Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Εξόρυξη άποψης και ανάλυση συναισθήματος σε μέσα κοινωνικής δικτύωσης

Δημήτριος Ε. Πιζάνιας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2018

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Ν. Πελέκης (Επιβλέπων)
- Καθηγητής Ι. Θεοδωρίδης
- Καθηγητής Ελ. Κοφίδης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Opinion Mining and Sentiment Analysis in Social Media

By

Dimitrios E. Pizanias

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2018

Περίληψη

Η ανάπτυξη των τεχνολογιών του διαδικτύου, η διεύρυνση του κοινού του και των χρήσεων του είχε σαν συνέπεια την αύξηση του όγκου των συνόλων δεδομένων που είναι διαθέσιμα μέσω του παγκοσμίου ιστού. Το γεγονός αυτό δημιούργησε την ανάγκη της ανάπτυξης αποδοτικών μηχανισμών εκμετάλλευσης των δεδομένων αυτών καθώς οι παραδοσιακές τεχνικές ήταν πλέον ανεπαρκείς στο να μπορέσουν να διαχειριστούν αυτόν τον όγκο δεδομένων. Η σημαντικότερη πηγή δεδομένων που σχετίζεται με τις τάσεις της κοινής γνώμης είναι τα κοινωνικά δίκτυα. Σε αυτά το μεγαλύτερο ποσοστό του παγκοσμίου πληθυσμού αναρτά την άποψη του, τις ανησυχίες του και τα σχόλια του για μία μεγάλη ποικιλία θεματολογίας. Στην παρούσα εργασία αναπτύσσονται σε θεωρητικό και πρακτικό επίπεδο μηχανισμοί που χρησιμοποιούνται για τον προσδιορισμό του συναισθηματικού προσανατολισμού των χρηστών των κοινωνικών δικτύων.

Abstract

The Internet technologies development, the widening of its audience and its uses has resulted in increasing the data sets' volume available through the world wide web. This fact emerged the need for developing new efficient mechanisms for exploiting these data, as traditional techniques were no longer sufficient to manage this volume of data. The most important source of data related to public opinion and tendencies is the social networks. To these, the largest proportion of the world's population is posting its opinion, concerns and comments on a wide range of subjects. In the present thesis the mechanisms used to determine the sentiment orientation of the social networks users are being presented at a theoretical and practical view.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

CRF	Conditional Random Fields
LSA	Latent Semantic Analysis
HAL	Hyperspace Analogue to Language
PMI	Positive, Minus, Interesting
RST	Rhetorical Structure Theory
SMO	Sequential Minimal Optimization
LMT	Logistic Model Trees
PART	Projective Adaptive Resonance Theory
RIP	Repeated Incremental Pruning to Produce Error Reduction
WEKA	Waikato Environment for Knowledge Analysis

Εισαγωγή

Η σύγχρονη εποχή χαρακτηρίζεται από την ραγδαία πρόοδο των τεχνολογιών του διαδικτύου. Η πρόσβαση σε αυτό έχει γίνει προσιτή για μεγάλο ποσοστό του πληθυσμού παγκοσμίως καθώς η ανάπτυξη της τεχνολογίας μείωσε κατά πολύ το απαιτούμενο κόστος των ποιοτικών συνδέσεων. Παράλληλα η πρόσβαση στις κάθε λογής διαδικτυακές υπηρεσίες έγινε εφικτή και από ασύρματα κανάλια επικοινωνίας ακόμα και από ασύρματες έξυπνες κινητές συσκευές. Αποτέλεσμα όλων αυτών ήταν να διευρυνθεί το διαδικτυακό κοινό τόσο ποσοτικά όσο και ποιοτικά. Πολύ περισσότεροι άνθρωποι χρησιμοποιούν διάφορες διαδικτυακές εφαρμογές για διαφορετικούς σκοπούς. Οι άνθρωποι αυτοί μπορεί να ανήκουν σε διαφορετικές κοινωνικές τάξεις, να έχουν διαφορετική οικονομική επιφάνεια και να είναι οποιασδήποτε ηλικίας.

Σημαντικός παράγοντας της διεύρυνσης του κοινού του διαδικτύου αποτέλεσαν και οι εφαρμογές των οποίων το περιεχόμενο διαμορφώνεται από τους ίδιους τους χρήστες. Τέτοιου είδους εφαρμογές είναι τα ιστολογία, τα κοινωνικά δίκτυα και δικτυακοί τόποι ηλεκτρονικού εμπορίου όπου οι χρήστες καλούνται να αξιολογήσουν το περιεχόμενο που προβάλλεται ή να καταθέσουν το σχόλιο τους. Η χρήση των εφαρμογών αυτών παρουσιάζει υψηλή δυναμική η οποία εκτιμάται ότι θα παραμείνει τουλάχιστον στα ίδια επίπεδα. Η φύση των εφαρμογών αυτών τις καθιστά πολύτιμες πηγές για την αναζήτηση των τάσεων της κοινής γνώμης. Οι αξιολογήσεις και τα σχόλια των χρηστών αποδίδουν την στάση των ανθρώπων απέναντι στα προ βληθέντα. Ο προσδιορισμός των τάσεων αυτών αποτελεί αντικείμενο της ανάλυσης συναισθήματος η οποία αναφέρεται σε μηχανισμούς επεξεργασίας και εκμετάλλευσης των δεδομένων που προαναφέρθηκαν. Τα αποτελέσματα αυτής αποδεικνύονται πολύτιμα για διάφορους σκοπούς όπως η λήψη αποφάσεων, οι οικονομικές αναλύσεις, η εκτίμηση τάσεων σε εμπορικό, κοινωνικό ή πολιτικό επίπεδο.

Η ανάλυση συναισθήματος χρησιμοποιεί δεδομένα κυρίως προερχόμενα από κοινωνικά δίκτυα. Η τεχνολογία παρέχει ισχυρά εργαλεία που διευκολύνουν τους μηχανισμούς της να εκτελούνται ταχύτερα και να παράγουν αποτελέσματα με την μέγιστη δυνατή ακρίβεια. Οι μελέτες που σχετίζονται με αυτήν έχουν αποδώσει αποδοτικούς αλγορίθμους προς την ίδια κατεύθυνση. Η επιλογή των εργαλείων και των τεχνικών που χρησιμοποιούνται κάθε φορά είναι συνάρτηση κυρίως του επιδιωκόμενου σκοπού.

Στην παρούσα εργασία γίνεται αρχικά μία παρουσίαση της προσέγγισης της ανάλυσης συναισθήματος καθώς και των μεθόδων που χρησιμοποιούνται. Στην συνέχεια παρουσιάζονται στην πράξη οι κυριότερες από τις μεθόδους και αξιολογούνται τα παραγόμενα αποτελέσματα. Το υπόλοιπο του παρόντος κειμένου έχει διαρθρωθεί ως εξής:

- Κεφάλαιο 1: Στο πρώτο κεφάλαιο περιγράφεται η έννοια της ανάλυσης συναισθήματος από δεδομένα που προέρχονται από διαδικτυακές εφαρμογές. Καταγράφονται επιγραμματικά οι μεθοδολογίες που χρησιμοποιούνται στις διαδικασίες της αλλά και οι σκοποί για τους οποίους μπορεί να είναι χρήσιμη.

- Κεφάλαιο 2: Στο κεφάλαιο αυτό παρουσιάζονται οι κυριότερες μεθοδολογίες και τεχνικές ανάλυσης συναισθήματος που στηρίζονται σε λεξικά.

- Κεφάλαιο 3: Σε αυτό παρουσιάζονται διαδικασίες ανάλυσης συναισθήματος που εκτελέστηκαν με τις μεθοδολογίες και τις τεχνικές της μηχανικής μάθησης. Για την υλοποίηση της ανάλυσης συναισθήματος χρησιμοποιήθηκες η γλώσσα R.

- Κεφάλαιο 4: Τα αποτελέσματα που προέκυψαν από την εκτέλεση των διαδικασιών αναλύονται και αξιολογούνται στο 4ο κεφάλαιο. Το ζητούμενο είναι να προσδιοριστούν οι καταλληλότερες μεθοδολογίες για κάθε σκοπό αλλά και να καθοριστούν οι προϋποθέσεις που πρέπει να πληρούνται σε κάθε περίπτωση ώστε η ανάλυση συναισθήματος να παράγει αποτελέσματα με την μεγαλύτερη δυνατή αξιοπιστία.

- Κεφάλαιο 5: Στο τελευταίο κεφάλαιο της εργασίας παρουσιάζονται τα συμπεράσματα που προέκυψαν τόσο από την θεωρητική μελέτη της ανάλυσης συναισθήματος όσο και από την πρακτική εφαρμογή των κυριότερων μεθοδολογιών.

Κεφάλαιο 1

Ανάλυση Συναισθήματος

Η ανάλυση των συναισθημάτων είναι ένας από τους ταχύτερα αναπτυσσόμενους τομείς έρευνας στον τομέα της πληροφορικής. Οι ρίζες της εντοπίζονται στις μελέτες για την ανάλυση της κοινής γνώμης στις αρχές του 20ου αιώνα καθώς και στην ανάλυση υποκειμενικότητας κειμένου που αναπτύχθηκε τη δεκαετία του 1990. Τα τελευταία χρόνια εντάθηκε το ενδιαφέρον για την ανάλυση συναισθημάτων ως αποτέλεσμα της υψηλής διαθεσιμότητας υποκειμενικών κειμένων στο διαδίκτυο.

Η ανάλυση συναισθήματος είναι ένα σύνολο μεθόδων, τεχνικών και εργαλείων για την ανίχνευση και την εξαγωγή πληροφοριών από υποκειμενικές πηγές. Τέτοιες πηγές είναι γνώμες, στάσεις, χρήση της γλώσσας. Αρχικά περιορίστηκε στην πολικότητα της κοινής γνώμης και αντικείμενο των σχετικών ερευνών ήταν η θετική ή αρνητική στάση απέναντι σε ζητήματα. Η διάκριση της ανάλυσης συναισθήματος και της εξόρυξης γνώμης δεν είναι ευδιάκριτη και συχνά οι έννοιες τους ταυτίζονται.

Το ενδιαφέρον για τη γνώση της γνώμης άλλων ανθρώπων εντοπίζεται στις πρώτες οργανωμένες κοινωνίες. Η μέριμνα των ηγετών για την διατήρηση ή αύξηση της δημοτικότητα τους είναι διαχρονική και ένα εργαλείο για τον σκοπό αυτό είναι το να αφουγκράζονται την διάθεση των υφισταμένων τους. Τον 5ο π.Χ. αιώνα εφαρμόζεται πρώτη φορά η ψηφοφορία για τον προσδιορισμό της κοινής επιθυμίας ενώ τα πρώτα ερωτηματολόγια καταμέτρησης τάσεων εμφανίστηκαν στις αρχές του 20ου αιώνα.

Οι πρώτες ακαδημαϊκές μελέτες για τον προσδιορισμό της κοινής γνώμης πραγματοποιήθηκαν κατά τον Β Παγκόσμιο Πόλεμο για πολιτικούς σκοπούς. Στα μέσα της προηγούμενης δεκαετίας εντάθηκε με βασικό στόχο την αξιολόγηση και μέσω αυτής την αναβάθμιση προϊόντων που διατίθενται μέσω του διαδικτύου. Η ανάπτυξη των τεχνολογιών του διαδικτύου, η εμφάνιση των τεχνικών διαμόρφωσης του διαδικτυακού περιεχομένου από τους ίδιους τους χρήστες και της κοινωνικής δικτύωσης επέφερε την επέκταση της χρήσης της ανάλυσης συναισθήματος και για άλλους σκοπούς όπως η πρόβλεψη εξέλιξης της κατάστασης των χρηματοπιστωτικών αγορών, η αποδοχή προϊόντων από το κοινό, η πρόβλεψη αντιδράσεων σε απότομα εναλλασσόμενες καταστάσεις και σε πολιτικοκοινωνικές αλλαγές. Τα τελευταία χρόνια οι μελέτες που σχετίζονται με την ανάλυση των συναισθημάτων

δοκιμάζονται από νέες προκλήσεις όπως η ανίχνευση πιο πολύπλοκών συναισθημάτων, ο εντοπισμός μεταφορικής χρήσης των λέξεων, η πολύ γλωσσική προσαρμογή και ο εντοπισμός σύνθετων συναισθημάτων που επεκτείνουν τον περιορισμένο εντοπισμό της πολικότητας. Οι πηγές των δεδομένων που χρησιμοποιούνται είναι τα έντυπα και ηλεκτρονικά μέσα ενημέρωσης, τα κοινωνικά δίκτυα, το διαδικτυακό περιεχόμενο και εικόνες (V.Mäntyläa, Graziotinb, & Kuutilaa, 2017). Οι μεθοδολογίες που χρησιμοποιούνται για την ανάλυση συναισθημάτων κατηγοριοποιούνται γενικά σε τρείς κατηγορίες:

- Μηχανική μάθηση
- Επεξεργασία φυσικής γλώσσας
- Ειδικές μέθοδοι ανάλυσης αισθήματος

Στις επόμενες παραγράφους του παρόντος κεφαλαίου περιγράφονται τα χαρακτηριστικά της ανάλυσης συναισθήματος, οι σκοποί που εξυπηρετεί καθώς και οι μεθοδολογίες που χρησιμοποιούνται.

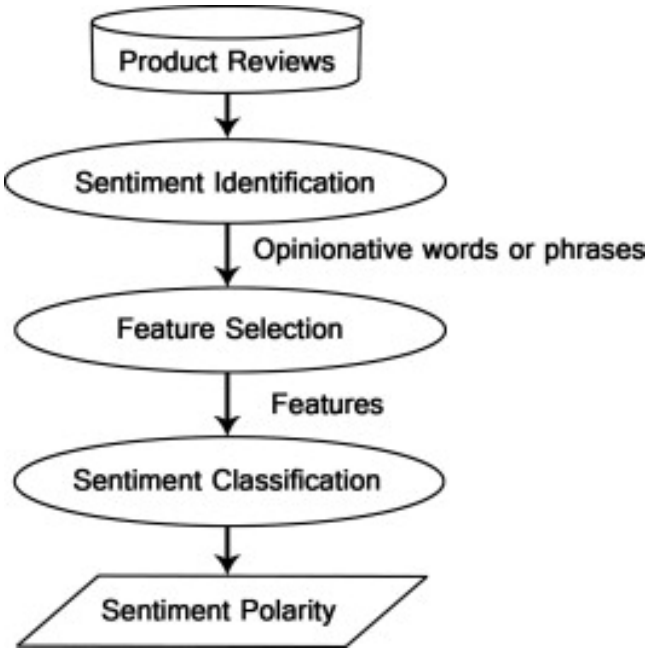
1.1 Περιγραφή

Η Ανάλυση Συναισθημάτων (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining) είναι η υπολογιστική μελέτη των απόψεων, των στάσεων και των συναισθημάτων των ανθρώπων με σημείο αναφοράς μια συγκεκριμένη οντότητα. Η οντότητα μπορεί να είναι άνθρωπος, σύνολο ανθρώπων, γεγονός, κοινωνικό – πολιτικό – οικονομικό ζήτημα, αντικείμενο. Πολύ συχνά οι οντότητες αυτές είναι αναθεωρήσιμες και οι διαδικασίες αναθεώρησης τους βασίζονται στην ανάλυση συναισθημάτων. Οι δύο εκφράσεις – Ανάλυση Συναισθημάτων ή Εξόρυξη Γνώμης είναι σχεδόν ταυτόσημες. Η διαφορά τους έγκειται στο ότι η εξόρυξη γνώμης αναζητά την στάση του ανθρώπου απέναντι στο αντικείμενο που εξετάζεται ενώ η ανάλυση συναισθήματος επεκτείνεται και στα συναισθήματα που αναδεικνύονται από τις προς μελέτη πηγές. Η γενική διαδικασία της ανάλυσης συναισθήματος φαίνεται στην παρακάτω εικόνα.

ΣΧΗΜΑ 1-1

Διαδικασία Ανάλυσης Συναισθήματος

(https://www.sciencedirect.com/science/article/pii/S2090447914000550)



Η Ανάλυση Συναισθήματος είναι στην πραγματικότητα μια διαδικασία ταξινόμησης που πραγματοποιείται σε τρία στάδια. Στο πρώτο στάδιο γίνεται η κατηγοριοποίηση σε επίπεδο εγγράφου, σε δεύτερο στάδιο γίνεται σε επίπεδο χαρακτηριστικών του λόγου και σε τρίτο στάδιο γίνεται σε επίπεδο προσδιορισμού συναισθήματος. Στο πρώτο στάδιο ένα έγγραφο κατατάσσεται ανάλογα με το περιεχόμενο του ως εκφράζον θετική, αρνητική γνώμη ή συναίσθημα. Στο στάδιο αυτό ολόκληρο το έγγραφο αντιμετωπίζεται ως οντότητα. Σε δεύτερο στάδιο η ανάλυση γίνεται σε επίπεδο προτάσεων και γίνεται προσπάθεια να ανιχνευθεί το συναίσθημα που εκφράζει. Αρχικά εντοπίζεται εάν είναι υποκειμενική ή αντικειμενική έκφραση. Στην πρώτη περίπτωση αναζητείται το αν έχει θετικό ή αρνητικό ύφος. Επειδή συχνά δεν υπάρχει διαφοροποίηση στην ταξινόμηση σε επίπεδο εγγράφων ή προτάσεων – καθώς στην πραγματικότητα οι προτάσεις είναι σύντομα έγγραφα - είναι απαραίτητο ένα ακόμα στάδιο. Στο τρίτο στάδιο της ανάλυσης όπου αναζητείται η ταξινόμηση σε επίπεδο έκφρασης συναισθήματος. Σε αυτό το στάδιο λαμβάνονται υπ' όψη τα προς κρίση χαρακτηριστικά της εξεταζόμενης οντότητας (το αντικείμενο της γνώμης των συμμετεχόντων) αλλά και – ενδεχομένως – τα χαρακτηριστικά των πηγών δεδομένων. Οι πηγές αυτές είναι κυρίως σχόλια σε κοινωνικά δίκτυα ή ιστολόγια, αξιολογήσεις προϊόντων, καταστάσεων ή λοιπών οντοτήτων (Medhat, Hassan, & Korashyb, 2014).

1.2 Μεθοδολογίες

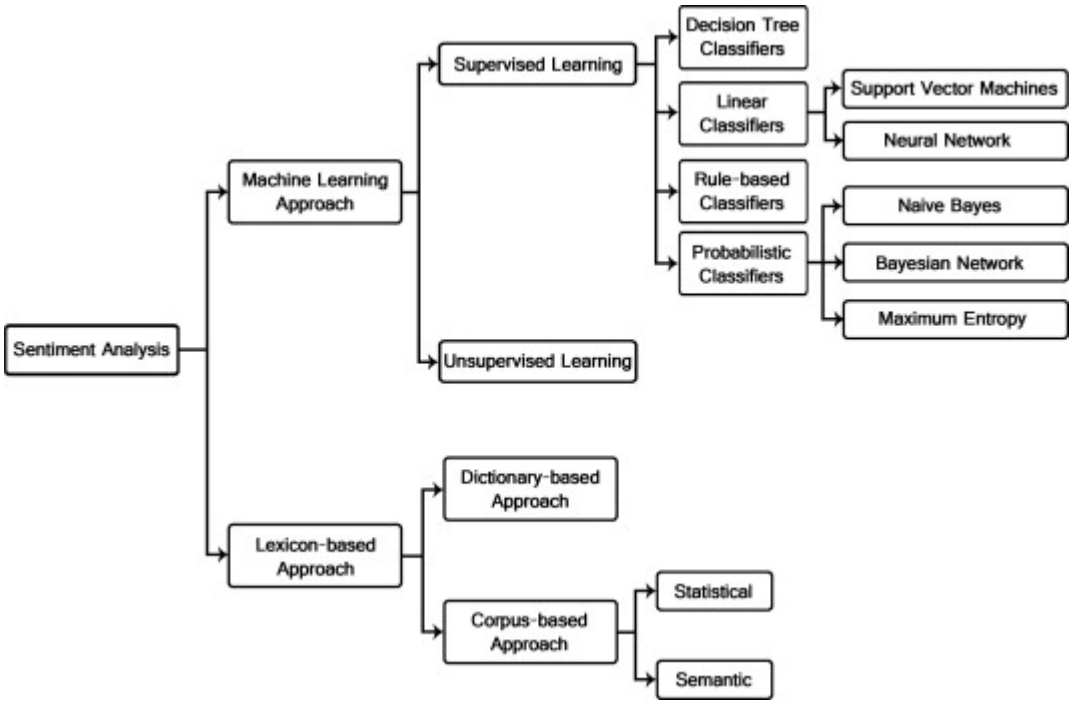
Για την ανάλυση συναισθήματος έχουν προταθεί πολλές μέθοδοι. Επιχειρώντας την κατηγοριοποίηση τους μπορεί να διαμορφωθεί η ακόλουθη ταξινόμηση:

- Διαδικασίες μηχανικής μάθησης: Πρόκειται για τεχνικές που μέσα από την επεξεργασία φυσικής γλώσσας αναζητείται να εντοπιστούν τα συναισθήματα που εκφράζουν τα κείμενα. Με τις διαδικασίες αυτές οι μηχανές είναι σε θέση να επεξεργάζονται ακολουθίες συμβολοσειρών και να επιστρέφουν τα συναισθήματα που αυτές αντιπροσωπεύουν. Οι τεχνικές αυτές χρησιμοποιούν ισχυρούς αλγορίθμους προκειμένου να εκπαιδεύονται οι μηχανές στην εκτίμηση συναισθημάτων όταν η είσοδος που δέχονται είναι κείμενα σε φυσική γλώσσα. Οι τεχνικές αυτές συχνά έχουν υψηλές απαιτήσεις σε προπαρασκευαστικές ενέργειες και υπολογιστικούς πόρους. Οι κατηγορίες των διαδικασιών που περιλαμβάνονται στην κατηγορία αυτή ακολουθούν την διάκριση της μηχανικής μάθησης σε εποπτευόμενη και μη εποπτευόμενη ενώ η εποπτευόμενη μηχανική μάθηση κατηγοριοποιείται περαιτέρω με βάση το είδος των αλγορίθμων που χρησιμοποιούνται (Llombart, 2017).
- Διαδικασίες που βασίζονται σε λεξικά: Πρόκειται για διαδικασίες που βασίζονται στην ανάπτυξη πεπερασμένης έκτασης λεξικών αντιστοίχησης λέξεων και εκφράσεων με συγκεκριμένα συναισθήματα. Η αποδοτικότητα των τεχνικών αυτών βασίζεται σε μεγάλο βαθμό στην πληρότητα και την ευστοχία των λεξικών αυτών. Η περαιτέρω διάκριση τους περιλαμβάνει δύο κατηγορίες, τις τεχνικές που εξετάζουν λέξεις και εκείνες που εξετάζουν φράσεις (Musto, Semeraro, & Polignano, 2015).

Στην επόμενη εικόνα φαίνεται σχηματικά και συνοπτικά η διάκριση των τεχνικών ανάλυσης συναισθήματος.

ΣΧΗΜΑ 1-2

Κατηγοριοποίηση τεχνικών ανάλυσης συναισθήματος
(https://www.sciencedirect.com/science/article/pii/S2090447914000550)



Στο $2^{ο}$ και $3^{ο}$ κεφάλαιο παρουσιάζονται τα χαρακτηριστικά των κυριότερων τεχνικών – μεθόδων ανάλυσης συναισθήματος ενώ στο $4^{ο}$ κεφάλαιο παρουσιάζεται η εφαρμογή τους σε μεγάλα σύνολα δεδομένων.

1.3 Χρήσεις

Η ανάλυση συναισθήματος και η εξόρυξη άποψης από κείμενα έχει ήδη εμφανίσει σημαντικές εφαρμογές σε πολλές και ποικίλες εκφάνσεις της ανθρώπινης δραστηριότητας. Σε αυτό έχει συμβάλει η υπερπληθώρα αποτύπωσης σε κείμενο απόψεων και συναισθημάτων μεγάλου μέρους του παγκοσμίου πληθυσμού οι οποίες διατίθενται στο διαδίκτυο. Κύρια πηγή αποτελούν οι εφαρμογές κοινωνικής δικτύωσης αλλά και τα σχόλια που καταθέτουν οι χρήστες του διαδικτύου σε διαφόρων ειδών διαδικτυακές εφαρμογές. Στις επόμενες παραγράφους αναφέρονται μερικοί από τους τομείς όπου η εξόρυξη άποψης και συναισθήματος αποτελεί έναν σημαντικό μηχανισμό για το προσωπικό που εμπλέκεται σε αυτούς.

1.3.1 ΦΙΛΟΞΕΝΙΑ ΚΑΙ ΤΑΞΙΔΙΑ

Η επιτυχία των επιχειρήσεων που επενδύουν στον τομέα του τουρισμού εξαρτάται σε μεγάλο βαθμό από το κατά πόσο οι πελάτες των επιχειρήσεων αυτών μένουν ευχαριστημένοι από τις υπηρεσίες που τους προσφέρονται. Είναι έτσι σημαντικό για τους επιχειρηματίες του χώρου να μπορούν να αφουγκράζονται τα συναισθήματα που δημιουργούνται στους υφισταμένους ώστε να λαμβάνουν επιχειρησιακές αποφάσεις που θα κάνουν τις υπηρεσίες τους περισσότερο ελκυστικές. Οι περισσότερες από τις επιχειρήσεις αυτές ζητούν από τους πελάτες τους να αξιολογήσουν τις υπηρεσίες που τους παρασχέθηκαν. Οι αποκρίσεις τους αποτελούν μία σημαντική πηγή προσδιορισμού της άποψης τους για την επιχείρηση τους. Την σύγχρονη εποχή που το μεγαλύτερο μέρος του παγκοσμίου πληθυσμού χρησιμοποιεί κατά μεγάλα χρονικά διαστήματα τα κοινωνικά δίκτυα η εξόρυξη της άποψης της κοινής γνώμης από δεδομένα που προέρχονται από αυτά είναι επίσης μία σημαντική πηγή όχι απλά για να ανιχνεύσουν την στάση της απέναντι στη επιχείρηση τους αλλά και για να μπορέσουν να προσδιορίσουν γενικότερα τις επικρατούσες τάσεις στις επιθυμίες της σχετικά με τα ταξίδια και την φιλοξενία.

Οι επιφορτισμένοι με την λήψη αποφάσεων στις τουριστικές επιχειρήσεις από την ανάλυσης συναισθημάτων και τη εξόρυξη άποψης προσδοκούν:

- Να κατανοήσουν τον τρόπο με τον οποίο οι χρήστες του διαδικτύου αξιολόγησαν τις υπηρεσίες που απήλαυσαν από την ίδια ή συναφείς επιχειρήσεις.
- Να προσδιορίσουν τις κατηγορίες υπηρεσιών που αξιολογούν πελάτες προερχόμενοι από διαφορετικές γεωγραφικές περιοχές ως σημαντικές ως προς την επιλογή της επιχείρησης που θα συνεργαστούν.
- Να ανιχνεύσουν τις υπηρεσίες και τις καταστάσεις που μπορεί να επηρεάσουν θετικά ή αρνητικά την στάση των δυνητικών πελατών απέναντι σε μία επιχείρηση του τουρισμού.
- Να εντοπίσουν ενδεχόμενες καινοτομίες οι οποίες θα τύχουν θετικής υποδοχής από το ταξιδιωτικό κοινό (Faggella, 2018).

1.3.2 ΑΞΙΟΛΟΓΗΣΗ ΠΑΡΟΧΗΣ ΥΠΗΡΕΣΙΩΝ ΠΡΟΣ ΠΕΛΑΤΕΣ

Σε ένα έντονα ανταγωνιστικό επιχειρηματικό περιβάλλον η ικανοποίηση των πελατών από τις υπηρεσίες που τους παρέχονται είναι ζωτικής σημασίας για την βιωσιμότητα τους. Οι φόρμες καταγραφής των εντυπώσεων είναι ένας τρόπος συλλογής στοιχείων που σχετίζονται με τον βαθμό ικανοποίησης του πελάτη. Ωστόσο ο όγκος τους είναι σχετικά περιορισμένος με συνέπεια και τα συμπεράσματα που θα μπορούσαν να προκύψουν από αυτά να έχουν επίσης περιορισμένη ακρίβεια. Σημαντικές πηγές για την εξόρυξη της άποψης της κοινής γνώμης για τις υπηρεσίες που παρέχει ένας οργανισμός μπορεί να είναι οι αναρτήσεις και τα σχόλια των χρηστών των κοινωνικών δικτύων αλλά και τα σχόλια των επισκεπτών επιλεγμένων διαδικτυακών τόπων. Τα περιεχόμενα αυτών συλλέγονται και με τεχνικές ανάλυσης συναισθήματος επιτυγχάνουν οι οργανισμοί να αντιλαμβάνονται τον βαθμό ικανοποίησης πελατών ποικίλων οργανισμών.

Μία σχετικά καινοτόμα τάση για την ανίχνευση της ικανοποίησης των πελατών από τις υπηρεσίες των οργανισμών είναι η ανάλυση των περιεχομένων των τηλεφωνικών συνομιλίων των καταναλωτών με τους χειριστές των τηλεφωνικών κέντρων. Βασικός στόχος της προσέγγισης αυτής είναι να εντοπίζεται η διάθεση του πελάτη ως προς την διατήρηση της προτίμησης του στον υπ' όψη οργανισμό. Τα φωνητικά δεδομένα κάθε κλήσης μετατρέπονται σε κείμενο χρησιμοποιώντας κατάλληλο λογισμικό. Το κείμενο αυτό αποστέλλεται στη συνέχεια αναλύεται ώστε να εντοπιστεί η στάση του πελάτη ως προς προϊόντα, υπηρεσίες ή την εξυπηρέτηση του. Για κάθε πελάτη καταγράφεται μία βαθμολογία που αντικατοπτρίζει τον βαθμό ικανοποίησης του. Σε τακτά χρονικά διαστήματα, μια ξεχωριστή διαδικασία ελέγχει την βάση δεδομένων των πελατών για να διαπιστώσει εάν κάποιοι από αυτούς έχουν βαθμό ικανοποίησης κάτω από ένα κατώφλι και για ποιο χρονικό διάστημα παραμένουν ανικανοποίητοι. Συνήθως οι επιχειρήσεις μεριμνούν ώστε να προβαίνουν σε κινήσεις υπέρ τους ώστε να ανακτήσουν την εμπιστοσύνη τους (πχ προσφορές αγαθών και υπηρεσιών σε μικρότερες ή μηδαμινές τιμές). Παράλληλα σε ενδεχόμενες μελλοντικές κλήσεις ο τηλεφωνητής της επιχείρησης ενημερώνεται σε πραγματικό χρόνο για το ιστορικό ικανοποιησιμότητας του πελάτη με τον οποίο συνομιλούν ώστε να προσαρμόσουν την συμπεριφορά τους απέναντι τους (repustate, 2018).

1.3.3 Πολιτικη

Στην πολιτική σκηνή η γνώση της κοινής γνώμης είναι πολύ σημαντική για όλους τους εμπλεκομένους. Τα επιτελεία των πολιτικών σε όλα τα επίπεδα της πολιτικής επενδύουν σε μεγάλο βαθμό στον ορθό προσδιορισμό των τάσεων του εκλογικού σώματος. Οι παραδοσιακοί τρόποι αφουγκρασμού της κοινής γνώμης περιλαμβάνουν δημοσκοπήσεις (τηλεφωνικές, γραπτές, ηχογραφημένες) και συνεντεύξεις. Η αποτελεσματικότητα των μεθόδων αυτών εξαρτάται από το πλήθος των συμμετεχόντων. Όσο διευρύνεται όμως το πλήθος των συμμετεχόντων τόσο αυξάνονται και οι απαιτήσεις σε οικονομικούς πόρους αλλά και σε χρόνο. Το διαδίκτυο πλέον δίνει την δυνατότητα στα επιτελεία των πολιτικών να έχουν μία συνεχή και πιο άμεση επαφή με το ευρύ κοινό των ψηφοφόρων. Οι τεχνολογίες του Web2.0 δίνουν την ευκαιρία σε όλους τους χρήστες του διαδικτύου να συμμετέχουν στην διαμόρφωση του περιεχομένου του. Αυτό γίνεται κατά κύριο λόγο στις εφαρμογές κοινωνικής δικτύωσης όπου κάθε κάτοχος λογαριασμού μπορεί να δημιουργήσει αναρτήσεις που να περιλαμβάνουν τις απόψεις του επί παντός επιστητού. Το περιεχόμενο των αναρτήσεων αυτών αναλύεται από τα επιτελεία των πολιτικών προκειμένου μεταξύ άλλων να εντοπιστούν:

- Ο αντίκτυπος της πολιτικής καμπάνιας στο σύνολο του εκλογικού σώματος αλλά σε κατηγορίες αυτού. Οι κατηγορίες του εκλογικού σώματος προσδιορίζονται με βάση ποικίλα κριτήρια (δημογραφικά, γεωγραφικά κα).
- Ο τρόπος με τον οποίο το εκλογικό σώμα τάσσεται απέναντι σε πολιτικές επιλογές.
- Οι προσδοκίες των πολιτών από τους πολιτικούς
- Καταστάσεις και συμπεριφορές που προκαλούν θετική ή αρνητική εντύπωση στους πολίτες.

1.3.4 ΠΡΟΩΘΗΤΙΚΕΣ ΕΝΕΡΓΕΙΕΣ

Οι προωθητικές ενέργειες απαιτούν την λήψη έγκαιρων και εύστοχων αποφάσεων σχετικά με κινήσεις τις οποίες θα υποδεχθεί το κοινό στόχος με θετική διάθεση. Βασικός παράγοντας για την εκπλήρωση των προϋποθέσεων αυτών αποτελεί η γνώση του κατά πόσο είναι θετικά διακείμενο το καταναλωτικό κοινό σε δεδομένες καταστάσεις που είτε έχουν διαμορφωθεί κατά το (πρόσφατο ή απώτερο) παρελθόν είτε είναι ενδεχόμενες. Απαραίτητο εργαλείο για αυτό είναι τα σχόλια και οι αναρτήσεις του καταναλωτικού κοινού στις διάφορες εφαρμογές του παγκοσμίου ιστού. Η ανάλυση των συναισθημάτων και η εξόρυξη άποψης από τις πηγές αυτές μπορεί να χρησιμοποιηθεί για έναν περισσότερους από τους παρακάτω σκοπούς (Gallantra Business Intelligence, 2018):

- Διαχείριση της φήμης του οργανισμού: Η ανάλυση των αναρτήσεων και των σχολίων του καταναλωτικού κοινού στις διαδικτυακές εφαρμογές είναι ικανή να αναδείξει ασθενή σημεία της σχέσης της επιχείρησης με το κοινό της. Ο έγκαιρος εντοπισμός τους δίνει της δίνει την ευκαιρία να προβαίνει σε αποτελεσματικές διορθωτικές κινήσεις ώστε να αποτρέψει την οριστική απώλεια της σύνδεσης με τους πελάτες της. Από την άλλη μεριά ο εντοπισμός των θετικά διακείμενων πελατών δίνει την ευκαιρία στην επιχείρηση να προβεί σε ενέργειες που θα ενισχύσουν τους δεσμούς τους.
- Αξιολόγηση της στάσης του κοινού απέναντι σε ένα νέο προϊόν ή υπηρεσία: Το διοικητικό προσωπικό ενός οργανισμού διακατέχεται από αγωνιώδη αισθήματα κάθε φορά που στην αγορά διατίθεται ένα νέο προϊόν η υπηρεσία καθώς όσο ενδελεχώς και να έχει εκτιμηθεί και προβλεφθεί η στάση του κοινού απέναντί της η πραγματική αντίδραση του κοινού μπορεί να είναι απρόβλεπτή. Όταν η αντίδραση του κοινού προσδιοριστεί με σαφήνεια και ακρίβεια γρήγορα, ο οργανισμός έχει την ευκαιρία να κάνει – ενδεχομένως αναγκαίες – διορθωτικές ενέργειες στην φύση της καινοτομίας ή τον τρόπο προώθησης της ώστε να μεγιστοποιηθεί ο θετικός αντίκτυπός της στο κοινό στόχο.
- Σύγκριση με τον ανταγωνισμό: Συχνά σκοπός της ανάλυσης συναισθημάτων είναι ο προσδιορισμός του αντίκτυπου που έχουν οι αντίστοιχες υπηρεσίες – προϊόντα των ανταγωνιστικών οργανισμών. Μέσα από τις αναλύσεις αυτές προσδιορίζονται τόσο τα ισχυρά όσο και τα ασθενή στοιχεία έναντι του ανταγωνισμού ώστε τα πρώτα να ενισχυθούν και τα δεύτερα να εξαλειφθούν.
- Έγκαιρος εντοπισμός προβλημάτων: Οι καταναλωτές είναι οι ασφαλέστεροι κριτές των υπηρεσιών και των προϊόντων της επιχείρησης. Καμία εκτίμηση δεν είναι ακριβέστερη όσο ο βαθμός ικανοποίησης των πελατών. Προβλήματα των προϊόντων της επιχείρησης που επηρεάζουν την ικανοποίηση των πελατών εντοπίζονται έγκαιρα με την ανάλυση συναισθήματος και την εξόρυξη άποψης με συνέπεια να παρέχεται πολύτιμος χρόνος στην επιχείρηση να προβεί στην λύση τους.
- Λήψη καινοτόμων ιδεών για της εξέλιξη της επιχείρησης: Στο σύγχρονο ανταγωνιστικό επιχειρηματικό περιβάλλον το οποίο παράλληλα μαστίζεται από την παρατεταμένη οικονομική κρίση, η καινοτομία είναι ένα σημαντικό στοιχείο το οποίο μπορεί να παρέχει στον οργανισμό συγκριτικό πλεονέκτημα. Η εξόρυξη άποψης μπορεί να φανερώσει προσδοκίες του καταναλωτικού κοινού οι οποίες μπορούν να μετουσιωθούν σε καινοτόμες λύσεις από την επιχείρηση.

- Εντοπισμός των τάσεων του καταναλωτικού κοινού συνολικά και κατά κατηγορίες: Οι αναρτήσεις του κοινού του διαδικτύου φανερώνει τάσεις σε διάφορους τομείς της ανθρώπινης δραστηριότητας. Οι τάσεις αυτές μπορεί να διαφοροποιούνται για ανθρώπους με διαφορετικά χαρακτηριστικά και είναι δυνατόν να ανιχνευτούν με ανάλυσης συναισθήματος και εξόρυξη άποψης. Όταν η επιχείρηση έχει οριοθετήσει το κοινό στόχο και κατέχει την γνώση των τάσεων του, είναι πολύ πιθανό να σχεδιάσει και να διαθέσει εύστοχες υπηρεσίες και προϊόντα.

Κεφάλαιο 2

Μηχανική Μάθηση

Η ανάλυση συναισθήματος χρησιμοποιεί τεχνικές που μέσα από την επεξεργασία φυσικής γλώσσας αναζητείται να εντοπιστούν τα συναισθήματα που εκφράζουν τα κείμενα. Οι τεχνικές αυτές συχνά ενεργοποιούν μηχανισμούς μηχανικής μάθησης που με την σειρά τους αποτελούν σημαντικό μέρος των διαδικασιών εξόρυξης γνώσης. Με τις διαδικασίες αυτές οι μηχανές είναι σε θέση να επεξεργάζονται ακολουθίες συμβολοσειρών και να επιστρέφουν τα συναισθήματα που αυτές αντιπροσωπεύουν. Οι τεχνικές αυτές χρησιμοποιούν ισχυρούς αλγορίθμους προκειμένου να εκπαιδεύονται οι μηχανές στην εκτίμηση συναισθημάτων όταν η είσοδος που δέχονται είναι κείμενα σε φυσική γλώσσα. Συχνά έχουν υψηλές απαιτήσεις σε προπαρασκευαστικές ενέργειες και υπολογιστικούς πόρους. Οι κατηγορίες των διαδικασιών που περιλαμβάνονται στην κατηγορία αυτή ακολουθούν την διάκριση της μηχανικής μάθησης σε εποπτευόμενη και μη εποπτευόμενη ενώ η εποπτευόμενη μηχανική μάθηση κατηγοριοποιείται περαιτέρω με βάση το είδος των αλγορίθμων που χρησιμοποιούνται (Llombart, 2017).

2.1 Εισαγωγή

Η ραγδαία ανάπτυξη του διαδικτύου τα τελευταία χρόνια έχει φέρει κοντά στις εφαρμογές του πολλούς ανθρώπους με ποικίλα χαρακτηριστικά. Οι περισσότεροι άνθρωποι που αποκτούν πρόσβαση στο διαδίκτυο, χρησιμοποιούν εφαρμογές του όπου το περιεχόμενο τους διαμορφώνεται από τους ίδιους τους χρήστες τους. Παράλληλα οι κάθε είδους οργανισμοί εντείνουν τις προσπάθειες τους στο να προσφέρουν στο κοινό τους όσο το δυνατόν περισσότερες διαδικτυακές υπηρεσίες. Αποτέλεσμα αυτών είναι να υπάρχουν διαθέσιμα στο διαδίκτυο μεγάλα σύνολα δεδομένων. Ενώ παλαιότερα το βασικό πρόβλημα για την προσέγγιση της γνώσης ήταν η διαθεσιμότητα πληροφοριών, τα τελευταία χρόνια το πρόβλημα έγκειται στην αποδοτική εκμετάλλευση μεγάλου όγκου διαθέσιμων πληροφοριών

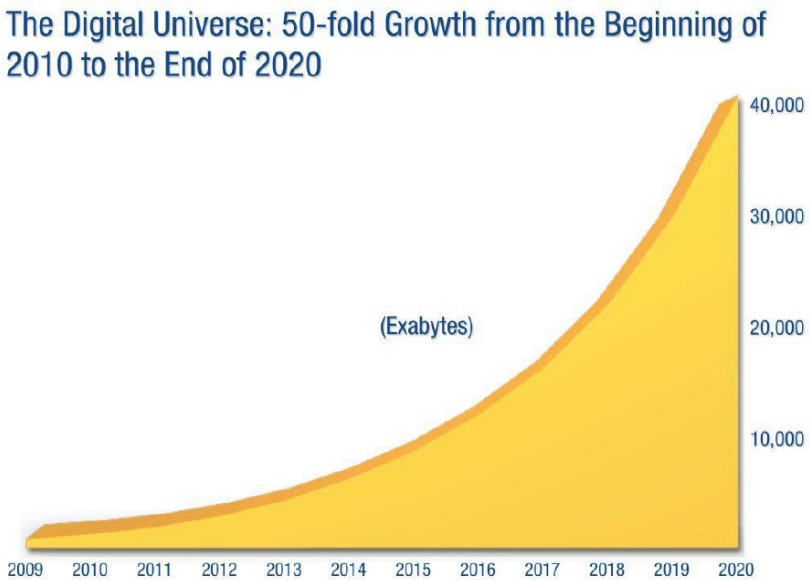
και κυρίως η διαλογή των ωφέλιμων από τα σύνολα αυτά. (Melinat, et al., 2014) Τα διαθέσιμα δεδομένα αξιολογούνται ως προς την εκμεταλλευσιμότητα τους ως εξής:

- Αν είναι ακριβή
- Αν έχουν ληφθεί έγκαιρα διατηρώντας την όποια αξία τους
- Αν είναι έτσι δομημένα που να εξυπηρετούν τον σκοπό συλλογής τους.
- Αν το πλαίσιο στο οποίο υπάρχουν είναι ικανό να αποδίδει σε αυτά το σωστό νόημα κατά την επεξεργασία τους.
- Αν είναι ικανά να περιορίζουν την αβεβαιότητα πάνω στο ζήτημα που με κάποιον τρόπο περιγράφουν.

Οι σύγχρονες τεχνολογίες πληροφορικής και τηλεπικοινωνιών εξυπηρετούν την διακίνηση και επεξεργασία μεγάλου όγκου δεδομένων. Αυτό προέκυψε σαν απαίτηση από την μεγάλη αύξηση στην διαθεσιμότητα της πληροφορίας. Χαρακτηριστικά αποτυπώνεται σε έρευνες ότι το μέγεθος του παγκόσμιου ιστού διπλασιάζεται κάθε 53 ημέρες από την έναρξη της έκρηξης τεχνολογίας Διαδικτύου (Nielsen, 1995) και ότι θα συνεχίσει να αυξάνεται κατά 40000 Exabyte ημερησίως τουλάχιστον μέχρι το 2020. Παράλληλα εκτιμάται ότι μόνο το 1/3 των διαθεσίμων διαδικτυακών δεδομένων θα είναι ωφέλιμα για επεξεργασία. (Gantz & Reinsel, 2012).

ΣΧΗΜΑ 2-1

Η ανάπτυξη του περιεχομένου του διαδικτύου (http://www.techgeeze.com/2016/03/digital-universe-2020.html)



This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Η λύση στο πρόβλημα που έχει ανακύψει είναι η εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων. Πρόκειται για τεχνικές και μεθοδολογίες κατά τις οποίες γίνεται επεξεργασία δεδομένων με τρόπο τέτοιο ώστε να ανιχνεύονται αν οι μεταξύ τους συσχετίσεις είναι ικανές να αναδείξουν μεταξύ τους συσχετίσεις. Οι συσχετίσεις που ενδεχομένως να προκύψουν αποτελούν τον οδηγό για την απόκτηση της γνώσης. Η μηχανική μάθηση συχνά αποτελεί μέρος της διαδικασίας της εξόρυξης δεδομένων.

2.2 Η μηχανική μάθηση ως μέρος των διαδικασιών εξόρυξης γνώσης

Ο βασικότερος παράγοντας ανάπτυξης της προσέγγισης της εξόρυξης γνώσης ήταν η υπερδιαθεσιμότητα πληροφοριών λόγω της τεχνολογικής προόδου και κυρίως του διαδικτύου. Η ανάγκη διαχείρισης της υπερπληθώρας δεδομένων δημιούργησε την ανάγκη για αποδοτικές μεθοδολογίες παραγωγής πληροφοριών από μεγάλα σύνολα – συχνά αδόμητων – δεδομένων. Λύση στο ζήτημα αυτό δίνει η εξόρυξη δεδομένων. Η εξόρυξη δεδομένων χρησιμοποιεί τεχνικές από διάφορους τομείς επιστήμης, που σχετίζονται με την ανάκτηση και εκμετάλλευση δεδομένων (Han & Kamber, 2006), όπως:

- Βάσεις Δεδομένων: Οι βάσεις δεδομένων παρέχουν μια εσωτερική αναπαράσταση του φυσικού κόσμου προσδιορίζοντας μια σταθερή σχέση μεταξύ των οργανωμένων αποθηκευμένων ψηφιακών δεδομένων με την πραγματικότητα που αντιλαμβάνεται ο άνθρωπος. Η προσέγγιση αυτή της αποθήκευσης των δεδομένων χαρακτηρίζεται από την οργανωμένη αποθήκευση δεδομένων, την ύπαρξη αποδοτικών δομών και λειτουργιών ελέγχου και ασφάλειας των δεδομένων καθώς και μηχανισμούς αποδοτικής ανάκτησης και εκμετάλλευσης των δεδομένων (Jeffery, 2012). Οι διαδικασίες της εξόρυξης δεδομένων χειρίζονται τις περισσότερες φορές περιεχόμενα μεγάλων βάσεων δεδομένων.

- Στατιστική: Η στατιστική χαρακτηρίζεται ως το σύνολο των μεθοδολογιών υπολογισμού μετρήσιμων μεγεθών που μπορεί να προσδιορίζουν τις ιδιότητες ενός συνόλου δεδομένων. Οι ιδιότητες αυτές προκύπτουν μετά από καθορισμένες διαδικασίες συλλογής, ταξινόμησης και επεξεργασίας των δεδομένων (Mehta, 2015). Κατά την εξόρυξη δεδομένων χρησιμοποιούνται στατιστικές μέθοδοι με σκοπό την κατασκευή προτύπων τα οποία θα χρησιμοποιηθούν ως σημεία αναφοράς για την παραγωγή γνώσης από μεγάλα σύνολα δεδομένων.

- Μηχανική μάθηση: Αποτελεί το σημαντικότερο μέρος στις περισσότερες διαδικασίες εξόρυξης γνώσης. Στο στάδιο αυτό επιχειρείται η κατασκευή του αναλυτικού μοντέλου το οποίο συγκρινόμενο με τα συλλεχθέντα δεδομένα θα οδηγήσει στην παραγωγή γνώσης. Η παραγωγή του μοντέλου βασίζεται σε σύνολα δεδομένων που τα οποία μπορεί να τύχουν επεξεργασίας με διαφόρους τρόπους. Στην συνέχεια του κεφαλαίου γίνεται εκτενέστερη αναφορά στις τεχνικές και τις μεθοδολογίες μηχανικής μάθησης (SAS, 2016).

- Πληροφορική: Οι διαδικασίες εξόρυξης γνώσης απαιτούν υψηλή επεξεργαστική ισχύ και την διαθεσιμότητα μεγάλου όγκου αποθηκευτικών πόρων. Αυτό συμβαίνει γιατί περιλαμβάνουν την επεξεργασία μεγάλου συνόλου δεδομένων. Για τον σκοπό αυτό θεωρείται απαραίτητα η διαθεσιμότητα υψηλών αποδόσεων εξοπλισμού πληροφορικής και η χρήση αποδοτικών αλγορίθμων κατά τις διεργασίες της εξόρυξης γνώσης (European Commission, 2013).

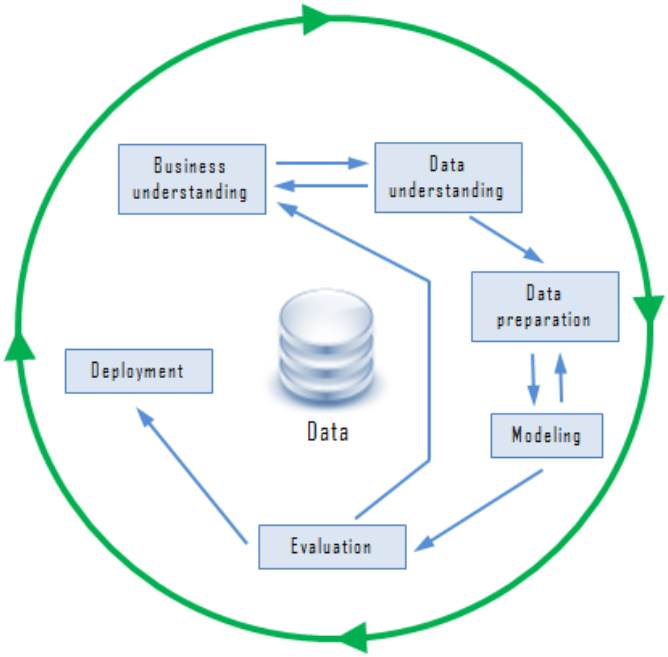
- Αναγνώριση προτύπων: Ο όρος αναγνώριση προτύπων αναφέρεται σε ένα σύνολο μηχανισμών για την αυτοματοποιημένη αναγνώριση ποικίλων αντικειμένων με σκοπό συνήθως την υποστήριξη διαδικασιών λήψης αποφάσεων. Σε γενικές γραμμές βασίζεται στην απόδοσης τα σε κάθε οντότητα μίας σειράς τιμών πεπερασμένου συνόλου παραμέτρων που μπορεί να τα χαρακτηρίζει. Οι τιμές αυτές καθορίζουν την κατάταξη της οντότητας σε μία ή περισσότερες κλάσεις. Η αντιστοίχηση της οντότητας σε μία ή περισσότερες κλάσεις γίνεται μετά από τον προσδιορισμό των κλάσεων αυτών καθώς και τα εύρη των τιμών των παραμέτρων που θα πρέπει να έχουν οι οντότητες προκειμένου να τοποθετηθούν σε κάθε μία από τις κλάσεις. Η δημιουργία αυτών των κανόνων γίνεται μετά από μία διαδικασία που καλείται εκπαίδευση και η οποία χρησιμοποιεί για τον σκοπό αυτό κατάλληλα σύνολα δεδομένων. (Polikar, 2006).

- Ανάλυση χωρικών δεδομένων: Η χωρική ανάλυση είναι η διαδικασία που με μετασχηματισμούς και επεξεργασίες γεωγραφικών δεδομένων παράγονται χρήσιμα συμπεράσματα και γνώση σε ζητήματα γεωγραφικής φύσεως. Η γνώση και τα συμπεράσματα αυτά μπορεί να επεκταθούν σε γεωγραφική ανάλυση η οποία έχει ως αντικείμενο την συσχέτιση διαμόρφωσης των χαρακτηριστικών οντοτήτων που μελετώνται σε σχέση με αντίστοιχα γεωγραφικά δεδομένα (Raju, 2008).

Η διαδικασία εξόρυξης δεδομένων αποτελείται τα εξής στάδια (Olson & Delen, 2008):

ΣΧΗΜΑ 2-2

Διαδικασία Εξόρυξης Δεδομένων (Πηγή: Data Mining Process, http://www.zentut.com/data-mining/data-mining-processes/)



• Προσδιορισμός των επιχειρησιακών απαιτήσεων

• Κατανόηση της δομής και της έννοιας των δεδομένων

• Προετοιμασία δεδομένων

• Μοντελοποίηση δεδομένων

• Κατασκευή Μοντέλου Αξιολόγησης

• Δεδομένα ανάπτυξης

Ολόκληρη η διαδικασία συνήθως δεν εξελίσσεται σε απολύτως διαδοχικά στάδια αλλά περιλαμβάνει αρκετές ανατροφοδοτήσεις.

2.2.1 Προσδιορισμος των επιχειρησιακων απαιτησεων

Στο πρώτο στάδιο της διαδικασίας προσδιορίζονται οι στόχοι που καλείται να εξυπηρετήσει η διαδικασία εξόρυξης γνώσης. Τα βήματα που ακολουθούνται στο στάδιο αυτό είναι τα εξής:

- Προσδιορισμός των επιχειρησιακών στόχων της όλης διαδικασίας από τα στελέχη του οργανισμού που είναι επιφορτισμένα με την λήψη αποφάσεων.

- Αξιολόγηση της τρέχουσας κατάστασης προκειμένου να εντοπιστούν τα χαρακτηριστικά της γνώσης που πρέπει να δημιουργηθεί.
- Προσδιορισμός της δομής και της μορφής των παραγώγων της διαδικασίας.
- Σχεδίαση και προσδιορισμός των σταδίων εξέλιξης της διαδικασίας.

2.2.2 Κατανοηση της δομης και της σημασια των δεδομενων

Στο στάδιο αυτό γίνεται η μελέτη των δεδομένων που πρόκειται να χρησιμοποιηθούν ώστε οι αναλυτές να κατανοήσουν την φύση τους, την δομή τους και την σημασία τους. Τα βήματα που περιλαμβάνει το στάδιο αυτό είναι τα εξής:

- Συλλογή δεδομένων: Εντοπίζονται οι διαθέσιμες πηγές δεδομένων, αξιολογούνται και επιλέγονται οι καταλληλότερες. Στην συνέχεια αναλύονται οι μορφές των δεδομένων και οι παράμετροι που τα προσδιορίζουν οι οποίοι πρέπει να είναι ανεξάρτητοι μεταξύ τους. Στο τέλος του βήματος αυτού συλλέγονται τα δεδομένα από τις πηγές τους. Οι πηγές αυτές μπορεί να είναι μελέτες και έρευνες χρησιμοποιώντας τα κατάλληλα εργαλεία (συλλογή στατιστικών στοιχείων, ερωτηματολόγια, πειραματικές μετρήσεις).
- Περιγραφή δεδομένων: Καθορίζεται η μορφή των δεδομένων που πρόκειται να χρησιμοποιηθούν, οι οποίες μπορεί να είναι:
    - Ποσοτικά δεδομένα
    - Ποιοτικά δεδομένα
    - Δεδομένα διακριτών τιμών
    - Δεδομένα συνεχών τιμών
    - Ονομαστικά δεδομένα
    - Σειριακά δεδομένα
- Διερεύνηση και αξιολόγηση δεδομένων: Στο βήμα αυτό εξετάζεται η καταλληλόλητα των δεδομένων να χρησιμοποιηθούν στην διαδικασία εξόρυξης γνώσης.

2.2.3 Προετοιμασια των δεδομενων

Τα δεδομένα που συλλέχθηκαν μπορεί να βρίσκονται διαφορετικές μορφές (κείμενο, αριθμοί, ήχος, βίντεο κλπ.), να περιέχουν θόρυβο ή αντικείμενα που είναι εφικτό να συσχετιστούν με τον κύριο όγκο του συνόλου των δεδομένων. Οι καταστάσεις αυτές μπορεί να οφείλονται σε σφάλματα ή σε ανεπιθύμητα συμβάντα κατά την συλλογή τους. Σε κάθε περίπτωση είναι αναγκαία η προσαρμογή των δεδομένων ώστε το σύνολο που θα προκύψει να είναι ικανό να τύχει κατάλληλης επεξεργασίας κατά την εξέλιξη της διαδικασίας της εξόρυξης γνώσης. Η προσαρμογή αυτή βασίζεται στο φιλτράρισμα των δεδομένων ώστε να απομακρυνθεί ο θόρυβος και τα μη κατάλληλα δεδομένα από το ωφέλιμο σύνολο αλλά και στην μετατροπή τους σε μορφές που να μπορούν να επεξεργαστούν με αυτόματο τρόπο ηλεκτρονικοί υπολογιστές.

Τα δεδομένα που χρησιμοποιούνται σε διαδικασίες εξόρυξης γνώσης μπορεί να είναι στις ακόλουθες μορφές:

- Πραγματικοί αριθμοί.
- Ακέραιοι αριθμοί
- Δυαδικά δεδομένα (true/false).
- Κατηγορίες που ορίζονται από εύρος ή σύνολο τιμών
- Ημερομηνία
- Συμβολοσειρές ή κείμενα.

Μετά την προσαρμογή των δεδομένων, αυτά υφίστανται προ επεξεργασία προκειμένου να μπορούν αποδοτικά να χρησιμοποιηθούν σαν είσοδοι από αλγορίθμους μηχανικής μάθησης. Συνηθέστερα πραγματοποιούνται οι παρακάτω διεργασίες:

- Συσσωμάτωση ή εξομάλυνση με την χρήση στατιστικών μέτρων όπως οι ακραίες τιμές, ο μέσος όρος, ο διάμεσος, η μέση τιμή.
- Φιλτράρισμα ακραίων τιμών με την χρήση διαγραμμάτων διασποράς.
- Ανάλυση παλινδρόμησης
- Ανάλυση συμπλέγματος
- Δέντρα αποφάσεων
- Ιεραρχική ανάλυση
- Προσαρμογή αριθμητική κλίμακας

2.2.4 ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η δημιουργία του μοντέλου εξόρυξης γνώσης είναι το βασικότερο στάδιο της διαδικασίας. Για τον σκοπό αυτό χρησιμοποιείται κατάλληλο λογισμικό για την ανάλυση των δεδομένων και την ανακάλυψη κανόνων σύνδεσης μεταξύ τους. Η προβολή των αποτελεσμάτων της ανάλυσης μπορεί να έχει την μορφή κειμένου, πινάκων, γραφημάτων ή συνδυασμό τους. Συνήθως τα σύνολα δεδομένων διαιρούνται σε υποσύνολα εκπαίδευσης και αξιολόγησης. Το πρώτο είναι το μεγαλύτερο σε όγκο και χρησιμοποιείται για την ανάπτυξη του μοντέλου ενώ το άλλο χρησιμοποιείται για τη δοκιμή του μοντέλου που δημιουργήθηκε ως προς το αν παράγει αξιόπιστη γνώση. Μετά από μία συνεχή διαδικασία δημιουργίας και αξιολόγησης το μοντέλο δεδομένων βελτιώνεται συνεχώς μέχρι να προκύψει το πιο ακριβές.

Οι τεχνικές που χρησιμοποιούνται για την κατασκευή μοντέλων είναι:

- Συσχετίσεις: Πρόκειται για την αναζήτηση της σχέσης μεταξύ ενός συγκεκριμένου χαρακτηριστικού των δεδομένων με ένα ή περισσότερα άλλα χαρακτηριστικά του ίδιου συνόλου δεδομένων. Η σχέση που προκύπτει χρησιμοποιείται για παραγωγή προτύπων πρόβλεψης.

- Ταξινόμηση: Οι μέθοδοι αυτοί εξετάζουν τις τιμές των χαρακτηριστικών των δεδομένων και με βάση αυτές τα κατατάσσουν σε κατηγορίες. Η κατάταξη αυτή γίνεται με την χρήση καταλλήλων συναρτήσεων που εφαρμόζονται σε ορισμένα ή όλα τα χαρακτηριστικά και η έξοδος του υποδεικνύει την κατάλληλη κατηγορία ταξινόμησης. Για την επιλογή των χαρακτηριστικών και την ανάπτυξη των συναρτήσεων χρησιμοποιείται ένα υποσύνολο των δεδομένων.

- Ομαδοποίηση: Αποτελεί κατηγορία μεθόδων που ομοιάζει με την ταξινόμησή αλλά διαφέρει ως προς τον σκοπό που είναι η κατάταξη μη ομαδοποιημένων δεδομένων χωρίς τη χρήση υποσυνόλων εκπαίδευσης αλλά με αυτόματες διαδικασίες.

- Πρόβλεψη: Με τις μεθόδους αυτές ανιχνεύεται η σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Οι μέθοδοι αυτοί συνήθως δημιουργούν μια καμπύλης παλινδρόμηση η οποία αποτελεί την βάση για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής από την τιμή των αντίστοιχων ανεξάρτητων.

- Ανάλυση διαδοχικών προτύπων: Οι μέθοδοι αυτής της κατηγορίας αναζητούν πρότυπα που χαρακτηρίζουν τα σύνολα δεδομένων σε μια συγκεκριμένη χρονική περίοδο ή συγκυρία. Τα πρότυπα που παράγονται χρησιμοποιούνται για την ασφαλή πρόβλεψη μελλοντικών τάσεων.

2.2.5 Αξιολογηση

Η αξιολόγηση ένα στάδιο που εκτελείται επαναλαμβανόμενα μετά την παραγωγή και την χρήση του μοντέλου. Συχνά – ακόμα αν ο αρχικός στόχος της εξόρυξης γνώσης έχει επιτευχθεί – χρειάζεται να αξιολογηθεί το μοντέλο καθώς οι συνθήκες του περιβάλλοντος μπορεί να έχουν μεταβληθεί μετά την παραγωγή του. Αυτό μπορεί να οδηγήσει σε μεταβολή των επιχειρησιακών στόχων ή αλλαγή της συμπεριφοράς των δεδομένων που παράγουν οι πηγές που χρησιμοποιήθηκαν. Για να ανιχνευθούν τέτοιου είδους καταστάσεις χρησιμοποιούνται κατάλληλα μέσα στατιστική απεικόνισης και μηχανισμοί τεχνητής νοημοσύνης. Στην φάση της αξιολόγησης συνεργάζονται αναλυτές δεδομένων και επιχειρησιακοί παράγοντες λήψης αποφάσεων. Προκειμένου από κοινού να προσδιορίζουν τις τρέχουσες απαιτήσεις, την εξέλιξη της διαδικασίας και τον βαθμό ικανοποίησης των επιχειρησιακών στόχων.

Κατά την αξιολόγηση αποτιμάται η επιχειρησιακή αξία των μοντέλων που δημιουργούνται και επιλέγεται ο τρόπος παρουσίασης των αποτελεσμάτων της ώστε να είναι αρκούντως κατανοητά από τους δρώντες της διακαίς.

2.2.6 Αναπτυξη

Το τελευταίο στάδιο της διαδικασίας είναι η χρήση των μοντέλων που προέκυψαν για την επίτευξη των επιχειρησιακών στόχων που έχουν τεθεί. Ο τρόπος που θα γίνει αυτό εξαρτάται από την φύση των στόχων, το είδος των δεδομένων και την εννοιολογική βάση της διαδικασίας.

2.3 Μηχανική Μάθηση

Ένας από τους τρόπους που οι ζώντες οργανισμοί αποκτούν γνώση είναι η εμπειρία. Οι ζώντες οργανισμοί με βάση τις παρελθούσες εμπειρίες παράγουν πρότυπα καταστάσεων τα οποία αναγνωρίζουν αν κατά κάποιο τρόπο επαναληφθούν στο μέλλον. Η αναγνώριση των προτύπων αυτών είναι ικανή να προσαρμόσει ανάλογα την συμπεριφορά τους. Η σύνδεση αιτιατού και αποτελέσματος με βάση την εμπειρία θα πρέπει να είναι δοκιμασμένη και τεκμηριωμένη ορθώς ώστε να μην οδηγεί στην παραγωγή εσφαλμένων προτύπων. Χαρακτηριστικό παράδειγμα εσφαλμένων προτύπων αποτελεί η πίστη των ανθρώπων σε διάφορες δεισιδαιμονίες και συμπτώσεις σαν παγιωμένες καταστάσεις αιτίων. Η ισχύς των προτύπων που δημιουργούνται είναι ευθέως ανάλογη με την ύπαρξη προηγούμενης εκμεταλλεύσιμης γνώσης.

Μηχανική μάθηση είναι τομέας της πληροφορικής που αναδύθηκε από την μελέτη της αναγνώρισης προτύπων. Πρόκειται για σύνολο προσεγγίσεων που καθιστά τους ηλεκτρονικούς υπολογιστές ικανούς να μαθαίνουν και να προσαρμόζουν την εξέλιξη των διαδικασιών που εκτελούν χωρίς να έχουν ρητά προγραμματιστεί προς μία συγκεκριμένη ροή αλλά με βάση τα δεδομένα εισόδου και προβλέψεις που κάνουν με βάση αυτά. Η δημιουργία των προτύπων που καθορίζει την

συμπεριφορά του ηλεκτρονικού υπολογιστή με βάση τα δεδομένα εισόδου ονομάζεται εκπαίδευση και για τον σκοπό αυτό χρησιμοποιούνται εξειδικευμένα σύνολα δεδομένων.

Η μηχανική μάθηση χρησιμοποιείται σε τρεις κυρίως περιπτώσεις:

- Εργασίες που πραγματοποιούνται από ανθρώπους (ή γενικότερα ζώντες οργανισμούς) αλλά είναι ανέφικτο να προγραμματιστεί ηλεκτρονικός υπολογιστής να τις εκτελέσει. Τέτοιες εργασίες είναι η οδήγηση, η αναγνώριση ομιλίας, η αναγνώριση εικόνας.

- Ανάλυση μεγάλων συνόλων σύνθετων δεδομένων όπου πρότυπα εφαρμόζονται σε αυτά προκειμένου να αναδυθεί η γνώση που εμπεριέχεται.

- Εργασίες των οποίων ο τρόπος διεκπεραίωσης εξαρτάται σε μεγάλο βαθμό από το περιβάλλον στο οποίο εξελίσσονται.

Οι μεθοδολογίες της μηχανικής μάθησης διακρίνονται σε τρεις κατηγορίες:

- **Μεθοδολογίες επιτηρούμενης μάθησης:** Πρόκειται για μεθοδολογίες κατά τις οποίες παρέχονται τόσο δεδομένα εισόδου όσο και επιθυμητά δεδομένα εξόδου. Τα δεδομένα εισόδου και εξόδου επισημαίνονται για ταξινόμηση για να παρέχουν μια βάση μάθησης για μελλοντική επεξεργασία δεδομένων. Τα δεδομένα εκπαίδευσης περιλαμβάνουν ένα σύνολο δειγμάτων ως ζεύγη εισόδου και επιθυμητού αποτελέσματος. Με βάση τα ζεύγη αυτά παράγεται το μοντέλο το οποίο θα αποτελέσει τον οδηγό για τις μελλοντικές εκτιμήσεις. Τα μοντέλα που παράγονται με τις μεθόδους της κατηγορίας αυτής χαρακτηρίζονται γενικά από υψηλή ακρίβεια. Το βασικό μειονέκτημα των μεθόδων αυτών είναι ότι απαιτούν τον προσδιορισμό προτύπων για κάθε κατηγορία δεδομένων. Οι μεθοδολογίες επιτηρούμενης μάθησης διακρίνονται περαιτέρω σε:

    - Μεθοδολογίες ταξινόμησης: Οι μεταβλητές εξόδου αποτελούν διακριτές τιμές που αντιστοιχούν σε κατηγορίες ταξινόμησης.

    - Μεθοδολογίες παλινδρόμησης: Οι μεταβλητές εξόδου αντιστοιχούν σε συνεχείς τιμές.

- **Μεθοδολογίες μάθησης χωρίς επιτήρηση:** Σε αντίθεση με τις μεθόδους επιτηρούμενης μάθησης, στις περιπτώσεις των μεθόδων αυτών δεν χρησιμοποιούνται ζεύγη εισόδου – επιθυμητής εξόδου για την κατασκευή των μοντέλων. Η μη επιτηρούμενη μηχανική μάθηση αποσκοπεί στην αποκάλυψη προηγουμένως άγνωστων προτύπων στα δεδομένα. Ελλείψει των προαναφερόμενων ζευγών, ο στόχος της μάθησης χωρίς επιτήρηση είναι να μοντελοποιήσει την υποκείμενη δομή ή τη διανομή στα δεδομένα προκειμένου να κατασκευάσει γνώση για αυτά. Οι μεθοδολογίες αυτές διακρίνονται σε:

    - Μεθοδολογίες συσταδοποίησης: Οι μεταβλητές εξόδου είναι διακριτές τιμές που αντιστοιχούν σε κατηγορίες οι οποίες όμως δεν είναι γνωστές εκ των προτέρων.

    - Μεθοδολογίες συσχέτισης: Με τις μεθοδολογίες αυτές αναζητούνται συσχετίσεις μεταξύ των χαρακτηριστικών των δεδομένων.

- **Ενισχυτική μάθηση:** Πρόκειται για μεθοδολογίες κατά τις οποίες επιδιώκεται η δημιουργία μοντέλων συμπεριφοράς που να οδηγούν σε έναν στόχο χωρίς όμως κατά την εξέλιξη των διαδικασιών να προσδιορίζεται η απόσταση από αυτόν.

Μία άλλη διάκριση των μεθοδολογιών μηχανικής μάθησης έχει να κάνει με το κατά πόσο ο εκτελών την διαδικασία μπορεί να επεμβαίνει στην ποιότητα των δεδομένων που χρησιμοποιούνται για την παραγωγή των μοντέλων. Έτσι διακρίνονται στις:

- Παθητικές μεθοδολογίες όπου τα δεδομένα που χρησιμοποιούνται για την παραγωγή του μοντέλου προέρχονται από απλή παρατήρηση του περιβάλλοντος.
- Ενεργητικές μέθοδοι όπου ο ενεργών την διαδικασία επεμβαίνει στην διαμόρφωση των δεδομένων που θα χρησιμοποιηθούν για την παραγωγή των μοντέλων.

2.4 Ανάλυση συναισθήματος με την χρήση τεχνικών μηχανικής μάθησης

Η επεξεργασία με μηχανική μάθηση των κειμένων που καταγράφονται σε εφαρμογές κοινωνικής δικτύωσης είναι μία διαδικασία η οποία με είσοδο τα κείμενα αυτά παράγουν στην έξοδο τους το συναίσθημα που εκφράζεται μέσω αυτών. Χαρακτηριστικό των μεθόδων αυτών είναι ότι βασίζονται σε ηλεκτρονικούς υπολογιστές οι οποίοι αφού εκπαιδευτούν να αναγνωρίζουν τα συναισθήματα που κρύβονται στα κείμενα, είναι πλέον σε θέση να τα εντοπίζουν. Η επεξεργασία με μηχανική μάθηση περιλαμβάνει τρία βασικά βήματα:

Αρχικοποίηση: Πρόκειται για το στάδιο εκείνο στο οποίο πραγματοποιούνται όλες εκείνες οι προπαρασκευαστικές ενέργειες που είναι απαραίτητες προκειμένου να είναι εφικτή η εφαρμογή του αλγορίθμου μηχανικής μάθησης που έχει επιλεγεί να χρησιμοποιηθεί. Αρχικά επιλέγεται η πηγή των δεδομένων (κειμένων) που θα χρησιμοποιηθούν για την δημιουργία και την αξιολόγηση του μοντέλου ταξινόμησης που πρόκειται να αναπτυχθεί. Το εντοπισμό τους ακολουθεί η διαδικασία λήψης τους. Το καθαρό κείμενο που συλλέγεται υφίσταται προ επεξεργασία προκειμένου να προκύψει μία μορφή κειμένου που να είναι επεξεργάσιμη αποδοτικά από ηλεκτρονικό υπολογιστή. Στην συνέχεια εντοπίζονται και επιλέγονται εκείνα τα χαρακτηριστικά των κειμένων τα οποία είναι ικανά να οδηγήσουν στην δημιουργία ενός ισχυρού και ακριβούς ταξινομητή του συναισθήματος που αναδύεται από αυτά.

Εκπαίδευση: Πρόκειται για το κυρίως μέρος της επεξεργασίας. Σε αυτό επιλέγεται ο αλγόριθμός ο οποίος θα χρησιμοποιήσει ένα σύνολο κειμένων τα οποία έχουν προηγουμένως υποστεί προ επεξεργασία και που θα οδηγήσει στην παραγωγή του μοντέλου ταξινόμησης. Ο αλγόριθμός αυτός εκπαιδεύει το μοντέλο που αναπτύσσεται ούτως ώστε να μπορεί να κατατάσσει τυχαία κείμενα σε συναισθήματα με βάση τα χαρακτηριστικά τους.

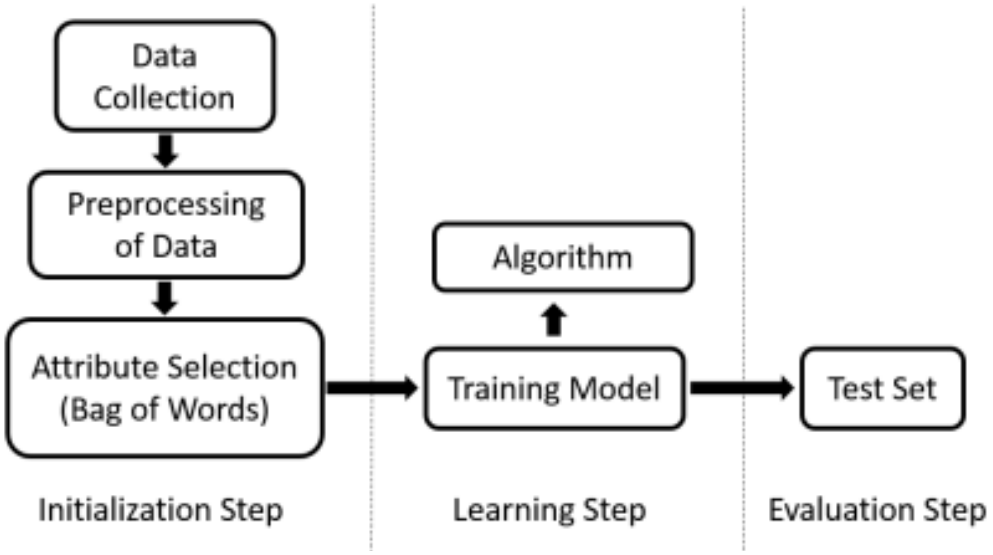
Αξιολόγηση: Πριν χρησιμοποιηθεί ένα μοντέλο ταξινόμησης θα πρέπει να αξιολογηθεί ως προς την αποτελεσματικότητα του. Για τον σκοπό αυτό επιλέγεται ένα σύνολο δεδομένων – συνήθως ξένο σε σχέση με αυτό που χρησιμοποιήθηκε για την ανάπτυξη του μοντέλου ταξινόμησης. Τα κείμενα τα οποία περιλαμβάνονται στο σύνολο αυτό είναι ήδη ταξινομημένα πέραν πάσης αμφιβολίας. Για το σύνολο αυτό υπολογίζεται το συναίσθημα τους με βάση τον δημιουργηθέντα ταξινομητή και συγκρίνεται με την πραγματική τους ταξινόμηση.

Στο επόμενο σχήμα παρουσιάζεται σχηματικά η διαδικασία της μηχανικής μάθησης για την

εξόρυξη συναισθήματος από κείμενα.

ΣΧΗΜΑ 2-3

Ανάλυση συναισθήματος με τεχνικές μηχανικής μάθησης
(http://article.sciencepublishinggroup.com/html/10.11648.j.ijdst.20160204.11.html)



2.4.1 Αρχικοποιηση

Συλλογή των δεδομένων

Πρόκειται για το στάδιο της αρχικοποίησης της διαδικασίας όπου εντοπίζονται οι πηγές άντλησης κατάλληλων δεδομένων. Κατάλληλα θεωρούνται τα δεδομένα όταν είναι αρκετά σε όγκο, είναι αντιπροσωπευτικά ως προς την ποιότητα τους (να προέρχονται από ποικιλία κατηγορίας ανθρώπων, η θεματολογία τους να συνάδει με τον σκοπό της έρευνας στην οποία θα χρησιμοποιηθούν και να έχουν σχετική ευρύτητα ως προς το ύφος που εκφράζουν). Στην συνέχεια γίνονται κτήμα των ανθρώπων που θα τα επεξεργαστούν με διάφορους τρόπους (πχ με την χρήση σχετικών προγραμματιστικών διεπαφών (Application Programming Interface – API) ή με την λήψη από το διαδίκτυο αρχείων εγγραφών σταθερού ή μεταβλητού μήκους). Η μορφή των συνόλων των δεδομένων θα πρέπει να είναι επεξεργάσιμη από ηλεκτρονικούς υπολογιστές. Αν αυτό δεν συμβαίνει τότε οι ερευνητές χρειάζεται να κάνουν τις απαραίτητες μεταβολές σε αυτά.

Προ επεξεργασία

Η προ επεξεργασία των κειμένων αφορά τον μετασχηματισμό τους σε τέτοια μορφή που να είναι εφικτή η επεξεργασία τους από ηλεκτρονικό υπολογιστή. Εκτός αυτού η τελική μορφή των κειμένων μετά την προ επεξεργασία θα είναι τέτοια που να επιτρέπει την αποτελεσματική εφαρμογή όλων των λειτουργιών της επεξεργασίας που πρόκειται να χρησιμοποιηθεί. Η προ επεξεργασία διαιρείται σε τέσσερις επί μέρους φάσεις οι οποίες αναλύονται στις επόμενες παραγράφους.

Βασικές λειτουργίες και καθαρισμός

Η πρώτη φάση της προ επεξεργασίας έχει σαν στόχο την απομάκρυνση στοιχείων, λέξεων και εκφράσεων που δεν επηρεάζουν το νόημα του κειμένου ως προς το συναίσθημα που αναδύεται. Στοιχεία τα οποία είναι περιττά ως προς την απόδοση νοήματος στο κείμενο είναι οι σύνδεσμοι σε υπερκείμενα και σύμβολα κωδικοποίησης λειτουργιών σε δίκτυα κοινωνικής δικτύωσης (πχ @ ή #). Επιπροσθέτως συνεχόμενα κενά ή αλλαγές γραμμής απλά επιβαρύνουν τις διαδικασίες της επεξεργασίας του κειμένου. Λέξεις που επίσης δεν προσφέρουν κατά κανέναν τρόπο στον προσδιορισμό του συναισθήματος που κρύβεται στο κείμενο είναι λέξεις οι οποίες είναι για κάποιο λόγο ανορθόγραφα. Επιπλέον χρειάζεται να αφαιρεθούν τα σημεία στίξεως εκτός από όσα έχουν συμμετοχή στο νόημα του κειμένου (όπως η απόστροφος για την αγγλική γλώσσα). Όπου εντοπίζονται διαδοχικά φωνήεντα περισσότερα των δύο τότε κανονικοποιούνται. Οι αλληλουχίες χαρακτήρων που κωδικοποιούν εικόνες που αντιστοιχούν σε συναισθήματα (emoticons) αντικαθίστανται από κατάλληλα tags όπως επίσης και διαδοχικές συλλαβές που αντιστοιχούν σε γέλωτα (πχ χαχαχαχαχα).

Όλες οι λειτουργίες σε αυτήν την φάση εκτελούνται με σκοπό να προκύψει ένα ομοιόμορφο κείμενο που να βασίζεται σε ένα συγκεκριμένο σύνολο προδιαγραφών. Αυτό είναι σημαντικό επειδή κατά τη διαδικασία της ταξινόμησης, τα χαρακτηριστικά επιλέγονται μόνο όταν υπερβαίνουν μια συγκεκριμένη συχνότητα στο σύνολο δεδομένων. Επομένως, μετά τις βασικές εργασίες προ επεξεργασίας, έχοντας διαφορετικές λέξεις γραμμένες στο κείμενο αλλά με τον ίδιο τρόπο βοηθά τον ηλεκτρονικό υπολογιστή να πραγματοποιήσει την ταξινόμηση ορθά.

Χειρισμός των emoticons

Ανάλογα με την φύση του κειμένου κατατάσσονται τα emoticons σε όσο το δυνατόν λιγότερες κατηγορίες. Έτσι κατά τον εντοπισμό τους στο κείμενο αντικαθίστανται από κατάλληλα tags. Όταν το πλήθος αυτών περιορίζεται στο απολύτως απαραίτητο τότε η συναισθηματική ταξινόμηση του συνόλου του κειμένου απλοποιείται. Συνήθως τα emoticons κατατάσσονται σε δύο κατηγορίες: σε αυτά που αντιστοιχούν σε ευχάριστη και σε αυτά που αντιστοιχούν σε δυσάρεστη διάθεση. Με τον τρόπο αυτό αυξάνεται η βαρύτητα τους στον εντοπισμό του συναισθήματος που αναδύεται από το κείμενο γεγονός που οδηγεί σε μεγαλύτερη ακρίβεια κατά την επεξεργασία.

Χειρισμός της άρνησης

Οι εκφράσεις που δηλώνουν άρνηση είναι ικανές κατά την επεξεργασία από ηλεκτρονικούς υπολογιστές να οδηγήσουν σε εσφαλμένη απόδοση του νοήματος σε κείμενα. Μια λέξη άρνησης μπορεί να επηρεάσει τον τόνο όλων των λέξεων γύρω από αυτήν και η αγνόηση τους είναι μία από τις κύριες αιτίες της εσφαλμένης ταξινόμησης του συνολικού κειμένου. Σε αυτή τη φάση, όλες οι εκφράσεις που δηλώνουν άρνηση (δεν μπορούν, δεν, δεν είναι, ποτέ κλπ.) αντικαθίστανται με την λέξη που αντιστοιχεί στην ρητή άρνηση ("όχι" στην Ελληνική γλώσσα ή "no" στην Αγγλική). Αυτή η τεχνική επιτρέπει την εμπλουτισμό του μοντέλου ταξινομητή με πολλές εκφράσεις αρνητικών n-grams που διαφορετικά θα αποκλείονταν λόγω της χαμηλής συχνότητάς τους.

Ορθογραφικό λεξικό

Στην φάση αυτή χρησιμοποιούνται κατάλληλα ορθογραφικά λεξικά τα οποία από την μία εντοπίζουν τις λέξεις οι οποίες είναι ορθογραφημένα παρούσες στο κείμενο και από την άλλη εντοπίζουν τις λέξεις που είναι παρούσες ανορθόγραφα. Οι τελευταίες μετά από τον εντοπισμό τους αντικαθίστανται από τις αντίστοιχες ορθές. Η διεργασία αυτή εμπλουτίζει την ανάπτυξη του μοντέλου του ταξινομητή με λέξεις οι οποίες σε διαφορετική περίπτωση θα είχαν αφαιρεθεί.

Εντοπισμός της ρίζας των λέξεων

Στην φάση αυτή επιχειρείται η κατηγοριοποίηση των λέξεων με βάση την ρίζα από την οποία και προέρχονται. Με τον τρόπο αυτό μειώνεται η εντροπία των λέξεων που περιλαμβάνονται σε ένα κείμενο και ταυτόχρονα αυξάνεται και το επίπεδο της σημασίας που προσφέρουν στον χαρακτηρισμό του κειμένου. Χρησιμοποιώντας διαφορετικούς όρους η διεργασία που πραγματοποιείται στην φάση αυτή αντιμετωπίζει με όμοιο τρόπο μία έννοια σε οποιοδήποτε μέρος του λόγου που μπορεί να εντοπιστεί στο κείμενο.

Χειρισμός των τερματικών λέξεων

Οι τερματικές λέξεις συνήθως είναι επιρρήματα ή επίθετα τα οποία δεν προσθέτουν κάποιο νέο νόημα στο κείμενο αλλά υπάρχουν στον λόγο για την ενίσχυσή ή την απόδοση έμφασης σε αυτό. Για τον λόγο αυτό είναι απαραίτητο να μην περιλαμβάνονται κατά την επεξεργασία του κειμένου για την δημιουργία του μοντέλου ταξινομητή. Έτσι κατά την φάση αυτή λαμβάνεται μέριμνα για την απομάκρυνση των λέξεων αυτών.

Επιλογή των χαρακτηριστικών

Το τελευταίο βήμα της αρχικοποίησης αφορά την επιλογή των χαρακτηριστικών των κειμένων τα οποία θεωρούνται κατάλληλα να χρησιμοποιηθούν για την ανάπτυξη του μοντέλου ταξινόμησης. Τα χαρακτηριστικά αυτά θα πρέπει να είναι ικανά να παράγουν αρκετή πληροφορία που να παίζει όσο το δυνατό σημαντικότερο ρόλο στην διαμόρφωση του μοντέλου και εν τέλει στην κατηγοριοποίηση των κειμένων (Angiani, et al., 2015).

2.4.2 Εκπαίδευση

Στο στάδιο της εκπαίδευσης για την παραγωγή του μοντέλου ταξινόμησης αρχικά εξάγονται τα χαρακτηριστικά που θα χρησιμοποιηθούν από κάθε κείμενο και από αυτά προκύπτει το συναίσθημα που αναδύεται. Ο τρόπος που αναδεικνύεται το ύφος του κειμένου εξαρτάται από τον αλγόριθμο που θα χρησιμοποιηθεί για τον σκοπό αυτό. Τα κείμενα που χρησιμοποιούνται για την εκπαίδευση περιλαμβάνουν εγγραφές που έχουν ήδη χαρακτηριστεί ως προς το ύφος τους και δεν υπάρχει αμφιβολία για την ορθότητα της κατάταξής τους. Επιδιώκεται να υπάρχει ικανός αριθμός για την αντιπροσώπευση κάθε κατηγορίας συναισθήματος που είναι επιθυμητό να εντοπιστεί. Βασική διεργασία που λαμβάνει χώρα στο στάδιο αυτό της επεξεργασίας είναι η εξαγωγή των χαρακτηριστικών η οποία γενικά μπορεί να γίνει με τους τρόπους που παρουσιάζονται στις επόμενες παραγράφους.

Εξαγωγή χαρακτηριστικών

Προκειμένου να είναι εφικτή η χρήση των τεχνικών μηχανικής μάθησης σε σύνολα δεδομένων που περιλαμβάνουν ελεύθερα κείμενα θα πρέπει να χρησιμοποιηθεί ένας μηχανισμός κωδικοποίησης τους ώστε να μετασχηματιστούν σε εισόδους κατάλληλες για τους αλγορίθμους που θα χρησιμοποιηθούν. Οι τεχνικές αναπαράστασης φυσικής γλώσσας διακρίνονται σε δύο κύριες κατηγορίες:

- Μη κατανεμημένες αναπαραστάσεις: Τα δείγματα αποτελούν σύνολο παραμέτρων που αντιστοιχούν σε έννοιες.
- Κατανεμημένες αναπαραστάσεις: Τα δείγματα αναπαρίστανται με διανύσματα κάθε παράμετρος της οποίας υποδηλώνει τον βαθμό συσχέτισης του δείγματος με μία έννοια.

Μη Κατανεμημένες αναπαραστάσεις

Αρχικά επιλέγεται ένας τρόπος διαχωρισμού των δειγμάτων σε παραμέτρους τέτοιων ώστε οι τιμές τους να είναι ικανές να τα χαρακτηρίσουν. Οι παράμετροι αυτοί τις περισσότερες φορές έχουν να κάνουν με τμήματα που χαρακτηρίζουν γενικότερα ένα κείμενο όπως λέξεις, εκφράσεις, μέρη του λόγου. Στην συνέχεια σε κάθε δείγμα αντιστοιχίζεται ένα διάνυσμα το οποίο σχετίζεται με την ύπαρξη της κάθε παραμέτρου στο κείμενο. Ανάλογα την απόφαση που θα πάρει ο αναλυτής μπορεί να καταγράφεται στο διάνυσμα η ύπαρξη της παραμέτρου στο κείμενο, η συχνότητα εμφάνισης της ή άλλα πολυπλοκότερα μέτρα.

Έστω τα κείμενα Κ1 και Κ2:

Κ1: «Το γεγονός αυτό ήταν θετική εξέλιξη»

Κ2: «Ήταν κακή παρένθεση η τελευταία του πράξη»

Στον παρακάτω πίνακα παρατίθενται οι λέξεις που χρησιμοποιούνται στις δύο προτάσεις και η μη κατανεμημένη αναπαράσταση τους.

ΠΙΝΑΚΑΣ 2-1

Μη κατανεμημένη αναπαράσταση λέξεων

	το	γεγονός	αυτό	ήταν	θετική	εξέλιξη	κακή	παρένθεση	η	τελευταία	του	πράξη
1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	1

Bag of words

Το μοντέλο συνόλου δεδομένων bag-of-words είναι ένας τρόπος κωδικοποίησης δεδομένων κειμένου προκειμένου να είναι εφικτή η μοντελοποίηση τους για χρήση σε λειτουργίες που βασίζονται σε αλγόριθμους μηχανικής μάθησης. Το μοντέλο bag-of-words είναι απλό στην εφαρμογή του και έχει παρουσιάζει αρκετά ικανοποιητικά αποτελέσματα σε εργασίες μοντελοποίησης γλωσσών και ταξινόμησης εγγράφων.

Το βασικό πρόβλημα στην επεξεργασία κειμένων με μηχανισμούς μηχανικής μάθησης είναι ότι δεν παρουσιάζουν σταθερή δομή και δεν μπορούν να αποτελέσουν αυτούσια είσοδο στους σχετικούς αλγορίθμους. Οι αλγόριθμοι μηχανικής μάθησης είναι σχεδιασμένοι να λειτουργούν με σαφώς καθορισμένες και ακριβείς εισόδους. Κατά συνέπεια θα πρέπει τα κείμενα που πρόκειται να χρησιμοποιηθούν σε τέτοιες λειτουργίες να μετατραπούν σε δομημένα αριθμητικά δεδομένα. Στην επεξεργασία των γλωσσών, οι φορείς των περιεχομένων τους περιλαμβάνουν αριθμητικές τιμές σε κατάλληλες παραμέτρους (συνήθως με την μορφή διανυσμάτων) προκειμένου να κωδικοποιούνται οι διάφορες γλωσσικές ιδιότητες του κειμένου. Η διεργασία αυτή ονομάζεται εξαγωγή χαρακτηριστικών ή κωδικοποίηση χαρακτηριστικών. Μια δημοφιλής και απλή μέθοδος εξόρυξης χαρακτηριστικών από δεδομένα κειμένου ονομάζεται μοντέλο κειμένου συνόλου-λέξεων (bag of words). Ένα μοντέλο bag-of-words, ή BoW για συντομία, είναι ένας τρόπος εξαγωγής χαρακτηριστικών από κείμενο για χρήση σε αλγορίθμους μηχανικής μάθησης. Η προσέγγιση αυτή είναι πολύ απλή και ευέλικτη και μπορεί να χρησιμοποιηθεί με πολλούς τρόπους ανάλογα με το είδος των κειμένων αλλά και τον σκοπό της λειτουργίας μηχανικής μάθησης.

Μια bag of words είναι μια αναπαράσταση ενός κειμένου που περιγράφει την εμφάνιση λέξεων μέσα σε ένα έγγραφο. Τα στοιχεία στα οποία βασίζεται είναι:

- Ένα λεξιλόγιο γνωστών λέξεων.

- Μέτρο της παρουσίας γνωστών λέξεων.

Χαρακτηριστικό της μεθόδου είναι ότι αφορά μόνο την συμμετοχή γνωστών λέξεων και όχι με το συντακτικό, την σημασιολογία ή την σειρά των λέξεων.

Η απλούστερη επεξεργασία κειμένων με την μέθοδο του bag of words γίνεται με την καταγραφή όλων των λέξεων που χρησιμοποιούνται σε ένα σύνολο κειμένων. Έστω N το πλήθος των λέξεων. Σε κάθε μία από αυτές αντιστοιχίζεται ένας ακέραιος αριθμός από 1 ως N. Κάθε κείμενο κωδικοποιείται με ένα διάνυσμα το οποίο αποτελείται από μία σειρά δυαδικών ψηφίων (0 ή 1) ανάλογα αν η λέξη που αντιστοιχεί στην θέση του διανύσματος εμφανίζεται έστω και μία φορά στο κείμενο. Άλλες παραλλαγές της μεθόδου αυτής καταγράφουν στις θέσεις του διανύσματος το πλήθος των εμφανίσεων της αντίστοιχης λέξης στο κείμενο, ή την πυκνότητα εμφάνισης. Σε κάθε περίπτωση τα διανύσματα που προκύπτουν είναι αρκετά αραιά. Η διαχείριση των αραιών διανυσμάτων γενικά απαιτεί σπατάλη πόρων και έτσι αναζητούνται τρόποι για τον περιορισμό των μεγεθών των διανυσμάτων που προκύπτουν. Για τον σκοπό μπορεί να χρησιμοποιηθούν διάφορες τεχνικές όπως:

- Αγνόηση του τρόπου γραφής των γραμμάτων της λέξης (μικρά ή κεφαλαία γράμματα).
- Παράλειψη των σημείων στίξης όταν κρίνεται ότι δεν παρέχουν πληροφορία ως προς την σημασιολογία του κειμένου.
- Παράλειψη σημείων του λόγου που χρησιμοποιούνται για να συνδέσουν νοήματα και δεν εμπεριέχουν κάποιου είδους ωφέλιμη πληροφορία.
- Αναγνώριση λέξεων με κακή ορθογραφία και αντιστοίχιση του με λέξεις ορθογραφημένες που ήδη έχουν καταχωρηθεί στο λεξιλόγιο.
- Εντοπισμός λέξεων γραμμένων σε διαφορετικές μορφές
- Ομαδοποίηση συνώνυμων λέξεων (Brownlee, 2017).

N-Grams

Με την μέθοδο αυτή εξετάζεται η ύπαρξη N διαδοχικών λέξεων σε ένα κείμενο. Αποτελεί μία εξέλιξη του bag-of-words. Συνήθως η προσέγγιση των bigrams (για N = 2) ή των trigrams (για N =3) παράγουν πιο αξιόπιστα αποτελέσματα σε σχέση με την προσέγγιση του bag-of-words (Text-Analytics, 2014).

ΣΧΗΜΑ 2-4

Παράδειγμα n-grams (http://www.grroups.com/blog/naive-bayes-and-text-classification-introduction-and-theory)

- unigram (1-gram):

a	swimmer	likes	swimming	thus	he	swims

- bigram (2-gram):

a swimmer	swimmer likes	likes swimming	swimming thus	...

- trigram (3-gram):

a swimmer likes	swimmer likes swimming	likes swimming thus	...

Skip-Gram

Με την μέθοδο αυτή ορίζεται ένα μέγεθος παραθύρου έστω W και ένα πλήθος λέξεων που παραλείπονται από το παράθυρο αυτό των λέξεων – έστω D. Το λεξιλόγιο των κειμένων περιλαμβάνει όλους τους συνδυασμούς που μπορεί να προκύπτουν από το μοτίβο αυτό (Minnaar, 2015).

Syntactic dependencies

Οι συντακτικές εξαρτήσεις παίζουν βασικό ρόλο στη διαδικασία της σημασιολογικής ερμηνείας. Αυτές ορίζονται ως επιλεκτικές συναρτήσεις στις λέξεις. Ανάμεσα στις ιδιότητές τους, ιδιαίτερη σημασία έχει η ικανότητά τους προσθέτουν στις λέξεις νοήματα και να τα εξελίσσουν βαθμιαία στην ροή των κειμένων.

Μια συντακτική εξάρτηση ορίζεται σημασιολογικά ως μια δυαδική λειτουργία που παίρνει ως παραμέτρους τις δηλώσεις των δύο σχετικών λέξεων (της βασικής και της εξαρτωμένης) και δίνει ως αποτέλεσμα μια σχετική διάταξη των εννοιών τους η οποία είναι κομβική για την απόδοση νοήματος στο κείμενο στο οποίο εντάσσονται. Οι κύριες ιδιότητες των συντακτικών εξαρτήσεων είναι οι ακόλουθες:

- Οι σχέσεις μεταξύ μεμονωμένων λέξεων. Οι λέξεις και οι εξαρτήσεις είναι οι δομικές μονάδες της Σύνταξης. Η συντακτική ανάλυση δεν χρησιμοποιεί πλέον φράσεις για να περιγράψει τη διάρθρωση της πρότασης, διότι ό, τι πρέπει να ειπωθεί, μπορεί να ειπωθεί με όρους εξαρτήσεων ανάμεσα σε μεμονωμένες λέξεις.
- Οι σχέσεις μεταξύ των λέξεων είναι ασύμμετρες εκτός από αυτές που δρουν συντονιστικά. Σε μια σχέση μία λέξη είναι πάντοτε εξαρτώμενη από την άλλη, που ονομάζεται βασική (head).
- Οι εξαρτήσεις εξατομικεύονται από επισημασμένους συνδέσμους που επιβάλλουν ορισμένους γλωσσικούς όρους στις συνδεδεμένες λέξεις.

Μια εξάρτηση συνδέεται στον σημασιολογικό χώρο με τη λειτουργία της ανάθεσης ενός επιχειρήματος σε ένα ρόλο λεξικής λειτουργίας. Μια λέξη μιας εξάρτησης θα θεωρηθεί ως η

λεξική λειτουργία και η άλλη σαν το επιχείρημά της. Η αντιστοίχιση επιτρέπεται μόνο και μόνο εάν το όρισμα ικανοποιεί ορισμένες σημασιολογικές συνθήκες που απαιτούνται από τη λειτουργία. Μέσα σε μια εξάρτηση τόσο το βασικό μέλος της σχέσης όσο και το εξαρτώμενο επιβάλλει περιορισμούς το ένα στο άλλο. Αυτό έχει ως συνέπεια μια εξάρτηση να έχει τουλάχιστον δύο συμπληρωματικές επιλεκτικές λειτουργίες, κάθε μία από τις οποίες επιβάλλει τους δικούς της περιορισμούς επιλογής που χρησιμοποιούνται στην κωδικοποίηση της γλωσσικής αποσαφήνισης.

Οι έννοιες που αποκτούν σημαντικό ρόλο στον προσδιορισμό των συντακτικών συσχετίσεων είναι:

- Λέξεις: Οι διαφορετικές κατηγορίες λέξεων (γνωστά και ως μέρη του λόγου), θεωρούνται σύνολα ιδιοτήτων και χρησιμοποιούνται για την περιγραφή του τρόπου με τον οποίο συνδυάζονται οι λέξεις με δυαδικές εξαρτήσεις (όταν οι δύο λέξεις θεωρούνται ότι είναι συμβατές σημασιολογικά). Δύο λέξεις είναι συμβατές και επομένως μπορούν να συνδυαστούν από μια εξάρτηση μόνο εάν τα σύνολα που ανήκουν μοιράζονται μερικές ιδιότητες. Συνήθως – χωρίς αυτό να αποτελεί κανόνα – τα ρήματα και τα ουσιαστικά αποτελούν τον φορέα της σύνθεσης. Σε κάθε περίπτωση επιλέγονται λεκτικές μονάδες που ανήκουν σε κατηγορίες με πλούσια σύνολα ιδιοτήτων. Αυτές οι ιδιότητες θα συσχετιστούν με κατάλληλες συντακτικές εξαρτήσεις. Είναι προφανές ότι οι σημασιολογικές διαφορές μεταξύ δύο λέξεων που ανήκουν σε διαφορετικές συντακτικές κατηγορίες δεν μπορούν να εξηγηθούν εξ ολοκλήρου απαριθμώντας τις ιδιότητες που δεν μοιράζονται. Επιπλέον οι συμμετέχουσες λέξεις οργανώνουν τις ιδιότητες που υποδηλώνουν με διαφορετικούς τρόπους πράγμα που περιπλέκει τον χειρισμό τους κατά την ανίχνευση των συντακτικών εξαρτήσεων.

- Εξάρτηση: Μια εξάρτηση είναι μια δυαδική συνάρτηση η οποία παίρνει ως ορίσματα δύο σύνολα ιδιοτήτων και δίνει ως αποτέλεσμα ένα πιο περιορισμένο σύνολο. Τα δύο σύνολα αντιστοιχούν στις δηλώσεις των δύο λέξεων που σχετίζονται με την εξάρτηση. Το σύνολο που προκύπτει είναι η διασταύρωση των δύο συνόλων εισόδου, υπό την προϋπόθεση ότι δεν υπάρχουν συγκεκριμένοι λεξικοί περιορισμοί που εμποδίζουν τη διασταύρωση. Το αποτέλεσμα μιας συνάρτησης εξάρτησης είναι στη συνέχεια από προεπιλογή ένα μία συνδυαστική ερμηνεία.

- Λεξικό – Συντακτικά πρότυπα: Οι λέξεις και οι μεταξύ τους εξαρτήσεις, προκειμένου να αποκαλύπτουν σημασιολογικές ερμηνείες θα πρέπει να συγκρίνονται με σαφώς καθορισμένα πρότυπα. Τα πρότυπα αυτά ονομάζονται λεξικοσυντακτικά και αυτά που χρησιμοποιούνται στην ανακάλυψη των συντακτικών εξαρτήσεων είναι αυτά που εμπεριέχουν δύο λεκτικές κατηγορίες.

Με βάση τα παραπάνω η αναζήτηση των συντακτικών εξαρτήσεων εξετάζει ζεύγη λεκτικών μονάδων προκειμένου να εντοπίσει εξαρτήσεις οι οποίες ταιριάζουν με κάποια συγκεκριμένα λεξικοσυντακτικά πρότυπα. Όταν η αναζήτηση καταλήγει σε θετικό αποτέλεσμα τότε η συντακτική εξάρτηση που πιστοποιείται καθορίζει την ανάδειξη

σημασιολογίας (Gamallo, 2008).

Κατανεμημένες αναπαραστάσεις

Πρόκειται για αναπαραστάσεις κατά τις οποίες οι μονάδες κειμένου (λέξεις, προτάσεις, παράγραφοι κοκ) αντιστοιχίζονται σε πυκνά διανύσματα. Οι παράμετροι των διανυσμάτων αντιστοιχούν σε συγκεκριμένες έννοιες και οι τιμές που λαμβάνουν οι παράμετροι προβάλλουν τον βαθμό κατά τον οποίο συσχετίζεται η μονάδα κειμένου με την αντίστοιχη έννοια.

Στο παράδειγμα των μη κατανεμημένων αναπαραστάσεων θεωρούμε τις έννοιες ΘΕΤΙΚΟΣ, ΑΡΝΗΤΙΚΟΣ, ΟΥΔΕΤΕΡΟΣ, ΥΠΟΚΕΙΜΕΝΙΚΟΣ, ΑΝΤΙΚΕΙΜΕΝΙΚΟΣ. Οι λέξεις στην περίπτωση αυτή αξιολογούνται ως προς την συνάφεια τους με τους όρους αυτούς όπως φαίνεται στον παρακάτω πίνακα (υπό κλίμακα από 0 έως 1).

ΠΙΝΑΚΑΣ 2-2

Αξιολόγηση λέξεων ως προς την συνάφεια τους με συγκεκριμένους όρους

	ΘΕΤΙΚΟ	ΑΡΝΗΤΙΚΟ	ΟΥΔΕΤΕΡΟ	ΥΠΟΚΕΙΜΕΝΙΚΟ	ΑΝΤΙΚΕΙΜΕΝΙΚΟ
ΤΟ	0,3	0,3	0,4	0,5	0,5
ΓΕΓΟΝΟΣ	0,3	0,3	0,4	0	1
ΑΥΤΟ	0,3	0,3	0,4	0,5	0,5
ΗΤΑΝ	0,3	0,3	0,4	0,5	0,5
ΘΕΤΙΚΗ	1	0	0	0,7	0,3
ΕΞΕΛΙΞΗ	0,5	0,2	0,3	0,5	0,5
ΚΑΚΗ	0	0	1	0,7	0,3
ΠΑΡΕΝΘΕΣΗ	0,3	0,3	0,4	0,7	0,3
Η	0,3	0,3	0,4	0,7	0,3
ΤΕΛΕΥΤΑΙΑ	0,2	0,5	0,3	0,3	0,7
ΤΟΥ	0,3	0,3	0,4	0,5	0,5
ΠΡΑΞΗ	0,3	0,3	0,4	0,3	0,7

Οι βασικότερες τεχνικές που χρησιμοποιούνται για την ανάπτυξη τέτοιου είδους αναπαραστάσεων είναι της Μείωσης διαστάσεων: Πρόκειται για τεχνικές κατά τις οποίες αναζητούνται τρόποι με τους οποίους επιτυγχάνεται η αναπαράσταση των ιδίων οντοτήτων που παριστάνονται με την βοήθεια διανυσμάτων διάστασης N, με διανύσματα διάστασης M, με $M < N$ (Shalev-Shwartz & Ben-David, 2014).

2.4.3 Αξιολογηση

Από τα προηγούμενα στάδια επεξεργασίας προκύπτει ένα μοντέλο ταξινόμησης το οποίο είναι ικανό να δέχεται σαν είσοδο ένα κείμενο και να επιστρέφει στην έξοδο του εκτίμηση για το συναίσθημα το οποίο αντιστοιχεί σε αυτό. Για τον σκοπό αυτό επιλέγεται ένα δεύτερο σύνολο κειμένων. Για κάθε στοιχείο του συνόλου αυτού είναι γνωστό το συναίσθημα που αντιστοιχίζεται. Κατά την εφαρμογή του μοντέλου, που έχει παραχθεί από την μέχρι εκείνη την στιγμή διαδικασία, προκύπτει στην έξοδο μία νέα εκτίμηση για το αναδυόμενο συναίσθημα. Η εκτίμηση αυτή συγκρίνεται με την συναίσθημα που πραγματικά αντιπροσωπεύει το κείμενο και έτσι η εκτίμηση χαρακτηρίζεται ως ορθή ή εσφαλμένη. Ο λόγος των ορθών προβλέψεων προς το σύνολο τους εκφράζει τον βαθμό επιτυχίας του

μοντέλου. Στις περιπτώσεις κατά τις οποίες ο βαθμός επιτυχίας του μοντέλου δεν είναι ικανοποιητικός γίνεται προσπάθεια να εντοπιστούν οι αιτίες της αστοχίας ώστε να γίνουν οι κατάλληλες αλλαγές που θα οδηγήσουν σε καλύτερης ποιότητας μοντέλο ταξινόμησης. Οι αλλαγές μπορεί να αφορούν στην διαφοροποίηση των ενεργειών κατά την προ επεξεργασία των δεδομένων εκπαίδευσης, την επιλογή διαφορετικών δεδομένων εκπαίδευσης ή και την επιλογή χρήσης διαφορετικών αλγορίθμων παραγωγής του μοντέλου ταξινόμησης. Η φάσης της αξιολόγησης επαναλαμβάνεται σε κάθε μεταβολή της διαδικασίας.

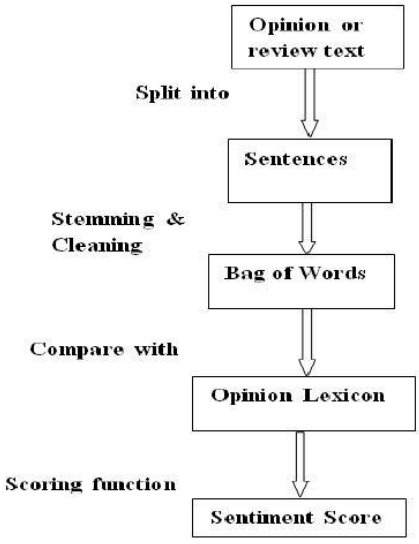
Κεφάλαιο 3

Τεχνικές με την Χρήση Λεξικών

Οι διαδικασίες που βασίζονται σε λεξικά δεν περιλαμβάνουν προπαρασκευαστικά στάδια για κάθε προσπάθεια ανάλυσης συναισθήματος. Πρόκειται για διαδικασίες που βασίζονται στην ανάπτυξη πεπερασμένης έκτασης λεξικών αντιστοίχησης λέξεων και εκφράσεων με συγκεκριμένα συναισθήματα. Η προσέγγιση με βάση το λεξικό περιλαμβάνει τον υπολογισμό του προσανατολισμού ενός εγγράφου από τον σημασιολογικό προσανατολισμό των λέξεων ή των φράσεων στο έγγραφο. Κάθε προσέγγιση που χρησιμοποιεί λεξικά επεκτείνει τις λειτουργίες της με διαφορετικό τρόπο εκμεταλλευόμενη τις πληροφορίες που μπορεί να αντλεί από αυτά προκειμένου εν τέλει να συγκρίνει το περιεχόμενο του προς εξέταση κειμένου με το περιεχόμενο των λεξικών και το αποτέλεσμα να είναι μία βαθμονόμηση του αναδυόμενου συναισθήματος σε μία δεδομένη αριθμητική κλίμακα. Στην επόμενη εικόνα φαίνεται η προσέγγιση εντοπισμού του συναισθηματικού προσανατολισμού των κειμένων με την χρήση λεξικών.

ΣΧΗΜΑ 3-1

Γενική διαδικασία εξόρυξης συναισθήματος με βάση λεξικά
(https://www.researchgate.net/figure/Lexicon-based-sentiment-analysis-approach_fig3_272463313)



Τα λεξικά των λέξεων χρησιμοποιούνται σχολιασμένα με τις λέξεις σημασιολογικό προσανατολισμό, ή πολικότητα. Τα λεξικά αυτά συνήθως παράγονται άπαξ και μπορούν να χρησιμοποιούνται με την ίδια απόδοση σε πολλαπλές αναλύσεις συναισθημάτων από κείμενα. Λεξικά για προσεγγίσεις που βασίζονται σε λεξικό μπορούν να δημιουργηθούν «με το χέρι» ή αυτόματα, χρησιμοποιώντας ρίζες λέξεων σπόρων για να επεκταθεί ο κατάλογος λέξεων που αναζητούνται σε ένα έγγραφο. Σημαντικό ρόλο στην αναζήτηση του συναισθηματικού

προσανατολισμού των εγγραφών – χρησιμοποιώντας τεχνικές που βασίζονται σε λεξικά – έχουν τα επίθετα. Κατά βάση ένας κατάλογος επιθέτων μεταγλωττίζεται και σε κάθε ένα από αυτά αποδίδεται μία βαθμολογία μίας κλίμακας που αντιστοιχεί στην ένταση των συναισθημάτων. Τα επίθετα που περιλαμβάνονται σε ένα κείμενο εντοπίζονται και εξάγεται η συνολική βαθμολογία των συναισθημάτων που αναδύονται από αυτό.
Η αποδοτικότητα των τεχνικών αυτών βασίζεται σε μεγάλο βαθμό στην πληρότητα και την ευστοχία των λεξικών αυτών. Οι αποδοτικότητα των τεχνικών που βασίζονται σε μηχανική μάθηση εξαρτάται από την ποιότητα των συνόλων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου πρόβλεψης. Αυτό θεωρητικά είναι και το βασικότερο κριτήριο που μπορεί να επηρεάσει τη επιλογή χρήσης της μίας ή της άλλης κατηγορίας τεχνικών για την εξόρυξη συναισθήματος από κείμενα. Οι βασισμένες σε λεξικά τεχνικές έχουν σταθερή απόδοση για κάθε πεδίο, θεματολογία, χρονικής συγκυρίας κείμενα που μπορεί να χρησιμοποιηθούν. Οι βασισμένες σε μηχανική μάθηση τεχνικές απαιτούν την χρήση συνόλου δεδομένων για εκπαίδευση που να παρουσιάζει χαρακτηριστικά που να είναι «κοντά» σε αυτά του συνόλου που πρόκειται να αξιολογηθεί. Οι τεχνικές που βασίζονται σε λεξικά μπορούν και αντιμετωπίζουν αποδοτικότερα και ειδικές καταστάσεις που επηρεάζονται από το γλωσσικό πλαίσιο που εντάσσονται τα κείμενα. Αντιμετωπίζονται με αυτές καλύτερα γλωσσικές ιδιαιτερότητες όπως η άρνηση ή εντατικοποίηση των εννοιών που αποδίδονται καθώς τα μοντέλα πρόβλεψης που παράγονται με μηχανική μάθηση δύσκολά μπορούν να εντοπίζουν τις διαφοροποιήσεις τέτοιών καταστάσεων στην χρήση ίδιων όρων (Taboada, et al., 2014). Η περαιτέρω διάκριση τους περιλαμβάνει δύο κατηγορίες, τις τεχνικές που εξετάζουν λέξεις και εκείνες που εξετάζουν φράσεις (Musto, et al., 2015).

3.1 Λεξικά

Η σημαντικότητα της ανάλυσης συναισθήματος που αναδύεται από κείμενα εντατικοποίησε τις σχετικές έρευνες. Συνέπεια αυτού ήταν να διατίθενται από το διαδίκτυο πηγές καταλόγων λέξεων και φράσεων που συνοδεύονται από τον προσδιορισμό τους ως προς τα συναισθήματα τα οποία μπορεί να αντιπροσωπεύουν. Ο τρόπος με τον οποίο δημιουργούνται τα εργαλεία αυτά ποικίλει:

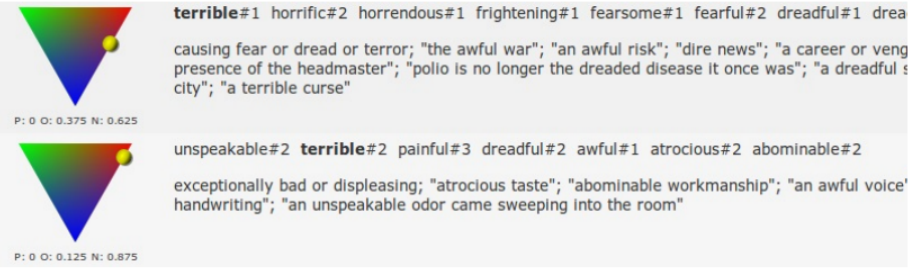
- Χειροκίνητα: Με τον τρόπο αυτό δημιουργούνται από ανθρώπους ρητές και σταθερές αντιστοιχίες μεταξύ λέξεων και εκφράσεων και συναισθημάτων.
- Corpus based: Με την τεχνική αυτή γίνεται διπλή – αμφίδρομη αντιστοίχηση μεταξύ όρων και αντικειμένων που αναφέρονται οι όροι αυτοί. Το συναίσθημα που εκλύεται από τους όρους δεν εξαρτάται αποκλειστικά από αυτούς αλλά και από τα αντίστοιχα αντικείμενα.
- Dictionary based: Οι όροι ομαδοποιούνται ανάλογα με την έννοια που περιγράφουν (πχ συνώνυμα) ή/και σχηματίζουν ιεραρχίες. Από τις δομές αυτές αναζητούνται οι λέξεις που εκφράζουν συναισθήματα και οι οποίες είναι ικανές να συμβάλουν στον συναισθηματικό προσδιορισμό των κειμένων. Οι όροι που εμπεριέχονται στα λεξικά εξετάζονται μεμονωμένα και όχι σε σχέση με το πλαίσιο στο οποίο εντάσσονται.

Κάποια από αυτά τα εργαλεία είναι τα εξής (Musto, et al., 2015):

- SentiWordNet: Το SentiWordNet είναι ένας λεξικογραφικό πόρος που σχεδιάστηκε για να υποστηρίζει εφαρμογές ανάλυσης συναισθημάτων. Παρέχει για κάθε λήμμα που περιέχει μία αντιστοιχία σε τρεις αριθμητικές κλίμακες για θετικότητα, αρνητικότητα και ουδετερότητα. Δεδομένου ότι παρέχει μια αναπαράσταση των συναισθημάτων με αριθμητικούς όρους κάθε όρος χαρακτηρίζεται από ένα διάνυσμα τριών παραμέτρων. Παράλληλα ο ίδιος όρος, ανάλογα με το πως χρησιμοποιείται σε ένα κείμενο μπορεί να χαρακτηρίζεται από διαφορετικό διάνυσμά όπως φαίνεται στην παρακάτω εικόνα. Για να προσδιοριστεί με ακρίβεια το συναίσθημα του όρου μέσα στο κείμενο μπορεί να υποστηριχθεί η διαδικασία από μηχανισμούς αποσαφήνισης του συναισθήματος που αναδύεται από αυτόν που συνήθως χρησιμοποιούν και τα συμφραζόμενα του όρου.

ΣΧΗΜΑ 3-2

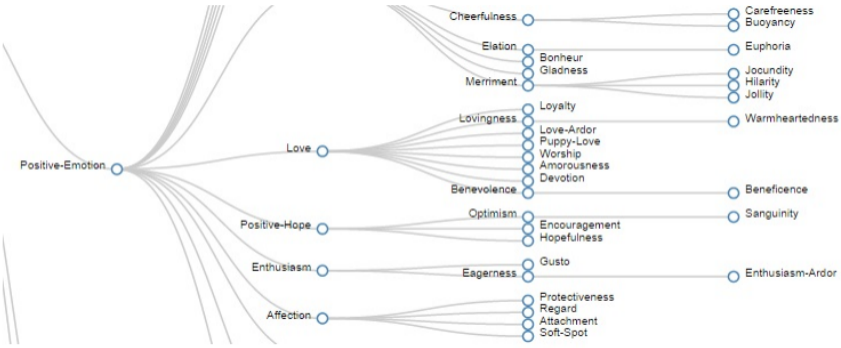
Διαφοροποιήσεις της λεξικογραφικής απεικόνισης του ιδίου όρου (https://www.researchgate.net/figure/SentiWordNet-visualization-of-the-opinion-related-properties-of-the_fig3_228399908)



- WordNet-Affect: Το WordNet-Affect είναι ένας γλωσσικός πόρος για μια λεξικογραφική αντιστοίχιση της συναισθηματικής γνώσης. Πρόκειται για μια επέκταση του WordNet που χαρακτηρίζει συναισθηματικές συσχετίσεις με συναισθηματικές έννοιες που ορίζονται ως A-labels – ετικέτες που χαρακτηρίζουν το συναίσθημα βάση δεδομένης κατηγοριοποίησης. Η αναζήτηση εκτελείται με βάση μια ιεραρχία ανεξάρτητης από το πεδίο που αναφέρονται τα κείμενα συναισθηματικών ετικετών που κατασκευάζονται αυτόματα χρησιμοποιώντας σχέσεις WordNet.

ΣΧΗΜΑ 3-3

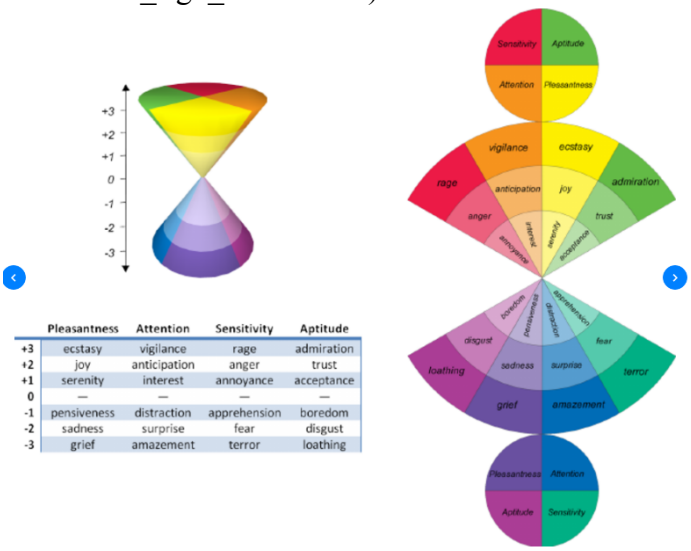
Παράδειγμα ιεραρχίας WordNet Affect (https://www.researchgate.net/figure/A-fragment-of-WordNet-Affect-hierarchy_fig2_287871786)



- MPQA: Πρόκειται για ένα εργαλείο με χαμηλή πολυπλοκότητα. Αποτελείται από μία λίστα όρων (περίπου 10000) οι οποίοι προέρχονται από ποικίλες πηγές και αναφέρονται επίσης σε ευρεία γκάμα θεματικών πεδίων. Κάθε όρος συνοδεύεται από μία ετικέτα (POS-tag) η οποία αντιστοιχεί στον συναισθηματικό χαρακτηρισμό του όρου (θετικός, αρνητικός, ουδέτερος) αλλά και στην ένταση του (ήπια, έντονη).
- SenticNet: Το SenticNet είναι μια λεξικολογική πηγή για ανάλυση συναισθημάτων σε επίπεδο ιδεών. Επικεντρώνεται στο “Sentic Computing”, ένα εκτεταμένο πολύ επιστημονικό κείμενο που αναφέρεται στην ανάλυση συναισθήματος. Σε αντίθεση με τα εργαλεία που έχουν ήδη αναφερθεί το SenticNet είναι ικανό να συσχετίζει την πολικότητα και τις συναισθηματικές πληροφορίες και σε πολύπλοκες έννοιες. Παρέχει βαθμολογίες συναισθήματος στο διάστημα [-3,3] για 14.000 περίπου έννοιες. Το συναίσθημα που ορίζει κάθε όρος αποτελεί σύνθεση της έντασης δεκαέξι βασικών συναισθημάτων, που ορίζονται σε ένα μοντέλο που ονομάζεται Κλεψύδρα Συναισθημάτων και η οποία απεικονίζεται στην επόμενη εικόνα.

ΣΧΗΜΑ 3-4

Η κλεψύδρα των συναισθημάτων (https://www.researchgate.net/figure/The-Hourglass-of-Emotions_fig2_228419623)



	Pleasantness	Attention	Sensitivity	Aptitude
+3	ecstasy	vigilance	rage	admiration
+2	joy	anticipation	anger	trust
+1	serenity	interest	annoyance	acceptance
0	—	—	—	—
-1	pensiveness	distraction	apprehension	boredom
-2	sadness	surprise	fear	disgust
-3	grief	amazement	terror	loathing

- WordStat Sentiment Dictionary: Το λεξικό WordStat Sentiment σχεδιάστηκε συνδυάζοντας αρνητικές και θετικές λέξεις από το λεξικό του Harvard IV, το λεξικό Regresive Imagery και το λεξικό λογοτεχνίας Pennebaker. Στην συνέχεια εφαρμόστηκαν κατάλληλοι αλγόριθμοι για την επέκταση της λίστας των όρων που περιέχει με συνώνυμα ή διαφοροποιημένες μορφές των αρχικά περιλαμβανομένων. Το αποτέλεσμα πλέον ήταν μία λίστα με περίπου 10000 αρνητικά και 5000 αρνητικά πρότυπα λέξεων. Στην πραγματικότητα, το συναίσθημα δεν υπολογίζεται με βάση αποκλειστικά αυτές τις δύο λίστες λέξεων και μοτίβων λέξεων, αλλά με δύο σύνολα κανόνων που χρησιμοποιούν αρνητικές εκφράσεις που μπορεί να προηγηθούν αυτών των λέξεων. Για παράδειγμα το αρνητικό συναίσθημα υπολογίζεται χρησιμοποιώντας τους ακόλουθους δύο κανόνες:
    - Αρνητικές λέξεις που δεν προηγούνται από μια άρνηση (όχι, όχι ποτέ) μέσα σε τέσσερις όρους στην ίδια πρόταση.
    - Θετικές λέξεις που προηγούνται από μια άρνηση μέσα σε τέσσερις όρους στην ίδια πρόταση.

  Το θετικό συναίσθημα μετράται με παρόμοιο τρόπο αναζητώντας θετικές λέξεις που δεν προηγούνται από μια άρνηση καθώς και αρνητικούς όρους μετά από μια άρνηση (Loughran & McDonald, 2011).

3.2 Διαδικασίες

3.2.1 Κατηγορίες τεχνικών

Οι λέξεις που εκφράζουν γνώμη είναι εκείνες που βασικά χρησιμοποιούνται σε διαδικασίες ταξινόμησης συναισθημάτων. Οι λέξεις θετικής γνώμης χρησιμοποιούνται για να εκφράσουν κάποιες επιθυμητές καταστάσεις, ενώ αρνητικές λέξεις γνώμης χρησιμοποιούνται για να εκφράσουν κάποιες ανεπιθύμητες καταστάσεις. Υπάρχουν επίσης φράσεις και ιδιώματα εκφράζοντα γνώμη. Όλα αυτά μαζί ονομάζονται λεξιλόγιο γνώμης. Υπάρχουν τρεις κύριες προσεγγίσεις για τη συγκέντρωση ή τη συλλογή της λίστας λέξεων γνώμης. Η χειρωνακτική προσέγγιση είναι πολύ χρονοβόρα και δεν είναι αποδοτική η αποκλειστική χρήση της σε τέτοιου είδους διαδικασίες. Συνήθως συνδυάζεται με τις άλλες δύο αυτοματοποιημένες προσεγγίσεις ως τελικό έλεγχο για να αποφευχθούν τα λάθη που προέκυψαν από τις τελευταίες. Γενικά οι διαδικασίες εντοπισμού του συναισθηματικού προσανατολισμού με την χρήση λεξικών έχουν ως βάση δύο παραδοχές:

- Κάθε μεμονωμένη λέξη εμπεριέχει μία ένδειξη του αν είναι θετική, αρνητική ή ουδέτερη
- Ο συναισθηματικός προσανατολισμός μπορεί να ποσοτικοποιηθεί με αριθμητικούς όρους.

3.2.2 Προσεγγισεις σε επιπεδο λεξεων

Κατά τις προσεγγίσεις του είδους αυτού ένα μικρό σύνολο λέξεων γνώμης συλλέγεται με ανθρώπινη παρέμβαση (χειρωνακτικά) με τον προσανατολισμό τους να είναι γνωστός. Στη συνέχεια, αυτό το σύνολο αναπτύσσεται με την αναζήτηση στα γνωστά σχετικά σύνολα που διατίθενται μέσω του διαδικτύου τα συνώνυμά τους και τα αντίθετα τους. Οι λέξεις που βρέθηκαν πρόσφατα προστίθενται στη λίστα των ριζών. Η διαδικασία επαναλαμβάνεται με την προσθήκη νέων λέξεων εκτός και αν δεν εντοπιστούν νέες λέξεις. Μετά τις επαναλήψεις αυτές επεμβαίνει άνθρωπος ο οποίος απομακρύνει ή διορθώνει εσφαλμένες καταχωρήσεις. Η προσέγγιση αυτή έχει ένα σημαντικό μειονέκτημα: την αδυναμία σε ορισμένες περιπτώσεις εντοπισμού λέξεων γνώμης που η εμβέλεια τους να αφορά έναν τομέα της ανθρώπινης δραστηριότητας. Η αποτελεσματικότητα της προσέγγισης αυτής βελτιώνεται αρκετά όταν εφαρμόζονται τεχνικές που λαμβάνουν υπ’ όψη και το «περιβάλλον» της κάθε λέξεις σε ένα κείμενο ώστε με την εφαρμογή κατάλληλων κανόνων να προκύπτει η γνώμη που μπορεί να εκφράζει.

3.2.3 ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΕ ΕΠΙΠΕΔΟ ΣΥΝΟΛΩΝ ΛΕΞΕΩΝ (CORPUS BASED)

Οι προσεγγίσεις που βασίζονται σε σύνολα λέξεων συμβάλουν στην άμβλυνση του προβλήματος της εύρεσης λέξεων γνώμης με συγκεκριμένους προσανατολισμούς σε ένα συγκεκριμένο πλαίσιο. Οι μέθοδοι τους εξαρτώνται από τα συντακτικά σχέδια ή τα μοτίβα που παρατηρούνται με μια λίστα ριζών λέξεων γνώμης που χρησιμοποιούνται για τον εντοπισμό άλλων λέξεων γνώμης από μία μεγάλη δεξαμενή. Μία από αυτές τις μεθόδους παρουσιάστηκε από τους Hatzivassiloglou και McKeown που στηρίζεται στο γεγονός ότι, στην περίπτωση της ταξινόμησης πολικότητας, οι δύο κατηγορίες ενδιαφέροντος αντιπροσωπεύουν αντιδιαμετρικές έννοιες και οι «περιορισμοί στην αντίθεση» μπορούν να χρησιμοποιηθούν για να οδηγήσουν στην ορθή ταξινόμηση του κειμένου (Hatzivassiloglou & McKeown, 1991). Η μέθοδος τους ξεκινά με μια λίστα με ρίζες από επίθετα γνώμης που μαζί με ένα σύνολο γλωσσικών περιορισμών προσδιορίσουν πρόσθετες λέξεις (κυρίως επίθετα ή επιρρήματα) και τους προσανατολισμούς τους. Για παράδειγμα ένας κανόνας αναφέρει ότι οι σύνδεσμοι «και» συνδέουν επίθετα με τον ίδιο προσανατολισμό. Αυτή η ιδέα ονομάζεται συνεκτικότητα συναισθημάτων, η οποία δεν είναι πάντα συνεπής πρακτικά. Υπάρχουν επίσης αντιφατικές εκφράσεις («αλλά», ωστόσο) που υποδεικνύονται ως αλλαγές γνώμης. Προκειμένου να προσδιοριστεί εάν δύο συνδυασμένα επίθετα έχουν τους ίδιους ή διαφορετικούς προσανατολισμούς, η μάθηση εφαρμόζεται σε ένα μεγάλο σύνολο λέξεων. Στη συνέχεια, οι δεσμοί μεταξύ των επίθετων σχηματίζουν ένα γράφημα και η ομαδοποίηση γίνεται με την βοήθεια αυτού σε δύο ομάδες λέξεων: θετικών και αρνητικών. Άλλες παρόμοιες τεχνικές βασίζονται σε ρίζες λέξεων και διαδίδουν το συναισθηματικό τους σθένος των, αφού η πολικότητα τους είναι γνωστή, σε όρους που συνυπάρχουν μαζί τους σε κείμενο, σε γλωσσικά λεξικά ή σε συνώνυμα και λέξεις που συνυπάρχουν μαζί τους σε άλλες σχέσεις που ορίζονται από μεγάλα γλωσσικά σύνολα (πχ WordNet).

Η μέθοδος Conditional Random Fields (CRF) είναι επίσης μία τεχνική της κατηγορίας αυτή που βασίζεται στην μελέτη αλληλουχίας λέξεων για την εξαγωγή εκφράσεων γνώμης. Με αυτήν διακρίνεται η πολικότητα του συναισθήματος με αλγόριθμους αντιστοίχησης μοτίβων πολλαπλών συμβολοσειρών. Εφαρμόστηκε σε κινεζικές online κριτικές και απέδωσε ικανοποιητικά συναισθηματικά λεξικά. Οι Xu και Liao χρησιμοποίησαν μοντέλο CRF δύο επιπέδων με μη σταθερές αλληλεξαρτήσεις για να εξαγάγουν τις συγκριτικές σχέσεις. Αυτό έγινε με τη χρήση των πολύπλοκων εξαρτήσεων μεταξύ των σχέσεων, των οντοτήτων και των λέξεων και των αδιατάρακτων αλληλεξαρτήσεων μεταξύ των σχέσεων. Σκοπός τους ήταν να δημιουργήσουν ένα γραφικό μοντέλο για την εξαγωγή και οπτικοποίηση των συγκριτικών σχέσεων μεταξύ των προϊόντων από τα σχόλια των πελατών. Τα αποτελέσματα παρουσιάστηκαν ως χάρτες συγκριτικής σχέσης για υποστήριξη αποφάσεων στη διαχείριση κινδύνων σε κερδοσκοπικούς οργανισμούς. Τα αποτελέσματά τους έδειξαν ότι η μέθοδος τους μπορεί να εξαγάγει τις συγκριτικές σχέσεις με μεγαλύτερη ακρίβεια από άλλες μεθόδους και ο συγκριτικός χάρτης σχέσεων τους είναι δυνητικά πολύ αποτελεσματικό εργαλείο για τη στήριξη της διαχείρισης επιχειρηματικών κινδύνων και τη λήψη αποφάσεων.

Η προσέγγιση αυτή θα πρέπει να συνδυάζει στατιστικούς ή σημασιολογικούς μηχανισμούς προκειμένου να καθίσταται αποδοτική καθώς σε διαφορετική περίπτωση

προκύπτουν μεγάλοι κατάλογοι εκφράσεων που είναι δύσκολα διαχειρίσιμοι αλλά και η χρήση σε διαδικασίες εντοπισμού της πολικότητας κειμένων είναι κοστοβόρα σε χρόνο και πόρους.

- Στατιστικοί μηχανισμοί: Η εύρεση συνόλων συσχέτισης ή ριζών λέξεων γνώμης μπορεί να πραγματοποιηθεί χρησιμοποιώντας στατιστικές τεχνικές. Αυτό μπορεί να γίνει με την παραγωγή παλαιών πολικοτήτων που προκύπτουν από την συνύπαρξη των επίθετων σε ένα σύνολο λέξεων. Μπορεί να χρησιμοποιηθεί ολόκληρο το σύνολο των ευρετηριασμένων εγγράφων από τον παγκόσμιο ιστό ως το σύνολο για την παραγωγή του λεξικού ώστε να ξεπεραστεί η αδυναμία της μη διαθεσιμότητας ορισμένων λέξεων εάν το χρησιμοποιούμενο σύνολο δεν είναι αρκετά μεγάλο. Η πολικότητα μιας λέξης μπορεί να προσδιοριστεί με τη μελέτη της συχνότητας εμφάνισης της σε ένα μεγάλο σχολιασμένο κορμό κειμένων. Αν η λέξη εμφανίζεται πιο συχνά μεταξύ θετικών κειμένων, τότε η πολικότητα είναι θετική. Εάν εμφανίζεται συχνότερα στα αρνητικά κείμενα, τότε η πολικότητα είναι αρνητική. Εάν έχει ίσες συχνότητες, τότε θεωρείται πως είναι μια ουδέτερη λέξη. Οι παρόμοιες λέξεις γνώμης συχνά εμφανίζονται μαζί σε ένα σύνολο οπότε αν δύο λέξεις εμφανίζονται συχνά μαζί στο ίδιο πλαίσιο, είναι πιθανό να έχουν την ίδια πολικότητα. Επομένως, η πολικότητα μιας άγνωστης λέξης μπορεί να προσδιοριστεί με τον υπολογισμό της σχετικής συχνότητας εμφάνισης της με άλλες λέξεις γνωστής πολικότητας. Οι στατιστικές μέθοδοι χρησιμοποιούνται σε πολλές εφαρμογές που σχετίζονται με τη ανάλυση συναισθήματος. Μία από αυτές είναι η ανίχνευση του χειρισμού των σχολιασμών των χρηστών του διαδικτύου με τη διεξαγωγή μιας στατιστικής δοκιμασίας τυχαίας έκφρασης που ονομάζεται δοκιμή Runs. Η Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis - LSA) είναι μια στατιστική προσέγγιση που χρησιμοποιείται για την ανάλυση των σχέσεων μεταξύ ενός συνόλου εγγράφων και των όρων που αναφέρονται σε αυτά προκειμένου να παραχθεί ένα σύνολο εννοιολογικών προτύπων σχεδίων που σχετίζονται με τα έγγραφα και τους όρους. Ο σημασιολογικός προσανατολισμός μιας λέξης είναι μια στατιστική προσέγγιση που χρησιμοποιείται μαζί με τη μέθοδο PMI[1]. Υπάρχει επίσης μια σημασιολογική εφαρμογή που ονομάζεται Hyperspace Analogue to Language (HAL) κατά την οποία ο σημασιολογικός χώρος είναι ο χώρος στον οποίο οι λέξεις αντιπροσωπεύονται από σημεία. Η θέση κάθε σημείου μαζί με κάθε άξονα σχετίζεται κατά κάποιο τρόπο με τη σημασία της λέξης. Η σημασιολογική πληροφορία προσανατολισμού των λέξεων χαρακτηρίζεται από ένα συγκεκριμένο διάνυσμα. Στην συνέχεια εκπαιδεύεται ένα μοντέλο για τον προσδιορισμό του σημασιολογικού προσανατολισμού των όρων.

[1] Το PMI (Positive, Minus, Interesting) είναι μια μεθοδολογία που ωθεί τους συμμετέχοντες σε μια αναζήτηση όλων των πτυχών γύρω από μία ιδέα. Στόχος της είναι να καταπολεμηθεί η τάση των ανθρώπων να εμμένουν στην αρχική τους εντύπωση σχετικά με ένα θέμα. Το PMI αναπτύχθηκε από τον Edward de Bono μια προσέγγιση για την επίλυση προβλημάτων που ενθαρρύνει τη σκέψη με δημιουργικό, μη παραδοσιακό τρόπο. Οι στόχοι του PMI είναι να βοηθήσουν τους συμμετέχοντες να δουν τις δύο πλευρές ενός επιχειρήματος και να σκεφτούν ευρύτερα ένα θέμα. Η δραστηριότητα είναι σύντομη και οι συμμετέχοντες χρειάζεται να εντοπίσουν τα θετικά, αρνητικά και ενδιαφέροντα στοιχεία σχετικά με κάποιο ζήτημα.

- Σημασιολογικοί μηχανισμοί: Η σημασιολογική προσέγγιση δίνει τιμές συναισθήματος άμεσα και βασίζεται σε διαφορετικά χαρακτηριστικά για τον υπολογισμό της ομοιότητας μεταξύ των λέξεων. Με τον τρόπο αυτό αποδίδονται παρόμοιες τιμές σε σημασιολογικά κοντινές λέξεις. Μία λίστα λέξεων (πχ προερχόμενη από το WordNet) θα μπορούσε να χρησιμοποιηθεί για την απόκτηση μιας λίστας λέξεων συναισθήματος με την διαδοχική επέκταση του αρχικού συνόλου με συνώνυμα και αντωνυμία. Η προκύπτουσα λίστα μπορεί στη συνέχεια να αποτελέσει την βάση για τον προσδιορισμό της πολικότητας των αισθήσεων για μια άγνωστη λέξη από τη σχετική καταμέτρηση θετικών και αρνητικών συνωνύμων αυτής της λέξης. Η Σημασιολογική προσέγγιση χρησιμοποιείται σε πολλές εφαρμογές για την κατασκευή ενός μοντέλου λεξικού για την περιγραφή ρημάτων, ουσιαστικών και επίθετων που θα χρησιμοποιηθούν στην ανάλυση συναισθήματος με την περιγραφή των σχέσεων υποκειμενικότητας μεταξύ των υποκειμένων σε μια πρόταση που εκφράζει ξεχωριστές στάσεις για κάθε υποκειμένου. Αυτές οι σχέσεις υποκειμενικότητας επισημαίνονται με πληροφορίες που αφορούν τόσο την ταυτότητα του όσο και τον προσανατολισμό της στάσης του και έτσι πραγματοποιείται η κατηγοριοποίηση σε σημασιολογικές ομάδες.

3.2.4 Τεχνικες επεξεργασιας φυσικης γλωσσας

Οι τεχνικές επεξεργασίας φυσικής γλώσσας συχνά χρησιμοποιούνται για τον καλύτερο εντοπισμό συντακτικών δομών που οδηγούν σε συγκεκριμένες σημασιολογικές συσχετίσεις με συναισθήματα. Οι τεχνικές αυτές τρέχουν σε προπαρασκευαστικό στάδιο της δημιουργία των λεξικών και μέσα από την συντακτική ανάλυση προτάσεων μπορούν να υπολογίσουν ποσοτικά το συναίσθημα που εκλύεται από κάθε γλωσσικό στοιχείο που μπορεί να βρίσκεται σε ένα κείμενο. Μια ακόμα τάση είναι η μελέτη τω κειμένων σε επίπεδο λόγου όπου οι πληροφορίες αναζητούνται ανάμεσα σε προτάσεις ή σε εκφράσεις μέσα σε προτάσεις. Ο εντοπισμός του αισθήματος στο επίπεδο του λόγου χρησιμοποιεί πέντε τύπους ρητορικών σχέσεων: Αντίθεση, Διόρθωση, Υποστήριξη, Αποτέλεσμα και Συνέχεια σε σχέση με το συναίσθημα που εντοπίζεται. Επίσης χρησιμοποιείται και η έννοια των πλαισίων γνώμης και αποτελούνται από απόψεις και τις σχέσεις με τους προσανατολισμούς τους. Η Θεωρία Ρητορικής Δομής (Rhetorical Strucutere Theory RST) περιγράφει τον τρόπο διαίρεσης ενός κειμένου σε διαστήματα, καθένα από τα οποία αντιπροσωπεύει ένα σημαντικό μέρος του κειμένου. Τα τμήματα που προκύπτουν αξιολογούνται ως προς την σπουδαιότητα τους με βάση τη δομή του λόγου. Με τον τρόπο αυτό γίνεται ακριβέστερη η εκτίμηση των συναισθημάτων που περικλείονται σε ένα κείμενο. Άλλες παρόμοιες τεχνικές εξετάσουν τις σχέσεις λόγου που εντοπίζονται στο κείμενο και που έχουν προηγουμένως εκτιμηθεί ως προς την πολικότητα τους ή ενσωματώνουν βαθμολογίες πολικότητας από διαφορετικά λεξικά συναισθημάτων χρησιμοποιώντας πληροφορίες για τις σχέσεις μεταξύ γειτονικών τμημάτων κειμένου. Συνήθως τα τμήματα κειμένων εκτείνονται σε μέγεθος προτάσεων.

Κεφάλαιο 4

Μελέτη Περίπτωσης

4.1 Χρήση της R

Η γλώσσα R είναι ένα περιβάλλον προγραμματισμού που χρησιμοποιείται κυρίως στην έρευνα. Είναι δωρεάν διαθέσιμο στο διαδίκτυο από όπου και παρέχεται πλήρης υποστήριξη από ισχυρή κοινότητα προγραμματιστών. Είναι ένα ισχυρό εργαλείο που χρησιμοποιείται σε μία ευρεία ποικιλία κατηγοριών ερευνητικών έργων. Αυτό επιτυγχάνεται με την κατάλληλη ενσωμάτωση πακέτων λειτουργιών που εξειδικεύονται στον προσανατολισμό της κάθε έρευνας. Στην παρούσα εργασία χρησιμοποιήθηκαν κατάλληλες βιβλιοθήκες για την ανάλυση συναισθήματος με μηχανική μάθηση. Αυτές οι βιβλιοθήκες περιλαμβάνουν λειτουργίες που μπορούν να χρησιμοποιηθούν σε όλες τις φάσεις της διαδικασίας εξόρυξης συναισθήματος από κείμενα.

4.2 Δεδομένα

Τα δεδομένα που θα χρησιμοποιηθούν προέρχονται από τον ετήσιο διαγωνισμό SemEval του έτους 2016 και το τμήμα εκείνο του διαγωνισμού που αφορά την ανάλυση συναισθήματος σε δεδομένα που προέρχονται από το Twitter (International Workshop on Semantic Evaluation 2016 / Task 4 - Sentiment Analysis in Twitter). Τα δεδομένα σε αυτό το τμήμα του διαγωνισμού διακρίνονται στις εξής κατηγορίες:

- Subtask A: Δεδομένα για την ταξινόμηση της συνολικής συναισθηματικής κατάστασης σε τρεις κατηγορίες (θετική, αρνητική, ουδέτερη).
- Subtask B: Δεδομένα για την ταξινόμηση του συναισθήματος (θετικό, αρνητικό) για tweets που σχετίζονται με συγκεκριμένο θέμα.
- Subtask C: Δεδομένα για την ταξινόμηση του συναισθήματος (πολύ θετικό, θετικό, ουδέτερο, αρνητικό, πολύ αρνητικό) για tweets που σχετίζονται με συγκεκριμένο θέμα.
- Subtask D: Δεδομένα για την κατανομή του συναισθήματος (θετικό, αρνητικό) για tweets που σχετίζονται με συγκεκριμένο θέμα.
- Subtask E: Δεδομένα για την ταξινόμηση του συναισθήματος (πολύ θετικό, θετικό, ουδέτερο, αρνητικό, πολύ αρνητικό) για tweets που σχετίζονται με συγκεκριμένο θέμα.

Κάθε ένα από τα αρχεία δεδομένων είναι γραμμογραφημένο ως εξής:

- Το μοναδικό αναγνωριστικό του tweet
- Τον χαρακτηρισμό του tweet

- Το περιεχόμενο του tweet

Στην παρούσα εργασία χρησιμοποιήθηκαν τα δεδομένα του Subtask A ως εξής:

- Ως training set χρησιμοποιείται η ένωση των εξής αρχείων: twitter-2013train-A, twitter-2014train-A, twitter-2015train-A, twitter-2016train-A.
- Ως test set χρησιμοποιείται η ένωση των αρχείων: twitter-2013test-A, twitter-2014test-A, twitter-2015test-A, twitter-2016test-A.

Στον παρακάτω πίνακα φαίνεται το πόσο θετικά, αρνητικά ή ουδέτερα tweets περιλαμβάνονται σε κάθε αρχείο

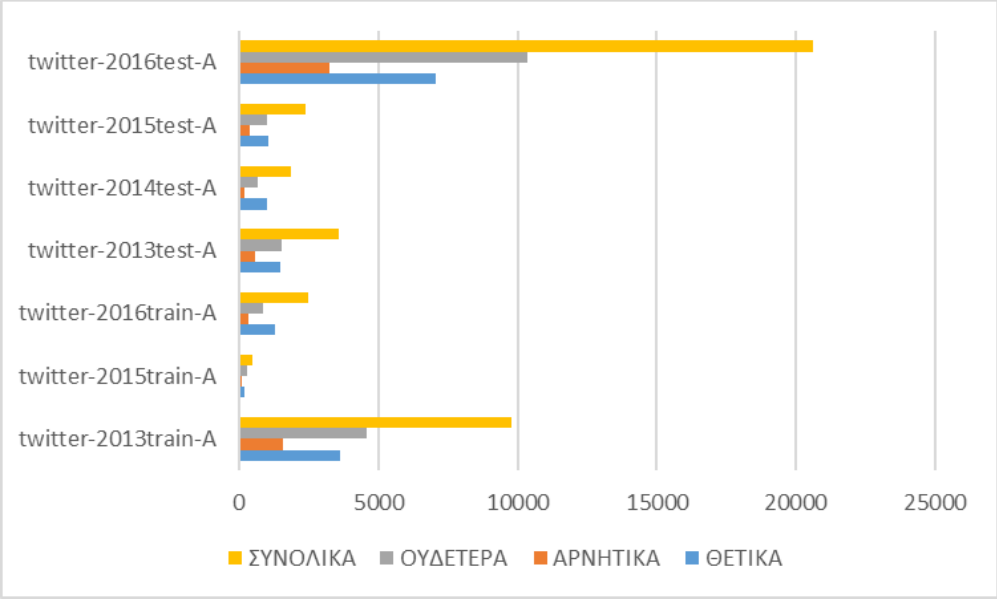
ΠΙΝΑΚΑΣ 4-1

Πίνακας κατανομής των δεδομένων

Αρχείο	Θετικά	Αρνητικά	Ουδέτερα	Σύνολο
twitter-2013train-A	3640	1548	4586	9774
twitter-2015train-A	170	66	253	2474
twitter-2016train-A	1287	347	840	3547
twitter-2013test-A	1475	559	1513	1853
twitter-2014test-A	982	202	669	2390
twitter-2015test-A	1038	365	987	20632
twitter-2016test-A	7059	3231	10342	9774

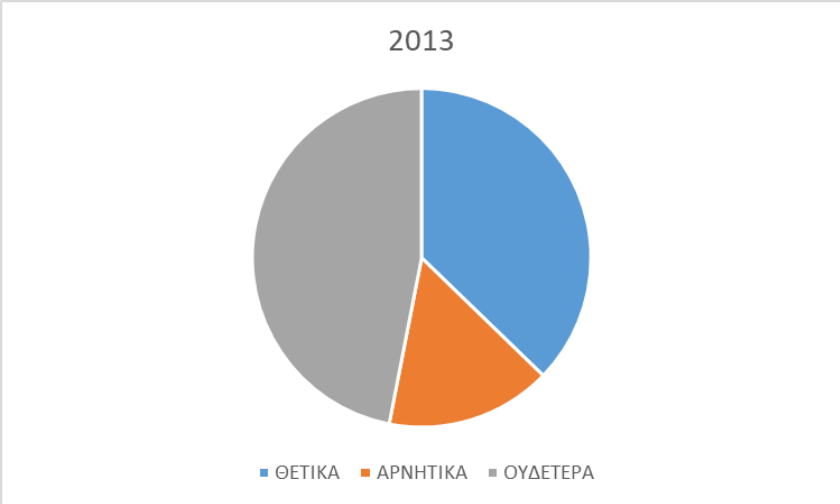
ΣΧΗΜΑ 4-11

Γράφημα κατανομής των δεδομένων



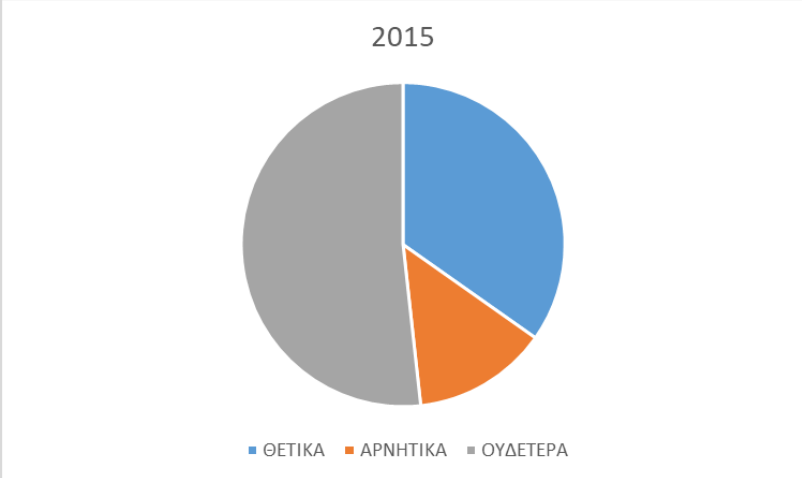
ΣΧΗΜΑ 4-2

Κατανομή των δεδομένων εκπαίδευσης για το έτος 2013



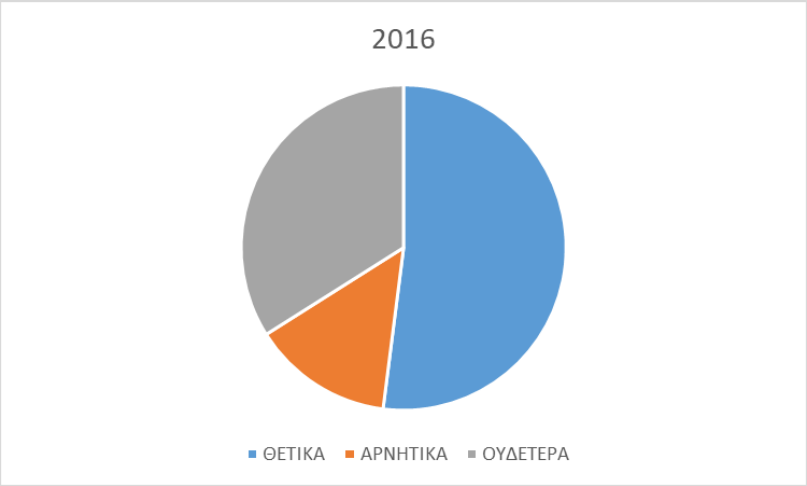
ΣΧΗΜΑ 4-3

Κατανομή των δεδομένων εκπαίδευσης για το έτος 2015



ΣΧΗΜΑ 4-4

Κατανομή των δεδομένων εκπαίδευσης για το έτος 2016



4.3 Μηχανική μάθηση

Η μελέτη της ανάλυσης συναισθήματος με τεχνικές μηχανικής μάθησης περιλαμβάνει τρεις κατηγορίες πειραματικών μετρήσεων.

- Πείραμα Α: Πειραματικές μετρήσεις με την χρήση συνόλων εκπαίδευσης και αξιολόγησης από tweets που κατεγράφησαν την ίδια χρονιά αλλά με την χρήση διαφορετικών αλγορίθμων κατασκευής μοντέλων πρόβλεψης. Με τα πειράματα αυτά επιχειρείται να εκτιμηθεί η διαφοροποίηση στην αποδοτικότητα για τους διαφόρους αλγορίθμους με την χρήση του ιδίου συνόλου δεδομένων.

- Πείραμα Β: Πειραματικές μετρήσεις με την χρήση συνόλων εκπαίδευσης και αξιολόγησης που προέρχονται από tweets που κατεγράφησαν σε διαφορετικές χρονιές χρησιμοποιώντας διαφορετικούς αλγορίθμους κατασκευής μοντέλων πρόβλεψης.

- Πείραμα Γ: Πειραματικές μετρήσεις με διαφορετικές ενέργειες προπαρασκευής των συνόλων εκπαίδευσης και αξιολόγησης των μοντέλων πρόβλεψης.

Στο τέλος των πειραματικών μετρήσεων συγκρίνονται τα αποτελέσματα που προκύπτουν προκειμένου να αξιολογηθεί:

- Η σημασία της προπαρασκευής των συνόλων δεδομένων για την ποιότητα των μοντέλων πρόβλεψης για τους χρησιμοποιούμενους αλγορίθμους.

- Η επίδραση της φύσης των αλγορίθμων στην ποιότητα των μοντέλων πρόβλεψης που προκύπτουν
- Η επίδραση της ομοιογένειας των δεδομένων εκπαίδευσης και αξιολόγησης των μοντέλων πρόβλεψης

4.3.1 Αλγοριθμοι

Οι αλγόριθμοι που χρησιμοποιήθηκαν για τις πειραματικές μετρήσεις είναι:

- **Naïve Bayes:** Πρόκειται για μία μέθοδο μάθησης που βασίζεται σε στατιστικές παρατηρήσεις που αφορούν την πιθανότητα εμφάνισης μίας τιμής της εξαρτημένης μεταβλητής σε σχέση με την τις τιμές του συνόλου ανεξαρτήτων και στο θεώρημα του Bayes. Χαρακτηριστικό του αλγορίθμου ταξινόμησης είναι ότι εκτιμά τις ανεξάρτητες μεταβλητές που επηρεάζουν την ταξινόμηση μεμονωμένα (Stecanella, 2017).
- **Logistic Regression:** Η λογική παλινδρόμηση είναι μια στατιστική μέθοδος για την ανάλυση ενός συνόλου δεδομένων στο οποίο υπάρχουν μία ή περισσότερες ανεξάρτητες μεταβλητές που καθορίζουν ένα αποτέλεσμα. Το αποτέλεσμα εξαρτάται και υπολογίζεται με βάση μία τιμή – διχοτόμο σε μεταβλητή. Στη λογική παλινδρόμηση, η εξαρτώμενη μεταβλητή είναι δυαδική ή διχοτομημένη, δηλ. Περιέχει μόνο δεδομένα που κωδικοποιούνται ως 1 ή 0. Ο στόχος της είναι να βρεθεί το καλύτερο και λογικό μοντέλο που θα περιγράψει τη σχέση ανάμεσα στο διχοτόμο χαρακτηριστικό του συνόλου δεδομένων και ένα σύνολο ανεξάρτητων μεταβλητών (προγνωστικών ή επεξηγηματικών). Η λογική παλινδρόμηση δημιουργεί τους συντελεστές (και τα τυπικά σφάλματα και τα επίπεδα σπουδαιότητάς τους) ενός τύπου για την πρόβλεψη ενός λογαριθμικού μετασχηματισμού της πιθανότητας παρουσίας του χαρακτηριστικού ενδιαφέροντος (MedCalc, 2018): Logit(p)=b0+b1x1+b2x2+…+bkxk (p είναι η πιθανότητα παρουσίας του χαρακτηριστικού που εξετάζεται) και μετασχηματίζεται σε Logit(p) = ln(p/(1-p))
- **Sequential Minimal Optimization (SMO):** Πρόκειται για αλγόριθμο που επιλύει προβλήματα τετραγωνικού προγραμματισμού που προκύπτουν κατά την εκπαίδευσης μηχανών διανυσμάτων υποστήριξης (Support Vector Machines – SVM). Πρόκειται για επαναληπτικό αλγόριθμο που επιλύει προβλήματα βελτιστοποίησης όπου δεδομένου ενός συνόλου δεδομένων εκπεφρασμένων σε διανύσματα n διαστάσεων κ και μίας μεταβλητής y που αντιστοιχεί σε κάθε ένα από αυτά, η τελευταία λαμβάνει τιμή 1 ή -1 ανάλογα με τις τιμές των παραμέτρων των διανυσμάτων. Ο αλγόριθμος γενικά διασπά το πρόβλημα σε μία σειρά από μικρότερα και η επίλυση του αρχικού προβλήματος πλέον προκύπτει από την γενίκευση της επίλυσης αυτών (Platt, 1998).
- **J48:** Το J48 είναι ένας αλγόριθμος ανάπτυξης δένδρων απόφασης που προέκυψε σαν μια επέκταση του επιτυχημένου ID3. Τα επιπρόσθετα χαρακτηριστικά του είναι η καταγραφή των ελλειπουσών τιμών, το κλάδεμα των δέντρων αποφάσεων

κατά την εξέλιξη του, οι σειρές χαρακτηριστικών συνεχών τιμών και η εξαγωγή κανόνων. Άλλοι αντίστοιχοι αλγόριθμοι πραγματοποιούν την ταξινόμηση αναδρομικά έως ότου κάθε φύλλο είναι καθαρό, δηλαδή η ταξινόμηση των δεδομένων πρέπει να είναι όσο το δυνατόν πιο τέλεια. Ο J48 δημιουργεί τους κανόνες από τους οποίους προκύπτει η ιδιαίτερη ταυτότητα αυτών των δεδομένων. Ο στόχος είναι η σταδιακή γενίκευση ενός δέντρου αποφάσεων έως ότου αποκτήσει ικανοποιητική ισορροπία ευελιξία και ακρίβεια (Kaur & Chhabra, 2014).

- **Logistic Model Trees (LMT):** Οι μέθοδοι επαγωγής των δέντρων και τα γραμμικά μοντέλα είναι δημοφιλείς τεχνικές για επιβλεπόμενες διαδικασίες εκπαίδευσης μοντέλων πρόβλεψης, τόσο για την πρόβλεψη των ονομαστικών τάξεων όσο και για τις αριθμητικές τιμές. Για την πρόβλεψη αριθμητικών ποσοτήτων, έχουν γίνει εργασίες για το συνδυασμό αυτών των δύο σχημάτων σε μοντέλα δένδρων (δέντρα που περιέχουν λειτουργίες γραμμικής παλινδρόμησης στα φύλλα). Ο αλγόριθμος αυτός προσαρμόζει αυτή την ιδέα για προβλήματα ταξινόμησης, χρησιμοποιώντας λογική παλινδρόμηση. Χρησιμοποιεί μια διαδικασία σταδιακής τοποθέτησης για την δημιουργία των μοντέλων λογικής παλινδρόμησης που μπορούν να επιλέξουν σχετικά χαρακτηριστικά από τα δεδομένα για την κατασκευή μοντέλων λογικής παλινδρόμησης στα φύλλα με διαδοχική εξομάλυνση εκείνων που κατασκευάζονται σε υψηλότερα επίπεδα στο δέντρο. Ένα πλεονέκτημα της χρήσης λογιστικής παλινδρόμησης είναι ότι παράγονται σαφείς εκτιμήσεις πιθανοτήτων κλάσης και όχι απλώς μια ταξινόμηση. Σε γενικές γραμμές ο αλγόριθμος επιλέγει τα χαρακτηριστικά που θα συμπεριληφθούν στα μοντέλα λογικής παλινδρόμησης και εφαρμόζει έναν τρόπο δημιουργίας των μοντέλων στα φύλλα, με τη βελτίωση των αντίστοιχων μοντέλων που έχουν εκπαιδευτεί σε υψηλότερα επίπεδα στο δέντρο, δηλαδή σε μεγαλύτερα υποσύνολα των δεδομένων εκπαίδευσης (Landwehr, et al., 2004).

- **Projective Adaptive Resonance Theory (PART):** Το νευρωνικό δίκτυο του θεωρητικού προσαρμοζόμενου θεωρητικής συντονισμού (PART) έχει αποδειχθεί ότι είναι αποτελεσματικό στη συσσώρευση συνόλων δεδομένων σε χώρους μεγάλων διαστάσεων. Ο αλγόριθμος PART βασίζεται στις παραδοχές ότι οι εξισώσεις μοντέλου του PART (μιας μεγάλης κλίμακας και ενός ιδιαιτέρως διαταραγμένου συστήματος διαφορικών εξισώσεων σε συνδυασμό με ένα μηχανισμό επαναφοράς) έχουν αρκετά κανονικές υπολογιστικές επιδόσεις. Η αρχιτεκτονική του βασίζεται στον αλγόριθμο ART που αναπτύχθηκε από τους Carpenter και Grossberg, με έναν μηχανισμό επιλεκτικής σηματοδοσίας εξόδου (SOS) για την αντιμετώπιση της εγγενούς ασυμμετρίας στον πλήρη χώρο των σημείων δεδομένων, προκειμένου να επικεντρωθούν στις διαστάσεις στις οποίες μπορούν να βρεθούν πληροφορίες. Το κύριο χαρακτηριστικό του νευρωνικού δικτύου PART είναι ένα κρυμμένο στρώμα νευρώνων που ενσωματώνει το SOS για να υπολογίσει την ανομοιογένεια μεταξύ της εξόδου ενός δεδομένου νευρώνα εισόδου και του αντίστοιχου στοιχείου του προτύπου (στατιστικός μέσος όρος) ενός υποψήφιου νευρώνα συστάδων και για να επιτρέψει στο σήμα να μεταδοθεί στον νευρώνα συστάδων μόνο όταν το μέτρο ομοιότητας είναι αρκετά μεγάλο (Wu, 2015) (GAN & YIN, 2014).

- **One Rule (OneR):** Πρόκειται για αλγόριθμο ο οποίος αναζητεί εκείνο το χαρακτηριστικό του συνόλου δεδομένων το οποίο μπορεί να φανερώσει την ταξινόμηση των στοιχείων του με την μέγιστη δυνατή ακρίβεια. Το OneR αποτελεί συντομογραφία του "One Rule - Ένας κανόνας". Είναι ένας απλός και ακριβής αλγόριθμος ταξινόμησης που παράγει έναν κανόνα για κάθε προγνωστικό στα δεδομένα και στη συνέχεια επιλέγει τον κανόνα με το μικρότερο συνολικό σφάλμα ως "έναν κανόνα". Για να δημιουργηθεί ένας κανόνας για έναν σύνολο, κατασκευάζεται ο πίνακας συχνότητας για κάθε προγνωστικό σε σχέση με τον στόχο. Έχει αποδειχθεί ότι η OneR παράγει κανόνες μόνο ελαφρώς λιγότερο ακριβείς από τους πλέον σύγχρονους αλγορίθμους ταξινόμησης (Sayad, 2018).
- **Repeated Incremental Pruning to Produce Error Reduction (RIP):** Βασίζεται σε κανόνες σύνδεσης με μειωμένο κλάδεμα σφάλματος (Reduced Error Pruning - REP) για την ανάπτυξη δένδρων αποφάσεων. Σε REP για αλγόριθμους κανόνων, τα δεδομένα εκπαίδευσης χωρίζονται σε ένα αναπτυσσόμενο σύνολο και ένα σύνολο που χρησιμοποιείται για το κλάδεμα του δένδρου. Αρχικά σχηματίζεται ένα σύνολο κανόνων με βάση το αναπτυσσόμενο σύνολο, χρησιμοποιώντας κάποια ευριστική μέθοδο. Αυτό το σύνολο απλοποιείται επανειλημμένα εφαρμόζοντας κάθε φορά ένα κανόνα κλαδέματος. Σε κάθε στάδιο απλοποίησης, ο επιλεγμένος κανόνας κλαδέματος είναι αυτός που αποδίδει τη μεγαλύτερη μείωση σφάλματος στο σετ κλάδεμα. Η απλοποίηση τερματίζεται στην κατάσταση εκείνη κατά την οποία όταν εφαρμοστεί οποιοσδήποτε κανόνας θα αυξήσει το σφάλμα (Cohen, 1995).

4.3.2 Προετοιμασια

Για την ανάλυση συναισθήματος με την χρήση τεχνικών μηχανικής μάθησης χρειάστηκε να ληφθούν και να εγκατασταθούν τα εξής πακέτα λειτουργιών για χρήση στο περιβάλλον της R:

- tm: Περιλαμβάνει λειτουργίες για την εξόρυξη κείμενου. Περιλαμβάνει χρήσιμες συναρτήσεις για την αναπαράσταση των κειμένων σαν διανύσματα ώστε να είναι δυνατή η περαιτέρω επεξεργασία ή μελέτη τους.
- RTextTools: Το RTextTools είναι ένα πακέτο λειτουργιών μηχανικής μάθησης για αυτόματη ταξινόμηση κειμένου. Το πακέτο περιλαμβάνει εννέα αλγόριθμους για την ταξινόμηση των συνόλων (svm, slda, boosting, bagging, τυχαία δάση, glmnet, δέντρα αποφάσεων, νευρωνικά δίκτυα, μέγιστη εντροπία), αναλυτικές αναλύσεις και λεπτομερή τεκμηρίωση.
- e1071: Λειτουργίες για latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier
- dplyr: Εργαλείο για τον χειρισμό μεγάλων πλαισίων δεδομένων.
- caret: περιλαμβάνει λειτουργίες για την εκπαίδευση και την γραφική αναπαράσταση μοντέλων ταξινόμησης και παλινδρόμησης.

- Rweka: Περιλαμβάνει λειτουργίες που προέρχονται από το εργαλείο που χρησιμοποιείται κατά κόρον σε εφαρμογές εξόρυξης γνώσης, το weka. Το *WEKA* (Waikato Environment for Knowledge Analysis) είναι λογισμικό μηχανικής μάθησης, που αναπτύχθηκε από το Πανεπιστήμιο του Waikato και το οποίο προσφέρει ένα σύνολο λειτουργιών για την γραφική παρουσίαση της εξέλιξης αλγορίθμων που στοχεύουν στην ανάλυση δεδομένων και την παράγωγή προβλεπτικών μοντέλων. Χαρακτηριστικό του είναι η διάθεση φιλικών προς τον τελικό χρήστη διεπαφών, κάτι που το καθιστά ιδιαίτερα δημοφιλές για χρήση σε διαφόρων τύπων έρευνες. Τα εργαλεία του (φίλτρα, κατηγοριοποιητικές, ταξινομητές, συσχετιστές και επιλογείς χαρακτηριστικών) χρησιμοποιούνται μέσω γραφικών διεπαφών. Είναι γραμμένο σε γλώσσα προγραμματισμού JAVA και περιέχει μεθόδους που μπορούν να ανταποκριθούν σε απαιτήσεις για
    - Προεπεξεργασία δεδομένων
    - Ταξινόμηση
    - Συσταδοποίηση
    - Ανακάλυψη Συσχετιστικών Κανόνων

  Είναι δωρεάν διαθέσιμο στην ιστοσελίδα από το διαδίκτυο και απαιτεί την εγκατάσταση Java στον εξοπλισμό που θα χρησιμοποιηθεί. Στο λογισμικό περιλαμβάνεται ένας αριθμός από αλγορίθμους κατηγοριοποίησης και πρόβλεψης. Το Rweka παρέχει ένα API για χρήση στην γλώσσα R.

4.4 Πείραμα Α

Κατά το πρώτο πείραμα εκτιμήθηκε η αποτελεσματικότητα μίας σειράς αλγορίθμων να παράγουν αξιόπιστα μοντέλα πρόβλεψης. Τα σύνολα εκπαίδευσης και αξιολόγησης είναι διαφορετικά και προέρχονται από την ίδια χρονιά.

ΠΙΝΑΚΑΣ 4-2

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο LMT

LMT

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 48,75%			2015 ΑΚΡΙΒΕΙΑ 43,55%			2016 ΑΚΡΙΒΕΙΑ 44,16%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	0	269		0	194	30	0	213	11
/	1	892	76	8	491	85	10	548	26
+	2	665	83	5	418	80	5	467	31

ΠΙΝΑΚΑΣ 4-3

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο Logistic Reggresion

Logistic

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 34%			2015 ΑΚΡΙΒΕΙΑ 39,65%			2016 ΑΚΡΙΒΕΙΑ 37,91%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	72	132	92	21	110	62	21	117	55
/	220	456	356	92	386	200	89	361	228
+	160	360	152	50	273	117	55	270	115

ΠΙΝΑΚΑΣ 4-4

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο SMO

SMO

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 40,65%			2015 ΑΚΡΙΒΕΙΑ 41,5%			2016 ΑΚΡΙΒΕΙΑ 34,93%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	28	151	102	23	115	86	24	107	93
/	105	532	332	57	328	199	102	253	229
+	87	410	253	65	245	193	98	224	181

ΠΙΝΑΚΑΣ 4-5

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο J48

J48

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 43,9%			2015 ΑΚΡΙΒΕΙΑ 41,8%			2016 ΑΚΡΙΒΕΙΑ 42,1%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+

-	8	196	77	24	108	92	10	136	78
/	26	661	282	45	312	227	34	363	187
+	20	521	209	35	256	212	22	302	179

ΠΙΝΑΚΑΣ 4-6

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο Decision Stump

Decision Stump

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 49,15%			2015 ΑΚΡΙΒΕΙΑ 44,24%			2016 ΑΚΡΙΒΕΙΑ 44,31%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	0	271	10	0	223	1	0	223	1
/	0	905	64	0	579	5	2	581	1
+	0	672	78	0	502	1	1	502	0

ΠΙΝΑΚΑΣ 4-7

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο PART

PART

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 44,7%			2015 ΑΚΡΙΒΕΙΑ 41,5%			2016 ΑΚΡΙΒΕΙΑ 39,96%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	12	173	96	12	132	80	16	125	83
/	43	625	301	53	328	203	46	335	203
+	24	469	257	38	261	204	38	292	173

ΠΙΝΑΚΑΣ 4-8

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο One Rule

One Rule

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 48,25%			2015 ΑΚΡΙΒΕΙΑ 43,4%			2016 ΑΚΡΙΒΕΙΑ 43,63%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	0	272	9	0	210	14	0	211	13
/	0	914	55	0	549	35	0	538	46
+	0	699	51	0	483	20	0	469	34

ΠΙΝΑΚΑΣ 4-9

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο Jrip

JRip

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 48,35%			2015 ΑΚΡΙΒΕΙΑ 44,09%			2016 ΑΚΡΙΒΕΙΑ 44,54%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	0	269	12	0	218	6	0	219	5
/	1	911	57	0	570	14	3	563	18
+	1	693	56	0	495	8	1	481	21

ΠΙΝΑΚΑΣ 4-10

Εκτίμηση πολικότητας συναισθήματος με τον αλγόριθμο Naïve Bayes

ΝΑΪVE BAYES

	Έτος συλλογής των στοιχείων των συνόλων εκπαίδευσης και αξιολόγησης								
	2013 ΑΚΡΙΒΕΙΑ 48,35%			2015 ΑΚΡΙΒΕΙΑ 44,09%			2016 ΑΚΡΙΒΕΙΑ 44,54%		
	ΕΚΤΙΜΗΣΗ								
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	-	/	+	-	/	+	-	/	+
-	46	0	63	0	262	58	339	509	533
/	13	211	92	0	740	65	509	2850	1705
+	33	700	447	0	65	307	198	1123	2234

4.5 Πείραμα Β

Στο δεύτερο πείραμα τα δεδομένα εκπαίδευσης και τα δεδομένα αξιολόγησης προέρχονται από διαφορετικές χρονολογίες. Εξετάστηκαν οι ίδιοι αλγόριθμοι με το πρώτο πείραμα. Οι πίνακες αλήθειας και η ακρίβεια πρόβλεψης για τον καθένα φαίνονται στους παρακάτω πίνακες. Σε κάθε στήλη του πίνακα καταγράφονται ο αριθμός των εγγραφών που εκτιμήθηκαν με βάση το μοντέλο πρόβλεψης που δημιουργήθηκε ως αρνητική, ουδέτερη ή θετική πολικότητα αντίστοιχη. Σε κάθε γραμμή του πίνακα καταγράφονται οι αντίστοιχες πολικότητες οι οποίες έχουν καταγραφεί ως πραγματικές. Η κύρια διαγώνιος κάθε 3Χ3 πίνακα που αντιστοιχεί σε κάθε αλγόριθμο περιλαμβάνει το πλήθος των εγγραφών που εκτιμήθηκαν σωστά από το μοντέλο πρόβλεψης.

ΠΙΝΑΚΑΣ 4-11

Εκτίμηση πολικότητας συναισθήματος με εφαρμογή διαφορετικών αλγορίθμων σε δεδομένα εκπαίδευσης και αξιολόγησης που προέρχονται από διαφορετικές χρονιές.

	ΕΚΤΙΜΗΣΗ		
	-	/	+
ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ	LMT:	ΑΚΡΙΒΕΙΑ 51,33	
-	2	174	5
/	1	599	10
+	0	395	16
	Logistic:	40	
-	24	117	64
/	74	373	202
+	50	279	128
	SMO:	41,68	
-	19	108	54
/	66	370	174

+	46	253	112
	J48:		46,1
-	11	134	36
/	22	477	111
+	22	322	67
	DecisionStump:		51
-	3	178	0
/	0	610	0
+	0	411	0
	PART:		44,34
-	18	117	46
/	39	415	156
+	34	277	100
	OneR:		50,33
-	0	178	3
/	0	597	13
+	0	403	8
	Jrip:		51,33
-	4	177	0
/	0	609	1
+	0	407	4
	NAÏVE BAYES:		56,55
-	56	112	118
/	56	497	197
+	34	265	465

Στον επόμενο πίνακα συγκρίνονται τα αποτελέσματα των πειραμάτων Α και Β. Από αυτόν φαίνεται ότι όταν τα δεδομένα εκπαίδευσης και αξιολόγησης προέρχονται από αναρτήσεις από διαφορετικές χρόνιες τότε προκύπτουν μοντέλα τα οποία αξιολογούνται ελαφρώς ακριβέστερα. Κάθε γραμμή του πίνακα αντιστοιχεί σε διαφορετικό αλγόριθμο. Στις τρεις πρώτες στήλες καταγράφεται η ακρίβεια κάθε αλγορίθμου για το σύνολο δεδομένων κάθε έτους. Στην τέταρτη στήλη υπολογίζεται ο μέσος όρος τους. Στην τελευταία στήλη φαίνεται η ακρίβεια που παρατηρήθηκε όταν το σύνολο εκπαίδευσης και το σύνολο αξιολόγησης του μοντέλου πρόβλεψης προέρχονται από παρατηρήσεις διαφορετικών ετών.

ΠΙΝΑΚΑΣ 4-12

Συγκριτικός Πίνακας αποτελεσμάτων εκτίμησης πολικότητας για σύνολα εκπαίδευσης και αξιολόγησης που προέρχονται από την ίδια χρονιά και από διαφορετικές χρονιές.

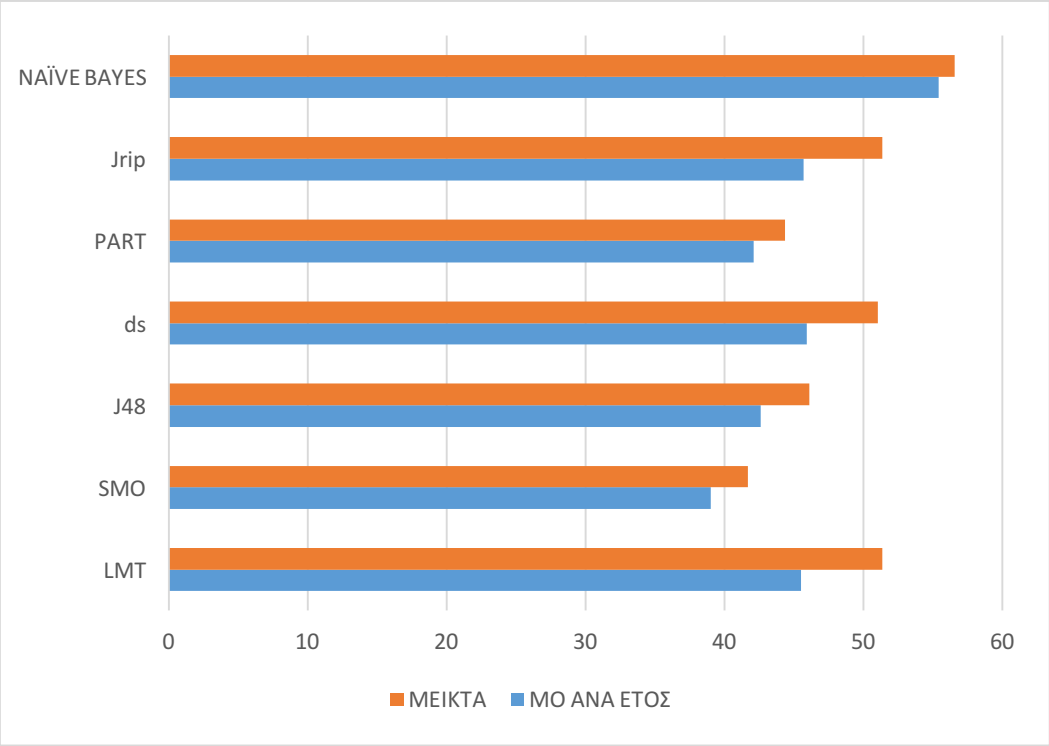
ΑΛΓΟΡΙΘΜΟΣ	2013	2015	2016	ΜΟ	ΜΕΙΚΤΑ
LMT	48,75	43,55	44,16	45,5	51,33
Logistic	34	39,65	37,91	38	40
SMO	40,65	41,5	34,93	39	41,68

J48	43,9	41,8	42,1	42,6	46,1
ds	49,15	44,24	44,31	45,9	51
PART	44,7	41,5	39,96	42,1	44,34
Jrip	48,35	44,09	44,54	45,7	51,33
NAÏVE BAYES	59,65	52,35	54,23	55,4	56,55

Τα δεδομένα του παραπάνω πίνακα φαίνονται σχηματικά στο επόμενο διάγραμμα.

ΣΧΗΜΑ 4-5

Συγκριτικό γράφημα αποτελεσμάτων εκτίμησης πολικότητας για σύνολα εκπαίδευσης και αξιολόγησης που προέρχονται από την ίδια χρονιά και από διαφορετικές χρονιές.



4.6 Πείραμα Γ

Στο πείραμα αυτό εκτελέστηκαν οι τέσσερις αποδοτικότεροι αλγόριθμοι (όπως προέκυψε από τα προηγούμενα πειράματα). Σε κάθε εκτέλεση διαφοροποιούταν το όριο των εμφανίσεων κάθε λέξεις στις αναρτήσεις προκειμένου να ληφθεί υπ' όψη ως προς την

εκτίμηση της πολικότητας. Έτσι έγινε εκτέλεση των αλγορίθμων για συχνότητες 15, 30, 60, 100 και 300 εμφανίσεων. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα. Κάθε γραμμή του πίνακα αντιστοιχεί και σε διαφορετικό αλγόριθμο ενώ κάθε στήλη αντιστοιχεί σε διαφορετικό κάτω όριο συχνοτήτων εμφανίσεων λέξεων (κριτήριο για να λαμβάνεται μία λέξη υπ' όψη κατά την κατασκευή του μοντέλου πρόβλεψης).

ΠΙΝΑΚΑΣ 4-13

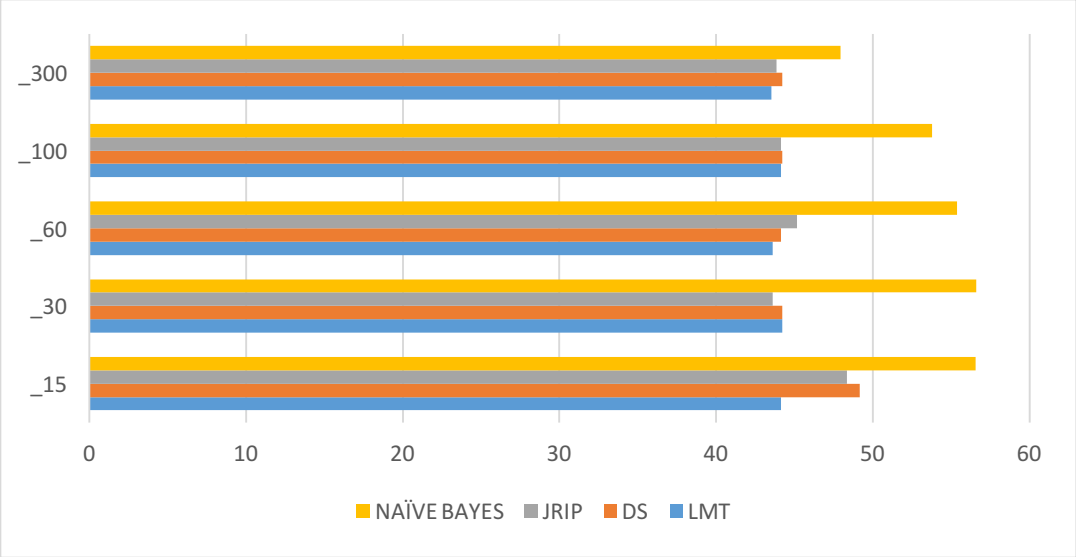
Συγκριτικός Πίνακας αποτελεσμάτων εκτίμησης πολικότητας με εφαρμογή διαφορετικών ορίων συχνότητας εμφάνισης λέξεων στα κείμενα

	ΣΥΧΝΟΤΗΤΑ ΕΜΦΑΝΙΣΗΣ (ΜΕΓΑΛΥΤΕΡΗ ΑΠΟ)				
ΑΛΓΟΡΙΘΜΟΣ	15	30	60	100	300
LMT	44,16	44,24	43,63	44,16	43,55
DS	49,15	44,24	44,16	44,24	44,24
JRIP	48,35	43,63	45,15	44,16	43,85
NAÏVE BAYES	56,55	56,6	55,38	53,77	47,94

Τα αποτελέσματα φαίνονται παραστατικά στο παρακάτω διάγραμμα.

ΣΧΗΜΑ 4-6

Συγκριτικό γράφημα αποτελεσμάτων εκτίμησης πολικότητας με εφαρμογή διαφορετικών ορίων συχνότητας εμφάνισης λέξεων στα κείμενα



Επίσης δοκιμάστηκε η αποδοτικότητα των αλγορίθμων χωρίς να έχει προηγηθεί επεξεργασία των κειμένων. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα και το διάγραμμα.

ΠΙΝΑΚΑΣ 4-14

Συγκριτικός Πίνακας αποτελεσμάτων εκτίμησης πολικότητας με επεξεργασία ή χωρίς επεξεργασία των κειμένων εκπαίδευσης

	ΧΩΡΙΣ ΕΠΕΞΕΡΓΑΣΙΑ	ΜΕ ΕΠΕΞΕΡΓΑΣΙΑ
LMT	48,75	48,3
DS	49,15	49,15
JRIP	48,35	47,9
NAÏVE BAYES	56,55	56,88

4.7 Πείραμα Δ

Στο πείραμα αυτό (Πείραμα Δ) εκτιμήθηκε η απόδοση της χρήσης λεξικών για την εύρεση της πολικότητας για το σύνολο αξιολόγησης. Στον επόμενο πίνακα φαίνονται τα αποτελέσματα. Στις γραμμές του πίνακα καταγράφονται οι πραγματικές πολικότητες που έχουν καταγραφεί ενώ στις στήλες φαίνονται οι εκτιμήσεις που έγιναν με βάση λεξικό όρων.

ΠΙΝΑΚΑΣ 4-15

Πίνακας εκτίμησης συναισθηματικής πολικότητας με την χρήση λεξικών

	ΠΡΑΓΜΑΤΙΚΕΣ		
ΠΡΟΒΛΕΨΗ	-	/	+
-	1499	0	0
/	0	4751	0
+	0	0	3750

Ακρίβεια : 100%

4.8 Ανάλυση Αποτελεσμάτων

Από τα αποτελέσματα που προέκυψαν από την πειραματική εκτέλεση των αλγορίθμων μηχανικής μάθησης στα σύνολα δεδομένων που χρησιμοποιήθηκαν προέκυψε ότι η μέση αποδοτικότητα τους κυμάνθηκε στα επίπεδα του 50%. Αυτό σημαίνει ότι οι τεχνικές που βασίζονται σε μηχανική μάθηση μία στις δύο φορές είναι ικανές να εντοπίζουν με ακρίβεια την συναισθηματική πολικότητα των μικρών κειμένων που αναρτώνται στο twiter. Η αποτελεσματικότητα των αλγορίθμων αυτών αυξάνεται σημαντικά όταν σκοπός της έρευνας είναι να εντοπιστεί η συναισθηματική πολικότητα που δεν ταιριάζει με ένα κείμενο. Η μελέτη των πινάκων αληθείας της εκτέλεσης των πειραμάτων καταδεικνύει ότι τα ερωτήματα της μορφής: "Το κείμενο δεν είναι Θετικό/Αρνητικό" μπορεί να απαντηθεί με πολύ μεγάλη ακρίβεια.

Τα αποτελέσματα της αξιολόγησης των μοντέλων πρόβλεψης που προκύπτουν από τις τεχνικές μηχανικής μάθησης φαίνεται να είναι ανεξάρτητα αν τα δεδομένα εκπαίδευσης και αξιολόγησης προέρχονται από διαφορετικές χρονιές. Αυτό είναι συνέπεια του ότι η γλώσσα και ο τρόπος που οι χρήστες των εφαρμογών κοινωνικής δικτύωσης γράφουν τις αναρτήσεις τους δεν διαφέρει θεαματικά από χρονιά σε χρονιά. Έτσι τα μοντέλα πρόβλεψης που δημιουργούνται επίσης δεν διαφέρουν πολύ μεταξύ τους.

Κάθε αλγόριθμος που χρησιμοποιήθηκε απέδωσε περίπου την ίδια ακρίβεια για δεδομένα που αντλήθηκαν από το twiter σε διαφορετικές χρονιές. Έτσι φαίνεται ότι η αποδοτικότητα της διαδικασίας συνδέεται κυρίως με την φύση του αλγορίθμου και λιγότερο με τα ίδια τα δεδομένα εκπαίδευσης.

Η προεπεξεργασία των δεδομένων που χρησιμοποιούνται για την δημιουργία του μοντέλου πρόβλεψης οδηγεί σε σημαντική μείωση του χρόνου που απαιτείται τόσο για την ίδια δημιουργία του όσο και για την αξιολόγηση του. Κατά συνέπεια μειώνεται ανάλογα και ο χρόνος που απαιτείται για την πρόβλεψη της συναισθηματικής πολικότητας των κειμένων με βάση το δημιουργθέν μοντέλο. Ασφαλή συμπεράσματα δεν προέκυψαν ως προς την αποτελεσματικότητα των μοντέλων που δημιουργήθηκαν με ή χωρίς επεξεργασία. Ωστόσο φαίνεται ότι η ακρίβεια των μοντέλων αυξάνεται αναλογικά με το πλήθος των λέξεων που ελέγχονται κατά την διαδικασία υπολογισμού της συναισθηματικής πολικότητας. Αυτό αν και επιβραδύνει την όλη διαδικασία ενισχύει την προσπάθεια ανίχνευσης της διάθεσης που εκλύεται από τα κείμενα με περισσότερα διακριτά στοιχεία που μπορεί να την προσδιορίζουν.

Ο πιο αξιόπιστος αλγόριθμος φαίνεται να είναι ο Naive Bayes. Προσφέρει σε κάθε περίπτωση ακριβέστερα μοντέλα πρόβλεψης σε σχέση με όσους χρησιμοποιήθηκαν. Ωστόσο όλων τα αποτελέσματα κυμαίνονται στο 50%. Η αποτελεσματικότητα αυτή υπολείπεται σαφώς από την απόδοση της πειραματικής εκτέλεσης με την χρήση λεξικού η οποία προβλέπει την συναισθηματική πολικότητα των κειμένων με ακρίβεια 100%. Στην τελευταία όμως περίπτωση θα πρέπει να σημειωθεί ότι χρησιμοποιήθηκαν λεξικά των οποίων η ανάπτυξη και επικαιροποίηση αποτελεί παράγωγο διαχρονικής μέριμνας ενώ η παραγωγή των μοντέλων πρόβλεψης που πραγματοποιήθηκε με τις τεχνικές μηχανικής μάθησης στα αντίστοιχα πειράματα βασίστηκε σε περιορισμένο αριθμό κειμένων.

Κεφάλαιο 5

Συμπεράσματα

Η επιστήμη της πληροφορικής εξελίσσεται διαρκώς (τα τελευταία χρόνια με γοργούς ρυθμούς) με αποτέλεσμα να παρουσιάζονται συνεχώς καινοτομίες που παρέχουν διευκολύνσεις και λύσεις σε πολλούς τομείς της ανθρώπινης δραστηριότητας. Στην εξέλιξη αυτή συνεισφέρουν και τα ζητήματα και οι προκλήσεις που η ίδια η εξέλιξη εγείρει. Από τέτοιου είδους ζητήματα προέκυψε τόσο η εξόρυξη γνώσης όσο και η εξόρυξη άποψης και συναισθημάτων από μεγάλα σύνολα δεδομένων. Η υπέρ διαθεσιμότητα δεδομένων και η ανάγκη επεξεργασίας τους προκειμένου να προκύψουν συμπεράσματα αποτέλεσαν το έναυσμα για την μελέτη των πτυχών αυτών της πληροφορικής. Οργανισμοί αλλά και φυσικά πρόσωπα πλέον έχουν πρόσβαση σε μεγάλα σύνολα δεδομένων τα οποία μέσω καταλλήλων μηχανισμών έχουν την δυνατότητα να παράγουν χρήσιμη – πολύτιμη πληροφορία. Η έρευνα στον τομές αυτό ανέδειξε τις τεχνικές και μεθοδολογίες που χρησιμοποιούνται σήμερα και οι οποίοι χαρακτηρίζονται γενικά αποδοτικοί με αποτέλεσμα να αποτελούν σημαντικό εργαλείο στις διαδικασίες λήψης αποφάσεων αφού μέσω αυτών φιλτράρονται τα δεδομένα, παράγονται μοντέλα πρόβλεψης που χρησιμοποιούνται για την εκτίμηση μελλοντικών καταστάσεων.

Καθώς η εξόρυξη γνώσης προσθέτει τόσο σημαντική αξία στα δεδομένα, θεωρείται ένας κλάδος πολύ σημαντικός για την εξέλιξη των πληροφοριακών συστημάτων κάθε είδους οργανισμού. Αυτό είναι και το σημαντικότερο κριτήριο για την επιλογή επένδυσης για την αναβάθμιση ενός επιστημονικού πεδίου. Η εξόρυξη γνώσης προσθέτει στα πληροφοριακά συστήματα ευφυΐα και τα δεδομένα από στατική περιουσία των οργανισμών μετατρέπονται σε δυναμικές πηγές γνώσης. Έτσι εκτιμάται ότι η ανακάλυψη νέων μεθοδολογιών – τεχνικών – μηχανισμών – διαδικασιών εξόρυξης γνώσης θα αποτελέσει και στο μέλλον αντικείμενο συστηματικής μελέτης. Νομοτελειακά αυτό θα οδηγήσει σε νέα εργαλεία και εφαρμογές. Πιο αναλυτικά η μελέτες αναμένεται να επικεντρωθούν στα εξής:

- Εξόρυξη γνώσης ποικίλων προσανατολισμών από μεγάλα κοινά σύνολα δεδομένων

- Η εφαρμογή δια δραστικότητας στις διαδικασίες εξόρυξης γνώσης με σκοπό η ανατροφοδότηση με τον ανθρώπινο παράγοντα να παράγονται πιο ακριβή αποτελέσματα και σε μικρότερο χρόνο.
- Η χρήση αποτελεσμάτων διαδικασιών εξόρυξης γνώσης σε νέες διαδικασίες προκειμένου είτε την παραγωγή νέας γνώσης είτε να ελέγχεται η εγκυρότητα της.
- Ανάπτυξη μηχανισμών με υψηλό βαθμό τυποποίησης που να μπορούν να χρησιμοποιούνται σε ετερογενή περιβάλλοντα.
- Η ανακάλυψη αποδοτικών και τυποποιημένων τρόπων παρουσίασης της γνώσης που προκύπτει από τους μηχανισμούς εξόρυξης γνώσης.
- Η ανακάλυψη νέων αλγόριθμων εξόρυξης γνώσης που να εξελίσσονται ταχύτερα και με απαίτηση λιγότερων υπολογιστικών πόρων.
- Καθώς πλέον οι διαδικασίες αυτές μπορεί να τρέχουν σε κατανεμημένα και διαμοιραζόμενα περιβάλλοντα, σημαντικό πεδίο μελέτης μπορεί να είναι η εξασφάλισης των διακινουμένων δεδομένων, των διαδικασιών και των αποτελεσμάτων τους.
- Η αντικειμενική αξιολόγηση της αποτελεσμάτων των διαδικασιών.

Η εξόρυξη γνώσης εφαρμόστηκε με ιδιαίτερη επιτυχία και στην αναζήτηση της συναισθηματικής κατάστασης ανθρώπων. Σε πολλές περιπτώσεις η γνώμη και η συναισθηματική – ψυχολογική κατάσταση μίας ή περισσοτέρων κατηγοριών ανθρώπων μπορεί να είναι πολύτιμη για υπεύθυνους λήψης αποφάσεων. Στις περιπτώσεις αυτές οι τεχνικές αναφέρονται ως εξόρυξη άποψης ή ανάλυση συναισθήματος. Πρόκειται για εφαρμογές με υψηλή δυναμική οι οποίες αναμένεται να διατηρήσουν αυτή την δυναμική τους και στο μέλλον κυρίως ως συνέπεια:

- Την έντονη επιθυμία διαφόρων ειδών οργανισμών και γενικότερα υπεύθυνων λήψης αποφάσεων για την συναισθηματική κατάσταση ομάδων ανθρώπων.
- Την διαθεσιμότητα πλούσιων συνόλων δεδομένων που περιγράφουν την άποψη και την συναισθηματική κατάσταση μεγάλων συνόλων των πληθυσμών, κυρίως μέσω δημοφιλών εφαρμογών του WEB2.0 που επιτρέπουν την διαμόρφωση του περιεχομένου τους από τον ίδιο τον χρήστη. Αυτά τα σύνολα μπορεί να προέχονται

από αναρτήσεις, σχόλια και αξιολογήσεις σε blogs ή εφαρμογές κοινωνικής δικτύωσης.

- Τα μεγάλα περιθώρια βελτίωσης του εξοπλισμού και του λογισμικού που απαιτούν οι μηχανισμοί εξόρυξης άποψης και ανάλυσης συναισθήματος.
- Το μεγάλο και πολυποίκιλο εύρος των εφαρμογών που μπορεί να έχει η εξόρυξη άποψης και η ανάλυση συναισθήματος. Η ώθηση στην ανάπτυξη των διαδικασιών αναμένεται να δίνεται από διαφορετικά επιστημονικά πεδία εκτός της πληροφορικής (οικονομικό τομέα, κοινωνιολογία, ψυχολογία κα).

Στην ανάλυση συναισθήματος και την εξόρυξη άποψης εφαρμόζονται κατά κύριο λόγο δύο ειδών τεχνικές:

- Τεχνικές βασιζόμενες σε λεξικά
- Τεχνικές βασιζόμενες στην μηχανική μάθηση.

Και στις δύο κατηγορίες αυτές περιλαμβάνονται μηχανισμοί που είναι αρκετά ώριμοι πλέον και έχουν αρκούντως δοκιμαστεί για την αποτελεσματικότητα τους. Για όλες τις σχετικές διαδικασίες παρέχονται αξιόλογα εργαλεία, πολλά από τα οποία είναι διαθέσιμα δωρεάν στο διαδίκτυο. Ένα τέτοιο εργαλείο είναι και το περιβάλλον της γλώσσας R. Πρόκειται για ένα περιβάλλον ανάπτυξης προγραμμάτων σε γλώσσα R το οποίο είναι αρθρωτό και ευέλικτο ικανό να ανταποκρίνεται με κατάλληλες ενέργειες προσαρμογής του, σε οποιοδήποτε είδος μελέτης. Το περιβάλλον αυτό χρησιμοποιήθηκε και στην πρακτική μελέτη που πραγματοποιήθηκε στα πλαίσια της παρούσας εργασίας. Μέσα από την μελέτη των δυνατοτήτων του περιβάλλοντος της γλώσσας R ως προς το αντικείμενο της εργασίας διαπιστώθηκε ότι δίνει την δυνατότητα στους ενεργώντες την να αναπτύξουν διαδικασίες που να μπορούν να επεξεργαστούν δεδομένα του από πηγές του διαδικτύου ακόμα και σε πραγματικό χρόνο. Μειονέκτημα αποτέλεσε η δυσκολία του να διαχειρίζεται μεγάλο όγκο, μεγάλου μήκους διανυσμάτων σε ηλεκτρονικούς υπολογιστές μέσων δυνατοτήτων. Αυτό το γεγονός περιορίζει την αποδοτικότητα κυρίως των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται για την ανάπτυξη των μοντέλων πρόβλεψης. Το περιβάλλον ανάπτυξης της γλώσσας R υποστηρίζεται επαρκώς τόσο από τον φορέα ανάπτυξης και συντήρησης του όσο και από πολυπληθή διαδικτυακή κοινότητα προγραμματιστών που έχει διαμορφωθεί γύρω από αυτό.

Ως προς τα αποτελέσματα της πρακτικής μελέτης που διενεργήθηκε σε δεδομένα που προήλθαν από αναρτήσεις στο tweeter διαπιστώθηκαν τα εξής:

- Οι τεχνικές που βασίζονται σε λεξικά ήταν αλάνθαστες στην εκτιμήσεις της πολικότητας που αναδύεται από τα κείμενα των αναρτήσεων. Αυτό φανερώνει την πληρότητα των λεξικών που έχουν διαχρονικά διαμορφωθεί.
- Οι τεχνικές που βασίζονται σε μηχανική μάθηση, στην προκειμένη περίπτωση δεν αποδείχθηκαν ικανοποιητικά αποδοτικές. Η απόδοση τους κυμάνθηκε γύρω στα επίπεδα του 50%. Αποδοτικότερος αλγόριθμος αποδείχθηκε ο Naïve Bayes.
- Κατά την εφαρμογή των μεθόδων μηχανικής μάθησης καταδείχθηκε ότι η αφαίρεση λέξεων που δεν εμφανίζονται συχνά στις αναρτήσεις επηρέασε ελαφρώς θετικά την απόδοση τους. Ωστόσο όταν η αφαίρεση αυτή ήταν υπερβολική τότε η αποδοτικότητα έπεσε σε ελαφρώς χαμηλότερα επίπεδα.
- Η επεξεργασία των αναρτήσεων (αφαίρεση stop words) στις τεχνικές μηχανικής μάθησης βελτίωσε ελαφρώς την απόδοση τους.
- Η απόδοση των μηχανισμών που βασίζονται σε λεξικά ήταν υψηλότατη σε κάθε περίπτωση επεξεργασίας των κειμένων.
- Αιτία για την αδυναμία των μηχανισμών μηχανικής μάθησης εκτιμάται ότι είναι η αδυναμία δημιουργίας αξιόπιστων μοντέλων πρόβλεψης από τα δεδομένα που διατέθηκαν. Στο φαινόμενο αυτό πιθανόν να διαδραματίζει ρόλο η φύση των αναρτήσεων του Twitter όπου πολλές φορές το νόημα τους μπορεί να ολοκληρώνεται από αντίστοιχες αναρτήσεις που συμμετέχουν σε μία συνομιλία ή που μπορεί να είναι εξαιρετικά σύντομες.

Συνοψίζοντας η ανάλυσης συναισθήματος και η εξόρυξη άποψης έχουν υψηλή δυναμική την οποία εκτιμάται ότι θα διατηρήσουν και στο μέλλον. Αυτό σημαίνει ότι έχουν προταθεί αξιόλογες μεθοδολογίες για τους σκοπούς τους, έχουν καθοριστεί σημαντικές εφαρμογές τους και διατίθενται ισχυρά εργαλεία. Η διατήρηση και η βελτίωση της παρούσας κατάστασης αναμένεται να υποστηριχθεί από την επιστημονική κοινότητα (για μεγάλο εύρος επιστημονικών κλάδων). Αν και οι τεχνικές μηχανικής μάθησης δεν λειτουργήσαν ικανοποιητικά στην παρούσα μελέτη στην βιβλιογραφία αναφέρεται ότι αν συνδυαστούν με τις τεχνικές που βασίζονται σε λεξικά μπορεί να οδηγήσουν σε ικανοποιητικά αποτελέσματα.

Αναφορές

Angiani, G. και συν., 2015. *A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter.* [Ηλεκτρονικό] Available at: http://ceur-ws.org/Vol-1748/paper-06.pdf [Πρόσβαση 4 8 2018].

Brownlee, J., 2017. *A Gentle Introduction to the Bag-of-Words Model.* [Ηλεκτρονικό] Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/ [Πρόσβαση 1 8 2018].

Cohen, W. W., 1995. *Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning.* [Ηλεκτρονικό] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.2612&rep=rep1&type=pdf [Πρόσβαση 1 8 2018].

Faggella, D., 2018. *Crowdsourced Sentiment Analysis – Applications in Social Media and Customer Service.* [Ηλεκτρονικό] Available at: https://www.techemergence.com/crowdsourced-sentiment-analysis-applications-social-media-customer-service/

Gallantra Business Intelligence, 2018. *Sentiment Analysis to Marketing.* [Ηλεκτρονικό] Available at: https://becominghuman.ai/sentiment-analysis-in-marketing-time-to-profit-155b5a1cca7a

Gamallo, P., 2008. *The Meaning of Syntactic Dependencies.* [Ηλεκτρονικό] Available at: https://www.researchgate.net/profile/Pablo_Gamallo/publication/242315755_The_Meaning_of_Syntactic_Dependencies/links/0deec529701202649000000/The-Meaning-of-Syntactic-Dependencies.pdf?origin=publication_detail [Πρόσβαση 1 8 2018].

GAN, G. & YIN, J., 2014. *COMPLEX DATA CLUSTERING: FROM NEURAL NETWORK ARCHITECTURE TO THEORY AND APPLICATIONS OF NONLINEAR DYNAMICS OF PATTERN RECOGNITION.* [Ηλεκτρονικό] Available at: https://pdfs.semanticscholar.org/5840/c44234027ce5a37bf6aa90f00b5edb292254.pdf [Πρόσβαση 18 8 2018].

Hatzivassiloglou, V. & McKeown, K., 1991. *Predicting the Semantic Orientation of Adjectives.* [Ηλεκτρονικό] Available at: http://www.aclweb.org/anthology/P97-1023

Kaur, G. & Chhabra, A., 2014. *Improved J48 Classification Algorithm for the Prediction of Diabetes.* [Ηλεκτρονικό] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.9273&rep=rep1&type=pdf [Πρόσβαση 19 8 2018].

Landwehr, N., Hall, M. & Frank, E., 2004. *Logistic Model Trees.* [Ηλεκτρονικό] Available at: https://www.cs.waikato.ac.nz/~ml/publications/2005/LMT.pdf [Πρόσβαση 20 8 2018].

Llombart, O. R., 2017. *Using Machine Learning Techniques for Sentiment Analysis.* [Ηλεκτρονικό]
Available at: https://ddd.uab.cat/pub/tfg/2017/tfg_70824/machine-learning-techniques.pdf
[Πρόσβαση 12 7 2018].

Loughran, T. & McDonald, B., 2011. *Sentiment Dictionaries.* [Ηλεκτρονικό]
Available at: https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/
[Πρόσβαση 13 8 2018].

MedCalc, 2018. *Logistic regression.* [Ηλεκτρονικό]
Available at: https://www.medcalc.org/manual/logistic_regression.php
[Πρόσβαση 18 8 2018].

Medhat, W., Hassan, A. & Korashyb, H., 2014. *Sentiment analysis algorithms and applications: A survey.* [Ηλεκτρονικό]
Available at: https://www.sciencedirect.com/science/article/pii/S2090447914000550
[Πρόσβαση 12 7 2018].

Minnaar, A., 2015. *Word2Dec Tutorial Part I: The Scip-Gram Model.* [Ηλεκτρονικό]
Available at: http://mccormickml.com/assets/word2vec/Alex_Minnaar_Word2Vec_Tutorial_Part_I_The_Skip-Gram_Model.pdf
[Πρόσβαση 1 8 2018].

Musto, C., Semeraro, G. & Polignano, M., 2015. *A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts.* [Ηλεκτρονικό]
Available at: http://ceur-ws.org/Vol-1314/paper-06.pdf
[Πρόσβαση 12 7 2018].

Musto, C., Semeraro, G. & Polignano, M., 2015. *A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts.* [Ηλεκτρονικό]
Available at: http://ceur-ws.org/Vol-1314/paper-06.pdf
[Πρόσβαση 17 8 2018].

Platt, J. C., 1998. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* [Ηλεκτρονικό]
Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf
[Πρόσβαση 18 8 2018].

Psychol, F., 2017. *A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior.* [Ηλεκτρονικό]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5524892/
[Πρόσβαση 22 8 2018].

repustate, 2018. *Reducing customer churn at a mobile carrier.* [Ηλεκτρονικό]
Available at: https://www.repustate.com/mobile-customer-service-sentiment-analysis-and-text-analytics/

Sayad, S., 2018. *OneR.* [Ηλεκτρονικό]
Available at: http://www.saedsayad.com/oner.htm
[Πρόσβαση 20 8 2018].

Shalev-Shwartz, S. & Ben-David, S., 2014. *Understanding Machine Learning: From Theory to Algorithms.* [Ηλεκτρονικό] Available at: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf [Πρόσβαση 5 8 2018].

Stecanella, B., 2017. *A practical explanation of a Naive Bayes classifier.* [Ηλεκτρονικό] Available at: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/ [Πρόσβαση 10 8 2018].

Taboada, M. και συν., 2014. *Lexicon-Based Methods for Sentiment Analysis.* [Ηλεκτρονικό] Available at: https://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI_a_00049 [Πρόσβαση 12 8 2018].

Text-Analytics, 2014. *What are N-Grams.* [Ηλεκτρονικό] Available at: http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html [Πρόσβαση 1 8 2018].

V.Mäntyläa, M., Graziotinb, D. & Kuutilaa, M., 2017. *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers.* [Ηλεκτρονικό] Available at: https://www.sciencedirect.com/science/article/pii/S1574013717300606 [Πρόσβαση 12 7 2018].

Wu, J., 2015. *Projective Adaptive Resonance Theory Revisited with Applications to Clustering Influence Spread in Online Social Networks.* [Ηλεκτρονικό] Available at: The Fourth International Conference on Data Analytics [Πρόσβαση 18 8 2018].

ΠΑΡΑΡΤΗΜΑ Α

Κώδικας R – Μηχανική Μάθηση

Εφαρμογή του αλγορίθμου Naïve Bayes

```
#ΑΡΧΕΙΑ ΕΓΓΡΑΦΩΝ ΠΟΥ ΘΑ ΧΡΗΣΙΜΟΠΟΙΗΘΟΥΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΤΩΝ ΜΟΝΤΕΛΩΝ ΚΑΙ
ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ

df_1<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2013train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))
```

```
df_1 = mutate(df_1,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_1 <- df_1 %>%

  filter(sentiment != "E")



df_3<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2015train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_3 <- df_3 %>%

  filter(sentiment != "E")



df_4<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2016train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_4 <- df_4 %>%

  filter(sentiment != "E")



df_5<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2013test-A.txt", "\t", col_name = c('id','sentiment','text'))
```

```
df_5 = mutate(df_5,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_5 <- df_5 %>%

  filter(sentiment != "E")



df_6<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2014test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_6 = mutate(df_6,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_6 = mutate(df_6,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_6 = mutate(df_6,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_6 = mutate(df_6,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_6 <- df_6 %>%

  filter(sentiment != "E")



df_7<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2015test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_7 <- df_7 %>%
```

```
  filter(sentiment != "E")



df_8<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2016test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_8 <- df_8 %>%

  filter(sentiment != "E")



#ΔΗΜΙΟΥΡΓΙΑ ΕΝΙΑΙΟΥ ΠΛΑΙΣΙΟΥ ΕΓΓΡΑΦΩΝ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΟΥ

df <-rbind(df_1,df_3,df_4,df_5,df_6,df_7,df_8)

  dim(df)



#ΜΕΤΑΡΟΠΗ ΤΟΥ ΠΕΔΙΟΥ ΠΟΥ ΑΝΤΟΙΧΕΙ ΣΤΟ ΣΥΝΑΙΣΘΗΜΑ ΣΕ ΠΑΡΑΓΟΝΤΑ

df$sentiment <- as.factor(df$sentiment)

#ΚΑΘΑΡΙΣΜΟΣ ΤΩΝ ΠΛΑΙΣΙΩΝ ΔΕΔΟΜΕΝΩΝ

#Remove all text within brackets (e.g. “It’s (so) cool” becomes “It’s
cool”)

#df$text <- bracketX(df$text);



#Replace abbreviations with their full text equivalents (e.g. “Sr” becomes
“Senior”)

df$text <- replace_abbreviation(df$text)

#Convert contractions back to their base words (e.g. “shouldn’t” becomes
“should not”)

df$text <- replace_contraction(df$text)

#Replace common symbols with their word equivalents (e.g. “$” becomes
“dollar”)
```

```
df$text <- replace_symbol(df$text)

#df$text <- removeWords(df$text, stopwords("en"))

corpus <- Corpus(VectorSource(df$text))

corpus.clean <- corpus %>%

  tm_map(content_transformer(tolower)) %>%

  tm_map(removePunctuation) %>%

  tm_map(removeNumbers) %>%

  #tm_map(removeWords, stopwords(kind="en")) %>%

  tm_map(stripWhitespace)


#ΔΗΜΙΟΥΡΓΙΑ ΠΙΝΑΚΑ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕΝΩΝ ΛΕΞΕΩΝ

dtm <- DocumentTermMatrix(corpus.clean)


#ΑΠΟΜΑΚΡΥΝΣΗ ΛΕΞΕΩΝ ΠΟΥ ΔΕΝ ΕΜΦΑΝΙΖΟΝΤΑΙ ΣΥΧΝΑ

Nfreq <- findFreqTerms(dtm.train, 10)

dtm.nb <- DocumentTermMatrix(corpus.clean, control=list(dictionary =
Nfreq))


df.train <- df[12601:15600,] #9670 11000, 1000 3000, 11000 12500

df.test <- df[18201:20000,] #18100 18600, 13000 13500, 30000 30500


dtm.train.nb <- dtm.nb[12601:15600,]

dtm.test.nb <- dtm.nb[18201:20000,]


corpus.clean.train <- corpus.clean[12601:15600]

corpus.clean.test <- corpus.clean[18201:20000]




#ΣΥΝΑΡΤΗΣΗ ΓΙΑ ΤΟΝ ΧΕΙΡΙΣΜΟ ΠΑΡΟΝΤΩΝ ΚΑΙ ΑΠΟΝΤΩΝ ΛΕΞΕΩΝ

convert_count <- function(x) {
```

```
  y <- ifelse(x > 0, 1,0)

  y <- factor(y, levels=c(0,1), labels=c(0, 1))

  y

}


#ΔΗΜΙΟΥΡΓΙΑ ΤΩΝ ΠΛΑΙΣΙΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ
ΤΟΥ ΜΟΝΤΕΛΟΥ

trainNB <- apply(dtm.train.nb, 2, convert_count)

testNB <- apply(dtm.test.nb, 2, convert_count)

#ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ

system.time( classifier <- naiveBayes(trainNB, df.train$sentiment, laplace
= 1) )

#ΑΞΙΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

system.time( pred <- predict(classifier, newdata=testNB) )

#ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ ΠΙΝΑΚΑ ΑΛΗΘΕΙΑΣ

table("Predictions"= pred,  "Actual" = df.test$sentiment )

#ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ CONFUSION MATRIX ΚΑΙ ΠΡΟΒΟΛΕΣ ΤΟΥ

conf.mat <- confusionMatrix(pred, df.test$sentiment)



conf.mat

conf.mat$byClass

conf.mat$overall

conf.mat$overall['Accuracy']
```

Λοιπές Μέθοδοι

```
library(tm)

library(RTextTools)

library(e1071)

library(dplyr)

library(caret)

library(RWeka)
```

```
library(qdap)

library(readr)

library(tidyr)

library(tidytext)



#ΑΡΧΕΙΑ ΕΓΓΡΑΦΩΝ ΠΟΥ ΘΑ ΧΡΗΣΙΜΟΠΟΙΗΘΟΥΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΤΩΝ ΜΟΝΤΕΛΩΝ ΚΑΙ
ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ

df_1<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2013train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_1 = mutate(df_1,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_1 = mutate(df_1,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_1 <- df_1 %>%

  filter(sentiment != "E")



  dim(df_1)



df_3<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2015train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_3 = mutate(df_3,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_3 <- df_3 %>%

  filter(sentiment != "E")
```

```
df_4<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2016train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_4 = mutate(df_4,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_4 <- df_4 %>%

  filter(sentiment != "E")




df_5<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2013test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_5 = mutate(df_5,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_5 = mutate(df_5,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_5 <- df_5 %>%

  filter(sentiment != "E")



df_6<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2014test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_6 = mutate(df_6,sentiment = ifelse(sentiment == "positive", "1",
sentiment))
```

```
df_6 = mutate(df_6,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_6 = mutate(df_6,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_6 = mutate(df_6,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_6 <- df_6 %>%

  filter(sentiment != "E")



df_7<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2015test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_7 = mutate(df_7,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_7 <- df_7 %>%

  filter(sentiment != "E")



df_8<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\
twitter-2016test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "positive", "1",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "negative", "-1",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse(sentiment == "neutral", "0",
sentiment))

df_8 = mutate(df_8,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df_8 <- df_8 %>%

  filter(sentiment != "E")
```

```
  dim(df_1)

  dim(df_3)

  dim(df_4)

  dim(df_5)

  dim(df_6)

  dim(df_7)

  dim(df_8)


#ΔΗΜΙΟΥΡΓΙΑ ΕΝΙΑΙΟΥ ΠΛΑΙΣΙΟΥ ΕΓΓΡΑΦΩΝ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΟΥ

df <-rbind(df_1,df_3,df_4,df_5,df_6,df_7,df_8)

  dim(df)


#ΜΕΤΑΡΟΠΗ ΤΟΥ ΠΕΔΙΟΥ ΠΟΥ ΑΝΤΟΙΧΕΙ ΣΤΟ ΣΥΝΑΙΣΘΗΜΑ ΣΕ ΠΑΡΑΓΟΝΤΑ

df$sentiment <- as.factor(df$sentiment)

#ΚΑΘΑΡΙΣΜΟΣ ΤΩΝ ΠΛΑΙΣΙΩΝ ΔΕΔΟΜΕΝΩΝ

#Remove all text within brackets (e.g. “It’s (so) cool” becomes “It’s
cool”)

df$text <- bracketX(df$text);


#Replace abbreviations with their full text equivalents (e.g. “Sr” becomes
“Senior”)

df$text <- replace_abbreviation(df$text)

#Convert contractions back to their base words (e.g. “shouldn’t” becomes
“should not”)

df$text <- replace_contraction(df$text)

#Replace common symbols with their word equivalents (e.g. “$” becomes
“dollar”)

df$text <- replace_symbol(df$text)

df$text <- removeWords(df$text, stopwords("en"))




xtrain <- 2501
```

```
ytrain <- 7500 #5000

xtest <- 13001

ytest <- 15000 #2000

t <- 5000

s <- 7000


dfSubSet1 <- df[xtrain:ytrain,];

dfSubSet2 <- df[xtest:ytest,];

dfSubSet <- rbind(dfSubSet1, dfSubSet2)



corpus <- Corpus(VectorSource(dfSubSet$text))

corpus.clean <- corpus %>%

  tm_map(content_transformer(tolower)) %>%

  tm_map(removePunctuation) %>%

  tm_map(removeNumbers) %>%

  tm_map(removeWords, stopwords(kind="en")) %>%

  tm_map(stripWhitespace)


#ΔΗΜΙΟΥΡΓΙΑ ΠΙΝΑΚΑ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕΝΩΝ ΛΕΞΕΩΝ

dtm <- DocumentTermMatrix(corpus.clean)


#ΑΠΟΜΑΚΡΥΝΣΗ ΛΕΞΕΩΝ ΠΟΥ ΔΕΝ ΕΜΦΑΝΙΖΟΝΤΑΙ ΣΥΧΝΑ

Nfreq <- findFreqTerms(dtm, 15)

dtm.nb <- DocumentTermMatrix(corpus.clean, control=list(dictionary =
Nfreq))


#df.train.nb <- dfSubSet[xtrain:ytrain,]

#df.test.nb <- dfSubSet[xtest:ytest,]


#ΣΥΝΑΡΤΗΣΗ ΓΙΑ ΤΟΝ ΧΕΙΡΙΣΜΟ ΠΑΡΟΝΤΩΝ ΚΑΙ ΑΠΟΝΤΩΝ ΛΕΞΕΩΝ
```

```
convert_count <- function(x) {

  y <- ifelse(x > 0, 1,0)

  y <- factor(y, levels=c(0,1), labels=c(0, 1))

  y

}



#ΔΗΜΙΟΥΡΓΙΑ ΤΩΝ ΠΛΑΙΣΙΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ
ΤΟΥ ΜΟΝΤΕΛΟΥ

allNB <- apply(dtm.nb, 2, convert_count)

allNBdf = tidy(allNB)

instanceconvertAll <- colnames(allNBdf[])

for (i in instanceconvertAll)

{

  allNBdf[[i]] <- as.numeric(allNBdf[[i]])

}

trainNBdf1 <- allNBdf[1:t,]

s0 <- t+1

testNBdf1 <- allNBdf[s0:s,]



sentiment <- df$sentiment[1:t]

trainNBdf1 <- cbind(sentiment,trainNBdf1)



sentiment <- df$sentiment[s0:s]

testNBdf1 <- cbind(sentiment,testNBdf1)



m1 <- LMT(sentiment ~ ., data = trainNBdf1)

e1 <- evaluate_Weka_classifier(m1,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e1



m2 <- Logistic(sentiment ~ ., data = trainNBdf1)
```

```
e2 <- evaluate_Weka_classifier(m2,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e2



m3 <- SMO(sentiment ~ ., data = trainNBdf1)

e3 <- evaluate_Weka_classifier(m3,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e3



m4 <- J48(sentiment ~ ., data = trainNBdf1)

e4 <- evaluate_Weka_classifier(m4,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e4



m6 <- DecisionStump(sentiment ~ ., data = trainNBdf1)

e6 <- evaluate_Weka_classifier(m6,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e6



m7 <- PART(sentiment ~ ., data = trainNBdf1)

e7 <- evaluate_Weka_classifier(m7,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e7



m8 <- OneR(sentiment ~ ., data = trainNBdf1)

e8 <- evaluate_Weka_classifier(m8,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e8



m9 <- JRip(sentiment ~ ., data = trainNBdf1)

e9 <- evaluate_Weka_classifier(m9,newdata=testNBdf1,numFolds = 10,
complexity = TRUE,class = TRUE)

e9
```

ΠΑΡΑΡΤΗΜΑ Β

Κώδικας R – Χρήση Λεξικών

```
library(tm)

library(RTextTools)

library(e1071)

library(dplyr)

library(caret)

library(RWeka)

library(qdap)

library(readr)

library(tidyR)

library(tidytext)



#ΑΡΧΕΙΑ ΕΓΓΡΑΦΩΝ ΠΟΥ ΘΑ ΧΡΗΣΙΜΟΠΟΙΗΘΟΥΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΤΩΝ ΜΟΝΤΕΛΩΝ ΚΑΙ ΤΗΝ
ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ

df_1<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2013train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_3<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2015train-A.txt", "\t", col_name = c('id','sentiment','text'))

#df_4<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2016train-A.txt", "\t", col_name = c('id','sentiment','text'))

df_5<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2013test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_6<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2014test-A.txt", "\t", col_name = c('id','sentiment','text'))

df_7<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2015test-A.txt", "\t", col_name = c('id','sentiment','text'))
```

```
#df_8<-
read_delim("C:\\Users\\user\\Desktop\\2017_English_final\\GOLD\\Subtask_A\\twitter
-2016test-A.txt", "\t", col_name = c('id','sentiment','text'))



#ΔΗΜΙΟΥΡΓΙΑ ΕΝΙΑΙΟΥ ΠΛΑΙΣΙΟΥ ΕΓΓΡΑΦΩΝ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΟΥ

df <-rbind(df_1,df_3,df_5,df_6,df_7)

df = mutate(df,sentiment = ifelse(sentiment == "positive", "1", sentiment))

df = mutate(df,sentiment = ifelse(sentiment == "negative", "-1", sentiment))

df = mutate(df,sentiment = ifelse(sentiment == "neutral", "0", sentiment))

df = mutate(df,sentiment = ifelse((sentiment != "1")&(sentiment != "-
1")&(sentiment != "0"), "E", sentiment))

df <- df %>%

  filter(sentiment != "E")



  dim(df)



sentLex<-analyzeSentiment(df[1:1000,]$text)

response <- as.numeric(df[1:1000,]$sentiment)

compareToResponse(sentLex, response)

plotSentimentResponse(sentLex, response)
```