

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Εξαγωγή πληροφορίας από σχόλια πελατών στα κοινωνικά δίκτυα - Ανίχνευση ειρωνείας σε tweets

Πτυχιακή Εργασία

Κοντούλης Χρυσοβαλάντης Γεώργιος, 2508

Επιβλέπων Καθηγητής: Βλαχάβας Ιωάννης

ΘΕΣΣΑΛΟΝΙΚΗ

ΙΟΥΛΙΟΣ 2018

Περίληψη

Η ταχύτατη ανάπτυξη της τεχνολογίας καθώς η αστάθεια που χαρακτηρίζει την αγορά καθιστά το επιχειρηματικό περιβάλλον πολύ ανταγωνιστικό. Αυτό το γεγονός οδηγεί πολλές επιχειρήσεις στην αναζήτηση μεθόδων που θα περιορίσουν την αβεβαιότητα και θα τους προσφέρει υποστήριξη για τη σωστή και την αποδοτική λήψη αποφάσεων. Μια ευρέως χρησιμοποιημένη μέθοδος είναι τα πληροφοριακά συστήματα επιχειρηματικής ευφυΐας για την λήψη αποφάσεων. Σε αυτή τη μέθοδο χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης για την παραγωγή πληροφορίας από δεδομένα κειμένου. Παράλληλα, στη σημερινή εποχή τα μέσα μαζικής επικοινωνίας αποτελούν θεμελιώδη πυλώνα της καθημερινότητας των ανθρώπων. Έτσι, πολλές επιχειρήσεις επικεντρώνονται στην εξόρυξη πληροφορίας από τα μέσα κοινωνικής δικτύωσης. Το Twitter παρέχει αρκετά χρήσιμη πληροφορία αφού λόγω του περιορισμού των λέξεων των tweets σε 140, θεωρείται ότι περιορίζει το θέμα σε ένα ανά tweet. Ωστόσο, ο τρόπος χρήσης της ανθρώπινης γλώσσας στα κοινωνικά δίκτυα χαρακτηρίζεται από θόρυβο. Ένα συχνό φαινόμενο που παρατηρείται στα κοινωνικά δίκτυα είναι η χρήση ειρωνείας που προκαλεί θόρυβο αφού μπορεί να αντιστρέψει τελείως την εννοιολογική σημασία μια πρότασης. Τον τελευταίο καιρό, η ανίχνευση ειρωνείας έχει αποκτήσει ιδιαίτερη σημασία στον τομέα της μηχανικής μάθησης. Στα πλαίσια της πτυχιακής, θα μελετηθούν και θα υλοποιηθούν αλγόριθμοι μηχανικής μάθησης για την ανίχνευση ειρωνείας σε tweets.

Abstract

The rapid development of technology as well as the market instability makes the business environment very competitive. This leads many businesses to look for ways to reduce uncertainty and provide them with support for right and efficient decision-making. A widely used method is business intelligence information systems for decision-making. Machine learning algorithms are used in this method to generate information from text data. At the same time, the media today constitute a fundamental pillar of people's everyday life. Thus, many businesses focus on extracting information from social media. Twitter provides quite useful information as it is considered to limit the topic to one per tweet because of the restriction to the word count of tweets to 140. However, the use of human language in social networks is noisy. A frequent phenomenon observed in social networks is the use of irony that causes noise as it can completely reverse the conceptual significance of a proposal. Lately, irony detection has become particularly important in the field of machine learning. In the framework of the present thesis, machine learning algorithms will be studied and implemented to detect irony in tweets.

Ευχαριστίες

Στο παρόν σημείο, θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου σε όλους εκείνους τους ανθρώπους που συνέβαλαν, με διαφορετικό τρόπο ο καθένας, στην εκπόνηση αυτής της διπλωματικής εργασίας.

Θα ήθελα να ευχαριστήσω θερμά τον κύριο Ιωάννη Βλαχάβα, για την εμπιστοσύνη που μου έδειξε και μου ανέθεσε το παρόν θέμα, για το όραμα που μου ενέπνευσε για την Τεχνητή Νοημοσύνη και την Μηχανική Μάθηση ως καθηγητής στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, καθώς και για την υποστήριξη που μου παρείχε μέσω του Νικόλαου Στυλιανού για θέματα δόμησης και παρουσίασης του γραπτού μέρους της πτυχιακής που κατά επέκταση μου προσέφερε πολύτιμες γνώσεις για την ορθολογική αρθρογράφηση επιστημονικών εγγράφων. Επίσης, θα ήθελα να ευχαριστήσω τον ίδιο τον Νικόλαο Στυλιανού που ήταν καταλυτικός παράγοντας για την διεκπεραίωση της πτυχιακής εργασίας με την καθοδήγηση, τις διορθώσεις και τις ιδέες που μου πρόσφερε. Ακόμα, θα ήθελα να τον ευχαριστώ για τον χρόνο που διέθεσε για τις εβδομαδιαίες συναντήσεις με σκοπό την επύλυση αποριών, καθώς και για την υποστήριξη που προσέφερε μέσω email.

Ένα μεγάλο ευχαριστώ μέσω καρδιάς στην οικογένεια μου που υπήρξε πυλώνας βοήθειας σε όλους τους τομείς κατά τη διάρκεια των φοιτητικών μου σπουδών. Τέλος, οφείλω να ευχαριστήσω τους φίλους μου για την υποστηρίξη και την συμπαράσταση που μου προσφέρουν σε ότι πρόβλημα προκύπτει.

Μέσω της διπλωματικής μου εργασίας, μου δόθηκε η δυνατότητα να εμβαθύνω τις γνώσεις μου τόσο σε θέματα μηχανικής μάθησης όσο και σε θέματα οργάνωσης και παρουσίασης επιστημονικών θεμάτων. Με την ολοκλήρωση αυτής της πτυχιακής σηματοδοτείται και το τέλος μιας πολυετούς περιόδου απόκτησης γνώσεων στα πλαίσια του θεσμού του προπτυχιακού πανεπιστημίου καθώς και την εκκίνηση μιας νέας σταδιοδρομίας ως επιστήμονας πληροφορικής.

Θεσσαλονίκη, Ιούλιος 2018
Χρυσοβαλάντης Γεώργιος Κοντούλης

Περιεχόμενα

Περίληψη	III
Abstract	V
Ευχαριστίες	VII
1 Εισαγωγή	1
2 Επιχειρηματική Ευφυΐα	5
2.1 Εισαγωγή στην Επιχειρηματική Ευφυΐα	5
2.2 Ο ρόλος της Επιχειρηματικής Ευφυΐας	6
2.2.1 Λήψη Επιχειρηματικών Αποφάσεων σε συνθήκες αβεβαιότητας	6
2.2.2 Νέες τεχνολογίες και μέθοδοι ανάλυσης	8
2.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας	9
2.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας	13
2.5 Η Επιχειρηματική Ευφυΐα στην Πράξη	15
2.5.1 Διοίκηση Επιχειρησιακής Απόδοσης	15
2.5.2 Χρηματοοικονομική ανάλυση και διαχείριση	16
2.5.3 Marketing & Πωλήσεις	16
2.5.4 Διαχείριση Ανθρωπίνων Πόρων & Διαχείριση Εφοδιαστικής Αλυσίδας	17
2.5.5 Χρηματοπιστωτικός τομέας	18
2.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας	18
2.7 Ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας	20
2.8 Ο Κύκλος Ζωής Ανάπτυξης Συστήματος Επιχειρηματικής Ευφυΐας	21
2.8.1 Αιτιολόγηση Έργου	23
2.8.2 Οργάνωση Έργου	24
2.8.3 Ανάλυση απαιτήσεων του έργου	27
2.8.4 Σχεδιασμός	30
2.8.5 Υλοποίηση	33
2.8.6 Εφαρμογή	34
2.8.7 Αξιολόγηση	35
2.8.8 Η Επιχειρηματική Ευφυΐα Ως Υπηρεσία	37

3 Κοινωνικά Δίκτυα και Μηχανική Μάθηση	39
3.1 Κοινωνικά Δίκτυα και επιχειρήσεις	39
3.1.1 Εισαγωγή στο Web και κοινωνικός ιστός	40
3.1.2 Επιρροή του Twitter στις σημερινές επιχειρήσεις	41
3.1.3 Opinion mining και είδη	42
3.1.4 Εισαγωγή στα APIs	44
3.1.5 Εξαγωγή πληροφορίας (information extraction)	45
3.2 Μηχανική Μάθηση	52
3.2.1 Ορισμός της Μηχανικής Μάθησης	52
3.2.2 Είδη Μηχανικής Μάθησης	54
3.2.3 Αλγόριθμοι Μηχανικής Μάθησης	55
3.3 Μετρικές αξιολόγησης	75
4 Πρόβλημα, υλοποίηση σε Python και αποτελέσματα	83
4.1 Διαγωνισμός SemEval 2018 για Irony Detection	83
4.1.1 Περιγραφή του data set και υπόσταση των tweets	85
4.2 Περιγραφή υλοποίησης	88
4.2.1 Feature Extraction	89
4.2.2 Preprocessing	91
4.2.3 Αλγόριθμοι για encoding του dataset	91
4.2.4 Αλγόριθμοι Feature Selection	98
4.2.5 Αλγόριθμοι Machine Learning	101
4.2.6 Τρόπος αξιολόγησης – Evaluation	124
4.3 Τελικά αποτελέσματα, συμπεράσματα και μελλοντικοί στόχοι	126
Παράρτημα	137
Βιβλιογραφία	161

Λίστα εικόνων

Εικόνα 2.1 Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας	9
Εικόνα 2.2 Μοντέλο Υδατόπτωσης	22
Εικόνα 2.3 Μοντέλο ανάπτυξης Συστημάτων E.E.	23
Εικόνα 3.1 Παράδειγμα μοντέλου Skip-grams	49
Εικόνα 3.2 Παράδειγμα μοντέλου CBOW	50
Εικόνα 3.3 Γενικός τρόπος λειτουργίας αλγορίθμων Μηχανικής Μάθησης	53
Εικόνα 3.4 Παράδειγμα στρωμάτων ή επιπέδων Νευρωνικού Δικτύου	57
Εικόνα 3.5 Συναρτήσεις ενεργοποίησης	58
Εικόνα 3.6 Μοντέλο Τεχνητού Νευρωνικού Δικτύου	59
Εικόνα 3.7 Δίκτυο εμπρός τροφοδότησης πολλών επιπέδων με λειτουργικά σήματα και σήματα λάθους	60
Εικόνα 3.8 Παράδειγμα Perceptron με 6 εισόδους και 4 νευρώνες εξόδου	63
Εικόνα 3.9 Παράδειγμα Perceptron σαν ταξινομητής για d-διάστατα δεδομένα	63
Εικόνα 3.10 Μετατόπιση του ορίου απόφασης από το κατώφλι	64
Εικόνα 3.11 Παράδειγμα Soft-Margin και Hard-Margin SVM	67
Εικόνα 3.12 Παράδειγμα Hard-Margin SVM	68
Εικόνα 3.13 Παράδειγμα Soft-Margin SVM	69
Εικόνα 3.14 Παράδειγμα Kernel Trick	70
Εικόνα 3.15 Confusion Matrix	76
Εικόνα 3.16 Precision, Recall, Accuracy και F1-score	77
Εικόνα 3.17 Μοντέλο τυχαίας πρόβλεψης	81
Εικόνα 3.18 Μοντέλο τέλειας πρόβλεψης	81
Εικόνα 3.19 Μοντέλο κακής αποτελεσματικότητας	81
Εικόνα 3.20 Μοντέλο μέτριας αποτελεσματικότητας	81
Εικόνα 3.21 Μοντέλο καλής αποτελεσματικότητας	81
Εικόνα 3.22 Area Under Curve – AUC	81
Εικόνα 4.1 Word Cloud με τις πιο χρησιμοποιημένες λέξεις	87
Εικόνα 4.2 Τρόπος λειτουργίας του μοντέλου	88
Εικόνα 4.3 Τρόπος λειτουργίας του μοντέλου Voting Ensembles	122
Εικόνα 4.4 Καμπύλη ROC-AUC για 10-fold-cross-validation (Neural Net με Word2Vec)	126
Εικόνα Δ.1 Καλύτερα αποτελέσματα διαγωνισμού SemEval 2018 Task 3 Part A	153

Εικόνα E.1 Καμπύλη ROC-AUC για το μοντέλο Gaussian Naive Bayes με word2vec	155
Εικόνα E.2 Καμπύλη ROC-AUC για το μοντέλο K-Neighbors με PCA και TF-IDF	155
Εικόνα E.3 Καμπύλη ROC-AUC για το μοντέλο SVM με doc2vec	156
Εικόνα E.4 Καμπύλη ROC-AUC για το μοντέλο Voting Ensembles με Univariate Selection και Bigrams	156
Εικόνα E.5 Καμπύλη ROC-AUC για το μοντέλο Voting Ensembles με doc2vec	157
Εικόνα E.6 Καμπύλη ROC-AUC για το μοντέλο SVM με Bigrams	157
Εικόνα E.7 Καμπύλη Learning Curve για το μοντέλο Gaussian Naive Bayes με word2vec	158
Εικόνα E.8 Καμπύλη Learning Curve για το μοντέλο K-Neighbors με PCA και TF-IDF	158
Εικόνα E.9 Καμπύλη Learning Curve για το μοντέλο SVM με doc2vec	159
Εικόνα E.10 Καμπύλη Learning Curve για το μοντέλο Voting Ensembles με Univariate Selection και Bigrams	159
Εικόνα E.11 Καμπύλη Learning Curve για το μοντέλο Voting Ensembles με doc2vec	160
Εικόνα E.12 Καμπύλη Learning Curve για το μοντέλο SVM με Bigrams	160

Λίστα πινάκων

Πίνακας 4.1 Στατιστικά του dataset	86
Πίνακας 4.2 Αξιολόγηση Gaussian Naive Bayes	104
Πίνακας 4.3 Αξιολόγηση multinomial Naive Bayes	105
Πίνακας 4.4 Αξιολόγηση Bernoulli Naive Bayes	106
Πίνακας 4.5 Αξιολόγηση K-Neighbors	107
Πίνακας 4.6 Αξιολόγηση Logistic Regression	108
Πίνακας 4.7 Αξιολόγηση MLP Νευρωνικού Δικτύου	112
Πίνακας 4.8 Αξιολόγηση LSTM Neural Net	115
Πίνακας 4.9 Αξιολόγηση Conv1D Neural Net	118
Πίνακας 4.10 Αξιολόγηση SVM	121
Πίνακας 4.11 Αξιολόγηση Voting Ensembles	123
Πίνακας 4.12 Αξιολόγηση των τριών καλύτερων μοντέλων με 10-fold-cross-validation	128
Πίνακας 4.13 Τελική αξιολόγηση των τριών καλύτερων μοντέλων με gold set	128
Πίνακας A.1 Πλήθος εμφανίσεων των λέξεων	137
Πίνακας B.1 Αποτελέσματα αξιολόγησης Gaussian Naive Bayes	139
Πίνακας B.2 Αποτελέσματα αξιολόγησης Multinomial Naive Bayes	140
Πίνακας B.3 Αποτελέσματα αξιολόγησης Bernoulli Naive Bayes	141
Πίνακας B.4 Αποτελέσματα αξιολόγησης K-Neighbors	142
Πίνακας B.5 Αποτελέσματα αξιολόγησης Logistic Regression	143
Πίνακας B.6 Αποτελέσματα αξιολόγησης MLP Neural Network	144
Πίνακας B.7 Αποτελέσματα αξιολόγησης Long Short-Term Memory Network (LSTM)	145
Πίνακας B.8 Αποτελέσματα αξιολόγησης 1-D Convolutional Neural Network	145
Πίνακας B.9 Αποτελέσματα αξιολόγησης SVM	146
Πίνακας B.10 Αποτελέσματα αξιολόγησης Voting Ensembles	147
Πίνακας Γ.1 Καλύτερα αποτελέσματα Gaussian Naive Bayes στο gold set	149
Πίνακας Γ.2 Καλύτερα αποτελέσματα Multinomial Naive Bayes στο gold set	149
Πίνακας Γ.3 Καλύτερα αποτελέσματα Bernoulli Naive Bayes στο gold set	149
Πίνακας Γ.4 Καλύτερα αποτελέσματα K-Neighbors στο gold set	149
Πίνακας Γ.5 Καλύτερα αποτελέσματα Logistic Regression στο gold set	150
Πίνακας Γ.6 Καλύτερα αποτελέσματα MLP Neural Network στο gold set	150
Πίνακας Γ.7 Καλύτερα αποτελέσματα Long Short-Term Memory Network στο gold set	150

Πίνακας Γ.8 Καλύτερα αποτελέσματα 1-D Convolutional Neural Network στο gold set	150
Πίνακας Γ.9 Καλύτερα αποτελέσματα SVM στο gold set	151
Πίνακας Γ.10 Καλύτερα αποτελέσματα Voting Ensembles στο gold set	151

Κεφάλαιο 1 – Εισαγωγή

ΑΝΤΙΚΕΙΜΕΝΟ ΤΗΣ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο επιχειρηματικός τομέας της σημερινής εποχής χαρακτηρίζεται από έντονη ανταγωνιστικότητα και ραγδαίες αλλαγές. Αποτέλεσμα αυτών των καταστάσεων είναι η αύξηση της αβεβαιότητας κατά τη λήψη αποφάσεων, η οποία μπορεί να οδηγήσει σε μη αποτελεσματικές αποφάσεις που είναι η κύρια αιτία της αποτυχίας πολλών επιχειρήσεων. Έτσι, πολλές επιχειρήσεις αναζητούν τρόπους και λύσεις που θα περιορίσουν την αβεβαιότητα και θα τους προσφέρουν μια πιο σφαιρική όψη των πραγμάτων που θα τους οδηγήσει στη λήψη πιο αποτελεσματικών αποφάσεων. Για να επιτευχθεί αυτός ο στόχος πρέπει να χρησιμοποιηθούν πολύπλοκες διαδικασίες και μέθοδοι που χρησιμοποιούν δεδομένα για να παράγουν χρήσιμες πληροφορίες. Τα πληροφοριακά συστήματα επιχειρηματικής ευφυΐας αποτελούν την πιο διαδεδομένη μέθοδο που χρησιμοποιείται στην αγορά, ακόμα και από επιχειρήσεις κολοσσούς όπως η Microsoft.

Ένα πληροφοριακό σύστημα (Information System) αποτελείται από ένα σύνολο διαδικασιών, ανθρώπινου δυναμικού και αυτοματοποιημένων υπολογιστικών συστημάτων, που προορίζονται για τη συλλογή, εγγραφή, ανάκτηση, επεξεργασία, αποθήκευση και ανάλυση πληροφοριών. Τα πληροφοριακά συστήματα χρησιμοποιούν δεδομένα που παράγονται από την ίδια την επιχείρηση αλλά και από το περιβάλλον της, με σκοπό την εξαγωγή χρήσιμης πληροφορίας για την λήψη αποφάσεων. Πρέπει να τονιστεί ότι αν και το αποτέλεσμα που παράγει ένα πληροφοριακό σύστημα μπορεί να είναι πολύ χρήσιμο, η τελική απόφαση πρέπει να ληφθεί από τον άνθρωπο με γνώμονα το αποτέλεσμα του συστήματος αλλά και με χρήση της εμπειρίας του και των γνώσεων του.

Η απότομη άνοδος της δημοτικότητας των μέσων κοινωνικής δικτύωσης αποτελεί μια σπουδαία πηγή άντλησης πληροφοριών. Συγκριμένα το Twitter, λόγου του περιορισμού του μεγέθους των tweet στις 140 λέξεις, αποτελεί σπουδαία πηγή πληροφόρησης αφού ο περιορισμός των λέξεων επιτρέπει την αρθρογράφηση tweet που εστιάζουν μόνο σε ένα θέμα. Ωστόσο, τα δεδομένα που πηγάζουν από τα κοινωνικά δίκτυα, όπως είναι το Twitter, περιέχουν πολύ θόρυβο που αποτελεί εμπόδιο στην αποδοτική επεξεργασία τους. Ιδιαίτερα, η χρήση ειρωνείας στα tweets προσθέτει θόρυβο στη σημασία τους καθώς μπορεί να αλλάξει

ολόκληρη την σημασία του tweet. Σε επιχειρησιακά περιβάλλοντα η επεξεργασία αυτών των δεδομένων από τον άνθρωπο δεν είναι εφικτή, λόγω του μεγάλου πλήθους δεδομένων.

Αυτό το πρόβλημα έρχεται να λύσει η μηχανική μάθηση. Μηχανική Μάθηση είναι η μελέτη αλγορίθμων υπολογιστών που μπορούν να βελτιωθούν μόνοι τους μέσω εμπειρίας που αποκτούν. Υπάρχουν δυο είδη μηχανικής μάθησης, μάθηση με επίβλεψη (supervised learning) και μάθηση χωρίς επίβλεψη (unsupervised learning). Στη μάθηση με επίβλεψη το σύστημα χρησιμοποιεί ένα σύνολο δεδομένων εκμάθησης με σκοπό να μάθει επαγωγικά μια συνάρτηση στόχο, η οποία θα είναι γενικευμένη για εισόδους με άγνωστη έξοδο, την οποία θα προβλέψει. Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει να δημιουργήσει μόνο του συσχετίσεις ή ομάδες από ένα σύνολο δεδομένων που δέχεται ως είσοδο και βασιζόμενο μόνο στις ιδιότητες τους να ανακαλύψει πρότυπα, χωρίς να ξέρει αν υπάρχουν και ποια είναι.

ΣΚΟΠΟΣ ΤΗΣ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Η πτυχιακή εστιάζει στη ανάπτυξη μοντέλων επιχειρηματικής ευφυΐας, τόσο σε θεωρητικό επίπεδο όσο και σε πρακτικό, και την εκμετάλλευση τους από τις επιχειρήσεις για την αποδοτική λήψη αποφάσεων. Αρχικά, γίνεται μια εκτενής περιγραφή της έννοιας επιχειρηματική ευφυΐας καθώς και αναλυτική περιγραφή του ρόλου που παίζει σε μια σύγχρονη επιχείρηση. Παράλληλα, αναλύονται τα δομικά επίπεδα της ανάπτυξης ενός μοντέλου επιχειρηματικής ευφυΐας στα πλαίσια μιας επιχείρησης και αναλύεται η χρήση της στην πράξη. Στη συνέχεια, αναπτύσσονται θεωρητικά κάποιοι διαδεδομένοι αλγόριθμοι μηχανικής μάθησης ενώ θα υλοποιηθούν κατά το πρακτικό κομμάτι.

Πιο συγκεκριμένα, σκοπός είναι να αναπτυχθεί ένα μοντέλο που χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για να αναγνωρίσει αν ένα tweet είναι ειρωνικό. Η μηχανική μάθηση αποτελεί έναν κλάδο με αυξανόμενη δημοτικότητα, το οποίο οφείλεται στην αποδοτικότητα των αποτελεσμάτων που παρέχει. Το πρακτικό κομμάτι της πτυχιακής θα επικεντρωθεί σε αλγορίθμους μηχανικής μάθησης με επίβλεψη. Ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη μοντελοποιεί και χρησιμοποιεί παλαιά πληροφορία, την οποία επεξεργάζεται με τη χρήση πολύπλοκων μαθηματικών μοντέλων, με σκοπό να παράξει προβλέψεις για μελλοντικές καταστάσεις που δεν μπορούμε να γνωρίζουμε εκ των προτέρων το αποτέλεσμα τους.

Το πρακτικό κομμάτι υλοποιείται στη γλώσσα προγραμματισμού Python. Πρέπει να σημειωθεί ότι χρησιμοποιήθηκαν δεδομένα, εκπαίδευσης (train set) και πρόβλεψης (test set), από το διαγωνισμό του SemEval 2018. Αν και δεν δηλώθηκε επίσημη συμμετοχή στον δια-

γωνισμό, λόγω ασυμφωνίας χρονικών περιόδων με την πτυχιακή, θα συγκριθούν οι μετρικές των μοντέλων του διαγωνισμού με τις μετρικές που παράγουν τα μοντέλα που υλοποιήθηκαν στα πλαίσια της πτυχιακής. Σκοπός αυτής της σύγκρισης είναι η διαπίστωση της αποδοτικότητας των μοντέλων της πτυχιακής σε σχέση με αυτών που παράχθηκαν σε ένα ανταγωνιστικό περιβάλλον όπως αυτό που θεσπίζει ο διαγωνισμός. Πιο συγκεκριμένα, θα γίνει αναλυτική περιγραφή όλων των αλγορίθμων που δοκιμάστηκαν, τόσο σε θεωρητικό επίπεδο όσο και σε πρακτικό. Τέλος, θα συλλεχθούν τα αποτελέσματα όλων των αλγορίθμων και θα γίνει παρουσίαση των καλύτερων αποτελεσμάτων.

ΔΟΜΗ ΤΗΣ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Το **2ο Κεφάλαιο** εισάγει την έννοια της επιχειρηματικής ευφυΐας και τη σημασία της στη σύγχρονη επιχείρηση. Στη συνέχεια, αναλύονται τα επίπεδα ενός μοντέλου επιχειρηματικής ευφυΐας, ενώ περιγράφεται και το πως χρησιμοποιείται στην πράξη. Τονίζονται, τα οφέλη και οι περιορισμοί, ενώ γίνεται μια γενική αναφορά των παρόχων λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας με σκοπό να δώσουν μια γενική εικόνα της αγοράς και των προϊόντων. Έπειτα, περιγράφεται η ανάπτυξη συστημάτων επιχειρηματικής ευφυΐας και ο κύκλος ζωής της ανάπτυξης των συστημάτων, καθώς και η σημαντικότητα και η άμεση σύνδεση τους.

Το **3ο Κεφάλαιο** κάνει μια εισαγωγή στα κοινωνικά δίκτυα και τον ρόλο που παίζουν στην σημερινή επιχείρηση. Ακολουθεί, μια εισαγωγή στον κλάδο της μηχανικής μάθησης και παρουσιάζονται βασικοί ορισμοί και έννοιες. Γίνεται ανάλυση των δύο ειδών μηχανικής μάθησης, δηλαδή της μάθησης με επίβλεψη και της μάθησης χωρίς επίβλεψη. Στη συνέχεια, το κεφάλαιο εστιάζει σε συγκριμένους αλγορίθμους που αφορούν την κωδικοποίηση των δεδομένων σε αριθμούς, όπως word embeddings, καθώς και στην ανάλυση αλγορίθμων μηχανικής μάθησης. Για την καλύτερη κατανόηση των παραπάνω παραθέτονται παραδείγματα.

Το **4ο Κεφάλαιο** περιγράφει αναλυτικά την υλοποίηση της εφαρμογής που στοχεύει στην αποτελεσματική πρόβλεψη της ειρωνείας ενός tweet. Πιο συγκεκριμένα, περιγράφονται όλα τα στάδια του προγραμματισμού της εφαρμογής, από την προετοιμασία των δεδομένων μέχρι και την ερμηνεία του αποτελέσματος. Επίσης, αναφέρονται οι βιβλιοθήκες των αλγορίθμων που χρησιμοποιήθηκαν με σκοπό την επεξήγηση της παραμετροποίησης που χρησιμοποιήθηκε και τον τρόπο που η κάθε παράμετρος επηρεάζει την λειτουργία του κάθε αλγορίθμου. Τέλος, παρουσιάζονται διάφορα στατιστικά για τα δεδομένα καθώς και διαγράμματα που είναι σχετικά με την αποδοτικότητα των αλγορίθμων.

Κεφάλαιο 2 - Επιχειρηματική Ευφυΐα

Περιεχόμενα κεφαλαίου

2.1 Εισαγωγή στην Επιχειρηματική Ευφυΐα	5
2.2 Ο ρόλος της Επιχειρηματικής Ευφυΐας	6
2.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας	9
2.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας	13
2.5 Η Επιχειρηματική Ευφυΐα στην Πράξη	15
2.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας	18
2.7 Ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας	20
2.8 Ο Κύκλος Ζωής Ανάπτυξης Συστήματος Επιχειρηματικής Ευφυΐας	21

2.1 Εισαγωγή στην Επιχειρηματική Ευφυΐα

Στην εποχή μας ο επιχειρηματικός τομέας χαρακτηρίζεται από τον πλούτο του σε νέες δυνατότητες και ευκαιρίες αλλά ταυτόχρονα και από τις δυσκολίες που είναι απόρροια της πρόσφατης οικονομικής κρίσης. Η αντιμετώπιση αυτών των νέων δυσκολιών απαιτεί την αναβάθμιση των διοικητικών μεθόδων και την βελτίωση των διαδικασιών λήψης αποφάσεων. Η βελτίωση των αποφάσεων καθιστά αναγκαίο την βαθιά κατανόηση και γνώση της επιχείρησης αλλά και του περιβάλλοντος της, καθώς και την έγκαιρη πληροφόρηση αφού η πληροφορία στις μέρες μας είναι ένα πολύ ισχυρό μέσο για την επιτυχία.

Αν και ο όρος Επιχειρηματική Ευφυΐα (Business Intelligence) δεν είναι πρόσφατος, αφού πρωτοεμφανίζεται το 1865, δεν είναι σαφώς ορισμένος. Έτσι ορίζουμε Επιχειρηματική Ευφυΐα ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Η Επιχειρηματική Ευφυΐα δίνει την δυνατότητα σε έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η εξαγόμενη πληροφορία μετατρέπεται σε γνώση που χρησιμοποιείται από τα διοικητικά στελέχη ώστε να καταστρώσουν ένα αποτελεσματικό σχέδιο

δράσης που θα οδηγήσει στην επιτυχία των επιχειρηματικών στόχων τους. Στο κεφάλαιο 2 περιγράφεται η Επιχειρηματική Ευφυΐα με βάση το βιβλίο “Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων”. [1]

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι εξειδικευμένα πληροφοριακά συστήματα, τα οποία χρησιμοποιούν ποιοτική πληροφορία η οποία δέχεται επεξεργασία από λογισμικό που διεξάγει ειδικές αναλύσεις. Αυτά τα συστήματα προσφέρουν ταχύτερη πρόσβαση στην πληροφορία, ευκολότερη υποβολή ερωτημάτων στο σύστημα, προχωρημένη ανάλυση των δεδομένων, καθώς και βελτίωση της ποιότητας των δεδομένων με αποτέλεσμα την βελτίωση της ποιότητας της πληροφορίας. Αυτό επιτυγχάνεται με Αποθήκες Δεδομένων (Data Warehouse) και με τεχνικές OLAP (OnLine Analytical Processing). Στις Αποθήκες Δεδομένων τα δεδομένα αφού υποστούν κατάλληλη επεξεργασία αποθηκεύονται σε συγκεντρωτική μορφή (πχ πωλήσεις ανά μήνα ή ανά κατηγορία προϊόντος). Με τις τεχνικές OLAP ο χρήστης μπορεί να προβάλει και να αναλύσει τα δεδομένα σε διάφορα επίπεδα γενίκευσης (π.χ. πωλήσεις ανά μήνα ή ανά τρίμηνο ή ανά έτος).

Στις μέρες μας λόγω μεγάλης αύξησης του όγκου δεδομένων δημιουργείται η ανάγκη για νέους μεθόδους μετατροπής των δεδομένων σε γνώση. Σε αυτή την ανάγκη δίνει λύση η Εξόρυξη Δεδομένων. Η Εξόρυξη Δεδομένων (Data Mining) ή Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) στοχεύει στην ανακάλυψη γνώσης που είναι κρυμμένη σε μεγάλους όγκους δεδομένων.

2.2 Ο ρόλος της Επιχειρηματικής Ευφυΐας

Όπως προαναφέρθηκε, η Επιχειρηματική Ευφυΐα παίζει πολύ βασικό ρόλο στην ανάπτυξη και λειτουργία των σύγχρονων μεγάλων επιχειρήσεων. Οι κυριότερες αιτίες γι' αυτό το γεγονός είναι οι ακόλουθες:

2.2.1 Λήψη Επιχειρηματικών Αποφάσεων σε συνθήκες αβεβαιότητας

Κατά την λειτουργία μιας επιχείρησης λαμβάνονται πολλές και ποικίλες αποφάσεις τόσο ως προς το είδος αλλά και τόσο ως προς το επίπεδο δυσκολίας. Γενικά μία απόφαση χαρακτηρίζεται δύσκολη όταν υπάρχει μεγάλη αβεβαιότητα για τα αποτελέσματα που θα αποφέρει. Έτσι αποφάσεις που σχετίζονται με ζητήματα καθημερινής λειτουργίας είναι συνήθως απλές και τυποποιημένες σε επίπεδο που μπορούν να ληφθούν με τη χρήση κατάλληλου λογισμικού. Άλλες αποφάσεις που αφορούν ευρύτερα τμήματα της επιχείρησης ή ζητήματα

στρατηγικού προσανατολισμού είναι πιο περίπλοκες καθώς περιέχουν μεγάλο βαθμό αβεβαιότητας. Για παράδειγμα, η δημιουργία ενός νέου καινοτόμου προϊόντος είναι ιδιαίτερα απαιτητική αφού πρέπει να ληφθούν υπόψη στοιχεία όπως το περιβάλλον του οργανισμού το οποίο μεταβάλλεται συνεχώς, οι τεχνολογικές εξελίξεις, οι προτιμήσεις και οι ανάγκες των πελατών, ο οικονομικός τομέας κτλ. Γενικότερα, υπάρχουν πολλοί παράγοντες που αυξάνουν τον βαθμό πολυπλοκότητας όπως το συνεχώς μεταβαλλόμενο περιβάλλον, η πίεση χρόνου ως αποτέλεσμα της αύξησης του ρυθμού λειτουργίας, το στοιχείο του ανταγωνισμού και η διεύρυνση σε μέγεθος των επιχειρήσεων.

Ένας ακόμα σημαντικός παράγοντας που αξίζει να αναφερθεί είναι η παγκοσμιοποίηση καθώς επηρεάζει το επιχειρηματικό περιβάλλον και το αλλάζει με έντονους ρυθμούς. Οι οργανισμοί πλέον δραστηριοποιούνται σε παγκόσμια κλίμακα πράγμα που κλιμακώνει τον ανταγωνισμό. Αυτό σημαίνει ότι οι επιχειρήσεις είναι διασκορπισμένες σε πολλές χώρες γεγονός που καθιστά δυσκολότερη την παρακολούθηση και την διοίκηση τους. Επίσης, η επέκταση των επιχειρήσεων σε παγκόσμια κλίμακα δημιουργεί προβλήματα που προκύπτουν από τις διαφορετικές κουλτούρες. Ανάμεσα σε διαφορετικούς πολιτισμούς οι ιδεολογίες μπορεί να διαφέρουν σε μεγάλο επίπεδο καθώς κάτι που θεωρείται ελκυστικό από έναν πολιτισμό μπορεί να θεωρείται κακόγουστο ή ακόμα και προσβλητικό από κάποιον άλλο.

Τέτοια προβλήματα μπορούν να συναντηθούν ακόμα και μέσα στην ίδια την επιχείρηση αφού το εργατικό δυναμικό μπορεί να έχει διαφορετικές κουλτούρες και θρησκευτικές αντιλήψεις με αποτέλεσμα την διαφορετική τους αντίδραση σε θέματα που τους επηρεάζουν άμεσα όπως είναι οι εργασιακές πολιτικές παρακίνησης των εργαζομένων. Ακόμα, οι εισαγωγή των επιχειρήσεων στην παγκόσμια αγορά έχει θιχτεί από την οικονομική κρίση πράγμα που περιορίζει τα οικονομικά κεφάλαια που διαθέτουν οι πελάτες και κατά επέκταση οι ίδιες οι επιχειρήσεις γεγονός που καθιστά ακόμα πιο δύσκολο το μάνατζμεντ τους. Επιπροσθέτως, με την ανάπτυξη του διαδικτύου ο καταναλωτής έχει εξελιχθεί και πλέον είναι καλύτερα πληροφορημένος και μορφωμένος ενώ παράλληλα έχει υψηλό εισόδημα το οποίο συνεπάγεται με υψηλότερες απαιτήσεις γεγονός που αποτελεί νέα πρόκληση για τις επιχειρήσεις.

Όλοι οι παραπάνω παράγοντες καθιστούν το επιχειρηματικό περιβάλλον ιδιαίτερα σύνθετο και αβέβαιο. Για την αντιμετώπιση αυτών το προκλήσεων χρειάζεται ιδιαίτερα αποτελεσματική διοίκηση. Τα διοικητικά στελέχη των επιχειρήσεων λαμβάνουν αποφάσεις στηριζόμενοι στην διαθέσιμη γνώση σχετικά με το αντικείμενο τους, την εμπειρία τους και τις διαθέσιμες πληροφορίες. Συνεπώς, είναι απαραίτητη η σωστή και κατάλληλη πληροφόρηση για

τη λήψη πετυχημένων αποφάσεων. Κατάλληλη πληροφόρηση σημαίνει ότι την κατάλληλη χρονική στιγμή το κατάλληλο άτομο έχει την σωστή πληροφορία. Με βελτιωμένες αποφάσεις επιτυγχάνεται η βελτίωση της επίδοσης του οργανισμού και κατά επέκταση την αύξηση της ανταγωνιστικότητας του απέναντι σε άλλους αντίπαλους οργανισμούς. Στην επίτευξη αυτού του στόχου μπορούν να συμβάλλουν τα συστήματα Επιχειρηματικής Ευφυΐας προσφέροντας πληροφόρηση και μειώνοντας τον βαθμό αβεβαιότητας κατά τη λήψη αποφάσεων.

2.2.2 Νέες τεχνολογίες και μέθοδοι ανάλυσης

Η ανάλυση δεδομένων και η εξαγωγή συμπερασμάτων από αυτά γινόταν παλαιότερα αποκλειστικά με χρήση στατιστικών μεθόδων, όμως η αύξηση του όγκου δεδομένων οδήγησε στην αναζήτηση νέων τεχνικών όπως είναι η πολυδιάστατη ανάλυση με χρήση Αποθηκών Δεδομένων και κύβων. Στην σημερινή εποχή η Εξόρυξη Δεδομένων δίνει λύση σε πολλά και πολύπλοκα προβλήματα που αφορούν την επεξεργασία δεδομένων και την ανακάλυψη της γνώσης. Πιο συγκεκριμένα δίνει λύση σε προβλήματα επεξεργασίας μεγάλου όγκου δεδομένων, χρησιμοποιεί συγκεκριμένες μεθοδολογίες για την διατύπωση των τελικών συμπερασμάτων και για την προεπεξεργασία των δεδομένων που καλύπτουν περιπτώσεις όπως οι χαμένες τιμές, ο θόρυβος, ο κατάλληλος μετασχηματισμός των δεδομένων κλπ.

Επίσης, οι μέθοδοι της Εξόρυξης Δεδομένων κάνουν χρήση μεθόδων Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και Αναγνώρισης Προτύπων το οποίο σύμφωνα με έρευνες είναι πιο αποδοτικό από την χρήση Στατιστικής και παράλληλα ενώ η μέθοδος της στατιστικής απαιτεί την εκ των προτέρων διατύπωση υποθέσεων οι μέθοδοι της Εξόρυξης Δεδομένων χρησιμοποιεί μοντέλα που προκύπτουν κατευθείαν από τα δεδομένα με κατάλληλη επεξεργασία. Επιπροσθέτως οι αναλυτές μπορούν να κάνουν χρήση των νέων μεθόδων για προγνωστικές αναλύσεις χρησιμοποιώντας δεδομένα του παρελθόντος για τη διατύπωση προβλέψεων για το μέλλον.

Σύμφωνα με όσα έχουν αναφερθεί ο σύγχρονος αναλυτής έχει στη διάθεση του τα εργαλεία που χρειάζονται για την επεξεργασία τεράστιων όγκων δεδομένων και την άντληση πληροφορίας από αυτά η οποία είναι πολύτιμη για τη λήψη αποφάσεων. Καταληκτικά, η λήψη επιχειρηματικών αποφάσεων πρέπει να γίνεται με τέτοιο τρόπο ώστε να διαχειρίζεται αποτελεσματικά το στοιχείο της αβεβαιότητας και τις νέες προκλήσεις που ήρθαν στο προσκήνιο με την παγκοσμιοποιημένη οικονομία και την πρόσφατη οικονομική κρίση. Συνεπώς, δη-

μιουργείται η ανάγκη για ποιοτική και έγκαιρη πληροφόρηση. Στο συγκεκριμένο ζήτημα δίνει λύση η μαζική χρήση της πληροφορικής που προσφέρει τα αναγκαία δεδομένα, ενώ για την επεξεργασία τους οι νέες μεθοδολογίες ανάλυσης είναι κατάλληλοι για την παραγωγή χρήσιμης πληροφορίας. Το αποτέλεσμα των παραπάνω είναι η άνθιση της Επιχειρηματικής Ευφυΐας.

2.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι δομημένα σε επίπεδα που συγκροτούν μια πυραμίδα. Στην βάση της πυραμίδας βρίσκονται τα ακατέργαστα δεδομένα ενώ στην κορφή η λήψη των τελικών αποφάσεων. Καθώς ανεβαίνουμε από την βάση προς την κορυφή της πυραμίδας αυξάνεται η υποστήριξη που παρέχετε για την λήψη επιχειρηματικών αποφάσεων. Παρακάτω παρουσιάζεται η πυραμίδα Συστημάτων της Επιχειρηματικής Ευφυΐας (Εικόνα 2.1).



Εικόνα 2.1 Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας

Δεδομένα και Μεταδεδομένα

Ως δεδομένα ορίζουμε ένα μη αξιολογημένο σύνολο διακριτών στοιχείων τα οποία προσδίδουν «τιμές» επί αντικειμένων, προσώπων, γεγονότων κλπ. Τα δεδομένα μπορεί να περι-

λαμβάνουν λέξεις, αριθμούς, σύμβολα, σχέδια, φωτογραφίες κλπ που περιγράφουν ή αντιπροσωπεύουν ποσότητες, έννοιες, ιδέες, αντικείμενα, γεγονότα, καταστάσεις και λειτουργίες.

Τα μεταδεδομένα είναι δεδομένα που περιγράφουν άλλα δεδομένα με σκοπό να εξηγεί, εντοπίζει ή κάνει ευκολότερη την ανάκτηση και διαχείριση πληροφορίας. Τα μεταδεδομένα διακρίνονται σε περιγραφικά (Descriptive metadata), τα οποία περιγράφουν ένα αντικείμενο ώστε να το αναγνωρίσουμε και να το ανακτήσουμε, σε δομικά (Structural metadata), τα οποία περιγράφουν περιγράφουν σύνθετα αντικείμενα από τα συστατικά τους και τέλος σε επιχειρησιακά (Administrative metadata), τα οποία περιλαμβάνουν άλλες βοηθητικές πληροφορίες όπως για παράδειγμα πότε δημιουργήθηκε ένα αντικείμενο, τύπος αρχείου ποιος έχει δικαιώματα προσπέλασης κλπ. Επίσης τα μεταδεδομένα προσφέρουν ευκολία αρχειοθέτησης (Archiving), συντήρησης (Preservation) και επιπλέον διαλειτουργικότητα αφού περιγράφουν τα αντικείμενα με τέτοιο τρόπο ώστε να είναι κατανοητά και από τον άνθρωπο αλλά και από τις μηχανές. Διαλειτουργικότητα ορίζεται η ικανότητα διαφορετικών συστημάτων να ανταλλάσουν δεδομένα με ελάχιστο κόστος λειτουργικότητας. [2]

Τα δεδομένα με τα μεταδεδομένα δεν έχουν ουσιαστική διαφορά αφού είναι κυρίως θέμα προοπτικής. Για παράδειγμα ένα τίτλος ενός κειμένου εκτός του ότι είναι τίτλος δηλαδή μεταδεδομένο είναι και κομμάτι του κειμένου δηλαδή δεδομένο. Επίσης τα δεδομένα και τα μεταδεδομένα μπορούν να αλλάξουν ρόλους καθώς και μπορούν να υπάρξουν μεταδεδομένα για τα μεταδεδομένα, το οποίο μπορεί να βοηθήσει στην αρχειοθέτηση για να υπάρχει έλεγχος όταν για παράδειγμα γίνεται συγχώνευση δύο αρχείων.

Πηγές Δεδομένων & Διαθεσιμότητα δεδομένων

Στη βάση της πυραμίδας βρίσκονται οι πηγές δεδομένων οι οποίες μπορεί να είναι συστήματα παρακολούθησης συναλλαγών, εταιρικές βάσεις δεδομένων, δικτυακοί servers, εσωτερικά έγγραφα ή εξωτερικές πηγές. Αυτά τα δεδομένα είναι σημαντικά για την καθημερινή λειτουργία της επιχείρησης όμως είναι ακατάλληλα για την λήψη αποφάσεων αφού είναι διάσπαρτα σε διάφορες πηγές και άρα πρέπει να ενοποιηθούν ενώ παράλληλα είναι και υπερβολικά αναλυτικά το οποίο καθιστά αναγκαίο την κατάλληλη επεξεργασία τους ώστε να αντιμετωπιστούν και τυχόν σφάλματα. Στην σημερινή εποχή, κάθε επιχείρηση διαθέτει συστήματα για την καταγραφή δεδομένων από συναλλαγές και άλλες δραστηριότητες.

Μια κατηγορία από αυτά τα συστήματα είναι τα Συστήματα Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning (ERP)) τα οποία επιτρέπουν την παρακολούθηση συ-

ναλλαγών σε όλες τις λειτουργικές περιοχές ενός οργανισμού μέσα από ένα ενιαίο περιβάλλον. Άλλα διαδεδομένα συστήματα παρακολούθησης συναλλαγών είναι τα Συστήματα Διαχείρισης Εφοδιαστικής Αλυσίδας (Supply Chain Management (SCM)) και τα Συστήματα Διαχείρισης Σχέσεων Πελατών (Customer Relationship Management (CRM)). Όλα αυτά τα συστήματα καταγράφουν καθημερινά τεράστιους όρκους δεδομένων που αφορούν τις δραστηριότητες της επιχείρησης. Επιπλέον καταγραφή δεδομένων γίνεται και από τη χρήση διάφορων συσκευών όπως είναι τα barcode readers, τα συστήματα ετικετών RFID, τα συστήματα GPS, οι κάμερες κλπ.

Η ανάπτυξη της τεχνολογίας και του διαδικτύου καθιστούν απαραίτητη την ύπαρξη ιστοσελίδας για κάθε επιχείρηση. Αυτές οι ιστοσελίδες παράγουν δεδομένα από διάφορους επισκέπτες αλλά κατά κύριο λόγο παράγουν δεδομένα από πελάτες τα οποία μπορεί να αφορούν σχόλια πελατών για τα προϊόντα της επιχείρησης ή το πλήθος που εισέρχεται χρηστών στην σελίδα. Ωστόσο, τα δεδομένα αυτά είναι κατά κανόνα αδόμητα που είναι και η βασική διαφορά από αυτά που αφορούν τις συναλλαγές, τα οποία είναι δομημένα.

Πέρα από τα δεδομένα που παράγονται από τα προαναφερθείσα συστήματα μια επιχείρηση λαμβάνει δεδομένα και από τρίτους φορείς όπως είναι οι κρατικές υπηρεσίες, τα μέσα ενημέρωσης, οι τράπεζες και άλλες επιχειρήσεις. Ακόμα, μια τεράστια και διαρκώς αυξανόμενη πηγή δεδομένων είναι το Web 2.0 κοινωνικής δικτύωσης, blogs, wikis και γενικά ιστοσελίδες που επιτρέπουν την λεκτική έκφραση των χρηστών.

Αυτή η υπερσυσσώρευση των δεδομένων καθώς οι τεχνικές επεξεργασίας τους και η άντληση πληροφορίας από αυτά ορίζεται ως «Big Data». Η επεξεργασία αυτού του τεράστιου όγκου πληροφορίας μπορεί να ανακαλύψει καταναλωτικές τάσεις και επιχειρηματικές ευκαιρίες. Όμως η επεξεργασία αυτών των δεδομένων δεν είναι εύκολη υπόθεση αφού πολλές φορές τα δεδομένα μπορεί να είναι διάσπαρτα σε διάφορες πηγές και να περιέχουν ελλιπή ή αντιφατικά στοιχεία. Συνεπώς αφού αυτές οι πληροφορίες μπορεί να περιέχουν σημαντική πληροφορία η επεξεργασία τους είναι απαραίτητη. Τα συστήματα Επιχειρηματικής Ευφυΐας στοχεύουν στην επεξεργασία όλων αυτών των δεδομένων και στην ανακάλυψη πολύτιμης πληροφορίας που θα χρησιμοποιηθεί για τη λήψη αποφάσεων.

Αποθήκες Δεδομένων

Οι Αποθήκες Δεδομένων αποτελούνται από βάσεις δεδομένων που περιέχουν συγκεντρωτικά και καθαρά δεδομένα τα οποία πρόκειται να αναλυθούν για την εξαγωγή συμπερα-

σμάτων. Οι Αποθήκες Δεδομένων είναι θεματικά προσανατολισμένες αφού επικεντρώνονται σε συγκεκριμένους τομείς και έτσι πρέπει να γίνει απομόνωση των σχετικών δεδομένων από τα μη σχετικά. Επίσης τα δεδομένα πρέπει να κατηγοριοποιηθούν σύμφωνα με θέματα ενδιαφέροντος της διοίκησης, για παράδειγμα πωλήσεις ανά περιοχή ή χρονική περίοδο ή ανά κατηγορία προϊόντος, καθώς πρέπει και να οριστεί ο βαθμός λεπτομέρειας, για παράδειγμα πωλήσεις ανά εβδομάδα ή μήνα κοκ. Σε τακτά χρονικά διαστήματα εφαρμόζονται εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων, γνωστές ως εργασίες ETL (Extract, Transform, Load).

Διερεύνηση Δεδομένων

Στο τρίτο επίπεδο ανεβαίνοντας από τη βάση της πυραμίδας βρίσκεται η Διερεύνηση Δεδομένων όπου εκτελούνται αρχικού επιπέδου επεξεργασίες στα δεδομένα. Ο χρήστης υποβάλλει ερωτήματα (queries) στη βάση δεδομένων, λαμβάνει απαντήσεις και συντάσσει αναφορές. Οι αναφορές μπορεί να περιέχουν αριθμητικές τιμές αλλά προτιμότερο είναι η οπτική παρουσίαση των αποτελεσμάτων αφού βοηθούν στην κατανόηση τους. Σε αυτό το επίπεδο μπορεί να γίνει μια αρχική στατιστική επεξεργασία δεδομένων ενώ κύριο χαρακτηριστικό είναι η δημιουργία υποθέσεων από τον χρήστη και στη συνέχεια η εκτέλεση ελέγχων με τη χρήση κατάλληλων εργαλείων ανάλυσης για να επιβεβαιωθεί ότι υποστηρίζονται από τα δεδομένα.

Εξόρυξη Δεδομένων

Στο τέταρτο στάδιο βρίσκεται η Εξόρυξη Δεδομένων. Σε αυτό το επίπεδο εκτελείται υψηλού επιπέδου ανάλυση δεδομένων με τη χρήση προχωρημένων στατιστικών μεθόδων αλλά και με τη χρήση μεθόδων της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης. Χρησιμοποιούνται μέθοδοι κατηγοριοποίησης (classification) οι οποίοι προβλέπουν την κατηγορία που ανήκει ένα αντικείμενο με βάση τα χαρακτηριστικά του. Επίσης χρησιμοποιούνται μέθοδοι ανάλυσης συστάδων (cluster analysis) οι οποίοι εντοπίζουν ομάδες παρόμοιων αντικειμένων. Ένα σημαντικό χαρακτηριστικό είναι ότι σε αυτό το επίπεδο ο χρήστης δεν χρειάζεται να διατυπώσει αρχικές υποθέσεις αφού ειδικοί αλγόριθμοι επεξεργάζονται τα δεδομένα και εξάγουν πληροφορία από αυτά η οποία συχνά αποτελεί ένα μοντέλο.

Βελτιστοποίηση

Το επίπεδο Βελτιστοποίησης ευθύνεται για τον εντοπισμό της βέλτιστης λύσης από το σύνολο των λύσεων που προέκυψαν από τις αναλύσεις των προηγούμενων επιπέδων. Για τη διεκπεραίωση αυτής της διεργασίας χρησιμοποιούνται διάφοροι μέθοδοι όπως είναι οι ευρετικές μέθοδοι (heuristics) ή ο Γραμμικός Προγραμματισμός. Γενικά τα προβλήματα χωρίζονται σε τρεις κατηγορίες ανάλογα με το πλήθος πιθανών λύσεων που έχουν. Τα διχοτόμα προβλήματα μπορούν να έχουν δύο δυνατές λύσεις, τα προβλήματα πολλαπλών λύσεων έχουν ένα πάνω όριο ως προς το πλήθος των λύσεων που μπορεί να έχουν και τέλος υπάρχουν τα προβλήματα απεριόριστου αριθμού λύσεων.

Λήψη απόφασης

Στην κορυφή της πυραμίδας γίνεται η τελική λήψη απόφασης. Είναι σημαντικό να τονιστεί ότι όλες οι μέθοδοι και τα συστήματα που αναφέρθηκαν παραπάνω έχουν στόχο την διευκόλυνση της λήψης απόφασης από τον άνθρωπο και όχι την αυτοματοποιημένη λήψη απόφασης από τον υπολογιστή. Η διευκόλυνση της λήψης απόφασης οφείλεται στην καλύτερη πληροφόρηση του ανθρώπου. Έτσι έχοντας υπόψη την επιπλέον πληροφόρηση που του παρέχεται θα χρησιμοποιήσει τη λογική του, τη γνώση του, τις ικανότητες του, τη φαντασία του, το ένστικτο του, τον χαρακτήρα του και τέλος τη διαίσθηση του για την τελική λήψη της απόφασης.

2.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας

Τα συστήματα Ε.Ε. επεξεργάζονται δεδομένα κάνοντας χρήση τεχνολογιών της Πληροφορικής και εξάγουν πληροφορία χρήσιμη για την καλύτερη λειτουργία της επιχείρησης. Βέβαια, όπως κάθε τεχνολογία, προσφέρει κάποια οφέλη αλλά έχει και κάποιους περιορισμούς.

Οφέλη της Επιχειρηματικής Ευφυΐας

Τα οφέλη της χρήσης της Επιχειρηματικής Ευφυΐας στις επιχειρήσεις είναι πολυάριθμα αλλά κάποια ξεχωρίζουν αφού συνδέονται άμεσα με την λειτουργία της. Η σωστή και έγκαιρη πληροφόρηση της διοίκησης βοηθάει στην βελτίωση της ποιότητας των αποφάσεων και κατά επέκταση στη διαμόρφωση στρατηγικών στόχων. Η σωστή πληροφόρηση συνδράμει στην καλύτερη κατανόηση των πελατών, του ανταγωνιστικού περιβάλλοντος αλλά και στο

χώρο των αγορών, των προμηθειών και των πόρων και συμπερασματικά θα υπάρξει αύξηση στις πωλήσεις, μείωση του κόστους εφοδιασμού και τέλος αύξηση των κερδών. Επιπλέον, η επίτευξη συγκριτικού πλεονεκτήματος είναι προϊόν της αύξησης του κέρδους, της μείωσης του κόστους τα οποία συνεισφέρουν στην αύξηση της αποτελεσματικότητας και τις αποδοτικότητας της διοίκησης και συνεπώς στην αύξησης της ανταγωνιστικότητας της επιχείρησης.

Επιπροσθέτως, στην αύξηση της ανταγωνιστικότητας της επιχείρησης παίζει σημαντικό ρόλο η πρόβλεψη γεγονότων και επιχειρηματικών ευκαιριών. Η βαθύτερη κατανόηση της αγοράς επιτρέπει την αύξηση των πιθανοτήτων πρόβλεψης αυτών των συμβάντων ενώ η χρήση μεθόδων προγνωστικής ανάλυσης (predictive analytics) επεξεργάζεται δεδομένα του παρελθόντος και βοηθά στην διατύπωση των προβλέψεων για τον εντοπισμό επιχειρηματικών ευκαιριών. Τέλος, η αξιοποίηση των δεδομένων μέσω τεχνολογιών πληροφορικής επιτρέπει την ανάπτυξη της επιχείρησης όπως περιγράφεται παραπάνω και έτσι είναι σημαντικό η επιχείρηση να επενδύει αποδοτικά σε καινούργιες τεχνολογίες πληροφορικής πράγμα το οποίο καταφέρνει κάνοντας χρήση των ήδη υπαρχόντων τεχνολογιών.

Περιορισμοί της Επιχειρηματικής Ευφυΐας

Η Επιχειρηματική Ευφυΐα παρά τα πολλά οφέλη που προσφέρει περιορίζεται σημαντικά από διάφορα προβλήματα και κινδύνους. Αρχικά το κόστος απόκτησης και λειτουργίας των Αποθηκών Δεδομένων και των συστημάτων Επιχειρηματικής Ευφυΐας μπορεί να αποτελέσει πρόβλημα για τις επιχειρήσεις. Επίσης, τα συστήματα ανάλυσης δεδομένων δεν έχουν νόημα όταν τα δεδομένα που επεξεργάζονται είναι χαμηλής ποιότητας, δηλαδή όταν τα δεδομένα είναι διάσπαρτα, λανθασμένα, ανομοιογενή, λανθασμένα ή αντιφατικά, με αποτέλεσμα η παραγόμενη πληροφορία να είναι ελλιπής ή ακόμα και λανθασμένη. Ακόμα, τα συστήματα Επιχειρηματικής Ευφυΐας χρησιμοποιούν διαφορετικά συστήματα για διεργασίες των δεδομένων πράγμα που μπορεί να δημιουργήσει προβλήματα συμβατότητας μεταξύ των διαφορετικών συστημάτων.

Ένα ακόμα πρόβλημα δημιουργείται όταν τα στελέχη είναι δύσπιστα, επιφυλακτικά και μη συνεργάσιμα ως προς τα συστήματα της Επιχειρηματικής Ευφυΐας γεγονός που δημιουργεί προβλήματα επικοινωνίας και συνεννόησης μεταξύ των στελεχών και των ειδικών πληροφορικής, ειδικά αφού έχουν και διαφορετικές οπτικές γωνίες επί του θέματος. Έτσι δημιουργείται η ανάγκη για ειδικά εκπαιδευμένο προσωπικό αλλά κυρίως την εκπαίδευση των στελεχών ώστε να μάθουν να χρησιμοποιούν αυτά τα συστήματα με βέλτιστο τρόπο. Από την άλλη

υπάρχει ο κίνδυνος υπερβολικής εμπιστοσύνης στα συστήματα Επιχειρηματικής Ευφυΐας με αποτέλεσμα τα στελέχη να μην τα χρησιμοποιούν ως εργαλείο για την υποβοήθηση λήψης αποφάσεων αλλά ως εργαλεία που παίρνουν μόνα τους αποφάσεις. Αυτό μπορεί να αποδειχθεί επικίνδυνο αφού τα συστήματα Επιχειρηματικής Ευφυΐας αποτυγχάνουν σε μεγάλα ποσοστά επειδή έχουν να αντιμετωπίσουν πολλές προκλήσεις κατά την επεξεργασία δεδομένων ειδικά όταν τα δεδομένα είναι προβληματικά.

2.5 Η Επιχειρηματική Ευφυΐα στην Πράξη

Κάθε δραστηριότητα των επιχειρήσεων απαιτεί λήψη αποφάσεων και η Επιχειρηματική Ευφυΐα μπορεί να προσφέρει αντίστοιχες εφαρμογές. Συνεπώς, η Επιχειρηματική Ευφυΐα έχει αμέτρητα πεδία εφαρμογής. Παρακάτω παρουσιάζονται χοντρικά τα συνηθέστερα πεδία εφαρμογής.

2.5.1 Διοίκηση Επιχειρησιακής Απόδοσης

“Η Διοίκηση Επιχειρησιακής Απόδοσης (ΔΕΑ) (Corporate Performance Management (CPM)) είναι ένα σύνολο μεθοδολογιών, μετρικών, διαδικασιών και συστημάτων”, τα οποία βοηθούν τα διευθυντικά στελέχη στον έλεγχο και την διαχείριση της απόδοσης του οργανισμού. Η ΔΕΑ στην σύγχρονη εποχή υλοποιείται με τη χρήση κατάλληλου λογισμικού αντιστοιχώντας τη στρατηγική πληροφορία στα επιχειρησιακά σχέδια και έτσι παράγονται συγκεντρωτικά αποτελέσματα. Σημαντικό στοιχείο είναι οι Κύριοι Δείκτες Επιδόσεων (ΚΔΕ) (Key Performance Indicators (KPI)) οι οποίοι αποτυπώνουν την επίδοση της επιχείρησης σε σχέση με κάποια δραστηριότητα του που συνήθως αφορά εκπλήρωση στρατηγικών στόχων ή παράγοντες ζωτικής σημασίας της επιχείρησης. Ο καθορισμός των κατάλληλων ΚΔΕ είναι διαφορετική εργασία για διαφορετικές επιχειρήσεις αφού χρησιμοποιείται για τον έλεγχο και την μέτρηση του βαθμού επίτευξης επιχειρησιακών στόχων και κάθε επιχείρηση μπορεί να έχει διαφορετικές δραστηριότητες και λειτουργίες στις οποίες θέλει να εφαρμόσει τους ΚΔΕ.

Στον καθορισμό κατάλληλων ΚΔΕ μπορούν να βοηθήσουν εταιρείες συμβούλων και πάροχοι λογισμικού Επιχειρηματικής Ευφυΐας σε συνεργασία με τα διευθυντικά στελέχη που ορίζουν τους στόχους και στη συνέχεια συγκρίνουν την τρέχουσα κατάσταση των ΚΔΕ με τους στόχους. Με αυτό τον τρόπο αν διαπιστωθεί ότι η τρέχουσα κατάσταση δεν είναι ικανοποιητική σε σχέση με τους στόχους τότε θα αναζητηθούν τα αίτια ώστε να εκπονηθούν εργασίες αποκατάστασης του προβλήματος ή ακόμα μπορεί να γίνει και αναθεώρηση των στόχων.

Έτσι γίνεται ο έλεγχος και η ρύθμιση των επιδόσεων του οργανισμού. Για την διεκπεραίωση των παραπάνω διαδικασιών χρησιμοποιούνται συστήματα Επιχειρηματικής Ευφυΐας τα οποία συγκεντρώνουν και προεπεξεργάζονται όλα τα δεδομένα που σχετίζονται με τους ΚΔΕ ώστε στη συνέχεια να μπορούν να προβούν στον υπολογισμό των τιμών που αξιολογούν την κατάσταση. Αυτή η διαδικασία πρέπει να εκτελείται με ταχύτητα, αποτελεσματικότητα και η παραγόμενη πληροφορία πρέπει να παρουσιάζεται με τρόπο κατανοητό ενώ ταυτόχρονα πρέπει να είναι ορθή και να αντιπροσωπεύει την πραγματική κατάσταση του υπό διερεύνηση ζητήματος.

2.5.2 Χρηματοοικονομική ανάλυση και διαχείριση

Στόχος είναι ο σχεδιασμός και η παρακολούθηση των χρηματοοικονομικών ροών και πιο συγκεκριμένα η παρακολούθηση των εσόδων, των εξόδων και της κατάστασης των αποθεμάτων το οποίο επιτρέπει την καλύτερη διαχείριση του κεφαλαίου κίνησης και τον έλεγχο των κινδύνων που αφορούν τις απαιτήσεις. Η εύκολη σύνταξη χρηματοοικονομικών καταστάσεων επιτρέπει στα στελέχη να εκτιμούν την επίδοση της επιχείρησης κάνοντας σύγκριση τα μεγέθη του προϋπολογισμού ώστε σε περίπτωση απόκλισης να ληφθούν τα κατάλληλα μέτρα. Τα συστήματα Επιχειρηματικής Ευφυΐας παρακολουθούν τα πάγια σε όλο τον κύκλο της ζωής τους για την πραγματοποίηση της ανάλυσης των χρηματοοικονομικών μεγεθών ενώ παράλληλα ελέγχουν την κερδοφορία ως σύνολο αλλά και ειδικότερα ανά χρονική περίοδο, περιοχή, πελάτες, κατηγορία προϊόντων κλπ. ώστε να εντοπίζονται με αυτόν τον τρόπο τάσεις και ευκαιρίες. Επιπλέον, η τρέχουσα κατάσταση συγκρίνεται με αυτή προηγούμενων χρόνων γεγονός που δίνει πληροφορία για την εκπλήρωση των στόχων ώστε να παρέχεται πληρέστερη εικόνα για την πορεία της επιχείρησης και των χρηματοοικονομικών επιδόσεων.

2.5.3 Marketing & Πωλήσεις

Τα Συστήματα Επιχειρηματικής Ευφυΐας δίνουν την δυνατότητα της εύκολης παρακολούθησης και ελέγχου των πωλήσεων με αποτέλεσμα την αύξηση της αποτελεσματικότητας και του ανταγωνισμού στην αγορά. Γίνεται ανάλυση των στοιχείων του αγωγού πωλήσεων από το στάδιο των αρχικών επαφών με τους εν' δυνάμει πελάτες μέχρι την τελική πώληση και στη συνέχεια γίνεται σύγκριση με τις τιμές στόχους ώστε να εκτιμηθεί η πορεία των πωλήσεων και να ληφθούν κατάλληλα μέτρα σε περίπτωση αστοχίας ενώ η ανάλυση στοιχείων του παρελθόντος μπορεί να επιτρέψει την πιο εύστοχη πρόβλεψη του ύψους των μελλοντικών

πωλήσεων. Άλλος ένας σημαντικός τομέας είναι η ανάλυση των στοιχείων της διαχείρισης δυναμικού του τμήματος πωλήσεων η οποία μπορεί να γίνει σε διάφορα επίπεδα που φτάνουν ακόμα και σε ατομικές επιδόσεις των πωλητών. Η διοίκηση χρησιμοποιεί τα αποτελέσματα αυτής της ανάλυσης για να εντοπίσει τα ισχυρά σημεία και τυχόν αδυναμίες ώστε να μπορεί να προβεί έγκαιρα σε κατάλληλες δράσεις για να εντοπιστούν τα προβλήματα που μπορεί να προκύψουν.

Ένας ακόμα σημαντικός τομέας της Επιχειρηματικής Ευφυΐας είναι η άντληση και ανάλυση πληροφορίας που οδηγεί ύστερα από μελέτη στην κατανόηση της καταναλωτικής συμπεριφοράς και πιο συγκεκριμένα κατανοώντας τις ανάγκες τους και τις προτιμήσεις του καταναλωτή δίνεται η δυνατότητα στην επιχείρηση να αξιοποιήσει νέες ευκαιρίες και να προωθήσει τις πωλήσεις. Πιο συγκεκριμένα, μπορεί να βοηθήσει στην δημιουργία διαφημιστικών εκστρατειών που στοχοποιούν συγκεκριμένο καταναλωτικό πληθυσμό αφού η παραπάνω ανάλυση επιτρέπει τον εντοπισμό ομάδων καταναλωτών που έχουν παρόμοια χαρακτηριστικά και καταναλωτική συμπεριφορά. Επιπλέον, επιτρέπεται η αξιολόγηση των διαφημιστικών εκστρατειών ως προς την απόδοση και πιο συγκεκριμένα ύστερα από τον υπολογισμό του κόστους και των οφελών γίνεται σύγκριση των πραγματικών αποτελεσμάτων με αυτά του προϋπολογισμού το οποίο επιτρέπει την βελτιστοποίηση των διαφημιστικών μεθόδων.

2.5.4 Διαχείριση Ανθρωπίνων Πόρων & Διαχείριση Εφοδιαστικής Αλυσίδας

Σκοπός είναι η βελτιστοποίηση της διαχείρισης της Εφοδιαστικής Αλυσίδας που περιέχει δραστηριότητες ελέγχου αποθεμάτων και πρώτης ύλης για τη δημιουργία προϊόντων. Έτσι επιτρέπεται η έγκαιρη αντιμετώπιση προβλημάτων όπως είναι οι ελλείψεις σε αποθέματα και καθυστερήσεις σε παραγγελίες με αποτέλεσμα να μην επηρεάζεται αρνητικά η παραγωγή και να αυξάνεται η ικανοποίηση του πελάτη με την έγκαιρη παράδοση και την αποφυγή άλλων πιθανών προβλημάτων σχετικά με την παραγγελία. Επιπλέον η Επιχειρηματική Ευφυΐα μπορεί να χρησιμοποιηθεί αποτελεσματικά για την επιλογή προμηθευτών. Αυτό επιτυγχάνεται με την ανάλυση στοιχείων του παρελθόντος που αφορούν την σχέση ποιότητα και τιμή των αγαθών και των υπηρεσιών που προσφέρουν, τους χρόνους παράδοσης και την συνέπεια ενώ μπορούν να ληφθούν υπόψη και η χρηματοοικονομική τους κατάσταση κλπ.

Ένας χαρακτηριστικός τομέας που χρησιμοποιούνται τα συστήματα Επιχειρηματικής Ευφυΐας είναι η στελέχωση ενός οργανισμού με ανθρώπινο δυναμικό αλλά και ο έλεγχος της παραγωγικότητας του το οποίο μπορεί να οδηγήσει σε χορήγηση αμοιβών. Διευκολύνεται ο

χειρισμός θεμάτων μισθοδοσίας ενώ παράλληλα ελέγχεται η παραγωγικότητα με τον υπολογισμό του παραγωγικού και του μη παραγωγικού χρόνου, τον έλεγχο προσέλευσης και αποχώρησης, την εύρεση των παραγωγικών εργαζομένων καθώς και την σύνταξη πολιτικών για την διατήρηση τους στην επιχείρηση αλλά και για περαιτέρω εξέλιξη τους. Επίσης, διευκολύνεται η συγκρότηση του πλάνου για την κάλυψη θέσεων εργασίας με διάφορους τρόπους, όπως πρόσληψη μόνιμου ή εποχιακού εργαζομένου, υπερωρίες, εσωτερική κινητικότητα, ώρες απασχόλησης κλπ. Οι πολιτικές διαχείρισης ανθρωπίνων πόρων μπορούν δεχθούν τιμοδότηση ώστε να επιτευχθεί η σύγκριση οικονομικών και λειτουργικών επιπτώσεων γεγονός που βοηθάει στην πρόβλεψη των αναγκών σε εργατικό δυναμικό χρησιμοποιώντας μεθόδους ανάλυσης απολύσεων, αποχωρήσεων, επαναπροσλήψεων κλπ.

2.5.5 Χρηματοπιστωτικός τομέας

Ο χρηματοπιστωτικός τομέας πρόσφατα δέχτηκε πλήγμα από την οικονομική κρίση, γεγονός που οδήγησε στην ανάγκη για τακτική επιτήρηση και στενότερο έλεγχο των επηρεαζόμενων φορέων. Αυτή η ανάγκη ικανοποιήθηκε από την θέσπιση κανονιστικών διατάξεων που διέπει την λειτουργία τους και επιβάλλουν αυστηρούς όρους όπως η η δημοσίευση αναφορών σχετικά με την λειτουργία τους. Τα μέτρα αυτά στοχεύουν τόσο στην καλύτερη διαχείριση του “επιχειρησιακού κινδύνου (operational risk management)” όσο και στην αντιμετώπιση του οικονομικού εγκλήματος ενώ η μη τήρηση αυτών των διατάξεων και κατά επέκταση η μειωμένη διαχείριση του κινδύνου μπορεί να οδηγήσει σε πρόστιμο και αποζημιώσεις ύψους δισεκατομμυρίων ευρώ.

Για την εφαρμογή των παραπάνω διατάξεων χρειάζεται να συγκεντρωθούν επιπλέον δεδομένα και αφού υποστούν σε κατάλληλη επεξεργασία θα πρέπει να αναλυθούν και το προϊόν της ανάλυσης να είναι χρήσιμο. Τα συστήματα Επιχειρηματικής Ευφυΐας είναι ικανά να δώσουν λύση στα παραπάνω προβλήματα, όπως η διαχείριση κινδύνου, η αντιμετώπιση του οικονομικού εγκλήματος κτλ. Επιπροσθέτως, η οργάνωση των δεδομένων και η συγκεντρωτική διαχείριση τους επιτρέπει την ευκολότερη σύνταξη των αναφορών που απαιτούνται από τις κανονιστικές διατάξεις της νομοθεσίας.

2.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας

Οι ανάγκες της επιχειρηματικής κοινότητας για πολύπλοκα συστήματα Επιχειρηματικής Ευφυΐας οδήγησαν στην δημιουργία μιας ευρείας αγοράς αξίας δισεκατομμυρίων ευρώ η

οποία απαρτίζεται από εταιρίες διαφορετικών μεγεθών ενώ μεγάλες εταιρίες της πληροφορικής έχουν κυρίαρχο ρόλο.

Η SAS (Statistical Analysis System) από την ίδρυση της ασχολήθηκε με λογισμικό στατιστικής ανάλυσης και πλέον ανήκει στην ομάδα των κυρίαρχων παρόχων συστημάτων Επιχειρηματικής Ευφυΐας. Τα λογισμικά που προσφέρει αντιμετωπίζουν συχνά εμφανιζόμενα προβλήματα στον χώρο της Επιχειρηματικής Ευφυΐας όπως είναι η διαχείριση των δεδομένων (Data Management), η ανάλυση δεδομένων μεγάλου όγκου (Big Data) και η λειτουργία σε περιβάλλον υπολογιστικού νέφους (SAS Cloud Analytics).

Η IBM έπαιξε πρωταγωνιστικό παράγοντα στην ιστορία της πληροφορικής, τόσο στον τομέα του υλικού όσου και του λογισμικού, ενώ έχει μακροχρόνια εμπειρία στον τομέα της τεχνητής νοημοσύνης. Επιπλέον, έχει δημιουργήσει καινοτόμα προϊόντα μεταξύ των οποίων το IBM Personal Computer που αποτέλεσε τα θεμέλια των προσωπικών υπολογιστών (Pcs).

Η Oracle δραστηριοποιείται στον χώρο του υλικού υπολογιστών, ιδιαίτερα μετά την αγορά της Sun Microsystems, αλλά και στο χώρο λογισμικού επιχειρησιακών συστημάτων ενώ αποτελεί ηγέτης στον χώρο των βάσεων δεδομένων. Επιπλέον, λογίζεται ως κορυφαίος πάροχος συστημάτων Επιχειρηματική Ευφυΐας ενώ κατέχει το μεγαλύτερο μερίδιο της αντίστοιχης αγοράς. Προσφέρει ένα ευρύ φάσμα συστημάτων που δίνουν λύση σε επιχειρηματικά προβλήματα όπως η χρηματοοικονομική διοίκηση, οι πωλήσεις και το μάρκετινγκ, η διαχείριση εφοδιαστικής αλυσίδας, η διαχείριση ανθρωπίνων πόρων, ο χρηματοπιστωτικός τομέας, η διαχείριση ρίσκου και η κανονιστική συμμόρφωση, η διαχείριση κοινωνικών σχέσεων κλπ.

Η SAP είναι μια ευρωπαϊκή εταιρεία που θεωρείται ένας από τους μεγαλύτερους παρόχους λογισμικού παγκοσμίως ενώ ειδικεύεται στα συστήματα Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning). Επίσης, δραστηριοποιείται στον τομέα της Επιχειρηματικής Ευφυΐας, ειδικά από το 2007 που προχώρησε σε αγορά της γαλλικής εταιρίας Business Objects και μέχρι σήμερα έχει κατακτήσει μια θέση στην κορυφή του χώρου.

Η Microsoft δραστηριοποιείται σε πολλούς τομείς, τόσο λογισμικού όσο και hardware. Αξίζει να αναφερθεί η παιχνιδομηχανή Xbox και τα tablets Microsoft Surface. Από πλευράς λογισμικού, τα Windows και το MS Office είναι χαρακτηριστικά παραδείγματα προϊόντων της εταιρίας ενώ προσφέρει πολλά ακόμα προϊόντα λογισμικού. Μέσα σε αυτά υπάρχουν και επιχειρησιακά προϊόντα όπως συστήματα ERP και εφαρμογές Επιχειρηματικής Ευφυΐας. Τα κύρια προϊόντα που την καθιέρωσαν ως βασικό πάροχο λογισμικού Επιχειρηματικής Ευφυΐας είναι η βάση δεδομένων SQL Server και το Microsoft Office και πιο συγκεκριμένα τα φύλλα

εργασίας Excel και λογισμικό δημιουργίας παρουσιάσεων Power Point. Ο συνδυασμός των παραπάνω κατέστησε την Microsoft ως τον μεγαλύτερο κατασκευαστή λογισμικού παγκοσμίως όσο αναφορά τα έσοδα.

Η Qlik είναι μια συνηδική εταιρεία παραγωγής λογισμικού με εξειδίκευση στα συστήματα Επιχειρηματικής Ευφυΐας. Ιδρύθηκε το 1993 και από τότε έχει αναπτυχθεί σε μια διεθνή εταιρεία με πελατολόγιο που απαρτίζεται από 100 και περισσότερες διαφορετικές χώρες. Τα κυριότερα προγράμματα της εταιρείας είναι το QlikView και το QlikSense. To QlikView είναι μια πλατφόρμα για την ανάπτυξη εφαρμογών Επιχειρηματικής Ευφυΐας ενώ το QlikSense είναι μια εφαρμογή οπτικοποίησης δεδομένων και δημιουργίας αναφορών. Και τα δύο προγράμματα προσφέρονται σε διαφορετικές εκδόσεις μερικές των οποίων είναι μέσω plug in σε browser, κινητά τηλέφωνα κ.α..

Οι παραπάνω εταιρίες προσφέρουν διάφορα λογισμικά Επιχειρηματικής Ευφυΐας. Αυτά που κάνουν την εμφάνιση τους σχεδόν σε κάθε μία από αυτές τις εταιρίες είναι λογισμικό της αναλυτικής των επιχειρήσεων (Business Intelligence and Analytics), λογισμικό για τη διαχείριση των πελατών και του μάρκετινγκ (Customer Intelligence), λογισμικό ασφάλειας και αντιμετώπισης απάτης (Fraud and Security Intelligence), λογισμικό διαχείρισης επιδόσεων (Performance Management), λογισμικό για τη διαχείριση του ρίσκου (Risk Management) και λογισμικό για τη διαχείριση της εφοδιαστικής αλυσίδας (Supply Chain Intelligence).

2.7 Ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας

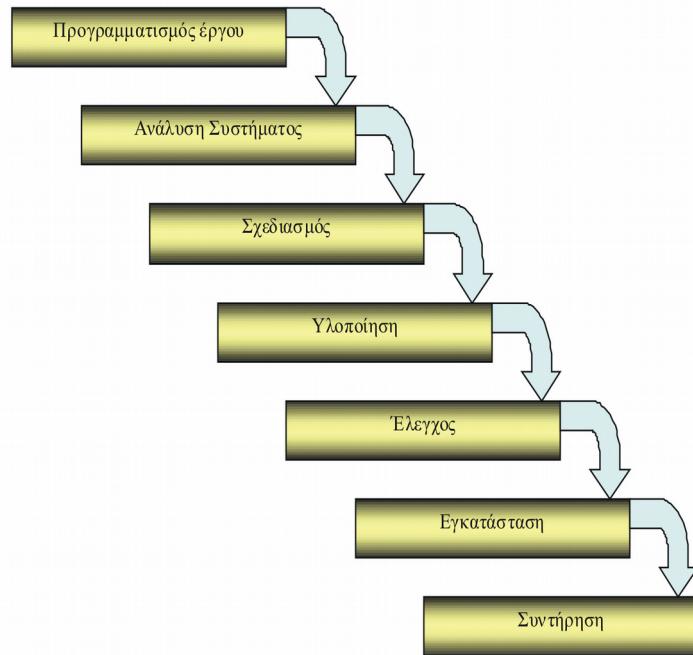
Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας είναι μια περίπλοκη και μακροχρόνια διαδικασία που συνδέεται άμεσα με την βαθιά κατανόηση των διοικητικών διαδικασιών της επιχείρησης και ιδιαίτερα αυτών που αφορούν τη λήψη αποφάσεων. Αξίζει να σημειωθεί ότι η ανάπτυξη ενός συστήματος Επιχειρηματικής Ευφυΐας είναι έργο που βρίσκεται σε διαρκή εξέλιξη αφού σε έναν οργανισμό οι ανάγκες για πληροφόρηση δεν είναι σταθερές αφού οι απαιτήσεις των πελατών αλλάζουν συχνά με την πάροδο του χρόνου και άρα το σύστημα οφείλει να προσαρμόζεται στις νέες αυτές ανάγκες. Σημαντικό ρόλο στην ανάπτυξη των συστημάτων Επιχειρηματικής Ευφυΐας παίζουν και οι Αποθήκες Δεδομένων. Η Αποθήκη Δεδομένων είναι μια βάση δεδομένων που περιέχει γιγάντιους όγκους δεδομένων από διαφορετικά συστήματα ενώ η δομή τους είναι περίτεχνη έτσι ώστε να επιτρέπεται η ταχεία πρόσβαση στα δεδομένα από διαφορετικά τμήματα μιας επιχείρησης και από στελέχη διαφορετικής ταξιαρχίας.

Τα τελευταία χρόνια η Επιχειρηματική Ευφυΐα έχει γνωρίσει μεγάλη ανάπτυξη αφού είναι αναγκαία η ενσωμάτωση της στις επιχειρήσεις, ειδικά στις μεγάλες που βρίσκεται στην κορυφή των τεχνολογικών προτεραιοτήτων τους. Υπάρχουν πολλοί λόγοι για την αναγκαιότητα τους. Η αναλογία μεταξύ ποσότητας δεδομένων και πληροφορίας είναι άνιση αφού υπάρχουν πολλά δεδομένα αλλά μικρή ποσότητα πληροφορίας που συνδέεται και με τη δημιουργία ανάγκης για ταχύτερη λήψη αποφάσεων βασισμένη σε πληροφορία. Επιπλέον, η ανάγκη εύρεσης πληροφορίας του παρελθόντος και η αδυναμία οργάνωσης των δεδομένων με τρόπο επιθυμητό οδηγούν στο πρόβλημα δαπάνης υπερβολικού χρόνου για τη συλλογή και την ανάλυση των δεδομένων. Επίσης, η αδυναμία σύνταξης αναφορών από το τμήμα πληροφορικής λόγω περιορισμένου χρόνου και η ανάγκη για βελτίωση των επιχειρηματικών διαδικασιών ώστε να αποδίδουν καλύτερα κέρδη είναι προβλήματα ζωτικής σημασίας μιας επιχείρησης που περιορίζουν την ανάπτυξη της.

2.8 Ο Κύκλος Ζωής Ανάπτυξης Συστήματος Επιχειρηματικής Ευφυΐας

Ο Κύκλος Ζωής Ανάπτυξης Συστήματος (ΚΖΑΣ) (System Development Life Cycle (SDLC)) καθορίζει τα στάδια κατασκευής ενός πληροφοριακού συστήματος. Ένα από τα πρώτα και πιο διαδεδομένα μοντέλα ήταν αυτό της «υδατόπτωσης» το οποίο θεωρεί ότι η ανάπτυξη ενός πληροφοριακού συστήματος είναι μια διαδικασία διαδοχικών σταδίων που τα αποτελέσματα ενός σταδίου χρησιμοποιούνται στο επόμενο. Τα στάδια είναι γραμμικά ενώ υπάρχει ένα σημείο αφετηρία και ένα τερματικό σημείο. Υπάρχουν πολλές παραλλαγές αυτού του μοντέλου που διαφέρουν στο πλήθος και στο ποια είναι τα στάδια, όμως το κεντρικό σκεπτικό του μοντέλου παραμένει το ίδιο. Το Εικόνα 2.2 παρακάτω παρουσιάζει το μοντέλο της υδατόπτωσης.

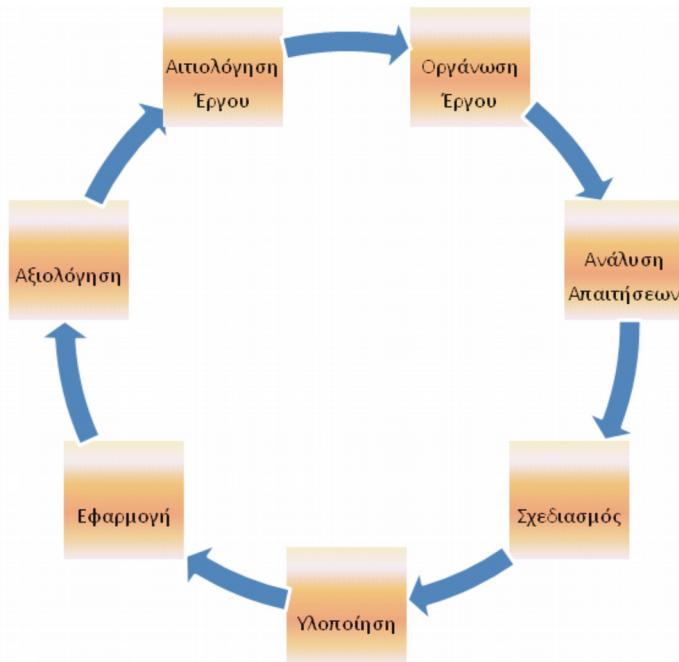
Γενικά το μοντέλο της υδατόπτωσης θεωρείται ανεπαρκές για την περιγραφή της πραγματικής διαδικασίας ανάπτυξης ενός συστήματος. Αυτό συμβαίνει για αρκετούς λόγους όπως για παράδειγμα η θεώρηση ότι οι χρήστες είχαν αποκλειστικό ρόλο τον καθορισμό των απαιτήσεων και ότι όλες οι απαιτήσεις μπορούν να διατυπωθούν εξ αρχής. Ακόμα στα πραγματικά συστήματα τα στάδια μπορούν να ξεκινούν πριν τελειώσει το προηγούμενο ενώ καθίσταται δυνατή η επιστροφή σε κάποιο προηγούμενο επίπεδο κάτι που έρχεται σε αντίθεση με το μοντέλο υδατόπτωσης. Αυτά τα προβλήματα οδήγησαν στην αναζήτηση εναλλακτικών μοντέλων.



Εικόνα 2.2 Μοντέλο Υδατόπτωσης

Το μοντέλο υδατόπτωσης είναι ακατάλληλο για τα Συστήματα Επιχειρηματικής Ευφυΐας. Αυτό συμβαίνει επειδή τα συστήματα Ε.Ε. αφορούν διοικητικές διαδικασίες λήψης αποφάσεων οι οποίες δεν είναι τυποποιημένες και αυστηρά καθορισμένες το οποίο έρχεται σε αντίθεση με το μοντέλο της υδατόπτωσης το οποίο θεωρεί ότι οι διαδικασίες είναι ουσιαστικά καθημερινής λειτουργίας δηλαδή είναι τυποποιημένες και σταθερές σε βάθος χρόνου. Δηλαδή το κλασικό μοντέλο υδατόπτωσης είναι κατάλληλο για την ανάπτυξη στατικών συστημάτων που επιλύουν απομονωμένα προβλήματα ενώ τα συστήματα Ε.Ε. είναι μεταβλητά και περιλαμβάνουν όλους τους τομείς της επιχείρησης.

Ωστόσο, το κλασικό μοντέλο υδατόπτωσης δεν καλύπτει το ζήτημα του στρατηγικού σχεδιασμού, ενώ δεν υποστηρίζει διαδοχικές εκδόσεις του συστήματος. Στα συστήματα Ε.Ε. τα δεδομένα και οι λειτουργικότητες τροποποιούνται σε διαδοχικές εκδόσεις ενώ κάθε έκδοση μπορεί να αναδείξει νέες απαιτήσεις χρηστών, οπότε η διαδικασία ανάπτυξης του συστήματος πρέπει να είναι κυκλική. Θεωρείται ότι το καταλληλοτερο μοντέλο για την ανάπτυξη συστημάτων Ε.Ε. έχει κυκλική δομή και αποτελείται από διαδοχικά βήματα. Ο κύκλος ζωής ενός συστήματος Ε.Ε. απεικονίζεται στην Εικόνα 2.3 και αποτελείται από την αιτιολόγηση του έργου, την οργάνωση του έργου, την ανάλυση απαιτήσεων του συστήματος, τον σχεδιασμό, την κατασκευή, την εφαρμογή και τέλος την αξιολόγηση.



Εικόνα 2.3 Μοντέλο ανάπτυξης Συστημάτων Ε.Ε.

2.8.1 Αιτιολόγηση Έργου

Το πρώτο στάδιο του κύκλου ζωής ενός συστήματος Ε.Ε. είναι η αιτιολόγηση υλοποίησης του έργου, όπου συλλέγονται επιχειρηματικές ευκαιρίες ή προβλήματα και γίνεται αναλυτική επεξήγηση για το πώς το έργο θα συμβάλει στην αντιμετώπιση τους. Γενικά, έργα που δεν έχουν συγκεκριμένο λόγω ύπαρξης και στοχεύουν στη γενική βελτίωση της πληροφόρησης στην επιχείρηση έχουν μεγάλες πιθανότητες να οδηγήσουν σε αποτυχία. Συνεπώς η ύπαρξη ενός συγκεκριμένου επιχειρηματικού στόχου είναι σημαντικός παράγοντας για την επιτυχία του συστήματος. Δηλαδή πρέπει να καθορίζεται εξ αρχής ποια επιχειρηματικά ζητήματα θα εξυπηρετηθούν και το πώς αυτά τα εξυπηρετηθούν από το έργο. Ακόμα, η ανάπτυξη ενός συστήματος Ε.Ε. από μη ειδικευμένα άτομα που αξιοποιούν την πληροφορική τεχνολογία μπορεί να οδηγήσει στη δημιουργία ενός περίτεχνου και ακριβού συστήματος που όμως δεν θα είναι ικανό να παράξει χρήσιμα αποτελέσματα για τον οργανισμό.

Σε αυτό το στάδιο παίρνει μέρος και η τεκμηρίωση της στρατηγικής του έργου εξυπηρετώντας τους στρατηγικούς στόχους του. Τα έργα Ε.Ε. θεωρούν την διοικητική πληροφόρηση ως ολότητα της επιχείρησης και έτσι ικανοποιούν ανάλογα τις ανάγκες της. Συνεπώς η αιτιολόγηση της αναγκαιότητας εκτέλεσης του έργου θα πρέπει να ορίζει με σαφή τρόπο το πώς γίνεται η αξιοποίηση της παραγόμενης πληροφορίας ώστε να επιλυθούν επιχειρηματικά ζητήματα και κατά επέκταση το πώς αυτά τα ζητήματα θα βοηθήσουν στην επίτευξη των στρα-

τηγικών στόχων του οργανισμού. Ένα σημαντικός παράγοντας της αιτιολόγησης του έργου είναι η ανάλυση του κόστους σε σχέση με το κέρδος.

Αρχικά, γίνεται εκτίμηση του αναγκαίου κόστους για την δημιουργία και τη συνεχή λειτουργία του συστήματος και στη συνέχεια γίνεται εκτίμηση της αξίας που θα παραχθεί από τη χρήση του χρησιμοποιώντας στοιχεία που από τη λειτουργία του συστήματος. Ο υπολογισμός του κέρδους δεν είναι πάντα εύκολη διαδικασία όμως σε περιπτώσεις που το πρόβλημα προκαλεί συγκεκριμένες ζημίες τότε έχει μετρήσιμα αποτελέσματα και άρα η διαδικασία ποσοτικοποίησης του κέρδους είναι πιο εύκολη.

Γενικά ένα σύστημα ΕΕ μπορεί να προσφέρει κέρδος αυξάνοντας τα έσοδα, μειώνοντας το κόστος, αυξάνοντας το μερίδιο της αγοράς και αυξάνοντας την ικανοποίηση του πελάτη. Τέλος στο στάδιο αυτό λαμβάνει χώρα η εκτίμηση του κινδύνου διαφόρων τομέων του έργου το οποίο μπορεί να οδηγήσει στην βελτίωση της διαχείρισης των κινδύνων αλλά και σε μια εκτίμηση των αποτελεσμάτων του έργου η οποία δεν θα αποκλίνει πολύ από την πραγματική. Ενδεικτικά κάποιοι από αυτούς τους τομείς μπορεί να είναι η τεχνολογία που θα χρησιμοποιηθεί, προβλήματα δεδομένων όπως είναι η ενοποίηση και ολοκλήρωση τους, η πολυπλοκότητα του έργου, ο χρηματοοικονομικός τομέας και τέλος προβλήματα με τη σύνθεση ομάδας του έργου. Είναι σημαντικό να τονιστεί ότι όλες οι παραπάνω διαδικασίας πρέπει να εφαρμόζονται σε κάθε επανάληψη του κύκλου ζωής του συστήματος.

2.8.2 Οργάνωση Έργου

Αντικείμενο αυτού του επιπέδου είναι η σχεδίαση και οργάνωση ενός σχεδίου που θα ακολουθηθεί για την εκτέλεση του έργου. Γενικά το σχέδιο θα πρέπει να περιγράφει τι θα παραδοθεί με την ολοκλήρωση του έργου, πότε θα ολοκληρωθεί, πόσο θα κοστίσει και ποιοι θα συμμετέχουν και ποιος θα είναι ο ρόλος τους. Αρχικά για την υλοποίηση του σχεδίου γίνεται καταγραφή και αποτίμηση των υπαρχουσών υποδομών που χωρίζονται σε τεχνικά και μη τεχνικά θέματα. Στα τεχνικά θέματα περιλαμβάνεται το υπάρχων υλικό υπολογιστών και το λογισμικό. Έτσι, γίνεται η καταγραφή και αξιολόγηση του διαθέσιμου υλικού υπολογιστών και παράλληλα γίνεται έλεγχος του λογισμικού για να διαπιστωθεί αν μπορεί να ανταποκριθεί στις απαιτήσεις του νέου συστήματος.

Αρχικά γίνεται καταγραφή και έπειτα η αξιολόγηση του υλικού των υπολογιστών που έχει η επιχείρηση ώστε να διαπιστωθεί αν μπορεί να ανταποκριθεί στις ανάγκες του νέου συστήματος. Επιπλέον, πρέπει να ληφθεί μέριμνα για την ικανοποίηση των μελλοντικών απαιτήσε-

ων επειδή η συνεχής εξέλιξη των συστημάτων ΕΕ τείνει να αυξάνει τον óγκο των δεδομένων και κατά επέκταση αυξάνονται οι ανάγκες για υπολογιστική ισχύ. Αν το υπάρχων υλικό δεν ανταποκρίνεται στις ανάγκες του νέου συστήματος τότε πρέπει να συγκεντρωθεί το κόστος αγοράς νέου συμβατού υλικού ώστε να προχωρήσει ο οργανισμός στην αγορά και ενσωμάτωση του νέου υλικού. Επίσης, γίνεται αξιολόγηση της δικτυακής υποδομής καθώς και του εύρους ζώνης ώστε να καλυφθούν πιθανές ανάγκες αναβάθμισης για την επιτάχυνση των διαδικασιών.

Οσο αναφορά το λογισμικό, γίνεται καταγραφή των συστημάτων διαχείρισης βάσεων δεδομένων που χρησιμοποιούνται στην επιχείρηση και αν εκτελούνται εργασίες ολοκλήρωσης και ανάλυσης δεδομένων καθώς και σύνταξης αναφορών τότε γίνεται καταγραφή του χρησιμοποιούμενου λογισμικού. Ακόμα, γίνεται έλεγχος της πληροφοριακής υποδομής που παράγει τα πηγαία δεδομένα, π.χ. ERP. Γίνεται έλεγχος πληρότητας της υποδομής του λογισμικού για να διαπιστωθεί αν υπάρχουν ανάγκες προσθήκης νέου λογισμικού και αν υπάρχουν τότε είτε γίνεται αγορά είτε συγγραφή νέου λογισμικού. Ένα ακόμα ζήτημα που προκύπτει είναι όταν τα συστήματα Ε.Ε. πρέπει να είναι προσβάσιμα με πολλούς εναλλακτικούς τρόπους, για παράδειγμα έκδοση κινητών τηλεφώνων, το οποίο απαιτεί την δημιουργία ενός portal.

Εκτός από την υλικοτεχνική υποδομή, υπάρχει μια άλλη υποδομή που έχει να κάνει με οργανωτικές δομές, διαδικασίες, τρόπους λειτουργίας, κανόνες, ρόλους κλπ. Αυτή η μη τεχνική υποδομή είναι πολύ σημαντική καθώς επηρεάζει την ανάπτυξη του συστήματος. Επίσης, είναι αναγκαία η ύπαρξη πολιτικής διαχείρισης αλλαγών (change management), που είναι πολύ πιθανό να προκύψουν κατά την εφαρμογή του συστήματος και μπορεί να προκαλέσουν αστάθεια και κατά επέκταση προβλήματα. Η τυποποίηση της δομής των πηγαίων δεδομένων και η αντιμετώπιση πιθανών προβλημάτων που μπορεί να έχουν, π.χ. ονοματοδοσία δεδομένων, είναι σημαντικές εργασίας που πρέπει να γίνει η καταγραφή τους. Πρέπει να γίνει αντιστοίχιση των δεδομένων στο αντίστοιχο λογικό μοντέλο αν υπάρχει, ενώ αν δεν υπάρχει πρέπει να οριστούν υπεύθυνοι που θα υλοποιήσουν ένα λογικό μοντέλο το οποίο θα επικυρωθεί από μια άλλη ομάδα, η οποία πρέπει να οριστεί και αυτή.

Ακόμα, αν γίνεται χρήση μεταδεδομένων θα πρέπει να γίνει καταγραφή του λογισμικού που χρησιμοποιείται καθώς και το περιεχόμενο τους και το πως γίνεται η ενημέρωση τους. Με την ολοκλήρωση της καταγραφής των υποδομών οργανώνεται το αναλυτικό σχέδιο της υλοποίησης του έργου, το οποίο είναι ζωτικής σημασίας αφού προσδιορίζει τη διαδικασία

εκτέλεσης του έργου ενώ καθορίζει και την εκτέλεση του απολογισμού του έργου μετά την ολοκλήρωση του αφού γίνεται σε συνδυασμό με το σχέδιο δράσης. Έτσι, πρώτα γίνεται καταγραφή των στόχων του έργου σε μορφή έκθεσης στην οποία προσδιορίζονται και οι λόγοι εκπόνησης του έργου και στη συνέχεια γίνεται η τεκμηρίωση της σε συνάρτηση με τη στρατηγική του οργανισμού.

Επίσης, προσδιορίζεται με ακρίβεια το πλαίσιο ενασχόλησης του έργου καθώς και η το μέγεθος του. Θεωρητικά, οι ανάγκες για πληροφόρηση στα πλαίσια μιας επιχείρησης είναι αμέτρητες και περιλαμβάνουν δεδομένα, λειτουργικότητες και δυνατότητες πληροφόρησης. Έτσι πρέπει σε κάποιο βαθμό να γίνει ρητή επιλογή των αναγκών που θα καλυφθούν και αυτών που δεν θα καλυφθούν ενώ πρέπει να δοθεί προσοχή στην επιλογή δεδομένων αφού τα συστήματα ΕΕ επηρεάζονται περισσότερο από τα δεδομένα σε σχέση με τις λειτουργικότητες.

Έπειτα, εκτελείται ο προγραμματισμός εκτέλεσης του έργου όπου εφαρμόζονται εργασίες όπως ο χωρισμός του έργου σε μικρότερα υποέργα τα οποία είναι καλά ορισμένα τμήματα του συνολικού, γίνεται εκτίμηση του απαραίτητου χρόνου και των απαραίτητων πόρων που απαιτούνται για τη δημιουργία του κάθε υποέργου και τέλος γίνεται εύρεση των σχέσεων αλληλεξάρτησης μεταξύ των υποέργων. Όλα τα υποέργα κοστολογούνται σε απαραίτητα κεφάλαια για την υλοποίηση τους καθώς και σε διαθεσιμότητα ανθρώπινου δυναμικού που θα τα υλοποιήσει που περιορίζεται από παράγοντες όπως ο χρόνος και η καταλληλότητα των ατόμων, το οποίο καθιστά αναγκαία την κατάλληλη κατανομή καθηκόντων και αρμοδιοτήτων.

Ακόμα, οι σχέσεις αλληλεξάρτησης μεταξύ των υποέργων μπορεί να περιορίσουν την εκτέλεση μερικών υποέργων καθώς μπορεί να είναι απαραίτητο για την υλοποίηση ενός υποέργου να πρέπει πρώτα να ολοκληρωθεί κάποιο άλλο ενώ άλλα έργα μπορούν να υλοποιηθούν παράλληλα. Συνεπώς κρίνεται απαραίτητη η καταγραφή αυτών των σχέσεων αφού έτσι βελτιώνεται η εποπτεία της πορείας του έργου καθώς βελτιώνεται και ο απολογισμός σε περίπτωση καθυστέρησης. Στο τέλος δημιουργείται το χρονοδιάγραμμα εκτέλεσης του έργου το οποίο περιέχει τον χρόνο εκτέλεσης των υποέργων, στον οποίο συμπεριλαμβάνονται τυχόν καθυστερήσεις αλλά και οι αλληλεξαρτήσεις των υποέργων. Επίσης, γίνεται σύνταξη του τελικού προϋπολογισμού που προϋποθέτει εξασφαλισμένη χρηματοδότηση, το οποίο συνεπάγεται την κατανόηση της αναγκαιότητας του έργου από τη διοίκηση και κατά επέκταση την διάθεση του απαιτούμενου κεφαλαίου.

2.8.3 Ανάλυση απαιτήσεων του έργου

Η ανάλυση απαιτήσεων του έργου αφορά κατά κύριο λόγο επιχειρηματικά ζητήματα και δεν αναφέρεται στην ανάλυση του συστήματος με την κλασική έννοια καθώς οι απαιτήσεις των συστημάτων ΕΕ είναι κυρίως στρατηγικής πληροφορίας και λιγότερο λειτουργικών θεμάτων. Η ανάλυση απαιτήσεων του έργου είναι το πιο σημαντικό επίπεδο του κύκλου ανάπτυξης του συστήματος αφού σε αυτό αποφασίζεται ποιες εργασίες θα εκτελεί το σύστημα και με ποιον τρόπο. Επίσης, πολλές αποτυχίες συστημάτων προέρχονται από σφάλματα, ελλείψεις και αβλεψίες στην ανάλυση. ενώ ειδικά οι αβλεψίες μπορούν να απορροσανατολίσουν το σύστημα καθιστώντας τη διαδικασία της διόρθωσης πιο απαιτητική.

Αρχικά πρέπει να καθοριστούν τα επιχειρηματικά ζητήματα και οι ανάγκες στρατηγικής πληροφόρησης που θα καλυφθούν από το σύστημα ώστε να γίνει δυνατός ο προσδιορισμός των εργασιών που πρέπει να εκτελεστούν καθώς και να γίνει εφικτός ο προσδιορισμός των δεδομένων που θα χρησιμοποιήσουν αυτές οι εργασίες. Η εύρεση των απαιτήσεων γίνεται με τη χρήση ερωτηματολογίων καθώς και με συνεντεύξεις ενώ η πληροφορία που συγκεντρώνεται προέρχεται από τις ειδικές γνώσεις του κάθε ατόμου που συμμετέχει. Τα ανώτατα διοικητικά στελέχη γνωρίζουν την στρατηγική διάσταση του έργου και κατά επέκταση γνωρίζουν το πως αυτό θα χρησιμοποιηθεί για την επίτευξη των στρατηγικών στόχων, το οποίο είναι πολύ σημαντικό για την αποτελεσματικότητα του έργου και συμπερασματικά καθίσταται απαραίτητη η συμμετοχή τους στην διαδικασία.

Επίσης, κατέχουν την τεχνογνωσία ως προς το ποιες θα πρέπει να είναι οι παραγόμενες πληροφορίες από το σύστημα καθώς και ποιες ερωτήσεις χρήζουν απάντηση και τι αναφορές πρέπει να συνταχθούν. Τα στελέχη πληροφορικής συνεισφέρουν με τεχνικού είδους γνώσεις και πιο συγκεκριμένα κατανοούν ζητήματα τεχνικής υποδομής ενώ γνωρίζουν και κατανοούν την υπάρχουσα κατάσταση των δεδομένων. Οι αναλυτές δεδομένων της επιχείρησης καταλαβαίνουν τα προβλήματα των δεδομένων και κατανοούν τις ανάγκες ανάλυσης. Το τελικό προϊόν της έκθεσης περιλαμβάνει τις απαιτήσεις του συστήματος καθώς και τι θα περιλαμβάνει το σύστημα και τι όχι. Τέλος, η έκθεση θα πρέπει να κάνει χρήση επιχειρηματικών όρων και όχι λειτουργικότητας, αφού η παράθεση των λειτουργιών δεν είναι προτεραιότητα.

Γενικά, η έκθεση θα πρέπει να προσδιορίζει το πρόβλημα που αντιμετωπίζει ο οργανισμός, το κόστος του προβλήματος, τον τρόπο επίλυσης του προβλήματος από το σύστημα Ε.Ε., πως συνδέεται η επίλυση του προβλήματος με τις στρατηγικές απαιτήσεις της επιχείρησης, ποιοι είναι οι χρήστες του συστήματος και πώς θα γίνει η απλούστευση της εργασία

τους. Επίσης, στην έκθεση προσδιορίζεται αν η επιχείρηση αξιοποιεί την πληροφορία ως περιουσιακό στοιχείο, ποιες αναφορές θα συντάσσονται και ποια ερωτήματα χρήζουν απάντηση ώστε να προσδιοριστεί τι πληροφορίες θα παραχθούν και τέλος πως θα γίνεται η παρουσίαση και διανομή της πληροφορίας. Τα συστήματα Ε.Ε. χαρακτηρίζονται «*data intensive*» αφού το μεγαλύτερο ποσοστό των εκτελούμενων εργασιών σχετίζεται με δεδομένα και όχι με λειτουργίες. Συνεπώς, θα πρέπει να γίνεται εκτενή περιγραφή των δεδομένων στην έκθεση ανάλυσης απαιτήσεων και πιο συγκεκριμένα θα πρέπει να περιγράφεται το λογικό μοντέλο των δεδομένων, ποια δεδομένα θα χρησιμοποιούνται, οι διαστάσεις των κύβων, οι διαδικασίες καθορισμού και μετασχηματισμού των δεδομένων, ο βαθμός λεπτομέρειας των δεδομένων και τέλος πόσο χρονικό διάστημα θα χρησιμοποιούνται τα δεδομένα.

Σε έναν οργανισμό τα συστήματα Ε.Ε. χρησιμοποιούνται από πολλά τμήματα τα οποία συνήθως έχουν τα δικά τους δεδομένα και έτσι αυτά ενοποιούνται, ολοκληρώνονται και συγχωνεύονται ώστε να μπορούν να χρησιμοποιηθούν με όσο πιο δυνατό βέλτιστο τρόπο. Οι διαδικασίες αυτές εξετάζονται στην ανάλυση απαιτήσεων. Όσο αναφορά τις λειτουργίες που θα εκτελεί το σύστημα, καταγράφονται οι εκθέσεις που θα συντάσσονται, τα ερωτήματα που θα υποβάλλονται στο σύστημα, οι εργασίες ανάλυσης που θα εκτελούνται (πχ εργασίες OLAP, στατιστικής ανάλυσης ή εργασίες προγνωστικής ανάλυσης με τεχνικές εξόρυξης δεδομένων), καθώς και οι τρόποι παρουσίασης και διανομής της πληροφορίας.

Τέλος, η ολοκλήρωση της έκθεσης απαιτήσεων γίνεται με την καταγραφή των απαιτήσεων για επέκταση της υλικοτεχνικής υποδομής (υλικό υπολογιστών, βάσεις δεδομένων, εργαλεία εξόρυξης δεδομένων, υποδομή δικτύωσης κλπ.), με την καταγραφή της μη υλικοτεχνικής υποδομής (διαδικασίες, πρότυπα, ρόλοι, οδηγίες) και με την καταγραφή ζητημάτων ασφαλείας του συστήματος (ποιοι έχουν πρόσβαση σε αυτό, αν υπάρχουν διαβαθμίσεις της πρόσβασης ανάλογα με το ποιος είναι ο χρήστης κλπ.).

Αφού ολοκληρωθούν οι παραπάνω διαδικασίες πρέπει να χρησιμοποιηθεί μια τακτική δοκιμής και ελέγχου. Αυτό συνήθως επιτυγχάνεται με τη χρήση πρωτοτύπου του συστήματος (prototyping), που ουσιαστικά είναι μια περιορισμένη τελική έκδοση του συστήματος η οποία περιλαμβάνει μερικές λειτουργίες που συνήθως είναι πλήρης αλλά μπορεί μερικές φορές να περιλαμβάνει και άλλες που δεν είναι πλήρεις. Η αξιολόγηση του πρωτοτύπου γίνεται από χρήστες που το αξιολογούν ύστερα από μια περίοδο χρήσης και αν εντοπίσουν λάθη, αδυναμίες ή παραλήψεις ενημερώνουν τους αναλυτές. Έτσι, επιτυγχάνεται η ανατροφοδότηση των αναλυτών από τους χρήστες σχετικά με παρατηρήσεις για το σύστημα από αρχικά

ακόμα στάδια που οι τροποποιήσεις είναι ευκολότερες. Επίσης, η μέθοδος του πρωτοτύπου επιτρέπει πειραματισμούς με διαφορετικές τεχνολογίες, όπως για παράδειγμα συστήματα βάσεων δεδομένων. Ακόμα, το πρωτότυπο μπορεί να επανεκδοθεί με διορθώσεις σε προβλήματα που εντοπίστηκαν σε προηγούμενες εκδόσεις ενώ παράλληλα μπορούν να προστεθούν και καινούργιες λειτουργίες. Οι επανεκδόσεις μπορούν να συνεχιστούν μέχρι και το σύστημα να πάρει την τελική και ολοκληρωμένη μορφή του.

Πριν την κατασκευή του πρωτοτύπου πρέπει να οριστεί το περιεχόμενο του, σε ποιους απευθύνεται και η αναμενόμενη ανατροφοδότηση. Σκοπός της ανατροφοδότησης είναι η διόρθωση του συστήματος το οποίο μπορεί να σημαίνει αλλαγές στις λειτουργίες, τα δεδομένα ή τα μεταδεδομένα του. Επίσης, το πρωτότυπο συνήθως δεν χρησιμοποιεί όλο το σύνολο των διαθέσιμων δεδομένων γιατί εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης δεδομένων απαιτούν αρκετό χρόνο και έτσι αυξάνεται το κόστος. Σκοπός του πρωτοτύπου είναι να παρουσιάσει στους χρήστες ένα υποσύνολο των αναλυτικών δυνατοτήτων του συστήματος καθώς και τον τρόπο παρουσίασης της πληροφορίας χρησιμοποιώντας μερικές από τις συνολικές λειτουργίες. Υπάρχουν διαφορετικά είδη πρωτοτύπων τα οποία διαθέτουν διαφορετικές δυνατότητες με αποτέλεσμα να υπάρχουν διαφορές σε χρονικές απαιτήσεις και στο απαιτούμενο κόστος.

Στην πιο απλή μορφή το πρωτότυπο χρησιμοποιεί μόνο τη διεπαφή (interface) και σκοπός του είναι να παρουσιάσει στους χρήστες τον τρόπο που γίνεται η παρουσίαση της πληροφορίας καθώς και να τους βοηθήσει να αντιληφθούν τις λειτουργίες του συστήματος και των δεδομένων του. Αυτό το είδος πρωτοτύπων βοηθάει στην εύρεση των απαιτήσεων των χρηστών. Ένα ακόμα είδος πρωτοτύπων είναι τα λειτουργικά τα οποία υλοποιούν πραγματικές λειτουργίες του συστήματος και συνεπώς η εξέλιξη τους σε τελικό συστήματα γίνεται σχετικά εύκολα. Έτσι, οι χρήστες χρησιμοποιούν το σύστημα για να πραγματοποιήσουν αναλύσεις ώστε να σχηματίσουν άποψη και να παραθέσουν τα σχόλια τους. Άλλα είδη πρωτοτύπων είναι της επίδειξης, το οποίο χρησιμοποιείται για παρουσιάσει των δυνατοτήτων του συστήματος στους πελάτες, και το διερευνητικό, το οποίο χρησιμοποιείται για την ανεύρεση κινδύνων και κατά επέκταση να αποφασιστεί ή να απορριφθεί η υλοποίηση του συστήματος.

Επίσης, κατά τη δημιουργία ενός πρωτοτύπου πρέπει να γίνει ανεύρεση των ατόμων που θα συμμετέχουν στον σχεδιασμό, στη χρήση και στην επικύρωση του πρωτοτύπου, θα πρέπει να οριστούν τα δεδομένα που θα χρησιμοποιηθούν καθώς και τι υλικό και λογισμικό υπολογιστών θα χρειαστεί. Ένα ακόμα σημαντικό ζήτημα είναι η χρήση μεταδεδομένων. Τα μετα-

δεδομένα είναι δεδομένα που χρησιμοποιούνται για την περιγραφή και την απόδοση νοήματος στα κανονικά δεδομένα και αφορούν τεχνικά και μη τεχνικά επιχειρησιακά θέματα. Τα επιχειρησιακά μεταδεδομένα βοηθούν τα διοικητικά στελέχη να κατανοήσουν το σύστημα, να πλοηγηθούν σε αυτό καθώς και να το χρησιμοποιήσουν αποτελεσματικά, πράγματα που δεν μπορούν να συμβούν χωρίς χρήση μεταδεδομένων αφού τα διοικητικά στελέχη συνήθως δεν έχουν μεγάλη τεχνική κατάρτιση επειδή η ειδίκευση τους είναι σε επιχειρησιακά θέματα.

Οι πληροφορίες που παρέχουν τα μεταδεδομένα μπορεί να αναφέρονται σε προβλήματα του οργανισμού που καλείται να επιλύσει το σύστημα E.E., στο περιεχόμενο και στη σημασία της παραγόμενης πληροφορίας, στις παραγόμενες εκθέσεις συστήματος και στα περιεχόμενα τους, στη δομή του συστήματος, στον καθορισμό των περιεχομένων των δεδομένων. Επίσης, τα μεταδεδομένα παρέχουν λεπτομέρειες και οδηγίες χρήσης των αναλυτικών μεθόδων, πληροφορίες σχετικά με πηγαία δεδομένα, πολιτικές καθαρισμού καθώς και μετασχηματισμού και φόρτωσης των πηγαίων δεδομένων, πληροφορίες σχετικά με εργασίες ETL (Extract, Transform, Load) που πραγματοποιήθηκαν, πληροφορίες σχετικά με την πρόσβαση στο σύστημα και τέλος σε ζητήματα δικαιωμάτων πρόσβασης στο σύστημα.

Η χρήση μεταδεδομένων σε μια επιχείρηση είναι πολύ σημαντική αφού επιτρέπουν τη σωστή ερμηνεία των δεδομένων και παράλληλα την παραγωγή χρήσιμης πληροφορίας ενώ επιβάλλουν την τυποποίηση των δεδομένων και των λειτουργιών με αποτέλεσμα να αφαιρούν περιπτώσεις αυθαίρετων παρεμβάσεων και ερμηνειών από τους χρήστες. Ακόμα, θέτουν μέτρα για την βελτίωση της ποιότητας των δεδομένων και καθιστούν ευκολότερη και αποτελεσματικότερη τη χρήση και τη συντήρηση του συστήματος. Οι επιχειρήσεις μπορούν είτε να αγοράσουν λογισμικό για τήρηση μεταδεδομένων είτε να κατασκευάσουν λογισμικό που ανταποκρίνεται στις δικές τους απαιτήσεις. Το λογισμικό αυτό πρέπει να διαθέτει λειτουργίες επικοινωνίας και να μπορεί να δεχθεί είσοδο από άλλα λογισμικά, όπως είναι λογισμικά ETL, OLAP, εξόρυξης δεδομένων κλπ. Τέλος, για την τήρηση των δεδομένων είναι απαραίτητη η επίβλεψη του από προσωπικό.

2.8.4 Σχεδιασμός

Στο στάδιο αυτό εκτελείται ο σχεδιασμός του πληροφοριακού συστήματος. Τα συστήματα E.E. εξαρτώνται πολύ από τα δεδομένα, για αυτό ένα βασικό κομμάτι τους είναι οι Αποθήκες Δεδομένων (ΑΔ). Οι Αποθήκες Δεδομένων διαφέρουν από τις κλασικές Σχεσιακές Βάσεις Δεδομένων (ΣΒΔ). Για παράδειγμα στις ΣΒΔ πρέπει να αποφεύγεται ο πλεονασμός των δεδο-

μένων, δηλαδή η εμφάνιση των ίδιων δεδομένων πολλές φορές γιατί δημιουργούνται προβλήματα ασυνέπειας των δεδομένων. Σε αντίθεση, η ΑΔ λόγω της ανάγκης για ταχεία πρόσβαση σε μεγάλους όγκους δεδομένων καθιστά αναγκαία την αποθήκευση ίδιων δεδομένων. Άλλες διαφορές είναι ότι στις ΑΔ αποθηκεύονται πολλά δεδομένα του παρελθόντος ενώ παράλληλα γίνεται ο υπολογισμός και η αποθήκευση των δεδομένων, τα οποία δεν εφαρμόζονται στις ΣΒΔ.

Επίσης, η δόμηση των δεδομένων πρέπει να γίνεται με τρόπο που επιτρέπει τα στελέχη ενός οργανισμού να κατανοήσουν και να χρησιμοποιήσουν την ΑΔ χωρίς τη βοήθεια κάποιου ειδικού ή ειδικού λογισμικού. Το λογικό σχήμα που χρησιμοποιείται στις ΑΔ είναι είτε το σχήμα του αστέρα (star schema) είτε μια παραλλαγή του, το σχήματα της χιονονιφάδας (snowflake schema). Σε αυτά τα λογικά σχήματα οι λογικές πληροφορίες χωρίζονται σε γεγονότα (facts) που είναι αριθμητικές ποσότητες συναλλαγών και σε διαστάσεις (dimensions) που είναι πληροφορίες που συνδέονται με γεγονότα. Ο σχεδιασμός των γεγονότων και των διαστάσεων πρέπει να γίνει με τέτοιο τρόπο ώστε να επιτρέπει τα διάφορα τμήματα του οργανισμού να παραθέσουν ερωτήματα και να πάρουν ορθές απαντήσεις. Επίσης, η αποθήκευση των δεδομένων γίνεται αποθηκεύοντας τα συνολικοποιημένα δεδομένα σε σχέση με τις διαστάσεις.

Ένα ακόμα πρόβλημα που πρέπει να αντιμετωπιστεί είναι ο βαθμός λεπτομέρειας της πληροφορίας. Αν ο βαθμός λεπτομέρειας είναι μεγάλος τότε το σύστημα θα είναι αργό, θα απαιτεί μεγάλο χώρο για την αποθήκευση της πληροφορίας και άρα θα είναι αναποτελεσματικό. Αντίθετα, αν ο βαθμός λεπτομέρειας είναι μικρός μπορεί να περιορίσει τις ανάγκες ανάλυσης, αφού υπάρχουν περιπτώσεις όπου η ζητούμενη λεπτομέρεια στη πληροφορία είναι σημαντική ακόμα και αν απαιτεί πολύ χρόνο η παραγωγή της.

Εκτός από ζητήματα λογικού σχεδιασμού της Αποθήκης Δεδομένων είναι διάφορα τεχνικά και άλλα προβλήματα. Για παράδειγμα οι ΑΔ αποθηκεύουν μεγάλο όγκο δεδομένων από διαφορετικά συστήματα τα οποία συνήθως αποθηκεύουν με διαφορετικό τρόπο τα δεδομένα ενώ μπορεί να είναι αντικρουόμενα, ελλιπή ή εσφαλμένα και έτσι δημιουργείται πρόβλημα κατά την Εξαγωγή, τον Μετασχηματισμό και την Φόρτωση (ΕΜΦ) (Extract, Transform, Load (ETL)) των δεδομένων.

Επίσης, πρέπει να οριστούν και πολιτικές ασφαλείας και πιο συγκεκριμένα πρέπει να οριστεί ποιος θα έχει πρόσβαση στα διαφορετικά τμήματα της ΑΔ καθώς και ποια πληροφορία θα είναι προσβάσιμη από κάποιον συγκεκριμένο χρήστη κλπ. Ακόμα, πρέπει να λυθούν και

προβλήματα που προκύπτουν κατά την συνεχή ενημέρωση παλιότερων δεδομένων αλλά και κατά την προσθήκη νέων δεδομένων κατά τη διάρκεια ζωής της ΑΔ. Γενικά, τα δεδομένα κατά την μεταφορά υπόκεινται σε κατάλληλη επεξεργασία, για παράδειγμα εκτελείται η συνολικοποίηση τους σύμφωνα με τα γεγονότα και τις διαστάσεις που ορίζονται στο σχήμα της ΑΔ. Στην αγορά προσφέρονται έτοιμα εργαλεία ΕΜΦ ενώ αν δεν βρεθεί κάποιο κατάλληλο μπορεί να γραφεί κώδικας που να καλύπτει τις απαιτούμενες ανάγκες.

Κατά τον σχεδιασμό του συστήματος εκτελείται και ο σχεδιασμός του λογισμικού ανάλυσης δεδομένων. Η Εξόρυξη Δεδομένων προσφέρει νέες πρωτόγνωρες αναλυτικές τεχνικές που κάνουν χρήση στατιστικών μεθόδων καθώς και μεθόδων τεχνητής νοημοσύνης για να παράξουν πληροφορία πολύτιμη για τη λήψη επιχειρηματικών αποφάσεων. Όμως, κάθε μια από αυτές τις τεχνικές έχει πλεονεκτήματα και μειονεκτήματα με αποτέλεσμα να καθίσταται απαραίτητη η καλή γνώση των δυνατοτήτων και των απαιτήσεων των τεχνικών αυτών από τους σχεδιαστές ώστε να μπορούν να επιλέξουν τις κατάλληλες. Ακόμα, πρέπει να έχουν βαθιά κατανόηση των αναγκών των χρηστών ώστε να μπορούν να επιλέξουν τις κατάλληλες τεχνικές που θα ικανοποιήσουν τις ανάγκες τους.

Επιπροσθέτως, επειδή αυτές οι τεχνικές είναι εξαιρετικά σύγχρονες και η εφαρμογή τους στα συστήματα Επιχειρηματικής Ευφυΐας βρίσκεται σε αρχικά στάδια, οι σχεδιαστές των συστημάτων μπορούν να επεκτείνουν τις υπάρχουσες τεχνικές είτε να προσθέσουν επιπλέον εργαλεία. Τέλος, κατά τον σχεδιασμό του συστήματος πρέπει να σχεδιαστεί και το σύστημα μεταδεδομένων. Τα μεταδεδομένα χρησιμοποιούνται κυρίως σε βάσεις δεδομένων αλλά και σε άλλα λογισμικά, όπως εργαλεία ΕΜΦ, εργαλεία καθαρισμού δεδομένων, λογισμικά OLAP κλπ.

Ακόμα, χρησιμοποιούνται και σε ερμηνείες επιχειρηματικών πληροφοριών που παράγονται από το σύστημα, τις εκθέσεις που συντάσσονται κλπ. Γενικά, τα συστήματα Ε.Ε. απαιτούν τη χρήση πολλών μεταδεδομένων, γεγονός που καθιστά απαραίτητη την ύπαρξη ικανού λογισμικού για τη διαχείριση τους. Το λογισμικό αυτό μπορεί είτε να αγοραστεί έτοιμο είτε να δημιουργηθεί. Το έτοιμο λογισμικό είναι φθηνότερο και άμεσα εφαρμόσιμο, όμως συνήθως δεν καλύπτει πλήρως τις ανάγκες των μεταδεδομένων της επιχείρησης. Με τη δημιουργία λογισμικού καλύπτονται όλες οι ανάγκες, όμως οι απαιτήσεις σε κόστος και χρόνο είναι αυξημένες. Σημαντικά ζητήματα που αφορούν το λογισμικό τήρησης των μεταδεδομένων είναι η αυτόματη ενημέρωση του από άλλα λογισμικά, θέματα προσβασιμότητας στα μεταδεδομένα, καθώς και το εάν το σύστημα θα είναι κεντρικό ή κατανεμημένο.

2.8.5 Υλοποίηση

Σε αυτό το στάδιο εκτελείται η υλοποίηση του συστήματος. Αρχικά γίνεται η εισαγωγή δεδομένων στην ΑΔ. Αυτή η διαδικασία είναι συνήθως περίπλοκη ειδικά αν δεν έχει γίνει αγορά κάποιου έτοιμου προϊόντος ΕΜΦ, στην οποία περίπτωση πρέπει να κατασκευαστεί κατάλληλο λογισμικό από την αρχή. Το λογισμικό ανάλυσης των δεδομένων αποτελείται από ένα ευρύ φάσμα δυνατοτήτων και φιλοσοφίας που μπορούν να επιστρατευτούν ανάλογα με το πρόβλημα που χρήζει επίλυση. Στη συνέχεια, γίνονται οι αρχικές εργασίες εισαγωγής δεδομένων στην ΑΔ. Αυτές περιλαμβάνουν διάφορες διορθώσεις των δεδομένων, όπως για παράδειγμα η κατασκευή νέων δεδομένων με συνδυασμό δεδομένων που ήδη υπάρχουν.

Μετά τη μεταφορά των δεδομένων εκτελούνται έλεγχοι ορθότητας στις διαδικασίες που χρησιμοποιούν τα λογισμικά, στην ροή εκτέλεσης των εργασιών, σε πιθανές αλληλεπιδράσεις που υπάρχουν μεταξύ διαφορετικών λογισμικών και στα τελικά δεδομένα. Κατά τον έλεγχο ορθότητας συνεισφέρουν τόσο τεχνικά στελέχη όσο και στελέχη εξειδικευμένα σε επιχειρηματικά θέματα. Συμπερασματικά, οι διαδικασίες ΕΜΦ απαιτούν πολύ χρόνο και έτσι είναι μεγάλης σημασίας να εκτελεστεί έλεγχος των επιδόσεων του συστήματος.

Εκτός από το λογισμικό ανάλυσης δεδομένων πρέπει να υπάρχει και λογισμικό που ευθύνεται για την οπτικοποίηση των αποτελεσμάτων και τη σύνταξη αναφορών. Τέτοια λογισμικά προσφέρονται από παρόχους συστημάτων Ε.Ε. και καλύπτουν σχεδόν όλες τις απαιτούμενες ανάγκες. Συνεπώς, πρέπει να γίνει αξιολόγηση των λογισμικών ώστε να γίνει η επιλογή αυτού που εξυπηρετεί καλύτερα τις απαιτούμενες ανάγκες και αν δεν υπάρχει τέτοιο τότε συνήθως είναι δυνατός και ο συνδυασμός διαφορετικών προϊόντων από διαφορετικούς κατασκευαστές. Σε περίπτωση που έχει αποφασιστεί η δημιουργία λογισμικού τότε σε αυτό το στάδιο εκτελείται η υλοποίηση του. Εάν στο στάδιο της ανάλυσης έγινε χρήση πρωτότυπου τότε εξελίσσεται σε ολοκληρωμένο σύστημα. Ειδικά σε περίπτωση που το πρωτότυπο ενσωματώνει πραγματικές λειτουργίες, δηλαδή είναι λειτουργικό, η μετατροπή του σε ολοκληρωμένο σύστημα είναι πιο εύκολη. Τέλος, το αγορασμένο αλλά και το δημιουργημένο λογισμικό υποβάλλεται σε εντατικές διαδικασίες ελέγχων.

Ένα ακόμα ζήτημα που πρέπει να οριστεί και να υλοποιηθεί ο τρόπος πρόσβασης στο σύστημα. Το διαδίκτυο αποτελεί ιδανικό περιβάλλον για την φιλοξενία του συστήματος αφού η ευρεία ανάπτυξη και διάδοση του προσφέρει ευχρηστία και μπορεί να αυξήσει την ανταγωνιστικότητα του διευκολύνοντας την απομακρυσμένη πρόσβαση, που μπορεί να γίνει ακόμα και μέσω κινητών συσκευών. Αυτή η λύση μπορεί να κρύβει αρκετούς κινδύνους που πρέπει

να αντιμετωπιστούν υλοποιώντας ένα σύστημα για την ελεγχόμενη και διαβαθμισμένη πρόσβαση στο σύστημα. Αν έχει γίνει επιλογή του διαδικτύου ως τρόπο πρόσβασης τότε σε αυτό το στάδιο γίνεται η δημιουργία της πύλης ΕΕ.

Τέλος, υλοποιείται το σύστημα μεταδεδομένων. Η επιχείρηση έχει δύο επιλογές είτε να αγοράσει έτοιμο εργαλείο τήρησης μεταδεδομένων είτε να κατασκευάσει το δικός της ώστε να ικανοποιεί όλες τις ανάγκες της. Τα μεταδεδομένα που θα εισαχθούν στο σύστημα προέρχονται κυρίως από άλλα λογισμικά, για αυτό το σύστημα μεταδεδομένων πρέπει να έχει δυνατότητες επικοινωνίας με τα συστήματα που το τροφοδοτούν. Ένα έτοιμο σύστημα που προσφέρεται στην αγορά έχει τέτοιες δυνατότητες, όμως δεν είναι πάντα αρκετές.

Επίσης, πρέπει να οριστεί κάποιος υπεύθυνος για την τήρηση του συστήματος μεταδεδομένων, αφού πρέπει να τηρείται διαρκώς ενημερωμένο και κάθε αλλαγή που γίνεται στα δεδομένα πρέπει να καταγράφεται στο σύστημα μεταδεδομένων ενώ πρέπει να περιγράφει και τα δεδομένα του συστήματος Ε.Ε. με ακρίβεια. Ακόμα, καλό είναι να οριστούν διαφορετικές διεπαφές πρόσβασης προσαρμοσμένες στις ανάγκες των διαφόρων χρηστών του συστήματος ώστε να διευκολύνεται το έργο τους, αφού εκτός από τον διαχειριστή, το σύστημα χρησιμοποιείται και από άλλα άτομα της επιχείρησης. Άλλη μια λύση είναι η οργάνωση των περιεχομένων του συστήματος μεταδεδομένων σε θεματικές ενότητες που εξυπηρετούν τις ανάγκες των διαφορετικών κατηγοριών χρηστών ενώ πρέπει να ληφθούν και μέτρα ασφάλειας για την πρόσβαση στα μεταδεδομένα.

2.8.6 Εφαρμογή

Στο στάδιο της εφαρμογής το σύστημα τίθεται σε λειτουργία. Αρχικά, γίνεται η εγκατάσταση του συστήματος καθώς και η εφαρμογή απαραίτητων τεχνικών ρυθμίσεων. Η παράδοση του συστήματος είναι προτιμότερο να γίνεται σταδιακά σε τμήματα κατά τη διάρκεια της εξέλιξης του, αφού έτσι περιορίζονται οι κίνδυνοι και διαχειρίζονται οι αλλαγές που επιφέρει η εγκατάσταση του συστήματος στον οργανισμό, καθώς οι χρήστες έρχονται σε επαφή και εξοικειώνονται με λειτουργίες του συστήματος πριν την ολοκλήρωση του.

Στη συνέχεια, ακολουθεί η διαδικασία της εκπαίδευσης των χρηστών. Η παρουσίαση των λειτουργιών του συστήματος βοηθάει στην κατανόηση του συστήματος από τους χρήστες ώστε να είναι σε θέση να εκτελέσουν βασικές εργασίες. Το σύστημα προσφέρει ποικιλία εργασιών από τις οποίες ο χρήστης πρέπει να επιλέξει τις κατάλληλες ώστε να ικανοποιήσει τις αναλυτικές του ανάγκες. Έτσι, για να εκτελέσουν τις εργασίες τους αποτελεσματικότερα

απαιτείται βαθύτερη κατανόηση των δυνατοτήτων του συστήματος. Η καλύτερη μέθοδος για την απόκτηση αυτής της ικανότητας είναι η εκπαίδευση των χρηστών σε πραγματικές συνθήκες όπου διαχειρίζονται πραγματικά δεδομένα και αντιμετωπίζουν πραγματικά προβλήματα. Επίσης, η διαδικασία της εκπαίδευσης πρέπει να επικεντρώνεται σε επιχειρησιακά ζητήματα, συνεπώς η συνεισφορά των στελεχών που συμμετείχαν στο σχεδιασμό του συστήματος κρίνεται σημαντική, αφού αυτοί έχουν ήδη κατανοήσει τον τρόπο που το σύστημα εξυπηρετεί τα επιχειρηματικά ζητήματα.

Μετά την εγκατάσταση του συστήματος, η χρήση του δημιουργεί ανάγκες συντήρησης ώστε να παραμένει διαρκώς ενημερωμένο και λειτουργικό. Σε συστήματα ΕΕ η πιο σημαντική εργασία συντήρησης είναι η μεταφορά δεδομένων (ΕΜΦ), αφού τα δεδομένα που χρησιμοποιούνται προέρχονται από άλλα πληροφοριακά συστήματα. Οι εργασίες ΕΜΦ εκτελούνται τακτικά για να τροφοδοτήσουν το σύστημα με δεδομένα. Συνεπώς, μετά από τις εργασίες ΕΜΦ πρέπει να γίνεται έλεγχος ορθότητας των δεδομένων και να ενημερώνεται το σύστημα μεταδεδομένων. Επίσης, η διαχείριση μεγέθυνσης του συστήματος επηρεάζει σημαντικά την λειτουργικότητα του.

Τα συστήματα ΕΕ μεγεθύνονται με ραγδαίους ρυθμούς αφού προστίθενται νέες λειτουργικότητες που απαιτούν νέα δεδομένα ή την επέκταση των ήδη υπαρχόντων. Ταυτόχρονα, αυξάνονται συνεχώς και οι χρήστες του συστήματος. Η μεγέθυνση του συστήματος πρέπει να έχει προβλεφθεί ώστε να αποφευχθούν τα προβλήματα που προκαλεί. Αυτό συνεπάγεται με την επιλογή κατάλληλου υλικού και λογισμικού ώστε να μπορεί να υποστηριχθεί η επέκταση του συστήματος. Ακόμα, η συντήρηση του συστήματος απαιτεί την δημιουργία και αποθήκευση αντιγράφων ασφαλείας αλλά και την παρακολούθηση χρήσης του.

2.8.7 Αξιολόγηση

Στο τέλος κάθε κύκλου ζωής ενός συστήματος Επιχειρηματικής Ευφυΐας εκτελούνται διαδικασίες αξιολόγησης του. Σε αυτό το στάδιο εκτελούνται διαδικασίες αξιολόγησης ως προς τον προϋπολογισμό και το χρονοδιάγραμμα εκτέλεσης του έργου. Πιο συγκεκριμένα, για το χρονοδιάγραμμα εκτέλεσης του συστήματος ελέγχεται αν τηρήθηκε ο προκαθορισμένος χρόνος που ορίστηκε για την δημιουργία κάθε υποέργου, τυχόν καθυστερήσεις που προέκυψαν καθώς και το πως αυτές επηρεάζουν την εκτέλεση των άλλων υποέργων κλπ. Ως προς τον προϋπολογισμό του έργου, γίνεται απογραφή του κόστους εκτέλεσης των υποέργων και προμήθειας εξοπλισμού καθώς ελέγχονται και αιτιολογούνται πιθανές υπερβάσεις. Επίσης,

ως μέσο αξιολόγησης χρησιμοποιείται η ανάλυση του κόστους και του οφέλους. Ο υπολογισμός του κόστους είναι σχετικά εύκολος, σε αντίθεση με τον υπολογισμό του οφέλους είναι αυξημένης δυσκολίας, ειδικά αν το έργο δεν είναι σχεδιασμένο να επιλύει προβλήματα με μετρήσιμα αποτελέσματα.

Επίσης, τα συστήματα Ε.Ε. προσφέρουν και άλλα λιγότερο προφανή οφέλη όπως η απόκτηση νέας επιχειρηματικής γνώσης, η βελτίωση των διαδικασιών, κυρίως στη λήψη αποφάσεων, και οι πιο αποτελεσματικές σχέσεις. Γενικά, τα έμμεσα οφέλη, όπως η μείωση του ρίσκου και η αύξηση της ανταγωνιστικότητας, χαρακτηρίζονται από μεγάλη δυσκολία ποσοτικοποίησης τους ενώ μπορεί να είναι και μη μετρήσιμα. Η αποτίμηση του έργου εκτελείται με κάθε ολοκλήρωση του κύκλου ζωής του συστήματος ακόμα και αν αυτό δεν λειτουργεί πλήρως σωστά. Αυτή η διαδικασία προσφέρει χρήσιμη γνώση για την βελτίωση του συστήματος. Εκτός από θέματα προϋπολογισμού και χρονοπρογραμματισμού, εκτελούνται έλεγχοι και ως προς τον βαθμό κάλυψης των απαιτήσεων του συστήματος, την καταλληλότητα της παραγόμενης πληροφορίας, την ικανοποίηση των χρηστών, την ποιότητα των δεδομένων, την αποτελεσματικότητα των αναλυτικών μεθόδων, τον βαθμό χρήσης του συστήματος, την ποιότητα των εκθέσεων και αναφορών και την ταχύτητα λειτουργίας του συστήματος.

Η αξιολόγηση του έργου εφαρμόζεται με την οργάνωση συνεντεύξεων και ερωτηματολογίων και εφόσον έχει περάσει επαρκή χρονικό διάστημα μετά την εγκατάσταση του συστήματος ώστε οι χρήστες να έχουν αποκτήσει αρκετή εμπειρία χρήσης του. Στην αξιολόγηση εκτός από τους χρήστες του συστήματος σημαντικό είναι να συμμετέχουν και όλα τα στελέχη που συνείσφεραν στην δημιουργία του. Επίσης, σημαντικό είναι οι ερωτήσεις να γίνονται γνωστές στους ενδιαφερόμενους αρκετό χρονικό διάστημα πριν ώστε να έχουν αρκετό χρόνο να προετοιμαστούν, ενώ καλό θα ήταν να ανατεθεί σε κάποιο στέλεχος ο συντονισμός και η διεξαγωγή της έρευνας.

Οι ανάγκες για ανάκτηση επιχειρηματικής γνώσης δεν είναι σταθερές, γεγονός που εξηγεί γιατί τα συστήματα Επιχειρηματικής Ευφυΐας είναι συστήματα σε διαρκή εξέλιξη. Οι χρήστες πρέπει να ανακαλύπτουν και να μοντελοποιούν διαρκώς νέα γνώση και να εκτελούν κατάλληλες τροποποιήσεις στα δεδομένα. Συμπερασματικά, το σύστημα είναι απαραίτητο να ακολουθεί μια λογική διαδοχικών εκδόσεων που παρέχουν κατάλληλες τροποποιήσεις. Η αναβάθμιση ενός έτοιμου συστήματος από παρόχους λογισμικού είναι επιθυμητή, ενώ η αναβάθμιση ενός συστήματος που έχει αναπτυχθεί στα πλαίσια κάποιας επιχείρηση προκαλεί

επιπλέον καθυστερήσεις στη λειτουργία της ενώ επιβαρύνει με έξοδα. Ωστόσο, η αναβάθμιση κρίνεται απαραίτητη παρά τις επιπλέον επιβαρύνσεις και δυσκολίες που έπονται.

Οι χρήστες του συστήματος, που είναι υψηλόβαθμα στελέχη, κατανοούν σε βάθος το σύστημα και τον στρατηγικό προσανατολισμό της επιχείρησης. Έτσι, είναι πολύ πιθανόν με την συνεχή χρήση του συστήματος να ανακαλύπτουν νέες ανάγκες που πρέπει να προστεθούν σε επόμενες εκδόσεις του συστήματος. Οι διαπιστώσεις διαφορετικών χρηστών μπορεί να είναι ταυτόσημες, επικαλυπτόμενες ή αντικρουόμενες. Ακόμα, σε μεγάλους οργανισμούς που είναι γεωγραφικά διασκορπισμένοι οι ιδέες για την βελτίωση του συστήματος περιορίζονται τοπικά και δεν μεταδίδονται σε όλο το εύρος του οργανισμού. Έτσι, είναι σημαντικό να οργανωθούν διάφορα μέσα επικοινωνίας και καταγραφής απόψεων. Αυτή η ανάγκη ικανοποιείται από εργαλεία που προσφέρει το Web 2.0, όπως είναι τα εταιρικά blogs και wikis, τα οποία μπορούν να χρησιμοποιηθούν αποτελεσματικά για επικοινωνία και μεταφορά απόψεων, που θα οργανώσουν τις ανάγκες που απαιτούνται για την αναβάθμιση του συστήματος.

Γενικά, έρευνες έχουν δείξει ότι οι παράγοντες που οδηγούν στην επιτυχία ενός συστήματος Ε.Ε. παίρνουν τρεις διαστάσεις, ως προς τον οργανισμό, τη διαδικασία και την τεχνολογία. Πιο συγκεκριμένα, είναι θεμιτή και αναγκαία η σταθερή υποστήριξη από τη διοίκηση καθώς και η ανάληψη του αναγκαίου κόστους, πρέπει να υπάρχει ένα καθαρό στρατηγικό όραμα και σαφές επιχειρηματικό ζητούμενο, πρέπει να υπάρχει οριοθέτηση του πεδίου και σταδιακή ανάπτυξη. Επίσης, η σύνθεση της ομάδας του έργου πρέπει να είναι σταθμισμένη κατάλληλα με τη συμμετοχή ειδικών σε επιχειρηματικά θέματα, ενώ στην ανάπτυξη του συστήματος θα πρέπει να συμμετέχουν και χρήστες. Τέλος, η ποιότητα των δεδομένων πρέπει να είναι καλή και το σύστημα πρέπει να είναι επεκτάσιμο και να έχει ιδιότητες προσαρμογής.

2.8.8 Η Επιχειρηματική Ευφυΐα Ως Υπηρεσία

Όπως έχει ήδη αναφερθεί, η ανάπτυξη συστημάτων Ε.Ε. απαιτεί πολλούς πόρους. Μεγάλες επιχειρήσεις που έχουν ανάγκες για πληροφόρηση μπορούν να διαθέσουν πιο εύκολα τέτοιο μέγεθος πόρων, σε αντίθεση με μικρές επιχειρήσεις που το αυξημένο κόστος λειτουργεί ως ανασταλτικός παράγοντας. Παράλληλα, ο συνδυασμός της μεγάλης πολυπλοκότητας του εγχειρήματος και η έλλειψη εμπειρίας αποτελούν προβλήματα που εμποδίζουν την αγορά συστημάτων Ε.Ε. από μικρές επιχειρήσεις. Λύση σε αυτά τα προβλήματα δίνει η Επιχειρηματική Ευφυΐα ως Υπηρεσία (ΕΕΩΥ) (Business Intelligence As A Service) που περιορίζει σημαντικά το κόστος και διευκολύνει την ανάπτυξη του συστήματος.

Η υπολογιστική νέφους (cloud computing) επιτρέπει παρόχους να προσφέρουν υπηρεσίες υπολογιστικών συστημάτων μέσω του διαδικτύου. Πιο συγκεκριμένα μπορεί να παρέχονται εφαρμογές (Software As A Service), εργαλεία ανάπτυξης (Platform As A Service), καθώς και υπολογιστικοί πόροι υποδομής (Infrastructure As A Service). Έτσι, οι διαδικασίες που απαιτούνται για την λειτουργία, την ανάπτυξη και γενικότερα την υποδομή του συστήματος είναι μέριμνα του παρόχου και όχι της εκάστοτε επιχείρησης που το αγοράζει. Οι χρήστες μοιράζονται υπολογιστικούς πόρους, τεχνογνωσία κλπ με αποτέλεσμα να επιτυγχάνεται σημαντική μείωση στο κόστος.

Ένα ακόμα ερώτημα που δημιουργείται είναι ποια από τα υποσυστήματα που απαρτίζουν ένα σύστημα Ε.Ε. Θα μεταφερθούν στο νέφος και ποια θα μείνουν ιδιόκτητα. Επίσης, η επιλογή του κατάλληλου παρόχου είναι ένα περίπλοκο πρόβλημα και πρέπει να γίνει με βάση την αξιολόγηση των υπηρεσιών που προσφέρει και της αξιοπιστίας του. Ένα ακόμα θέμα που πρέπει να ληφθεί υπόψη είναι η παροχή ασφάλειας. Τα συστήματα Ε.Ε. περιέχουν απόρρητα δεδομένα, οπότε η μεταφορά τους μέσω του διαδικτύου σε εξωτερικούς υπολογιστές μπορεί να αποκρύπτει κινδύνους. Έτσι, οι πάροχοι ανατρέχουν σε τρίτους φορείς για την πιστοποίηση της ασφάλειας των δεδομένων των πελατών. Ένα ακόμα σημαντικό ζήτημα είναι οι επιδόσεις του συστήματος. Η παροχή της υπηρεσίας μέσω διαδικτύου μπορεί να καθυστερήσει την εκτέλεση των εργασιών. Αν το πρόβλημα οφείλεται στον πάροχο τότε ο χρήστης δεν μπορεί να παρέμβει, όμως μπορεί να υπογράψει σύμβαση με τον πάροχο που τον υποχρεώνει να προβεί σε κατάλληλες ρυθμίσεις για την εξασφάλιση ικανοποιητικών επιδόσεων.

Συστήματα Ε.Ε. ως υπηρεσία παρέχονται από πολλούς μεγάλους κατασκευαστές της αγοράς. Η Microsoft παρέχει το Power Business Intelligence for Office 365, το οποίο προσφέρει αυξημένες δυνατότητες συνεργασίας, μοντελοποίηση στο EXCEL, διαχείριση δεδομένων, καθώς και δημιουργία dashboards και τρισδιάστατων απεικονίσεων. Η Oracle παρέχει το λογισμικό Business Intelligence Cloud Service, το οποίο προσφέρει συλλογή και μεταφόρτωση δεδομένων, εργασίες ΕΜΦ, προχωρημένες δυνατότητες ανάλυσης κλπ. Η IBM παρέχει μια μεγάλη γκάμα διαθέσιμων προϊόντων μεταξύ των οποίων τα λογισμικά Cognos, SPSS και Watson Analytics τα οποία προσφέρονται σε εκδόσεις υπολογιστικής νέφους. Αυτά τα συστήματα παρέχουν ανάλυση κινδύνου για τον υπολογισμό και τη διαχείριση του ρίσκου, τη διεξαγωγή προγνωστικών αναλύσεων και τη λήψη αποφάσεων από άτομα και ομάδες ατόμων, την ανάλυση των χρηματοοικονομικών στοιχείων καθώς και τη σύνταξη των χρηματοοικονομικών καταστάσεων.

Κεφάλαιο 3 – Κοινωνικά Δίκτυα και Μηχανική Μάθηση

Περιεχόμενα κεφαλαίου

3.1 Κοινωνικά Δίκτυα και επιχειρήσεις	39
3.1.1 Εισαγωγή στο Web και κοινωνικός ιστός	40
3.1.2 Επιρροή του Twitter στις σημερινές επιχειρήσεις	41
3.1.3 Opinion mining και είδη	42
3.1.4 Εισαγωγή στα APIs	44
3.1.5 Εξαγωγή πληροφορίας (information extraction)	45
3.2 Μηχανική Μάθηση	52
3.2.1 Ορισμός της Μηχανικής Μάθησης	52
3.2.2 Είδη Μηχανικής Μάθησης	54
3.2.3 Αλγόριθμοι Μηχανικής Μάθησης	55
3.3 Μετρικές αξιολόγησης	75

3.1 Κοινωνικά Δίκτυα και επιχειρήσεις

Τα τελευταία χρόνια, τα social media δίνουν την δυνατότητα στους ανθρώπους να εκφέρουν άφοβα την άποψή τους για πολλά σημαντικά ζητήματα που απασχολούν τους ίδιους ή τις κοινωνίες στις οποίες ανήκουν. Στην εποχή της πληροφορίας στην οποία ανήκουμε, η δημιουργία ενός εργαλείου με το οποίο θα μπορούσε κάποιος να αντιληφθεί τις συσχετίσεις και τις απόψεις του κόσμου γύρω από τα τρέχοντα ζητήματα, θα μπορούσε να βοηθήσει στην ανάπτυξη εφαρμογών ή την κατανόηση των ζητημάτων που απασχολούν την κοινή γνώμη. Η αύξηση της δημοτικότητας των κοινωνικών μέσων, όπως τα blogs και τα social networks, μεγέθυναν το ενδιαφέρον για την ανάλυση συναισθήματος (sentiment analysis).

Η ραγδαία αύξηση των reviews, των ratings, των recommendations αλλά και άλλων μορφών ηλεκτρονικής έκφρασης γνώμης, ουσιαστικά μετέτρεψε την ηλεκτρονική γνώμη σε κάποιου είδος εικονικού νομίσματος για επιχειρήσεις που επιθυμούν να πουλήσουν τα προϊόντα τους, να αναγνωρίσουν καινούργιες ευκαιρίες καθώς και να διαχειριστούν τη φήμη τους. Οι επιχειρήσεις προσπαθούν να εκμεταλλευτούν αυτή την κατάσταση προς όφελος

τους. Για την επίτευξη αυτού του στόχου πρέπει να αυτοματοποιηθούν οι διαδικασίες κατανόησης συζητήσεων, η αναγνώριση του σχετικού περιεχομένου και η κατάλληλη χρήση του καθώς και οι μέθοδοι που επεξεργάζονται τον θόρυβο με σκοπό την εξάλειψη του. Σε αυτά τα προβλήματα δίνει λύση το πεδίο της ανάλυσης των συναισθημάτων (sentiment analysis).

3.1.1 Εισαγωγή στο Web και κοινωνικός ιστός

To Web αποτελείται από τρία συστατικά, το Web 1.0, το Web 2.0 και το Web 3.0. To Web 1.0 ονομάζεται Document Web και αποτελείται από συνδέσμους μεταξύ εγγράφων. Πιο συγκεκριμένα, είναι ένα κατανεμημένο σύστημα παράδοσης εγγράφων που χρησιμοποιεί πρωτόκολλα του Internet. Βασίζεται στο μοντέλο client-server και ενώνει έγγραφα που αποθηκεύονται σε υπολογιστικά συστήματα που επικοινωνούν μέσω Διαδικτύου. Η παροχή των πληροφοριών είναι μονόδρομη από τους σχεδιαστές-ιδιοκτήτες των ιστοσελίδων προς τους επισκέπτες-καταναλωτές του περιεχομένου. To Web 2.0 ονομάζεται Social Web και αποτελείται από συνδέσμους μεταξύ ανθρώπων. Είναι ένα σύνολο από οικονομικές, κοινωνικές και τεχνολογικές τάσεις που αθροιστικά διαμορφώνουν τη βάση για την επόμενη γενιά του Διαδικτύου. Δεν είναι καινούργιο πρωτόκολλο του Παγκοσμίου Ιστού, αλλά διαφέρει στον τρόπο που χρησιμοποιεί ήδη υπάρχουσες τεχνολογίες, καθώς στον τρόπο με τον οποίο χρησιμοποιούν οι σχεδιαστές πληροφοριακών συστημάτων και οι χρήστες το Διαδίκτυο.

Επίσης, είναι το μέσο στο οποίο πλέον τρέχουν εφαρμογές και υπηρεσίες, κάτι που μέχρι τώρα γινόταν τοπικά στους υπολογιστές των χρηστών. Ακόμα, χαρακτηρίζεται ως δυναμικός Παγκόσμιος Ιστός, ο οποίος οργανώνεται με κυρίαρχη αρχή την σύνδεση μεταξύ ανθρώπων και χρηστών και δίνει έμφαση στην παροχή υπηρεσιών στους χρήστες. To Web 3.0 ονομάζεται Semantic Web και αποτελείται από συνδέσμους μεταξύ δεδομένων.

To Web 2.0 χαρακτηρίζεται από ένα σύνολο βασικών αρχών. Όπως προαναφέρθηκε, η πλατφόρμα ανάπτυξης εφαρμογών είναι ο Παγκόσμιος Ιστός. Οι χρήστες δεν έχουν παθητικό ρόλο, αλλά είναι οι δημιουργοί της πληροφορίας, αφού ανεβάζουν δεδομένα, τα περιγράφουν και αναπτύσσουν εφαρμογές. Ο βασικός πυρήνας των εφαρμογών είναι τα δεδομένα, ενώ οι εφαρμογές δεν έχουν μορφή προσφερόμενου προϊόντος αλλά προσφερόμενων υπηρεσιών στους χρήστες. To Web 2.0 βασίζεται στη συμμετοχή, στα πρότυπα, στη αποκέντρωση, στην ιδιωτικότητα, στην δυνατότητα διαμόρφωσης, στον έλεγχο χρήστη και στην ταυτοποίηση. Ακόμα, εκφράζεται σε δύο βασικές περιοχές, τον ευρύτερο ιστό και την επιχείρηση. Ο κοινωνικός ιστός χρησιμοποιείται για την κοινωνικοποίηση των ανθρώπων και για την μεταξύ

τους αλληλεπίδραση μέσω του Παγκοσμίου Ιστού δημιουργώντας και εξελίσσοντας ψηφιακά επικοινωνιακά ανθρωποδίκτυα. Τα δεδομένα που χρησιμοποιούνται είναι κοινωνικού τύπου, δηλαδή δεδομένα που σχετίζονται άμεσα ή μπορούν να χρησιμοποιηθούν έμμεσα για προσδιορισμό σχέσεων μεταξύ ανθρώπων.

Επίσης, τα δεδομένα καθορίζονται από τις πράξεις των χρηστών και μπορεί να συνοδεύονται από μεταδεδομένα (metadada) και επισημειώσεις (tagging). Ο κοινωνικός ιστός περιλαμβάνει ένα σύνολο από κοινωνικές συσχετίσεις οι οποίες διασυνδέουν ανθρώπους δια μέσου του Παγκοσμίου Ιστού και βασίζεται στην κατάλληλη υποδομή ιστοτόπων που αξιοποιούν Web 2.0 τεχνολογίες. Οι τεχνολογίες του κοινωνικού ιστού είναι πολυάριθμες και χωρίζονται σε τεχνολογίες υποδομής, τεχνολογίες διάθεσης περιεχομένου και τεχνολογίες που αναδεικνύουν το χρήστη. Οι πιο διαδεδομένες είναι τα wikis, τα ιστολόγια (blogs), το tagging, το social networking, τα RSS feeds και το podcasting και webcasting. Τα κοινωνικά δίκτυα (social networks) είναι μια δομή που παράγεται από άτομα ή οργανισμούς, οι οποίοι συνδέονται μεταξύ τους με έναν ή περισσότερους τύπους αλληλεξάρτησης. Τα κοινωνικά δίκτυα ουσιαστικά είναι ένας γράφος όπου τα άτομα και οι οργανισμοί είναι οι κόμβοι, ενώ ο τρόπος σύνδεσης τους είναι οι σύνδεσμοι του γράφου.

3.1.2 Επιρροή του Twitter στις σημερινές επιχειρήσεις

Η άνοδος της δημοτικότητας των κοινωνικών δικτύων έφερε στην επιφάνεια νέες μορφές ατομικής έκφρασης και επικοινωνίας. Στη σημερινή κοινωνία, τα κοινωνικά δίκτυα κατέχουν πρωτεύον θέση στην μεταφορά πληροφορίας, αφού καθημερινά δημιουργούνται και διαχειρίζονται μεγάλα ποσά δεδομένων από τα κοινωνικά δίκτυα, όπως είναι τα blogs και υπηρεσίες όπως το Facebook και το Twitter. To Twitter είναι ένα μέσο κοινωνικής δικτύωσης, στο οποίο οι χρήστες συμμετέχουν άμεσα στα τρέχοντα ζητήματα και εκφράζουν την άποψη τους με πολύ περιεκτικά μηνύματα, λόγω του περιορισμού των χαρακτήρων σε 140.

Επιπλέον, θα πρέπει να αναφερθεί ότι το δεύτερο τρίμηνο του 2017 οι ενεργοί χρήστες του twitter αριθμούν 328 εκατομμύρια οι οποίοι παράγουν περίπου 500 εκατομμύρια tweets ανά ημέρα. To Twitter επιτρέπει στους χρήστες να διατυπώσουν μεγάλο πλήθος μηνυμάτων μικρού μεγέθους με σκοπό την επικοινωνία, τον σχολιασμό και την συνομιλία για σχετικά γεγονότα οποιοδήποτε είδους, για παράδειγμα πολιτικά, προϊόντα κ.α. Η τάση για χρήση των κοινωνικών δικτύων έχει αυξηθεί σημαντικά από την παγκόσμια διάδοση των κινητών τηλεφώνων, ενώ παράλληλα η πρόσβαση είναι εύκολη και διαπλατφορμική (cross-platform).

Είναι σημαντικό να σημειωθεί ότι τα κοινωνικά δίκτυα σε πολλές περιπτώσεις προσφέρουν πληροφορία που είναι πιο ενημερωμένη από τις συμβατικές πηγές πληροφόρησης, όπως είναι τα online νέα, που τα καθιστούν ιδιαίτερα ελκυστική πηγή πληροφοριών. Ως επακόλουθο, η αυτοματοποιημένη ανάλυση περιεχομένου καθώς και η εξαγωγή πληροφορίας από τα κοινωνικά δίκτυα είναι ένας περιζήτητος τομέας. Η εξόρυξη δεδομένων από τα κοινωνικά δίκτυα είναι μια πιο περίπλοκη διαδικασία από την κλασσική εξόρυξη δεδομένων, αφού τα κείμενα στα κοινωνικά δίκτυα δεν περιέχουν καλά μορφοποιημένα γραμματικά κείμενα.

Τα κυριότερα προβλήματα που συναντώνται κατά την εξαγωγή πληροφορίας από τα κοινωνικά δίκτυα είναι: τα κείμενα είναι μικρά σε μέγεθος, αφού το Twitter επιτρέπει μέχρι 140 χαρακτήρες ενώ το Facebook μέχρι 255 χαρακτήρες. Επίσης, τα κείμενα περιέχουν θόρυβο και είναι διατυπωμένα σε ανεπίσημη μορφή γλώσσας, περιλαμβάνονταν ορθογραφικά λάθη, έλλειψη σημείων στίξης και κεφαλαίων, χρήση μη τυπικών συντομογραφιών και δεν περιέχουν γραμματικά ορθές προτάσεις. Ακόμα, περιέχουν υψηλή αβεβαιότητα της αξιοπιστίας των πληροφοριών που μεταδίδονται στα μηνύματα κειμένου σε σύγκριση με τα μέσα ενημέρωσης. Τα προαναφερθέντα προβλήματα δημιουργησαν την ανάγκη για μια νέα κατηγορία έρευνας στον τομέα της εξόρυξης δεδομένων. Αυτή η κατηγορία επικεντρώνεται στην ανάπτυξη μεθόδων για εξαγωγή πληροφορίας από σύντομα μηνύματα κειμένου που περιέχουν θόρυβο. Η έρευνα για την εξαγωγή πληροφορίας από τα κοινωνικά μέσα βρίσκεται ακόμη στα αρχικά της στάδια και επικεντρώνεται κυρίως στην επεξεργασία αγγλικών.

3.1.3 Opinion mining και είδη

Οι περισσότερες αγοραστικές αποφάσεις στον εικονικό κόσμο εκτελούνται μετά από την επιρροή των αναθεωρητών (reviewers) και άλλων αντίστοιχων ατόμων που μεταδίδουν τη γνώση τους σχετικά με κάποιο προϊόν ή υπηρεσία. Αυτός είναι ο λόγος για τον οποίο οι επιχειρήσεις αναγκάζονται να αναζητήσουν και να αναλύσουν δεδομένα ώστε να πάρουν χρήσιμη πληροφορία σχετικά με τα ενδιαφέροντα των ατόμων στο διαδίκτυο. Από την οπτική γωνία των εταιρειών, οι κριτικές και τα σχόλια είναι πολύ κρίσιμα. Επομένως, η ανάλυση των σχολίων και των αναθεωρήσεων είναι σημαντική εργασία για οποιοδήποτε οργανισμό. Αυτά τα σχόλια, οι απόψεις και οι αναθεωρήσεις είναι γνωστά ως «δεδομένα αισθήσεων» (sentiment data), ενώ η διαδικασία που προσδιορίζει εάν τα σχόλια και οι αναθεωρήσεις είναι θετικά ή αρνητικά είναι γνωστά ως «ανάλυση δεδομένων αισθήματος» (sentiment data analysis) ή «ανάλυση συναισθημάτων» (sentiment analysis) ή «Εξόρυξη γνώμης» (opinion mining).

Η εξόρυξη γνώμης (opinion mining) μπορεί να είναι χρήσιμη με διάφορους τρόπους. Μπορεί να βοηθήσει τους διαφημιζόμενους να αξιολογήσουν την επιτυχία μιας διαφημιστικής καμπάνιας ή να προωθήσουν ένα νέο προϊόν, να καθορίσουν ποιες εκδόσεις ενός προϊόντος ή μιας υπηρεσίας είναι δημοφιλείς και να προσδιορίσουν ποια δημογραφικά χαρακτηριστικά είναι ποηθητά ή αντιτίθενται σε συγκεκριμένες λειτουργίες του προϊόντος. Σε γενικές γραμμές, το συναίσθημα μπορεί να περιλαμβάνει πολικότητα ή σθένος (π.χ. θετικό, αρνητικό, ουδέτερο), αισθήματα (θυμωμένος, χαρούμενος, λυπημένος, υπερήφανος, απογοητευμένος, κ.λπ.). και άλλες συναισθηματικές καταστάσεις.

Υπάρχουν διάφορες μέθοδοι για την ανάλυση των δεδομένων «αισθήσεων» (sentiment data), όπως η ανάλυση συναισθηματικού επιπέδου εγγράφου (Document-level of sentiment analysis), προτασιακού επιπέδου ανάλυση συναισθημάτων (Sentence-level of sentiment analysis), ανάλυση αισθήσεων βασισμένη σε διάσταση (Aspect based sentiment analysis), ανίχνευση υποκειμενικότητας / αντικειμενικότητας (Subjectivity/objectivity identification), ανάλυση συγκριτικού συναισθήματος (Comparative sentiment analysis) κ.α.

Document-level of sentiment analysis

Στην ανάλυση συναισθηματικού επιπέδου εγγράφου, κάθε έγγραφο επικεντρώνεται σε μια ενιαία οντότητα ή συμβάν και περιέχει γνώμη από μόνο ένα άτομο. Η άποψη εδώ μπορεί να ταξινομηθεί σε δύο απλές κατηγορίες: θετική ή αρνητική (πιθανώς ουδέτερη). [5]

Sentence-level of sentiment analysis

Σε αυτό το επίπεδο φιλτράρονται οι προτάσεις που δεν περιέχουν κάποια γνώμη και καθορίζεται κατά πόσο η γνώμη για την οντότητα που περιγράφεται είναι θετική ή αρνητική. [5]

Aspect based sentiment analysis

Η ανάλυση βασισμένη σε διάσταση (aspect) επικεντρώνεται στην αναγνώριση όλων των εκφράσεων που έχουν να κάνουν με συναισθήματα μέσα σε ένα συγκεκριμένο έγγραφο αλλά και στις διαστάσεις στις οποίες αναφέρονται οι απόψεις. [5]

Subjectivity/objectivity identification

Σε αυτό το επίπεδο εκτελείται συνήθως ταξινόμηση ενός συγκεκριμένου κειμένου (συνήθως μια πρόταση) σε μία από τις δύο κατηγορίες: αντικειμενική ή υποκειμενική. [6]

Comparative sentiment analysis [5]

Σε πολλές περιπτώσεις, οι χρήστες εκφράζουν τις απόψεις τους σε σύγκριση με ένα παρόμοιο προϊόν ή μάρκα. Επομένως, ο στόχος εδώ είναι να προσδιοριστούν προτάσεις που περιέχουν συγκριτικές απόψεις. Για την υλοποίηση αυτής της ανάλυσης χρησιμοποιείται Sentiment lexicon acquisition που περιγράφεται παρακάτω.

Sentiment lexicon acquisition [5]

Αυτή η μέθοδος ανάλυσης συναισθημάτων χρησιμοποιεί μια λίστα λέξεων και εκφράσεων που χρησιμοποιούνται από τους ανθρώπους για να εκφράσουν τα υποκειμενικά αισθήματα και τα συναισθήματα ή τις απόψεις τους. Οι άλλοι τύποι ανάλυσης συναισθημάτων κάνουν χρήση θετικών και αρνητικών λέξεων. Οι προτάσεις, οι οποίες δεν εκφράζουν άποψη για το αν κάτι είναι καλό ή κακό, αναλύονται χρησιμοποιώντας 3 προσεγγίσεις: χειροκίνητη προσέγγιση (Manual Approach), προσέγγιση βασισμένη σε λεξικό (Dictionary based approach) και προσέγγιση με βάση κάποια συλλογή (Corpus-based approach).

Manual Approach: Δεν χρησιμοποιείται γιατί είναι χρονοβόρα και άρα δεν είναι αποδοτική.

Dictionary based approach: Αυτή η προσέγγιση χρησιμοποιεί το "Word Net" (μία βάση δεδομένων που περιέχει αγγλικό λεξιλόγιο) για να βρει κατάλληλες λέξεις που αφορούν συναίσθημα για να πραγματοποιήσει την ανάλυση.

Corpus-based approach: Δημιουργήσει ένα λεξικό που περιέχει λέξεις που αφορούν συναίσθηματα για κάποιο συγκεκριμένο τομέα για να πραγματοποιήσει την ανάλυση.

3.1.4 Εισαγωγή στα APIs

Ένα Application Programming Interface (API) ορίζει ένα σύνολο κανόνων που επιτρέπει διαφορετικά προγράμματα να επικοινωνούν μεταξύ τους. Ουσιαστικά ορίζεται μια διεπαφή μεταξύ τους. Η λειτουργία τους είναι παρόμοια με τις διευθύνσεις του Παγκόσμιου Ιστού (web addresses), δηλαδή όταν ο χρήστης ζητά μια ενέργεια που απαιτεί δεδομένα, να παρέχονται ανάλογα με τα δικαιώματα πρόσβασης τους. Ουσιαστικά, τα APIs αποτελούν μια τεχνολογία που επιτρέπει στις εφαρμογές να “επικοινωνούν” μεταξύ τους. Τα APIs αποκαλούνται και “machine readable interfaces”. Τα APIs εξελίσσονται με ιδιαίτερα γρήγορους ρυθμούς και αφορούν κοινότητες των πελατών, των επιχειρήσεων και των προγραμματιστών. Τα API είναι πολύ χρήσιμα και σε περιπτώσεις όπου κάποια εταιρία θέλει να δημιουργήσει μια εφαρμογή για πολλές πλατφόρμες, όπως desktop software και κινητές συσκευές, αφού επιτρέπει

την χρήση κοινής βάσης δεδομένων σε όλες τις υλοποιήσεις στις διάφορες πλατφόρμες. Σε αυτή την περίπτωση με τη χρήση του API endpoint επιτρέπεται η πρόσβαση στα δεδομένα της κοινής βάσης δεδομένων.

Ένα API αποτελείται από δύο πλευρές, τον server και τον client. Ο server παρέχει το API και υποστηρίζει την διατήρηση του API ως ένα πρόγραμμα που εκτελείται στον server και αναμένει αιτήματα δεδομένων. Ο client είναι λογισμικό που γνωρίζει τα δεδομένα που διατίθενται μέσω του API και μπορεί να τα διαχειριστεί, μετά από το αίτημα του χρήστη. Τα ανοιχτά (open) API διατίθεται στο Internet και μπορεί να χρησιμοποιηθεί χωρίς κόστος. Ένα ανοιχτό API προσφέρει οφέλη και στον κάτοχο αλλά και στον χρήστη. Από την πλευρά του κατόχου, με τη χρήση του API τα προϊόντα και οι υπηρεσίες του κερδίζουν δημοτικότητα, ενώ για τον χρήστη διευκολύνουν τους developers για την ανάπτυξη λογισμικού.

Ένα χρήσιμο API για τις επιχειρήσεις είναι αυτό του Twitter. Στο Twitter υπάρχουν τρεις βασικές οντότητες: τα tweets, τους χρήστες και τα timelines. Τα tweets παρέχουν πληροφορίες όπως κείμενο, συγγραφέα και διαθέσιμα μεταδεδομένα. Για τους χρήστες υπάρχει η πληροφορία για το username, το screen name καθώς και το avatar που χρησιμοποιούν. Τα timelines παρέχουν ταξινομημένα tweets του home για έναν χρήστη ή tweets που αναφέρουν έναν χρήστη. Ακόμα, μπορούν να ανακτηθούν direct messages, friends και followers, suggested user, favorites, list of tweets και local trends. Επίσης, μπορούν να ανακτηθούν από συγκεκριμένη τοποθεσία στον κόσμο. Τέλος, μπορούν να αναζητηθούν tweets με βάση λέξεις κλειδιά.

3.1.5 Εξαγωγή πληροφορίας (information extraction)

Η εξαγωγή πληροφοριών (information extraction - IE) είναι διαδικασία της αυτόματης εξαγωγής δομημένων πληροφοριών από μη δομημένα ή και ημι-δομημένα έγγραφα, τα οποία είναι αναγνώσιμα από μια μηχανή. Στις περισσότερες περιπτώσεις, η δραστηριότητα αυτή αφορά την επεξεργασία κειμένων της ανθρώπινης γλώσσας μέσω επεξεργασίας της φυσικής γλώσσας (natural language processing - NLP). Παρόλο που τα διάφορα συστήματα IE έχουν κατασκευαστεί για διαφορετικούς σκοπούς και ενδέχεται να διαφέρουν σημαντικά μεταξύ τους, υπάρχουν ορισμένα βασικά κοινά στοιχεία. Τα ποιο συνηθισμένα από αυτά είναι τα ακόλουθα: [3] [4]

- Η προεπεξεργασία του κειμένου (pre-processing of the text) επεξεργάζεται το κείμενο με τη βοήθεια εργαλείων υπολογιστικής γλωσσολογίας (computational linguistics

tools), όπως tokanization, η διάσπαση των προτάσεων (sentence splitting), η μορφολογική ανάλυση (morphological analysis) κλπ.

- Η εύρεση και ταξινόμηση των εννοιών, ανιχνεύει και ταξινομεί αναφορές σε ανθρώπους, πράγματα, τοποθεσίες, γεγονότα και άλλους προκαθορισμένους τύπους εννοιών.
- Η ενοποίηση αφορά την παρουσίαση των εξαγόμενων δεδομένων σε πρότυπη μορφή (standard form).
- Η αφαίρεση του θορύβου (noise) περιλαμβάνει την εξάλειψη διπλών δεδομένων (duplicate data).
- Η ανάλυση μεταδεδομένων (metadata analysis) περιλαμβάνει την εξαγωγή του τίτλου, του σώματος, της δομής του σώματος (προσδιορισμός των παραγράφων) και της ημερομηνίας του εγγράφου.
- Το tokenization περιλαμβάνει την τμηματοποίηση του κειμένου σε μονάδες τύπου λέξης που ονομάζονται tokens. Στη συνέχεια γίνεται ταξινόμηση του τύπου τους, για παράδειγμα αναγνώριση λέξεων που είναι γραμμένες με κεφαλαία γράμματα, λέξεις γραμμένες με πεζά γράμματα, σημεία στίξης, αριθμοί κ.λπ.
- Η μορφολογική ανάλυση (morphological analysis) περιλαμβάνει την εξαγωγή μορφολογικών πληροφοριών από tokens που αποτελούν πιθανές μορφές λέξεων (word forms). Μερικά παραδείγματα είναι η βασική μορφή (ή το lemma), κομμάτι ομιλίας, άλλες μορφολογικές ετικέτες (morphological tags) ανάλογα με το μέρος της ομιλίας.
- Η ανίχνευση ορίων προτάσεων/εκφράσεων (sentence/utterance boundary detection) περιλαμβάνει την τμηματοποίηση κειμένου σε μια ακολουθία προτάσεων ή εκφράσεων, εκ των οποίων η κάθε μια αναπαρίσταται ως ακολουθία λεξικών στοιχείων μαζί με τα χαρακτηριστικά τους.
- Η αναγνώριση φράσης (phase recognition) περιλαμβάνει αναγνώριση τοπικών δομών μικρής κλίμακας, όπως φράσεις που περιέχουν ουσιαστικά (nouns), ομάδες ρημάτων, φράσεις προθέσεων, ακρωνύμια και συντμήσεις.
- Η συντακτική ανάλυση (syntactic analysis) περιλαμβάνει τον υπολογισμό μιας δομής εξάρτησης (parse tree) της πρότασης με βάση την ακολουθία λεξικών στοιχείων και δομών μικρής κλίμακας.
- Το πρώτο βήμα στις περισσότερες εργασίες του IE είναι να βρεθούν τα σωστά ονόματα ή ονομαστικές οντότητες που αναφέρονται σε ένα κείμενο. Ο στόχος της

αναγνώρισης της οντότητας ονόματος (named entity recognition - NER) είναι να εντοπίσει για κάθε οντότητα ονόματος που αναφέρεται στο κείμενο τις αναφορές της και να δώσει ετικέτα στον τύπο της. Αυτό που αποτελεί έναν τύπο οντότητας εξαρτάται από την εκάστοτε εφαρμογή. Γενικά, ο τύπος μιας οντότητας μπορεί να περιλαμβάνει άτομα, μέρη, οργανισμούς, αλλά και πιο συγκεκριμένες οντότητες όπως ονόματα γονιδίων και πρωτεϊνών έως και ονόματα πανεπιστημιακών μαθημάτων.

- Έχοντας εντοπίσει όλες τις αναφορές των οντοτήτων των ονομάτων σε ένα κείμενο, είναι χρήσιμο να συνδέσουμε ή να συμπλέξουμε (cluster) τις αναφορές σε σύνολα που αντιστοιχούν τις αναφορές στις αντίστοιχες οντότητες.
- Η εξόρυξη σχέσεων (relation extraction) έχει ως στόχο την εύρεση και την ταξινόμηση των σημασιολογικών σχέσεων μεταξύ της εξαγωγής σχέσεων των οντοτήτων του κειμένου. Συνήθως, τέτοιες σχέσεις είναι οι δυαδικές σχέσεις (binary relations), όπως spouse-of, child-of, employment, partwhole, membership και geospatial relations.
- Ο στόχος της εξαγωγής συμβάντων (event extraction) είναι να βρεθούν τα γεγονότα στα οποία συμμετέχουν αυτές οι οντότητες. Επιπλέον, πρέπει να εκτελεστεί συσχέτιση γεγονότων για να διαπιστώσουμε ποιες από τις αναφορές γεγονότων σε ένα κείμενο αναφέρονται στο ίδιο συμβάν.
- Για να καταλάβουμε πότε συνέβησαν τα γεγονότα σε ένα κείμενο, πρέπει να αναγνωριστούν οι χρονικές εκφράσεις, όπως είναι οι ημέρες της εβδομάδας (Παρασκευή και Πέμπτη), οι μήνες, οι διακοπές, οι προσωρινές εκφράσεις κλπ. Παραδείγματα σχετικών εκφράσεων είναι: “δύο μέρες από τώρα” ή “το επόμενο έτος” και ώρες όπως “στις 3:30 μ.μ.” ή “το μεσημέρι”.
- Πολλά κείμενα περιγράφουν επαναλαμβανόμενες στερεότυπες καταστάσεις (recurring stereotypical situations). Ο στόχος της συμπλήρωσης προτύπων (template) είναι ο εντοπισμός των καταστάσεων που προαναφέρθηκαν μέσα στα έγγραφα. Στη συνέχεια μόλις εντοπιστούν αντιστοιχούνται στο πρότυπο τα αντίστοιχα δεδομένα και αποθηκεύονται.

Η διαδικασία της εξόρυξης πληροφορίας μπορεί να είναι τελείως αυτοματοποιημένη ή μπορεί να εκτελεστεί με τη βοήθεια του ανθρώπινου παράγοντα. Γενικά, η καλύτερη μέθοδος είναι αυτή που συνδυάζει τον αυτοματισμό της διαδικασίας και την επεξεργασία των αποτελεσμάτων από τον άνθρωπο.

Με δεδομένο ένα κείμενο εισόδου ή μια συλλογή κειμένων, η αναμενόμενη έξοδος ενός συστήματος εξόρυξης πληροφορίας μπορεί να οριστεί επακριβώς. Αυτό διευκολύνει την αξιολόγηση των διαφορετικών συστημάτων εξόρυξη πληροφορίας. Πιο συγκεκριμένα, οι μετρικές (metrics) ορθότητα (precision) και ανάκλησης (recall) χρησιμοποιούνται για την αξιολόγηση τέτοιων συστημάτων. Μετράνε την αποτελεσματικότητα του συστήματος από την πλευρά του χρήστη. Για παράδειγμα, σε τι βαθμό παράγει το σύστημα την κατάλληλη έξοδο (recall) και σε τι βαθμό παράγει μόνο την κατάλληλη έξοδο (precision). Έτσι, η ορθότητα και η ανάκληση μπορούν να θεωρηθούν ως μέτρα ορθότητας και πληρότητας αντίστοιχα. Στην συνέχεια, αυτές οι μετρικές παρουσιάζονται πιο αναλυτικά (ενότητα 3.3 Μετρικές Αξιολόγησης). [3]

Η εξαγωγή πληροφοριών μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα πηγών κειμένου, όπως emails, web pages, αναφορές, παρουσιάσεις, νομικά έγγραφα και επιστημονικές εργασίες. Μερικοί ακόμα τομείς που μπορεί να εφαρμοστεί με επιτυχία είναι η επιχειρηματική ευφυΐα που επιτρέπει στους αναλυτές να συλλέγουν δομημένες πληροφορίες από πολλαπλές πηγές, η χρηματοοικονομική έρευνα για ανάλυση και ανακάλυψη κρυφών σχέσεων, η επιστημονική έρευνα για ανακάλυψη αυτοματοποιημένων αναφορών ή πρόταση σχετικών εγγράφων, η παρακολούθηση των μέσων ενημέρωσης για αναφορές επιχειρήσεων, εμπορικών σημάτων και ανθρώπων. Επίσης, μπορεί να εφαρμοστεί με επιτυχία και σε διαχείριση αρχείων ιατρικής περίθαλψης για τη διάρθρωση και την περίληψη των αρχείων των ασθενών και για φαρμακευτική έρευνα για ανακάλυψη φαρμάκων, ανακάλυψη ανεπιθύμητων ενεργειών και αυτοματοποιημένη ανάλυση κλινικών δοκιμών. [5]

Υπάρχουν διάφορες προσεγγίσεις και υλοποίησεις για το πρόβλημα της εξόρυξης δεδομένων. Δύο γνωστά μοντέλα είναι ο Word2vec, που έχει διάφορα μοντέλα υλοποίησης, και ο tf-idf.

Word2vec

To Word2vec είναι μια ομάδα σχετικών μοντέλων που χρησιμοποιούνται για την παραγωγή ενσωματώσεων λέξεων (word embeddings), με άλλα λόγια συμβάλλει στην εξαγωγή σχέσεων μεταξύ μιας λέξης και των συμφραζόμενων λέξεων της. Αυτά τα μοντέλα είναι ρηχά νευρωνικά δίκτυα με δύο επίπεδα που εκπαιδεύονται για να ανοικοδομήσουν τα γλωσσικά πλαίσια των λέξεων. To Word2vec παίρνει ως είσοδο ένα μεγάλο κορμό του κειμένου και παράγει ένα διανυσματικό χώρο, συνήθως από αρκετές εκατοντάδες διαστάσεις, με κάθε μονα-

δική λέξη στο σώμα να αντιστοιχεί σε ένα αντίστοιχο διάνυσμα (vector) στον χώρο. Τα διανύσματα λέξεων τοποθετούνται στον διανυσματικό χώρο έτσι ώστε οι λέξεις που μοιράζονται κοινά συμφραζόμενα στη συλλογή να βρίσκονται κοντά μεταξύ τους στο χώρο. [7] Τα δύο πιο σημαντικά μοντέλα του Word2Vec είναι: Skip-grams και CBOW (Continuous Bag of Words model), τα οποία περιγράφονται παρακάτω.

Skip-grams [8]

Στο μοντέλο Skip-gram, παίρνουμε μια κεντρική λέξη και ένα “παράθυρο” συμφραζόμενων (context) (γειτονικών) λέξεων και προσπαθούμε να προβλέψουμε τις συμφραζόμενες (context) λέξεις σε κάποιο μέγεθος παραθύρου για κάθε κεντρική λέξη. Έτσι, το μοντέλο πρόκειται να ορίσει μια πιθανότητα κατανομής, δηλαδή θα οριστεί η πιθανότητα μιας λέξης που εμφανίζεται στα συμφραζόμενα (context) μιας κεντρικής λέξης, έτσι ώστε να μεγιστοποιείται η πιθανότητα με την κατάλληλη επιλογή των αναπαραστάσεων του διανύσματος. Ένα παράδειγμα του μοντέλου Skip-gram απεικονίζεται στην εικόνα 3.1.

Source Text	Training Samples
The quick brown fox jumps over the lazy dog. ➔	(the, quick) (the, brown)
The quick brown fox jumps over the lazy dog. ➔	(quick, the) (quick, brown) (quick, fox)
The quick brown fox jumps over the lazy dog. ➔	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown fox jumps over the lazy dog. ➔	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

Εικόνα 3.1 Παράδειγμα μοντέλου Skip-grams

Continuous Bag of Words model (CBOW) [8]

Ουσιαστικά πρόκειται για το αντίθετο μοντέλο του skip-gram. Στο CBOW, προσπαθούμε να προβλέψουμε την κεντρική λέξη αθροίζοντας διανύσματα των γύρω λέξεων. Σε αυτό το μοντέλο οι λέξεις μετατρέπονται σε διανύσματα. Η διαδικασία ξεκινάει με τυχαία αρχικοποίηση των διανυσμάτων των λέξεων. Το προγνωστικό μοντέλο “μαθαίνει” τα διανύσματα ελα-

χιστοποιώντας τη συνάρτηση απώλειας (loss function). Στο Word2vec, αυτό συμβαίνει με ένα νευρωνικό δίκτυο με προς τα εμπρός τροφοδότηση (feed forward) και με τεχνικές βελτιστοποίησης όπως η Stochastic gradient descent. Υπάρχουν επίσης μοντέλα βασισμένα σε μετρήσεις τα οποία δημιουργούν έναν πίνακα μετρήσεων συνεμφάνισης (co-occurrence) λέξεων που υπάρχουν στη συλλογή. Δημιουργείται ένα μεγάλος πίνακας, όπου κάθε σειρά του χρησιμοποιείται για τις "λέξεις" και κάθε στήλη για τα "συμφραζόμενα". Το πλήθος των "συμφραζόμενων" είναι μεγάλο, δεδομένου ότι είναι ουσιαστικά συνδυαστικό σε μέγεθος. Για την επιδιόρθωση αυτού του ζητήματος του μεγέθους, εφαρμόζουμε SVD στον πίνακα. Αυτό μειώνει τις διαστάσεις του πίνακα διατηρώντας το μέγιστο δυνατό ποσοστό πληροφοριών. Παράδειγμα μοντέλου CBOW απεικονίζεται στην εικόνα 3.2.

1. I enjoy flying.
2. I like NLP.
3. I like deep learning.

The resulting counts matrix will then be:

$$X = \begin{matrix} & \begin{matrix} I & like & enjoy & deep & learning & NLP & flying & . \end{matrix} \\ \begin{matrix} I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{matrix} & \left[\begin{matrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{matrix} \right] \end{matrix}$$

Εικόνα 3.2 Παράδειγμα μοντέλου CBOW

TF-IDF

To Tf-idf σημαίνει term frequency - inverse document frequency και το βάρος του tf-idf είναι ένα βάρος που χρησιμοποιείται συχνά στην ανάκτηση πληροφοριών και την εξόρυξη κειμένου. Αυτό το βάρος είναι ένα στατιστικό μέτρο που χρησιμοποιείται για την αξιολόγηση της σπουδαιότητας μιας λέξης σε ένα έγγραφο μέσα σε μια συλλογή. Η σπουδαιότητα αυξάνεται ανάλογα με το πλήθος που εμφανίζεται μια λέξη στο έγγραφο, αλλά αντισταθμίζεται από τη συχνότητα της λέξης στο σώμα. Οι μηχανές αναζήτησης χρησιμοποιούν συχνά παραλλαγές του βάρους του tf-idf ως κεντρικό εργαλείο για τη βαθμολόγηση και την ταξινόμη-

ση της σχετικότητας ενός εγγράφου με δεδομένο ένα ερώτημα χρήστη. Το Tf-idf μπορεί να χρησιμοποιηθεί με επιτυχία για φιλτράρισμα stop-words σε διάφορα θεματικά πεδία, όπως η περίληψη κειμένου και η ταξινόμηση. Το βάρος tf-idf αποτελείται από δύο όρους. [9]

- Ο πρώτος όρος μετράει το πόσο συχνά ένας όρος εμφανίζεται σε ένα έγγραφο και υπολογίζει την κανονικοποιημένη συχνότητα των όρων (term frequency - TF), δηλαδή το πλήθος που εμφανίζεται μια λέξη σε ένα έγγραφο, διαιρούμενο με το συνολικό αριθμό των λέξεων σε αυτό το έγγραφο. Δεδομένου ότι κάθε έγγραφο έχει διαφορετικό μέγεθος, είναι πιθανό ότι ένας όρος μπορεί να εμφανιστεί πολύ περισσότερες φορές σε μεγάλα έγγραφα από ότι σε πιο σύντομα. Έτσι, ο όρος συχνότητα συχνά διαιρείται με το μέγεθος του εγγράφου (που είναι το συνολικό πλήθος των όρων στο έγγραφο) για κανονικοποίηση.

$$TF(t) = \frac{\text{Number Of Times Term } t \text{ Appears In A Document}}{\text{Total Number Of Terms In The Document}} \quad [9]$$

- Ο δεύτερος όρος είναι η αντίστροφη συχνότητα εγγράφων (inverse document frequency – IDF) και μετράει το πόσο σημαντικός είναι ένας όρος. Κατά τον υπολογισμό TF, όλοι οι όροι θεωρούνται εξίσου σημαντικοί. Ωστόσο, είναι γνωστό ότι ορισμένοι όροι, όπως "είναι", "από" και "ότι", μπορεί να εμφανίζονται πολλές φορές αλλά έχουν μικρή σημασία. Επομένως, πρέπει να σταθμίσουμε τους συχνούς όρους και να αυξήσουμε τους σπάνιους. Υπολογίζεται ως ο λογάριθμος του πλήθους των εγγράφων στη συλλογή διαιρούμενο με τον αριθμό των εγγράφων στα οποία εμφανίζεται ο συγκεκριμένος όρος.

$$IDF(t) = \ln\left(\frac{\text{Total Number Of Documents}}{\text{Number Of Documents With Term } t \text{ In It}}\right) \quad [9]$$

Στη συνέχεια συνδυάζουμε τους ορισμούς της συχνότητας όρων και της αντίστροφης συχνότητας εγγράφων, για να παράγουμε ένα σύνθετο βάρος για κάθε όρο σε κάθε έγγραφο. Το σχήμα βάρους του tf-idf αποδίδει στον όρο t ένα βάρος σε ένα έγγραφο d που δίνεται από τον τύπο:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

Με άλλα λόγια, το $tf - idf_{t,d}$ αντιστοιχεί τον όρο t σε ένα βάρος στο έγγραφο d που είναι:

1. υψηλότερο, όταν ο όρος t εμφανίζεται πολλές φορές σε μικρό πλήθος εγγράφων δινοντας έτσι μεγάλο βαθμό διακριτικής ισχύς σε αυτά τα έγγραφα)

2. χαμηλότερο, όταν ο όρος εμφανίζεται λιγότερες φορές σε ένα έγγραφο ή εμφανίζεται σε πολλά έγγραφα (προσφέροντας έτσι ένα λιγότερο έντονο σήμα ομοιότητας).
3. όσο το δυνατόν πιο χαμηλό, όταν ο όρος εμφανίζεται σχεδόν σε όλα τα έγγραφα. [9]

3.2 Μηχανική Μάθηση

Στην σημερινή εποχή η Μηχανική Μάθηση γνωρίζει μεγάλη ανάπτυξη και καθημερινά χρησιμοποιείται σε μεγάλο αριθμό εφαρμογών. Το εύρος αυτών των εφαρμογών εκτείνεται από προγράμματα εξόρυξης δεδομένων, που ανακαλύπτουν γενικούς κανόνες σε μεγάλα σετ δεδομένων, έως και σε συστήματα που φιλτράρουν πληροφορία ώστε να μάθουν τα ενδιαφέροντα των χρηστών. Επίσης, η Μηχανική Μάθηση έχει εφαρμοστεί σε άλλους τομείς και έχει ως αποτέλεσμα τη δημιουργία αυτοκινήτων που μπορούν να πλοηγηθούν στον χώρο χωρίς ανθρώπινη επίβλεψη, αποτελεσματική αναγνώριση φωνής, αποδοτική αναζήτηση στο διαδίκτυο κ.α.

3.2.1 Ορισμός της Μηχανικής Μάθησης

Η Μηχανική Μάθηση (Machine Learning) έχει οριστεί με διάφορους τρόπους. Μηχανική Μάθηση είναι η μελέτη αλγορίθμων υπολογιστών που μπορούν να βελτιωθούν μόνοι τους μέσω εμπειρίας που αποκτούν. Επίσης, Μηχανική Μάθηση είναι η επιστήμη που προγραμματίζει υπολογιστές οι οποίοι είναι ικανοί να εκτελούν ενέργειες μόνοι τους χωρίς να έχουν προγραμματιστεί ρητά.

Μερικοί ακόμα ορισμοί: [10]

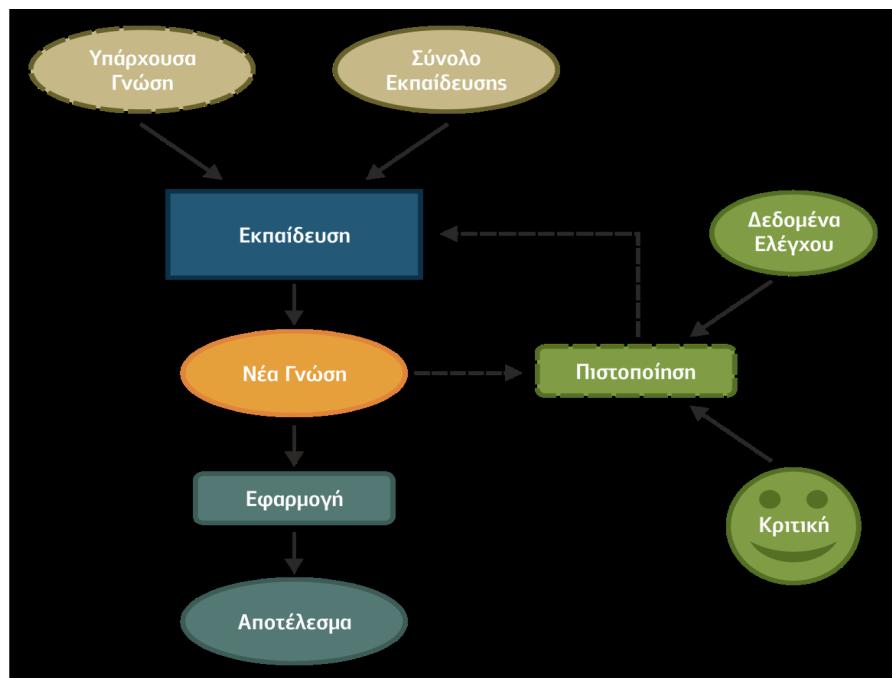
- Carbonell (1987), "... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης".
- Mitchell (1997), "Ενα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E".
- Witten & Frank (2000), "Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον".

Γενικά, σε ένα γνωστικό σύστημα η έννοια της μάθησης προσδιορίζεται από δύο κύριες ιδιότητες. Το σύστημα πρέπει να είναι ικανό να αποκτάει επιπλέον γνώση μέσω της αλληλε-

πίδρασης του με το περιβάλλον που δραστηριοποιείται. Επίσης, πρέπει να είναι ικανό να βελτιώνει τον τρόπο που εκτελεί μια ενέργεια μέσω της επανειλημμένης εκτέλεσης της.

Οι αλγόριθμοι της Μηχανικής Μάθησης έχουν κάποιες κοινές φάσεις που μας επιτρέπουν να θεωρήσουμε ένα κοινό γενικό τρόπο λειτουργίας. Ένα παράδειγμα γενικού τρόπου λειτουργίας απεικονίζεται στην εικόνα 3.3. Η κυριότερη φάση κάθε αλγορίθμου είναι η εκπαίδευση. Σε αυτή τη φάση ο αλγόριθμος δέχεται ως είσοδο ένα σύνολο δεδομένων εκπαίδευσης (training set) για τη δημιουργία νέας γνώσης. Αξίζει να σημειωθεί ότι υπάρχουν περιπτώσεις όπου δεν είναι απαραίτητο να χρησιμοποιηθεί όλη η προ-υπάρχουσα γνώση. Επίσης, σε άλλες περιπτώσεις η προ-υπάρχουσα γνώση δεν είναι αρκετή ή ακόμα υπάρχουν περιπτώσεις όπου δεν χρειάζεται να χρησιμοποιηθεί καθόλου προ-υπάρχουσα γνώση.

Στη συνέχεια, ακολουθεί η φάση της πιστοποίησης, η οποία πραγματοποιείται συνήθως από τον ίδιο τον αλγόριθμο κάνοντας χρήση διαδικασιών ανάκλησης (recall) με τη βοήθεια δεδομένων ελέγχου (test data). Στο τέλος της διαδικασίας της πιστοποίησης ο χρήστης ελέγχει το αποτέλεσμα με τις γνώσεις που διαθέτει περί του προβλήματος που προσπαθεί να λύσει ο αλγόριθμος. Τέλος, η νέα παραγόμενη γνώση τροφοδοτείται στις εφαρμογές που την χρειάζονται, ώστε να επιλυθούν πραγματικά προβλήματα. [11]



Εικόνα 3.3 Γενικός τρόπος λειτουργίας αλγορίθμων Μηχανικής Μάθησης

3.2.2 Είδη Μηχανικής Μάθησης

Στα πλαίσια της Μηχανικής Μάθησης έχουν αναπτυχθεί πολλές τεχνικές που χρησιμοποιούνται ανάλογα με το είδος του προβλήματος που καλούνται να επιλύσουν. Αυτές οι τεχνικές ανήκουν σε τρεις κατηγορίες, την μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples), την μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation) και την ενισχυτική μάθηση (reinforcement learning).

A) Μάθηση με Επίβλεψη

Στη μάθηση με επίβλεψη (supervised learning) το σύστημα χρησιμοποιεί ένα σύνολο δεδομένων με σκοπό να μάθει επαγωγικά μια συνάρτηση. Αυτή η συνάρτηση ονομάζεται συνάρτηση στόχος (target function) και αποτελεί στοιχείο του μοντέλου που περιγράφει τα δεδομένα. Το σύνολο δεδομένων ονομάζεται δεδομένα εκμάθησης ή ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου ή χαρακτηριστικά και αποτελούν δεδομένα που έχουν γνωστή είσοδο αλλά και γνωστή έξοδο. Συνεπώς, η μέθοδος έχει σκοπό να χρησιμοποιήσει τα δεδομένα εκμάθησης ώστε να γενικεύσει την συνάρτηση στόχο για εισόδους με άγνωστη έξοδο, την οποία θα προβλέψει. Επίσης, η επαγωγική μάθηση χρησιμοποιεί την υπόθεση επαγωγικής μάθησης (inductive learning hypothesis). Η υπόθεση της επαγωγικής μάθησης υποστηρίζει ότι αν μια υπόθεση προσεγγίζει καλά τη συνάρτηση στόχο για ένα επαρκή σύνολο παραδειγμάτων, τότε θα προσεγγίζει εξίσου καλά τη συνάρτηση στόχο για περιπτώσεις που δεν έχουν εξεταστεί ακόμα. [10][11]

Στην τεχνική της μάθησης με επίβλεψη τα προβλήματα κατηγοριοποιούνται σε δύο κατηγορίες προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης (classification) και τα προβλήματα παρεμβολής (regression). Στα προβλήματα ταξινόμησης δημιουργούνται μοντέλα πρόβλεψης διακριτών τάξεων (κλάσεων ή κατηγοριών). Ένα τέτοιο παράδειγμα προβλήματος είναι η ομάδα αίματος. Στα προβλήματα παρεμβολής δημιουργούνται μοντέλα πρόβλεψης αριθμητικών τιμών. Ένα τέτοιο παράδειγμα προβλήματος είναι η πρόβλεψη ισοτιμίας νομίσματος ή η πρόβλεψη της τιμής μιας μετοχής. [10][11]

B) Μάθηση Χωρίς Επίβλεψη

Στη μάθηση χωρίς επίβλεψη (unsupervised learning) το σύστημα πρέπει να δημιουργήσει μόνο του συσχετίσεις ή ομάδες από ένα σύνολο δεδομένων που δέχεται ως είσοδο και βασι-

ζόμενο μόνο στις ιδιότητες τους να ανακαλύψει πρότυπα (περιγραφές) χωρίς να γνωρίζει τις επιθυμητές εξόδους. Τα πρότυπα περιγράφουν ένα κομμάτι των δεδομένων και το σύστημα δεν γνωρίζει εκ των προτέρων πόσα πρότυπα υπάρχουν, αν υπάρχουν, και ποια είναι. Παραδείγματα προτύπων είναι οι κανόνες συσχέτισης (association rules) και οι ομάδες (clusters) που είναι προϊόν της διαδικασίας της ομαδοποίησης (clustering). Από τα παραπάνω πρότυπα προκύπτουν δύο αντίστοιχες κατηγορίες προβλημάτων, τα προβλήματα ανάλυσης συσχετισμών (association analysis) και τα προβλήματα ομαδοποίησης (clustering). [10][11]

Γ) Ενισχυτική Μάθηση

Στην ενισχυτική μάθηση (reinforcement learning) ο αλγόριθμος μέσω της αλληλεπίδρασης του με το περιβάλλον μαθαίνει μια στρατηγική ενεργειών. Πιο συγκεκριμένα, το σύστημα πρέπει μόνο του, με την αλληλεπίδραση του με το περιβάλλον, να ανακαλύψει ποιες ενέργειες θα οδηγήσουν στην μεγιστοποίηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή). Αυτή η τεχνική χρησιμοποιείται κατά κύριο λόγο σε προβλήματα σχεδιασμού (planning). Τέτοια προβλήματα είναι η κίνηση ρομπότ, η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους, μάθηση επιτραπέζιων παιχνιδιών κ.α.

Στο βασικό πλαίσιο της ενισχυτικής μάθησης, η οντότητα που μαθαίνει και παίρνει αποφάσεις ονομάζεται πράκτορας (agent), ενώ οτιδήποτε άλλο εκτός του πράκτορα ονομάζεται περιβάλλον. Ο πράκτορας και το περιβάλλον αλληλεπιδρούν σε μια ακολουθία διακριτών χρονικών στιγμών, όπου ο πράκτορας εκτελεί ενέργειες και το περιβάλλον αποκρίνεται σε αυτές παρουσιάζοντας στον πράκτορα καινούργιες καταστάσεις και προσφέροντας του ειδικές αριθμητικές τιμές που ονομάζονται ανταμοιβές (rewards). Ο πράκτορας προσπαθεί σε βάθος χρόνο να μεγιστοποιήσει τις ανταμοιβές που λαμβάνει από το περιβάλλον. [10][11]

3.2.3 Αλγόριθμοι Μηχανικής Μάθησης

Γενικά, υπάρχουν πολλοί αλγόριθμοι Μηχανικής Μάθησης που ανήκουν στις κατηγορίες που προαναφέρθηκαν παραπάνω. Κάθε κατηγορία αλγορίθμων είναι ικανά να επιλύσουν συγκεκριμένα είδη προβλημάτων. Παρακάτω θα αναφερθούν τρία είδη αλγορίθμων, τα Νευρωνικά Δίκτυα (neural networks), οι Μηχανές Διανυσμάτων Υποστήριξης (MΔΥ) ή Support Vector Machines (SVMs) και οι αλγόριθμοι που στηρίζονται στην πιθανοτική θεωρία του Bayes (Bayesian probability).

A) Νευρωνικά Δίκτυα (Neural Networks)

Ο όρος Νευρωνικά Δίκτυα (Neural Networks, Connectionist Networks, Parallel Distributed Processing Models) περιγράφει ένα σύνολο διαφορετικών μαθηματικών μοντέλων, εμπνευσμένα από αντίστοιχα βιολογικά μοντέλα, δηλαδή μοντέλα που προσπαθούν να προσεγγίσουν τη συμπεριφορά των νευρώνων του ανθρώπινου εγκεφάλου. [11] Ένα Τεχνητό Νευρωνικό Δίκτυο μοιάζει με ένα φυσικό ως προς τον τρόπο απόκτησης γνώσης που γίνεται μέσα από τη διαδικασία της μάθησης, καθώς και για την αποθήκευση της γνώσης χρησιμοποιούνται οι δυνάμεις σύνδεσης των νευρώνων, που ονομάζονται συναπτικά (synaptic) βάρη. Τα Νευρωνικά Δίκτυα εκπαιδεύονται με τη βοήθεια παραδειγμάτων ώστε να μαθαίνουν από το περιβάλλον τους. Υπάρχουν πολλές κατηγορίες που διαφοροποιούνται από την αρχιτεκτονική τους αλλά και τον τρόπο εκπαίδευσης.

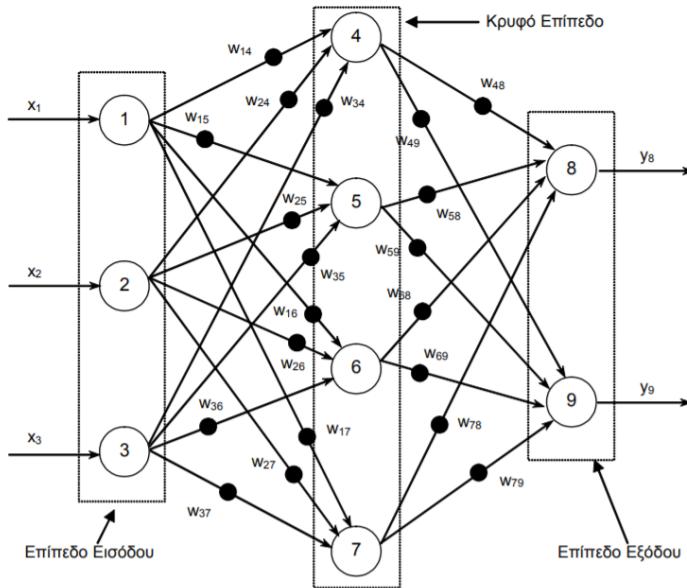
Συνοπτικά, ένα Νευρωνικό Δίκτυο οργανώνεται σε επίπεδα (layers), τα οποία ονομάζονται και στρώματα. Τα ενδιάμεσα επίπεδα ονομάζονται κρυμμένα επίπεδα (hidden layers) και δεν είναι απαραίτητο να υπάρχουν. Κάθε επίπεδο αποτελείται από μονάδες (units) ή κόμβους (nodes) που είναι συνδεδεμένα μεταξύ τους, ώστε κάθε μία μονάδα να έχει συνδέσμους με πολλές άλλες μονάδες του ίδιου ή άλλου επιπέδου. Ένας κόμβος μπορεί να διεγείρει ή να προκαλέσει την αναστολή ενεργοποίησης ενός άλλου κόμβου με την παραγωγή κατάλληλης εξόδου. Για την επίτευξη αυτού, ο κόμβος χρησιμοποιεί το σταθμισμένο άθροισμα όλων των εισόδων μέσω των συνδέσμων που καταλήγουν σε αυτόν και παράγει μια μοναδική έξοδο μέσω της συνάρτησης μετάβασης (transfer function), εάν το ζυγισμένο άθροισμα των εισόδων είναι μεγαλύτερο μιας ορισμένης τιμής κατωφλίου (threshold value) θ , δηλαδή όταν:

$$\sum_{i=1}^n x_i w_i - \theta > 0$$

Το επίπεδο εισόδου (input layer) εισάγει τις εισόδους στο δίκτυο και επικοινωνεί με ένα ή περισσότερα κρυμμένα επίπεδα. Το επίπεδο εξόδου (output layer) συνδέει τα κρυμμένα επίπεδα και παράγει την έξοδο. [11][12] Ένα παράδειγμα των στρωμάτων ή επιπέδων ενός Νευρωνικού Δικτύου απεικονίζεται στην εικόνα 3.4.

Γενικά, τα Νευρωνικά Δίκτυα προσφέρουν έναν πρακτικό και εύκολο τρόπο για την εκμάθηση αριθμητικών και διανυσματικών συναρτήσεων ορισμένων σε συνεχή ή διακριτά μεγέθη. Στα θετικά, έχουν ανοχή στον θόρυβο που υπάρχει στα δεδομένα εκπαίδευσης, δηλαδή δεδομένα που περιέχουν κάποιες λανθασμένες τιμές (π.χ. λάθη καταχώρησης). Χρησιμο-

ποιούνται για γραμμική και μη γραμμική παρεμβολή, αλλά και για ταξινόμηση. Στα αρνητικά, δεν έχουν την ικανότητα να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν. [12]



Εικόνα 3.4 Παράδειγμα στρωμάτων ή επιπέδων Νευρωνικού Δικτύου

Τα Τεχνητά Νευρωνικά Δίκτυα αποτελούνται από μια συλλογή νευρώνων (Processing Units - PUs) που είναι συνδεδεμένοι μεταξύ τους. Κάθε νευρώνας αποτελείται από ένα σύνολο εισόδων αλλά μόνο μία έξοδο, η οποία χρησιμοποιείται ως είσοδος σε άλλους νευρώνες. Οι συνδέσεις διαφοροποιούνται ως προς τη σημαντικότητα τους και για τον προσδιορισμό αυτής της διαφοροποίησης χρησιμοποιείται ένα συντελεστής βάρους, η σύναψη. Το επεξεργαστικό κομμάτι των νευρώνων καθορίζεται από τη συνάρτηση μεταφοράς, η οποία ορίζει την κάθε έξοδο σε σχέση με τις εισόδους και τους συντελεστές βάρους. [12]

Κατά τη δημιουργία ενός Τεχνικού Νευρωνικού Δικτύου πρέπει να καθοριστούν ένα σύνολο από βασικά στοιχεία. Πρέπει να οριστεί το πλήθος των ενδιάμεσων κρυφών επιπέδων, το πλήθος των κόμβων ανά επίπεδο, το πως συνδέονται οι κόμβοι μεταξύ τους. Επίσης, πρέπει να οριστεί η τιμή ενεργοποίησης ή κατωφλίου, η μορφή της συνάρτησης μετάβασης, οι τιμές των αρχικών βαρών μεταξύ των κόμβων και οι αλγόριθμοι (κανόνες εκπαίδευσης) που χρησιμοποιούνται ώστε να ενισχυθούν οι σύνδεσμοι μεταξύ των κόμβων κατά τη διαδικασία της εκπαίδευσης. [12]

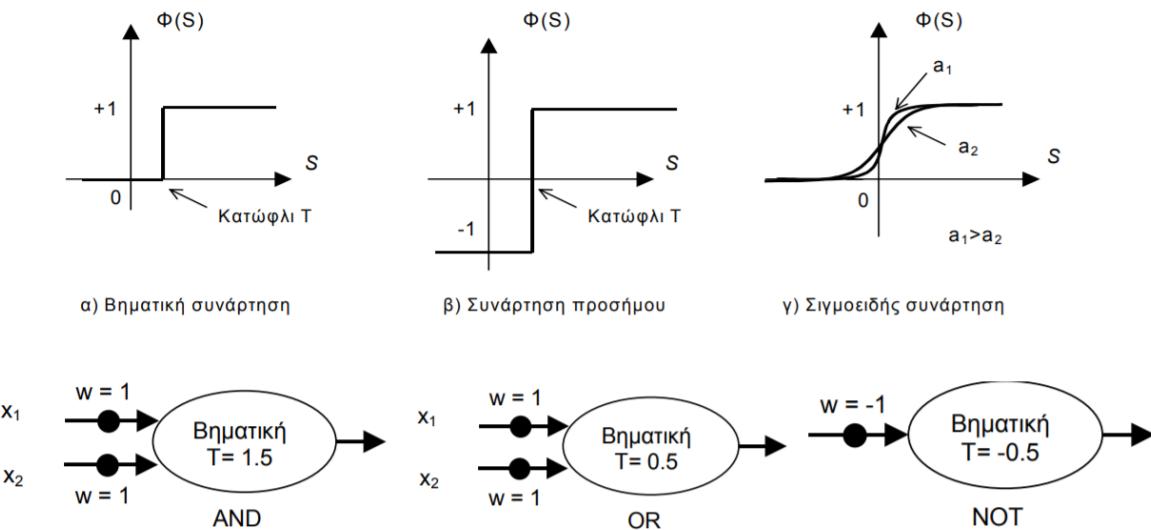
Ένας τεχνητός νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας και το μοντέλο του αποτελείται από τρία βασικά χαρακτηριστικά. Αυτά τα χαρακτηριστικά είναι: ένα σύνολο

από συνάψεις ή συνδετικούς κρίκους, ένας αθροιστής και μια συνάρτηση ενεργοποίησης. Δηλαδή, ένας τεχνητός νευρώνας αποτελείται από πολλές εισόδους x_i και μία μόνο έξοδο y . Σε κάθε είσοδο x_i προσδίδεται ένα βάρος w_i και τα αποτελέσματα αθροίζονται μέσω της συνάρτησης αθροίσματος (summation function). Η συνάρτηση ενεργοποίησης ή κατωφλίου (activation ή threshold function) είναι ένα μη γραμμικό φίλτρο που διαμορφώνει το σήμα εξόδου y , σε συνάρτηση με το ζυγισμένο άθροισμα των εισόδων την ποσότητα S . [12]

Η συνάρτηση ενεργοποίησης μπορεί να είναι βηματική (step transfer function), γραμμική (linear transfer function), μη γραμμική (non-linear transfer function) ή στοχαστική (stochastic transfer function). Παραδείγματα συναρτήσεων ενεργοποίησης απεικονίζονται στην εικόνα 3.5. Η βηματική μπορεί να είναι με ή χωρίς μετατόπιση. [12][13] Χωρίς μετατόπιση έχει τη μορφή:

$$\phi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Συναρτήσεις Ενεργοποίησης



Εικόνα 3.5 Συναρτήσεις ενεργοποίησης

Η βηματική συνάρτηση δεν θεωρείται χρήσιμη ως συνάρτηση ενεργοποίησης, αφού από τον απειροστικό λογισμό έχει το κύριο μειονέκτημα ότι η παράγωγος της απειρίζεται. Γενικά, θέλουμε η συνάρτηση ενεργοποίησης να έχει γραφική παράσταση παρόμοια της βηματικής αλλά να είναι συνεχείς και παραγωγίσιμες σε όλο το πεδίο ορισμού. Ένα παράδειγμα τέτοιας

συνάρτησης είναι η σιγμοειδής. Η γραμμική συνάρτηση ενεργοποίησης μπορεί να είναι οποιαδήποτε γραμμική συνάρτηση. Ένα παράδειγμα γραμμικής συνάρτησης: [12][13]

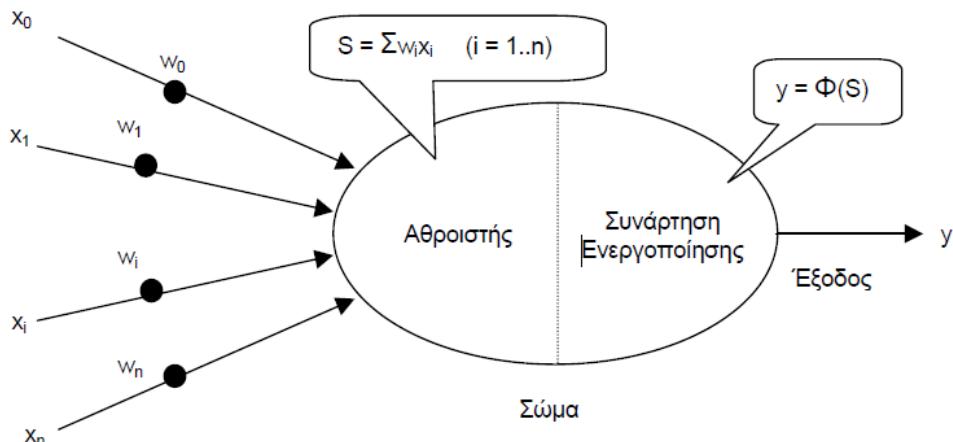
$$\varphi(x) = x$$

Οι πιο συνηθισμένη μη γραμμική συνάρτηση ενεργοποίησης που χρησιμοποιείται στα Νευρωνικά Δίκτυα είναι η σιγμοειδής συνάρτηση. Οι τυπικές σιγμοειδείς είναι δύο: [12][13]

$$\text{Λογιστική σιγμοειδής: } \varphi(x) = \frac{1}{1+e^x} \quad \text{Υπερβολική εφαπτομένη: } \varphi(x) = \tanh(x)$$

Η πιο απλή μορφή ενός Νευρωνικού Δικτύου αποτελείται από έναν μόνο νευρώνα και ονομάζεται στοιχειώδης Perceptron (basic Perceptron). Δεδομένου ενός διανύσματος εισόδου $x = (x_1, x_2, \dots, x_n)$ και μιας συνάρτησης μετάβασης g , η έξοδος α του Perceptron δίνεται από τη σχέση (σχηματική αναπαράσταση στην εικόνα 3.6):

$$\alpha = g\left(\sum_{i=1}^n x_i w_i\right)$$



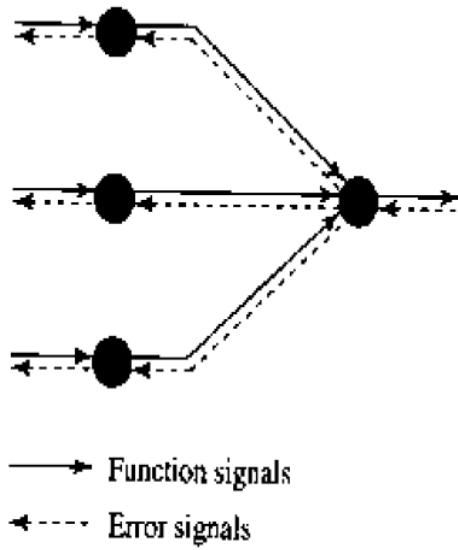
Εικόνα 3.6 Μοντέλο Τεχνητού Νευρωνικού Δικτύου

Ένα Νευρωνικό Δίκτυο μπορεί να χαρακτηριστεί ως κατευθυνόμενος γράφος, που αποτελείται από κόμβους με συναπτικές διασυνδέσεις καθώς και συνδέσεις ενεργοποίησης. Κάθε νευρώνας προσομοιώνεται από ένα σύνολο γραμμικών συναπτικών συνδέσεων, από ένα εξωτερικά εφαρμοζόμενο κατώφλι καθώς και από μια μη γραμμική σύνδεση ενεργοποίησης. Για την αναπαράσταση του κατωφλίου χρησιμοποιούνται συναπτικές συνδέσεις που τους ορίζεται τιμή σήματος εισόδου -1. Στη συνέχεια, ζυγίζονται τα αντίστοιχα σήματα εισόδου ενός νευρώνα από τις συναπτικές συνδέσεις και το συνολικό εσωτερικό επίπεδο ενεργοποίησης

του νευρώνα που ζητείται καθορίζεται από το άθροισμα των βαρών των σημάτων εισόδου καθορίζει. Το εσωτερικό επίπεδο ενεργοποίησης περιορίζεται από την σύνδεση ενεργοποίησης κατά την διαδικασία της παραγωγής της εξόδου που παριστάνει την κατάσταση του νευρώνα. [12]

Τα Νευρωνικά Δίκτυα πολλαπλών επιπέδων εμπρός τροφοδότησης, που ονομάζονται και Perceptrons πολλών επιπέδων (MLPs), αποτελούνται από το επίπεδο εισόδου, που αποτελείται από ένα σύνολο αισθητήρων (πηγαίοι κόμβοι). Επίσης, αποτελούνται από ένα ή περισσότερα κρυφά επίπεδα των οποίων οι κόμβοι υπολογισμού ονομάζονται “κρυφοί νευρώνες” (hidden layers) και τέλος αποτελούνται από ένα επίπεδο υπολογιστικών κόμβων εξόδου.

Η διάδοση του σήματος εισόδου μέσα στο δίκτυο γίνεται με προς τα εμπρός κατεύθυνση από επίπεδο σε επίπεδο. Από πλευρά αρχιτεκτονικής, στο ίδιο επίπεδο δεν υπάρχουν συνδέσεις καθώς δεν υπάρχουν και απευθείας συνδέσεις μεταξύ εισόδου και εξόδου. Μεταξύ επιπέδων το δίκτυο είναι πλήρως συνδεδεμένο. Ακόμα, δεν υπάρχει εξάρτηση μεταξύ του πλήθους των εξόδων και του πλήθους των εισόδων, ενώ ανά επίπεδο το πλήθος των κόμβων είναι ανεξάρτητο. Οι νευρώνες κάθε επιπέδου δέχονται ως είσοδο μόνο σήματα εξόδων του προηγούμενου επιπέδου. Κάθε μονάδα είναι ένα Perceptron. Επιπλέον, υπάρχουν και τα αναδρομικά δίκτυα τα οποία είναι υλοποιημένα με τουλάχιστον έναν βρόγχο ανάδρασης. [12]



Εικόνα 3.7 Δίκτυο εμπρός τροφοδότησης πολλών επιπέδων με λειτουργικά σήματα και σήματα λάθους

Ένα σήμα λάθους (error signal) δημιουργείται σε έναν νευρώνα εξόδου του δικτύου και διαδίδεται προς τα πίσω μέσω του δικτύου. Η ονομασία του (error signal) προκύπτει επειδή

για να υπολογιστεί από τους νευρώνες του δικτύου πρέπει να χρησιμοποιηθεί μια συνάρτηση εξαρτώμενη από το λάθος. Ένα παράδειγμα δικτύου που χρησιμοποιούνται δύο ειδών σήματα, τα λειτουργικά (function signals) και τα σήματα λάθους (error signals), απεικονίζεται στην εικόνα 3.7. Ένα λειτουργικό σήμα είναι ένα σήμα εισόδου που διαδίδεται προς τα εμπρός μέσω του δικτύου σαν σήμα εξόδου. Ονομάζεται έτσι γιατί επιτελεί μια χρήσιμη συνάρτηση στην έξοδο του δικτύου, καθώς και ο υπολογισμός του γίνεται σε συνάρτηση με τις εισόδους και τα συσχετιζόμενα βάρη που εφαρμόζονται στο νευρώνα από τον οποίο διέρχεται το λειτουργικό σήμα. [12]

Η εκπαίδευση των MLPs γίνεται με έναν επιβλεπόμενο τρόπο που υλοποιείται με τον αλγόριθμο πίσω διάδοσης του λάθους (Error Back Propagation Algorithm – BP). Ο αλγόριθμος βασίζεται στον κανόνα διόρθωσης του λάθους (error correction learning rule). Η πίσω διάδοση του λάθους υλοποιείται με δύο περάσματα μέσω διαφορετικών επιπέδων του δικτύου, ένα προς τα εμπρός (forward pass) και ένα προς τα πίσω (backward pass).

Στο εμπρός πέρασμα γίνεται η εφαρμογή ενός διανύσματος εισόδου (input vector) στους νευρώνες εισόδου του δικτύου. Η διάδοση της επίδρασης του γίνεται μέσα στο δίκτυο από επίπεδο σε επίπεδο, ενώ η πραγματική απόκριση του δικτύου αποτελείται από ένα σύνολο παραγόμενων εξόδων. Αξίζει να σημειωθεί ότι κατά τη διάρκεια του εμπρός περάσματος τα βάρη του δικτύου είναι σταθερά. Κατά την πίσω διάδοση, η παραγωγή ενός σήματος λάθους γίνεται με την αφαίρεση της πραγματικής απόκρισης από την επιθυμητή. Επίσης, η διάδοση του σήματος του λάθους γίνεται προς τα πίσω στο δίκτυο, ενώ τα βάρη προσαρμόζονται σύμφωνα με τον κανόνα διόρθωσης λάθους.

Ο αλγόριθμος Backpropagation έχει θετικά και αρνητικά. Στα θετικά, είναι εύκολος στη χρήση, αφού έχει λίγες παραμέτρους προς ρύθμιση και είναι εύκολος να υλοποιηθεί, μπορεί να εφαρμοστεί σε ευρεία περιοχή δεδομένων και έχει ευρέως διαδεδομένη χρήση. Στα μειονεκτήματα, η διαδικασία της εκμάθησης είναι αργή, η εισαγωγή νέων στοιχείων υπερκαλύπτει τα παλιά εκτός αν συνεχίσουν να παρέχονται, η διαρκής ενημέρωση του δικτύου είναι δύσκολη διαδικασία, το δίκτυο μπορεί να χαρακτηριστεί ως black box και δεν υπάρχει εγγύηση γενίκευσης ούτε με ελάχιστο σφάλμα. [12]

Ένα Νευρωνικό Δίκτυο για να χρησιμοποιηθεί πρέπει πρώτα να εκπαιδευτεί ώστε να μάθει το περιβάλλον του. Κατά την διαδικασία της μάθησης προσδιορίζονται οι κατάλληλοι συντελεστές βάρους ενώ για την υλοποίηση της χρησιμοποιούνται αλγόριθμοι που ονομάζονται κανόνες μάθησης ή αλγόριθμοι εκπαίδευσης. Γενικά, τα Νευρωνικά Δίκτυα

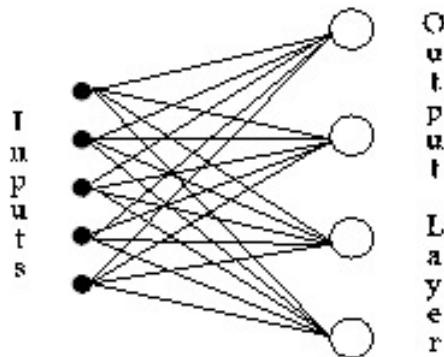
έχουν την ικανότητα να βελτιώνονται μόνα τους μέσω της μάθησης χρησιμοποιώντας το περιβάλλον τους. Η διαδικασία της βελτίωσης εκτελείται σε στάδια, σύμφωνα με ένα προκαθορισμένο μέτρο, ενώ η διαδικασία της μάθησης υλοποιείται μέσω της επαναληπτικής ρύθμισης των τιμών των συναπτικών βαρών και των κατωφλίων. Μετά από κάθε επανάληψη το Νευρωνικό Δίκτυο αυξάνει τα επίπεδα γνώσης του.

Ένας ποιο τυπικός ορισμός της μάθησης δόθηκε από τους Mendel και McLaren, ο οποίος ορίζει την μάθηση “ως μια διαδικασία με την οποία προσαρμόζονται οι ελεύθερες παράμετροι ενός νευρωνικού δικτύου μέσω μιας συνεχούς διαδικασίας διέγερσης από το περιβάλλον στο οποίο βρίσκεται το δίκτυο”, ενώ ο τρόπος με τον οποίο πραγματοποιούνται οι αλλαγές των παραμέτρων καθορίζει το είδος της μάθησης. Η παραπάνω διαδικασία μπορεί να μεταφραστεί σε μια αλληλουχία τριών βημάτων, όπου το πρώτο βήμα είναι η διέγερση του Νευρωνικού Δικτύου από το περιβάλλον, το δεύτερο βήμα είναι οι αλλαγές που προκύπτουν ως αποτέλεσμα αυτής της διέγερσης και τέλος το τρίτο βήμα αφορά την αλλαγή της συμπεριφοράς του δικτύου στο περιβάλλον ως αποτέλεσμα των αλλαγών που συνέβησαν στην εσωτερική του δομή.

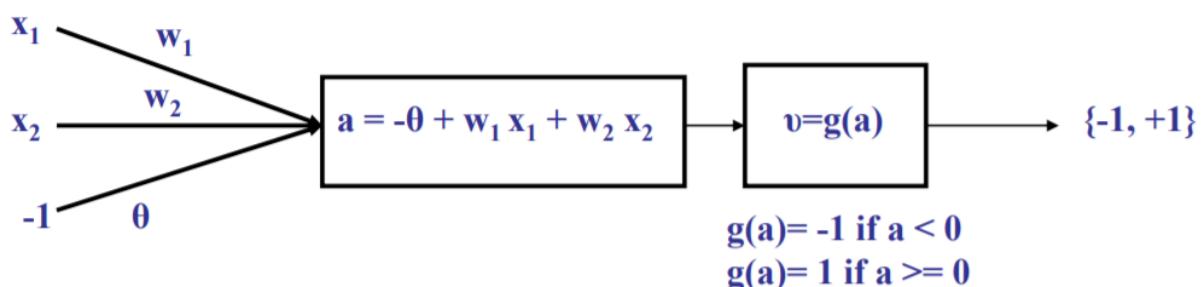
Γενικότερα, υπάρχουν πολλοί αλγόριθμοι που έχουν στόχο την προσαρμογή των τιμών των βαρών ενός Νευρωνικού Δικτύου, όμως όλοι οι αλγόριθμοι χωρίζονται σε δύο κατηγορίες, τη μάθηση με επίβλεψη (supervised learning) και τη μάθηση χωρίς επίβλεψη (unsupervised learning). Η μάθηση με επίβλεψη συνδυάζει έναν παρατηρητή, που έχει σκοπό την εκπαίδευση του δικτύου, και τη συνολική πληροφορία. Η μάθηση με διόρθωση σφάλματος και η στοχαστική μάθηση είναι δύο μέθοδοι που ανήκουν σε αυτή την κατηγορία.

Η μάθηση με επίβλεψη χωρίζεται σε δύο ακόμα κατηγορίες, την δομική (structural) και την προσωρινή (temporal). Οι αλγόριθμοι που ανήκουν στην κατηγορία της δομικής εκμάθησης χρησιμοποιούνται για την εύρεση της βέλτιστης σχέσης μεταξύ εισόδων και εξόδων για κάθε ξεχωριστό ζευγάρι προτύπων. Η αναγνώριση και η κατηγοριοποίηση προτύπων είναι παραδείγματα της δομικής μάθησης, ενώ η πρόβλεψη και ο έλεγχος είναι παραδείγματα της προσωρινής μάθησης. Η μάθηση χωρίς επίβλεψη χαρακτηρίζεται ως αυτό-οργανομένη (self-organized) αφού δεν χρειάζονται κάποιον να επιβλέπει το δίκτυο και χρησιμοποιούν μόνο τοπική πληροφορία κατά τη διαδικασία της εκπαίδευσης. Το συγκεκριμένο είδος αλγορίθμων οργανώνει τα δεδομένα και ανακαλύπτει σημαντικές συλλογικές ιδιότητες. Ο αλγόριθμος Hebbian, ο διαφορικός αλγόριθμος Hebbian και ο Min-Max αλγόριθμος είναι παραδείγματα της εκπαίδευσης χωρίς επίβλεψη. [12]

Ο αισθητήρας (Perceptron) είναι η απλούστερη μορφή Νευρωνικού Δικτύου και είναι ένα δίκτυο με δύο επίπεδα. Το πρώτο επίπεδο αποτελείται από τις εισόδους του δικτύου και δεν εκτελείται καμία επεξεργασία πληροφορίας σε αυτό αφού δεν έχει νευρώνες. Το δεύτερο επίπεδο είναι το επίπεδο εξόδου του δικτύου και αποτελείται από νευρώνες. Ένα παράδειγμα Perceptron με έξι εισόδους και τέσσερις νευρώνες απεικονίζεται στην εικόνα 3.8. Επίσης, χρησιμοποιείται για την ταξινόμηση γραμμικά διαχωριζόμενων προτύπων. Ως γραμμικά διαχωριζόμενα πρότυπα ορίζονται τα πρότυπα που χωρίζονται στο δειγματοχώρο με γραμμικές συναρτήσεις (γραμμές ή επίπεδα). Σαν ταξινομητής, το Perceptron, για d-διάστατα δεδομένα αποτελείται από d βάρη, ένα κατώφλι και μία συνάρτηση (παράδειγμα στην εικόνα 3.9). Αν ομαδοποιήσουμε τα βάρη σε διάνυσμα w τότε έχουμε: $v = g(w \cdot x - \theta)$. [12] Στη συνέχεια, ορίζεται ο αλγόριθμος μάθησης του Perceptron.



Εικόνα 3.8 Παράδειγμα Perceptron με 6 εισόδους και 4 νευρώνες εξόδου



Εικόνα 3.9 Παράδειγμα Perceptron σαν ταξινομητής για d-διάστατα δεδομένα

Αλγόριθμος μάθησης του Perceptron [12]

1. ΔΕΝ γίνεται διόρθωση στο $w(n)$:

$$\text{αν } w^T(n)x(n) \geq 0 \text{ και } x(n) \in l_1 \Rightarrow w(n+1) = w(n)$$

$$\text{αν } w^T(n)x(n) < 0 \text{ και } x(n) \in l_2 \Rightarrow w(n+1) = w(n)$$

2. ΑΛΛΙΩΣ, το διάνυσμα βαρών του Perceptron ενημερώνεται σύμφωνα με τον κανόνα:

$$\text{αν } w^T(n)x(n) \geq 0 \text{ και } x(n) \in l_2 \Rightarrow w(n+1) = w(n) - \eta(n)x(n)$$

$$\text{αν } w^T(n)x(n) < 0 \text{ και } x(n) \in l_1 \Rightarrow w(n+1) = w(n) + \eta(n)x(n)$$

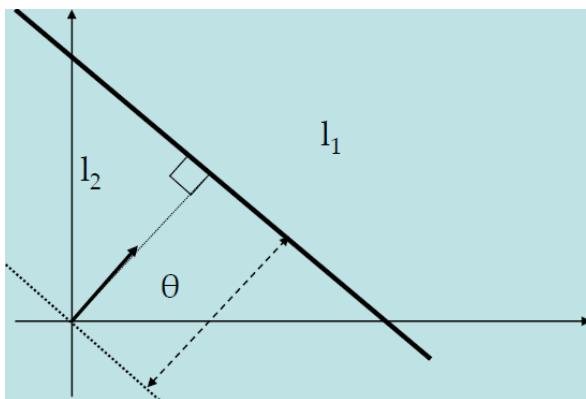
Σκοπός του Perceptron είναι η ταξινόμηση ενός συνόλου εισόδων σε μία από τις κλάσεις l_1 και l_2 . Για την υλοποίηση της ταξινόμησης γίνεται η υπόθεση ότι αν $y=+1$ τότε ανέθεσε το σημείο που αναπαριστούν τα x_1, x_2 στην κλάση l_1 , ενώ αν $y=-1$ ανέθεσε τα σημεία στην κλάση l_2 . Η έξοδος του γραμμικού συνδυαστή είναι:

$$u = \sum_{i=1}^p w_i x_i - \theta,$$

ενώ οι περιοχές απόφασης διαχωρίζονται από το υπερεπίπεδο που ορίζεται από τη σχέση:

$$u = \sum_{i=1}^p w_i x_i - \theta = 0 \Leftrightarrow w_1 x_1 + w_2 x_2 - \theta = 0 \quad [12]$$

Το όριο απόφασης μετατοπίζεται από την αρχή των αξόνων εξαιτίας του κατωφλίου, όπως φαίνεται στην εικόνα 3.10.



Εικόνα 3.10 Μετατόπιση του ορίου απόφασης από το κατώφλι

Τα συναπτικά βάρη του Perceptron προσαρμόζονται επαναληπτικά ενώ το διάνυσμα των βαρών ω προσαρμόζεται με την χρήση του κανόνα σύγκλισης του Perceptron.

Κανόνας σύγκλισης του Perceptron [12]

Αν το διανύσματα εισόδου είναι:

$$x(n) = [-1, x_1(n), x_2(n), \dots, x_p(n)]^T$$

και το διάνυσμα βαρών είναι:

$$w(n) = [\theta(n), w_1(n), w_2(n), \dots, w_p(n)]^T,$$

και η έξοδος του γραμμικού συνδυαστή είναι:

$$u(n) = w^T(n)x(n)$$

τότε υπάρχει ένα διάνυσμα βαρών για το οποίο ισχύει:

$$w^T x \geq 0 \quad \forall x \in l_1 \quad \text{και} \quad w^T x < 0 \quad \forall x \in l_2$$

B) Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines - SVMs)

Στο πεδίο της Μηχανικής Μάθησης οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines – SVMs) στηρίζονται στη θεωρία της Στατιστικής Μάθησης (Statistical Learning Theory) και στα νευρωνικά δίκτυα τύπου Perceptron. Είναι μία από τις πιο διαδεδομένες μεθόδους γραμμικής και μη γραμμικής παρεμβολής και ταξινόμησης, ενώ χρησιμοποιείται ακόμα και για ανίχνευση ακραίων τιμών (outliers). Μερικά παραδείγματα εφαρμογών είναι: αναγνώριση γραφής (handwriting recognition), ταξινόμηση κειμένων (text categorization), ταξινόμηση δεδομένων έκφρασης γονιδίων (gene expression data), ταξινόμηση εικόνων (image classification) κτλ. [14]

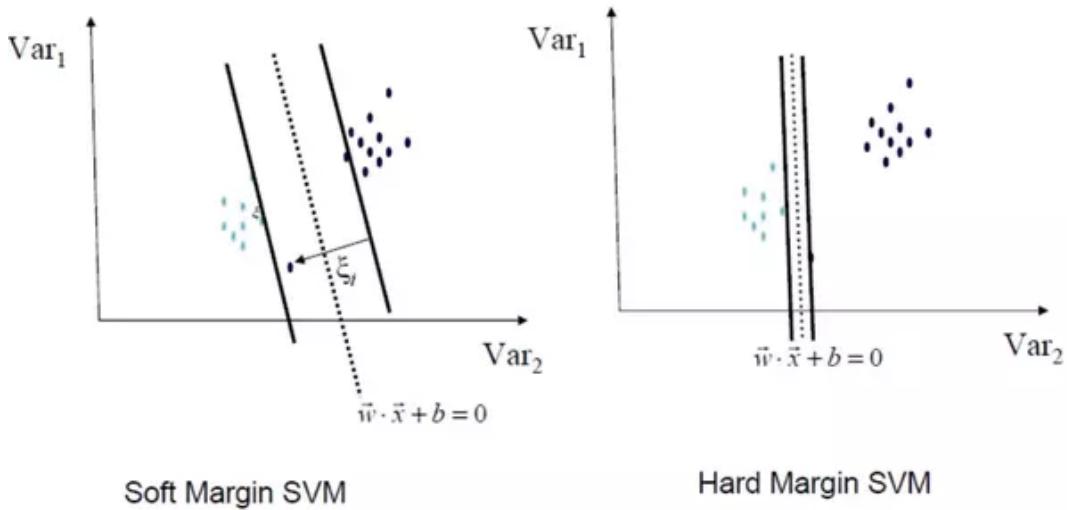
Οι ΜΔΥ, στην περίπτωση της ταξινόμησης, προσπαθούν να εντοπίσουν μια υπερεπιφάνεια (hypersurface), η οποία διαχωρίζει τα θετικά και τα αρνητικά παραδείγματα σε ένα χώρο παραδειγμάτων. Η επιλογή της υπερεπιφάνειας γίνεται με τέτοιο τρόπο ώστε η απόσταση από τα κοντινότερα θετικά και αρνητικά παραδείγματα να μεγιστοποιείται (maximum margin hypersurface). Έστω ότι δίνουμε στον αλγόριθμο εκπαίδευσης ΜΔΥ ένα σύνολο παραδειγμάτων εκπαίδευσης, τα στοιχεία του οποίου είναι χωρισμένα σε δύο κατηγορίες και κάθε

στοιχείο είναι γνωστό σε ποια από τις δύο ομάδες ανήκει. Ο αλγόριθμος δημιουργεί ένα μοντέλο που μπορεί να τοποθετήσει νέα στοιχεία σε μία από τις δύο ομάδες, ενώ αρχικά δεν είναι γνωστό σε ποια κατηγορία ανήκουν. [14][17]

Συνεπώς, είναι ένας μη-πιθανοτικός δυαδικός γραμμικός ταξινομητής (non-probabilistic binary linear classifier), αν και υπάρχουν μέθοδοι που χρησιμοποιούν τις ΜΔΥ με πιθανοτική ταξινόμηση, όπως είναι η μέθοδος Platt scaling. Ένα μοντέλο ΜΔΥ αναπαριστά τα στοιχεία του συνόλου παραδειγμάτων ως σημεία σε ένα χώρο. Τα σημεία τοποθετούνται με τρόπο κατάλληλο ώστε τα στοιχεία των διαφορετικών κατηγοριών να είναι χωρισμένα από ένα κενό που είναι όσο πιο ευρύ γίνεται. Στη συνέχεια, τα νέα στοιχεία που εισάγονται μετατρέπονται σε σημεία του χώρου που προαναφέρθηκε και προβλέπεται σε ποια κατηγορία ανήκουν ανάλογα με την πλευρά του διαχωρισμού ανήκουν. [14]

Συνήθως, οι ΜΔΥ είναι μοντέλα της μάθησης με επίβλεψη, αλλά όταν τα δεδομένα δεν μας πληροφορούν για την ομάδα που ανήκουν (unlabeled data) τα στοιχεία τότε δεν μπορεί να χρησιμοποιηθεί η μάθηση με επίβλεψη και άρα η χρήση της μάθησης χωρίς επίβλεψη είναι αναγκαστική. Η μάθηση χωρίς επίβλεψη προσπαθεί να βρει φυσική ομαδοποίηση των δεδομένων σε ομάδες και στη συνέχεια να τοποθετήσει νέα στοιχεία σε αυτές τις ομάδες που σχημάτισε. Ο αλγόριθμος support vector clustering χρησιμοποιεί στατιστικά των διανυσμάτων υποστήριξης (support vectors) για να τοποθετήσει σε κατηγορίες τα στοιχεία που δεν παρέχουν πληροφορία για την ομάδα που ανήκουν. Αυτός ο αλγόριθμος είναι από τους πιο ευρέως χρησιμοποιούμενους αλγορίθμους ομαδοποίησης σε βιομηχανικές εφαρμογές. [14]

Τα ΜΔΥ μπορούν να εκτελέσουν γραμμική ταξινόμηση. Εστω ότι έχουμε δεδομένα εκπαίδευσης για n σημεία της μορφής $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$, όπου το y_i είναι είτε 1 είτε -1 και υποδηλώνει σε ποια κλάση ανήκει το \vec{x}_i . Τα \vec{x}_i είναι p-διάστατα πραγματικά διανύσματα. Στόχος είναι να βρεθεί το μέγιστο περιθώριο του υπερ-επιπέδου (maximum-margin hyperplane) που διαχωρίζει την ομάδα των \vec{x}_i για τα οποία ισχύει $y_i=1$ από την ομάδα των \vec{x}_i για τα οποία ισχύει $y_i=-1$. Το μέγιστο περιθώριο του υπερ-επιπέδου ορίζεται ως η μέγιστη απόσταση μεταξύ του υπερ-επιπέδου και του κοντινότερου σημείου \vec{x}_i από κάθε ομάδα. Οποιοδήποτε υπερ-επίπεδο μπορεί να εκφραστεί ως ένα σύνολο σημείων \vec{x} που ικανοποιούν τη σχέση $\vec{w} \cdot \vec{x} - b = 0$, όπου το \vec{w} είναι κανονικοποιημένο διάνυσμα ως προς το υπερ-επίπεδο (δεν είναι απαραίτητα κανονονικοποιημένο). Στα γραμμικά ΜΔΥ έχουμε δύο κατηγορίες ανάλογα με το περιθώριο (margin), την hard-margin και την soft-margin. Ένα παράδειγμα hard-margin και soft-margin ΜΔΥ απεικονίζεται στην εικόνα 3.11. [14]



Εικόνα 3.11 Παράδειγμα Soft-Margin και Hard-Margin SVM

Για την hard-margin ισχύει ότι αν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρίσιμα, τότε μπορούμε να επιλέξουμε δύο παράλληλα υπερ-επίπεδα που διαχωρίζουν τις δύο κλάσεις των δεδομένων με τέτοιο τρόπο ώστε να μεγιστοποιείται η απόσταση μεταξύ τους. Η περιοχή που δεσμεύεται από τα δύο υπερ-επίπεδα ονομάζεται περιθώριο (margin) και το μέγιστο περιθώριο του υπερ-επιπέδου είναι το υπερ-επίπεδο που βρίσκεται στην μέση της απόστασης των δύο υπερ-επιπέδων. Ένα παράδειγμα hard-margin ΜΔΥ απεικονίζεται στην εικόνα 3.12. Με την κατάλληλη αντιστοίχηση των δεδομένων, τα υπερ-επίπεδα μπορούν να περιγραφούν από τις εξισώσεις:

$$1) \vec{w} \cdot \vec{x} - b = 1$$

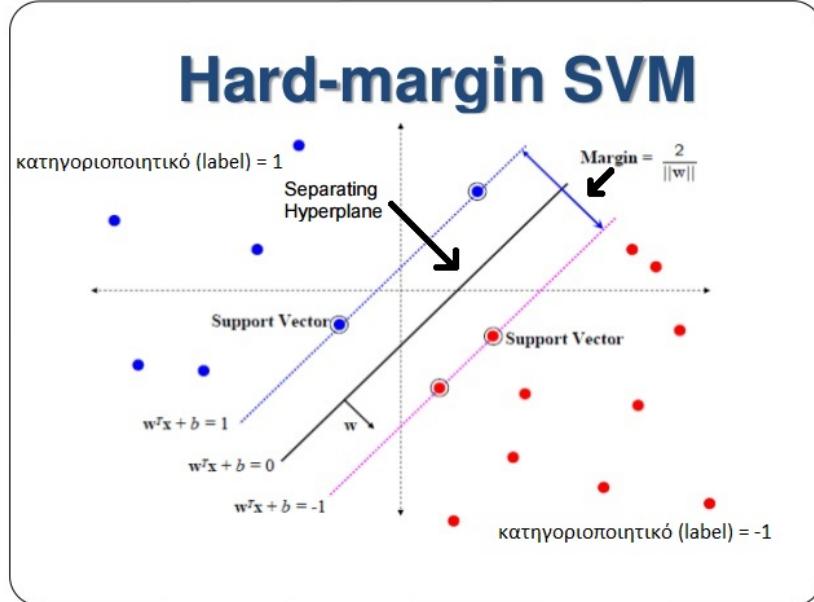
$$2) \vec{w} \cdot \vec{x} - b = -1$$

Για την πρώτη εξίσωση, οποιοδήποτε σημείο πάνω ή μεγαλύτερο από αυτό το όριο ανήκει στην κλάση με κατηγοριοποιητικό 1, ενώ για την δεύτερη εξίσωση οποιοδήποτε σημείο πάνω ή μικρότερο από αυτό το όριο ανήκει στην κλάση με κατηγοριοποιητικό -1. [14]

Γεωμετρικά, η απόσταση μεταξύ των δύο υπερ-επιπέδων ισούται με $\frac{2}{\|\vec{w}\|}$, οπότε για να μεγιστοποιηθεί η απόσταση πρέπει να ελαχιστοποιηθεί το $\|\vec{w}\|$. Ο υπολογισμός της απόστασης γίνεται με τη χρήση του τύπου της απόστασης σημείου από επίπεδο. Ακόμα, πρέπει να αποφύγουμε καταχώρηση δεδομένων στο περιθώριο. Για να επιτευχθεί αυτό θέτουμε τους εξής περιορισμούς: για κάθε i ισχύει είτε η σχέση $\vec{w} \cdot \vec{x}_i - b \geq 1$, αν $y_i = 1$ είτε η σχέση $\vec{w} \cdot \vec{x}_i - b \leq -1$, αν $y_i = -1$. Με αυτούς του περιορισμούς ορίζουμε ότι κάθε σημείο

πρέπει να βρίσκεται στη σωστή πλευρά του περιθωρίου, δηλαδή στη σωστή ομάδα. Οι παραπάνω σχέσεις μπορούν να γραφούν ως μία:

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \text{ για κάθε } 1 \leq i \leq n \quad [14]$$

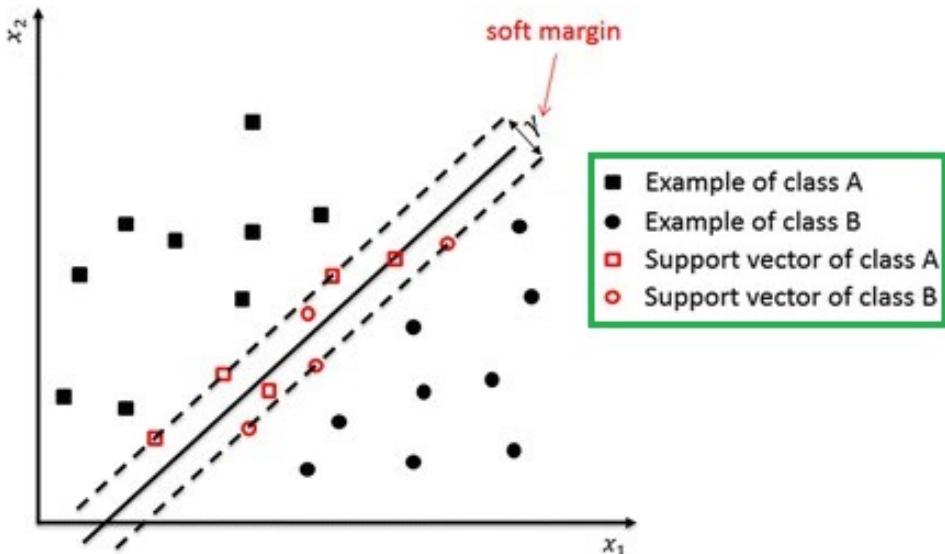


Εικόνα 3.12 Παράδειγμα Hard-Margin SVM

Η κατηγορία soft-margin επεκτείνει τις δυνατότητες των MΔΥ, ώστε να λειτουργούν για δεδομένα που δεν είναι γραμμικώς διαχωριζόμενα. Ένα παράδειγμα soft-margin MΔΥ απεικονίζεται στην εικόνα 3.13. Σε αυτό το πρόβλημα δίνει λύση η hinge loss function που ορίζεται ως: $\max(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b))$. Αυτή η συνάρτηση παίρνει τιμή 0 όταν ισχύει ο περιορισμός $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$, για κάθε $1 \leq i \leq n$, δηλαδή παίρνει τιμή 0 όταν το \vec{x}_i είναι στη σωστή πλευρά του περιθωρίου. Για τα δεδομένα που είναι στη λάθος πλευρά του περιθωρίου, η τιμή της συνάρτησης είναι ανάλογη με την απόσταση από το περιθώριο. Επίσης, θέλουμε να ελαχιστοποιήσουμε την τιμή της παράστασης:

$$\left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b)) \right] + \lambda \|\vec{w}\|^2$$

Η παράμετρος λ ορίζει τη σχέση μεταξύ της αύξησης του μεγέθους του περιθωρίου και της σωστής τοποθέτησης του \vec{x}_i στη σωστή πλευρά του περιθωρίου. Επιπροσθέτως, για αρκετά μικρές τιμές του λ , οι soft-margin MΔΥ συμπεριφέρονται παρόμοια με την κατηγορία των hard-margin MΔΥ αν τα δεδομένα που εισάγονται μπορούν να ταξινομηθούν γραμμικά. [14]



Εικόνα 3.13 Παράδειγμα Soft-Margin SVM

Τα ΜΔΥ, εκτός από την εκτέλεση γραμμικής ταξινόμησης, μπορούν να εκτελούν αποτελεσματικά και μη γραμμική ταξινόμηση. Αυτό μπορεί να επιτευχθεί με την χρήση του “τρικ” του πυρήνα (kernel trick). Η μέθοδος αυτή μετατρέπει τον αρχικό χώρο πεπερασμένων διαστάσεων σε ένα πολύ μεγαλύτερο σε διαστάσεις χώρο, υποθέτοντας ότι θα διευκολυνθεί η ταξινόμηση στον μεγαλύτερο χώρο. Ένα παράδειγμα του kernel trick απεικονίζεται στην εικόνα 3.14, στο οποίο μετασχηματίζεται ο αρχικός χώρος δύο διαστάσεων σε χώρο τριών διαστάσεων όπου και βρίσκουμε το πως ταξινομούνται τα στοιχεία και στη συνέχεια επιστρέφουμε στον δισδιάστατο χώρο. Αυτοί οι μετασχηματισμοί ονομάζονται πυρήνες (kernels) και αυξάνουν τον υπολογιστικό φόρτο. Συνεπώς, για να παραμείνει αποδοτική η διαδικασία, οι αντιστοιχίσεις που γίνονται στην εικόνα των ΜΔΥ θα πρέπει να σχεδιαστούν με τρόπο τέτοιο ώστε τα dot products να μπορούν να υπολογιστούν εύκολα από πλευράς των μεταβλητών στον αρχικό χώρο. Αυτό επιτυγχάνεται με τη χρήση μια συνάρτησης πυρήνα (kernel function) που είναι κατάλληλη για το πρόβλημα.

Στην περίπτωση της μη γραμμικής ταξινόμησης (Non-linear Classification) γίνεται χρήση του kernel trick στα υπερ-επίπεδα μέγιστου περιθωρίου. Ο αλγόριθμος αντικαθιστά όλα τα dot products με μια μη γραμμική συνάρτηση πυρήνα (nonlinear kernel function). Αυτό επιτρέπει στον αλγόριθμο να ταιριάξει τα υπερ-επίπεδα μεγίστου περιθωρίου σε ένα μετασχηματισμένο χώρο χαρακτηριστικών (transformed feature space). Ο μετασχηματισμός μπορεί να είναι μη γραμμικός και οι διαστάσεις του μετασχηματισμένου χώρου πολλές. Αν και ο ταξινομητής είναι ένα υπερ-επίπεδο στο μετασχηματισμένο χώρο χαρακτηριστικών, μπορεί να είναι μη γραμμικός στον αρχικό χώρο δεδομένων. Αξίζει να σημειωθεί ότι η χρήση ενός

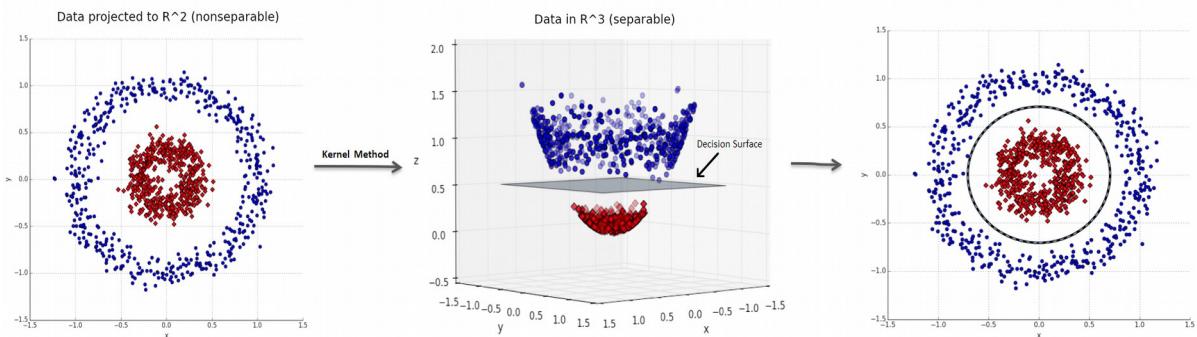
χώρου χαρακτηριστικών που είναι μεγάλος και πολυδιάστατος μπορεί να αυξήσει το γενικευμένο σφάλμα των ΜΔΥ, στο τέλος όμως αφού χρησιμοποιηθούν αρκετά παραδείγματα μάθησης ο αλγόριθμος αποδίδει καλά. [14][16][17] Μερικοί συνηθισμένοι πυρήνες (kernels) είναι οι εξής: [14][15]

- Polynomial (homogeneous): $k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j)^d$
- Polynomial (inhomogeneous): $k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^d$
- Gaussian radial basis function: $k(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \|\vec{x}_i - \vec{x}_j\|^2)$, για $\gamma > 0$
- Hyperbolic tangent:

$$k(\vec{x}_i, \vec{x}_j) = \tanh(\kappa \vec{x}_i \cdot \vec{x}_j + c), \text{ για κάποια } \kappa > 0 \text{ και } c < 0 \text{ (όχι όμως για όλα)}$$

Η εξίσωση $k(\vec{x}_i, \vec{x}_j) = \varphi(\vec{x}_i) \cdot \varphi(\vec{x}_j)$ συνδέει τον πυρήνα με το μετασχηματισμό $\varphi(\vec{x}_i)$.

Η τιμή w ορίζεται ως $\vec{w} = \sum_i a_i y_i \varphi(\vec{x}_i)$ και είναι μετασχηματισμένος χώρος. Τα dot products με κατηγοριοποιητή το w μπορούν να υπολογιστούν με τη χρήση του kernel trick (για παράδειγμα $\vec{w} \cdot \varphi(\vec{x}) = \sum_i a_i y_i k(\vec{x}_i, \vec{x})$). [14]



Εικόνα 3.14 Παράδειγμα Kernel Trick

Η μέθοδος των ΜΔΥ έχει αρνητικά και θετικά. Στα θετικά, δουλεύει πολύ καλά με σαφώς ορισμένο περιθώριο διαχωρισμού (clear margin of separation), είναι αποδοτική σε μεγάλους πολυδιάστατους χώρους, είναι αποδοτική σε περιπτώσεις όπου οι διαστάσεις είναι περισσότερες από το πλήθος των δειγμάτων και είναι αποδοτικό με τη χρήση μνήμης καθώς χρησιμοποιεί ένα υποσύνολο σημείων του συνόλου εκπαίδευσης στην συνάρτηση απόφασης (decision function). Το υποσύνολο των σημείων ονομάζεται διάνυσμα υποστήριξης (support vectors). Στα αρνητικά, δεν λειτουργεί καλά όταν έχουμε μεγάλο σύνολο δεδομένων αφού ο απαιτούμενος χρόνος μάθησης θα είναι μεγάλος. Επίσης, δεν είναι αποδοτική όταν το σύνολο δεδομένων έχει πολύ θόρυβο, π.χ. οι κλάσεις στόχοι (target classes) είναι αναμεμιγμένες.

Τα ΜΔΥ δεν παρέχουν άμεσα εκτιμήσεις πιθανοτήτων. Για τον υπολογισμό τους πρέπει να χρησιμοποιηθεί five-fold cross-validation, που έχει μεγάλο κόστος. [16]

Γ) Πιθανοτική Θεωρία του Bayes (Bayesian Probability)

Η πιθανοτική θεωρία του Bayes (Bayesian Probability) μας επιτρέπει να εξετάσουμε την πιθανότητα ενός γεγονότος με την χρήση ήδη γνωστής γνώσης που αφορά οποιοδήποτε γεγονός σχετικό με το εξεταζόμενο γεγονός που προαναφέρθηκε. Αυτό επιτυγχάνεται με τον τύπο:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

όπου το A και το B είναι δύο γεγονότα, η πιθανότητα $P(A|B)$ είναι η υπό συνθήκη πιθανότητα ότι συμβαίνει το γεγονός A δεδομένο ότι το B έχει ήδη συμβεί. Αυτή η πιθανότητα ονομάζεται και μεταγενέστερη πιθανότητα (posterior probability). Το $P(A)$ και το $P(B)$ είναι η πιθανότητα να συμβεί το A και η πιθανότητα να συμβεί το B αντίστοιχα, χωρίς να ληφθεί υπόψη η οποιαδήποτε σχέση μεταξύ τους. Η πιθανότητα $P(B|A)$ είναι η υπό συνθήκη πιθανότητα ότι συμβαίνει το γεγονός B δεδομένο ότι το A έχει ήδη συμβεί. Με τη χρήση του θεωρήματος του Bayes μπορούμε να δημιουργήσουμε ένα σύστημα που προβλέπει την πιθανότητα της μεταβλητής απόκρισης που ανήκει σε κάποια κατηγορία, δεδομένου ενός νέου συνόλου χαρακτηριστικών. [19]

Στην μηχανική μάθηση συνήθως ενδιαφερόμαστε να κάνουμε την καλύτερη επιλογή της υπόθεσης (hypothesis – H) δεδομένου τα δεδομένα (data – D). Σε ένα πρόβλημα κατηγοριοποίησης η υπόθεση H μπορεί να είναι η κλάση που θα εκχωρηθεί σε ένα καινούργιο σύνολο δεδομένων D. Έστω ότι στην προηγούμενη σχέση όπου είχαμε A θα έχουμε H και όπου B θα έχουμε D. Από την εξίσωση που προαναφέρθηκε, αν υποθέσουμε ότι το H είναι η μεταβλητή απόκρισης (response variable) και το D είναι η είσοδος, τότε έχουμε ότι η πιθανότητα $P(H|D)$ είναι η υπό συνθήκη πιθανότητα της υπόθεσης ότι η μεταβλητή απόκρισης H ανήκει σε μια συγκεκριμένη τιμή, δεδομένου τα χαρακτηριστικά εισόδου D. Αυτή η πιθανότητα ονομάζεται μεταγενέστερη πιθανότητα (posterior probability). Η πιθανότητα $P(D)$ είναι η πιθανότητα κατάρτισης δεδομένων ή αποδεικτικών στοιχείων (evidence). Η πιθανότητα $P(H)$ είναι η πιθανότητα της προηγούμενης πιθανότητας της μεταβλητής απόκρισης. Η πιθανότητα $P(D|H)$ είναι η πιθανότητα των δεδομένων D δεδομένου ότι η υπόθεση H είναι αληθής, δηλαδή είναι η πιθανότητα των δεδομένων εκμάθησης. Έτσι, η παραπάνω ισότητα μπορεί να γραφεί ως:

$$\text{posterior} = \frac{\text{prior} * \text{likelihood}}{\text{evidence}} \quad [18][19][20]$$

Έστω ένα πρόβλημα όπου το πλήθος των γνωρισμάτων είναι ίσο με n και η απόκριση είναι τύπου boolean, για παράδειγμα μπορεί να ανήκει σε μια από δύο κλάσεις. Επίσης, τα γνωρίσματα είναι κατηγορηματικά (categorical) και στην περίπτωση του προβλήματος έχουνε δύο κατηγορίες. Για την εκπαίδευση του ταξινομητή (classifier) πρέπει να υπολογιστεί η πιθανότητα $P(D|H)$ για όλες τις τιμές του χώρου των εισόδων και του χώρου των μεταβλητών απόκρισης. Έτσι πρέπει να υπολογίσουμε $2^*(2^n - 1)$ παραμέτρους για την εκμάθηση του μοντέλου. Αυτό συνήθως δεν είναι δυνατό σε πραγματικά περιβάλλοντα, αφού αν είχαμε για παράδειγμα 30 boolean γνωρίσματα, τότε θα έπρεπε να υπολογιστούν πάνω από 3 δισεκατομμύρια μεταβλητές.

Σε αυτό το πρόβλημα δίνει λύση ο αλγόριθμος του αφελή Bayes (Naive Bayes) κάνοντας την υπόθεση ότι υπάρχει υπό συνθήκη ανεξαρτησία στο σύνολο δεδομένων εκμάθησης. Έτσι, μειώνεται δραστικά η πολυπλοκότητα του παραπάνω προβλήματος σε $2n$. Η υπόθεση της υπό συνθήκης ανεξαρτησίας (conditional independence) ορίζει ότι δεδομένου των μεταβλητών X , Y και Z λέμε ότι η X είναι υπό συνθήκη ανεξάρτητη της Y δεδομένου της Z , αν και μόνο αν η κατανομή της πιθανότητας που διέπει το X είναι ανεξάρτητη των τιμών της Y δεδομένου της Z . Με άλλα λόγια, η X και η Y είναι υπό συνθήκη ανεξάρτητες δεδομένου της Z , αν και μόνο αν δεδομένου ότι η Z συμβαίνει, τότε η γνώση του αν συμβαίνει η X δεν μας προσφέρει καμία γνώση για την πιθανότητα που συμβαίνει και η Y . Αντίστροφά, αν και μόνο αν δεδομένου ότι η Z συμβαίνει, τότε η γνώση του αν συμβαίνει η Y δεν μας προσφέρει καμία γνώση για την πιθανότητα που συμβαίνει και η X . Αυτή η υπόθεση κάνει τον αλγόριθμο του Bayes αφελή. Σε αυτήν την περίπτωση, δεδομένου η διαφορετικών τιμών των χαρακτηριστικών, η πιθανότητα μπορεί γραφτεί ως:

$$P(X_1 \dots X_n | Y) = \prod_{i=1}^n P(X_i | Y)$$

όπου το X αναπαριστά τα χαρακτηριστικά (attributes or features) και η Y είναι η μεταβλητή απόκρισης (response variable). Η πιθανότητα $P(X|Y)$ γίνεται ίση με το προϊόν της κατανομής πιθανότητας κάθε χαρακτηριστικού X δεδομένου του Y . [19] Ο Naive Bayes έχει αρκετές χρήσεις. Μπορεί να χρησιμοποιηθεί για προβλέψεις σε πραγματικό χρόνο, για πρόβλεψη πολλαπλών κατηγοριών, για ταξινόμηση κειμένου (text classification), φιλτράρισμα ανεπιθύμητων μηνυμάτων (spam filtering), ανάλυση συναισθημάτων (sentiment analysis) αλλά και για σύστημα σύστασης (recommendation system). [18]

Στα πλαίσια του Bayes υπάρχει η πεποίθηση που ονομάζεται προηγούμενη (prior). Στη συνέχεια, με την απόκτηση δεδομένων ανανεώνουμε αυτήν την πεποίθηση. Το αποτέλεσμα ονομάζεται μεταγενέστερη (posterior). Αν αποκτηθούν και άλλα δεδομένα, τότε η παλιά posterior γίνεται η καινούργια prior και ο κύκλος επαναλαμβάνεται. Η πιθανότητα $P(A|B)$ διαβάζεται ως η πιθανότητα του A δεδομένου του B και είναι η πιθανότητα να συμβεί το A αν συμβεί το B, δηλαδή είναι μια υπό συνθήκη πιθανότητα.

Η χρησιμοποίηση πιθανοτήτων για την εκτίμηση παραμέτρων μοντέλων ονομάζεται μέγιστη εκτίμηση πιθανοτήτων (maximum likelihood estimation – MLE). Αν ληφθεί υπόψη και η prior, τότε ονομάζεται μέγιστη μεταγενέστερη εκτίμηση (maximum a posteriori estimation – MAP). Αν η prior είναι ομογενής (uniform), τότε η MLE και η MAP είναι το ίδιο. Επίσης, μετά τον υπολογισμό της posterior για ένα πλήθος διαφορετικών υποθέσεων, μπορούμε να επιλέξουμε τις υποθέσεις που έχουν την μεγαλύτερη πιθανότητα. Αυτό ονομάζεται μέγιστη πιθανή υπόθεση (maximum probable hypothesis) και ονομάζεται υπόθεση maximum a posteriori (MAP). Τα συμπεράσματα (inference) αναφέρονται στον τρόπο που το μοντέλο μαθαίνει τις παραμέτρους. Ένα μοντέλο διαφοροποιείται από τον τρόπο που εκπαιδεύεται. [19]

Στο φάσμα των μεθόδων του Bayes υπάρχουν δύο κατηγορίες, η στατιστική μοντελοποίηση (statistical modelling) και η πιθανοτική μηχανική μάθηση (probabilistic machine learning). Η πιθανοτική μηχανική μάθηση χρησιμοποιεί μη παραμετρικές προσεγγίσεις (nonparametric approaches). Η διαδικασία της μοντελοποίησης εκτελείται όταν τα δεδομένα είναι αραιά, σημαντικά και δύσκολο να αποκτηθούν, αφού το δείγμα είναι μικρό και δεν υπάρχει χώρος για την μείωση του μεγέθους του. Για παράδειγμα σε κοινωνικά πειράματα όπου είναι δύσκολο να εκτελεστούν μεγάλα ελεγχόμενα πειράματα. Επίσης, όταν το δείγμα είναι μικρό είναι σημαντικό να ποσοτικοποιηθεί η αβεβαιότητα, κάτι που η μέθοδος του Bayes μπορεί να προσδιορίσει αποτελεσματικά. [19]

Αν υποθέσουμε ότι η μεταβλητή απόκρισης είναι ομοιόμορφα κατανεμημένη, τότε είναι εξίσου πιθανό να ληφθεί οποιαδήποτε απόκριση και έτσι μπορούμε να απλοποιήσουμε ακόμα περισσότερο τον αλγόριθμο. Με αυτή την υπόθεση η priori ή $P(Y)$ γίνεται σταθερή τιμή, η οποία ισούται με το 1/κατηγορίες της απόκρισης. Αφού η priori και η απόδειξη (evidence - $P(D)$) είναι ανεξάρτητες από την μεταβλητή απόκρισης, μπορούν να απαλειφθούν από την εξίσωση. Έτσι, το πρόβλημα της μεγιστοποίησης της posteriori ουσιαστικά περιορίζεται στη μεγιστοποίηση της πιθανότητας (likelihood), αφού ο τύπος της πιθανότητας γίνεται posterior = likelihood. [19]

Όπως προαναφέρθηκε, πρέπει να εκτιμήσουμε την κατανομή της μεταβλητής της απόκρισης από τα δεδομένα εκπαίδευσης ή να υποθέσουμε ότι είναι ομοιόμορφα κατανεμημένη. Παρομοίως, για την εκτίμηση των παραμέτρων για την κατανομή ενός χαρακτηριστικού, πρέπει να υποθέσουμε μια κατανομή ή να παράγουμε μη παραμετρικά μοντέλα για τα χαρακτηριστικά που προκύπτουν από τα δεδομένα εκπαίδευσης. Αυτές οι υποθέσεις ονομάζονται μοντέλα γεγονότων (event models). Η διαφοροποίηση των υποθέσεων παράγει διαφορετικούς αλγορίθμους για διαφορετικούς σκοπούς.

Για συνεχής κατανομές συνήθως χρησιμοποιείται ο αλγόριθμος Gaussian naive Bayes, ενώ για διακριτά χαρακτηριστικά γίνεται χρήση του multinomial και των κατανομών Bernoulli. Οι ταξινομητές του Naive Bayes λειτουργούν αποδοτικά σε περίπλοκες περιπτώσεις, παρά των απλοποιημένων υποθέσεων και την αφέλεια. Το πλεονέκτημα αυτών των κατηγοριοποιητών είναι ότι απαιτούν μικρό μέγεθος δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που απαιτούνται για την κατηγοριοποίηση. Ο αλγόριθμος Naive Bayes προτιμάτε για κατηγοριοποίηση κειμένου (text categorization). [19]

Ο αλγόριθμος του Naive Bayes έχει θετικά και αρνητικά. Στα θετικά, είναι εύκολο και γρήγορο να προβλεφθεί μια κλάση από ένα σύνολο δεδομένων δοκιμών, ενώ είναι καλός και στην πρόβλεψη πολλαπλών τάξεων. Επίσης, όταν έχει γίνει η υπόθεση ότι υπάρχει ανεξαρτησία, τότε ο κατηγοριοποιητής του Naive Bayes αποδίδει καλύτερα σε σύγκριση με άλλα μοντέλα, όπως η λογική παλινδρόμηση (logistic regression), ενώ απαιτεί και μικρότερο σύνολο δεδομένων εκπαίδευσης. Αποδίδει καλά σε περίπτωση κατηγορικών μεταβλητών εισόδου σε σχέση με αριθμητικές μεταβλητές. Για αριθμητικές μεταβλητές, θεωρούμε ότι η κατανομή είναι κανονική.

Στα αρνητικά, αν για μια κατηγορική μεταβλητή του συνόλου δεδομένων δοκιμών προκύψει ότι έχει μια κατηγορία που δεν είχε παρατηρηθεί στο σύνολο δεδομένων εκπαίδευσης, τότε το μοντέλο θα θέσει μηδέν την πιθανότητα και δεν θα είναι δυνατό να δημιουργήσει μια πρόβλεψη. Αυτό το πρόβλημα είναι γνωστό ως μηδενική συχνότητα (zero frequency). Για την επίλυση του, πρέπει να χρησιμοποιηθεί τεχνική εξομάλυνσης (smoothing technique), όπως είναι για παράδειγμα η εκτίμηση Laplace (Laplace estimation). Επίσης, ο Naive Bayes θεωρείται κακός εκτιμητής (estimator), οπότε οι πιθανότητες που παράγει ως έξοδο δεν είναι πολύ αξιόπιστες. Τέλος, η υπόθεση του Naive Bayes ότι οι προγνωστικοί παράγοντες είναι ανεξάρτητοι δεν ισχύει στην πραγματική ζωή, αφού είναι σχεδόν αδύνατο να βρεθεί ένα σύνολο από προγνωστικούς παράγοντες που είναι τελείως ανεξάρτητοι. [18]

3.3 Μετρικές αξιολόγησης

Όπως είδαμε παραπάνω υπάρχουν πολλοί αλγόριθμοι Μηχανικής Μάθησης που εξυπηρετούν διαφορετικούς σκοπούς. Παράλληλα, για ένα συγκεκριμένο πρόβλημα μπορεί να υπάρχουν πολλοί αλγόριθμοι που βελτιστοποιούν το όφελος ανάλογα με το πιο είναι το σημείο επικέντρωσης του χρήστη, για παράδειγμα επικέντρωση στην ταχύτητα εκτέλεσης του αλγορίθμου σε βάρος εύρεσης της βέλτιστης λύσης. Συνεπώς, η επιλογή του κατάλληλου αλγορίθμου είναι πολύ σημαντική, αφού ο αλγόριθμος που θα χρησιμοποιηθεί για την επίλυση του προβλήματος πρέπει να βελτιστοποιεί το κέρδος που προσφέρει σε σχέση με τους υπόλοιπους αλγορίθμους.

Για την επιλογή του κατάλληλου αλγορίθμου ο χρήστης πρέπει να λάβει υπόψη του τα οφέλη των σχετικών αλγορίθμων καθώς και τα αρνητικά τους, ώστε η επιλογή του να του αποφέρει όσο το δυνατόν μεγαλύτερο κέρδος συγκριτικά με τους υπόλοιπους. Για την εξυπηρέτηση αυτού του σκοπού, δηλαδή την επιλογή του καλύτερου αλγορίθμου για την επίλυση ενός συγκεκριμένου προβλήματος, χρησιμοποιούνται Μετρικές Αξιολόγησης (Metrics). Οι Μετρικές Αξιολόγησης αναθέτουν ποσοτικές τιμές στα οφέλη των αλγορίθμων με αποτέλεσμα να απλουστεύουν ως ένα βαθμό την διαδικασία σύγκρισης και επιλογής του κατάλληλου αλγορίθμου.

Ποιο συγκεκριμένα, οι Μετρικές Αξιολόγησης οργανώνονται σε έναν πίνακα που ονομάζεται **Confusion Matrix** (Εικόνα 3.15), γνωστός και ως Error Matrix. Ο Confusion Matrix επιτρέπει την αναπαράσταση της απόδοσης ενός αλγορίθμου. Κάθε σειρά του πίνακα αντιπροσωπεύει τις περιπτώσεις της προβλεπόμενης κλάσης (Predicted Value) ενώ κάθε στήλη αντιπροσωπεύει τις περιπτώσεις της πραγματικής κλάσης (Actual Value) ή αντίστροφα. Η ονομασία του προέρχεται από το γεγονός ότι καθιστά εύκολο τον έλεγχο για το αν το σύστημα μπερδεύει δύο κλάσεις.

Ο Confusion Matrix είναι ένας πίνακας με 2 γραμμές και 2 στήλες που καταγράφει το πλήθος των **False Positives (FP)**, **False Negatives (FN)**, **True Positives (TP)** και **True Negatives (TN)**. Ο όρος False Positive αναφέρεται σε ένα αποτέλεσμα που υποδεικνύει ότι μια κατάσταση υπάρχει ενώ στη πραγματικότητα δεν υπάρχει. Ο όρος False Negative αναφέρεται σε ένα αποτέλεσμα που υποδεικνύει ότι μια κατάσταση δεν υπάρχει ενώ στην πραγματικότητα υπάρχει. Ο όρος True Positive αναφέρεται σε ένα αποτέλεσμα που υποδεικνύει ότι μια κατάσταση υπάρχει, το οποίο ισχύει και στην πραγματικότητα. Ο όρος True Negative

αναφέρεται σε ένα αποτέλεσμα που υποδεικνύει ότι μια κατάσταση δεν υπάρχει, το οποίο ισχύει και στην πραγματικότητα. [21] [23]

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

Εικόνα 3.15 Confusion Matrix

Για την περαιτέρω κατανόηση των παραπάνω ορισμών παραθέτονται τα παρακάτω παραδείγματα:

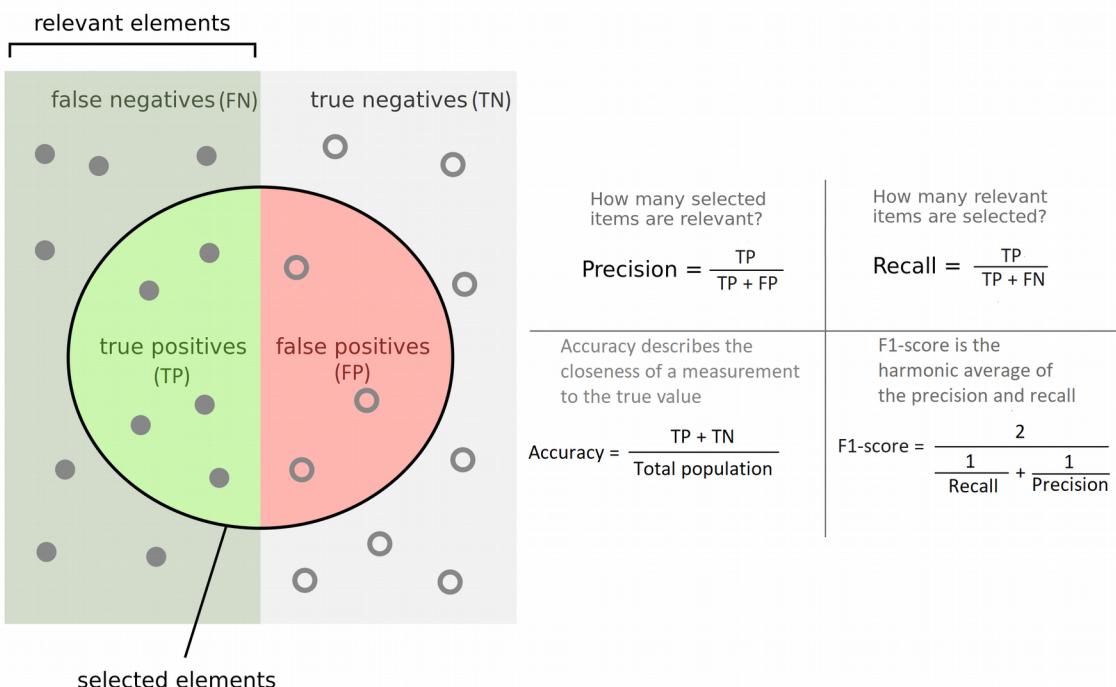
- **True positive (TP):** όταν άρρωστοι άνθρωποι αναγνωρίζονται ορθός ότι είναι άρρωστοι
- **False positive (FP):** όταν υγιής άνθρωποι λανθασμένα αναγνωρίζονται ως άρρωστοι
- **True negative (TN):** όταν υγιής άνθρωποι ορθός αναγνωρίζονται ως υγιής
- **False negative (FN):** όταν άρρωστοι άνθρωποι λανθασμένα αναγνωρίζονται ως υγιής

Γενικά αν θεωρήσουμε ότι το Positive ορίζεται ως αναγνωρισμένο ενώ το negative ως μη αναγνωρισμένο, τότε έχουμε: [22]

- True positive = ορθός αναγνωρισμένο
- False positive = λανθασμένα αναγνωρισμένο
- True negative = ορθός μη αναγνωρισμένο
- False negative = λανθασμένα μη αναγνωρισμένο

Άλλες χρήσιμες Μετρικές Αξιολόγησης είναι η ορθότητα (Precision ή positive predictive value), η ανάκληση (Recall ή Sensitivity ή true positive rate), η ακρίβεια (Accuracy) και ο αρμονικός μέσος (F1-score). Γενικά, η επιλογή και η χρήση των μετρικών γίνεται ανάλογα με το επιστημονικό πεδίο που τις χρησιμοποιεί. Στο πεδίο της πληροφορικής οι μετρικές που χρησιμοποιούνται πιο πολύ είναι το Precision και το Recall. Το F1-score είναι χρήσιμο αλλά γενικά χρησιμοποιείται πιο σπάνια σε σχέση με τις άλλες βασικές μετρικές. Η ορθότητα μπορεί να θεωρηθεί ως μέτρο ακρίβειας ή πιστότητας, ενώ η ανάκληση είναι μέτρο πληρότητας.

Επίσης, για την επιλογή μιας μετρικής παίζουν ρόλο διάφοροι παράγοντες, όπως το βάρος που αντιστοιχούμε στο κόστος σφάλματος των positives σε σχέση με το βάρος των negatives. Για παράδειγμα αν θέλουμε να αποφύγουμε περισσότερο τα false negatives από τα false positives, και το αντίστροφο, τότε η ανάκληση είναι πιο ενημερωτική σε σχέση με την ακρίβεια. Οι τέσσερις μετρικές που αναφέρθηκαν ορίζονται στο Εικόνα 3.16.. [22]



Εικόνα 3.16 Precision, Recall, Accuracy και F1-score

Ορθότητα (Precision)

Η ορθότητα (Precision) ορίζεται ως $TP / (TP + FP)$ και αντιπροσωπεύει το πόσα αντικείμενα από αυτά που επιλέχθηκαν είναι σχετικά. [22]

Ακρίβεια (Accuracy)

Η ακρίβεια (Accuracy) ορίζεται ως $(TP + TN) / (\text{Total Population})$, όπου το Total Population ισούται με $TP+FP+FN+TN$, και περιγράφει την εγγύτητα της μέτρησης μιας ποσότητας με την πραγματική τιμή της ποσότητας. [21][22]

Σε αντίθεση με την καθομιλουμένη, όπου οι όροι Precision και Accuracy είνοι συνώνυμες, στο επιστημονικό πεδίο είναι αντίθετες. Ιδανικά, ένας αλγόριθμος είναι και ορθός και ακριβής, με όλες τις μετρήσεις να είναι πολύ κοντά στις πραγματικές τιμές. Γενικά, ένας αλγόριθμος μπορεί να είναι ακριβής αλλά όχι ορθός ή ορθός αλλά όχι ακριβής ή τίποτα από τα δύο ή και τα δύο. Για παράδειγμα, αν ένα πείραμα περιέχει ένα συστηματικό σφάλμα, τότε αυξάνο-

ντας το μέγεθος του δείγματος αυξάνεται η ορθότητα αλλά δεν επηρεάζει την ακρίβεια. Εξαλείφοντας το συστηματικό σφάλμα βελτιώνεται η ακρίβεια αλλά δεν αλλάζει η ορθότητα. Ένας αλγόριθμος θεωρείται έγκυρος αν είναι ορθός και ακριβής. [22]

Ανάκληση (Recall)

Η **ανάκληση (Recall)** ορίζεται ως $TP / (TP + FN)$ και αντιπροσωπεύει το πόσα σχετικά αντικείμενα επιλέχθηκαν, δηλαδή είναι η αναλογία των ατόμων που η εξέταση έδειξε ορθός ότι είναι θετικοί ως προς το σύνολο των ανθρώπων που είναι όντος θετικοί. Με άλλα λόγια, η ανάκληση είναι η ικανότητα ενός τεστ να προσδιορίζει σωστά το ποσοστό των ασθενών με ασθένεια. Μπορεί να θεωρηθεί ως η πιθανότητα ότι το τεστ είναι θετικό δεδομένο ότι ο ασθενής είναι άρρωστος. Επίσης, η ανάκληση μπορεί να οριστεί ως ο βαθμός των true positives (TP) που δεν έχει αγνοηθεί και ουσιαστικά ποσοτικοποιεί την αποφυγή των false negatives (FN). Ένα αρνητικό αποτέλεσμα σε ένα τεστ με υψηλή ανάκληση είναι χρήσιμο για να αποκλειστεί μια ασθένεια. Ένα τεστ υψηλής ανάκλησης είναι αξιόπιστο, άρα και πιο χρήσιμο, όταν το αποτέλεσμα του είναι αρνητικό, αφού σπάνια κάνει λάθος για αυτούς που έχουν την ασθένεια. Ένα τεστ με 100% ανάκληση θα αναγνωρίσει όλους τους ασθενής με ασθένεια ελέγχοντας τα τεστ που ήταν θετικά. Έτσι οι περιπτώσεις ασθένειας που δεν ανιχνεύονται είναι λιγότερες. Ένα αρνητικό αποτέλεσμα στο τεστ θα απέκλειε σίγουρα την ύπαρξη ασθένειας σε έναν ασθενή. Ένα θετικό αποτέλεσμα σε ένα τεστ με μεγάλη ανάκληση δεν είναι χρήσιμο για τον αποκλεισμό μιας ασθένειας.

Έστω για παράδειγμα ένα κίβδηλο τεστ υγείας το οποίο είναι σχεδιασμένο να παράγει πάντα θετικά αποτελέσματα. Όταν χρησιμοποιείται από άρρωστους ανθρώπους όλα τα αποτελέσματα θα είναι θετικά και άρα θα έχουμε 100% ανάκληση. Όμως η ανάκληση από τον ορισμό της δεν λαμβάνει υπόψη τα false positives. Το κίβδηλο τεστ παράγει θετικό σε όλους τους υγιής ανθρώπους, το οποίο παράγει ποσοστό 100 % για τα false positive, καθιστώντας το άχρηστο για τον αποκλεισμό μιας ασθένειας. Αξίζει να σημειωθεί ότι η ανάκληση δεν είναι το ίδιο με την ορθότητα. [22]

Αρμονικός μέσος (F1-score)

Ο **αρμονικός μέσος (F1-score)** ορίζεται ως $2 / (1 / Recall + 1 / Precision)$ και εκφράζει τον αρμονικό μέσο της ανάκλησης με την ορθότητα. Όταν ο αρμονικός μέσος παίρνει τιμή 1 τότε έχουμε τέλεια ανάκληση και τέλεια ορθότητα, ενώ όταν παίρνει τιμή 0 τότε έχουμε την

χειρότερη δυνατή ανάκληση και ορθότητα. Ο αρμονικός μέσος είναι ειδική περίπτωση του F_{β} -measure που ορίζεται ως $(1 + \beta^2) * \text{precision} * \text{recall} / ((\beta^2) * \text{precision} + \text{recall})$. Για $\beta=1$ προκύπτει ο τύπος του αρμονικού μέσου, ο οποίος θεωρεί την ανάκληση και την ορθότητα ως ισοζυγισμένα. [22]

Receiver Operating Characteristic – Area Under Curve (ROC-AUC)

Πέρα από τις τέσσερεις μετρικές που προαναφέρθηκαν, για την αξιολόγηση ενός μοντέλου μπορεί να φανεί χρήσιμος ο υπολογισμός του Receiver Operating Characteristic - Area Under Curve (ROC-AUC) όπως και η αντίστοιχη καμπύλη για την οπτικοποίηση του ποσοστού της μετρικής. Η καμπύλη ROC είναι ένας οπτικός τρόπος για την επιθεώρηση της απόδοσης ενώ δυαδικού ταξινομητή, δηλαδή που παράγει 0 και 1 σαν έξοδο.

Η καμπύλη δημιουργείται με γραφική παράσταση της πραγματικής θετικής συχνότητας (true positive rate - TPR) έναντι της ψευδώς θετικής συχνότητας (false positive rate - FPR) σε διάφορες τιμές κατωφλίου. [24] Πιο συγκεκριμένα, συγκρίνει το ρυθμό που ο ταξινομητής παράγει σωστές προβλέψεις (True Positive ή TP) με το ρυθμό που ο ταξινομητής παράγει false alarm (False Positives ή FP). Ο όρος true positive rate (TPR) ορίζει πόσα σωστά θετικά αποτελέσματα προκύπτουν σε σχέση με όλα τα θετικά δείγματα που είναι διαθέσιμα κατά το τεστ, ενώ το false positive rate (FPR) ορίζει πόσα λανθασμένα θετικά αποτελέσματα προκύπτουν σε σχέση με όλα τα αρνητική δείγματα που είναι διαθέσιμα κατά το τεστ.

$$TPR = \text{TruePositives} / (\text{TruePositives} + \text{FalseNegatives})$$

$$FPR = \text{FalsePositives} / (\text{FalsePositives} + \text{TrueNegatives})$$

Τα True Positive Rate και True Negative Rate αναφέρονται και ως Sensitivity και Specificity, αντίστοιχα. Ουσιαστικά, μετράται ο συμβιβασμός που γίνεται ανάμεσα στο ρυθμό που προβλέπεται κάτι σωστά με το ρυθμό που προβλέπεται κάτι που δεν ισχύει. Με άλλα λόγια, ο χώρος ROC ορίζεται από το FPR και το TPR ως X και Y άξονες, αντίστοιχα, που απεικονίζει τους σχετικούς περιορισμούς που υπάρχουν μεταξύ των true positives (benefits) και των false positive (costs). [24] Η ανάλυση ROC παρέχει εργαλεία για την επιλογή πιθανών βέλτιστων μοντέλων και την απόρριψη των υποβέλτιστων μοντέλων ανεξάρτητα από το πλαίσιο κόστους ή την κατανομή της κλάσης. Η ανάλυση ROC σχετίζεται με άμεσο και φυσικό τρόπο με την ανάλυση κόστους / οφέλους της λήψης διαγνωστικών αποφάσεων.

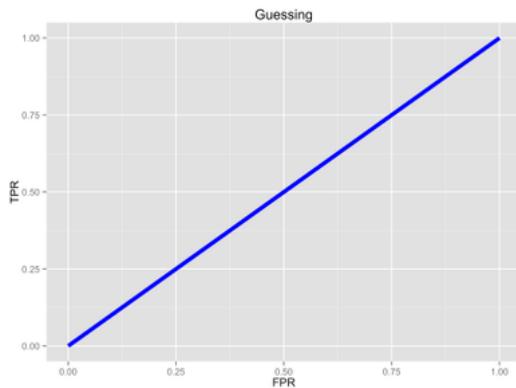
Στην εικόνα 3.17 η καμπύλη είναι μια διαγώνιος γραμμή, η οποία απεικονίζει ένα μοντέλο το οποίο κάνει τυχαίες προβλέψεις. Οι προβλέψεις του ταξινομητή που θα είναι σωστές θα έχουν ποσοστό 50%, το οποίο σημαίνει ότι η τιμές των TPR και FPR θα είναι ίσες. Συχνά, στα σχήματα ROC συμπεριλαμβάνεται η καμπύλη του ταξινομητή με τις τυχαίες προβλέψεις για να ορίσει ένα σημείο αναφοράς. Οι καμπύλες πάνω από την διαγώνιο γραμμή είναι καλύτερες από τυχαίες προβλέψεις ενώ οι καμπύλες που βρίσκονται χαμηλότερα δίνουν κάκιστα αποτελέσματα. Στην εικόνα 3.18 απεικονίζεται ένα τέλειος ταξινομητής, δηλαδή ένας ταξινομητής που βρίσκει όλες τις προβλέψεις σωστά. Αυτός ο ταξινομητής έχει το τέλειο trade off μεταξύ του TPR και του FPR, δηλαδή το TPR έχει τιμή 1 ενώ το FPR τιμή 0.

Στην εικόνα 3.19 απεικονίζεται ένας ταξινομητής που δίνει χειρότερα αποτελέσματα από τον τυχαίο ταξινομητή, αφού η καμπύλη του είναι κάτω από αυτή του τυχαίου. Στην εικόνα 3.20 απεικονίζεται ένα ταξινομητής που δίνει αποτελέσματα καλύτερα από αυτά του τυχαίου ταξινομητή, που μπορεί να χαρακτηριστεί ως μέτριος ταξινομητής. Αξίζει να σημειωθεί πως όταν μια καμπύλη κάνει απότομη πρώτη μπορεί να δηλώνει μια ανωμαλία στα δεδομένα ή κακή υπόθεση του μοντέλου. Στην εικόνα 3.21 απεικονίζεται ένα μοντέλο που δίνει αρκετά καλά αποτελέσματα.

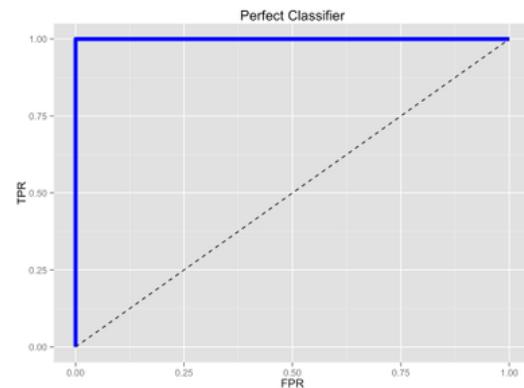
Στην εικόνα 3.22 απεικονίζεται η περιοχή κάτω από την καμπύλη (area under curve – AUC), η οποία είναι ισοδύναμη με την πιθανότητα ένας ταξινομητής να ταξινομήσει ένα τυχαία επιλεγμένο θετικό στοιχείο υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό στοιχείο, υποθέτοντας ότι τα θετικά ταξινομούνται υψηλότερα από τα αρνητικά. [24] Η AUC μπορεί να υπολογιστεί από τον τύπο:

$$A = \int_{-\infty}^{\infty} TPR(T)FPR'(T)dT = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(T' > T)f_1(T')f_0(T)dT'dT = P(X_1 > X_0)$$

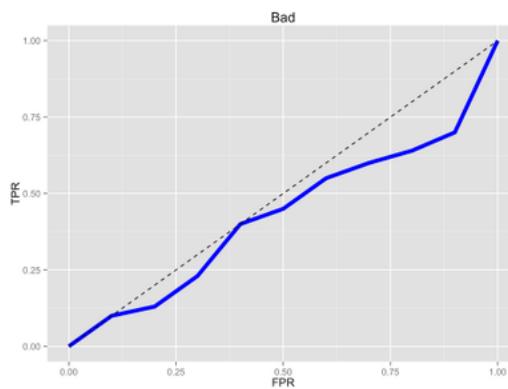
όπου X_1 είναι το score για ένα θετικό στοιχείο και X_0 είναι το score για ένα αρνητικό στοιχείο, ενώ το f_0 και το f_1 είναι πυκνότητες πιθανοτήτων (probability densities). Τα όρια αντιστρέφονται επειδή το T έχει χαμηλότερη τιμή στον άξονα X. Όσο πιο κοντά είναι η τιμή του AUC στο 1 τόσο πιο καλό είναι το μοντέλο, ενώ όσο πιο κοντά στο 0 τόσο πιο κακό είναι. Το ROC και το AUC είναι αλληλένδετα αφού όσο πιο πάνω και αριστερά βρίσκεται η καμπύλη ROC τόσο πιο μεγάλη τιμή θα έχει η AUC και άρα τόσο καλύτερος θα είναι ο ταξινομητής. Η τιμή του AUC μας δίνει τη δυνατότητα να συγκρίνουμε μοντέλα, αφού μπορούμε εύκολα να επιλέξουμε το μοντέλο με την μεγαλύτερη τιμή AUC.



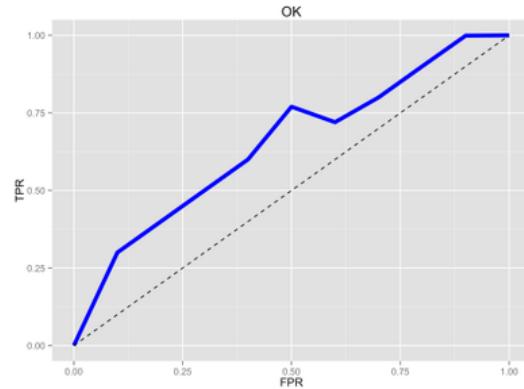
Εικόνα 3.17 Μοντέλο τυχαίας πρόβλεψης



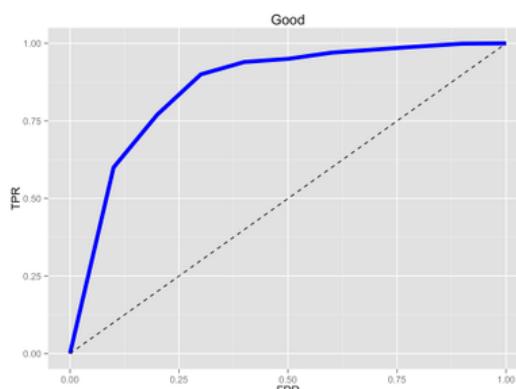
Εικόνα 3.18 Μοντέλο τέλειας πρόβλεψης



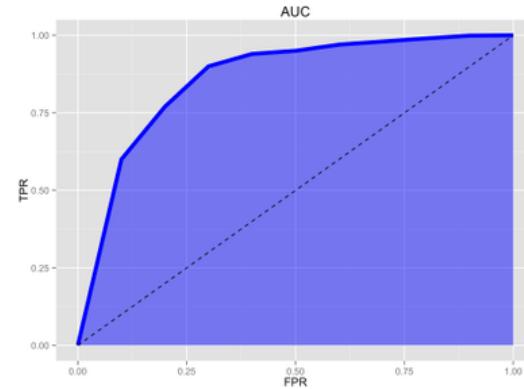
Εικόνα 3.19 Μοντέλο κακής αποτελεσματικότητας



Εικόνα 3.20 Μοντέλο μέτριας αποτελεσματικότητας



Εικόνα 3.21 Μοντέλο καλής αποτελεσματικότητας



Εικόνα 3.22 Area Under Curve - AUC

Σε αντίθεση, αν είχαμε μόνο την καμπύλη ROC δεν θα μπορούσαμε τόσο εύκολα να συγκρίνουμε τα μοντέλα, αφού δεν είναι κάποιο νούμερο που την χαρακτηρίζει ολόκληρη. Μια αξιόπιστη και έγκυρη εκτίμηση AUC μπορεί να ερμηνευτεί ως η πιθανότητα ο ταξινομητής να εκχωρήσει υψηλότερη βαθμολογία σε ένα τυχαία επιλεγμένο θετικό στοιχείο από ένα τυχαία επιλεγμένο αρνητικό στοιχείο. Επίσης, μερικές φορές είναι πιο χρήσιμο να ελεγχθεί μια

συγκεκριμένη περιοχή της καμπύλης ROC αντί ολόκληρη την καμπύλη, αφού είναι πιθανός ο υπολογισμός μερικής AUC. [25] Για παράδειγμα, μπορούμε να επικεντρωθούμε στην περιοχή της καμπύλης που έχει χαμηλό false positive rate, που είναι συνήθως πολύ σημαντικό για population screening tests. [26]

Το στατιστικό ROC AUC χρησιμοποιείται συχνά για σύγκριση μοντέλων. [27] Ωστόσο, αυτή η πρακτική έχει αμφισβητηθεί πρόσφατα με βάση μια νέα έρευνα μηχανικής μάθησης που δείχνει ότι η AUC είναι αρκετά θορυβώδης ως μέτρο ταξινόμησης [28] και έχει κάποια άλλα σημαντικά προβλήματα στη σύγκριση των μοντέλων. [29][30] Μια αξιόπιστη και έγκυρη εκτίμηση AUC μπορεί να ερμηνευτεί ως η πιθανότητα ο ταξινομητής να εκχωρήσει υψηλότερη βαθμολογία σε ένα τυχαία επιλεγμένο θετικό παράδειγμα από ένα τυχαία επιλεγμένο αρνητικό παράδειγμα. Ωστόσο, η έρευνα [28][29] υποδηλώνει συχνές αποτυχίες στην απόκτηση αξιόπιστων και έγκυρων εκτιμήσεων AUC. Έτσι, αμφισβητήθηκε η πρακτική αξία του μέτρου AUC [30], αυξάνοντας έτσι το ενδεχόμενο η AUC να εισαγάγει στην πράξη περισσότερη αβεβαιότητα στις συγκρίσεις της ταξινόμησης της μηχανικής μάθησης σε σχέση με τις λύσεις που προσφέρει.

Αξίζει να σημειωθεί ότι οι τιμές των μετρικών είναι αντικειμενικές και εξαρτώνται από το πρόβλημα που χρήζει επίλυση, αφού μπορεί να αντιμετωπίζεται πρόβλημα το οποίο είναι δύσκολο και οι καλύτερες βαθμολογίες που μπορούμε να πετύχουμε σε αυτό να είναι μέτριες σε σχέση κάποιο πιο εύκολο που πετυχαίνει αρκετά μεγαλύτερες βαθμολογίες. Συμπερασματικά, στα δύσκολα προβλήματα η καμπύλη ROC δεν θα βρίσκεται πολύ υψηλά και αριστερά, αλλά θα μοιάζει περισσότερο σε αυτή της εικόνας 3.20.

Κεφάλαιο 4 - Πρόβλημα, υλοποίηση σε Python και αποτελέσματα

Περιεχόμενα κεφαλαίου

4.1 Διαγωνισμός SemEval 2018 για Irony Detection	83
4.1.1 Περιγραφή του data set και υπόσταση των tweets	85
4.2 Περιγραφή υλοποίησης	88
4.2.1 Feature Extraction	89
4.2.2 Preprocessing	91
4.2.3 Αλγόριθμοι για encoding του dataset	91
4.2.4 Αλγόριθμοι Feature Selection	98
4.2.5 Αλγόριθμοι Machine Learning	101
4.2.6 Τρόπος αξιολόγησης – Evaluation	124
4.3 Τελικά αποτελέσματα, συμπεράσματα και μελλοντικοί στόχοι	126

4.1 Διαγωνισμός SemEval 2018 για Irony Detection

Η ανάπτυξη των κοινωνικών δικτύων σηματοδότησε διάφορους τρόπους χρήσης της φυσικής γλώσσας, όπως είναι για παράδειγμα η ειρωνεία (irony). Η ειρωνεία χρησιμοποιείται τακτικά στα κοινωνικά μέσα δικτύωσης, κάτι που καθιστά δύσκολη την επεξεργασία της φυσικής γλώσσας σε αλγορίθμικό επίπεδο. Αν και υπάρχουν διαφορετικά είδη ειρωνείας, η γενική ιδέα είναι ότι το πραγματικό νόημα της πρότασης διαφέρει από αυτό που διατυπώνεται στη φυσική γλώσσα. Συνεπώς, η μοντελοποίηση της ειρωνείας ανήκει σε ένα μεγάλο εύρος ερευνητικών πεδίων, όπως για παράδειγμα εξόρυξη κειμένου (text mining), author profiling, ανίχνευση διαδικτυακής παρενόχλησης και sentiment analysis.

Μέθοδοι τύπου rule-based χρησιμοποιήθηκαν στο παρελθόν και στηρίζονται σε πληροφορίες που αντλούνται από το λεξιλόγιο. Επίσης, χρησιμοποιήθηκαν απλοί μέθοδοι machine learning οι οποίοι χρησιμοποιούν δεδομένα εκπαίδευσης και εκμεταλλεύονται πληροφορίες που παρέχονται από διάφορες πηγές, όπως είναι το bag of words, συντακτικά μοτίβα, πληροφορία από συναισθήματα (sentiment information) ή σημασιολογική συγγένεια (semantic re-

latedness). Ωστόσο, αυτοί οι μέθοδοι δεν μπορούν να συγκριθούν με τους πιο πρόσφατους μεθόδους εντοπισμού ειρωνείας αφού αποδίδουν καλύτερα. Τέτοιοι μέθοδοι είναι αλγόριθμοι deep learning που ενσωματώνουν τη σημασιολογική συγγένεια με προηγμένες μεθόδους, όπως είναι για παράδειγμα τα word embeddings.

Στα πλαίσια της πτυχιακής, θα διερευνηθούν αλγόριθμοι Μηχανικής Μάθησης και άλλοι μέθοδοι που χρησιμοποιούνται για Sentiment Analysis. Ποιο συγκεκριμένα, η πτυχιακή επικεντρώνεται σε Irony Detection σε μικρού μεγέθους κείμενα που πραγματεύονται ένα μόνο θέμα στην Αγγλική γλώσσα. Τέτοια κείμενα, όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, θεωρούνται τα tweets του Twitter, αφού περιορίζονται από το άνω όριο των 140 λέξεων και άρα επικεντρώνονται σε ένα θέμα. Ωστόσο, η επαρκής εξόρυξη των tweets από το Twitter είναι χρονοβόρα και εκτός του αντικειμένου που πραγματεύεται η πτυχιακή. Συνεπώς, χρησιμοποιήθηκε dataset από το μέρος Α' του Task 3 του διαγωνισμού SemEval 2018. [31] Ποιο συγκεκριμένα, το dataset ονομάζεται SemEval2018-T3-train-taskA_emoji. Το dataset απαρτίζεται από τρεις στήλες. Η πρώτη στήλη απαριθμεί το σύνολο των tweets, η δεύτερη στήλη είναι το label του tweet και παίρνει δύο τιμές, μηδέν αν το tweet δεν είναι ειρωνικό και ένα αν είναι ειρωνικό. Τέλος, η τρίτη στήλη είναι το ίδιο το tweet.

Το dataset αποτελείται από 3834 tweets στην Αγγλική γλώσσα. Για τη διευκόλυνση της συλλογής των δεδομένων αλλά και την κατηγοριοποίηση τους σε ειρωνικά και μη, χρησιμοποιήθηκαν hashtags που εκφράζουν ειρωνεία, και πιο συγκεκριμένα χρησιμοποιήθηκαν τα #irony, #not και #sarcasm. Πολλές φορές όμως τέτοια hashtags μπορεί να αυξήσουν τον θόρυβο, αφού δεν είναι δεδομένο ότι όταν ένα tweet περιλαμβάνει τα παραπάνω hashtags θα είναι ειρωνικό, για παράδειγμα όταν σε ένα tweet έχουμε τα hashtags #not και #irony μπορεί να σημαίνει είτε ότι δεν έχουμε ειρωνεία (not irony) είτε ότι χρησιμοποιούνται και τα δύο hashtags για να δηλώσουν υπερβολή ή εκνευρισμό. Έτσι, για τη συλλογή των δεδομένων χρησιμοποιήθηκαν τα hashtags #irony, #not και #sarcasm και τέλος η κατηγοριοποίηση τους σε ειρωνικά και μη έγινε με το χέρι χρησιμοποιώντας ένα λεπτομερές σχήμα κατηγοριοποίησης από του φορείς του διαγωνισμού.

Στόχος του Task 3 μέρος Α του διαγωνισμού είναι να χρησιμοποιηθούν μοντέλα που θα προβλέπουν μια δυαδική (binary) μεταβλητή η οποία θα παίρνει τιμή 1 όταν ένα tweet είναι ειρωνικό και τιμή 0 αν δεν είναι ειρωνικό. Υπάρχουν διάφοροι τρόποι ορισμού της ειρωνείας. Για την καλύτερη κατανόηση του task θα χρησιμοποιηθεί ένα παράδειγμα που στοχεύει στην καλύτερη κατανόηση της διαφοράς μεταξύ της ειρωνείας και της μη ειρωνείας. Ένας τρόπος

ορισμού του προβλήματος είναι να ορίσουμε πότε ένα tweet δεν είναι ειρωνικό και να αποφύγουμε τον ορισμό του να είναι ειρωνικό επειδή το πρόβλημα είναι δυαδικό (binary). [31] Έτσι, θεωρούμε ότι δεν έχουμε ειρωνεία όταν το tweet είναι ξεκάθαρα μη ειρωνικό ή δεν είναι ξεκάθαρο αν είναι ειρωνικό. Για παράδειγμα:

- And then my sister should be home from college by time I get home from babysitting.
And it's payday. THIS IS A GOOD FRIDAY
- Please dont fuck with me when I first wake up #not a morning person!

Οι παρακάτω προτάσεις είναι παραδείγματα ενός ειρωνικού και ενός μη ειρωνικού tweet, αντίστοιχα.

- I just love when you test my patience!! #not
- Had no sleep and have got school now #not happy

Ο σκοπός της πτυχιακής δεν είναι η συμμετοχή στον διαγωνισμό λόγω τον προθεσμιών που δεν συμπίπτουν με αυτών της πτυχιακής, Ωστόσο, η διαδικασία υλοποίησης ακολουθεί τα πρότυπα του διαγωνισμού. Πιο συγκεκριμένα, ο διαγωνισμός παρέχει train set (SemEval2018-T3-train-taskA_emoji, αναφέρθηκε παραπάνω) και test set, αλλά το test set δεν έχει labels που δηλώνουν αν ένα tweets είναι ειρωνικό. Η δημοσιοποίηση του test set με τα irony labels θα γίνει μετά το πέρας της ανάπτυξης των μοντέλων.

Έτσι, για να προσομοιωθεί αυτή η κατάσταση, κατά την ανάπτυξη του μοντέλου της πτυχιακής θα χρησιμοποιηθεί μόνο το train set και μετά το πέρας όλης της διαδικασίας τα μοντέλα θα δοκιμαστούν με ολόκληρο το train set και το test set (SemEval2018-T3_gold_test_taskA_emoji) για να προσομοιωθούν πραγματικές καταστάσεις και να παραχθεί ένα γενικευμένο μοντέλο που έχει καλή απόδοση με οποιοδήποτε test set, δηλαδή να αποφευχθεί το overfitting από παραμετροποίηση. Για να εκπαιδευτεί και να μπορεί να κάνει προβλέψεις το μοντέλο, θα χρησιμοποιηθεί 10-fold-cross-validation που χωρίζει το αρχικό train set σε δύο κομμάτια, όπου το ένα κομμάτι χρησιμοποιείται για train set ενώ το δεύτερο χρησιμοποιείται για test set. Περισσότερες λεπτομέρειες θα αναφερθούν παρακάτω.

4.1.1 Περιγραφή data set και υπόσταση των tweets

Τα tweets που περιέχει το data set (SemEval2018-T3-train-taskA_emoji) αποτελούνται από γράμματα του αγγλικού αλφάριθμου, κεφαλαία και μικρά, καθώς και σημεία στίξης και μη αλφαριθμητικούς χαρακτήρες. Στην κατηγορία των μη αλφαριθμητικών χαρακτήρων ανήκουν τα emojis, για παράδειγμα ☺, ❤, 😊. Υπάρχουν όμως και emojis που σχηματίζονται από

συνδυασμό γραμμάτων της αλφαβήτας με σημεία στίξης αλλά και συνδυασμό σημείων στίξης μεταξύ τους, όπως για παράδειγμα :p, ;). Επιπλέον, ένα tweet μπορεί να περιέχει διαδικτυακά links ή αναφορές (references) χρηστών, για παράδειγμα @user_name. Ένα πολύ σημαντικό στοιχείο των tweets είναι το hashtag (#). Το hashtag είναι θεμελιώδης χαρακτηριστικό του Twitter αφού πολλές φορές προσδιορίζει το θέμα αλλά και το περιεχόμενο του tweet.

Το Twitter εκμεταλλευόμενο αυτό το γεγονός έχει δημιουργήσει αναζήτηση των tweets με τη χρήση των hashtags. Επίσης, με τη χρήση των hashtags τα tweets οργανώνονται σε κατηγορίες, όπως είναι οι τάσεις (trending). Όσο αναφορά τα προαναφερθέντα, στα πλαίσια της πτυχιακής θα μελετηθεί η σημαντικότητα των κεφαλαίων λέξεων, των θαυμαστικών (!), των ερωτηματικών (?), των αποσιωπητικών (...), των hashtags (#), των αναφορών - mentions (@user_name), των παραθέσεων – quotes (“ ”), των διαδικτυακών links (URLs) καθώς και των emoticons – emojis. Περισσότερες λεπτομέρειες θα αναφερθούν παρακάτω.

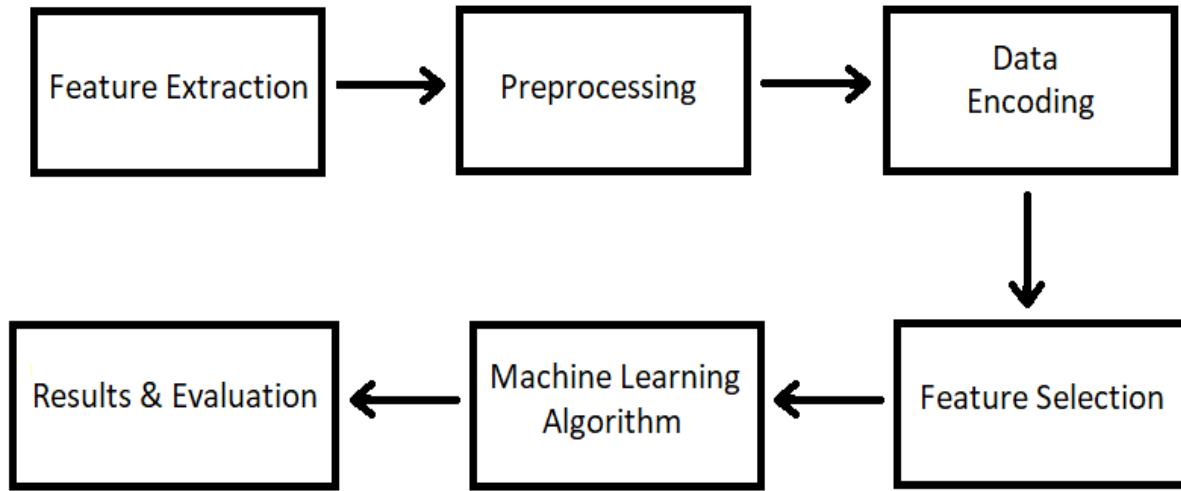
Μετά την περιγραφή της υπόστασης των tweets, το επόμενο βήμα είναι να γνωρίσουμε το data set. Για την επίτευξη αυτού υπολογίστηκαν κάποια στατιστικά τα οποία παρουσιάζονται στο πίνακα 4.1. Από τα στατιστικά βλέπουμε ότι το συνολικό πλήθος των hashtags είναι σχεδόν ίσο με το πλήθος των συνολικών tweets, κάτι που σε συνδυασμό με τη σημασιολογία τους υποδεικνύει τη σημαντικότητα τους. Αν και τα hashtags παρέχουν χρήσιμη πληροφορία για το περιεχόμενο και το θέμα ενός tweet, η επιρροή τους στο αν είναι ειρωνικό ή όχι θα πρέπει να μελετηθεί περαιτέρω. Επίσης, παρατηρούμε ότι οι κεφαλαίες λέξεις και οι αναφορές (mentions) χρησιμοποιούνται τακτικά, αλλά θα πρέπει να μελετηθεί περαιτέρω η σημαντικότητα τους.

Συνολικό πλήθος tweets: 3834	Αθροισμα από όλα τα tweet	Μέσος όρος ανά tweet
Κεφαλαίες λέξεις	2636	0.6875326030255607
Θαυμαστικά	1120	0.2921231090245175
Ερωτηματικά	532	0.1387584767866458
Αποσιωπητικά	553	0.1442357850808555
Hashtags	3709	0.967396974439228
Αναφορές (mentions)	2419	0.6309337506520605
Παραθέσεις (quotes)	525	0.13693270735524257
URLs	945	0.24647887323943662
Emoticons	915	0.23865414710485133

Πίνακας 4.1 Στατιστικά του data set

4.2 Περιγραφή υλοποίησης

Για την υλοποίηση του μοντέλου χρησιμοποιήθηκαν αλγόριθμοι κωδικοποίησης κειμένου σε αριθμούς, αλγόριθμοι μηχανικής μάθησης, αλγόριθμοι για feature selection κ.α. Ο γενικός τρόπος λειτουργίας του μοντέλου απεικονίζεται στην εικόνα 4.2.



Εικόνα 4.2 Τρόπος λειτουργίας του μοντέλου

Πιο συγκεκριμένα, το πρόγραμμα ξεκινάει από το αρχείο main.py όπου δίνει τον έλεγχο στο αρχείο ClassRead.py το οποίο περιέχει την κλάση Reader η οποία είναι υπεύθυνη για την ανάγνωση των δεδομένων από το αρχείο αλλά και για το preprocessing καθώς και για το feature extraction και το feature selection. Στη κλάση Reader πρώτα διαβάζονται τα δεδομένα από το αρχείο SemEval2018-T3-train-taskA_emoji.txt και αποθηκεύονται σε ένα data frame που ονομάζεται train_A. Αμέσως μετά, καλείται η συνάρτηση pre_processing στην οποία εκτελείται η λειτουργία του feature extraction και στη συνέχεια εκτελούνται διαδικασίες καθαρισμού των δεδομένων. Στη συνέχεια, επιστρέφεται ο έλεγχος στο αρχείο main.py όπου καλείται ο αλγόριθμος μηχανικής μάθησης. Όλα τα αρχεία που περιέχουν τους αλγορίθμους μηχανικής μάθησης ακολουθούν το ίδιο μοτίβο όσο αναφορά την δομή τους. Πιο συγκεκριμένα, στην αρχή γίνεται αρχικοποίηση κάποιων μεταβλητών και στη συνέχεια υλοποιείται 10-fold-cross-validation.

Μέσα στο 10-fold-cross-validation καλείται δύο φορές η συνάρτηση get_enc της κλάσης Reader του αρχείου ClassRead.py, μία φορά για τα train δεδομένα και μια φορά για τα test δεδομένα έτσι ώστε τα χώρισε το 10-fold-cross-validation. Η συνάρτηση get_enc είναι υπεύθυνη για την κωδικοποίηση των δεδομένων κειμένου σε αριθμούς αλλά και την λειτουργία του feature selection. Έπειτα, δημιουργείται ο machine learning classifier και γίνεται train

στα δεδομένα που επέστρεψε η συνάρτηση `get_enc`. Μόλις γίνει το `train` του classifier υπολογίζονται κάποιες μετρικές για τον classifier και στη συνέχεια γίνεται το `predict` του label των test δεδομένων που δηλώνει αν ένα tweet είναι ειρωνικό ή όχι. Τέλος, υπολογίζονται οι μετρικές αξιολόγησης του αλγορίθμου. Οι διαδικασίες, που απεικονίζονται και στην εικόνα 4.2, θα περιγραφούν πιο αναλυτικά στα παρακάτω υποκεφάλαια.

Για την προγραμματιστική υλοποίηση χρησιμοποιήθηκαν δεκατέσσερις βιβλιοθήκες. Για την δημιουργία σχεδιαγραμμάτων χρησιμοποιήθηκαν οι βιβλιοθήκες ‘matplotlib’ και ‘scipy’, ενώ για τη δημιουργία εικόνων world cloud χρησιμοποιήθηκε η βιβλιοθήκη ‘wordcloud’. Οι βιβλιοθήκες ‘emoji’ και ‘re’ χρησιμοποιούνται στο Feature Extraction. Επίσης, χρησιμοποιούνται οι βιβλιοθήκες ‘numpy’, ‘sklearn’, ‘keras’, ‘ntlk’, ‘string’, ‘gensim’, ‘pandas’, ‘os’ και ‘pathlib’. Ως back-end χρησιμοποιείται το tensorflow με υποστήριξη κάρτας γραφικών.

4.2.1 Feature Extraction

Η διαδικασία του Feature Extraction είναι το πρώτο πράγμα που εκτελείται όσο αναφορά την επεξεργασία των δεδομένων. Σκοπός του Feature Extraction είναι η καταμέτρηση διάφορων στοιχείων που μπορεί να περιέχει ένα tweet. Κατά το Feature Extraction προστίθενται 9 καινούργια features, που υπάρχουν και στον πίνακα 4.1. Τα νέα features κανονικοποιούνται με τη χρήση MinMaxScaler της sklearn στο (0, 1) πριν προστεθούν στα features του αλγορίθμου encoding που χρησιμοποιείται, ώστε να αντιμετωπιστούν ισόβαθμα από τους classifiers. Τα features με τη σειρά που προστίθενται στα δεδομένα είναι: πλήθος κεφαλαίων λέξεων (upper case words), πλήθος θαυμαστικών (!), πλήθος ερωτηματικών (?), πλήθος αποσιωπητικών (...), πλήθος hashtags (#), πλήθος αναφορών (@ - mentions), πλήθος παραθέσεων (”, “” - quotes), πλήθος URL και τέλος πλήθος emoticons. Τα πλήθη των παραπάνω στοιχείων υπολογίζονται για κάθε ένα tweet ξεχωριστά, δηλαδή έχουν μορφή (3834,9) όπου το 3834 είναι το πλήθος των tweets και το 9 είναι το σύνολο των προστιθέμενων features. Η καταμέτρηση των παραθέσεων - quotes, των URL και των emoticons απαιτούσαν ειδική διαχείριση.

Για τα quotes έπρεπε να καταμετρηθούν τα “” και τα “”, οπότε γίνεται καταμέτρηση όλων των μονών quotes που υπάρχουν σε ένα tweet, δηλαδή υπολογίζεται το πλήθος των ‘ και των “ που υπάρχουν σε ένα tweet, και στο τέλος διαιρείται το πλήθος με το 2 για να πάρουμε το κανονικό πλήθος των quotes (αφού τα quotes περιέχουν δύο κομμάτια, ένα για να ανοίξουν και ένα για να κλείσουν). Όμως ο υπολογισμός του πλήθους των ‘ προκαλεί πρόβλημα αφού τα ‘ χρησιμοποιούνται και σε προτάσεις σαν απόστροφοι, για παράδειγμα “It’s getting late”.

Έτσι, στο τέλος, για κάθε tweet γίνεται έλεγχος του πλήθους των παραθέσεων και αν βρεθεί ότι δεν είναι ζυγός αριθμός τότε αφαιρούμε ένα από το πλήθος. Αν και αυτή η στρατηγική δεν βρίσκει ακριβώς το πλήθος των quotes για όλα τα tweets αλλά μια κοντινή προσέγγιση, λειτουργεί γιατί κατά μέσο όρο ένα tweet αποτελείται από 15 λέξεις, οπότε δεν υπάρχουν πολλά περιθώρια ένα tweet να περιέχει πολλές παραθέσεις και άρα μετά την κανονικοποίηση των νέων features το σφάλμα περιορίζεται ακόμα περισσότερο σε αμελητέα επίπεδα.

Επίσης, έγιναν αναζητήσεις στα δεδομένα που είχαν quotes και διαπιστώθηκε ότι δεν συναντάται τακτικά αυτή περίπτωση. Έτσι, ο συνδυασμός του μικρού πλήθους παραθέσεων ανά tweet, με τις αναζητήσεις και τους ελέγχους και με τον περιορισμό που θέσαμε, δηλαδή το πλήθος των μονών παραθέσεων (“και”) να είναι ζυγός αλλιώς αφαιρούμε ένα από το πλήθος, περιορίζει το σφάλμα σε πολύ μικρά νούμερα που θεωρούνται αμελητέα.

Για τα URL χρησιμοποιήθηκε regular expression που εντόπιζε τα “`http\S+`” ή “`www.\S+`”, δηλαδή εντόπιζε τους όρους που ξεκινούσαν με “`http`” ή “`www.`” και τελείωναν με οποιαδήποτε συμβολοσειρά.

Τέλος, για τον εντοπισμό των emoticons αρχικά καθαρίστηκε το data set από τα URLs γιατί εντοπίζόταν το emoji “`:`” στο “`https://...`”. Στη συνέχεια, χρησιμοποιήθηκε η βιβλιοθήκη ‘emoji’ για τον εντοπισμό των emoticons. Το αρχείο δεδομένων που χρησιμοποιείται περιέχει emoticons σε σχήμα (π.χ.) αλλά και emoticons με συνδυασμό γραμμάτων και σημεία στίξης (π.χ. `:D`). Η βιβλιοθήκη ‘emoji’ διαχειρίζεται μόνο τα emoticons με σχήμα, δηλαδή με μη αλφαριθμητικούς χαρακτήρες. Συνεπώς, για τον εντοπισμό των emoticons με συνδυασμό γραμμάτων και σημείων στίξης χρησιμοποιείται η κλάση EmoticonDetector του αρχείου EmoticonDetector.py. Η κλάση αυτή στο initialize φορτώνει στη μνήμη το αρχείο emoticons.txt το οποίο περιέχει emoticons δημιουργημένα με αλφαριθμητικούς χαρακτήρες και σημεία στίξης. Κάθε emoji υπάρχει δύο φορές στο αρχείο, μία φορά με κενό στο τέλος του και μια φορά με το “`|`” που χρησιμοποιείται σαν διαχωριστικό αντί για κενό, για να αποφευχθεί ο εντοπισμός emoticons που σχηματίζονται με απλή χρήση λόγου. Η συνάρτηση count_emoticons της κλάσης EmoticonDetector καταμετράει τα emoticons που αντλούνται από το αρχείο emoticons.txt και υπάρχουν μέσα στα tweets. Το τελικό πλήθος των emoticons για κάθε tweet προκύπτει από το άθροισμα των δύο πληθών που υπολογίστηκαν.

Αν είχαμε διαφορετικό data set θα μπορούσαμε να προσθέσουμε σαν feature το πλήθος των hashtags που δηλώνουν ειρωνεία για κάθε tweet, για παράδειγμα `#not`, `#irony` και `#sarcasm`, και να δούμε πως επηρεάζει την αποτελεσματικότητα. Στο συγκεκριμένο data set κάτι

τέτοιο δεν είναι εφικτό αφού η συλλογή των tweets έγινε με την αναζήτηση των #not, #irony και #sarcasm στο Twitter, οπότε τα μοντέλα θα επικέντρωναν όλο το βάρος της πρόβλεψης πάνω σε αυτό το feature και θα πετύχαιναν πάρα πολύ υψηλά αποτελέσματα. Αυτός είναι και ο λόγος που δεν θα δημοσιοποιηθεί gold set που θα περιέχει τα #not, #irony και #sarcasm.

4.2.2 Preprocessing

Η διαδικασία του preprocessing εκτελείται στην συνάρτηση pre_processing της κλάσης Reader του αρχείου ClassRead.py. Αρχικά, από τα tweet αφαιρούνται τα URLs, τα mentions των usernames και τα hashtags χωρίς όμως να αφαιρείται και η λέξη που βρίσκεται μετά το hashtag. Στη συνέχεια, γίνεται tokenization του κάθε tweet με τη χρήση της βιβλιοθήκης “nltk”. Αφαιρούνται τα σημεία στίξης (punctuation) από κάθε tweet καθώς και όλοι οι μη αλφαριθμητικοί χαρακτήρες. Επίσης, γίνεται η διαγραφή των αγγλικών stop-words εκτός από τις λέξεις “not” και “n’t”. Τέλος, εκτελείται stemming στις λέξεις των tweets, δηλαδή αφαιρούνται οι καταλήξεις από τις λέξεις και το τελικό αποτέλεσμα είναι ουσιαστικά ένα είδος ρίζας της λέξης.

4.2.3 Αλγόριθμοι για encoding του dataset

Οι αλγόριθμοι μηχανικής μάθησης απαιτούν τα δεδομένα να είναι σε αριθμητική μορφή. Έτσι, πρέπει να μετατρέψουμε τα tweets από τα δεδομένα κειμένου σε αντίστοιχη αριθμητική αναπαράσταση. Για την επίτευξη αυτού του στόχου χρησιμοποιούνται έξι αλγόριθμοι, οι οποίοι είναι: TF-IDF, One-Hot-Encoding, Bigrams με One-Hot-Encoding, word2vec, doc2vec και GloVe (Global Vectors).

Περιγραφή αλγορίθμου TF-IDF

Η διαδικασία κωδικοποίησης με TF-IDF περιγράφηκε αναλυτικά στο κεφάλαιο 3.1.5 αλλά παρακάτω θα δοθεί ένα παράδειγμα για την καλύτερη κατανόηση της διαδικασίας. Το TF-IDF υπολογίζεται από τον τύπο: $TF-IDF = tf * idf$, όπου το tf είναι η συχνότητα του όρου (term frequency) και το idf είναι η αντίστροφη συχνότητα εγγράφου (inverse document frequency). Έστω ότι έχουμε ένα έγγραφο το οποίο περιέχει 100 λέξεις και ότι η λέξη “γάτα” εμφανίζεται 3 φορές. Η συχνότητα του όρου (term frequency - tf) για την λέξη “γάτα” είναι ίση με $(3/100) = 0.03$. Επίσης, έστω ότι έχουμε 10 εκατομμύρια έγγραφα και η λέξη “γάτα” εμφανίζεται σε χίλια από αυτά. Τότε, η αντίστροφη συχνότητα εγγράφου (inverse document

frequency – idf) υπολογίζεται ως $\log(10.000.000 / 1.000) = 4$. Τελικά, το TF-IDF για τον όρο “γάτα” είναι ίσο με το γινόμενο των δύο αυτών τιμών, δηλαδή είναι ίσο με: $0.03 * 4 = 0.12$.

Προγραμματιστικά από την βιβλιοθήκη sklearn εισάγεται το TfidfVectorizer το οποίο γίνεται μία φορά fit_transform για τα train δεδομένα και μια φορά σκέτο transform για τα test δεδομένα. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform ο TfidfVectorizer από την αρχή για τα νέα train δεδομένα και στην συνέχεια transform για τα νέα test δεδομένα. Στην δήλωση του TfidfVectorizer ορίζονται τέσσερα ορίσματα, τα οποία είναι τα εξής: lowerCase=False, analyzer='word', tokenizer με τιμή dummy_fun και preprocessor=dummy_fun. Το πρώτο όρισμα είναι το lowerCase που ορίζεται σε False ώστε να μην μετατραπούν όλοι οι χαρακτήρες του tweet σε πεζά γράμματα (lower case). Το δεύτερο όρισμα είναι το analyzer, το οποίο ορίζεται σε ‘word’ και σηματοδοτεί ότι το TF-IDF feature θα παραχθεί από λέξεις.

Το τρίτο όρισμα λειτουργεί μόνο στην περίπτωση του analyzer='word' και είναι το tokenzier, το οποίο δέχεται την συνάρτηση dummy_fun. Η συνάρτηση dummy_fun επιστρέφει αυτό που της στάλθηκε σαν όρισμα, δηλαδή το όρισμα tokenzier με αυτή τη συνάρτηση έχει ως σκοπό να παρακάμψει το tokenzier που παρέχεται από το TfidfVectorizer γιατί τα tweets έχουν ήδη περάσει από τη διαδικασία του tokenization κατά το preprocessing. Το τέταρτο όρισμα είναι το preprocessor, το οποίο δέχεται την συνάρτηση dummy_fun. Η συνάρτηση dummy_fun, όπως αναφέρθηκε πάνω, επιστρέφει αυτό που της στάλθηκε σαν όρισμα, και έχει σκοπό την παράκαμψη του preprocessing που παρέχεται από το TfidfVectorizer με το όρισμα preprocessor, γιατί τα tweets έχουν ήδη περάσει από τη διαδικασία του preprocessing.

Περιγραφή αλγορίθμου One-Hot-Encoding

Η κωδικοποίηση One-Hot-Encoding είναι ουσιαστικά η αναπαράσταση κατηγορικών μεταβλητών (categorical variables) ως δυαδικά διανύσματα (binary vectors). Αυτό όμως απαιτεί πρώτα οι κατηγορικές τιμές να αντιστοιχηθούν σε ακέραιες τιμές. Στη συνέχεια, κάθε ακέραια τιμή αναπαριστάται από ένα δυαδικό διάνυσμα που έχει σε όλες τις θέσεις του μηδενικά εκτός από την θέση που αντιστοιχεί στην συγκεκριμένη ακέραια τιμή, η οποία ισούται με ένα. Με άλλα λόγια, η κεντρική ιδέα της κωδικοποίησης με One-Hot-Encoding είναι να σχηματιστεί ένας πίνακας που περιέχει μηδέν και ένα. Σε αυτόν τον πίνακα οι γραμμές θα συμβολίζουν τα tweet ενώ οι στήλες θα συμβολίζουν κάθε μια λέξη που υπάρχει στα tweets.

Πιο συγκεκριμένα, όταν μια λέξη υπάρχει μέσα σε ένα tweet, η θέση που αντιστοιχεί στο tweet και στην λέξη θα έχει τιμή ένα ενώ για αυτές τις λέξεις που δεν υπάρχουν στο συγκεκριμένο tweet θα έχουν τιμή μηδέν.

Προγραμματιστικά, για την κωδικοποίηση One-Hot-Encoding από την βιβλιοθήκη sklearn χρησιμοποιείται ο CountVectorizer, στον οποίο γίνεται μία φορά fit_transform για τα train δεδομένα και μια φορά σκέτο transform για τα test δεδομένα. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform ο CountVectorizer από την αρχή για τα νέα train δεδομένα και στην συνέχεια transform για τα νέα test δεδομένα. Σαν όρισμα στον CountVectorizer δίνουμε τέσσερα ορίσματα, το analyzer='word', tokenizer=dummy_fun, lowercase=False και το binary=True. Το όρισμα analyzer ορίζει ότι τα χαρακτηριστικά θα δημιουργηθούν από λέξεις.

Το όρισμα tokenizer χρησιμοποιείται για το tokenization των λέξεων, αλλά επειδή έχει γίνει ήδη κατά το preprocessing χρησιμοποιούμε την συνάρτηση dummy_fun η οποία επιστρέφει το όρισμα που δέχεται, δηλαδή δεν εφαρμόζεται καμία διεργασία στο token. Το όρισμα lowercase χρησιμοποιείται για να μετατρέψει όλες τις λέξεις σε λέξεις με πεζά γράμματα, αλλά επειδή γίνεται ήδη αυτό κατά το preprocessing του δίνουμε την τιμή False. Ο CountVectorizer υπολογίζει για κάθε tweet το πλήθος εμφάνισης της κάθε λέξης στο συγκεκριμένο tweet, έτσι δίνουμε το όρισμα binary το οποίο θέτει σε όλα τα μη μηδενικά αθροίσματα των λέξεων την τιμή 1, ώστε να δημιουργηθεί η one-hot αναπαράσταση. Το παρακάτω παράδειγμα δείχνει αναλυτικά την διαδικασία που περιγράφηκε για την καλύτερη κατανόηση του One-Hot-Encoding. Έστω ότι έχουμε δύο tweets:

- “I like the warm weather of Brazil”
- “The weather in Brazil is tropical”

Το πρώτο tweet θεωρούμε ότι ανήκει στα δεδομένα εκπαίδευσης, άρα το περνάμε στον CountVectorizer με fit_transform, ενώ το δεύτερο tweet θεωρούμε ότι ανήκει στα testing δεδομένα, άρα το περνάμε στον CountVectorizer με transform. Στη συνέχεια, ο CountVectorizer από το fit_transform με το πρώτο tweet σχηματίζει λεξικό (vocabulary) από τις λέξεις του, όπως φαίνεται από τις στήλες του παρακάτω πίνακα. Το δεύτερο tweet από το transform δεν αλλάζει το λεξικό. Έτσι με σταθερό λεξικό ο αλγόριθμος τοποθετεί 1 στις θέσεις των λέξεων του δεύτερου tweet που εμφανίζονται στο λεξικό που σχηματίστηκε από τα δεδομένα εκπαίδευσης. Οι λέξεις του δεύτερου tweet που δεν υπάρχουν στο λεξικό αγνοούνται. Αυτό συμ-

βαίνει γιατί ανήκει στα test δεδομένα και άρα εφαρμόζεται απλό transform πάνω του. Έτσι παράγεται ο παρακάτω πίνακας, αφού μετατραπούν όλα τα γράμματα των λέξεων σε πεζά. Ο πίνακας έχει ως στήλες τις μοναδικές λέξεις των δύο tweets, δηλαδή υπάρχουν 7 στήλες, και οι γραμμές συμβολίζουν τα δύο tweets, δηλαδή έχουμε 2 γραμμές.

	I	like	the	warm	weather	of	Brazil
Tweet 1	1	1	1	1	1	1	1
Tweet 2	0	0	1	0	1	0	1

Δηλαδή, η One-Hot-Encoding αναπαράσταση του πρώτου tweet είναι “1 1 1 1 1 1” ενώ η αναπαράσταση του δεύτερου tweet είναι “0 0 1 0 1 0 1”.

Περιγραφή αλγορίθμου Bigrams με One-Hot-Encoding

Η απλή κωδικοποίηση με One-Hot-Encoding δεν λαμβάνει υπόψη τις σχέσεις μεταξύ των λέξεων ενός tweet. Έτσι, η κωδικοποίηση με Bigrams με One-Hot-Encoding δίνει μερική λύση σε αυτό το πρόβλημα. Η μόνη διαφορά με την απλή κωδικοποίηση με One-Hot-Encoding είναι ότι πρώτα σχηματίζονται Bigrams από τις λέξεις ενός tweet και στη συνέχεια εφαρμόζεται ο CountVectorizer. Bigrams ενός tweet είναι ουσιαστικά μια πλειάδα που σχηματίζεται από δύο λέξεις που βρίσκονται η μια μετά την άλλη στο tweet. Δηλαδή, έστω τα tweet από το παραπάνω παράδειγμα:

- “I like the warm weather of Brazil”
- “The weather in Brazil is tropical”

Τα Bigrams για το πρώτο tweet είναι: (I, like), (like, the), (the, warm), (warm, weather), (weather, of), (of, Brazil) και για το δεύτερο tweet είναι: (The, weather), (weather, in), (in, Brazil), (Brazil, is), (is, tropical). Η διαδικασία είναι η ίδια απλά αντί να δίνουμε στον CountVectorizer μία λέξη τη φορά δίνουμε μία πλειάδα τη φορά. Στο παραπάνω παράδειγμα όλες οι πλειάδες είναι μοναδικές, οπότε ο CountVectorizer θα δώσει τιμή ένα σε όλες τις πλειάδες του πρώτου tweet, ενώ σε όλες τις πλειάδες του δεύτερου tweet θα δώσει 0. Δηλαδή για το πρώτο tweet το CountVectorizer θα παράγει (1, 1, 1, 1, 1, 1) ενώ για το δεύτερο tweet θα παράγει (0, 0, 0, 0, 0, 0). Ο παραπάνω πίνακας για το τρέχον παράδειγμα και την μέθοδο με τα bigrams με One-Hot-Encoding σχηματίζεται ως:

	(I, like)	(like, the)	(the, warm)	(warm, weather)	(weather, of)	(of, Brazil)
Tweet 1	1	1	1	1	1	1
Tweet 2	0	0	0	0	0	0

Προγραμματιστικά για την παραγωγή των Bigrams χρησιμοποιείται το ngrams της βιβλιοθήκης nltk. Στη συνέχεια γίνεται χρήση του CountVectorizer από τη βιβλιοθήκη sklearn. Σαν όρισμα στον CountVectorizer δίνουμε τέσσερα ορίσματα, το analyzer='word', tokenizer=dummy_fun, lowercase=False και το binary=True. Το όρισμα analyzer ορίζει ότι τα χαρακτηριστικά θα δημιουργηθούν από λέξεις. Το όρισμα tokenizer χρησιμοποιείται για το tokenization των λέξεων, αλλά επειδή έχει γίνει ήδη κατά το preprocessing χρησιμοποιούμε την συνάρτηση dummy_fun η οποία επιστρέφει το όρισμα που δέχεται, δηλαδή δεν εφαρμόζεται καμία διεργασία στο token. Το όρισμα lowercase χρησιμοποιείται για να μετατρέψει όλες τις λέξεις σε λέξεις με πεζά γράμματα, αλλά επειδή γίνεται ήδη αυτό κατά το preprocessing του δίνουμε την τιμή False. Επίσης, στη συνάρτηση ngrams της βιβλιοθήκης nltk δίνουμε σαν όρισμα τα tweets και τον αριθμό 2 που σηματοδοτεί ότι θα παραχθούν bigrams.

Ο CountVectorizer γίνεται μία φορά fit_transform για τα train δεδομένα και μια φορά σκέτο transform για τα test δεδομένα. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform ο CountVectorizer από την αρχή για τα νέα train δεδομένα και στην συνέχεια transform για τα νέα test δεδομένα. Σε κάποια προβλήματα εκτός από bigrams χρησιμοποιούνται και trigrams ή γενικότερα ngrams. Στο πρόβλημα μας όμως αυτά δεν θα είχαν καλή απόδοση γιατί τα tweet έχουν κατά μέσο όρο 15 λέξεις και το train set δεν είναι αρκετά μεγάλο ώστε να υπάρχουν αρκετοί συνδυασμοί λέξεων για trigrams ή ngrams. Δηλαδή, ο πίνακας του One-Hot-Encoding θα ήταν αρκετά αραιός αφού τέτοιοι συνδυασμοί των λέξεων είναι σπάνιοι σε τέτοιο μικρό δείγμα δεδομένων, ενώ παράλληλα και ο πίνακας του One-Hot-Encoding που σχηματίζεται από τα bigrams είναι και αυτός αρκετά αραιός.

Περιγραφή αλγορίθμου Word2Vec

Η κωδικοποίηση με word2vec [7] περιγράφηκε αναλυτικά στο κεφάλαιο 3.1.5. Προγραμματιστικά χρησιμοποιείται το Word2Vec της βιβλιοθήκης gensim με το μοντέλο skip-gram [32], το οποίο ορίζεται με την παράμετρο sg=1, ενώ τα διανύσματα που παράγει για τις λέξεις έχουν μέγεθος 100, το οποίο ορίζεται με την παράμετρο size=100. Στο Word2Vec χρησιμοποιείται άλλη μια παράμετρος, το min_count=0, το οποίο χρησιμοποιείται για να αγνοηθούν οι σπάνιες λέξεις που έχουν συνολικό πλήθος εμφάνισης μικρότερο του min_count, αλλά επειδή του δίνουμε την τιμή μηδέν δεν αγνοείται καμιά λέξη. Επίσης, χρησιμοποιείται

η παράμετρος window με default τιμή 5. Η παράμετρος window ορίζει την μέγιστη απόσταση μεταξύ της τρέχουσας και της προβλεπόμενης λέξης μέσα στην πρόταση.

Το μοντέλο word2vec παράγει διανύσματα για κάθε μια λέξη ενός tweet ξεχωριστά. Για να χρησιμοποιηθούν τα διανύσματα που παράγει το word2vec από αλγορίθμους μηχανικής μάθησης πρέπει οι λέξεις σε κάθε ένα tweet να έχουν το ίδιο πλήθος. Υπάρχουν διάφοροι μέθοδοι που χρησιμοποιούνται, αλλά εμείς υπολογίζουμε το μέσο όρο των διανυσμάτων κάθε λέξης. Αυτό θα μας δώσει σαν αποτέλεσμα το που βρίσκεται γενικά ολόκληρο το tweet στο χώρο, κρατώντας τις ιδιότητες που έχουν τα διανύσματα των λέξεων ξεχωριστά.

Για να βελτιώσουμε ακόμα περισσότερο το αποτέλεσμα υπολογίζεται το TF-IDF της κάθε λέξης του tweet, το οποίο το πολλαπλασιάζουμε με το vector που παράχθηκε από το word2vec και στην συνέχεια υπολογίζουμε τον μέσο όρο κάθε διανύσματος που ανήκει σε ένα tweet, όπως περιγράφηκε παραπάνω. Ουσιαστικά, με τον πολλαπλασιασμό με το TF-IDF της κάθε λέξης δίνουμε βάρος στην κάθε λέξη, το οποίο δηλώνει πόσο σημαντική είναι η λέξη στο σύνολο δεδομένων. Για να επιτευχθούν τα παραπάνω αρχικά πρέπει να φέρουμε τα tweets σε μορφή συμβατή με το word2vec. Για να γίνει αυτό χρησιμοποιείται η συνάρτηση labelizeTweets η οποία αλλάζει την μορφή του tweet από (word_1, word_2, ..., word_n) σε ((word_1, word_2, ..., word_n), Label_x).

Στη συνέχεια, γίνεται μία φορά build_vocab και train στο Word2Vec για τα train δεδομένα. Επίσης, τα train δεδομένα γίνονται fit_transform σε έναν TfidfVectorizer για τον υπολογισμό του μέσου όρου όλων των διανυσμάτων των λέξεων ενός tweet που αναφέρθηκε παραπάνω. Για τα test δεδομένα εφαρμόζεται μόνο transform στον TfidfVectorizer. Τέλος και για τις δύο περιπτώσεις, δηλαδή για το train set και το test set, καλείται η συνάρτηση build-WordVector η οποία υπολογίζει το μέσο όρο των διανυσμάτων των λέξεων ενός tweet με τη βοήθεια του TF-IDF κάθε λέξης, όπως περιγράφηκε παραπάνω.

Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται build_vocab και train στο Word2Vec και fit_transform στο TfidfVectorizer από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο TfidfVectorizer για τα νέα test δεδομένα. Στις λέξεις που τροφοδοτούνται στο μοντέλο word2vec και δεν υπάρχουν, δηλαδή δεν υπήρχαν στο train set, αντιστοιχίζεται ένα vector που αποτελείται μόνο από μηδενικά. Το word2vec του test set υπολογίζεται κανονικά όπως περιγράφηκε παραπάνω με τη μόνη διαφορά να είναι ότι οι λέξεις που δεν υπήρχαν στο train set θα έχουν μηδενικά διανύσματα.

Περιγραφή αλγορίθμου Doc2Vec

Η κωδικοποίηση doc2vec [33] είναι η αναπαράσταση ενός ολόκληρου εγγράφου ή πρότασης, στο πρόβλημα μας η αναπαράσταση ενός tweet, από ένα διάνυσμα στο χώρο. Προγραμματιστικά χρησιμοποιείται το Doc2Vec της βιβλιοθήκης gensim με το μοντέλο distributed bag of words (PV-DBOW), το οποίο ορίζεται με το όρισμα $dm=0$, ενώ τα διανύσματα που παράγει για τα tweet έχουν μέγεθος 100, το οποίο ορίζεται με το όρισμα $vector_size=100$. Στο Doc2Vec χρησιμοποιείται άλλο ένα όρισμα, το $min_count=0$, το οποίο χρησιμοποιείται για να αγνοηθούν οι σπάνιες λέξεις που έχουν συνολικό πλήθος εμφάνισης μικρότερο του min_count , αλλά επειδή του δίνουμε την τιμή μηδέν δεν αγνοείται καμιά λέξη. Επίσης, χρησιμοποιείται η παράμετρος $window$ με default τιμή 5. Η παράμετρος $window$ ορίζει την μέγιστη απόσταση μεταξύ της τρέχουνσας και της προβλεπόμενης λέξης μέσα στην πρόταση.

Προγραμματιστικά, η διαδικασία παραγωγής διανυσμάτων με doc2vec είναι παρόμοια με αυτή που περιγράφηκε παραπάνω για το word2vec. Πιο συγκεκριμένα, πρέπει πάλι να φέρουμε σε κατανοητή μορφή από το μοντέλο Doc2Vec τα tweets, το οποίο επιτυγχάνεται πάλι με την συνάρτηση `labelizeTweets` που αναφέρθηκε παραπάνω.

Στη συνέχεια, γίνεται μία φορά `build_vocab` και `train` στο Doc2Vec για τα `train` δεδομένα. Τα διανύσματα που παράγονται για τα `train` δεδομένα αποθηκεύονται με τη χρήση του `docvecs` σε κατάλληλη μεταβλητή. Για τα `test` δεδομένα χρησιμοποιείται η συνάρτηση `infer_vector` της βιβλιοθήκης gensim για το μοντέλο Doc2Vec, η οποία υπολογίζει που βρίσκονται τα tweets του `test set` στο χώρο σε σχέση με τα tweets του `train set` στα οποία και έγινε το `train`. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό `train` και `test set` έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται `build_vocab` και `train` στο Doc2Vec από την αρχή για τα νέα `train` δεδομένα και στη συνέχεια `infer_vector` για τα νέα `test` δεδομένα.

Περιγραφή αλγορίθμου GloVe

Τέλος, το μοντέλο GloVe (Global Vectors) [34] είναι ουσιαστικά ένα μοντέλο word2vec που έχει γίνει ήδη `train` σε κάποια δεδομένα. Για την υλοποίηση του GloVe χρησιμοποιήθηκε το αρχείο “`glove.twitter.27B.200d.txt`” [34], το οποίο έγινε `train` σε 2 δισεκατομμύρια tweets και περιέχει 27 δισεκατομμύρια tokens και λεξιλόγιο (vocabulary) της τάξης του 1.2 εκατομμύρια. Το μέγεθος των διανυσμάτων είναι 200 διαστάσεις. Το μοντέλο φορτώνεται μια φορά στη μνήμη καθώς δεν απαιτεί `training` και άρα δεν επηρεάζεται από τα διαφορετικά `train set`

που παράγει το 10-fold-cross-validation. Η διαδικασία υπολογισμού των διανυσμάτων για ένα ολόκληρο tweet είναι η ίδια που περιγράφηκε για το word2vec παραπάνω. Δηλαδή, γίνεται fit_transform ένας TfidfVectorizer για τα train δεδομένα και στη συνέχεια transform για τα test δεδομένα. Έπειτα, καλείται η συνάρτηση buildWordVector για να υπολογίζει το μέσο όρο των διανυσμάτων που αντιστοιχούν στις λέξεις ενός tweet με τη βοήθεια του TF-IDF της κάθε λέξης, το οποίο και εδώ χρησιμοποιείται ως βάρος που δηλώνει το πόσο σημαντική είναι η λέξη στο σύνολο δεδομένων.

Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο TfidfVectorizer από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο TfidfVectorizer για τα νέα test δεδομένα. Όταν τροφοδοτούνται λέξεις που δεν υπάρχουν στο μοντέλο GloVe τότε το vector που τους αντιστοιχίζεται αποτελείται μόνο από μηδενικά. Έτσι, η κωδικοποίηση word2vec των train και test δεδομένων υπολογίζεται κανονικά όπως περιγράφηκε παραπάνω με τη μόνη διαφορά να είναι ότι οι λέξεις που δεν υπήρχαν στο pre-trained μοντέλο GloVe θα έχουν μηδενικά διανύσματα.

4.2.4 Αλγόριθμοι Feature Selection

Γενικά, η απόδοση ενός αλγορίθμου machine learning εξαρτάται κατά κύριο παράγοντα από την είσοδο του. Τα δεδομένα περιέχουν πολύ θόρυβο και ειδικά αυτά που προέρχονται από social medias. Έτσι, πρέπει να απομονωθούν τα δεδομένα που προκαλούν θόρυβο και να χρησιμοποιηθούν τα όσο πιο δυνατό γίνεται ποιοτικά. Οι αλγόριθμοι για feature selection μπορεί να φανούν ιδιαίτερα χρήσιμοι για την επίτευξη αυτού του στόχου. Σκοπός αυτών των αλγορίθμων είναι η μείωση των features που παράχθηκαν από τους αλγορίθμους των encoders. Τα features που θεωρούνται λιγότερο χρήσιμα αφαιρούνται από τους αλγορίθμους του feature selection, με αποτέλεσμα να μείνουν μόνο τα features που οι αλγόριθμοι κρίνουν πιο σημαντικά, δηλαδή αυτά που επηρεάζουν θετικά την απόδοση.

Επίσης, η μείωση των features από το feature selection βοηθάει στη μείωση του overfitting και βελτιώνει τη γενικότητα των αλγορίθμων μηχανικής μάθησης. Το feature selection έχει νόημα μόνο στους εξής encoders: TF-IDF, One-Hot-Encoding και Bigrams με One-Hot-Encoding. Στα πλαίσια της πτυχιακής δοκιμάστηκαν πέντε αλγόριθμοι feature selection, οι οποίοι είναι οι εξής: Univariate Selection, Recursive Feature Elimination, Principal Component Analysis, Truncated SVD και τέλος η μέθοδος Feature Importance.

Περιγραφή αλγορίθμου Univariate Selection

Ο αλγόριθμος Univariate Selection εξετάζει κάθε feature ξεχωριστά για να καθορίσει τη δύναμη της σχέσης του feature με την αντίστοιχη μεταβλητή. Ο αλγόριθμος λειτουργεί επιλέγοντας τα καλύτερα features βασιζόμενος στα univariate statistical tests, δηλαδή αφαιρεί όλα τα features εκτός από τα K features με το υψηλότερο score, που του ορίσαμε. Προγραμματιστικά χρησιμοποιείται η SelectKBest από τη βιβλιοθήκη sklearn. Στην SelectKBest χρησιμοποιούνται δύο ορίσματα, το score_func=chi2 και το k=100.

Το πρώτο όρισμα είναι το score_func που ορίζεται με το chi squared (chi2) statistical test, το οποίο είναι η δήλωση της μεθόδου που βαθμολογεί τα features. Το δεύτερο όρισμα είναι το k που ορίζεται με την τιμή 100 και είναι ο αριθμός των κορυφαίων σε score features που θα επιλεγούν από τον αλγόριθμο. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο SelectKBest από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο SelectKBest για τα νέα test δεδομένα.

Περιγραφή αλγορίθμου Recursive Feature Elimination (RFE)

Ο αλγόριθμος Recursive Feature Elimination (RFE) λειτουργεί αφαιρώντας αναδρομικά features και χτίζοντας ένα μοντέλο στα features που περισσεύουν. Χρησιμοποιεί την ακρίβεια (accuracy) του μοντέλου για να προσδιορίσει ποια features και ποιοι συνδυασμοί από features συνεισφέρουν περισσότερο στην πρόβλεψη της τιμής στόχου (target attribute) ενός μοντέλου μηχανικής μάθησης. Χρησιμοποιεί έναν εξωτερικό estimator ο οποίος δίνει βάρη στα features ώστε να μπορεί να επιλέξει τα πιο σημαντικά. Αρχικά, ο estimator εκπαιδεύεται στο αρχικό σετ από features για να βρεθεί πόσο σημαντικά είναι τα features. Στη συνέχεια, τα λιγότερο σημαντικά features κλαδεύονται από το τρέχον σύνολο με τα features. Αυτή η διαδικασία συνεχίζεται αναδρομικά στο κλαδεμένο σύνολο των features μέχρι να επιτευχθεί ο επιθυμητός αριθμός των features. Γενικά, η επιλογή του estimator δεν κάνει μεγάλη διαφορά.

Χρησιμοποιήθηκε ο RandomForestClassifier της βιβλιοθήκης sklearn. Στη συνέχεια, στο RFE της βιβλιοθήκης sklearn δίνεται σαν όρισμα ο estimator που δημιουργήσαμε μαζί με το νούμερο 100 που είναι το νούμερο των πιο σημαντικών features που θέλουμε να επιλεγούν. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο RFE από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform

για τα νέα test δεδομένα. Ο RFE δεν χρησιμοποιήθηκε για παραγωγή των τελικών αποτελεσμάτων, αλλά χρησιμοποιήθηκε για παραγωγή συμπερασμάτων που αναφέρονται παρακάτω.

Περιγραφή αλγορίθμου Principal Component Analysis (PCA)

Ο αλγόριθμος Principal Component Analysis (PCA) εκτελεί γραμμική μείωση διαστάσεων (linear dimensionality reduction) χρησιμοποιώντας Singular Value Decomposition των δεδομένων για να προβληθεί σε χαμηλότερο χώρο διαστάσεων. Με άλλα λόγια, Principal Component Analysis χρησιμοποιεί γραμμική άλγεβρα για να μετασχηματίζει τα δεδομένα σε μια συμπιεσμένη μορφή. Γενικά, το PCA θεωρείται τεχνική μείωσης των δεδομένων. Μια ιδιότητα του PCA είναι ότι επιτρέπει την επιλογή του αριθμού των διαστάσεων στο μετασχηματισμένο αποτέλεσμα.

Προγραμματιστικά, χρησιμοποιείται το PCA της βιβλιοθήκης sklearn με όρισμα n_components=100, που ορίζει ότι το μετασχηματισμένο αποτέλεσμα θα έχει 100 features. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο PCA από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο PCA για τα νέα test δεδομένα.

Περιγραφή αλγορίθμου Truncated Singular Value Decomposition (SVD)

Ο αλγόριθμος Truncated Singular Value Decomposition (SVD) είναι ένας αλγόριθμος για μείωση των διαστάσεων, σαν το PCA. Η κύρια διαφορά τους είναι ότι ο αλγόριθμος Truncated SVD δεν συγκεντρώνει τα δεδομένα πριν υπολογίσει την μοναδιαία τιμή αποσύνθεσης (singular value decomposition - SVD), το οποίο σημαίνει ότι μπορεί να λειτουργήσει με sparse matrices αποδοτικά. Ο αλγόριθμος SVD εκτελεί γραμμική μείωση των διαστάσεων μέσω της αποκομμένης μοναδικής τιμής αποσύνθεσης (truncated singular value decomposition – SVD).

Προγραμματιστικά, χρησιμοποιείται το TruncatedSVD της βιβλιοθήκης sklearn με όρισμα n_components=100, που ορίζει ότι το μετασχηματισμένο αποτέλεσμα θα έχει 100 features. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο TruncatedSVD από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο TruncatedSVD για τα νέα test δεδομένα.

Περιγραφή αλγορίθμου Feature Importance

Η μέθοδος Feature Importance ουσιαστικά βασίζεται στον υπολογισμό της σημαντικότητας των features από αλγορίθμους τύπου bagged decision trees, όπως ο Random Forest και ο Extra Trees. Προγραμματιστικά, για τον υπολογισμό της σημαντικότητας των features χρησιμοποιείται ο αλγόριθμος RandomForestClassifier της βιβλιοθήκης sklearn. Στον RandomForestClassifier υπάρχουν τρία ορίσματα. Το πρώτο όρισμα είναι το n_estimators=250, που είναι τα δένδρα (trees) στο Forest. Το δεύτερο όρισμα είναι το max_features=7, που είναι το πλήθος των features σε κάθε υποδιαίρεση (split). Το τρίτο όρισμα είναι το max_depth=30, το οποίο είναι το μέγιστο βάθος (depth) του δένδρου. Στη συνέχεια, μόλις γίνει train ο αλγόριθμος στα train δεδομένα δίνεται σαν όρισμα στο SelectFromModel της βιβλιοθήκης sklearn.

Ο SelectFromModel επιλέγει features βασιζόμενος στο βάρος που τους δόθηκε από τον RandomForestClassifier. Το δεύτερο όρισμα του SelectFromModel είναι το threshold = "9*mean", όπου το mean είναι η διάμεσος (median) από τα βάρη των features. Το threshold είναι το κατώφλι που ορίζει ποια features θα επιλεχθούν. Πιο συγκεκριμένα, όσα features έχουν σημαντικότητα μεγαλύτερη ή ίση από την τιμή κατωφλίου διατηρούνται, ενώ τα υπόλοιπα απορρίπτονται. Επειδή χρησιμοποιείται 10-fold-cross-validation και ουσιαστικά σε κάθε ένα από τα 10 folds έχουμε διαφορετικό train και test set έχουν ληφθεί κατάλληλα μέτρα ώστε να γίνεται fit_transform στο SelectFromModel από την αρχή για τα νέα train δεδομένα και στη συνέχεια transform στο SelectFromModel για τα νέα test δεδομένα.

4.2.5 Αλγόριθμοι Machine Learning

Η χρήση αλγορίθμων μηχανικής μάθησης για την επίλυση προβλημάτων που αφορούν sentiment analysis και irony detection τον τελευταίο καιρό είναι ιδιαίτερα δημοφιλής, ιδίως η χρήση αλγορίθμων τύπου deep learning. Στα πλαίσια της πτυχιακής υλοποιήθηκαν 9 αλγόριθμοι μηχανικής μάθησης και ένας αλγόριθμος επιλογής αποτελεσμάτων από πολλά μοντέλα. Η λίστα αυτών των αλγορίθμων περιλαμβάνει νευρωνικά δίκτυα (MLP Neural Networks), Conv1D νευρωνικά δίκτυα, LSTM νευρωνικά δίκτυα, Gaussian Naive Bayes, Support Vector Machine (SVM), Bernoulli, Logistic Regression, K-Neighbors, Multinomial Naive Bayes και τέλος την μέθοδο Voting Ensembles που συνδυάζει τους καλύτερους αλγορίθμους σε έναν classifier.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set. Αυτό επιτυγχάνεται με το Kfold της βιβλιοθήκης sklearn.

To Kfold έχει τρία ορίσματα, το n_splits=10, το random_state=41 και το shuffle=True. Με το όρισμα n_splits ορίζεται πόσα fold θα έχει το cross validation. Με το όρισμα random_state ορίζεται το seed του random number generator. Με το όρισμα suffle ορίζεται αν θα ανακατευτούν τα δεδομένα πριν χωριστούν σε train set και test set, όταν ορίζεται ως true τότε ανακατεύονται.

Επειτα, το train set και το test set, μαζί με άλλες μεταβλητές σχετικές με τα δεδομένα, δινονται ως ορίσματα στην συνάρτηση get_enc της κλάσης Reader του αρχείου ClassRead.py ώστε να περάσουν από αλγορίθμους για encoding και feature selection. Στη συνέχεια, ορίζονται οι classifiers με τις κατάλληλες παραμέτρους, οι οποίες επιλέχθηκαν μετά από δοκιμές. Το μοντέλο γίνεται train με τα train δεδομένα και στη συνέχεια γίνεται η πρόβλεψη των ironic labels των tweets του test set που ορίζουν αν ένα tweet είναι ειρωνικό. Με τη χρήση των προβλέψεων και των πραγματικών ironic labels υπολογίζεται το receiving operating characteristic (ROC) για κάθε ένα από τα 10 folds με τη συνάρτηση roc_auc_score της βιβλιοθήκης sklearn. Επίσης, υπολογίζεται και ο συνεχής μέσος όρος (continued average) του ROC για κάθε ένα fold του 10-fold-cross-validation.

Στη συνέχεια, υπολογίζονται τέσσερεις μετρικές, που έχουν ως σκοπό να αξιολογήσουν τις προβλέψεις του classifier. Αυτές οι μετρικές είναι το precision, το accuracy, το recall καθώς και το f1-score, που αναπτύχθηκαν αναλυτικά στο κεφάλαιο 3.3. Ο υπολογισμός των μετρικών γίνεται πάλι με τη χρήση των προβλέψεων και των πραγματικών ironic labels, μέσω του confusion_matrix της βιβλιοθήκης sklearn. Από τον confusion matrix, που υπολογίστηκε, χρησιμοποιούνται τα true positives, true negatives, false positives και false negatives για τον υπολογισμό των τεσσάρων αυτών μετρικών. Οι μετρικές αυτές υπολογίζονται για κάθε επανάληψη και στο τέλος του 10-fold-cross-validation υπολογίζεται ο μέσος όρος αυτών των τιμών, που είναι ουσιαστικά αυτό που χρησιμοποιείται για την γενική αξιολόγηση του αλγορίθμου. Η κατάταξη των αλγορίθμων θα γίνει με βάση το f1-score, όπως ορίζεται και από τον διαγωνισμό. Πιο συγκεκριμένα, όσο μεγαλύτερο είναι το f1-score τόσο καλύτερος είναι ο αλγόριθμος.

Περιγραφή μοντέλου Gaussian Naive Bayes

Το μοντέλο Gaussian Naive Bayes αναπτύχθηκε αναλυτικά στο κεφάλαιο 3.2.3. Ο αλγόριθμος Naive Bayes είναι μια διαισθητική μέθοδος που χρησιμοποιεί τις πιθανότητες κάθε χαρακτηριστικού που ανήκει σε κάθε κλάση για να κάνει μια πρόβλεψη. Το μοντέλο Gauss-

ian Naive Bayes απλοποιεί τον υπολογισμό των πιθανοτήτων υποθέτοντας ότι η πιθανότητα κάθε χαρακτηριστικού που ανήκει σε μια δεδομένη τιμή κλάσης είναι ανεξάρτητη από όλα τα άλλα χαρακτηριστικά. Αυτή είναι μια ισχυρή παραδοχή, αλλά οδηγεί σε μια γρήγορη και αποτελεσματική μέθοδο. Η πιθανότητα μιας τιμής κλάσης που δίνεται σε μια τιμή ενός χαρακτηριστικού ονομάζεται υπό όρους πιθανότητα (conditional probability). Πολλαπλασιάζοντας τις υπό συνθήκη πιθανότητες μαζί για κάθε χαρακτηριστικό για μια δεδομένη τιμή κλάσης, έχουμε μια πιθανότητα ενός στιγμιότυπου δεδομένων που ανήκει σε αυτή την κλάση. Για να κάνουμε μια πρόβλεψη μπορούμε να υπολογίσουμε τις πιθανότητες ενός στιγμιοτύπου από κάθε κλάση και να επιλέξουμε την τιμή κλάσης με την υψηλότερη πιθανότητα.

Το μοντέλο Naive Bayes περιγράφεται συχνά χρησιμοποιώντας κατηγορηματικά δεδομένα επειδή είναι εύκολο να περιγραφεί και να υπολογιστεί χρησιμοποιώντας αναλογίες. Μια πιο χρήσιμη έκδοση του αλγορίθμου υποστηρίζει αριθμητικά χαρακτηριστικά και υποθέτει ότι οι τιμές κάθε αριθμητικού χαρακτηριστικού ακολουθούν κανονική κατανομή, δηλαδή πέφτουν κάπου σε καμπύλη καμπάνας. Αυτή η παραδοχή εξακολουθεί να δίνει πολύ καλά αποτελέσματα. Παρά τις υπερβολικά απλουστευμένες υποθέσεις τους, οι ταξινομητές Naive Bayes δουλεύουν αρκετά καλά σε πολλές πραγματικές καταστάσεις, όπως στην ταξινόμηση εγγράφων και στο φίλτραρισμα ανεπιθύμητων μηνυμάτων.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set και στη συνέχεια τα δεδομένα χρησιμοποιούνται από το μοντέλο GaussianNB της βιβλιοθήκης sklearn. Το μοντέλο GaussianNB υλοποιεί τον αλγόριθμο Gaussian Naive Bayes που υποθέτει ότι η πιθανότητα των χαρακτηριστικών είναι Gaussian, δηλαδή:

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Οι παράμετροι σ_y και μ_y υπολογίζονται χρησιμοποιώντας της μέγιστη πιθανότητα. Για το μοντέλο του Naive Bayes χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.2, ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	53.24	54.95	79.69	55.08	63.77
-	word2vec	53.23	55.63	90.93	55.74	67.10
-	doc2vec	58.08	60.79	77.51	60.91	66.30
-	GloVe	55.80	57.82	73.60	57.87	63.43

Πίνακας 4.2 Αξιολόγηση Gaussian Naive Bayes

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει το word embedding word2vec, ενώ το doc2vec πετυχαίνει και αυτό παρόμοιο score.

Περιγραφή μοντέλου Multinomial Naive Bayes

Το μοντέλο Multinomial Naive Bayes υλοποιεί τον αλγόριθμο Naive Bayes για multinomially distributed data. Είναι μία από τις δύο κλασσικές παραλλαγές του naive Bayes που χρησιμοποιούνται για ταξινόμηση κειμένου (text classification), όπου τα δεδομένα αναπαριστώνται συνήθως με word vector count, ενώ στην πράξη δουλεύουν καλά και τα tf-idf διανύσματα. Η κατανομή παραμετροποιείται από διανύσματα του τύπου $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$ για κάθε κλάση y , όπου n είναι ο αριθμός των χαρακτηριστικών (στην ταξινόμηση κειμένου είναι το μέγεθος του λεξιλογίου) και θ_{yi} είναι η πιθανότητα $P(x_i|y)$ των χαρακτηριστικών να εμφανιστούν σε ένα δείγμα που ανήκει στην κλάση y . Οι παράμετροι θ_y εκτιμώνται από μια πιο απλοποιημένη εκδοχή της μέγιστης πιθανότητας, όπως για παράδειγμα το relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

όπου $N_{yi} = \sum_{x_i \in T} x_i$ είναι το πλήθος των φορών που το χαρακτηριστικό i εμφανίζεται σε ένα δείγμα της κλάση y στο train set T , και $N_y = \sum_{i=1}^{|T|} N_{yi}$ είναι το συνολικό πλήθος των χαρακτηριστικών για την κλάση y . Η διαδικασία του smoothing με $\alpha \geq 0$ αναφέρεται σε χαρακτηριστικά που δεν υπάρχουν στα δείγματα εκμάθησης και αποτρέπει μηδενικές πιθανότητες σε μελλοντικούς υπολογισμούς. Όταν $\alpha = 1$ τότε έχουμε Laplace smoothing, ενώ όταν $\alpha < 1$ τότε έχουμε Lidstone smoothing. Ο ταξινομιτής multinomial Naive Bayes είναι κατάλληλος για ταξινόμηση διακριτών χαρακτηριστικών, για παράδειγμα πλήθος λέξεων για ταξινόμηση κειμένου. Η multinomial κατανομή συνήθως απαιτεί χαρακτηριστικά με ακέραιους αριθμούς, στην πράξη όμως, μπορεί να λειτουργήσει και με κλασματικούς αριθμούς όπως το tf-idf.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set και στη συνέχεια τα δεδομένα χρησιμοποιούνται από το μοντέλο MultinomialNB της βιβλιοθήκης sklearn. Για το μοντέλο του multinomial Naive Bayes χρησιμοποιήθηκαν μόνο συγκεκριμένοι συνδυασμοί των encoding αλγορίθμων με τους feature selection αλγορίθμους, γιατί το μοντέλο δεν δέχεται αρνητικές τιμές σαν χαρακτηριστικά. Οι συνδυασμοί που χρησιμοποιήθηκαν είναι οι εξής: Feature Improtance με One-Hot-Encoding, Univariate Selection με Bigrams, Univariate Selection με One-Hot-Encoding, Univariate Selection με TF-IDF, Feature Improtance με Bigrams, Feature Improtance με TF-IDF, One-Hot-Encoding, Brigrams και TF-IDF. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.3, ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (οι μετρικές είναι οι μέσοι όροι για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	62.06	63.04	66.70	63.10	64.21
-	TF-IDF	63.63	63.27	61.87	63.37	62.59

Πίνακας 4.3 Αξιολόγηση multinomial Naive Bayes

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει ο αλγόριθμος one-hot encoding, ενώ οι υπόλοιποι αλγόριθμοι έχουν αισθητή διαφορά.

Περιγραφή μοντέλου Bernoulli Naive Bayes

Το μοντέλο Bernoulli Naive Bayes υλοποιεί τους αλγορίθμους εκπαίδευσης και ταξινόμησης του Naive Bayes για δεδομένα που είναι κατανεμημένα σύμφωνα με multivariate Bernoulli distributions, δηλαδή μπορεί υπάρχουν πολλά χαρακτηριστικά αλλά κάθε ένα από αυτά θεωρείται ότι είναι μια μεταβλητή δυαδικής τιμής (Bernoulli, boolean). Επομένως, αυτό το μοντέλο απαιτεί τα δείγματα να αναπαριστώνται ως χαρακτηριστικά διανύσματα δυαδικών τιμών. Αν δοθεί οτιδήποτε άλλο είδος δεδομένων ως είσοδος, τότε το μοντέλο μπορεί να μετατρέψει την είσοδο σε δυαδική μορφή, το οποίο εξαρτάται από την παράμετρο binarize. Ο κανόνας απόφασης που χρησιμοποιείται από το μοντέλο Bernoulli Naive Bayes είναι: $P(x_i | y) = P(i | y)x_i + (1 - P(i | y))(1 - x_i)$. Η διαφορά του κανόνα από το μοντέλο multinomial Naive Bayes είναι ότι ‘τιμωρεί’ ρητά την μη ύπαρξη ενός χαρακτηριστικού i που είναι

δείκτης μιας κλάσης γ, ενώ ο κανόνας του multinomial μοντέλου θα αγνοούσε τη μη ύπαρξη ενός χαρακτηριστικού.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set και στη συνέχεια τα δεδομένα χρησιμοποιούνται από το μοντέλο BernoulliNB της βιβλιοθήκης sklearn. Για το μοντέλο του Bernoulli Naive Bayes χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.4, ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	62.21	63.01	65.97	63.10	63.93
-	TF-IDF	62.21	63.01	65.97	63.10	63.93
PCA	One-Hot	58.62	59.93	67.14	60.0	62.51
Univariate Selection	TF-IDF	62.65	62.75	62.86	62.83	62.64

Πίνακας 4.4 Αξιολόγηση Bernoulli Naive Bayes

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει ο αλγόριθμος one-hot encoding, ενώ ο αλγόριθμος TF-IDF πετυχαίνει και αυτός το ίδιο score.

Περιγραφή μοντέλου K-Neighbors

Η ταξινόμηση με το μοντέλο K-Neighbors υλοποιείται με απλή πλειοψηφία των πλησιέστερων γειτόνων κάθε σημείου, δηλαδή η τιμή που κάνει πρόβλεψη το μοντέλο προκύπτει από την τιμή της κλάσης δεδομένων που έχει του περισσότερους κοντινούς γείτονες στο προβλεπόμενο σημείο. Γενικά, η απόσταση που χρησιμοποιείται για να βρεθεί η κοντινότητα μπορεί να είναι οποιαδήποτε, αλλά η Ευκλείδεια απόσταση είναι η πιο διαδεδομένη. Τα μοντέλα που βασίζονται σε γείτονες είναι μοντέλα μηχανικής μάθησης που δεν μπορούν να γενικευτούν, αφού μπορούν να ‘θυμηθούν’ όλα τα δεδομένα εκπαίδευσης. Παρά την απλότητα του, οι κοντινότεροι γείτονες είναι χρήσιμοι για μεγάλα προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression). Επίσης, επειδή είναι μη παραμετρική μέθοδος, είναι συχνά χρήσιμη σε ταξινόμηση όπου το όριο απόφασης είναι ακανόνιστο. Η επιλογή της τι-

μής του κ εξαρτάται από τα δεδομένα, γενικά όσο μεγαλύτερο είναι το κ τόσο λιγότερες είναι οι επιπτώσεις του θορύβου.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set και στη συνέχεια τα δεδομένα χρησιμοποιούνται από το μοντέλο KNeighborsClassifier της βιβλιοθήκης sklearn. Χρησιμοποιείται ένα μόνο όρισμα, το n_neighbors=140, το οποίο ορίζει ότι για την πρόβλεψη θα χρησιμοποιηθούν οι 140 κοντινότεροι γείτονες. Κάτω υπό συνθήκες, η αποδοτικότητα μπορεί να αυξηθεί αν προσθέσουμε βάρη στους γείτονες, ώστε οι πιο κοντινοί γείτονες να συνεισφέρουν περισσότερο στην πρόβλεψη. Υστερα από δοκιμές, για το πρόβλημα της πτυχιακής, χρησιμοποιήθηκε η προεπιλεγμένη (default) τιμή της παραμέτρου weights η οποία είναι η ‘uniform’, η οποία αντιστοιχεί ομοιόμορφα τα βάρη σε κάθε γείτονα δηλαδή όλα τα σημεία έχουν το ίδιο βάρος.

Για το μοντέλο του K-Neighbors χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding μαζί με όλους τους αλγορίθμους για το feature selection. Σε όλους τους αλγορίθμους κωδικοποίησης, εκτός από τα word embeddings word2vec και doc2vec, χρησιμοποιήθηκε η συνάρτηση Normalizer της βιβλιοθήκης sklearn στα train και στα test δεδομένα με τη χρήση fit_transform και transform αντίστοιχα. Στο GloVe έγινε κανονικοποίηση (normalize) γιατί είχε καλύτερα αποτελέσματα. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.5, ενώ όλα τα αποτελέσματα παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	*Doc2vec	60.76	61.73	66.57	61.83	* 63.38
PCA	Bigrams	62.13	62.77	65.41	62.88	63.61
PCA	TF-IDF	64.44	64.42	64.60	64.56	64.36
Univariate Selection	TF-IDF	61.31	62.54	67.73	62.62	64.28

Πίνακας 4.5 Αξιολόγηση K-Neighbors

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει ο συνδυασμός του PCA με το TF-IDF, αν και οι συνδυασμοί Univariate Selection με TF-IDF και PCA με Bigrams έχουν μικρή διαφορά. Τα αποτελέσματα του αλγορίθμου με το word embedding doc2vec προέκυψαν χωρίς scaling στα δεδομένα, ενώ για τους υπόλοιπους αλγορίθμους έγινε scaling των δεδομένων στο διάστημα (0,1).

Περιγραφή μοντέλου Logistic Regression

Το μοντέλο Logistic Regression είναι ένα μοντέλο μηχανικής μάθησης που χρησιμοποιείται για να προβλέψει την πιθανότητα μιας κατηγορικής μεταβλητής. Στο μοντέλο Logistic Regression η μεταβλητή που θα προβλεφθεί είναι μια δυαδική μεταβλητή που περιέχει δεδομένα του κωδικοποιούνται ως 1 που περιγράφει επιτυχία, λογικό αληθές, ή ως 0 που περιγράφει αποτυχία, λογικό ψευδές κτλ. Στο πρόβλημα της πτυχιακής το 1 περιγράφει ότι ένα tweet είναι ειρωνικό ενώ το 0 περιγράφει ότι ένα tweet δεν είναι ειρωνικό. Με άλλα λόγια, το μοντέλο προβλέπει την πιθανότητα $P(Y=1)$ ως συνάρτηση του X. Για να είναι το μοντέλο δυαδικό πρέπει η προβλεπόμενη μεταβλητή να είναι και αυτή δυαδική.

Προγραμματιστικά, χρησιμοποιείται 10-fold-cross validation για να χωριστούν τα δεδομένα σε train και σε test set και στη συνέχεια τα δεδομένα χρησιμοποιούνται από το μοντέλο LogisticRegression της βιβλιοθήκης sklearn. Χρησιμοποιούνται δύο ορίσματα, το solver = "liblinear" [35], το οποίο είναι ο αλγόριθμος που θα χρησιμοποιηθεί στο πρόβλημα της βελτιστοποίησης, και το C=0.1, το οποίο είναι το πόσο ισχυρή είναι η κανονικοποίηση (regularization) και όσο πιο μικρή τιμή έχει τόσο πιο ισχυρή κανονικοποίηση (regularization) εφαρμόζεται.

Επίσης, το όρισμα penalty έχει προκαθορισμένη (default) τιμή 'l2' και χρησιμοποιείται για να καθορίσει τον κανόνα που χρησιμοποιείται κατά το penalization. Για το μοντέλο Logistic Regression χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.6 ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
Feature Importance	TF-IDF	60.37	62.64	73.87	62.79	66.32
Univariate Selection	TF-IDF	59.66	62.10	74.88	62.24	66.29
PCA	TF-IDF	60.17	62.44	73.72	62.58	66.14
SVD	TF-IDF	60.44	62.67	73.57	62.81	66.23
-	TF-IDF	61.19	63.45	73.85	63.61	66.79

Πίνακας 4.6 Αξιολόγηση Logistic Regression

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει ο αλγόριθμος TF-IDF και στη συνέχεια ακολουθούν οι υπόλοιποι συνδυασμοί στο πινακάκι, όλοι τους με μικρή διαφορά.

Περιγραφή μοντέλου MLP Neural Network

Το μοντέλο Multilayer Perceptron (MLP) Neural Network αναπτύχθηκε αναλυτικά στο κεφάλαιο 3.2.3. Σε ένα νευρωνικό δίκτυο το πρώτο επίπεδο (layer) ονομάζεται επίπεδο εισόδου (input layer) και αποτελείται από ένα σύνολο νευρώνων που αναπαριστούν τα χαρακτηριστικά της εισόδους (input features). Έπειτα, υπάρχει, προαιρετικά, το κρυφό επίπεδο (hidden layer) όπου κάθε νευρώνας μετατρέπει τις τιμές του προηγούμενου επιπέδου με τη χρήση ενός σταθμισμένου γραμμικού αθροίσματος του τύπου $w_1x_1 + w_2x_2 + \dots + w_mx_m$, ακολουθούμενο από μια μη γραμμική συνάρτηση ενεργοποίησης (non linear activation function), όπως η υπερβολική εφαπτομένη (hyperbolic tangent function). Το επίπεδο εξόδου (output layer) λαμβάνει τιμές από το τελευταίο κρυφό επίπεδο (last hidden layer) και τις μετατρέπει σε τιμές εξόδου (output values).

Προγραμματιστικά, χρησιμοποιήθηκε η βιβλιοθήκη Keras με back-end το TensorFlow. Το TensorFlow παρέχει αποδοτικές αριθμητικές βιβλιοθήκες, ενώ επιλέγει αυτόματα τον καλύτερο τρόπο να αναπαραστήσει το δίκτυο στο διαθέσιμο hardware, είτε με χρήση GPU είτε με CPU είτε με κατανεμημένη χρήση και στα δύο, για τις λειτουργίες της εκπαίδευσης και της εκτέλεσης προβλέψεων. Τα μοντέλα στο Keras ορίζονται ως μια ακολουθία (sequence) από επίπεδα. Αυτό δηλώνεται με την συνάρτηση Sequential. Το νευρωνικό δίκτυο που χρησιμοποιείται είναι ένα πλήρως συνδεδεμένο δίκτυο με τρία επίπεδα. Τα πλήρως συνδεδεμένα επίπεδα ορίζονται με την χρήση της κλάσης Dense.

Στο επίπεδο εισόδου το όρισμα input_dim ορίζει το μέγεθος της εισόδου. Το πλήθος των νευρώνων σε κάποιο επίπεδο είναι το πρώτο όρισμα. Για το πρώτο επίπεδο χρησιμοποιήθηκαν 20 νευρώνες, 10 νευρώνες για το δεύτερο επίπεδο και ένας νευρώνας για το τρίτο επίπεδο. Το δεύτερο όρισμα είναι η μέθοδος αρχικοποίησης (initialization method) που ορίζεται με το όρισμα kernel_initializer. Επίσης, υπάρχει και η συνάρτηση ενεργοποίησης (activation function) η οποία ορίζεται με τη χρήση του ορίσματος activation. Χρησιμοποιήθηκε η προεπιλεγμένη (default) αρχικοποίηση των βαρών του δικτύου, η οποία χρησιμοποιεί ομοιόμορφη κατανομή (uniform distribution) και θέτει τυχαίες μικρές τιμές στα βάρη ανάμεσα στο 0 και το 0.05.

Στα πρώτα δύο επίπεδα θα χρησιμοποιηθεί η συνάρτηση ενεργοποίησης ‘softsign’ [36], ενώ στο επίπεδο εξόδου θα χρησιμοποιηθεί η ‘sigmoid’ συνάρτηση ενεργοποίησης για να βεβαιωθούμε ότι η έξοδος του νευρωνικού δικτύου θα είναι είτε 0 είτε 1, αφού όλα τα δεδομένα για το ironic label είναι και αυτά 0 ή 1. Και στα τρία επίπεδα χρησιμοποιήθηκε ο kernel initializer ‘glorot_uniform’ [37]. Στα πρώτα δύο επίπεδα χρησιμοποιήθηκε kernel_constraint = maxnorm(2), το οποίο είναι η συνάρτηση περιορισμών (constraint function) που εφαρμόζεται στον πυρήνα του πίνακα με τα βάρη. Μετά το επίπεδο εισόδου και το κρυφό επίπεδο χρησιμοποιήθηκαν δύο μέθοδοι Dropout [38] οι οποίοι θέτουν τυχαία 0 σε ένα πλήθος μονάδων εισόδου κατά τη διάρκεια της εκπαίδευσης, το οποίο συμβάλλει στην αποφυγή του overfitting. Αυτό το πλήθος των μονάδων εισόδου εξαρτάτε από το rate που παίρνει τιμές ανάμεσα στο 0 και το 1 και συμβολίζει το κομμάτι των μονάδων εισόδου που θα αγνοηθεί. Και τα δύο Dropout ορίζονται με rate ισοδύναμο με 0.2.

Επίσης, κατά το compile του δικτύου πρέπει να οριστούν ο optimizer που χρησιμοποιείται για την αναζήτηση διαφόρων βαρών για το δίκτυο καθώς και την αναζήτηση προαιρετικών μετρικών που πιθανός θα θέλαμε να συλλεχθούν και να εκτυπωθούν κατά την εκπαίδευση, το loss function που χρησιμοποιείται για να αξιολογήσει ένα σύνολο από βάρη και το metrics. Θα χρησιμοποιηθεί λογαριθμικό loss, που στο Keras για δυαδική ταξινόμηση (binary classification) ορίζεται με το loss με τιμή ‘binary_crossentropy’. Επίσης, ο optimizer που θα χρησιμοποιηθεί είναι ο RMSprop [39] με learning rate 0.001. Τέλος, επειδή έχουμε πρόβλημα ταξινόμησης θα υπολογιστεί και θα εκτυπωθεί το accuracy της ταξινόμησης σαν μετρική.

Πριν να γίνει fit το δίκτυο στα δεδομένα εκπαίδευσης ορίζεται το EarlyStopping callback, το οποίο σταματάει την διαδικασία της εκπαίδευσης αν η ποσότητα που παρακολουθείται έχει σταματήσει να βελτιώνεται. Η ποσότητα που παρακολουθείτε ορίζεται με το όρισμα monitor και είναι η ‘val_loss’. Ορίζονται άλλα τρία ορίσματα, τα οποία είναι το min_delta=0, το patience=2, το verbose=0 και το mode=’auto’. Το όρισμα min_delta είναι η ελάχιστη μεταβολή στην ποσότητα που παρακολουθείται που χαρακτηρίζεται ως βελτίωση, δηλαδή για μια απόλυτη αλλαγή μικρότερη του min_delta θεωρείται ότι δεν υπάρχει βελτίωση.

Το όρισμα patience είναι το πλήθος των εποχών (epochs) στο οποίο δεν υπάρχει βελτίωση και μετά το οποίο η εκπαίδευση θα σταματήσει. Το όρισμα verbose είναι οι πληροφορίες που εμφανίζονται σχετικές με τη διαδικασία, όταν ορίζεται στο 0 δεν εμφανίζεται καμία πληροφορία. Το όρισμα mode ορίζει αν η εκπαίδευση θα σταματήσει όταν η ποσότητα που παρακολουθείται σταματήσει να μειώνεται ή να αυξάνεται. Πιο συγκεκριμένα, όταν το όρισμα

παίρνει τιμή ‘min’ τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να μειώνεται, όταν παίρνει τιμή ‘max’ τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να αυξάνεται και όταν παίρνει τιμή ‘auto’ τότε το πότε θα σταματήσει η εκπαίδευση εξαρτάτε από το είδος της ποσότητας που παρακολουθείται, δηλαδή της monitor = ‘val_loss’.

Στη συνέχεια, γίνεται fit του μοντέλου στα δεδομένα εκπαίδευσης. Αυτό επιτυγχάνεται με τη συνάρτηση fit() που δέχεται τέσσερα ορίσματα. Τα δύο πρώτα ορίσματα είναι τα δεδομένα εκπαίδευσης, τα κωδικοποιημένα tweets και τα ironic labels. Το τρίτο όρισμα είναι το batch_size=20, το οποίο είναι το νούμερο των στιγμιοτύπων που αξιολογούνται πριν από κάθε ανανέωση των βαρών στο μοντέλο. Το τέταρτο όρισμα είναι το epochs=50, που ορίζει το πόσες φορές θα τρέξει η διαδικασία της εκπαίδευσης πάνω στα δεδομένα εκπαίδευσης.

Μετά την ολοκλήρωση της εκπαίδευσης, μπορεί να αξιολογηθεί η απόδοση του μοντέλου με τα ίδια δεδομένα εκπαίδευσης. Αυτό γίνεται με την συνάρτηση evaluate() που δέχεται σαν ορίσματα τα δεδομένα εκπαίδευσης που χρησιμοποιήσαμε προηγουμένως για την εκπαίδευση του μοντέλου. Η συνάρτηση αυτή παράγει προβλέψεις για τα tweets που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο και τις συγκρίνει με τα ironic labels που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο. Από αυτή τη σύγκριση παράγει βαθμολογίες, όπως ο μέσος όρος του loss και άλλες μετρικές που ορίσαμε όπως το accuracy. Επίσης, χρησιμοποιείται και η συνάρτηση summary() που εκτυπώνει τη σύνοψη του μοντέλου. Η πρόβλεψη πάνω στο test set γίνεται απλώς με την συνάρτηση predict(). Όπως αναφέρθηκε παραπάνω οι προβλεπόμενες τιμές είναι είτε 1 είτε 0, επειδή χρησιμοποιήθηκε η ‘sigmoid’ συνάρτηση στο επίπεδο εξόδου.

Για το μοντέλο νευρωνικό δίκτυο χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Για την παραμετροποίηση του μοντέλου χρησιμοποιήθηκε GridSearch σε όλες τις παραμέτρους που αναφέρθηκαν παραπάνω ώστε να βρεθούν οι πιο αποτελεσματικές. Επίσης, έγινε και χειροκίνητη παραμετροποίηση για να δοκιμαστούν διαφορετικοί συνδυασμοί παραμέτρων, για παράδειγμα 20 νευρώνες στο πρώτο layer και 10 στο δεύτερο ή στο πρώτο layer init_mode ‘uniform’ και ‘glorot_normal’ στο δεύτερο. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.7 ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
PCA	TF-IDF	65.47	64.52	61.50	70.92	63.32
PCA	Bigrams	60.97	62.02	66.70	67.49	63.60
Feature Importance	Bigrams	62.36	62.96	67.11	67.95	64.22
Feature Importance	One-Hot	64.16	63.95	62.82	68.96	63.39
TruncatedSVD	Bigrams	61.47	62.70	67.99	67.32	64.47
TruncatedSVD	TF-IDF	64.17	63.77	62.34	70.72	63.11

Πίνακας 4.7 Αξιολόγηση MLP Νευρωνικού Δικτύου

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται από τον διαγωνισμό. Το καλύτερο f1-score το πετυχαίνει ο συνδυασμός του Turned SVD με το Bigrams, ενώ πολύ κοντά είναι και ο συνδυασμός του Feature Importance με το Bigrams.

Περιγραφή μοντέλου Long Short-Term Memory Network (LSTM)

Ένα από τα προτερήματα ενός επαναλαμβανόμενου νευρωνικού δικτύου (recurrent Neural Network - RNN) είναι η ικανότητα του να μπορεί να συνδέει πληροφορίες που αποκτήθηκαν προηγουμένως στην παρόν διεργασία. Ωστόσο, αυτό δεν είναι εφικτό σε όλες τις περιπτώσεις. Όταν ένα RNN χρειάζεται να ανακτήσει πληροφορίες που αποκτήθηκαν προσφάτως για να εκτελέσει την παρούσα διαδικασία, τότε τα καταφέρνει με επιτυχία. Όμως, αυτό δεν ισχύει στην περίπτωση όπου η πληροφορία που χρειάζεται προέρχεται από ένα γενικότερο πλαίσιο, δηλαδή πληροφορία που αποκτήθηκε σε μεγαλύτερο χρονικό διάστημα. Αυτό το πρόβλημα το επιλύει το μοντέλο long short term memory networks (LSTMs). Ένα δίκτυο LSTM είναι ένα ειδικό ειδικό είδος RNN, ικανό να μαθαίνει μακροπρόθεσμες εξαρτήσεις (long-term dependencies). Το μοντέλο LSTM προτάθηκε από τον Sepp Hochreiter και τον Jürgen Schmidhuber το 1997 [40] και βελτιώθηκε από την ομάδα του Felix Ger το 2000. [41]

Αυτά τα δίκτυα λειτουργούν αποδοτικά σε ένα μεγάλο εύρος προβλημάτων και η χρήση τους είναι διαδεδομένη. Τα δίκτυα LSTM είναι ρητά σχεδιασμένα ώστε να αποφύγουν το πρόβλημα των μακροπρόθεσμων εξαρτήσεων και να ‘θυμούνται’ πληροφορίες για μεγάλα χρονικά διαστήματα. Όλα τα επαναλαμβανόμενα νευρωνικά δίκτυα έχουν τη μορφή μίας αλυσίδας επαναλαμβανόμενων ενοτήτων (module) του νευρικού δικτύου. Στα απλά RNN, αυτή η επαναλαμβανόμενη ενότητα έχει μια πολύ απλή δομή, όπως ένα μονό επίπεδο που χρησιμοποιεί την συνάρτηση tanh. Η σύνθεση ενός μπλοκ το κάνει πιο έξυπνο από έναν κλασσικό νευρώνα, ενώ έχει και μνήμη για την αποθήκευση πρόσφατων ακολουθιών. Κάθε

μονάδα είναι σαν μια μικρή μηχανή αποθήκευσης κατάστασης όπου οι πύλες των μονάδων έχουν βάρη που μαθαίνονται κατά τη διάρκεια της εκπαίδευσης. [43]

Οι μονάδες της long short-term μνήμης (LSTM) (ή blocks) είναι μια δομική μονάδα για στρώματα επαναλαμβανόμενου νευρικού δικτύου (RNN). Ένα RNN που αποτελείται από μονάδες LSTM ονομάζεται δίκτυο LSTM. Τα δίκτυα LSTM αντί για τους νευρώνες έχουν block μνήμης που συνδέονται μέσω επιπέδων. Ένα block λειτουργεί με μια ακολουθία εισόδου (input sequence) και κάθε πύλη εντός ενός block χρησιμοποιεί τις μονάδες ενεργοποίησης σιγμοειδούς (sigmoid activation units) για να ελέγξει εάν ενεργοποιούνται ή όχι, δηλαδή η αλλαγή κατάστασης και η προσθήκη πληροφοριών σε ένα block γίνεται κάτω από συγκεκριμένες συνθήκες. [43]

Υπάρχουν διάφορες αρχιτεκτονικές μονάδων LSTM. Μια κοινή μονάδα LSTM αποτελείται από ένα cell, μια πύλη εισόδου (input gate), μια πύλη εξόδου (output gate) και μια πύλη forget. Ένα cell μνήμης LSTM αποθηκεύει μια τιμή (ή κατάσταση), είτε μακροπρόθεσμα είτε για σύντομες χρονικές περιόδους. Αυτό επιτυγχάνεται χρησιμοποιώντας μια συνάρτηση ενεργοποίησης ταυτότητας (identity activation function) για το cell μνήμης. Έτσι, όταν ένα δίκτυο LSTM (που είναι ένα RNN που αποτελείται από μονάδες LSTM) εκπαιδεύεται με back-propagation, τότε με την πάροδο του χρόνου η κλίση (gradient) δεν τείνει να εξαφανιστεί.

Οι πύλες LSTM υπολογίζουν μια ενεργοποίηση, συνήθως με τη χρήση της συνάρτησης logistic. Η πύλη εισόδου ελέγχει την έκταση στην οποία μια νέα τιμή εισρέει μέσα στο cell, η πύλη forget ελέγχει την έκταση στην οποία παραμένει μια τιμή στο cell και η πύλη εξόδου ελέγχει την έκταση στην οποία η τιμή στο cell χρησιμοποιείται για τον υπολογισμό της ενεργοποίησης της εξόδου της μονάδας LSTM. [43]

Υπάρχουν συνδέσεις μέσα και έξω από αυτές τις πύλες. Λίγες συνδέσεις είναι επαναλαμβανόμενες. Τα βάρη αυτών των συνδέσεων μιας μονάδας LSTM χρησιμοποιούνται για να κατευθύνουν τη λειτουργία των πυλών. Αυτά τα βάρη υπολογίζονται κατά τη διάρκεια της εκπαίδευσης. Κάθε μια από τις πύλες έχει τις δικές της παραμέτρους, όπως είναι τα βάρη και το bias, σε σχέση με άλλες μονάδες εκτός της μονάδας LSTM. Η έκφραση long short-term αναφέρεται στο γεγονός ότι το LSTM είναι ένα μοντέλο για τη βραχυπρόθεσμη (short-term) μνήμη που μπορεί να διαρκέσει για μεγάλο (long) χρονικό διάστημα. [43]

Προγραμματιστικά, χρησιμοποιήθηκε η βιβλιοθήκη Keras με back-end το Tensorflow. Πιο συγκεκριμένα χρησιμοποιείται ένα στρώμα εισόδου LSTM με 10 units, που είναι η διαστάσεις του χώρου της εξόδου. Η είσοδος του στρώματος είναι τρισδιάστατη, όπου η πρώτη

διάσταση είναι ίση με το πλήθος των tweets, η δεύτερη ίση με 1 και η τρίτη ίση με το πλήθος των features που προέκυψαν κατά την κωδικοποίηση των δεδομένων. Επίσης, αυτό το στρώμα έχει την παράμετρο `return_sequences=True`, η οποία ορίζει ότι πρέπει να επιστραφεί η τελευταία έξοδος στην ακολουθία εξόδου είτε την πλήρη ακολουθία. Η τελευταία παράμετρος είναι η `activation` με τιμή '`softplus`' [44], που ορίζει την συνάρτηση ενεργοποίησης που θα χρησιμοποιηθεί. Χρησιμοποιείται ακόμα ένα στρώμα LSTM με παραμέτρους το `units=20` και την παράμετρο `activation` με τιμή '`softplus`'.

Κάτω από κάθε στρώμα LSTM χρησιμοποιείται `Dropout` [38] ίσο με 0.2. Το στρώμα εξόδου έχει ένα νευρώνα, `kernel_initializer` με τιμή '`glorot_uniform`' [37] και `activation` με τιμή '`sigmoid`'. Πριν το στρώμα εξόδου χρησιμοποιείται ένα στρώμα `Dense` με 500 νευρώνες, με `kernel_initializer` με τιμή '`glorot_uniform`', με `activation` με τιμή '`softsign`' [36] και με `kernel_constraint` με τιμή `maxnorm(2)`. Για το `compile` χρησιμοποιείται ο `optimizer RMS` [39] με learning rate 0.001, ενώ για το `loss` χρησιμοποιείται το `binary_crossentropy` και με `metrics` το `accuracy`. Τα παραπάνω αναλύθηκαν στην περιγραφή του Νευρωνικού Δικτύου.

Πριν να γίνει `fit` το δίκτυο στα δεδομένα εκπαίδευσης ορίζεται το `EarlyStopping callback`, το οποίο σταματάει την διαδικασία της εκπαίδευσης αν η ποσότητα που παρακολουθείται έχει σταματήσει να βελτιώνεται. Η ποσότητα που παρακολουθείτε ορίζεται με το όρισμα `monitor` και είναι η '`val_loss`'. Ορίζονται άλλα τρία ορίσματα, τα οποία είναι το `min_delta=0`, το `patience=2`, το `verbose=0` και το `mode='auto'`. Το όρισμα `min_delta` είναι η ελάχιστη μεταβολή στην ποσότητα που παρακολουθείται που χαρακτηρίζεται ως βελτίωση, δηλαδή για μια απόλυτη αλλαγή μικρότερη του `min_delta` θεωρείται ότι δεν υπάρχει βελτίωση.

Το όρισμα `patience` είναι το πλήθος των εποχών (`epochs`) στο οποίο δεν υπάρχει βελτίωση και μετά το οποίο η εκπαίδευση θα σταματήσει. Το όρισμα `verbose` είναι οι πληροφορίες που εμφανίζονται σχετικές με τη διαδικασία, όταν ορίζεται στο 0 δεν εμφανίζεται καμία πληροφορία. Το όρισμα `mode` ορίζει αν η εκπαίδευση θα σταματήσει όταν η ποσότητα που παρακολουθείται σταματήσει να μειώνεται ή να αυξάνεται. Πιο συγκεκριμένα, όταν το όρισμα παίρνει τιμή '`min`' τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να μειώνεται, όταν παίρνει τιμή '`max`' τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να αυξάνεται και όταν παίρνει τιμή '`auto`' τότε το πότε θα σταματήσει η εκπαίδευση εξαρτάτε από το είδος της ποσότητας που παρακολουθείται, δηλαδή της `monitor = 'val_loss'`.

Στη συνέχεια, γίνεται `fit` του μοντέλου στα δεδομένα εκπαίδευσης. Αυτό επιτυγχάνεται με τη συνάρτηση `fit()` που δέχεται τέσσερα ορίσματα. Τα δύο πρώτα ορίσματα είναι τα δεδο-

μένα εκπαίδευσης, τα κωδικοποιημένα tweets και τα ironic labels. Το τρίτο όρισμα είναι το batch_size=20, το οποίο είναι το νούμερο των στιγμιοτύπων που αξιολογούνται πριν από από κάθε ανανέωση των βαρών στο μοντέλο. Το τέταρτο όρισμα είναι το epochs=50, που ορίζει το πόσες φορές θα τρέξει η διαδικασία της εκπαίδευσης πάνω στα δεδομένα εκπαίδευσης.

Μετά την ολοκλήρωση της εκπαίδευσης, μπορεί να αξιολογηθεί η απόδοση του μοντέλου με τα ίδια δεδομένα εκπαίδευσης. Αυτό γίνεται με την συνάρτηση evaluate() που δέχεται σαν ορίσματα τα δεδομένα εκπαίδευσης που χρησιμοποιήσαμε προηγουμένως για την εκπαίδευση του μοντέλου. Η συνάρτηση αυτή παράγει προβλέψεις για τα tweets που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο και τις συγκρίνει με τα ironic labels που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο. Από αυτή τη σύγκριση παράγει βαθμολογίες, όπως ο μέσος όρος του loss και άλλες μετρικές που ορίσαμε όπως το accuracy. Επίσης, χρησιμοποιείται και η συνάρτηση summary() που εκτυπώνει τη σύνοψη του μοντέλου. Η πρόβλεψη πάνω στο test set γίνεται απλώς με την συνάρτηση predict(). Όπως αναφέρθηκε παραπάνω οι προβλεπόμενες τιμές είναι είτε 1 είτε 0, επειδή χρησιμοποιήθηκε η ‘sigmoid’ συνάρτηση στο επίπεδο εξόδου.

Για το LSTM NN χρησιμοποιήθηκαν μόνο τα word embeddings, word2vec, doc2vec και GloVe. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.8 ενώ όλα τα αποτελέσματα παρουσιάζονται και σε πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	63.32	62.88	61.85	68.17	62.25
-	doc2vec	58.39	61.13	77.73	66.67	66.51
-	Glove	60.86	61.34	63.02	64.95	61.84

Πίνακας 4.8 Αξιολόγηση LSTM Neural Net

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει το word embedding doc2vec με αισθητή διαφορά από τα υπόλοιπα.

Περιγραφή μοντέλου 1-D Convolutional Neural Network (Conv1D NN)

To convolutional neural network μίας διάστασης είναι ουσιαστικά ένα νευρωνικό δίκτυο, όπως αυτό που περιγράφηκε παραπάνω, που χρησιμοποιεί στρώματα που εκτελούν την

πράξη της συνέλιξης. Πιο συγκεκριμένα, το μονοδιάστατο στρώμα (1-D layer) δημιουργεί έναν πυρήνα συνέλιξης που εκτελεί την πράξη της συνέλιξης με το στρώμα εισόδου πάνω σε μία μόνο διάσταση, χωρική ή χρονική, για να παράγει ένα tensor εξόδων. Τα φίλτρα, γνωστά και ως πυρήνες, μπορεί να έχουν οποιοδήποτε μήκος. Το μήκος αναφέρεται στον αριθμό των γραμμών του φίλτρου. Το πλάτος του πυρήνα σε περίπτωση χαρακτήρων και αναπαραστάσεων λέξεων είναι η διάσταση ολόκληρου του word embedding ή ολόκληρης της αναπαράστασης χαρακτήρων. [45] Έτσι, η μόνη διάσταση που έχει σημασία στην περίπτωση των συνελίξεων (convolutions) σε εργασίες NLP, είναι το μήκος του φίλτρου ή το μέγεθος του φίλτρου.

Τα φίλτρα πρέπει να εκτελούν την πράξη της συνέλιξης με την είσοδο και να παράγουν την έξοδο. Ουσιαστικά, η συνέλιξη είναι ο πολλαπλασιασμός των βαρών στα φίλτρα με την αντίστοιχη αναπαράσταση των λέξεων ή των χαρακτήρων. [45] Καθένα από τα αποτελέσματα του πολλαπλασιασμού στη συνέχεια αθροίζεται και παράγει μία έξοδο. Αυτό ποικίλλει ανάλογα με συντελεστές όπως το βήμα (stride), δηλαδή πόσο το φίλτρο κινείται σε κάθε στάδιο, και το μήκος του φίλτρου. Η έξοδος της συνέλιξης εξαρτάται άμεσα από αυτούς τους δύο συντελεστές. [45]

Προγραμματιστικά, χρησιμοποιήθηκε η βιβλιοθήκη Keras με back-end το Tensorflow. Πιο συγκεκριμένα χρησιμοποιείται ένα στρώμα εισόδου Dense με 20 νευρώνες, με kernel_initializer='glorot_uniform' [37], με activation με τιμή 'softsign' [36] και με kernel_constraint με τιμή maxnorm(2). Η είσοδος του του στρώματος είναι τρισδιάστατη, όπου η πρώτη διάσταση είναι ίση με το πλήθος των tweets, η δεύτερη ίση με 1 και η τρίτη ίση με το πλήθος των features που προέκυψαν κατά την κωδικοποίηση των δεδομένων. Επίσης, πριν το στρώμα εξόδου υπάρχει ακόμα ένα Dense στρώμα με 500 νευρώνες, με kernel_initializer = 'glorot_uniform', με activation = 'softsign' και με kernel_constraint = maxnorm(2).

Το στρώμα εξόδου έχει ένα νευρώνα, kernel_initializer με τιμή 'glorot_uniform' και activation με τιμή 'sigmoid'. Για το compile χρησιμοποιείται ο optimizer RMS [39] με learning rate 0.001, ενώ για το loss χρησιμοποιείται το binary_crossentropy και με metrics το accuracy. Όλα τα παραπάνω αναλύθηκαν στην περιγραφή του μοντέλου Νευρωνικό Δίκτυο.

Ανάμεσα στο στρώμα εισόδου και το Dense στρώμα που βρίσκεται πριν το στρώμα εξόδου υπάρχουν πέντε στρώματα Conv1D. Αυτά τα στρώματα απαιτούν τρισδιάστατη είσοδο και υπολογίζουν συνέλιξη μιας διάστασης. Όλα αυτά τα στρώματα έχουν τις ίδιες παραμέτρους. Πιο συγκεκριμένα, χρησιμοποιούνται τέσσερεις παράμετροι, το filters με τιμή 32, το kernel_size με τιμή 3, το padding με τιμή 'same' και το activation με τιμή 'relu'. Η παράμε-

τρος filters είναι οι διαστάσεις του χώρου εξόδου, δηλαδή ο αριθμός των φίλτρων εξόδου στην συνέλιξη. Η παράμετρος kernel_size καθορίζει το μήκος του παραθύρου της 1D συνέλιξης. Η παράμετρος padding με την τιμή ‘same’ συμπληρώσει (padding) την είσοδο ώστε η έξοδος να έχει το ίδιο μήκος με την αρχική είσοδο. Η παράμετρος activation ορίζει την συνάρτηση ενεργοποίησης που θα χρησιμοποιηθεί.

Κάτω από το τελευταίο στρώμα Conv1D χρησιμοποιείται GlobalAveragePooling1D [42], το οποίο εκτελεί τη διαδικασία του global average pooling για τα δεδομένα και δέχεται ως είσοδο ένα τρισδιάστατο tensor της μορφής (batch_size, steps, features) και επιστρέφει ένα δισδιάστατο tensor της μορφής (batch_size, features). Τέλος, χρησιμοποιούνται δύο Dropout [38] με τιμή 0.2, ένα κάτω από πρώτο και ένα κάτω από το τελευταίο Conv1D στρώμα.

Πριν να γίνει fit το δίκτυο στα δεδομένα εκπαίδευσης ορίζεται το EarlyStopping callback, το οποίο σταματάει την διαδικασία της εκπαίδευσης αν η ποσότητα που παρακολουθείται έχει σταματήσει να βελτιώνεται. Η ποσότητα που παρακολουθείται ορίζεται με το όρισμα monitor και είναι η ’val_loss’. Ορίζονται άλλα τρία ορίσματα, τα οποία είναι το min_delta=0, το patience=2, το verbose=0 και το mode=’auto’. Το όρισμα min_delta είναι η ελάχιστη μεταβολή στην ποσότητα που παρακολουθείται που χαρακτηρίζεται ως βελτίωση, δηλαδή για μια απόλυτη αλλαγή μικρότερη του min_delta θεωρείται ότι δεν υπάρχει βελτίωση.

Το όρισμα patience είναι το πλήθος των εποχών (epochs) στο οποίο δεν υπάρχει βελτίωση και μετά το οποίο η εκπαίδευση θα σταματήσει. Το όρισμα verbose είναι οι πληροφορίες που εμφανίζονται σχετικές με τη διαδικασία, και όταν παίρνει τιμή 0 δεν εμφανίζεται καμία πληροφορία. Το όρισμα mode ορίζει αν η εκπαίδευση θα σταματήσει όταν η ποσότητα που παρακολουθείται σταματήσει να μειώνεται ή να αυξάνεται. Πιο συγκεκριμένα, όταν το όρισμα παίρνει τιμή ‘min’ τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να μειώνεται, όταν παίρνει τιμή ‘max’ τότε η εκπαίδευση σταματάει όταν η ποσότητα σταματήσει να αυξάνεται και όταν παίρνει τιμή ‘auto’ τότε το πότε θα σταματήσει η εκπαίδευση εξαρτάτε από το είδος της ποσότητας που παρακολουθείται, δηλαδή της monitor = ’val_loss’.

Στη συνέχεια, γίνεται fit του μοντέλου στα δεδομένα εκπαίδευσης. Αυτό επιτυγχάνεται με τη συνάρτηση fit() που δέχεται τέσσερα ορίσματα. Τα δύο πρώτα ορίσματα είναι τα δεδομένα εκπαίδευσης, τα κωδικοποιημένα tweets και τα ironic labels. Το τρίτο όρισμα είναι το batch_size=20, το οποίο είναι το νούμερο των στιγμιοτύπων που αξιολογούνται πριν από κάθε ανανέωση των βαρών στο μοντέλο. Το τέταρτο όρισμα είναι το epochs=50, που ορίζει το πόσες φορές θα τρέξει η διαδικασία της εκπαίδευσης πάνω στα δεδομένα εκπαίδευσης.

Μετά την ολοκλήρωση της εκπαίδευσης, μπορεί να αξιολογηθεί η απόδοση του μοντέλου με τα ίδια δεδομένα εκπαίδευσης. Αυτό γίνεται με την συνάρτηση `evaluate()` που δέχεται σαν ορίσματα τα δεδομένα εκπαίδευσης που χρησιμοποιήσαμε προηγουμένως για την εκπαίδευση του μοντέλου. Η συνάρτηση αυτή παράγει προβλέψεις για τα tweets που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο και τις συγκρίνει με τα ironic labels που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο. Από αυτή τη σύγκριση παράγει βαθμολογίες, όπως ο μέσος όρος του `loss` και άλλες μετρικές που ορίσαμε όπως το `accuracy`. Επίσης, χρησιμοποιείται και η συνάρτηση `summary()` που εκτυπώνει τη σύνοψη του μοντέλου. Η πρόβλεψη πάνω στο test set γίνεται απλώς με την συνάρτηση `predict()`. Όπως αναφέρθηκε παραπάνω οι προβλεπόμενες τιμές είναι είτε 1 είτε 0, επειδή χρησιμοποιήθηκε η ‘sigmoid’ συνάρτηση στο επίπεδο εξόδου.

Για το Conv1D NN χρησιμοποιήθηκαν μόνο τα word embeddings, word2vec, doc2vec και GloVe. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.9 ενώ όλα τα αποτελέσματα παρουσιάζονται και σε πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	62.20	61.84	59.87	66.88	60.85
-	doc2vec	61.14	60.69	62.29	65.68	60.77
-	Glove	59.83	60.27	61.55	64.28	60.62

Πίνακας 4.9 Αξιολόγηση Conv1D Neural Net

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει το word embedding word2vec, ενώ οι υπόλοιποι αλγόριθμοι δεν έχουν μεγάλη διαφορά στο score.

Περιγραφή μοντέλου SVM

Το μοντέλο Support Vector Machine (SVM) αναπτύχθηκε αναλυτικά στο κεφάλαιο 3.2.3. Το μοντέλο SVM είναι ένας εποπτευόμενος (supervised) αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα παλινδρόμησης (regression) αλλά και για προβλήματα ταξινόμησης. Η χρήση του σε προβλήματα ταξινόμησης είναι πιο διαδεδομένη. Το μοντέλο SVM είναι ένας ταξινομητής που χρησιμοποιεί ένα hyperplane για να διαχωρίσει τα δεδομένα σε κλάσεις. Με άλλα λόγια, αν έχουμε δεδομένα εκπαίδευσης με ετικέτες (labels) που τα χαρα-

κτηρίζουν, τότε ο αλγόριθμος παράγει ένα hyperplane που κατηγοριοποιεί νέα δεδομένα. Σε δισδιάστατο χώρο το hyperplane είναι μια γραμμή, η μορφολογία της οποίας εξαρτάτε από το kernel του μοντέλου το οποίο μπορεί να είναι linear ή μη γραμμικό, που χωρίζει τον χώρο σε δύο κομμάτια με βέλτιστο τρόπο, όπου κάθε ομάδα βρίσκεται σε διαφορετική πλευρά. Το μοντέλο μπορεί να κάνει προβλέψεις χρησιμοποιώντας αυτή τη γραμμή. Χρησιμοποιώντας τα δεδομένα εισόδου στην εξίσωση της γραμμής, υπολογίζεται αν το νέα σημείο είναι πάνω ή κάτω από τη γραμμή. Η απόσταση μεταξύ της γραμμής και των κοντινότερων σημείων δεδομένων ονομάζεται περιθώριο (margin). Η πιο βέλτιστη γραμμή που μπορεί να διαχωρίσει τις δύο κλάσεις είναι αυτή που πετυχαίνει το μεγαλύτερο περιθώριο. Αυτή η γραμμή ονομάζεται Maximal-Margin hyperplane.

Το περιθώριο υπολογίζεται ως η κάθετη απόσταση από τη γραμμή με τα κοντινότερα σημεία σε αυτή. Μόνο αυτά τα σημεία είναι χρήσιμα για τον προσδιορισμό της γραμμής και της κατασκευή του ταξινομητή. Αυτά τα σημεία ονομάζονται support vectors, αφού υποστηρίζουν ή προσδιορίζουν το hyperplane. Το hyperplane προσδιορίζεται από τα δεδομένα εκπαίδευσης χρησιμοποιώντας μια διαδικασία βελτιστοποίησης που μεγιστοποιεί το περιθώριο. Η εύρεση του μέγιστου περιθωρίου είναι η κύρια διαφορά του μοντέλου από τους άλλους αλγορίθμους ταξινόμησης, δηλαδή δεν βρίσκει απλά ένα όριο της απόφασης αλλά βρίσκει το βέλτιστο. Αξίζει να σημειωθεί ότι το μοντέλο επιλέγει το hyperplane δίνοντας προτεραιότητα στην πιο ακριβή ταξινόμηση των κλάσεων αντί της μεγιστοποίησης του περιθωρίου.

Όμως, το μοντέλο SVM όταν συναντήσει ένα στοιχείο που βρίσκεται σε παράξενη θέση σε σχέση με τα υπόλοιπα της ομάδας του, τότε το θεωρεί απόκλιση (outliers) και το αγνοεί. Επιπροσθέτως, εκτός από γραμμική ταξινόμηση το μοντέλο μπορεί να εκτελέσει αποτελεσματικά και μη γραμμική ταξινόμηση με τη χρήση του kernel trick. Το kernel trick χαρτογραφεί έμμεσα την είσοδο από χώρο με λίγες διαστάσεις σε χώρο με πολλές διαστάσεις (high-dimensional feature spaces) και στη συνέχεια εκτελείται η διαδικασία διαχωρισμού των δεδομένων που περιγράφηκε παραπάνω.

Προγραμματιστικά, από το `svm` της βιβλιοθήκης `sklearn` χρησιμοποιήθηκε το μοντέλο SVC με τρία ορίσματα, το `kernel='rbf'`, το `C=100` και το `gamma=0.1`. Όταν το `C` παίρνει μεγάλες τιμές, η βελτιστοποίηση θα γίνει με μικρότερο περιθώριο μεταξύ του hyperplane και των support vectors αν αυτό το hyperplane ταξινομεί σωστά όλα τα σημεία εκπαίδευσης. Με άλλα λόγια, το περιθώριο είναι αυστηρό και σημεία δεν μπορούν να βρίσκονται μέσα του. Αντιθέτως, όταν το `C` παίρνει πολύ μικρές τιμές τότε ο βελτιστοποιητής θα ψάξει για μεγαλύ-

τερο περιθώριο μεταξύ του hyperplane και των support vectors, ακόμα και αν το hyperplane ταξινομεί εσφαλμένα περισσότερα σημεία. Επίσης, το C επηρεάζει το πλήθος των support vectors που χρησιμοποιεί το μοντέλο, επειδή το C επηρεάζει το πλήθος των σημείων που επιτρέπεται να βρίσκονται μέσα στο περιθώριο. Όσο πιο μικρή είναι η τιμή του C, τόσο πιο ευαίσθητος είναι ο αλγόριθμος στα δεδομένα εκπαίδευσης, δηλαδή υπάρχει υψηλότερη διακύμανση (higher variance) και χαμηλότερο bias. Όσο πιο μεγάλη είναι η τιμή του C, τόσο λιγότερο ευαίσθητος είναι ο αλγόριθμος σε δεδομένα εκπαίδευσης, δηλαδή υπάρχει χαμηλότερη διακύμανση (lower variance) και υψηλότερο bias.

Η παράμετρος gamma ορίζει πόσο μακριά φτάνει η επιρροή ενός στοιχείου εκπαίδευσης. Όταν παίρνει χαμηλότερες τιμές τότε φτάνει μακριά ενώ όταν παίρνει υψηλότερες τιμές τότε φτάνει κοντά. Με άλλα λόγια όταν το gamma έχει χαμηλή τιμή, τα σημεία που βρίσκονται μακριά από την πιθανή γραμμή διαχωρισμού συμπεριλαμβάνονται κατά τον υπολογισμό την γραμμής διαχωρισμού (hyperplane). Σε αντίθεση, όταν το gamma έχει υψηλή τιμή τότε τα σημεία κοντά στην πιθανή γραμμή διαχωρισμού συμπεριλαμβάνονται κατά τον υπολογισμό την γραμμής διαχωρισμού (hyperplane). Το μοντέλο SVM στην πράξη υλοποιείται με τη χρήση ενός πυρήνα (kernel), ο οποίος υπολογίζει το hyperplane ανάλογα με το είδος του. Ο kernel αλγόριθμος που χρησιμοποιείται είναι ο Radial Basis Function (RBF).

Στους συνδυασμούς αλγορίθμων PCA με One-Hot-Encoding, Univariate Selection με One-Hot-Encoding και SVD με One-Hot-Encoding πριν χρησιμοποιηθούν τα δεδομένα εκπαίδευσης και τα δεδομένα πρόβλεψης από το μοντέλο μετασχηματίζονται με το StandardScaler της βιβλιοθήκης sklearn, γιατί δίνουν καλύτερα αποτελέσματα σε σύγκριση με τα αποτελέσματα χωρίς scaling. Όλοι οι υπόλοιποι συνδυασμοί αλγορίθμων δεν χρησιμοποιούν scaling στα δεδομένα. Το StandardScaler κανονικοποιεί (standardize) τα χαρακτηριστικά (features) αφαιρώντας τη διακύμανση μέσου όρου (mean) και κλιμακώνοντας σε μοναδιαία διακύμανση (scaling to unit variance). Το κεντράρισμα και η κλιμάκωση γίνεται ανεξάρτητα για κάθε χαρακτηριστικό υπολογίζοντας τα σχετικά στατιστικά στα δείγματα των δεδομένων εκπαίδευσης. Η διακύμανση μέσου όρου (mean) και η τυπική απόκλιση (standard deviation) αποθηκεύονται για να χρησιμοποιηθούν και στο test set με την μέθοδο transform.

Η κανονικοποίηση είναι σημαντική γιατί πολλά στοιχεία που χρησιμοποιούνται στην αντικειμενική συνάρτηση (objective function) ενός αλγορίθμου μάθησης, όπως το RBF kernel που χρησιμοποιούμε στο μοντέλο, υποθέτουν ότι όλα τα χαρακτηριστικά έχουν κέντρο γύρω από το μηδέν και ότι έχουν αντίστοιχη διακύμανση (variance). Αν ένα χαρακτηριστικό έχει

διακύμανση που είναι πολύ μεγαλύτερη από αυτή που έχουν τα άλλα, μπορεί να κυριαρχήσει στην αντικειμενική συνάρτηση με αποτέλεσμα να κάνει τον εκτιμητή να μην μπορεί μπορεί να μάθει σωστά από τα υπόλοιπα χαρακτηριστικά.

Για το SVM χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.10 ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds).

		Precision	Accuracy	Recall	ROC	F1-score
-	Bigrams	58.35	60.98	76.24	61.11	66.03
-	doc2vec	59.07	61.34	74.06	61.50	65.60
PCA	TF-IDF	65.12	65.25	65.72	65.35	65.32
PCA	Bigrams	59.61	61.68	72.24	61.79	65.24
Feature Importance	Bigrams	60.03	61.94	72.00	62.12	65.31
Univariate Selection	Bigrams	58.85	61.16	74.13	61.29	65.52
TruncatedSVD	Bigrams	59.54	61.65	72.62	61.78	65.34

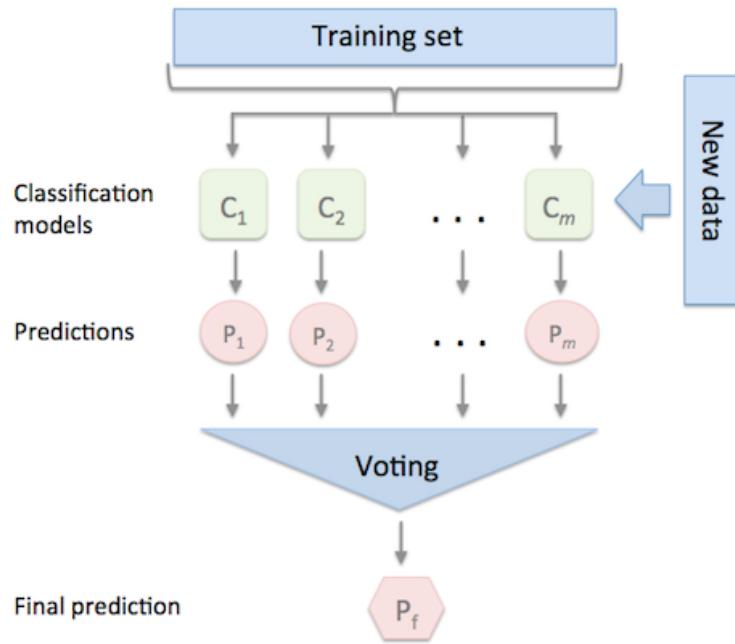
Πίνακας 4.10 Αξιολόγηση SVM

Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει ο αλγόριθμος Bigrams, και ακολουθεί το word embedding doc2vec με μικρή διαφορά.

Περιγραφή μοντέλου Voting Ensembles

Το μοντέλο Voting Ensembles δημιουργεί πολλά μοντέλα τα οποία τα συνδυάζει για να παράγει ένα καλύτερο αποτέλεσμα. Πιο συγκεκριμένα, συνδυάζει παρόμοια ή εννοιολογικά διαφορετικά μοντέλα μηχανικής μάθησης για ταξινόμηση και παράγει αποτελέσματα μέσω ψηφοφορίας όπου επικρατεί η πλειοψηφία. Στην εικόνα 4.3 απεικονίζεται ο γενικός τρόπος λειτουργίας του μοντέλου. Το μοντέλο υλοποιεί ‘σκληρή’ (‘hard’) και ‘μαλακή’ (‘soft’) ψηφοφορία (voting). Στη μέθοδο της ‘σκληρής’ ψηφοφορίας, η ετικέτα της τελικής κλάσης (test set) παίρνει την τιμή της ετικέτας που έχει δώσει το μεγαλύτερο πλήθος των μοντέλων ταξινόμησης, δηλαδή της πλειοψηφίας. Στην ‘μαλακή’ ψηφοφορία, προβλέπονται οι ετικέτες των

κλάσεων με βάση το argmax των αθροισμάτων των προβλεπόμενων πιθανοτήτων, δηλαδή παίρνει την τιμή της ετικέτας που έχει το μέγιστο μέσο όρο των πιθανοτήτων που έχουν οι ετικέτες που προέβλεψαν οι ταξινομητές. Οι μέσοι όροι υπολογίζονται ξεχωριστά από τις πιθανότητες των ταξινομητών για κάθε ετικέτα, δηλαδή υπολογίζεται ο μέσος όρος των πιθανοτήτων για την ετικέτα 1 που αντιστοιχεί σε κάθε ταξινομητή και έπειτα για την ετικέτα 0.



Εικόνα 4.3 Τρόπος λειτουργίας του μοντέλου Voting Ensembles

Πιο αναλυτικά, το hard voting (σκληρή ψηφοφορία) ή αλλιώς majority voting προβλέπει την ετικέτα της κλάσης μέσω της πλειοψηφίας των ετικετών που έχουν προβλέψει οι ταξινομητές. Για παράδειγμα, έστω ότι έχουμε τρεις ταξινομητές που ταξινομούν ένα δείγμα δεδομένων εκπαίδευσης με τον εξής τρόπο:

- classifier 1 → label class 0
- classifier 2 → label class 0
- classifier 3 → label class 1

Παρατηρούμε ότι δύο ταξινομητές προέβλεψαν την ετικέτα 0 και ένας την ετικέτα 1, οπότε μέσω της ψήφου της πλειοψηφίας προκύπτει ότι η τελική ετικέτα του δείγματος θα είναι 0.

Προγραμματιστικά, χρησιμοποιείται το VotingClassifier της βιβλιοθήκης sklearn. Το VotingClassifier έχει δύο ορίσματα, τα μοντέλα μηχανικής μάθησης και το voting=hard. Τα μοντέλα της μηχανικής μάθησης που χρησιμοποιούνται είναι συνολικά έξι, τέσσερα SVM, ένα K-Neighbors, με μόνο παράμετρο το n_neighbors με τιμή 140, και ένα Bernoulli Naive

Bayes, με default παραμέτρους. Και τα έξι μοντέλα υλοποιούνται όπως περιγράφηκε παραπάνω και έχουν ακριβώς την ίδια παραμετροποίηση που αναφέρθηκε παραπάνω, εκτός από τα SVM που μόνο το ένα από τα τρία έχει την ίδια παραμετροποίηση.

Τα άλλα τρία μοντέλα SVM έχουν παραμέτρους τέτοιες ώστε ένα μοντέλο τείνει να πάσχει από overfitting, ένα μοντέλο τείνει να πάσχει από underfitting και ένα μοντέλο να είναι σε λίγο πιο fitted από αυτό της ενδιάμεσης κατάστασης. Οι παράμετροι του SVM που χρησιμοποιήθηκαν στο μοντέλο που παρουσιάστηκε παραπάνω, καθιστούν το μοντέλο ως ενδιάμεσης κατάστασης. Το μοντέλο που τείνει να πάσχει από overfitting έχει $C=10000$, το μοντέλο που τείνει να πάσχει από underfitting έχει $C=10$, το μοντέλο ενδιάμεσης κατάστασης έχει $C=100$, το μοντέλο που είναι λίγο πιο fitted από αυτό της ενδιάμεσης κατάστασης έχει $C=1000$, ενώ και τα τέσσερα μοντέλα έχουν ίδιο gamma ίσο με 0,1 και ίδιο kernel ίσο με ‘rbf’. Η λειτουργία του ορίσματος voting περιγράφηκε παραπάνω.

		Precision	Accuracy	Recall	ROC	F1-score
PCA	Bigrams	61.27	62.33	66.96	62.44	63.89
PCA	TF-IDF	64.95	64.50	63.32	64.63	63.97
-	doc2vec	60.90	62.28	68.82	62.42	64.49
Univariate Selection	Bigrams	62.28	62.83	64.71	62.89	63.40
TruncatedSVD	Bigrams	61.67	62.72	67.87	62.88	64.41
TruncatedSVD	TF-IDF	65.22	64.76	63.56	64.89	64.22
Feature Importance	Bigrams	61.52	62.62	67.87	62.78	64.36
Feature Importance	TF-IDF	64.64	64.39	64.05	64.56	64.14

Πίνακας 4.11 Αξιολόγηση Voting Ensembles

Για το Voting Ensembles χρησιμοποιήθηκαν όλοι οι συνδυασμοί των αλγορίθμων encoding καθώς και όλοι οι αλγόριθμοι για το feature selection. Τα καλύτερα αποτελέσματα χρησιμοποιώντας 10-fold-cross-validation στο train set παρουσιάζονται στον πίνακα 4.11 ενώ όλα τα αποτελέσματα με όλους τους συνδυασμούς των αλγορίθμων encoding και feature selection παρουσιάζονται στον πίνακα που υπάρχει στο Παράρτημα (τα νούμερα είναι οι μέσοι όροι των μετρικών και για τα 10 folds). Η αξιολόγηση των μοντέλων γίνεται από την μετρική f1-score, όπως ορίζεται και από τον διαγωνισμό. Όπως παρατηρούμε το καλύτερο f1-score το πετυχαίνει το word embedding doc2vec, ενώ οι υπόλοιποι συνδυασμοί αλγορίθμων πετυχαίνουν αντίστοιχα καλά αποτελέσματα με τους περισσότερους να έχουν αμελητέα διαφορά.

4.2.6 Τρόπος αξιολόγησης - Evaluation

Για την αξιολόγηση των αλγορίθμων χρησιμοποιούνται τέσσερεις μετρικές που αναφέρθηκαν στο κεφάλαιο 3.3. Αυτές οι μετρικές είναι η ορθότητα (precision), η ακρίβεια (accuracy), η ανάκληση (recall) και ο αρμονικός μέσος (F1-score). Η κατάταξη των αποτελεσμάτων των αλγορίθμων θα γίνει με βάση το F1-score, όπου μεγαλύτερο F1-score σημαίνει καλύτερο και πιο αποδοτικό μοντέλο. Πρέπει να τονιστεί ότι η μετρική accuracy υπολογίζεται για όλες τις κλάσεις, δηλαδή προκύπτει από τον μέσο όρο του accuracy των ετικετών 0 και 1. Σε αντίθεση, οι μετρικές precision, recall και F1-score υπολογίζονται μόνο για τη θετική κλάση, δηλαδή υπολογίζονται μόνο για τις ετικέτες που έχουν τιμή 1. Οι παραπάνω υπολογισμοί των μετρικών για τις συγκεκριμένες κλάσεις ορίζονται από τον διαγωνισμό του SemEval2018. Επίσης, χρησιμοποιείται μία πρόσθετη μετρική, που δεν απαιτείται από τον διαγωνισμό, το Receiver Operating Characteristic - Area Under Curve (ROC-AUC) το οποίο αναλύθηκε και αυτό στο κεφάλαιο 3.3.

Προγραμματιστικά, για τον υπολογισμό της ROC-AUC χρησιμοποιείται η συνάρτηση roc_auc_score της βιβλιοθήκης sklearn. Για την μετρική accuracy χρησιμοποιείται η συνάρτηση accuracy_score της βιβλιοθήκης sklearn, ενώ για τις υπόλοιπες μετρικές, δηλαδή για το precision, recall και το F1-score, χρησιμοποιείται η συνάρτηση precision_recall_fscore_support της βιβλιοθήκης sklearn. Η συνάρτηση precision_recall_fscore_support έχει το όρισμα average='binary' που σε συνδυασμό με το default όρισμα pos_label=1 δίνει τις αντίστοιχες τιμές των μετρικών για την ετικέτα με τιμή 1.

Η συνάρτηση precision_recall_fscore_support επιστρέφει τέσσερεις τιμές εκ των οποίων οι πρώτες τρεις είναι με τη σειρά το precision, recall και το F1-score. Η τέταρτη τιμή είναι το support το οποίο δεν χρησιμοποιείται. Όλες οι μετρικές που προαναφέρθηκαν υπολογίζονται για κάθε fold του 10-fold-cross-validation, αλλά η κατάταξη σχηματίζεται μόνο από το μέσο όρο των μετρικών και για τα 10 folds. Έτσι, σε κάθε fold υπολογίζονται οι μετρικές και αθροίζονται σε αντίστοιχες μεταβλητές που θα χρησιμοποιούν για τον υπολογισμό του μέσου όρου των μετρικών και για τα 10 folds.

Πέρα από τις μετρικές χρησιμοποιήθηκαν και δύο καμπύλες (curves) για την οπτικοποίηση της καμπύλης learning curve καθώς και της καμπύλης ROC. Οι καμπύλες αυτές χρησιμοποιούνται μόνο για τα τρία καλύτερα μοντέλα που πέτυχαν τα καλύτερα F1-score. Οι εικόνες των καμπυλών βρίσκονται στο Παράρτημα μέρος E.

Περιγραφή Learning Curve

Η learning curve χρησιμοποιήθηκε για να ελεγχθεί αν το μοντέλο κάνει overfit στα δεδομένα εκπαίδευσης. Πιο συγκεκριμένα, η learning curve έχει στο άξονα των Y το score του μοντέλου ενώ στον άξονα των X το πλήθος των στοιχείων εκπαίδευσης. Στο σχήμα υπάρχουν δύο καμπύλες εκ των οποίων η κόκκινη είναι για το training score ενώ η πράσινη είναι για το cross-validation score. Έτσι, για να αποκομίσουμε πληροφορίες από την καμπύλη κοιτάμε την δεξιά πλευρά όπου υπάρχουν αρκετά στοιχεία εκπαίδευσης για την αξιολόγηση.

Αν οι δύο καμπύλες είναι κοντά η μία στην άλλη και ταυτόχρονα έχουν και οι δύο χαμηλό score, τότε το μοντέλο πάσχει από το πρόβλημα του under fitting (High Bias). Αν η training curve έχει πολύ καλύτερο score από την cross-validation curve, δηλαδή αν υπάρχει μεγάλο κενό ανάμεσα στις δύο καμπύλες, τότε το μοντέλο πάσχει από πρόβλημα του over fitting (High Variance). Επίσης, χρησιμοποιείται για την επιλογή παραμέτρων στα μοντέλα [46], για τη σύγκριση μοντέλων [47], καθώς και για τον καθορισμό της ποσότητας δεδομένων που χρειάζεται για την εκπαίδευση. [48]

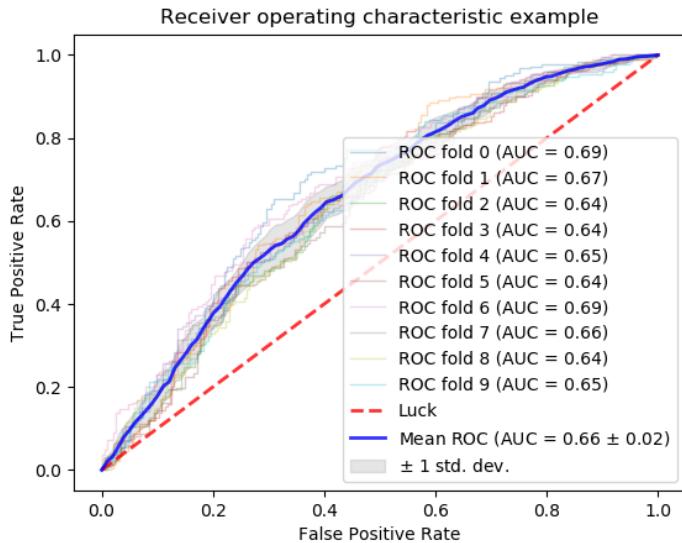
Περιγραφή ROC-AUC Curve

Όπως προαναφέρθηκε η ROC-AUC Curve αναπτύχθηκε αναλυτικά στο κεφάλαιο 3.3. Η καμπύλη χρησιμοποιείται για να αξιολογήσει την ποιότητα εξόδου του ταξινομητή που πρέκυψε χρησιμοποιώντας cross-validation. Οι καμπύλες ROC συνήθως έχουν στον Y άξονα το true positive rate και στον X άξονα το false positive rate. Αυτό σημαίνει ότι η πάνω αριστερή γωνία του σχήματος είναι το “ιδανικό” σημείο, γιατί εκεί έχουμε μηδέν false positive rate και ένα true positive rate. Αυτό σε πραγματικά δεδομένα δεν είναι εφικτό, αλλά υποδεικνύει ότι μια μεγαλύτερη περιοχή κάτω από την καμπύλη (area under the curve - AUC) είναι συνήθως καλύτερη. Ένα σημαντικό χαρακτηριστικό που είναι καλό να έχουν οι καμπύλες ROC είναι η ‘απότομη’ κλίση, αφού το ιδανικό είναι να μεγιστοποιηθεί το true positive rate ενώ ελαχιστοποιείται το false positive rate.

Έστω για παράδειγμα η καμπύλη της εικόνας 4.4 (η καμπύλη είναι αποτέλεσμα του μοντέλου Neural Net με Word2Vec). Το παράδειγμα δείχνει την ανταπόκριση της μετρικής ROC σε διαφορετικά dataset που δημιουργήθηκαν από 10-fold-cross-validation. Με τη χρήση όλων αυτών των καμπυλών μπορούμε να υπολογίσουμε τη μέση περιοχή κάτω από την καμπύλη (mean area under curve) και να δούμε τη διακύμανση της καμπύλης όταν τα δεδομένα εκπαίδευσης χωρίζονται σε διαφορετικά υπο-σετ δεδομένων. Αυτό δείχνει κατά προσέγ-

γιση πως η έξοδος του ταξινομητή επηρεάζεται από τις αλλαγές στα δεδομένα εκπαίδευσης, δηλαδή μια προσέγγιση του πόσο σταθερό είναι το μοντέλο, καθώς και πόσο διαφορετικά είναι τα υπο-σετ που παράγονται από το 10-fold-cross-validation μεταξύ τους. Στην εικόνα η κόκκινη γραμμή, που ονομάζεται Luck, αναπαριστά ένα ‘φανταστικό’ μοντέλο το οποίο παράγει προβλέψεις στην τύχη. Συνεπώς, κάτω από την κόκκινη γραμμή το μοντέλο είναι πολύ ‘κακό’, ενώ όσο πιο πάνω είναι από την γραμμή τόσο καλύτερο και πιο αποδοτικό είναι.

Το ιδανικό μοντέλο θα έχει καμπύλη που τείνει στην πάνω αριστερή γωνία, το οποίο συνεπάγεται ότι θέλουμε ο χώρος κάτω από την καμπύλη να είναι όσο το δυνατό πιο μεγάλος γίνεται. Στην εικόνα απεικονίζονται οι καμπύλες ROC για κάθε ένα από τα 10 folds του 10-fold-cross-validation με διαφορετικό χρώμα. Η καμπύλη με το βαθύ μπλε χρώμα, που ονομάζεται Mean ROC, είναι ο μέσος όρος των καμπυλών του 10-fold-cross-validation και αναπαριστά μια γενική προσέγγιση του πως αντιδρά το μοντέλο σε διάφορα σετ δεδομένων.



Εικόνα 4.4 Καμπύλη ROC-AUC για 10-fold-cross-validation (Neural Net με Word2Vec)

4.3 Τελικά αποτελέσματα, συμπεράσματα και μελλοντικοί στόχοι

Για την συλλογή των αποτελεσμάτων χρησιμοποιήθηκαν δύο προσεγγίσεις. Η πρώτη είναι στα πλαίσια του διαγωνισμού του SemEval 2018, όπου δοκιμάστηκαν όλοι οι αλγόριθμοι στο 10-fold-cross-validation και από αυτούς επιλέχθηκαν μόνο οι τρεις καλύτεροι όπως υποδεικνύει το F1-score, αλλά και κάποιες παρατηρήσεις που θα αναφερθούν παρακάτω. Πέρα από τους τρεις καλύτερους αλγορίθμους, στόχος ήταν να χρησιμοποιηθεί και ο αλγόριθμος Voting Ensembles με συνδυασμό των καλύτερων μοντέλων, αλλά λόγω περιορισμένου χρόνου και

έλλειψη ελέγχων μέχρι την καταληκτική ημερομηνία του διαγωνισμού δεν συμπεριλήφθηκε στα αποτελέσματα του διαγωνισμού. Ωστόσο, εκπρόθεσμα του διαγωνισμού δοκιμάστηκε και όπως θα δούμε παρακάτω, πέτυχε το καλύτερο F1-score στο gold set με σημαντική διαφορά. Επίσης, εκπρόθεσμα δοκιμάστηκαν και οι αλγόριθμοι LSTM και Conv1D, οι οποίοι σύμφωνα με έρευνες δουλεύουν πολύ καλά σε τέτοιου είδους προβλήματα.

Οι αλγόριθμοι που επιλέχθηκαν, δηλαδή μόνο οι τρεις καλύτεροι, χρησιμοποιήθηκαν για την παραγωγή των τελικών αποτελεσμάτων με δεδομένα εκπαίδευσης ολόκληρο το αρχείο “SemEval2018-T3-train-taskA_emoji.txt” και δεδομένα πρόβλεψης το gold set αρχείο με όνομα “SemEval2018-T3_gold_test_taskA_emoji.txt” [31]. Η δεύτερη προσέγγιση είναι στα πλαίσια της πτυχιακής, όπου επιλέγονται όλα τα μοντέλα που είχαν τα υψηλότερα F1-score ανά αλγόριθμο και όχι μόνο οι τρεις καλύτεροι. Αυτά τα μοντέλα είναι αυτά που υπάρχουν στους πίνακες των αποτελεσμάτων που παρουσιάζονται παραπάνω στην περιγραφή του κάθε αλγορίθμου. Τα τελικά αποτελέσματα με τα gold set δεδομένα ως δεδομένα πρόβλεψης προκύπτουν όπως περιγράφηκε και στην προσέγγιση του διαγωνισμού παραπάνω.

Επειδή χρησιμοποιήθηκαν τρεις εκδοχές του αλγορίθμου Naive Bayes, για την επιλογή των καλύτερων αλγορίθμων στα πλαίσια του διαγωνισμού λήφθηκαν υπόψη τα καλύτερα F1-score μόνο από μια εκδοχή του αλγορίθμου Naive Bayes, αφού τα μεγαλύτερα F1-score ήταν παραπλήσια και στους τρεις αλγορίθμους. Γενικότερα, στα πλαίσια της πτυχιακής δοκιμάστηκαν πολλοί αλγόριθμοι για ερευνητικούς λόγους και πιο συγκεκριμένα για να διαπιστωθεί ποιοι δουλεύουν και με τι ποσοστό επιτυχίας στο συγκεκριμένο πρόβλημα.

Έτσι, οι αλγόριθμοι που είναι παραλλαγές του Naive Bayes και ο αλγόριθμος Logistic Regression, παρά το γεγονός ότι δεν φημίζονται για αποτελέσματα που παράγουν σε τέτοιου είδους προβλήματα, δοκιμάστηκαν για να διαπιστωθεί αυτή η παρατήρηση. Στη συνέχεια, ερευνήθηκαν όλοι οι συνδυασμοί των αλγορίθμων κωδικοποίησης, επιλογής χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης σε συνδυασμό με την προσθήκη επιπλέον χαρακτηριστικών μέσα από τη διαδικασία της εξαγωγής χαρακτηριστικών. Επιπλέον, για την βελτίωση των αποτελεσμάτων των αλγορίθμων χρησιμοποιήθηκε η μέθοδος Grid Search για την εύρεση της καλύτερης τιμής των παραμέτρων. Παράλληλα, έγιναν και πολλές χειροκίνητες δοκιμές για να δοκιμαστούν πιο πολύπλοκοι συνδυασμοί τιμών των παραμέτρων σε όλους τους αλγορίθμους μηχανικής μάθησης, καθώς και διαφορετικές τοπολογίες σε όλα τα είδη νευρωνικών δικτύων. Γενικότερα, οι εκτενείς δοκιμές και η παραμετροποίηση για την βελτίωση της απόδοσης των μοντέλων αποτέλεσαν χρονοβόρο και σχετικά μεγάλο κομμάτι της πτυχιακής.

Εκτός από δοκιμές για τις παραμέτρους των αλγορίθμων μηχανικής μάθησης, έγιναν και δοκιμές για τις παραμέτρους των αλγορίθμων επιλογής χαρακτηριστικών και κωδικοποίησης των δεδομένων. Παρακάτω παραθέτονται τα αποτελέσματα των αλγορίθμων για το 10-fold-cross-validation καθώς και τα τελικά αποτελέσματα που προκύπτουν από το gold set για τις δύο προσεγγίσεις που αναφέρθηκαν παραπάνω. Αξίζει να σημειωθεί ότι αλγόριθμος RFE απαιτεί πάρα πολλές ώρες να παράξει αποτέλεσμα, σε επεξεργαστή Intel i7-4710MQ και με κάρτα γραφικών Nvidia 840M κάνει πάνω από δέκα ώρες για το 10-fold-cross-validation.

Έτσι, κρίθηκε μη πρακτικός και δεν χρησιμοποιήθηκε για την παραγωγή αποτελεσμάτων. Γενικότερα, ο αλγόριθμος δεν είναι πρακτικός σε προβλήματα με μεγάλο data set, καθώς όπως είδαμε σε ένα πρόβλημα με σχετικά μικρό data set απαιτεί πολλές ώρες για παράξει αποτέλεσμα. Όλα τα αποτελέσματα των αλγορίθμων για το 10-fold-cross-validation καθώς και τα τελικά αποτελέσματα στο gold set υπάρχουν στο μέρος B του Παραρτήματος.

Αποτελέσματα καλύτερων αλγορίθμων στα πλαίσια του διαγωνισμού SemEval 2018

Παρακάτω παρουσιάζονται, στον πίνακα 4.12, τα αποτελέσματα των τριών καλύτερων μοντέλων στο 10-fold-cross-validation και τα αντίστοιχα αποτελέσματα που πέτυχαν στο gold set, στον πίνακα 4.13.

Classifier	Feature Selection	Encoder	Precision	Accuracy	Recall	ROC	F1-score
SVM	-	doc2vec	59.07	61.34	74.06	61.50	65.60
K-Neighbors	PCA	TF-IDF	64.44	64.42	64.60	64.56	64.36
Gaussian	-	word2vec	53.23	55.63	90.93	55.74	67.10
Naive Bayes							

Πίνακας 4.12 Αξιολόγηση των τριών καλύτερων μοντέλων με 10-fold-cross-validation

Classifier	Feature Selection	Encoder	Precision	Accuracy	Recall	ROC	F1-score
SVM	-	doc2vec	51.05	61.60	78.13	64.43	61.75
K-Neighbors	PCA	TF-IDF	54.32	64.54	65.25	64.66	59.35
Gaussian	-	word2vec	46.09	54.71	83.60	59.66	59.42
Naive Bayes							

Πίνακας 4.13 Τελική αξιολόγηση των τριών καλύτερων μοντέλων με gold set

Στην εικόνα Δ.1, που βρίσκεται στο μέρος Δ του Παραρτήματος, παρουσιάζονται τα αποτελέσματα του διαγωνισμού του SemEval 2018 για το Task 3 Part A.[31] Στα πλαίσια του διαγωνισμού τα μοντέλα που δοκιμάστηκαν στο gold set είναι ένα Gaussian Naive Bayes, ένα K-Neighbors και ένα SVM. Επιλέχθηκαν έτσι ώστε να δοκιμαστεί ένα μοντέλο από κάθε είδος, ενώ τα μοντέλα που δεν δέχονται παραμέτρους που επηρεάζουν την απόδοση τους ή που σύμφωνα με πειράματα σε αντίστοιχα προβλήματα δεν πετυχαίνουν κορυφαία αποτελέσματα, θεωρήθηκε ως ένα είδος. Στην κατηγορία των μοντέλων που δεν δέχονται ορίσματα που επηρεάζουν την απόδοση του αλγορίθμου και που δεν πετυχαίνουν κορυφαία αποτελέσματα ανήκουν όλα τα είδη αλγορίθμων Naive Bayes, καθώς και μοντέλα που σύμφωνα με έρευνες και πειράματα δεν πετυχαίνουν κορυφαία αποτελέσματα, όπως Logistic Regression.

Από κάθε κατηγορία επιλέγεται το μοντέλο με το πιο υψηλό F1-score. Η επιλογή του τελικού αλγορίθμου έγινε με γνώμονα τα αποτελέσματα στο 10-fold-cross-validation αλλά και την παρατήρηση του ότι το μοντέλο SVM πετυχαίνει καλά αποτελέσματα σε τέτοιου είδους προβλήματα, το οποίο πηγάζει από δημοσιεύσεις αποτελεσμάτων προηγούμενων ετών του διαγωνισμού. Αξίζει να σημειωθεί ότι το μοντέλο SVM με τη χρήση των Bigrams πάσχει από overfitting, το οποίο προέκυψε από την Learning Curve, όπως θα δούμε παρακάτω.

Έτσι, για το μοντέλο SVM επιλέχθηκε ο δεύτερος καλύτερος συνδυασμός αλγορίθμων που είναι ο doc2vec. Σύμφωνα, με τα παραπάνω αποτελέσματα και τα αποτελέσματα του διαγωνισμού αν λάμβανα μέρος με το μοντέλο SVM θα έφτανα στην δέκατη τρίτη με δέκατη τέταρτη θέση, αφού το μοντέλο SVM με doc2vec πετυχαίνει ακριβώς το ίδιο F1-score με τον δέκατο τρίτο διαγωνιζόμενο. Επίσης, παρατηρούμε ότι οι επόμενοι έξι διαγωνιζόμενοι, μέχρι και τη θέση οκτώ, έχουν πετύχει αποτέλεσμα που διαφέρει μόλις κατά δύο δεκαδικά ψηφία, ενώ οι επόμενοι δύο, δηλαδή θέση έντεκα και δώδεκα, διαφέρουν μόνο κατά 0,09 και 0,27 μονάδες αντίστοιχα. Παρακάτω αναφέρεται και η θέση που πέτυχαν τα μοντέλα που υλοποιήθηκαν εκπρόθεσμα του διαγωνισμού, τα οποία πέτυχαν και πολύ καλύτερη θέση.

Αποτελέσματα καλύτερων αλγορίθμων στα πλαίσια της πτυχιακής και συμπεράσματα

Στο Γ μέρος του Παραρτήματος παρουσιάζονται τα αποτελέσματα στο gold set που αντιστοιχούν στα καλύτερα αποτελέσματα από κάθε μοντέλο που πέτυχαν στο 10-fold-cross-validation, που αναφέρθηκαν στους παραπάνω πίνακες κατά την περιγραφή των αλγορίθμων. Από τα αποτελέσματα παρατηρούμε ότι δεν υπάρχει συγκεκριμένος συνδυασμός αλγορίθμου κωδικοποίησης με αλγορίθμου επιλογής χαρακτηριστικών, αλλά εξαρτάται από το είδος του

αλγορίθμουν. Κάποιοι συνδυασμοί αλγορίθμων μπορεί να οδηγήσουν σε καλά αποτελέσματα σε ένα συγκεκριμένο μοντέλο μηχανικής μάθησης, αλλά σε ένα άλλο μοντέλο μπορεί να οδηγήσουν σε πολύ κακά αποτελέσματα. Αυτό εξαρτάται από την ευαισθησία που έχει το κάθε μοντέλο στα δεδομένα καθώς και τον τρόπο με τον οποίο τα επεξεργάζεται. Επίσης, παρατηρούμε ότι ενώ στο 10-fold-cross-validation οι περισσότεροι αλγόριθμοι πετυχαίνουν καλά αποτελέσματα, στο gold set πέφτουν κατά ένα σχετικά μεγάλο ποσοστό που συνήθως κυμαίνεται στο 5%, εκτός εξαιρέσεων.

Οπως παρατηρούμε από τα αποτελέσματα που προέκυψαν από τον διαγωνισμό, καθώς και από τα αποτελέσματα που προέκυψαν από την πτυχιακή, η ανίχνευση ειρωνείας είναι ένα δύσκολο πρόβλημα, που όμως βρίσκεται ακόμα στα αρχικά στάδια της εξέλιξης του. Από τα αποτελέσματα παρατηρούμε ότι το καλύτερο F1-score στο gold set το πέτυχε ο αλγόριθμος Voting Ensemble που συνδυάζει μοντέλα που πέτυχαν υψηλό αποτέλεσμα στο 10-fold-cross-validation. Το μοντέλο Voting Ensembles έχει αυξημένη δημοτικότητα λόγω των καλών αποτελεσμάτων που πετυχαίνει σε τέτοιου είδους προβλήματα. Το κύριο προτέρημα του μοντέλου Voting Ensembles είναι ότι ψηφίζει το τελικό αποτέλεσμα μέσα από αποτελέσματα διάφορων αλγορίθμων, οπότε μπορούμε να συνδυάσουμε πολλά μοντέλα SVM με διαφορετικές παραμέτρους για καλύτερα αποτελέσματα. Ο αλγόριθμος πέρα από τους διαφορετικούς αλγορίθμους που χρησιμοποιεί, χρησιμοποιεί και τέσσερεις εκδοχές του αλγορίθμου SVM οι οποίες έχουν παραμέτρους ειδικές έτσι ώστε το μοντέλο SVM τείνει να έχει τέσσερεις καταστάσεις που αφορούν το overfitting, το underfitting και άλλες δύο ενδιάμεσες καταστάσεις.

Επίσης, παρατηρούμε ότι πολλοί αλγόριθμοι είχαν παραπλήσιο F1-score κατά το 10-fold-cross-validation αλλά στο gold set η διαφορά στα αποτελέσματα που πέτυχαν ήταν μεγαλύτερη από την προσδοκούμενη. Παράλληλα, το μοντέλο που πέτυχε το καλύτερο F1-score στο 10-fold-cross-validation δεν πέτυχε και το μεγαλύτερο F1-score στο gold set. Αυτά μπορεί να συμβαίνουν για διάφορους λόγους, ένας εκ των οποίων μπορεί να είναι ότι κάποιο μοντέλο που πέτυχε χαμηλότερο F1-score μπορεί να πάσχει από overfitting και γενικότερα δεν μπορεί να γενικευτεί αποδοτικά για οποιαδήποτε δεδομένα.

Τα μοντέλα μηχανικής μάθησης που πέτυχαν τους περισσότερους συνδυασμούς αλγορίθμων σε πλήθος που είχαν τα υψηλότερα F1-score στο 10-fold-cross-validation είναι το SVM και το Logistic Regression, ενώ στο gold set είναι το μοντέλο K-Neighbors, το μοντέλο SVM και το μοντέλο Voting Ensembles. Το μοντέλο που πέτυχε το μεγαλύτερο F1-score στο 10-fold-cross-validation είναι το Gaussian Naive Bayes, ενώ το μοντέλο Logistic Regression και

το μοντέλο LSTM έχουν πολύ μικρή διαφορά. Ο αλγόριθμος που πέτυχε το μεγαλύτερο F1-score στο gold set είναι ο Voting Ensembles με Univariate Selection και Bigrams, και ακολουθούν ο K-Neighbors με doc2vec και το SVM με doc2vec. Έτσι, αν το μοντέλο Voting Ensembles με Univariate Selection και Bigrams ήταν αυτό που είχε επιλεχθεί στον διαγωνισμό θα μας τοποθετούσε στην τρίτη θέση με F1-score 66.03%, ενώ αν το μοντέλο είχε ολοκληρωθεί πριν την λήξη του διαγωνισμού θα είχε επιλεγεί ο αλγόριθμος Voting Ensembles με doc2vec που πετυχαίνει F1-score 63.10% και μας τοποθετεί στην έκτη θέση.

Από αυτές τις παρατηρήσεις συμπεραίνουμε ότι αυτά τα μοντέλα είναι πιο συνεπής στο συγκεκριμένο πρόβλημα με τα συγκεκριμένα δεδομένα, αφού όπως φαίνεται και από τα αποτελέσματα είχαν αντίστοιχα καλά αποτελέσματα και στο gold set. Αξίζει να σημειωθεί ότι οι αλγόριθμοι τύπου Naive Bayes καθώς και Logistic Regression δεν πέτυχαν κορυφαία αποτελέσματα στο gold set αν και δεν είχαν γενικά κακή απόδοση, όπως σωστά προβλέφθηκε, το οποίο φαίνεται από τα τελικά αποτελέσματα που πέτυχαν στο gold set. Επίσης, παρατηρούμε ότι στον ίδιο αλγόριθμο με διαφορετική είσοδο υπάρχουν μεγάλες διακυμάνσεις στο F1-score, το οποίο υποδεικνύει το πόσο σημαντικό ρόλο παίζει η επεξεργασία που υπόκεινται τα δεδομένα πριν χρησιμοποιηθούν από κάποιον classifier.

Για τους τρεις αλγορίθμους που πέτυχαν τα καλύτερα αποτελέσματα στο 10-fold-cross-validation, που αναφέρθηκαν παραπάνω, υπολογίστηκε η καμπύλη ROC-AUC καθώς και η καμπύλη μάθησης (learning curve), οι οποίες βρίσκονται στο Ε μέρος του Παραρτήματος. Αυτά τα τρία μοντέλα είναι το Gaussian Naive Bayes με word2vec, το K-Neighbors με PCA και TF-IDF και το SVM με doc2vec. Επίσης, αυτές οι καμπύλες υπολογίστηκαν και για τα μοντέλα που πέτυχαν κορυφαία αποτελέσματα στο gold set που αναφέρθηκαν παραπάνω.

Αυτά τα μοντέλα είναι το SVM με doc2vec, το Voting Ensembles με Univariate Selection και Bigrams, το Voting Ensembles με doc2vec καθώς και το SVM με Bigrams, το οποίο όπως αναφέρθηκε κάνει overfit. Ο υπολογισμός των καμπυλών ROC-AUC γίνεται στο 10-fold-cross-validation για να δούμε το πως αντιδρά το μοντέλο σε διαφορετικά δεδομένα, ενώ ο υπολογισμός της Learning Curve γίνεται μόνο στο train set στα αρχεία που έχουν τροποποιείται κατάλληλα για την εξέταση των μοντέλων στο gold set επειδή για τον υπολογισμό της χρειάζεται όλο το train set και όχι κομμάτια του. Η Learning Curve μπορεί να μας δώσει πληροφορίες για το αν ένα μοντέλο κάνει overfit ή underfit στα δεδομένα εκπαίδευσης.

Από την καμπύλη ROC-AUC του Gaussian Naive Bayes με word2vec, εικόνα E.1 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει μικρό μέσο όρο AUC, οπότε δεν αναμένονται

πολύ καλά αποτελέσματα. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι έχει διακύμανση άρα τα αποτελέσματα σε τυχαία δεδομένα πρόβλεψης μπορεί να έχουν κάποια διαφορά μεταξύ τους. Από την καμπύλη ROC-AUC του K-Neighbors με PCA και TF-IDF, εικόνα E.2 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει πολύ καλό μέσο όρο AUC σε σχέση με τα υπόλοιπα μοντέλα, οπότε αναμένονται αρκετά καλά αποτελέσματα σε σχέση με υπόλοιπα μοντέλα. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι έχει διακύμανση άρα τα αποτελέσματα σε τυχαία δεδομένα πρόβλεψης μπορεί να έχουν σημαντική διαφορά μεταξύ τους.

Από την καμπύλη ROC-AUC του SVM με doc2vec, εικόνα E.3 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει καλό μέσο όρο AUC, οπότε αναμένονται σχετικά καλά αποτελέσματα. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι υπάρχει μικρή διακύμανση, αλλά υπάρχουν δύο folds που έχουν σημαντική διαφορά, εκ των οποίων το ένα έχει πολύ καλό AUC ενώ το άλλο έχει αρκετά χαμηλό AUC σε σχέση με υπόλοιπα του μοντέλου.

Αρα συμπεραίνουμε ότι υπάρχουν δεδομένα στα οποία σπάνια μπορεί να έχουμε πολύ καλό ή πολύ κακό αποτέλεσμα, αλλά στη μέση περίπτωση θα έχουμε σχετικά καλά αποτελέσματα. Από την καμπύλη ROC-AUC του Voting Ensembles με Univariate Selection και Bigrams, εικόνα E.4 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει καλό μέσο όρο AUC, οπότε αναμένονται σχετικά καλά αποτελέσματα. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι υπάρχει σημαντική διακύμανση, οπότε αναμένεται να δίνει πολύ καλά ή πολύ κακά αποτελέσματα που τείνουν στη απόδοση του base μοντέλου.

Από την καμπύλη ROC-AUC του Voting Ensembles με doc2vec, εικόνα E.5 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει καλό μέσο όρο AUC, οπότε αναμένονται σχετικά καλά αποτελέσματα. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι υπάρχει μικρή διακύμανση, οπότε αναμένεται να δίνει σταθερά καλά αποτελέσματα. Επίσης, υπάρχει και ένα fold το οποίο έχει πολύ καλό AUC, που σημαίνει ότι σε σπάνιες περιπτώσει με κατάλληλα δεδομένα το μοντέλο θα παράγει πολύ καλά αποτελέσματα. Από την καμπύλη ROC-AUC του SVM με Bigrams, εικόνα E.6 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει καλό μέσο όρο AUC αλλά λίγο μικρότερο από αυτό που έχουν τα άλλα μοντέλα, οπότε αναμένονται σχετικά καλά αποτελέσματα που είναι ίσως λίγο μικρότερα από αυτά των υπόλοιπων μοντέλων. Στο 10-fold-cross-validation, από fold σε fold παρατηρούμε ότι υπάρχει πάρα πολύ μικρή διακύμανση, αφού κάθε fold έχει περίπου ίδιο AUC με μικρές αποκλίσεις. Έτσι, το μοντέλο θεωρείται πολύ σταθερό και αναμένεται να δίνει σταθερά καλά αποτελέσματα.

Από την Learning Curve του Gaussian Naive Bayes με word2vec, εικόνα E.7 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει σχετικά χαμηλά αποτελέσματα που δεν φαίνεται να επηρεάζονται ιδιαίτερα από το πλήθος των δεδομένων εκπαίδευσης. Στο δεξί μέρος του σχήματος, οι δύο καμπύλες βρίσκονται πολύ κοντά η μία στην άλλη, άρα το μοντέλο δεν πάσχει από overfit. Επίσης, στο δεξί μέρος παρατηρούμε ότι έχουμε αποτελέσματα καλύτερα του μετρίου άρα δεν πάσχει και από underfit. Από την Learning Curve του K-Neighbors με PCA και TF-IDF, εικόνα E.8 στο Παράρτημα, παρατηρούμε ότι το μοντέλο όταν έχει λίγα δεδομένα εκπαίδευσης δεν έχουμε καλή απόδοση, ενώ όσο τα δεδομένα αυξάνονται η απόδοση βελτιώνεται. Στο δεξί μέρος του σχήματος, οι δύο καμπύλες βρίσκονται πολύ κοντά η μία στην άλλη, άρα το μοντέλο δεν πάσχει από overfit. Επίσης, στο δεξί μέρος παρατηρούμε ότι τα αποτελέσματα κυμαίνονται σε καλό επίπεδο άρα δεν πάσχει και από underfit.

Από την Learning Curve του SVM με doc2vec, εικόνα E.9 στο Παράρτημα, παρατηρούμε ότι το μοντέλο έχει καλά αποτελέσματα που δεν φαίνεται να επηρεάζονται από το μέγεθος των δεδομένων εκπαίδευσης. Στο δεξί μέρος του σχήματος, οι δύο καμπύλες βρίσκονται πολύ κοντά η μία στην άλλη ενώ κυμαίνονται και σε καλό αποτέλεσμα, άρα το μοντέλο δεν πάσχει από overfit και underfit αντίστοιχα.

Από την Learning Curve του Voting Ensembles με Univariate Selection και Bigrams, εικόνα E.10 στο Παράρτημα, παρατηρούμε ότι όταν δίνουμε στο μοντέλο μικρό μέγεθος των δεδομένων εκπαίδευσης οι καμπύλες έχουν σχετική μικρή απόσταση, ενώ όσο αυξάνονται τα δεδομένα εκπαίδευσης που του δίνουμε η διαφορά περιορίζεται σημαντικά. Στο δεξί μέρος του σχήματος, οι δύο καμπύλες βρίσκονται πολύ κοντά η μία στην άλλη ενώ κυμαίνονται και σε καλό αποτέλεσμα, άρα το μοντέλο δεν πάσχει από overfit και underfit αντίστοιχα. Από την Learning Curve του Voting Ensembles με doc2vec, εικόνα E.11 στο Παράρτημα, παρατηρούμε ότι όταν δίνουμε στο μοντέλο μικρό μέγεθος των δεδομένων εκπαίδευσης οι καμπύλες έχουν αισθητή απόσταση, ενώ όσο αυξάνονται τα δεδομένα εκπαίδευσης που του δίνουμε η διαφορά περιορίζεται σημαντικά. Στο δεξί μέρος του σχήματος, οι δύο καμπύλες βρίσκονται πολύ κοντά η μία στην άλλη ενώ κυμαίνονται και σε καλό αποτέλεσμα, άρα το μοντέλο δεν πάσχει από overfit και underfit αντίστοιχα.

Από την Learning Curve του SVM με Bigrams, εικόνα E.12 στο Παράρτημα, παρατηρούμε ότι οι δύο καμπύλες έχουν πολύ μεγάλη απόσταση και άρα το μοντέλο πάσχει από overfitting. Έτσι, περιμένουμε το μοντέλο να μην έχει καλά αποτέλεσμα στο gold set, ο οποίος είναι και ο λόγος που δεν επιλέχθηκε σαν τελικό μοντέλο που θα πάρει μέρος στον διαγωνισμό.

Ωστόσο, στα πλαίσια της πτυχιακής δοκιμάστηκε για να δούμε τι αποτελέσματα θα παράγει και όπως φαίνεται από το αποτέλεσμα στον πίνακα Γ.9 του Παραρτήματος πέτυχε F1-score 61.41%, το οποίο ανήκει στην ομάδα μοντέλων με τα υψηλότερα F1-score. Η παρατήρηση από το Learning Curve έρχεται σε αντίθεση με την αντίστοιχη παρατήρηση που πηγάζει από την καμπύλη ROC-AUC, δηλαδή ότι το μοντέλο φαίνεται πολύ σταθερό στις προβλέψεις του σε διαφορετικά δεδομένα εκπαίδευσης. Από τις δύο παρατηρήσεις που πηγάζουν από την καμπύλη ROC-AUC και την Learning Curve συμπεραίνουμε ότι τα δεδομένα πρόβλεψης είναι όμοια σε κάποιο βαθμό με τα δεδομένα εκπαίδευσης.

Αξίζει να σημειωθεί ότι οι πιο γρήγοροι αλγόριθμοι όσο αναφορά την ταχύτητα εκτέλεσης τους ήταν o Gaussian Naive Bayes, o Bernoulli Naive Bayes, o multinomial Naive Bayes, o Logistic Regression και o K-Neighbors. Αν και αυτοί οι αλγόριθμοι, εκτός του K-Neighbors, δεν ήταν στην κορυφή των καλύτερων αποτελεσμάτων, πέτυχαν ποσοστά αρκετά κοντά στο καλύτερο αποτέλεσμα με απόκλιση περίπου 5%. Έτσι, αν δεν μας ενδιέφερε η ακρίβεια του αποτελέσματος αλλά θέλαμε να κάνουμε κάποια γρήγορη πρόβλεψη, όπως σε κάποια online εφαρμογή που το αποτέλεσμα πρέπει να είναι διαθέσιμο μέσα σε δευτερόλεπτα, τότε αυτοί οι αλγόριθμοι θα ήταν ιδανικοί για αυτή τη δουλειά, ιδίως ο αλγόριθμος K-Neighbors που πέτυχε το δεύτερο καλύτερο F1-score στο gold set.

Επίσης, από τα αποτελέσματα παρατηρούμε ότι ο Gaussian Naive Bayes παράγει καλά αποτελέσματα όταν χρησιμοποιούνται μόνο οι αλγόριθμοι κωδικοποίησης δεδομένων χωρίς την διαδικασία της επιλογής χαρακτηριστικών. Ακόμα, παρατηρούμε ότι ενώ όλοι οι απλοί αλγόριθμοι κωδικοποίησης δεδομένων δουλεύουν καλά, τα word embeddings παράγουν τα καλύτερα αποτελέσματα. Τα bigrams με one-hot-encoding και εξαγωγή χαρακτηριστικών παράγουν τα χειρότερα αποτελέσματα και αυτό πιθανόν οφείλεται στο ότι ο πίνακας είναι ιδιαίτερα μεγάλος, επειδή πολλά από τα bigrams που σχηματίζονται είναι μοναδικά ή δεν εμφανίζονται σε πολλά διαφορετικά tweets, το οποίο καθιστά τον πίνακα αραιό με αποτέλεσμα να μην προσφέρει ιδιαίτερα χρήσιμη πληροφορία που να μπορεί να εκμεταλλευτεί ο συγκεκριμένος αλγόριθμος. Κάτι αντίστοιχο συμβαίνει και στον multinomial Naive Bayes με τα bigrams και την επιλογή χαρακτηριστικών.

Γενικότερα, τα αποτελέσματα που ήταν κοντά ή πιο κάτω από το baseline μοντέλο μπορεί να οφείλονται σε διάφορους παράγοντες. Οι πιο συνηθισμένοι είναι το overfitting και το underfitting. Στην περίπτωση του overfitting ο αλγόριθμος μαθαίνει πολύ καλά τα δεδομένα και τον θόρυβο που υπάρχει σε αυτά με αποτέλεσμα να μειώνεται η απόδοση του δραστικά. [49]

To overfitting εμφανίζεται συνήθως σε μοντέλα που δεν έχουν παραμέτρους και είναι μη γραμμικά. Μια μέθοδος για την αποφυγή του overfitting είναι η προσθήκη επιπλέον δεδομένων εκπαίδευσης. Στην περίπτωση του underfitting ο αλγόριθμος δεν μπορεί να μάθει τα δεδομένα εκπαίδευσης και άρα δεν μπορεί να γενικευτεί σε καινούργια δεδομένα. [49]

Το μοντέλο SVM δουλεύει πολύ καλά με bigrams, είτε χωρίς αλγόριθμο επιλογής χαρακτηριστικών είτε με αλγόριθμο επιλογής χαρακτηριστικών. Το μοντέλο Gaussian Naive Bayes δουλεύει καλύτερα με word embeddings, αν και τα αποτελέσματα του στο gold set δεν κυμαίνονται στο επίπεδο των κορυφαίων μοντέλων. Το μοντέλο Logistic Regression λειτουργεί καλύτερα με τον αλγόριθμο TF-IDF καθώς και με συνδυασμό TF-IDF και αλγορίθμων επιλογής χαρακτηριστικών, αλλά γενικά η αποδοτικότητα του δεν κυμαίνεται στα κορυφαία επίπεδα.

Μελλοντικοί στόχοι

Με την ολοκλήρωση της πτυχιακής αξίζει να αναφερθούν διάφοροι τομείς που θα μπορούσαν να αποτελέσουν αντικείμενο έρευνας στο μέλλον. Τον τελευταίο καιρό, από διάφορες έρευνες σε παρόμοια προβλήματα καθώς και από τα αποτελέσματα παρατηρείται ότι αλγόριθμοι όπως το LSTM καθώς και το Conv1D πετυχαίνουν αρκετά υψηλή βαθμολογία. Έτσι, ένα μελλοντικός τομέας έρευνας είναι η περαιτέρω ανάπτυξη αυτών των μοντέλων καθώς και ο συνδυασμός αυτών των δύο για καλύτερο αποτέλεσμα. Επιπλέον, πρέπει να ερευνηθούν διάφοροι τρόποι χρήσης των word embeddings, word2vec και GloVe, στα νευρωνικά δίκτυα, όπως είναι η χρήση της συνάρτησης pad_sequences του Keras η οποία προσθέτει μηδενικά στην αρχή των προτάσεων ώστε να έχουν όλες το ίδιο μήκος. Η χρήση αυτής της συνάρτησης μας επιτρέπει να εκμεταλλευτούμε τις ιδιότητες κάθε λέξης ξεχωριστά καθώς και την σειρά (sequence) με την οποία υπάρχουν στην πρόταση, κάτι που απουσιάζει από την μέθοδο που χρησιμοποιείται και κατά της οποία υπολογίζεται ο μέσος όρος από τα διανύσματα των λέξεων μιας πρότασης.

Επίσης, ένας ακόμα ενδιαφέρον τομέας έρευνας είναι η ανάπτυξη ενός μοντέλου Voting Ensembles καθώς και η έρευνα της αποτελεσματικότητας του. Το μοντέλο Voting Ensembles θα δέχεται τις προβλέψεις από διαφορετικά μοντέλα που εκπαιδεύτηκαν στα ίδια δεδομένα αλλά με διαφορετικούς αλγορίθμους κωδικοποίησης και επιλογής χαρακτηριστικών, κάτι που δεν γίνεται με τον αλγόριθμο της βιβλιοθήκης sklearn. Με αυτόν τον τρόπο βελτιώνεται το αποτέλεσμα αφού στην ψηφοφορία πλειοψηφίας χρησιμοποιούνται οι καλύτερες προβλέψεις

των μοντέλων, το οποίο επιτυγχάνεται επιλέγοντας τους πιο αποδοτικούς συνδυασμούς αλγορίθμων κωδικοποίησης δεδομένων και επιλογής χαρακτηριστικών για το κάθε μοντέλο. Τα μοντέλα που θα δέχεται ο αλγόριθμος δεν θα επηρεάζονται από την βιβλιοθήκη που τα υλοποιεί, κάτι που το μοντέλο Voting Ensembles της βιβλιοθήκης sklearn δεν το επιτρέπει καθώς περιορίζει τα μοντέλα που μπορούν να χρησιμοποιηθούν μόνο σε αυτά που υλοποιούνται από την ίδια την βιβλιοθήκη sklearn. Στη συνέχεια το μοντέλο θα εξάγει την τελική πρόβλεψη μέσο ψηφοφορίας η οποία θα χρησιμοποιεί την μέθοδο της πλειοψηφίας των προβλέψεων από όλα τα μοντέλα.

Τέλος, αξίζει να ερευνηθούν τεχνικές και αλγόριθμοι που έχουν καλή απόδοση και χρησιμοποιούνται την παρούσα χρονική περίοδο σε τέτοιου είδους προβλήματα. Επίσης, η έρευνα νέων τεχνικών και μεθόδων που δεν χρησιμοποιούνται ευρέως μπορεί να αποδειχθεί καρποφόρα, είτε από την άποψη ότι οι τεχνικές και οι μέθοδοι μπορεί να είναι αποδοτικοί είτε από τη άποψη ότι μπορεί να γίνει συνεισφορά στην ανάπτυξη και στην βελτίωση τους. Παράλληλα, η έρευνα νέων μεθόδων, αλλά και παλαιών, μπορεί να οδηγήσει στην ανάπτυξη καινούργιων αποδοτικών μεθόδων που δεν έχουν ακόμα υλοποιηθεί ούτε στην πράξη αλλά ούτε και στην θεωρία. Επιπλέον, η παρακολούθηση των εξελίξεων στον τομέα του προβλήματος, καθώς και ευρύτερα στον τομέα της ανάλυσης συναισθημάτων (sentiment analysis), μπορεί να προσφέρει πολύτιμη πληροφορία για την κατάσταση του προβλήματος καθώς και για νέες μεθόδους πιο αποδοτικές από τις ήδη υπάρχουσες.

Πέρα από το πρόβλημα της ανίχνευσης ειρωνείας, οι αλγόριθμοι και γενικότερα τα μοντέλα που χρησιμοποιούνται για την επίλυση αυτού του προβλήματος χρησιμοποιούνται για πολλά άλλα διαφορετικά προβλήματα. Ένα τέτοιο πρόβλημα, που είναι και πιο δημοφιλής την παρούσα περίοδο, είναι η ανάλυση συναισθημάτων (sentiment analysis). Ουσιαστικά η ανίχνευση ειρωνείας αποτελεί υποτομέας της ανάλυσης συναισθημάτων, με τα δύο αυτά προβλήματα να έχουν πάρα πολλές ομοιότητες όσο αφορά τις προσεγγίσεις των μοντέλων.

Συνεπώς, με σχετικά μικρές αλλαγές στα μοντέλα, μπορεί να γίνει εύκολα η μετατροπή των μοντέλων που χρησιμοποιήθηκαν για την επίλυση του προβλήματος της ανίχνευσης ειρωνείας σε μοντέλα που έχουν σκοπό την επίλυση του προβλήματος της ανάλυσης συναισθήματος. Έτσι, ένας γενικός μελλοντικός στόχος, πέρα από την βελτίωση των μοντέλων και του αποτελέσματος του συγκεκριμένου προβλήματος, είναι η μετατροπή των μοντέλων ώστε να είναι ικανά να επιλύσουν όσο το δυνατόν πιο αποδοτικά το πρόβλημα της ανάλυσης συναισθημάτων.

Παράρτημα

ΜΕΡΟΣ Α

Πλήθος Εμφανίσεων	Λέξεις	Πλήθος Εμφανίσεων	Λέξεις
47	always	41	another
76	back	39	bad
41	best	55	better
34	big	95	ca
95	christmas	35	come
166	day	36	done
57	even	37	every
46	feel	37	first
32	free	77	fun
44	game	45	getting
88	go	86	going
104	good	87	got
116	great	47	happy
37	hours	35	keep
96	know	47	last
33	let	63	life
68	lol	50	look
214	love	56	make
37	man	53	morning
77	much	46	na
69	need	53	never
80	new	36	next
45	nice	51	night
38	nothing	380	nt
82	oh	127	one
117	people	89	really
73	right	53	rt
37	say	32	school
94	see	42	sleep
46	someone	51	start
70	still	34	take
66	thanks	52	thing
36	things	78	think

98	time	99	today
35	tomorrow	39	tonight
45	twitter	33	two
45	u	46	us
50	via	54	wait
61	want	66	way
38	week	75	well
80	work	37	world
35	wow	49	yeah
74	year	252	not

Πίνακας A.1 Πλήθος εμφανίσεων των λέξεων

ΜΕΡΟΣ Β

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6777	0.5889	0.3656	0.5875	0.4523
Univariate Selection	One-Hot	0.5997	0.5693	0.5936	0.5624	0.5431
Univariate Selection	Bigrams	0.7274	0.5133	0.3588	0.5290	0.3098
PCA	TF-IDF	0.5965	0.5696	0.4233	0.5690	0.4948
PCA	One-Hot	0.6062	0.5766	0.4316	0.5764	0.5035
PCA	Bigrams	0.6157	0.5542	0.4023	0.5542	0.4210
SVD	TF-IDF	0.6151	0.5873	0.4614	0.5869	0.5269
SVD	One-Hot	0.6048	0.5759	0.4281	0.5755	0.5006
SVD	Bigrams	0.6630	0.5636	0.2886	0.5597	0.3493
Feature Importance	TF-IDF	0.6830	0.5912	0.3339	0.5907	0.4459
Feature Importance	One-Hot	0.7161	0.5779	0.3177	0.5769	0.4010
Feature Importance	Bigrams	0.7351	0.5216	0.1725	0.5281	0.1977
-	TF-IDF	0.5329	0.5471	0.7413	0.5478	0.6196
-	One-Hot	0.5324	0.5495	0.7969	0.5508	0.6377
-	Bigrams	0.5489	0.5511	0.7691	0.5435	0.6178
-	word2vec	0.5323	0.5563	0.9093	0.5574	0.6710
-	doc2vec	0.5808	0.6079	0.7751	0.6091	0.6630
-	GloVe	0.5580	0.5782	0.7360	0.5787	0.6343

Πίνακας B.1 Αποτελέσματα αξιολόγησης Gaussian Naive Bayes

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6713	0.6233	0.4997	0.6259	0.5654
Univariate Selection	One-Hot	0.6514	0.6103	0.4857	0.6119	0.5521
Univariate Selection	Bigrams	0.7472	0.5657	0.2353	0.5691	0.3304
Feature Importance	TF-IDF	0.6577	0.6304	0.5483	0.6315	0.5957
Feature Importance	One-Hot	0.6537	0.6105	0.4689	0.6109	0.5447
Feature Importance	Bigrams	0.6850	0.5860	0.3236	0.5874	0.4343
-	TF-IDF	0.6363	0.6327	0.6187	0.6337	0.6259
-	One-Hot	0.6206	0.6304	0.6670	0.6310	0.6421
-	Bigrams	0.6402	0.5946	0.4317	0.5956	0.5135

Πίνακας B.2 Αποτελέσματα αξιολόγησης Multinomial Naive Bayes

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6265	0.6275	0.6286	0.6283	0.6264
Univariate Selection	One-Hot	0.6219	0.6220	0.6184	0.6224	0.6193
Univariate Selection	Bigrams	0.6345	0.6236	0.5821	0.6244	0.6061
PCA	TF-IDF	0.6131	0.6147	0.6203	0.6158	0.6155
PCA	One-Hot	0.5862	0.5993	0.6714	0.6002	0.6251
PCA	Bigrams	0.5951	0.5905	0.5650	0.5912	0.5784
SVD	TF-IDF	0.6052	0.6103	0.6347	0.6110	0.6187
SVD	One-Hot	0.5846	0.5915	0.6259	0.5921	0.6039
SVD	Bigrams	0.5853	0.5957	0.6510	0.5962	0.6156
Feature Importance	TF-IDF	0.6469	0.6345	0.5895	0.6350	0.6160
Feature Importance	One-Hot	0.6503	0.6382	0.5972	0.6392	0.6212
Feature Importance	Bigrams	0.6302	0.6199	0.5814	0.6212	0.6031
-	TF-IDF	0.6221	0.6301	0.6597	0.6310	0.6393
-	One-Hot	0.6221	0.6301	0.6597	0.6310	0.6393
-	Bigrams	0.6410	0.6218	0.5622	0.6243	0.5949
-	word2vec	0.5317	0.5286	0.4417	0.5280	0.4823
-	doc2vec	0.5943	0.5980	0.6162	0.5993	0.6034
-	GloVe	0.6015	0.5959	0.5615	0.5960	0.5802

Πίνακας B.3 Αποτελέσματα αξιολόγησης Bernoulli Naive Bayes

*: Οι συνδυασμοί αλγορίθμων με αστερίσκο προέκυψαν χωρίς τη χρήση Normalizer στα δεδομένα, ενώ οι υπόλοιποι προέκυψαν με τη χρήση Normalizer

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6705	0.6236	0.4933	0.6247	0.5648
Univariate Selection	One-Hot	0.6227	0.6160	0.5975	0.6179	0.6047
Univariate Selection	Bigrams	0.6131	0.6254	0.6773	0.6262	0.6428
PCA	TF-IDF	0.6444	0.6442	0.6460	0.6456	0.6436
PCA	One-Hot	0.6313	0.6098	0.5211	0.6098	0.5700
PCA	Bigrams	0.6213	0.6277	0.6541	0.6183	0.6361
SVD	TF-IDF	0.6707	0.6317	0.5216	0.6329	0.5848
SVD	One-Hot	0.6681	0.6095	0.4521	0.6103	0.5343
SVD	Bigrams	0.6297	0.6006	0.5012	0.6026	0.5539
Feature Importance	TF-IDF	0.6457	0.6283	0.5812	0.6299	0.6080
Feature Importance	One-Hot	0.6558	0.6249	0.5278	0.6249	0.5835
Feature Importance	Bigrams	0.6195	0.6233	0.6431	0.6245	0.6294
-	TF-IDF	0.7008	0.6304	0.4576	0.6312	0.5515
-	One-Hot	0.6361	0.6139	0.5272	0.6138	0.5762
-	Bigrams	0.7049	0.5644	0.2228	0.5643	0.3350
-	*word2vec	0.5328	0.5386	0.6223	0.5402	*0.5725
-	*doc2vec	0.6076	0.6173	0.6657	0.6183	*0.6338
-	GloVe	0.6098	0.5978	0.5392	0.5981	0.5714

Πίνακας B.4 Αποτελέσματα αξιολόγησης K-Neighbors

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.5966	0.6210	0.7488	0.6224	0.6629
Univariate Selection	One-Hot	0.6303	0.6236	0.6016	0.6248	0.6136
Univariate Selection	Bigrams	0.5885	0.6116	0.7413	0.6129	0.6552
PCA	TF-IDF	0.6017	0.6244	0.7372	0.6258	0.6614
PCA	One-Hot	0.6337	0.6270	0.6013	0.6278	0.6159
PCA	Bigrams	0.5954	0.6163	0.7286	0.6179	0.6539
SVD	TF-IDF	0.6044	0.6267	0.7357	0.6281	0.6623
SVD	One-Hot	0.6257	0.6236	0.6150	0.6244	0.6191
SVD	Bigrams	0.5932	0.6155	0.7375	0.6171	0.6562
Feature Importance	TF-IDF	0.6037	0.6264	0.7387	0.6279	0.6632
Feature Importance	One-Hot	0.6386	0.6330	0.6131	0.6342	0.6240
Feature Importance	Bigrams	0.5932	0.6100	0.7089	0.6120	0.6435
-	TF-IDF	0.6119	0.6345	0.7385	0.6361	0.6679
-	One-Hot	0.6544	0.6486	0.6272	0.6491	0.6396
-	Bigrams	0.6050	0.6249	0.7239	0.6266	0.6576
-	word2vec	0.6193	0.6210	0.6300	0.6221	0.6230
-	doc2vec	0.5857	0.6100	0.7533	0.6115	0.6579
-	GloVe	0.6192	0.6194	0.6157	0.6198	0.6169

Πίνακας B.5 Αποτελέσματα αξιολόγησης Logistic Regression

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6392	0.6288	0.6140	0.6781	0.6195
Univariate Selection	One-Hot	0.6267	0.6207	0.5904	0.6682	0.6074
Univariate Selection	Bigrams	0.6233	0.6116	0.6083	0.6652	0.6057
PCA	TF-IDF	0.6547	0.6452	0.6150	0.7092	0.6332
PCA	One-Hot	0.6285	0.6254	0.6111	0.6823	0.6185
PCA	Bigrams	0.6097	0.6202	0.6670	0.6749	0.6360
SVD	TF-IDF	0.6417	0.6377	0.6234	0.7072	0.6311
SVD	One-Hot	0.6322	0.6278	0.6101	0.6855	0.6199
SVD	Bigrams	0.6147	0.6270	0.6799	0.6732	0.6447
Feature Importance	TF-IDF	0.6472	0.6377	0.6063	0.6893	0.6241
Feature Importance	One-Hot	0.6416	0.6395	0.6282	0.6896	0.6339
Feature Importance	Bigrams	0.6236	0.6296	0.6711	0.6795	0.6422
-	TF-IDF	0.6197	0.6178	0.6101	0.6613	0.6123
-	One-Hot	0.6204	0.6160	0.5956	0.6612	0.6061
-	Bigrams	0.6146	0.6105	0.6139	0.6595	0.6091
-	word2vec	0.6173	0.6210	0.6364	0.6796	0.6245
-	doc2vec	0.6143	0.6147	0.6241	0.6572	0.6162
-	GloVe	0.6132	0.6144	0.6206	0.6712	0.6156

Πίνακας B.6 Αποτελέσματα αξιολόγησης MLP Neural Network

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	0.6332	0.6288	0.6185	0.6817	0.6225
-	doc2vec	0.5839	0.6113	0.7773	0.6667	0.6651
-	GloVe	0.6086	0.6134	0.6302	0.6495	0.6184

Πίνακας B.7 Αποτελέσματα αξιολόγησης Long Short-Term Memory Network (LSTM)

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	0.6220	0.6184	0.5987	0.6688	0.6085
-	doc2vec	0.6114	0.6069	0.6229	0.6568	0.6077
-	GloVe	0.5983	0.6027	0.6155	0.6428	0.6062

Πίνακας B.8 Αποτελέσματα αξιολόγησης 1-D Convolutional Neural Network

*: Οι συνδυασμοί αλγορίθμων με αστερίσκο προέκυψαν με χρήση StandarScaler στα δεδομένα, ενώ οι υπόλοιποι χωρίς StandarScaler

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6363	0.6259	0.5959	0.6270	0.6113
*Univariate Selection	*One-Hot	0.6194	0.6171	0.6183	0.6190	* 0.6152
Univariate Selection	Bigrams	0.5885	0.6116	0.7413	0.6129	0.6552
PCA	TF-IDF	0.6512	0.6525	0.6572	0.6535	0.6532
*PCA	*One-Hot	0.5406	0.5553	0.7296	0.5565	* 0.6201
PCA	Bigrams	0.5961	0.6168	0.7224	0.6179	0.6524
SVD	TF-IDF	0.6454	0.6455	0.6489	0.6466	0.6457
*SVD	*One-Hot	0.5383	0.5542	0.7486	0.5552	* 0.6257
SVD	Bigrams	0.5954	0.6165	0.7262	0.6178	0.6534
Feature Importance	TF-IDF	0.6365	0.6371	0.6408	0.6382	0.6373
Feature Importance	One-Hot	0.6226	0.6212	0.6118	0.6218	0.6161
Feature Importance	Bigrams	0.6003	0.6194	0.7200	0.6212	0.6531
-	TF-IDF	0.6218	0.6220	0.6213	0.6229	0.6202
-	One-Hot	0.6438	0.6358	0.6035	0.6362	0.6223
-	Bigrams	0.5835	0.6098	0.7624	0.6111	0.6603
-	*word2vec	0.6057	0.5993	0.5668	0.5996	* 0.5846
-	doc2vec	0.5907	0.6134	0.7406	0.6150	0.6560
-	GloVe	0.5590	0.5046	0.5278	0.5197	0.4287

Πίνακας B.9 Αποτελέσματα αξιολόγησης SVM

		Precision	Accuracy	Recall	ROC	F1-score
Univariate Selection	TF-IDF	0.6501	0.6301	0.5695	0.6308	0.6026
Univariate Selection	One-Hot	0.6667	0.6202	0.4913	0.6218	0.5590
Univariate Selection	Bigrams	0.6228	0.6283	0.6471	0.6289	0.6340
PCA	TF-IDF	0.6495	0.6450	0.6332	0.6463	0.6397
PCA	One-Hot	0.6516	0.6241	0.5327	0.6243	0.5852
PCA	Bigrams	0.6127	0.6233	0.6696	0.6244	0.6389
SVD	TF-IDF	0.6522	0.6476	0.6356	0.6489	0.6422
SVD	One-Hot	0.6482	0.6178	0.5160	0.6182	0.5733
SVD	Bigrams	0.6117	0.6233	0.6754	0.6246	0.6407
Feature Importance	TF-IDF	0.6464	0.6439	0.6405	0.6456	0.6414
Feature Importance	One-Hot	0.6453	0.6267	0.5653	0.6278	0.6006
Feature Importance	Bigrams	0.6152	0.6262	0.6787	0.6278	0.6436
-	TF-IDF	0.6427	0.6343	0.6021	0.6348	0.6207
-	One-Hot	0.6574	0.6343	0.5587	0.6347	0.6027
-	Bigrams	0.6578	0.6197	0.4964	0.6204	0.5644
-	word2vec	0.5702	0.5670	0.5408	0.5678	0.5541
-	doc2vec	0.6090	0.6228	0.6882	0.6242	0.6449
-	GloVe	0.6068	0.5500	0.4043	0.5559	0.4203

Πίνακας B.10 Αποτελέσματα αξιολόγησης Voting Ensembles

ΜΕΡΟΣ Γ

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	43.99	51.65	80.06	56.52	56.78
-	word2vec	46.09	54.71	83.60	59.66	59.42
-	doc2vec	48.80	59.05	65.59	60.17	55.96
-	GloVe	47.95	57.78	75.24	60.77	58.57

Πίνακας Γ.1 Καλύτερα αποτελέσματα Gaussian Naive Bayes στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	52.68	63.01	66.23	63.56	58.68
-	TF-IDF	53.73	63.64	60.12	63.04	56.75

Πίνακας Γ.2 Καλύτερα αποτελέσματα Multinomial Naive Bayes στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
-	One-Hot	53.31	63.64	67.20	64.25	59.45
-	TF-IDF	53.31	63.64	67.20	64.25	59.45
PCA	One-Hot	50.00	60.33	67.20	61.50	57.33
Univariate Selection	TF-IDF	55.85	65.30	59.80	64.36	57.76

Πίνακας Γ.3 Καλύτερα αποτελέσματα Bernoulli Naive Bayes στο gold set

***doc2vec**: δεν χρησιμοποιήθηκε Normalizer

Για τους υπόλοιπους συνδυασμούς χρησιμοποιήθηκε Normalizer

		Precision	Accuracy	Recall	ROC	F1-score
-	*doc2vec	55.39	66.07	74.27	67.47	63.46
PCA	Bigrams	52.19	62.88	76.52	65.21	62.05
PCA	TF-IDF	53.97	64.03	63.34	63.91	58.28
Univariate Selection	TF-IDF	61.14	68.75	5819	66.94	59.63

Πίνακας Γ.4 Καλύτερα αποτελέσματα K-Neighbors στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
Feature Importance	TF-IDF	50.51	60.84	63.66	61.32	56.33
Univariate Selection	TF-IDF	50.25	60.58	62.37	60.89	55.66
PCA	TF-IDF	51.75	62.11	66.23	62.82	58.11
SVD	TF-IDF	50.90	61.22	63.34	61.58	56.44
-	TF-IDF	52.74	63.01	64.95	63.34	58.21

Πίνακας Γ.5 Καλύτερα αποτελέσματα Logistic Regression στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
PCA	TF-IDF	56.45	64.41	45.01	66.98	50.08
PCA	Bigrams	52.00	62.11	58.52	66.53	55.06
Feature Importance	Bigrams	53.61	63.64	62.05	67.58	57.52
Feature Importance	One-Hot	56.73	65.17	51.44	68.92	53.96
TruncatedSVD	Bigrams	53.27	63.39	62.70	66.98	57.60
TruncatedSVD	TF-IDF	53.64	62.88	47.26	65.94	50.25

Πίνακας Γ.6 Καλύτερα αποτελέσματα MLP Neural Network στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	53.20	63.39	63.98	70.98	58.10
-	doc2vec	48.92	58.92	80.38	65.24	60.82
-	Glove	50.95	61.35	68.48	65.60	58.43

Πίνακας Γ.7 Καλύτερα αποτελέσματα Long Short-Term Memory Network στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
-	word2vec	54.82	64.66	62.05	69.14	58.22
-	doc2vec	56.65	66.32	64.30	67.53	60.24
-	Glove	50.79	61.09	61.41	65.60	55.60

Πίνακας Γ.8 Καλύτερα αποτελέσματα 1-D Convolutional Neural Network στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
-	Bigrams	51.29	61.86	76.52	64.37	61.41
-	doc2vec	51.05	61.60	78.13	64.43	61.75
PCA	TF-IDF	54.51	64.03	56.27	62.70	55.37
PCA	Bigrams	51.91	62.50	73.95	64.46	61.00
Feature Importance	Bigrams	46.58	56.88	59.16	57.27	52.12
Univariate Selection	Bigrams	49.69	59.94	78.77	63.17	60.94
TruncatedSVD	Bigrams	51.58	62.11	73.31	64.03	60.55

Πίνακας Γ.9 Καλύτερα αποτελέσματα SVM στο gold set

		Precision	Accuracy	Recall	ROC	F1-score
PCA	Bigrams	52.75	63.26	70.73	64.54	60.43
PCA	TF-IDF	57.51	66.19	56.59	64.55	57.05
-	doc2vec	55.02	65.68	73.95	67.10	63.10
Univariate Selection	Bigrams	57.34	68.23	77.81	69.87	66.03
TruncatedSVD	Bigrams	51.96	62.50	72.34	64.18	60.48
TruncatedSVD	TF-IDF	55.24	64.66	57.55	63.45	56.37
Feature Importance	Bigrams	55.93	66.07	68.16	66.43	61.44
Feature Importance	TF-IDF	54.88	64.28	55.94	62.85	55.41

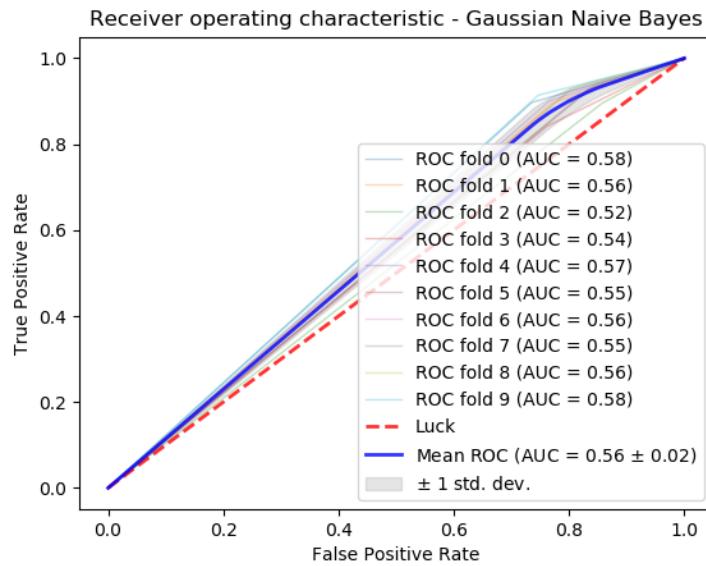
Πίνακας Γ.10 Καλύτερα αποτελέσματα Voting Ensembles στο gold set

ΜΕΡΟΣ Α

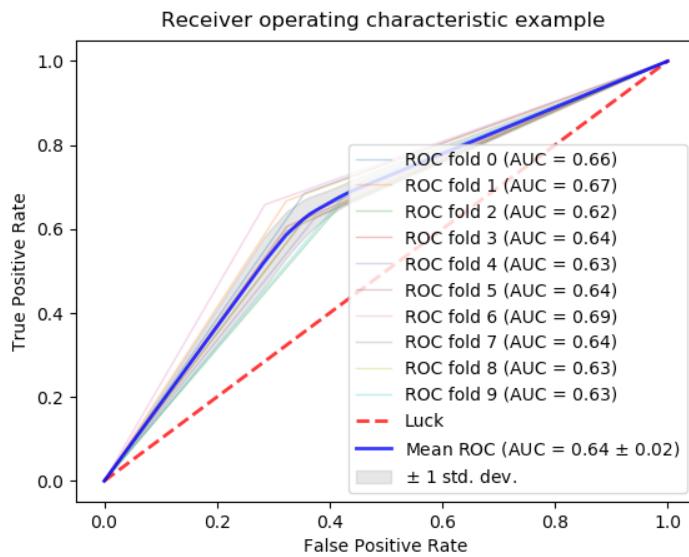
Results									
#	User	Entries	Date of Last Entry	Team Name	Accuracy ▲	Precision ▲	Recall ▲	F1-score ▲	
1	ChuhanWu	1	01/19/18	THU_NGN	0.7347 (1)	0.6304 (4)	0.8006 (4)	0.7054 (1)	
2	cbaziotis	2	01/21/18	NTUA-SLP	0.7321 (2)	0.6535 (2)	0.6913 (13)	0.6719 (2)	
3	omidrohanian	1	01/21/18	WLV	0.6429 (15)	0.5317 (20)	0.8360 (2)	0.6500 (3)	
4	rangwani_harsh	4	01/21/18		0.6607 (10)	0.5506 (13)	0.7878 (7)	0.6481 (4)	
5	thanhuu	2	01/21/18	NIHARIO, NCL	0.7015 (3)	0.6091 (5)	0.6913 (13)	0.6476 (5)	
6	Shuangqian	2	01/19/18	DLUTNLP-1	0.6276 (19)	0.5199 (23)	0.7974 (5)	0.6294 (6)	
7	jogonba2	3	01/17/18	ELiRF-UPV	0.6110 (23)	0.5059 (27)	0.8328 (3)	0.6294 (7)	
8	liangxh16	2	01/20/18		0.6594 (11)	0.5550 (11)	0.7138 (10)	0.6245 (8)	
9	CJ	2	01/18/18	CJ	0.6671 (8)	0.5654 (9)	0.6945 (12)	0.6234 (9)	
10	dadangewp	4	01/19/18	#NonDicevoSulSerio	0.6786 (7)	0.5831 (8)	0.6656 (15)	0.6216 (10)	
11	tigi	1	01/18/18	UWB	0.6875 (4)	0.5988 (7)	0.6431 (19)	0.6202 (11)	
12	dirazuherfa	2	01/21/18	INAOE-UPV	0.6505 (12)	0.5455 (15)	0.7138 (10)	0.6184 (12)	
13	zswvivi	3	01/21/18	RM@IT	0.6492 (13)	0.5441 (16)	0.7138 (10)	0.6175 (13)	
14	qshuang	1	01/21/18		0.6008 (25)	0.4980 (29)	0.7942 (6)	0.6121 (14)	
15	biggoka	1	01/21/18		0.5651 (31)	0.4731 (33)	0.8489 (1)	0.6076 (15)	
16	vpatti	5	01/21/18	emotIDM	0.5982 (26)	0.4959 (30)	0.7814 (8)	0.6067 (16)	
17	nishnik	3	01/21/18	binarizer	0.6659 (9)	0.5528 (12)	0.6471 (18)	0.5962 (17)	
18	YinLi	3	01/21/18	SIRIUS_LC	0.6837 (5)	0.6040 (6)	0.5884 (22)	0.5961 (18)	

Εικόνα Δ.1 Καλύτερα αποτελέσματα διαγωνισμού SemEval 2018 Task 3 Part A

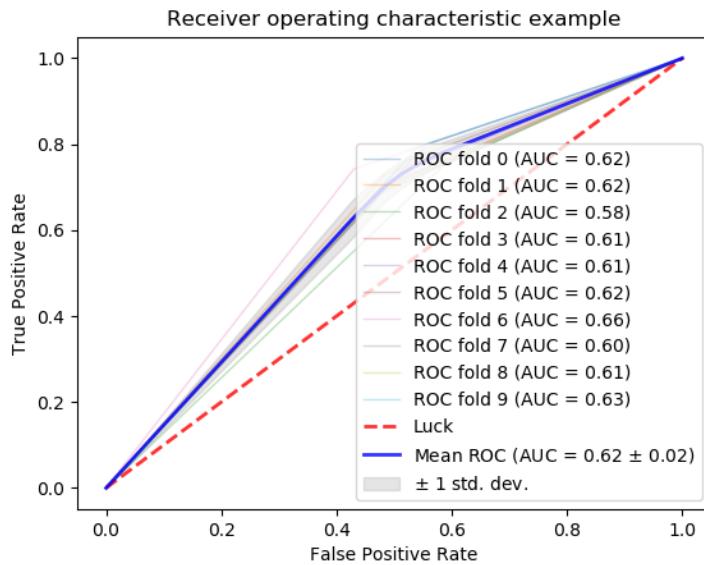
ΜΕΡΟΣ Ε



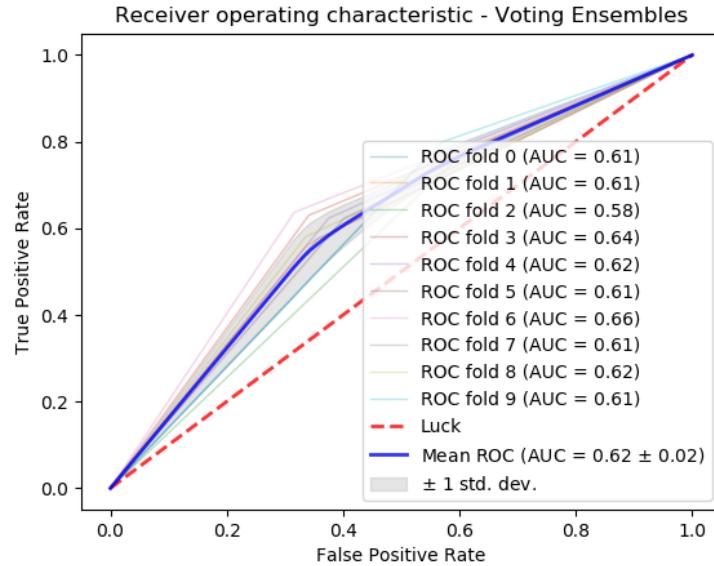
Εικόνα E.1 Καμπύλη ROC-AUC για το μοντέλο Gaussian Naive Bayes με word2vec



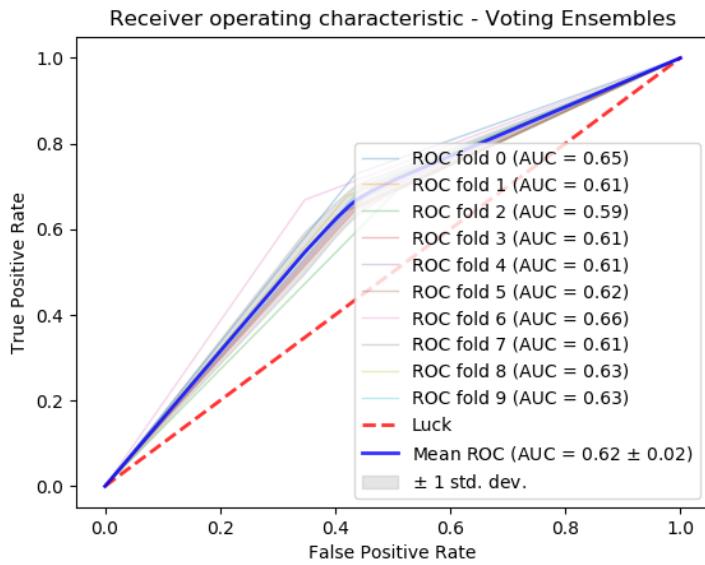
Εικόνα E.2 Καμπύλη ROC-AUC για το μοντέλο K-Neighbors με PCA και TF-IDF



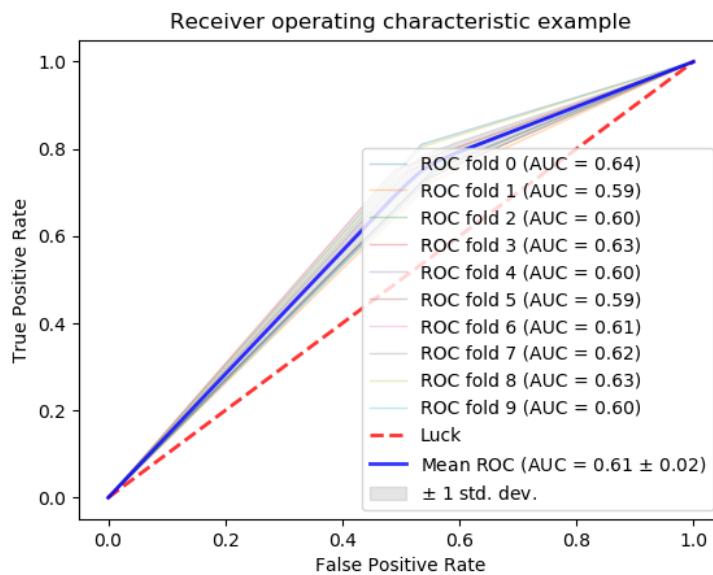
Εικόνα E.3 Καμπύλη ROC-AUC για το μοντέλο SVM με doc2vec



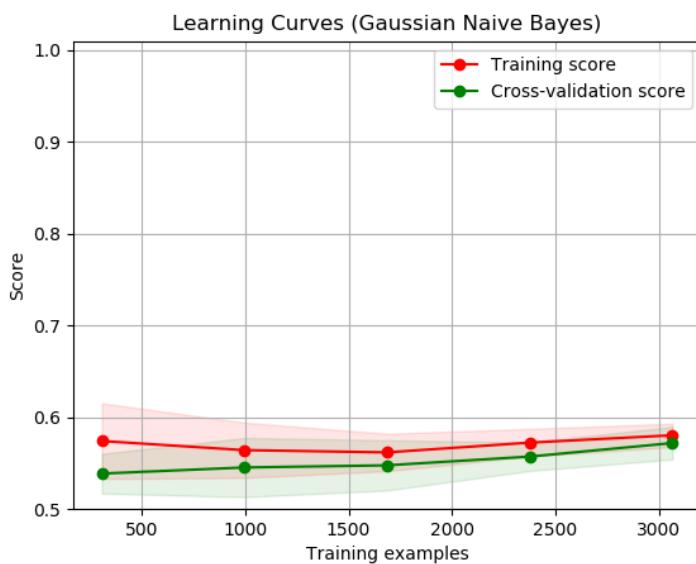
Εικόνα E.4 Καμπύλη ROC-AUC για το μοντέλο Voting Ensembles με Univariate Selection και Bigrams



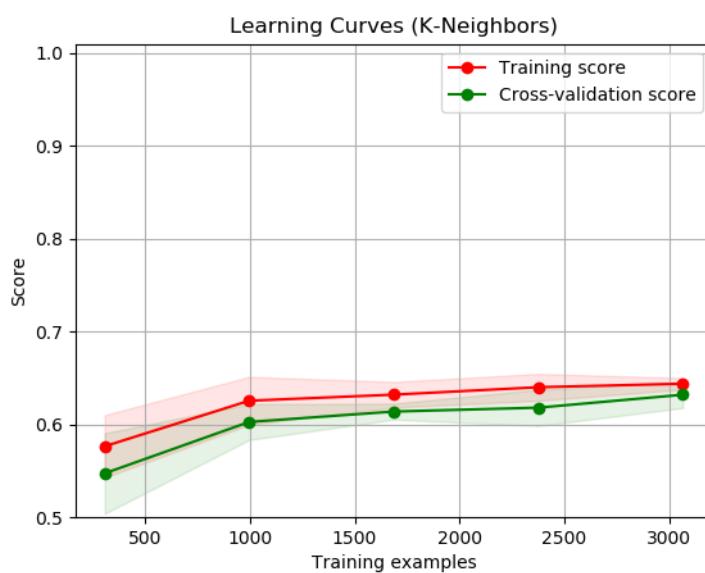
Εικόνα E.5 Καμπύλη ROC-AUC για το μοντέλο Voting Ensembles με doc2vec



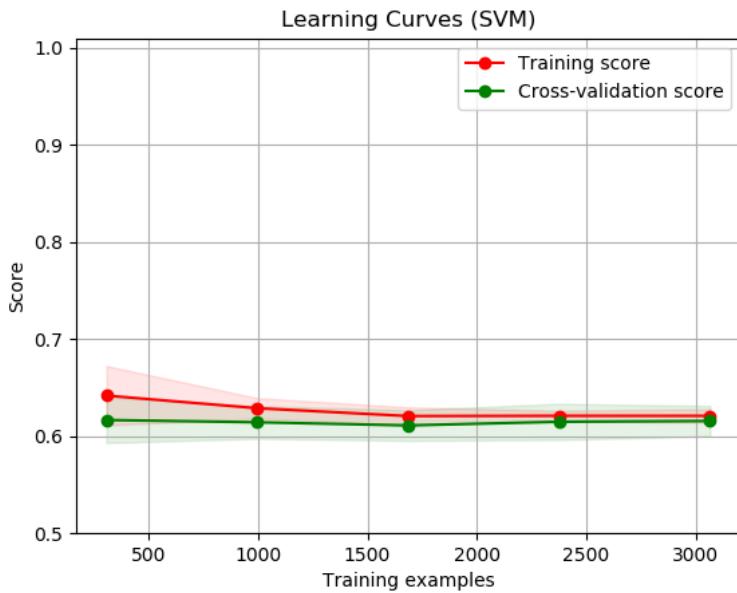
Εικόνα E.6 Καμπύλη ROC-AUC για το μοντέλο SVM με Bigrams



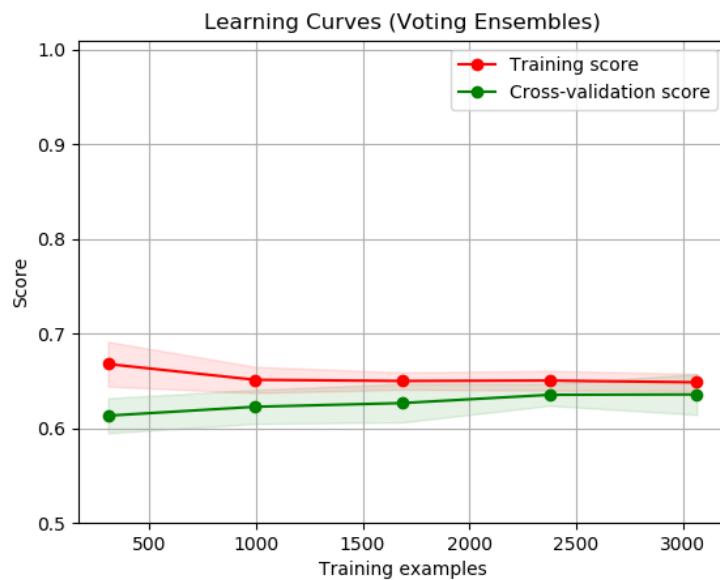
Εικόνα E.7 Καμπύλη Learning Curve για το μοντέλο Gaussian Naive Bayes με word2vec



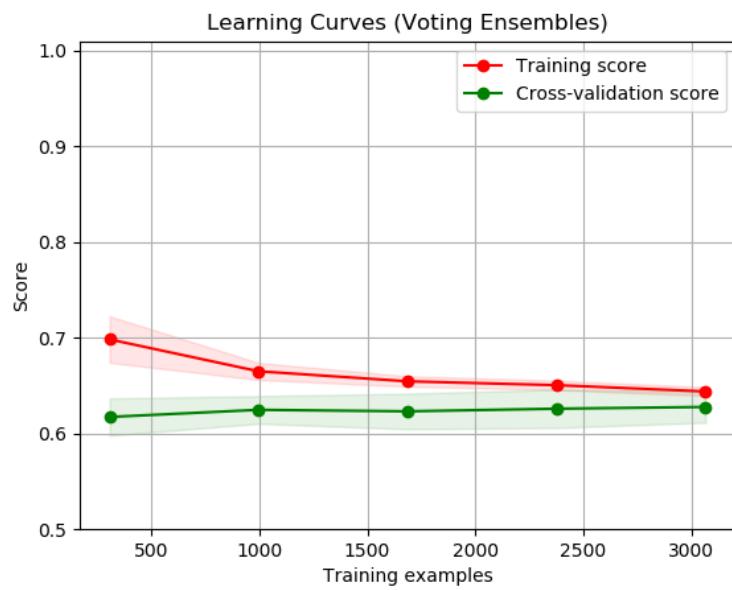
Εικόνα E.8 Καμπύλη Learning Curve για το μοντέλο K-Neighbors με PCA και TF-IDF



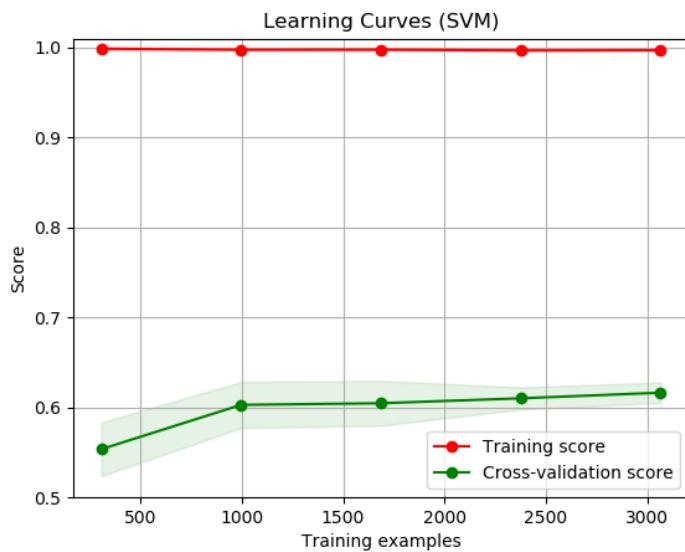
Εικόνα E.9 Καμπύλη Learning Curve για το μοντέλο SVM με doc2vec



Εικόνα E.10 Καμπύλη Learning Curve για το μοντέλο Voting Ensembles με Univariate Selection και Bigrams



Εικόνα E.11 Καμπύλη Learning Curve για το μοντέλο Voting Ensembles με doc2vec



Εικόνα E.12 Καμπύλη Learning Curve για το μοντέλο SVM με Bigrams

Βιβλιογραφία

- [1] Ε. Κύρκος (2015), “Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων”, Καλλίπος.
- [2] Jenn Riley (2004). “Understanding Metadata”, NISO Press, National Information Standards Organization
- [3] Piskorski J., Yangarber R. (2013). “Information Extraction: Past, Present and Future”. In: Poibeau T., Saggion H., Piskorski J., Yangarber R. (eds) Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg
- [4] Daniel Jurafsky & James H. Martin (2006). “Speech and Language Processing”. Stanford University
- [5] Sudhaa Gopinath (2014). “Types of Sentiment Analysis”. Available at: <https://www.edureka.co/blog/types-of-sentiment-analysis/> [Accessed 21 July 2018].
- [6] Walaa Medhat, Ahmed Hassan, Hoda Korashy. (2014). “Sentiment analysis algorithms and applications: A survey”. Ain Shams Eng. J.
- [7] Mikolov T., Chen K., Corrado G.S., & Dean J. (2013). “Efficient Estimation of Word Representations in Vector Space”. CoRR, abs/1301.3781.
- [8] Shubham Agarwal. (2017). “Word to Vectors—Natural Language Processing”. Available at: <https://towardsdatascience.com/word-to-vectors-natural-language-processing-b253d-d0b0817> [Accessed 21 July 2018].
- [9] Juan Ramos. (2003), “Using tf-idf to determine word relevance in document queries”. In First International Conference on Machine Learning, New Brunswick: NJ, USA, Rutgers University.
- [10] Βλαχάβας, Ι., Κεφαλάς, Π., Βασιλειάδης, Ν., Ρεφανίδης, Ι., Κοκκοράς, Φ. & Σακελλαρίου, Η. (2011). “Τεχνητή Νοημοσύνη (3η έκδοση)”. Θεσσαλονίκη: Εκδόσεις Πανεπιστήμιου Μακεδονίας.

- [11] Κατερίνα Γεωργούλη (2015), “Τεχνητή Νοημοσύνη Μια εισαγωγική προσέγγιση”. Κάλλιπος. Pages: 127-131, 155-168, 175-177
- [12] Αργυράκης, Πάνος (2001). “Νευρωνικά Δίκτυα και Εφαρμογές ”. Ελληνικό Ανοιχτό Πανεπιστήμιο. Pages: 15-106, 165-172
- [13] Haykin, Simon (1999). Neural Networks, A Comprehensive Foundation. Prentice Hall International, Inc, New Jersey
- [14] Alex J. Smola and Bernhard Schölkopf (2004). “A tutorial on support vector regression”. NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK.
- [15] Jason Brownlee (2016). "Support Vector Machines for Machine Learning". Available at: <https://machinelearningmastery.com/support-vector-machines-for-machine-learning/> [Accessed 21 July 2018].
- [16] Sunil Ray (2015). “Understanding Support Vector Machine algorithm from examples”. Available at: <https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/> [Accessed 21 July 2018].
- [17] Savan Patel (2017). “Chapter 2: SVM (Support Vector Machine)-Theory”. Available at: <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72> [Accessed 21 July 2018].
- [18] Sunil Ray (2015). “6 Easy Steps to Learn Naive Bayes Algorithm”. Available at: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> [Accessed 21 July 2018].
- [19] Prashant Gupta (2017). “Naive Bayes in Machine Learning”. Available at: <https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4> [Accessed 21 July 2018].
- [20] Jason Brownlee (2016). “Naive Bayes for Machine Learning”. Available at: <https://machinelearningmastery.com/naive-bayes-for-machine-learning/> [Accessed 21 July 2018].

- [21] Shaffi Ahamed Shaikh (2011). "Measures Derived from a 2 x 2 Table for an Accuracy of a Diagnostic Test". Department of Family & Community Medicine, College of Medicine, KSU, Riyadh, Kingdom of Saudi Arabia.
- [22] Powers David M W (2011)."Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. Pages: 1-4.
- [23] Abhishek Sharma. (2017). "Confusion Matrix in Machine Learning". Available at: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/> [Accessed 21 July 2018].
- [24] Fawcett Tom. (2006). "An introduction to ROC analysis". Pattern Recognition Letters 27. Pages: 861-874.
- [25] McClish Donna Katzman (1989). "Analyzing a Portion of the ROC Curve". Medical Decision Making. Pages: 190–195.
- [26] Dodd Lori E., Pepe Margaret S. (2003)."Partial AUC Estimation and Regression". Biometrics. Pages: 614–623.
- [27] Hanley James A., McNeil Barbara J. (1983). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". Radiology. Pages: 839–843.
- [28] Hanczar Blaise, Hua Jianping, Sima Chao, Weinstein John, Bittner Michael, Dougherty Edward R. (2010). "Small-sample precision of ROC-related estimates". Bioinformatics. Pages: 822–830.
- [29] Lobo Jorge M., Jiménez-Valverde Alberto, Real Raimundo. (2008). "AUC: a misleading measure of the performance of predictive distribution models". Global Ecology and Biogeography. Pages: 145–151.
- [30] Hand David J. (2009). "Measuring classifier performance: A coherent alternative to the area under the ROC curve". Machine Learning. Pages: 103–123.

- [31] Cynthia Van Hee, Els Lefever, and Véronique Hoste. (2018). “Semeval-2018 Task 3: Irony detection in English Tweets”. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018.
- [32] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). "Distributed Representations of Words and Phrases and their Compositionality". NIPS.
- [33] Le Q.V., & Mikolov T. (2014). “Distributed Representations of Sentences and Documents”. ICML.
- [34] Pennington J., Socher R., & Manning C.D. (2014). “Glove: Global Vectors for Word Representation”. EMNLP.
- [35] Fan R., Chang K., Hsieh C., Wang X., & Lin C. (2008). “LIBLINEAR: A Library for Large Linear Classification”. Journal of Machine Learning Research, 9, 1871-1874.
- [36] Glorot X., Bordes A., & Bengio Y. (2011). “Deep Sparse Rectifier Neural Networks”. AISTATS.
- [37] Glorot X., & Bengio Y. (2010). “Understanding the difficulty of training deep feed-forward neural networks”. AISTATS.
- [38] Srivastava N., Hinton G.E., Krizhevsky A., Sutskever I., & Salakhutdinov R. (2014). “Dropout: a simple way to prevent neural networks from overfitting”. Journal of Machine Learning Research, 15, 1929-1958.
- [39] T. Tieleman and G. E. Hinton. (2012). “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude”.
- [40] Sepp Hochreiter, Jürgen Schmidhuber. (1997)."Long short-term memory". Neural Computation. Pages: 1735–1780.
- [41] Felix A. Gers, Jürgen Schmidhuber, Fred Cummins. (2000). "Learning to Forget: Continual Prediction with LSTM". Neural Computation. Pages: 2451–2471.
- [42] Min Lin, Qiang Chen, & Shuicheng Yan. (2013). “Network in network”. arXiv preprint arXiv:1312.4400. Page: 4, 7

- [43] Christopher Olah. (2015). "Understanding LSTM Networks". Available at: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> [Accessed 21 July 2018].
- [44] Glorot X., Bordes A., & Bengio Y. (2011). "Deep Sparse Rectifier Neural Networks". AISTATS.
- [45] Debajyoti Datta. (2016). "Understanding Convolutions in Text". Available at: <http://debajyotidatta.github.io/nlp/deep/learning/word-embeddings/2016/11/27/Understanding-Convolutions-In-Text/> [Accessed 21 July 2018].
- [46] Madhavan, P.G. (1997). "A New Recurrent Neural Network Learning Algorithm for Time Series Prediction". Journal of Intelligent Systems. Page: 113 Fig. 3.
- [47] Jacob VanderPlas. (2012). "Astronomy with scikit-learn Release Scipy2012". Page: 28
- [48] Meek Christopher, Thiesson Bo, Heckerman David. (2002). "The Learning-Curve Sampling Method Applied to Model-Based Clustering". Journal of Machine Learning Research. Page: 397.
- [49] Jason Brownlee. (2016). "Overfitting and Underfitting With Machine Learning Algorithms". Available at: <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/> [Accessed 21 July 2018].