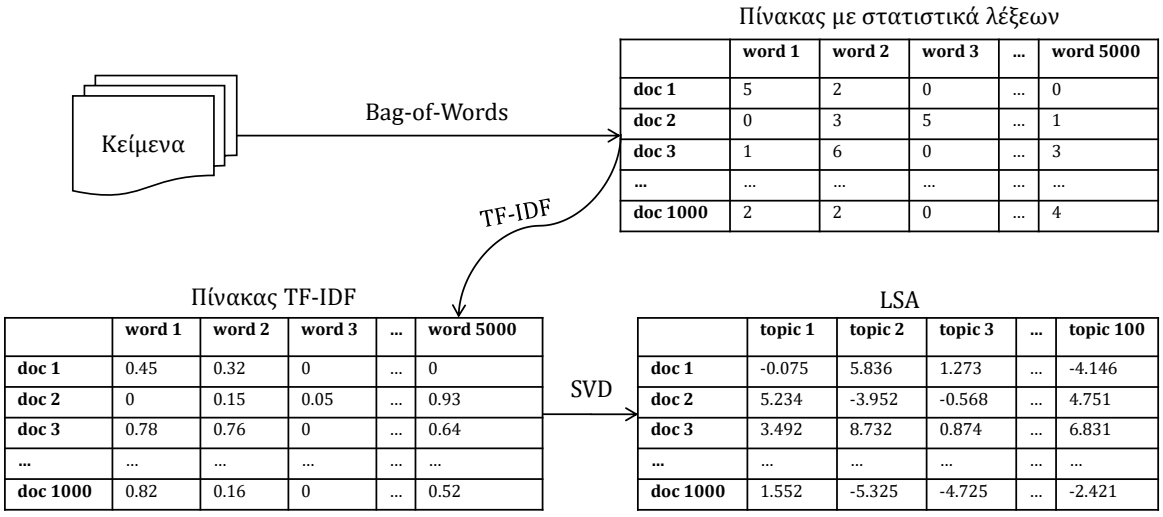
   i) Αρχικά επιλέγουμε το λεξιλόγιο. Συνήθως επιλέγονται οι N πιο συχνές λέξεις που παρατηρούνται σε όλα τα κείμενα, αφού αφαιρεθούν λέξεις όπως

"the", "and", "or" και σημεία στίξης (οι λέξεις αυτές αναφέρονται ως stop-words στη βιβλιογραφία).

ii) Σχηματίζουμε το διάνυσμα χαρακτηριστικών κάθε κειμένου, με το πλήθος εμφάνισης κάθε λέξης του λεξιλογίου στο κείμενο.

2. Κανονικοποιούμε τα διανύσματα βάση της TF-IDF τιμής κάθε λέξης.

3. Εφαρμόζουμε μία τεχνική μείωσης διαστάσεων όπως SVD.



	word 1	word 2	word 3	...	word 5000
doc 1	5	2	0	...	0
doc 2	0	3	5	...	1
doc 3	1	6	0	...	3
...	...	...	...	...	...
doc 1000	2	2	0	...	4

	word 1	word 2	word 3	...	word 5000
doc 1	0.45	0.32	0	...	0
doc 2	0	0.15	0.05	...	0.93
doc 3	0.78	0.76	0	...	0.64
...	...	...	...	...	...
doc 1000	0.82	0.16	0	...	0.52

	topic 1	topic 2	topic 3	...	topic 100
doc 1	-0.075	5.836	1.273	...	-4.146
doc 2	5.234	-3.952	-0.568	...	4.751
doc 3	3.492	8.732	0.874	...	6.831
...	...	...	...	...	...
doc 1000	1.552	-5.325	-4.725	...	-2.421

Σχήμα 3.4: Latent Semantic Analysis (LSA)

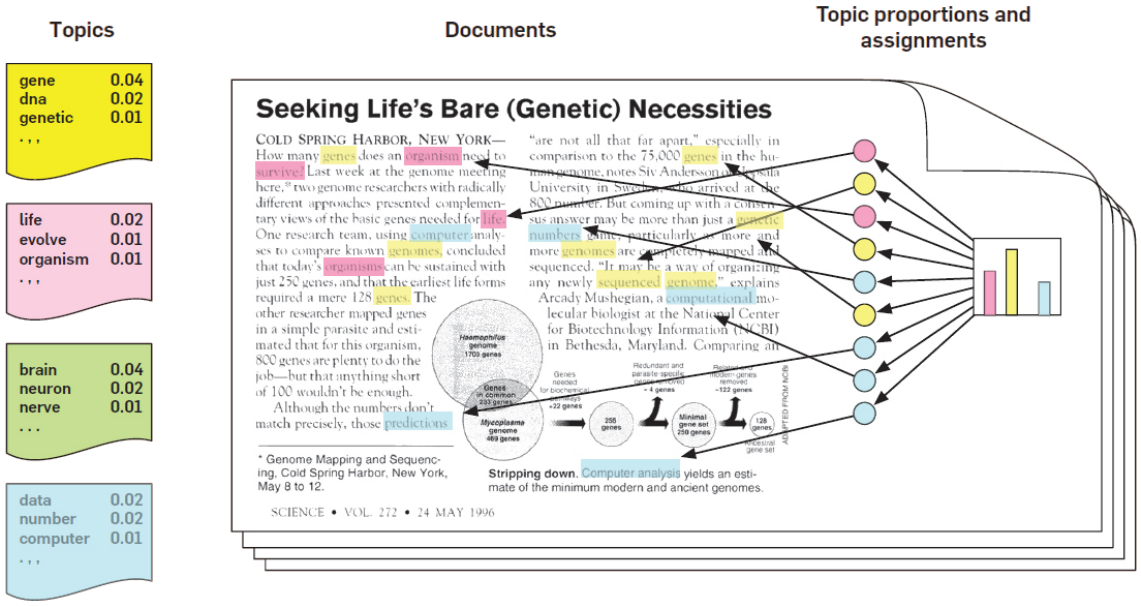
Κάθε διάσταση του τελικού διανύσματος αντιστοιχεί σε ένα λανθάνον (latent) θέμα.

3.2.2.2 Πιθανολογικές Τεχνικές

Μια πολύ δημοφιλής πιθανολογική τεχνική για την παραγωγή κατανεμημένων αναπαραστάσεων είναι η Latent Dirichlet Allocation (LDA)[1] (Blei κ.ά. 2003; Blei 2012) το οποίο πετυχαίνει καλύτερα αποτελέσματα σε σχέση με το LSA. Το LDA είναι ένα στατιστικό μοντέλο, το οποίο υποθέτει ότι υπάρχει μία φανταστική γεννητική διαδικασία η οποία παράγει κείμενα (όπως εννοείται στο statistical inference), ως ένα μείγμα προεπιλεγμένων θεμάτων (topics). Κάνοντας την παραδοχή ότι τα κείμενα υπό εξέταση έχουν παραχθεί από μία τέτοια διαδικασία, στόχος είναι να συμπεράνουμε ποια είναι η κατανομή των θεμάτων για κάθε κείμενο.

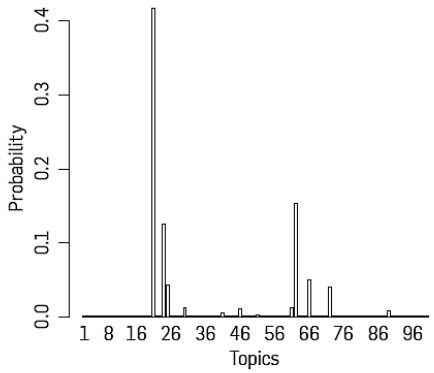
Πιο αναλυτικά: Ένα θέμα – topic, ορίζεται ως μία κατανομή λέξεων πάνω σε ένα προκαθορισμένο λεξιλόγιο. Όλα τα κείμενα εκφράζονται ως ένας συνδυασμός των ίδιων θεμάτων και κάθε κείμενο "περιέχει" αυτά τα θέματα σε διαφορετικές αναλογίες. Υποθέτουμε ότι κάθε λέξη σε κάθε κείμενο, έχει ληφθεί από ένα από τα προκαθορισμένα θέματα (όπως εννοείται στην στατιστική, δηλαδή σαν να κάθε θέμα να διατηρεί ένα "κουβάς" με τις λέξεις του), όπου το θέμα με τη σειρά του έχει επιλεχθεί από την κατανομή των προεπιλεγμένων θεμάτων.

[1]Είναι εντελώς διαφορετική τεχνική με την Linear Discriminant Analysis (LDA) η οποία και αυτή χρησιμοποιείται στη Στατιστική και τη Μηχανική Μάθηση αλλά σε άλλο πλαίσιο. Ένα ακόμη παράδειγμα όπου υπάρχει σύγκρουση όρων.



Σχήμα 3.5: Παράδειγμα κατανομής θεμάτων ανά κείμενο (Blei 2012)

Αυτό το οποίο παρατηρούμε εμείς είναι μόνο τα κείμενα (και οι λέξεις που περιέχουν). Η δομή των θεμάτων και η αντιστοιχία κάθε λέξης με ένα από τα θέματα, είναι κρυφές δομές (latent). Συνεπώς αρχικά επιλέγουμε πόσα θα είναι τα k θέματα από τα οποία θα αποτελείται κάθε κείμενο. Στη συνέχεια εφαρμόζοντας το LDA, συμπεραίνουμε την αντιστοιχία κάθε λέξης με ένα θέμα, και την αναλογία των θεμάτων σε κάθε κείμενο.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Σχήμα 3.6: Αντιστοιχία λέξεων με θέματα (Blei 2012)

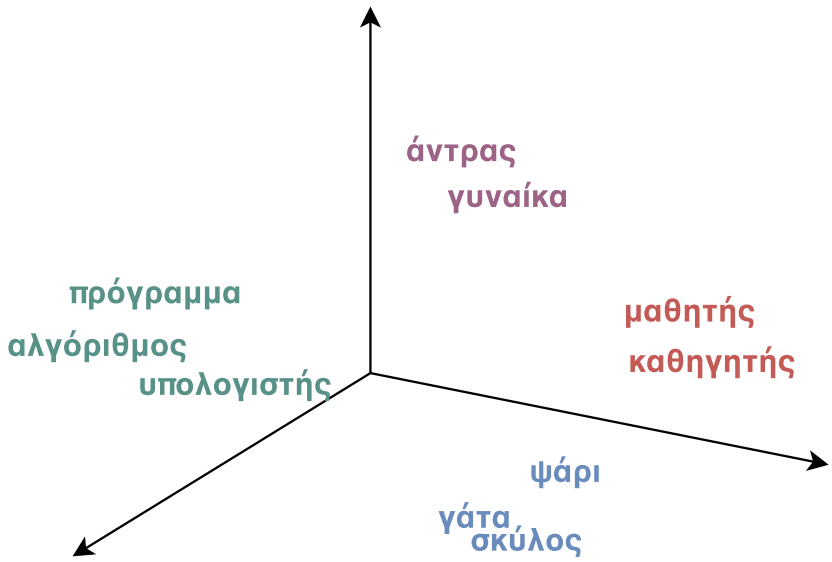
Στην Εικόνα 3.6 φαίνεται ένα παράδειγμα αντιστοίχισης λέξεων με θέματα. Το κάθε θέμα δεν έχει κάποιο όνομα. Μπορούμε μονάχα να συμπεράνουμε ένα όνομα για κάθε θέμα, βάση των λέξεων που περιέχει. Οι λέξεις έχουν ομαδοποιηθεί σε k θέματα. Θα μπορούσε κάποιος να φανταστεί ότι το αποτέλεσμα είναι αντίστοιχο

με αυτού μίας τεχνικής Συσταδοποίησης (K-Means). Το τελικό αποτέλεσμα είναι ότι κάθε κείμενο αναπαρίσταται ως ένα πυκνό διάνυσμα k διαστάσεων.

3.2.3 Διανύσματα Λέξεων (Word Embeddings)

Τα τελευταία χρόνια έχει σημειωθεί σημαντική πρόοδος στις κατανεμημένες αναπαραστάσεις κειμένων, με το επίκεντρο του ενδιαφέροντος να συγκεντρώνεται στις κατανεμημένες αναπαραστάσεις λέξεων. Τα διανύσματα λέξεων parencitecollobert2008, mikolov2013, γνωστά και ως word embeddings, αφορούν αναπαραστάσεις λέξεων, όπου μία λέξη χαρτογραφείται σε ένα συνεχές διάνυσμα πραγματικών αριθμών και λίγων διαστάσεων (συνήθως 100 – 500 διαστάσεων). Ο στόχος είναι λέξεις με κοινό νόημα να χαρτογραφηθούν κοντά στον διανυσματικό χώρο. Αναπαριστώντας μία λέξη ως ένα διάνυσμα και όχι μονοσήμαντα όπως με την one-hot αναπαράσταση, μπορούμε να συγκρατήσουμε τις αλληλεπιδράσεις της με τις υπόλοιπες λέξεις. Κάθε διάσταση στον διανυσματικό χώρο των λέξεων, αντιστοιχεί κατά κάποιον τρόπο σε μία αφηρημένη έννοια και η τιμή που έχει κάθε λέξη σε μία διάσταση, αντικατοπτρίζει το βαθμό στον οποίο σχετίζεται μαζί της.

Η ιδέα της αναπαράστασης μίας λέξης ως διάνυσμα είναι αρκετά παλιά (Osgood 1964). Επιπλέον, υπάρχουν αρκετά παραδείγματα λεξικών (Mohammad κ.ά. 2013; Staiano κ.ά. 2014; Cambria κ.ά. 2016) στα οποία λέξεις συσχετίζονται, με χειροκίνητο τρόπο από ειδικούς (Γλωσσολόγους, Ψυχολόγους κλπ.), με ορισμένα συναισθήματα ή συναισθηματικές καταστάσεις (“φόβος”, “χαρά”, “θυμός”, ...). Όμως, αντικείμενο αυτής της ενότητας είναι η επισκόπηση των τεχνικών για αυτόματη παραγωγή διανυσμάτων λέξεων. Υπάρχουν διάφορες προσεγγίσεις για την παραγωγή διανυσμάτων λέξεων, όπως με ΤΝΔ ή με τεχνικές μείωσης διαστάσεων σαν το LSA που παρουσιάστηκε παραπάνω. Ένα σημαντικό πλεονέκτημα των περισσότερων μεθόδων είναι ότι δεν απαιτούν την χρήση σημειωμένων δεδομένων (μάθηση χωρίς επίβλεψη).



Σχήμα 3.7: Διανύσματα Λέξεων (Word Embeddings). Λέξεις οι οποίες συνεμφανίζονται συχνά, βρίσκονται κοντά στον χώρο.

Οι πρώτες τεχνικές παραγωγής διανυσμάτων λέξεων με χρήση ΤΝΔ, εμφανίζονται στις αρχές τις δεκαετίας του 2000 (Bengio κ.ά. 2003) μέσω της παραγωγής γλωσσικών μοντέλων (language models). Στη συνέχεια βρίσκουν εφαρμογή σε αρκετά προβλήματα επεξεργασίας φυσικής γλώσσας (Collobert κ.ά. 2008) και αποκτούν μεγάλη απήχηση ύστερα από την εργασία (Mikolov κ.ά. 2013a) όπου παρουσιάζεται το word2vec, το οποίο κάνει χρήση ΤΝΔ. Μία δημοφιλής εναλλακτική προσέγγιση για παραγωγή διανυσμάτων λέξεων είναι με την τεχνική GloVe (Pennington κ.ά. 2014) η οποία βασίζεται σε τεχνικές αντίστοιχες του LSA μόνο που το πλαίσιο (context) μίας λέξης δεν είναι ολόκληρο το έγγραφο, αλλά ένα παράθυρο λέξεων (πιο στοχευμένη αναπαράσταση).

Το ερευνητικό ενδιαφέρον για την παραγωγή κατανεμημένων αναπαραστάσεων με χρήση ΤΝΔ είναι πολύ έντονο τα τελευταία χρόνια. Έχουν παρουσιαστεί τεχνικές με παραλλαγές του word2vec (Levy κ.ά. 2014) αλλά και νέες ιδέες όπως (Ji κ.ά. 2015; Joulin κ.ά. 2016; Nickel κ.ά. 2017). Γενικεύοντας την παραπάνω μεθοδολογία είναι εφικτό να παραχθούν κατανεμημένες αναπαραστάσεις όχι μόνο για λέξεις, αλλά και για φράσεις, προτάσεις, ακόμη και για ολόκληρα κείμενα, όπως με τα Skip-Thought Vectors (Kiros κ.ά. 2015) και Doc2vec (Le κ.ά. 2014). Στη συνέχεια ακολουθεί μία επιγραμματική παρουσίαση της προσέγγισης του word2vec, καθώς είναι η πιο δημοφιλής προσέγγιση.

3.2.3.1 Word2Vec

Το Word2Vec χρησιμοποιεί ένα Ρηχό Νευρωνικό Δίκτυο (Shallow Neural Network), δηλαδή με μόνο ένα κρυφό επίπεδο (hidden layer). Η τεχνική αυτή εκτελεί το εξής "κόλπο": αρχικά εκπαιδεύουμε ένα νευρωνικό δίκτυο ώστε να μάθει να εκτελεί μία εργασία, όμως αφού το εκπαιδεύσουμε, δεν το χρησιμοποιούμε για τον σκοπό που το εκπαιδεύσαμε. Αυτό το οποίο θα χρησιμοποιήσουμε είναι τα βάρη που θα έχουν σχηματιστεί στο κρυφό επίπεδο του δικτύου! Τα βάρη που θα έχουν σχηματιστεί για κάθε λέξη θα είναι τα τελικά διανύσματα των λέξεων.

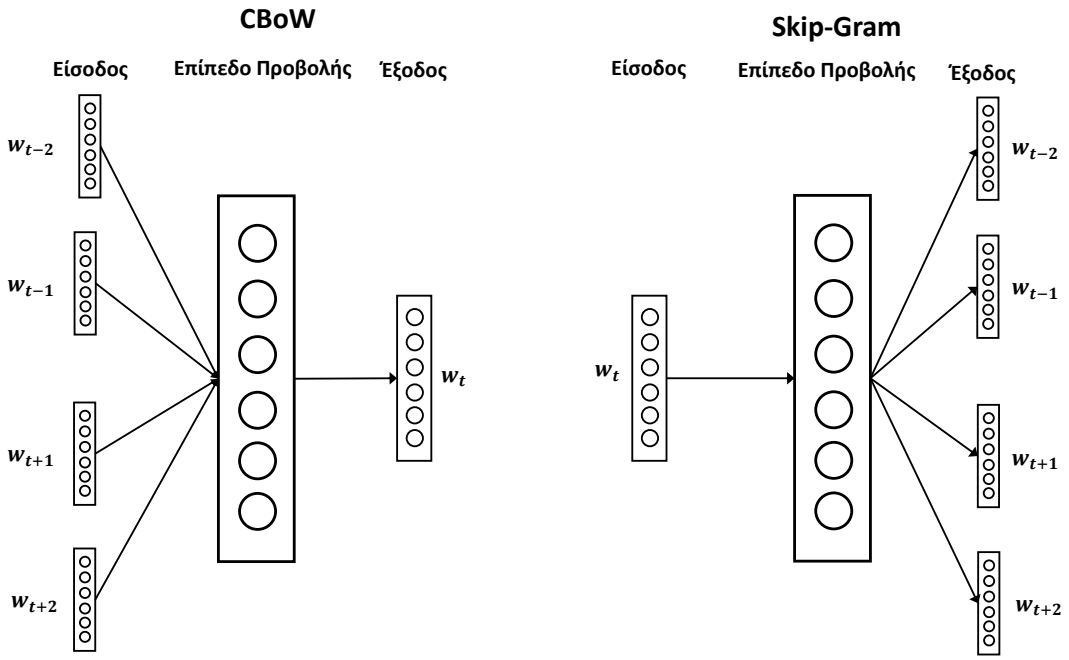
Ο όρος word embeddings (τον οποίο συναντάμε συχνά στη βιβλιογραφία) προέρχεται από το γεγονός ότι το διάνυσμα κάθε λέξης σχηματίζεται από τα βάρη της λέξης στο πρώτο επίπεδο του δικτύου, το οποίο ονομάζεται και επίπεδο Εμφύτευσης (Embedding). Στα μαθηματικά εμφύτευση ονομάζεται η διαδικασία στην οποία μία δομή χαρτογραφείται (ή προβάλλεται-ενσωματώνεται-εμφυτεύεται) από ένα διανυσματικό χώρο σε ένα άλλο. Έτσι στην περίπτωση των λέξεων, το διάνυσμα κάθε λέξης χαρτογραφείται από τον πολυδιάστατο αρχικό χώρο (one-hot διανύσματα), σε ένα χώρο λιγότερων διαστάσεων.

Αυτό το οποίο μαθαίνει στην ουσία το ΤΝΔ στο Word2Vec είναι ένα γλωσσικό μοντέλο (Language Model). Ένα γλωσσικό μοντέλο είναι ένα στατιστικό μοντέλο, το οποίο δεδομένης μίας ακολουθίας λέξεων, παράγει μία κατανομή πιθανοτήτων σχετικά με τις πιθανές ακόλουθες λέξεις (ποια είναι η πιθανότερη λέξη που ακολουθεί?). Για παράδειγμα, έστω ότι έχουμε μία ακολουθία n λέξεων, τότε μπορούμε να αναθέσουμε μία πιθανότητα σε όλη την ακολουθία $P(w1, w2, ..., wn)$. Πρακτικά, αυτό το οποίο κάνει ένα γλωσσικό μοντέλο είναι να αποτυπώνει πόσο πιθανό είναι να σχηματιστεί μία ακολουθία λέξεων (πρόταση) σε μία γλώσσα. Έτσι, η ακολουθία "happy birthday John", θα έχει μεγαλύτερη πιθανότητα από την ακολουθία "happy dog table" (η οποία δεν βγάζει κανένα νόημα).

Τα μοντέλα αυτά είναι πολύ χρήσιμα σε διάφορους τομείς στην ΕΦΓ, όπως για την παραγωγή κειμένων ή την πρόβλεψη της επόμενης λέξης σε μία ακολουθία

λέξεων (χαρακτηριστικό παράδειγμα είναι οι εφαρμογές πληκτρολογίου στα smartphones, όπου προτείνουν στον χρήστη πιθανές επόμενες λέξεις, βάση των λέξεων που έχει ήδη πληκτρολογήσει).

Η διαδικασία δημιουργίας του γλωσσικού μοντέλου γίνεται χωρίς επίβλεψη. Το μόνο που απαιτείται είναι μία αρκετά μεγάλη συλλογή από κείμενα, από τα οποία το ΤΝΔ θα "μάθει" τα διανύσματα των λέξεων. Τα διανύσματα αυτά είναι συνέπεια της διαδικασίας μάθησης και όχι το αντικείμενο της διαδικασίας. Στο word2vec παρουσιάζονται 2 διαφορετικές τεχνικές, η *Continuous Bag of Words (CBOW)* και η *Skip-Gram*. Για να περιγράψουμε τις δύο τεχνικές ας υποθέσουμε αρχικά ότι η τρέχουσα λέξη υπό εξέταση είναι η w_i.



(α′) CBoW μοντέλο. Δεδομένου ενός παραθύρου από λέξεις, στόχος είναι να προβλέψει την πιθανότητα να βρεθεί κάθε λέξη σε αυτό το πλαίσιο.

(β′) Skip-Gram μοντέλο. Δεδομένης μίας λέξης, στόχος είναι η πρόβλεψη για το πόσο πιθανό είναι οι άλλες λέξεις να βρεθούν στο πλαίσιο που την περιβάλλουν.

Σχήμα 3.8: Σύγκριση μεταξύ του Continuous Bag of Words (CBOW) και Skip-Gram (Mikolov κ.ά. 2013a).

Continuous Bag of Words (CBOW) Η είσοδος στο νευρωνικό δίκτυο είναι ένα παράθυρο λέξεων πριν και μετά την τρέχουσα λέξη w_i της μορφής $w_i-2, w_i-1, w_i+1, w_i+2$. Η έξοδος του δικτύου είναι η λέξη w_i. Αυτό που μαθαίνει το νευρωνικό δίκτυο είναι: *να προβλέπει την λέξη δεδομένου του πλαισίου (context)*. Το πόσο μεγάλο θα είναι το παράθυρο των λέξεων πριν και μετά είναι κάτι που καθορίζεται βάση μίας παραμέτρου (window). Για παράδειγμα για *window=3*, έχουμε ως είσοδο τις λέξεις $w_i-3, w_i-2, w_i-1, w_i+1, w_i+2, w_i+3$.

Skip-gram Η είσοδος στο νευρωνικό δίκτυο είναι η τρέχουσα λέξη w_i και η έξοδος είναι ένα σύνολο λέξεων της μορφής $w_i-2, w_i-1, w_i+1, w_i+2$. Η λειτουργία που

μαθαίνει το νευρωνικό δίκτυο είναι: να προβλέπει το περιεχόμενο (context) δεδομένης της λέξης. Το πόσο μεγάλο θα είναι το παράθυρο, αλλά και πόσες λέξεις θα παραλειφθούν πριν και μετά την τρέχουσα λέξη είναι παράμετροι. Για παράδειγμα αν skip=2 και window=3, έχουμε ως είσοδο τις λέξεις $w_i-5, w_i-4, w_i-3, w_i+3, w_i+4, w_i+5$.

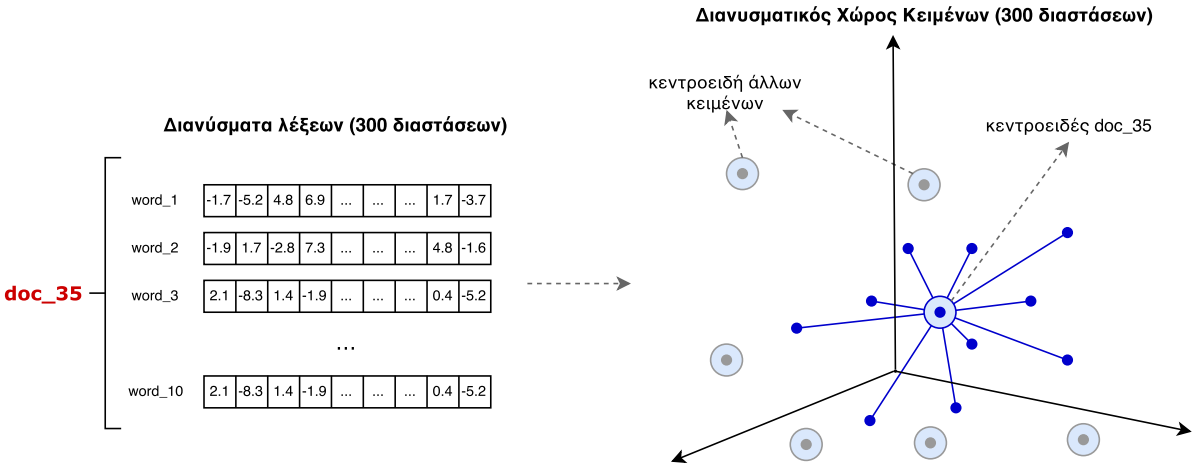
Δηλαδή έχουμε:

- CBOW = context -> Neural Network -> word
- Skip-gram= word -> Neural Network -> context

3.2.3.2 Χρήση Διανυσμάτων Λέξεων

Για να δοθεί ένα κείμενο σε έναν αλγόριθμο μηχανικής μάθησης πρέπει να έχει προηγουμένως αναπαρασταθεί ως ένα διάνυσμα. Στις BoW αναπαραστάσεις καταλήγουμε αμέσως σε ένα διάνυσμα για κάθε κείμενο. Στις κατανεμημένες αναπαραστάσεις κειμένων που παρουσιάστηκαν πιο πάνω και πάλι μπορούμε να χρησιμοποιήσουμε το διάνυσμα του κειμένου ως έχει για είσοδο στον αλγόριθμο μηχανικής μάθησης. Όμως στην περίπτωση των word embeddings έχουμε την αναπαράσταση για κάθε λέξη, αλλά όχι ακόμα για ολόκληρο το κείμενο. Τα διανύσματα λέξεων χρησιμοποιούνται συνήθως με τους εξής δύο τρόπους για την παραγωγής της τελικής αναπαράστασης του κειμένου:

Neural Bag Of Words (NBOW) Η προσέγγιση αυτή είναι αρκετά απλή και συχνά πετυχαίνει ικανοποιητικά αποτελέσματα. Για να καταλήξουμε σε ένα ενιαίο διάνυσμα (αναπαράσταση) για κάθε κείμενο, γίνεται μία συσσωμάτωση των διανυσμάτων των λέξεων του κειμένου. Συνήθως υπολογίζεται το κέντρο βάρους (κεντροειδές) των διανυσμάτων. Επίσης, έχουν δοκιμαστεί και άλλες απλές μαθηματικές συναρτήσεις (min, max, sum, tf-idf mean κλπ.), οι οποίες πετυχαίνουν παρόμοια αποτελέσματα (De Boom κ.ά. 2016). Η προσέγγιση αυτή αγνοεί την σειρά των λέξεων.



Σχήμα 3.9: Συσσωμάτωση των διανυσμάτων λέξεων 300 διαστάσεων, με υπολογισμό του κέντρου βάρους (κεντροειδές) τους.

Στο παράδειγμα στην Εικόνα 3.9 έχουμε διανύσματα λέξεων 300 διαστάσεων. Αφού υπολογίσουμε το κέντρο βάρους των λέξεων κάθε κειμένου, μπορούμε να αναπαραστήσουμε το κείμενο ως ένα διάνυσμα 300 διαστάσεων. Με αυτό τον εύκολο τρόπο παράγουμε μία αριθμητική αναπαράσταση για κάθε κείμενο. Στη συνέχεια μπορούμε να στείλουμε τα διανύσματα των κειμένων (κέντρα βάρους) σε έναν αλγόριθμο μηχανικής μάθησης για ταξινόμηση.

Τεχνητά Νευρωνικά Δίκτυα Τα ΤΝΔ αφορούν τεχνικές για την δημιουργία στατιστικών μοντέλων, οι οποίες είναι εμπνευσμένες από τον τρόπο λειτουργίας του ανθρώπινου νευρώνα. Έχουν το πλεονέκτημα, ότι μπορούν να κατασκευάσουν ή να μάθουν χαρακτηριστικά που να μοντελοποιούν μία είσοδο (όπως ένα κείμενο), χωρίς να τους έχουν σχεδιαστεί εκ των προτέρων από έναν άνθρωπο. Αυτή η ευελιξία τα έχει κάνει πολύ δημοφιλή σε προβλήματα επεξεργασίας φυσικής γλώσσας, καθώς μπορούν να μοντελοποιήσουν ως ένα βαθμό την ασάφεια της φυσικής γλώσσας.

Οι πιο δημοφιλείς αρχιτεκτονικές ΤΝΔ που εφαρμόζονται σε προβλήματα ΕΦΓ είναι τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN), Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN) και Αναδρομικά Νευρωνικά Δίκτυα (Recursive Neural Networks - TreeNN[2]). Ο τρόπος λειτουργίας τους θα αναλυθεί στο επόμενο κεφάλαιο, όπου θα παρουσιαστούν οι αλγόριθμοι μηχανικής μάθησης. Ακολουθεί μία σύνοψη των τριών αρχιτεκτονικών.

Convolutional Neural Networks (CNN) Οι αρχιτεκτονικές αυτές είναι εμπνευσμένες από τον τρόπο λειτουργίας του οπτικού συστήματος του εγκεφάλου και έχουν σχεδιαστεί με στόχο προβλήματα Υπολογιστικής Όρασης (Yann LeCun κ.ά. 1998). Το γεγονός ότι εκτελούν πολλούς υπολογισμούς παράλληλα, τα κάνει ιδιαίτερα γρήγορα. Η ταχύτητα τους, αλλά και η ευκολία στην εκπαίδευση τέτοιων αρχιτεκτονικών, τα έχει κάνει δημοφιλή στην ΕΦΓ , παρόλο που δεν σχεδιάστηκαν για τέτοιου είδους προβλήματα. Όμως, ένα σημαντικό μειονέκτημα για τα CNN είναι ότι δεν αναγνωρίζουν την έννοια της σειράς. Έτσι, δεν μπορούν να αξιοποιήσουν τη σημαντική πληροφορία της σειράς των λέξεων, αλλά και της ιεραρχικής δομής της γλώσσας.

Recurrent Neural Networks (RNN) Τα RNN είναι σχεδιασμένα για να λειτουργούν πάνω σε ακολουθίες. Αυτό ακριβώς είναι και ένα κείμενο, μία ακολουθία λέξεων. Συνεπώς, είναι η πιο "φυσική" επιλογή, καθώς επεξεργάζονται την είσοδό τους (κείμενο) όπως και ο άνθρωπος, δηλαδή σειριακά. Αυτό τους δίνει την δυνατότητα να αξιοποιούν την πληροφορία της σειράς των λέξεων. Όμως, ακριβώς επειδή η επεξεργασία της εισόδου γίνεται σειριακά, είναι αρκετά πιο αργά από τα CNN . Επιπλέον, έχουν κάποια τεχνικά προβλήματα σχετικά με τον τρόπο εκπαίδευσής τους, για τα οποία όμως έχουν προταθεί διάφοροι τρόποι για την αντιμετώπισή τους(Κεφάλαιο 4.3.3).

Recursive Neural Networks (TreeNN) Τα TreeNN είναι σχεδιασμένα για την επεξεργασία ιεραρχιών. Για να εφαρμοστούν σε προβλήματα επεξεργασίας φυσι-

[2]Επειδή υπάρχει σύγκρουση όρων ανάμεσα σε Recurrent και Recursive Neural Networks, στο πλαίσιο αυτής της εργασίας θα αναφερόμαστε στα Recursive Neural Networks ώς TreeNN. Ο λόγος για την επιλογή της συντόμευσης είναι, ότι κυρίως αυτά τα δίκτυα εφαρμόζονται σε δέντρα/ιεραρχίες

κής γλώσσας, απαιτούν να έχει γίνει συντακτική ανάλυση του κειμένου. Ο λόγος είναι γιατί εφαρμόζονται πάνω στο συντακτικό δέντρο του κειμένου και όχι στην αρχική ακολουθία των λέξεων. Αυτό τους επιτρέπει να αξιοποιούν τις εξαρτήσεις ανάμεσα στις λέξεις. Επίσης στο πλαίσιο της ΑΣ, παράγουν αποτελέσματα, τα οποία επιτρέπουν σε έναν άνθρωπο να ερμηνεύσει τον τρόπο λειτουργίας τους. Ένα TreeNN, στην ουσία είναι ένα RNN, στο οποίο οι λέξεις του δίνονται ως είσοδος όχι με την σειρά με την οποία εμφανίζονται στο κείμενο, αλλά σύμφωνα με την δενδρική δομή του κειμένου.

Όμως, το γεγονός ότι απαιτούν το συντακτικό δέντρο των κειμένων, είναι σημαντικό μειονέκτημα. Οι λόγοι είναι ότι, (1) αυτή είναι μία αρκετά χρονοβόρα διαδικασία και (2) οι ιεραρχίες που παράγονται εξαρτώνται σε μεγάλο βαθμό από την δομή του κειμένου. Αυτό σημαίνει ότι, σε κείμενα όπως σε αυτά που συναντώνται σε κοινωνικά δίκτυα, σαν το Twitter, τα οποία περιέχουν αρκετά συντακτικά και γραμματικά λάθη, τα αντίστοιχα συντακτικά δέντρα δεν είναι αρκετά καλά. Αυτό συμβαίνει διότι η δομή του κειμένου δεν ακολουθεί τους σωστούς ή συμβατικούς γλωσσικούς κανόνες. Το αποτέλεσμα είναι να επηρεάζονται σημαντικά οι επιδόσεις του μοντέλου που βασίζεται σε αυτά.

3.2.3.3 Ιδιότητες Διανυσμάτων Λέξεων

Μέχρι τώρα είδαμε πως παράγονται τα διανύσματα λέξεων και διάφορους τρόπους για την αξιοποίησή τους από έναν αλγόριθμο μηχανικής μάθησης. Σε αυτό το σημείο θα γίνει αναφορά σε ορισμένες ενδιαφέρουσες ιδιότητες των διανυσμάτων λέξεων. Οι ιδιότητες αυτές βοηθούν στην εκτέλεση διάφορων χρήσιμων εργασιών και ξεκαθαρίζουν τι πληροφορίες κωδικοποιούνται σε ένα τέτοιο διάνυσμα.

Ένα διάνυσμα λέξης, στην ουσία κωδικοποιεί το πλαίσιο (context) μίας λέξης. Το πλαίσιο μίας λέξης σε ένα κείμενο, είναι οι λέξεις οι οποίες εμφανίζονται γύρω της. Λέξεις οι οποίες εμφανίζονται συχνά στο ίδιο πλαίσιο θα έχουν παρόμοια διανύσματα. Ας πάρουμε για παράδειγμα το ακόλουθο πλαίσιο λέξεων:

```
... ο Άλμπερτ Αϊνστάιν ήταν ο __, ο οποίος δημιούργησε
την θεωρία της σχετικότητας ...
```

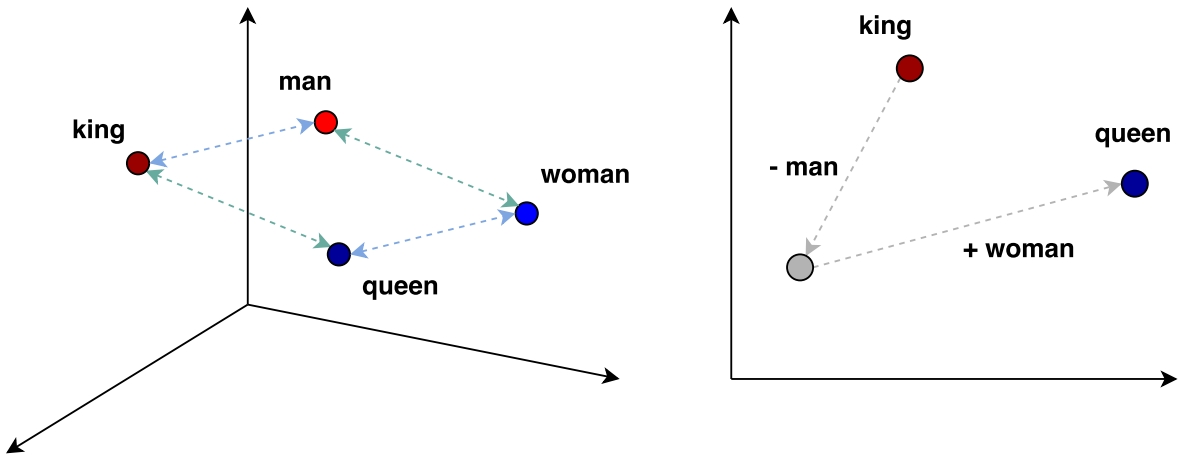
Δεδομένου αυτού του πλαισίου, είναι πιο πιθανό να συναντήσει κανείς στη θέση της λέξης που λείπει, τη λέξη "*φυσικός*" ή "*επιστήμονας*", παρά τη λέξη "*τραπέζι*".

Αυτή η ιδιότητα, καθώς και το γεγονός ότι οι λέξεις βρίσκονται σε ευκλείδειο χώρο, επιτρέπει την εκτέλεση αριθμητικών πράξεων (αναλογιών), όπως στο διάσημο πλέον παράδειγμα (Mikolov κ.ά. 2013b):

```
vec(''king'') - vec(''man'') + vec(''woman'') = vec(''queen'')
```

Αν και οι παραπάνω πράξεις έχουν μεγάλο ενδιαφέρον και είναι πολύ χρήσιμη ιδιότητα για αρκετά προβλήματα επεξεργασίας φυσικής γλώσσας, υπάρχει ένα πρόβλημα σε ότι αφορά την ΑΣ. Όπως αναφέρθηκε, λέξεις οι οποίες βρίσκονται συχνά στο ίδιο πλαίσιο, θα βρίσκονται και κοντά στον χώρο. Όμως, αυτό σημαίνει ότι κοντά θα βρίσκονται εκτός από συνώνυμα όπως "έξυπνος"-"ευφυής" και αντώνυμα όπως "καλός"-"κακός", τα οποία έχουν αντίθετο συναισθηματικό προσανατολισμό, παρόλο που συχνά εμφανίζονται στο ίδιο πλαίσιο. Για παράδειγμα, στην πρόταση:

```
Δίαβασα ένα πολύ ___  βιβλίο στις διακοπές.
```



Σχήμα 3.10: Πράξεις με διανύσματα λέξεων (Mikolov κ.ά. 2013b)

είναι το ίδιο πιθανό να εμφανίζονται οι λέξεις "καλό" και "κακό". Αυτό καταδεικνύει, ότι στον χώρο δεν αποτυπώνονται πλήρως τα σημασιολογικά χαρακτηριστικά των λέξεων. Είναι σημαντικό να τονιστεί αυτό, διότι αρκετές φορές αναφέρεται στην βιβλιογραφία, ότι τα διανύσματα λέξεων με τεχνικές όπως το word2vec, περιέχουν τα σημασιολογικά χαρακτηριστικά των λέξεων.

Χαρακτηριστικά, σε ένα πρόβλημα ΑΣ, το ιδανικό θα ήταν οι λέξεις να ήταν κατανεμημένες στο χώρο, ανάλογα με το συναισθηματικό προσανατολισμό τους, ή τουλάχιστον με τέτοιο τρόπο, ώστε να αποτυπώνεται και αυτή οι πληροφορία. Για να επιτευχθεί αυτό έχουν προταθεί διάφορες λύσεις (Maas κ.ά. 2011; Tang κ.ά. 2014). Όμως, αυτές οι τεχνικές απαιτούν την ύπαρξη σημειωμένων δεδομένων. Ο λόγος είναι ότι πρέπει να είναι γνωστός ο συναισθηματικός προσανατολισμός κάθε κειμένου, στο σύνολο δεδομένων, ώστε να μετακινηθούν κατάλληλα οι λέξεις.

Στο πλαίσιο των τεχνικών που κάνουν χρήση ΤΝΔ, η συνηθέστερη πρακτική είναι η εκπαίδευση του επιπέδου εμφύτευσης (embedding layer) του δικτύου. Πρώτα αρχικοποιούνται τα βάρη του επιπέδου εμφύτευσης, με τις τιμές των διανυσμάτων των λέξεων. Στη συνέχεια το επίπεδο αυτό εκπαιδεύεται μαζί με τα υπόλοιπα επίπεδα του δικτύου. Αυτό έχει σαν αποτέλεσμα την αλλαγή των βαρών των λέξεων και την μετακίνηση των λέξεων στο χώρο.

Όμως, είναι σημαντικό να γίνει κατανοητό τι συνέπειες έχει αυτή η διαδικασία. Κατά τη φάση της εκπαίδευσης, αυτό το οποίο θα κάνει εμμέσως το ΤΝΔ, είναι να συσχετίσει τις διάφορες περιοχές του διανυσματικού χώρου των λέξεων, με ένα συναίσθημα ή συναισθηματικό προσανατολισμό. Όμως, λέξεις οι οποίες δεν έχουν συναντηθεί κατά τη φάση της εκπαίδευσης, θα έχουν μείνει στην αρχική τους θέση. Αυτό σημαίνει ότι η θέση τους στο χώρο, μπορεί να μην αντικατοπτρίζει πλέον το "σωστό" συναίσθημα των λέξεων. Αυτό μπορεί να "μπερδέψει" το μοντέλο και να οδηγήσει σε λάθη.

Έτσι, αν το σύνολο δεδομένων δεν είναι αρκετά μεγάλο, ώστε να περιέχει την πλειοψηφία των πιθανών λέξεων που μπορεί να συναντήσει ένα μοντέλο σε ένα άγνωστο κείμενο, ίσως είναι προτιμότερο να διατηρηθούν οι λέξεις στις αρχικές τους θέσεις, ακόμα και αν αυτή δεν είναι η ιδανική σε ότι αφορά τον συναισθηματικό τους προσανατολισμό. Ένα σύνθετο νευρωνικό δίκτυο, έχει την ικανότητα να αξιοποιήσει τα συντακτικά και σημασιολογικά χαρακτηριστικά των λέξεων (ακόμα και τα λίγα διαθέσιμα) και να πετύχει καλά αποτελέσματα σε προβλήματα αναγνώ-

ρισης συναισθήματος, ακόμη και χωρίς να είναι κωδικοποιημένος ευθέως ο συναισθηματικός προσανατολισμός. Το συμπέρασμα είναι ότι η φύση του προβλήματος, το μέγεθος του συνόλου δεδομένων, η ποιότητα των αρχικών διανυσμάτων και η αρχιτεκτονική του δικτύου, είναι όλοι παράγοντες οι οποίοι πρέπει να ληφθούν υπόψη τους για την επιλογή της σωστής προσέγγισης.

3.3 Προεπεξεργασία Κειμένων

Ένα κείμενο σε ηλεκτρονική μορφή, αναπαριστάται ως μία σειρά από χαρακτήρες. Πριν την επεξεργασία των κειμένων για την εξαγωγή των χαρακτηριστικών, συνήθως απαιτείται μία προ-επεξεργασία του κειμένου. Η διαδικασία αυτή, ονομάζεται *προεπεξεργασία κειμένου* (text preprocessing) και αποτελείται από μία σειρά βημάτων, μερικά από τα οποία είναι προαπαιτούμενα για την εκτέλεση των υπολοίπων. Ανάλογα με το είδος της ανάλυσης την οποία επιθυμούμε να εφαρμόσουμε στη συνέχεια και τα χαρακτηριστικά που σκοπεύουμε να εξάγουμε, εκτελούνται ένα η περισσότερα από αυτά τα βήματα. Σε αυτό το σημείο θα δούμε τα βασικότερα από αυτά.

ekphrasis Στα πλαίσια της έρευνας για την ανάπτυξη καλύτερων μοντέλων ΑΣ, αναπτύχθηκε ένα εργαλείο προεπεξεργασίας κειμένων, το *ekphrasis*[3]. Το ekphrasis, εκτελεί: (1) λεκτική ανάλυση (tokenization) η οποία διατηρεί εκφράσεις οι οποίες είναι χρήσιμες για τον προσδιορισμό του συναισθήματος, (2) ορθογραφική διόρθωση, (3) κανονικοποίηση λέξεων και φράσεων (text normalization), (3) σημείωση λέξεων και φράσεων (word annotation), (4) διαχωρισμό ενοποιημένων λέξεων (text segmentation), στις επιμέρους λέξεις (για τον διαχωρισμό των hashtags). Σχεδιάστηκε με γνώμονα τις απαιτήσεις στην ανάλυση μηνυμάτων από κοινωνικά δίκτυα (Twitter, Facebook κλπ.), όμως μπορεί να χρησιμοποιηθεί για την ανάλυση οποιουδήποτε είδους κειμένων, καθώς η λειτουργικότητά του υπερκαλύπτει τις ανάγκες των απλών κειμένων. Πιο αναλυτική παρουσίαση γίνεται στο Κεφάλαιο 5.5.

3.3.1 Λεκτική ανάλυση (Tokenization)

Στην διαδικασία αυτή ένα κείμενο, μετατρέπεται από μία ακολουθία χαρακτήρων, σε μία ακολουθία από *λεκτικές μονάδες ή όρους* (tokens ή terms), όπως λέξεις, σημεία στίξης, αριθμούς κλπ. Αυτό το βήμα είναι απαραίτητο για την εκτέλεση των υπολοίπων, καθώς ενεργούν πάνω στους όρους του κειμένου. Σε γλώσσες με το Ελληνικό ή το Λατινικό (Αγγλικά) αλφάβητο, η διαδικασία αυτή είναι πιο απλή, καθώς οι λέξεις χωρίζονται με το κενό. Όμως, ο διαχωρισμός των λέξεων απλά στα κενά δεν είναι πάντα αρκετός. Για παράδειγμα, τα "Νέα Ιωνία", "εκ περιτροπής", "εν όψει" ή στα Αγγλικά "rock 'n' roll", αν και αποτελούνται από πολλές λέξεις οι οποίες χωρίζονται με κενό ή ορισμένα σημεία στίξης, αναφέρονται σε μία συγκεκριμένη έννοια και συνεπώς θα ήταν προτιμότερο να διατηρηθούν ως ένας όρος. Αντίθετα, τα "I'm" και "doesn't" αν και δεν χωρίζονται με κενό, αποτελούνται από ξεχωριστές λέξεις-έννοιες ("I am" και "does not").

Συνεπώς, για την καλύτερη λεκτική ανάλυση, ακόμα και σε γλώσσες όπως τα Αγγλικά, θα πρέπει να δοθεί σημασία σε αυτό το βήμα. Ένας ακόμα λόγος για

[3] https://github.com/cbaziotis/ekphrasis

την σημασία της καλής λεκτικής ανάλυσης, είναι ότι όλα τα επόμενα βήματα βασίζονται σε αυτή. Έτσι, τα όποια λάθη θα προωθηθούν στα επόμενα βήματα της επεξεργασίας του κειμένου, αλλά και στην εξαγωγή των χαρακτηριστικών και την δημιουργία του μοντέλου. Ειδικά η ΑΣ είναι αρκετά ευαίσθητη στην λεκτική ανάλυση, διότι υπάρχουν εκφράσεις οι οποίες είναι καθοριστικές για τον προσδιορισμό του ύφους του κειμένου. Τέτοιες εκφράσεις είναι:

- Λέξεις χωρισμένες με παύλες. Για παράδειγμα, "over-consumption", "anti-american", "mind-blowing".
- Emoticons, όπως ">:(", ":))", ":-D".
- Λογοκριμένες λέξεις, όπως "f**k", "s**t".
- Λέξεις με έμφαση, όπως "a *great* time", "I don't *think* I know...".

Επιπρόσθετα, υπάρχουν εκφράσεις οι οποίες δεν προσφέρουν επιπλέον πληροφορίες για τον προσδιορισμό του συναισθήματος, αλλά αποτελούν απλά θόρυβο για το μοντέλο. Τέτοιες εκφράσεις είναι:

- Ημερομηνίες, όπως "Feb 18th", "December 2, 2016", "December 2-2016", "10/17/94", "3 December 2016", "April 25, 1995", "11.15.16", "November 24th 2016", "January 21st".
- Ώρες, όπως "5:45pm", "11:36 AM", "2:45 pm", "5:30".
- Συναλλάγματα, όπως "$220M", "$2B", "$65.000", "€10", "$50K".
- Τηλεφωνικοί αριθμοί.
- Ηλεκτρονικοί σύνδεσμοι (URLs), όπως "http://www.cs.unipi.gr", ή "https://t.co/Wfw5Z1iSEt".

Υπάρχουν αναρίθμητοι συνδυασμοί ημερομηνιών ή ηλεκτρονικών συνδέσμων. Με τον διαχωρισμό των όρων αυτών στους επιμέρους αριθμούς και λέξεις, αυξάνεται το λεξιλόγιο των όρων, χωρίς να προσφέρουν χρήσιμη πληροφορία. Η αναγνώριση και διατήρηση αυτών των εκφράσεων ως έναν όρο, επιτρέπει σε επόμενο βήμα την κανονικοποίηση τους, δηλαδή αντικατάσταση όλων των ημερομηνιών με έναν συγκεκριμένο όρο, όπως `<date>`. Με αυτό τον τρόπο, (1) μειώνεται σημαντικά το συνολικό λεξιλόγιο, αλλά και οι όροι σε μία παρατήρηση (έγγραφο ή μήνυμα) και (2) το μοντέλο κερδίζει πληροφορία, καθώς στο πρόβλημα της ΑΣ δεν μας ενδιαφέρει ποια είναι η ημερομηνία ή ο τηλεφωνικός αριθμός που περιέχεται στο κείμενο, αλλά το γεγονός ότι η ακολουθία γραμμάτων αναφέρεται σε μία ημερομηνία.

3.3.2 Αναγνώριση Μερών του Λόγου

Η διαδικασία της *Αναγνώρισης των Μερών του Λόγου* (Part-of-Speech Tagging), αντιστοιχεί σε κάθε έναν από τους όρους που έχουν εξαχθεί από την λεκτική ανάλυση, σε ένα και μόνο μέρος του λόγου από μία προκαθορισμένη λίστα (ρήμα, επίθετο, επίρρημα κλπ.). Η λίστα αυτή μπορεί να είναι πολύ γενική, με λίγα μόνο μέρη του λόγου, η αρκετά αναλυτική, διακρίνοντας ανάμεσα στα διάφορα είδη επιρρημάτων ή ρημάτων και των άλλων μερών του λόγου. Οι επιδόσεις της αναγνώρισης, επηρεάζεται από την λεκτική ανάλυση, διότι εκτελείται επί των όρων που έχουν εξαχθεί.

3.3.3 Αποκατάληξη (Stemming)

Η *αποκατάληξη* (stemming) είναι η διαδικασία της περικοπής των καταλήξεων των όρων. Το αποτέλεσμα της αποκατάληξη μίας λέξης, είναι το στέλεχος της λέξης (stem). Το στέλεχος μίας λέξης δεν είναι πάντα και αυτό κανονική λέξη. Η αποκατάληξη απλοποιεί τις λέξεις, κάνοντας τα συστήματα μηχανικής μάθησης λιγότερο επιρρεπής στις μορφολογικές διαφορές των λέξεων. Ένας αρκετά δημοφιλής αλγόριθμος για αποκατάληξη είναι ο Porter (Porter 1980). Στον Πίνακα 3.1, παρουσιάζονται παραδείγματα αποκατάληξης ρημάτων με την χρήση του αλγορίθμου Porter. Επίσης, ένα η πρόταση,

> Somewhere, something incredible is waiting to be known.
> - Carl Sagan

μετά την αποκατάληξη γίνεται:

> Somewher someth incred is wait to be known

Λέξεις	Stem	Stem	Stem
connection connections connective connected connecting	connect	study studies studied studying	studi

Πίνακας 3.1: Παραδείγματα Αποκατάληξης με τον αλγόριθμο Porter (Porter 1980). Λέξεις με κοινή ρίζα χαρτογραφούνται στην ίδια λέξη-ρίζα (stem). Όπως φαίνεται και στην εικόνα, το stem μπορεί να μην είναι και το ίδιο κανονική λέξη (studi).

3.3.4 Λημματοποίηση (Lemmatization)

Η *Λημματοποίηση* (Lemmatization) αφορά την αναγωγή των όρων στις ρίζες τους. Όπως η αποκατάληξη, στοχεύει στην μείωση των μορφολογικών διαφορών των λέξεων, όμως είναι λιγότερο επιθετική διαδικασία από την αποκατάληξη. Επίσης το λήμμα μίας λέξης είναι *πάντα* μία κανονική λέξη. Για να μπορέσει να εκτελεστεί απαιτείται να έχει γίνει προηγουμένως αναγνώριση των μερών του λόγου, καθώς να για να μπορέσει να βρεθεί το σωστό λήμμα πρέπει να είναι γνωστό τι μέρος του λόγου είναι. Ορισμένα παραδείγματα αποκατάληξης είναι:

```
am, are, is -> be
car, cars, car's, cars' -> car
```

3.3.5 Αφαίρεση Τερματικών Όρων (Stopwords)

Η *αφαίρεση τερματικών όρων*, είναι η διαδικασία αφαίρεσης των λέξεων με μικρό σημασιολογικό περιεχόμενο, οι οποίες ονομάζονται τερματικοί όροι (stopwords). Για αυτή την διαδικασία αρχικά καθορίζεται μία λίστα με τους τερματικούς όρους (λεξικό) και αφαιρούνται από το κείμενο όλοι οι όροι που ανήκουν στη λίστα. Με τον τρόπο αυτό μειώνεται το συνολικό λεξιλόγιο του συνόλου δεδομένων

(εγγράφων) και αποτρέπεται η δημιουργία περιττών χαρακτηριστικών. Ορισμένοι τερματικοί όροι είναι:

```
a, an, and, are, as, at, be, by, for, from, has, he, in, is, it,
   its, of, on, that, the, to, was, were, will, with
```

Αν και σε πολλά συστήματα κατηγοριοποίησης κειμένων γίνεται αφαίρεση των τερματικών όρων, στη ΑΣ έχει δειχθεί ότι έχει αρνητικά αποτελέσματα (Saif κ.ά. 2014).

3.3.6 Διαχείριση Αρνήσεων (Negation Handling)

Ο συναισθηματικός προσανατολισμός μίας λέξης συνήθως αντιστρέφεται όταν βρίσκεται στο πλαίσιο μίας άρνησης. Για την αξιοποίηση αυτής της πληροφορίας, μία συνήθης πρακτική είναι η προσθήκη του επιθέματος _NEG σε κάθε λέξη η οποία βρίσκεται στο πλαίσιο μίας άρνησης. Αρχικά προσδιορίζεται ένα σύνολο από λέξεις ο οποίες θεωρούνται ότι αποτελούν την αφετηρία μίας άρνησης, όπως:

```
wouldn't, no, hasn't, shouldn't, hadn't, haven't, doesnt, don't,
   arent, hadnt, havent, couldnt, cant, ain't, hasnt, couldn't,
   never, wouldnt, isn't, aren't, doesn't, not, none, didn't,
   nothing
```

Στη συνέχεια, σε όσες λέξεις βρίσκονται μετά την άρνηση, μέχρι και το τέλος της πρότασης, προστίθεται η κατάληξη _NEG. Έτσι για παράδειγμα, η πρόταση:

```
No one enjoyed this movie.
```

μετατρέπεται σε:

```
No one_NEG enjoyed_NEG this_NEG movie_NEG.
```

Κεφάλαιο 4

Μηχανική Μάθηση

Σε αυτό το κεφάλαιο θα γίνει παρουσίαση και σύγκριση των δημοφιλέστερων προσεγγίσεων για προβλήματα κατηγοριοποίησης κειμένου, οι οποίες κάνουν χρήση τεχνικών μηχανικής μάθησης. Αρχικά θα δοθεί μία σύντομη εισαγωγή στις έννοιες της μηχανικής μάθησης. Στην συνέχεια θα γίνει σύγκριση ανάμεσα στις "παραδοσιακές" προσεγγίσεις, με αυτές που κάνουν χρήση Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ). Η σύγκριση θα γίνει στο πλαίσιο των προβλημάτων ταξινόμησης κειμένων, καθώς τα περισσότερα προβλήματα Ανάλυσης Συναισθήματος (ΑΣ) ανήκουν σε αυτή την κατηγορία.

Οι "παραδοσιακές" προσεγγίσεις απαιτούν την διαδικασία της *μηχανικής χαρακτηριστικών* (feature engineering), όπου τα χαρακτηριστικά θα πρέπει να σχεδιαστούν εκ των προτέρων από τον άνθρωπο και στη συνέχεια να δοθούν στον αλγόριθμο μηχανικής μάθησης. Αντίθετα, τα ΤΝΔ εκτελούν *μάθηση χαρακτηριστικών* (feature learning), αυτοματοποιώντας την διαδικασία σχεδιασμού των χαρακτηριστικών, αλλά και επιτρέποντας την δημιουργία αφηρημένων χαρακτηριστικών, τα οποία είναι πολύ δύσκολο να σχεδιαστούν από τον άνθρωπο. Εκτός από τους κύριους αλγορίθμους μηχανικής μάθησης, θα περιγραφούν και ορισμένες τεχνικές μηχανικής χαρακτηριστικών, οι οποίες είναι απαραίτητες για την δημιουργία των καλών (αντιπροσωπευτικών) χαρακτηριστικών για πολλά από τα μοντέλα.

Στόχος αυτού του κεφαλαίου δεν είναι η αναλυτική παρουσίαση κάθε αλγορίθμου, διότι αυτό ξεφεύγει από τον σκοπό αυτής της εργασίας. Θα παρουσιαστούν τα βασικά χαρακτηριστικά, οι ιδιότητες και ιδιαιτερότητες κάθε αλγορίθμου. Μεγαλύτερη έμφαση θα δοθεί στα ΤΝΔ, καθώς η έρευνα σε αυτά είναι ένα από τα αντικείμενα της εργασίας. Στόχος είναι να εξετάσουμε ποιες τεχνικές είναι καταλληλότερες για να προσεγγίσουμε ένα πρόβλημα ΑΣ, αλλά και για ποιους λόγους.

4.1 Θεωρητικό Υπόβαθρο

Η μηχανική μάθηση είναι το επιστημονικό πεδίο, το οποίο ερευνά τον τρόπο με τον οποίο μπορούμε να σχεδιάσουμε ένα μηχάνημα, ώστε να εκτελεί μία εργασία ή να κάνει προβλέψεις, *χωρίς* όμως να το έχουμε ρητά προγραμματίσει εκ των προτέρων. Αυτό κατορθώνεται σχεδιάζοντας αλγορίθμους οι οποίοι μαθαίνουν να εκτελούν την εργασία που μας ενδιαφέρει, μέσα από δεδομένα (εμπειρία). Η διαδικασία κατά την οποία ο αλγόριθμος μηχανικής μάθησης μαθαίνει να εκτελεί την συγκεκριμένη εργασία, αξιοποιώντας τα δεδομένα που του παρέχουμε, ονομάζεται *εκπαίδευση*.

4.1.1 Τύποι Μηχανικής Μάθησης

Η μηχανική μάθηση χωρίζεται σε τρεις γενικές κατηγορίες ή προσεγγίσεις, ανάλογα με την φύση των δεδομένων που διαθέτουμε και το είδος του προβλήματος που θέλουμε να λύσουμε.

4.1.1.1 Επιβλεπόμενη Μάθηση

Στην *επιβλεπόμενη μάθηση* (supervised learning) *ή μάθηση με επιτήρηση*, διαθέτουμε ένα σύνολο δεδομένων (dataset), το οποίο αποτελείται από *παρατηρήσεις ή παραδείγματα* (training examples), για κάθε ένα από τα οποία γνωρίζουμε το σωστό αποτέλεσμα. Σε αυτή την περίπτωση λέμε ότι το σύνολο δεδομένων είναι *επισημασμένο* (labeled). Στόχος αυτού του είδους μάθησης, είναι να ανακαλύψουμε την *σχέση* ανάμεσα στις παρατηρήσεις και στα αποτελέσματα. Η σχέση η οποία ανακαλύπτουμε είναι ένα μοντέλο τους προβλήματος. Το μοντέλο είναι το αποτέλεσμα της μάθησης και το χρησιμοποιούμε για να κάνουμε προβλέψεις σε νέες, άγνωστες παρατηρήσεις. Υπάρχουν δύο είδη επιβλεπόμενης μάθησης.

Ταξινόμηση (Classification) Αφορά τα προβλήματα στα οποία θέλουμε να ταξινομήσουμε τις παρατηρήσεις σε ένα σύνολο προεπιλεγμένων *διακριτών τιμών ή κλάσεων*. Για παράδειγμα, οι παρατηρήσεις μπορούν να είναι ένα σύνολο από κείμενα με κριτικές ταινιών, για κάθε μία από τις οποίες, γνωρίζουμε τον συναισθηματικό προσανατολισμό της. Ο προσανατολισμός κάθε κειμένου μπορεί να ανήκει σε μία από τις κλάσεις του συνόλου {*θετικός*, *αρνητικός*, *ουδέτερος*}, το οποίο σύνολο ορίζεται εξαρχής. Στόχος είναι να σχεδιάσουμε έναν σύστημα μηχανικής μάθησης, το οποίο παρατηρώντας τα επισημασμένα δεδομένα που του παρέχουμε, να ανακαλύψει μία σχέση, ανάμεσα στα χαρακτηριστικά των παρατηρήσεων και τις κλάσεις στις οποίες ανήκουν, ώστε να είναι σε θέση να προβλέψει την κλάση νέων κειμένων.

Παλινδρόμηση (Regression) Αφορά προβλήματα στα οποία τα αποτελέσματα των παρατηρήσεων είναι συνεχή. Ένα παράδειγμα είναι η πρόβλεψη του κόστους ασφάλισης ενός ατόμου. Έστω μία ασφαλιστική εταιρία, η οποία διαθέτει ένα σύνολο δεδομένων με το ιστορικό διάφορων πρώην πελατών της. Για κάθε έναν από τους πελάτες της γνωρίζει το συνολικό κόστος του σε έξοδα υγείας. Χρησιμοποιώντας τις παρατηρήσεις (ατομικά ιστορικά), μπορούμε να σχεδιάσουμε ένα μοντέλο παλινδρόμησης, το οποίο θα αντανακλά τη σχέση ανάμεσα στα χαρακτηριστικά των πελατών (ηλικία, ιατρικό ιστορικό, δημογραφικά στοιχεία) και στο τελικό κόστος. Με αυτό το μοντέλο, μπορούμε στη συνέχεια να εκτιμήσουμε το ασφαλιστικό κόστος ενός νέου πελάτη.

4.1.1.2 Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Στην *μη επιβλεπόμενη μάθηση* (unsupervised learning) *ή μάθηση χωρίς επιτήρηση*, δεν διαθέτουμε επισημασμένα δεδομένα. Στόχος είναι η ανακάλυψή δομών στα δεδομένα, χωρίς όμως να γνωρίζουμε αν και πόσες δομές υπάρχουν. Ένα από τα είδη μη επιβλεπόμενης μάθησης είναι η *συσταδοποίηση* (clustering). Ένα παράδειγμα είναι η ανακάλυψη ομάδων στις οποίες ανήκουν καταναλωτές με κοινή

συμπεριφορά, χρησιμοποιώντας ένα σύνολο δεδομένων με την καταναλωτική συμπεριφορά των πελατών ενός καταστήματος.

Ωστόσο, οι τεχνικές μη επιβλεπόμενης μάθησης δεν είναι πάντα αυτοσκοπός, καθώς μπορούν να χρησιμοποιηθούν ως μέσω για την δημιουργία χαρακτηριστικών για χρήση σε ένα άλλο πρόβλημα. Όπως παρουσιάστηκε και στο προηγούμενο κεφάλαιο μπορούν να χρησιμοποιηθούν για την δημιουργία κατανεμημένων αναπαραστάσεων για λέξεις (Κεφ. 3.2.3).

4.1.1.3 Ενισχυτική Μάθηση (Reinforcement Learning)

Στην *ενισχυτική μάθηση* (reinforcement learning), δεν διαθέτουμε επισημασμένα δεδομένα, όμως διαθέτουμε (ή δημιουργούμε) ένα μέτρο αξιολόγησης των αποτελεσμάτων. Πιο συγκεκριμένα, ένας αλγόριθμος ενισχυτικής μάθησης, μαθαίνει να εκτελεί μία εργασία μέσω ενός μηχανισμού επιβράβευσης, ο οποίος λειτουργεί με ανάδραση. Συνήθως υποθέτουμε ότι ο αλγόριθμος περιγράφει έναν ευφυή πράκτορα, ο οποίος προσπαθεί να μάθει να εκτελεί την ζητούμενη εργασία (π.χ. να περπατάει). Κάθε φορά που ο πράκτορας εκτελεί μία κίνηση ή κάνει μία πρόβλεψη, λαμβάνει μία "επιβράβευση" η οποία είναι μεγαλύτερη όσο καλύτερα εκτελεί την ζητούμενη εργασία. Στόχος του αλγορίθμου είναι μέσω της επανάληψης να μεγιστοποιήσει την ανταμοιβή που λαμβάνει, με αποτέλεσμα να μάθει να εκτελεί την ζητούμενη εργασία.

Η ενισχυτική μάθηση έχει αρκετές εφαρμογές στη ρομποτική. Ένα παράδειγμα είναι η διαδικασία κατά την οποία ένα ρομπότ μαθαίνει να περπατάει, η οποία είναι αντίστοιχη με αυτή που μαθαίνει ο άνθρωπος ή ένα άλλο ζώο. Δοκιμάζοντας διάφορες κινήσεις, ορισμένες είναι επιτυχημένες με αποτέλεσμα να ανταμείβεται, ενώ άλλες είναι αποτυχημένες με αποτέλεσμα να πέσει (αρνητική ανταμοιβή - πόνος). Ύστερα από ένα χρονικό διάστημα αλληλεπίδρασης με το περιβάλλον του, το ρομπότ μαθαίνει να περπατάει.

4.1.2 Η Διαδικασία της Μάθησης

Το πρόβλημα το οποίο θα μας απασχολήσει είναι η ταξινόμηση κειμένων. Όπως αναφέρθηκε ήδη, είναι ένα πρόβλημα επιβλεπόμενης μάθησης. Ο στόχος είναι η εκπαίδευση ενός μοντέλου, το οποίο θα κατηγοριοποιεί κείμενα, ανάλογα με τον συναισθηματικό προσανατολισμό τους (θετικός ή αρνητικός). Η είσοδος στον αλγόριθμο είναι:

- Ένα προκαθορισμένο σύνολο κλάσεων C, αποτελούμενο από K διαφορετικές κλάσεις.
$$C = \{c_1, c_2, ..., c_K\}$$
- Ένα σύνολο δεδομένων D, αποτελούμενο από N κείμενα για τα οποία είναι γνωστή η κλάση στην οποία ανήκουν.
$$D = \{(d_1, c_1), (d_2, c_2), (d_3, c_3), ..., (d_N, c_N)\}$$

Επίσης, κάθε κείμενο d_i αναπαριστάται από ένα σύνολο από F χαρακτηριστικά (features). Δηλαδή:
$$d_i = \{x_1, x_2, ..., x_F\}$$

Έτσι, το σύνολο δεδομένων αποτελείται από ζευγάρια από κείμενα $d_i \in R^F$, και την τιμή τους c_i (κλάση), όπου $i = 1 \dots N$ και $c_i \in 1 \dots K$. Θέλουμε να βρούμε μία συνάρτηση f η οποία θα χαρτογραφεί κείμενα σε κλάσεις, δηλαδή:

$$f_W : R^F \mapsto R^K$$

όπου W είναι τα βάρη ή οι παράμετροι της συνάρτησης.

Μέσω της διαδικασίας της εκπαίδευσης, θα γίνει προσαρμογή των παραμέτρων της συνάρτησης, ώστε να "ταιριάξει" στις παρατηρήσεις που διαθέτουμε. Η συνάρτηση αυτή θα είναι ο ταξινομητής ή το μοντέλο, με το οποίο θα μπορούμε στη συνέχεια να ταξινομήσουμε άγνωστα έγγραφα στις επιλεγμένες κλάσεις. Για την διαμόρφωση των παραμέτρων της συνάρτησης, απαιτούνται δύο διαδικασίες:

Συνάρτηση κόστους Η *συνάρτηση κόστους* (loss function) *ή αντικειμενική συνάρτηση* (objective function) αξιολογεί το πόσο καλά το μοντέλο ταιριάζει στα δεδομένα. Στόχος είναι η ελαχιστοποίηση του κόστους.

Ενημέρωση του Μοντέλου Χρησιμοποιώντας τη συνάρτηση κόστους, εφαρμόζουμε έναν μηχανισμό ο οποίος ενημερώνει το μοντέλο, αυξομειώνοντας τα βάρη της f_W.

Υπάρχουν πολλές συναρτήσεις κόστους, και αρκετοί μηχανισμοί για την ενημέρωση του μοντέλου. Ανάλογα με τον αλγόριθμο μηχανικής μάθησης και το πρόβλημα, αλλάζουν και οι διαθέσιμες επιλογές μας.

4.1.2.1 Προβλήματα Εκπαίδευσης

Υπάρχουν δύο προβλήματα τα οποία μπορούν να προκύψουν αναφορικά με την ικανότητα του μοντέλου να μοντελοποιήσει το πρόβλημα. Ιδανικά θέλουμε το μοντέλο, να μάθει τις δομές ή τις αρχές που διέπουν το πρόβλημα και να αγνοήσει τον θόρυβο. Τις περισσότερες φορές αυτό απαιτεί την εύρεση μίας λεπτής ισορροπίας, ανάμεσα σε δύο φαινόμενα.

Υποπροσαρμογή Η υποπροσαρμογή (underfitting) προκύπτει, όταν το μοντέλο δεν είναι αρκετά σύνθετο ώστε να περιγράψει το πρόβλημα. Σε αυτή την περίπτωση σημαίνει ότι το πρόβλημα είναι πιο περίπλοκο από το μοντέλο. Το πρόβλημα αυτό μπορεί να ξεπεραστεί με τους εξής τρόπους:

- Αύξηση της πολυπλοκότητας του μοντέλου. Για παράδειγμα στο πλαίσιο των ΤΝΔ, αυτό μπορεί να επιτευχθεί με την αύξηση των παραμέτρων των επιπέδων, ή του βάθους του δικτύου.
- Αύξηση της ευαισθησίας του μοντέλου. Αυτό επιτυγχάνεται με την επιλογή των κατάλληλων τιμών για ορισμένες από τις υπερπαραμέτρους του αλγορίθμου (παράμετρος C σε γραμμικά μοντέλα.)
- Εξαγωγή περισσότερων χαρακτηριστικών. Μπορεί η αδυναμία του αλγορίθμου να μοντελοποιήσει τα δεδομένα, να οφείλεται στο ότι δεν είναι αρκετά αντιπροσωπευτικός ο τρόπος με τον οποίο τα αναπαριστούμε. Εξάγοντας περισσότερα ή και πιο σύνθετα χαρακτηριστικά, επιτρέπουμε στον αλγόριθμο να ανακαλύψει πιο σύνθετες σχέσεις στα δεδομένα.

Υπερπροσαρμογή Η *υπερπροσαρμογή* (overfitting) είναι πιο συνηθισμένο πρόβλημα και προκύπτει όταν το μοντέλο είναι πάρα πολύ σύνθετο και είναι ικανό να ταιριάξει τέλεια στα δεδομένα εκπαίδευσης. Αυτό σημαίνει ότι έχει μάθει ακόμα και τον "θόρυβο" στα δεδομένα, δηλαδή ακόμη και τις μικρές ιδιαιτερότητές τους οι οποίες δεν αντιστοιχούν σε κάποια πραγματική δομή (δεν φέρουν χρήσιμη πληροφορία).

Το πρόβλημα τις υπερπροσαρμογής είναι πιο συχνό σήμερα, καθώς η διαθέσιμη υπολογιστική ισχύ επιτρέπει την εύκολη δημιουργία πολύ σύνθετων μοντέλων. Αν και είναι ένα πρόβλημα το οποίο μπορεί να προκύψει στους περισσότερους αλγορίθμους μηχανικής μάθησης, είναι αρκετά πιο έντονο στα ΤΝΔ, τα οποία συνήθως διαθέτουν εκατομμύρια παραμέτρους. Υπάρχουν διάφοροι τρόποι αντιμετώπισης ή περιορισμού του προβλήματος. Ορισμένες απλές λύσεις είναι:

- Αύξηση των δεδομένων εκπαίδευσης. Με αυτό τον τρόπο είναι πιο δύσκολο για το μοντέλο να ταιριάξει στο θόρυβο των δεδομένων, αλλά γίνονται και πιο ξεκάθαρες οι πραγματικές δομές στα δεδομένα. Είναι η καλύτερη λύση, αλλά και συχνά η πιο δύσκολη.

- Ενίσχυση των δεδομένων εκπαίδευσης (data augmentation). Με αυτό τον τρόπο προσθέτουμε στο σύνολο δεδομένων, τεχνητές παρατηρήσεις. Τις παρατηρήσεις αυτές μπορούμε να τις δημιουργήσουμε, κάνοντας μικρές τροποποιήσεις ή μετασχηματισμούς στις υπάρχουσες.

- Περιορισμός μοντέλου. Με αυτό τον τρόπο μειώνουμε τις παράμετροι του μοντέλου, μειώνοντας κατά συνέπεια τους βαθμούς ελευθερίας του μοντέλου. Έτσι, ο αλγόριθμος δεν έχει την ευκαιρία να μάθει σχέσεις, πέρα από τις πραγματικές σχέσεις στα δεδομένα.

- Μείωση χαρακτηριστικών. Με αυτό τον τρόπο απλοποιείται η αναπαράσταση των παρατηρήσεων. Αυτή πολλές φορές δεν είναι καλή λύση. Ο λόγος είναι ότι αν οι παρατηρήσεις δεν περιέχουν περιττά χαρακτηριστικά, τότε πετάμε χρήσιμη πληροφορία. Όπως θα δούμε και στη συνέχει υπάρχουν καλύτερες λύσεις.

Ένας άλλος τρόπος αντιμετώπισης του προβλήματος είναι με τεχνικές *εξομάλυνσης* (regularization). Η εξομάλυνση έχει πολλές μορφές ανάλογα με τον είδος του αλγορίθμου. Μία συνηθισμένη λύση, η οποία χρησιμοποιείται σε πολλούς αλγορίθμους, αφορά την αποθάρρυνση μεγάλων βαρών στο μοντέλο. Με τον τρόπο αυτό εμποδίζουμε την υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης, καθώς δεν επιτρέπουμε στο μοντέλο να δώσει υπερβολική σημασία σε συγκεκριμένα χαρακτηριστικά. Όπως θα δούμε και στη συνέχεια, κυρίως αναφορικά με τα ΤΝΔ, υπάρχουν και άλλες μορφές για να εξομαλύνουμε το μοντέλο μας.

4.2 Παραδοσιακές Τεχνικές

Σε αυτό το σημείο θα παρουσιαστούν οι βασικότερες προσεγγίσεις για ΑΣ. Επειδή το πεδίο γεννήθηκε μέσα από την Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ), όπως ήταν αναμενόμενο εφαρμόστηκαν τεχνικές οι οποίες ήταν δημοφιλείς σε τέτοια προβλήματα. Οι τεχνικές αυτές βασίζονται σε ένα σύνολο από χαρακτηριστικά

τα οποία εξάγονται από το κείμενο, τα οποία στη συνέχεια χρησιμοποιούνται από αλγορίθμους μηχανικής μάθησης, για την δημιουργία του αντίστοιχου στατιστικού μοντέλου.

Σαν πρώτο βήμα, οι τεχνικές αυτές απαιτούν τον σχεδιασμό των χαρακτηριστικών. Στόχος είναι τα χαρακτηριστικά αυτά να αντιπροσωπεύουν το συναίσθημα το οποίο περιέχει ένα κείμενο. Τα χαρακτηριστικά αυτά μπορεί να είναι από πολύ απλά (bag-of-words), έως πιο εξεζητημένα, τα οποία να βασίζονται σε θεωρητικά μοντέλα από επιστήμες όπως η Γλωσσολογία ή η Ψυχολογία. Σε αυτή την περίπτωση μπορεί να απαιτείται η συμμετοχή ειδικών στην διαδικασία.

4.2.1 Μηχανική Χαρακτηριστικών (Feature Engineering)

Στο προηγούμενο κεφάλαιο έγινε παρουσίαση των πιο συνηθισμένων χαρακτηριστικών τα οποία χρησιμοποιούνται σε προβλήματα ΕΦΓ. Σε αυτό το σημείο θα ασχοληθούμε με την διαδικασία επεξεργασίας αυτών των χαρακτηριστικών, για την επιλογή των πιο αντιπροσωπευτικών ή την δημιουργία νέων βελτιωμένων. Τα βήματα αυτά μπορούν να εφαρμοστούν και συνδυαστικά. Επίσης, στόχος είναι να ξεκαθαριστεί και η σημασία ορισμένων όρων, καθώς συχνά υπάρχουν παρανοήσεις, κυρίως λόγω της παραπλανητικής ονομασίας των τεχνικών.

4.2.1.1 Αποσαφήνιση Όρων

Σαν πρώτο βήμα είναι σημαντικό να ξεκαθαρίσουμε την σημασία των όρων. Στο πλαίσιο της παραγωγής των τελικών χαρακτηριστικών, τα οποία θα δοθούν σε έναν αλγόριθμο μηχανικής μάθησης, μεσολαβούν ορισμένα βήματα. Ένα τέτοιο βήμα είναι και αυτό της *εξαγωγής χαρακτηριστικών* (feature extraction). Ορισμένες φορές ο όρος αυτός χρησιμοποιείται για να υποδηλώσει την διαδικασία παραγωγής των αρχικών χαρακτηριστικών ή εναλλακτικά χρησιμοποιείται ως αναφορά στη συνολική διαδικασία κατά την οποία καταλήγουμε στα τελικά χαρακτηριστικά.

Άλλες φορές όμως, ο όρος αυτός χρησιμοποιείται ως αναφορά σε τεχνικές συνδυασμού χαρακτηριστικών, τα οποία έχουν ήδη εξαχθεί, για την παραγωγή νέων. Ένα τέτοιο παράδειγμα είναι με την εφαρμογή μείωσης διαστάσεων, όπου αφού έχουν εξαχθεί τα αρχικά αραιά διανύσματα, από μία τεχνική όπως bag-of-words (Κεφ. 3.1.1), εφαρμόζονται σε αυτά τεχνικές όπως γραμμικής μείωσης διαστάσεων (Κεφ. 3.2.1). Στόχος αυτών των τεχνικών είναι να καταλήξουμε σε ένα διανυσματικό χώρο πολύ λιγότερων διαστάσεων, όπου κάθε διάσταση (χαρακτηριστικό) είναι συνδυασμός των αρχικών. Με αυτό τον τρόπο συχνά καταλήγουμε με λιγότερα αλλά "καλύτερα" χαρακτηριστικά. Σε αυτή την διαδικασία αναφερόμαστε, σε αυτή την εργασία, με τον όρο *εξαγωγή χαρακτηριστικών* (feature extraction)

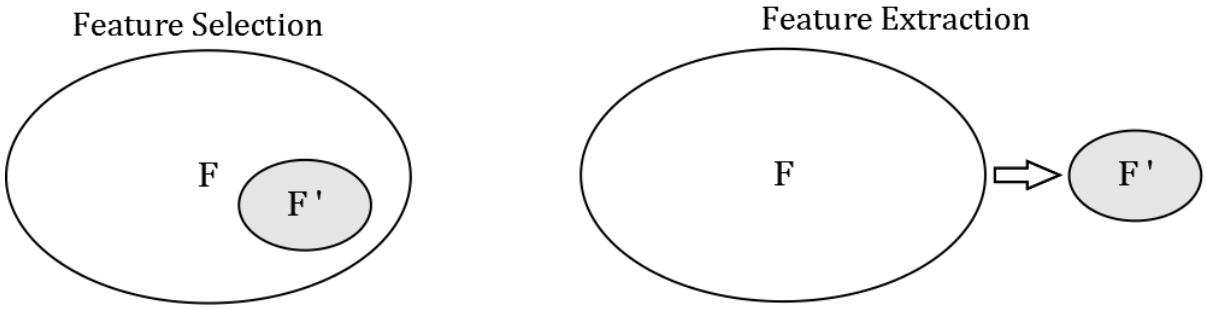
Ένα άλλο βήμα, αφορά την διαδικασία της *επιλογής χαρακτηριστικών* (feature selection). Στο βήμα αυτό επιλέγουμε ένα υποσύνολο των αρχικών χαρακτηριστικών, βάση κάποιου κριτηρίου χρησιμότητας. Ο λόγος είναι ότι συνήθως, ειδικά όταν έχουμε αραιά διανύσματα μεγάλων διαστάσεων, μεγάλο ποσοστό αυτών των χαρακτηριστικών δεν είναι αντιπροσωπευτικά του προβλήματος και εισάγουν θόρυβο στο μοντέλο. Έτσι, οι τεχνικές *επιλογής χαρακτηριστικών* στόχο έχουν να αποβάλλουν τα περιττά χαρακτηριστικά.

Οπότε, και στις δυο τεχνικές, οδηγούμαστε σε διανύσματα χαρακτηριστικών λιγότερων διαστάσεων. Αυτός είναι ένας από τους λόγους που συχνά *επιλογής*

χαρακτηριστικών και *εξαγωγής χαρακτηριστικών* συγχέονται. Επίσης αυτά τα βήματα είναι δυνατό να συνδυαστούν. Δηλαδή είναι δυνατό να επιλεχθούν τα χρήσιμα χαρακτηριστικά με την χρήση μίας τεχνικής *επιλογής χαρακτηριστικών* και στη συνέχεια να εξαχθούν *νέα* χαρακτηριστικά από αυτά με μία τεχνική *εξαγωγής χαρακτηριστικών*. Συνοψίζοντας:

1. Στη διαδικασία επιλογής χαρακτηριστικών (feature selection) διαλέγουμε, με κάποιο κριτήριο (score function), ένα *υποσύνολο* των αρχικών χαρακτηριστικών. Τα χαρακτηριστικά στα οποία καταλήγουμε, είναι χαρακτηριστικά και του αρχικού συνόλου χαρακτηριστικών.

2. Στη διαδικασία εξαγωγής χαρακτηριστικών (feature extraction), παράγονται νέα χαρακτηριστικά από τον συνδυασμό των αρχικών, μέσω της προβολής των παρατηρήσεων σε χαμηλότερες διαστάσεις. Λόγω της προβολής και της απώλειας πληροφορίας, δεν μπορούμε να ανακατασκευάσουμε τα αρχικά χαρακτηριστικά τα οποία οδήγησαν στην δημιουργία των νέων. Τα νέα χαρακτηριστικά *δεν* βρίσκονται και στο αρχικό σύνολο χαρακτηριστικών.



Σχήμα 4.1: Σύγκριση Feature Selection με Feature Extraction

4.2.1.2 Επιλογή Χαρακτηριστικών (Feature Selection)

Στο σημείο αυτό, θα δούμε ορισμένες από τις πιο δημοφιλείς τεχνικές επιλογής χαρακτηριστικών στο πλαίσιο των προβλημάτων επεξεργασία φυσικής γλώσσας (Y. Yang κ.ά. 1997; Forman 2003; Zheng κ.ά. 2004; Novovicova κ.ά. 2005; Y. Xu κ.ά. 2007). Οι τεχνικές αυτές εφαρμόζονται σε παρατηρήσεις στις οποίες χρησιμοποιούνται τοπικές αναπαραστάσεις (Κεφ. 3.1). Για τις τεχνικές που θα δούμε στη συνέχεια, θεωρούμε τα εξής:

- Οι παρατηρήσεις (κείμενα), περιγράφονται από ένα σύνολο από χαρακτηριστικά ή όρους T. Σε κάθε παρατήρηση εμφανίζονται ορισμένοι από τους όρους. Ένας όρος t μπορεί να είναι μία λέξη ή ακολουθία λέξεων (n-gram).

- Έχει προ-επιλεχθεί ένα σύνολο από κλάσεις C στις οποίες μπορεί να ανήκει μία παρατήρηση. Για παράδειγμα σε ένα πρόβλημα κατηγοριοποίησης εγγράφων βάση του συναισθηματικού τους προσανατολισμού, οι επιλεγμένες κλάσεις μπορεί να είναι $C = \{\textit{θετικός}, \textit{αρνητικός}, \textit{ουδέτερος}\}$.

Για κάθε χαρακτηριστικό, ο αλγόριθμος επιλογής χαρακτηριστικών, αποδίδει μία τιμή (score), η οποία αντιστοιχεί στην χρησιμότητά του. Αφού αξιολογηθεί κάθε χαρακτηριστικό, τότε επιλέγονται τα πιο χρήσιμα, με έναν από τους παρακάτω τρόπους:

- Τα N καλύτερα χαρακτηριστικά. Τα χαρακτηριστικά ταξινομούνται βάση της χρησιμότητάς τους και επιλέγονται τα N καλύτερα.
- Τα $N\%$ καλύτερα χαρακτηριστικά. Τα χαρακτηριστικά ταξινομούνται βάση της χρησιμότητάς τους και επιλέγεται το $N\%$ με τα καλύτερα.
- Κατώφλι. Επιλέγονται τα χαρακτηριστικά όπου η χρησιμότητά τους, είναι πάνω (ή κάτω) από ένα επιλεγμένο κατώφλι.

Πλήθος Εμφανίσεων Στην τεχνική αυτή, το η τιμή χρησιμότητας κάθε όρου, αντιστοιχεί στο πλήθος εμφανίσεων του όρου σε όλα τα έγγραφα του συνόλου δεδομένων. Εναλλακτικά, η τιμή χρησιμότητας αντιστοιχεί στο πλήθος των εγγράφων στα οποία εμφανίζεται ο όρος.

Συχνότητα Εμφανίσεων Στην τεχνική αυτή, η τιμή χρησιμότητας κάθε όρου, αντιστοιχεί στη συχνότητα (ποσοστό) των εμφανίσεων του όρου σε όλα τα έγγραφα του συνόλου δεδομένων. Εναλλακτικά, η τιμή χρησιμότητας αντιστοιχεί στο ποσοστό των εγγράφων στα οποία εμφανίζεται ο όρος.

TF-IDF Ο όρος TF-IDF είναι συντόμευση για το *term frequency–inverse document frequency* (συχνότητα όρου-αντίστροφη συχνότητα εγγράφων). Είναι ένα στατιστικό μέτρο το οποίο υπολογίζει πόσο σημαντικός είναι ένας όρος για ένα κείμενο, στο πλαίσιο ενός συνόλου κειμένων (corpus).

TF (Term frequency): Είναι ο λόγος του πλήθους των εμφανίσεων ενός όρου t σε ένα κείμενο, προς το πλήθος των λέξεων στο κείμενο.

TF-IDF: Είναι ο λογάριθμος του λόγου, του πλήθους των κειμένων $|D|$, προς το πλήθος των κειμένων στα οποία εμφανίζεται ο όρος.

Έτσι έχουμε ότι η $tf\text{-}idf$ τιμή ενός όρου t είναι:

$$tf(t,d) = \frac{count(t,d)}{\sum_k count(k,d)}, \qquad idf(t,d,D) = \log\frac{|D|}{|\{d : t \in d\}|}$$
$$tf\text{-}idf(t,d,D) = tf(t,d) * idf(t,d,D)$$

Σημείωση: Υπάρχουν διάφορες παραλλαγές για την αρίθμηση των εμφανίσεων ενός όρου, οι οποίες επηρεάζουν τη συμπεριφορά τόσο της τεχνικής του πλήθους εμφανίσεων, όσο και της TF-IDF. Για παράδειγμα, μία συνήθης πρακτική είναι η καταμέτρηση μόνο μίας φοράς ενός όρου σε ένα κείμενο.

Τεστ χ^2 (chi-squared) Αυτό είναι ένα στατιστικό τεστ το οποία μετρά την ανεξαρτησία ανάμεσα σε δύο γεγονότα. Συγκεκριμένα εξετάζει κατά πόσο η εμφάνιση ενός όρου t σε ένα κείμενο κλάσης c, είναι τυχαίο γεγονός ή όχι. Αν ένα χαρακτηριστικό δεν έχει ισχυρή εξάρτηση με καμία από τις κλάσεις, αυτό σημαίνει ότι δεν είναι χρήσιμο για το συγκεκριμένο πρόβλημα, άρα είναι περιττό.

Το τεστ χ^2, υπολογίζει για κάθε όρο t, πόσο έντονα σχετίζεται με κάθε κλάση c. Έτσι, για να υπολογίσουμε την ανεξαρτησία/εξάρτηση ανάμεσα σε έναν όρο t και μία κλάση c_i, σε ένα σύνολο από N κείμενα, κάνουμε:

$$\chi^2(t,c_i) = \frac{N[P(t,c_i)P(\bar{t},\bar{c}_i) - P(t,\bar{c}_i)P(\bar{t},c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \tag{4.1}$$

Τέλος, για να υπολογίσουμε το συνολική χρησιμότητα κάθε όρου t, αθροίζουμε τις χ^2 τιμές του σε κάθε κλάση:

$$\chi^2_{score}(t) = \sum_{i=1}^{|C|} \chi^2(t, c_i) \tag{4.2}$$

Αμοιβαία Πληροφορία (Mutual Information) Η τεχνική αυτή υπολογίζει την αμοιβαία πληροφορία μεταξύ ενός όρου t και μίας κλάσης c (Y. Xu κ.ά. 2007; Manning κ.ά. 2008). Το μέτρο αυτό υπολογίζει το μέγεθος της πληροφορίας που περιέχει ένα χαρακτηριστικό για κάθε κλάση. Ή αλλιώς, μετρά πόση πληροφορία συνεισφέρει η παρουσία ενός χαρακτηριστικού στην σωστή ταξινόμηση. Έτσι, για μία κλάση c και για έναν όρο t, με πιθανότητες $P(t)$, $P(c)$ αντίστοιχα, η αμοιβαία πληροφορία τους $MI(t, c)$ υπολογίζεται ως εξής:

$$MI(t, c) = \log \frac{t \wedge c}{P(t) \times P(c)} = \log \frac{P(t, c)}{P(t) \times P(c)} \tag{4.3}$$

όπου συγκρίνεται η πιθανότητα εμφάνισης των t και c μαζί, με τις πιθανότητες εμφάνισης των t και c ανεξάρτητα. Αν η κοινή πιθανότητα $P(t, c)$ είναι μεγαλύτερη από την τύχη $P(t)P(c)$, τότε υπάρχει σχέση ανάμεσα στον όρο t και στην κλάση c.

Όπως είναι εμφανές, η τιμή της $MI(t, c)$ μπορεί να είναι αρνητική σε ορισμένες περιπτώσεις, το οποίο δεν επιτρέπεται από τη θεωρία της πληροφορίας. Στην θεωρία της πληροφορίας ο όρος "αμοιβαία πληροφορία", αναφέρεται σε δύο τυχαίες μεταβλητές. Στο πλαίσιο όμως των προβλημάτων που μας αφορούν, ενδιαφερόμαστε για γεγονότα. Συνεπώς το παραπάνω μέτρο, ονομάζεται *Σημειακή Αμοιβαία Πληροφορία* (Pointwise Mutual Information - PMI).

Σημειακή Αμοιβαία Πληροφορία (Pointwise Mutual Information) Είναι η αμοιβαία πληροφορία ανάμεσα σε δύο γεγονότα. Για μία κατηγορία c και έναν όρο t, η σημειακή αμοιβαία πληροφορία τους $PMI(t, c)$ υπολογίζεται ως εξής:

$$PMI(t, c) = \log \frac{t \wedge c}{P(t) \times P(c)} \approx \log \frac{A \times N}{(A + C) \times (A + B)} \tag{4.4}$$

όπου:

- A είναι οι φορές, που ο όρος t και η κλάση c εμφανίζονται μαζί.
- B είναι οι φορές, που ο όρος t εμφανίζεται χωρίς την κλάση c.
- C είναι οι φορές, που η κλάση c εμφανίζεται χωρίς τον όρο t.
- N είναι το πλήθος των εγγράφων που ανήκουν στην κλάση c.

Για τον υπολογισμό του καθολικού βαθμού χρησιμότητας κάθε όρου, πρέπει να συγκεντρώσουμε τα αποτελέσματα $PMI(t, c)$ για κάθε κλάση. Συνήθως αυτό γίνεται, υπολογίζοντας τον μέσο όρο ή το άθροισμα των επιμέρους αποτελεσμάτων:

$$PMI_{avg}(t) = \sum_{i=1}^{|C|} P(c_i)PMI(t, c_i) \tag{4.5}$$

$$PMI_{max}(t) = \max_{i=1}^{|C|}\{PMI(t, c_i)\} \tag{4.6}$$

Information Gain Το μέτρο αυτό συνδέεται με την αμοιβαία πληροφορία. Όμως, είναι ένα μέτρο της πληροφορίας που κερδίζουμε, γνωρίζοντας την παρουσία ή την απουσία ενός όρου σε ένα έγγραφο.

$$\begin{aligned} IG(t) = &-\sum_{i=1}^{|C|} P(c_i)logP(c_i) \\ &+P(t)\sum_{i=1}^{|C|} P(c_i|t)logP(c_i|t) \\ &+P(\bar{t})\sum_{i=1}^{|C|} P(c_i|\bar{t})logP(c_i|\bar{t}) \end{aligned} \tag{4.7}$$

Διαφορά MI (PMI) με IG Όπως δείχνεται και στο (Y. Yang κ.ά. 1997) η IG είναι στην πραγματικότητα η *μέση (ή αναμενόμενη) αμοιβαία πληροφορία*. Είναι το σταθμισμένο άθροισμα των $PMI(t, c)$ και $PMI(\bar{t}, c)$, με συντελεστές τα $P(t|c)$ και $P(\bar{t}|c)$ αντίστοιχα.

$$IG(t) = \sum_{i=1}^{|C|} P(t, c_i)PMI(t, c_i) + \sum_{i=1}^{|C|} P(\bar{t}, c_i)PMI(\bar{t}, c_i) \tag{4.8}$$

Δύο σημαντικές διαφορές είναι:

- Η IG αξιοποιεί την πληροφορία της απουσίας ενός όρου, ενώ η PMI όχι.
- Η IG κανονικοποιεί τα PMI με τη χρήση των κοινών πιθανοτήτων $P(t, c_i)$ και $P(\bar{t}, c_i)$, ενώ η PMI όχι.

4.2.1.3 Εξαγωγή Χαρακτηριστικών (Feature Extraction)

Αυτή η μέθοδος αφορά τεχνικές οι οποίες δημιουργούν παράγωγα χαρακτηριστικά από τα αρχικά. Στόχος είναι η δημιουργία χαρακτηριστικών τα οποία δεν περιέχουν πλεονάζουσα πληροφορία, τα οποία αντικατοπτρίζουν συσχετίσεις ανάμεσα στα αρχικά χαρακτηριστικά. Για να επιτευχθεί αυτό χρησιμοποιούνται τεχνικές μείωσης διαστάσεων.

Οι πιο συνηθισμένες τεχνικές σε προβλήματα επεξεργασίας φυσικής γλώσσας είναι η *ανάλυση κύριων συνιστωσών* (PCA) και η *ανάλυση σε ιδιάζουσες τιμές* (SVD), οι οποίες όπως είδαμε και στο Κεφάλαιο 3.2.1 χρησιμοποιούνται για την παραγωγή κατανεμημένων αναπαραστάσεων. Η SVD, υπολογίζει μία προσέγγιση R διανυσμάτων (χαρακτηριστικών), ως τον γραμμικό συνδυασμό K διανυσμάτων

(χαρακτηριστικών). Η PCA, βρίσκει τις κύριες συνιστώσες ("κατευθύνσεις") στις οποίες υπάρχει η μέγιστη διακύμανση (πληροφορία). Οι δύο τεχνικές συνδέονται, καθώς ένας από τους δύο τρόπους εκτέλεσης της PCA, έχει σαν πρώτο βήμα την εκτέλεση SVD.

4.2.1.4 Στάθμιση Χαρακτηριστικών (Feature Weighting)

Αυτή η μέθοδος σταθμίζει τα χαρακτηριστικά βάση της σημαντικότητάς τους. Η λογική είναι αντίστοιχη με αυτή της επιλογής χαρακτηριστικών. Κάθε τεχνική *στάθμισης* χαρακτηριστικών, υπολογίζει την χρησιμότητα κάθε χαρακτηριστικού και στην συνέχεια σταθμίζει τις τιμές τους με την χρήση των κατάλληλων βαρών.

Για τον υπολογισμό της σημαντικότητας μπορούν να χρησιμοποιηθούν οι ίδιες τεχνικές που χρησιμοποιούνται για την επιλογή χαρακτηριστικών. Μόνο που σε αυτή την περίπτωση, αντί να επιλεχθεί ένα υποσύνολο των χαρακτηριστικών, αυξομειώνονται οι τιμές των υπαρχόντων. Το πιο συνηθισμένο παράδειγμα είναι η στάθμιση των τιμών των χαρακτηριστικών βάση της TF-IDF τιμής τους.

4.2.2 Αλγόριθμοι

Σε αυτό το σημείο θα παρουσιάσουμε ορισμένους δημοφιλείς αλγορίθμους μηχανικής μάθησης, για τη ΑΣ.

4.2.2.1 Naive Bayes

Ο Naive Bayes ("Αφελής" Μπέυζ) είναι ένας πολύ δημοφιλής ταξινομιτής σε προβλήματα επεξεργασίας φυσικής γλώσσας. Ένα από τα πρώτα προβλήματα στα οποία εφαρμόστηκε ήταν αυτό της αναγνώρισης μηνυμάτων ανεπιθύμητης ηλεκτρονικής αλληλογραφίας (spam) (Sahami κ.ά. 1998; Androutsopoulos κ.ά. 2000). Διακρίνεται για την απλότητα και την αποδοτικότητά του. Όπως υποδηλώνει και το όνομά της μεθόδου, βασίζεται στο θεώρημα πιθανότητας του Bayes.

$$P(c \mid d) = \frac{P(d \mid c)\, P(c)}{P(d)} \tag{4.9}$$

- $P(c)$, είναι η πιθανότητα εμφάνισης της κλάση c. Καλείται *εκ των προτέρων πιθανότητα* (prior probability) της c. Υπολογίζεται ως εξής: $P(c_i) = \frac{N_{c_i}}{N_c}$, όπου N_{c_i}, το πλήθος των παρατηρήσεων που ανήκουν στην κλάση c_i και N_c το πλήθος των κλάσεων.

- $P(d)$, είναι η πιθανότητα εμφάνισης του εγγράφου d. Καλείται *εκ των προτέρων πιθανότητα* (prior probability) του d.

- $P(c \mid d)$, είναι η πιθανότητα το έγγραφο d να ανήκει στην κλάση c. Καλείται *εκ των υστέρων πιθανότητα* (posterior probability) της c. Αυτό είναι το ζητούμενο.

- $P(d \mid c)$, είναι η πιθανότητα να παρατηρηθεί το έγγραφο d, με δεδομένο ότι ανήκει στην κλάση c. Αυτό καλείται *πιθανοφάνεια* (likelihood). Διαισθητικά, αυτή η τιμή είναι η απάντηση στην ερώτηση: "ποια είναι η πιθανότητα να παρατηρηθεί ένα κείμενο με τα χαρακτηριστικά (λέξεις) του κειμένου d, με δεδομένο ότι ανήκει στην κλάση c".

Σύμφωνα με τα παραπάνω, για να υπολογίσουμε σε ποια κλάση ανήκει ένα έγγραφο d το οποίο αποτελείται από ένα σύνολο από χαρακτηριστικά $d = \{x_1, x_2, ..., x_n\}$ (δηλαδή λέξεις $\{word_1, word_2, ..., word_n\}$), πρέπει να υπολογίσουμε την κλάση με τη *μέγιστη εκ των υστέρων πιθανότητα* (Maximum A-posteriori Probability – MAP):

$$
\begin{aligned}
c_{MAP} &= \underset{c \in C}{argmax}\ P(c \mid d) && \text{(η πιο πιθανή κλάση)} \\
&= \underset{c \in C}{argmax}\ \frac{P(d \mid c)\, P(c)}{P(d)} && \text{(θεώρημα Bayes)} \\
&= \underset{c \in C}{argmax}\ P(d \mid c)\, P(c) && \text{(απλοποίηση)} \\
&= \underset{c \in C}{argmax}\ P(x_1, x_2, ..., x_n \mid c)\, P(c) && \text{(Εγγραφο d = } \{x_1, x_2, ..., x_n\}\text{)}
\end{aligned}
$$

Το πρόβλημα είναι ο υπολογισμός του $P(x_1, x_2, ..., x_n \mid c)$. Για να υπολογιστεί, απαιτείται ο υπολογισμός όλων των συνδυασμών των επιμέρους δεσμευμένων πιθανοτήτων, το οποίο (1) είναι πάρα πολύ χρονοβόρο και (2) αυτό σημαίνει ότι χρειάζονται πολύ περισσότερα δεδομένα για την εκτίμηση των πιθανοτήτων. Κάνοντας όμως τις παραδοχές, (1) ότι η σειρά των λέξεων δεν έχει σημασία (bag-of-words) και (2) ότι οι τα ενδεχόμενα είναι ανεξάρτητα μεταξύ τους, μπορούμε να υπολογίσουμε την παράσταση ως απλά το γινόμενο των επιμέρους δεσμευμένων πιθανοτήτων. Αυτή η παραδοχή είναι ο λόγος που η τεχνική ονομάζεται "Αφελής" (Naive) Μπέυζ.

$$
P(x_1, x_2, ..., x_n \mid c) = P(x_1 \mid c) \times P(x_2 \mid c) \times ... \times P(x_n \mid c) \tag{4.10}
$$

Έτσι, για τον ταξινόμηση ενός εγγράφου κάνουμε:

$$
c_{MAP} = \underset{c \in C}{argmax}\ P(x_1, x_2, ..., x_n \mid c)\, P(c) \tag{4.11}
$$

$$
= \underset{c \in C}{argmax}\ P(c) \prod_{x \in X} P(x_i \mid c) \tag{4.12}
$$

Αυτό πρακτικά σημαίνει ότι η ποσότητα $P(x_i \mid c)$, μετράει πόσο ισχυρή ένδειξη είναι ο όρος t_i στην σωστή ταξινόμηση της κλάσης c.

4.2.2.2 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM), είναι μία οικογένεια αλγορίθμων επιτηρούμενης μάθησης, για την ταξινόμηση παρατηρήσεων σε δύο ή περισσότερες κλάσεις. Ο τρόπος με τον οποίο κατηγοριοποιούν τις παρατηρήσεις, είναι με την εύρεση ενός (ή και περισσότερων) υπερεπίπεδου (γενίκευση της ευθείας για διαστάσεις) στον χώρο των χαρακτηριστικών, το οποίο να τις διαχωρίζει γραμμικά. Επιπλέον, ένα SVM προσπαθεί να βρει εκείνο το υπερεπίπεδο, το οποίο θα έχει την μέγιστη δυνατή απόσταση από τις παρατηρήσεις κάθε κλάσης.

Έστω, ότι ο διανυσματικός χώρος είναι 2 διαστάσεων και θέλουμε να κατηγοριοποιήσουμε τις παρατηρήσεις σε δύο κλάσεις. Ένα πολύ απλό παράδειγμα τις λειτουργίας του αλγορίθμου είναι το εξής: Αρχικά βρίσκουμε μία οποιαδήποτε ευθεία η οποία να διαχωρίζει γραμμικά τα σημεία στις δύο κλάσεις. Στη συνέχεια

επιλέγονται δύο σημεία, ένα από κάθε κλάση, τα οποία έχουν την μικρότερη απόσταση από αυτή. Τα σημεία αυτά ονομάζονται διανύσματα υποστήριξης (support vectors), καθώς σε αυτά βασίζεται η εύρεση της ιδανικής ευθείας. Στη συνέχεια το SVM σχηματίζει μία ευθεία η οποία ενώνει αυτά τα δύο σημεία και επιλέγει ως την ιδανική ευθεία διαχωρισμού, την ευθεία που τέμνει κάθετα το μέσο της ευθείας που ενώνει τα δύο σημεία.

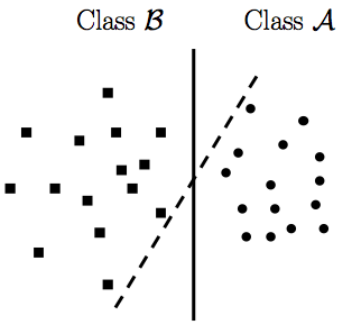


Figure 1. Which plane is best?

(α′) Ευθεία η οποία διαχωρίζει γραμμικά τις παρατηρήσεις

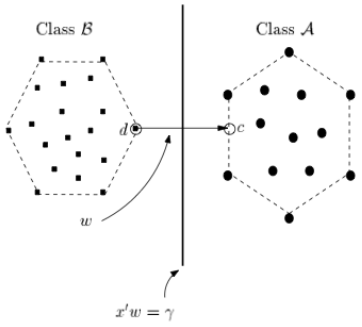


Figure 2. The two closest points of the convex hulls determine the separating plane.

(β′) Χρήση διανυσμάτων υποστήριξης

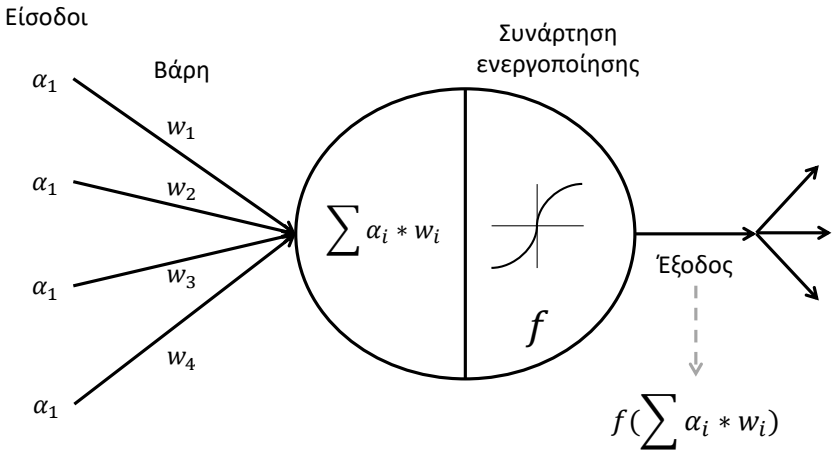
Σχήμα 4.2: Αρχικά βρίσκουμε μία ευθεία που να χωρίζει γραμμικά τις παρατηρήσεις και στη συνέχεια χρησιμοποιούμε τα διανύσματα υποστήριξης, για την εύρεση της βέλτιστης ευθείας. Εικόνα από (Bennett κ.ά. 2000)

Στην περίπτωση που οι παρατηρήσεις δεν είναι γραμμικά διαχωρίσιμες, χρησιμοποιείται μία τεχνική για προβολή των δεδομένων σε ένα διανυσματικό χώρο περισσότερων διαστάσεων. Η τεχνική αυτή ονομάζεται *kernel trick*. Όμως, σε προβλήματα κατηγοριοιποήσης κειμένων, όπως στη ΑΣ, τα διανύσματα χαρακτηριστικών είναι αραιά και μεγάλων διαστάσεων, με συνέπεια οι παρατηρήσεις να είναι γραμμικά διαχωρίσιμες.

4.3 Τεχνητά Νευρωνικά Δίκτυα

Τα ΤΝΔ είναι μία οικογένεια αλγορίθμων μηχανικής μάθησης, εμπνευσμένα από τον τρόπο λειτουργίας των νευρώνων του εγκεφάλου. Ο νευρώνας είναι ένα κύτταρο, το οποίο αποτελεί δομικό μέρος του εγκεφάλου και κατά συνέπεια του νευρικού συστήματος. Ένας νευρώνας συνδέεται με άλλους νευρώνες μέσω ειδικών συνδέσεων, τις *συνάψεις*. Κάθε νευρώνας μπορεί να έχει πολλές εισόδους, δηλαδή να λαμβάνει σήματα από πολλούς άλλους νευρώνες, αλλά έχει μία μόνο έξοδο. Όταν το άθροισμα των σημάτων ενός νευρώνα, ξεπεράσει ένα συγκεκριμένο κατώφλι, τότε ο νευρώνας ενεργοποιείται, βγάζοντας ένα σήμα εξόδου, το οποίο δίνεται ως είσοδο σε άλλους νευρώνες οι οποίοι είναι συνδεδεμένοι με αυτόν. Αυτό είναι ένα υπεραπλουστευμένο μοντέλο της λειτουργίας του ανθρώπινου νευρωνικού δικτύου. Όμως, αρκετές φορές μοντέλα εμπνευσμένα από τη βιολογία , ακόμα και πολύ απλοϊκά, μπορούν να φανούν χρήσιμα σε πολλά προβλήματα.

Στο πλαίσιο των ΤΝΔ, ο νευρώνας είναι μία υπολογιστική μονάδα. Ο πρώτος τέτοιος αλγόριθμος είναι ο Perceptron (Rosenblatt 1958), ο οποίος είναι ένας γραμ-



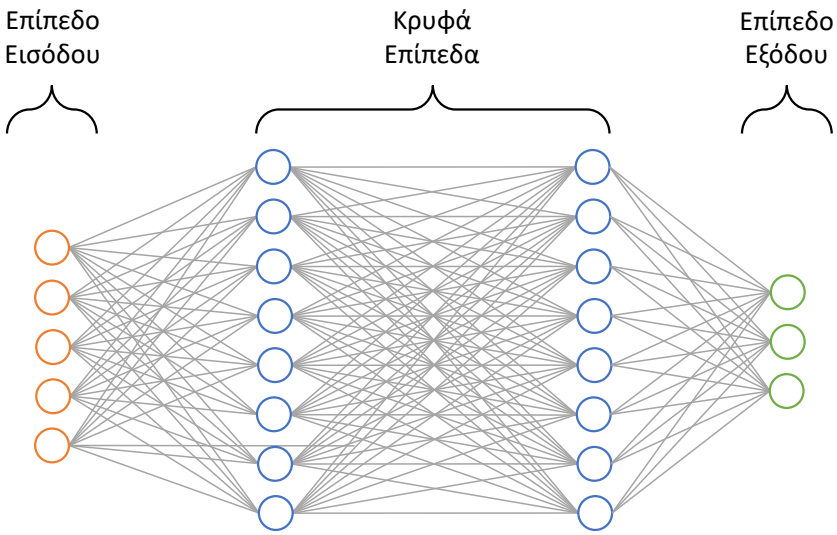
Σχήμα 4.3: Λειτουργία Τεχνητού Νευρώνα

μικός ταξινομητής. Ένας νευρώνας δέχεται ως είσοδο σήματα από τις συνάψεις (συνδέσεις εισόδου), ορίζοντας κάποια βάρη σε κάθε μία από αυτές. Τα σήματα σταθμισμένα ως προς τα βάρη, δίνονται ως είσοδο σε μία συνάρτηση ενεργοποίησης (activation function). Όταν η τιμή της συνάρτησης ενεργοποίησης πάρει τιμή πάνω από ένα ορισμένο κατώφλι, τότε βγάζει ως έξοδο την τιμή 1 (ο νευρώνας ενεργοποιείται). Ένας νευρώνας μπορεί να έχει διάφορες συναρτήσεις ενεργοποίησης που επηρεάζουν την συμπεριφορά του.

Με στόχο την μάθηση σύνθετων μη-γραμμικών συναρτήσεων, είναι δυνατό να σχεδιαστούν αρχιτεκτονικές που συνδυάζουν πολλούς νευρώνες, στοιχισμένους σε επίπεδα. Οι αρχιτεκτονικές αυτές είναι γνωστές και ως Multi-Layer Perceptron (MLP). Αντί για τον νευρώνα τύπου Perceptron, μπορεί να χρησιμοποιηθεί κάποια άλλη υπολογιστική μονάδα στους κόμβους του δικτύου. Μία συνήθης αρχιτεκτονική είναι αυτή του Feed-Forward Neural Network (FFNN), στο οποίο κάθε νευρώνας συνδέεται με όλους τους νευρώνες του προηγούμενου επιπέδου. Στην αρχιτεκτονική αυτή δεν υπάρχουν συνδέσεις μεταξύ των νευρώνων του ίδιου επιπέδου. Τα επίπεδα τα οποία μεσολαβούν ανάμεσα στα επίπεδα εισόδου και εξόδου, ονομάζονται κρυφά επίπεδα. Αν ένα δίκτυο έχει περισσότερα από ένα κρυφά επίπεδα, τότε το δίκτυο λέγεται ότι είναι *βαθύ*. Τα δίκτυα αυτά ονομάζονται Βαθιά Νευρωνικά Δίκτυα ή Deep Neural Networks (DNN). Αυξάνοντας το βάθος του δικτύου και με την χρήση μη-γραμμικών συναρτήσεων ενεργοποίησης το δίκτυο μπορεί να μάθει να εκτελεί πιο σύνθετες εργασίες (μαθαίνοντας πιο σύνθετες συναρτήσεις).

Ένα FFNN με τουλάχιστον ένα κρυφό επίπεδο και πεπερασμένο πλήθος νευρώνων, έχει αποδειχθεί ότι είναι ένας *καθολικός προσεγγιστής* (universal approximator) (Hornik κ.ά. 1989). Αυτό σημαίνει ότι μπορεί να προσεγγίσει οποιαδήποτε συνεχή συνάρτηση. Στην πράξη όμως, για δίκτυα με ένα μόνο κρυφό δίκτυο, αυτό δεν ισχύει για τα προβλήματα που μας ενδιαφέρουν. Για τον λόγο αυτό τα τελευταία χρόνια η επιστημονική κοινότητα έχει στραφεί στη δημιουργία βαθιών νευρωνικών δικτύων.

Στο πλαίσιο της εργασίας θα ασχοληθούμε με δύο σύνθετες αρχιτεκτονικές οι οποίες τα τελευταία χρόνια έχουν πετύχει πολύ καλά αποτελέσματα σε προβλήματα ΕΦΓ, με ένα χαρακτηριστικό παράδειγμα τα προβλήματα ΑΣ. Τα είδη δικτύων τα οποία θα εξετάσουμε είναι τα Convolutional Neural Networks (CNN) και



Σχήμα 4.4: Αρχιτεκτονική ενός δικτύου με δύο κρυφά επίπεδα.

Recurrent Neural Networks (RNN).

4.3.1 Εκπαίδευση

Η εκπαίδευση αφορά την διαδικασία προσαρμογής του μοντέλου στα δεδομένα. Αυτό επιτυγχάνεται με την επαναλαμβανόμενη ενημέρωση των βαρών του δικτύου. Σε αυτό το σημείο θα παρουσιαστούν οι βασικές διαδικασίες για την εκπαίδευση ενός ΤΝΔ. Οι διαδικασίες αυτές με μικρές διαφορές, είναι κοινές σε όλες τις αρχιτεκτονικές ΤΝΔ.

4.3.1.1 Συνάρτηση Κόστους

Η *συνάρτηση κόστους ή αντικειμενική συνάρτηση*, αξιολογεί τις επιδόσεις του μοντέλου. Δηλαδή, μετράει το πόσο καλά ένα σύνολο παραμέτρων W, μπορεί να προσεγγίσει τα αληθινά αποτελέσματα των παρατηρήσεων. Μπορούν να χρησιμοποιηθούν διάφορες συναρτήσεις, ανάλογα με τη φύση του προβλήματος. Το πρόβλημα το οποίο μας ενδιαφέρει είναι αυτό της ταξινόμησης κειμένων σε N κλάσεις. Η συνάρτηση υπολογίζει το μέσο σφάλμα του μοντέλου για όλες τις παρατηρήσεις στο σύνολο εκπαίδευσης. Αφού υπολογιστεί το σφάλμα, δηλαδή η απόκλιση του μοντέλου από τα πραγματικά δεδομένα, χρησιμοποιείται για την διόρθωση των τιμών των παραμέτρων του μοντέλου.

Σε ένα πρόβλημα ταξινόμησης με ΤΝΔ, στο επίπεδο εξόδου (τελευταίο) εφαρμόζουμε την συνάρτηση ενεργοποίησης $softmax$ (Εξίσωση (4.13)). Η συνάρτηση αυτή συμπιέζει τις τιμές ενός $K-$διάστατου διανύσματος z στο εύρος τιμών (0,1), όπου οι τιμές του αθροίζονται στο 1. Έτσι οι τιμές των κλάσεων κανονικοποιούνται και αυτό μας επιτρέπει να ερμηνεύσουμε τις τιμές αυτές σαν πιθανότητες για κάθε μία κλάση.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{4.13}$$

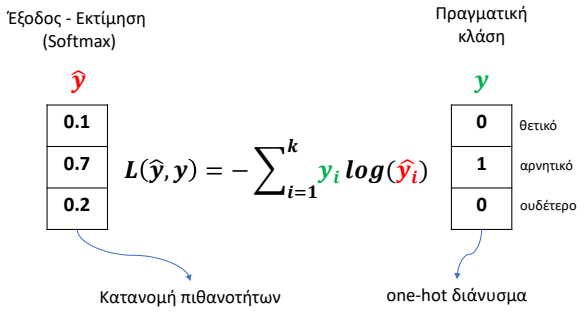
Η πιο συνηθισμένη συνάρτηση κόστους για προβλήματα ταξινόμησης είναι αυτή του *σφάλματος διεντροπίας* (cross-entropy loss). Η συνάρτηση αυτή υπολογίζει

την απόσταση δύο κατανομών πιθανοτήτων. Σε ένα πρόβλημα ταξινόμησης, όπως είδαμε ένα ΤΝΔ βγάζει ως έξοδο για κάθε μία παρατήρηση, μία κατανομή πιθανοτήτων στις κλάσεις του προβλήματος. Η κατανομή αυτή συγκρίνεται με το one-hot διάνυσμα το οποίο αντιστοιχεί στην πραγματική κλάση της παρατήρησης.

$$L_i = -\log\left(\frac{e^{y_i}}{\sum_j e^{f_j}}\right) \tag{4.14}$$

$$\mathcal{L}(W) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} y_{ij}\log(p_{ij}) \tag{4.15}$$

όπου W είναι όλα τα βάρη (παράμετροι) του δικτύου, i είναι οι δείκτες για τις παρατηρήσεις, j είναι οι δείκτες για τις κλάσεις, $\hat{y}$ είναι η εκτιμώμενη πιθανότητα και y η πραγματική πιθανότητα.

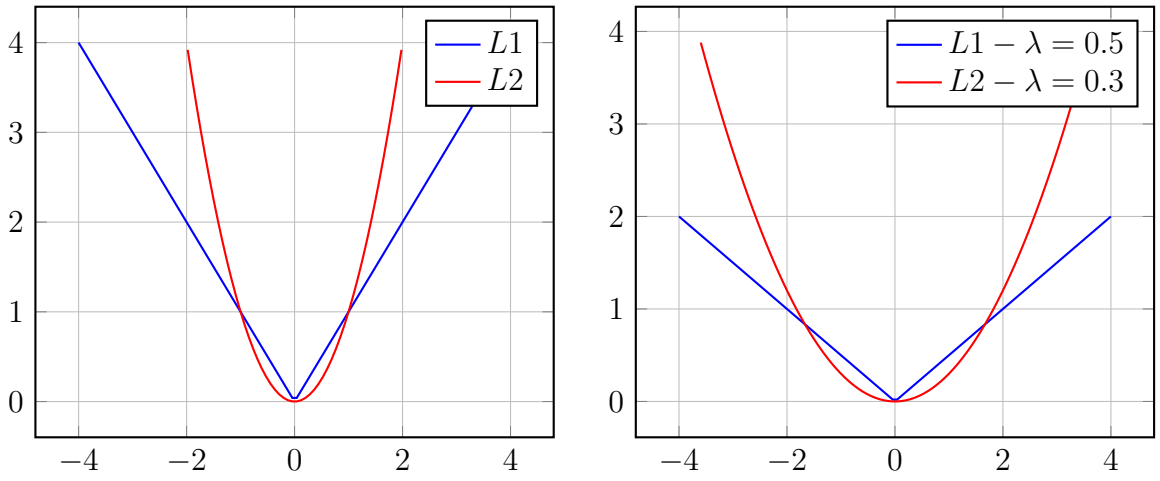


Σχήμα 4.5: Συνάρτηση κόστους διεντροπίας για μία παρατήρηση. Το γενικό κόστος του ΤΝΔ είναι ο μέσος όρος όλων των σφαλμάτων.

4.3.1.2 Εξομάλυνση

Η εξομάλυνση περιορίζει κατά κάποιο τρόπο την πολυπλοκότητα ενός μοντέλου, με στόχο να το εμποδίσει να μάθει τον θόρυβο στα δεδομένα. Υπάρχουν διάφοροι τρόποι για να επιτευχθεί αυτό. Εδώ θα εξετάσουμε τους πιο συνηθισμένους στα ΤΝΔ.

Αποσύνθεση Βαρών (Weight Decay) Αυτή η τεχνική δυσκολεύει τον σχηματισμό μεγάλων βαρών στο δίκτυο, εισάγοντας έναν μηχανισμό ο οποίος τα "τραβάει" προς το μηδέν. Για κάθε βάρος το οποίο θέλουμε να περιορίσουμε, προσθέτουμε έναν όρο, ο οποίος είναι ο συντελεστής εξομάλυνσής του. Τις περισσότερες φορές η εξομάλυνση εφαρμόζεται σε όλα τα βάρη του δικτύου, εισάγοντας τους όρους εξομάλυνσης κατευθείαν στην συνάρτηση κόστους. Όμως, είναι δυνατό να εφαρμόσουμε εξομάλυνση και σε συγκεκριμένα μόνο επίπεδα του δικτύου. Οι δύο όροι εξομάλυνσης οι οποίοι χρησιμοποιούνται πιο συχνά είναι οι $L1$ και $L2$. Έτσι, για τα βάρη w_i του δικτύου, έχουμε:



(α΄) Χωρίς συντελεστή εξομάλυνσης. (β΄) Με συντελεστή εξομάλυνσης.

Σχήμα 4.6: Γραφικές παραστάσεις $L1$ και $L2$ εξομάλυνσης. Όπως φαίνεται και στην εικόνα, μειώνοντας τον συντελεστή λ, μειώνουμε και την ένταση της εξομάλυνσης.

$$weight\ decay \quad L1 = \sum_{i}^{n} |w_i| \tag{4.16}$$

$$weight\ decay \quad L2 = \sum_{i}^{n} w_i^2 \tag{4.17}$$

L2 Αυτός ο όρος εξομάλυνσής, περιορίζει την τετραγωνική τιμή κάθε βάρους. Είναι η πιο δημοφιλής επιλογή. Διαισθητικά, η προσθήκη του $L2$ όρου, δυσκολεύει την δημιουργία μεγάλων βαρών στο δίκτυο. Αυτό σημαίνει ότι αποτρέπουμε το δίκτυο από το να δώσει υπερβολική βαρύτητα σε συγκεκριμένα χαρακτηριστικά. Επίσης η εισαγωγή $L2$ εξομάλυνσής κάνει πιο ομαλή την συνάρτηση κόστους, το οποίο βοηθάει στην βελτιστοποίηση της. Για κάθε βάρος w, ο $L2$ όρος είναι $\frac{1}{2}\lambda w^2$, όπου λ είναι ο συντελεστής έντασης της εξομάλυνσης και το $\frac{1}{2}$ το προσθέτουμε και την δημιουργία πιο απλών παραγώγων.

L1 Αυτός ο όρος εξομάλυνσής, περιορίζει την απόλυτη τιμή κάθε βάρους. Διαισθητικά, η προσθήκη του $L1$ όρου, αυξάνει την αραιότητα στα βάρη του δικτύου, καθώς πολλά από αυτά γίνονται σχεδόν μηδέν. Στην $L2$ εξομάλυνσή έχουμε πολλά μικρά βάρη, ενώ στην $L1$ εξομάλυνσή έχουμε ορισμένα μηδενικά βάρη. Για κάθε βάρος w, ο $L1$ όρος είναι $\lambda|w|$, όπου λ είναι ο συντελεστής έντασης της εξομάλυνσης.

Οι παραπάνω όροι μπορούν να εφαρμοστούν και συνδυαστικά. Έτσι, για την εφαρμογή $L1$ και $L2$ εξομάλυνσης κατευθείαν στην συνάρτηση κόστους, κάνουμε:

$$L = \underbrace{data}_{\text{σφάλμα δεδομένων}} + \underbrace{\sum\sum \lambda|w|}_{\text{L1 εξομάλυνση}} \tag{4.18}$$

$$L = \underbrace{data}_{\text{σφάλμα δεδομένων}} + \underbrace{\frac{1}{2}\sum\sum \lambda w^2}_{\text{L2 εξομάλυνση}} \tag{4.19}$$

4.3.1.3 Βελτιστοποίηση

Για την διόρθωση των βαρών στο δίκτυο, πρέπει να εφαρμοστεί μία διαδικασία βελτιστοποίησης. Η βελτιστοποίηση συναρτήσεων μπορεί να γίνει με πολλούς τρόπους. Στόχος της διαδικασίας είναι η εύρεση τιμών για τα βάρη του δικτύου, οι οποίες θα ελαχιστοποιούν το σφάλμα της συνάρτησης κόστους.

Μία δημοφιλής τεχνική είναι η Κατάβαση Κλίσης ή Gradient Descent (GD). Αυτή η τεχνική διορθώνει επαναληπτικά τα βάρη του δικτύου. Αρχικά υπολογίζουμε την μερική παράγωγο κάθε βάρους w_i ως προς τη συνάρτηση κόστους. Στη συνέχεια αυξομειώνουμε την τιμή του βάρους, ανάλογα με την κλίση του w_i, δηλαδή $\nabla_{w_i} L(w_i)$. Το μέγεθος της διόρθωσης (βήμα), ονομάζεται ρυθμός μάθησης και συμβολίζεται με η.

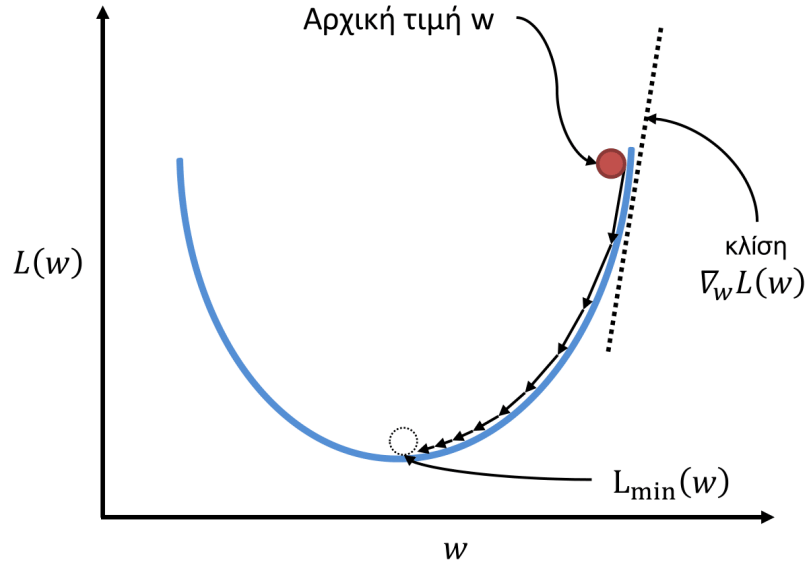
$$w = w - \eta \cdot \nabla_w L(w) \tag{4.20}$$

Στην πράξη, χρησιμοποιείται η Στοχαστική Κατάβαση Κλίσης ή Stochastic Gradient Descent (SGD), στην οποία τα βάρη ενημερώνονται όχι όλα μαζί κάθε φορά, αλλά ένα-ένα ή σε ομάδες. Ο λόγος που το κάνουμε αυτό είναι για να μειώσουμε τις απαιτήσεις σε μνήμη.

Εκτός από την απλή SGD, υπάρχουν και πιο εξεζητημένες παραλλαγές της, οι οποίες πετυχαίνουν ταχύτερη σύγκλιση, όπως με Nesterov Momentum (Nesterov 1983). Επιπλέον, τα τελευταία χρόνια χρησιμοποιούνται σε όλο και μεγαλύτερο βαθμό τεχνικές αυτόματης βελτιστοποίησης, στις οποίες γίνεται αυτόματη ρύθμιση του ρυθμού μάθησης, όπως Adagrad (Duchi κ.ά. 2011), Adadelta (Zeiler 2012) και Adam (Kingma κ.ά. 2014), η οποία είναι αυτή τη στιγμή η πιο δημοφιλής τεχνική.

4.3.1.4 Οπισθοδιάδοση (Backpropagation)

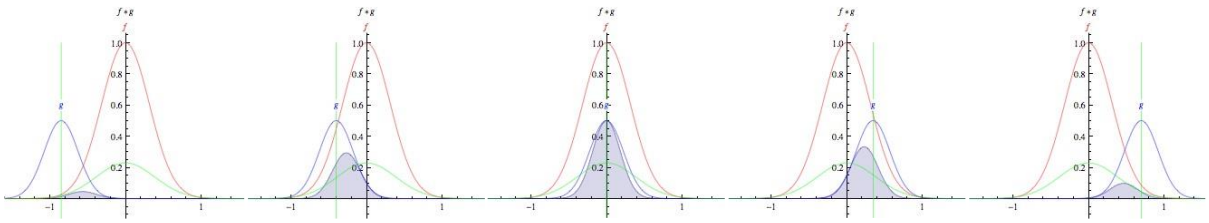
Ένα ΤΝΔ, μπορεί να αναπαρασταθεί ως ένας Κατευθυνόμενος Ακυκλικός Γράφος, όπου κάθε κόμβος αντιστοιχεί σε ένα βάρος του δικτύου. Για την ενημέρωση των βαρών του δικτύου χρησιμοποιείται η τεχνική της οπισθοδιάδοσης (backpropagation) (Rumelhart κ.ά. 1988) . Η τεχνική αυτή είναι ένας από τους σημαντικότερους λόγους για την ευρεία αποδοχή των ΤΝΔ, καθώς σε αντίθεση με άλλες τεχνικές, το υπολογιστικό κόστος για τον υπολογισμό των μερικών παραγώγων των βαρών του δικτύου, αυξάνεται γραμμικά σε σχέση με το μέγεθος του δικτύου. Με την χρήση της οπισθοδιάδοσης είναι δυνατό να εκπαιδεύσουμε πολύ μεγάλα δίκτυα.



Σχήμα 4.7: Συνάρτηση κόστους διεντροπίας για μία παρατήρηση. Το γενικό κόστος του ΤΝΔ είναι ο μέσος όρος όλων των σφαλμάτων.

4.3.2 Convolutional Neural Networks (CNN)

Τα CNN (Y. LeCun κ.ά. 1989) αφορούν αρχιτεκτονικές οι οποίες είναι εμπνευσμένες από τον τρόπο λειτουργίας του οπτικού συστήματος του εγκεφάλου. Ένα από τα πρώτα πετυχημένα CNN είναι το LeNet (Yann LeCun κ.ά. 1998) για αναγνώριση χαρακτήρων από κείμενα. Σχεδιάστηκαν στοχεύοντας προβλήματα υπολογιστικής όρασης, όπως αναγνώριση προσώπων, αντικειμένων ή χαρακτήρων σε εικόνες ή βίντεο. Τα CNN εφαρμόζουν μία σειρά από φίλτρα (ή μάσκες) σε μία εικόνα. Κάθε ένα από αυτά τα φίλτρα είναι ευαίσθητο σε συγκεκριμένα χαρακτηριστικά, όπως σχήματα ή υφές. Τα φίλτρα αυτά είναι μέρη του νευρωνικού δικτύου και το χαρακτηριστικό στο οποίο ανταποκρίνεται το καθένα, διαμορφώνεται κατά την εκπαίδευση του δικτύου. Συνήθως τα δίκτυα αυτά είναι αρκετά βαθιά, αποτελούμενα από μία ιεραρχία από φίλτρα, τα οποία μαθαίνουν προοδευτικά (αυξάνοντας το βάθος) πιο σύνθετα χαρακτηριστικά, όπως χαρακτηριστικά προσώπων (μάτια, μύτη) ή σχήματα (παράθυρα, έπιπλα).



Σχήμα 4.8: Συνέλιξη δύο συναρτήσεων $f * g$. Εικόνα από (Weisstein 2003)

Εκτελούν δύο βασικές πράξεις, τη *συνέλιξη* (convolution) και την *συγκέντρωση* (pooling ή sub-sampling). Η *συνέλιξη* είναι μία μαθηματική πράξη, η οποία εφαρμόζεται σε δύο συναρτήσεις και παράγει μία τρίτη, η οποία αντικατοπτρίζει, το πόσο επικαλύπτονται οι δύο αρχικές συναρτήσεις. Το CNN εκτελεί μία ακολουθία

βημάτων, η οποία μπορεί να επαναληφθεί πολλές φορές, προσθέτοντας βάθος στο δίκτυο:

1. Συνέλιξη (Convolution). Ένα σύνολο φίλτρων σύρονται πάνω από μία εικόνα, υπολογίζοντας την *συνέλιξη* του κάθε φίλτρου, με τα διάφορα τμήματα της εικόνας. Κάθε ένα φίλτρο παράγει μία νέα εικόνα, τον *χάρτη χαρακτηριστικών* (feature map).

2. Μη-γραμμική συνάρτηση ενεργοποίησης.Εφαρμογή μία μη-γραμμικής συνάρτησης ενεργοποίησης στον *χάρτη χαρακτηριστικών*.

3. Συγκέντρωση (pooling). Εφαρμογή μίας πράξης συγκέντρωσης των τιμών στον *χάρτη χαρακτηριστικών*.

Συνέλιξη (Convolution) Όπως ήδη αναφέρθηκε, αυτό το βήμα εφαρμόζει διάφορα φίλτρα σε μία εικόνα, παράγοντας μία νέα εικόνα. Τα φίλτρα αυτά αρχικοποιούνται τυχαία, σαν οποιοδήποτε επίπεδο ενός ΤΝΔ. Μέσω της εκπαίδευσης κάθε φίλτρο γίνεται ευαίσθητο σε ένα χαρακτηριστικό. Ένα επίπεδο CNN έχει πολλά φίλτρα και κάθε φίλτρο αντιστοιχεί σε ένα χαρακτηριστικό. Η διαδικασία είναι αρκετά απλή. Έστω ότι μία εικόνα αντιπροσωπεύεται ως μία μήτρα, όπου κάθε κελί της, αντιστοιχεί σε ένα εικονοστοιχείο (με τιμές 0-255). Ένα φίλτρο έχει διαστάσεις $N \times N$, όπου το N συνήθως είναι αρκετά μικρό (2-5). Το φίλτρο αυτό σύρεται πάνω από την εικόνα, εφαρμόζοντας έναν υπολογισμό σε κάθε τμήμα της εικόνας. Ο υπολογισμός είναι:

1. Πολλαπλασιασμός, στοιχείο προς στοιχείο, του φίλτρου με το αντίστοιχο τμήμα της εικόνας.

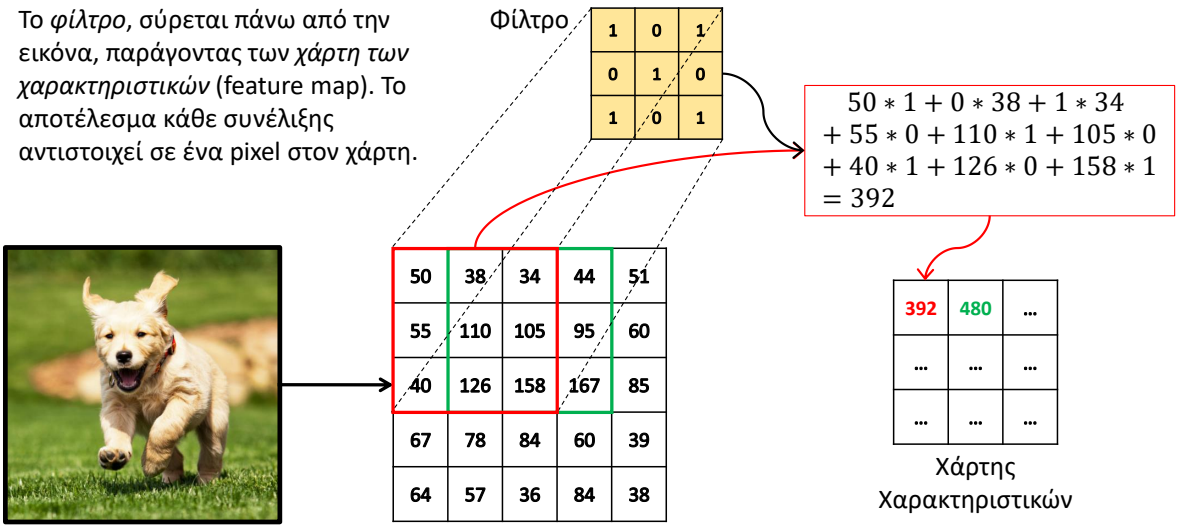
2. Άθροιση, των στοιχείων του νέου πίνακα και παραγωγή ενός μοναδικού στοιχείου.

Αφού ολοκληρωθεί αυτή η διαδικασία, παράγεται μία νέα εικόνα, η οποία ονομάζεται χάρτης χαρακτηριστικών (feature map). Η νέα εικόνα είναι μικρότερων διαστάσεων της αρχικής, καθώς κάθε φίλτρο παράγει ένα στοιχείο στη νέα εικόνα. Ο χάρτης αυτός είναι το αποτέλεσμα της συνέλιξης και έχει μεγάλες τιμές όταν στην εικόνα υπάρχει το χαρακτηριστικό (σχήμα ή υφή), στο οποίο είναι ευαίσθητο το αντίστοιχο φίλτρο.

Συνάρτηση Ενεργοποίησης Σε αυτό το βήμα, εφαρμόζεται μία μη-γραμμική συνάρτηση ενεργοποίησης, στον χάρτη χαρακτηριστικών. Αυτό επιτρέπει στο CNN να μάθει πιο σύνθετα χαρακτηριστικά. Σήμερα πλέον στα CNN χρησιμοποιείται κυρίως η ReLU $f(x) = \max(0, x)$, καθώς είναι φτηνή υπολογιστικά και κάνει πιο διακριτά τα χαρακτηριστικά. Άλλες δημοφιλείς συναρτήσεις για τα CNN είναι οι:

- Leaky ReLU, $f(x) = 1(x < 0)(\alpha x) + 1(x >= 0)(x)$.

- Maxout, $\max(w_1^T x + b_1, w_2^T x + b_2)$. Είναι μία γενική μορφή των δύο προηγούμενων συναρτήσεων, αλλά χρησιμοποιεί και περισσότερες παραμέτρους.



Σχήμα 4.9: Η διαδικασία της συνέλιξης σε ένα CNN. Στην περίπτωση που το CNN εφαρμόζεται σε μία εικόνα, δεν μας ενδιαφέρει οι τιμές του χάρτη να είναι συνεπείς με τις κανονικές εικόνες (0-255).

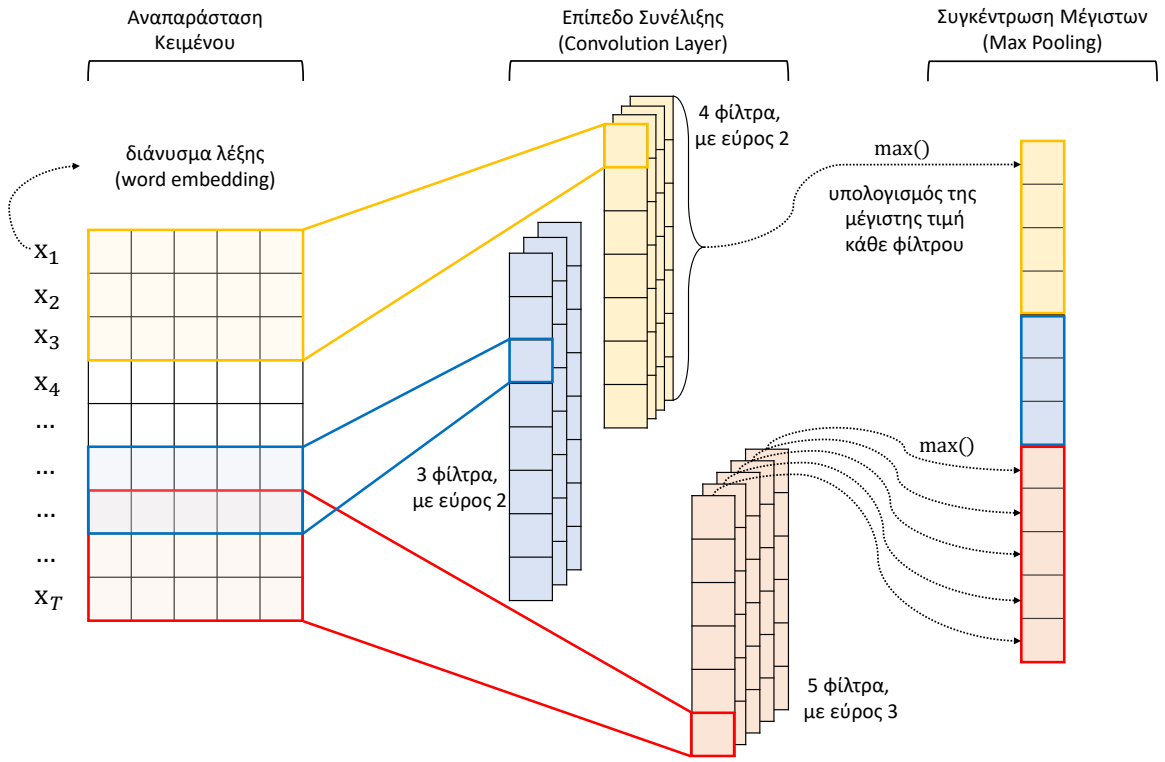
Συγκέντρωση (Pooling) Αυτό το βήμα μειώνει δραστικά τις διαστάσεις, διατηρώντας μόνο τα πιο σημαντικά χαρακτηριστικά. Η διαδικασία αυτή επιτρέπει στο δίκτυο να αναγνωρίζει ένα σχήμα, ακόμα και αν αυτό μεταβάλλεται, δηλαδή αν έχει κλίση, αν έχει μεγεθυνθεί κλπ. Ο λόγος που αυτό είναι σημαντικό, είναι γιατί μας νοιάζει αν ένα χαρακτηριστικό ή αντικείμενο, βρίσκεται στην εικόνα ή όχι, παρά τι μορφή έχει σε αυτή. Στο βήμα αυτό εφαρμόζεται μία πράξη ανά ορισμένα μικρά τμήματα του χάρτη χαρακτηριστικών και παράγεται ένα αποτέλεσμα, για κάθε τμήμα. Η συνηθέστερες πράξεις είναι:

- Μέγιστο (max). Επιλέγεται το μεγαλύτερο στοιχείο του εκάστοτε τμήματος.
- Μέσος όρος (mean). Επιλέγεται ο μέσος όρος των στοιχείων του εκάστοτε τμήματος.
- Άθροισμα (sum). Επιλέγεται το άθροισμα των στοιχείων του εκάστοτε τμήματος.

Συνήθως η πράξη που επιλέγεται είναι το *μέγιστο*, διότι όπως αναφέραμε μας ενδιαφέρει να αναγνωρίσουμε τα πιο σημαντικά χαρακτηριστικά.

4.3.2.1 CNN για Επεξεργασία Φυσικής Γλώσσας

Όταν εφαρμόζουμε ένα CNN σε προβλήματα επεξεργασίας φυσικής γλώσσας, πρέπει να προσαρμόσουμε την λειτουργία του. Αρχικά, πρέπει να αναπαραστήσουμε το κείμενο σαν μία εικόνα. Αυτό το κάνουμε, στοιβάζοντας τα διανύσματα κάθε λέξης, σχηματίζοντας μία μήτρα $M \times N$, όπου M το πλήθος των λέξεων και N οι διαστάσεις των διανυσμάτων των λέξεων. Επίσης, λαμβάνοντας υπόψη μας ότι κάθε σειρά πλέον αντιστοιχεί σε μία λέξη, ορίζουμε πάντα το πλάτος των φίλτρων να ισούται με το πλάτος της εικόνας. Το ύψος του φίλτρου ρυθμίζεται και καθορίζει πόσες λέξεις θα καλύπτει. Έτσι ένα φίλτρο είναι ένα παράθυρο W λέξεων, με διαστάσεις $W \times N$, όπου το W το ύψος ή εύρος του φίλτρου.



Σχήμα 4.10: Εφαρμογή ενός CNN σε πρόβλημα επεξεργασίας φυσικής γλώσσας. Το κείμενο αναπαριστάται ως ένας πίνακας, με στοιβαγμένα τα διανύσματα των λέξεων. Εφαρμόζουμε ένα επίπεδο *συνέλιξης* (convolution), με διάφορα φίλτρα, όπου κάθε ομάδα φίλτρων έχει διαφορετικό εύρος. Το εύρος κάθε φίλτρου αντιστοιχεί σε ένα n-gram. Κάθε φίλτρο μαθαίνει ένα χαρακτηριστικό. Στη συνέχεια εφαρμόζουμε ένα επίπεδο *συγκέντρωσης* (pooling), το οποίο μειώνει δραστικά τις διαστάσεις και διατηρεί τα πιο σημαντικά χαρακτηριστικά του κειμένου. Το διάνυσμα μετά την συγκέντρωση, αποτελεί την διανυσματική αναπαράσταση ολόκληρου του κειμένου.

CNN ως Neural N-grams Όπως είναι πλέον προφανές, ένα CNN στην ουσία μαθαίνει να αναγνωρίζει n-grams (Κεφάλαιο 3.1.2). Όμως, έχει το πλεονέκτημα ότι χρησιμοποιώντας διανύσματα λέξεων, αναγνωρίζει ακολουθίες εννοιών και όχι απλά λέξεων (όπως στις one-hot αναπαραστάσεις). Επίσης, το γεγονός ότι υπάρχει η δυνατότητα σχηματισμού βαθιών CNN, σημαίνει ότι το δίκτυο μπορεί να μάθει ακόμα πιο σύνθετα χαρακτηριστικά.

4.3.3 Recurrent Neural Networks (RNN)

Τα RNNs σε αντίθεση με άλλα δίκτυα έχουν την ιδιότητα ότι οι συνδέσεις τους σχηματίζουν κύκλους. Είναι συνδέσεις με ανάδραση (feedback). Επίσης έχουν δυναμική αρχιτεκτονική, το οποίο τους επιτρέπει να ξεπεράσουν ορισμένους από τους περιορισμούς άλλων δικτύων, όπως CNN ή FFNN. Ενδεικτικά:

- Μπορούν να επεξεργάζονται δεδομένα μεταβλητού μήκους. Τα άλλα δίκτυα απαιτούν εισόδους σταθερών διαστάσεων.

- Μπορούν να εκτελούν δυναμικό πλήθος υπολογιστικών βημάτων. Στα υπόλοιπα δίκτυα το πλήθος των βημάτων είναι συνήθως ανάλογο των επιπέδων τους.
- Έχουν "μνήμη", που τους επιτρέπει να ανακαλύπτουν εξαρτήσεις στα δεδομένα.

Ο τρόπος λειτουργίας τους μοιάζει με τον τρόπο που ο άνθρωπος επεξεργάζεται τις πληροφορίες, δηλαδή σειριακά. Αυτή η ευελιξία, τα κάνει ιδανικά για την επεξεργασία ακολουθιών, όπως σε προβλήματα επεξεργασίας φυσικής γλώσσας. Η βασική λειτουργία ενός RNN είναι εξαιρετικά απλή: δέχεται ως είσοδο ένα διάνυσμα x και παράγει ως έξοδο ένα διάνυσμα y. Όμως, η κρίσιμη διαφορά είναι ότι σε κάθε βήμα, για την παραγωγή του αποτελέσματος, λαμβάνεται υπόψη το αποτέλεσμα του προηγούμενου βήματος. Το απλό RNN, το οποίο εξετάζουμε τώρα, έχει ορισμένες παραλλαγές(Hopfield 1982; Elman 1990; Jordan 1997). Η παραλλαγή η οποία εξετάζουμε εδώ είναι το δίκτυο Elman (Elman 1990).

Λειτουργία. Πιο συγκεκριμένα ένα RNN, επεξεργάζεται τα στοιχεία μίας ακολουθίας σειριακά (ένα-προς-ένα), εκτελώντας τον *ίδιο* υπολογισμό, σε κάθε στοιχείο της ακολουθίας. Επίσης, διατηρεί μία εσωτερική κατάσταση, η οποία λειτουργεί σαν ένα είδος μνήμης. Στην απλούστερη μορφή αυτή η κατάσταση έχει τη μορφή ενός διανύσματος h, το οποίο ονομάζεται *κρυφή κατάσταση* (hidden state). Σε κάθε βήμα, το RNN ενημερώνει το h, λαμβάνοντας υπόψη του την τιμή του τρέχοντος στοιχείου x_t και της προηγούμενης τιμής του h. Έτσι για κάθε χρονική στιγμή t, έχουμε:

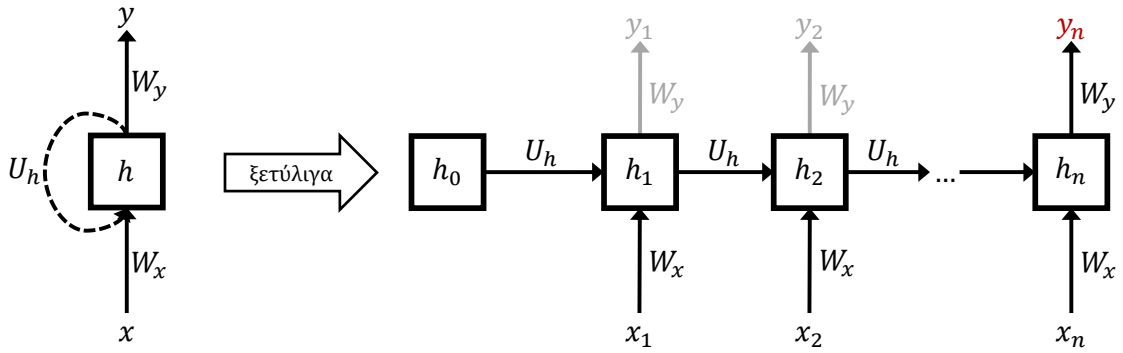
$$h_t = f_h(W_h x_t + U_h h_{t-1} + b_h) \tag{4.21}$$
$$y_t = f_y(W_y h_t + b_y) \tag{4.22}$$

όπου

- h_t η κρυφή κατάσταση (ή κρυφό διάνυσμα), τη χρονική στιγμή t.
- x_t το διάνυσμα του στοιχείου της ακολουθίας τη χρονική στιγμή t.
- y_t η έξοδος (διάνυσμα) τη χρονική στιγμή t.
- W_x, U_h, W_y τα βάρη[1] του δικτύου για τα h, x και y αντίστοιχα.
- b_h το συστηματικό σφάλμα (bias) για το h.
- f_x, f_h οι συναρτήσεις ενεργοποίησης. Συνήθως χρησιμοποιείται η ίδια συνάρτηση για h και y.

Αρχικά, η συνηθέστερη συνάρτηση ενεργοποίησης ήταν η *σιγμοειδής* συνάρτηση (Sigmoid Function), όμως πλέον θεωρείται προτιμότερη η *υπερβολική εφαπτομένη* (tanh). Η σιγμοειδής συνάρτηση συμπιέζει τις τιμές ενός διανύσματος, στο εύρος $[0, 1]$ ενώ η υπερβολική εφαπτομένη (tanh) συμπιέζει τις τιμές ενός διανύσματος, στο εύρος $[-1, 1]$. Το διαφορετικό πεδίο τιμών τους επηρεάζει την συμπεριφορά του δικτύου κατά την εκπαίδευση με οπισθοδιάδοση.

[1] Συνήθως χρησιμοποιείται το W για να δηλώσει τα βάρη στην κανονικές συνδέσεις ενός δικτύου και το U για να δηλώσει τα βάρη σε αναδρομικές συνδέσεις.



Σχήμα 4.11: Διάγραμμα λειτουργίας ενός RNN. Υπάρχουν δύο τρόποι ώστε να σκέφτεται κανείς το πως λειτουργεί ένα RNN. Στην αριστερή εικόνα φαίνεται η αναδρομική λειτουργία του RNN. Το RNN δέχεται το ένα μετά το άλλο τα στοιχεία της ακολουθίας και ενημερώνει την εσωτερική ή κρυφή του κατάσταση. Ένας άλλος τρόπος αναπαράστασης είναι με το "ξετύλιγμα" του δικτύου στο χρόνο. Ουσιαστικά ένα RNN είναι ένα FFNN με ανάδραση. Στην ξετυλιγμένη μορφή, ένα RNN μοιάζει με ένα πολυεπίπεδο FFNN.

Η Εικόνα 4.11 ξεκαθαρίζει τον τρόπο λειτουργίας ενός RNN. Όπως φαίνεται και από το σχήμα, ένα RNN μπορεί να δεχθεί μία ακολουθία οποιουδήποτε μήκους. Αφού επεξεργαστεί ένα-προς-ένα τα στοιχεία, παράγει το τελικό αποτέλεσμα y_n, το οποίο είναι μία *σταθερή διανυσματική αναπαράσταση* για όλη την ακολουθία. Σε αυτή την περίπτωση, εκτελούνται μόνο οι ενημερώσεις της κρυφής κατάστασης (Εξίσωση (4.21)) και στο τέλος παράγεται η έξοδος (Εξίσωση (4.22)).

Παράδειγμα Ας δούμε ένα συνηθισμένο παράδειγμα εφαρμογής ενός RNN στην επεξεργασία φυσικής γλώσσας. Σε ένα πρόβλημα κατηγοριοποίησης κειμένου, το RNN επεξεργάζεται τις λέξεις του εγγράφου, τη μία μετά την άλλη, και στο τέλος παράγει την *διανυσματική αναπαράσταση του εγγράφου* y_n. Η αναπαράσταση αυτή χρησιμοποιείται σαν διάνυσμα χαρακτηριστικών για την κατηγοριοποίηση του κειμένου, βάση του συναισθηματικού του προσανατολισμού.

Πιο συγκεκριμένα, το έγγραφο αναπαριστάται από μία ακολουθία λέξεων. Κάθε λέξη αναπαριστάται από ένα διάνυσμα (word embedding) x_i, με $x_i \in R^E$, όπου E οι διαστάσεις των διανυσμάτων λέξεων. Έτσι έχουμε την ακολουθία $X = (x_1, x_2, ..., x_T)$, όπου T το πλήθος των λέξεων στο έγγραφο. Το RNN επεξεργάζεται σειριακά τις λέξεις, διατηρώντας στο εσωτερικό του, μία περίληψη όσων έχει διαβάσει μέχρι τη χρονική στιγμή t. Στο τέλος, περιέχει την περίληψη όλης της πληροφορίας του εγγράφου και από αυτή παράγει την τελική διανυσματική αναπαράσταση για το έγγραφο.

4.3.3.1 Αμφίδρομο RNN

Ένα αμφίδρομο RNN (bidirectional RNN ή BiRNN) αποτελείται από τον συνδυασμό δύο διαφορετικών RNN, όπου το κάθε ένα επεξεργάζεται την ακολουθία με διαφορετική φορά. Το κίνητρο αυτής της τεχνικής, είναι η δημιουργία μία περίληψης του εγγράφου και από τις δύο κατευθύνσεις, ώστε να σχηματιστεί μία καλύτερη αναπαράσταση. Έτσι έχουμε ένα δεξιόστροφο RNN $\overrightarrow{f}$, το οποίο διαβά-

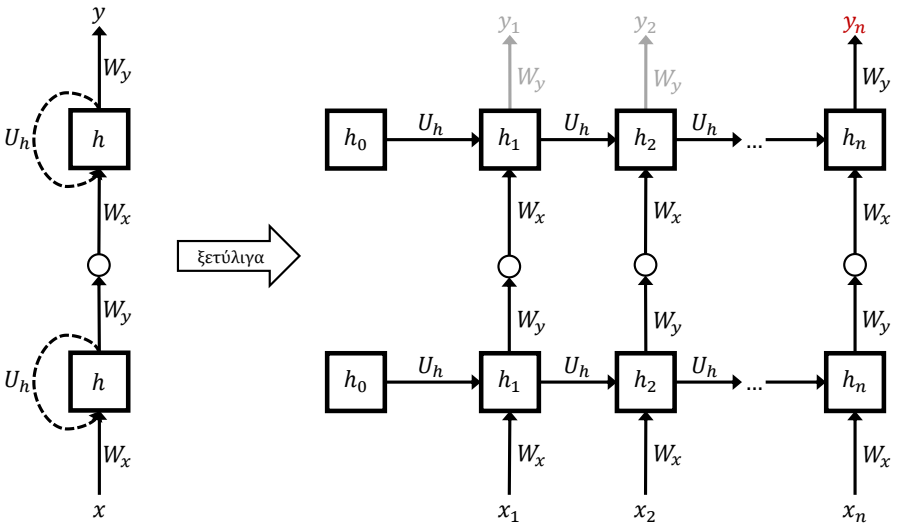
ζει μία πρόταση από το x_1 προς x_T και ένα αριστερόστροφο RNN $\overleftarrow{f}$, το οποίο διαβάζει μία πρόταση από το x_T προς x_1. Έτσι, κάθε χρονική στιγμή t, έχουμε:

$$h_i = \overrightarrow{h_i} \parallel \overleftarrow{h_i}, \quad h_i \in R^{2N} \tag{4.23}$$

όπου το $\parallel$ συμβολίζει την πράξη της ένωσης δύο διανυσμάτων και N είναι οι διαστάσεις του κάθε RNN.

4.3.3.2 Βαθιά RNN

Όπως και με τα απλά FFNN μπορούμε να στοιχίσουμε ένα RNN σε επίπεδα για την δημιουργία βαθιών δικτύων. Όπως έχει δειχθεί και στο (Karpathy κ.ά. 2015), όσο περισσότερα επίπεδα, τόσο το καλύτερο, σε ότι αφορά την ακρίβεια του δικτύου. Όμως από τα 3 επίπεδα και πάνω, τα κέρδη είναι σχεδόν αμελητέα.

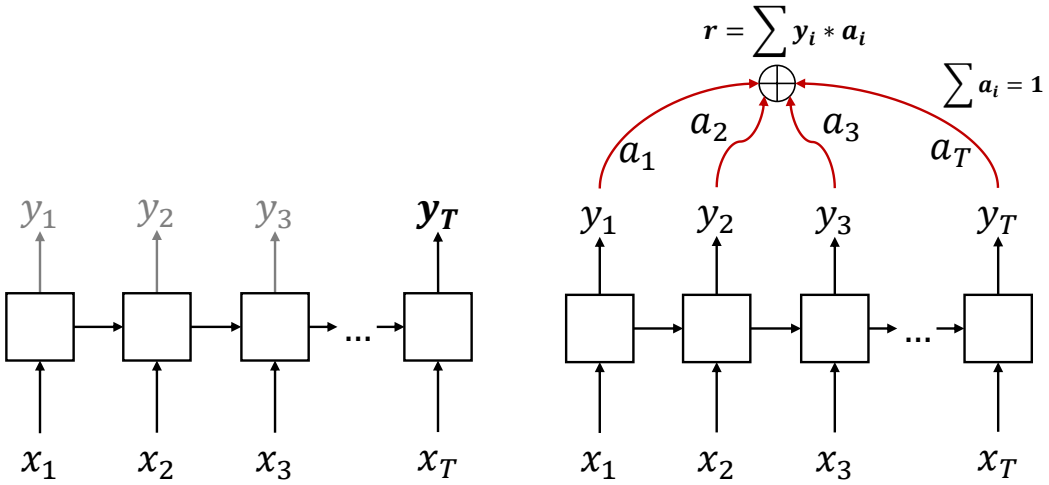


Σχήμα 4.12: Ένα βαθύ RNN με δύο επίπεδα.

4.3.3.3 Μηχανισμός Προσοχής

Όπως είδαμε και παραπάνω, ένα RNN χρησιμοποιεί την τελευταία τιμή της εσωτερικής του κατάστασης ως την διανυσματική αναπαράσταση όλης της ακολουθίας. Η αναπαράσταση αυτή ενημερώνεται καθώς το RNN διαβάζει την ακολουθία και στο τέλος περιέχει μία σύνοψη της ακολουθίας. Όμως, ειδικά όταν η ακολουθία είναι αρκετά μεγάλη, υπάρχει η περίπτωση το δίκτυο να μην μπορεί να συγκρατήσει όλες τις σημαντικές πληροφορίες στην εσωτερική του κατάσταση. Για να αντιμετωπίσουμε αυτό το πρόβλημα, μπορούμε να χρησιμοποιήσουμε μία τεχνική η οποία επιχειρεί να ενισχύσει την συνεισφορά των σημαντικών στοιχείων στην τελική αναπαράσταση.

Η τεχνική αυτή εφαρμόζει ένα μηχανισμό *προσοχής* (attention), ο οποίος δίνει μεγαλύτερη βαρύτητα (προσοχή) στα σημαντικά στοιχεία της ακολουθίας. Αυτό το πετυχαίνει αξιοποιώντας όλες τις ενδιάμεσες καταστάσεις του RNN. Έτσι κάθε χρονική στιγμή t, η εσωτερική κατάσταση του RNN y_i, χρησιμοποιείται ως την ερμηνεία της λέξης x_i. Για την παραγωγή της διανυσματικής αναπαράστασης ολόκληρου του



(α′) Απλό RNN. Η τελική αναπαράσταση της ακολουθίας είναι η τελευταία έξοδος του RNN y_n.

(β′) RNN με μηχανισμό attention. Η τελική αναπαράσταση της ακολουθίας είναι το σταθμισμένο άθροισμα των ενδιάμεσων εξόδων y_i.

Σχήμα 4.13: Σύγκριση μεταξύ του απλού RNN και ενός RNN με attention. Στο RNN με attention, Η τελική αναπαράσταση r είναι το σταθμισμένο άθροισμα όλων των εξόδων του RNN. Τα βάρη κάθε βήματος a_i, ορίζονται από το attention layer.

κειμένου, χρησιμοποιούμε το άθροισμα των ερμηνειών των λέξεων, σταθμισμένο ως προς την σημαντικότητά τους (Graves 2013; Bahdanau κ.ά. 2014).

Ο μηχανισμός προσοχής, αποτελεί ένα επίπεδο του δικτύου, το οποίο εκπαιδεύεται μαζί με τα υπόλοιπα και μαθαίνει να αποδίδει μεγαλύτερη βαρύτητα στις σημαντικές λέξεις, για το επίμαχο πρόβλημα. Έτσι, όταν η ακολουθία είναι ένα κείμενο, αυτό σημαίνει ότι ο μηχανισμός δίνει μεγαλύτερη προσοχή στις πιο σημαντικές λέξεις του κειμένου. Στο πλαίσιο της Συναισθηματικής Ανάλυσης, μαθαίνει να αποδίδει μεγαλύτερα βάρη στις λέξεις οι οποίες είναι καθοριστικές για τον προσδιορισμό του συναισθήματος σε ένα κείμενο.

Ο μηχανισμός προσοχής δεν είναι μία συγκεκριμένη τεχνική αλλά μία γενικότερη προσέγγιση. Χρησιμοποιείται όταν σε μία είσοδο, η οποία μπορεί να είναι ένα κείμενο ή μία εικόνα, θέλουμε μία δεδομένη χρονική στιγμή να δώσουμε μεγαλύτερη έμφαση σε διαφορετικά σημεία της και όχι να την λάβουμε υπόψη μας ως σύνολο. Για παράδειγμα, μπορεί στο πρόβλημα της παραγωγής περιγραφών για εικόνες (K. Xu κ.ά. 2015), να θέλουμε να δώσουμε έμφαση σε διαφορετικά σημεία μίας εικόνας, ανάλογα με την έννοια που θέλουμε να περιγράψουμε. Ομοίως, στο πλαίσιο της μηχανικής μετάφρασης (Bahdanau κ.ά. 2014), όταν παράγουμε το μεταφρασμένο κείμενο, θέλουμε μετά από κάθε λέξη να δώσουμε έμφαση σε διαφορετικές λέξεις του πηγαίου κειμένου.

4.3.3.4 LSTM

Το Long short-term memory (LSTM) δίκτυο (Hochreiter κ.ά. 1997; Gers κ.ά. 2002), είναι μία παραλλαγή του RNN. Εισάγει έναν εξεζητημένο μηχανισμό, ο οποίος του επιτρέπει να ξεπεράσει το πρόβλημα του RNN, σχετικά με την αναγνώριση απομακρυσμένων εξαρτήσεων. Το LSTM έχει δύο βασικές διαφορές από το απλό RNN:

- *Δεν* εφαρμόζει συνάρτηση ενεργοποίησης στις αναδρομικές συνδέσεις. Αυτό σημαίνει ότι οι ενημερώσεις θα είναι γραμμικές. Έτσι εγγυάται ότι τα σφάλματα, δεν θα εξαφανίζονται από την επαναληπτική εφαρμογή των ενημερώσεων. Συνεπώς είναι σίγουρο ότι θα υπάρχει σωστή ροή της πληροφορίας στο δίκτυο.

- Μηχανισμός με θύρες. Ο μηχανισμός αυτός, εισάγει ορισμένες θύρες, οι οποίες ρυθμίζουν το πόσο θα ενημερώνεται κάθε διάνυσμα του δικτύου (εσωτερική κατάσταση, έξοδος κλπ.). Με αυτό τον τρόπο το δίκτυο αφομοιώνει και διατηρεί τις πιο σημαντικές πληροφορίες.

Ο μηχανισμός με τις θύρες είναι μια σημαντική ιδέα, η οποία προσθέτει μεγάλη ευελιξία στο δίκτυο. Με τον τρόπο αυτό έχει την δυνατότητα να ρυθμίζει την ροή της πληροφορίας. Επίσης, επειδή όλος ο μηχανισμός είναι παραγωγίσιμος, σημαίνει ότι το δίκτυο μπορεί να *μάθει* να ρυθμίζει την λειτουργία των θυρών, με τεχνικές όπως το GD. Όπως και το RNN, έχει αρκετές παραλλαγές, οι οποίες κυρίως αφορούν τον τρόπο με τον οποίο ενημερώνεται η μνήμη του (cell state). Για αυτό το λόγο είναι σημαντικό να δίνεται προσοχή στον μαθηματικό ορισμό σε κάθε εργασία που το χρησιμοποιεί. Αναλυτικές καταγραφές και συγκρίσεις των διαφόρων παραλλαγών παρουσιάζονται στα (Greff κ.ά. 2015; Lipton κ.ά. 2015).

Ορισμός Αυτή είναι ίσως η πιο συνηθισμένη παραλλαγή του LSTM. Οι αρχικές τιμές είναι: $c_0 = 0$ και $h_0 = 0$. Όλα τα διανύσματα έχουν τις ίδιες διαστάσεις. Η πράξη $\circ$ δηλώνει το γινόμενο Hadamard (στοιχείο προς στοιχείο).

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) && \text{θύρα λήθης (forget)} \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) && \text{θύρα εισόδου (input)} \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) && \text{θύρα εξόδου (output)}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c) && \text{υποψήφια μνήμη (candidate cell state)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{μνήμη (cell state)} \\
h_t &= o_t \circ \sigma_h(c_t) && \text{έξοδος (output)}
\end{aligned}
$$

- Τα διανύσματα i, f, o, είναι οι θύρες εισόδου (input), λήθης (forget) και εξόδου (output) αντίστοιχα. Ο τρόπος ενημέρωσης του καθενός είναι πανομοιότυπος, μόνο που για το καθένα χρησιμοποιούνται τα αντίστοιχα βάρη. Ονομάζονται θύρες διότι έχουν ως συνάρτηση ενεργοποίησης την σιγμοειδή, η οποία συμπιέζει τις τιμές των διανυσμάτων στο εύρος $[0, 1]$. Εναλλακτικά θα μπορούσε κάποιος να τα ερμηνεύσει σαν μάσκες ή φίλτρα, πάνω σε άλλα διανύσματα. Πολλαπλασιάζοντας μία θύρα με ένα άλλο διάνυσμα, καθορίζουμε σε τι βαθμό θα διατηρηθούν οι τιμές αυτού του διανύσματος.

  - Η θύρα εισόδου i, ρυθμίζει πόσο θέλουμε η νέα είσοδος x_t να περάσει μέσα στο κελί.
  - Η θύρα λήθης f, ρυθμίζει πόσο θέλουμε η προηγούμενη κατάσταση h_{t-1} να περάσει μέσα στο κελί.
  - Η θύρα εξόδου o, ρυθμίζει πόσο θέλουμε να εκθέσουμε τη νέα κατάσταση h_t στο επόμενο βήμα.

- Το διάνυσμα $\tilde{c}_t$, είναι η *υποψήφια* νέα τιμή για τον πυρήνα του LSTM. Είναι στην ουσία η ίδια πράξη με αυτή που εκτελούμε στο απλό RNN.
- Το διάνυσμα c_t, είναι η νέα τιμή για τον πυρήνα του LSTM. Όπως φαίνεται και από την σχετική εξίσωση, η νέα τιμή θα είναι ένας συνδυασμός της προηγούμενης τιμής (φιλτραρισμένης από την θύρα λήθης) και από την *υποψήφια* τιμή (φιλτραρισμένης από την θύρα εισόδου).
- Το διάνυσμα h_t, είναι η νέα κατάσταση του LSTM. Υπολογίζεται φιλτράροντας την τιμή του πυρήνα του LSTM, με την θύρα εξόδου.

Peephole LSTM Το Peephole LSTM (LSTM με "ματάκι") (Gers κ.ά. 2002) είναι ίδιο με το απλό LSTM, μόνο που αντί οι θύρες να βασίζονται στην προηγούμενη τιμή του h_{t-1}, βασίζονται στην προηγούμενη τιμή του c_{t-1}.

4.3.3.5 GRU

Μία πρόσφατη παραλλαγή του LSTM είναι το Gated Recurrent Units (GRU) (Cho κ.ά. 2014) δίκτυο, το οποίο έχει πιο απλή αρχιτεκτονική. Έχει δύο θύρες αντί για τρεις και δεν διατηρεί εσωτερική μνήμη (c_t).

Ορισμός Η πράξη $\circ$ δηλώνει το γινόμενο Hadamard (στοιχείο προς στοιχείο). Οι αρχικές τιμές είναι: $h_0 = 0$.

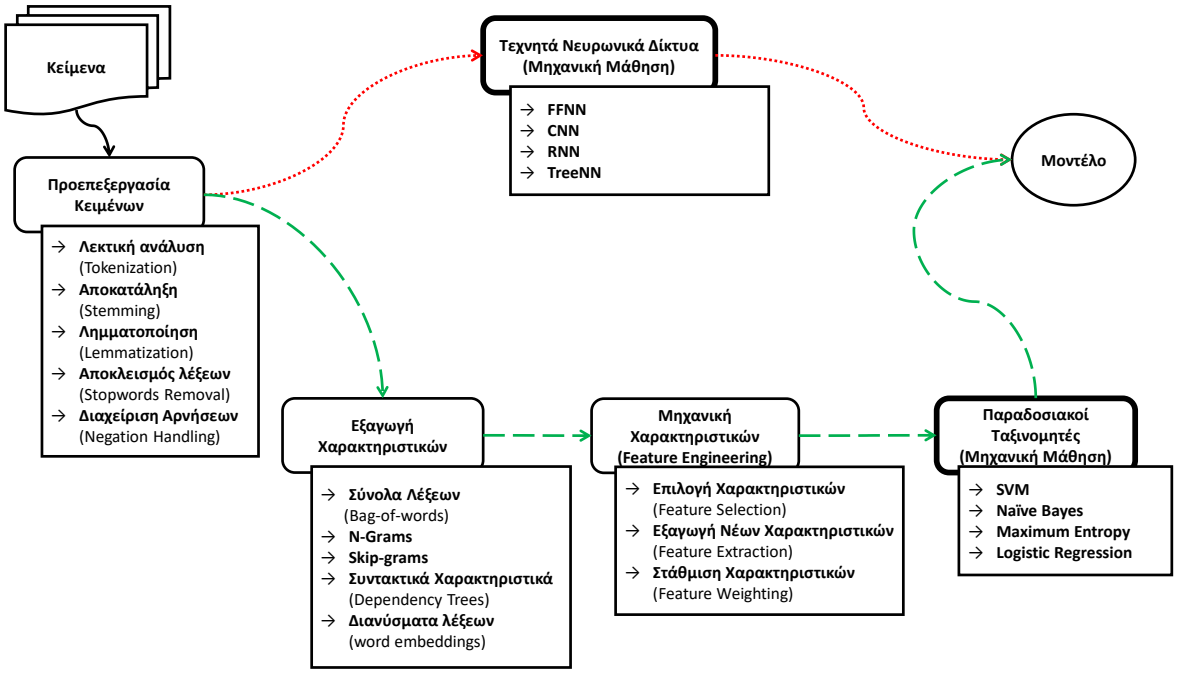
$$
\begin{aligned}
z_t &= \sigma_g(W_z x_t + U_z h_{t-1} + b_z) && \text{θύρα ενημέρωσης (update)} \\
r_t &= \sigma_g(W_r x_t + U_r h_{t-1} + b_r) && \text{θύρα επαναφοράς (reset)} \\
\tilde{h}_t &= \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) && \text{υποψήφια έξοδος (candidate output)} \\
h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t && \text{έξοδος (output)}
\end{aligned}
$$

- Δεν υπάρχει θύρα λήθης. Αντιθέτως, χρησιμοποιείται μία θύρα ενημέρωσης, η οποία ρυθμίζει τον βαθμό στον οποίο ο συνδυασμός της νέας εισόδου x_t, με την προηγούμενη έξοδο h_{t-1}, θα συνεισφέρουν στην νέα έξοδο h_t.
- Δεν υπάρχει διάνυσμα μνήμης. Χρησιμοποιείται κατευθείαν η έξοδος (ή κρυφή κατάσταση, όπως στο RNN).
- Όταν υπολογίζουμε την έξοδο, δεν εφαρμόζουμε μία συνάρτηση ενεργοποίησης.

4.3.3.6 Σύγκριση

Το απλό RNN πλέον δεν χρησιμοποιείται συχνά. Αντίθετα τα περισσότερα RNNs χρησιμοποιούν μία από τις πιο εξεζητημένες παραλλαγές. Οι πιο δημοφιλείς είναι το LSTM (με τις παραλλαγές του) και το GRU. Μέχρι αυτή τη στιγμή δεν είναι ξεκάθαρο ποια από τις δύο αρχιτεκτονικής είναι καλύτερη. Το GRU, έχοντας πιο απλή αρχιτεκτονική, όπως είναι λογικό είναι πιο γρήγορο. Τα αποτελέσματα σε σχετικές μελέτες είναι μικτά (Chung κ.ά. 2014; Greff κ.ά. 2015; Zaremba 2015) και οι όποιες διαφορές είναι μικρές.

Σε γενικές γραμμές, η απλότητα του GRU, το καθιστά προτιμότερο όταν τα δεδομένα είναι λίγα, διότι οι λιγότερες παράμετροι που έχει, το προστατεύουν από



Σχήμα 4.14: Οι δύο κυρίαρχες προσεγγίσεις για την δημιουργία ενός μοντέλου επεξεργασίας φυσικής γλώσσας.

πιθανή υπερπροσαρμογή. Από την άλλη μεριά, όταν τα δεδομένα είναι πολλά ή οι εξαρτήσεις ανάμεσα στα στοιχεία των ακολουθιών είναι πιο σύνθετες, το LSTM συνήθως αποδίδει καλύτερα.

4.4 Σύνοψη διαφορετικών προσεγγίσεων

Στο κεφάλαιο αυτό έγινε καταγραφή και σύγκριση των κυριότερων προσεγγίσεων μηχανικής μάθησης, για την επίλυση προβλημάτων ΕΦΓ. Όπως φαίνεται και στην Εικόνα 4.14, μπορούμε να διακρίνουμε δύο βασικές προσεγγίσεις.

Αρχικά, η "παραδοσιακή" προσέγγιση απαιτεί μεγάλη συμμετοχή του ανθρώπου στην διαδικασία του σχεδιασμού του μοντέλου. Τα χαρακτηριστικά πρέπει να σχεδιαστούν χειροκίνητα. Η διαδικασία αυτή είναι εξαιρετικά δύσκολη, διότι σε σύνθετα προβλήματα όπως αυτά που συναντάμε στην ΕΦΓ, υπάρχει μεγάλη ασάφεια στα δεδομένα. Επίσης, είναι μία πολύ χρονοβόρα διαδικασία, επειδή συχνά απαιτείται αλλαγή στα χαρακτηριστικά που έχουν εξαχθεί, ώστε να περιγράφουν με καλύτερο τρόπο τα δεδομένα. Η γενικότερη διαδικασία της μηχανικής χαρακτηριστικών, συχνά απορροφά την πλειοψηφία του χρόνου για την μοντελοποίηση ενός προβλήματος.

Αντιθέτως, με την χρήση ΤΝΔ μπορούμε να παρακάμψουμε τη δύσκολη και χρονοβόρα διαδικασία της μοντελοποίησης των κειμένων και να συγκεντρώσουμε την προσοχή μας στη διαδικασία της μηχανικής μάθησης. Ο λόγος που δεν απαιτείται αυτό από τα ΤΝΔ, είναι γιατί έχουν την ιδιότητα να "μαθαίνουν" τα χαρακτηριστικά. Αυτή η ιδιότητα των ΤΝΔ τα κάνει πολύ δελεαστικά, καθώς εκτός από το γεγονός ότι συχνά ξεπερνούν σε επιδόσεις τα παραδοσιακά συστήματα, αποτελούν ένα πολύ γενικό εργαλείο. Συχνά μία αρχιτεκτονική μπορεί να χρησιμοποιηθεί αυτούσια για την επίλυση πολλών προβλημάτων, λόγω της προσαρμοστικότητάς τους.

Κεφάλαιο 5

Μοντέλα Ανάλυσης Συναισθήματος

Σε αυτό το κεφάλαιο θα παρουσιαστούν τα αποτελέσματα της έρευνας στα πλαίσια της εργασίας. Αναπτύχθηκαν ένα σύνολο από μοντέλα για Ανάλυσης Συναισθήματος (ΑΣ), με κύρια τα μοντέλα που αναπτύχθηκαν στα πλαίσια της συμμετοχής, στον διεθνή διαγωνισμό σημασιολογικής αξιολόγησης Semeval 2017. Κατατέθηκαν μοντέλα σε δύο κατηγορίες:

- Task 4: "Sentiment Analysis in Twitter" (Baziotis κ.ά. 2017a). Στόχος της κατηγορίας ήταν η ανάπτυξη μοντέλων για την επίλυση μίας σειράς προβλημάτων ΑΣ, σε μηνύματα του κοινωνικού δικτύου Twitter. Τα μοντέλα σχεδιάστηκαν για την ανάλυση μηνυμάτων της Αγγλικής γλώσσας. Το μοντέλο στην υποκατηγορία A (Subtask A), η οποία έχει την μεγαλύτερη συμμετοχή, πέτυχε την 1η θέση.

- Task 6: "#HashtagWars: Learning a Sense of Humor" (Baziotis κ.ά. 2017b). Στόχος της κατηγορίας ήταν η ανάπτυξη μοντέλων για την επίλυση προβλημάτων Υπολογιστικού Χιούμορ (Computational Humor), σε μηνύματα του κοινωνικού δικτύου Twitter. Το μοντέλο στην υποκατηγορία A, πέτυχε την 2η θέση, αλλά εκ των υστέρων βελτιώσεις, ξεπερνούν κατά πολύ τα μέχρι τώρα καλύτερα αποτελέσματα, τόσο στα δεδομένα του διαγωνισμού, όσο και στο σχετικό σύνολο δεδομένων.

Στο κεφάλαιο αυτό θα αναλυθούν τα μοντέλα και τα αποτελέσματα για τον διαγωνισμό Semeval 2017 Task 4: "Sentiment Analysis in Twitter". Τα μοντέλα μας συμμετείχαν σε όλες τις υποκατηγορίες του Task 4, για την Αγγλική γλώσσα. Βασίζονται σε αμφίδρομα LSTMs ενισχυμένα με δύο διαφορετικούς μηχανισμούς *attention* (Κεφ. 4.3.3.3). Εκπαιδεύτηκαν πάνω σε διανύσματα λέξεων, τα οποία σχηματίστηκαν χρησιμοποιώντας μία μεγάλη συλλογή από μηνύματα από το Twitter, τα οποία συλλέξαμε. Επίσης, για την προετοιμασία των κειμένων, αναπτύχθηκε ένα εργαλείο προεπεξεργασίας κειμένων, σχεδιασμένο ώστε να μπορεί να διαχειριστεί τις ιδιαιτερότητες των κειμένων σε μηνύματα από κοινωνικά δίκτυα. Το μοντέλο του Subtask A, πέτυχε την 1η θέση ανάμεσα σε 39 ομάδες. Επιτυχημένα ήταν και τα μοντέλα στις υπόλοιπες υποκατηγορίες, πετυχαίνοντας σε όλες την 2η θέση (με εξαίρεση το Subtask E).

5.1 Περιγραφή του Προβλήματος

Αντικείμενο του Task 4: "Sentiment Analysis in Twitter" ήταν η ανάπτυξη μοντέλων για την αντιμετώπιση μίας σειράς από προβλήματα ΑΣ, όπως προσδιορισμός και ποσοτικοποίηση του συναισθηματικού προσανατολισμού σε μηνύματα από το Twitter, καθώς και για επιλεγμένα θέματα στα μηνύματά αυτά.

5.1.1 Κατηγοριοποίηση Μηνυμάτων

Προσανατολισμός Μηνύματος (Message-level SA) Το πρόβλημα αυτό αφορά την κατηγοριοποίηση των μηνυμάτων, βάση του γενικού συναισθηματικού προσανατολισμού τους, σε ένα σύνολο από προεπιλεγμένες κλάσεις, όπως {*θετικό*, *αρνητικό*, *ουδέτερο*}.

Προσανατολισμός Θέματος (Topic-based SA) Συνεχίζοντας την αντίστοιχη υποκατηγορία από το Semeval 2016, το ζητούμενο του προβλήματος αφορά την κατηγοριοποίηση ενός μηνύματος, σε προεπιλεγμένες κλάσεις, βάση του συναισθηματικού προσανατολισμού όχι γενικά για ένα μήνυμα, αλλά συγκεκριμένα προς ένα δοσμένο θέμα, το οποίο αναφέρεται στο μήνυμα.

5.1.2 Ποσοτικοποίηση Μηνυμάτων

Σε αυτό το πρόβλημα το ζητούμενο ήταν η πρόβλεψη της *κατανομής* των μηνυμάτων για ένα θέμα, σε ένα σύνολο από προεπιλεγμένες κλάσεις. Στις πρακτικές εφαρμογές, το τελικό ζητούμενο δεν είναι ο υπολογισμός του συναισθήματος των διαφόρων μηνυμάτων ή εγγράφων. Το ζητούμενο είναι να συμψηφιστούν όλες αυτές, ώστε να υπολογιστεί η κατανομή του συναισθήματος. Για παράδειγμα, σε μία εφαρμογή εύρεσης του συναισθηματικού προσανατολισμού προς διάφορα πολιτικά κόμματα, αυτό που πραγματικά έχει ενδιαφέρον είναι να βρεθεί ποιο ποσοστό των μηνυμάτων είναι *θετικά ή αρνητικά* απέναντι σε κάθε κόμμα.

5.1.3 Ορισμός Κατηγοριών

Το Semeval 2017 αποτελείται από 5 υποκατηγορίες, για κάθε γλώσσα (Αγγλικά, Αραβικά):

- **Subtask A**: Δοθέντος ενός μηνύματος, να ταξινομηθεί το μήνυμα σε μία από τις κλάσεις {*θετικό*, *αρνητικό*, *ουδέτερο*}.

Subtask	Στόχος	Λεπτομέρεια	Θέμα
A	Ταξινόμηση	3-κλάσεων	-
B	Ταξινόμηση	2-κλάσεων	ΝΑΙ
C	Ταξινόμηση	5-κλάσεων	ΝΑΙ
D	Ποσοτικοποίηση	2-κλάσεων	ΝΑΙ
E	Ποσοτικοποίηση	5-κλάσεων	ΝΑΙ

Πίνακας 5.1: Υποκατηγορίες του Semeval 2017 - Task 4: "Sentiment Analysis in Twitter".

- **Subtask B**: Δοθέντος ενός μηνύματος και ενός θέματος, να ταξινομηθεί το συναίσθημα που εκφράζεται στο μήνυμα προς το θέμα, σε μία από τις δύο κλάσεις {*θετικό*, *αρνητικό*}.
- **Subtask C**: Δοθέντος ενός μηνύματος και ενός θέματος, να ταξινομηθεί το συναίσθημα που εκφράζεται στο μήνυμα προς το θέμα, σε μία από τις πέντε κλάσεις {*πολύ_θετικό*, *θετικό*, *ουδέτερο*, *αρνητικό*, *πολύ_αρνητικό*}.
- **Subtask D**: Δοθέντος ενός συνόλου μηνυμάτων σχετικά με ένα θέμα, να εκτιμηθεί η *κατανομή* των μηνυμάτων στις δύο κλάσεις {*θετικό*, *αρνητικό*}.
- **Subtask E**: Δοθέντος ενός συνόλου μηνυμάτων σχετικά με ένα θέμα, να εκτιμηθεί η *κατανομή* των μηνυμάτων στις πέντε κλάσεις {*πολύ_θετικό*, *θετικό*, *ουδέτερο*, *αρνητικό*, *πολύ_αρνητικό*}.

Το Subtask A εκτελείται κάθε χρόνο τα τελευταία 5 χρόνια και είναι με διαφορά η πιο δημοφιλής κατηγορία και μία από τις δημοφιλέστερες ολόκληρου του Semeval. Το 2016 ήταν πρώτη σε συμμετοχές και το 2017 δεύτερη. Αυτό δείχνει το ενδιαφέρον της επιστημονικής κοινότητας για το πρόβλημα. Στην Εικόνα 5.1 παρουσιάζεται η σύνοψη με τις περιγραφές των υποκατηγοριών του διαγωνισμού.

5.2 Σύνολα Δεδομένων

Μία από τις σημαντικότερες προσφορές του Semeval είναι η συλλογή και ο διαμοιρασμός δεδομένων, για την εκπαίδευση μοντέλων μηχανικής μάθησης. Αυτό είναι πολλές φορές το μεγαλύτερο εμπόδιο για την έρευνα. Η συλλογή αρκετών και ποιοτικών δεδομένων είναι μία από τις προϋποθέσεις για την δημιουργία ενός καλού στατιστικού μοντέλου.

Φέτος, προσφέρθηκαν τα δεδομένων από τους διαγωνισμούς προηγούμενων ετών και για πρώτη φορά επιτράπηκε να συμπεριληφθούν στα δεδομένα εκπαίδευσης, τα δεδομένα των δοκιμών των προηγούμενων ετών. Έτσι, τα δεδομένα τα οποία ήταν διαθέσιμα φέτος, ήταν σχεδόν διπλάσια από άλλες χρονιές.

Τα στατιστικά των δεδομένων φαίνονται στην εικόνα 5.2. Αποτελούνται από μηνύματα στο Twitter από το 2013 έως το 2016. Τα μηνύματα σημειώθηκαν από ανθρώπους και όχι αυτοματοποιημένα, με την χρήση της υπηρεσίας Crowdflower[1]. Για την στοχευμένη ΑΣ, επιλέχθηκαν τα πιο δημοφιλή θέματα σύμφωνα με το Twitter trends[2]. Επίσης αυτή τη χρονιά συμπεριλήφθηκε η δυνατότητα αξιοποιήσης δεδομένων σχετικά με τους χρήστες των μηνυμάτων: αναγνωριστικό χρήση (user id), πλήθος ακολούθων, περιγραφή, πλήθος φίλων, τοποθεσία, γλώσσα, χώρα, όνομα, ζώνη ώρας.

5.3 Μέτρα Αξιολόγησης

Σε αυτό το τμήμα θα παρουσιαστούν τα μέτρα αξιολόγησης, για κάθε υποκατηγορία του *SemEval-2017 Task 4* (Rosenthal κ.ά. 2017). Σε μεγάλο βαθμό είναι ίδια με τα μέτρα του 2016, τα οποία περιγράφονται πολύ αναλυτικά στο (Nakov κ.ά. 2016a).

[1] https://www.crowdflower.com/
[2] https://trends24.in/

Dataset	Task	Θετικά 2	Θετικά 1	Ουδέτερα 0	Αρνητικά -1	Αρνητικά -2	Σύνολο
Train	A	19652 (39.64%)		22195 (44.78%)	7723 (15.58%)		49570
	B,D	14897 (78.85%)		-	3997 (21.15%)		18894
	C,E	1016 (3.34%)	12852 (42.23%)	12888 (42.35%)	3380 (11.11%)	296 (0.97%)	30432
Test	A	2375 (19.33%)		5937 (48.33%)	3972 (32.33%)		12284
	B,D	2463 (39.82%)		-	3722 (60.18%)		6185
	C,E	131 (1.06%)	2332 (18.84%)	6194 (50.04%)	3545 (28.64%)	177 (1.43%)	12379

Πίνακας 5.2: Στατιστικά των συνόλων δεδομένων για το Task 4. Προσέξτε την μεγάλη διαφορά στους λόγους των θετικών-αρνητικών κλάσεων, ανάμεσα στην προηγούμενη και την φετινή χρονιά.

		Πραγματική		
		Pos	Neg	Neu
Πρόβλεψη	Pos	PP	PU	PN
	Neg	UP	UU	UN
	Neu	NP	NU	NN

Πίνακας 5.3: Πίνακας Σύγχυσης, για το Subtask A.

5.3.1 Subtask A: Ταξινόμηση - Γενικό Συναίσθημα

Τα μέτρα αξιολόγησης είναι: *macro-averaged recall* ή ρ_{macro} ή ρ, macro-averaged $F1^{PN}$ (μόνο για θετικά, αρνητικά μηνύματα) και ακρίβεια acc.

Macro-averaged Recall Το ρ_{macro} ή ρ είναι το μέσο *recall*, ανάμεσα σε όλες τις κλάσεις, *θετικό* (P), *αρνητικό* (N), *ουδέτερο* (U).

$$\rho_{macro} = \frac{1}{3}(\rho^{\mathrm{P}} + \rho^{\mathrm{N}} + \rho^{\mathrm{U}}) \tag{5.1}$$

Το ρ^P είναι το ποσοστό από τα μηνύματα που ανήκουν στην θετική κλάση και κατηγοριοποιήθηκαν σωστά σε αυτή:

$$\rho^P = \frac{PP}{PP + UP + NP} \tag{5.2}$$

F1 Το $F1^{PN}$ ήταν το βασικό μέτρο της προηγούμενης χρονιάς. Το $F1^{PN}$ είναι ο μέσος όρος ανάμεσα στο $F1$ σκορ, των θετικών και των αρνητικών μηνυμάτων.

$$F1^{PN} = \frac{1}{2}(F1^P + F1^N) \tag{5.3}$$

και για να υπολογίσουμε το $F1^P$, αρχικά πρέπει να υπολογίσουμε το π^P, δηλαδή το ποσοστό των μηνυμάτων που ανήκουν στην θετική κλάση και είναι όντως θετικά:

$$\pi^P = \frac{PP}{PP + PU + NP}, \tag{5.4}$$

$$F1^P = \frac{2 \times \pi^P \times \rho^P}{\pi^P \times \rho^P} \tag{5.5}$$

Με αντίστοιχο τρόπο υπολογίζονται τα $F1^N$, ρ^N, ρ^U. Τα $F1$ και ρ είναι προτιμότερα από την απλή ακρίβεια acc, διότι είναι λιγότερο επιρρεπή σε δυσανάλογες κλάσεις.

5.3.2 Subtask B: Ταξινόμηση - Στοχευμένο Συναίσθημα (2 κλάσεις)

Όπως και στο Subtask A, το μέτρο αξιλόγησης είναι το *macro-averaged recall* ή ρ_{macro} ή ρ, το $F1$ και η ακρίβεια acc. Οι υπολογισμοί προσαρμόζονται στις δύο κλάσεις του συγκεκριμένου προβλήματος. Δηλαδή για το ρ έχουμε:

$$\rho_{macro} = \frac{1}{2}(\rho^{\mathrm{P}} + \rho^{\mathrm{N}}) \tag{5.6}$$

5.3.3 Subtask C: Σειριακή Ταξινόμηση - Στοχευμένο Συναίσθημα (5 κλάσεις)

Επειδή το πρόβλημα αυτό αφορά σειριακή ταξινόμηση, θα πρέπει να χρησιμοποιηθούν μέτρα τα οποία λαμβάνουν υπόψη τους τη σειρά των κλάσεων. Οι κλάσεις του προβλήματος είναι οι εξής: $C =$*{πολύ_θετικό, θετικό, ουδέτερο, αρνητικό, πολύ_αρνητικό}*, οι οποίες αντιστοιχούν στις τιμές $C = \{-2, -1, 0, +1, +2\}$ οι οποίες αντιπροσωπεύουν την σειρά των κλάσεων. Αυτό σημαίνει ότι αν ταξινομηθεί εσφαλμένα ένα μήνυμα ως *αρνητικό* (-1), το σφάλμα θα είναι μεγαλύτερο αν η πραγματική κλάση του μηνύματος είναι *πολύ_θετικό* (+2) από *θετικό* (+1). Τα μέτρα τα οποία χρησιμοποιήθηκαν είναι το *macroaveraged mean absolute error* (MAE) ως το βασικό μέτρο και το *microaveraged mean absolute error* (MAE^{μ}).

macroaveraged mean absolute error (MAE) Το MAE έχει το πλεονέκτημα απέναντι στο κανονικό MAE^{μ}, ότι διαχειρίζεται καλύτερα τις δυσαναλογίες των κλάσεων. Ο λόγος είναι ότι υπολογίζει τον μέσο όρο στο επίπεδο των κλάσεων και όχι των παρατηρήσεων:

$$MAE(h, T_e) = \frac{1}{\mid C \mid} \sum_{j=1}^{|C|} \frac{1}{\mid T_{e_j} \mid} \sum_{x_i \in T_{e_j}} \mid h(x_i) - y_i \mid \tag{5.7}$$

microaveraged mean absolute error (MAE^{μ}) Το MAE^{μ} είναι το "κανονικό" MAE και είναι απλά ο μέσος όρος των σφαλμάτων για όλες τις παρατηρήσεις του test set. Δεν κάνει διάκριση ανάμεσα στις κλάσεις:

$$MAE^{\mu}(h, T_e) = \frac{1}{\mid T_{e_j} \mid} \sum_{x_i \in T_{e_j}} \mid h(x_i) - y_i \mid \tag{5.8}$$

Για τα παραπάνω ισχύουν τα εξής:

- y_i είναι το πραγματικό σήμα της παρατήρησης x_i
- το $h(x_i)$ δηλώνει το εκτιμημένο σήμα
- το T_{e_j} δηλώνει το σύνολο από τις παρατηρήσεις στο test set, για τις οποίες οι πραγματική κλάση είναι η c_j

- η απόλυτη διαφορά $\mid h(x_i) - y_i \mid$ αντιστοιχεί στην "απόσταση" της πρόβλεψης από την πραγματική τιμή της παρατήρησης.

Επίσης, τα δύο αυτά μέτρα υπολογίζουν σφάλματα και κατά συνέπεια μικρότερες τιμές είναι καλύτερες.

5.3.4 Subtask D: Ποσοτικοποίηση (2 κλάσεις)

Το πρόβλημα αυτό είναι διαφορετικής φύσης από τα υπόλοιπα (ταξινόμηση) και θα πρέπει να χρησιμοποιηθεί ένα μέτρο το οποίο να συγκρίνει κατανομές. Το μέτρο αυτό είναι το κόστος της *κανονικοποιημένης διεντροπίας* (normalized cross-entropy), το οποίο είναι γνωστό ως *Kullback-Leibler Divergence* (KLD). Το μέτρο αυτό υπολογίζει την απόκλιση της πραγματική κατανομής p με την εκτιμημένη κατανομή $\hat{p}$ σε ένα σύνολο κλάσεων C.

$$KLD(\hat{p}, p, C) = \sum_{c_j \in C} p(c_j) log \frac{p(c_j)}{\hat{p}(c_j)} \tag{5.9}$$

Επειδή στις περιπτώσεις που η τιμή του $\hat{p}(c_j)$ είναι σχεδόν μηδέν, το KLD μπορεί να γίνει άπειρο, προστίθεται ένας πολύ μικρός όρος ϵ ο οποίος εξομαλύνει τα αποτελέσματα. Ορίζεται $\epsilon = \frac{1}{2|T_e|}$. Έτσι η εξομαλυμένη τιμή του $p(c_j)$, ή $p_s(c_j)$, υπολογίζεται ως εξής:

$$p_s(c_j) = \frac{p(c_j) + \epsilon}{(\sum_{c_j \in C} p(c_j)) + \epsilon \times |C|} = \frac{p(c_j) + \epsilon}{1 + \epsilon \times |C|} \tag{5.10}$$

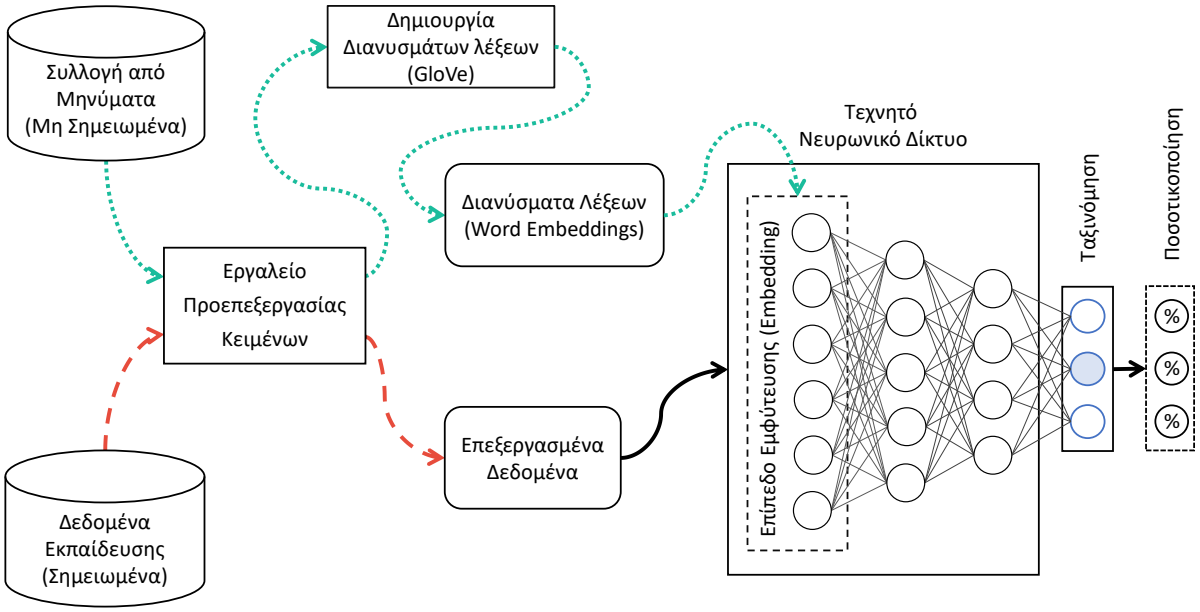
Αντίστοιχα υπολογίζεται και το $\hat{p}_s(c_j)$. Επειδή το μέτρο υπολογίζει την απόσταση από την πραγματική κατανομή, μικρότερες τιμές είναι καλύτερες.

5.3.5 Subtask E: Σειριακή Ποσοτικοποίηση (5 κλάσεις)

Σε αυτό το πρόβλημα, πρέπει να εκτιμήσουμε μία κατανομή για τις κλάσεις, για τις οποίες όμως υπάρχει συγκεκριμένη σειρά. Άρα, πρέπει το μέτρο που θα χρησιμοποιηθεί (1) να είναι κατάλληλο για σύγκριση κατανομών (2) να λαμβάνει υπόψη του τη σειρά των κλάσεων. Το μέτρο το οποίο υοθετήθηκε στο διαγωνισμό είναι το *Earth Mover's Distance* (Rubner κ.ά. 2000). Έτσι για τις κλάσεις του προβλήματος C ={*πολύ_θετικό*, *θετικό*, *ουδέτερο*, *αρνητικό*, *πολύ_αρνητικό*}, με την σειρά $C = \{-2, -1, 0, +1, +2\}$, υπολογίζουμε το *EMD* ως εξής:

$$EMD(\hat{p}, p) = \sum_{j=1}^{|C|-1} \mid \sum_{i=1}^{j} \hat{p}(c_i) - \sum_{i=1}^{j} p(c_i) \mid \tag{5.11}$$

Το μέτρο αυτό υπολογίζει την απόσταση της εκτιμημένης κατανομής από την πραγματική, άρα είναι μέτρο σφάλματος, συνεπώς μικρότερες τιμές είναι καλύτερες. Το εύρος του *EMD* είναι από 0 (καλύτερο) έως $|C| - 1$ (χειρότερο).



Σχήμα 5.1: Σύνοψη της αρχιτεκτονικής των συστημάτων μηχανικής μάθησης για το Semeval 2017 - Task 4.

5.4 Επισκόπηση Προσέγγισης

Η Εικόνα 5.1 παρουσιάζει μία σύνοψη της αρχιτεκτονικής των συστημάτων. Τα συστήματα σχηματίζονται, από 2 βήματα συν ένα ακόμη προαιρετικό ανάλογα την υποκατηγορία:

(1) Προεπεξεργασία των κειμένου. Σε αυτό το βήμα γίνεται προετοιμασία των κειμένων, με στόχο την μετατροπή τους σε μία μορφή την οποία θα μπορεί να αξιοποιήσει καλύτερα το Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ). Εκτελείται λεκτική ανάλυση, διόρθωση ορθογραφικών λαθών, τμηματοποίηση λέξεων (για τον διαχωρισμό των hashtags) και κανονικοποίηση λέξεων και φράσεων.

(2) Εκπαίδευση. Ένα ΤΝΔ εκπαιδεύεται σε ένα πρόβλημα, αξιοποιώντας τα επεξεργασμένα μηνύματα σε συνδυασμό με το προ-εκπαιδευμένα διανύσματα λέξεων.

(3) Ποσοτικοποίηση (Για τα Subtask D, E). Σε αυτό το βήμα γίνεται εκτίμηση της κατανομής των μηνυμάτων στις κλάσεις του αντίστοιχου προβλήματος, χρησιμοποιώντας τον ταξινομητή του προηγούμενου βήματος.

5.4.1 Συλλογή Μηνυμάτων από Twitter

Συλλέξαμε ένα αρκετά μεγάλο σύνολο δεδομένων αποτελούμενο από δισεκατομμύρια μηνύματα στο Twitter από τις 12/2012 έως 07/2016. Από τα μηνύματα αυτά, ξεχωρίστηκαν εκείνα τα οποία είναι στην Αγγλική γλώσσα και τα οποία έχουν τουλάχιστον 3 λέξεις και δεν είναι αναμετάδοση (retweet) άλλου μηνύματος. Ο λόγος για τον οποία επιλέχθηκαν μόνο όσα δεν είναι retweet άλλων, είναι

γιατί ορισμένες φορές ένα μήνυμα μπορεί να αναμεταδοθεί αυτούσιο ή τροποποιημένο χιλιάδες φορές, το οποίο μπορεί να αλλοιώσει την αντικειμενικότητα των στατιστικών των λέξεων.

Το τελικό σύνολο δεδομένων, αποτελείται από 330 εκατομμύρια μηνύματα. Αυτά τα μηνύματα χρησιμοποιήθηκαν (1) για την παραγωγή διανυσμάτων λέξεων (word embeddings) και (2) για την συλλογή στατιστικών για τις εμφανίσεις των λέξεων (unigrams και bigrams). Τα στατιστικά αξιοποιούνται από τον επεξεργαστή κειμένων για την εκτέλεση ορθογραφικής διόρθωσης και διαχωρισμού συμπτυγμένων λέξεων.

5.4.2 Διανύσματα Λέξεων

Τα διανύσματα λέξεων (Κεφ. 3.2.3) είναι πυκνές διανυσματικές αναπαραστάσεις, οι οποίες περιέχουν σημασιολογικές και συντακτικές πληροφορίες των λέξεων. Αξιοποιώντας τη μεγάλη και ποιοτική συλλογή μηνυμάτων, εκπαιδεύτηκαν διανύσματα λέξεων με την χρήση της τεχνικής GloVe (Pennington κ.ά. 2014), με τις προεπιλεγμένες τιμές για όλες τις παραμέτρους. Για την παραγωγή των διανυσμάτων, χρησιμοποιήθηκαν λέξεις οι οποίες εμφανίζονται τουλάχιστον 10 φορές, οδηγώντας στον σχηματισμό λεξιλογίου με 660.000 όρους. Επίσης, τα κείμενα τα οποία δόθηκαν ως είσοδο στο GloVe, είχαν αρχικά επεξεργαστεί με το δικό μας εργαλείο επεξεργασίας κειμένου. Με αυτό τον τρόπο, σχηματίστηκαν διανύσματα για τους ειδικούς όρους (ετικέτες όπως <hashtag>, <elongated> κλπ.) και διαμορφώθηκαν πιο ποιοτικά στατιστικά. Τα διανύσματα λέξεων χρησιμοποιούνται για την αρχικοποίηση των βαρών του πρώτου επιπέδου (επίπεδο εμφύτευσης) των Νευρωνικών Δικτύων.

5.5 Προετοιμασία Δεδομένων (ekphrasis)

Η προεπεξεργασία των κειμένων είναι ένα πολύ σημαντικό βήμα, καθώς οι όροι (tokens/terms) οι οποίοι θα παραχθούν από αυτό θα αποτελέσουν τη βάση για τον σχηματισμό των χαρακτηριστικών, πάνω στα οποία θα χτιστεί το μοντέλο. Για τον λόγο αυτό αναπτύχθηκε ένα εργαλείο επεξεργασίας κειμένων (*ekphrasis*[3]), το οποίο εκτελεί: (1) λεκτική ανάλυση (tokenization) η οποία διατηρεί εκφράσεις οι οποίες είναι χρήσιμες για τον προσδιορισμό του συναισθήματος, (2) διόρθωση ορθογραφικών λαθών, (3) κανονικοποίηση λέξεων και φράσεων (text normalization), (3) σημείωση λέξεων και φράσεων (word annotation), (4) διαχωρισμό ενοποιημένων λέξεων (text segmentation), στις επιμέρους λέξεις (για τον διαχωρισμό των hashtags). Το εργαλείο αποτελείται από δύο ξεχωριστά τμήματα, αυτό του λεκτικού αναλυτή και αυτό της μετα-επεξεργασίας των όρων. Ο λεκτικός αναλυτής αναλαμβάνει να αναγνωρίσει τους όρους στο κείμενο στην συνέχεια τους δίνει ως είσοδο στο μετα-επεξεργαστή.

5.5.1 Λεκτικός Αναλυτής (Tokenizer)

Η διαδικασία της λεκτικής ανάλυσης είναι αρκετά σημαντική καθώς το αποτέλεσμά της, θα αποτελέσει βάση για όλη την υπόλοιπη διαδικασία σχηματισμού του

[3] https://github.com/cbaziotis/ekphrasis

Αρχικό	@SentimentSymp: can't wait for the Nov 9 #Sentiment talks! YAAAAAAY!!! >:-D http://sentimentsymposium.com/.
(Potts 2011)	@sentimentsymp : can't wait for the Nov_09 #sentiment talks ! YAAAY ! ! ! >:-D http://sentimentsymposium.com/ .
ekphrasis (tok)	@sentimentsymp : can not wait for the nov 9 sentiment talks ! yay ! >:-d http://sentimentsymposium.com/.
ekphrasis (full)	<user> : can not wait for the **<date>** <hashtag> sentiment </hashtag> talks ! **yay** **<allcaps>** **<elongated>** ! **<repeated>** <devil> <url>

Πίνακας 5.4: Σύγκριση του *ekphrasis* με τον λεκτικό αναλυτή του (Potts 2011), ο οποίος είναι ειδικά σχεδιασμένος για ΑΣ στο Twitter. Στο παράδειγμα φαίνεται το αποτέλεσμα τόσο της λεκτικής ανάλυσης, όσο και το αποτέλεσμα της χρήσης όλων των δυνατοτήτων του *ekphrasis*.

μοντέλου. Στην διαδικασία αυτή χωρίζουμε την ακολουθία χαρακτήρων από την οποία αποτελείται ένα μήνυμα, στα μέρη που την αποτελούν. Η τυπικός διαχωρισμός γίνεται χωρίζοντας ένα κείμενο στα κενά και τα σημεία στίξης, με σκοπό να εξαχθούν οι λέξεις που το αποτελούν. Όμως αυτή η απλοϊκή προσέγγιση οδηγεί συχνά σε ανεπιθύμητα αποτελέσματα. Είναι πολλές οι φορές, όπου μπορεί να είναι χρήσιμο για το πρόβλημά μας, να διατηρήσουμε μία ακολουθία χαρακτήρων ως έναν όρο ακόμα και αν μεσολαβούν ειδικοί χαρακτήρες ανάμεσα του (ακόμη και κενά). Αυτό το πρόβλημα είναι αρκετά πιο έντονο στην περίπτωση κειμένων με πιο ανεπίσημο ύφος όπως στο Twitter, στο οποίο συχνά συναντά κανείς αργκό, λανθασμένη σύνταξη και γενικότερα "δημιουργική" γραφή.

Έτσι, είναι σημαντικό η διαδικασία της λεκτικής ανάλυσης να μπορεί να αναγνωρίσει όλους τους χρήσιμους όρους, και να τους διατηρήσει σαν έναν όρο. Για τον λόγο αυτό έχουν αναπτυχθεί ορισμένοι λεκτικοί αναλυτές οι οποίοι ειδικεύονται στην επεξεργασία κειμένων από μηνύματα στο Twitter, όπως οι (Gimpel κ.ά. 2011; Potts 2011). Αυτοί οι αναλυτές αναγνωρίζουν τη γλώσσα σήμανσης του Twitter (#hashtag, user κλπ), ορισμένες συναισθηματικές εκφράσεις και τα βασικά emoticons. Όμως, υπάρχει αρκετή χρήσιμη πληροφορία η οποία χάνεται ακόμα και με την χρήση αυτών των αναλυτών.

Για τον λόγο αυτό αναπτύχθηκε ένα εργαλείο επεξεργασίας κειμένων (*ekphrasis*[4])) το οποίο καλύπτει αρκετά περισσότερες περιπτώσεις. Το *ekphrasis* μπορεί να αναγνωρίσει πρακτικά όλα τα emoticons, όλα τα emojis, αλλά και αδόμητες εκφράσεις όπως ημερομηνίες (07/11/2011, April 23rd), ώρες (4:30pm, 11:00 am), συναλλάγματα ($10, 25mil, 50€), ακρωνύμια, λογοκριμένες λέξης (e.g. s**t), τονισμένες λέξεις (*very*) και πολλές άλλες εκφράσεις. Στην Εικόνα 5.4 φαίνεται ένα παράδειγμα σύγκρισης του με τον λεκτικό αναλυτή στο (Potts 2011).

5.5.2 Μετα-Επεξεργασία

Η *μετα-επεξεργασία* εκτελείται επί των όρων, οι οποίοι έχουν εξαχθεί από τον λεκτικό αναλυτή. Σε αυτό το βήμα εκτελούνται τα βήματα της ορθογραφικής διόρθωσης, της κανονικοποίησης και σημείωσης λέξεων και φράσεων, καθώς και του διαχωρισμού ενοποιημένων λέξεων.

[4]Για την λεκτική ανάλυση, το *ekphrasis* χρησιμοποιεί ένα μεγάλο σύνολο από σύνθετες *Κανονικές **Εκφράσεις*** (Regular Expressions).

Και για τις δύο λειτουργίες χρησιμοποιείται ο αλγόριθμος δυναμικού προγραμματισμού Viterbi, με τον οποίο βρίσκεται η πιο πιθανή διόρθωση για μία ανορθόγραφη λέξη και ομοίως ο πιο πιθανός συνδυασμός λέξεων για ενοποιημένες λέξεις. Για τον υπολογισμό των σχετικών πιθανοτήτων, αξιοποιούνται στατιστικά λέξεων από τη συλλογή με μηνύματα (Κεφ. 5.4.1). Η υλοποίηση του μηχανισμού ορθογραφικής διόρθωσης ακολουθεί την προσέγγιση του (Jurafsky κ.ά. 2000) και η υλοποίηση του διαχωρισμού ενοποιημένων λέξεων ακολουθεί την προσέγγιση του (Segaran κ.ά. 2009).

Η κανονικοποίηση των λέξεων ή φράσεων αφορά την αντικατάσταση τους με συγκεκριμένες ετικέτες. Για παράδειγμα εκφράσεις όπως ημερομηνίες (May 21, 2017), τηλεφωνικοί αριθμοί ή emails, αντικαθίστανται από ειδικές ετικέτες, όπως <date>, <phone>, <email>. Με τον τρόπο αυτό, εκφράσεις όπως οι παραπάνω, χαρτογραφούνται σε ένα μικρό σύνολο νέων λέξεων, κρύβοντας περιττές (για το συγκεκριμένο πρόβλημα) λεπτομέρειες. Ο λόγος που αυτό είναι επιθυμητό, είναι διότι για να προσδιοριστεί το συναίσθημα σε ένα μήνυμα δεν χρειάζεται να ξέρουμε ακριβώς μία ημερομηνία ή ένα email, αρκεί απλά να ξέρουμε τι έννοια αντιπροσωπεύει κάθε έκφραση. Έτσι, μειώνεται δραστικά το λεξιλόγιο και διευκολύνεται το τεχνητό νευρωνικό δίκτυο ώστε να μάθει πιο γενικά χαρακτηριστικά.

Η σημείωση των λέξεων ή φράσεων αφορά την διαδικασία κατά την οποία, σε συγκεκριμένες λέξεις και φράσεις εισάγονται ειδικές ετικέτες γύρω από αυτές. Για παράδειγμα, ένα hashtag όπως το #helloworld περιστοιχίζεται (και διαχωρίζεται) με ειδικές ετικέτες ως εξής: <hashtag> hello world </hashtag>. Με αυτόν τον τρόπο, το νευρωνικό δίκτυο μπορεί να αξιοποιήσει αυτή την πληροφορία και να ανακαλύψει συσχετίσεις, όπως δηλαδή ότι το νόημα ή η ένταση μία λέξης έχει ειδική σημασία όταν πλαισιώνεται από τις διάφορες ετικέτες. Άλλα παραδείγματα αφορούν λογοκριμένες λέξεις (w**d), λέξεις με όλα τα γράμματα κεφαλαία (WORD) ή λέξεις με έμφαση (*word*).

Η λειτουργικότητα η οποία παρουσιάστηκε παραπάνω μπορεί σε μεγάλο βαθμό να προσαρμοστεί, καθώς σχεδόν όλα τα βήματα είναι παραμετρικά. Μπορεί δηλαδή κανείς να επιλέξεις ποιες λέξεις θα κανονικοποιηθούν, ποιες θα σημειωθούν αλλά και με ποιο τρόπο (με ετικέτες γύρω από τη λέξη, μόνο μετά τη λέξη κλπ.). Η αναλυτική τεκμηρίωση του εργαλείου είναι διαθέσιμη στην ιστοσελίδα του project, το οποίο είναι ανοιχτού κώδικα. Για την εκτέλεση της επεξεργασίας αρκούν λίγες γραμμές κώδικα σε Python.

```
text_processor = TextPreProcessor(
# όροι για κανονικοποίηση
normalize=['url', 'email', 'percent', 'money', 'phone', 'user',
'time', 'url', 'date', 'number'],
# όροι για σημείωση
annotate={"hashtag", "allcaps", "elongated", "repeated",
'emphasis', 'censored'},
fix_html=True,  # διόρθωση χαρακτήρων από HTML

# πηγή στατιστικών για διαχωρισμό λέξεων
segmenter="twitter",

# πηγή στατιστικών για ορθογραφική διόρθωση λέξεων
corrector="twitter",

unpack_hashtags=True,  # διαχωρισμός hashtags
```

```
unpack_contractions=True,  # ανάπτυξη αρνήσεων (can't -> can not)
spell_correct_elong=False,  # ορθογραφική διόρθωση σε επιμηκείς λέξεις

# επιλογή Tokenizer. Εκτός από αυτόν που παρέχεται με το ekphrasis,
# προσφέρεται η δυνατότητα να χρησιμοποιηθεί και κάποιος τρίτος
tokenizer=SocialTokenizer(lowercase=True).tokenize,

# λίστα με λεξικά (dictionaries) για αντικατάσταση όρων
dicts=[emoticons]
)
```

Αρχικό	The *new* season of #TwinPeaks is coming on May 21, 2017. CANT WAIT \o/ !!! #tvseries #davidlynch :D
Επεξεργασμένο	the new <emphasis> season of <hashtag> twin peaks </hashtag> is coming on <date> . cant <allcaps> wait <allcaps> <happy> ! <repeated> <hashtag> tv series </hashtag> <hashtag> david lynch </hashtag> <laugh>

Πίνακας 5.5: Παράδειγμα λειτουργίας του εργαλείου επεξεργασίας κειμένου. Η εισαγωγή ειδικών ετικετών έχει ως στόχο να βοηθήσει το RNN να μάθει καλύτερα χαρακτηριστικά, αξιοποιώντας τις εξαρτήσεις των λέξεων με τις ετικέτες. Έτσι για παράδειγμα, μπορεί να μάθει ότι μία λέξη έχει ειδική σημασία όταν βρίσκεται ανάμεσα στις ετικέτες <hashtag> </hashtag>.

5.6 Μοντέλα

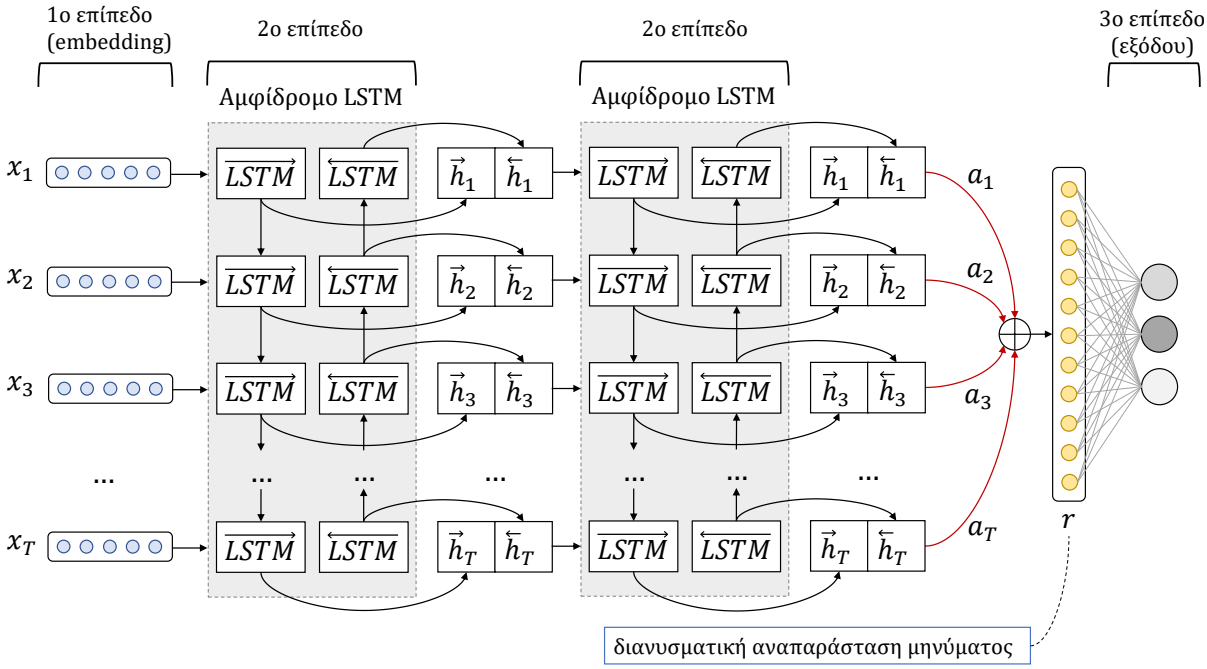
Στο σημείο αυτό θα παρουσιαστούν οι αρχιτεκτονικές των Τεχνητών Νευρωνικών Δικτύων τα οποία συμμετείχαν στις κατηγορίες του Semeval 2017 Task 4. Σχεδιάστηκαν δύο μοντέλα. Το πρώτο σχεδιάστηκε για το Subtask A, με στόχο την ταξινόμηση μηνυμάτων ανάλογα των γενικό συναισθηματικό προσανατολισμό τους (Message-level Sentiment Analysis - MSA). Το δεύτερο μοντέλο σχεδιάστηκε για τις υπόλοιπες υποκατηγορίες, οι οποίες αφορούν την ταξινόμηση και ποσοτικοποίηση βάση του συναισθήματος προς ένα θέμα δοσμένα θέμα (Topic-based Sentiment Analysis - TSA).

5.6.1 Συναίσθημα Μηνύματος - MSA

Το Νευρωνικό Δίκτυο για την ανάλυση του συναισθήματος στο επίπεδο του μηνύματος, αποτελείται από ένα βαθύ (2 επίπεδα) αμφίδρομο LSTM, εξοπλισμένο με έναν μηχανισμό attention, για την ενίσχυση των σημαντικών λέξεων σε ένα μήνυμα.

5.6.1.1 Επίπεδο Εμφύτευσης (Embedding)

Η είσοδος του δικτύου είναι ένα μήνυμα (tweet), το οποίο επεξεργάζεται, ως μία ακολουθία λέξεων. Χρησιμοποιώντας ένα επίπεδο εμφύτευσης (embedding), προβάλουμε τις λέξεις $X = (x_1, x_2, ..., x_T)$ σε ένα διανυσματικό χώρο R^E λίγων διαστάσεων, όπου E οι διαστάσεις (νευρώνες) του επιπέδου εμφύτευσης. Αρχικοποιούμε τα βάρη του επιπέδου με τα προ-εκπαιδευμένα διανύσματα λέξεων (Κεφ. 5.4.2).



Σχήμα 5.2: The MSA model: A 2-layer bidirectional LSTM with attention over that last layer.

5.6.1.2 Επίπεδα BiLSTM

Ένα LSTM δέχεται ως είσοδο τις λέξεις (embeddings) του μηνύματος και παράγει νέες διανυσματικές αναπαραστάσεις για κάθε μία $H = (h_1, h_2, ..., h_T)$, όπου κάθε h_i είναι η έξοδος του LSTM στο βήμα i, η οποία αποτελεί μία σύνοψη της πρότασης μέχρι και τη λέξη x_i. Χρησιμοποιούμε ένα *αμφίδρομο LSTM* (BiLSTM) ώστε να αποκτήσουμε αναπαραστάσεις οι οποίες συνοψίζουν την πρόταση και από τις δύο κατευθύνσεις.

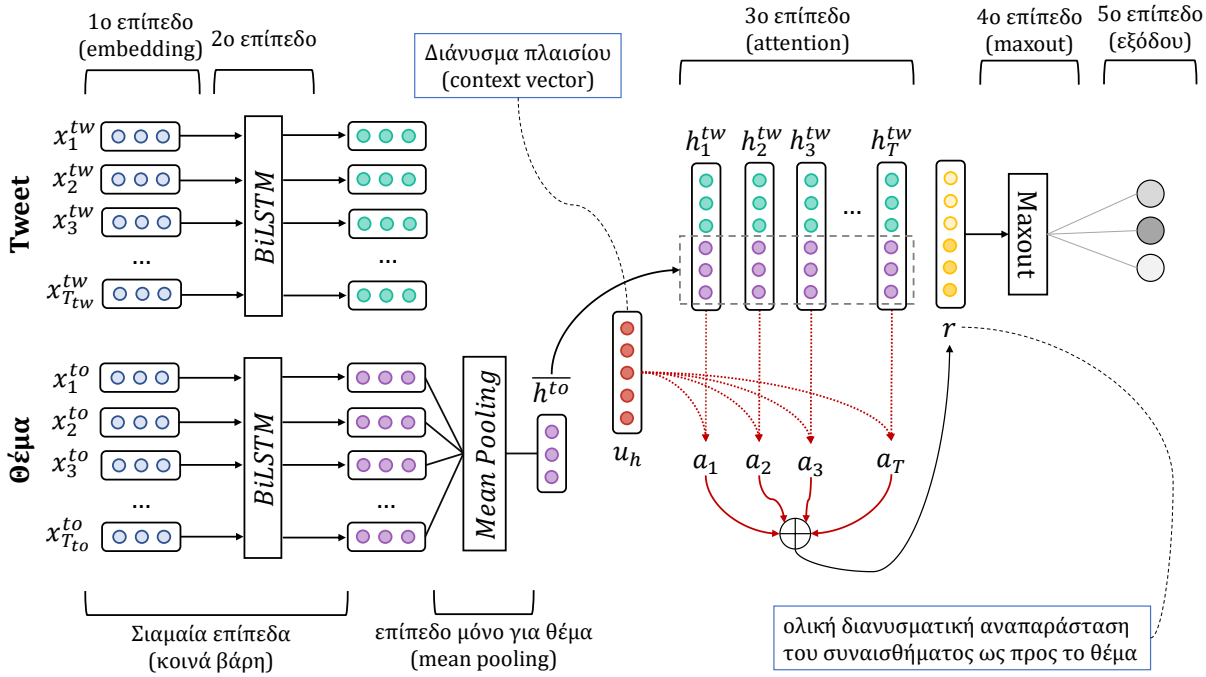
Ένα BiLSTM αποτελείται από ένα δεξιόστροφο LSTM $\overrightarrow{f}$ το οποίο διαβάζει την πρόταση από τη λέξη x_1 έως την x_T και ένα αριστερόστροφο $\overleftarrow{f}$ το οποίο διαβάζει την πρόταση από τη λέξη x_T έως τη x_1. Έτσι, για κάθε λέξη δημιουργούμε την αμφίδρομη αναπαράστασή της, ενώνοντας τις επιμέρους διανυσματικές αναπαραστάσεις:

$$h_i = \overrightarrow{h_i} \parallel \overleftarrow{h_i}, \quad h_i \in R^{2L} \tag{5.12}$$

όπου ο τελεστής $\parallel$ δηλώνει την ένωση διανυσμάτων και το L είναι οι διαστάσεις του LSTM. Επιπλέον, στοιβάζουμε δύο επίπεδα BiLSTM, ώστε το δίκτυο να μάθει πιο σύνθετα και αφηρημένα χαρακτηριστικά.

5.6.1.3 Επίπεδο Attention

Σε ένα μήνυμα, δεν συνεισφέρουν στον ίδιο βαθμό όλες οι λέξεις, για την έκφραση του συναισθήματος που φέρεται σε αυτό. Συνήθως, είναι λίγες λέξεις οι οποίες καθορίζουν το ύφος και την συναισθηματικό προσδιορισμό του μηνύματος. Για τον λόγο αυτό, χρησιμοποιούμε ένα μηχανισμό attention, ο οποίος βρίσκει την σχετική σημαντικότητα των λέξεων σε ένα μήνυμα. Ο μηχανισμός υπολογίζει για κάθε λέξη ένα συντελεστή βαρύτητας, τον οποίο χρησιμοποιούμε για να σταθμίσουμε την συμμετοχή των λέξεων στην τελική αναπαράσταση του μηνύματος. Συγ-



Σχήμα 5.3: The MSA model: A 2-layer bidirectional LSTM with attention over that last layer.

κεκριμένα, υπολογίζουμε την τελική αναπαράσταση ως το σταθμισμένο άθροισμα των κρυφών αναπαραστάσεων των λέξεων h_i:

$$e_i = tanh(W_h h_i + b_h), \quad e_i \in [-1, 1] \tag{5.13}$$

$$a_i = \frac{exp(e_i)}{\sum_{t=1}^{T} exp(e_t)}, \quad \sum_{i=1}^{T} a_i = 1 \tag{5.14}$$

$$r = \sum_{i=1}^{T} a_i h_i, \quad r \in R^{2L} \tag{5.15}$$

όπου W_h, b_h τα βάρη του επιπέδου, τα οποία μέσω της εκπαίδευσης του δικτύου μαθαίνουν να αποδίδουν μεγαλύτερη βαρύτητα στις πιο σημαντικές λέξεις. Το αποτέλεσμα του σταθμισμένου αθροίσματος είναι η διανυσματική αναπαράσταση ολόκληρου του μηνύματος r, η οποία χρησιμοποιείται ως διάνυσμα χαρακτηριστικών για ταξινόμηση.

5.6.1.4 Επίπεδο Εξόδου - Softmax

Το επίπεδο αυτό δέχεται ως είσοδο το διάνυσμα χαρακτηριστικών του μηνύματος και παράγει μία κατανομή πιθανοτήτων για τις κλάσεις (θετικό, αρνητικό, ουδέτερο). Το μήνυμα ταξινομείται στην κλάση με την μεγαλύτερη πιθανότητα. Η ταξινόμηση εκτελείται με την χρήση ενός FFNN με έναν νευρώνα για κάθε κλάση, το οποίο εκτελεί Λογιστική Παλινδρόμηση εφαρμόζοντας την *softmax* συνάρτηση ενεργοποίησης.

5.6.2 Συναίσθημα Θέματος - TSA

Αυτό το δίκτυο σχεδιάστηκε για την ταξινόμηση μηνυμάτων ως προς το συναισθηματικό προσανατολισμό τους, προς ένα δοσμένο θέμα, το οποίο αναφέρεται στο μήνυμα. Η αρχιτεκτονική του δικτύου αποτελείται από ένα σύνολο από *σιαμαία επίπεδα* για την παραγωγή των διανυσματικών αναπαραστάσεων του μηνύματος και του θέματος.

Σιαμαία λέγονται τα δίκτυα τα οποία έχουν όμοιες αρχιτεκτονικές και τα βάρη τους είναι συνδεδεμένα κατά τη φάση της εκπαίδευσης. Με αυτό τον τρόπο, τα σιαμαία δίκτυα παράγουν την ίδια έξοδο για μία συγκεκριμένη είσοδο. Ο λόγος για τον οποίο επιλέγεται η αρχιτεκτονική αυτή, είναι γιατί σε αυτό το πρόβλημα, το δίκτυο έχει δύο εισόδους, το μήνυμα και το θέμα. Έτσι, χρησιμοποιούμε μία σιαμαία αρχιτεκτονική, ώστε οι λέξεις του μηνύματος και του θέματος, να προβληθούν στο ίδιο διανυσματικό χώρο και με αυτό τον τρόπο να μπορεί να γίνει σύγκριση ανάμεσά τους. Σιαμαία είναι τα επίπεδα εμφύτευσης και biLSTM.

5.6.2.1 Επίπεδο Εμφύτευσης (Embedding)

Το δίκτυο δέχεται ως είσοδο δύο ακολουθίες: (1) τις λέξεις του μηνύματος $X^{tw} = (x_1^{tw}, x_2^{tw}, ..., x_{T_{tw}}^{tw})$, όπου T_{tw} το πλήθος των λέξεων του μηνύματος και (2) τις λέξεις του θέματος $X^{to} = (x_1^{to}, x_2^{to}, ..., x_{T_{to}}^{to})$, όπου T_{to} το πλήθος των λέξεων του θέματος. Οι λέξεις προβάλλονται στον χώρο R^E, όπου E οι διαστάσεις του επιπέδου εμφύτευσης. Τα βάρη του επιπέδου αρχικοποιούνται με τα προ-εκπαιδευμένα διανύσματα λέξεων (Κεφ. 5.4.2).

5.6.2.2 Σιαμαίο BiLSTM

Χρησιμοποιούμε ένα σιαμαίο BiLSTM, με *κοινά βάρη* και για τις δύο εισόδους, ώστε να προβάλουμε και τις δύο ακολουθίες σε *κοινό* διανυσματικό χώρο. Το BiLSTM παράγει τις κρυφές αναπαραστάσεις για τις λέξεις του μηνύματος $H^{tw} = (h_1^{tw}, h_2^{tw}, ..., h_{T_{tw}}^{tw})$ και του θέματος $H^{to} = (h_1^{to}, h_2^{to}, ..., h_{T_{to}}^{to})$, όπου κάθε αναπαράσταση αποτελείται από την ένωση των αναπαραστάσεων και από τις δύο κατευθύνσεις ($\overrightarrow{LSTM}$, $\overleftarrow{LSTM}$):

$$h_i^j = \overrightarrow{h_i^j} \parallel \overleftarrow{h_i^j}, \quad h_i^j \in R^{2L}, \quad j \in \{tw, to\} \tag{5.16}$$

όπου ο τελεστής $\parallel$ δηλώνει την ένωση διανυσμάτων και το L είναι οι διαστάσεις του LSTM.

5.6.2.3 Επίπεδο Συγκέντρωσης (Pooling)

Στο επίπεδο αυτό συνδυάζονται οι κρυφές αναπαραστάσεις των λέξεων του θέματος, για τον σχηματισμό μίας ενιαίας διανυσματικής αναπαράστασης. Ο τρόπος με τον οποίο συγκεντρώνονται όλα τα επιμέρους διανύσματα, είναι με τον υπολογισμό του *μέσου όρου στο χρόνο* (mean over time), παράγοντας το διάνυσμα του θέματος $\overline{h^{to}}$:

$$\overline{h^{to}} = \frac{1}{T_{to}} \sum_{1}^{T_{to}} h_i^{to} \tag{5.17}$$

5.6.2.4 Αναπαραστάσεις Πλαισίου

Μέχρι αυτό το σημείο έχουμε μία ακολουθία με τις κρυφές αναπαραστάσεις H^{tw} (διανύσματα) των λέξεων του μηνύματος και την ενιαία πλέον αναπαράσταση (διάνυσμα) του θέματος $\overline{h^{to}}$. Σαν επόμενο βήμα, συμπληρώνουμε (ενώνουμε) το διάνυσμα του θέματος, σε *κάθε* ένα από τα διανύσματα των λέξεων. Με αυτό τον τρόπο δημιουργούμε νέες αναπαραστάσεις για κάθε λέξη, οι οποίες περιέχουν το πλαίσιο στο οποίο αναφέρονται (το θέμα). Συγκεκριμένα:

$$h_i = h_i^{tw} \parallel \overline{h^{to}}, \quad h_i^j \in R^{4L} \tag{5.18}$$

5.6.2.5 Επίπεδο Context-Attention

Αυτό το επίπεδο υλοποιεί ένα μηχανισμό attention όπως στο (Z. Yang κ.ά. 2016), με στόχο την ανακάλυψη των πιο σημαντικών λέξεων, οι οποίες καθορίζουν τον συναισθηματικό προσανατολισμό, προς το δοθέν θέμα. Ο μηχανισμός αυτός χρησιμοποιεί ένα διάνυσμα u_h το οποίο κωδικοποιεί το πλαίσιο στο οποίο λειτουργεί ο μηχανισμός. Αφού έχουν υπολογιστεί οι κρυφές αναπαραστάσεις των λέξεων, εκτελεί ένα "ερώτημα" πάνω σε αυτές, της μορφής "ποιες είναι οι συναισθηματικά φορτισμένες λέξεις προς το θέμα". Το ερώτημα έχει την έννοια του ερωτήματος στην Ανάκτηση Πληροφορίας (Information Retrieval) και με τον τρόπο αυτό παράγεται ένα νέο διάνυσμα για κάθε λέξη ($e_i^\top u_h$):

$$e_i = tanh(W_h h_i + b_h), \quad e_i \in [-1, 1] \tag{5.19}$$

$$a_i = \frac{exp(e_i^\top u_h)}{\sum_{t=1}^{T_{tw}} exp(e_t^\top u_h)}, \quad \sum_{i=1}^{T_{tw}} a_i = 1 \tag{5.20}$$

$$r = \sum_{i=1}^{T_{tw}} a_i h_i, \quad r \in R^{4L} \tag{5.21}$$

5.6.2.6 Επίπεδο Maxout

Αφού έχουμε παραγάγει την ενιαία διανυσματική αναπαράσταση r, την προωθούμε ως είσοδο σε ένα επίπεδο Maxout (Goodfellow κ.ά. 2013), ώστε να εκτελέσει την σύγκριση ανάμεσα στις πτυχές του θέματος και του μηνύματος. Επιλέγουμε το Maxout, καθώς ενισχύει τα αποτελέσματα του dropout 5.8.

5.6.2.7 Επίπεδο Εξόδου - Softmax

Το επίπεδο αυτό δέχεται ως είσοδο το διάνυσμα το οποίο αποτελεί το αποτέλεσμα της σύγκρισης και παράγει μία κατανομή πιθανοτήτων για τις κλάσεις του αντίστοιχου προβλήματος (2 ή 5). Το μήνυμα ταξινομείται στην κλάση με την μεγαλύτερη πιθανότητα. Η ταξινόμηση εκτελείται με την χρήση ενός FFNN με έναν νευρώνα για κάθε κλάση, το οποίο εκτελεί Λογιστική Παλινδρόμηση εφαρμόζοντας την *softmax* συνάρτηση ενεργοποίησης.

5.7 Ποσοτικοποίηση

Σκοπός της ποσοτικοποίησης (quantification) είναι η εκτίμηση της κατανομής των μηνυμάτων σε ένα σύνολο κλάσεων. Για την εκτίμηση της κατανομής γίνεται χρήση ενός ταξινομητή ο οποίος έχει εκπαιδευτεί στο αντίστοιχο πρόβλημα. Αυτό σημαίνει ότι για την εκτίμηση της κατανομής των μηνυμάτων για το Subtask D, θα χρησιμοποιηθεί ο ταξινομητής του Subtask B και αντίστοιχα για το Subtask E θα χρησιμοποιηθεί ο ταξινομητής του Subtask C. Επίσης, το ζητούμενο των Subtask D και E είναι η εκτίμηση της κατανομής των μηνυμάτων, για κάθε θέμα ξεχωριστά. Δηλαδή, για το Subtask D, για κάθε θέμα, θα πρέπει να εκτιμήσουμε τι ποσοστό των μηνυμάτων που αναφέρονται σε αυτό είναι θετικά και τι ποσοστό είναι αρνητικά.

Η πιο απλή προσέγγιση είναι η *Ταξινόμηση και Καταμέτρηση* (Classify & Count - CC) (Forman 2008), στην οποία ταξινομούμε όλες τις άγνωστες παρατηρήσεις (μηνύματα), χρησιμοποιώντας έναν εκπαιδευμένο ταξινομητή και στη συνέχεια υπολογίζουμε τι ποσοστό από τις παρατηρήσεις ανήκει σε κάθε κλάση. Όμως, όπως έχει σημειώνεται και στο (Gao κ.ά. 2016), ένας καλός ταξινομητής δεν συνεπάγεται ότι θα είναι και ένας καλός ποσοτικοποιητής. Αυτό είναι γίνεται εμφανές με το εξής παράδειγμα: έστω ότι υπάρχουν δύο ταξινομητές σε ένα πρόβλημα δυαδικής ταξινόμησης (2 κλάσεις). Ο ταξινομητής A έχει FP[5] $= 10$ και $FN = 10$ και ο ταξινομητής B έχει $FP = 8$ και $FN = 10$. Αν και ο B είναι καλύτερος ταξινομητής, καθώς κάνει λιγότερα λάθη, ο A είναι καλύτερος ποσοτικοποιητής. Ο λόγος είναι ότι τα λάθη του A συμπληρώνουν το ένα το άλλο, με αποτέλεσμα να δίνει καλύτερη εκτίμηση της κατανομής. Το συμπέρασμα είναι ότι ένας καλό ποσοτικοποιητής θα πρέπει να έχει μία μικρή μεροληψία (bias).

Μία άλλη προσέγγιση είναι η χρήση ενός ταξινομητή, ο οποίος παράγει μία κατανομή πιθανοτήτων για κάθε παρατήρηση, η οποία ονομάζεται *Πιθανοτική Ταξινόμηση και Καταμέτρηση* (Probabilistic Classify & Count - PCC) (Gao κ.ά. 2016). Με αυτή την τεχνική, παράγουμε την συνολική εκτίμηση, υπολογίζοντας τον μέσο όρο των επιμέρους κατανομών. Με την ίδια λογική υπολογίζουμε την κατανομή των κλάσεων για τα μηνύματα ενός θέματος. Συγκεκριμένα, έστω T το σύνολο των θεμάτων στο training set και $p(c|tweet)$ η εκ των υστέρων πιθανότητα ότι ένα μήνυμα (tweet) ανήκει στην κλάση c, όπως έχει εκτιμηθεί από τον ταξινομητή. Στη συνέχεια, εκτιμούμε το ποσοστό (πιθανότητα) των μηνυμάτων του θέματος, τα οποία ανήκουν στην κλάση c ως εξής:

$$\hat{p}_T(c) = \frac{1}{|T|} \sum_{tweet \in T} p(c|tweet) \tag{5.22}$$

Όπως δείχνεται στην μελέτη (Gao κ.ά. 2016), όπου συγκρίνονται διάφορες τεχνικές ποσοτικοποίησης για ΑΣ στο Twitter, η PCC πετυχαίνει καλύτερα αποτελέσματα από την CC. Βάση αυτών των αποτελεσμάτων, επιλέχθηκε χρήση της PCC για την ποσοτικοποίηση στα Subtask D και E.

[5]FP (False Positive) είναι το πλήθος των παρατηρήσεων που λανθασμένα έχουν ταξινομηθεί στη θετική κλάση και ομοίως FN (False Negative) το πλήθος των παρατηρήσεων που λανθασμένα έχουν ταξινομηθεί στην αρνητική κλάση.

5.8 Εξομάλυνση - Regularization

Για την εξομάλυνση των μοντέλων, προσθέτουμε θόρυβο (ακολουθώντας κανονική κατανομή) στο επίπεδο εμφύτευσης. Επίσης μετά την εισαγωγή του θορύβου, εφαρμόζουμε dropout στο επίπεδο εμφύτευσης (Srivastava κ.ά. 2014). Το dropout είναι μία τεχνική, η οποία απενεργοποιεί σε κάθε παράδειγμα εκπαίδευσης, τυχαία ένα ποσοστό των νευρώνων ενός δικτύου. Αυτό έχει ως συνέπεια την μάθηση πιο "ανεξάρτητων" χαρακτηριστικών από κάθε νευρώνα, καθώς κάθε νευρώνας μαθαίνει να μην βασίζεται σε μεγάλο βαθμό σε άλλους. Αυτά τα δύο βήματα, αποτελούν κατά κάποιο τρόπο ένα βήμα τυχαίας ενίσχυσης δεδομένων (random data augmentation), το οποίο κάνει το μοντέλο πιο ανθεκτικό στην υπερπροσαρμογή. Ο λόγος είναι ότι το δίκτυο δεν συναντά ποτέ την ίδια ακριβώς πρότασης, κατά την εκπαίδευση. Έτσι μαθαίνει να γενικεύει καλύτερα. Επιπλέον, εφαρμόζουμε dropout στις επαναλαμβανόμενες συνδέσεις των LSTM όπως στο (Gal κ.ά. 2016). Ακόμη, προσθέτουμε έναν όρο αποσύνθεσης βαρών $L2$, ώστε να αποτρέψουμε τον σχηματισμό μεγάλων βαρών στο δίκτυο, κάτι το οποίο οδηγεί σε υπερπροσαρμογή.

Τέλος, ένα ακόμη βήμα για την καλύτερη γενίκευση των μοντέλων ήταν ότι "παγώσαμε" τα επίπεδα εμφύτευσης κατά την εκπαίδευσης. Αυτό σημαίνει ότι τα επίπεδα εμφύτευσης έμειναν σταθερά, με τις τιμές τους να αντιστοιχούν σε αυτές των διανυσμάτων λέξεων τα οποία χρησιμοποιήθηκαν για να τα αρχικοποιήσουν. Ο λόγος που έγινε αυτό, ήταν για να μην μετακινηθούν οι λέξεις στον διανυσματικό χώρο σε νέες θέσεις, οι οποίες δεν θα ανταποκρίνονται στον πραγματικό προσανατολισμό όλων των λέξεων. Αυτό αναλύθηκε σε μεγαλύτερο βάθος στο Κεφάλαιο 3.2.3.3.

5.9 Training

Όλα τα δίκτυα εκπαιδεύονται για την ελαχιστοποίηση του κόστους διεντροπίας (cross-entropy loss). Τα βάρη ενημερώνονται με back-propagation με SGD και mini-batches μεγέθους 128. Επίσης χρησιμοποιούμε τον βελτιστοποιητή Adam (Kingma κ.ά. 2014) για την αυτόματη ρύθμιση του ρυθμού μάθησης.

5.9.1 Στάθμιση Κλάσεων

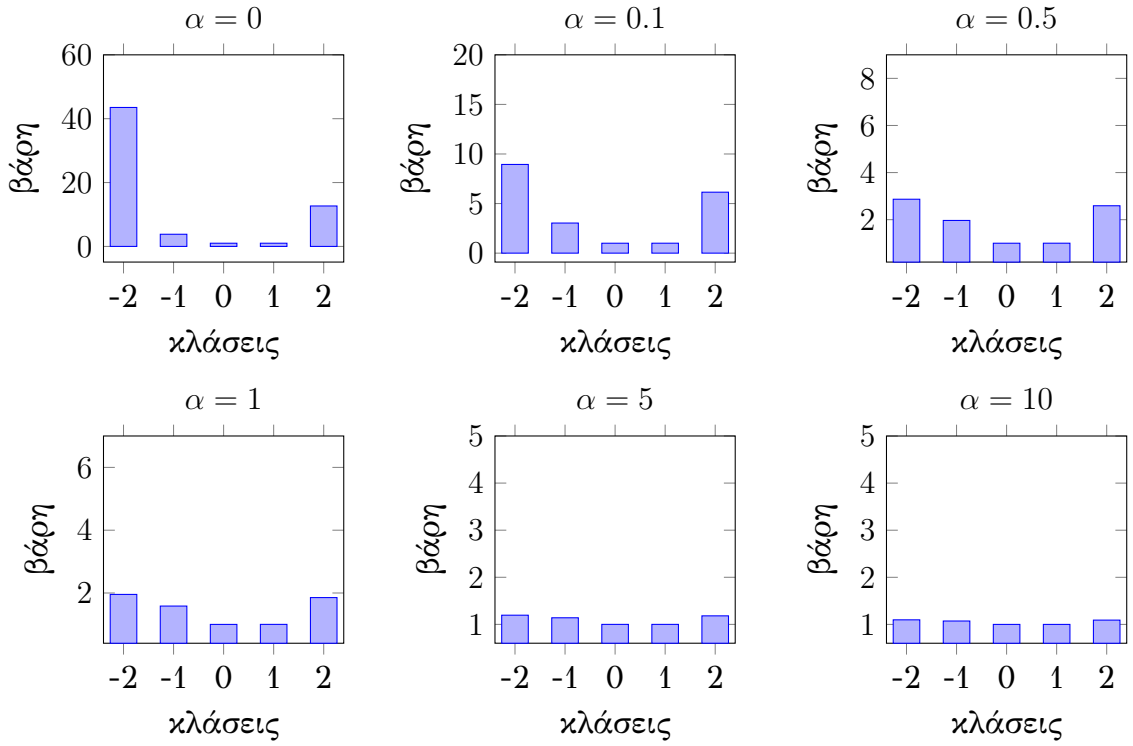
Σε όλα τα σύνολα δεδομένων των υποκατηγοριών, οι κλάσεις είναι δυσανάλογες, με αποτέλεσμα το μοντέλο να μεροληπτεί υπέρ των κλάσεων με το μεγαλύτερο πλήθος. Για παράδειγμα, επειδή συναντά πιο συχνά μηνύματα της κλάσης *ουδέτερο*, είναι πιο πιθανό να ταξινομήσει ένα νέο μήνυμα ως ουδέτερο. Για να αντιμετωπίσουμε αυτό το πρόβλημα, εισάγουμε βάρη στην αντικειμενική συνάρτηση, με σκοπό να τιμωρήσουμε πιο έντονα τη λάθος ταξινόμηση των μηνυμάτων των υποεκπροσωπούμενων κλάσεων. Αυτό σημαίνει, ότι αν ο ταξινομητής κάνει λάθος ταξινόμηση σε ένα μήνυμα το οποίο ανήκει στην κλάση *αρνητικό*, η αντικειμενική συνάρτηση θα παράγει μεγαλύτερο σφάλμα, από ότι αν κάνει λάθος ταξινόμηση σε ένα μήνυμα της κλάσης *ουδέτερο*.

Επιπλέον, εισάγουμε έναν συντελεστή εξομάλυνσης (additive smoothing factor), ώστε να περιορίσουμε τα υπερβολικά μεγάλα βάρη. Αυτό το πρόβλημα είναι πολύ έντονο στην περίπτωση των Subtask C και E, όπου οι διαφορές των κλάσεων είναι πολύ μεγάλες (Πίνακας 5.2). Χωρίς την εισαγωγή του συντελεστή, κλάσεις όπως

η *πολύ_αρνητικό*, αποκτούν 40 φορές μεγαλύτερη βαρύτητα από την *ουδέτερο*. Η εξίσωση για τον υπολογισμό των βαρών των κλάσεων υπολογίζεται ως εξής:

$$w_i = \frac{max(x)}{x_i + \alpha \times max(x)} \tag{5.23}$$

όπου x, το διάνυσμα με τον αριθμό των παρατηρήσεων για κάθε κλάση και α ο συντελεστής εξομάλυνσης. Σε όλα τα Subtasks θέτουμε $\alpha = 0$ και στα Subtasks C και E θέτουμε $\alpha = 0.1$.



Σχήμα 5.4: Επίδραση του συντελεστή εξομάλυνσης α, στις τιμές των βαρών κάθε κλάσης, για τα δεδομένα των Subtask C και E. Όταν το $\alpha = 0$, δεν γίνεται καθόλου εξομάλυνση. Όσο η τιμή του α αυξάνεται, τόσο μειώνονται οι διαφορές.

5.9.2 Υπέρ-παράμετροι

Οι υπέρ-παράμετροι του δικτύου είναι αυτές οι οποίες ρυθμίζουν την αρχιτεκτονική του δικτύου, όπως το πλήθος των νευρώνων (διαστάσεις) κάθε επιπέδου, ο ρυθμός μάθησης ή η εξομάλυνση. Οι περισσότερες από αυτές τις τιμές αλληλοεξαρτώνται. Αυτό σημαίνει ότι η εύρεση των ιδανικών τιμών για όλες τις παραμέτρους, είναι ένα πολυδιάστατο πρόβλημα. Η δυσκολία γίνεται ακόμα μεγαλύτερη, με δεδομένο ότι ο χρόνος εκπαίδευσης ενός ΤΝΔ είναι πολλαπλάσιος αυτού ενός άλλου αλγορίθμου μηχανικής μάθησης. Επίσης, οι επιδόσεις των ΤΝΔ επηρεάζονται πολύ από τις τιμές των υπέρ-παραμέτρων.

Για την εύρεση των ιδανικών τιμών, συνήθως επιλέγονται δύο προσεγγίσεις: η *αναζήτηση πλέγματος* (grid search) και η *τυχαία αναζήτηση* (random search). Η αναζήτηση πλέγματος δοκιμάζει *κάθε* συνδυασμό από ένα σύνολο προκαθορισμένων τιμών για κάθε παράμετρο. Αυτό σημαίνει ότι ο χρόνος εκτέλεσης όλων

των συνδυασμών είναι υπερβολικά μεγάλος. Έτσι, όταν επιλέγεται αυτή η προσέγγιση επιλέγεται ένα σχετικά μικρό πλήθος τιμών για κάθε παράμετρο ώστε να περιοριστούν οι συνδυασμοί. Συνεπώς, τις περισσότερες φορές είναι πολύ δύσκολο να βρεθούν καλές τιμές σε μικρό χρόνο. Από την άλλη μεριά, όπως έχει δειχθεί (Bergstra κ.ά. 2012) η τυχαία αναζήτηση πετυχαίνει καλύτερα αποτελέσματα. Αν και δεν εγγυάται την εύρεση του βέλτιστου συνδυασμού, βρίσκει έναν αρκετά καλό συνδυασμό τιμών σε αρκετά μικρότερο χρόνο. Ειδικά στην περίπτωση των ΤΝΔ, είναι η πιο συνηθισμένη προσέγγιση.

Όμως, και οι δύο αυτές τεχνικές κάνουν μία τυφλή αναζήτηση στον χώρο των παραμέτρων. Η εναλλακτική αφορά τεχνικές οι οποίες σε κάθε δοκιμή λαμβάνουν υπόψη τους την πορεία των αποτελεσμάτων, για την επιλογή του νέου συνδυασμού. Μπορεί για παράδειγμα, η μείωση του πλήθους των νευρώνων σε ένα επίπεδο, να επιδεινώνει δραστικά τις επιδόσεις του μοντέλου. Σε μία τέτοια περίπτωση, δεν έχει νόημα να δοκιμαστούν άλλοι συνδυασμοί προς αυτή την κατεύθυνση.

Μία από αυτές τις τεχνικές είναι η *Μπεϋζιανή βελτιστοποίηση* (Bergstra κ.ά. 2013). Η τεχνική αυτή χτίζει ένα πιθανοτικό μοντέλο για την αντικειμενική συνάρτηση. Τα αποτελέσματα κάθε δοκιμής, χρησιμοποιούνται ως δεδομένα για την ενημέρωση των πεποιθήσεων του μοντέλου. Με αυτό τον τρόπο, αν για παράδειγμα μικρές τιμές μίας παραμέτρου οδηγούν σε κακά αποτελέσματα, τότε θα είναι πιο δύσκολο να επιλεχθεί μία μικρή τιμή για την παράμετρο αυτή σε επόμενες δοκιμές.

Το πλεονέκτημα αυτής της Μπεϋζιανής βελτιστοποίηση, είναι ότι αρκετές φορές βρίσκει έναν καλό συνδυασμό παραμέτρων σε ακόμη μικρότερο χρόνο. Υπάρχει όμως ο κίνδυνος το μοντέλο να κατευθυνθεί προς μία μη ιδανική θέση και να εγκλωβιστεί σε αυτή (τοπικό ελάχιστο). Για τον λόγο αυτό, η προσέγγιση που ακολουθήσαμε ήταν η εξής: στην αρχή εκτελέσαμε έναν αριθμό δοκιμών με τυχαία αναζήτηση και στη συνέχεια, χρησιμοποιώντας τα αποτελέσματα αυτά, συνεχίσαμε με Μπεϋζιανή βελτιστοποίηση. Με αυτό τον τρόπο μειώνουμε τον κίνδυνο το μοντέλο να παγιδευτεί, καθώς θα έχει ήδη "δει" ένα σύνολο τυχαίων συνδυασμών.

MSA Model Για το MSA μοντέλο, οι καλύτεροι παράμετροι που βρήκαμε είναι οι εξής:

- Το επίπεδο εμφύτευσης έχει μέγεθος 300.
- Τα επίπεδα με τα αμφίδρομα LSTM έχουν μέγεθος 150 για κάθε LSTM (300 σύνολο).
- Εισάγουμε θόρυβο (με Γκαουσιανή κατανομή) με $\sigma = 0.2$ στο επίπεδο εμφύτευσης.
- Εφαρμόζουμε dropout με συντελεστή 0.3 στο επίπεδο εμφύτευσης, 0.5 στο LSMT και 0.25 στις αναδρομικές του συνδέσεις.
- Προσθέτουμε ένα L_2 εξομάλυνση στην αντικειμενική συνάρτηση, με συντελεστή 0.0001.

TSA Model Για το TSA μοντέλο, οι καλύτεροι παράμετροι που βρήκαμε είναι οι εξής:

- Το επίπεδο εμφύτευσης έχει μέγεθος 300.

- Τα LSTM έχουν μέγεθος 64 το καθένα (128 ως αμφίδρομα).
- Εισάγουμε θόρυβο (με Γκαουσιανή κατανομή) με $\sigma = 0.2$ στο επίπεδο εμφύτευσης.
- Εφαρμόζουμε dropout με συντελεστή 0.3 στο επίπεδο εμφύτευσης, 0.2 στο LSMT, τόσο στις εξόδους του, όσο και στις αναδρομικές συνδέσεις, 0.3 στο επίπεδο για το attention και 0.3 στο επίπεδο Maxout.
- Προσθέτουμε ένα L_2 εξομάλυνση στην αντικειμενική συνάρτηση, με συντελεστή 0.001.

5.10 Αποτελέσματα

Τα αποτελέσματά των μοντέλων στην επίσημη κατάταξη του Semeval-2017 ήταν:

- Subtask A: 1/38 (ισοβαθμία).
- Subtask B: 2/24.
- Subtask C: 2/16.
- Subtask D: 2/16.
- Subtask E: 11/12.

Όπως φαίνεται όλα τα μοντέλα πέτυχαν πολύ καλά αποτελέσματα με εξαίρεση το μοντέλο του Subtask E. Το αξιοσημείωτο είναι ότι για τον υπολογισμό των προβλέψεων του Subtask E, το οποίο αφορούσε ποσοτικοποίηση, έγινε χρήση του εκπαιδευμένου ταξινομητή του Subtask C, το οποίο πέτυχε της 2η θέση. Στην ανάλυση των αποτελεσμάτων, δίνονται ορισμένες ερμηνείες και πιθανές αιτίες για αυτό το αποτέλεσμα.

5.10.1 Σύγκριση με άλλα μοντέλα

Αρχικά, θα δούμε μία σύγκριση του μοντέλου το οποίο συμμετείχε στον διαγωνισμό, με άλλα μοντέλα τα οποία δεν κάνουν χρήση ΤΝΔ. Σε όλα τα μοντέλα θα εφαρμοστεί η ίδια ακριβώς προεπεξεργασία για τη κείμενα, ώστε να είναι συγκρίσιμα τα αποτελέσματα. Τα μοντέλα, θα συγκριθούν στο Subtask A, το οποίο είναι πιο απλό και έτσι είναι πιο εύκολη η σύγκριση. Για την εκπαίδευση θα χρησιμοποιηθούν όλα τα δεδομένα εκπαίδευσης του Subtask και για την αξιολόγηση θα χρησιμοποιηθούν τα επίσημα δεδομένα αξιολόγησης του διαγωνισμού.

BoW: Είναι το απλό Bag-of-Word μοντέλο. Για το μοντέλο αυτό αφαιρέθηκαν όλες οι λέξεις οι οποίες δεν εμφανίζονται σε τουλάχιστον πέντε tweets. Δοκιμάστηκαν δύο παραλλαγές:

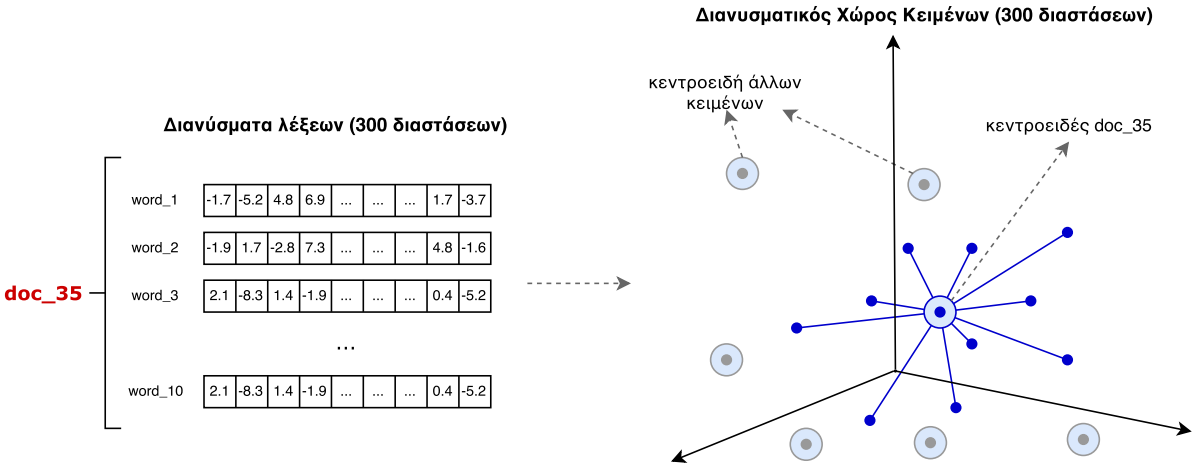
- μία με επιλογή των 5000 καλύτερων λέξεων, βάση της αμοιβαίας πληροφορίας τους (MI)
- και μία στην οποία οι τιμή κάθε λέξης (διάστασης), ορίζεται βάση της TF-IDF τιμής της λέξης

Features	$recall^M$	$F1^{pn}$	$precision^M$
BoW +{MI}	0.615	0.595	0.604
BoW +{TF-IDF}	0.618	0.598	0.596
Neural-BoW	0.644	0.636	0.621
Neural-BoW +{TF-IDF}	0.642	0.632	0.62
Neural-BoW (NegContext)	0.645	0.636	0.622
Neural-BoW (NegContext+diff)	0.645	0.636	0.622
CNN (Kim 2014)	0.659	0.650	0.624
AttentionRNN	0.681	0.677	0.651

Πίνακας 5.6: Σύγκριση μοντέλων, στα δεδομένα του Subtask A

Neural-BoW: Το μοντέλο αυτό χρησιμοποιεί για να αναπαραστήσει ένα μήνυμα, το κεντροειδές (κέντρο βάρους) των διανυσματικών αναπαραστάσεων των λέξεων του μηνύματος. Για τις διανυσματικές αναπαραστάσεις κάθε λέξης, χρησιμοποιούνται τα προεκπαιδευμένα embeddings. Σχηματικά, τα παραπάνω φαίνονται στην Εικόνα 5.5. Δοκιμάστηκαν δύο παραλλαγές:

- κεντροειδές διανυσμάτων λέξεων
- κεντροειδές διανυσμάτων λέξεων, όπου το διάνυσμα κάθε λέξης σταθμίζεται βάση της TF-IDF τιμής της λέξης.

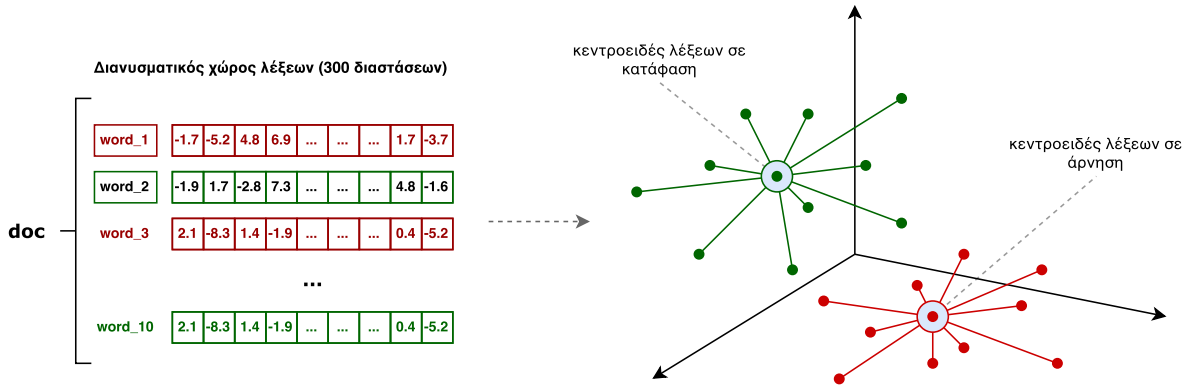


Σχήμα 5.5: Neural-BoW μοντέλο.

Neural-BoW (NegContext) : Το μοντέλο αυτό αναπαριστά ένα tweet χρησιμοποιώντας δύο διαφορετικά κεντροειδή. Το ένα κεντροειδές υπολογίζεται από λέξεις οι οποίες βρίσκονται εντός κατάφασης και το άλλο υπολογίζεται από λέξεις οι οποίες βρίσκονται εντός άρνησης. Τα δύο επιμέρους κεντροειδή συνδυάζονται (με ένωση) για την δημιουργία της διανυσματικής αναπαράστασης του tweet. Σχηματικά, τα παραπάνω φαίνονται στην Εικόνα 5.6. Δοκιμάστηκαν δύο παραλλαγές:

- κεντροειδές διανυσμάτων λέξεων, από ένωση επιμέρους κεντροειδών (κατάφαση, άρνηση). Δηλαδή $doc_i = (cent_{pos}, cent_{neg})$.

- κεντροειδές διανυσμάτων λέξεων, από ένωση επιμέρους κεντροειδών, αλλά και της απόλυτης διαφοράς τους. Δηλαδή $doc_i = (cent_{pos}, cent_{neg}, |cent_{pos} - cent_{neg}|)$.



Σχήμα 5.6: Neural-BoW (NegContext) μοντέλο.

5.10.2 Μηχανισμός Προσοχής

Για να αξιολογήσουμε την συνεισφορά του μηχανισμού προσοχής, αξιολογήσαμε τις επιδόσεις των μοντέλων με και χωρίς τον μηχανισμό. Στον Πίνακα 5.7 καταγράφονται τα αποτελέσματα με τον μέσο όρο για κάθε κριτήριο, από την εκπαίδευση 10 μοντέλων, στο επίσημο σύνολο δεδομένων προς αξιολόγηση (test set). Ο λόγος που εκπαιδεύτηκε κάθε δίκτυο από 10 φορές είναι για να αυξηθεί η στατιστική σημαντικότητα των αποτελεσμάτων.

RNN	Subtask A (MSA)		Subtask B (TSA)	
	ρ	$F1^{pn}$	ρ	$F1^{pn}$
Regular	0.678	0.673	0.856	0.817
Attention	**0.682**	**0.675**	**0.863**	**0.82**

Πίνακας 5.7: Αποτελέσματα της συνεισφοράς του μηχανισμού προσοχής. Το μέτρο ρ είναι η μέση ανάκλιση και $F1^{pn}$ macro-μέσο $F1$ σκορ για τα μηνύματα των κλάσεων *{θετικό, αρνητικό}*.

Όπως φαίνεται και στον πίνακα, ο μηχανισμός προσοχής όντως βελτιώνει τις επιδόσεις των μοντέλων. Όμως, η βελτίωση δεν ήταν αρκετά σημαντική. Ένας πιθανός λόγος ίσως είναι ότι το υπόλοιπο δίκτυο είναι ήδη αρκετά ικανό να περιγράψει τα δεδομένα, με αποτέλεσμα να μην αφήνει μεγάλο περιθώριο βελτίωσης.

5.10.3 Ποσοτικοποίηση

Για να αποκτήσουμε μία καλύτερη εικόνα σχετικά με τις προσεγγίσεις για ποσοτικοποίηση, συγκρίνουμε τις τεχνικές CC και PCC. Όπως φαίνεται, τόσο από την βιβλιογραφία, όσο και από τα αποτελέσματα των μοντέλων στον διαγωνισμό, δεν είναι ξεκάθαρο ποια από τις δύο τεχνικές είναι προτιμότερη. Η PCC ξεπερνά την CC στο (Bella κ.ά. 2010) αλλά υστερεί στη CC στο (Esuli κ.ά. 2015).

Ακολουθώντας τα αποτελέσματα του (Gao κ.ά. 2016), τα οποία αφορούν ΑΣ στο Twitter, αποφασίσαμε να επιλέξουμε την PCC. Θεωρήσαμε πως η ομοιότητα του πεδίου εφαρμογής της τεχνικής, ανάμεσα στη μελέτη και στον διαγωνισμό, ήταν το πιο σημαντικό κριτήριο. Ο Πίνακας 5.8, παρουσιάζει τα αποτελέσματα των τεχνικών ποσοτικοποίησης, στα δεδομένα αξιολόγησης του διαγωνισμού.

Method	Subtask D			Subtask E
	KLD	AE	RAE	EMD
CC	0.060	**0.093**	**0.608**	**0.359**
PCC	**0.048**	0.095	0.848	0.595

Πίνακας 5.8: Αποτελέσματα τεχνικών ποσοτικοποίησης. Το μέτρο KLD αναφέρεται στη Kullback-Leibler Divergence, το EMD στη Earth Mover's Distance, το AE στο απόλυτο σφάλμα (Absolute Error) και το RAE στο σχετικό απόλυτο σφάλμα (Relative Absolute Error). Για όλες τις μετρικές, οι χαμηλότερες τιμές είναι καλύτερες.

Η PCC ξεπερνά την CC στο Subtask D, αλλά είναι κατά πολύ χειρότερη στο Subtask E. Ορισμένες πιθανές αιτίες για την αναντιστοιχία ανάμεσα στις σχετικές επιδόσεις των δύο τεχνικών είναι:

- Η διαφορά στο πλήθος των κλάσεων. Στο Subtask D έχουμε 2 κλάσεις, ενός στο Subtask E έχουμε 5 κλάσεις.
- Η αλλαγή στην αναλογία των κλάσεων *{θετικό, αρνητικό}*. Είναι πιθανό η PCC, η οποία βασίζεται στις κατανομές πιθανότητας κάθε παρατήρησης και όχι στην τελική κλάση που προβλέφθηκε, να είναι επιρρεπής σε αυτή τη διαφορά.
- Η εισαγωγή του όρου εξομάλυνσης στα Subtasks C και E. Όπως και πριν, το γεγονός ότι η PCC βασίζεται στις κατανομές πιθανότητας, μπορεί να επηρεάστηκε από την μειωμένη μεροληψία (bias) και να έπεσε πολύ έξω στις προβλέψεις της.

5.10.4 Επίσημα Αποτελέσματα Semeval-2017

Ακολουθούν τα επίσημα αποτελέσματα του διαγωνισμού SemEval-2017 Task 4: "Sentiment Analysis in Twitter", όπως παρουσιάζονται και στο (Rosenthal κ.ά. 2017).

#	System	AvgRec	AvgF1	Acc
1	DataStories	$\mathbf{0.681}_{1}$	0.677_{2}	0.651_{5}
	BB_twtr	$\mathbf{0.681}_{1}$	0.685_{1}	0.658_{3}
3	LIA	$\mathbf{0.676}_{3}$	0.674_{3}	0.661_{2}
4	Senti17	$\mathbf{0.674}_{4}$	0.665_{4}	0.652_{4}
5	NNEMBs	$\mathbf{0.669}_{5}$	0.658_{5}	0.664_{1}
6	Tweester	$\mathbf{0.659}_{6}$	0.648_{6}	0.648_{6}
7	INGEOTEC	$\mathbf{0.649}_{7}$	0.645_{7}	0.633_{11}
8	SiTAKA	$\mathbf{0.645}_{8}$	0.628_{9}	0.643_{9}
9	TSA-INF	$\mathbf{0.643}_{9}$	0.620_{11}	0.616_{17}
10	UCSC-NLP	$\mathbf{0.642}_{10}$	0.624_{10}	0.565_{30}
11	HLP@UPENN	$\mathbf{0.637}_{11}$	0.632_{8}	0.646_{8}
12	YNU-HPCC	$\mathbf{0.633}_{12}$	0.612_{15}	0.647_{7}
	SentiME	$\mathbf{0.633}_{12}$	0.613_{13}	0.601_{23}
14	ELiRF-UPV	$\mathbf{0.632}_{14}$	0.619_{12}	0.599_{24}
15	ECNU	$\mathbf{0.628}_{15}$	0.613_{13}	0.630_{12}
16	TakeLab	$\mathbf{0.627}_{16}$	0.607_{16}	0.628_{14}
17	DUTH	$\mathbf{0.621}_{17}$	0.605_{17}	0.640_{10}
18	CrystalNest	$\mathbf{0.619}_{18}$	0.593_{19}	0.629_{13}
19	deepSA	$\mathbf{0.618}_{19}$	0.587_{20}	0.616_{17}
20	NILC-USP	$\mathbf{0.612}_{20}$	0.595_{18}	0.617_{16}
21	Ti-Senti	$\mathbf{0.607}_{21}$	0.577_{22}	0.627_{15}
22	BUSEM	$\mathbf{0.605}_{22}$	0.587_{20}	0.603_{22}
23	EICA	$\mathbf{0.595}_{23}$	0.555_{24}	0.599_{24}
24	OMAM	$\mathbf{0.590}_{24}$	0.542_{26}	0.615_{19}
25	Adullam	$\mathbf{0.589}_{25}$	0.552_{25}	0.614_{20}
26	NileTMRG	$\mathbf{0.578}_{26}$	0.515_{32}	0.606_{21}
27	Amobee-C-137	$\mathbf{0.575}_{27}$	0.520_{30}	0.587_{27}
28	ej-za-2017	$\mathbf{0.571}_{28}$	0.539_{27}	0.582_{28}
	LSIS	$\mathbf{0.571}_{28}$	0.561_{23}	0.521_{34}
30	XJSA	$\mathbf{0.556}_{30}$	0.519_{31}	0.575_{29}
31	Neverland-THU	$\mathbf{0.555}_{31}$	0.507_{33}	0.597_{26}
32	MI&T-Lab	$\mathbf{0.551}_{32}$	0.522_{29}	0.561_{31}
33	diegoref	$\mathbf{0.546}_{33}$	0.527_{28}	0.540_{33}
34	xiwu	$\mathbf{0.479}_{34}$	0.365_{34}	0.547_{32}
35	SSN_MLRG1	$\mathbf{0.431}_{35}$	0.344_{35}	0.439_{35}
36	YNU-1510	$\mathbf{0.340}_{36}$	0.201_{37}	0.387_{36}
37	WarwickDCS	$\mathbf{0.335}_{37}$	0.221_{36}	0.382_{37}
	Avid	$\mathbf{0.335}_{37}$	0.163_{38}	0.206_{38}

Πίνακας 5.9: Αποτελέσματα για SemEval-2017 Task 4, subtask A "Message Polarity Classification", για Αγγλικά. Τα συστήματα είναι ταξινομημένα βάση του AvgRec (οι μεγαλύτερες τιμές είναι καλύτερες).

#	System	AvgRec	AvgF1	Acc
1	BB_twtr	$\mathbf{0.882}_{1}$	0.890_{1}	0.897_{1}
2	DataStories	$\mathbf{0.856}_{2}$	0.861_{2}	0.869_{2}
3	Tweester	$\mathbf{0.854}_{3}$	0.856_{3}	0.863_{3}
4	TopicThunder	$\mathbf{0.846}_{4}$	0.847_{4}	0.854_{4}
5	TakeLab	$\mathbf{0.845}_{5}$	0.836_{5}	0.840_{6}
6	funSentiment	$\mathbf{0.834}_{6}$	0.824_{8}	0.827_{8}
	YNU-HPCC	$\mathbf{0.834}_{6}$	0.816_{10}	0.818_{10}
8	WarwickDCS	$\mathbf{0.829}_{8}$	0.834_{6}	0.843_{5}
9	CrystalNest	$\mathbf{0.827}_{9}$	0.822_{9}	0.827_{8}
10	Ti-Senti	$\mathbf{0.826}_{10}$	0.830_{7}	0.838_{7}
11	Amobee-C-137	$\mathbf{0.822}_{11}$	0.801_{12}	0.802_{12}
12	SINAI	$\mathbf{0.818}_{12}$	0.806_{11}	0.809_{11}
13	NRU-HSE	$\mathbf{0.798}_{13}$	0.787_{13}	0.790_{13}
14	EICA	$\mathbf{0.790}_{14}$	0.775_{14}	0.777_{16}
15	OMAM	$\mathbf{0.779}_{15}$	0.762_{17}	0.764_{17}
16	NileTMRG	$\mathbf{0.769}_{16}$	0.774_{15}	0.789_{15}
17	ELiRF-UPV	$\mathbf{0.766}_{17}$	0.773_{16}	0.790_{13}
18	DUTH	$\mathbf{0.663}_{18}$	0.600_{18}	0.607_{18}
19	ej-za-2017	$\mathbf{0.594}_{19}$	0.486_{21}	0.518_{19}
20	SSN_MLRG1	$\mathbf{0.586}_{20}$	0.494_{20}	0.518_{19}
21	YNU-1510	$\mathbf{0.516}_{21}$	0.499_{19}	0.499_{21}
22	TM-Gist	$\mathbf{0.499}_{22}$	0.428_{22}	0.444_{22}
23	SSK_JNTUH	$\mathbf{0.483}_{23}$	0.372_{23}	0.412_{23}

Πίνακας 5.10: Αποτελέσματα για SemEval-2017 Task 4, subtask B "Tweet classification according to a two-point scale", για Αγγλικά. Τα συστήματα είναι ταξινομημένα βάση του AvgRec (οι μεγαλύτερες τιμές είναι καλύτερες).

#	System	$\mathbf{MAE}^M$	$\mathbf{MAE}^\mu$
1	BB_twtr	$\mathbf{0.481}_1$	0.554_6
2	DataStories	$\mathbf{0.555}_2$	0.543_4
3	Amobee-C-137	$\mathbf{0.599}_3$	0.582_{10}
4	Tweester	$\mathbf{0.623}_4$	0.734_{13}
5	TwiSe	$\mathbf{0.640}_5$	0.616_{12}
6	CrystalNest	$\mathbf{0.698}_6$	0.571_9
7	ELiRF-UPV	$\mathbf{0.806}_7$	0.586_{11}
8	EICA	$\mathbf{0.823}_8$	0.509_2
9	funSentiment	$\mathbf{0.842}_9$	0.530_3
10	DUTH	$\mathbf{0.895}_{10}$	0.544_5
	OMAM	$\mathbf{0.895}_{10}$	0.475_1
12	YNU-HPCC	$\mathbf{0.925}_{12}$	0.567_8
13	NRU-HSE	$\mathbf{0.928}_{13}$	0.557_7
14	YNU-1510	$\mathbf{1.262}_{14}$	0.764_{14}
15	SSN_MLRG1	$\mathbf{1.325}_{15}$	0.985_{15}

Πίνακας 5.11: Αποτελέσματα για SemEval-2017 Task 4, subtask C "Tweet classification according to a five-point scale", για Αγγλικά. Τα συστήματα είναι ταξινομημένα βάση του MAE^M (οι μικρότερες τιμές είναι καλύτερες).

#	System	KLD	AE	RAE
1	BB_twtr	$\mathbf{0.036}_{1}$	0.080_{1}	0.598_{1}
2	DataStories	$\mathbf{0.048}_{2}$	0.095_{2}	0.848_{2}
3	TakeLab	$\mathbf{0.050}_{3}$	0.096_{3}	1.057_{5}
4	CrystalNest	$\mathbf{0.056}_{4}$	0.104_{5}	1.202_{6}
5	Tweester	$\mathbf{0.057}_{5}$	0.103_{4}	1.051_{4}
6	funSentiment	$\mathbf{0.060}_{6}$	0.109_{6}	0.939_{3}
7	NileTMRG	$\mathbf{0.077}_{7}$	0.120_{7}	1.228_{7}
8	NRU-HSE	$\mathbf{0.078}_{8}$	0.132_{8}	1.528_{8}
9	ecnucsy	$\mathbf{0.092}_{9}$	0.143_{9}	1.922_{9}
10	THU_HCSI_IDU	$\mathbf{0.129}_{10}$	0.179_{10}	2.428_{11}
11	Amobee-C-137	$\mathbf{0.149}_{11}$	0.179_{10}	2.168_{10}
12	OMAM	$\mathbf{0.164}_{12}$	0.204_{12}	2.790_{12}
13	SSK_JNTUH	$\mathbf{0.421}_{13}$	0.314_{13}	2.983_{13}
14	ELiRF-UPV	$\mathbf{1.060}_{14}$	0.593_{15}	7.991_{15}
15	YNU-HPCC	$\mathbf{1.142}_{15}$	0.592_{14}	7.859_{14}

Πίνακας 5.12: Αποτελέσματα για SemEval-2017 Task 4, subtask D "Tweet quantification according to a two-point scale", για Αγγλικά. Τα συστήματα είναι ταξινομημένα βάση του KLD (οι μικρότερες τιμές είναι καλύτερες).

#	System	EMD
1	BB_twtr	0.245
2	TwiSe	0.269
3	funSentiment	0.273
4	ELiRF-UPV	0.306
5	NRU-HSE	0.317
6	Amobee-C-137	0.345
7	OMAM	0.350
8	Tweester	0.365
9	THU_HCSI_IDU	0.385
10	YNU-HPCC	0.447
11	DataStories	0.595
12	ecnucsy	1.461

Πίνακας 5.13: Αποτελέσματα για SemEval-2017 Task 4, subtask E "Tweet quantification according to a two-point scale", για Αγγλικά. Τα συστήματα είναι ταξινομημένα βάση του EMD (οι μικρότερες τιμές είναι καλύτερες).

Κεφάλαιο 6

Συμπεράσματα

Όπως είδαμε, η Ανάλυση Συναισθήματος (ΑΣ) είναι ένα από τα νέα πεδία της Επεξεργασίας Φυσικής Γλώσσας (ΕΦΓ), το οποίο έχει συγκεντρώσει μεγάλο ενδιαφέρον, τόσο επιστημονικά, όσο και για τις εφαρμογές του. Τις περισσότερες φορές, η ΑΣ προσεγγίζεται ως πρόβλημα κατηγοριοποίησης κειμένων. Για την υλοποίηση συστημάτων ΑΣ, αρχικά χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης όπως SVM, πάνω σε χαρακτηριστικά τα οποία είχαν σχεδιαστεί ακολουθώντας την θεωρία της υπολογιστικής γλωσσολογίας (BoW, n-grams, dependecy trees κλπ.). Όμως, η φυσική γλώσσα δεν ακολουθεί ένα σύνολο από αυστηρά ορισμένους κανόνες, αλλά είναι αδόμητη και κατά συνέπεια είναι δύσκολο να μοντελοποιηθεί. Ιδιαίτερη δυσκολία έχει η επεξεργασία κειμένων σε δίκτυα κοινωνικής δικτύωσης, όπως το Twitter. Έτσι, είναι προφανές ότι θα πρέπει να χρησιμοποιηθούν τεχνικές, οι οποίες θα μπορούν να αξιοποιήσουν πιο αφηρημένα χαρακτηριστικά από τα κείμενα.

Αυτό το πετυχαίνουν οι τεχνικές βαθιάς μάθησης (deep learning). Οι τεχνικές αυτές, αντίθετα με τις πιο "παραδοσιακές" τεχνικές, δεν βασίζονται στον σχεδιασμό χαρακτηριστικών από τον άνθρωπο, αλλά μαθαίνουν αυτόματα τα χαρακτηριστικά. Με τον τρόπο αυτό, ένα μοντέλο το οποίο βασίζεται σε τεχνικές βαθιάς μάθησης, μπορεί να ανακαλύψει χρήσιμα χαρακτηριστικά, τα οποία δεν θα μπορούσε ένα άνθρωπος να σχεδιάσει χειροκίνητα. Η ανωτερότητα των μοντέλων αυτών να περιγράψουν καλύτερα αφηρημένα προβλήματα, όπως αυτό της ΑΣ στο Twitter, φαίνεται ξεκάθαρα στα αποτελέσματα του διαγωνισμού Semeval, στην κατηγορία "Sentiment Analysis in Twitter".

Ήδη από την προηγούμενη χρονιά (Nakov κ.ά. 2016b), τα μοντέλα βαθιάς μάθησης ξεπέρασαν κατά πολύ τις άλλες προσεγγίσεις. Το ίδιο φαινόμενο παρατηρήθηκε και την φετινή χρονιά (Rosenthal κ.ά. 2017), όπου και με την δική μας συμμετοχή (Baziotis κ.ά. 2017a) πετύχαμε την πρώτη θέση. Τα μοντέλα βαθιάς μάθησης έχουν κυριαρχήσει σχεδόν σε όλα τα προβλήματα ΕΦΓ, όπως και σε άλλα επιστημονικά πεδία (Υπολογιστική Όραση).

Ακόμη, μία συνέπεια της αυτόματης εξαγωγής χαρακτηριστικών, είναι η απλοποίηση και η επιτάχυνση της διαδικασίας της μοντελοποίησης. Αυτό πρακτικά σημαίνει, ότι μπορεί κάποιος με μικρές αρχιτεκτονικές αλλαγές, να σχεδιάσει σε μικρό χρονικό διάστημα αρκετά μοντέλα, τα οποία στοχεύουν τελείως διαφορετικά προβλήματα. Αξιοποιώντας αυτή την ικανότητα των τεχνητών νευρωνικών δικτύων, λάβαμε μέρος και στον διαγωνισμό Semeval-2017 Task 6: "#HashtagWars: Learning a Sense of Humor", με αντίστοιχη επιτυχία (Baziotis κ.ά. 2017b). Συνεπώς, προκύπτει ότι τα τεχνητά νευρωνικά δίκτυα και συγκεκριμένα οι τεχνικές βαθιάς μάθησης, αποτελούν ένα τελείως διαφορετικό παράδειγμα μοντελοποίησης, το οποίο

ξεπερνά πολλά από τα συνηθισμένα εμπόδια στην διαδικασία σχηματισμού μοντέλων μηχανικής μάθησης. Όμως, παρά τα πλεονεκτήματα τα οποία προσφέρουν, υπάρχουν πολλά περιθώρια βελτίωσης.

Ορισμένες ιδέες για μελλοντική εργασία στην ΑΣ με βαθιά νευρωνικά δίκτυα, είναι οι εξής:

Μορφολογία των λέξεων. Στα περισσότερα μοντέλα ΕΦΓ, ο δομικός λίθος των κειμένων είναι οι λέξεις. Σε κάθε λέξη αντιστοιχεί ένα διαφορετικό διάνυσμα. Όμως, αυτή η προσέγγιση αγνοεί την μορφολογία (την εσωτερική δομή) των λέξεων. Για παράδειγμα, σε πολλές γλώσσες αρκετά ρήματα μπορεί να έχουν πολλές διαφορετικές μορφές και οι περισσότερες προσεγγίσεις για αναπαραστάσεις λέξεων δεν αξιοποιούν αυτές τις ιδιότητες των λέξεων. Ένα σημαντικό πρόβλημα αφορά τις λέξεις εκτός λεξιλογίου (Out of Vocabulary - OVV - words), όπου μία λέξη η οποία δεν έχει συναντηθεί κατά την εκπαίδευση του μοντέλου αναπαραστάσεων λέξεων, δεν θα έχει ένα διάνυσμα που να την περιγράφει. Συχνά όμως, οι OVV λέξεις μπορεί να μοιράζονται κοινά μορφολογικά χαρακτηριστικά με ήδη γνωστές λέξεις.

Η αξιοποίηση της μορφολογίας των λέξεων, πιθανότατα θα έχει μεγάλη επίπτωση σε κείμενα όπως αυτά των μηνυμάτων στο Twitter, όπου συχνά συναντά κανείς λέξεις οι οποίες είναι προϊόν "δημιουργικής" γραφής και βασίζονται στην μορφολογία άλλων, γνωστών λέξεων. Μερικές πιθανές κατευθύνσεις είναι η εκπαίδευση μοντέλων σε επίπεδο χαρακτήρων και όχι λέξεων (Kim κ.ά. 2015; Xiang Zhang κ.ά. 2015; Miyamoto κ.ά. 2016; Vosoughi κ.ά. 2016) ή η αξιοποίηση αναπαραστάσεων λέξεων οι οποίες έχουν εκπαιδευτεί λαμβάνοντας υπόψη τους την μορφολογία των λέξεων (Bojanowski κ.ά. 2016).

Βελτίωση Ποσοτικοποίησης. Ένα σημαντικό πρόβλημα, όπως φάνηκε και από τα αποτελέσματα του Semeval, είναι οι επιδόσεις των μοντέλων στο πρόβλημα της ποσοτικοποίησης (quantification). Τα αποτελέσματα ήταν αντιφατικά, με ορισμένες προσεγγίσεις να δουλεύουν καλά σε ένα Task και άλλες όχι. Η ποσοτικοποίηση είναι ουσιαστικά το τελικό αποτέλεσμα το οποίο θα ήθελε να δει κάποιος από μία πρακτική εφαρμογή ΑΣ. Διότι, πρακτικά αυτό που έχει αξία είναι το ποσοστό των θετικών και των αρνητικών σχολίων για ένα προϊόν, μία αντίληψη ή έναν άνρθωπο και όχι το αποτέλεσμα ενός συγκεκριμένου σχολίου. Το ζητούμενο είναι η παραγωγή μίας σύνοψης, ώστε να χρησιμοποιηθεί αυτή η πληροφορία για την λήψη αποφάσεων.

Για την βελτίωση των τεχνικών ποσοτικοποίησης, θα μπορούσε κανείς να συνεχίσει την έρευνα του (Gao κ.ά. 2016). Μερικές περαιτέρω περιπτώσεις που θα μπορούσε να εξετάσει μία τέτοια έρευνα είναι:

- Διαφορετικό πλήθος κλάσεων. Το ζητούμενο είναι, αν αλλάζει η συμπεριφορά ενός μοντέλου όταν αλλάζει το πλήθος των κλάσεων. Επίσης, ένα ακόμη ενδιαφέρον ερώτημα είναι, το πως θα μπορούσε κανείς να προσεγγίσει την ποσοτικοποίηση, αν η ΑΣ είχε γίνει ως παλινδρόμηση (regression) και όχι ως κατηγοριοποίηση (classification).

- Στοχευμένη ΑΣ. Αλλάζουν τα αποτελέσματα μίας τεχνικής, όταν αλλάζει το αντικείμενο της ΑΣ? Δηλαδή, όταν η ανάλυση γίνεται με στόχο το συναίσθημα που εκφράζεται ως προς διαφορετικά πρόσωπα ή προϊόντα, τι συνέπειες έχει αυτό στις επιδόσεις των διάφορων τεχνικών. Δεν είναι προφανές ότι θα υπάρχει κάποια διαφορά, αλλά θα είχε ενδιαφέρον αν ίσχυε κάτι τέτοιο.

Στοχευμένη ΑΣ. Όπως φάνηκε και από τα αποτελέσματα του Semeval στα Subtasks (B και D), οι επιδόσεις του δικού μας μοντέλου με της ομάδας *BB_twtr* (Cliche 2017), ήταν πολύ κοντά. Όμως, το ενδιαφέρον είναι πως ενώ το δικό μας μοντέλο αξιοποιούσε την πληροφορία του θέματος (στόχου), η ομάδα *BB_twtr* την αγνοούσε. Αντίθετα, χρησιμοποίησαν την ίδια προσέγγιση, με αυτή για την εύρεση του γενικού συναισθηματικού προσανατολισμού των μηνυμάτων. Το γεγονός ότι ένα μοντέλο το οποίο λαμβάνει υπόψη του την πληροφορία του θέματος-στόχου, με ένα άλλο που την αγνοεί, είναι συγκρίσιμα σε ένα πρόβλημα στοχευμένης ΑΣ, απαιτεί περισσότερη μελέτη. Ορισμένες πιθανές αιτίες μπορεί να είναι:

- Προβληματικά δεδομένα εκπαίδευσης. Στην προκειμένη περίπτωση αυτό σημαίνει, ότι το συναίσθημα το οποίο απορρέει γενικά από το κείμενο, μπορεί να ταυτίζεται σε μεγάλο ποσοστό, με αυτό του συναισθήματος που εκφράζεται ως προς τον ζητούμενο στόχο. Έτσι, δεν έχει πολλά να κερδίσει ένα μοντέλο το οποίο αξιοποιεί την επιπλέον πληροφορία. Αυτό είναι ένα σημαντικό πρόβλημα, καθώς τα περισσότερα datasets τα οποία αφορούν στοχευμένη ΑΣ, δεν περιέχουν αντικρουόμενες απόψεις στα κείμενά τους. Συνεπώς, δεν είναι ξεκάθαρο ότι το γενικό συναίσθημα, διαφέρει από το στοχευμένο.

- Αδύναμα μοντέλα για στοχευμένη ΑΣ. Ένας προφανής λόγος θα μπορούσε να είναι, ότι απλώς τα μοντέλα μας, τα οποία αξιοποιούσαν την πληροφορία του θέματος, δεν ήταν αρκετά καλά. Όμως, αυτό είναι σχετικά δύσκολο, με δεδομένο ότι στον διαγωνισμό συμμετείχαν πολλές ομάδες και η δική μας δεν ήταν η μοναδική που ακολούθησε αντίστοιχη προσέγγιση. Θα ήταν περίεργο να έκαναν όλες οι υπόλοιπες ομάδες σχεδιαστικά λάθη, πόσο μάλλον από την στιγμή που είχαμε καλύτερα αποτελέσματα από αυτές.

Μία άλλη ενδιαφέρουσα κατεύθυνση θα ήταν η αυτόματη ανακάλυψη των θεμάτων σε ένα κείμενο, πριν την εφαρμογή της ΑΣ. Αυτό είναι αρκετά δύσκολο πρόβλημα. Έχει μελετηθεί πολύ στο πλαίσιο της εύρεσης όρων (λέξεων και φράσεων), που μπορεί να προσδιορίζουν ορισμένες οντότητες, ως προς τις οποίες εκφράζονται απόψεις (Pavlopoulos κ.ά. 2014). Όμως, με την χρήση τεχνικών deep learning, θα μπορούσε κανείς να ανακαλύψει (να συμπεράνει την ύπαρξή τους) ακόμη και οντότητες οι οποίες δεν αναφέρονται ρητά στο κείμενο.

Βιβλιογραφία

Turing, A. M. (1950). "Computing Machinery and Intelligence". Στο: *Mind* 59.236. 09508, σσ. 433–460. ISSN: 0026-4423.

Rosenblatt, Frank (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." Στο: *Psychological review* 65.6, σ. 386.

Osgood, Charles E. (1964). "Semantic Differential Technique in the Comparative Study of Cultures". Στο: *American Anthropologist* 66.3. 00000, σσ. 171–200.

Porter, Martin F. (1980). "An Algorithm for Suffix Stripping". Στο: *Program* 14.3, σσ. 130–137.

Hopfield, John J. (1982). "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". Στο: *Proceedings of the national academy of sciences* 79.8. 18706, σσ. 2554–2558.

Nesterov, Yurii (1983). "A Method of Solving a Convex Programming Problem with Convergence Rate O (1/K2)". Στο: *Soviet Mathematics Doklady*. Τόμ. 27. 01390, σσ. 372–376.

Salton, Gerard, Edward A. Fox και Harry Wu (1983). "Extended Boolean Information Retrieval". Στο: *Communications of the ACM* 26.11, σσ. 1022–1036.

Rumelhart, David E., Geoffrey E. Hinton και Ronald J. Williams (1988). "Learning Representations by Back-Propagating Errors". Στο: *Cognitive modeling* 5.3. 00000, σ. 1.

Hornik, Kurt, Maxwell Stinchcombe και Halbert White (1989). "Multilayer Feedforward Networks Are Universal Approximators". Στο: *Neural Networks* 2.5, σσ. 359–366. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

LeCun, Y. κ.ά. (1989). "Backpropagation Applied to Handwritten Zip Code Recognition". Στο: *Neural Computation* 1.4. 00000, σσ. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

Elman, J (1990). "Finding Structure in Time". en. Στο: *Cognitive Science* 14.2. 00000, σσ. 179–211. ISSN: 03640213. DOI: 10.1016/0364-0213(90)90002-E.

Miller, George A. (1995). "WordNet: A Lexical Database for English". Στο: *Communications of the ACM* 38.11. 09274, σσ. 39–41.

Hochreiter, Sepp και Jürgen Schmidhuber (1997). "Long Short-Term Memory". Στο: *Neural computation* 9.8, σσ. 1735–1780.

Jordan, Michael I. (1997). "Serial Order: A Parallel Distributed Processing Approach". en. Στο: *Advances in Psychology*. Τόμ. 121. 01033. Elsevier, σσ. 471–495. ISBN: 978-0-444-81931-4. DOI: 10.1016/S0166-4115(97)80111-2.

Yang, Yiming και Jan O. Pedersen (1997). "A Comparative Study on Feature Selection in Text Categorization". Στο: *Icml*. Τόμ. 97, σσ. 412–420.

LeCun, Yann κ.ά. (1998). "Gradient-Based Learning Applied to Document Recognition". Στο: *Proceedings of the IEEE* 86.11. 00000, σσ. 2278–2324.

Sahami, Mehran κ.ά. (1998). "A Bayesian Approach to Filtering Junk E-Mail". Στο: *Learning for Text Categorization: Papers from the 1998 Workshop*. Τόμ. 62, σσ. 98–105.

Androutsopoulos, Ion κ.ά. (2000). "An Evaluation of Naive Bayesian Anti-Spam Filtering". Στο: *arXiv:cs/0006013*. arXiv: cs/0006013.

Bennett, Kristin P. και Erin J. Bredensteiner (2000). "Duality and Geometry in SVM Classifiers". Στο: *ICML*, σσ. 57–64.

Jurafsky, Daniel και H. James (2000). "Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech". Στο: 00025.

Rubner, Yossi, Carlo Tomasi και Leonidas J. Guibas (2000). "The Earth Mover's Distance as a Metric for Image Retrieval". Στο: *International journal of computer vision* 40.2, σσ. 99–121.

Gers, Felix A., Nicol N. Schraudolph και Jürgen Schmidhuber (2002). "Learning Precise Timing with LSTM Recurrent Networks". Στο: *Journal of machine learning research* 3.Aug, σσ. 115–143.

Pang, Bo, Lillian Lee και Shivakumar Vaithyanathan (2002). "Thumbs up?: Sentiment Classification Using Machine Learning Techniques". Στο: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*. Association for Computational Linguistics, σσ. 79–86.

Turney, Peter D. (2002). "Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews". Στο: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 00000. Association for Computational Linguistics, σσ. 417–424.

Basuroy, Suman, Subimal Chatterjee και S. Abraham Ravid (2003). "How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets". Στο: *Journal of Marketing* 67.4, σσ. 103–117. ISSN: 0022-2429. DOI: 10.1509/jmkg.67.4.103.18692.

Bengio, Yoshua κ.ά. (2003). "A Neural Probabilistic Language Model". Στο: *Journal of machine learning research* 3.Feb, σσ. 1137–1155.

Blei, David M., Andrew Y. Ng και Michael I. Jordan (2003). "Latent Dirichlet Allocation". Στο: *Journal of machine Learning research* 3.Jan. 19414, σσ. 993–1022.

Forman, George (2003). "An Extensive Empirical Study of Feature Selection Metrics for Text Classification". Στο: *Journal of machine learning research* 3.Mar. 00000, σσ. 1289–1305.

Weisstein, Eric W (2003). *Convolution*. Αδημοσίευτη ερευνητική εργασία. Wolfram Research, Inc.

Hu, Minqing και Bing Liu (2004). "Mining and Summarizing Customer Reviews". Στο: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 04571. New York, NY, USA: ACM, σσ. 168–177. ISBN: 978-1-58113-888-7. DOI: 10.1145/1014052.1014073.

Pang, Bo και Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". Στο: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. 02393. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI: 10.3115/1218955.1218990.

Zheng, Zhaohui, Xiaoyun Wu και Rohini Srihari (2004). "Feature Selection for Text Categorization on Imbalanced Data". Στο: *SIGKDD Explor. Newsl.* 6.1. 00446, σσ. 80–89. ISSN: 1931-0145. DOI: 10.1145/1007730.1007741.

Novovicova, Jana και Antonin Malik (2005). "Information-Theoretic Feature Selection Algorithms for Text Classification". Στο: *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference On*. Τόμ. 5. IEEE, σσ. 3272–3277.

Pang, Bo και Lillian Lee (2005). "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales". Στο: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, σσ. 115–124.

Popescu, Ana-Maria και Oren Etzioni (2005). "Extracting Product Features and Opinions from Reviews". Στο: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, σσ. 339–346. DOI: 10.3115/1220575.1220618.

Chevalier, Judith A και Dina Mayzlin (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews". Στο: *Journal of Marketing Research* 43.3. 04120, σσ. 345–354. ISSN: 0022-2437. DOI: 10.1509/jmkr.43.3.345.

Lemmon, Michael και Evgenia Portniaguina (2006). "Consumer Confidence and Asset Prices: Some Empirical Evidence". Στο: *The Review of Financial Studies* 19.4. 00626, σσ. 1499–1529. ISSN: 0893-9454. DOI: 10.1093/rfs/hhj038.

Liu, Yong (2006). "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue". Στο: *Journal of Marketing* 70.3. 01855, σσ. 74–89. ISSN: 0022-2429. DOI: 10.1509/jmkg.70.3.74.

Mishne, Gilad, Natalie S. Glance κ.ά. (2006). "Predicting Movie Sales from Blogger Sentiment." Στο: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, σσ. 155–158.

Zhuang, Li, Feng Jing και Xiao-Yan Zhu (2006). "Movie Review Mining and Summarization". Στο: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. 00753. ACM, σσ. 43–50.

Blitzer, John, Mark Dredze, Fernando Pereira κ.ά. (2007). "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification". Στο: *ACL*. Τόμ. 7, σσ. 440–447.

Esuli, Andrea και Fabrizio Sebastiani (2007). "SENTIWORDNET: A High-Coverage Lexical Resource for Opinion Mining". Στο: *Evaluation*. 02232, σσ. 1–26.

Liu, Yang κ.ά. (2007). "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs". Στο: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 00219. New York, NY, USA: ACM, σσ. 607–614. ISBN: 978-1-59593-597-7. DOI: 10.1145/1277741.1277845.

Tetlock, Paul C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". en. Στο: *The Journal of Finance* 62.3. 01820, σσ. 1139–1168. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2007.01232.x.

Xu, Yang κ.ά. (2007). "A Study on Mutual Information-Based Feature Selection for Text Categorization". en. Στο: *Journal of Computational Information Systems* 3.3. 00000, σσ. 1007–1012. ISSN: 1553-9105.

Collobert, Ronan και Jason Weston (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning". Στο: *Proceedings of the 25th International Conference on Machine Learning*. ACM, σσ. 160–167.

De Marneffe, Marie-Catherine και Christopher D. Manning (2008). *Stanford Typed Dependencies Manual*. Αδημοσίευτη ερευνητική εργασία. Technical report, Stanford University.

Forman, George (2008). "Quantifying Counts and Costs via Classification". Στο: *Data Mining and Knowledge Discovery* 17.2, σσ. 164–206.

Han, Bing (2008). "Investor Sentiment and Option Prices". Στο: *The Review of Financial Studies* 21.1, σσ. 387–414. ISSN: 0893-9454. DOI: 10.1093/rfs/hhm071.

Manning, Christopher D., Prabhakar Raghavan και Hinrich Schütze (2008). *Introduction to Information Retrieval*. 00191. New York: Cambridge University Press. ISBN: 978-0-521-86571-5.

Jansen, Bernard J. κ.ά. (2009). "Twitter Power: Tweets As Electronic Word of Mouth". Στο: *J. Am. Soc. Inf. Sci. Technol.* 60.11. 01920, σσ. 2169–2188. ISSN: 1532-2882. DOI: 10.1002/asi.v60:11.

Joshi, Mahesh και Carolyn Penstein-Rosé (2009). "Generalizing Dependency Features for Opinion Mining". Στο: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. 00000. Association for Computational Linguistics, σσ. 313–316.

Ritterman, Joshua, Miles Osborne και Ewan Klein (2009). "Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic". Στο: *1st International Workshop on Mining Social Media*. Τόμ. 9. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), σσ. 9–17.

Segaran, Toby και Jeff Hammerbacher (2009). *Beautiful Data: The Stories Behind Elegant Data Solutions*. en. 00114. "O'Reilly Media, Inc." ISBN: 978-1-4493-7929-2.

Asur, S. και B. A. Huberman (2010). "Predicting the Future with Social Media". Στο: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Τόμ. 1. 01621, σσ. 492–499. DOI: 10.1109/WI-IAT.2010.63.

Baccianella, Stefano, Andrea Esuli και Fabrizio Sebastiani (2010). "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." Στο: *LREC*. Τόμ. 10. 01551, σσ. 2200–2204.

Bella, Antonio κ.ά. (2010). "Quantification via Probability Estimators". Στο: *Data Mining (ICDM), 2010 IEEE 10th International Conference On*. 00000. IEEE, σσ. 737–742.

Culotta, Aron (2010). "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages". Στο: *Proceedings of the First Workshop on Social Media Analytics*. New York, NY, USA: ACM, σσ. 115–122. ISBN: 978-1-4503-0217-3. DOI: 10.1145/1964858.1964874.

Gilbert, Eric και Karrie Karahalios (2010). "Widespread Worry and the Stock Market." Στο: *ICWSM*, σσ. 59–65.

O'Connor, Brendan κ.ά. (2010). "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." Στο: *ICWSM* 11.122-129, σσ. 1–2.

Özgür, Levent και Tunga Güngör (2010). "Text Classification with the Support of Pruned Dependency Patterns". en. Στο: *Pattern Recognition Letters* 31.12, σσ. 1598–1607. ISSN: 01678655. DOI: 10.1016/j.patrec.2010.05.005.

Sakaki, Takeshi, Makoto Okazaki και Yutaka Matsuo (2010). "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors". Στο: *Proceedings of the 19th International Conference on World Wide Web*. 00000. ACM, σσ. 851–860.

Tumasjan, Andranik κ.ά. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." Στο: *ICWSM* 10.1. 01804, σσ. 178–185.

Zhu, Feng και Xiaoquan (Michael) Zhang (2010). "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics".

Στο: *Journal of Marketing* 74.2. 01186, σσ. 133–148. ISSN: 0022-2429. DOI: 10.1509/jmkg.74.2.133.

Aramaki, Eiji, Sachiko Maskawa και Mizuki Morita (2011). "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter". Στο: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, σσ. 1568–1576.

Bollen, Johan, Huina Mao και Alberto Pepe (2011a). "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena." Στο: *ICWSM* 11, σσ. 450–453.

Bollen, Johan, Huina Mao και Xiaojun Zeng (2011b). "Twitter Mood Predicts the Stock Market". Στο: *Journal of Computational Science* 2.1. 00000, σσ. 1–8. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2010.12.007.

Duchi, John, Elad Hazan και Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". Στο: *Journal of Machine Learning Research* 12.Jul. 02185, σσ. 2121–2159. ISSN: ISSN 1533-7928.

Gimpel, Kevin κ.ά. (2011). "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments". Στο: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, σσ. 42–47.

Maas, Andrew L. κ.ά. (2011). "Learning Word Vectors for Sentiment Analysis". Στο: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 00594. Association for Computational Linguistics, σσ. 142–150.

Maynard, Diana και Adam Funk (2011). "Automatic Detection of Political Opinions in Tweets". en. Στο: *The Semantic Web: ESWC 2011 Workshops*. Springer, Berlin, Heidelberg, σσ. 88–99. DOI: 10.1007/978-3-642-25953-1_8.

Metaxas, P. T., E. Mustafaraj και D. Gayo-Avello (2011). "How (Not) to Predict Elections". Στο: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, σσ. 165–171. DOI: 10.1109/PASSAT/SocialCom.2011.98.

Oh, Chong και Olivia Sheng (2011). "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement". Στο: *ICIS 2011 Proceedings*.

Potts, Christopher (2011). *Sentiment Symposium Tutorial: Tokenizing*. http://sentiment.christopherpotts.net/tokenizing.html.

Shakhnarovich, Gregory και Baback Moghaddam (2011). "Face Recognition in Subspaces". en. Στο: *Handbook of Face Recognition*. Επιμέλεια υπό Stan Z. Li και Anil K. Jain. 00001. London: Springer London, σσ. 19–49. ISBN: 978-0-85729-931-4 978-0-85729-932-1. DOI: 10.1007/978-0-85729-932-1_2.

Thelwall, Mike, Kevan Buckley και Georgios Paltoglou (2011). "Sentiment in Twitter Events". en. Στο: *Journal of the American Society for Information Science and Technology* 62.2, σσ. 406–418. ISSN: 1532-2890. DOI: 10.1002/asi.21462.

Weng, Jianshu και Bu-Sung Lee (2011). "Event Detection in Twitter." Στο: *ICWSM* 11. 00533, σσ. 401–408.

Zhang, Xue, Hauke Fuehres και Peter A. Gloor (2011). "Predicting Stock Market Indicators Through Twitter "I Hope It Is Not as Bad as I Fear"". Στο: *Procedia - Social and Behavioral Sciences* 26. 00336, σσ. 55–62. ISSN: 1877-0428. DOI: 10.1016/j.sbspro.2011.10.562.

Bergstra, James και Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". Στο: *Journal of Machine Learning Research* 13.Feb, σσ. 281–305.

Blei, David M. (2012). "Probabilistic Topic Models". Στο: *Communications of the ACM* 55.4, σσ. 77–84.

Chamlertwat, Wilas κ.ά. (2012). "Discovering Consumer Insight from Twitter via Sentiment Analysis." Στο: *J. UCS* 18.8. 00000, σσ. 973–992.

Dougal, Casey κ.ά. (2012). "Journalists and the Stock Market". Στο: *The Review of Financial Studies* 25.3, σσ. 639–679. ISSN: 0893-9454. DOI: 10.1093/rfs/hhr133.

Jungherr, Andreas, Pascal Jürgens και Harald Schoen (2012). "Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment"". en. Στο: *Social Science Computer Review* 30.2. 00215, σσ. 229–234. ISSN: 0894-4393. DOI: 10.1177/0894439311404119.

Kavanaugh, Andrea L. κ.ά. (2012). "Social Media Use by Government: From the Routine to the Critical". Στο: *Government Information Quarterly* 29.4, σσ. 480–491.

Roy, Asim (2012). "A Theory of the Brain: Localist Representation Is Used Widely in the Brain". Στο: *Frontiers in Psychology* 3. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00551.

Sidorov, Grigori κ.ά. (2012). "Syntactic Dependency-Based n-Grams as Classification Features". Στο: *Mexican International Conference on Artificial Intelligence*. Springer, σσ. 1–11.

Stieglitz, S. και L. Dang-Xuan (2012). "Political Communication and Influence through Microblogging–An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior". Στο: *2012 45th Hawaii International Conference on System Sciences*, σσ. 3500–3509. DOI: 10.1109/HICSS.2012.476.

Wang, Hao κ.ά. (2012). "A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle". Στο: *Proceedings of the ACL 2012 System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, σσ. 115–120.

Zeiler, Matthew D. (2012). "ADADELTA: An Adaptive Learning Rate Method". Στο: *arXiv:1212.5701 [cs]*. 01126. arXiv: 1212.5701 [cs].

Bergstra, James, Daniel Yamins και David D. Cox (2013). "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures." Στο: *ICML (1)* 28, σσ. 115–123.

Goodfellow, Ian J. κ.ά. (2013). "An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks". Στο: *arXiv preprint arXiv:1312.6211*.

Graves, Alex (2013). "Generating Sequences With Recurrent Neural Networks". Στο: *arXiv:1308.0850 [cs]*. 00570. arXiv: 1308.0850 [cs].

Hu, Yuheng, Fei Wang και Subbarao Kambhampati (2013). "Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment." Στο: *IJCAI*. 00000.

Makrehchi, Masoud, Sameena Shah και Wenhui Liao (2013). "Stock Prediction Using Event-Based Sentiment Analysis". Στο: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences On*. Τόμ. 1. 00029. IEEE, σσ. 337–342.

Mikolov, Tomas κ.ά. (2013a). "Distributed Representations of Words and Phrases and Their Compositionality". Στο: *Advances in Neural Information Processing Systems*. 05119, σσ. 3111–3119.

Mikolov, Tomas κ.ά. (2013b). "Efficient Estimation of Word Representations in Vector Space". Στο: *arXiv preprint arXiv:1301.3781*. 04344.

Mohammad, Saif M. και Peter D. Turney (2013). "Crowdsourcing a Word–emotion Association Lexicon". Στο: *Computational Intelligence* 29.3. 00322, σσ. 436–465.

Rui, Huaxia, Yizao Liu και Andrew Whinston (2013). "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales". Στο: *Decision Support Systems* 55.4, σσ. 863–870. ISSN: 0167-9236. DOI: 10.1016/j.dss.2012.12.022.

Si, Jianfeng κ.ά. (2013). "Exploiting Topic Based Twitter Sentiment for Stock Prediction." Στο: *ACL (2)* 2013. 00087, σσ. 24–29.

Smailović, Jasmina κ.ά. (2013). "Predictive Sentiment Analysis of Tweets: A Stock Market Application". Στο: *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. 00031. Springer, σσ. 77–88.

Yu, Yang, Wenjing Duan και Qing Cao (2013). "The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach". Στο: *Decision Support Systems* 55.4. 00000, σσ. 919–926. ISSN: 0167-9236. DOI: 10.1016/j.dss.2012.12.028.

Zhou, X. κ.ά. (2013). "Sentiment Analysis on Tweets for Social Events". Στο: *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 00044, σσ. 557–562. DOI: 10.1109/CSCWD.2013.6581022.

Arias, Marta, Argimiro Arratia και Ramon Xuriguera (2014). "Forecasting with Twitter Data". Στο: *ACM Trans. Intell. Syst. Technol.* 5.1. 00000, 8:1–8:24. ISSN: 2157-6904. DOI: 10.1145/2542182.2542190.

Bahdanau, Dzmitry, Kyunghyun Cho και Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". Στο: *arXiv:1409.0473 [cs, stat]*. 02127. arXiv: 1409.0473 [cs, stat].

Cambria, Erik, Daniel Olsher και Dheeraj Rajagopal (2014). "SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis". Στο: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, σσ. 1515–1521.

Ceron, Andrea κ.ά. (2014). "Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France". en. Στο: *New Media & Society* 16.2. 00000, σσ. 340–358. ISSN: 1461-4448. DOI: 10.1177/1461444813480466.

Cho, Kyunghyun κ.ά. (2014). "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation". Στο: *arXiv:1406.1078 [cs, stat]*. 01612. arXiv: 1406.1078 [cs, stat].

Chung, Junyoung κ.ά. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". Στο: *arXiv preprint arXiv:1412.3555*.

Khadjeh Nassirtoussi, Arman κ.ά. (2014). "Text Mining for Market Prediction: A Systematic Review". Στο: *Expert Systems with Applications* 41.16. 00000, σσ. 7653–7670. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2014.06.009.

Kim, Yoon (2014). "Convolutional Neural Networks for Sentence Classification". Στο: *arXiv preprint arXiv:1408.5882*. 01131.

Kingma, Diederik και Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". Στο: *arXiv preprint arXiv:1412.6980*.

Le, Quoc V. και Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents." Στο: *ICML*. Τόμ. 14, σσ. 1188–1196.

Levy, Omer και Yoav Goldberg (2014). "Dependency-Based Word Embeddings." Στο: *ACL (2)*. 00321. Citeseer, σσ. 302–308.

Pavlopoulos, John και Ion Androutsopoulos (2014). "Aspect Term Extraction for Sentiment Analysis: New Datasets, New Evaluation Measures and an Improved Unsupervised Method". Στο: *Proceedings of LASMEACL*. 00014, σσ. 44–52.

Pennington, Jeffrey, Richard Socher και Christopher D. Manning (2014). "Glove: Global Vectors for Word Representation." Στο: *EMNLP*. Τόμ. 14. 02413, σσ. 1532–1543.

Saif, Hassan κ.ά. (2014). "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter". Στο: 00044.

Smailović, Jasmina κ.ά. (2014). "Stream-Based Active Learning for Sentiment Analysis in the Financial Domain". Στο: *Information Sciences* 285, σσ. 181–203. ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.04.034.

Sprenger, Timm O. κ.ά. (2014). "Tweets and Trades: The Information Content of Stock Microblogs". en. Στο: *European Financial Management* 20.5, σσ. 926–957. ISSN: 1468-036X. DOI: 10.1111/j.1468-036X.2013.12007.x.

Srivastava, Nitish κ.ά. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." Στο: *Journal of Machine Learning Research* 15.1, σσ. 1929–1958.

Staiano, Jacopo και Marco Guerini (2014). "DepecheMood: A Lexicon for Emotion Analysis from Crowd-Annotated News". Στο: *arXiv preprint arXiv:1405.1605*. 00000.

Tang, Duyu κ.ά. (2014). "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." Στο: *ACL (1)*. 00309, σσ. 1555–1565.

Choi, Jinho D., Joel R. Tetreault και Amanda Stent (2015). "It Depends: Dependency Parser Comparison Using A Web-Based Evaluation Tool." Στο: *ACL (1)*, σσ. 387–396.

Esuli, Andrea και Fabrizio Sebastiani (2015). "Optimizing Text Quantifiers for Multivariate Loss Functions". Στο: *ACM Trans. Knowl. Discov. Data* 9.4, 27:1–27:27. ISSN: 1556-4681. DOI: 10.1145/2700406.

Greff, Klaus κ.ά. (2015). "LSTM: A Search Space Odyssey". Στο: *arXiv:1503.04069 [cs]*. 00312. arXiv: 1503.04069 [cs].

Ji, Shihao κ.ά. (2015). "WordRank: Learning Word Embeddings via Robust Ranking". Στο: *arXiv:1506.02761 [cs, stat]*. 00005. arXiv: 1506.02761 [cs, stat].

Karpathy, Andrej, Justin Johnson και Li Fei-Fei (2015). "Visualizing and Understanding Recurrent Networks". Στο: *arXiv preprint arXiv:1506.02078*. 00162.

Kim, Yoon κ.ά. (2015). "Character-Aware Neural Language Models". Στο: *arXiv preprint arXiv:1508.06615*. 00274.

Kiros, Ryan κ.ά. (2015). "Skip-Thought Vectors". Στο: *Advances in Neural Information Processing Systems*. 00363, σσ. 3294–3302.

Lipton, Zachary C., John Berkowitz και Charles Elkan (2015). "A Critical Review of Recurrent Neural Networks for Sequence Learning". Στο: *arXiv preprint arXiv:1506.00019*.

Liu, Bing (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. 00174. New York, NY: Cambridge University Press. ISBN: 978-1-107-01789-4.

Xu, Kelvin κ.ά. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." Στο: *ICML*. Τόμ. 14. 01037, σσ. 77–81.

Zaremba, Wojciech (2015). "An Empirical Exploration of Recurrent Network Architectures". Στο: 00247.

Zhang, Xiang, Junbo Zhao και Yann LeCun (2015). "Character-Level Convolutional Networks for Text Classification". Στο: *Advances in Neural Information Processing Systems*. 00268, σσ. 649–657.

Bojanowski, Piotr κ.ά. (2016). "Enriching word vectors with subword information". Στο: *arXiv preprint arXiv:1607.04606*.

Cambria, Erik κ.ά. (2016). "SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives". Στο: *The 26th International Conference on Computational Linguistics (COLING), Osaka*. 00042.

De Boom, Cedric κ.ά. (2016). "Learning Representations for Tweets through Word Embeddings". Στο: *Benelearn*. 00000.

Gal, Yarin και Zoubin Ghahramani (2016). "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks". Στο: *Advances in Neural Information Processing Systems*. 00079, σσ. 1019–1027.

Gao, Wei και Fabrizio Sebastiani (2016). "From Classification to Quantification in Tweet Sentiment Analysis". en. Στο: *Social Network Analysis and Mining* 6.1. ISSN: 1869-5450, 1869-5469. DOI: 10.1007/s13278-016-0327-z.

Isaac, Mike και Sydney Ember (2016). "For Election Day Influence, Twitter Ruled Social Media". Στο: *The New York Times*. ISSN: 0362-4331.

Joulin, Armand κ.ά. (2016). "Bag of Tricks for Efficient Text Classification". Στο: *arXiv preprint arXiv:1607.01759*.

Jungherr, Andreas (2016). "Twitter Use in Election Campaigns: A Systematic Literature Review". Στο: *Journal of Information Technology & Politics* 13.1, σσ. 72–91. ISSN: 1933-1681. DOI: 10.1080/19331681.2015.1132401.

Miyamoto, Yasumasa και Kyunghyun Cho (2016). "Gated Word-Character Recurrent Language Model". Στο: *arXiv preprint arXiv:1606.01700*. 00018.

Nakov, Preslav κ.ά. (2016a). "Evaluation Measures for the Semeval-2016 Task 4: Sentiment Analysis in Twitter (Draft: Version 1.12)". Στο: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics*. 00001.

— (2016b). "SemEval-2016 Task 4: Sentiment Analysis in Twitter". Στο: *Proceedings of SemEval*. 00116, σσ. 1–18.

Vosoughi, Soroush, Prashanth Vijayaraghavan και Deb Roy (2016). "Tweet2Vec: Learning Tweet Embeddings Using Character-Level CNN-LSTM Encoder-Decoder". en. Στο: 00014. ACM Press, σσ. 1041–1044. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914762.

Yang, Zichao κ.ά. (2016). "Hierarchical Attention Networks for Document Classification". Στο: *Proceedings of NAACL-HLT*. 00095, σσ. 1480–1489.

Baziotis, Christos, Nikos Pelekis και Christos Doulkeridis (2017a). "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-Level and Topic-Based Sentiment Analysis". Στο: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 00000. Vancouver, Canada: Association for Computational Linguistics, σσ. 747–754.

— (2017b). "DataStories at SemEval-2017 Task 6: Siamese LSTM with Attention for Humorous Text Comparison". Στο: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 00000. Vancouver, Canada: Association for Computational Linguistics, σσ. 390–395.

Cliche, Mathieu (2017). "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs". Στο: *arXiv preprint arXiv:1704.06125*. 00002.

Nickel, Maximilian και Douwe Kiela (2017). "Poincar\'e Embeddings for Learning Hierarchical Representations". Στο: *arXiv:1705.08039 [cs, stat]*. 00004. arXiv: 1705.08039 [cs, stat].

Rosenthal, Sara, Noura Farra και Preslav Nakov (2017). "SemEval-2017 Task 4: Sentiment Analysis in Twitter". Στο: *Proceedings of the 11th International Workshop on Semantic Evaluation*. 00209. Vancouver, Canada: Association for Computational Linguistics.