ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ
ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Ανάλυση Συναισθήματος σε κείμενα με τεχνικές μηχανικής μάθησης από Twitter και YouTube για το θέμα της μετανάστευσης

(Sentiment Analysis on short texts from Twitter and YouTube concerning immigration using Machine Learning)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλέξανδρος Σπυρίδων Κοτοβός
321/2010080

Επιβλέπων : Μαραγκουδάκης Εμμανουήλ
Αναπληρωτής Καθηγητής

Σάμος, Φεβρουάριος 2018

Αλέξανδρος – Σπυρίδων Κοτοβός

Διπλωματούχος Μηχανικός Πληροφοριακών και Επικοινωνιακών Συστημάτων

Πανεπιστημίου Αιγαίου

Η παρούσα διπλωματική εργασία αφορά υλοποίηση που έγινε με βάση μια αποκλειστική ιδέα του επιβλέποντα. Ο κώδικας για τη συλλογή δεδομένων, η δημιουργία της ιστοσελίδας και το παρόν κείμενο αποτελεί 100% δική μου εργασία. Δηλώνω υπεύθυνα ότι δεν έχω κάνει καμία ενέργεια λογοκλοπής και όπου χρειάστηκε να χρησιμοποιήσω δουλειά άλλων ερευνητών το έχω αναφέρει με σαφήνεια μέσα στο κείμενο.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Πανεπιστημίου Αιγαίου.

Περίληψη

Εκατοντάδες εκατομμύρια ανθρώπων καθημερινά χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης προκειμένου να ανταλλάξουν μηνύματα, να σχολιάσουν κάποιο θέμα και να εκφράσουν τις απόψεις τους σχετικά με την επικαιρότητα. Το φαινόμενο αυτό προσελκύει πολύ μεγάλο ενδιαφέρον στον τομέα της ανάλυσης των συναισθημάτων που κρύβονται πίσω από αυτές τις αλληλεπιδράσεις. Η ραγδαία εξέλιξη του Internet παρέχει στην ερευνητική κοινότητα ένα τεράστιο όγκο δεδομένων που μπορούν να χρησιμοποιηθούν για τη καλύτερη κατανόηση των αναγκών των χρηστών αλλα και την άντληση πληροφοριών με σκοπό να επιτευχθεί ένα γενικό συμπέρασμα για διάφορα θέματα πολιτικής, κοινωνικής, οικονομικής φύσεως και όχι μόνο.

Η ανάλυση των συναισθημάτων (sentiment analysis), γνωστή και ως εξόρυξη γνώμης (opinion mining) είναι ο τομέας της επιστήμης που μελετά αξιολογήσεις, συναισθήματα, εκτιμήσεις, γνώμες, τις στάσεις και τα συναισθήματα των ανθρώπων προς οντότητες, όπως τις υπηρεσίες, οργανισμούς, προϊόντα, άτομα, τα διάφορα κοινωνικά θέματα και τα χαρακτηριστικά τους. Ο τομέας της ανάλυσης των συναισθημάτων ως τομέας της έρευνας είναι μεγάλος και σχετικά νέος και ανεξερεύνητος. Η πρώτη ίσως αναφορά του όρου εντοπίζεται στο όχι και τόσο μακρινό, για τα ερευνητικά δεδομένα 2003 (Jeonghee, 2003).

Το έργο της εξαγωγής συμπερασμάτων από αλληλεπιδράσεις χρηστών σε διαδικτυακά κοινωνικά δίκτυα γίνεται ακόμα πιο δύσκολο αν αναλογιστούμε το εύρος των διαθέσιμων γλωσσών και διαλέκτων και την έλλειψη οργανωμένων και ολοκληρωμένων συνόλων δεδομένων, σε συνδυασμό με τη πληθώρα των συναισθημάτων που μπορεί να κρύβονται σε ένα μικρό κείμενο 140 χαρακτήρων (Twitter) ή σε ένα σχόλιο σε ένα video (YouTube).

Στόχος της παρούσας διπλωματικής είναι η σχεδίαση και υλοποίηση ενός συνόλου εφαρμογών που θα δίνουν την δυνατότητα στον χρήστη να αναζητά συγκεκριμένες “λέξεις-κλειδιά” μέσα σε tweets και σχόλια από συγκεκριμένα βίντεο που τον ενδιαφέρουν στο YouTube, στη συνέχεια θα τα αποθηκεύει σε μία βάση δεδομένων (tweets και σχόλια) και τέλος θα τα κατηγοριοποιεί είτε σε θετικά είτε σε αρνητικά. Πρόκειται λοιπόν για ένα ολοκληρωμένο σύστημα αξιολόγησης σύντομων κειμένων που περιλαμβάνει μία εφαρμογή για την αναζήτηση και αποθήκευση των επιθυμητών κειμένων, μία εφαρμογή που διαχωρίζει τα ήδη βαθμολογημένα κείμενα σε θετικά και αρνητικά για την εκπαίδευση του μοντέλου και ένα σύστημα που φέρνει τα αποθηκευμένα κείμενα από την βάση δεδομένων και εφαρμόζει πάνω τους ένα μοντέλο αναγνώρισης συναισθημάτων.

Abstract

Millions of people around the world use, on a daily basis, social networking sites in order to exchange texts, comment on a subject and express their opinion about timeliness. This phenomenon has attracted the interest of the research community in the realm of sentiment analysis of these interactions. The evolution of the Internet produces a huge amount of data that can be used to better understand the user needs and to gain information in order to achieve a general conclusion for political, social, economic matters and many more.

Sentiment analysis, also known as opinion mining is the area of science that studies the evaluations, feelings, assessments, opinions, attitudes and feelings of the people to entities such as services, organizations, products, people, social issues and their characteristics. The analysis of emotions as an area of research, is enormous and relatively new and unexplored. It is possible that the first mention of the term was not a long time ago concerning research data (Jeonghee, 2003).

The task of inference from user interactions in online social networks becomes even harder considering the range of available languages and dialects and the lack of organized and comprehensive data sets in combination with the variety of emotions that can be hidden in a short text of 140 characters (Twitter) or a comment in a YouTube video.

The aim of this thesis is to design and implement a set of applications that will enable the user to search certain keywords in tweets and comments from YouTube videos of his interest, store them in a database (tweets and comments) and categorize them either positively or negatively. It is therefore an integrated short text evaluation system that includes a search and storage application for the desired texts, an application that separates the already graded texts into positive and negative for model training and an application that fetches the stored texts from the database and applies a sentiment analyzing model on them.

Ευχαριστίες - Αφιερώσεις

Αρχικά οφείλω να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα αναπληρωτή Καθηγητή κ.Μαραγκουδάκη Εμμανουήλ, που μου επέτρεψε να εργαστώ πάνω στο θέμα της διπλωματικής εργασίας που επέλεξα.

Θερμές και ιδιαίτερες ευχαριστίες θέλω να δώσω στον πρώην συμφοιτητή και νυν συνάδελφο Ιωάννη Παρασκάκη για την καθοριστική του συμβολή στην παρούσα διπλωματική. Η βοήθεια του ήταν διαρκής καθ' όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας και η συνεργασία μας άψογη.

Τη διπλωματική αυτή εργασία αφιερώνω στην οικογένειά μου για την διαρκή στήριξη όλα τα χρόνια των σπουδών μου.

Τέλος θα ήθελα να ευχαριστήσω τους Μουλασιώτη Λάζαρο, Κυριάκο Παναγιώτη, Πιτσέλη Ελευθέριο, Παπασπύρο Παναγιώτη, Μπότο Στέργιο, Τσιπά Νικόλαο, Πορή Δημήτριο, Καγιάννη Κωνσταντίνο, Γιαννιώτη Χρήστο, Μπουλούμπαση Δημοσθένη, Στρατή Αντώνιο καθώς και τον πνευματικό αδερφό μου Κουρούνη Γεώργιο για την ηθική υποστήριξή τους όλα αυτά τα χρόνια καθώς και για τις ατελείωτες ώρες που περάσαμε μαζί κατά την διάρκεια των φοιτητικών μας σπουδών.

Περιεχόμενα

Κατάλογος Εικόνων

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή στην έννοια της εξόρυξης γνώμης

Στη σύγχρονη εποχή παρατηρείται άνθιση των ανοιχτών εργαλείων επεξεργασίας φυσικής γλώσσας και των ερευνητικών προγραμμάτων. Οι πρόσφατες εξελίξεις τόσο στο θεωρητικό υπόβαθρο όσο και στις αναπαραστάσεις της γλώσσας λειτουργούν ως η κινητήριος δύναμη που θα συνδράμει την κατανόηση της γλώσσας και την περαιτέρω εκβιομηχάνιση του γλωσσικού τοπίου. Η συνεργασία διαφορετικών επιστημονικών πεδίων όπως η μηχανική μάθηση, η γνωσιακή ψυχολογία και η υπολογιστική γλωσσολογία καθιστούν αναγκαία την αλλαγή των καθιερωμένων τρόπων έρευνας και δραστηριοποίησης.

Η εξόρυξη γνώμης (opinion mining) είναι ένας τύπος επεξεργασίας φυσικής γλώσσας για την παρακολούθηση της γνώμης του κοινού σχετικά με ένα συγκεκριμένο θέμα ή προϊόν. Η εξόρυξη γνώμης, η οποία είναι σχεδόν όμοια με την ανάλυση συναισθημάτων, περιλαμβάνει την οικοδόμηση ενός συστήματος συλλογής και κατηγοριοποίησης απόψεων σχετικά με ένα θέμα ή προϊόν.

1.2 Χρησιμότητα της εξόρυξης γνώμης

Η ανάλυση των σχολίων που αναρτώνται καθημερινά στο διαδίκτυο είναι εξαιρετικά σημαντική καθώς επιτυγχάνουμε γρήγορη ανάλυση της γλώσσας, ερμηνεία σε σχόλια,ψηφιακή επεξεργασία, χαρτογράφηση και οπτικοποίηση. Ωστόσο, το πιο σημαντικό χαρακτηριστικό είναι η ικανότητά μας να κατανοούμε ποσοτικά την πληροφορία σε μια κλίμακα που μέχρι πρότινος ήταν αδιανόητη. Για παράδειγμα, δεδομένου του τεράστιου όγκου δεδομένων των ακαδημαϊκών δημοσιεύσεων που διατίθενται σήμερα, θα χρειαζόταν αρκετό χρόνο σε ένα ερευνητή να αναλύσει όλες τις σχετικές γνώμες για ένα συγκεκριμένο πρόβλημα. Η χρήση εξόρυξης κειμένου θα μπορούσε να μειώσει δραστικά τον απαιτούμενο χρόνο.

Ένα βασικό πλεονέκτημα της εξόρυξης γνώμης είναι ότι επιτρέπει την αποτελεσματικότερη ανάλυση της υπάρχουσας γνώσης. Η δυνατότητα εξαγωγής πληροφοριών, μειώνει αυτόματα τον χρόνο που αφιερώνεται στην εξασφάλιση των απαραίτητων γνώσεων σε ένα ευρύ φάσμα πηγών ηλεκτρονικής έρευνας, που κανονικά γίνεται με τη χρονοβόρα διαδικασία ανασκόπησης της βιβλιογραφίας. Η αποτελεσματικότητα που επιτυγχάνεται ‘’ξεκλειδώνοντας’’ πληροφορίες μπορεί να οδηγήσει σε ευρύτερη γνώση και βαθύτερη κατανόηση.

Οι άνθρωποι πλέον, εκφράζουν επί το πλείστον τη γνώμη τους σχετικά με προϊόντα, την οργάνωση, την υγειονομική τους κατάσταση, τις υπηρεσίες του δημόσιου και ιδιωτικού τομέα καθώς και κοινωνικά φαινόμενα χωρίς κανένα δισταγμό με αποτέλεσμα να δημιουργείται μία τεράστια “δεξαμενή” πληροφορίας, δηλαδή δεδομένων, τα οποία βοηθούν τους ειδικούς να εξάγουν χρήσιμα συμπεράσματα. Αυτά τα συμπεράσματα μπορεί να είναι κοινωνιολογικής,

ιστορικής, πολιτιστικής, πολιτικής σημασίας και να αφορούν την άποψη των πολιτών σε μείζοντα επίκαιρα θέματα ή και σε γενικότερα διαχρονικά θέματα.

1.3 Σκοπός και στόχος της διπλωματικής

Σκοπός της διπλωματικής εργασίας είναι να μπορέσει να αξιοποιηθεί περισσότερο η τεχνολογία στον τομέα της αναγνώρισης συναισθημάτων καθώς αποτελεί ένα ορόσημο στη σύγχρονη εποχή και συνδέεται άμεσα με το μέλλον. Στην ουσία σκοπός μας είναι να υπάρχει όσο το δυνατόν πιο ασφαλής εξαγωγή συμπεράσματος ως προς το συναίσθημα.

Στόχος της παρούσας διπλωματικής είναι η σχεδίαση και υλοποίηση ενός συνόλου εφαρμογών που θα δίνουν την δυνατότητα στον χρήστη να αναζητά συγκεκριμένες "λέξεις-κλειδιά" μέσα σε tweets και σχόλια από συγκεκριμένα βίντεο που τον ενδιαφέρουν στο YouTube και θα τα κατηγοριοποιεί είτε σε θετικά είτε σε αρνητικά με ένα υψηλό ποσοστό ακρίβειας.

1.4 Διάρθρωση της διπλωματικής

Η παρούσα διπλωματική αποτελείται από 7 βασικά κεφάλαια.

Στο πρώτο κεφάλαιο κάνουμε μία γενική εισαγωγή στην έννοια και χρησιμότητα της εξόρυξης δεδομένων, αναλύουμε τον σκοπό και τον στοχο της διπλωματικής και αναφερόμαστε στα κοινωνικά δίκτυα.

Στο δεύτερο κεφάλαιο αναφερόμαστε στην εξόρυξη δεδομένων, εξηγούμε τους τρόπους μηχανικής μάθησης και στις τεχνικές εξόρυξης δεδομένων και στην συνέχεια εστιάζουμε στην κατηγοριοποίηση (classification).

Στο τρίτο κεφάλαιο εξηγούμε τα οφέλη και τις δυσκολίες της επεξεργασίας φυσικής γλώσσας, αναλύουμε τι είναι η εξόρυξη γνώμης (opinion mining) και η ανάλυση συναισθήματος (sentiment analysis) και αναφέρουμε τα βασικά στάδια προεπεξεργασίας των δεδομένων.

Στο τέταρτο κεφάλαιο συνοψίζουμε την πορεία της ερευνητικής κοινότητας στον τομέα της ανάλυσης συναισθήματος (sentiment analysis) και της εξόρυξης γνώμης (opinion mining) με χρήση μηχανισμών μηχανικής μάθησης και εξηγούμε τους λόγους που αποφασίσαμε να αναπτύξουμε αυτήν την εφαρμογή προκειμένου να συνεισφέρουμε στον τομέα αυτό.

Στο πέμπτο κεφάλαιο αναφερόμαστε στην μετανάστευση, αναλύουμε τα αίτια που συντελούν στην ύπαρξή της και παρουσιάζουμε την εικόνα της στις μέρες μας.

Στο έκτο κεφάλαιο περιγράφουμε αναλυτικά το συνολικό των εφαρμογών που υποστηρίζουν την αναγνώριση των συναισθημάτων και παρέχουμε κομμάτια κώδικα για την κατανόηση των διαδικασιών. Επίσης εξηγούμε την προσαρμογή μοντέλων αναγνώρισης συναισθημάτων στην υλοποίησή μας. Τέλος εξηγούμε τον τρόπο με τον οποίο ανιχνεύουμε και αποθηκεύουμε τα tweets και τα σχόλια απο τα videos του YouTube που διαλέγουμε.

Στο έβδομο κεφάλαιο της εργασίας μας παρουσιάζουμε συγκεντρωτικά τα αποτελέσματα της χρήσης της εφαρμογής σε πραγματικό χρόνο και αναλύουμε τα συμπεράσματα που προέκυψαν κατά την υλοποίηση.

1.5 Κοινωνικά δίκτυα

Τα κοινωνικά δίκτυα είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Ο όρος σήμερα χρησιμοποιείται επίσης για να περιγράψει ιστοσελίδες οι οποίες επιτρέπουν την διεπαφή ανάμεσα στους χρήστες, πχ. με σχόλια, φωτογραφίες, άλλες πληροφορίες από σχετική βιβλιογραφία. Οι πιο γνωστές από αυτές τις ιστοσελίδες είναι το Facebook, Twitter, Instagram και Linkedin.

Οι ιστότοποι αυτοί αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνούν και να αναπτύσσουν επαφές μέσα από αυτές.

Ένα κοινωνικό δίκτυο είναι μια κοινωνική δομή που αποτελείται από ένα σύνολο παραγόντων, όπως άτομα ή οργανισμούς. Στο διαδίκτυο, τα κοινωνικά δίκτυα είναι μία πλατφόρμα που συντηρείται για την δημιουργία κοινωνικών σχέσεων μεταξύ των ανθρώπων, που συνήθως αποτελούν ενεργά μέλη του κοινωνικού δικτύου, με κοινά ενδιαφέροντα ή δραστηριότητες.

Οι ιστότοποι κοινωνικής δικτύωσης είναι οργανωμένες ιστοσελίδες στο διαδίκτυο με περισσότερο ομαδοκεντρικό χαρακτήρα που παρέχουν, στην συντριπτική τους πλειοψηφία, μία σειρά από βασικές και δωρεάν υπηρεσίες όπως τη δημιουργία προφίλ, το ανέβασμα εικόνων και βίντεο, τον σχολιασμό σε ενέργειες που γίνονται από άλλα μέλη του δικτύου ή μίας ομάδας, την άμεση ανταλλαγή μηνυμάτων και πολλά άλλα.

Κεφάλαιο 2ο: Σχετική Βιβλιογραφία

Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques

Το 2003 οι Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu και Wayne Niblacki δημοσιεύουν την μελέτη τους σχετικά με τον "Αναλυτή Συναισθημάτων" (Sentiment Analyzer) που είχαν δημιουργήσει. Πρόκειται για μία από τις πρώτες ερευνητικές αναφορές στην επιστήμη του Sentiment Analysis και Opinion Mining. Ο εν λόγω αναλυτής κατηγοριοποιούσε τα συναισθήματα ανάλογα με το εκάστοτε θέμα ενός κειμένου χρησιμοποιώντας τεχνολογίες επεξεργασίας φυσικής γλώσσας NLP. Η εργασία τους χωριζόταν σε τρία επιμέρους συστήματα τα οποία ήταν 1) Ο αναλυτής της θεματικής ενότητας, 2) Το σύστημα εξόρυξης συναισθημάτων και 3) Ο αναλυτής συναισθημάτων που συνδύαζε τα 2 προαναφερθέντα υποσυστήματα. Ο εν λόγω αναλυτής εφαρμόστηκε σε κείμενα αξιολόγησης προϊόντων. Τα αποτελέσματα της έρευνας ήταν ιδιαίτερα θετικά και ενθαρρυντικά για την ερευνητική κοινότητα καθώς επιτεύχθηκε ποσοστό ακρίβειας που άγγιζε το 91.0% με 93.0%. Κλείνοντας τη δημοσίευση τους οι Nasukawa et al. σημειώνουν τη σημασία του ανθρώπινου παράγοντα (expert) στη διαδικασία της επικύρωσης των αποτελεσμάτων (validation process) και αναφέρουν πως ένα από τα θέματα που θα μπορούσε να αποτελέσει μελλοντική εργασία είναι η περαιτέρω αυτοματοποίηση αυτής της διαδικασίας.

Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis

Το 2005 οι Theresa Wilson, Janyce Wiebe και Paul Hoffmann δημοσιεύουν την μελέτη τους που παρουσιάζει μία νέα προσέγγιση στην phrase-level ανάλυση συναισθήματος η οποία αρχικά καθορίζει αν μία έκφραση είναι ουδέτερη ή πολική (θετική ή αρνητική) και στη συνέχεια αποσαφηνίζει την πολικότητα των πολικών εκφράσεων. Παρουσιάζουν νέες τεχνικές όσον αφορά την βάση-περιεχομένου πολικότητα. Συγκεκριμένα αρχίζοντας με ένα μεγάλο σύνολο στοιχείων που είναι ήδη αναγνωρισμένα ως προς την πολικότητά τους, προσδιορίζουν την βάση-περιεχομένου πολικότητα των φράσεων που περιέχουν στιγμιότυπα των συγκεκριμένων στοιχείων. Χρησιμοποιούν μία διαδικασία δύο σταδίων που αφορούν τη μηχανική μαθηση και μία ποικιλία χαρακτηριστικών. Το πρώτο βήμα ταξινομεί κάθε φράση που περιέχει μια ένδειξη ως ουδέτερη ή πολική. Το δεύτερο βήμα λαμβάνει όλες τις φράσεις που έχουν χαρακτηριστεί στο πρώτο βήμα ως πολικές και αποσαφηνιζει την βάση-περιεχομένου πολικότητά τους. Καταλήγοντας σημειώνουν πως με αυτό τον τρόπο το σύστημα είναι σε θέση να προσδιορίσει αυτόματα την βάση-περιεχομένου πολικότητα για ένα πολύ μεγάλο υποσύνολο απο sentiment expressions επιτυγχάνοντας σημαντικά βελτιωμένα αποτελέσματα σε σχέση με τον μέσο όρο.

A Review of Opinion Mining Methods for Analyzing Citizens' Contributions in Public Policy Debate

Το 2011 οι Εμμανουήλ Μαραγκουδάκης, Ευριπίδης Λουκής και Ιωάννης Χαραλαμπίδης εξετάζουν ένα ευρύ φάσμα μεθόδων εξόρυξης γνώσης που έχουν αναπτυχθεί για την ανάλυση των απόψεων που αφορούν εμπορικά προϊόντα και των αξιολογήσεων που δημοσιεύονται στο διαδίκτυο, όπως και τις προοπτικές που μπορούν να προσφέρουν. Προτείνουν μία βασική δομή για την χρησιμοποίηση τους που αποτελείται από πέντε στάδια: 1) Ταξινόμηση κάθε συγκεκριμένης δημοσίευσης στην υπό συζήτηση πολιτική/απόφαση σε θετική, ουδέτερη ή αρνητική, χρησιμοποιώντας document-level ανάλυση συναισθήματος 2) Προσδιορισμός για κάθε δημοσίευση των υποκειμενικών προτάσεων (εκφραζόμενες απόψεις) και ταξινόμησή τους σε σε θετικές, ουδέτερες ή αρνητικές υπολογίζοντας στην συνέχεια την σχετική συχνότητα εμφάνισης τους 3) Σύγκριση και ενσωμάτωση των ευρημάτων των παραπάνω σταδίων 1,2 καθώς και των ευρημάτων από την ανάλυση άλλου τύπου feedback (e-votes) που επιτρέπει την εξαγωγή συμπερασμάτων ως προς το γενικό συναίσθημα του πολίτη σχετικα με συγκεκριμέμη πολιτική/απόφαση 4) Με περαιτέρω επεξεργασία όλων των δημοσιεύσεων προσδιόρση των κύριων ζητημάτων που τέθηκαν και σχολιάστηκαν από τους πολίτες και 5) Για κάθε θέμα που προκύπτει ταξινόμηση των προτάσεων που το περιέχουν σε θετική, ουδέτερη ή αρνητική χρησιμοποιώντας sentence-level μεθόδους ανάλυσης συναισθήματος-προσδιορισμού απόψεων και υπολογίζοντας εκ νέου την σχετική συχνότητα θετικών, ουδέτερων ή αρνητικών υποκειμενικών προτάσεων. Καταλήγουν πως τόσο οι παραδοσιακές όσο και οι αναδυόμενες eParticipation φόρμες οδηγούν στην παραγωγή τεράστιων ποσοτήτων από δημοσιεύσεις πολιτών που αφορούν πολιτικές/αποφάσεις.

Opinion Mining and Sentiment Analysis in Policy Formulation Initiatives: The EU-Community Approach

Το 2015 οι Ιωάννης Χαραλαμπίδης, Εμμανουήλ Μαραγκουδάκης και Ευριπίδης Λουκής δημοσιεύουν το συγκεκριμένο άρθρο και κάνουν μία προσέγγιση σε ICT-based μεθόδους και ειδικότερα σε Opinion mining και Sentiment Analysis τεχνικές προκειμένου να επεξεργαστούν εκτεταμένο πολιτικό περιεχόμενο που συλλέχθηκε από ένα μεγάλο πλήθος πηγών. Συγκεκριμένα χωρίζουν τις ICT μεθόδους σε τέσσερις γενιές: 1) Η πρώτη γενιά περιλαμβάνει τη δημιουργία ιστοτόπων και λογαριασμών κοινωνικών μέσων των κρατικών υπηρεσιών 2) Η δεύτερη γενιά περιλαμβάνει μεθόδους που ανακτούν αυτόματα το πολιτικό περιεχόμενο από διάφορες πηγές (διάφορους λογαριασμούς κοινωνικών μέσων και ιστοτόπους) χρησιμοποιώντας τα APIs τους 3) Η τρίτη γενιά περιλαμβάνει μεθόδους που προσανατολίζονται προς την αυτόματη ανάκτηση εξωτερικού περιεχομένου χρησιμοποιώντας εκτός από τα APIs τους και πιο προηγμένη επεξεργασία και 4) Η τέταρτη γενιά επικεντρώνεται στην ανάκτηση και επεξεργασία περιεχομένου υψηλής ποιότητας που δημιουργήθηκε από εμπειρογνώμονες. Συνοψίζοντας καταλήγουν πως παρόλο που τα πρώτα αποτελέσματα είναι ενθαρρυντικά

χρειάζεται περαιτέρω αξιολόγηση χρησιμοποιώντας περισσότερα δεδομένα πραγματικής ζωής η οποία θα εντοπίσει προτερήματα και αδυναμίες και θα οδηγήσει σε βελτιώσεις.

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Το 2010 οι Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge και Noah A. Smith δημοσιεύουν τη μελέτη τους στην οποία συνδυάζουν αποτελέσματα εκλογών με την επεξεργασία και ανάλυση κειμένων από το Twitter και άλλες micro-blogging εφαρμογές. Στην εν λόγω δημοσίευση οι ερευνητές χρησιμοποιούν αποτελέσματα τηλεφωνικών και άλλων ερευνών σχετικά με πολιτικά θέματα όπως την εκλογή του προέδρου των Ηνωμένων πολιτειών (εκλογές 2009 ανάμεσα σε Obama και McCain) αλλά και μηνυμάτων χρηστών στο Twitter και σε άλλα παρεμφερή μέσα σε συνδυασμό με τα δημογραφικά στοιχεία που τα εν λόγω συστήματα προσφέρουν. Τα αποτελέσματα αυτής της πρωτότυπης έρευνας τόσο σε ανάλυση κειμένου όσο και στην ανάλυση προβλέψεων αναδεικνύουν την ανάγκη για τη δημιουργία συστημάτων ανάλυσης συναισθημάτων που απαντούν σε συγκεκριμένες ερωτήσεις. Τέλος οι συγγραφείς τονίζουν τη σημασία των στοιχείων που παρέχουν οι, κάθε είδους έρευνες (surveys), στην εξέλιξη του τομέα της ανάλυσης συναισθημάτων και στη δημιουργία πιο εκλεπτυσμένων εφαρμογών.

Lexical Normalization for Social Media Text

Το 2013 οι Han, B., Cook, P., and Baldwin, T. (Han, 2013) δημοσιεύουν τη μελέτη τους σχετικά με τις μεθόδους κανονικοποίησης των λεξικών και την αναγνώριση λέξεων και παραλλαγών που θα μπορούσαν να παρεμποδίσουν τη λειτουργικότητα των τεχνικών της επεξεργασίας φυσικής γλώσσας. Η κανονικοποίηση του κειμένου είναι μία πολύπλοκη διαδικασία και μοιάζει με εκείνη του ορθογραφικού ελέγχου αλλά διαφέρει στο γεγονός ότι συνήθως γίνεται σκόπιμα για λόγους ευκολίας γραφής ή περιορισμών του συστήματος (140 χαρακτήρες όριο για κάθε Tweet). Η διαδικασία της σύνθεσης της λέξης από συντομογραφία (deabbreviation - π.χ. b4 σε before), της ανάκτησης της λέξης ύστερα από ελεύθερη γραφή (π.χ. goooooood σε good) και η αναγνώριση παραλλαγών σε λέξεις που δεν ανήκουν στα υπάρχοντα λεξικά καθιστούν το αντικείμενο αυτής της έρευνας ακόμα πιο δύσκολο. Στη μακροσκελή δημοσίευση τους οι Han et al. παρουσιάζουν τα αποτελέσματα της δοκιμής διάφορων τεχνικών κανονικοποίησης αλλά και μία συγκριτική μελέτη ανάμεσα σε υπάρχοντα συστήματα και στην προτεινομένη προσέγγιση. Τα αποτελέσματα της μελέτης είναι ιδιαίτερα ενθαρρυντικά καθώς η προτεινόμενη λύση υπερτερεί όλων των άλλων λύσεων σε όλους τους τομείς της συγκριτικής αξιολόγησης.

Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Το 2010 οι Alexander Pak, Patrick Paroubek *(Alexander Pak, 2010)* δημοσιεύουν τη μελέτη τους στην οποία περιγράφουν τη διαδικασία της αυτοματοποιημένης συλλογής των Tweets προκειμένου να δημιουργήσουν ένα σώμα δεδομένων (corpus) για την ανάλυση συναισθήματος και τους σκοπούς της εξόρυξης γνώμης. Στη μελέτη αυτή οι συγγραφείς πραγματοποιούν γλωσσολογική ανάλυση των συλλεχθέντων δεδομένων και εξηγούν τις ανακαλύψεις τους. Χρησιμοποιώντας το σώμα (corpus), οι συγγραφείς δημιούργησαν ένα ταξινομητή συναισθημάτων, που είναι σε θέση να προσδιορίσει θετικά, αρνητικά και ουδέτερα συναισθήματα για ένα κείμενο. Πειραματικές αξιολογήσεις και δοκιμές δείχνουν ότι οι προτεινόμενες τεχνικές που παρουσιάζονται στην εν λόγω μελέτη είναι αποδοτικές και μάλιστα πιο αποδοτικές από παλαιότερες μεθόδους. Στην παρούσα έρευνά η επιλεγμένη γλώσσα ήταν τα αγγλικά , ωστόσο, η προτεινόμενη τεχνική μπορεί να χρησιμοποιηθεί με οποιαδήποτε άλλη γλώσσα.

Sentiment Analysis of Short Informal Texts

Το 2014 οι Svetlana Kiritchenko, Xiaodan Zhu και Saif M. Mohammad (*Svetlana Kiritchenko, 2014)* παρουσιάζουν τη μελέτη τους στην οποία περιγράφουν ένα state-of-the-art σύστημα ανάλυσης συναισθημάτων σε σύντομα μηνύματα όπως SMSs και Tweets τόσο σε επίπεδο μηνύματος όσο και σε επίπεδο όρων (message-level task και term-level task). Το εν λόγω σύστημα βασίζεται σε στατιστική κατηγοριοποίηση με επίβλεψη (supervised statistical text classification) το οποίο εκμεταλλεύεται μία πληθώρα σημασιολογικών και συναισθηματικών λειτουργιών με τη χρήση νέων, ειδικών λεξικών μεγάλης κάλυψης. Τα εν λόγω λεξικά δημιουργούνται αυτόματα από τα Tweets και μπορούν να περιέχουν hashtags (θεματικές ενότητες) και emoticons. Το σύστημα που προέκυψε από αυτή τη μελέτη και διακρίθηκε στο Conference on Semantic Evaluation Exercises (SemEval-2013) *(Saif M. Mohammad, 2013)* αποδεικνύει ότι τεχνικές αντιστροφής της πολικότητας των λέξεων δεν είναι πάντα ακριβείς ούτε κατάλληλες για τη διαδικασία της ανάλυσης των συναισθημάτων. Πιο συγκεκριμένα η εν λόγω μελέτη αποδεικνύει ότι όταν οι θετικοί όροι αναιρούνται (αντιστρέφονται), τείνουν να μεταφέρουν ένα αρνητικό συναίσθημα. Αντίθετα, όταν οι αρνητικοί όροι αναιρούνται (αντιστρέφονται), τείνουν να εξακολουθούν να μεταφέρουν ένα αρνητικό συναίσθημα. Επιπλέον, η ένταση αξιολόγησης για τόσο τους θετικούς όσο και για τους αρνητικούς όρους αλλάζει στο πλαίσιο τις αντιστροφής, και το ποσό της μεταβολής κυμαίνεται από όρο σε όρο. Προκειμένου να καταλάβουμε επαρκώς τις επιπτώσεις της αντιστροφής για μεμονωμένους όρους, οι συγγραφείς προτείναν της εμπειρική εκτίμηση των αποτελεσμάτων της ανάλυσης και τη δημιουργία δύο λεξικών για όρους σε αρνητικό πλαίσιο και σε θετικό πλαίσιο αντίστοιχα. Το εν λόγω σύστημα μπορεί να διαχειριστεί 100 Tweets ανά δευτερόλεπτο ενώ έχει καταφέρει να αξιολογήσει 135 εκατομμύρια Tweets σε ένα cluster 50 υπολογιστών σε 11 ώρες λειτουργίας.

Sentiment Analysis on YouTube: A Brief Survey

Το 2015 οι Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat και Fazal Masud Kundi επικεντρώνονται στην ακολούθηση των εξής πιθανών προβλημάτων με σκοπό να βρουν την πολικότητα των σχολίων που δημοσιεύουν οι χρήστες του YouTube: 1) Τους περιορισμούς των τρέχοντων λεξικών συναισθήματος 2) Την χρήση της μη επίσημης γραφής (slang) από τους χρήστες 3) Την εκτίμηση της πολικότητας των συναισθημάτων απο τους community-created όρους 4) Την αντιστοίχιση των κατάλληλων ετικετών στα συμβάντα (events) 5) Την επίτευξη ικανοποιητικών αποτελεσμάτων στο classification και 6) Προκλήσεις που περιλαμβάνει η ανάλυση συναισθημάτων στα μέσα κοινωνικής δικτύωσης.Επίσης εξετάζονται και διαφορετικές τεχνικές για την αναγνώριση της πολικότητας των σχολίων π.χ User Sentiment Detection. Επιπλέον υποστηρίζουν πως μελλοντικά η βελτίωση του κοινωνικού λεξικού και η σωστή ταξινόμηση των συμβάντων μπορεί να βοηθήσει στην αύξηση της απόδοσης για την πρόβλεψη της πολκότητας των σχολίων. Συνοψίζοντας καταλήγουν πως η ανίχνευση της πολικότητας των συναισθημάτων των σχολίων του χρήστη στο YouTube παρά την πολύ δουλειά που γίνεται είναι μία εργασία που προκαλεί αρκετές δυσκολίες στους ερευνητές μέχρι στιγμής.

How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings

Το 2010 οι Stefan Siesdorfer, Sergiu Chelaru, Wolfgang Nejdi και Jose San Pedro παρουσιάζουν μια διεξοδική μελέτη σχολιασμού και συμπεριφοράς βαθμολόγησης σχολίων από χρήστες του YouTube σε ένα δείγμα περισσότερων απο 67.000.000 σχολίων από 67.000 videos. Επιπλεόν μελέτησαν την επιρροή του συναισθήματος που εκφράζεται στα σχόλια στις αξιολογήσεις των σχολίων χρησιμοποιώντας το SentiWordNet. Η συγκεκριμένη μελέτη είχε αρκετούς συμπληρωματικούς στόχους.Από την μία πλευρά, μελετούν τη βιωσιμότητα της χρησιμοποίησης σχολίων και community feedback για την εκπαίδευση μοντέλων κατηγοριοποίησης με σκοπό να είναι σε θέση να καθοριστεί η πιθανότητα αποδοχής νέων σχολίων απο την κοινότητα. Τέτοιου τύπου μοντέλα έχουν άμεση εφαρμογή στον εμπλουτισμό της αναζήτησης σχολίων, με το να προωθούν τα ενδιαφεροντα σχόλια ακόμα και παρά την απουσία αρκετού feedback από την κοινότητα. Από την άλλη πλευρά, διεξάγουν μία ανάλυση σε βάθος της διανομής της βαθμολόγησης των σχολίων, περιλαμβάνοντας ποιοτικές και ποσοτικές μελέτς σχετικά με την αξία των συναισθηματικών όρων και τις διαφορές μεταξύ των διάφορων κατηγοριών.Καταλήγουν πως οι προτεινόμενες τεχνικές έχουν άμεση εφαρμογή για την αναζήτηση σχολίων κατά την αναζήτηση επιπλέον πληροφοριών στα σχόλια άλλων χρηστών.

Twitter Sentiment Analysis: The Good The Bad and the OMG!

Το 2011 οι Ευθύμιος Κουλουμπής, Theresa Wilson, Johanna Moore διερευνούν τη χρησιμότητα των γλωσσικών χαρακτηριστικών για την ανίχνευση του συναισθήματος των μηνυμάτων του Twitter.Αξιολογούν τη χρησιμότητα των υπαρχόντων λεξικών καθώς και των χαρακτηριστικών που συλλαμβάνουν πληροφορίες σχετικά με την ανεπίσημη και δημιουργική

γλώσσα που συναντιέται στο microblogging. Παραθέτουν μία εποπτευμένη προσέγγιση για το πρόβλημα, αλλά παράλληλα αξιοποιούν υπάρχοντα hashtags στα δεδομένα του Twitter με σκοπό την δημιουργία δεδομένων εκπαίδευσης. Τα πειράματά τους δείχνουν πως το part-of-speech tagging (διαδικασία σήμανσης μίας λέξης σε ένα κείμενο που αντιστοιχεί σε ένα συγκεκριμένο μέρος του λόγου με βάση τον ορισμό και το περιεχόμενό του) δεν είναι τόσο χρήσιμο όσον αφορά τον τομέα του microblogging. Χρειάζεται περαιτέρω έρευνα για να καθοριστεί αν τα POS χαρακτηριστικά είναι απλα χαμηλής ποιότητας ή απλά δεν είναι τόσο χρήσιμα στον συγκεκριμένο τομέα. Χρησιμοποιώντας ένα υπάρχον λεξικό σε συνδυασμό με τα χαρακτηριστικά του microblogging ήταν σχετικά χρήσιμο, αλλα τα εργαλεία που χρησιμοποιήθηκαν αποκλειστικά (παρουσία intensifiers, θετικά/αρνητικά/ουδέτερα emoticons και συντομογραφίες) ήταν σαφώς τα πιο χρήσιμα. Η χρήση των hashtags για τη συλλογή δεδομένων εκπαίδευσης αποδείχθηκε χρήσιμη, όπως και τα δεδομένα που συλλέχθηκαν με βάση τα θετικά και αρνητικά emoticons. Ωστόσο, καταλήγουν πως ποια μέθοδος παράγει τα καλύτερα δεδομένα εκπαίδευσης και κατα πόσο οι δύο πηγες αυτών των δεδομένων είναι συμπληρωματικές μπορεί να εξαρτάται από τον τύπο των χαρακτηριστικών που χρησιμοποιήθηκαν.

Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets

Το 2014 οι Pablo Gamallo και Marcos Garcia περιγράφουν στο άρθρο τους μία στρατηγική βασισμένη σε έναν Naive-Bayes classifier (ταξινομητή) για την ανίχνευση της πολικότητας σε αγγλικά tweets. Τα πειράματα που διεξήγαγαν έδειξαν πως η καλύτερη απόδοση επιτυγχάνεται με τη χρησιμοποίηση ενός δυαδικού ταξινομητή εκπαιδευμένο να ανιχνεύει μόνο δύο κατηγορίες: θετική και αρνητική. Με σκοπό να ανιχνεύσουν tweets με και χωρίς πολικότητα χρησιμοποίησαν μία βασική στρατηγική βασισμένη στον εντοπισμό λημμάτων πολικότητας μέσα στα κείμενα/tweets. Εφόσον το tweet δεν περιέχει το λιγότερο ένα λήμμα που βρίσκεται επίσης σε λεξικό εξωτερικής πολικότητας, τότε δεν έχει καμία πολικότητα και, επομένως, κατηγοριοποιείται με ουδέτερη τιμή.Η χρήση τόσο ενός λεξικού πολικότητας όσο και πολλαπλών λέξεων (multiwords) βελτιώνουν σημαντικά τα αποτελέσματα. Το σύστημα τους χρησιμοποιείται από την εταιρεια Cilenis S.L, μία εξειδικευμένη εταρεία στην τεχνολογία της επεξεργασίας φυσικής γλώσσας, και έχει εφαρμοστεί σε τέσσερις γλώσσες: Αγγλικά, Ισπανικά, Πορτογαλικα και Γαλικιανά.

Learning Document-Level Semantic Properties from Free-text Annotations

Το 2008 οι S.R.K Branavan, Harr Chen, Jacob Eisenstein και Regina Barzilay παρουσιάζουν μία νέα μέθοδο αξιοποιώντας τους σχολιασμούς ελεύθερων κειμένων για να συμπεράνουν τις ιδιότητες των εγγράφων. Ο αριθμός των σχολίων στα ελεύθερα κείμενα αυξάνεται συνεχώς, λόγω της πρόσφατης δραματικής αύξησης στο γενικό ηλεκτρονικό περιεχόμενο που δημιουργείται από χρήστες. Ένα παράδειγμα τέτοιου περιεχομένου είναι κριτικές προϊόντων που συνήθως χαρακτηρίζονται με φράσεις κλειδιά όπως ‘’real bargain’’ ή ‘’good value’’. Έχοντας

ως σκοπό να εκμεταλλευτούν τέτοιου τύπου ‘’θορυβώδεις’’ σχολιασμούς, βρίσκουν ταυτόχρονα κρυμμένες παραφράσεις στη δομή των φράσεων κλεδί, ένα μοντέλο των εγγράφων και βαθύτερες σημασιολογικές ιδιότητες που συνδέουν αυτά τα δύο. Αυτό τους επιτρέπει να προβλέψουν τις ιδιότητες των μη σχολιασμένων κειμένων με την εφαρμογή ενός ιεραρχικού Bayesian μοντέλου. Εκτελούν πολλές εκδοχές του μοντέλου τους και διαπιστώνουν πως ουσιαστικά αποδίδει καλύτερα από τις υπόλοιπες εναλλακτικές προσεγγίσεις.

PageRanking WordNet Synsets: An Application to Opinion Mining

Το 2007 οι Andrea Esuli και Fabrizio Sebastiani παρουσιάζουν μια εφαρμογή του PageRank, ένα αλλαγής μεταβλητών χωρίς προκαθορισμένη πορεία μοντέλο που αρχικά σχεδιάστηκε για να κατατάσσει τα αποτελέσματα αναζήτησης στο διαδίκτυο, προς την κατάταξη των synsets (ρήματα, ουσιαστικά, επίθετα, επιρρήματα που έχουν την ίδια σημασία π.χ οι λέξεις ‘’boat’’ και ‘’ship’’) του WordNet, ως προς το πόσο διαθέτουν μία δεδομένη σημασιολογική ιδιότητα. Οι σημασιολογικές ιδιότητες που χρησιμοποιούν για την παραδειγματοποίηση της προσέγγισης είναι η θετικότητα και η αρνητικότητα, δύο ιδιότητες κεντρικής σημασίας στην ανάλυση συναισθήματος. Η ιδέα τους προέρχεται από την παρατήρηση πως το WordNet μπορέι να θεωρηθεί ως ένα γράφημα στο οποίο είναι συνδεδεμένα τα synsets μέσω της δυαδικής σχέσης: <<ένας όρος που ανήκει στο synset s_k εμφανίζεται στον σχολιασμό του synset s_i>> και στην υπόθεση ότι αυτή η σχέση μπορεί να θεωρηθεί ως πομπός των εν λόγω σημασιολογικών ιδιοτήτων. Σε γενικές γραμμές, ερεύνησαν τη δυνατότητα εφαρμογής ενός τυχαίου μοντέλου στο πρόβλημα της ταξινόμησης των synsets σύμφωνα με τη θετικότητα και την αρνητικότητα, ωστόσο κατέληξαν πως το μοντέλο τους χρησιμεύει για μια πιο γενική χρήση δηλαδή για τον προσδιορισμό άλλων ιδιοτήτων τέτοιων όρων.

International Sentiment Analysis for News and Blogs

Το 2008 οι Mikhail Bautin, Lohit Vijayarenu και Steven Skiena εξερευνούν μία προσέγγιση που χρησιμοποιεί μία μηχανή μεταγλώττισης τελευταίας τεχνολογίας και πραγματοποιούν μία ανάλυση συναισθήματος στην μετάφραση μία ξένης γλώσσας σε αγγλική. Διεξάγουν τα πειράματά τους σε εφημερίδες με εννέα διαφορετικές γλώσσες και σε ένα παράλληλο corpus με πέντε διαφορετικές γλώσσες και εν τέλει τους οδηγούν στα εξής συμπεράσματα: (α) Οι βαθμολογίες συναισθήματος των οντοτήτων που αποκτήθηκαν με τη δική τους μέθοδο είναι στατιστικά αρκετά πιο συσχετιζόμενες μεταξύ εννέα γλωσσών από νέες πηγές και πέντε γλωσσών ενός παράλληλου corpus (β) Η ποιότητα της μεθόδου ανάλυσης συναισθήματός τους είναι σε μεγάλο βαθμό ανεξάρτητη με τον μεταφραστή και (γ) Μετά την εφαρμογή ορισμένων τεχνικών, οι βαθμολογίες συναισθήματος των οντοτήτων μπορούν να χρησιμοποιηθούν για την διεξαγωγή ουσιαστικών διαπολιτισμικών συγκρίσεων.

Designing Novel Review Ranking Systems: Predicting usefulness and Impact of Reviews

Το 2007 οι Anindya Ghose και Panagiotis G. Ipeirotis προτείνουν δύο μηχανισμούς κατάταξης για την ταξινόμηση των κριτικών των προϊόντων: Ένα μηχανισμό κατάταξης με γνώμονα τον καταναλωτή που κατατάσσει τις κριτικές σύμφωνα με την βοήθεια που αναμένεται να προσφέρουν και ένα μηχανισμό κατάταξης με γνώμονα τον κατασκευαστή που κατατάσσει τις κριτικές σύμφωνα με την αναμενόμενη επίδρασή τους στις πωλήσεις. Οι μηχανισμοί κατάταξής τους συνδυάζουν την οικονομετρική ανάλυση με τεχνικές εξόρυξης κειμένου και συγκεκριμένα με ανάλυση υποκειμενικότητας. Δείχνουν πως η ανάλυση υποκειμενικότητας μπορεί να δώσει χρήσιμες ενδείξεις σχετικά με τη χρησιμότητα μιας κριτικής και τον αντίκτυπό της στις πωλήσεις. Η συγκεκριμένη έρευνα είναι μοναδική ως προς το τρόπο ππου παρακολουθεί τον επηρεασμό του προϊόντος και των πωλήσεων από το συναίσθημα σε μία κριτική και το βαθμό στον οποίο αυτές οι κριτικές είναι ενημερωτικές. Επιπλέον, συμπεραίνουν πως οι κριτικές που γίνονται από τους χρήστες και τείνουν να περιλαμβάνουν ένα μείγμα υποκειμενικών και αντικειμενικών στοιχείων είναι σαφώς πιο ενημερωτικές άρα και χρήσιμες. Ωστόσο, παρόλο που τα αποτελέσματά τους έχουν διάφορες εφαρμογές στο σχεδιασμό της αγοράς σε διαδικτυακά forums γνώμης, καταλήγουν πως για τις επιπτώσεις των κριτικών στις πωλήσεις ώστε να υπάρχουν πιο ασφαλή συμπεράσματα πρέπει να διενεργηθούν περαιτέρω έρευνες.

Large-Scale Sentiment Analysis for News and Blogs

Το 2007 οι Namrata Godbole, Manjunath Srinivasaiah και Steven Skiena παρουσιάζουν ένα σύστημα που εκχωρεί βαθμολογίες που αφορούν θετική ή αρνητική γνώμη σε κάθε ξεχωριστή οντότητα στο σώμα του κειμένου. Το σύστημά αποτελείται από μία φάση προσδιορισμού συναισθήματος, η οποία συνδέει τις εκφρασμένες απόψεις με κάθε σχετική οντότητα και μία φάση συσσωμάτωσης και βαθμολόγησης συναισθημάτων, η οποία βαθμολογεί ξεχωριστά κάθε οντότητα με παρόμοιες στην ίδια κλάση. Επιπλέον, υποστηρίζουν πως υπάρχουν πολλές ενδιαφέρουσες κατευθύνσεις που μπορούν να εξερευνηθούν αλλά στην συγκεκριμένη έρευνα τους ενδιαφέρει με ποιο τρόπο το συναίσθημα μπορεί να διαφέρει ανάλογα με τη δημογραφική ομάδα, την πηγή ειδήσεων ή τη γεωγραφική θέση. Καταλήγουν πως επεκτείνοντας τη χωρική ανάλυση των οντοτήτων-ειδήσεων σε χάρτες συναισθήματος, μπορούμε να εντοπίσουμε γεωγραφικές περιοχές που να γνωρίζουμε εκ των προτέρων αν είναι θετικά ή αρνητικά προδιαθετειμένες για συγκεκριμένες οντότητες.

Sampling Search-Engine Results

Το 2006 οι Aris Anagnostopoulos, Andrei Z. Broder και David Carmel ερευνούν το πρόβλημα της αποδοτικής δειγματοληψίας σε αποτελέσματα αναζήτησης από διαδικτυακές

μηχανές. Χρησιμοποιούν ένα μικρό τυχαίο δείγμα αντί για το πλήρες σύνολο των αποτελεσμάτων και τους οδηγεί σε αποδοτικές προσεγγίσεις αλγορίθμων για διάφορες εφαρμογές όπως ο υπολογισμός του μεγέθους του result set και ο προσδιορισμός του συνόλου κατηγοριών σε μία δεδομένη ταξινόμηση που καλύπτει τα αποτελέσματα αναζήτησης. Παρουσιάζουν και αναλύουν αποδοτικούς αλγόριθμους για την απόκτηση ομοιόμορφων τυχαίων δειγμάτων εφαρμόσιμα σε οποιαδήποτε μηχανή αναζήτησης που βασίζεται σε λίστες καταχώρησης και εκτίμηση εγγράφων τύπου document-at-a-time (για παράδειγμα το Google και το Yahoo Search). Επιπλέον, ο αλγόριθμός τους μπορεί να τροποποιηθεί για να ακολουθήσει τη σύγχρονη αντικειμενοστραφή προσέγγιση, όπου οι λίστες καταχώρησης θεωρούνται streams που περιέχουν μία μέθοδο τύπου next. Επίσης, αναλύουν πως μπορεί να κατασκευαστεί ένα δείγμα από μία βασική next(p) μέθοδο, η οποία περιλαμβάνει δείκτες απόδοσης δειγμάτων με πιθανότητα p και ένα αντίστοιχο δείγμα μεθόδου για Boolean τελεστές. Καταλήγοντας δοκιμάζουν την αποτελεσματικότητα και την ποιότητα της προσέγγισής τους τόσο σε συνθετικά όσο και σε πραγματικά δεδομένα.

Multi-Document Summarization of Evaluative Text

Το 2006 οι Giuseppe Carenini, Raymond Ng και Adam Pauls παρουσιάζουν και συγκρίνουν δύο προσεγγίσεις για τον σκοπό της συνοπτικής αξιολόγησης των επιχειρημάτων. Η πρώτη είναι μία προσέγγιση που βασίζεται στην εξαγωγή προτάσεων ενώ η δεύτερη είναι μία προσέγγιση βασισμένη στην παραγωγή γλώσσας. Αξιολογώντας αυτές τις προσεγγίσεις σε μία μελέτη χρηστών, διαπιστώνουν ότι ποσοτικά αποδίδουν εξίσου καλά. Ωστόσο, ποιοτικά διαπιστώνουμε πως αποδίδουν καλά για διαφορετικούς αλλά συμπληρωματικούς λόγους. Οι τάσεις που εντόπισαν στα αποτελέσματα καθώς και τα ποιοτικά σχόλια από τους συμμετέχοντες στη μελέτη χρησης, δείχνουν πως οι συνοψιστές έχουν ο καθένας διαφορετικά πλεονεκτήματα και αδυναμίες. Αντίστοιχα, ποσοτικά οι συνοψιστές είχαν εξίσου καλή απόδοση ενώ ξεπέρασαν σημαντικά σε απόδοση βασική προσέγγιση για την περίληψη πολλαπλών εγγράφων. Καταλήγουν στο συμπέρασμα πως εξετάζοντας τις δύο αυτές προσεγγίσεις ξεχωριστά κάποιες περιλήψεις έχουν έλλειψη συνολικής ακρίβειας με αποτέλεσμα είτε να αποτυγχάνεται η σωστή επισκόπηση των απόψεων που εκφράζονται στο κείμενο είτε οι επισκοπήσεις να μοιάζουν ‘’ρομποτικές’’ και μάλλον ασυνάρτητες. Συνεπώς, για να είναι μία μέθοδος αποτελεσματική για την περίληψη των αξιολογητικών επιχειρημάτων πρέπει να περιλαμβάνει μία σύνθεση και από τις δύο προσεγγίσεις.

Opinion Spam and Analysis

Το 2008 οι Nitin Jindal και Bing Liu μελετούν το φαινόμενο του spam και την αξιοπιστία των διαδικτυακών απόψεων. Αρχικά, εντόπισαν τρεις τύπους spam: Ο τύπος 1 αφορά παραπλανητικές απόψεις, ο τύπος 2 αφορά κριτικές που γίνονται με βάση την μάρκα και όχι αυτό καθαυτό το προϊόν και ο τύπος 3 αφορά μη-κριτικές όπως διαφημίσεις, ερωτήσεις, τυχαία κείμενα και άλλα. Η ανίχνευση τέτοιων ανεπιθύμητων μηνυμάτων γίνεται πρώτα ανιχνεύοντας πανομοιότυπες κριτικές. Στη συνέχεια, ανιχνεύουν spams τύπου 2 και 3 χρησιμοποιώντας επιβλεπόμενη μάθηση με παραδείγματα εκπαίδευσης όπου η επισήμανση έγινε από τους ίδιους. Τα αποτελέσματα έδειξαν ότι το μοντέλο της λογιστικής παλινδρόμησης (logistic regression) είναι εξαιρετικά αποτελεσματικό. Ωστόσο, για να ανιχνευθούν τα spams τύπου 1, είναι διαφορετική η ιστορία, καθώς είναι πολύ δύσκολο να επισημανθούν τα παραδείγματα εκπαίδευσης χειροκίνητα για τα spams του συγκεκριμένου τύπου. Τέλος, προτείνουν να χρησιμοποιηθούν διπλές κριτικές spam ως θετικά παραδείγματα εκπαίδευσης και άλλες κριτικές ως αρνητικά παραδείγματα και οδηγώντας τους στην απόδειξη της αποτελεσματικότητας του μοντέλου αυτού.

Παρακάτω θα περιγράψουμε με λίγα λόγια τρία σύγχρονα εμπορικά συστήματα που ως στόχο έχουν την ανάλυση συναισθήματος (sentiment analysis):

QDegrees Services:

Η QDegrees Services είναι μία ραγδαία αναπτυσσόμενη εταιρεία στον τομέα της εξυπηρέτησης πελατών στην Ινδία. Παρέχουν υπηρεσίες σε διάφορα πεδία όπως data analytics, web & app development, market research και management consulting. Ένα μεγάλο μέρος των πελατών της βασίζεται στο σύστημα ανάλυσης συναισθήματός τους και αντιμετωπίζεται (η διαδικασία της ανάλυσης συναισθήματος) από την εταιρεία ως μία αναγκαία διαδικασία για κάθε οργανισμό ώστε να προοδεύσει. Μέσω αυτού παρέχουν υπηρεσίες όπως η καταγραφή της στάσης των πελατών των εταιρειών που απευθύνονται σε εκείνη και η κατανόηση των συναισθημάτων τους, αναλύοντας κριτικές, έρευνες και απευθείας διαδικτυακές συζητήσεις σε πλατφόρμες κοινωνικών δικτύων. Η μεθοδολογία που χρησιμοποιούν στην ανάλυση συναισθήματος είναι η εξής:

- Text input: Αρχικά παρέχουν δεδομένα αισθήματος για την διαδικασία.

- Text mining: Εφαρμόζουν τεχνικές εξόρυξης δεδομένων και εξόρυξης κειμένου στο σύνολο δεδομένων (dataset). Αφαιρούν και ‘’καθαρίζουν’’ τις λέξεις και τα γράμματα που δεν έχουν σημασία και δεν είναι σχετικά (π.χ &, #, @ και κενά διαστήματα).

- Text Categorization: Κατηγοριοποιούν τα δεδομένα με τη βοήθεια τεχνικών μηχανικής μάθησης.

- Text Clustering: Ομαδοποιούν τα δεδομένα με την ίδια σημασία και κατηγορία.

- Classification/Relation Modelling: Εφαρμόζουν ένα ταξινομητή Naïve Bayes και τεχνικές SVM.

- Application: Εφαρμόζουν το μοντέλο στα δεδομένα.

Η συγκεκριμένη μεθοδολογία παρατίθεται στην συνέχεια με ένα απλό σχήμα:

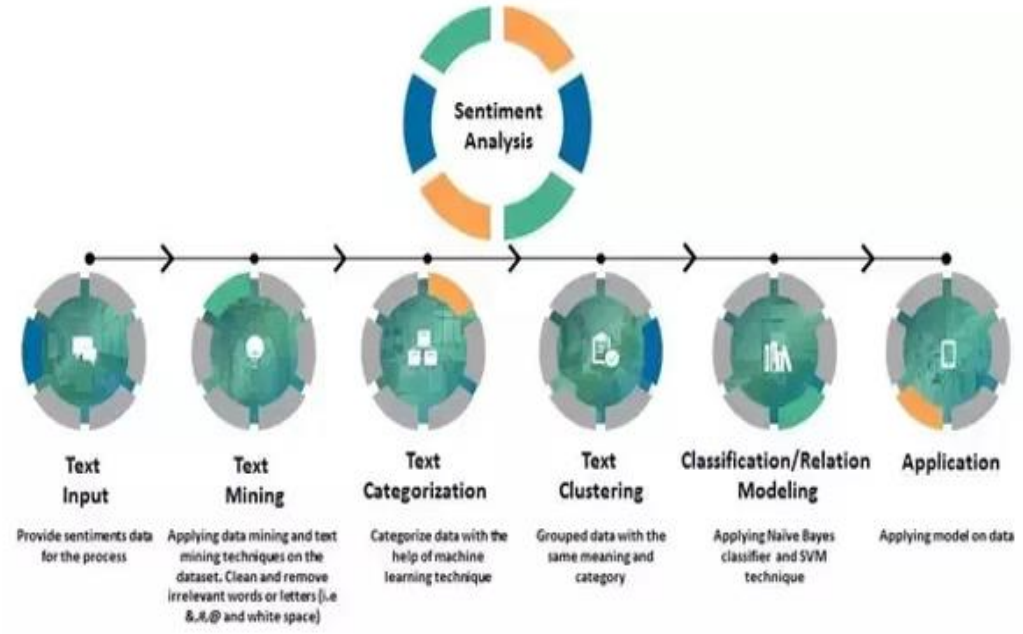


Figure 2.1: QDegrees SA Methodology

Netowl:

Η εταιρεία Netowl συμμετείχε αρχικά σε προγράμματα χρηματοδοτημένα από την κυβέρνηση των ΗΠΑ που αφορούσαν εξερεύνηση προηγμένων πολυγλωσσικών αναλύσεων κειμένων. Στη συνέχεια δημιούργησε το δικό της προϊόν, το NameTag, το οποίο ήταν απλά ένα πολυγλωσσικό εργαλείο για εξόρυξη οντοτήτων και μετά πρόσθεσε άλλες διάφορες λειτουργίες όπως relationship and event extraction, geotagging, identity resolution και πιο πρόσφατα από όλες την ανάλυση συναισθήματος. Πλέον, προσφέρει στους πελάτες της ανάλυση συναισθημάτων βασισμένη σε οντότητες καθώς και ανάλυση συναισθημάτων βασισμένη σε χαρακτηριστικά. Σε επίπεδο οντοτήτων, εντοπίζει τα συναισθήματα προς άλλους διάφορους τύπους οντοτήτων όπως οι άνθρωποι, οι οργανισμοί, τα εμπορικά σήματα και τα προϊόντα. Σε επίπεδο πτυχών, καταγράφει τις συγκεκριμένες πτυχές της οντότητας που σχετίζονται με τα συναισθήματα όπως την τιμή ενός προϊόντος, τη νέα πολιτική μίας χώρας, την προεκλογική εκστρατεία ενός υποψηφίου και άλλα. Με την ικανότητα να εντοπίζει τι αφορά το συναίσθημα η Netowl χρησιμοποιεί μία μεθοδολογία η οποία:

- Δεν αρκείται μόνο στο να εξάγει δυαδικού τύπου απαντήσεις, δηλαδή μόνο αν το συναίσθημα είναι θετικο ή αρνητικό, αλλά προσφέρει μία πιο ειδικευμένη μέθοδο έτσι ώστε να διακρίνει διαφορετικές απόψεις, στάσεις, προθέσεις και συμπεριφορές.

- Επιτρέπει πολύ πιο λεπτομερή ανάλυση συναισθημάτων σε κάθε οντότητα και ως αποτέλεσμα καταγράφει σε βάθος πολλαπλά αντιφατικά συναισθήματα που εκφράζονται σε ένα ενιαίο έγγραφο ή πρόταση.

- Εκχωρεί κανονικοποιημένες και σταθερές τεχνικές σε εξορυγμένες οντότητες και εκφράσεις συναισθημάτων. Λαμβάνει υπόψη την κεφαλαιοποίηση, τα ψευδώνυμα, τα ακρωνύμια, τις συντομογραφίες, τις μορφολογικές παραλλαγές (π.χ αριθμούς) και άλλα. Με αυτό τον τρόπο είναι σε θέση να ταξινομεί, να συγκεντρώνει και να ποσοτικοποιεί τις μυριάδες συναισθημάτων που εκφράζονται σε ένα μεγάλο σύνολο δεδομένων παράγοντας ταυτόχρονα διαγράμματα, γραφήματα και πίνακες ελέγχου για την καλύτερη παρακολούθηση.

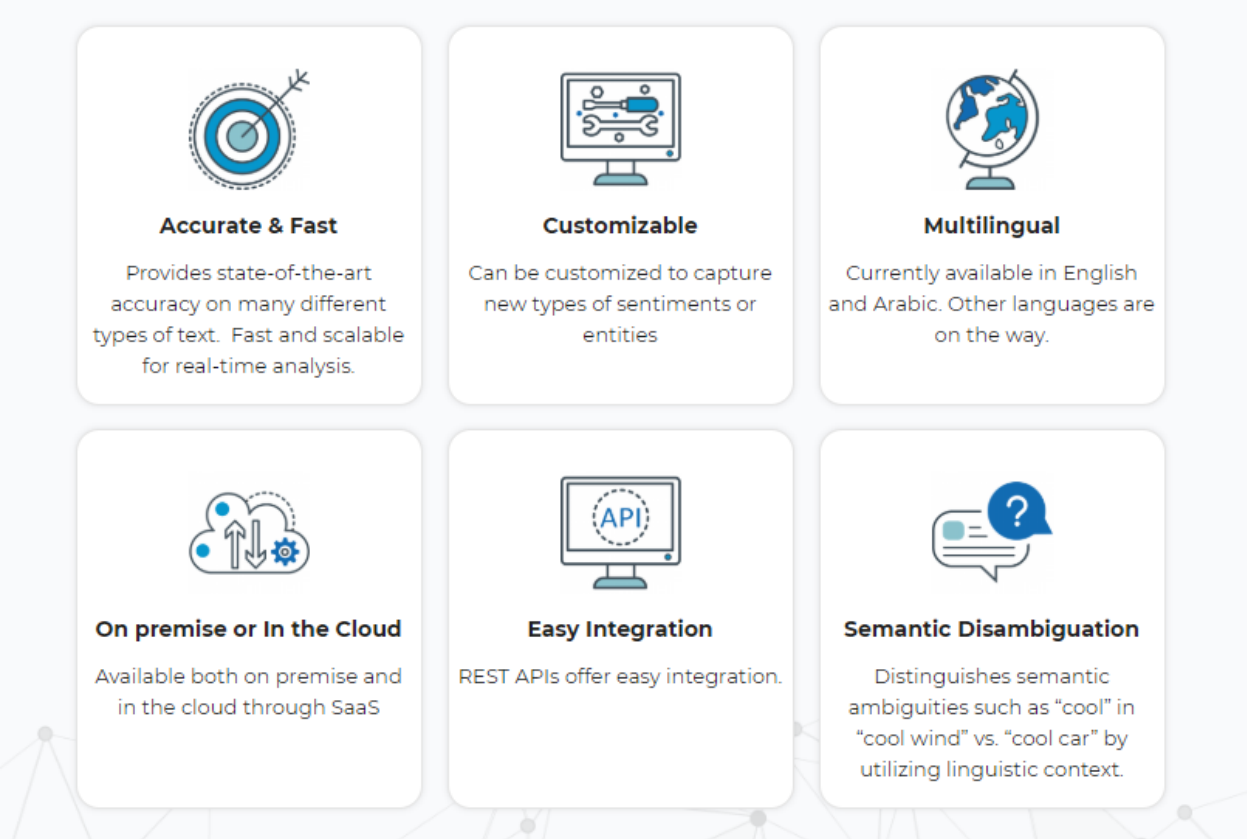


Figure 2.2: Netowl SA System advantages

3i Data Scraping:

Η εταιρεία 3i Data Scraping βασίζεται εξ ολοκλήρου σε μία cloud based υποδομή, διασφαλίζοντας ότι η διαδικασία απόρριψης διαδικτυακών δεδομένων ειναι κλιμακωτή και εξασφαλίζοντας με αυτό τον τρόπο ότι τα projects από πελάτες θα ολοκληρωθούν σε εύλογο χρονικό διάστημα. Ασχολούνται με διάφορα πεδία της εξόρυξης γνώμης και τελευταία κυρίαρχο ρόλο όσον αφορά την εταιρεία αποτελεί η ανάλυση συναισθημάτων.

Το κύριο χαρακτηριστικό της ανάλυσης συναισθημάτων της είναι η αξιολόγηση των προσβάσεων (εκτιμήσεις πελατών, σχόλια στοιχείων, τύποι σχολίων που παράγουν λύσεις κ.ο.κ) καθώς και η ανάλυση για τα συναισθήματα που αποκαλύπτονται (χαρά, δυσαρέσκεια κ.ο.κ). Αυτό το επιτυγχάνουν με την ανάπτυξη ενός σημειακού συστήματος βαθμολογώντας από το 1 έως το 10, με το 10 να είναι το πιο ευνοϊκό, όπου κάθε λέξη συνδέεται ξεχωριστά με κάποιο συναίσθημα. Κάθε βαθμολογία λέξης, καθώς και συνολικά η βαθμολογία του κειμένου, υπολογίζεται για να διαπιστωθεί ποια είναι ακριβώς η πεποίθηση ή η σκοπιά του συγγραφέα. Μία ακόμα τεχνική που χρησιμοποιούν είναι η αναγνώριση της υποκειμενικότητας-αντικειμενικότητας ενός συγγραφέα. Παρόλο που γενικά είναι εξαιρετικά δύσκολο, η 3i Data Scraping συνδέει την κάθε άποψη με τον συγγραφέα οπότε έχει μία πιο καθαρή εικόνα σχετικά

με το τι και ποιον εξυπηρετεί η άποψή του και κατά πόσο έχει γραφτεί αμερόληπτα. Συνεπώς με το συγκεκριμενο σύστημα ανάλυσης συναισθημάτων επιτυγχάνει τα εξής:

- Ελαχιστοποίηση των ιστοσελίδων με περίπλοκα σχόλια τα οποία συνήθως δεν έχουν ευδιάκριτο νόημα, εξασφαλίζοντας πλήρη και εύκολη πρόσβαση σε ‘’δεξαμενές’’ δεδομένων.

- Άμεση ενημέρωση για μαζική δημοσίευση ροών πληροφοριών και ένα διαδραστικό API για την άμεση απόσπαση πληροφοριών και συναισθημάτων σχετικά με αυτές.

- Παρακολούθηση τυχόν τροποποιήσεων σε ιστότοπους έτσι ώστε να επιτευχθεί η αδιατάρακτη παροχή real-time πληροφοριών.

- Ευελιξία που είναι εξαιρετικά αποδοτική από πλευράς κόστους.

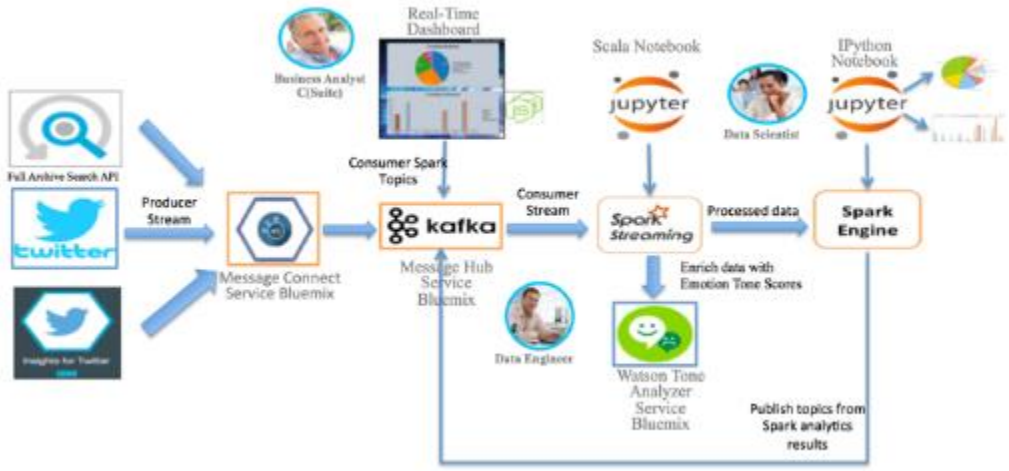


Figure 2.3: 3i Data Scraping SA System example

Κεφάλαιο 3ο: Εξόρυξη Δεδομένων

3.1 Ορισμός

Η εξόρυξη δεδομένων είναι η μη τετριμμένη εξαγωγή υποκρυπτόμενης, άγνωστης και εν δυνάμει χρήσιμης πληροφορίας με αυτόματο ή ημι-αυτόματο τρόπο σε μεγάλες ποσότητες δεδομένων για τον προσδιορισμό προτύπων και τη δημιουργία σχέσεων όσον αφορά την επίλυση προβλημάτων μέσω της ανάλυσης δεδομένων. Με απλά λόγια, η εξόρυξη δεδομένων ορίζεται ως μια διαδικασία που χρησιμοποιείται για την εξαγωγή χρησιμοποιήσιμων και χρήσιμων δεδομένων από ενα μεγαλύτερο σύνολο ακατέργαστης πληροφορίας-δεδομένων.

Σε ορισμένους τομείς, η εξόρυξη δεδομένων μετασχηματίζει όχι μόνο τον τρόπο με τον οποίο γίνεται η έρευνα, αλλά και το αντικείμενο της έρευνας. Νέοι ορίζοντες και ερευνητικά ερωτήματα αναδύονται. Για παράδειγμα, έχει προκύψει ένας εντελώς νέος τομέας ψηφιακών ανθρωπιστικών επιστημών. Η έρευνα σε αυτόν τον τομέα δεν οδηγεί μόνο στην καλύτερη κατανόηση της πληροφόρησης και της κοινωνικοπολιτιστικής σημασίας που είναι ενσωματωμένη στα ιστορικά κείμενα αλλά παρέχει επίσης καλύτερα εργαλεία και μεθοδολογίες για τη βελτίωση της κατανόησης του κόσμου των πολυμέσων στον οποίο ζούμε.

3.2 Τρόποι μηχανικής μάθησης

Εν γένει, ο τομέας της Μηχανικής Μάθησης αναπτύσσει τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση και ενισχυτική μάθηση. Πιο αναλυτικά:

- Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα:
    1. Ταξινόμησης (Classification)
    2. Πρόγνωσης (Prediction)
    3. Διερμηνείας (Interpretation)

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα:

    1. Ανάλυσης Συσχετισμών (Association Analysis)
    2. Ομαδοποίησης (Clustering)

- Ενισχυτική Μάθηση (Reinforcement Learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον.

Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

Οι κανόνες συσχέτισης δημιουργούνται με την ανάλυση δεδομένων όσον αφορά συχνά εμφανιζόμενα if/then πρότυπα, χρησιμοποιώντας τα κριτήρια υποστήριξης και εμπιστοσύνης για τον εντοπισμό των σημαντικότερων σχέσεων μέσα στα δεδομένα. Η υποστήριξη είναι το πόσο συχνά εμφανίζονται τα στοιχεία στη βάση δεδομένων, ενώ η εμπιστοσύνη είναι ο αριθμός των περιπτώσεων που τα if/then statements είναι ακριβή.

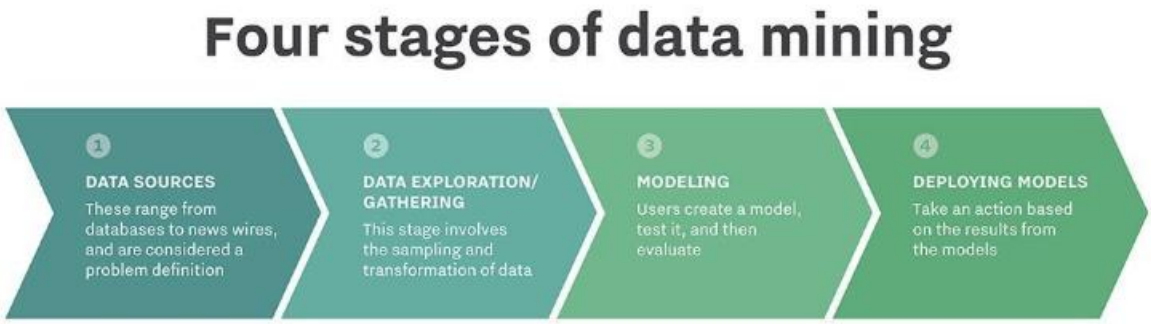


Figure 3.1: Πηγή https://www.searchsqlserver.techtarget.com

3.3 Τεχνικές εξόρυξης δεδομένων

Τεχνικές εξόρυξης δεδομένων είναι οι ανάλυση ακολουθίας ή διαδρομής (Sequence or Path Analysis), η κατηγοριοποίηση (Classification), η συσταδοποίηση (Clustering) και η παλινδρόμηση (Regression). Με τη τεχνική Sequence or Path Analysis αναζητούμε μοτίβα στα οποία ένα γεγονός οδηγεί σε μεταγενέστερο συμβάν. Με τη τεχνική Classification δοσμένης μίας συλλογής εγγραφών (σώμα εκπαίδευσης-training set), κάθε εγγραφή περιέχει ένα σύνολο ιδιοτήτων-attributes, μιας εκ των οποίων είναι η κλάση-class. Στόχος μας είναι οι προηγουμένως αθέατες εγγραφές να χαρακτηρισθούν με μία κλάση όσο ακριβέστερα γίνεται. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί ένα μοντέλο ταξινόμησης για τον προσδιορισμό των αιτούντων δανείων ως χαμηλού, μεσαίου ή υψηλού πιστωτικού κινδύνου. Με τη τεχνική Clustering έχουμε ως στόχο την ομαδοποίηση ενός συγκεκριμένου συνόλου αντικειμένων με βάση τα χαρακτηριστικά τους, συγκεντρώνοντας τα ανάλογα με τις ομοιότητές τους. Για παράδειγμα, η συσταδοποίηση μπορεί να βοηθήσει τις εταιρείες να ανακαλύψουν υπο-ομάδες στη βάση των πελατών τους και να χρησιμοποιήσει αυτές τις γνώσεις για ανάπτυξη συγκεκριμένου marketing που αφορά μόνο αυτές. Η τεχνική Regression χρησιμοποιείται για την πρόβλεψη της τιμής μιας δοσμένης συνεχούς μεταβλητής (έχει ως τιμή έναν πραγματικό αριθμό), με βάση τις τιμές άλλων μεταβλητών, υποθέτοντας μια γραμμική ή όχι εξάρτηση του μοντέλου. Για παράδειγμα, η

παλινδρόμηση μπορεί να χρησιμοποιηθεί για την πρόβλεψη του κόστους ενός προϊόντος ή υπηρεσίας, δεδομένων άλλων μεταβλητών.

Οι τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται σε πολλούς τομείς έρευνας, συμπεριλαμβανομένων των μαθηματικών, της κυβερνητικής , της γενετικής ,του marketing και άλλων. Αποτελούν ένα μέσο για την προώθηση της αποτελεσματικότητας και την πρόβλεψη της συμπεριφοράς και αν χρησιμοποιηθούν σωστά μία επιχείρηση μπορεί να ξεχωρίσει από τον ανταγωνισμό της μέσω των στοχευμένων κινήσεων που προέρχονται από τη σωστή πρόβλεψη. Για παράδειγμα οι εταιρείες του χρηματοπιστωτικού κλάδου χρησιμοποιούν εργαλεία εξόρυξης δεδομένων για τη δημιουργία μοντέλων κινδύνου και την ανίχνευση απάτης.

3.4 Classification

Το classification (κατηγοριοποίηση) είναι μία τεχνική εξόρυξης δεδομένων η οποία δοσμένης μιας συλλογής εγγραφών (σώμα εκπαίδευσης-training set), η κάθε εγγραφή περιέχει ένα σύνολο ιδιοτήτων-attributes, μιας εκ των οποίων είναι η κλάση-class. Ο στόχος του classification είναι η κατάταξη με ακρίβεια στο κατάλληλο target class, για κάθε περίπτωση στα δεδομένα δηλαδή οι προηγουμένωνς αθέατες εγγραφές να χαρακτηρισθούν με μία κλάση όσο ακριβέστερα γίνεται. Για παράδειγμα, τέτοιου τύπου εφαρμογές είναι η κατηγοριοποίηση συναλλαγών με πιστωτική κάρτα για το αν είναι νόμιμες ή μη, η κατηγοριοποίηση άρθρων εφημερίδων ως οικονομικά, αθλητικά, κοινωνικά και τα λοιπά, η κατηγοριοποίηση δορυφορικών εικόνων για την εύρεση αυθαίρετων πισίνων σε σπίτια και άλλα.

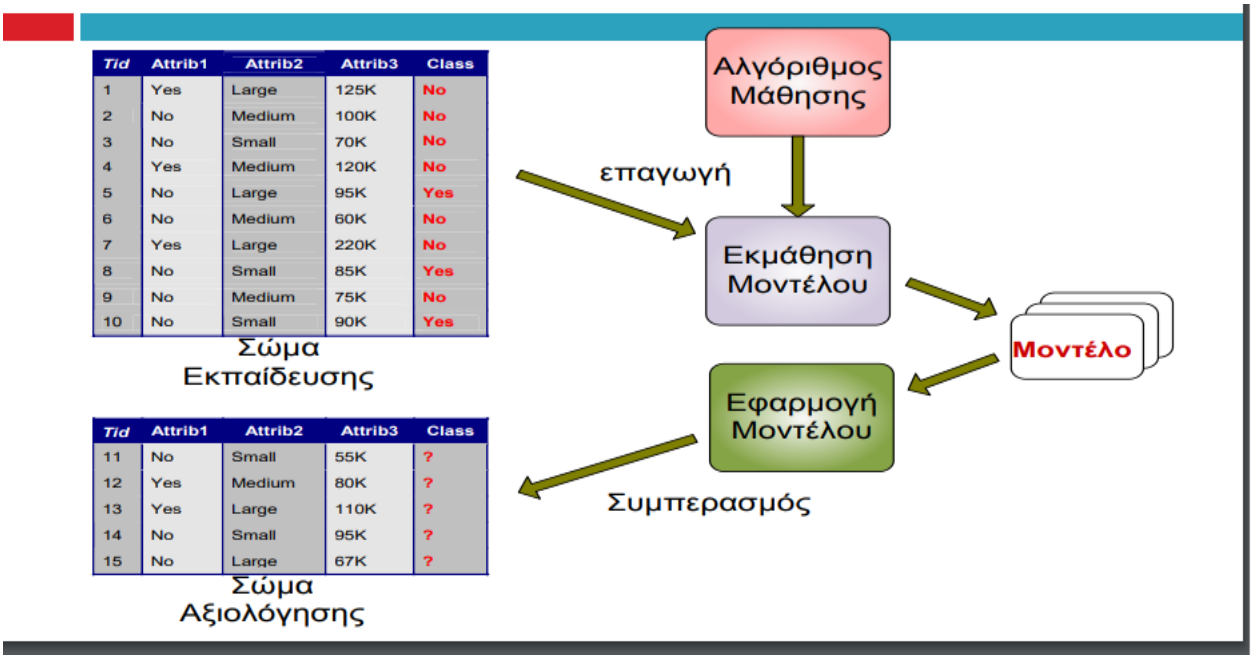


Figure 3.2: Γραφική απεικόνιση της κατηγοριοποίησης

3.4.1 Τεχνικές κατηγοριοποίησης

Οι τεχνικές κατηγοριοποίησης είναι οι εξής:

- Μέθοδοι με δέντρα αποφάσεων
- Μέθοδοι βασισμένοι σε κανόνες
- Μέθοδοι βασισμένοι στη μνήμη
- Νευρωνικά δίκτυα
- Δίκτυα Bayes και απλοϊκή μέθοδος Bayes
- Μηχανές διανυσμάτων υποστήριξης

Στη συνέχεια εξηγούμε τους αλγόριθμους που μας έδωσαν τα καλύτερα αποτελέσματα στην συγκεκριμένη διπλωματική εργασία:

<u>Δέντρα απόφασης</u>:

Τα δέντρα απόφασης χρησιμοποιούνται ευρέως για την κατηγοριοποίηση και πρόβλεψη δεδομένων. Ένα δέντρο απόφασης κατασκευάζεται σύμφωνα με ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων. Κάθε εσωτερικός κόμβος προσδιορίζει τον έλεγχο των γνωρισμάτων και κάθε κλαδί που συνδέει τους εσωτερικούς με τους απόγονους αντιστοιχεί σε μία πιθανή τιμή για το γνώρισμα. Τα βασικά κριτήρια για την δημιουργία ενός δέντρου απόφασης είναι τα εξής;

- Information Gain
- Gini Index
- Misclassification error

Στην συγκεκριμένη διπλωματική εργασία χρησιμοποιήσαμε το κριτήριο information gain:

Το πιο σημαντικό βήμα για την δημιουργία ενός δέντρου απόφασης είναι να προσδιοριστεί η ρίζα του. Για να επιτευχθεί αυτό χρησιμοποιείται το μέτρο της Εντροπίας, το οποίο βαθμολογεί από 0 (σίγουρη απάντηση) εώς 1 (τελείως αβέβαιη απάντηση ίση με το στρίψιμο ενός νομίσματος) την ποιότητα μίας μεταβλητής (attribute) για να μπει ως ρίζα του δέντρου.

Εντροπία ενός κόμβου t:

$$Εντροπία(t) = -\sum_{j} p(j|t) log p(j|t)$$

Συνεπώς, χρησιμοποιεί την βαθμολογία εντροπίας (information gain) για να αξιολογήσει όλες τις μεταβλητές και αυτή με την μεγαλύτερη βαθμολογία θα χρησιμοποιηθεί ως βάση στο δέντρο.

$$I.G(Ai, S) = E(S) - \sum_{\nu E Ai} \frac{n_\nu}{n} - E(S_{Ai=\nu})$$

όπου E(s) είναι η εντροπία, Ai είναι τα attributes, S είναι οι εγγραφές, n το πλήθος των συγκεκριμένων επιλογών στις μεταβλητές και v οι επιλογές που παρέχει μία μεταβλητή.

Για παράδειγμα, τα βασικά βήματα για την λύση του σεναρίου αν θα γίνει blockbuster μία ταινία είναι αρχικά ο καθορισμός της πρόβλεψης (αν θα γίνει δηλαδή ή όχι) και ότι είναι πρόβλημα τύπου κατηγοριοποίησης (classification). Στην συνέχεια καθορίζουμε τα attributes μας π.χ διαθέσιμα χρήματα, σκηνοθέτης, είδος, ηθοποιοί και φέρνουμε το πρόβλημα σε μορφή πίνακα (tabular). Έπειτα χωρίζουμε το data set μας σε training και test και κατασκευάζουμε το δέντρο απόφασης. Τέλος κάνουμε το απαραίτητο validation, δηλαδή κρύβουμε τα πραγματικά αποτελέσματα και μας δίνει το σύστημα τις προβλέψεις του και χρησιμοποιώντας το μητρώο σύγχυσης (confusion matrix) που εξηγούμε στο κεφαλαιο 2.4.2 συγκρίνουμε τις τιμές που μας έδωσε το σύστημα με τα πραγματικά αποτελέσματα για να βρούμε τις συνολική ακρίβεια, ακρίβεια και ανάκληση του συστήματος.

Ενδεικτικά οι μαθηματικοί τύποι για τα άλλα δύο κριτήρια είναι οι εξής:

Gini index:

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2$$

όπου p(j|t) είναι η πιθανότητα εμφάνισης της κλάσης j στον κόμβο t.

Misclassification Error:

$$Error(t) = 1 - \max P(i|t)$$

όπου μετράει το σφάλμα στην κατηγοριοποίηση που κάνει ένας κόμβος.

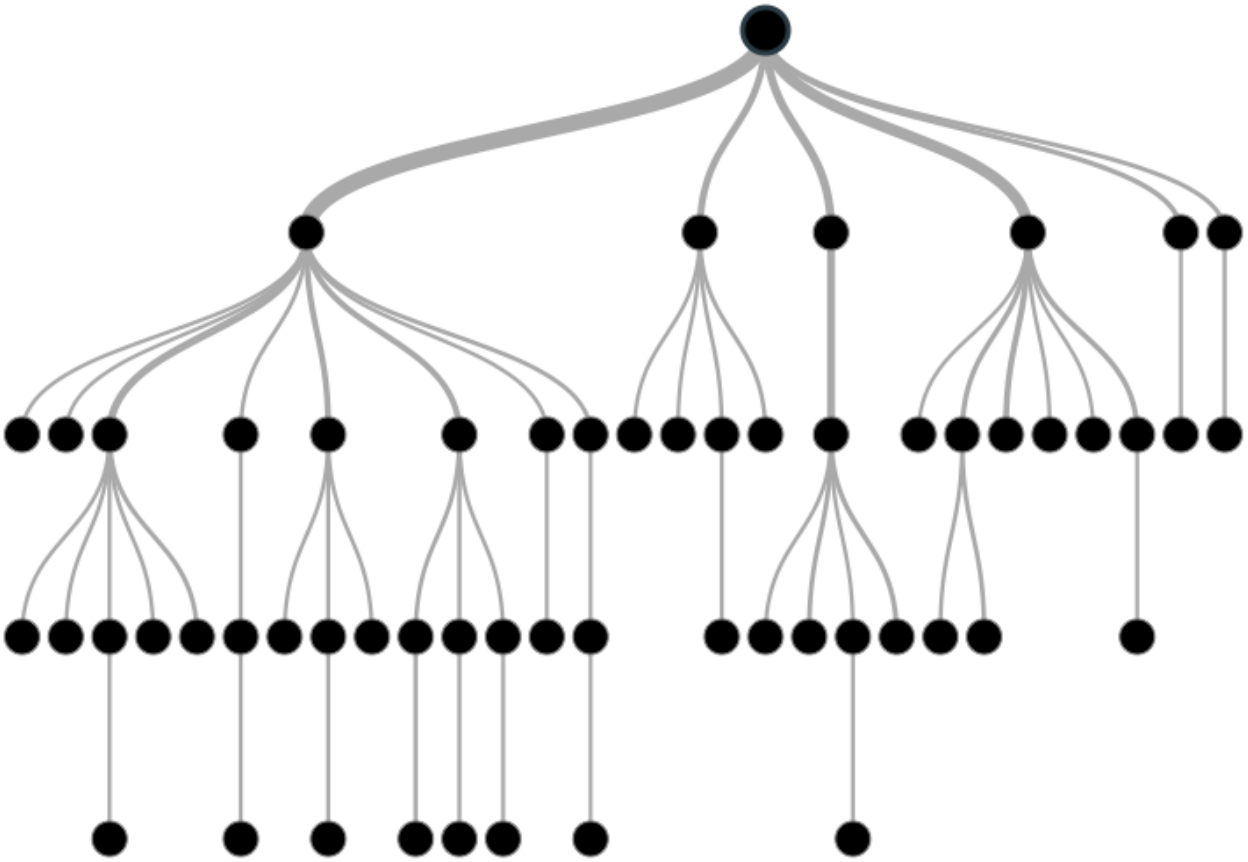


Figure 3.3: Decision Tree Example

<u>Naïve Bayes</u>:

Στην μηχανική μάθηση ενδιαφερόμαστε συχνά να διαλέξουμε την καλύτερη υπόθεση (h) δοσμένων κάποιων δεδομένων (d). Στα προβλήματα κατηγοριοποίησης, η υπόθεση μπορεί να είναι η κλάση που θα αναθέσουμε σε κάποια νέα δεδομένα. Ένας από τους ευκολότερους τρόπους για να διαλέγουμε την πιο πιθανή υπόθεση δοσμένων κάποιων δεδομένων είναι να χρησιμοποιήσουμε την πρότερη γνώση για το πρόβλημα. Το Θεώρημα του Bayes μας παρέχει ένα τρόπο να το πετύχουμε αυτό και είνα το εξής:

$$P(h|d) = \frac{\mathrm{P(d|h)} * P(h)}{P(d)}$$

όπου P(h|d) είναι η πιθανότητα η υπόθεση h δοσμένων των δεδομένων d δηλαδή η posterior probability, P(d|h) είναι η πιθανότητα βάσει των δεδομένων d ότι η υπόθεση h είναι

true, P(h) είναι η πιθανότητα της υπόθεσης h να είναι true ανεξάρτητα από τα δεδομένα δηλαδή η prior probability, P(d) είναι η πιθανότητα των δεδομένων ανεξάρτητα από την υπόθεση. Ενδιαφερόμαστε να υπολογίσουμε την posterior probability του P(h|d) από την prior probability p(h) me P(D) και P(d|h). Αφού υπολογίσουμε τη posterior probability για μια σειρά διαφορετικών υποθέσεων, επιλέγουμε εκείνη με την υψηλότερη πιθανότητα.

Ο Naïve Bayes είναι ένας αλγόριθμος κατηγοριοποίησης για δυαδικά και πολυκλασικά προβλήματα. Έχει αυτή την ονομασία επειδή ο υπολογισμός των πιθανοτήτων για κάθε υπόθεση είναι απλοποιημένος για να καταστήσει τον υπολογισμό τους ευκολότερο. Αντί να επιχειρήσουμε να υπολογίσουμε τις τιμές κάθε τιμής χαρακτηριστικού P(d1, d2, d3 | h), υποθέτουμε ότι είναι υπό όρους ανεξάρτητοι, δεδομένης της τιμής στόχου και υπολογίζονται με συντομία ως P (d1 | h) * P (d2 | H).

Συνεπώς με τον Naïve Bayes εξοικονομούμε αρκετή ταχύτητα όσον αφορά τον υπολογισμό της λύσης του προβλήματος που εξετάζουμε.

<u>K-Nearest Neighbors (k-nn) Classification</u>:

Ο k-nn είναι ένας μη παραμετρικός lazy learning αλγόριθμος δηλαδή δεν χρησιμοποιεί τα δεδομένα εκπαίδευσης για να κάνει οποιαδήποτε γενίκευση. Ο συγκεκριμένος αλγόριθμος αποθηκεύει όλες τις διαθέσιμες περιπτώσεις και ταξινομεί νέες με πλειοψηφία των γειτόνων του και το αντικείμενο να ανατίθεται στην κλάση που είναι κοντινέστερη στους πλησιέστερους γείτονες με βάση ένα μέτρο ομοιότητας (π.χ διανυσματική απόσταση).

Ο αλγόριθμος σε γενικές γραμμές δουλεύει ως εξής:

- Προσδιορίζεται ένας θετικός ακέραιος αριθμός k μαζί με ένα νέο δείγμα.
- Επιλέγουμε τις k καταχωρήσεις στη βάση δεδομένων μας που είναι πιο κοντά στο δείγμα.
- Βρίσκουμε την πιο κοινή ταξινόμηση αυτών των καταχωρήσεων.
- Αυτή είναι η κατηγοριοποίηση που παρέχουμε στο νέο δείγμα.

Ο πιο συνήθης τρόπος για να μετρηθεί η διανυσματική απόσταση είναι η ευκλείδια απόσταση:

$$Ευκλείδια\ απόσταση\ = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Μερικά από τα θετικά του k-nn είναι πως είναι ένας αρκετά απλός αλγόριθμος, ο οποίος εξηγείται εύκολα, παρέχει σχετικά μεγάλη ακρίβεια στις προβλέψεις του παρόλο που δεν είναι ανταγωνιστικές σε σύγκριση με καλύτερα ελεγχόμενα μοντέλα μάθησης, το στάδιο προπόνησης είανι αρκετά γρήγορο, είναι ευέλικτο και δεν υπάρχουν υποθέσεις σχετικά με τα δεδομένα. Στα αρνητικά, είναι υπολογιστικά ακριβό επειδή ο αλγόριθμος αποθηκεύει όλα τα δεδομένα εκπαίδευσης, απαιτεί αρκετά υψηλή δέσμευση μνήμης, το στάδιο πρόβλεψης μπορεί να είναι αργό, αποθηκεύει όλα ή σχεδόν όλα τα δεδομένα εκπαίδευσης και είναι ευαίσθητος σε άσχετα χαρακτηριστικά ανάλογα με την κλίμακα των δεδομενων.

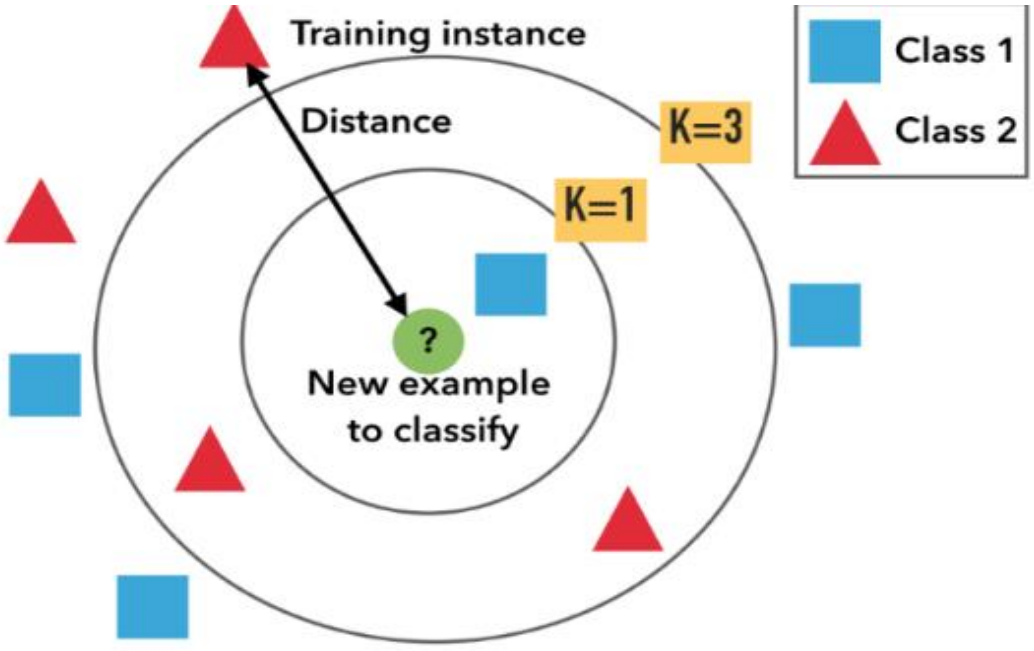


Figure 3.4: K-nn example

3.4.2 Παράμετροι αξιολόγησης

Μητρώο σύγχυσης:

Το μητρώο σύγχυσης (confusion matrix) είναι ένας πίνακας που επιτρέπει την οπτικοποίηση της απόδοσης ενός ταξινομητή. Κάθε στήλη του πίνακα αντιπροσωπεύει τα στιγμιότυπα μιας προβλεπόμενης κλάσης, ενώ κάθε γραμμή αντιπροσωπεύει τα στιγμιότυπα μιας πραγματικής κλάσης. Συνεπώς το στοιχείο στη θέση (i,j) αντιπροσωπεύει τον αριθμό των σημείων δεδομένων των οποίων η πραγματική ετικέτα κλάσης ήταν i και ταξινομήθηκαν στην κλάση j. Αν θεωρήσουμε την απλή περίπτωση ενός δυαδικού προβλήματος ταξινόμησης με κατηγορίες C1 και C2, τότε το μητρώο σύγχυσης αναπαρίσταται ως εξής:

Actual class	Predicted class	
	C_1	C_2
C_1	TP	FN
C_2	FP	TN

όπου:

TP (True Positive): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C1 και ταξινομήθηκαν στην κλάση C1.

TN (True Negative): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C2 και ταξινομήθηκαν στην κλάση C2.

FP (False Positive): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C2 και ταξινομήθηκαν στην κλάση C1.

FN (False Negative): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C1 και ταξινομήθηκαν στην κλάση C2.

Από το μητρώο σύγχυσης μπορούν να εξαχθούν άμεσα οι τιμές της συνολικής ακρίβειας (accuracy), της ακρίβειας (precision) και της ανάκλησης (recall). Σημειώνεται οτι στην περίπτωση πολλών κλάσεων η ακρίβεια και η ανάκληση υπολογίζονται για κάθε μια από τις κλάσεις ξεχωριστά.

Συνολική ακρίβεια:

Η συνολική ακρίβεια (accuracy), είναι το ποσοστό των δεδομένων που ταξινομήθηκαν σωστά, δηλαδή:

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

Ακρίβεια:

Είναι το ποσοστό των σημείων δεδομένων που ταξινομήθηκαν στην κλάση i, και των οποίων η πραγματική ετικέτα κλάσης ήταν i. Δηλαδή:

$$\text{Precision}(\text{c1}) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Στην πραγματικότητα η ακρίβεια μας δείχνει από όλα τα δείγματα που ταξινομήθηκαν στην C1, πόσα πραγματικά ανήκαν στη C1, δηλαδή απαντάει στην ερώτηση «Δοθείσας μιας εκτίμησης του ταξινομητή, ποιά η πιθανότητα να είναι σωστή;»

Ανάκληση:

Είναι το ποσοστό των σημείων δεδομένων με πραγματική ετικέτα κλάσης i, τα οποία ταξινομήθηκαν επιτυχώς στην κλάση αυτή. Δηλαδή:

$$\text{Recall}(\text{c1}) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Η ανάκληση μας δείχνει από όλα τα δείγματα που ανήκαν πραγματικά στην κλάση C1, πόσα από αυτά ταξινομήσαμε επιτυχώς, δηλαδή απαντάει στην ερώτηση «Δοθέντος ενός δείγματος με πραγματική ετικέτα C1, ποιά η πιθανότητα να το ταξινομήσω σωστά; »

Κεφάλαιο 4ο: Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας είναι είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων (φυσικών) γλωσσών. Η ιδέα του να επικοινωνεί κάποιος με τον υπολογιστή και να τον ελέγχει μιλώντας τη μητρική του γλώσσα ή κάποια ευρύτερα ομιλούμενη, όπως τα αγγλικά, είναι πολύ ελκυστική. Όμως, η φυσική γλώσσα έχει διττή φύση (ως προς τη σύνταξη και ως προς τη σημασιολογία), γεγονός που δεν εμποδίζει μεν την επεξεργασία της, αλλά δημιουργεί προβλήματα στην κατανόησή της, με αποτέλεσμα να καθίσταται το εγχείρημα της επεξεργασίας και παράλληλα της κατανόησής της ιδιαίτερα δύσκολο.

4.1 Προεπεξεργασία δεδομένων

Το στάδιο της προεπεξεργασίας δεδομένων είναι απαραίτητο για τον καθαρισμό και την προετοιμασία των δεδομένων για ταξινόμηση και μπορεί να περιλαμβάνονται τα εξής διαδεδομένα στάδια:

- Μετατροπή των κεφαλαίων γραμμάτων σε πεζά:

  Συνηθίζεται κατά την προεπεξεργασία, η μετατροπή όλων των γραμμάτων σε πεζά, ώστε να ταυτίζονται οι διαφορετικές εμφανίσεις κάποιας λέξης.

- Αφαίρεση σημείων στίξης:

  Σε πολλές έρευνες κατά την προεπεξεργασία συνηθίζεται η αφαίρεση των σημείων στίξης.

- Αφαίρεση αριθμών:

  Στις περισσότερες περιπτώσεις, οι αριθμοί σε ένα κείμενο δε σχετίζονται με το συναίσθημα που εκφράζεται και επομένως πολλοί ερευνητές θεωρούν την ανάλυσή τους περιττή.

- Αφαίρεση stop words:

  Ως stop words ορίζονται κάποιες συνηθισμένες λέξεις που δε φέρουν ιδιαίτερη πληροφορία (π.χ at, on, the, which, is), οπότε η ανάλυσή τους μπορεί να οδηγήσει τον ταξινομητή σε λανθασμένα συμπεράσματα.

- Εντοπισμός θέματος λέξεων (stemming):

  Κατά το stemming αφαιρούνται από τις λέξεις οι καταλήξεις ώστε να εντοπιστεί η ρίζα της καθεμίας, με στόχο τη μείωση της πολυπλοκότητας της ανάλυσης χωρίς απώλεια σημαντικής πληροφορίας.

- Εντοπισμός n-grams:

  Τα n-grams είναι ακολουθίες n στοιχείων κειμένου (χαρακτήρων, γραμμάτων συλλαβών ή λέξεων) που προκύπτουν από ένα δεδομένο κείμενο και η χρήση τους βοηθάει στην αναγνώριση φράσεων που μπορεί να περιέχουν κάποιο νόημα το οποίο δεν μπορέι να εντοπιστεί σε περίπτωση ατομικής μελέτης των στοιχείων. Το μήκος των n-grams εξαρτάται από την εκάστοτε εφαρμογή.

- Αναγνώριση μέρους του λόγου (Part of Speech Tagging ή POS tagging):

  Στο στάδιο αυτό γίνεται αναγνώριση και σημείωση του μέρους του λόγου (ρήμα, επίθετο,επίρρημα,ουσιαστικό κ.τ.λ) για κάθε λέξη του κειμένου με στόχο την αποκάλυψη της γραμματικής και επομένως τη βαθύτερη ανάλυσή του.

- Λημματοποίηση (lemmatization) λέξεων:

  Κατά τη λημματοποίηση, οι διάφορες μορφές μιας λέξης (παράγωγα, κλίση) αντιστοιχούνται στο ίδιο λήμμα (π.χ το κοινό λήμμα των ‘’κάνοντας’’ και ‘’έκανα’’ είναι το ‘’κάνω’’). Με τον τρόπο αυτό οι λέξεις γενικεύονται και η ταξινόμησή τους γίνεται πιο εύκολα.

Κάποιοι ακόμα σημαντικοί παράγοντες στην επεξεργασία δεδομένων είναι τα κριτήρια TF-IDF και cosine similarity.

TF-IDF:

Η Συχνότητα Όρου - Αντίστροφη Συχνότητα Όρου είναι ένας αριθμός που δείχνει πόσο σημαντικός είναι ένας όρος (λέξη) για ένα αρχείο μέσα σε μια συλλογή από αρχεία. Η τιμή TF-IDF αυξάνεται ανάλογα με τον αριθμό εμφανίσεων μιας λέξης μέσα στο αρχείο, αλλά αντισταθμίζεται από τη συχνότητα με την οποία η λέξη εμφανίζεται στη συλλογή όλων των αρχείων, ώστε λέξεις που εμφανίζονται σε μεγάλο αριθμό αρχείων να μη θεωρούνται τόσο σημαντικές όσο οι πιο σπάνιες. Αν μια σπάνια λέξη εμφανιστεί σε ένα αρχείο, σημαίνει ότι το αρχείο πολύ πιθανώς χαρακτηρίζεται (και) από τη λέξη αυτή.

Ομοιότητα συνημιτόνων (cosine similarity):

Η ομοιότητα του συνημιτόνου μεταξύ δύο διανυσμάτων (ή δύο εγγράφων που απεικονίζονται ως διανύσματα στον διανυσματικό χώρο) είναι ένα μέτρο που υπολογίζει το συνημίτονο της γωνίας μεταξύ τους. Αποτελεί μία μέτρηση προσανατολισμού και όχι μεγέθους και μπορεί να θεωρηθεί ως μία σύγκριση μεταξύ εγγράφων σε κανονικοποιημένο χώρο επειδή δεν λαμβάνουμε υπόψη μόνο την εμφάνιση κάθε λέξης σε κάθε έγγραφο, αλλά τη γωνία μεταξύ

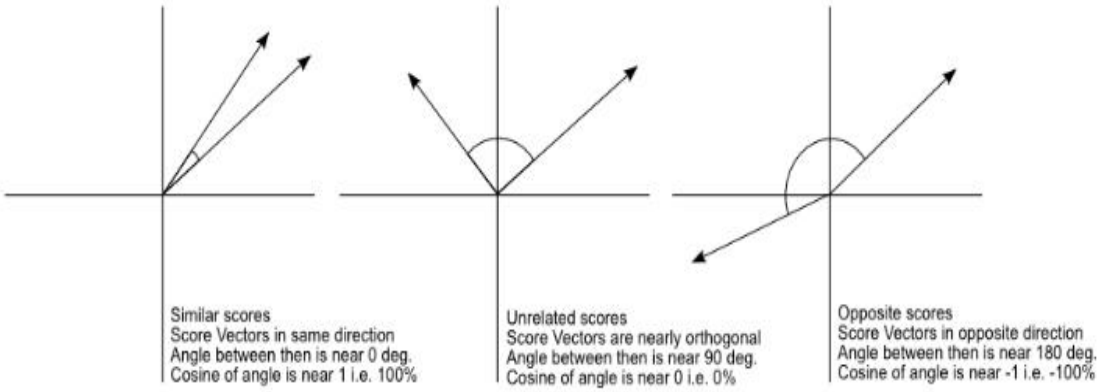
αυτών. Για να παράξουμε την ομοιότητα συνημιτόνων πρέπει να λύσουμε την παρακάτω εξίσωση ως προς το cos*ϑ:*

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Figure 4.1 : Cosine Similarity math type

Με την χρήση του συγκεκριμένου τύπου παράγεται μια μέτρηση που μας δείχνει κατά πόσο σχετίζονται, δηλαδή πόσο όμοια είναι δύο έγγραφα μεταξύ τους. Για παράδειγμα:



The Cosine Similarity values for different documents, 1 (same direction), 0 (90 deg.), -1 (opposite directions).

Figure 4.2: Cosine Similarity example

4.2 Οφέλη επεξεργασίας

Πλήθος τομέων μπορούν να επωφεληθούν από τη χρήση της, με κυριότερο, καταρχάς, την επικοινωνία ανθρώπου-μηχανής(human-computer interaction). Στο χώρο αυτό, η χρήση φυσικής γλώσσας επιτρέπει στους χρήστες να χρησιμοποιούν το φυσικό τρόπο επικοινωνίας τους και όχι τεχνητές γλώσσες (προγραμματισμού, μηχανής, κ.ά.) ή δομημένα μενού. Μια τέτοια προσέγγιση έχει και προτερήματα και μειονεκτήματα. Ναι μεν δεν απαιτείται εκπαίδευση στη χρήση της γλώσσας, αλλά αυτό διευκολύνει περισσότερο τους περιστασιακούς χρήστες και λιγότερο τους εξειδικευμένους, όπως είναι για παράδειγμα οι προγραμματιστές ή οι υπάλληλοι γραφείου που εισάγουν στοιχεία σε φόρμες.

Μια δεύτερη περιοχή είναι αυτή της διαχείρισης πληροφορίας (information management), όπου η NLP θα μπορούσε να ενεργοποιήσει διαδικασίες αυτόματης διαχείρισης και επεξεργασίας της πληροφορίας με βάση τη διερμηνεία της. Αν, για παράδειγμα, ένα σύστημα μπορούσε να κατανοήσει το νόημα ενός εγγράφου, θα μπορούσε να το αρχειοθετήσει μαζί με τα άλλα αντίστοιχα έγγραφα.

Τρίτη περιοχή είναι αυτή της αναζήτησης σε βάσεις δεδομένων (database searching). Οι συνήθεις τρόποι έκφρασης μιας επιθυμητής πληροφορίας είναι μέσω επιλογής από λίστες, συμπλήρωσης μενού ή σύνταξης του αιτήματος σε τεχνητή γλώσσα (special query language-SQL). Η χρήση τεχνητής γλώσσας επιτρέπει μεν την ανάπτυξη απλών μηχανισμών αναζήτησης, αλλά και πάλι ο χρήστης πρέπει να έχει κάποια γνώση σχετικά με τη δομή της βάσης. Από την άλλη πλευρά, ο χρήστης είναι πιο εξοικειωμένος με το περιεχόμενο ή την περιοχή ενδιαφέροντος της βάσης παρά με τη δομή της. Με τη χρήση φυσικής γλώσσας, τα αιτήματα μπορεί να περιοριστούν σε όρους σχετικούς με το περιεχόμενο και την περιοχή ενδιαφέροντος.

4.3 Δυσκολίες στην επεξεργασία

Η μεγαλύτερη δυσκολία στην επεξεργασία φυσικής γλώσσας είναι η διφορούμενη ερμηνεία που προκαλεί ασάφεια στη γλώσσα (ambiguity of language) σε πολλά επίπεδα:

- Καταρχήν, ασάφεια σε επίπεδο σύνταξης (ambiguity at syntactic level) της γλώσσας. Κάποιες συντακτικά ορθά προτάσεις επιδέχονται πάνω από μια διερμηνεία, ανάλογα με το πώς θα αναλυθούν συντακτικά, καθιστάμενες συντακτικά ασαφείς. Για παράδειγμα:

  Χτύπησα τον κλέφτη με το τσεκούρι.

Το τσεκούρι ήταν το όπλο με το οποίο χτύπησα τον κλέφτη ή χτύπησα τον κλέφτη που κρατούσε το τσεκούρι;

- Δευτερευόντως, ασάφεια σε επίπεδο λεξιλογικό (ambiguity at lexical level), όταν το νόημα μιας λέξης είναι διφορούμενο. Για παράδειγμα:

Το πρώτο γράμμα του Γιώργου.

Εννοεί το πρώτο γράμμα που έγραψε ο Γιώργος ή το γράμμα του αλφαβήτου από το οποίο αρχίζει το όνομα «Γιώργος»; Η λέξη γράμμα έχει δυο έννοιες, της επιστολής και του γράμματος του αλφαβήτου.

- Τρίτον, ασάφεια σε αναφορικό επίπεδο (ambiguity at referential level), όταν δεν είναι ευκρινές το σε ποιον, πού ή σε τι η πρόταση αναφέρεται. Για παράδειγμα:

Ο Γιάννης χτύπησε τον Πέτρο, γιατί του αρέσει η Μαίρη.

Σε ποιον αρέσει η Μαίρη, στο Γιάννη ή στον Πέτρο;

- Τέταρτον, ασάφεια σε σημασιολογικό επίπεδο (ambiguity at semantic level), όταν, με διατήρηση της ίδιας συντακτικής ανάλυσης, η πρόταση επιδέχεται τουλάχιστον δυο διαφορετικές ερμηνείες. Για παράδειγμα:

Τον άφησε στα κρύα του λουτρού.

Η πρόταση κυριολεκτεί ότι κάποιος άφησε κάποιον άλλον στα κρύα ενός λουτρού ή παρουσιάζει μεταφορικά ότι τον παράτησε και έφυγε στη μέση κάποιας συνεργασίας;

- Τέλος, ασάφεια σε πραγματολογικό επίπεδο (pragmatic level), κατά τη διερμηνεία μιας πρότασης, όταν λαμβάνουμε υπόψη το πλαίσιο του κειμένου που την περιέχει. Στην παρακάτω πρόταση δεν είναι εύκολο να λυθεί η πραγματολογική ασάφεια:

Οι δεινόσαυροι έχουν εξαφανιστεί πολλά χρόνια.

Πόσα χρόνια είναι τα πολλά χρόνια;

4.4 Εξόρυξη γνώμης και κατηγοριοποιήσεις

Η εξόρυξη γνώμης (opinion mining) είναι ένας τύπος επεξεργασίας φυσικής γλώσσας για την παρακολούθηση της γνώμης του κοινού σχετικά με ένα συγκεκριμένο θέμα ή προϊόν. Η εξόρυξη γνώμης, η οποία είναι σχεδόν όμοια με την ανάλυση συναισθημάτων, περιλαμβάνει την οικοδόμηση ενός συστήματος συλλογής και κατηγοριοποίησης απόψεων σχετικά με ένα θέμα ή προϊόν.

Η ανάλυση συναισθήματος και η εξόρυξη γνώμης είναι σχεδόν το ίδιο πράγμα, ωστόσο υπάρχει μικρή διαφορά μεταξύ αυτών. Η εξόρυξης γνώμης αποσπά και αναλύει τη γνώμη των ανθρώπων για μια οντότητα, ενώ η ανάλυση συναισθήματος ψάχνει πρώτα για το συναίσθημα σε ένα κείμενο και στη συνέχεια το αναλύει.

Οι κύριοι τομείς της εξόρυξης γνώμης είναι οι εξής:

- opinions classification

  Ανάθεση συναισθήματος σε ολόκληρη τη γνώμη ή διαχωρισμό των γνώμεων σε ομάδες με βάση την πολικότητά τους (συνήθως χρησιμοποιούνται δύο ή τρεις ομάδες - θετικές, αρνητικές και μερικές φορές ουδέτερες).

- feature based opinion mining

  Ανακάλυψη σε μία γνώμη των διάφορων πτυχών που αφορούν ένα συγκεκριμένο θέμα, προϊόν, οντότητα και άλλα που αρέσει ή δεν αρέσει στον χρήστη.

- comparative sentences analysis

  Ανάλυση προτάσεων με στόχο να συγκρίνουν απευθείας το ένα αντικείμενο με το άλλο.

Στη συγκεκριμένη διπλωματική εργασία εστιάζουμε στη feature based εξόρυξη γνώμης την οποία αναλύουμε στην συνέχεια.

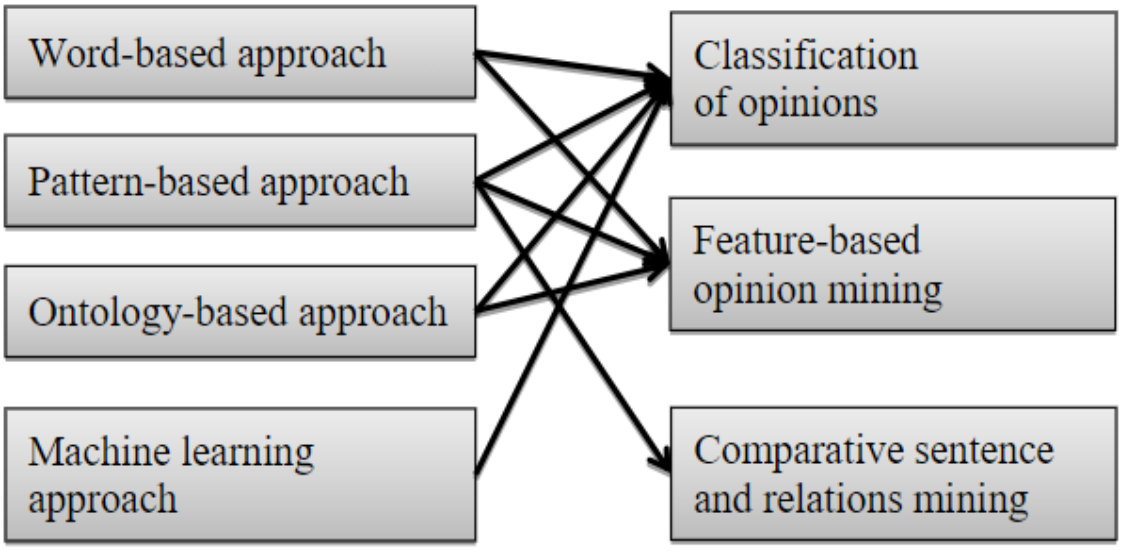


Figure 4.3: Different approaches used in opinion mining analysis

Υπάρχουν 3 κύριες προσεγγίσεις feature based εξόρυξης γνώμης που μπορούν να χρησιμοποιηθούν στην ανάλυση συναισθήματος:

- Word-based approach
- Pattern-based approach
- Ontology-based approach

Word-based approach:

Σε αυτή την προσέγγιση υποθέτουμε πως η σημασία κάθε γνώμης (όπως και το συναίσθημα που υπάρχει σε αυτή) εμπεριέχεται στις λέξεις ξεχωριστά, έτσι ώστε το συναίσθημα να ανατίθεται σε κάθε ξεχωριστή λέξη που υπάρχει μέσα στην γνώμη. Η συγκεκριμένη μέθοδος είναι γενικά αρκετά εύχρηστη. Ωστόσο, απαιτούνται γνώμες που τα πλεονεκτήματα και τα μειονεκτήματα έχουν ξεκαθαριστεί σε κάθε εξετασθείσα γνώμη. Συγκεκριμένα με την βοήθεια του cloud tag βλέπουμε πόσοι χρήστες αναφέρουν συγκεκριμένα χαρακτηριστικά ως πλεονεκτήματα ή μειονεκτήματα. Παρόλα αυτά, δεν μπορούμε να παρατηρήσουμε ποια είναι η σημαντικότητα της γνώμης του συγκεκριμένου χρήστη ή ομάδα χρηστών.

Το cloud tag (ή σύννεφο λέξεων) είναι μια οπτική αναπαράσταση δεδομένων κειμένου, που συνήθως χρησιμοποιείται για την απεικόνιση των μεταδεδομένων (ετικετών) λέξεων-κλειδιών σε ιστότοπους ή για την απεικόνιση ελεύθερου κειμένου. Οι ετικέτες είναι συνήθως μεμονωμένες λέξεις και η σημασία κάθε ετικέτας εμφανίζεται με μέγεθος γραμματοσειράς ή χρώμα.

Pattern-based approach:

Σε αυτή την προσέγγιση υποθέτουμε πως τα συναισθήματα μεταφέρονται από φράσεις/εκφράσεις αντί από ξεχωριστές λέξεις, έτσι ώστε το συναίσθημα να ανατίθεται σε αναγνωρισμένες-προσδιορισμένες φράσεις. Ουσιαστικά, επιτρέπει την αναγνώριση ορισμένων φράσεων στις γνώμες στις οποίες μπορούν να αποδοθούν αισθήματα. Το πλεονέκτημα αυτής της προσέγγισης είναι η δυνατότητα ανίχνευσης φράσεων που τροποποιούν το συναίσθημα όπως η άρνηση, η ενίσχυση και άλλες, δηλαδή η εντόπιση των χαρακτηριστικών στις γνώμες που συνδυάζονται με πολωμένες λέξεις που συνδέονται μέσα σε αυτές. Βασικές προϋποθέσεις είναι ο καθορισμός κανόνων και η απαίτηση ενός λεξικού συναισθήματος. Το μειονέκτημα που πρέπει να ληφθεί υπόψη είναι ότι ορισμένες λέξεις έχουν θετικό νόημα σε ένα περιεχόμενο ενώ αρνητικό σε ένα άλλο.

Ontology-based approach:

Σε αυτή την προσέγγιση η οντολογία χρησιμοποιείται για να παρουσιαστεί προϋπάρχουσα και συγκεκριμένη γνώση όσον αφορά το θέμα που αφορά μία συγκεκριμένη γνώμη. Η οντολογία είναι μία τυπική και κοινή προδιαγραφή ενός τομέα ενδιαφέροντος και μπορέι να χρησιμιποιηθεί για να τον περιγράψει και να αιτιολογήσει τις οντότητες εντός του. Εκπροσωπεί τη γνώση ως ένα συνολο εννοιών εντός ενός τομέα μαζί με τις σχέσεις μεταξύ αυτών των εννοιών. Οι κλάσεις (έννοιες) στην οντολογία μπορούν να έχουν ιεραρχική δομή και περιέχουν ένα σύνολο αντικειμένων (άτομα, στιγμιότυπα εννοιών) που αντιπροσωπεύουν πραγματικά αντικείμενα ή όντα από ένα συγκεκριμένο τομέα. Οι έννοιες μπορεί να έχουν ιδιότητες που εκφράζουν την σημασία τους. Ωστόσο, για την κατασκευή της οντολογίας απαιτείται γνώση σχετικά με τον συγκεκριμένο τομέα ενδιαφέροντος, δηλαδή δεν μπορεί να εφαρμοστεί η ίδια οντολογία σε διαφορετικό τομέα ενδιαφέροντος. Επιπλέον, η δυσκολία αυτής της προσέγγισης έγκειται στη διαδικασία εκχώρησης του συναισθήματος σε κάθε χαρακτηριστικό ξεχωριστά, αλλά μπορεί να επιτευχθεί χειροκίνητα με την επισήμανση ορισμένων μερών απόψεων και την ανάθεση τους σε κατάλληλα features ως χαρακτηριστικά (attributes). Τέλος, απαιτείται η προετοιμασία ειδικών λεξικών που περιέχουν θετικές και αρνητικές λέξεις ή εκφράσεις οι οποίες περιέχο εκ των προτέρων συναισθήματα.

Κατόπιν, είναι σημαντικό να αναφερθεί πως αρκετές φορές παρατηρείται και συνδυασμός των συυκεκριμένων προσεγγίσεων σε projects που αφορούν feature-based εξόρυξη γνώμης (multi-model approach).

4.5 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis) είναι μια διαδικασία υπολογιστικής ταυτοποίησης και κατηγοριοποίησης των απόψεων που εκφράζονται σε ένα κομμάτι κειμένου, ειδικά προκειμένου να καθοριστεί εάν η στάση του συγγραφέα έναντι ενός συγκεκριμένου θέματος, προϊόντος κ.λπ. είναι θετική, αρνητική ή ουδέτερη. Τα δεδομένα που αναλύονται ποσοτικοποιούν τα συναισθήματα ή τις αντιδράσεις του ευρύτερου κοινού προς συγκεκριμένα προϊόντα, ανθρώπους, αντιλήψεις ή ιδέες, κοινωνικά φαινόμενα και αποκαλύπτουν την πολικότητα των πληροφοριών.

Επιπρόσθετα, είναι ένας αναδυόμενος κλάδος στον τομέα της επεξεργασίας φυσικής γλώσσας και της εξόρυξης γνώσης από το διαδίκτυο, ο οποίος παρέχει έναν τρόπο για τη διαδικασία λήψης αποφάσεων σε διάφορους τομείς όπως το marketing, η εκπαίδευση, η υγειονομική περίθαλψη, η διαχείριση οικονομικών και ανθρώπινων πόρων και άλλων. Οι άνθρωποι που ασχολούνται με την επιχειρηματική ανάπτυξη και εργάζονται σε διάφορα τμήματα εντοπίζουν τα συναισθήματα των πελατών σε κοινωνικά δίκτυα, κοινότητες ιστού, ιστολόγια και άλλα μέσα επικοινωνίας μέσω διαδικτύου. Οι άνθρωποι πλέον, εκφράζουν επί το πλείστον τη γνώμη τους σχετικά με προϊόντα, την οργάνωση, την υγειονομική τους κατάσταση, τις υπηρεσίες του δημόσιου και ιδιωτικού τομέα καθώς και κοινωνικά φαινόμενα χωρίς κανένα δισταγμό με αποτέλεσμα να δημιουργείται μία τεράστια “δεξαμενή” πληροφορίας, δηλαδή

δεδομένων, τα οποία βοηθούν τους ειδικούς να εξάγουν χρήσιμα συμπεράσματα. Στη συγκεκριμένη εργασία ασχολούμαστε με την εξόρυξη κειμένου. Η εξόρυξη κειμένου ασχολείται με προβλήματα επεξεργασίας εγγράφων κειμένων και εξαγωγή γνώσης από τα κείμενα επεξεργασίας.

Παρόλα αυτά η ανάλυση συναισθήματος δεν είναι μια τέλεια επιστήμη. Η ανθρώπινη γλώσσα είναι περίπλοκη με αποτέλεσμα η διδασκαλία ενός μηχανήματος όσον αφορά την κατανόηση ενός κειμένου είναι μία δύσκολη διαδικασία. Η διδασκαλία μιας μηχανής για να κατανοήσει το ύφος ενός κειμένου είναι ακόμη πιο δύσκολη.

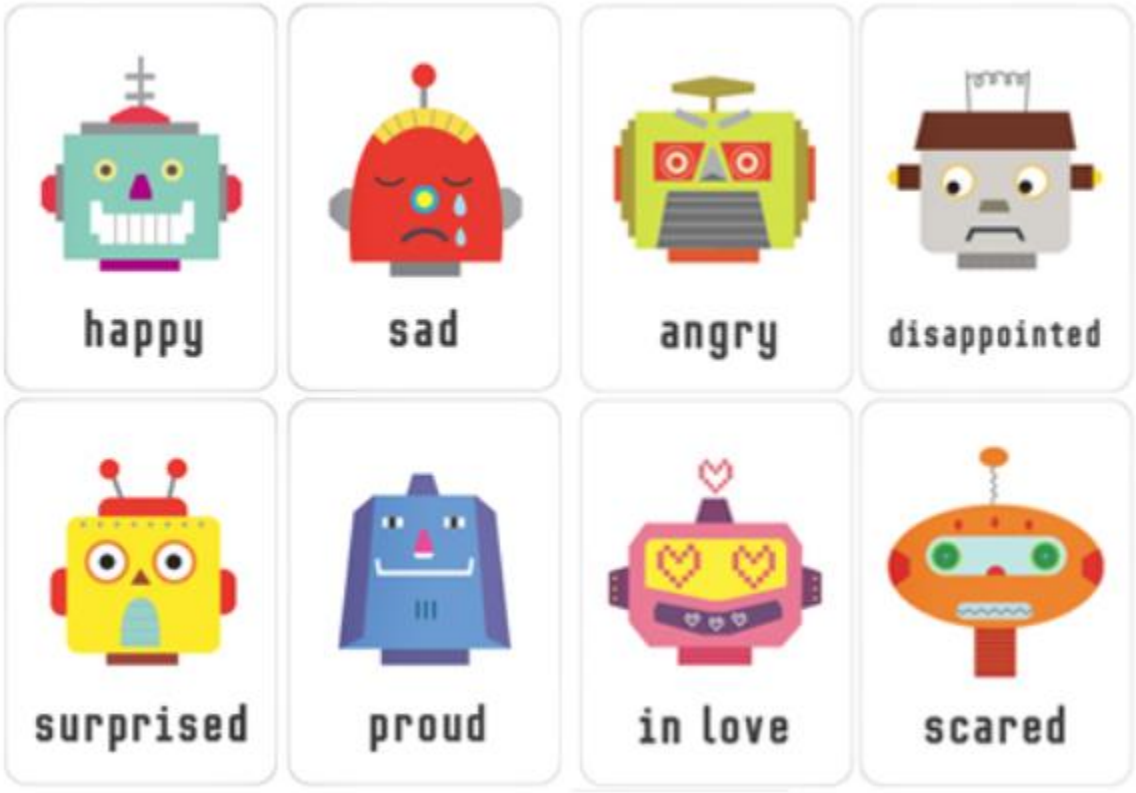


Figure 4.4: Πηγή https://www.brandwatch.com

Οι άνθρωποι είναι αρκετά διαισθητικοί όταν πρόκειται να ερμηνεύσουν τον τόνο ενός κομματιού γραφής. Για παράδειγμα ας εξετάσουμε την ακόλουθη πρόταση: "Η πτήση μου καθυστέρησε. Τέλεια!". Οι περισσότεροι θα μπορούσαν γρήγορα να καταλάβουν ότι το ύφος της πρότασης είναι σαρκαστικό. Χωρίς συμφραζόμενη κατανόηση όμως, μια μηχανή που κοιτάζει την παραπάνω πρόταση μπορεί να δει τη λέξη "τέλεια" και να την κατηγοριοποιήσει ως θετική με συνέπεια να εκλάβει όλο το ύφος της πρότασης ως τέτοιο. Το παραπάνω παράδειγμα δείχνει πως η ανάλυση συναισθήματος έχει τους περιορισμούς της και δεν πρέπει να χρησιμοποιηθεί ως ένας δείκτης ακρίβειας ποσοστού 100%. Όπως και με οποιαδήποτε αυτοματοποιημένη διαδικασία, είναι επιρρεπής σε σφάλματα και συχνά χρειάζεται ένα ανθρώπινο μάτι για να το παρακολουθήσει.

Κεφάλαιο 5ο: Μετανάστευση

5.1 Εισαγωγή

Η μετανάστευση (immigration), τόσο κατά τις κοινωνικές επιστήμες όσο και κατά το Διεθνές Δίκαιο, είναι η μετακίνηση ανθρώπων σε μία χώρα της οποίας δεν έχουν την ιθαγένεια, προκειμένου να εγκατασταθούν εκεί, ιδιαίτερα ως μόνιμοι κάτοικοι ή μελλοντικοί πολίτες της χώρας. Αποτελεί μια δυναμική διαδικασία, οι μορφές της οποίας ποικίλλουν και μεταβάλλονται σε συνάρτηση με τις ευρύτερες πολιτικές, κοινωνικές και οικονομικές αλλαγές. Η μετανάστευση ενυπάρχει και αναδύεται από την ιστορική και κοινωνική οργάνωση και στο σύνολό της είναι μια μορφή κοινωνικής σχέσης που καθορίζεται από την αγορά, το έθνος, το κράτος, το φύλο, τις κοινωνικές κατηγορίες και τον τρόπο επαφής και επικοινωνίας μεταξύ τους.

Οι μετανάστες λοιπόν, είναι οι άνθρωποι που εγκαταλείπουν την πατρίδα τους είτε με τη θέλησή τους είτε χωρίς αυτή, προς αναζήτηση νέων ευκαιριών καθώς και ασφαλέστερων και καλύτερων προοπτικών διαβίωσης. Μετανάστης θεωρείται ένας άνθρωπος που διαμένει τουλάχιστον 6 μήνες μακριά από τον συνήθη τόπο κατοικίας του. Ο 20^{ος} αιώνας έχει χαρακτηρισθεί ως «ο αιώνας της μετανάστευσης» διότι σημειώθηκαν οι σημαντικότερες πληθυσμιακές μετακινήσεις και αλλαγές στη σύνθεση του πληθυσμού (βίαιες ή ειρηνικές).

Figure 5.1: Πηγή https://www.zougla.gr

5.2 Αίτια της μεταναστευτικής κίνησης

Οι λόγοι για τους οποίους μεταναστεύει κανείς είναι πολλοί, διαφορετικοί και ενδεχομένως συνδυαστικοί. Είναι αλήθεια πως οι κοινωνικό-οικονομικές ανισότητες βορρά και νότου καθώς και οι σοβαρές παραβιάσεις ανθρωπίνων δικαιωμάτων υποβόσκουν ως οι σημαντικότερες αιτίες στις οποίες προστίθενται τις τελευταίες δεκαετίες η παγκοσμιοποίηση και η εξέλιξη τεχνολογίας των πληροφοριών, αν όχι σαν αιτίες σίγουρα σαν σημαντικοί παράγοντες. Μπορούμε λοιπόν να διακρίνουμε τις ακόλουθες ενδεικτικές γενικές κατηγορίες:

- Φυσικοί παράγοντες

Μεταβολές στο φυσικό περιβάλλον που καθιστούν δύσκολη την επιβίωση, όπως ξηρασία, πλημμύρες, σεισμοί κ.λπ. Η σημασία των παραγόντων αυτών είναι μεγαλύτερη όσο χαμηλότερο είναι το επίπεδο της τεχνολογίας και επομένως η εξάρτηση ενός πληθυσμού από τη φύση και την επιτόπια παραγωγή.

- Οικονομικοί παράγοντες

Σαν τέτοιοι θα μπορούσαν να αναφερθούν η έλλειψη επαρκών δυνατοτήτων απασχόλησης, η υποαπασχόληση, το χαμηλό εισόδημα σε συνδυασμό πολλές φορές με την υπέρμετρη χρονικά εργασιακή απασχόληση, η αναγκαστική μετακίνηση σαν αναπόσπαστο στοιχείο για την άσκηση συγκεκριμένης επαγγελματικής δραστηριότητας και άλλοι.

- Πολιτικοί παράγοντες

Σ' αυτή την κατηγορία εντάσσονται οι διώξεις λόγω φυλής, θρησκείας, εθνικότητας, κοινωνικής τάξης, πολιτικών ή άλλων πεποιθήσεων, καθώς και η διακριτική μεταχείριση που πολλές φορές ακολουθείται από το καθεστώς μίας χώρας σε βάρος μεμονωμένων ατόμων ή και συγκεκριμένης κατηγορίας ενός πληθυσμού. Είναι δυνατόν πολλές φορές ταυτόχρονα πολιτικοί και οικονομικοί παράγοντες να ωθούν στη φυγή και τη μετανάστευση πολίτες μίας χώρας.

- Κοινωνικοί παράγοντες

Εδώ θα πρέπει να αναφερθεί η αξία της μετανάστευσης ως προϋπόθεση κοινωνικής προκοπής ή ανόδου στον συγκεκριμένο κοινωνικό ιστό μίας "εθνικής" κοινωνίας και τα οφέλη που σε ατομικό επίπεδο συνοδεύουν μία τέτοια ανέλιξη.

- Ψυχολογικοί παράγοντες

Πολλές φορές άνθρωποι ωθούνται στη μετανάστευση από τυχοδιωκτισμό ή φιλαποδημία ή τέλος μιμητικά, ακολουθώντας την κρατούσα τάση μίας συγκεκριμένης χρονικής περιόδου και τους γνωστούς, φίλους, συγγενείς, ομοπάτριους κ.λπ. που προηγήθηκαν.

Κατόπιν, είναι πολύ σημαντικό να τονιστεί ότι σημαντική επίδραση ασκούν οι προσδοκίες από τον τόπο προορισμού, οι οποίες πολλές φορές είναι υπερμεγέθεις λόγω εσφαλμένης (πολύ συχνά σκόπιμης) πληροφόρησης.

Figure 5.2: Πηγή https://www.inred.gr

5.3 Η μετανάστευση στις μέρες μας

Η μετανάστευση ήταν πάντοτε ένα κεντρικό τμήμα της ανθρώπινης Ιστορίας -αρχίζοντας φυσικά με την έξοδο του Homo Sapiens από την Αφρική πριν από 125.000 χρόνια και τη διασπορά του στη Μέση Ανατολή, τη Μικρά Ασία, την Κεντρική και Νότια Ασία και τελικά τον Νέο Κόσμο.

Και σε πιο πρόσφατες σελίδες της ανθρώπινης Ιστορίας, η μετανάστευση έπαιζε κεντρικό ρόλο, με αποκορύφωμα την αναγκαστική μεταφορά 12 εκατομμυρίων ανθρώπων κυρίως από τη Δυτική Αφρική στον Νέο Κόσμο, όπου θα χρησίμευαν ως δούλοι. Πολύ σημαντικό επίσης ρόλο στην ιστορία των μαζικών μετακινήσεων έπαιξε η άνοδος των ΗΠΑ ως βιομηχανικής δύναμης. Μεταξύ του 1850 και της Μεγάλης Κρίσης της δεκαετίας του '30, πάνω από 12 εκατομμύρια εργαζόμενοι έφυγαν από τις χώρες της βόρειας, νότιας και ανατολικής Ευρώπης για τις ΗΠΑ σε αναζήτηση καλύτερων συνθηκών διαβίωσης.

Η μετανάστευση, λοιπόν, δεν είναι άγνωστο φαινόμενο στην Ιστορία μας. Ωστόσο, από το 1960 μέχρι το 2005 ο αριθμός των μεταναστών ναι μεν διπλασιάστηκε, αλλά σε ποσοστά επί του συνολικού πληθυσμού περίπου παραμένει στο 3%. Πράγμα που σημαίνει πως ενώ η

μετανάστευση είναι τόσο παλιά όσο και η ανθρωπότητα, για κάποιο λόγο σήμερα είναι πιο ορατή από ποτέ.

Παρ' όλα αυτά υπάρχουν σήμερα ορισμένα στοιχεία που διαφοροποιούν τη σημερινή μετανάστευση από ανάλογα φαινόμενα του παρελθόντος. Οπως υποστηρίζει ο Khalid Kosher σε άρθρο του στο «Current History» δύο από τα νέα στοιχεία είναι:

- Ο μεγάλος αριθμός των μεταναστών.
- Η εκρηκτική αύξηση της λαθρομετανάστευσης.

Οσον αφορά τον αριθμό των μεταναστών, αν προσδιορίσουμε ως μετανάστη κάποιον που μένει έξω από τη χώρα του πάνω από ένα έτος, υπολογίζεται ότι υπάρχουν σήμερα πάνω από 200 εκατομμύρια μετανάστες παγκοσμίως -όσο περίπου ο πληθυσμός της Βραζιλίας. Σήμερα 1 στα 35 άτομα είναι ένας διεθνής μετανάστης. Η μετανάστευση είναι επίσης ένα πολύ πιο παγκοσμιοποιημένο φαινόμενο σε σχέση με άλλες περιόδους της Ιστορίας, καθώς οι μετανάστες ταξιδεύουν σε όλα τα μήκη και πλάτη του κόσμου. Το 2005 υπήρχαν 60 εκατομμύρια διεθνείς μετανάστες στην Ευρώπη, 44 εκατομμύρια στην Ασία, 41 εκατομμύρια στη Βόρειο Αμερική, 16 εκατομμύρια στην Αφρική και από 6 εκατομμύρια στη Λατινική Αμερική και στην Αυστραλία. Το μεγαλύτερο ποσοστό των μεταναστών -35 εκατομμύρια- ζουν στις ΗΠΑ και ακολουθεί η Ρωσική Ομοσπονδία με 13 εκατομμύρια και η Γερμανία, η Ουκρανία και η Ινδία, που η κάθε μια φιλοξενεί περίπου 7 εκατομμύρια μετανάστες.

Η προέλευση των μεταναστών είναι δύσκολο να υπολογιστεί, καθώς οι χώρες από τις οποίες φεύγουν δεν κρατούν αρχεία σχετικά με τον αριθμό των πολιτών τους που ζουν εκτός των συνόρων. Πάντως, υπολογίζεται ότι τουλάχιστον 35 εκατομμύρια Κινέζοι ζουν έξω από τη χώρα τους, καθώς και 20 εκατομμύρα Ινδοί και 8 εκατομμύρια Φιλιππινέζοι.

Το δεύτερο πολύ σημαντικό χαρακτηριστικό της σύγχρονης μετανάστευσης είναι η τεράστια αύξηση των λαθρομεταναστών. Φυσικά, για ευνόητους λόγους, είναι πολύ δύσκολο να υπολογιστεί ο ακριβής αριθμός των λαθρομεταναστών. Πάντως, σύμφωνα με κοινώς αποδεκτούς υπολογισμούς υπάρχουν σήμερα περίπου 40 εκατομμύρια λαθρομετανάστες, το ένα τρίτο των οποίων ζουν σήμερα στις ΗΠΑ. Επίσης υπάρχουν 5 εκατομμύρια στη Ρωσική Ομοσπονδία και πιθανώς 5 εκατομμύρια στην Ευρώπη. Υπολογίζεται ότι ετησίως 4 εκατομμύρια άτομα περνούν λαθραία τα σύνορα των διαφόρων χωρών.

Είναι γεγονός ότι η λαθρομετανάστευση είναι το στοιχείο εκείνο στο οποίο επικεντρώνεται κάθε δημόσια συζήτηση για τη μετανάστευση. Αυτό οφείλεται σε πολλούς λόγους. Ενας λόγος είναι ο αυξημένος κίνδυνος για την ασφάλεια που συνεπάγεται η λαθρομετανάστευση -οι τρομοκρατικές ενέργειες, η μεταφορά μεταδοτικών ασθενειών και, φυσικά, η εγκληματικότητα. Επίσης, η λαθρομετανάστευση θεωρείται ότι παραβιάζει άμεσα τα κυριαρχικά δικαιώματα των κρατών. Τέλος, τα παράνομα δίκτυα διακίνησης λαθρομεταναστών συνιστούν άμεσο κίνδυνο για την εθνική ασφάλεια της χώρας, καθώς ενισχύουν τη διαφθορά των αρχών και το οργανωμένο έγκλημα.

Ομως ίσως τις χειρότερες επιπτώσεις της λαθρομετανάστευσης να τις βιώνουν οι νόμιμοι μετανάστες. Στον βαθμό που οι λαθρομετανάστες παίρνουν θέσεις εργασίας από τους ντόπιους (προφέροντας την εργασία τους φτηνότερα), προκαλούν την έκρηξη ξενοφοβίας μεταξύ του ντόπιου πληθυσμού που στρέφεται εναντίον πάντων -τόσο των λαθρομεταναστών όσο και των νόμιμων μεταναστών.

«Οταν η λαθρομετανάστευση -γράφει ο Koser- έχει ως αποτέλεσμα τον ανταγωνισμό για σπάνιες θέσεις εργασίας, μπορεί να δημιουργήσει αισθήματα ξενοφοβίας. Αυτό που είναι σημαντικό είναι ότι τα ξενοφοβικά αισθήματα σ' αυτές τις περιπτώσεις δεν έχουν στόχο μόνο τους λαθρομετανάστες, αλλά επίσης και τους νόμιμους μετανάστες, τους πρόσφυγες και τα μέλη των εθνικών μειονοτήτων».

Figure 5.3: Πηγή Πηγή https://www.kepsy.gr

Κεφάλαιο 6ο: Περιγραφή του συστήματος αναγνώρισης συναισθημάτων

6.1 Απαραίτητο θεωρητικό υπόβαθρο

Προκειμένου να κατανοήσει ο αναγνώστης το περιεχόμενο αυτής της διπλωματικής εργασίας και τη χρήση διαφόρων εργαλείων, θα πρέπει να έχει ένα βασικό επίπεδο γνώσεων στα παρακάτω θέματα.

6.1.1 Εξόρυξη γνώσης από δεδομένα

Για την υλοποίηση αυτού του συστήματος χρησιμοποιούμε την πλατφόρμα του RapidMiner για την κατηγοριοποίηση κάθε κειμένου σε θετικό ή αρνητικό. Ο αναγνώστης πρέπει να έχει τις βασικές γνώσεις εξόρυξης γνώσεων από δεδομένα που συμπεριλαμβάνουν την επεξεργασία κειμένου για εξαγωγή διανύσματος των χαρακτηριστικών του και τους αλγορίθμους μάθησης και δημιουργίας μοντέλων.

6.1.2 Διαχείριση βάσεων δεδομένων

Στην εφαρμογή μας χρησιμοποιούμε βάση δεδομένων έτσι ώστε να διατηρήσουμε σε ασφαλές μέρος τα δεδομένα που απαιτούνται για την ανάλυση συναισθήματος. Ο αναγνώστης χρειάζεται να έχει βασικες γνώσεις για τη συσχέτιση,διαχείριση πινάκων καθώς και την εισαγωγή και εξαγωγή δεδομένων από αυτούς.

6.1.3 Γλώσσα προγραμματισμού JAVA

Οι εφαρμογές που υποστηρίζουν την ανίχνευση και αποθήκευση των κειμένων που χρειαζόμαστε είναι γραμμένες σε γλώσσα προγραμματισμού JAVA. Ο αναγνώστης της παρούσας διπλωματικής θα πρέπει να διαθέτει βασικές γνώσεις των frameworks που διαχειρίζονται την σύνδεση της JAVA με βάσεις δεδομένων(MySQL Workbench).Επίσης χρειάζεται να έχει βασικές γνώσεις όσον αφορά τα API’s γενικότερα αλλά και την χρησιμοποίησή τους όσον αφορά τα Twitter και YouTube ειδικότερα.

6.2 Λογισμικό

Το λογισμικό που χρησιμοποιήθηκε για την υλοποίηση των συγκεκριμένων εφαρμογών είναι το εξής:

- Η υλοποίηση των εφαρμογών σε γλώσσα JAVA επιτεύχθηκε με την βοήθεια του προγραμματιστικού περιβάλλοντος NetBeans IDE 8.2.
- Η αποθήκευση των tweets και των σχολίων από τα βίντεο του YouTube έγινε χρησιμοποιώντας το MySQL Workbench 6.3 CE.
- Η αναγνώριση συναισθημάτων έγινε με την βοηθεία του Rapidminer 8.1 και ειδικότερα με την προέκταση του Text Processing 8.1.0 για τις διάφορες διεργασίες που αφορούν την επεξεργασία κειμένου.

6.3 NetBeans IDE

Για να καταστεί εφικτή η δημιουργία ενός Data Set ώστε να γίνει η επιθυμητή αναγνώριση συναισθήματος έπρεπε να δημιουργήσουμε ένα σύνολο εφαρμογών το οποίο αρχικά θα κάνει το διαχωρισμό του Training Set. Στην συνέχεια θα μας επιτρέπει να αναζητούμε και να αποθηκεύουμε σε μία βάση δεδομένων τα tweets και τα σχόλια από το βίντεο του YouTube που επιθυμεί ο χρήστης και τέλος θα ανακτά αυτά τα κείμενα απο την βάση δεδομένων και θα εφαρμόζει πάνω τους ένα μοντέλο αναγνώρισης συναισθημάτων..

Συνεπώς για να είμαστε σε θέση να αποθηκεύσουμε τα απαραίτητα tweets και comments πρέπει να εξασφαλίσουμε αρχικά άδεια και στη συνέχεια μοναδικά credentials τόσο από το Twitter όσο και από το YouTube το οποίο ανήκει στην Google.

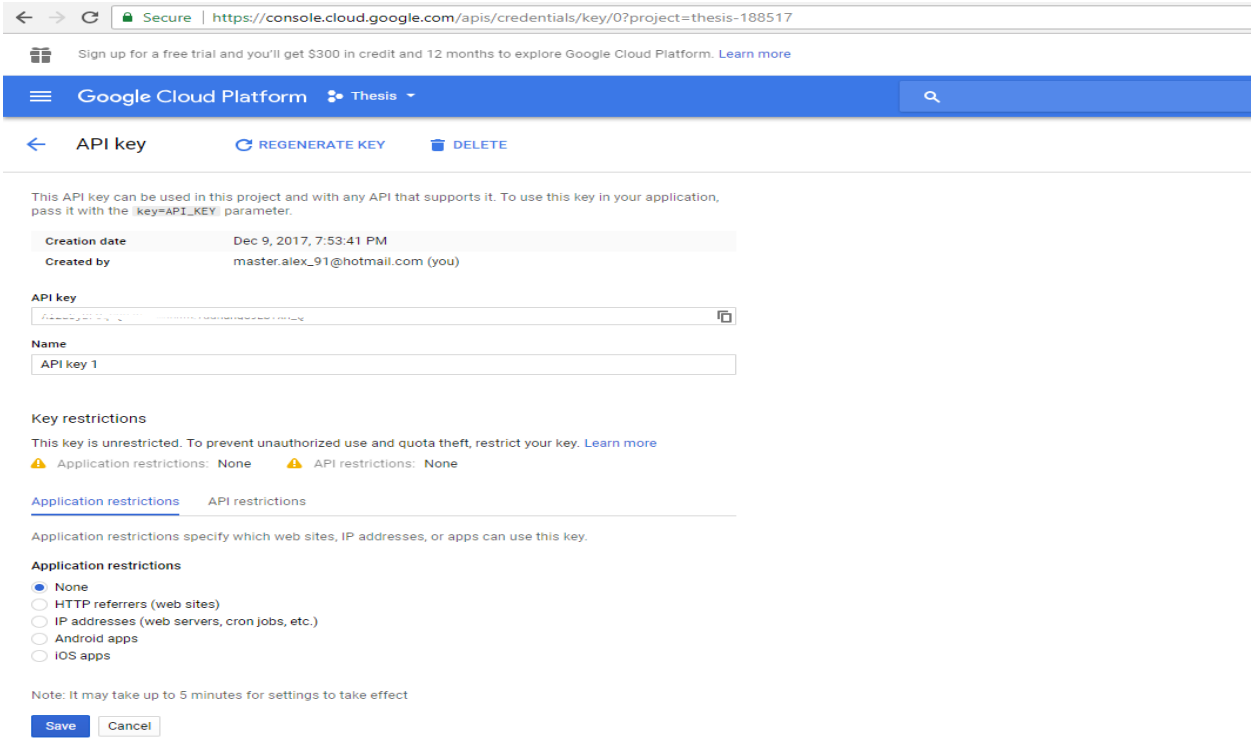


Figure 6.1: Google API Credentials

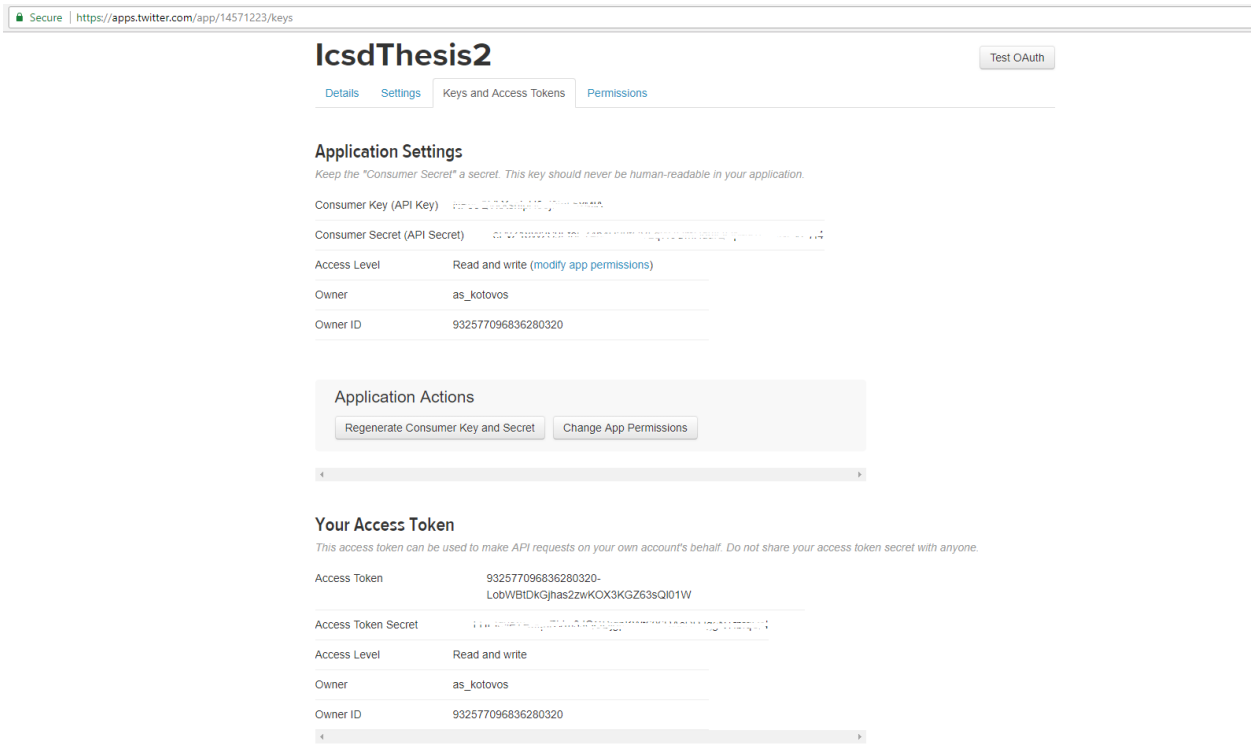


Figure 6.2: Twitter API Credentials

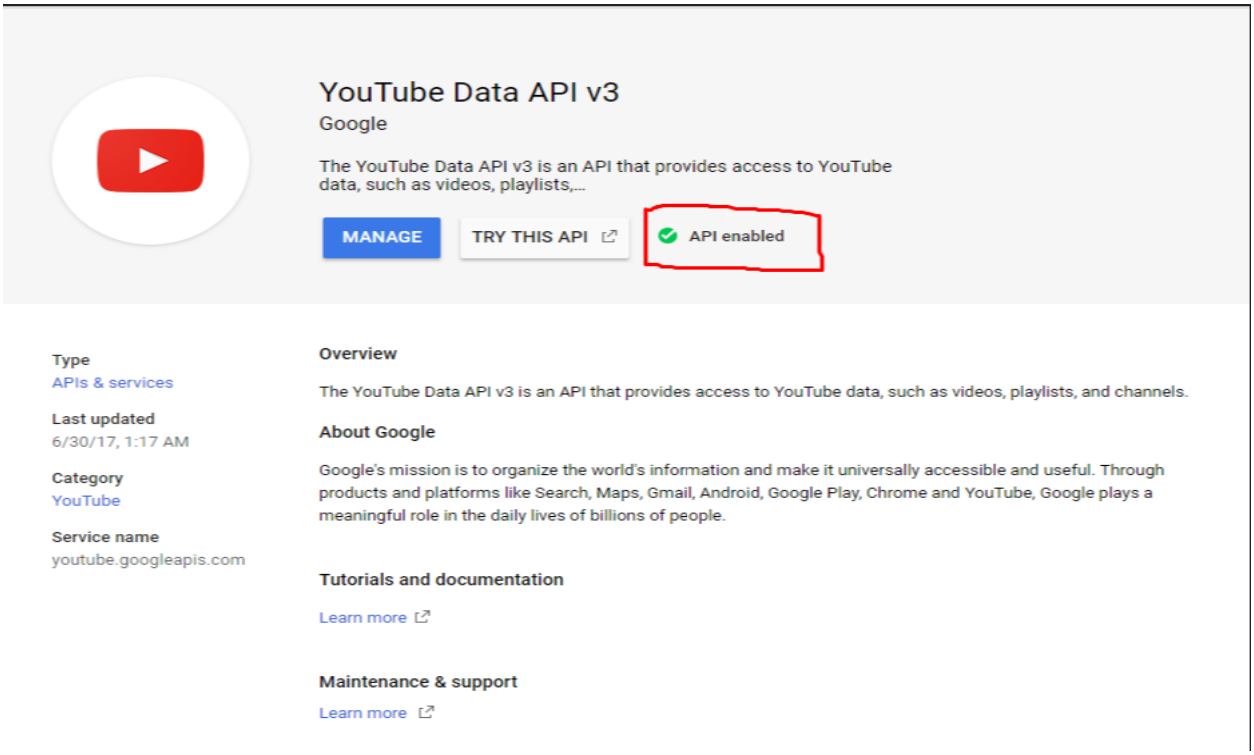


Figure 6.3: YouTube Data API v3 Enabled

6.3.1 ThesisCorpus

Η πρώτη εφαρμογή που δημιουργήσαμε σε JAVA είναι υπεύθυνη για το διαχωρισμό των κειμένων στις διαθέσιμες κλάσεις συναισθημάτων. Πιο συγκεκριμένα χρησιμοποιεί ως είσοδο το Excel που περιέχει τα rated κείμενα, τα χωρίζει σε θετικά και αρνητικά και δημιουργεί τα αρχεία κειμένου που είναι απαραίτητα για την διαδικασία της δημιουργίας του μοντέλου.

```
/*
 * @author Alex Kotovos
 */
public class ThesisCorpus {

    /**
     * @param args the command line arguments
     */
    public static void main(String[] args) {
        try {
            FileInputStream file = new FileInputStream(new File("C:\\Users\\Alex Kotovos\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\Sentiments\\Tweets.xlsx"));

            //Create Workbook instance holding reference to .xlsx file
            XSSFWorkbook workbook = new XSSFWorkbook(file);

            //Get first/desired sheet from the workbook
            XSSFSheet sheet = workbook.getSheetAt(0);
```

Figure 6.4: Άνοιγμα αρχείου Excel με χρήση του XSSFWorkBook

```
        //Iterate through each rows one by one
        Iterator<Row> rowIterator = sheet.iterator();
        int counter = 0;
        while (rowIterator.hasNext())
        {
            Row row = rowIterator.next();

            if(row.getCell(0) != null && row.getCell(1) != null && row.getCell(0).getCellType() == Cell.CELL_TYPE_STRING && row.getCell(1).getCellType() == Cell.CELL_TYPE_STRING){
                String tweet =  row.getCell(0).getStringCellValue();
                String annotation =  row.getCell(1).getStringCellValue();


                if(annotation.equals("positive")){
                    PrintWriter out = new PrintWriter("C:\\Users\\Alex Kotovos\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\Sentiments\\Positive\\"+counter+".txt");
                    out.println(tweet);
                    out.close();
                }else if(annotation.equals("negative")){
                    PrintWriter out = new PrintWriter("C:\\Users\\Alex Kotovos\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\Sentiments\\Negative\\"+counter+".txt");
                    out.println(tweet);
                    out.close();
                }

                counter++;
            }

        }
        file.close();
    } catch (Exception e) {
        e.printStackTrace();
    }
```

Figure 6.5: Διαχωρισμός των εγγραφών του Excel σε Positive και Negative και αποθήκευση σε αρχεία κειμένου

6.3.2 Thesis-ThesisCrawler

Το project Thesis-ThesisCrawler δημιουργήθηκε για να μας προσφέρει ένα παραθυρικό περιβάλλον το οποίο μας επιτρέπει να χρησιμοποιήσουμε το API του YouTube και του Twitter προκειμένου να τραβήξουμε και να αποθηκεύσουμε τα comments και τα tweets αντίστοιχα.

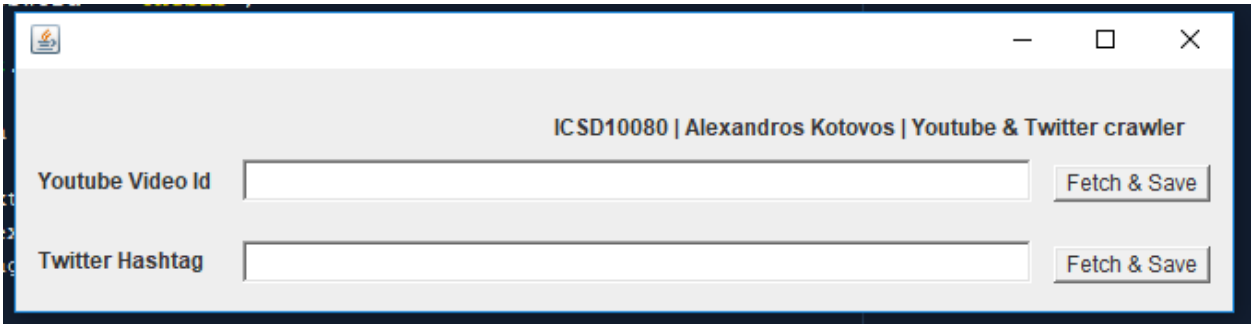


Figure 6.6: Το παραθυρικό περιβάλλον του ThesisCrawler

Εισάγωντας ένα video Id και πατώντας το “Fetch & Save” εκτελείται το παρακάτω κομμάτι κώδικα. Πιο συγκεκριμένα δημιουργούμε τη συνδεση με τη βάση δεδομένων και στη συνέχεια καλούμε τη συνάρτηση getCommentsByURL η οποία «διαβάζει» τα comments και επιστρέφει το token για την επόμενη σελίδα αποτελεσμάτων.

```
private void button1ActionPerformed(java.awt.event.ActionEvent evt) {//GEN-FIRST:event_button1Actio
    String youtube_video_id = this.textField1.getText();

    try {

        String url = "jdbc:mysql://62.210.36.88:3306/thesis";
        String username = "thesis";
        String password = "thesis";

        System.out.println("Connecting database...");

        Connection connection = (Connection) DriverManager.getConnection(url, username, password);

        String nextPage = this.getCommentsByURL(youtube_video_id, "", connection);
        while (!nextPage.isEmpty()) {
            nextPage = this.getCommentsByURL(youtube_video_id, nextPage, connection);
        }

        connection.close();
```

Figure 6.7: Call To Action Listener - YouTube Fetch & Save

Στο παρακάτω κομμάτι κώδικα δημιουργούμε δυναμικά τα URL που απαιτούνται για τη κλήση του YouTube comments API της Google. Σε αυτά τα URL ενσωματώνουμε το Video ID από την εισαγωγή του χρήστη και το nextPageToken όταν αυτό απαιτείται. Το αποτέλεσμα της κλήσης του API είναι ένα JSON αρχείο με τα δεδομένα των διαφόρων comments. Για το λόγο αυτό χρησιμοποιούμε το JSONParser ώστε να μετατρέψουμε το HTTPResponse σε JSONObject προκειμένου να ανακτήσουμε τη πληροφορία που θέλουμε.

```
public String getCommentsByURL(String youtube_video_id, String nextPageToken, Connection connection) throws MalformedURLException, IOException, ParseException, SQLException {
    String url_string = "";
    if (nextPageToken.isEmpty()) {
        url_string = "https://www.googleapis.com/youtube/v3/commentThreads?key=AIzaSyBFOqCQVZUCsm8bRnz7ddhuHgG9EbTXH_Q&textFormat=plaintext&part=snippet&videoId="
                + youtube_video_id + "&maxResults=100";
    } else {
        url_string = "https://www.googleapis.com/youtube/v3/commentThreads?key=AIzaSyBFOqCQVZUCsm8bRnz7ddhuHgG9EbTXH_Q&textFormat=plaintext&part=snippet&videoId="
                + youtube_video_id + "&maxResults=100&pageToken=" + nextPageToken;
    }

    System.out.println(url_string);
    URL url = new URL(url_string);

    String response = "";
    try (BufferedReader reader = new BufferedReader(new InputStreamReader(url.openStream(), "UTF-8"))) {
        for (String line; (line = reader.readLine()) != null;) {
            response += line;
        }
    }

    JSONParser parser = new JSONParser();
    JSONObject json = (JSONObject) parser.parse(response);
```

Figure 6.8: Setting up Connection with Google API for YouTube Comments and JSONParser

Στο παρακάτω κομμάτι κώδικα λαμβάνουμε το nextPageToken από την απάντηση του API ώστε να το επιστρέψουμε για την αμέσως επόμενη κλήση. Στη συνέχεια κάνουμε ένα for-loop ανατρέχουμε σε κάθε ένα από τα comments και διαχειριζόμαστε τη JSON δομή ώστε να αποσπάσουμε το "COMMENT" από την «άχρηστη» πληροφορία. Τέλος δημιουργούμε τα Insert Statements για την εισαγωγή των εγγραφών στη βάση δεδομένων[1].

```
String returnNextPageToken = (String) json.get("nextPageToken");

JSONArray ja = (JSONArray) json.get("items");
for (int i = 0; i < ja.size(); i++) {
    JSONObject current = (JSONObject) ja.get(i);
    JSONObject snippet = (JSONObject) current.get("snippet");
    JSONObject topLevelComment = (JSONObject) snippet.get("topLevelComment");
    JSONObject snippet2 = (JSONObject) topLevelComment.get("snippet");

    this.jLabel5.setText("Comment Fetched and stored : " + (this.COUNT + i));

    Statement st = connection.createStatement();

    java.util.Date dt = new java.util.Date();

    java.text.SimpleDateFormat sdf
            = new java.text.SimpleDateFormat("yyyy-MM-dd HH:mm:ss");

    String currentTime = sdf.format(dt);

    String COMMENT = (String) snippet2.get("textOriginal");
    COMMENT = COMMENT.replaceAll("'", "");
    COMMENT = COMMENT.replaceAll("\\\\", "");
    // note that i'm leaving "date_created" out of this insert statement
    System.out.println("INSERT INTO youtube_comments (comment_text, added_ts, video_id) "
            + "VALUES ('" + COMMENT + "', '" + currentTime + "', '" + youtube_video_id + "')");    //H8i5OOs1FZo
    st.executeUpdate("INSERT INTO youtube_comments (comment_text, added_ts, video_id) "
            + "VALUES ('" + COMMENT + "', '" + currentTime + "', '" + youtube_video_id + "')");
}
```

Figure 6.9: Comment ''Extraction'' and Insert Statements Creation

Εισάγωντας ένα twitter hashtag και πατώντας το "Fetch & Save" εκτελείται το παρακάτω κομμάτι κώδικα. Πιο συγκεκριμένα δημιουργούμε τη συνδεση με τη βάση δεδομένων και στη συνέχεια καλούμε τη συνάρτηση getTweetsByURL η οποία «διαβάζει» τα tweets και επιστρέφει τις παραμέτρους για την κλήση της επόμενης σελίδας αποτελεσμάτων.

[1] Δε χρησιμοποιήσαμε το ήδη υπάρχον connection στη βάση και προτιμήσαμε να εισάγουμε με batch insert τα comments για λόγους εξοικονόμησης χρόνου καθώς το runtime είχε μεγαλώσει αρκετά με τα διαδοχικά calls στο Google API.

```
private void button2ActionPerformed(java.awt.event.ActionEvent evt) {//GEN-FIRST:event_button2Actior
    try {

        String url = "jdbc:mysql://62.210.36.88:3306/thesis";
        String username = "thesis";
        String password = "thesis";

        System.out.println("Connecting database...");

        Connection connection = (Connection) DriverManager.getConnection(url, username, password);

        String nextPage = this.getTweetsByURL( "?q="+this.textField2.getText(), connection);
        while (!nextPage.isEmpty()) {
            nextPage = this.getTweetsByURL( nextPage, connection);
        }

        connection.close();
        System.out.println("ALL CLEAR");

    } catch (Exception ex) {
        System.out.println(ex.toString());
    }
}//GEN-LAST:event_button2ActionPerformed
```

Figure 6.10: Call To Action Listener - Twitter Fetch & Save

Στο παρακάτω κομμάτι κώδικα δημιουργούμε τη σύνδεση με το twitter API χρησιμοποιώντας το OAuth1.0 πρωτόκολλο αυθεντικοποίησης κατά το οποίο εισάγουμε τα token, consumer key και secret που μας έχουν δωθεί κατά τη δημιουργία του developer account στο twitter.

```
public String getTweetsByURL(String nextPageSearchTerms, Connection connection) {
  try {

      // Setup the variables necessary to create the OAuth 1.0 signature and make the request
      String httpMethod = "GET";
      String consumerKey = "NF95QVkXsnipH0oj0thL5xM1A";
      String secret = "ePvZ18W2S9FtoE74h4R2utGVLqWJGmHduiQOpwKbaGtNFVP7f4";

      String signatureMethod = "HMAC-SHA1";
      byte[] requestBody = null;
      HttpsURLConnection request = null;
      BufferedReader in = null;
      URL url = new URL("https://api.twitter.com/1.1/search/tweets.json"+nextPageSearchTerms);

      String nonce = getNonce();
      String timestamp = getTimestamp();

      Map<String, String> oauthParams = new LinkedHashMap<String, String>();
      oauthParams.put("oauth_consumer_key", consumerKey);
      oauthParams.put("oauth_signature_method", signatureMethod);
      oauthParams.put("oauth_token", "932577096836280320-LobWBtDkGjhas2zwKOX3KGZ63sQ101W");
      oauthParams.put("oauth_version", "1.0");
      oauthParams.put("oauth_timestamp", timestamp);
      oauthParams.put("oauth_nonce", nonce);

      // Get the OAuth 1.0 Signature
      String signature = getSignature(url.toString(), "", secret);
      System.out.println( signature);
      System.out.println(timestamp);

      String requestString = "OAuth oauth_consumer_key=\"NF95QVkXsnipH0oj0thL5xM1A\",oauth_token=\"932577096836280320-LobWBtDkGjhas2zwKOX3KGZ63sQ101W\","
              + "oauth_signature_method=\"HMAC-SHA1\",oauth_timestamp=\"1512865916\","
              + "oauth_nonce=\""+nonce+"\",oauth_version=\"1.0\",oauth_signature=\""+signature+"\"";
      System.out.println(requestString);
      request = (HttpsURLConnection) url.openConnection();
      request.setRequestMethod(httpMethod);
      request.addRequestProperty("Authorization", requestString);

      InputStreamReader reader = new InputStreamReader(request.getInputStream());
```

Figure 6.11: Setting up Connection with Twitter API using OAuth1.0

Αφού λάβουμε το response από το twitter API, ακολουθούμε την ίδια διαδικασία με το Youtube ώστε να «διαβάσουμε» το JSON και να απομονώσουμε τα διάφορα tweets. Τέλος επιστρέφουμε τα nextPageSearchTermsReturn για να γίνει η κλήση της επόμενης σελίδας.

```
// Get the response stream
String response = in.readLine();

JSONParser parser = new JSONParser();
JSONObject json = (JSONObject) parser.parse(response);

JSONObject metadata = (JSONObject) json.get("search_metadata");
String nextPageSearchTermsReturn = (String) metadata.get("next_results");

JSONArray ja = (JSONArray) json.get("statuses");

for (int i = 0; i < ja.size(); i++) {
    JSONObject jsonObject = (JSONObject) ja.get(i);
    String COMMENT = (String) jsonObject.get("text");
    System.out.println(COMMENT);
    System.out.println("------------------");
}

return nextPageSearchTermsReturn;
```

Figure 6.12: Tweet ''Extraction''

6.3.3 ThesisPrediction

Στο project ThesisPrediction χρησιμοποιούμε τα μοντέλα και τα processes που έχουμε δημιουργήσει στο RapidMiner προκειμένου να περάσουμε κάθε μία εγγραφή της βάσης από το classification model για να πάρουμε το prediction label. Αρχικά δημιουργούμε τη σύνδεση στη βάση δεδομένων προκειμένου να «τραβήξουμε» όλες τις εγγραφές από τη βάση και στη συνέχεια να αποθηκεύσουμε το αποτέλεσμα του classification στην εκάστοτε εγγραφή.

```
String url = "jdbc:mysql://62.210.36.88:3306/thesis";
String username = "thesis";
String password = "thesis";

System.out.println("Connecting database...");

Connection conn = (Connection) DriverManager.getConnection(url, username, password);

  String table_name = "youtube_comments";
String table_name = "twitter_comments";
String query = "SELECT * FROM "+table_name;

// create the java statement
Statement st = conn.createStatement();

// execute the query, and get a java resultset
ResultSet rs = st.executeQuery(query);
```

Figure 6.13: Database Connection and Queries

Με το παρακάτω κομμάτι κώδικα αρχικοποιούμε το RapidMiner χρησιμοποιώντας το process Thesis_Sentiment_Prediction.

```
System.out.println("Step 1: Initiating Rapidminer....");
File z = new File("C:\\Users\\Alex Kotovos\\.RapidMiner\\repositories\\Local Repository\\data\\Sentiment_Prediction\\Thesis_Sentiment_Prediction.rmp");
System.setProperty("rapidminer.home", "C:\\Program Files\\RapidMiner\\RapidMiner Studio");
RapidMiner.setExecutionMode(RapidMiner.ExecutionMode.COMMAND_LINE);

RapidMiner.init();
Process process1 = new Process(z);
```

Figure 6.14: RapidMiner Initiation using Thesis_Sentiment_Prediction process

Για κάθε μία εγγραφή της βάσης δεδομένων ανανεώνουμε το txt αρχείο που χρησιμοποιεί το RapidMiner process και στο οποίο θα χρησιμοποιήσει το classification model που έχουμε δημιουργήσει.

```
// iterate through the java resultset
while (rs.next())
{
    System.out.println("----------------------------------------
    int id = rs.getInt("id");
    String commentText = rs.getString("comment_text");
    System.out.println("ID : "+id+" - Comment "+commentText);

    BufferedWriter bw = null;
    FileWriter fw = null;

    fw = new FileWriter(FILENAME);
    bw = new BufferedWriter(fw);
    bw.write(commentText);
    bw.close();
    fw.close();

    System.out.println("Step 2: Wrote in TXT");
```

Figure 6.15: Updating txt file

Τέλος διαβάζουμε από την εκτέλεση του process το ResultSet και κάνουμε extract το prediction(label) και ανανεώνουμε τη βάση δεδομένων εκτελώντας το Update Query.

```
IOContainer ioResult1=process1.run();

ExampleSet resultSet1=(ExampleSet)ioResult1.getElementAt(0);
Example example1=resultSet1.getExample(0);
Attribute Prediction=example1.getAttributes().get("prediction(label)");
String resultString1=example1.getValueAsString(Prediction);
System.out.println("Step 3: PREDICTION = "+resultString1);

Statement s = conn.createStatement();
s.executeUpdate("UPDATE "+table_name+" SET prediction_label = '"+resultString1+"' WHERE id = "+id);
```

Figure 6.16: Retrieving prediction(label) and Updating Database

6.4 MySQL Workbench

Για να μπορέσουμε να αποθηκεύσουμε τα tweets και τα σχόλια από τα βίντεο του YouTube ώστε να τα χρησιμοποιήσουμε στην διαμόρφωση του Data Set μας χρησιμοποιούμε μία βάση δεδομένων MySQL η οποία βρίσκεται σε απομακρυσμένο εξυπηρετητή (62.210.36.88) και τη διαχειριζόμαστε από το περιβάλλον του MySQL Workbench.

Με τις παρακάτω εντολές δημιουργήσαμε, αρχικά το λογαριασμό του χρήστη, στη συνέχεια τη βάση δεδομένων και τέλος δώσαμε τα απαραίτητα permissions ώστε ο εν λόγω χρήστης να μπορεί να συνδεθεί στο database server.

```
Your MySQL connection id is 7
Server version: 5.7.20-0ubuntu0.16.04.1 (Ubuntu)

Copyright (c) 2000, 2017, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> CREATE USER 'newuser'@'localhost' IDENTIFIED BY 'password';
Query OK, 0 rows affected (0.00 sec)

mysql> CREATE USER 'thesis'@'%' IDENTIFIED BY 'thesis';
Query OK, 0 rows affected (0.00 sec)

mysql> CREATE DATABASE thesis;
Query OK, 1 row affected (0.00 sec)

mysql> GRANT ALL PRIVILEGES ON * . * TO 'newuser'@'localhost';
Query OK, 0 rows affected (0.01 sec)

mysql> GRANT ALL PRIVILEGES ON thesis. * TO 'thesis'@'%';
```

Figure 6.17: User and Database Creation

Στην παρακάτω εικόνα φαίνεται ο τρόπος σύνδεσης στην απομακρυσμένη βάση δεδομενων μέσω του περιβάλλοντος του Mysql Workbench.

Η βάση δεδομένων περιέχει τα κείμενα μικρού μήκους που χρειαζόμαστε καθώς και μερικές σημαντικές πληροφορίες για το καθένα ξεχωριστά.

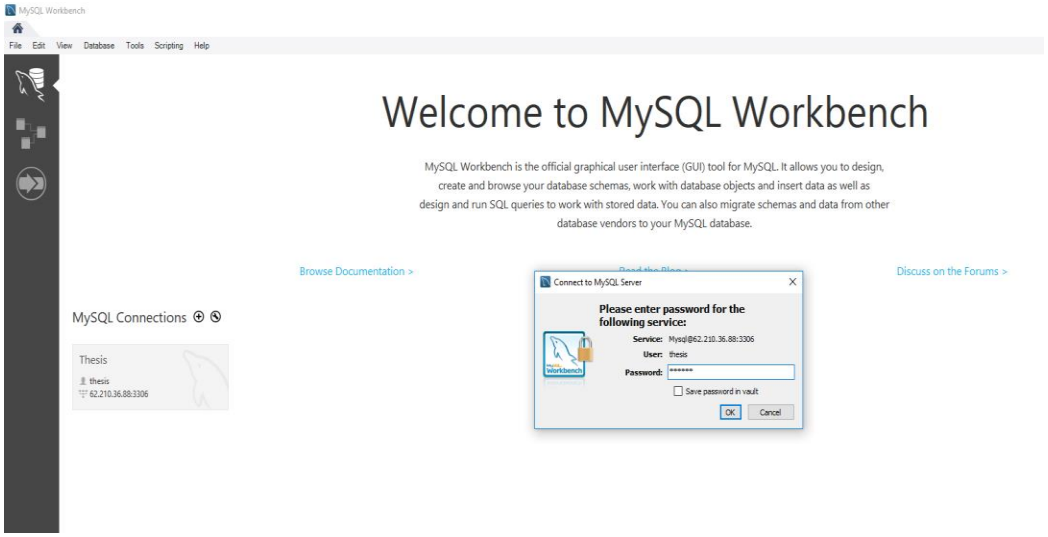


Figure 6.18: Connecting with Database

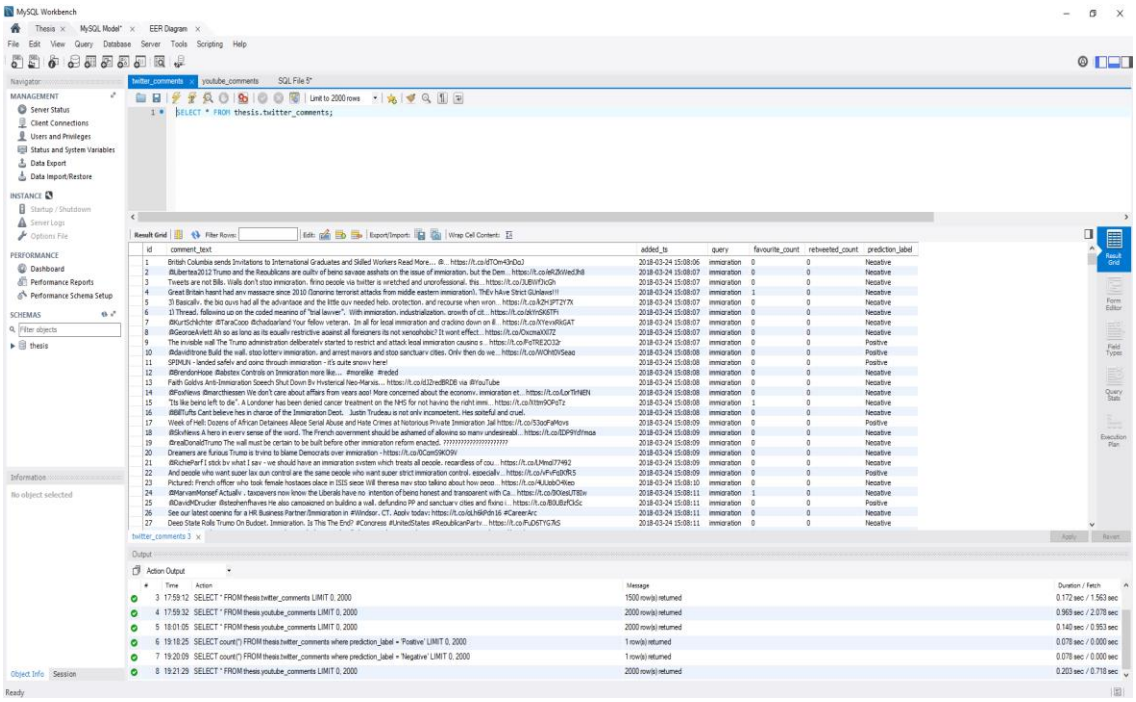


Figure 6.19: Database Example

Οι πίνακες που την απαρτίζουν είναι οι ακόλουθοι:

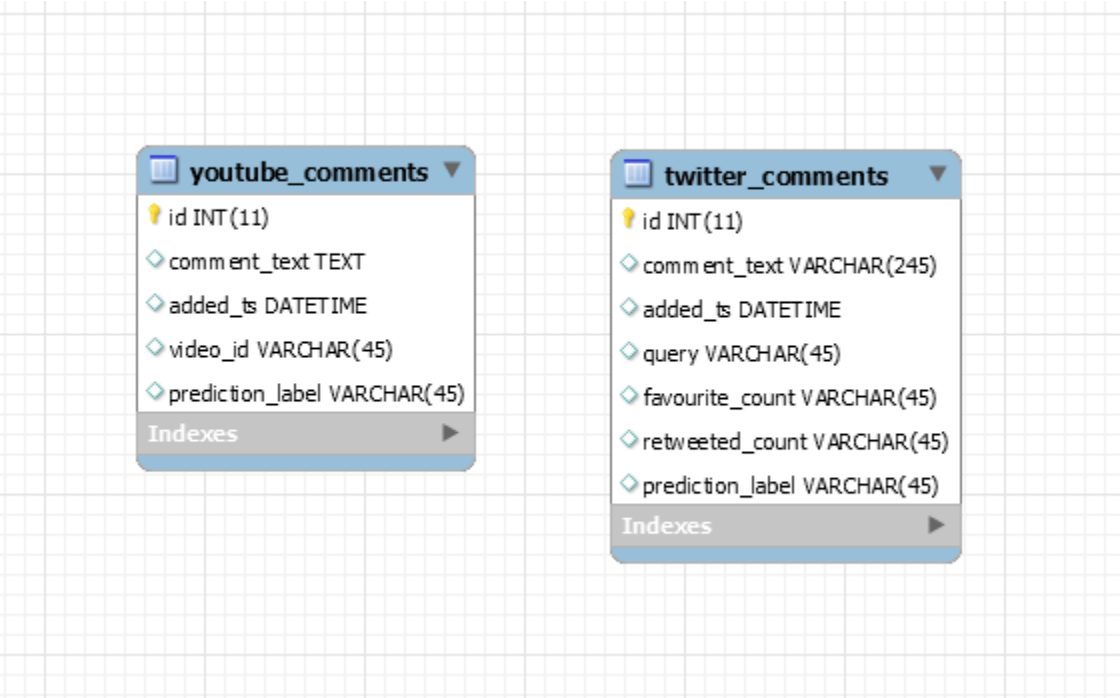


Figure 6.20: Database Tables

6.5 Rapid Miner

Σε αυτό το κεφάλαιο θα αναλύσουμε τον τρόπο με τον οποιο χρησιμοποιώντας το RapidMiner και τις προεκτάσεις του(extensions) δημιουργήσαμε το μοντέλο που χρειάζεται ώστε να δουλέψει η εφαρμογή μας για να κάνει το classification για το unlabeled κείμενο που ανακτούμε μεσω αυτής. Θα αναλύσουμε τον τρόπο με τον οποίο γίνεται το validation του μοντέλου και τα ποσοστά επιτυχίας του.

6.5.1 Φάση 1η – Model Training: Επιλογή Training Set

Τα ήδη rated κείμενα μικρού μήκους που απαρτίζουν το training set μας αποτελούνται από μία συγχώνευση από διάφορα προυπάρχουσα rated sets τα οποία βρήκαμε από τις ιστοσελίδες:https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=text&sort=nameUp&view=table και http://www.cs.cornell.edu/people/pabo/movie-review-data/. Η συγχώνευση αυτή έγινε έτσι ώστε να έχουμε όσο το δυνατόν καλύτερα ποσοστά επιτυχίας τόσο στα θετικά σχόλια όσο και στα αρνητικά. Για τον λόγο αυτό δημιουργήσαμε μία εφαρμογή σε JAVA όπως εξηγήσαμε

αναλυτικά στο κεφάλαιο 6.3.1 η οποία είναι υπεύθυνη για το διαχωρισμό των κειμένων στις διαθέσιμες κλάσεις συναισθημάτων.

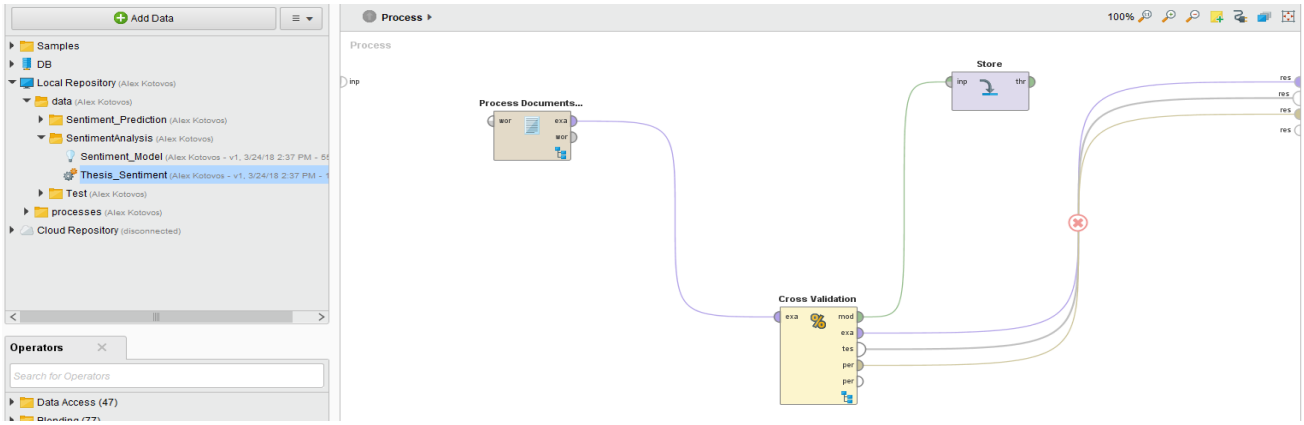


Figure 6.21: Thesis_Sentiment Process

6.5.2 Φάση 1η – Model Training: Δημιουργία και Εκτέλεση Μοντέλου

Η δημιουργία μοντέλου περιλαμβάνει όλες εκείνες τις απαραίτητες δραστηριότητες που πρέπει να γίνουν προκειμένου τα κείμενα που αποτελούν την κάθε κλάση να μετατραπούν σε διάνυσμα, που θα περιέχει μόνο την ωφέλιμη πληροφορία. Στη συνέχεια το διάνυσμα αυτό θα περάσει από ένα validation με κάποιον αλγόριθμο και θα δημιουργηθεί ένα μοντέλο. Το μοντέλο αυτό θα δέχεται ένα άλλο διάνυσμα και θα το κατηγοριοποιεί ανάλογα με τις κλάσεις στις οποίες έχει εκπαιδευτεί. Το διάνυσμα που αναφέρουμε δημιουργείται με το κριτήριο του TF-IDF το οποίο ουσιαστικά απεικονίζει τη βαρύτητα της κάθε λέξης (όπως αυτές δημιουργούνται από την επεξεργασία των κειμένων) σε κάθε ένα κείμενο με βάση τον αριθμό εμφάνισης (συχνότητα) της κάθε λέξης. Η εξαγωγή του ωφέλιμου κειμένου γίνεται με μία σειρά από επεξεργασίες (τελεστές του Rapid Miner) οι οποίοι εντός του τελεστή process documents λαμβάνουν τα αρχεία των δύο κλάσεων και:

- Μετατρέπουν όλους τους χαρακτήρες σε lower case.

Figure 6.22: Transform Cases Operator

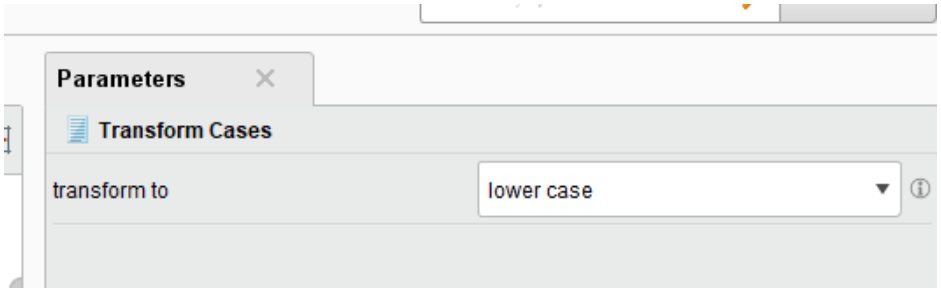


Figure 6.23: Transform Cases Operator parameters

- Αφαιρούν τους ειδικούς χαρακτήρες.

Figure 6.24: Tokenize Operator

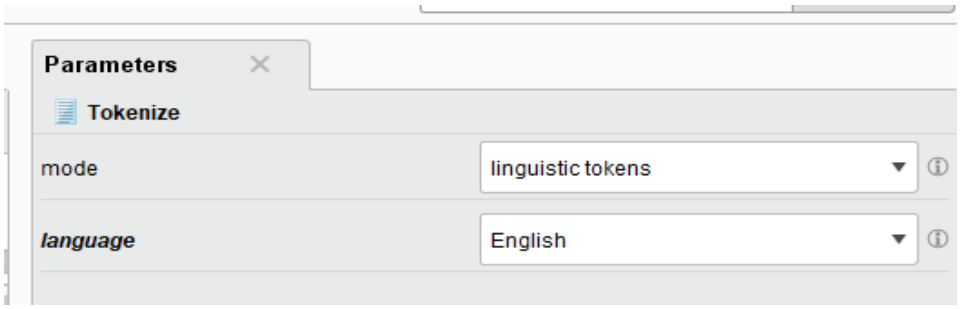


Figure 6.25: Tokenize Operator parameters

- Κόβουν τις λέξεις μέχρι να επιτευχθεί το μικρότερο μήκος(κορμός κάθε λέξης).

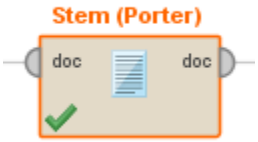


Figure 6.26: Stem (Porter) Operator

- Αφαιρούν τις λέξεις με πολύ μεγάλο (18 χαρακτήρες) ή μικρό (3 χαρακτήρες) μήκος που πιθανότατα δεν θα προσδίδουν κάτι ουσιαστικό στο γενικό νόημα του κειμένου.

Figure 6.27: Filter Tokens (by Length) Operator

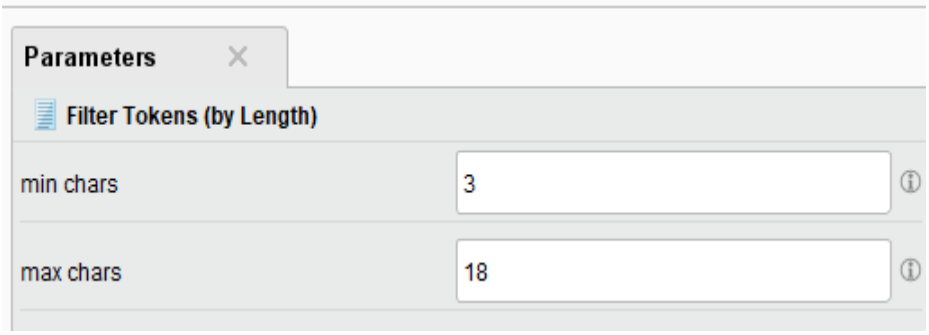


Figure 6.28: Filter Tokens (by Length) Operator parameters

- Χωρίζουν το σύνολο των λέξεων σε συμβολοσειρές 6 χαρακτήρων.

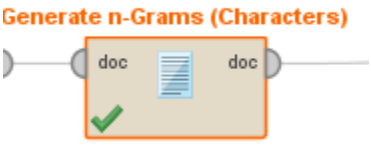


Figure 6.29: Generate n-Grams (Characters) Operator

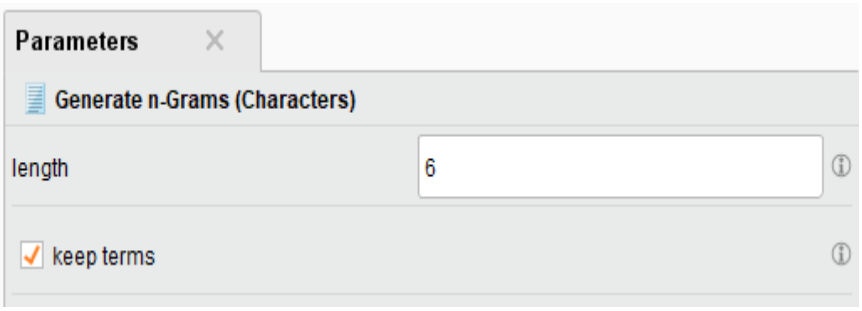


Figure 6.30: Generate n-Grams (Characters) Operator parameters

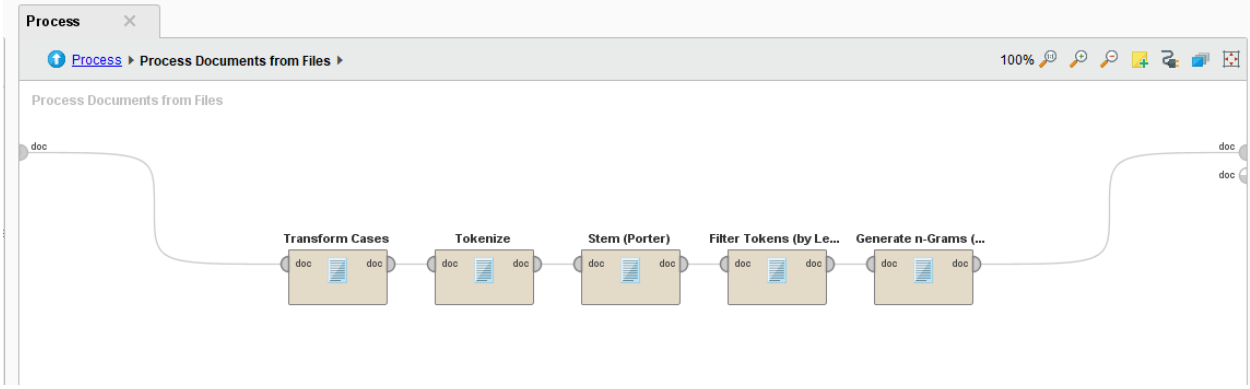


Figure 6.31: Extracting Text concerning Training Data

Αφού λοιπόν γίνει η παραπάνω επεξεργασία και υπολογιστεί το διάνυσμα που περιγράφουμε παραπάνω το training set μας είναι έτοιμο να ελεγχθεί από έναν ten-fold nominal cross validation. Αυτό το validation επιτυγχάνεται χωρίζοντας το training set σε training set και test set (στο test set αποκρύπτεται η κλάση/label) και δοκιμάζεται ο αλγόριθμος που επιλέγουμε στο υποσύνολο του αρχικού training set. Χρησιμοποιούμε stratified sampling για πιο ομοιογενές δείγμα (δηλ. να περιέχονται δείγματα από όλες τις κλάσεις ομοιόμορφα).Ο αλγόριθμος του validation που επιλέξαμε για την δημιουργία του μοντέλου μας είναι ο Decision tree με κριτήριο το information gain και παραμέτρους maximal depth 20 και minimal leaf size 2. Ο αλγόριθμος αυτός μας έβγαλε τα καλύτερα αποτελέσματα από τους αλγορίθμους που χρησιμοποιήσαμε.

Figure 6.32: Ten-Fold Nominal Cross Validation using Decision Tree Algorithm

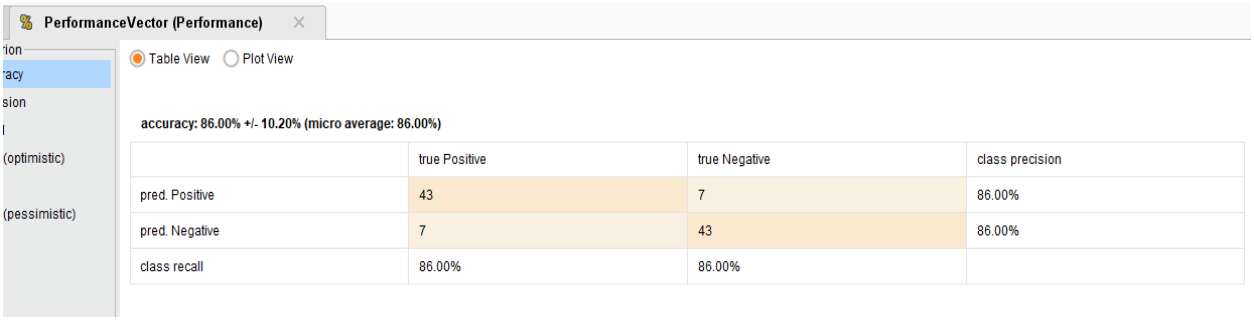


Figure 6.33: Training Results

Παρατηρούμε αρκετά μεγάλο class recall (επιτυχίες προβλεψης) αλλά και μεγάλο class precision (μεγάλη ακρίβεια των προβλέψεων).

Εκτός από τον παραπάνω αλγόριθμο χρησιμοποιήσαμε και:

- K-nn Classification(50 κείμενα ανά κλάση)
- Naive Bayes(50 κείμενα ανά κλάση)
- SVM(50 κείμενα ανά κλάση)

Οι παραπάνω αλγόριθμοι απορρίφθηκαν λόγω υπερβολικά χαμηλών ποσοστών όσον αφορά και την επιτυχία αλλά και την ακρίβεια στις προβλέψεις τους.

6.5.3 Φάση 2η – Sentiment Classification: Επιλογή Data Set

Τα 3000 κείμενα που απαρτίζουν το Data Set μας είναι απόρροια του συνόλου των εφαρμογών που έχουμε δημιουργήσει και προέρχονται απευθείας από το Twitter και το YouTube σε αγγλική γλώσσα. Χωρίζονται σε 2 κλάσεις οι οποίες είναι η θετική και η αρνητική.

6.5.4 Φάση 2η – Sentiment Classification: Δημιουργία και Εκτέλεση Μοντέλου

Η λογική δημιουργίας του πρώτου process ακολουθήθηκε και στη δημιουργία του δεύτερου μοντέλου.

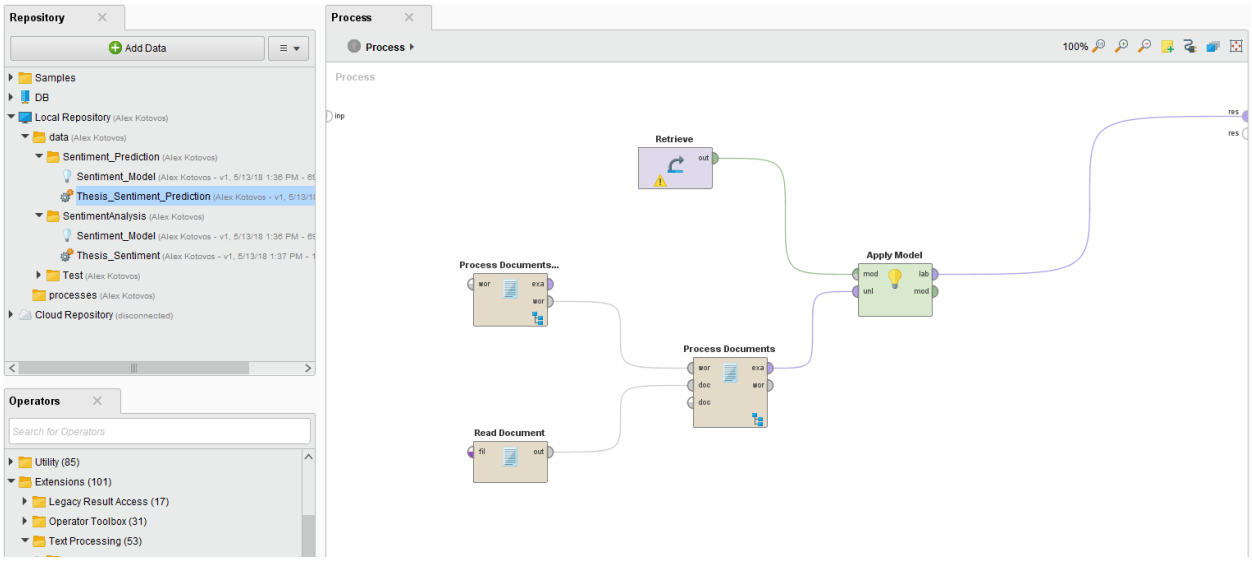


Figure 6.34: Thesis_Sentiment_Prediction Process

Παρατηρούμε πως το process μας αξιολογεί πλέον μόνο του τα κείμενα και μας επιστρέφει το prediction(label) δηλαδή αν είναι θετικά ή αρνητικά με ικανοποιητικό ποσοστό όσον αφορά το confidence του πάνω απο 80% (83,3% για την ακρίβεια).

ExampleSet (Apply Model)

ExampleSet (1 example, 8 special attributes, 51 regular attributes) Filter (1 / 1 examples): all

Row No.	prediction(la...	confidence(...	confidence(...	file_type	metadata_file	metadata_d...	metadata_p...	metadata_si...	...	@user	ablanc	all
1	Negative	0.167	0.833	txt	\use_this.txt	May 15, 2018 ...	C:\Users\Alex...	7	0	0	0	0

Figure 6.35: Classification Results

Κεφάλαιο 7ο: Αποτελέσματα και Συμπεράσματα

7.1 Αποτελέσματα

Το σύστημά μας με την βοήθεια του αλγόριθμου decision tree και κριτήριο το information gain κατόρθωσε να προβλέψει την πολικότητα των σχολίων με ποσοστό όσον αφορά την συνολική ακρίβεια 86%, την επιτυχία πρόβλεψης 86% σε θετικά σχόλια και 86% σε αρνητικά και την ακρίβεια των προβλέψεων 86% σε θετικά σχόλια και 86% σε αρνητικά όπως φαίνεται στην εικόνα 6.33. Επίσης μας επιστρέφει το prediction(label) δηλαδή αν είναι θετικά ή αρνητικά τα σχόλια με ποσοστό όσον αφορά το confidence του 83,3% όπως φαίνεται στην εικόνα 6.35.

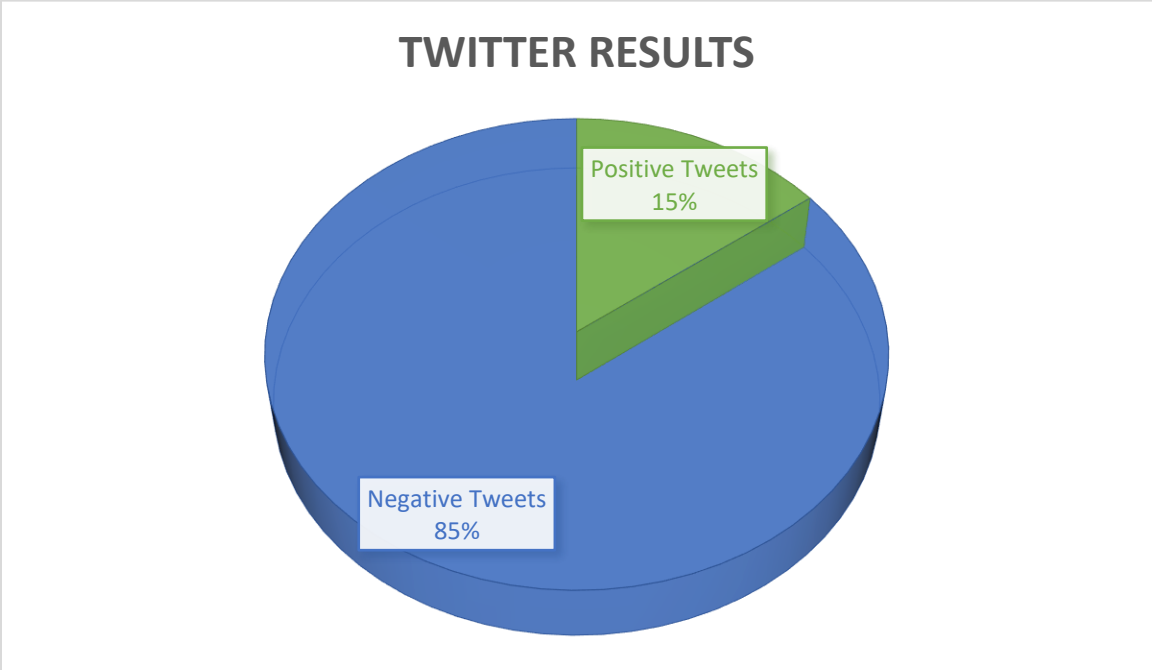
Τα συγκεντρωτικά αποτελέσματα απόδοσης της πλατφόρμας Thesis είναι:

- Για το Twitter από τα 1500 tweets που συλλέξαμε τα 220 είναι θετικά ενώ τα υπόλοιπα 1280 αρνητικά.

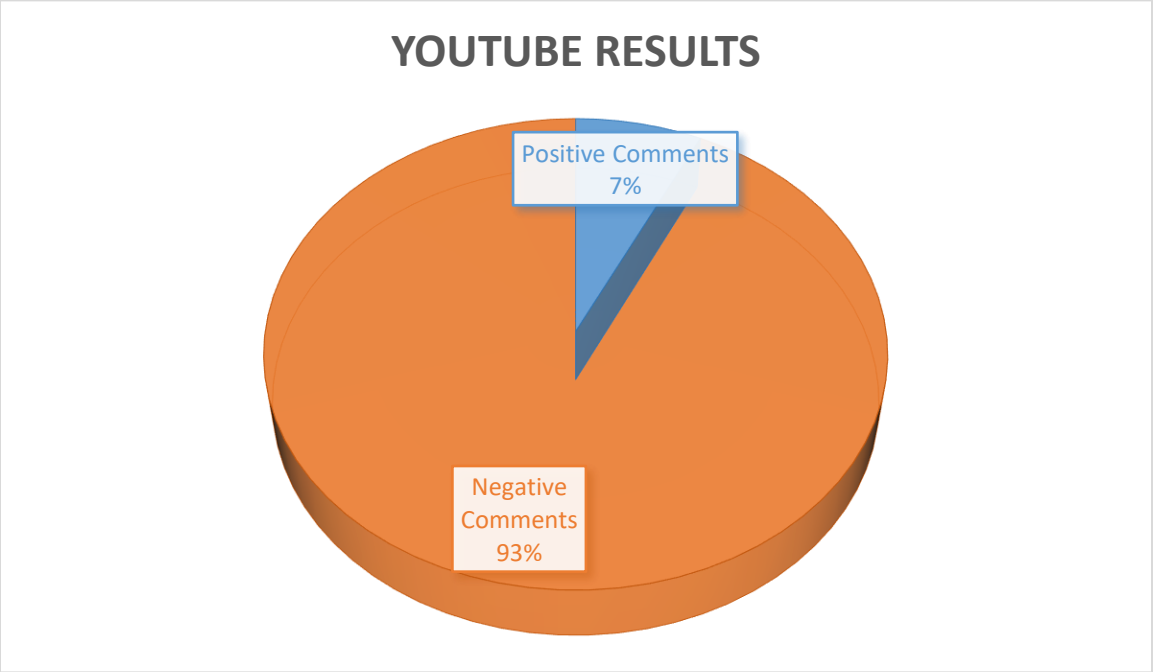
- Για το YouTube από τα 1500 σχόλια που συλλέξαμε τα 108 είναι θετικά ενώ τα υπόλοιπα 1392 είναι αρνητικά.

7.2 Στατιστικά Γραφήματα προβλεψη μοντελου

Από τα 1500 tweets που συλλέξαμε το 15% των απόψεων όσον αφορά το φαινόμενο της μετανάστευσης είναι θετικές ενώ το 85% είναι αρνητικές.



Από τα 1500 σχόλια στο YouTube που συλλέξαμε το 7% των απόψεων όσον αφορά το φαινόμενο της μετανάστευσης είναι θετικές ενώ το 93% είναι αρνητικές.



7.3 Συμπεράσματα

Από την εκπόνηση της παρούσας διπλωματικής εργασίας παρατηρούμε πως εν έτει 2018 και παρά το γεγονός ότι οι περισσότερες χώρες είναι διατεθειμένες να δεχτούν μετανάστες, η σημαντική αύξησή τους συνιστά μια μεγάλη δοκιμασία για τις χώρες υποδοχής και δεν αντιμετωπίζεται με ιδιαίτερα θετική διάθεση από τους γηγενείς, καθώς η πλειοψηφία των σχολίων ήταν αρνητική.

Το σύστημά μας με την βοήθεια του αλγόριθμου decision tree και κριτήριο το information gain κατόρθωσε να προβλέψει την πολικότητα των σχολίων με ικανοποιητική ακρίβεια.

Κατόπιν είναι σημαντικό να αναφερθεί πως παρά το γεγονός πως τα ποσοστά ακρίβειας είναι υψηλά στο RapidMiner, τα training sets μας πάρθηκαν από sites που αφορούσαν θέματα όχι τόσο παρεμφερή με την μετανάστευση (αξιολογήσεις ταινιών,πολιτικά φαινόμενα κτλ) με συνέπεια τα αποτελέσματα να είναι μεν αξιόπιστα, αλλά αν υπήρχε ένα πιο εξειδικευμένο

training set θα παρηγάγαμε αρκετά καλύτερα αποτελέσματα όσον αφορά την φερεγγυότητά τους.

Figure 7.5: Πηγή https://www.variety.com

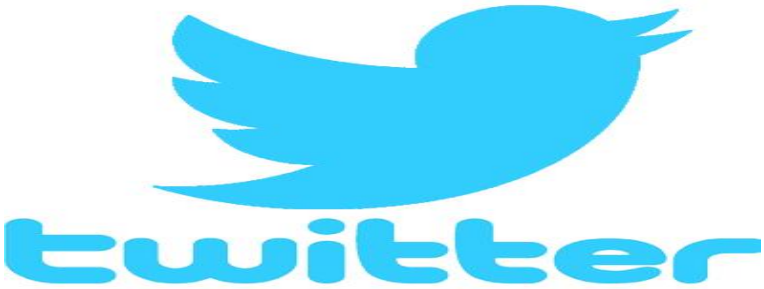


Figure 7.6: Πηγή https://www.mysocialaccounts.com

Βιβλιογραφικές Αναφορές

1. Jeonghee Yi, T. N. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing technique. Proceedings of the Third IEEE International Conference on Data Mining, (pp. 427-434).

2. Theresa Wilson, Janyce Wiebe, Paul Hoffman (2005). Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005. c 2005 Association for Computational Linguistics

3. Manolis Maragoudakis, Euripides Loukis, Ioannis Charalabidis (2011). Proceedings of the 2011 International Conference on Informatics, Cybernetics, and Computer Engineering (ICEE2011) November 19-20, 2001 Melbourne, Australia

4. Ioannis Charalabidis, Manolis Maragoudakis, Euripides Loukis (2015). Proceedings of the 7th IFIP 8.5 International Conference on Electronic Participation – Volume 9249, pages 147-160, New York, August 30 – September 02, Springer-Verlag New York, Inc 2015

5. Brendan O'Connor, R. B. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Association for the Advancement of Artificial Intelligence.

6. Han, B. C. (2013). Lexical Normalization for Social Media Text. ACM Trans. Intell. Syst. Technol. 4, 1.

7. Alexander Pak, P. P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Conference: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010.

8. Svetlana Kiritchenko, X. Z. (2014). Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research 50. (2014). Journal of Artificial Intelligence Research 50.

9. Zubair Asghar (2015). Sentiment Analysis on YouTube: A Brief Survey, MAGNT Research Report (ISSN. 1444-8939), Vol.3 (1). (pp. 1250-1257).

10. Stefan Siesdofer (2010). How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings, WW '10 Proceedings of the 19th International Conference on World Wide Web, pages 891-900, Raleigh, North Carolina, USA.

11. Efthimios Kouloubis (2011) Twitter Sentiment Analysis: The Good The Bad and the OMG!, Conference: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21.

12. Pablo Gamallo (2014) Citius: a Naïve-Bayes Strategy for Sentiment Analysis on English Tweets, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets” Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171-175, Dublin, Ireland, August 23-24.

13. S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay (2008) “Learning document-level semantic properties from free-text annotations,” in Proceedings of the Association for Computational Linguistics (ACL).

14. A. Esuli and F. Sebastiani (2007) “PageRanking WordNet synsets: An application to opinion mining,” in Proceedings of the Association for Computational Linguistics (ACL).

15. M. Bautin, L. Vijayarenu, and S. Skiena (2008) “International sentiment analysis for news and blogs,” in Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

16. N. Godbole, M. Srinivasaiah, and S. Skiena (2007) “Large-scale sentiment analysis for news and blogs,” in Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

17. A. Ghose and P. G. Ipeirotis (2007) “Designing novel review ranking systems: Predicting usefulness and impact of reviews,” in Proceedings of the International Conference on Electronic Commerce (ICEC). (Invited paper).

18. A. Anagnostopoulos, A. Z. Broder, and D. Carmel (2006) “Sampling search-engine results,” World Wide Web, vol. 9, pp. 397–429.

19. G. Carenini, R. Ng, and A. Pauls (2006) “Multi-document summarization of evaluative text,” in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp. 305–312.

20. N. Jindal and B. Liu (2008) “Opinion spam and analysis,” in Proceedings of the Conference on Web Search and Web Data Mining (WSDM), pp. 219–230.

21. K. Wojcik, J. Tuchowski (2014) “Feature based sentiment analysis”, Conference: 3rd International Scientific Conference on Contemporary Issues in Economics, Business and Management EBM

Links

https://www.brandwatch.com/blog/understanding-sentiment-analysis/

http://repfiles.kallipos.gr/html_books/93/04a-main.html#_idTextAnchor075

http://www.enet.gr/?i=news.el.article&id=52505

http://cyprusnews.eu/kiriakostriantafillidis/1009430-2013-03-09-10-59-33.html

http://www.astynomia.gr/index.php?option=ozo_content&perform=view&id=1852

https://searchsqlserver.techtarget.com/definition/data-mining

https://www.preceden.com/timelines/285766-data-mining-history

https://www.ceid.upatras.gr/webpages/courses/cplusplus/dm/4_Bayesian%20Networks_Neural%20Networks.pdf

https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html#%2EWrVpJ_RWToI%2Elinkedin

http://www.cs.cornell.edu/home/llee/data/

http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

https://machinelearningmastery.com/naive-bayes-for-machine-learning/

https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=text&sort=nameUp&view=table

http://www.cs.cornell.edu/people/pabo/movie-review-data/

https://www.solver.com/k-nearest-neighbors-k-nn-classification-intro

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746

https://ieeexplore.ieee.org/document/6726842/?reload=true

http://www.pcstaffing.co.za/26502_mining-and-sentiment.html

https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

https://economictimes.indiatimes.com/definition/data-mining

Ακρωνύμια

API Application Programming Interface

URL Uniform Resource Locator

NLP Natural Language Processing

ICT Information and Communication Technology

OAUTH Open Authorization

SVM Support Vector Machine

TF-IDF Term Frequency–Inverse Document Frequency

Γλωσσάρι

OAuth1.0 Το πρωτόκολλο αυθεντικοποίησης OAuth για ανάθεση πρόσβασης, επιτρέπει την ειδοποίηση ενός resource provider (π.χ το Twitter) ότι ο resource owner (π.χ ο εαυτός μου) χορηγεί άδεια σε κάποιο τρίτο (π.χ ένα Twitter Application) για πρόσβαση στις πληροφορίες του (π.χ τη λίστα των ακολούθων μου).

Batch Insert Με τον όρο batch insert εννοούμε την εκτέλεση πολλών εντολών insert στη βάση σε ένα μόνο SQL script.

SentiWordNet Η λεξική πηγή SentiWordNet βασίζεται, στη περιέχων με ήδη labelled συναισθήματα, λεξική βάση δεδομένων WordNet.