ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μηχανική Μάθηση σε δυο στάδια για την εξαγωγή συναισθήματος από τον Κοινωνικό Ιστό

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΚΑΛΛΙΟΠΗΣ ΓΡΑΜΜΕΝΟΥ

Επιβλέπουσα : Κάτια Κερμανίδου
Καθηγήτρια Ι.Π.

Κέρκυρα, Φεβρουάριος 2019

Η σελίδα αυτή είναι σκόπιμα λευκή.

Περίληψη

Στη σύγχρονη εποχή, στην οποία ζούμε, ο όγκος των δεδομένων που δημιουργείται σε καθημερινή βάση είναι τεράστιος και εξελίσσεται με ραγδαίους ρυθμούς. Η ανάγκη για συλλογή, επεξεργασία και κατανόηση του πλήθους δεδομένων που συγκεντρώνονται καθημερινά είναι επιτακτική, καθώς η εξαγωγή πληροφορίας είναι μονόδρομος. Σκοπός της παρούσας διπλωματικής εργασίας είναι η εξαγωγή πληροφορίας με τη βοήθεια ανάλυσης συναισθήματος, δηλαδή η εξαγωγή συμπερασμάτων σχετικά με τις απόψεις που κοινωνούνται μέσω ενός κειμένου. Συγκεκριμένα, γίνεται έρευνα στον επιστημονικό χώρο της εξαγωγής συναισθήματος από κοινωνικά δίκτυα και ειδικότερα σε απόψεις και σχόλια χρηστών της διαδικτυακής πλατφόρμας TripAdvisor.

Λέξεις Κλειδιά: ανάλυση συναισθήματος, Tripadvisor, Weka, μηχανική μάθηση

Πίνακας περιεχομένων

Κατάλογος εικόνων

Κατάλογος πινάκων

1

Εισαγωγή

Στόχος της εργασίας είναι η έρευνα στο πεδίο της Ανάλυσης Συναισθήματος και η ανάπτυξη ενός μοντέλου, για την εύρεση του συναισθηματικού προσανατολισμού σε απόψεις από το Tripadvisor. Η εξαγωγή συναισθήματος είναι ένα ερευνητικό πεδίο με μεγάλο ενδιαφέρον στην επιστημονική κοινότητα και στη βιομηχανία λόγω της πληθώρας των εφαρμογών. Η εξαγωγή συναισθήματος από μία πλατφόρμα όπως το Tripadvisor είναι ακόμα δυσκολότερο πρόβλημα λόγω της ιδιαιτερότητας του μέσου. Η γλώσσα είναι ένα από τα σημαντικότερα χαρακτηριστικά που κάνουν τον άνθρωπο να ξεχωρίζει από τα υπόλοιπα ζώα. Η δημιουργία επομένως, αλγορίθμων ή μοντέλων που μπορούν να κατανοήσουν φυσική γλώσσα είναι ένα από τα καθοριστικά βήματα για τη δημιουργία ευφυών μηχανών όπου θα μπορούν να εξάγουν συναίσθημα από κείμενο χωρίς την ανθρώπινη παρέμβαση και χωρίς λάθη.

Βασικοί στόχοι της εργασίας είναι:

- Η παρουσίαση σε θεωρητικό πλαίσιο της Ανάλυσης Συναισθήματος και των Τεχνικών Ανάλυσης Συναισθήματος.
- Αναφορά και μελέτη της πλατφόρμας tripadvisor.
- Αναφορά εφαρμογής WEKA
- Υπηρεσίες πλατφόρμας AYLIEN και εφαρμογής της στις εκφράσεις χρηστών στο tripadvisor για την εξαγωγή συναισθήματος.
- Τεχνικές Μηχανικής Μάθησης για την ανάπτυξη μοντέλων δημοφιλή σε προβλήματα επεξεργασίας φυσικής γλώσσας.
- Ανάπτυξη μοντέλου εξαγωγής συναισθήματος από την πλατφόρμα Tripadvisor με χρήση της πλατφόρμας Aylien και εφαρμογή μεταμάθησης. Σύγκριση αποτελεσμάτων σε επίπεδο Ανάκλησης (Recall) και Ακρίβειας (Precision) μετά την μεταμάθηση

του Aylien και των αλγορίθμων Μηχανικής Μάθησης Naive Bayes και Support Vector Machine στο WEKA.
- Αξιολόγηση των αποτελεσμάτων.

1.1 Εισαγωγή στην Ανάλυση Συναισθημάτων - Ανάλυση Συναισθήματος

Ανάλυση Συναισθήματος (Sentiment Analysis) ή αλλιώς Εξόρυξη Γνώμης (Opinion Mining) ενός κειμένου είναι η διαδικασία κατά την οποία ανακτάται η γνώμη του συγγραφέως σχετικά με κάποιο συγκεκριμένο θέμα ή αλλιώς οντότητες. Πρόκειται για ένα ιδιαίτερα σημαντικό ερευνητικό πεδίο καθώς η διαδικασία την οποία ακολουθεί κανείς για να καταλήξει σε μια απόφαση, επηρεάζεται σε όλες τις φάσεις της από τις γνώμες του περιγύρου. Ιδιαίτερα όταν πρόκειται για απόφαση σχετικά με κάποια αγορά, είναι σύνηθες να πραγματοποιείται κάποια έρευνα πριν την ολοκλήρωση της αγοράς δείχνοντας εμπιστοσύνη και αναζητώντας κριτικές και απόψεις άλλων καταναλωτών που έκαναν την ίδια αγορά (Lyons, 1981).

Η ανάλυση συναισθήματος αναφέρεται στην γενική μέθοδο εξαγωγής πολικότητας και υποκειμενικότητας από το κείμενο (δυνητικά και από την ομιλία). Ο όρος Σημασιολογικός Προσανατολισμός (Semantic Orientation ή SO) αναφέρεται στην πολικότητα και την ένταση των λέξεων, των φράσεων ή των κειμένων (Taboada etal, 2011).

Σε γενικές γραμμές η Ανάλυση Συναισθήματος, η οποία είναι γνωστή και ως Εξόρυξη Γνώμης (Opinion Mining), αναφέρεται στη χρήση τεχνικών επεξεργασίας φυσικής γλώσσας (Natural Language Processing-NLP) και υπολογιστικής γλωσσολογίας (computational linguistics) για την ανάλυση κειμένου με σκοπό τον εντοπισμό και την εξαγωγή υποκειμενικής πληροφορίας. Είναι δηλαδή το επιστημονικό πεδίο που εξετάζει τον εντοπισμό των απόψεων και των συναισθημάτων που εκφράζονται σε μορφή κειμένου, όπως για παράδειγμα στις ειδήσεις, στα ιστολόγια (blogs), στις συζητήσεις, στα microblogs και στα κοινωνικά δίκτυα. Ο τομέας αυτός μπορεί να εντοπιστεί με διάφορες ονομασίες όπως για παράδειγμα sentiment analysis (Pang & Lee, 2008), subjectivity (Lyons, 1981), (Langacker, 1985), opinion mining (Pang & Lee, 2008), analysis of stance (Biber & Finegan, 1988), (Conrad & Biber, 2000), appraisal (Martin & White, 2005), point of view (Wiebe, 1994), (Scheibman, 2002), evidentiality (Chafe & Nichols, 1986), (Liu, 2012).

Μεταξύ των δύο κύριων τύπων πληροφορίας κειμένου (textual information) – των γεγονότων και των απόψεων- ένα μεγάλο μέρος των σημερινών μεθόδων επεξεργασίας πληροφοριών, όπως η αναζήτηση στο διαδίκτυο και η εξόρυξη κειμένου (text mining), χρησιμοποιούν τον πρώτο. Η η Ανάλυση Συναισθήματος (Opinion Mining) αναφέρεται στην ευρύτερη έν-

νοια/περιοχή της επεξεργασίας φυσικής γλώσσας, στην υπολογιστική γλωσσολογία και εξόρυξη κειμένου που αφορούν την μελέτη των απόψεων, συναισθημάτων και αισθημάτων που εκφράζονται στο κείμενο. Μια σκέψη, άποψη ή συμπεριφορά βασισμένη σε ένα συναίσθημα και όχι στην λογική, αναφέρεται συχνά ως συναίσθημα (Taboada etal, 2011).

Ο εναλλακτικός ορισμός του Opinion Mining, δηλαδή η Ανάλυση Συναισθήματος, έχει πολύ μεγάλη σημασία σε τομείς στους οποίους οργανισμοί ή ιδιώτες επιθυμούν να μάθουν το γενικό συναίσθημα, την κοινή γνώμη σχετικά με μια οντότητα, όπως ένα προϊόν, μια ταινία, ένα άτομο, και έχει εφαρμογή σε πολλούς τομείς, συμπεριλαμβανομένων της επιστήμης και της τεχνολογίας, στην εκπαίδευση, στην διασκέδαση, στην πολιτική, στο marketing, στη λογιστική, στη νομοθεσία και στην έρευνα και ανάπτυξη (Research and development - R&D). Η προσέγγιση για την κατηγοριοποίηση του κειμένου, εμπλέκει τους ταξινομητές (classifiers) σε κατηγοριοποιημένα (labeled) σώματα κειμένου ή προτάσεων, κάτι που σημαίνει ότι είναι διαδικασία επιβλεπόμενης μάθησης (Taboada etal, 2011). Ένας ακόμα όρος που σχετίζεται με την ανάλυση συναισθήματος είναι η πολικότητα της γνώμης (opinion orientation, sentiment orientation, polarity of opinion, semantic orientation), ο οποίος αναφέρεται στον συναισθηματικό προσανατολισμό ενός κειμένου, μιας πρότασης ή μιας λέξης. Για παράδειγμα η πολικότητα ενός κειμένου μπορεί να είναι «θετική», «αρνητική» ή «ουδέτερη», που σημαίνει ότι επικρατεί ενός είδους συναίσθημα (Indurkhya & Damerau, 2010).

Αν το Web 2.0 αφορούσε απλά την δημοκρατικοποίηση των δημοσιεύσεων στον ιστό, τότε το επόμενο στάδιο του διαδικτύου ίσως πρέπει να βασίζεται στην δημοκρατικοποίηση της εξόρυξης γνώσης από το περιεχόμενο όλου αυτού του όγκου πληροφορίας που δημοσιεύεται (Medhat etal, 2014).

Ένα βήμα πιο κοντά στον στόχο αυτό επιτυγχάνεται μέσω της έρευνας. Πολλές ερευνητικές ομάδες σε πανεπιστήμια σε όλον τον κόσμο επικεντρώνονται στην κατανόηση της δυναμικότητας του συναισθήματος στις e-κοινωνίες μέσων των τεχνικών της ανάλυσης συναισθημάτων (Medhat etal, 2014). Το πρόβλημα είναι ότι οι περισσότεροι αλγόριθμοι χρησιμοποιούν απλούς όρους για να εκφράσουν το συναίσθημα σχετικά με κάποιο προϊόν ή υπηρεσία. Ωστόσο πολιτισμικοί παράγοντες, γλωσσολογικές αποχρώσεις και ιδιαιτερότητες καθιστούν εξαιρετικά δύσκολο τον χαρακτηρισμό ενός τμήματος γραπτού κειμένου ως προς θετικό ή αρνητικό συναίσθημα. Το γεγονός ότι ακόμα και οι άνθρωποι συχνά διαφωνούν για το συναίσθημα των κειμένων, καταδεικνύει το πόσο δύσκολο έργο είναι για έναν υπολογιστή να εξάγει το σωστό συναίσθημα. Κατά συνέπεια, τίθεται όριο επιτυχίας για τη μηχανή ό,τι μπορεί να κατηγοριοποιήσει ο άνθρωπος.

Οι πρώτες προσπάθειες ανάλυσης συναισθήματος παρουσιάστηκαν τη δεκαετία του ‘70 αλλά ουσιαστικά απογειώθηκαν με τη χρήση του Web 2.0 Οι πιο συνηθισμένες προσεγγίσεις της

ανάλυσης συναισθήματος είναι η μηχανική μάθηση (machine learning), το λεξικό (lexicon based method) και οι υβριδικές προσεγγίσεις (hybrid approach).

Εφαρμογές Ανάλυσης Συναισθήματος

Παρά τη σχετικά πρόσφατη άνθηση του τομέα της υπολογιστικής γλωσσολογικής έρευνας, η Ανάλυση Συναισθήματος έχει αναδυθεί ως ένας ιδιαίτερα δραστήριος ερευνητικός τομέας, κυρίως λόγω των πολλών και σημαντικών εφαρμογών της. Η ανάλυση κειμένων στο διαδίκτυο διαδραματίζει σημαντικό ρόλο τόσο στην κατανόηση των κοινωνικών φαινομένων όσο και στην αποτύπωση των κοινωνικών τάσεων. Τα αποτελέσματα της Ανάλυσης Συναισθήματος και Εξόρυξης Γνώσης αποτελούν αντικείμενο μελέτης και έρευνας πολλών επιστημονικών πεδίων όπως η οικονομία, η κοινωνιολογία, η πολιτική, η ψυχολογία.

Η ανάλυση συναισθήματος μπορεί να αναδείξει τη συνολική αντίληψη των χρηστών αναφορικά με ένα θέμα/ζήτημα. Επιπλέον έχει τη δυνατότητα να αναδείξει ομάδες χρηστών ανάμεσα στο συνολικό πληθυσμό, να παρατηρήσει τη διαχρονική πορεία και εξέλιξη των ομάδων, ακόμα και να συστήσει ένα προϊόν ή μία δραστηριότητα σε ένα άτομο. Τόσο τα κοινωνικά δίκτυα και τα προσωπικά ιστολόγια (blogs) όσο και οι ομάδες συζητήσεων (discussion forums), εκτός από «χώρους» ανταλλαγής και παράθεσης ιδεών και απόψεων για τους χρήστες συγκροτούν μία πλούσια πηγή δεδομένων για την ανάλυση συναισθήματος και την εξόρυξη γνώσης (Aggarwal & Zhai, 2012).

Αντίστοιχα σημαντική είναι η αξιοποίηση των κριτικών/ αξιολογήσεων των χρηστών για προϊόντα και υπηρεσίες. Η επεξεργασία και η ανάλυσή τους αποκρυσταλλώνει τις απόψεις των χρηστών για προϊόντα και υπηρεσίες, οδηγώντας σταδιακά στην αντικατάσταση των παραδοσιακών δημοσκοπήσεων και ερευνών μέσω ερωτηματολογίων. Η εξόρυξη γνώσης των καταναλωτών/χρηστών αποκαλύπτει τις ευκαιρίες για την προώθηση νέων προϊόντων/υπηρεσιών και υπογραμμίζει τα περιθώρια βελτίωσης των προϊόντων και υπηρεσιών που ήδη κυκλοφορούν στην αγορά. Αποσαφηνίζοντας τις τάσεις της αγοράς και τις προτιμήσεις των καταναλωτών, η εξόρυξη γνώσης προσφέρει στις επιχειρήσεις που αξιοποιούν αυτές τις πληροφορίες σημαντικό ανταγωνιστικό πλεονέκτημα.

1.1.1 Το πρόβλημα της Ανάλυσης Συναισθήματος

Από τις πρώτες μελέτες που έγιναν από ερευνητές στον τομέα της Ανάλυσης Συναισθήματος (Aggarwal & Zhai, 2012) είχε γίνει αντιληπτό ότι η κατηγοριοποίηση συναισθήματος (sentiment classification) διαφέρει από την απλή κατηγοριοποίηση κειμένου.

«Η κατηγοριοποίηση κειμένου – γνωστή και ως ταξινόμηση κειμένου ή ανίχνευση θέματος – αναφέρεται στην αντιστοίχιση κειμένων φυσικής γλώσσας σε θεματικές κατηγορίες ή κλάσεις οι οποίες ανήκουν σε ένα προκαθορισμένο σύνολο» (Aghaei etal, 2012).

Οι κατηγορίες, στην κατηγοριοποίηση συναισθήματος, καθορίζονται με βάση τα θέματα στόχους του εκάστοτε προβλήματος. Το πλήθος των κατηγοριών ποικίλει. Μπορεί να αποτελείται από ένα μικρό σύνολο μιας ή δύο κατηγοριών αλλά μπορεί και να είναι ένα μεγάλο σύνολο. Για παράδειγμα οι κατηγορίες που υπάρχουν για την ταξινόμηση ενός άρθρου ενημερωτικής ιστοσελίδας με βάση τη θεματολογία του. Οι κατηγορίες που υπάρχουν είναι Αθλητικά, Οικονομικά, Επικαιρότητα κτλ. Ένα άρθρο μπορεί να ανήκει σε πάνω από μία κατηγορίες.

Αντίθετα, η Ανάλυση Συναισθήματος αναφέρεται σε ένα μικρό σύνολο κατηγοριών για παράδειγμα: θετικό, αρνητικό, ουδέτερο, 1 αστέρι… 5 αστέρια. Οι κατηγορίες αυτές είναι ανεξάρτητες της θεματολογίας του προβλήματος, αλλά και αμοιβαία αποκλειόμενες.

Το πρόβλημα που προσπαθεί να επιλύσει η Ανάλυση Συναισθήματος είναι ένα από τα πιο απλά προβλήματα με τα οποία ασχολείται η Επεξεργασία Φυσικής Γλώσσας (Ahmed, 2004). Ένας υπολογιστής δεν είναι απαραίτητο να καταλαβαίνει πλήρως την σημασιολογία της κάθε λέξης αλλά θα πρέπει να εντοπίζει τη συνολική στάση του ανθρώπου που γράφει ένα σχόλιο ή μία κριτική. Οι παραπάνω απαιτήσεις απλοποιούν σε μεγάλο βαθμό τις απαιτήσεις κατανόησης της Φυσικής Γλώσσας από τους υπολογιστές αλλά δεν παύει να είναι μεγάλο πρόβλημα η ανίχνευση της πολικότητας στο κείμενο, ακόμη και για τον άνθρωπο (Biber & Finegan, 1988).

1.1.2 Ανασκόπηση τεχνικών Ανάλυσης Συναισθήματος

Παρά το γεγονός πως η ανάλυση συναισθήματος και η εξόρυξη γνώμης έχει παρουσιάσει ένα τεράστιο ενδιαφέρον για έρευνα, υπήρξε ένα υπόγειο ρεύμα ενδιαφέροντος για αρκετό διάστημα. Θα μπορούσε κανείς να μετρήσει πρώιμα έργα σχετικά με το αντικείμενο ως πρόδρομους της εποχής (Carbonell, 1979), (Wilks & Bien, 1983). Αργότερα η έρευνα επικεντρώθηκε κυρίως στην ερμηνεία της μεταφοράς, της αφήγησης, της άποψης και συναφείς τομείς ενδεικτικά (Hearst, 1992), (Huettner & Subasic, 2000), (Kantrowitz, 2003). Το 2001 σηματοδοτείται η έναρξη της ευρείας ευαισθητοποίησης των ερευνητών γύρω από την ανάλυση συναισθήματος ενδεικτικά (Turney, 2002), (Wiebe etal, 2003), (Yu & Hatzivassiloglou, 2003).

Η ανάλυση συναισθήματος και η εξόρυξη γνώμης έχουν σημαντικό ρόλο ως τεχνολογίες γενικής εφαρμογής και για άλλα συστήματα. Μια χρησιμότητα είναι σε συστήματα σύστασης (recommendation systems) (Tatemura, 2000), (Terveen etal, 1997) όπου όταν ένα σύστημα λαμβάνει πολλά αρνητικά σχόλια δεν είναι εφικτό να προταθεί. Επίσης χρησιμοποιείται στην ανίχνευση των απειλών στο ηλεκτρονικό ταχυδρομείο, όταν κατηγοριοποιείται η επικοινωνία

ανάλογα με το θέμα. Τέλος, η ανάλυση συναισθήματος χρησιμοποιείται σε ηλεκτρονικές σελίδες όπου εμφανίζονται διαφημίσεις σε πλαϊνές μπάρες. Σε αυτές τις περιπτώσεις είναι χρήσιμη για την ανίχνευση ιστοσελίδων που περιέχουν ευαίσθητο περιεχόμενο ακατάλληλο για τοποθέτηση.

Οι Ganu, Elhadad, Marian (2009) πραγματοποίησαν μια μελέτη σχετικά με τη καλύτερη αξιολόγηση των προϊόντων χρησιμοποιώντας κείμενα καταναλωτών αντίθετα από αξιολογήσεις με αστέρια. Στόχος τους είναι να αναγνωρίσουν την πληροφορία που υπάρχει στα ελεύθερα και μη δομημένα σχόλια πελατών με σκοπό να βελτιώσουν την ακρίβεια των αποτελεσμάτων. Συγκεκριμένα η έρευνα εστιάζεται στη βελτίωση της προτροπής κάποιου να επισκεφθεί κάποιο εστιατόριο με μεγαλύτερη ακρίβεια. Οι ερευνητές ασχολούνται με την κατηγοριοποίηση και την ανάλυση συναισθήματος στο επίπεδο της πρότασης καθώς τα σχόλια που υπάρχουν στο διαδίκτυο μεταφέρουν πολλές φορές λεπτομερείς πληροφορίες σε λίγες μόνο λέξεις. Το project των ερευνητών ονομάζεται URSA (User Review Structure Analysis). Με αυτό θέλουν να δημιουργήσουν καλύτερα εργαλεία για την αναγνώριση και ανάλυση των σχολίων των χρηστών.

Τα δεδομένα προέρχονται από το Citysearch New York και αφορούν 50.000 κριτικές σε εστιατόρια. Πραγματοποιείται μία ανάλυση στα δεδομένα για να γίνει η αναγνώριση των κατηγοριών και των συναισθημάτων, καθώς και ο σχολιασμός ενός ταξινομημένου dataset. Προτείνουν νέες μετρικές πρόβλεψης όπου λαμβάνουν υπόψη τους τις κριτικές που είναι γραπτές φράσεις. Γι' αυτό μετέφρασαν το κατηγοριοποιημένο dataset σε κείμενο το οποίο έχει επισημειωθεί. Αυτό το κείμενο χρησιμοποιείται για να γίνονται προβλέψεις. Τα αποτελέσματα της έρευνας αυτής δείχνουν ότι υπάρχουν καλύτερες κριτικές και προβλέψεις των εστιατορίων όταν οι σχολιασμοί των χρηστών είναι κείμενο παρά αν είναι αριθμητικές κριτικές (π.χ 1 Αστέρι…. 5 Αστέρια).

Οι Παυλόπουλος και Ανδρουτσόπουλος (2014) στην έρευνά τους το 2014 στηρίχθηκαν στην ανάλυση των σχολίων (reviews) των χρηστών με βάση το προϊόν προς αξιολόγηση. Η τεχνική με την οποία ασχολούνται ονομάζετε ABSA (Aspect-Based Sentiment Analysis). Η τεχνική ABSA χρησιμοποιείται για κάθε ξεχωριστό αντικείμενο (πχ. Μπαταρία) μιας οντότητας (πχ. Κινητό τηλέφωνο). Διαφορετικές λέξεις ή φράσεις μπορεί να χρησιμοποιηθούν για να αναφερθούν στο ίδιο αντικείμενο. Οι ερευνητές χωρίζουν την έρευνα σε δύο φάσεις. Στην πρώτη φάση όπου η σημασιολογική ομοιότητα των αντικειμένων είναι ορατή και στη δεύτερη φάση που τα αντικείμενα είναι χωρισμένα.

Σε ένα ABSA σύστημα δίνονται ένα πλήθος αξιολογήσεων για ένα προϊόν ή υπηρεσία (reviews από υπολογιστή). Το σύστημα προσπαθεί να αναγνωρίσει το κυριότερο χαρακτηριστικό του αντικειμένου (πχ motherboard) και το μέσο όρο αξιολόγησης (πχ. 1 έως 5 αστέρια).

Στην έρευνα αυτή οι ερευνητές χρησιμοποιούν 3.710 σχόλια χρηστών στην Αγγλική γλώσσα από το (Ganu etal, 2009). Υπάρχουν επίσης 3.085 σχόλια για φορητούς υπολογιστές από 394 πελάτες. Σαν αποτέλεσμα των ερευνών καταλήγουν στο ότι ένας καλός διαχωρισμός των χαρακτηριστικών ενός προϊόντος ή μιας υπηρεσίας με τη μέθοδο ABSA μπορεί να παρουσιάσει σημαντικά αποτελέσματα.

Οι Minqing Hu και Bing Liu (2004) πραγματοποιούν την ερευνά τους πάνω στην εξόρυξη γνώσης και συγκέντρωση από διάφορες κριτικές πελατών. Καθώς το ηλεκτρονικό εμπόριο ολοένα και μεγαλώνει, οι κριτικές, θετικές και αρνητικές, που λαμβάνει κάποιο προϊόν αυξάνονται με μεγάλο ρυθμό. Για παράδειγμα σε ένα δημοφιλές προϊόν οι κριτικές των πελατών είναι από εκατοντάδες μέχρι και χιλιάδες. Έτσι καταλαβαίνουμε πως υπάρχει μεγάλη δυσκολία σε κάποιον που θέλει να αγοράσει κάποιο προϊόν να διαβάσει όλες τις κριτικές και να αποφασίσει. Ακριβώς το ίδιο πρόβλημα αλλά σε μεγαλύτερο βαθμό υπάρχει και για τις ίδιες τις εταιρίες που εμπορεύονται κάποιο προϊόν το οποίο πωλείται σε πάνω από έναν ιστότοπο. Σε αυτή την έρευνα έχουν σαν κύριο σκοπό την εξόρυξη γνώσης και τη συγκέντρωση των κριτικών διάφορων πελατών για ένα προϊόν. Πιο συγκεκριμένα συγκεντρώνουν τα χαρακτηριστικά των προϊόντων που αξιολογήθηκαν από τους αγοραστές και στη συνέχεια αναγνωρίζουν με βάση το συναίσθημα σε κάθε μία κριτική αν είναι θετική ή αρνητική. Τέλος συγκεντρώνουν όλα τα αποτελέσματα μαζί.

Οι Ποντίκη, Γαλάνης, Παυλόπουλος, Παπαγεωργίου, Ανδρουτσόπουλος και Manandhar (2014) πραγματοποίησαν το 2014 μια έρευνα πάνω στο ABSA (Aspect Based Sentiment Analysis) όπου την χωρίζουν σε τέσσερα επίπεδα. Στα δύο πρώτα επίπεδα υπάρχουν τα δεδομένα της έρευνας για δύο κατηγορίες προϊόντων (εστιατόρια και φορητοί υπολογιστές), ενώ στο τρίτο και τέταρτο επίπεδο της έρευνας υπάρχουν τα δεδομένα μόνο για τα εστιατόρια. Το πρώτο επίπεδο της έρευνας αφορά την εξαγωγή κατάταξης (Aspect term extraction). Μέσα από ένα σύνολο δεδομένων από αξιολογήσεις χρηστών για τα εστιατόρια και τους φορητούς υπολογιστές βρίσκουν όλες τις λέξεις κλειδιά (aspects) που χαρακτηρίζουν το προϊόν ή την υπηρεσία. Αναγνωρίζουν σε αυτό το επίπεδο όλα τα aspects ακόμη και αυτά που έχουν ουδέτερη σημασία (ούτε θετικά, ούτε αρνητικά). Στο δεύτερο επίπεδο της έρευνας ασχολούνται με την πολικότητα του όρου (aspect term polarity). Χρησιμοποιούν τα aspects του πρώτου επιπέδου και τα κατηγοριοποιούν με βάση την πόλωσή τους (θετικά, αρνητικά, ουδέτερα). Στο τρίτο επίπεδο της έρευνάς τους ασχολούνται με την ανίχνευση κατηγορίας όρου (aspect category detection). Υπάρχει σε αυτό το επίπεδο ένα προκαθορισμένο σύνολο aspect, όπως είναι η τιμή και το φαγητό, και ένα σύνολο από σχόλια χρηστών. Να σημειώσουμε πως τα σχόλια χρηστών δεν έχουν κατηγοριοποιηθεί με βάση τη πολικότητά τους. Τέλος, στο τέταρτο μέρος της έρευνας οι ερευνητές μελετούν την πολικότητα της κατηγορίας εμφάνισης (aspect category polarity). Εδώ οι κατηγορίες για κάθε σχόλιο χρήστη παρέχονται. Σκοπός

του επιπέδου αυτού είναι να καθορίσουν τη πολικότητα των απόψεων (θετική, αρνητική, ουδέτερη) για κάθε κατηγορία aspect που συζητείται σε κάθε σχόλιο. Τα δεδομένα που χρησιμοποίησαν για τα σχόλια των χρηστών στην κατηγορία εστιατόρια ήταν από το (Ganu etal, 2009) και αφορούν 3.041 σχόλια, ενώ στη κατηγορία των φορητών υπολογιστών χρησιμοποίησαν 3.845 σχόλια.

Οι Suresh, Roodi και Ειρηνάκη (Suresh etal, 2014) το 2014 στην έρευνά τους ασχολούνται με ένα σύστημα το οποίο εξάγει aspects με ακριβείς χαρακτηρισμούς από σχόλια χρηστών και προτείνει σχόλια σε άλλους χρήστες με βάση τα πρότυπα αξιολόγησης. Οι ερευνητές προτείνουν ένα σύστημα που παράγει αυτόματα εξατομικευμένες προτάσεις σχολίων στους χρήστες χρησιμοποιώντας δύο διαφορετικές προσεγγίσεις. Αρχικά υπάρχουν τα παραδοσιακά συστήματα φιλτραρίσματος όπου το σύστημα παράγει τη βαθμολόγηση των προφίλ του κάθε χρήστη και προσαρμόζει τη λίστα με τις αξιολογήσεις του καθενός χρήστη. Η δεύτερη προσέγγιση που απασχολεί τους ερευνητές έχει να κάνει με την εξόρυξη γνώμης ώστε να προσδιορίσει τα σημαντικά χαρακτηριστικά σε κάθε κριτική. Το σύστημα που αναλύουν στην συγκεκριμένη έρευνα έχει σύνολο δεδομένων από το Yelp (πλατφόρμα reviews όπως το Tripadvisor, https://www.yelp.ie).

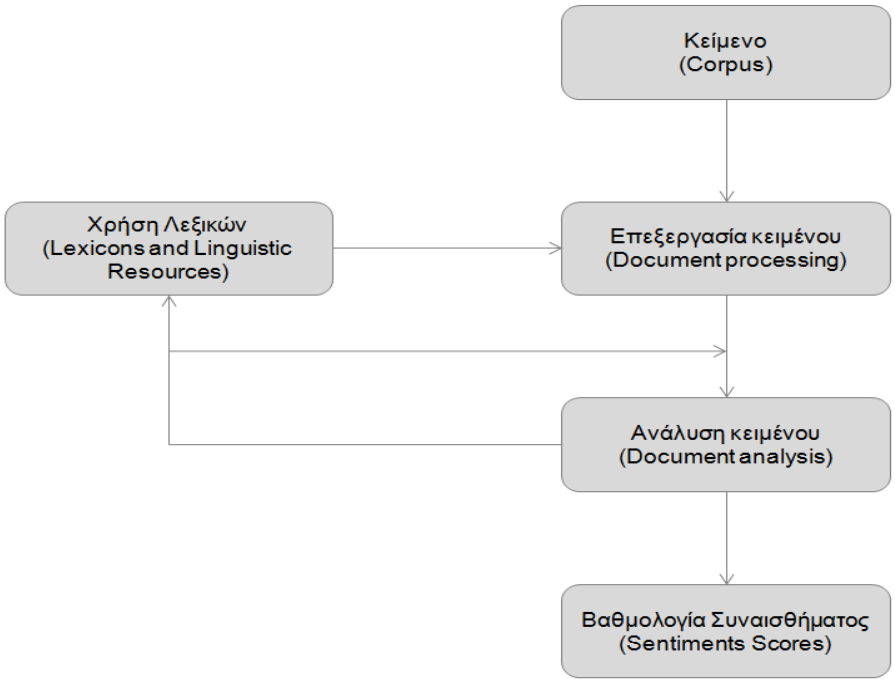
Lexicon- Based Method

Η προσέγγιση βασισμένη σε λεξικό (lexicon –based method) υποθέτει ότι ο συναισθηματικός προσανατολισμός ενός κειμένου μπορεί να συναχθεί από το συναισθηματικό προσανατολισμό των επιμέρους λέξεων και φράσεων του. Σε αντίθεση με τις βασισμένες σε μηχανική μάθηση μεθόδους, η προσέγγιση βάση λεξικού δεν απαιτεί την εκπαίδευση ενός ταξινομητή. Αντιθέτως, χρησιμοποιεί λεξικά συναισθήματος για να αποδώσει το συναίσθημα των συναισθηματικά φορτισμένων λέξεων στο κείμενο. Η απόδοση μιας μεθόδου Ανάλυσης Συναισθήματος βασισμένης σε λεξικό συνήθως καθορίζεται από τον τύπο του λεξικού συναισθήματος και από τον αλγόριθμο ανίχνευσης συναισθήματος (sentiment detection algorithms), δηλαδή από τον αλγόριθμο που χρησιμοποιείται για τον εντοπισμό των συναισθηματικά φορτισμένων λέξεων του κειμένου και τον υπολογισμό του συνολικού συναισθήματος (Taboada etal, 2011).

Οι βασισμένες σε λεξικό μέθοδοι έχουν το πλεονέκτημα της απλότητας και της ταχύτητας, στοιχεία που αποτελούν απαραίτητη προϋπόθεση, ιδίως όταν απαιτείται η ανάλυση τεράστιων όγκων δεδομένων. Οι συμβατικές μέθοδοι αδυνατούν να ανιχνεύσουν σύνθετους τύπους, όπως για παράδειγμα την άρνηση ή το ενισχυμένο συναίσθημα. Επιπλέον, εξαιτίας του γεγονότος ότι στα λεξικά συναισθήματος οι λέξεις επισημειώνονται με μια σταθερή πολικότητα δεν λαμβάνεται υπόψιν το ευρύτερο σημασιολογικό πλαίσιο (context).

Τα παραδοσιακά λεξικά, όπως τα MPQA και SentiWordnet για την αγγλική γλώσσα είναι σχεδιασμένα για να χειρίζονται επίσημα και καλογραμμένα κείμενα. Στις διαδικτυακές πλατφόρμες όπως το TripAdvisor ο ιδιαίτερος και ανεπίσημος τρόπος γραφής οδηγεί στον εντοπισμό ανάκτησης ελάχιστων μη-συμβατικών ή ασυνήθιστων λέξεων, αφού αυτές σπάνια είναι καταχωρημένες στα παραδοσιακά λεξικά. Κατά συνέπεια οι μέθοδοι βασισμένες σε λεξικό μπορούν να εφαρμοστούν άμεσα σε πλατφόρμες κοινωνικής δικτύωσης (TripAdvisor) χωρίς να χάνεται πολύτιμος χρόνος για την εκπαίδευση ταξινομητών με το βασικό περιορισμό ότι η απόδοση και αποτελεσματικότητα όλου του συστήματος συνδέεται αναπόσπαστα με τη λεξική πηγή στην οποία βασίζεται. Μια μέθοδος βασισμένη σε λεξικό είναι τόσο καλή όσο και το λεξικό που χρησιμοποιεί. Τέλος η απόδοση τέτοιων συστημάτων σε επίπεδο ακρίβειας και χρονικής πολυπλοκότητας επιδεινώνεται δραστικά με την εκθετική αύξηση του μεγέθους του λεξικού.

Θα πρέπει να επισημάνουμε ότι σε οποιαδήποτε εφαρμογή ανάλυσης κειμένου υπάρχει η δυνατότητα επιλογής στατιστικών ή συντακτικών τεχνικών. Οι συντακτικές τεχνικές (syntactic techniques) μπορούν να οδηγήσουν σε καλύτερη ακρίβεια γιατί κάνουν χρήση των συντακτικών κανόνων μιας γλώσσας με σκοπό να ανιχνεύσουν τα ρήματα, τα επίθετα και τα ουσιαστικά. Δυστυχώς, αυτού του είδους οι τεχνικές είναι άρρηκτα συνδεδεμένες με τη γλώσσα του κειμένου και συνεπώς δε μπορούν να εφαρμοστούν σε άλλες γλώσσες. Από την άλλη οι στατιστικές τεχνικές (statistical techniques) έχουν πιθανό υπόβαθρο και εστιάζουν στις σχέσεις μεταξύ των λέξεων και των κατηγοριών. Οι τεχνικές αυτές έχουν δύο σημαντικά πλεονεκτήματα σε σχέση με τις συντακτικές. Μπορούμε να τις χρησιμοποιήσουμε σε διάφορες γλώσσες τροποποιώντας τες ελάχιστα ή και καθόλου και μπορούμε να χρησιμοποιήσουμε μετάφραση μηχανής (machine translation) του αρχικού σετ δεδομένων και να έχουμε και πάλι ικανοποιητικά αποτελέσματα. Οι παραπάνω μέθοδοι μπορούν να συνδυαστούν τόσο σε τεχνικές βασισμένες σε λεξικό όσο και με τεχνικές μηχανικής μάθησης.



Εικόνα 1 Διαδικασία Ανάλυσης Συναισθήματος με χρήση λεξικού (Χριστοπούλου, 2017)

Δυσκολίες /Προκλήσεις

Η μέθοδος αυτή προσεγγίζει το συναισθηματικό χαρακτηρισμό του κειμένου από τον χαρακτηρισμό των επιμέρους λέξεων ή φράσεων. Αυτές, οι λέξεις ή φράσεις, έχουν επισημειωθεί ως προς το συναισθηματικό τους περιεχόμενο και βρίσκονται κατοχυρωμένες σε ένα λεξικό, το οποίο χαρακτηρίζεται συναισθηματικό λεξικό (sentiment lexicon) σε επίπεδο λέξεων (word-level) ή έννοιας (concept-level). Στα λεξικά αυτά οι λέξεις ή φράσεις συνοδεύονται από αντίστοιχες βαθμολογίες οι οποίες εκφράζουν κατά πόσο αυτές ταιριάζουν με μία συγκεκριμένη κατηγορία συναισθήματος: συνήθως οι βασικές αυτές κατηγορίες είναι οι δύο (θετική και αρνητική) και οι βαθμολογίες έχουν το αντίστοιχο πρόσημο, με την απόλυτη τιμή να εκφράζει την κλίμακα (ή «βεβαιότητα») της κατηγορίας.

Η μέθοδος αυτή παρότι απλή παρουσιάζει κάποιες αδυναμίες που περιορίζουν την απόδοσή της στην ταξινόμηση του κειμένου (Taboada etal, 2011), (Kennedy & Inkpen, 2006), (Polanyi & Zaenen, 2006). Βασικοί περιορισμοί:

Άρνηση: Στα αγγλικά ανιχνεύεται με τις λέξεις not, never, nothing, nobody κτλ. και οι οποίες αντιστρέφουν την πολικότητα (polarity) των λέξεων που ακολουθούν.

Μετατόπιση Έντασης /Σθένους: Τέτοιες λέξεις είναι very, truly, really, slightly, more, less κ.α. Αυτές οι λέξεις αυξάνουν ή μειώνουν την ένταση της επόμενης λέξης.

Το πρόβλημα αυτής της προσέγγισης είναι ότι δεν λαμβάνει υπόψη τις διαφορές μεταξύ των λέξεων μετατόπισης έντασης.

Σειρά Λέξεων: Η μέθοδος αυτή αγνοεί τη σειρά των λέξεων που εμφανίζονται στο κείμενο. Η σειρά των λέξεων μπορεί να αντιστρέψει την πολικότητα μιας πρότασης

π.χ. That's not true, I'm a fan of this restaurant.

That's true, I'm not a fan of this restaurant.

Αντιθετικοί/ Εναντιωματικοί Σύνδεσμοι: Οι αντιθετικοί σύνδεσμοι σε μία πρόταση, όπως υποδηλώνει το όνομά τους, συνδέουν δύο φράσεις αντίθετης πολικότητας. Παραδείγματα αντιθετικών συνδέσμων στα αγγλικά είναι οι λέξεις but, although, however κ.ά. Συνήθως, η πολικότητα της πρότασης καθορίζεται από το δεύτερο συστατικό της πρότασης

π.χ. The restaurant is expensive but nice

The restaurant is nice but expensive

Ιδιωματισμοί: Μία άλλη αστοχία της μεθόδου είναι ότι μελετώντας κάθε λέξη ξεχωριστά, αγνοούμε την ύπαρξη φράσεων των οποίων οι επιμέρους λέξεις προσδίδουν ένα ιδιαίτερο συνολικό νόημα.

Ειρωνεία: Πολλές φορές οι άνθρωποι επιστρατεύουν το χιούμορ και τον σαρκασμό για να εκφράσουν την συνήθως αρνητική άποψή τους. Η ανίχνευση της ειρωνείας είναι πολλές φορές δύσκολη από τον άνθρωπο, πόσο μάλλον από μία μηχανή.

Για παράδειγμα, η πρόταση "The restaurant was great in that it will make all future meals seem more delicious"

Αμφισημία: Το φαινόμενο κατά το οποίο μία λέξη ή φράση έχει διαφορετικό νόημα ανάλογα με το ευρύτερο νοηματικό πλαίσιο στο οποίο χρησιμοποιείται ονομάζεται αμφισημία. Για παράδειγμα, η λέξη "unpredictable" έχει θετική έννοια όταν αναφέρεται σε μία ταινία και συνήθως αρνητική όταν αναφέρεται στον καιρό.

Οι Poria, Cambria et al (2014), εξετάζουν το πρόβλημα του sentiment analysis σε επίπεδο εννοιών (concept-level), σύμφωνα με το οποίο κάθε πρόταση εισόδου αναλύεται σε έννοιες και αυτές αναζητούνται σε ένα ειδικά σχεδιασμένο λεξικό, στο λεξικό SenticNet. Στο (Poria

etal, 2015), η ίδια ομάδα ερευνητών, πέτυχε ακόμη καλύτερα αποτελέσματα πάνω στα ίδια dataset, πάλι διεξάγοντας concept-level sentiment analysis αλλά χρησιμοποιώντας πιο σύνθετους κανόνες για την εύρεση των εξαρτήσεων μεταξύ των εννοιών και γλωσσολογικά patterns.

Πολλαπλοί Στόχοι: Πολλές φορές σε μία κριτική γίνεται αναφορά σε παραπάνω από μία οντότητες (πρόσωπα, προϊόντα, γεγονότα) ή ακόμα και σε διαφορετικά χαρακτηριστικά (aspects) της ίδιας οντότητας. Στην περίπτωση αυτή συνήθως δεν ενδιαφέρει η ταξινόμηση του κειμένου συνολικά ως μία θετική ή αρνητική άποψη αλλά εξετάζεται το κείμενο σε επίπεδο χαρακτηριστικών (aspect-level). Η ανάλυση συναισθήματος σε επίπεδο χαρακτηριστικών (Aspect-level SA) στοχεύει στον προσδιορισμό του συναισθήματος ως προς συγκεκριμένα aspects των οντοτήτων και διαφέρει από την ανάλυση συναισθήματος σε επίπεδο κειμένου (Document-level SA). Το πρώτο βήμα είναι η αναγνώριση των οντοτήτων και των χαρακτηριστικών τους. Στη συνέχεια εξάγονται οι γλωσσικές εκφράσεις που αναφέρονται στα ζεύγη (Entity, Aspect) και με χρήση συνωνύμων και λεξικού αποδίδεται το αντίστοιχο συναίσθημα. Για παράδειγμα, στην πρόταση:

"The food was delicious and the menu was perfect with something to suit eneryone"

{(food), positive}

{(menu), positive}

Υβριδική Προσέγγιση - Hybrid Approach

Η συνεχώς αυξανόμενη πρόοδος που παρατηρείται στον τομέα της Ανάλυσης συναισθήματος έχει οδηγήσει τους ερευνητές στη μελέτη πιθανότητας μιας υβριδικής προσέγγισης η οποία συνολικά θα παρουσίαζε την ακρίβεια των μεθόδων μηχανικής μάθησης και την ταχύτητα των μεθόδων βασισμένων σε λεξικά συναισθήματος. Στην ουσία αυτό που προσπαθούν οι υβριδικές προσεγγίσεις είναι να εκμεταλλευτούν τα πλεονεκτήματα των δύο μεθόδων και να αποφύγουν τα μειονεκτήματα. Στη γενική περίπτωση στις υβριδικές προσεγγίσεις η μία από τις δύο βασικές προσεγγίσεις χρησιμοποιείται για να τονώσει την απόδοση της άλλης προσέγγισης. Για παράδειγμα από τη μία μεριά οι επιβλεπόμενοι ταξινομητές μπορεί να κάνουν χρήση μεθόδων βασισμένων σε λεξικό για να μειώσουν την εξάρτηση από χειροκίνητες επισημειωμένες συλλογές δεδομένων εκπαίδευσης, ενώ από την άλλη οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για να «στηρίξουν» τα λεξικά συναισθήματος στις μεθόδους βασισμένες σε λεξικό (Taboada etal, 2011).

1.2 Κοινωνικός Ιστός - Κοινωνικά Δίκτυα στον Παγκόσμιο Ι-στό

Ένα κοινωνικό δίκτυο βασίζεται σε ένα λογισμικό το οποίο επιτρέπει στους ανθρώπους να συνδέονται, να δημιουργούν δυνητικές κοινότητες και να συνεργάζονται μέσω του υπολογιστή και του Διαδικτύου. Οι ιστοσελίδες κοινωνικής δικτύωσης παρέχουν αυτή τη δυνατότητα για αλληλεπίδραση. Ιστοσελίδες κοινωνικής δικτύωσης είναι εκείνες που επιτρέπουν στους επισκέπτες να στέλνουν email, να δημοσιεύουν σχόλια, να παίρνουν μέρος σε ζωντανές συζητήσεις και να χτίζουν διαδικτυακό περιεχόμενο (YALSA, 2009).

Η εποχή του Web 2.0, η οποία χαρακτηρίζεται από τον Κοινωνικό Ιστό και την πρόσβαση πληροφοριών σε πραγματικό χρόνο, έχει δύο όψεις που αναπτύσσονται ταυτόχρονα και επηρεάζουν η μία την εξέλιξη της άλλης. Η πρώτη όψη αφορά την ύπαρξη εφαρμογών και υπηρεσιών που προσφέρονται στους χρήστες του Παγκόσμιου Ιστού, με στόχο τη διευκόλυνσή τους για ανάρτηση υλικού, αξιολογήσεων, κρίσεων, σχολίων και γενικότερα για δημιουργία κλίματος 'κοινοτήτων' χρηστών (όπως είναι πχ. τα ιστολόγια ή τα κοινωνικά δίκτυα). Η δεύτερη όψη είναι οι απαιτούμενες τεχνολογίες. Για να λειτουργήσουν αυτές οι εφαρμογές αναπτύσσονται αντίστοιχα τεχνολογίες για να υποστηρίξουν τις απαιτήσεις ανάπτυξης και λειτουργίας τους (όπως είναι οι οντολογίες ή οι τεχνολογίες ανάπτυξης διαδικτυακών εφαρμογών).

Πλέον με την εξάπλωση του Σημασιολογικού Ιστού (Semantic Web) βρισκόμαστε στην αρχή της εποχής του Web 3.0 και κύριο χαρακτηριστικό αυτού είναι ότι τα δεδομένα που υπάρχουν στον Παγκόσμιο Ιστό αναλύονται και έτσι μπορούν να προσφέρονται εξατομικευμένες απαντήσεις/υπηρεσίες στους τελικούς χρήστες.

Οδεύοντας πλέον προς έναν «ευφυή» Ιστό (Intelligent Web), που όπως προμηνύεται σε μερικά χρόνια θα αποτελεί το Web 4.0 (Aghaei etal, 2012) μπορούμε να παρατηρήσουμε ότι η τωρινή κατάσταση του Διαδικτύου τροφοδοτεί την εξέλιξή του. Πιο συγκεκριμένα έχουμε την άνθηση του Διαδικτύου των Αντικειμένων (Internet of Things, IoT), όπου «αντικείμενα» (όπως είναι οι φυσικοί αισθητήρες) και άνθρωποι, μέσα από τις εικονικές προσωπικότητες που διαμορφώνουν στον Παγκόσμιο Ιστό, παράγουν δεδομένα τα οποία μέσα από την αξιοποίησή τους παρέχουν υπηρεσίες και εξατομικευμένες λύσεις για την εξυπηρέτηση διαφόρων τομέων της καθημερινής ζωής. Στα πλαίσια του IoT μια μεγάλη πρόκληση είναι τα μεγάλα δεδομένα και η διαχείρισή τους για την εξαγωγή ποιοτικών πληροφοριών.

Στην ίδια εν μέρει λογική, εμφανίζεται και μια νέα μορφή, το «Διαδίκτυο από Εσένα» (Internet of You, IoY) όπου κύριο χαρακτηριστικό του είναι τα μικρά δεδομένα που υπάρ-

χουν ξεχωριστά από κάθε έναν χρήστη. Μέσα από την ανάλυση των μικρών δεδομένων στόχος είναι να παρατηρείται η δραστηριότητα κάθε χρήστη στο Διαδίκτυο κατά τη διάρκεια της ημέρας, να συλλέγονται και να αναλύονται τα «προσωπικά ίχνη» τους (Aghaei etal, 2012).

Δίκτυα Και Γράφοι

Ο όρος κοινωνικό δίκτυο χρησιμοποιείται για να περιγράψει υπηρεσίες παγκόσμιου ιστού (webbased services) οι οποίες επιτρέπουν στους χρήστες τους να δημιουργήσουν προφίλ, δημόσια (public profiles) ή ημι-δημόσια (semi-public profiles), ούτως ώστε να επικοινωνούν με άλλους χρήστες του ίδιου δικτύου (Chelmis & Prasanna, 2011). Τα κοινωνικά δίκτυα αναπτύχθηκαν ιδιαίτερα μετά τη μετάβαση από το Web 1.0, όπου ο χρήστης είχε κυρίως παθητικό ρόλο χωρίς μεγάλες δυνατότητες δικής του συνεισφοράς, στο Web 2.0 (Kaplan & Haenlein, 2010).

Το Web 2.0 αποτελεί την εξέλιξη του Web (ή Web 1.0) και σηματοδοτεί τη μετάβαση από το «Read Only Web» στο «Read/Write Web» όπου η συνεισφορά των πληροφοριών και της γνώσης είναι τόσο εύκολη όσο και η «κατανάλωσή» τους (Tim, 2005). Η μετάβαση στο Web 2.0 χαρακτηρίζεται από τη μεταβολή του ρόλου του χρήστη από «καταναλωτή» σε «συμπαραγωγό». Ενώ στο Web 1.0 οι χρήστες μπορούσαν απλώς να διαβάσουν το περιεχόμενο ιστοσελίδων, στο Web 2.0 μπορούν να συνεισφέρουν, δημιουργώντας οι ίδιοι περιεχόμενο (user-generated content).

Παράλληλα με τη διαμόρφωση του περιεχομένου διαφόρων μορφών, όπως κείμενο, ήχος, εικόνα, βίντεο, στους χρήστες επαφίεται και η κατηγοριοποίηση, η αξιολόγηση και η κατάταξη του περιεχομένου, όπως για παράδειγμα ποια είδηση θεωρείται από αυτούς ως η πιο σημαντική. Ένα κοινωνικό δίκτυο μπορεί να αναπαρασταθεί με αρκετούς τρόπους. Ένας από τους περισσότερο χρήσιμους είναι μέσω της θεωρίας των γράφων. Στη μαθηματική θεωρία των γράφων, ένας γράφος αποτελείται από κόμβους (ή κορυφές), κάποιοι από τους οποίους (ενδεχομένως όλοι) συνδέονται μεταξύ τους. Στην περίπτωση που όλοι οι κόμβοι συνδέονται μεταξύ τους έχουμε έναν πλήρη γράφο. Οι συνδέσεις (links) μεταξύ των κόμβων, όταν υπάρχουν, μπορεί να είναι είτε τόξα ή γραμμές, αναλόγως του αν υπάρχει ή δεν υπάρχει κάποια κατεύθυνση στις συνδέσεις μεταξύ των κόμβων του γράφου.

Έτσι, σύμφωνα με τη θεωρία των γράφων, ένα κοινωνικό δίκτυο είναι μία κοινωνική δομή που αποτελείται από κόμβους, οι οποίοι συνδέονται με έναν ή περισσότερους τύπους αλληλεξάρτησης, όπως φιλία, συγγένεια, κοινά ενδιαφέροντα, οικονομικές συναλλαγές, αντιπάθεια, κοινές πεποιθήσεις, γνώση ή κύρος. Οι κόμβοι ανταλλάσσουν μεταξύ τους πόρους, οι οποίοι τους κρατούν συνδεδεμένους. Οι πόροι μπορεί να περιλαμβάνουν πληροφορίες, δεδομένα, κοινωνική υποστήριξη, ή οικονομική υποστήριξη. Κάθε είδος ανταλλασσόμενου πόρου θεωρείται σαν μια σχέση στο κοινωνικό δίκτυο, και τα άτομα που διατηρούν μια τέτοια σχέση

λέγεται ότι διατηρούν έναν δεσμό. Η ισχύς κάθε δεσμού μπορεί να ποικίλλει, από αδύνατο προς δυνατό, και εξαρτάται από τον αριθμό και τον τύπο των πόρων που ανταλλάσσονται, τη συχνότητα της ανταλλαγής καθώς και την σχετικότητα αυτών (Marsden & Campbell, 1984).

Η μορφή του κοινωνικού δικτύου βοηθά στον καθορισμό της χρησιμότητάς του προς τα διάφορα μέλη του. Μικρότερα, πιο στενά συνδεδεμένα δίκτυα, μπορεί να είναι λιγότερο χρήσιμα στα μέλη τους από ότι δίκτυα με πιο ασθενείς δεσμούς, όπως είναι τα ανοικτά δίκτυα με νέες ιδέες και ευκαιρίες για τα μέλη τους. Με άλλα λόγια ένα γκρουπ φίλων που έχουν δραστηριότητες μόνο μεταξύ τους, δεν έχει να προσφέρει νέα γνώση και προοπτική. Είναι σημαντικότερο για ένα άτομο να διαθέτει συνδέσεις σε διαφορετικά δίκτυα, παρά να έχει έναν αριθμό συνδέσεων μέσα σε ένα δίκτυο. Τα άτομα μπορούν να ασκούν μεγαλύτερη επιρροή στα κοινωνικά δίκτυα που ανήκουν με το να γεφυρώνουν δύο ή περισσότερα δίκτυα που δεν είναι άμεσα συνδεδεμένα και να συμπληρώνουν τις δομικές οπές.

Μέσα Κοινωνικής Δικτύωσης

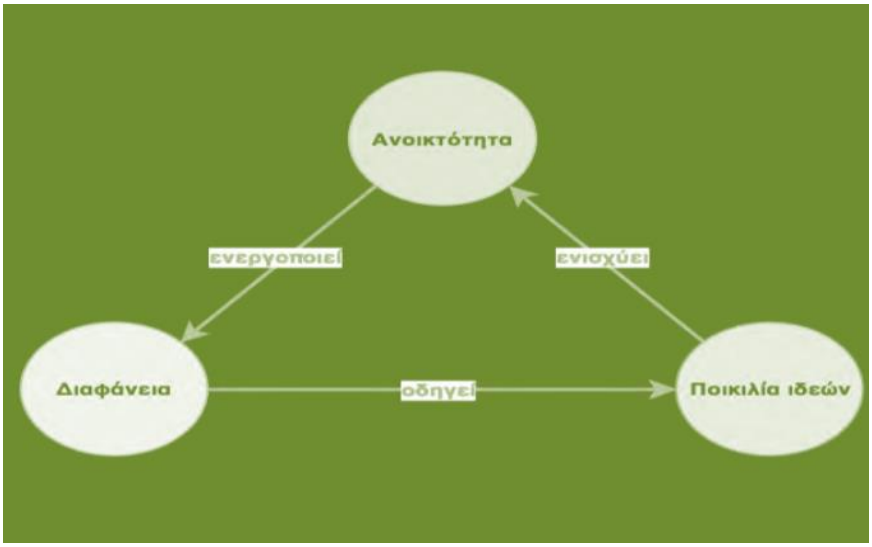
Ανατρέχοντας στη σχετική βιβλιογραφία, προκύπτει μια πληθώρα ορισμών για τα μέσα κοινωνικής δικτύωσης, γεγονός που καταδεικνύει το τεράστιο ενδιαφέρον της ακαδημαϊκής κοινότητας για το φαινόμενο αυτό. Ένας βασικός ορισμός μπορεί να θεωρηθεί ο ακόλουθος από τους συγγραφείς στο (Chelmis & Prasanna, 2011).

«Τα μέσα κοινωνικής δικτύωσης ορίζονται σαν ένα σύνολο από διαδικτυακές εφαρμογές που βασίζονται στα ιδεολογικά και τεχνολογικά θεμέλια του Web 2.0 και επιτρέπουν την δημιουργία και την ανταλλαγή περιεχομένου από τους χρήστες»

Επειδή πλέον τα δίκτυα συνολικά είναι πολλά και οι χρήστες σχεδόν αμέτρητοι, υπάρχει μεγάλη ανάγκη από όλο και περισσότερους χρήστες για αξιόπιστα δίκτυα. Θα λέγαμε ότι σε αξιόπιστα δίκτυα, η ανοικτότητα εξασφαλίζει τη διαφάνεια, η οποία με τη σειρά της προωθεί την ποικιλομορφία των ιδεών. Αυτό φαίνεται και σχηματικά στην εικόνα 2.

Υποστηρίζοντας τη δημιουργία κοινωνικών δικτύων μπορούμε να αυξήσουμε τη διάδοση της γνώσης που μπορεί να οδηγήσει σε περισσότερη καινοτομία, ιδίως όταν τα δίκτυα αυτά βασίζονται στην εμπιστοσύνη.

Αν και ο όρος "μέσα κοινωνικής δικτύωσης" είναι αρκετά γενικός και καλύπτει διάφορες διαδικτυακές πλατφόρμες με διαφορετικές ιδιότητες, μορφές επικοινωνίας, και λειτουργίες, υπάρχουν ορισμένα χαρακτηριστικά τα οποία όλα τα μέσα ουσιαστικά μοιράζονται.



Εικόνα 2 Αξιόπιστα κοινωνικά δίκτυα (http://jarche.com/)

Κοινωνικά Δίκτυα

Σήμερα τα μέσα κοινωνικής δικτύωσης, αποτελούν ένα ιδιαίτερα δημοφιλές επικοινωνιακό εργαλείο, μεταξύ των χρηστών του διαδικτύου. Ένα κοινωνικό δίκτυο, είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Οι δημοφιλέστερες ιστοσελίδες κοινωνικών δικτύων είναι το Facebook, το Twitter, το Linkedin και το Instagram.

Τα κοινωνικά δίκτυα, ως δυνητικές κοινότητες, συνιστούν «χώρους» ανταλλαγής και παράθεσης ιδεών και απόψεων για τους χρήστες, παρέχοντας μία πλούσια πηγή δεδομένων για Ανάλυση Συναισθήματος και Εξόρυξη Γνώμης. Η Εξόρυξη Γνώμης, αναδεικνύει την συνολική άποψη των χρηστών αναφορικά με ένα θέμα, που συζητείται στα μέσα κοινωνικής δικτύωσης, εντοπίζει ομάδες χρηστών στο γενικό πληθυσμό και μπορεί να συστήσει προϊόντα ή δραστηριότητες στους χρήστες, είτε βάσει των προτιμήσεων τους, είτε με κριτήριο προηγούμενες επιλογές τους.

Η online κοινωνική δικτύωση ταξιδιού είναι ένας εναλλακτικός τρόπος που οι τουρίστες προγραμματίζουν τα ταξίδια τους. Αυτές οι ιστοσελίδες επιτρέπουν στους χρήστες να αλληλεπιδράσουν και να αναρτήσουν τα σχόλιά τους για τα ξενοδοχεία, εστιατόρια ή τα τοπικά αξιοθέατα στο διαδίκτυο. Το 2000 ιδρύθηκε ένας ιστότοπος βασισμένος στην ιδέα ότι οι ταξιδιώτες στηρίζονται στα σχόλια των άλλων ταξιδιωτών για να προγραμματίσουν τις διακοπές τους και τι θα επισκεφτούν.

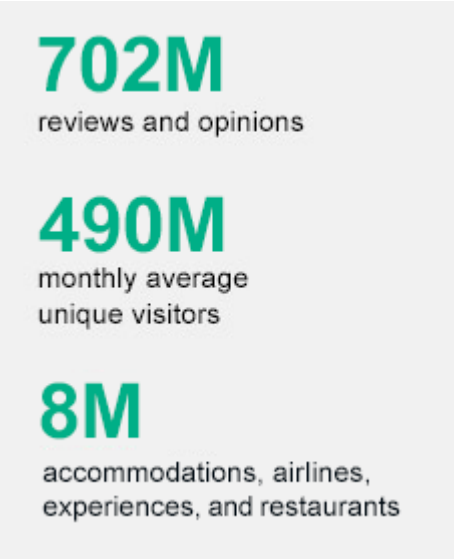
Το TripAdvisor είναι μια ιστοσελίδα όπου οι περισσότερες πληροφορίες που αναρτώνται είναι πρωτοβουλία των χρηστών του. Αναρτούν τα σχόλια και τις εκτιμήσεις τους για ένα προορισμό ένα ξενοδοχείο, ένα εστιατόριο, ένα αξιοθέατο ή για οποιοδήποτε άλλο τουριστικό αντικείμενο ή υπηρεσία.. Επιτρέπει στους χρήστες να προσθέτουν τις εμπειρίες τους από τα ταξίδια τους, με σύνδεση από τις προϋπάρχουσες πηγές (π.χ. διευθύνσεις ηλεκτρονικού ταχυ-

δρομείου, άλλα blogs κ.λπ.). Αυτό μπορεί να βοηθήσει στην οργάνωση του ταξιδιού με έναν πιο αποτελεσματικό τρόπο.

1.2.1 Πλατφόρμα TripAdvisor

Εικόνα 3 TripAdvisor (https://www.tripadvisor.com/)

Στα πλαίσια του πειράματος ανάλυσης συναισθημάτων χρησιμοποιήθηκαν σχόλια χρηστών από την διαδικτυακή πλατφόρμα Tripadvisor (https://www.tripadvisor.com/). Το Tripadvisor είναι ένας από τους μεγαλύτερους ταξιδιωτικούς ιστοτόπους παγκοσμίως. Φιλοξενεί απόψεις/κριτικές χρηστών σχετικά με προϊόντα και υπηρεσίες εστιατορίων, μουσείων, ξενοδοχείων και γενικότερα οργανισμών και επιχειρήσεων που σχετίζονται με τα ταξίδια. Αποτελεί ένα φόρουμ απόψεων και πληροφόρησης.



Εικόνα 4 Στατιστικά Νοεμβρίου 2018 (https://www.tripadvisor.com)

Το TripAdvisor αποτελεί μία από τις μεγαλύτερες ταξιδιωτικές ιστοσελίδες στον κόσμο και δίνει τη δυνατότητα στους ταξιδιώτες να αξιοποιήσουν όλες τις δυνατότητες κάθε ταξιδιού. Με 702 εκατομμύρια κριτικές (όπως παρουσιάζεται μέχρι τον Νοέμβριο του 2018) και απόψεις που καλύπτουν τη μεγαλύτερη λίστα ταξινομήσεων στον κόσμο παγκοσμίως - καλύπτουν 8 εκατομμύρια καταλύματα, αεροπορικές εταιρείες, εμπειρίες και εστιατόρια - το TripAdvisor παρέχει στους ταξιδιώτες τη γνώση του πλήθους για να βοηθήσει να αποφασίσουν πού θα μείνουν, τι να κάνουν και πού να φάνε σε κάθε προορισμό ανάλογα με τις προτιμήσεις τους. Το TripAdvisor συγκρίνει επίσης τιμές από περισσότερες από 200 ιστοσελίδες κρατήσεων ξενοδοχείων, ώστε οι ταξιδιώτες να μπορούν να βρουν τη χαμηλότερη τιμή στο ξενοδοχείο που τους ταιριάζει. Οι ιστοσελίδες με εμπορικό σήμα TripAdvisor διατίθενται σε 49 αγορές και φιλοξενούν τη μεγαλύτερη ταξιδιωτική κοινότητα στον κόσμο με 490 εκατομμύρια μέσους μηνιαίους μοναδικούς επισκέπτες καθώς όλοι θέλουν να αξιοποιήσουν στο έπακρο κάθε ταξίδι.

Το TripAdvisor, Inc. είναι μια αμερικανική εταιρεία ταξιδιών και εστιατορίων που παρουσιάζει κριτικές ξενοδοχείων και εστιατορίων, κρατήσεις καταλυμάτων και άλλο ταξιδιωτικό περιεχόμενο. Περιλαμβάνει επίσης διαδραστικά ταξιδιωτικά φόρουμ. Οι υπηρεσίες ιστοτόπου που παρέχει είναι δωρεάν στους χρήστες, οι οποίοι παρέχουν το μεγαλύτερο μέρος του περιεχομένου, και ο ιστότοπος υποστηρίζεται από μια ξενοδοχειακή μονάδα κρατήσεων και ένα επιχειρηματικό μοντέλο διαφήμισης. Το TripAdvisor ιδρύθηκε από τον Stephen Kaufer, τον

Langley Steinert, τον Nick Shanny και τον Thomas Palka τον Φεβρουάριο του 2000. Ο Kaufer δήλωσε ότι η αρχική ιδέα δεν ήταν μια ιστοσελίδα που δημιουργήθηκε από τους χρήστες των κοινωνικών μέσων για να ανταλλάξουν κριτικές αλλά «Ξεκινήσαμε ως ένα site όπου συγκεντρώναμε περισσότερο τα σχόλια από οδηγούς ή εφημερίδες ή περιοδικά.» (BBC Radio 4: The Bottom Line. 2014) Σιγά σιγά οι επισκέπτες πρόσθεταν τα δικά τους σχόλια και κατέληξε στη σημερινή μορφή. Η αρχική χρηματοδότηση ελήφθη από την εταιρία Flagship Ventures, τον όμιλο Bollard και από ιδιώτες επενδυτές.

Αμφιλεγόμενα Και Κακόβουλα Σχόλια

Το TripAdvisor έχει αποτελέσει αντικείμενο διαμάχης για την ανάρτηση ανεπιβεβαίωτων ανώνυμων αξιολογήσεων για οποιοδήποτε ξενοδοχείο, πανσιόν ή εστιατόριο. Περίπου 30 ξενοδοχεία είχαν μπει στη «μαύρη λίστα» του TripAdvisor για ύποπτες κριτικές, όπως ένα ξενοδοχείο στην Κορνουάλη που δωροδόκησε τους επισκέπτες να αφήσουν θετικές κριτικές για το ξενοδοχείο.

Το TripAdvisor έχει δηλώσει ότι οι κριτικές δεν δημοσιεύονται αμέσως στον ιστότοπο, αλλά υπόκεινται σε διαδικασία επαλήθευσης που εξετάζει τη διεύθυνση IP και τη διεύθυνση ηλεκτρονικού ταχυδρομείου του συγγραφέα και προσπαθεί να εντοπίσει τυχόν ύποπτα πρότυπα ή άσεμνη ή καταχρηστική γλώσσα. Ο ιστότοπος επιτρέπει επίσης στην κοινότητα χρηστών να αναφέρει ύποπτο περιεχόμενο, το οποίο στη συνέχεια αξιολογείται από το προσωπικό του Tripadvisor.

Τον Σεπτέμβριο του 2011, η Αρχή Διαφημιστικών Προτύπων του Ηνωμένου Βασιλείου (ASA) ξεκίνησε επίσημη έρευνα σχετικά με το TripAdvisor αφού έλαβε μια καταγγελία που υπέβαλε η εταιρεία KwikChex σχετικά με τις έρευνες στο διαδίκτυο ότι οι ισχυρισμοί της να παρέχουν αξιόπιστες και ειλικρινείς κριτικές από ταξιδιώτες είναι ψευδείς. Η ASA διαπίστωσε ότι το TripAdvisor δεν πρέπει να ισχυρίζεται ή να υπονοεί ότι όλες οι κριτικές του ήταν από πραγματικούς ταξιδιώτες ή ήταν ειλικρινείς, πραγματικές ή έμπιστες και ως αποτέλεσμα της έρευνας, το TripAdvisor διατάχθηκε να καταργήσει το σύνθημα από τον ιστοχώρο του στο Ηνωμένο Βασίλειο. Έχει αλλάξει το σύνθημα του τμήματος αναθεώρησης του ξενοδοχείου στις «κριτικές από την κοινότητά μας».

Το TripAdvisor δήλωσε ότι η αλλαγή του branding είχε προγραμματιστεί για κάποιο χρονικό διάστημα και ότι οι αλλαγές άρχισαν τον Ιούνιο του 2011, πριν από την έρευνα ASA. Η ASA δήλωσε ότι «ανησυχούσε ότι οι καταναλωτές θα μπορούσαν να ξεγελαστούν από ψευδείς καταχωρήσεις δεδομένου ότι οι καταχωρήσεις θα μπορούσαν να πραγματοποιηθούν χωρίς καμία μορφή επαλήθευσης», αλλά αναγνώρισε ότι το TripAdvisor χρησιμοποίησε «προηγμένα και εξαιρετικά αποτελεσματικά συστήματα απάτης» σε μια προσπάθεια εντοπισμού και κατάργησης πλαστών περιεχομένων (BBC Radio 4: The Bottom Line. 2014).

Πως Λειτουργεί η Κατάταξη Δημοτικότητας (Popularity Ranking)

Η κατάταξη δημοτικότητας βασίζεται στην ποιότητα, την ποσότητα των σχολίων και το πόσο πρόσφατα είναι, που λαμβάνει μια επιχείρηση από τους χρήστες - και τη συνέπεια αυτών των αξιολογήσεων με την πάροδο του χρόνου.

Ποιότητα

Οι αξιολογήσεις φυσαλίδων που παρέχουν οι χρήστες ως μέρος των αξιολογήσεών τους χρησιμοποιούνται για να ταξινομήσουν την ποιότητα της εμπειρίας σε κάθε επιχείρηση. Όταν όλα τα υπόλοιπα πράγματα είναι ίσα, η επιχείρηση με τις περισσότερες βαθμολογίες των 5 φυσαλίδων κατατάσσεται υψηλότερα από μια επιχείρηση με χαμηλότερες βαθμολογίες φυσαλίδων.

Επίκαιρο

Οι πρόσφατες κριτικές είναι πιο πολύτιμες από τις παλαιότερες κριτικές. Δίνουν μια πιο ακριβή αναπαράσταση της τρέχουσας εμπειρίας στην επιχείρηση. Αυτό σημαίνει ότι οι κριτικές - καλές ή κακές - που είναι παλαιότερες δεν υπολογίζονται τόσο πολύ στην κατάταξη μιας επιχείρησης, όπως μια επισκόπηση που γράφτηκε πιο πρόσφατα. Αν και οι παλαιότερες κριτικές δεν έχουν τόσο σημαντικό βάρος στην κατάταξη, εξακολουθούν να είναι ορατές στην ενότητα Επισκόπηση κάθε καταχώρισης και στο ιστορικό αναθεώρησης της επιχείρησης.

Ποσότητα

Ο αριθμός των κριτικών αποτελεί έναν κρίσιμο δείκτη για τους χρήστες του TripAdvisor σχετικά με μια επιχείρηση. Οι χρήστες του TripAdvisor διαβάζουν συνήθως πολλές κριτικές για να διαμορφώσουν μια ισορροπημένη γνώμη για μια επιχείρηση και να έχουν περισσότερη εμπιστοσύνη στις αποφάσεις τους όταν βλέπουν συμφωνία σε ένα μεγάλο σύνολο σχολίων.

Όταν μιλάμε για την ποσότητα κριτικών, είναι σημαντικό να σημειώσουμε ότι μια επιχείρηση πρέπει απλώς να έχει αρκετές κριτικές για να παρέχει στατιστική σημασία και να επιτρέπει τη σύγκριση με άλλες επιχειρήσεις. Παρόλα αυτά, περισσότερες κριτικές δεν σημαίνει ότι μια επιχείρηση θα κατατάσσεται υψηλότερα από τους ανταγωνιστές της. Για παράδειγμα, μια επιχείρηση με 1.000 κριτικές δεν θα είναι αναγκαστικά υψηλότερη από μία με 500 κριτικές που ελήφθησαν περίπου κατά την ίδια χρονική περίοδο. Αυτό οφείλεται στο γεγονός ότι και οι δύο έχουν αρκετές κριτικές για να μας κάνουν να είμαστε σίγουροι για την πιθανή ταξιδιωτική εμπειρία που μπορούν να προσφέρουν.

Ο αλγόριθμος κατάταξης δημοτικότητας έχει σχεδιαστεί για να παρέχει ένα στατιστικό μέτρο εμπιστοσύνης σχετικά με την τρέχουσα εμπειρία σε μια επιχείρηση. Καθώς συσσωρεύουμε περισσότερες κριτικές σχετικά με μια επιχείρηση με την πάροδο του χρόνου, έχουμε περισσότερες πληροφορίες σχετικά με τη δυνητική εμπειρία που μπορούν να περιμένουν οι κατα-

ναλωτές. Μόλις φτάσουμε σε μια κρίσιμη μάζα κριτικών, είμαστε σε θέση να προβλέψουμε με μεγαλύτερη ακρίβεια την κατάταξη της επιχείρησης.

Συνοψίζοντας:

- Οι καλές κριτικές είναι καλύτερες από τις κακές κριτικές
- Οι πρόσφατες κριτικές έχουν περισσότερο βάρος από τις παλαιότερες κριτικές
- Περισσότερες κριτικές βοηθούν στην οικοδόμηση της εμπιστοσύνης πιο γρήγορα

Αυτοί οι παράγοντες αλληλεπιδρούν με την πάροδο του χρόνου για να καθορίσουν την κατάταξη δημοτικότητας μιας επιχείρησης. Για παράδειγμα, η ποιότητα και η ποσότητα των κριτικών με την πάροδο του χρόνου μας παρέχει μια άποψη για τη συνοχή κάθε επιχείρησης. Μια επιχείρηση που έχει συνεχώς καλές κριτικές θα είναι υψηλότερη από μία με παρόμοιο αριθμό καλών και κακών κριτικών. Ομοίως, τα επίκαιρα σχόλια και η ποσότητα είναι στενά συνδεδεμένα - ένας μεγάλος αριθμός πρόσφατων αξιολογήσεων θα εκτιμηθεί υψηλότερα από εκείνες που υπάρχουν αρκετά χρόνια.

Με την πάροδο του χρόνου, το Tripadvisor συνεχίζει να βελτιστοποιεί και να βελτιώνει την κατάταξη δημοτικότητας για να βελτιώνεται η εμπειρία στον ιστότοπο, τόσο για τους χρήστες όσο και για τις επιχειρήσεις. Η κατάταξη είναι ιδιαίτερα σημαντική στο TripAdvisor για επιχειρήσεις. Οποιεσδήποτε αλλαγές είναι προσεκτικά σχεδιασμένες και δοκιμασμένες για να βελτιώσουν τον αλγόριθμο κατάταξης με πολύ συγκεκριμένους τρόπους, διατηρώντας παράλληλα την ακριβή κατάταξη των υφιστάμενων επιχειρήσεων στο TripAdvisor.

Οι πρόσφατες αλλαγές, το 2016 και το 2018, επικεντρώθηκαν στην προώθηση μεγαλύτερης συνεκτικότητας στην κατάταξη. Ο στόχος με κάθε αλλαγή ήταν να αντικατοπτρίζει με μεγαλύτερη ακρίβεια την απόδοση μιας επιχείρησης σε σχέση με άλλους στην τοποθεσία της με την πάροδο του χρόνου, ανεξάρτητα από το μέγεθός της ή τον ρυθμό με τον οποίο συλλέγει κριτικές.

Επειδή η Ταξινόμηση Δημοτικότητας βασίζεται σε ανατροφοδότηση των χρηστών, η συνεχής συλλογή νέων, υψηλής ποιότητας κριτικών - που αντανακλούν επίπεδα υπηρεσιών και αξίας που ικανοποιούν τις προσδοκίες - είναι ο καλύτερος τρόπος για τις επιχειρήσεις να βελτιώσουν τη θέση τους με την πάροδο του χρόνου.

Είναι σημαντικό να σημειωθεί ότι η Κατάταξη δημοτικότητας λαμβάνει υπόψη τις επιδόσεις μιας επιχείρησης σε σχέση με τις επιδόσεις άλλων επιχειρήσεων στην περιοχή. Καθώς μια επιχείρηση ανεβαίνει στην κατάταξη, επηρεάζει άλλους άμεσα γύρω της. Οι κινήσεις στην κατάταξη δημοτικότητας μπορεί να είναι αποτέλεσμα των αξιολογήσεων μιας συγκεκριμένης επιχείρησης - ή των αξιολογήσεων που έχουν προκύψει για άλλες επιχειρήσεις στην περιοχή.

Το TripAdvisor προσφέρει μια ποικιλία εργαλείων όπως το Review Express και τα εργαλεία υπενθύμισης, καθώς και τις Συλλογές Αναθεώρησης Συλλογής. Οι επιχειρήσεις που χρησιμοποιούν τα εργαλεία συλλογής αναθεωρήσεων μπορούν να μετρήσουν με συνέπεια την απόδοσή τους και να κάνουν βελτιώσεις βάσει των σχολίων που λαμβάνουν.

Οι θέσεις στην κατάταξη δεν επηρεάζονται αναγκαστικά από μία μόνο επισκόπηση - θετική ή αρνητική. Οι επιχειρήσεις ενδέχεται να παρατηρούν αλλαγές στις θέσεις τους ανάλογα με τις δικές τους επιδόσεις ή με βάση τον τρόπο με τον οποίο εκτελούν άλλες επιχειρήσεις γύρω τους. Δεδομένου ότι ο αλγόριθμος λαμβάνει υπόψη τη συνολική ποιότητα των σχολίων που λαμβάνει μια επιχείρηση με την πάροδο του χρόνου, η τακτική συλλογή σχολίων από τους χρήστες είναι ο καλύτερος τρόπος για να διατηρηθεί ή να βελτιωθεί η θέση της επιχείρησης με την πάροδο του χρόνου, σε σχέση με τις επιδόσεις άλλων επιχειρήσεων στην περιοχή.

Οι απαντήσεις των διαχειριστών δεν λαμβάνονται υπόψη στην κατάταξη δημοτικότητας. Ωστόσο, η έρευνα δείχνει ότι όταν ένας ιδιοκτήτης απαντά άμεσα και επαγγελματικά σε ανασκόπηση, αντιμετωπίζοντας οποιεσδήποτε συγκεκριμένες καταγγελίες καθώς και τα θετικά σχόλια, μπορεί να έχει μεγάλο αντίκτυπο στους υποψήφιους πελάτες. Μια μελέτη δείχνει ότι το 85% των χρηστών λένε ότι μια προσεκτική απάντηση σε μια κριτική βελτιώνει την εντύπωση για ένα ξενοδοχείο και το 65% είναι πιο πιθανό να κλείσει ένα ξενοδοχείο που ανταποκρίνεται σε κριτικές έναντι ενός συγκρίσιμου ξενοδοχείου που δεν το κάνει (Custom Survey Research Engagement, 2015).

Οι εμπορικές σχέσεις μιας επιχείρησης με το TripAdvisor δεν έχουν καμία απολύτως επίδραση στην κατάταξη δημοτικότητας. Η κατάταξη υπολογίζεται καθημερινά με βάση όλες τις δημοσιευμένες κριτικές, συμπεριλαμβανομένων των νέων αξιολογήσεων που ελήφθησαν εκείνη την ημέρα.

1.2.2 Πλατφόρμα Aylien

Εικόνα 5 Aylien (https://aylien.com/)

Το AYLIEN είναι μια πλατφόρμα τεχνητής νοημοσύνης η οποία επικεντρώνεται στη δημιουργία τεχνολογιών οι οποίες βοηθούν τις μηχανές να κατανοήσουν καλύτερα τις ανθρώπινες γλώσσες. Η εταιρία παρέχει ανάλυση κειμένου και API τα οποία επιτρέπουν στους χρήστες να κατανοούν ανθρώπινο περιεχόμενο. Η AYLIEN προσφέρει επίσης μια σειρά λύσεων ανάλυσης περιεχομένου σε προγραμματιστές, επιστήμονες δεδομένων, εμπόρους και ακαδημαϊκούς. Βασικές υπηρεσίες περιλαμβάνουν πακέτα ανάκτησης πληροφοριών, μηχανική μάθηση, επεξεργασία φυσικής γλώσσας και API αναγνώρισης εικόνων. Το AYLIEN χρησιμοποιεί βαθιά μάθηση για την εξαγωγή συναισθήματος. Το AYLIEN Text API είναι ένα πακέτο εργαλείων Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing-NLP), Ανάκτησης Πληροφοριών

και Εκμάθησης Μηχανών για την εξαγωγή σημασίας από κειμενογραφικό και οπτικό περιεχόμενο. Αυτό το API μπορεί να χρησιμοποιηθεί για διάφορες εργασίες όπως:

Το API ανάλυσης κειμένου AYLIEN αποτελείται από οκτώ διαφορετικά API φυσικής επεξεργασίας γλώσσας, ανάκτησης πληροφοριών και μηχανικής μάθησης, τα οποία, όταν συνδυαστούν, επιτρέπουν στους προγραμματιστές να εξάγουν νόημα και διορατικότητα από οποιοδήποτε έγγραφο.

Το API ανάλυσης κειμένου AYLIEN API είναι ένα πακέτο API για την επεξεργασία φυσικής γλώσσας και την εκμάθηση μηχανών μάθησης για την ανάλυση και την εξαγωγή διαφόρων ειδών πληροφοριών από περιεχόμενο κειμένου. Παρέχονται λειτουργίες ταξινόμησης κειμένου, ανάλυσης αισθητήρων, εξατομίκευσης οντοτήτων και συγκεντρωτικών σχηματισμών για την εξαγωγή ουσιαστικής γνώσης και κατανόησης του περιεχομένου του κειμένου. Για ακόμα καλύτερα αποτελέσματα, συνίσταται η χρήση δύο ή περισσότερων APIs (https://developer.aylien.com).

Ταξινόμηση

Η γνώση της σημασιολογικής κατηγορίας υψηλού επιπέδου ενός μη επισημειωμένου εγγράφου, όπως μιας ιστοσελίδας ή ενός άρθρου, μπορεί να είναι εξαιρετικά χρήσιμη σε διάφορες εφαρμογές. Το τελικό σημείο ταξινόμησης βοηθά στην κατηγοριοποίηση οποιουδήποτε κειμένου ή διεύθυνσης URL σύμφωνα με μια προκαθορισμένη ταξινόμηση. Επίσης, προσφέρεται η δυνατότητα στους χρήστες να δημιουργήσουν τα δικά τους προσαρμοσμένα συστήματα ταξινόμησης ή να τροποποιήσουν έτοιμα του Aylien σύμφωνα με τις απαιτήσεις τους.

Classification By Taxonomy

Taxonomies

Το τελικό σημείο κατάταξης βάσει ταξινόμησης είναι ικανό να ταξινομεί περιεχόμενο σύμφωνα με πολλαπλές ταξινομίες που μπορούν να επιλεγούν με την προσθήκη του αναγνωριστικού της ταξινόμησης στο τέλος του ταξινομημένου τελικού σημείου.

Traversing Taxonomies

Όλες οι υποστηριζόμενες ταξινομίες είναι τυποποιημένες σε μια δομή που μοιάζει με δέντρο, η οποία επιτρέπει την εύκολη μετάβαση από κατηγορίες παιδιών σε κατηγορίες γονέων, αναδρομικά. Κάθε αποτέλεσμα ταξινόμησης περιέχει μια σειρά συνδέσμων, η οποία περιέχει συνδέσμους προς την τρέχουσα ετικέτα ταξινόμησης (rel = self) καθώς και για τους γονείς της, αν υπάρχουν (rel = parent).

Sentiment Analysis

Η εξαγωγή συναισθημάτων από ένα κείμενο, όπως ένα tweet, μια κριτική ή ένα άρθρο, μπορεί να μας δώσει μια πολύτιμη γνώση σχετικά με τα συναισθήματα και την προοπτική του συγγραφέα: εάν ο τόνος είναι θετικός, ουδέτερος ή αρνητικός και αν το κείμενο είναι υποκειμενικό (δηλαδή εκφράζει τη γνώμη του συντάκτη) ή αντικειμενικό (δηλαδή εκφράζει ένα γεγονός).

Στην ανάλυση Sentiment, γίνεται χρήση της κατάλληλης παραμέτρου λειτουργίας ανάλογα με το κείμενο εισόδου: tweet για σύντομο κείμενο, όπως ενημερώσεις κοινωνικών μέσων και έγγραφο για μεγαλύτερο κείμενο, όπως μια κριτική ή ένα άρθρο. Η ανάλυση της υποκειμενικότητας είναι προς το παρόν διαθέσιμη μόνο σε λειτουργία tweet και συνήθως λειτουργεί καλύτερα με είσοδο σε επίπεδο προτάσεων.

Entity Level Sentiment Analysis

Το τελικό σημείο της Ανάλυσης Συναισθημάτων στο επίπεδο οντοτήτων (Entity-level Sentiment Analysis - ELSA) παρέχει το συναίσθημα που σχετίζεται με την οντότητα που αναφέρεται σε ένα έγγραφο.

Για κάθε οντότητα που αναφέρεται σε ένα κείμενο, η ELSA θα επιστρέψει:

• μια πρόβλεψη της πολικότητας των αισθήσεων που εκφράζεται προς αυτήν την οντότητα (θετική, αρνητική ή ουδέτερη)

• το σκορ εμπιστοσύνης της πρόβλεψης

• τον τύπο της οντότητας (για παράδειγμα Πρόσωπο, Οργανισμός, Τοποθεσία, Προϊόν)

• URI DBPedia, όπου ισχύει

Το τελικό σημείο ELSA είναι ιδιαίτερα χρήσιμο για την ανάλυση άρθρων που αναφέρονται σε πολλαπλές οντότητες, ενώ εκφράζουν διαφορετικά συναισθήματα για κάθε μία. Μπορεί να χρησιμοποιηθεί για την ανάλυση μιας οντότητας σε ένα ενιαίο έγγραφο ή σε πολλά έγγραφα. Το τελικό σημείο δέχεται τόσο κείμενο όσο και urls ως παραμέτρους και η μόνη γλώσσα που υποστηρίζεται αυτή τη στιγμή είναι η αγγλική.

Aspect-Based Sentiment Analysis

Ορισμένοι τύποι εγγράφων, όπως ανατροφοδότηση από τους πελάτες ή σχόλια, ενδέχεται να περιέχουν με λεπτό τρόπο συναίσθημα σχετικά με διάφορες πτυχές των οντοτήτων (π.χ. προϊόντων ή υπηρεσιών) που αναφέρονται στο έγγραφο. Για παράδειγμα, μια αναθεώρηση σχετικά με ένα ξενοδοχείο μπορεί να περιέχει προτάσεις για το προσωπικό, τα κρεβάτια και

την τοποθεσία του. Αυτές οι πληροφορίες μπορούν να είναι πολύτιμες για την κατανόηση της άποψης των πελατών σχετικά με μια συγκεκριμένη υπηρεσία ή προϊόν.

Χρησιμοποιώντας το τελικό σημείο της ανάλυσης συναισθήματος ως προς την πτυχή (Aspect-based Sentiment Analysis - ABSA) γίνεται ανάκτηση μιας λίστας πτυχών που αναφέρονται σε ένα έγγραφο το οποίο ανήκει σε ένα συγκεκριμένο τομέα και το συναίσθημα του συγγραφέα προς κάθε μία από αυτές τις πτυχές.

Supported Domains

Οι ακόλουθες τιμές που αντιστοιχούν σε διαφορετικούς τομείς υποστηρίζονται επί του παρόντος και είναι αποδεκτές για την παράμετρο τομέα:

- Ξενοδοχεία
- Εστιατόρια
- Αυτοκίνητα
- Αεροπορικές εταιρείες

Entity Extraction

Τα έγγραφα περιέχουν συχνά αναφορές οντοτήτων όπως οι άνθρωποι, οι τόποι, τα προϊόντα και οι οργανισμοί, τους οποίους ονομάζουμε ομαδοποιημένες οντότητες συλλογικά. Επιπλέον, ενδέχεται να περιέχουν συγκεκριμένες τιμές ή στοιχεία, όπως συνδέσμους, αριθμούς τηλεφώνου, διευθύνσεις ηλεκτρονικού ταχυδρομείου, ποσά νομισμάτων και ποσοστά. Για την εξαγωγή αυτών των οντοτήτων και τιμών από ένα τμήμα κειμένου, καθώς και τις λέξεις-κλειδιά, μπορεί να χρησιμοποιηθεί το τελικό σημείο εξαγωγής οντοτήτων. Η εξαγωγή οντοτήτων εξετάζει τα δομικά μοτίβα ενός εγγράφου για την εύρεση και εξαγωγή οντοτήτων και επομένως μπορεί να είναι επιρρεπής σε σφάλματα (https://developer.aylien.com).

Concept Extraction

Το τελικό σημείο εξαγωγής Concept Extracts αποκαλύπτει διαφορετικούς τύπους αξιοσημείωτων οντοτήτων από ένα έγγραφο, χρησιμοποιώντας την Wikipedia (και ενδεχομένως άλλες βάσεις γνώσεων) ως πλαίσιο. Επίσης, συνδέεται με τα Συνδεδεμένα Ανοικτά Δεδομένα για να παρέχει δομημένα δεδομένα γύρω από τις εξαγόμενες οντότητες, όπως LOD URIs που μπορούν να χρησιμοποιηθούν για την ανάκτηση πρόσθετων πληροφοριών σχετικά με μια οντότητα, όπως το ύψος ενός ατόμου ή την τιμή της μετοχής μιας εταιρείας, καθώς και σημασιολογικούς τύπους μιας οντότητας (DBpedia, Schema.org, κ.λπ.) που μπορούν να χρησιμοποιη-

θούν για φιλτράρισμα οντοτήτων με βάση τον τύπο τους. Για εξαγωγή γνωστών οντοτήτων με μεγαλύτερη ακρίβεια συνίσταται η ταυτόχρονη εξόρυξη οντότητας και πλαισίου.

Text Analysis Platform (TAP)

Το TAP αποτελεί τον ευκολότερο τρόπο για τη δημιουργία προσαρμοσμένων μοντέλων NLP στο σύννεφο. Μερικές φορές, δεν είναι αρκετό ένα γενικό μοντέλο επεξεργασίας φυσικής γλώσσας. Με το TAP, μπορούμε να δημιουργήσουμε, να εκπαιδεύσουμε, να αξιολογήσουμε και να αναπτύξουμε τα δικά μας προσαρμοσμένα μοντέλα NLP μέσα σε λίγα λεπτά, χωρίς μηχανική μάθηση, επεξεργασία φυσικής γλώσσας ή απαιτούμενες γνώσεις κωδικοποίησης. Πρόκειται για ένα δωρεάν πρόγραμμα Beta στο οποίο μπορούμε να δημιουργήσουμε το σύνολο δεδομένων από την αρχή ή να χρησιμοποιήσουμε τα υπάρχοντα σύνολα δεδομένων για να εκπαιδεύσουμε το προσαρμοσμένο μοντέλο μας. Δίνεται η δυνατότητα ελέγχου και αξιολόγησης των επιδόσεων κάθε μοντέλου και προσαρμογών όπου χρειάζεται. Όταν η απόδοση του μοντέλου είναι ικανοποιητική, είναι πλέον προς παραγωγή και μπορούν να γίνουν κλήσεις σε αυτό μέσω του API TAP.

1.3 Περιβάλλον WEKA

Η εφαρμογή WEKA είναι ένα εργαλείο εξόρυξης γνώσης το οποίο είναι υλοποιημένο σε JAVA, από το Πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Το WEKA περιέχει διάφορους αλγόριθμους μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν στο πεδίο της εξόρυξης γνώσης και της Μηχανικής Μάθησης. Οι αλγόριθμοι αυτοί μπορούν είτε να κληθούν απ' ευθείας είτε να κληθούν μέσα από πρόγραμμα σε JAVA. Περιέχει εργαλεία για:

- προεπεξεργασία δεδομένων,
- ταξινόμηση,
- παλινδρόμηση,
- κανόνες συσχέτισης,
- γραφική παράσταση όλων των παραπάνω
- δημιουργία μοντέλων αναπαράστασης δεδομένων με την χρήση μεθοδολογίας Μηχανικής Μάθησης

Εικόνα 6 Weka (https://www.cs.waikato.ac.nz/ml/weka/)

Είναι το ιδανικό περιβάλλον για ανάπτυξη νέων μηχανισμών μάθησης. Είναι ανοικτού κώδικα κάτω από το πλαίσιο GNU και βρίσκεται στην ηλεκτρονική διεύθυνση (http://www.cs.waikato.ac.nz/ml/weka). Η εφαρμογή πήρε το όνομά της από το

πτηνό Weka το οποίο δεν πετάει και βρίσκεται μόνο στα νησιά της Νέας Ζηλανδίας. Η χρησιμοποιούμενη έκδοση του weka είναι 3.8.0.

Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα απλό αρχείο ή συσχέτιση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό των χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά και κάποιοι άλλοι τύποι χαρακτηριστικών υποστηρίζονται επίσης). Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων, χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφονται από ένα ερώτημα βάσης δεδομένων. Δεν είναι ικανό για εξόρυξη από πολυσχεσιακές βάσεις δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων της βάσης δεδομένων σε έναν πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka. Άλλη μια σημαντική περιοχή που δεν καλύπτεται προς το παρόν από τους αλγορίθμους που περιλαμβάνονται στο Weka είναι η μοντελοποίηση αλληλουχιών.

Εκτός από τους δημοφιλέστερους αλγορίθμους ταξινόμησης, περιλαμβάνει διάφορα σχήματα αριθμητικής πρόβλεψης (π.χ. μοντέλα γραμμικής παλινδρόμησης, μοντέλα μάθησης βασισμένα σε στιγμιότυπα, κ.α.), σχήματα Μετά –Μάθησης (π.χ. Bagging, Boosting, Stacking κ.α.), μεθόδους ομαδοποίησης (clustering), καθώς και μια πληθώρα εργαλείων για την προεπεξεργασία σωμάτων δεδομένων, όπως μεθόδους διακριτοποίησης και φιλτραρίσματος αλλά και για την αξιολόγηση των παραγόμενων μοντέλων.

2

Εξόρυξη Δεδομένων - Data Mining

Ο όρος εξόρυξη δεδομένων εισήχθη το 1990, όμως η έννοια της εξόρυξης δεδομένων έχει τις ρίζες της εδώ και πολλά χρόνια. Η εξόρυξη δεδομένων έφθασε στη σημερινή της μορφή αφού πρώτα πέρασε από διάφορες φάσεις έρευνας και μελετών Η ανάπτυξη αυτή ξεκίνησε όταν τα δεδομένα των επιχειρήσεων άρχισαν να αποθηκεύονται στους υπολογιστές. Η διαδικασία συνεχίστηκε με τις προόδους στην τεχνολογία των υπολογιστών συμπεριλαμβανομένης και της αποθήκευσης δεδομένων, της επεξεργαστικής ισχύος, των νέων λογισμικών και των αλγορίθμων (Han etal, 2011). Καθοριστικό επίσης ήταν για την εξέλιξη της, ότι στο σημερινό ανταγωνιστικό κόσμο των πληροφοριών, όλοι προσπαθούν να κάνουν την καλύτερη χρήση των δεδομένων τους για να κάνουν τις επιχειρήσεις τους να έχουν κέρδος και επιτυχία.

Η συλλογή και αποθήκευση δεδομένων σε υπολογιστές, ταινίες και δίσκους ξεκίνησε το 1960. Το επόμενο εξελικτικό βήμα στην εξόρυξη δεδομένων συνέβη κατά τη διάρκεια της δεκαετίας του 1980 με την εισαγωγή των σχεσιακών βάσεων δεδομένων και τη δομημένη γλώσσα ερωτημάτων. Αυτό βοήθησε τους χρήστες να μάθουν περισσότερα σχετικά με τα δεδομένα που αποθηκεύονται σε σχεσιακές βάσεις δεδομένων και χρησιμοποιούν δομημένη γλώσσα με ερωτήματα. Έτσι τα δεδομένα έγιναν διαθέσιμα σε επίπεδο ρεκόρ για την τότε εποχή. Στην συνέχεια, η εξέλιξη της αποθήκευσης δεδομένων συνέβη κατά τη διάρκεια της δεκαετίας του 1990, όπου γίνεται η αναλυτική επεξεργασία και οι πολυδιάστατες βάσεις δεδομένων συνέβαλαν στην αύξηση της αποθήκευσης δεδομένων. Εάν αναλύσουμε κάθε βήμα της εξέλιξης, είναι πολύ σαφές ότι κάθε βήμα βασίζεται στο προηγούμενο βήμα.

Κατά τη διάρκεια της δεκαετίας του 1960 τα δεδομένα δεν ήταν κάτι σύνθετο. Πλέον όμως η κατάσταση είναι εντελώς διαφορετική. Τα επιχειρηματικά δεδομένα έχουν μετατραπεί σε επιχειρηματικές πληροφορίες και είναι αρκετά ισχυρά ώστε να απαντήσουν σε πολλές πολύπλοκες επιχειρηματικές ερωτήσεις ακόμη και να προβλέψουν το μέλλον της επιχείρησης. Η ανάπτυξη των βάσεων δεδομένων συμβαίνει σε τεράστια ποσοστά που απαιτούν μεθόδους για να παρέχουν χρήσιμες πληροφορίες από αυτό το τεράστιο όγκο δεδομένων. Η τεχνολογία εξόρυξης δεδομένων έχει περάσει με την πάροδο των χρόνων από την ανάπτυξη της διαδικασίας σε τρεις διαφορετικούς τομείς που συνέβαλαν στην τρέχουσα μορφή της. Οι τομείς αυτοί είναι η στατιστική, η τεχνητή νοημοσύνη και η μηχανική μάθηση (Harrington, 2012).

Data Mining – Ορισμός

Εξόρυξη δεδομένων ή αλλιώς data mining ονομάζεται η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, μη προφανούς, άγνωστης μέχρι τώρα και δυνητικά ωφέλιμης γνώσης από μεγάλες βάσεις δεδομένων. Για να επιτευχθεί αυτό γίνεται χρήση αλγορίθμων ομαδοποίησης, κατηγοριοποίησης καθώς και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Εναλλακτικά, θεωρείται και ως η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από σύνολο ή βάσεις δεδομένων μεγάλου μεγέθους. Είναι ένα σημαντικό εργαλείο, το οποίο βοηθάει τον άνθρωπο μέσα από τη διαδικασία της εξερεύνησης και της ανάλυσης πολλών δεδομένων με αυτόματα ή ημιαυτόματα μέσα. Οι νέες πληροφορίες που προκύπτουν μπορούν να χρησιμοποιηθούν σε διάφορους τομείς, όπως για παράδειγμα στην υποστήριξη της λήψης αποφάσεων, στις προβλέψεις και στις εκτιμήσεις σημαντικών επιχειρηματικών αποφάσεων. Γενικά, έχουν χρησιμότητα σε τομείς οι οποίοι μπορούν να βοηθήσουν μια επιχείρηση να αποκτήσει και να διατηρήσει σημαντικό ανταγωνιστικό πλεονέκτημα (Ahmed, 2004) .

Αίτια Δημιουργίας Data Mining

Η συνεχής πρόοδος της τεχνολογίας στον τομέα της πληροφορικής παρέχει τη δυνατότητα αποθήκευσης τεράστιου όγκου δεδομένων σε αρχεία, βάσεις δεδομένων, το διαδίκτυο και άλλα μέσα. Οι περισσότερες επιχειρήσεις πλέον χρησιμοποιούν την δυνατότητα αυτή και καταγράφουν το μεγαλύτερο πλήθος των πληροφοριών τους σε ηλεκτρονική μορφή. Αυτό έχει σαν αποτέλεσμα τον διπλασιασμό του όγκου αποθηκευμένων δεδομένων (Μαστρογιάννης, 2009). Έτσι δημιουργήθηκε η ανάγκη για την ανάλυση και την ερμηνεία της σημαντικής πληροφορίας που υπάρχει στις αποθήκες δεδομένων (data warehouse) των επιχειρήσεων.

Στόχος Της Εξόρυξης Δεδομένων

Στόχος της εξόρυξης δεδομένων είναι να εξαχθεί πληροφορία η οποία θα βοηθήσει να παρθούν κατάλληλες αποφάσεις. Για να γίνει αυτό χρειάζεται να υπάρξει η περιγραφή και η πρόβλεψη στα σύνολα δεδομένων. Σκοπός της πρόβλεψης είναι ο υπολογισμός της μελλοντικής αξίας ή συμπεριφοράς των μεταβλητών που μας ενδιαφέρουν και εξαρτώνται από τη συμπεριφορά άλλων μεταβλητών. Σκοπός της περιγραφής είναι η ανακάλυψη προτύπων και η αναπαράσταση των δεδομένων μιας πολύπλοκης βάσης δεδομένων με κατανοητό και αξιοποιήσιμο τρόπο. Ανάλογα με τις εφαρμογές εξόρυξης διαφοροποιείται η σημαντικότητα αυτών των δύο. Η περιγραφή είναι πιο σημαντική από τη πρόβλεψη όσων αφορά την εξόρυξη γνώσης, ενώ η πρόβλεψη είναι πιο σημαντική για την αναγνώριση προτύπων και την εφαρμογή μηχανικής μάθησης (Han etal, 2011).

Μια βασική ανάγκη που δημιουργείται είναι η αξιοποίηση του ασύλληπτα μεγάλου ποσού δεδομένων που είναι διαθέσιμο και η ανακάλυψη γνώσεων που κρύβονται πίσω από αυτά. Η Εξόρυξη Γνώσης ή Ανακάλυψη Γνώσης από Βάσεις Δεδομένων ή Εξόρυξη Δεδομένων (Data Mining-DM) (Klösgen & Zytkow, 2002) είναι η εξεύρεση μιας ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

2.1 Εξόρυξη γνώσης από κείμενο

Μια υποκατηγορία της εξόρυξης γνώσης είναι η εξόρυξη κειμένου (Text Mining-TM), η οποία ορίζεται σαν μια διαδικασία εξαγωγής με αυτόματο τρόπο νέας, έγκυρης, και χρήσιμης γνώσης από διαφορετικούς γραπτούς πόρους, καθώς επίσης και η όσο το δυνατόν καλύτερη οργάνωση αυτής της νέας γνώσης-πληροφορίας για την όποια μελλοντική αναφορά (Hearst, 1992). Μια υποκατηγορία της εξόρυξης κειμένου είναι η εξόρυξη γνώμης (opinion mining-OM) ή ανάλυση συναισθήματος (sentimental analysis) η οποία στοχεύει να εξάγει γνωρίσματα και συστατικά ενός αντικειμένου που έχει σχολιαστεί και καθορίζει αν το σχόλιο είναι θετικό, αρνητικό ή ουδέτερο (Liu, 2012).

Εφαρμόζοντας τις τεχνικές εξόρυξης γνώσης στα Κοινωνικά Δίκτυα μπορούμε να ανακαλύψουμε ενδιαφέρουσες πλευρές της ανθρώπινης συμπεριφοράς και της ανθρώπινης αλληλεπίδρασης, να βελτιώσουμε την αντίληψη που έχουν οι άνθρωποι σχετικά με ένα θέμα, να προσ-

διορίσουμε ομάδες ανθρώπων ανάμεσα στις μάζες του πληθυσμού, να μελετήσουμε ομάδες που αλλάζουν με το χρόνο, να βρεθούν άνθρωποι με επιρροή, ή ακόμα και να γίνει η σύσταση ενός προϊόντος ή μιας δραστηριότητας σε ένα άτομο.

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση (machine learning) είναι μια περιοχή της τεχνητής νοημοσύνης η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν». Με τη μηχανική μάθηση καθίσταται εφικτή η κατασκευή προσαρμόσιμων (adaptable) προγραμμάτων υπολογιστών τα οποία λειτουργούν με βάση την αυτοματοποιημένη ανάλυση συνόλων δεδομένων και όχι τη διαίσθηση των μηχανικών που τα προγραμμάτισαν. Το πεδίο της μηχανικής μάθησης ασχολείται με το ζήτημα της δημιουργίας προγραμμάτων ηλεκτρονικών υπολογιστών που βελτιώνονται αυτόματα, αποκτώντας εμπειρία (Harrington, 2012). Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με την υπολογιστική στατιστική, μια επιστημονική αρχή η οποία επίσης ειδικεύεται στην πρόβλεψη αποφάσεων (prediction-making) (Bishop, 2006), αφού και τα δύο πεδία χρησιμοποιούν ανάλυση δεδομένων, επικαλύπτονται όμως και με την εξόρυξη δεδομένων (data mining), ωστόσο η δεύτερη εστιάζει περισσότερο στη διερευνητική ανάλυση δεδομένων (Lyons, 1981).

Η μηχανική μάθηση έχει σαν σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν. Οι μηχανές αυτές, αξιοποιώντας προηγούμενη γνώση και εμπειρία μπορούν να μαθαίνουν και να βελτιώνουν την απόδοσή τους σε κάποιους τομείς. Υπάρχουν πάρα πολλοί αλγόριθμοι μηχανικής μάθησης που δίνουν ικανοποιητικά αποτελέσματα ανάλογα με το είδος της εφαρμογής που θα τους χρησιμοποιήσουμε. Στην περίπτωση της εξόρυξης γνώσης από κείμενο και πιο συγκεκριμένα στην εξόρυξη γνώμης και τη συναισθηματική ανάλυση κειμένου, την καλύτερη απόδοση έχουν οι αλγόριθμοι που ανήκουν στις κατηγορίες των Επαγωγικών Αλγορίθμων Μηχανικής Μάθησης, και ειδικότερα στις κατηγορίες των Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machines - SVMs) και των Bayesian μεθόδων (Aggarwal & Zhai, 2012), (Liu, 2012).

Το 1959, ο πρωτοπόρος σχεδιαστής παιχνιδιών Arthur Samuel όρισε ως μηχανική μάθηση: «Το πεδίο μελέτης που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί γι' αυτό το σκοπό» (Samuel, 1959). Το 1997 ο Tom M. Mitchell έδωσε ένα πιο επίσημο ορισμό ο οποίος χρησιμοποιείται ευρέως: «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E» (Mitchell, 1997).

Η μηχανική μάθηση είναι ο επιστημονικός κλάδος που διερευνά την δημιουργία και μελέτη των αλγορίθμων που μπορούν να εκπαιδευτούν από δεδομένα. Τέτοιοι αλγόριθμοι λειτουργούν μέσω της δημιουργίας ενός μοντέλου από τυχαίες εισόδους (example inputs) χρησιμοποιώντας το για προβλέψεις ή αποφάσεις, αντί να ακολουθήσουν αυστηρές στατικές οδηγίες.

2.2.1 Τεχνικές Μηχανικής Μάθησης

Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη (supervised learning), ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένα εισόδους σε γνωστές, επιθυμητές εξόδους (σύνολο εκπαίδευσης), με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο (σύνολο ελέγχου). Στον υπολογιστή παρουσιάζονται παραδείγματα εισόδων με τα αντίστοιχα επιθυμητά αποτελέσματα (κατηγορίες) και ο στόχος είναι να μάθει ένα γενικό κανόνα που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους, καταγράφοντας δομικά στοιχεία των εισόδων αυτών (Radovanović & Ivanović, 2008).

Στην επιβλεπόμενη μάθηση η εκπαίδευση των ταξινομητών γίνεται με επισημειωμένα δείγματα (labeled samples). Οι Naïve Bayes, Maximum Entropy και Support Vector Machines (SVM) είναι οι πιο γνωστοί ταξινομητές αυτής της κατηγορίας (Kotsiantis etal, 2007).

Μη Επιβλεπόμενη Μάθηση

Στην ανεπίβλεπτη μάθηση ή μάθηση χωρίς επίβλεψη (unsupervised learning), ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει επιθυμητές εξόδους για το σύνολο εκπαίδευσης (Radovanović & Ivanović, 2008).

Η μη επιβλεπόμενη μάθηση είναι μια διεργασία μηχανικής μάθησης κατά την οποία εξάγεται μια συνάρτηση για να περιγράψει κρυφές δομές από δεδομένα χωρίς ετικέτες. Εφόσον τα παραδείγματα που δίνονται στον μηχανισμό εκμάθησης είναι άγνωστης κατηγορίας, δεν υπάρχει λάθος-σφάλμα καθώς και ούτε κάποιο σήμα αξιολόγησης ώστε να γίνει εκτίμηση της πιθανής λύσης. Αυτό είναι άλλωστε και που ξεχωρίζει τις μη επιβλεπόμενες διαδικασίες από τις επιβλεπόμενες τεχνικές ανάλυσης. Η μη επιβλεπόμενη μάθηση είναι στενά συνδεδεμένη με το πρόβλημα της εκτίμησης πυκνότητας στην στατιστική. Ωστόσο η μη επιβλεπόμενη μάθηση εμπλέκει και πολλές άλλες τεχνικές σύμφωνα με τις οποίες ψάχνει να συνοψίσει και να εξηγήσει χαρακτηριστικά κλειδιά των δεδομένων. Πολλές μέθοδοι που αναμιγνύονται σε αυτές τις τεχνικές βασίζονται σε μεθόδους εξόρυξης γνώσης που χρησιμοποιούνται κατά την προεπεξεργασία των δεδομένων. Την πιο σημαντική προσέγγιση της μη επιβλεπόμενης μάθησης αποτελεί η συσταδοποίηση ή αλλιώς clustering. Στο πρόβλημα της συσταδοποίησης

μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις και χρειαζόμαστε κάποιον αλγόριθμο ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες (clusters).Οι συστάδες που δημιουργούνται θέλουμε να διαχωρίζουν όσο το δυνατόν πιο ορθά τα δεδομένα. Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο να είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ' ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας (Radovanović & Ivanović, 2008).

2.2.2 Μοντέλα Επιβλεπόμενης Μάθησης

Naïve Bayes

Στην μηχανική μάθηση, οι ταξινομητές Naïve Bayes αποτελούν μια οικογένεια από απλούς πιθανοτικούς ταξινομητές που βασίζονται στην εφαρμογή του θεωρήματος Bayes (Thomas Bayes 1702-1761) με αυστηρή (αφελή–naïve) υπόθεση ανεξαρτησίας ανάμεσα στα χαρακτηριστικά.

Ο αλγόριθμος Naïve Bayes έχει μελετηθεί εκτενώς στο παρελθόν σχεδόν από το 1950. Αποτελεί μια απλή τεχνική και κατασκευή ταξινομητών, μοντέλων δηλαδή που ορίζουν ετικέτες κλάσης σε οντότητες προβλημάτων που αναπαρίστανται από διανύσματα με τιμές χαρακτηριστικών, όπου οι ετικέτες αυτές καθορίζονται από ένα πεπερασμένο σύνολο. Δεν είναι ένας μεμονωμένος αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζονται σε ένα κοινό πρότυπο.

Οι Naive Bayes ταξινομητές υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού δεδομένης της μεταβλητής κλάσης. Υποθέτει ένα υποκείμενο πιθανοτικό μοντέλο και μας επιτρέπει να καταγράψουμε την αβεβαιότητα σχετικά με το μοντέλο με τρόπο που βασίζεται στην αρχή, καθορίζοντας τις πιθανότητες των αποτελεσμάτων. Μπορεί να λύσει προβλήματα διάγνωσης και πρόβλεψης (Mc Callum & Nigam, 1998).

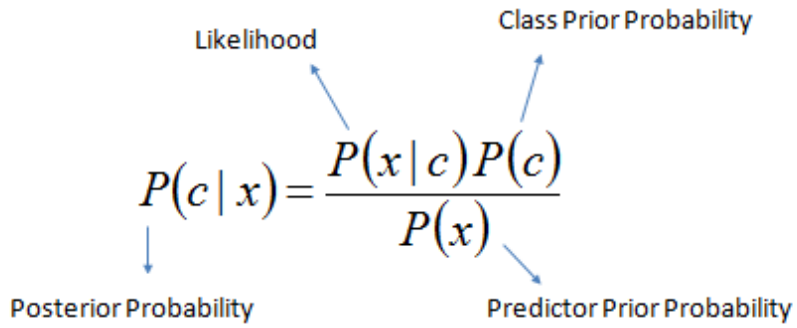
Έχουμε άλλες εναλλακτικές λύσεις όταν αντιμετωπίζουμε προβλήματα με το NLP, όπως το Vector Machine Support (SVM) και τα νευρωνικά δίκτυα. Ωστόσο, ο απλός σχεδιασμός των ταξινομητών Naive Bayes τους καθιστά πολύ ελκυστικούς για αυτούς τους ταξινομητές. Επιπλέον, έχουν αποδειχθεί ότι είναι γρήγορες, αξιόπιστες και ακριβείς σε πολλές εφαρμογές του NLP.

$$P(A_i \mid B) = \frac{P(A_i)\,P(B \mid A_i)}{\sum_{i=1}^{k} P(A_i)\,P(B \mid A_i)}$$

Εικόνα 7 Θεώρημα του Bayes

Το **Θεώρημα του Bayes** δεν είναι ούτε ακατανόητο, ούτε δύσμορφο όπως οι περισσότεροι τύποι. Με λίγη μαθηματική διαίσθηση, αυτό που διατύπωσε και απέδειξε ο Άγγλος μαθηματικός στις αρχές του 18ου αιώνα, γίνεται πλήρως αντιληπτό.

ή αλλιώς:



Εικόνα 8 Θεώρημα του Bayes (2)

Ο Naive Bayes έχει τον εξής τύπο:

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Εικόνα 9 Naive Bayes

όπου:

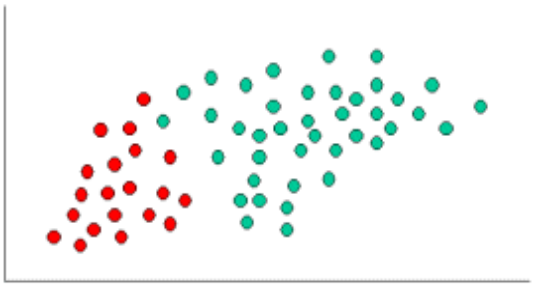
- P (c|x) είναι η δευσμευμένη πιθανότητα της τάξης (στόχου) που δίνεται πρόβλεψη (χαρακτηριστικό).
- P(c) είναι η προηγούμενη πιθανότητα της κλάσης.
- P (x|c) είναι η δεσμευμένη πιθανότητα της προβλεπόμενης κατηγορίας προγνωστικού.
- P (x) είναι η προηγούμενη πιθανότητα του προγνωστικού.

Το θεώρημα του Bayes ανήκει στην μεγάλη οικογένεια της «Θεωρίας των Πιθανοτήτων». Δεν είναι όμως ένα απλό μέλος της. Ο Harold Jeffreys, ένας από τους σπουδαιότερους Άγγλους μαθηματικούς, που ασχολήθηκε με τις «μπεϊζιανές» πιθανότητες και εισήγαγε την έννοια του αλγορίθμου Bayes, είχε δηλώσει πως ο συγκεκριμένος τύπος «είναι στη Θεωρία Πιθανοτήτων όπως αντίστοιχα το Πυθαγόρειο Θεώρημα στη Γεωμετρία» (Jeffreys, 1998).

Στον μη-αφελή τρόπο του Bayes, εξετάζουμε συνολικά τις προτάσεις, οπότε μόλις η πρόταση δεν εμφανιστεί στο εκπαιδευτικό σύνολο, θα έχουμε μια μηδενική πιθανότητα, καθιστώντας δύσκολη την περαιτέρω υπολογισμό. Ενώ για το Naive Bayes, υπάρχει μια υπόθεση ότι κάθε

λέξη είναι ανεξάρτητη η μία από την άλλη. Τώρα, εξετάζουμε μεμονωμένες λέξεις σε μια πρόταση, αντί για ολόκληρη την πρόταση.

Ο Bayes ασχολήθηκε με τις **δεσμευμένες πιθανότητες** ή αλλιώς τις «υπό συνθήκη» πιθανότητες. Ο νόμος του, μέσα στην συνοπτική του γραφή, κρύβει μυστικά τα οποία βασανίζουν από μαθηματικούς και φυσικούς, μέχρι γιατρούς και φιλοσόφους. Γράφοντας το **«P(A|B) = P(B|A)*P(A)/ P(B)»** ο Άγγλος μαθηματικός δεν περίμενε πως ο τύπος του, μετά από 300 χρόνια, θα έχει δημιουργήσει ολόκληρα φιλοσοφικά ερωτήματα. Πως θα έχει επηρεάσει την τεχνητή νοημοσύνη ή πως θα θέσει τις βάσεις για την εξέλιξη κλάδων όπως η κβαντομηχανική. Ο συγκεκριμένος τύπος ουσιαστικά σημαίνει: «Η πιθανότητα να συμβεί το ενδεχόμενο Α, δεδομένου του Β, ισούται με την πιθανότητα να συμβεί το Β δεδομένου του Α, επί την πιθανότητα να συμβεί το Α, διά την πιθανότητα να συμβεί το Β».
Η τεχνική Classifier Naive Bayes βασίζεται στο λεγόμενο Bayesian θεώρημα και είναι ιδιαίτερα κατάλληλη όταν η διαστατικότητα των εισροών είναι υψηλή. Παρά την απλότητα της, οι Naive Bayes μπορούν συχνά να ξεπεράσουν τις πιο εξελιγμένες μεθόδους ταξινόμησης (Καρποδίνης, 2016).



Εικόνα 10 practical Naive Bayes (Καρποδίνης, 2016)

Όπως υποδεικνύεται στην εικόνα, τα αντικείμενα μπορούν να ταξινομηθούν είτε ως ΠΡΑΣΙΝΟ είτε ΚΟΚΚΙΝΟ. Το καθήκον του αλγορίθμου είναι να κατατάξει νέες περιπτώσεις κατά την άφιξή τους, δηλαδή να αποφασίσουμε σε ποια ετικέτα κατηγορίας ανήκουν, με βάση τα εξερχόμενα αντικείμενα.

Naïve Bayes Multinomial αλγόριθμος

Οι κατηγοριοποιητές Naive Bayes έχουν εφαρμοστεί επιτυχώς σε πολλούς τομείς, ιδιαίτερα στην επεξεργασία φυσικής γλώσσας (NLP). Το μοντέλο αυτό καθορίζει ότι ένα έγγραφο αντιπροσωπεύεται από το σύνολο των συμβάντων λέξης από το έγγραφο. Η σειρά των λέξεων έχει χαθεί, ωστόσο, ο αριθμός των εμφανίσεων κάθε λέξης στο έγγραφο καταγράφεται. Κατά

τον υπολογισμό της πιθανότητας ενός εγγράφου, πολλαπλασιάζεται η πιθανότητα των λέξεων που εμφανίζονται. Εδώ μπορούμε να κατανοήσουμε τα γεγονότα κάθε λέξης που είναι τα γεγονότα και το έγγραφο να είναι η συλλογή των λέξεων συμβάντων. Ονομάζουμε αυτό το μοντέλο πολυωνυμικών γεγονότων. Ο Naïve Bayes Multinomial αλγόριθμος λειτουργεί σε σύνολα κειμένων όπου κάθε στιγμιότυπο x αποτελείται από τις τιμές χαρακτηριστικών και η συνάρτηση στόχος *f(x)* μπορεί να πάρει οποιαδήποτε τιμή από ένα προκαθορισμένο πεπερασμένο σύνολο. Η ταξινόμηση αγνώστων στιγμιότυπων περιλαμβάνει τον υπολογισμό της πιο πιθανής τιμής στόχου και ορίζεται ως εξής (McCallum & Nigam, 1998):

$$\mathrm{P}(d_i|c_j;\theta) = \mathrm{P}(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{\mathrm{P}(w_t|c_j;\theta)^{N_{it}}}{N_{it}!}$$

Εικόνα 11 Multinomial Naive Bayes

Support Vector Machine Αλγόριθμος (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM), είναι μοντέλα επιβλεπόμενης μηχανικής μάθησης, που χρησιμοποιούνται τόσο για την κατηγοριοποίηση κειμένων, όσο και για την κατηγοριοποίηση συναισθήματος. Οι Μηχανές Διανυσμάτων Υποστήριξης αποτελούν μη πιθανοτικούς γραμμικούς ταξινομητές, καθώς συνδυάζουν τα γραμμικά μοντέλα, με τεχνικές μάθησης σε στιγμιότυπα. Τα μοντέλα SVM, επιλέγουν έναν μικρό πλήθος στιγμιοτύπων εκπαίδευσης, από κάθε κατηγορία, που ονομάζονται διανύσματα υποστήριξης (support vectors), τα οποία ορίζουν το μέγιστο περιθώριο (margin), μεταξύ των δύο κατηγοριών. Τα διανύσματα υποστήριξης αξιοποιούνται για την κατασκευή μίας γραμμικής συνάρτησης διάκρισης (discriminant function), η οποία διαχωρίζει τα δεδομένα με βέλτιστο τρόπο (Cortes & Vapnik, 1995).

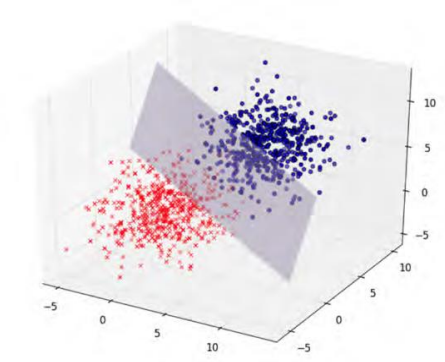
Τα μοντέλα κατηγοριοποίησης SVM, είναι από τα δημοφιλέστερα στην κατηγοριοποίηση συναισθήματος, λόγω της αποτελεσματικότητας, της ταχύτητας, και της ικανότητάς τους να παράγουν μη γραμμικά υπερεπίπεδα απόφασης, καθιστώντας εφικτή την επίλυση προβλημάτων, που δεν είναι δυνατό να επιλυθούν με γραμμικά μοντέλα.

Τα μοντέλα SVM, εκτός από την εκτέλεση γραμμικής κατηγοριοποίησης, μπορούν να πραγματοποιήσουν μη γραμμική κατηγοριοποίηση, εφαρμόζοντας το τέχνασμα του πυρήνα (kernel trick), μετασχηματίζοντας το χώρο των χαρακτηριστικών του προβλήματος, σε έναν χώρο μεγαλύτερης διάστασης.

Ο αλγόριθμος Support Vector Machines (SVM) είναι ένα σύνολο μεθόδων εκμάθησης που χρησιμοποιούνται για προβλήματα ταξινόμησης και παλινδρομικής ανάλυσης. Η κύρια ιδέα του SVM είναι να κατασκευαστεί ένα υπερεπίπεδο, έτσι ώστε η απόσταση διαχωρισμού με-

ταξύ των θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Τα διανύσματα των πιο κοντινών στοιχείων στο υπερεπίπεδο αυτό είναι και τα υποστηρικτικά διανύσματα (support vectors). Πιο αναλυτικά, καθώς ο SVM είναι ένας αλγόριθμος large-margin και όχι πιθανοτικός η βασική ιδέα πέρα από την διαδικασία εκπαίδευσης του αλγορίθμου είναι να βρεθεί το υπερεπίπεδο μεγίστου περιθωρίου (maximum margin hyperplane) το οποίο αναπαρίσταται από το διάνυσμα και το οποίο όχι μόνο διαχωρίζει τα διανύσματα των εγγράφων της μιας κλάσης από αυτά της άλλης αλλά ταυτόχρονα φροντίζει ο διαχωρισμός (ή το όριο) να είναι και όσο μεγαλύτερος γίνεται. Αυτό καταλήγει να είναι λοιπόν ένα πρόβλημα βελτιστοποίησης (Cortes & Vapnik, 1995).

Η ταξινόμηση των δεδομένων είναι μια κοινή ανάγκη στο πεδίο της μηχανικής μάθησης. Ας υποθέσουμε ότι δίνονται κάποια σημεία δεδομένων που ανήκουν στα δύο σύνολα και ο στόχος είναι να αποφασίσουμε σε πιο σύνολο θα μπει ένα νέο στιγμιότυπο. Στην περίπτωση των SVM ένα στιγμιότυπο θεωρείται σαν ένα διάνυσμα p-διαστάσεων και θέλουμε να ξέρουμε αν μπορούμε να χωρίσουμε αυτά τα σημεία με ένα (p-1)-διάστατο υπερεπίπεδο που ονομάζεται γραμμικός ταξινομητής.



Εικόνα 12 Διαχωρισμός συμβάντων με τον SVM στον χώρο (Καρποδίνης, 2016)

Πρόκειται για μια προηγμένη τεχνική κατηγοριοποίησης (classification) της Υπολογιστικής Νοημοσύνης. Η SVM ανήκει και αυτή στην κατηγορία των μεθόδων της επιβλεπόμενης μάθησης και μπορεί εκτός από κατηγοριοποίηση να χρησιμοποιηθεί και για περιπτώσεις παλινδρόμησης (regression). Τα τελευταία χρόνια έχει αναπτυχθεί σε μια πολύ ενεργητική ερευνητική περιοχή και έχει εφαρμοστεί σε πολλών ειδών επιστημονικά προβλήματα. Τα αντίστοιχα της SVM στα Νευρωνικά Επιπρόσθετα, οι Μηχανές Διανυσμάτων Υποστήριξης έχουν την ικανότητα να επεκτείνουν τη χρήση τους και στο διαχωρισμό κλάσεων οι οποίες δεν μπορούν να διαχωριστούν μέσω ενός γραμμικού ταξινομητή. Σε αυτήν την περίπτωση, οι συντεταγμένες των αντικειμένων αντιστοιχίζονται στο χώρο, χρησιμοποιώντας μη γραμμικές συναρτήσεις. Ο χώρος στον οποίο κάθε αντικείμενο αντιστοιχίζεται είναι χώρος υψηλού αριθμού δια-

στάσεων στον οποίο δύο κλάσεις μπορούν να διαχωριστούν με ένα γραμμικό ταξινομητή (Carbonell, 1979). Πολύ σημαντική είναι και η επέκταση των κλασσικών SVM και η μετατροπή τους σε πρόβλημα παλινδρόμησης όπως προαναφέρθηκε. Ένα χαρακτηριστικό τέτοιο παράδειγμα αποτελεί η πρόβλεψη χρηματοοικονομικών δεικτών. Αυτό έγινε εφικτό από την εισαγωγή της ε − ευαίσθητης συνάρτησης απωλειών (ε − sensitive loss function) από τον Vapnick, η οποία καθιέρωσε τα διανύσματα μη-γραμμικής παλινδρόμησης. Παρά το γεγονός ότι οι SVM και τα SVR έχουν ήδη εφαρμοστεί σε πολλά χρηματοοικονομικά προβλήματα, ακόμα δεν έχουν καθιερωθεί ευρέως στα πεδία αυτά. Πρόκειται για πολλά υποσχόμενες τεχνικές οι οποίες, όμως, έως τώρα δεν έχουν καταφέρει να υπερκεράσουν τους περιορισμούς της "κατάρας" της ύπαρξης πολλών διαστάσεων (curse of dimensionality), η οποία θα εξηγηθεί παρακάτω. Και για το λόγο αυτό σε ορισμένες περιπτώσεις τα απλά Νευρωνικά Δίκτυα έχουν αποδειχθεί, για την ώρα, αποδοτικότερα στην επίλυση προβλημάτων του πραγματικού κόσμου (Wilks & Bien, 1983).

2.3 Μεταμάθηση

Η meta learning - όπως είναι ο διεθνώς γνωστός όρος- περιγράφηκε για πρώτη φορά απ' τον Donald Maudsley το 1979. Το 1985, ο John Biggs αναφέρθηκε εκτενέστερα στη μεταμάθηση, ορίζοντάς την ως την ανάπτυξη της αυτεπίγνωσης σε κάθε μαθητή ή εκπαιδευόμενο, η οποία τον βοηθά να αποκτήσει έλεγχο πάνω στις συνήθειες που αφορούν στο πώς αντιλαμβάνεται, ερευνά, μαθαίνει και τελικά αναπτύσσεται ο ίδιος (Biggs, 1985).

Η αποδοτικότητα ενός μοντέλου που παράγεται από αλγορίθμους μάθησης καθορίζεται τόσο από το μέγεθος και την ποιότητα του σώματος εκπαίδευσης, όσο και από την καταλληλότητα του χρησιμοποιούμενου αλγορίθμου μάθησης, παράγοντες οι οποίοι είναι κατά γενική ομολογία δύσκολο να προσδιορισθούν. Μια εναλλακτική προσέγγιση επιδιώκει να αυξήσει την αξιοπιστία ενός συστήματος μηχανικής μάθησης επιστρατεύοντας την «εμπειρία» περισσοτέρων του ενός μοντέλων – «ειδικών» (experts), από τον κατάλληλο συνδυασμό των οποίων προκύπτει η τελική απόφαση αναφορικά με ένα άγνωστο στιγμιότυπο του προβλήματος. Σε αυτή την κατηγορία οι διαφορετικοί ταξινομητές προκύπτουν από εκπαίδευση ενός αλγορίθμου σε διαφορετικά υποσύνολα των διαθέσιμων δεδομένων. Αυτή η κατηγορία μεθόδων εκμεταλλεύεται την αστάθεια που παρουσιάζουν διάφοροι αλγόριθμοι μάθησης, την υπερευαισθησία δηλαδή που επιδεικνύουν στις μικρές μεταβολές των δεδομένων της εισόδου. Σκοπός είναι η διαδοχική δημιουργία μοντέλων ικανών να αλληλοσυμπληρώνονται, υπό την έννοια ότι το κάθε ένα αποδίδει τα μέγιστα σε ένα υποσύνολο του σώματος εκπαίδευσης το οποίο τα υπόλοιπα δεν μπορούν να αξιοποιήσουν αποτελεσματικά. Όπως είναι αναμενόμενο, η μέθοδος αποδίδει ιδιαίτερα καλά για ασταθείς αλγορίθμους μάθησης-αλγορίθμους οι οποίοι παρά-

γουν ταξινομητές αρκετά διαφορετικούς με μικρές αλλαγές του συνόλου εκπαίδευσης. Τα δέντρα απόφασης και τα νευρωνικά δίκτυα είναι παραδείγματα ασταθών αλγορίθμων, ενώ αντίθετα ο K-NN και ο απλός Bayes ταξινομητής (NB) είναι γενικά πολύ σταθεροί. Στην περιοχή αυτή της μηχανικής μάθησης η οποία ονομάζεται Μετά-μάθηση (meta-learning) συγκαταλέγονται οι ακόλουθες μεθοδολογίες συνδυασμού μοντέλων:

Bagging: (Breiman, 1996) Η πιο απλή μέθοδος αυτής της κατηγορίας λέγεται "σάκιασμα" (bagging ή Bootstrap Aggregation). Η μέθοδος αυτή συνίσταται στην παραγωγή ενός αριθμού μοντέλων (υπό-ταξινομητών), προερχόμενων από έναν κοινό αλγόριθμο μάθησης, χρησιμοποιώντας όμως διαφορετική διαμέριση του σώματος εκπαίδευσης (επιλογή με επανατοποθέτηση) για κάθε ένα εξ' αυτών. Η σχηματική αναπαράσταση της μεθόδου παρατίθεται στην Εικόνα 6. Για τη λήψη της τελικής απόφασης ακολουθείται συνήθως η πλειοψηφική λογική. Δηλαδή η τελική απόφαση του συστήματος συμπίπτει με την απόφαση της πλειοψηφίας. Παρόμοια τεχνική δανεισμένη από την τεχνική της στατιστικής επικύρωσης είναι η επιτροπή διασταυρωμένης επικύρωσης (cross-validated committees), όπου τα υποσύνολα των δεδομένων στα οποία εκπαιδεύονται οι ταξινομητές καθορίζονται από τη γνωστή μέθοδο διασταυρωμένης επικύρωσης.

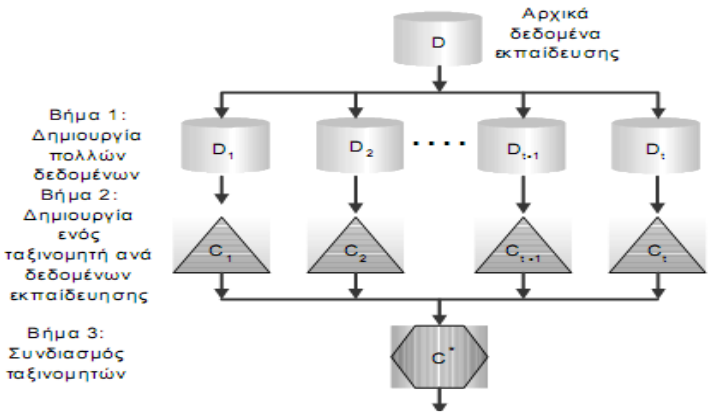
Boosting: (Breiman, 1996), (Freund & Schapire, 1996) Παρόμοια διαδικασία με την προηγούμενη εφαρμόζεται και στην περίπτωση της Προώθησης (Boosting), με τη διαφορά ότι τα μοντέλα που συστήνουν την επιτροπή των ειδικών παράγονται διαδοχικά, προκειμένου κάθε καινούργιο μοντέλο να επηρεάζεται άμεσα από την απόδοση των προηγούμενων του, επιδιώκοντας να αποφύγει λανθασμένες αποφάσεις που ενδεχομένως προηγήθηκαν. Αποτέλεσμα της διαδικασίας αποτελεί η αύξηση του βάρους των λανθασμένων στιγμιότυπων και η αντίστοιχη μείωση του βάρους εκείνων που ταξινομήθηκαν σωστά.

Αρχικά λοιπόν όλα τα στιγμιότυπα εκπαίδευσης φέρουν το ίδιο βάρος w=1/N, όπου N το πλήθος του σώματος εκπαίδευσης. Εν συνεχεία το πρώτο μοντέλο εκπαιδεύεται και η απόδοσή του αξιολογείται, υπολογίζοντας το ζυγισμένο σφάλμα του e1 επί του σώματος εκπαίδευσης ως το άθροισμα των βαρών των εσφαλμένως ταξινομηθέντων στιγμιότυπων προς το συνολικό άθροισμα των βαρών. Παράλληλα τα βάρη των στιμιότυπων που ταξινομήθηκαν λανθασμένα πολλαπλασιάζονται με τον παράγοντα e/(1-e) και ακολουθεί κανονικοποίηση όλων των βαρών, ώστε να αθροίζουν στη μονάδα.

Αυτή η ακολουθία βημάτων επαναλαμβάνεται έως τη δημιουργία ενός καθορισμένου αριθμού μοντέλων, έστω b. Στην περίπτωση που το σφάλμα του τρέχοντος μοντέλου ξεπεράσει την τιμή 0,5 ή εξισωθεί με το μηδέν, τότε ο αλγόριθμος τερματίζει και το μοντέλο αυτό δε λαμβάνεται υπ' όψη. Κατά την ταξινόμηση ενός άγνωστου στιγμιότυπου, κάθε ένα από τα παραγόμενα προβαίνει στην εκτίμησή του, η οποία ωστόσο συμμετέχει με διαφορετική βαρύτητα στην τελική απόφαση. Πιο συγκεκριμένα, σε κάθε μοντέλο ανατίθεται ένας συντελεστής

βαρύτητας ανάλογα με το σφάλμα του μοντέλου. Αυτός ο συντελεστής πολλαπλασιάζεται με την εκτίμηση του μοντέλου ώστε να αναδειχθεί η τελική απόφαση της επιτροπής.

Ένα αξιοσημείωτο χαρακτηριστικό του αλγορίθμου της προώθησης αποτελεί το γεγονός ότι η εκτέλεση ενός αρκετά μεγάλου αριθμού επαναλήψεων του δείχνει να αποφέρει ευεργετικά αποτελέσματα, ακόμα και όταν το συνολικό σφάλμα του τελικού μοντέλου επί των δεδομένων εκπαίδευσης έχει προ πολλού ελαχιστοποιηθεί.



Εικόνα 13 Γραφική αναπαράσταση (bagging) (Breiman, 1996)

Stacking: (Parmanto etal, 1996) Η μέθοδος της Συσσωρευμένης Γενίκευσης (Stacked Generalization ή Stacking) κάνει χρήση ενός συνόλου μοντέλων που σε αντίθεση με τις προσεγγίσεις που παρουσιάστηκαν ως τώρα, προέρχονται από διαφορετικούς αλγορίθμους μάθησης. Επίσης η λήψη της τελικής απόφασης δεν προϋποθέτει πλέον την υιοθέτηση της απόφασης της πλειοψηφίας ή τη ζυγισμένη εκτίμηση των επιμέρους αποφάσεων αλλά κάνει χρήση ενός μοντέλου - προέδρου, το οποίο μαθαίνει ποιο από τα μέλη της επιτροπής θα πρέπει να εμπιστεύεται σε κάθε περίπτωση. Πρόκειται ουσιαστικά για την επίλυση ενός νέου προβλήματος μάθησης, με δεδομένα τις αποφάσεις των μελών της επιτροπής (που ονομάζονται μοντέλα μηδενικού επιπέδου - level 0 inducers) καθώς και την πραγματική τιμή της συνάρτησης στόχου, για τα στιγμιότυπα ενός υποσυνόλου του σώματος εκπαίδευσης του αρχικού προβλήματος που δε χρησιμοποιήθηκαν κατά την εκπαίδευση των μοντέλων αυτών. Το μοντέλο που παράγεται κατά το δεύτερο αυτό στάδιο το οποίο εκτελεί χρέη προέδρου ονομάζεται μοντέλο πρώτου επιπέδου (level 1 inducer).

2.4 Deep Learning

Η βαθιά εκμάθηση (Deep Learning) είναι μια τεχνική μάθησης μηχανών που διδάσκει υπολογιστές να κάνουν αυτό που έρχεται φυσικά στον άνθρωπο: εκμάθηση με παραδείγματα. Η βαθιά εκμάθηση είναι η βασική τεχνολογία πίσω από τα αυτοκίνητα χωρίς οδηγό, επιτρέπο-

ντάς τους να αναγνωρίσουν ένα σημάδι στάσης ή να διακρίνουν έναν πεζό από ένα φανοστάτη. Είναι το κλειδί του φωνητικού ελέγχου σε καταναλωτικές συσκευές όπως τηλέφωνα, tablet, τηλεοράσεις και ηχεία ανοιχτής ακρόασης.

Στη βαθιά εκμάθηση, ένα μοντέλο υπολογιστή μαθαίνει να εκτελεί εργασίες ταξινόμησης απευθείας από εικόνες, κείμενο ή ήχο. Τα μοντέλα βαθιάς μάθησης μπορούν να επιτύχουν την ακρίβεια της τεχνολογίας και μερικές φορές υπερβαίνουν τις επιδόσεις σε ανθρώπινο επίπεδο. Τα μοντέλα εκπαιδεύονται χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων με σήμανση και αρχιτεκτονικές νευρωνικών δικτύων που περιέχουν πολλά στρώματα.

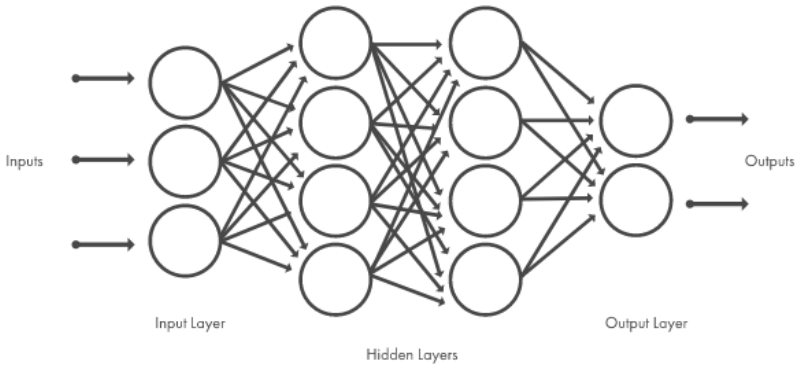
Η βαθιά εκμάθηση επιτυγχάνει την ακρίβεια αναγνώρισης σε υψηλότερα επίπεδα από ποτέ. Αυτό βοηθά τα ηλεκτρονικά είδη ευρείας κατανάλωσης να ανταποκρίνονται στις προσδοκίες των χρηστών και είναι κρίσιμης σημασίας για εφαρμογές για την ασφάλεια. Οι πρόσφατες εξελίξεις στη βαθιά εκμάθηση έχουν βελτιωθεί στο σημείο όπου η βαθιά μάθηση ξεπερνά τους ανθρώπους σε ορισμένες εργασίες όπως η ταξινόμηση αντικειμένων σε εικόνες.

Οι πιο βαθιές μέθοδοι μάθησης χρησιμοποιούν αρχιτεκτονικές νευρωνικών δικτύων, και γι''αυτό τα μοντέλα βαθιάς μάθησης συχνά αναφέρονται ως δίκτυα νευρώνων.

Ο όρος «βαθιά» συνήθως αναφέρεται στον αριθμό των κρυφών επιπέδων στο νευρικό δίκτυο. Τα παραδοσιακά νευρωνικά δίκτυα περιέχουν μόνο 2-3 κρυμμένα στρώματα, ενώ τα βαθιά δίκτυα μπορούν να έχουν έως και 150.

Τα μοντέλα βαθιάς μάθησης εκπαιδεύονται με τη χρήση μεγάλων συνόλων ετικετών δεδομένων και αρχιτεκτονικών νευρωνικών δικτύων που μαθαίνουν τις λειτουργίες απευθείας από τα δεδομένα χωρίς να χρειάζεται χειροκίνητη εξαγωγή χαρακτηριστικών (LeCun etal, 2015).

Στην εικόνα βλέπουμε Νευρωνικά δίκτυα, τα οποία είναι οργανωμένα σε στρώματα που αποτελούνται από ένα σύνολο διασυνδεδεμένων κόμβων. Τα δίκτυα μπορούν να έχουν δεκάδες ή εκατοντάδες κρυμμένα στρώματα.



Εικόνα 14 Νευρωνικά Δίκτυα (Schmidhuber, 2015)

Ένας από τους πιο δημοφιλείς τύπους βαθιών νευρωνικών δικτύων είναι γνωστός ως συνελικτικά νευρωνικά δίκτυα (CNN ή ConvNet). Το CNN περιγράφει τις γνωστές λειτουργίες με

δεδομένα εισόδου και χρησιμοποιεί 2D στρώματα περιστροφής, κάνοντας αυτή την αρχιτεκτονική κατάλληλη για την επεξεργασία δεδομένων 2D, όπως είναι οι εικόνες.

Ποια είναι η διαφορά μεταξύ μηχανικής μάθησης και βαθιάς μάθησης;

Η βαθιά μάθηση είναι μια εξειδικευμένη μορφή μηχανικής μάθησης. Μια ροή εργασίας μηχανής εκμάθησης ξεκινά με τα σχετικά χαρακτηριστικά που εξάγονται χειροκίνητα. Οι λειτουργίες στη συνέχεια χρησιμοποιούνται για τη δημιουργία ενός μοντέλου που κατηγοριοποιεί τα αντικείμενα. Με μια ροή εργασίας υψηλής εμβέλειας, οι σχετικές λειτουργίες εξάγονται αυτόματα. Επιπλέον, η βαθιά εκμάθηση εκτελεί «μάθηση από άκρο σε άκρο» - όπου ένα δίκτυο λαμβάνει δεδομένα και ένα έργο που πρέπει να εκτελέσει, όπως ταξινόμηση, και μαθαίνει πώς να το κάνει αυτομάτως.

Μια άλλη βασική διαφορά είναι η κλίμακα αλγορίθμων βαθιάς μάθησης με δεδομένα, ενώ η ρηχή μάθηση συγκλίνει. Η ρητή εκμάθηση αναφέρεται σε μεθόδους μηχανικής μάθησης που πλησιάζουν σε ένα ορισμένο επίπεδο απόδοσης όταν προστίθενται περισσότερα παραδείγματα και εκπαιδευτικά δεδομένα στο δίκτυο.

Ένα βασικό πλεονέκτημα των δικτύων βαθιάς μάθησης είναι ότι συχνά συνεχίζουν να βελτιώνονται όσο αυξάνεται το μέγεθος των δεδομένων.

Στη μηχανική μάθηση, επιλέγουμε χειροκίνητα τις λειτουργίες και έναν ταξινομητή για να ταξινομήσουμε. Με τη βαθιά γνώση, τα στάδια εξαγωγής και μοντελοποίησης είναι αυτόματα (LeCun etal, 2015).

Επιλογή μεταξύ μηχανικής μάθησης και βαθιάς μάθησης

Στην πράξη, η βαθιά μάθηση είναι ένα υποσύνολο της μηχανικής μάθησης. Είναι τεχνικά μηχανική μάθηση και λειτουργεί με παρόμοιο τρόπο, αλλά οι δυνατότητές της είναι διαφορετικές.

Τα βασικά μοντέλα μηχανικής μάθησης γίνονται σταδιακά καλύτερα, όποια και αν είναι η λειτουργία τους, αλλά εξακολουθούν να είναι μια σειρά από οδηγίες. Εάν ένας αλγόριθμος επιστρέφει μια ανακριβή πρόβλεψη, τότε ένας μηχανικός πρέπει να εισέλθει και να κάνει προσαρμογές. Αλλά με ένα μοντέλο βαθιάς μάθησης, οι αλγόριθμοι μπορούν να καθορίσουν μόνοι τους εάν μια πρόβλεψη είναι ακριβής ή όχι.

Η μηχανική μάθηση προσφέρει μια ποικιλία τεχνικών και μοντέλων ανάλογα με την εφαρμογή, το μέγεθος των δεδομένων που επεξεργαζόμαστε και το είδος του προβλήματος που θέλουμε να λύσουμε. Μια επιτυχημένη εφαρμογή βαθιάς εκμάθησης απαιτεί πολύ μεγάλο όγκο δεδομένων για την εκπαίδευση του μοντέλου, καθώς και μονάδες GPU ή μονάδες επεξεργασίας γραφικών για την ταχεία επεξεργασία των δεδομένων.

2.5 Επεξεργασία Φυσικής Γλώσσας (NLP)

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) είναι ένας κλάδος της τεχνητής νοημοσύνης που βοηθά τους υπολογιστές να κατανοούν, να ερμηνεύουν και να χειρίζονται την ανθρώπινη γλώσσα. Το NLP αντλεί από πολλούς κλάδους, συμπεριλαμβανομένης της επιστήμης των υπολογιστών και της υπολογιστικής γλωσσολογίας, στην προσπάθειά του να καλύψει το χάσμα μεταξύ της ανθρώπινης επικοινωνίας και της κατανόησης των υπολογιστών.

Ενώ η επεξεργασία της φυσικής γλώσσας δεν είναι μια νέα επιστήμη, η τεχνολογία προχωράει γρήγορα χάρη στο αυξημένο ενδιαφέρον για τις επικοινωνίες από άνθρωπο σε μηχανή, καθώς και τη διαθεσιμότητα μεγάλων δεδομένων, ισχυρών υπολογιστών και ενισχυμένων αλγορίθμων (Nasuwaka & Yi, 2003).

Ως άνθρωποι μπορούμε να μιλήσουμε και να γράψουμε διάφορες γλώσσες. Αλλά η μητρική γλώσσα ενός υπολογιστή - γνωστή ως κωδικός μηχανής ή γλώσσα μηχανής - είναι σε μεγάλο βαθμό ακατανόητη για τους περισσότερους ανθρώπους. Στα χαμηλότερα επίπεδα κάθε συσκευής, η επικοινωνία δεν συμβαίνει με λέξεις αλλά με εκατομμύρια μηδενικά και άσσους που παράγουν λογικές ενέργειες.

Πράγματι, οι προγραμματιστές χρησιμοποίησαν κάρτες διάτρησης για να επικοινωνήσουν με τους πρώτους υπολογιστές πριν από 70 χρόνια. Αυτή η χειρωνακτική και επίπονη διαδικασία έγινε κατανοητή από σχετικά μικρό αριθμό ανθρώπων. Πλέον μπορούμε να μιλάμε στις συσκευές μας, "Alexa, μου αρέσει αυτό το τραγούδι" και η συσκευή που παίζει μουσική στο σπίτι μειώνει την ένταση και απαντάει με ανθρώπινη φωνή. Ο αλγόριθμός του προσαρμόζεται για να παίξει αυτό το τραγούδι την επόμενη φορά που θα ακούσουμε αυτόν τον μουσικό σταθμό.

Ουσιαστικά, η συσκευή ενεργοποιήθηκε όταν άκουσε να μιλάτε, εκτέλεσε μια ενέργεια και παρείχε ανατροφοδότηση σε μια καλά διαμορφωμένη αγγλική πρόταση, όλα μέσα σε περίπου πέντε δευτερόλεπτα. Η πλήρης αλληλεπίδραση έγινε δυνατή από το NLP, μαζί με άλλα στοιχεία της Τεχνητής Νοημοσύνης, όπως η μηχανική μάθηση και η βαθιά μάθηση (Yi etal, 2003).

Γιατί είναι σημαντικό το NLP;

Η επεξεργασία της φυσικής γλώσσας βοηθά τους υπολογιστές να επικοινωνούν με τους ανθρώπους στη γλώσσα τους και να κλιμακώνουν άλλες εργασίες σχετικές με τη γλώσσα. Για παράδειγμα, το NLP καθιστά δυνατό στους υπολογιστές να διαβάζουν κείμενο, να ακούν λόγο, να ερμηνεύουν, να μετρούν το συναίσθημα και να καθορίζουν ποια μέρη είναι σημαντικά.

Οι σημερινές μηχανές μπορούν να αναλύσουν περισσότερα δεδομένα βασισμένα στη γλώσσα από ότι οι άνθρωποι, χωρίς κόπο και με συνεπή, αμερόληπτο τρόπο. Λαμβάνοντας υπόψη την εκπληκτική ποσότητα των αδόμητων δεδομένων που παράγονται καθημερινά, από τα ιατρικά αρχεία μέχρι τα κοινωνικά μέσα, η αυτοματοποίηση θα είναι κρίσιμη για την αποτελεσματική ανάλυση των δεδομένων κειμένου και λόγου.

Η ανθρώπινη γλώσσα είναι εκπληκτικά πολύπλοκη και ποικίλη. Εκφράζουμε τον εαυτό μας με άπειρους τρόπους, τόσο προφορικά όσο και γραπτά. Όχι μόνο υπάρχουν εκατοντάδες γλώσσες και διάλεκτοι, αλλά σε κάθε γλώσσα υπάρχει ένα μοναδικό σύνολο κανόνων γραμματικής και σύνταξης, όρων και slang. Όταν γράφουμε, συχνά γράφουμε λάθος ή συντομεύουμε λέξεις ή παραλείπουμε τη στίξη. Όταν μιλάμε, έχουμε περιφερειακές προθέσεις, και δανειζόμαστε όρους από άλλες γλώσσες.

Εικόνα 15 Natural Language Processing (https://www.sas.com/en_nz/insights/analytics/what-is-natural-language-processing-nlp.html)

Ενώ η επιβλεπόμενη και η μη επιβλεπόμενη μάθηση, και συγκεκριμένα η βαθιά εκμάθηση, χρησιμοποιούνται σήμερα ευρέως για τη μοντελοποίηση της ανθρώπινης γλώσσας, υπάρχει επίσης ανάγκη για συντακτική και σημασιολογική κατανόηση και εμπειρογνωμοσύνη πεδίου σε αυτές τις προσεγγίσεις μηχανικής μάθησης. Το NLP είναι σημαντικό επειδή βοηθά στην επίλυση της ασάφειας στη γλώσσα και προσθέτει χρήσιμη αριθμητική δομή στα δεδομένα για πολλές εφαρμογές, όπως η αναγνώριση ομιλίας ή η ανάλυση κειμένου (Yu & Hatzivassiloglou, 2003).

Η επεξεργασία φυσικής γλώσσας περιλαμβάνει πολλές διαφορετικές τεχνικές για την ερμηνεία της ανθρώπινης γλώσσας, που κυμαίνονται από τις μεθόδους στατιστικής και μηχανικής μάθησης έως τις βασισμένες σε κανόνες και αλγοριθμικές προσεγγίσεις. Είναι αναγκαία μια ευρεία σειρά προσεγγίσεων, διότι τα δεδομένα που βασίζονται σε κείμενο και φωνή ποικίλλουν ευρέως, όπως και οι πρακτικές εφαρμογές.

Βασικές εργασίες NLP είναι η ομαδοποίηση, ανάλυση, λημματοποίηση και επισημείωση (tokenization, parsing, lemmatization / stemming, tagging) μερών ομιλίας, ανίχνευση γλώσσας και αναγνώριση σημασιολογικών σχέσεων.

Σε γενικές γραμμές, οι αλγόριθμοι NLP κατανέμουν τη γλώσσα σε μικρότερα, στοιχειώδη κομμάτια, προσπαθούν να κατανοήσουν τις σχέσεις μεταξύ των κομματιών και να διερευνήσουν πώς τα κομμάτια συνεργάζονται για να δημιουργήσουν νόημα.

Αυτά τα βασικά καθήκοντα χρησιμοποιούνται συχνά σε ικανότητες NLP υψηλότερου επιπέδου, όπως:

Κατηγοριοποίηση περιεχομένου: Μια σύνοψη εγγράφων βασισμένη στη γλωσσολογία, συμπεριλαμβανομένης της αναζήτησης και της ευρετηρίασης, των ειδοποιήσεων περιεχομένου και της ανίχνευσης διπλότυπων.

Ανεύρεση και μοντελοποίηση θεμάτων: Συλλογή νοήματος και θεμάτων με ακρίβεια σε συλλογές κειμένων και εφαρμογή προηγμένων αναλύσεις σε κείμενο, όπως βελτιστοποίηση και πρόβλεψη.

Συγκεντρωτική εξαγωγή: Αυτόματη εξαγωγή δομημένων πληροφοριών από πηγές που βασίζονται σε κείμενο.

Ανάλυση συναισθημάτων: Προσδιορισμός της διάθεσης ή των υποκειμενικών απόψεων μέσα σε μεγάλες ποσότητες κειμένου, συμπεριλαμβανομένου του μέσου συναίσθηματος και της εξόρυξης γνώμης.

Μετατροπή ομιλίας σε κείμενο και μετατροπή κειμένου σε ομιλία: Μετατροπή φωνητικών εντολών σε γραπτό κείμενο και αντίστροφα.

Συνοπτική παρουσίαση εγγράφου: Δημιουργία αυτόματων συνόψεων μεγάλων κειμένων κειμένου.

Μηχανική μετάφραση: Αυτόματη μετάφραση κειμένου ή ομιλίας από τη μια γλώσσα στην άλλη (Yi etal, 2003).

Σε όλες αυτές τις περιπτώσεις, ο πρωταρχικός στόχος είναι να ληφθούν ακατέργαστα δεδομένα γλώσσας και να χρησιμοποιηθούν γλωσσολογία και αλγόριθμοι για να μετατραπεί ή να εμπλουτιστεί το κείμενο με τέτοιο τρόπο ώστε να αποδίδει μεγαλύτερη αξία.

2.6 Μετρικές Αξιολόγησης

Προκειμένου να αξιολογηθεί η επίδοση ενός ταξινομητή, έχει προταθεί πληθώρα μετρικών αξιολόγησης. Στη συνέχεια, παρουσιάζονται οι δημοφιλέστερες μετρικές αξιολόγησης αλγορίθμων μηχανικής μάθησης.

Η πιο απλή και αντιπροσωπευτική μετρική είναι η Γενική Ορθότητα Πρόβλεψης. Η **ορθότητα** (accuracy) υπολογίζεται ως το ποσοστό των στιγμιότυπων του συνόλου ελέγχου που ταξινομήθηκαν στην σωστή κατηγορία.

Accuracy =TP+TN/TP+FP+TN+FN

όπου,

TP = το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία positive και ταξινομήθηκαν στην κατηγορία positive. (σωστή ταξινόμηση)

TN = το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία negative και ταξινομήθηκαν στην κατηγορία negative. (σωστή ταξινόμηση)

FP = το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία negative και ταξινομήθηκαν στην κατηγορία positive. (λανθασμένη ταξινόμηση)

FN = το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία positive και ταξινομήθηκαν στην κατηγορία negative. (λανθασμένη ταξινόμηση)

Η **ευαισθησία ή ανάκληση** (sensitivity ή recall) υπολογίζεται ως εξής:

Recall = TP/TP +FN

Η **ακρίβεια** (precision) υπολογίζεται ως εξής:

Precision = TP/TP+FP

Η **εξειδίκευση** (specificity) υπολογίζεται ως εξής:

Specificity=TN/TN+FP

Τέλος, η μετρική **F-Measure** παρέχει μία συνολική εκτίμηση των μοντέλων, καθώς συνδυάζει δύο άλλες μετρικές, την ανάκληση και την ακρίβεια. Η μετρική F-Measure στην ουσία είναι ο αρμονικός μέσος όρος (harmonic mean) της ανάκλησης και της ακρίβειας, και υπολογίζεται ως εξής:

FMeasure = 2*recall *precision/recall + precision

Σε κάθε αλγόριθμο που θα εφαρμόσουμε το WEKA μας δίνει ένα πίνακα, ο οποίος ονομάζεται Confusion Matrix. Ο πίνακας έχει την εξής μορφή:

Examples		
Class	True Positive(tp)	False Negative(fn)
	False Positive(fp)	True Negative(tn)

Οι παραπάνω μετρικές υπολογίζονται στα πειράματα που ακολουθούν για κάθε αλγόριθμο Μηχανικής Μάθησης που χρησιμοποιείται.

Καμπύλες ROC

Ένα χρήσιμο εργαλείο για την πρόβλεψη της πιθανότητας ενός δυαδικού αποτελέσματος είναι η καμπύλη χαρακτηριστικών λειτουργίας του δέκτη ή η καμπύλη ROC.

Πρόκειται για μια γραφική παράσταση του ψευδοθετικού ποσοστού (άξονας x) έναντι του πραγματικού θετικού ρυθμού (άξονας y) για έναν αριθμό διαφορετικών υποψήφιων οριακών

τιμών μεταξύ 0,0 και 1,0. Με άλλα λόγια, καταγράφει το ποσοστό ψευδούς συναγερμού έναντι του ποσοστού επιτυχίας.

Η καμπύλη ROC είναι ένα χρήσιμο εργαλείο καθώς:

- Οι καμπύλες διαφορετικών μοντέλων μπορούν να συγκριθούν άμεσα γενικά ή για διαφορετικά κατώφλια.
- Η περιοχή κάτω από την καμπύλη (AUC) μπορεί να χρησιμοποιηθεί ως περίληψη της ικανότητας του μοντέλου.

Το σχήμα της καμπύλης περιέχει πολλές πληροφορίες, συμπεριλαμβανομένου του τι μπορούμε να ενδιαφέρουμε περισσότερο για ένα πρόβλημα, το αναμενόμενο ψευδώς θετικό ποσοστό και το ψευδώς αρνητικό ποσοστό.

Για να καταστεί σαφές:

- Οι μικρότερες τιμές στον άξονα x του σχεδίου υποδεικνύουν χαμηλότερα ψευδώς θετικά και υψηλότερα πραγματικά αρνητικά.
- Μεγαλύτερες τιμές στον άξονα y της γραφικής παράστασης υποδεικνύουν υψηλότερα αληθινά θετικά και χαμηλότερα ψευδώς αρνητικά.

Ένα ικανό μοντέλο θα αποδώσει υψηλότερη πιθανότητα σε ένα τυχαία επιλεγμένο πραγματικό θετικό περιστατικό από ένα αρνητικό περιστατικό κατά μέσο όρο. Αυτό εννοούμε όταν λέμε ότι το μοντέλο έχει δεξιότητες. Γενικά, τα επιδέξια μοντέλα αντιπροσωπεύονται από καμπύλες που ανεβαίνουν στο πάνω αριστερό μέρος του οικοπέδου.

Ένα μοντέλο χωρίς ικανότητες αντιπροσωπεύεται στο σημείο [0.5, 0.5]. Ένα μοντέλο χωρίς εξειδίκευση σε κάθε κατώφλι αντιπροσωπεύεται από μια διαγώνια γραμμή από το κάτω αριστερό μέρος του διαγράμματος στην άνω δεξιά και έχει AUC 0,5.

Ένα μοντέλο με τέλεια δεξιότητα αντιπροσωπεύεται σε ένα σημείο [0.0, 1.0]. Αντιπροσωπεύεται από μια γραμμή που ταξιδεύει από το κάτω αριστερό μέρος του οικοπέδου προς τα επάνω αριστερά και έπειτα σε όλη την κορυφή προς τα επάνω δεξιά.

Ο χειριστής μπορεί να σχεδιάσει την καμπύλη ROC για το τελικό μοντέλο και να επιλέξει ένα κατώτατο όριο που δίνει μια επιθυμητή ισορροπία μεταξύ των ψευδών θετικών και των ψευδών αρνητικών.

- Οι καμπύλες ROC συνοψίζουν το αντιστάθμισμα μεταξύ του πραγματικού θετικού ποσοστού και του ψευδώς θετικού ποσοστού για ένα μοντέλο πρόβλεψης χρησιμοποιώντας διαφορετικά όρια πιθανότητας.

- Οι καμπύλες ακρίβειας-ανάκλησης συνοψίζουν την αντιστάθμιση μεταξύ του πραγματικού θετικού ποσοστού και της θετικής πρόβλεψης για ένα μοντέλο πρόβλεψης χρησιμοποιώντας διαφορετικά όρια πιθανότητας.
- Οι καμπύλες ROC είναι κατάλληλες όταν οι παρατηρήσεις είναι ισορροπημένες μεταξύ κάθε κατηγορίας, ενώ οι καμπύλες ανάκτησης ακριβείας είναι κατάλληλες για μη ισορροπημένα σύνολα δεδομένων (Hanley & McNeil, 1982).

3

Πείραμα

3.1 Περιγραφή δεδομένων - Διαδικασία Συλλογής

Τα δεδομένα που συλλέχθηκαν έχουν κοινή θεματολογία. Πρόκειται για κριτικές χρηστών του Tripadvisor σχετικά με τα ελληνικά εστιατόρια και πιο συγκεκριμένα πρόκειται για κάποια επιλεγμένα εστιατόρια της Κέρκυρας. Καθώς η Κέρκυρα αποτελεί έναν άκρως τουριστικό προορισμό, έχει μεγάλη σημασία να διαπιστώσουμε αν οι επισκέπτες του νησιού μένουν ικανοποιημένοι από το φαγητό που του προσφέρεται ή όχι.

Αρχικά, έγινε αναζήτηση των κερκυραϊκών εστιατορίων που υπάρχουν στην πλατφόρμα του Tripadvisor και στη συνέχεια επιλέχθηκαν τα παρακάτω:

- Strofilia Taverna-Kassiopi
- Du Lac-Limni
- La Veranda di Corfu-Corfu
- Karnveer Indian Restaurant-Gouvia
- The Square Steak Burger Bar-Kassiopi
- The Hive-Sidari
- The Flower Garden Restaurant-Benitses
- Flavor-Paleokastritsa
- Porto Timoni Restaurant Cafe-Afionas

- Nolita-Corfu
- Toula’s-Corfu
- Thraka Grill House-Kassiopi
- Bikolis taverna-Peroulades
- Big Max Diner-Kavos
- The Three Brothers-Astrakeri
- Anthos Restaurant-Corfu
- Squirrel Restaurant-Ipsos
- Taverna Kalami-Boukari
- Archontiko-Corfu
- The Tandoori Bites-Corfu
- George Elena’s Taverna-Corfu
- Etrusco-Kato Korakiana
- Pomo d’Oro wine restaurant-Corfu
- Ladokolla-Corfu
- The Venetian Well-Corfu
- La Cucina-Corfu
- Foros-Peritheia
- Ortholithi-Agios Gordios
- Apovrado-Kommeno
- Pane e Souvlaki-Corfu
- Porta Remounda-Corfu
- Ta Kokoria-Corfu
- Aegli-Corfu
- Avli-Corfu
- Tsipouradiko Corfu -Corfu
- Sabbia-Agios Gordios
- Taverna Elizabeth-Doukades
- Ambelonas Corfu-Pelekas
- Mamma’s Taverna-Moraitika
- Pavlos Taverna Steakout-Kavos
- The Village Taverna-Moraitika

- Akrotiri Café north west Corfu-Arillas
- To Alatopipero-Corfu

με βάση την πληθώρα των σχολίων που περιέχουν. Πρόκειται για συνολικά 43 εστιατόρια στο νησί της Κέρκυρας. Η λίστα περιλαμβάνει εστιατόρια όλου του νησιού. Τα σχόλια που ανακτήθηκαν και χρησιμοποιήθηκαν στην παρούσα έρευνα αφορούν αποκλειστικά κριτικές του έτους 2018. Οπότε, η επιλογή των εστιατορίων έγινε και με το σκεπτικό να έχουν κριτικές που τα αφορούν στο παρόν έτος.

Συνολικά πρόκειται για ένα σετ από 600 κριτικές οι οποίες αφορούν κερκυραϊκά εστιατόρια για το έτος 2018. Τα εστιατόρια επιλέχθηκαν με βάση τις λίστες του Tripadvisor για Fine Dining, Local cuisine, Moderately priced, Lunch και Cheap Eats. Η λίστα δεν αφορά τα καλύτερα εστιατόρια από κάθε κατηγορία, αλλά τα εστιατόρια με ικανοποιητικό αριθμό σχολίων για το έτος 2018.

Εστιατόριο	Αριθμός Σχολίων
Strofilia Taverna-Kassiopi	16
Du Lac-Limni	3
La Veranda di Corfu-Corfu	5
Karnveer Indian Restaurant-Gouvia	16
The Square Steak Burger Bar-Kassiopi	16
The Hive-Sidari	16
The Flower Garden Restaurant-Benitses	13
Flavor-Paleokastritsa	16
Porto Timoni Restaurant Cafe-Afionas	16
Nolita-Corfu	16
Toula's-Corfu	16

Thraka Grill House-Kassiopi	16
Bikolis taverna-Peroulades	16
Big Max Diner-Kavos	16
The Three Brothers-Astrakeri	16
Anthos Restaurant-Corfu	16
Squirrel Restaurant-Ipsos	16
Taverna Kalami-Boukari	13
Archontiko-Corfu	16
The Tandoori Bites-Corfu	16
George Elena's Taverna-Corfu	16
Etrusco-Kato Korakiana	11
Pomo d'Oro wine restaurant-Corfu	10
Ladokolla-Corfu	16
The Venetian Well-Corfu	16
La Cucina-Corfu	5
Foros-Peritheia	16
Ortholithi-Agios Gordios	16
Apovrado-Kommeno	16
Pane e Souvlaki-Corfu	16

Porta Remounda-Corfu	10
Ta Kokoria-Corfu	16
Aegli-Corfu	16
Avli-Corfu	16
Tsipouradiko Corfu -Corfu	16
Sabbia-Agios Gordios	9
Taverna Elizabeth-Doukades	9
Ambelonas Corfu-Pelekas	3
Mamma's Taverna-Moraitika	16
Pavlos Taverna Steakout-Kavos	16
The Village Taverna-Moraitika	16
Akrotiri Café north west Corfu-Arillas	16
To Alatopipero-Corfu	13

Πίνακας 1 Εστιατόρια – Αριθμός Σχολίων

Στη συνέχεια, δημιουργήθηκε ένα σετ δεδομένων/λογιστικό φύλλο με τα 600 σχόλια ανά εστιατόριο. Τα δεδομένα αυτά εισήχθησαν στο Aylien με σκοπό να εξάγουμε την ανάλυση συναισθήματος. Η συγκεκριμένη πλατφόρμα δίνει τη δυνανότητα να κατατμήσει κάθε σχόλιο και το κατατάσσει σε θεματολογίες σχετικές με τα εστιατόρια.

Ενδεικτικά η πλατφόρμα χωρίζει τα σχόλια σε προτάσεις που αφορούν (Aspects):

a) Ambience
b) Busyness
c) Cleanliness
d) Desserts
e) Drinks
f) Facilities
g) Food
h) GroundService
i) Menu

j) Location
k) Payment
l) Quietness
m) Reservations
n) Staff
o) Value

και αξιολογεί με negative, neutral ή positive (Sentiment) την πρόταση της κάθε κατηγορίας με σκοπό να αποδώσει μια αξιολόγηση σε όλο το σχόλιο συνολικά. Πολλές φορές γίνεται και παράλειψη κάποιου μέρους του σχολίου ή παραμένει το σχόλιο αυτούσιο και αξιολογείται. Επίσης, κάποιο σχόλιο μπορεί να έχει διαφορετικά Aspects και όχι μόνο ένα.

Τα 600 σχόλια που εισήχθησαν στην πλατφόρμα επέφεραν 1.696 διατμημένα σχόλια με την κατάλληλη, σύμφωνα με την πλατφόρμα πάντα, αξιολόγηση. Το καινούριο dataset που προέκυψε, είχε 1.332 πλέον παραδείγματα, καθώς αφαιρέθηκαν όλες οι διπλές προτάσεις που μπορεί να προέκυψαν από την κατάτμηση στο σύστημα. Κάθε παράδειγμα χαρακτηρίστηκε με ένα μοναδικό χαρακτηριστικό (ID) όπου το πρώτο σχόλιο από τα 600 χωρίστηκε σε τέσσερις καινούριες γραμμές με τα αντίστοιχα υπο-χαρακτηριστικά. Προέκυψαν τέσσερις στήλες από αυτή τη διαδικασία, η στήλη ID, η στήλη Sentence που περιέχει το κείμενο, η στήλη Aspects και η στήλη Sentiment.

Ενδεικτικά για το σχόλιο:

«We visited this restaurant after a recommendation from a local. The location was magnificent, in the center of old town but quite and romantic. The service was excellent (really fast and always kindly around). The food was really tasty!!! We had as starters the tart and the goat cheese salad. For main courses the lamb and the veal (both delicious). We were really satisfied from our visit to Venetian Well restaurant and we will visit it again for sure!!!»

προκύπτει ο παρακάτω πίνακας:

ID	Sentence	Aspect	Sentiment
11	The food was really tasty!!!	Food	positive
12	We had as starters the tart and the goat cheese salad.	Desserts	positive
13	The location was magnificent, in the center of old town but quite and romantic.	Location	positive
14	The service was excellent (really fast and always kindly around).	Staff	positive

Ωστόσο, ενώ ένα μεγάλο ποσοστό του dataset αξιολογήθηκε σωστά (91,8%) υπήρξε ένας μικρός αριθμός σχολίων τα οποία χαρακτηρίστηκαν λανθασμένα (8,2%). Συγκεκριμένα, 109 από τα 1332. Σε δεύτερη αξιολόγηση εντοπίστηκαν όλες οι εσφαλμένες αξιολογήσεις και χαρακτηρίστηκαν στο dataset με ένα κόκκινο θαυμαστικό. Προστέθηκε μία επιπλέον στήλη

Sentiment Annotator Class με την σωστή αξιολόγηση για όλες τις προτάσεις και επισημάνθηκαν με κόκκινο οι αξιολογήσεις που ήταν λανθασμένες από το σύστημα.

Οι λανθασμένες προτάσεις συγκεντρώθηκαν σε ξεχωριστό αρχείο για να αναγνωριστούν πιθανές λέξεις ή εκφράσεις που μπορεί να δυσκόλευαν το σύστημα και να έβγαζε λανθασμένη απόφαση. Καταλήξαμε σε 40 τέτοιες «επίφοβες» λέξεις/εκφράσεις που πιθανώς να αντέστρεφαν την πολικότητα.

Στη συνέχεια, οι εκφράσεις αυτές προστέθηκαν ως χαρακτηριστικά:

1. But
2. Not
3. Very
4. Well
5. So
6. Avoid
7. Disappointment
8. Over
9. Without
10. Worth
11. Limited
12. Really
13. Tasty
14. Offered
15. Second To None
16. Overpriced
17. Interesting
18. Everything
19. Even
20. Bargain
21. Great
22. Spent
23. Romantic
24. Good
25. More Than
26. Nothing
27. Special
28. Only
29. Fine
30. Ready
31. Any
32. low
33. quite
34. expensive
35. lovely
36. comfortable
37. excellent
38. bad
39. reasonable
40. enough

Τα 40 επιπλέον χαρακτηριστικά που προστέθηκαν είναι λέξεις οι οποίες θα μπορούσαν να επηρεάσουν την «κρίση» της πλατφόρμας Aylien και να βγήκαν λανθασμένα αποτελέσματα. Στο λογιστικό φύλλο, σε αυτές τις στήλες συμπληρώθηκαν οι λέξεις TRUE ή FALSE ανάλογα αν η λέξη υπήρχε στην κάθε πρόταση που εξετάσαμε.

Στη συνέχεια, αποθηκεύτηκε ως CSV αρχείο και εισήχθη στην πλατφόρμα του WEKA.

(Ένα αρχείο CSV (Τιμές διαχωρισμένες με κόμματα) είναι ένας ειδικός τύπος αρχείου. Αντί για την αποθήκευση πληροφοριών σε στήλες, τα αρχεία CSV αποθηκεύουν πληροφορίες διαχωρισμένες με κόμματα. Όταν αποθηκεύουμε κείμενο και αριθμούς σε ένα αρχείο CSV, είναι εύκολο να τα μετακινήσουμε από ένα πρόγραμμα σε ένα άλλο.)

3.2 Διαδικασία Πειράματος

Στην πλατφόρμα του Weka εφαρμόστηκαν στα δεδομένα δύο αλγορίθμους ώστε να συγκριθούν τα αποτελέσματά τους σε σχέση με τα αποτελέσματα που έδωσε η πλατφόρμα του aylien.

Στην εξόρυξη γνώσης από κείμενο και πιο συγκεκριμένα στην εξόρυξη γνώμης και συναισθηματική ανάλυση κειμένου, οι πιο αποτελεσματικοί αλγόριθμοι στις κατηγορίες των Αλγορίθμων Μηχανικής Μάθησης ειδικότερα είναι οι Bayesian μέθοδοι και των Μηχανών Διανυσματικής Υποστήριξης (Aggarwal & Zhai Cheng, 2012).

Ο Naive Bayes είναι ένας πολύ δημοφιλής ταξινομητής σε προβλήματα επεξεργασίας φυσικής γλώσσα. Ένα από τα πρώτα προβλήματα στα οποία εφαρμόστηκε ήταν αυτό της αναγνώρισης μηνυμάτων ανεπιθύμητης αλληλογραφίας (spam)(Sahami, 1998). Διακρίνεται για την απλότητα και τη αποδοτικότητά του.

Τα μοντέλα κατηγοριοποίησης SVM, είναι από τα δημοφιλέστερα, στην κατηγοριοποίηση συναισθήματος, λόγω της αποτελεσματικότητας, της ταχύτητας και της ικανότητάς τους να παράγουν μη γραμμικά υπερεπίπεδα απόφασης, καθιστώντας εφικτή την επίλυση προβλημάτων που δεν είναι δυνατόν να επιλυθούν με γραμμικά μοντέλα. Στα προβλήματα κατηγοριοποίησης κειμένων όπως στην ανάλυση συναισθήματος τα διανύσματα χαρακτηριστικών είναι αραιά και μεγάλων διαστάσεων με συνέπεια οι παρατηρήσεις να είναι γραμμικά διαχωρίσιμες. (Cortes & Vapnik, 1995)

Επιλέχθηκαν οι συγκεκριμένοι αλγόριθμοι, οι οποίοι σύμφωνα με τα παραπάνω μπορεί να θεωρηθεί ότι είναι πολύ αποδοτικοί και είναι ήδη υλοποιημένοι στο WEKA.

Για τη μέτρηση της απόδοσης της κατηγοριοποίησης επιλέγουμε τη μέθοδο της Διασταυρωμένης επικύρωσης (k-fold Cross Validation) γιατί είναι η πιο αξιόπιστη. Σαν k επιλέγεται το 10 γιατί είναι το πιο διαδεδομένο. (McLachland etal, 2004)

3.2.1 Μεταμάθηση με αλγορίθμους Bayes

Αρχικά, εφαρμόστηκε στα δεδομένα τον αλγόριθμο Naïve Bayes από την κατηγορία "bayes" των διαθέσιμων αλγορίθμων του Weka. Τα αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου φαίνονται στον παρακάτω πίνακα:

Correctly Classified Instances	**1217**	**91.3664 %**
Incorrectly Classified Instances	115	8.6336 %

Kappa statistic	0.5888
Mean absolute error	0.0841
Root mean squared error	0.2155
Relative absolute error	56.7772 %
Root relative squared error	79.3635 %
Total Number of Instances	1332
Ignored Class Unknown Instances	1

=== Detailed Accuracy By Class ===								
Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
positive	0,964	0,362	0,950	0,964	0,957	0,632	0,890	0,979
neutral	0,100	0,011	0,263	0,100	0,145	0,143	0,789	0,198
negative	0,752	0,034	0,669	0,752	0,708	0,681	0,912	0,621
Weighted Avg.	0,914	0,321	0,901	0,914	0,906	0,617	0,888	0,919

=== Confusion Matrix ===			
a	**b**	**c**	**<-- classified as**
1127	12	30	a = positive
33	5	12	b = neutral
26	2	85	c = negative

3.2.2 S.V.M.

Στη συνέχεια, εφαρμόσαμε στα δεδομένα μας τον αλγόριθμο Support Vector Machine (S.V.M.). Ο αλγόριθμος αυτός στο Weka εμφανίζεται ως SMO στην κατηγορία "functions" των διαθέσιμων αλγορίθμων του Weka. Τα αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου φαίνονται στον παρακάτω πίνακα:

Correctly Classified Instances	**1213**	**91.0661 %**
Incorrectly Classified Instances	119	8.9339 %

Kappa statistic	0.5935
Mean absolute error	0.2479
Root mean squared error	0.3158
Relative absolute error	167.3008 %
Root relative squared error	116.2967 %
Total Number of Instances	1332
Ignored Class Unknown Instances	1

=== Detailed Accuracy By Class ===								
Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
positive	0,956	0,325	0,955	0,956	0,955	0,632	0,813	0,950
neutral	0,260	0,011	0,481	0,260	0,338	0,336	0,581	0,153
negative	0,753	0,043	0,615	0,735	0,669	0,639	0,868	0,489
Weighted Avg.	0,911	0,289	0,908	0,911	0,908	0,621	0,809	0,881

=== Confusion Matrix ===			
a	**b**	**c**	<-- classified as
1117	13	39	a = positive
24	13	13	b = neutral
29	1	83	c = negative

4

Αποτελέσματα

Οι πίνακες που ακολουθούν είναι το CONFUSION MATRIX μετά την μεταμάθηση που εφαρμόστηκαν στα χαρακτηριστικά που εντοπίστηκαν και θεωρήθηκαν υπεύθυνα για τη λάθος κατηγοριοποίηση του συναισθήματος.

CONFUSION MATRIX που αφορά τη θετική ταξινόμηση

Actual Class Positive	**Predicted**	
	Negative	**Positive**
Positive	Tpositive =1113 (a)	FNegative=39(b)
Negative	FPositive=52(c)	TNegative=92 (d)

RECALL=TP/TP+FN=1113/1113+39=0, 96

PRECISION=TP/TP+FP=1113/1113+52=0, 95

ACCURACY= TPositive+ Tnegative/ Tpositive+ Tnegative+ Fpositive+ Fnegative= 0, 92

F=2 X (Precision x Recall) / (Precision +Recall)=0,95

CONFUSION MATRIX που αφορά την αρνητική ταξινόμηση

Actual Class Negative	**Predicted**	
	Negative	**Positive**
Negative	Tnegative=92 (a)	Fpositive=52 (b)
Positive	Fnegative=39 (c)	Tpositive=1113 (d)

RECALL=TN/TN+FP=92/92+52= 0, 63

PRECISION=TN/TN+FN=92/92+39= 0, 70

ACCURACY= Tnegative + Tpositive/ Tpositive+ Tnegative+ Fpositive+ Fnegative=0, 92

F=2 X (Precision x Recall)/(Precision +Recall)=0,66

CONFUSION MATRIX που αφορά την ουδέτερη ταξινόμηση

Actual Class Neutral	**Predicted**	
	Neutral	**NO Neutral**
Neutral	TPNeutral=19 (a)	FNnoNeutral=109 (b)
NO Neutral	FPnoNeutral=1132 (c)	TNneutral=17 (d)

RECALL=TPneutral/TPneutral+ FNNeutral=19/19+109= 0, 14

PRECISION= TPNeutral/TPNeutral+FPNeutra=19/19+1132=0, 01

ACCURACY= TPNeutral+ TNneutral / TPNeutral+ TNneutral+FNNeutral+ FPNeutral =0,02

F=2 X (Precision x Recall) / (Precision +Recall)=0,01

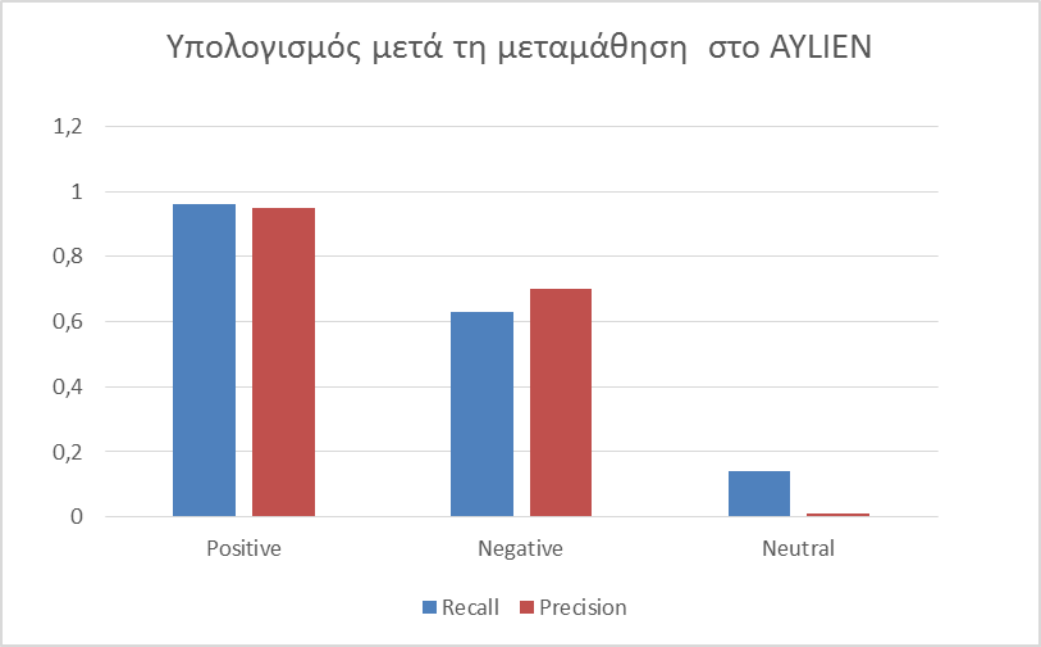
Τα αποτελέσματα αυτά προέκυψαν από το αρχικό σετ δεδομένων μετά τη μεταμάθηση στο Aylien και αφού διαγράψαμε τα διπλότυπα. Οι παραπάνω πίνακες προέκυψαν μετά από υπολογισμούς των μετρικών αξιολόγησης από το αρχείο csv, όπως αυτό διαμορφώθηκε μετά την εισαγωγή των σχολίων στο Aylien και σύμφωνα με τα αποτελέσματα του Aylien και τη δική μας παρέμβαση στη λάθος ταξινόμηση και πριν την εισαγωγή του αρχείου csv στο Weka.

Μετά το Confusion Matrix υπολογίζεται το Recall, Precision και το F-Measure της θετικής αρνητικής και ουδέτερης ταξινόμησης τα οποία δημιουργούν τον παρακάτω πίνακα:

CLASS	RECALL	PRECISION	ACCURACY	F-MEASURE
Positive	0,96	0,95	0,92	0,95
Negative	**0,63**	**0,70**	0,92	**0,66**
Neutral	0,14	0,01	0,01	0,02

Πίνακας 2 Αποτελέσματα RECALL, PRECISION, ACCURACY και F-MEASURE στις κλάσεις Positive Negative & Neutral μετά την μεταμάθηση στο Aylien

Στα παρακάτω διαγράμματα έχουμε οπτικοποίηση των αποτελεσμάτων Recall, Precision και F-Measure στις κλάσεις θετικού, αρνητικού και ουδέτερου συναισθήματος μετά το δικό μας υπολογισμό και των αποτελεσμάτων που μας δίνει το Weka εφαρμόζοντας τους αλγορίθμους Naïve Bayes και SVM.

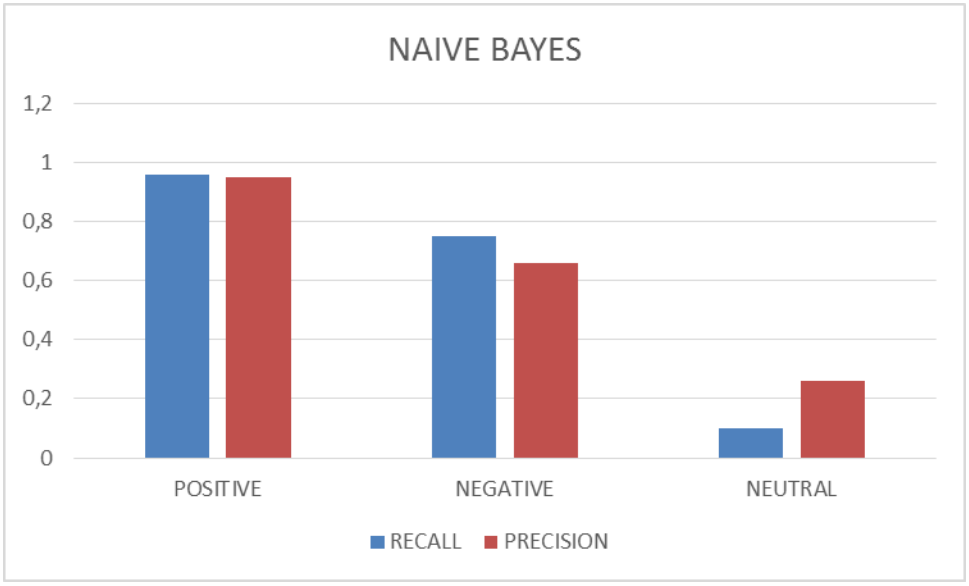


Εικόνα 16 Precision και Recall θετικών, ουδέτερων και αρνητικών συναισθημάτων μετά τους δικούς μας υπολογισμούς

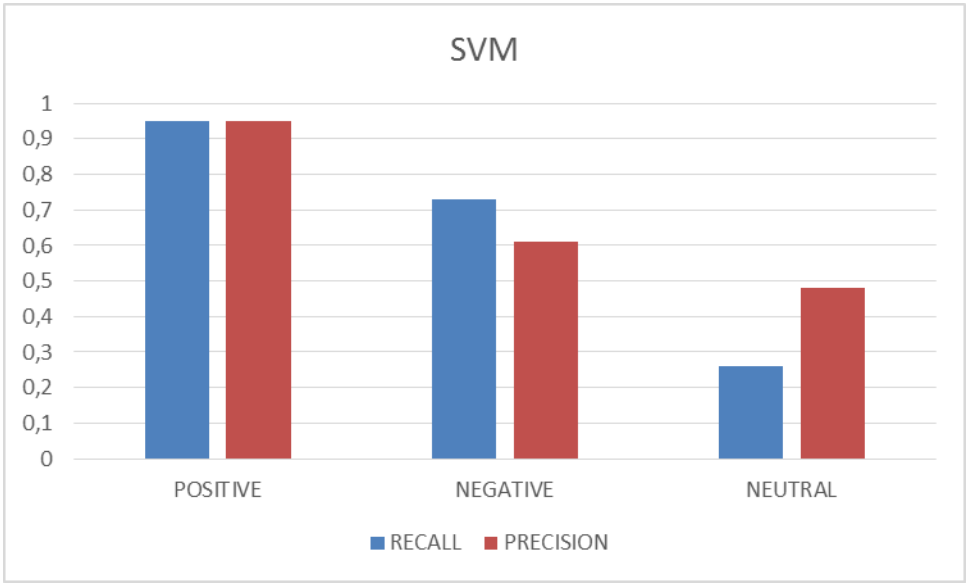
Οι θετικές απαντήσεις βγάζουν ικανοποιητικά αποτελέσματα (Recall: 0,96 και Precision: 0,95) και επομένως θεωρούμε ότι είναι αξιόπιστα τα αποτελέσματά τους.

Οι αρνητικές απαντήσεις δεν είναι τόσο αξιόπιστες, λόγω των αποτελεσμάτων των Recall και Precision τα οποία είναι 0,63 και 0,70 αντίστοιχα.

Στο ουδέτερο συναίσθημα το ποσοστό είναι πολύ μικρό και το Recall σε σχέση με το Precision μας δίνει καλύτερα αποτελέσματα.



Εικόνα 17 Precision και Recall θετικών, αρνητικών και ουδέτερων συναισθημάτων του Weka εφαρμόζοντας τον αλγόριθμο Naïve Bayes



Εικόνα 18 Precision και Recall θετικών, αρνητικών και ουδέτερων συναισθημάτων του Weka εφαρμόζοντας τον αλγόριθμο SVM

Τα αποτελέσματα του Weka μετά την εφαρμογή των αλγορίθμων Naïve Bayes και SVM σε 10 fold- validation έχουν πολύ μικρή απόκλιση στις μετρικές Ανάκληση και Ακρίβεια στο αρνητικό και θετικό συναίσθημα. Θα μπορούσαμε να πούμε ότι μας δίνουν τις ίδιες τιμές. Στο ουδέτερο συναίσθημα έχουμε υψηλότερα αποτελέσματα στο Precision και στους δύο αλγορίθμους.

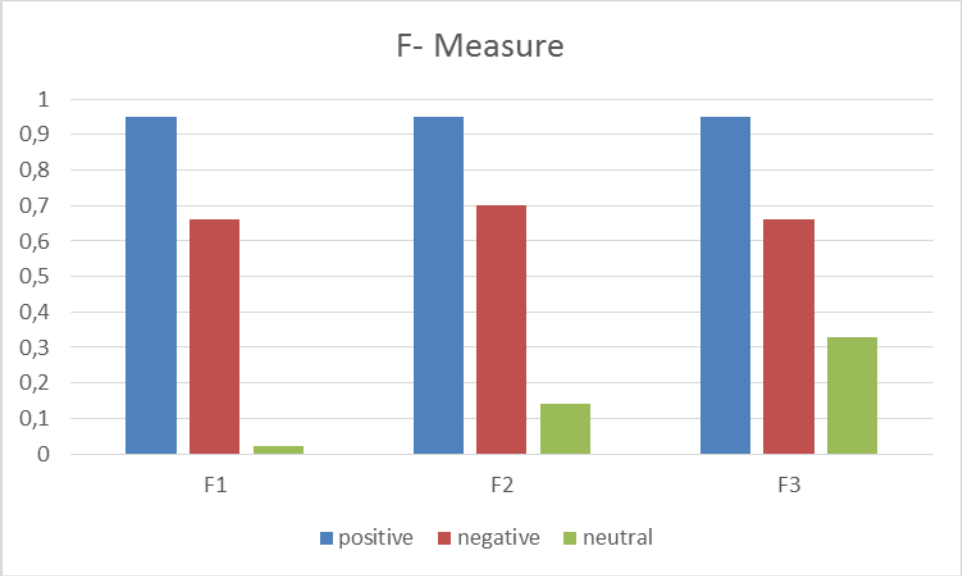
Αν υποθέσουμε ότι F1 είναι η τιμή του F-Measure μετά τους υπολογισμούς μας:

F2: τιμή F-Measure μετά την εφαρμογή στο Weka του αλγόριθμου Naïve Bayes

F3: τιμή F-Measure μετά την εφαρμογή στο Weka του αλγόριθμου SVM τότε αποτυπώνεται το παρακάτω διάγραμμα. Παρατηρούμε ότι τα F-Measure είναι περίπου στις ίδιες τιμές ανεξάρτητα τη μέθοδο και τον αλγόριθμο στο θετικό συναίσθημα **στο αρνητικό συναίσθημα** παρότι είναι μικρότερη τιμή σε σχέση με το θετικό συναίσθημα μας δίνει τα **καλύτερα αποτελέσματα** ο **αλγόριθμος Naïve Bayes στο Weka και παρατηρούμε ότι στο αρνητικό συναίσθημα η τιμή του F- Measure είναι ίδια μετά τη μεταμάθηση στο Aylien και στο S.V.M στο Weka** . Στο ουδέτερο συναίσθημα έχουμε πολύ χαμηλό ποσοστό F-Measure αλλά την καλύτερη τιμή μας τη δίνει αλγόριθμος S.V.M. του Weka.

F-MEASURE	POSITIVE	NEGATIVE	NEUTRAL
F1	0,95	0,66	0,02
F2	0,95	**0,70**	0,14
F3	0,95	0,66	0,33

Πίνακας 3 Αποτελέσματα F-MEASURE στις κλάσεις Positive, Negative & Neutral μετά την μεταμάθηση στο Aylien και στους αλγορίθμους Naive Bayes και S.V.M. στο WEKA

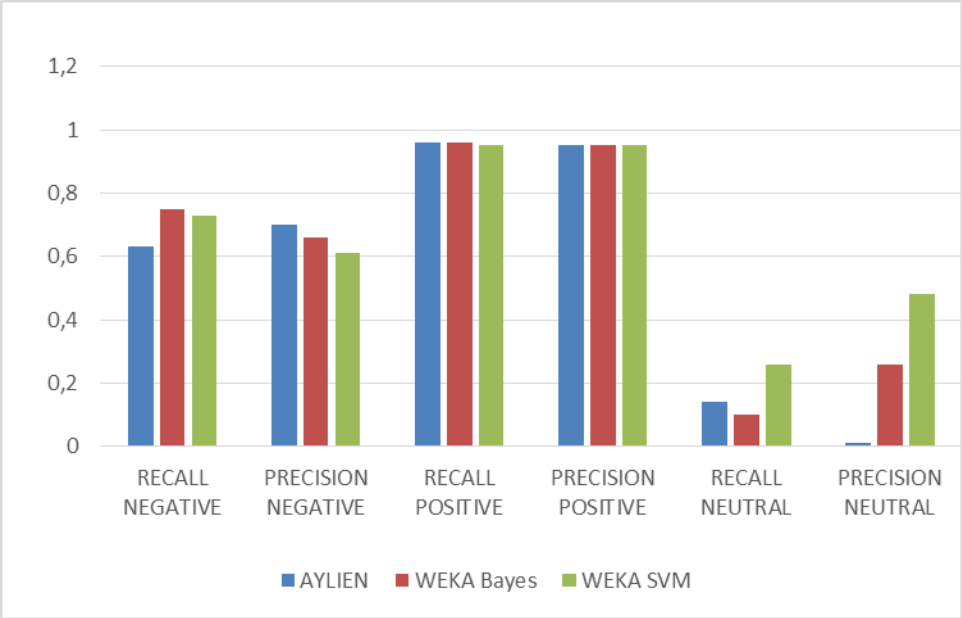


Εικόνα 19 Αποτελέσματα θετικής, αρνητικής και ουδέτερης αξιολόγησης F-Measure

Συγκεντρωτικός Πίνακας Αποτελεσμάτων θετικού και αρνητικού συναισθήματος των μετρικών RECALL και PRECISION του AYLIEN και των αλγορίθμων Bayes και SVM του WEKA

	RECALL NEGATIVE	PRECISION NEGATIVE	RECALL POSITIVE	PRECISION POSITIVE	RECALL NEUTRAL	PRECISION NEUTRAL
AYLIEN	0,63	0,70	0,96	0,95	0,14	0,01
WEKA Bayes	0,75	0,66	0,96	0,95	0,10	0,26
WEKA SVM	0,73	0,61	0,95	0,95	0,26	0,48

Πίνακας 4 Αποτελέσματα RECALL και PRECISION στο NEGATINE, POSITIVE και NEUTRAL



Εικόνα 20 Ποσοστά RECALL και PRECISON αρνητικού, θετικού και ουδέτερου συναισθήματος στο AYLIEN και στους αλγορίθμους Naïve Bayes στο WEKA και SVM στο WEKA

Παρατηρούμε ότι στο θετικό συναίσθημα τα αποτελέσματα είναι σε υψηλό ποσοστό και περίπου το ίδιο στους αλγορίθμους Naive Bayes και S.V.M. του WEKA και στο Aylien γύρω στο μηδέν 0,95. Άρα τα αποτελέσματά μας είναι αξιόπιστα και μας δίνουν την ίδια τιμή.

Σύμφωνα με το παραπάνω διάγραμμα τα ποσοστά Ανάκλησης και Ακρίβειας του αρνητικού συναισθήματος παρουσιάζουν μικρές τιμές. Όσον αφορά την Ακρίβεια η οποία είναι η μετρι-

κή που απαντά στο ερώτημα αν τα αρνητικά είναι πραγματικά σωστά διαπιστώνουμε ότι στο Aylien η τιμή είναι υψηλότερη σε σχέση με τους αλγορίθμους στο WEKA.

Επομένως τα αποτελέσματα που παίρνουμε στο Aylien μετά την μεταμάθηση είναι πιο αξιόπιστα στο Precision από αυτά των αλγορίθμων του WEKA.

Στη μετρική της Ανάκλησης τα αποτελέσματα των αλγορίθμων του WEKA έχουν υψηλότερο ποσοστό σε σχέση με τα αποτελέσματα του Aylien μετά την μεταμάθηση. Μάλιστα το μεγαλύτερο ποσοστό μας το δίνει ο αλγόριθμος Bayes (0,75).

5

Συμπεράσματα-Συζήτηση

5.1 Συμπεράσματα

Στόχος της εργασίας είναι η έρευνα στο πεδίο της Ανάλυσης Συναισθήματος και η ανάπτυξη ενός μοντέλου, για την εύρεση του συναισθηματικού προσανατολισμού σε απόψεις από το Tripadvisor. Η εξαγωγή συναισθήματος είναι ένα ερευνητικό πεδίο με μεγάλο ενδιαφέρον στην επιστημονική κοινότητα και στη βιομηχανία λόγω της πληθώρας των εφαρμογών. Η εξαγωγή συναισθήματος από μία πλατφόρμα όπως το Tripadvisor είναι ακόμα δυσκολότερο πρόβλημα λόγω της ιδιαιτερότητας του μέσου. Η γλώσσα είναι ένα από τα σημαντικότερα χαρακτηριστικά που κάνουν τον άνθρωπο να ξεχωρίζει από τα υπόλοιπα ζώα. Η δημιουργία επομένως, αλγορίθμων ή μοντέλων που μπορούν να κατανοήσουν φυσική γλώσσα είναι ένα από τα καθοριστικά βήματα για τη δημιουργία ευφυών μηχανών όπου θα μπορούν να εξάγουν συναίσθημα από κείμενο χωρίς την ανθρώπινη παρέμβαση και χωρίς λάθη.

Βασικοί στόχοι της εργασίας είναι:

- Η παρουσίαση σε θεωρητικό πλαίσιο της Ανάλυσης Συναισθήματος και των Τεχνικών Ανάλυσης Συναισθήματος.
- Αναφορά και μελέτη της πλατφόρμας tripadvisor.
- Αναφορά εφαρμογής WEKA

- Υπηρεσίες πλατφόρμας AYLIEN και εφαρμογής της στις εκφράσεις χρηστών στο tripadvisor για την εξαγωγή συναισθήματος.
- Τεχνικές Μηχανικής Μάθησης για την ανάπτυξη μοντέλων δημοφιλή σε προβλήματα επεξεργασίας φυσικής γλώσσας.
- Ανάπτυξη μοντέλου εξαγωγής συναισθήματος από την πλατφόρμα Tripadvisor με χρήση της πλατφόρμας Aylien και εφαρμογή μεταμάθησης. Σύγκριση αποτελεσμάτων σε επίπεδο Ανάκλησης (Recall) και Ακρίβειας (Precision) μετά την μεταμάθηση του Aylien και των αλγορίθμων Μηχανικής Μάθησης Naive Bayes και Support Vector Machine στο WEKA.
- Αξιολόγηση των αποτελεσμάτων.

Σκοπός αυτής της εργασίας ήταν η αξιολόγηση σε επίπεδο Ανάκλησης (Recall) και Ακρίβειας (Precision) της εξαγωγής συναισθήματος από την διαδικτυακή πλατφόρμα Tripadvisor σχετικά με εστιατόρια της Κέρκυρας το έτος 2018 για τα οποία έχουν εκφραστεί απόψεις μέσω της συγκεκριμένης πλατφόρμας. Στα αποτελέσματα που είχαμε από το Aylien εφαρμόστηκε μεταμάθηση, με άλλα λόγια stacking. Δηλαδή στα αποτελέσματα που αφορούν συναίσθημα μετά την εξαγωγή τους από την πλατφόρμα AYLIEN και είχαν λανθασμένη ταξινόμηση, επισημάνθηκαν λέξεις και εκφράσεις οι οποίες θεωρήθηκαν υπεύθυνες για τη λανθασμένη ταξινόμηση. Κάνοντας εκπαίδευση με αυτές τις λέξεις δημιουργήθηκε μια στήλη με την ορθή ταξινόμηση. Όλα τα προηγούμενα δεδομένα μετατράπηκαν σε csv και μετά την εισαγωγή των δεδομένων αυτών στο Weka εφαρμόστηκαν δύο διαφορετικοί αλγόριθμοι (Naïve Bayes και SVM). Υπολογίστηκαν σε όλες τις παραπάνω περιπτώσεις η Ανάκληση, η Ακρίβεια και το F- Measure κάθε φορά στο αρνητικό, θετικό και ουδέτερο συναίσθημα.

Παρατηρήθηκε ότι τα αποτελέσματα που υπήρχαν μετά την μεταμάθηση στις παραπάνω μετρικές δεν ήταν τα ίδια με τα αποτελέσματα των αλγορίθμων στο WEKA.

Τα ποσοστά Ανάκλησης και Ακρίβειας που αφορούσαν το θετικό συναίσθημα ήταν σε υψηλό επίπεδο και περίπου στην ίδια τιμή ενώ αντίθετα αυτά που αφορούσαν το αρνητικό συναίσθημα είχαν μικρότερο ποσοστό και διακυμάνσεις στις τιμές.

Σύμφωνα με τα αποτελέσματα που αφορούν την ουδέτερη κλάση θα μπορούσαμε να πούμε ότι τα ουδέτερα ταξινομημένα αποτελέσματα μας δίνουν πολύ μικρές τιμές τέτοιες που δεν επηρεάζουν το ολικό αποτέλεσμα. Ως ουδέτερες απόψεις δε προσφέρουν κάποιο συναίσθημα (δεν έχουν συναίσθημα).

Μετά την μεταμάθηση τα αποτελέσματα που έχουμε στο Aylien και εφαρμόζοντας τους αλγόριθμούς μας έχουν μικρό ποσοστό Recall και Precision στο αρνητικό συναίσθημα. Επιβεβαιώθηκε επομένως ένα φαινόμενο το οποίο συναντάμε στην έρευνα και στη βιβλιογραφία και έχει να κάνει με τις αρνητικές εκφράσεις και κατ' επέκταση με το αρνητικό συναίσθημα.

Ένα άλλο σημαντικό συμπέρασμα είναι το γεγονός ότι η συλλογή και η προεπεξεργασία των δεδομένων παίζει μεγάλο ρόλο στα αποτελέσματα των αλγορίθμων ταξινόμησης. Όταν δεδομένα που θα αναλυθούν είναι πολύ συγκεκριμένα γύρω από ένα θέμα τόσο μεγαλύτερη πιθανότητα υπάρχει να φτιαχτεί ένας classification αλγόριθμος με υψηλή ακρίβεια και ανάκληση.

Στην παρούσα εργασία κάναμε μια εκτενή έρευνα σχετικά τις απόψεις και το συναίσθημα που εξάγουμε από κοινωνικά δίκτυα και συγκεκριμένα τις απόψεις που αποτυπώθηκαν στο TripAdvisor για τα εστιατόρια στην Κέρκυρα για το έτος 2018.

Χρησιμοποιήθηκαν δύο διαφορετικές πλατφόρμες για την εξαγωγή του συναισθήματος, καθώς και διάφοροι τρόποι εξαγωγής, όπως είναι το deep learning, μέθοδοι NLP και κάποιοι αλγόριθμοι του Weka.

Εφαρμόστηκε μεταμάθηση στα αποτελέσματα που είχαμε από το Aylien και στη συνέχεια με αυτά τα δεδομένα τρέξαμε στο Weka τους αλγορίθμους Naïve Bayes και S.V.M.

Τα αποτελέσματα που παρουσιάστηκαν στο προηγούμενο κεφάλαιο σχετικά με τις θετικές απόψεις που έχουμε από το πείραμά μας με τη μεταμάθηση, τον αλγόριθμο Bayes στο Weka και τον SMO στο Weka όσον αφορά στο Recall και το Precision είναι στην ίδια τιμή.

Τα αποτελέσματα σχετικά με τις αρνητικές απόψεις που έχουμε από το πείραμά μας με τη μεταμάθηση, τον αλγόριθμο bayes στο Weka και τον SMO στο Weka όσον αφορά στο Recall και το Precision δεν είναι στην ίδια τιμή, υπάρχουν αποκλίσεις.

Από τις εικόνες των διαγραμμάτων και τις τιμές που έχουμε φαίνεται ότι για τις θετικές απόψεις τα αποτελέσματά μας είναι ίδια σε όλες τις εκφάνσεις του πειράματος και για όλους τους αλγορίθμους που εφαρμόσαμε. Αντίθετα, τα αποτελέσματα από τις αρνητικές εκφράσεις στους αλγορίθμους Bayes και SVM είναι παρεμφερή ενώ στους δικούς μας υπολογισμούς πριν την εισαγωγή των δεδομένων στο Weka παρατηρούμε ότι το Precision έχει μικρό ποσοστό (0,70) και ανάλογα το Recall (0,63).

Παρατηρήσαμε ότι **τα ποσοστά Ανάκλησης και Ακρίβειας του αρνητικού συναισθήματος** παρουσιάζουν μικρές τιμές. Όσον αφορά την **Ακρίβεια** η οποία είναι η μετρική που απαντά στο ερώτημα αν τα αρνητικά είναι πραγματικά σωστά διαπιστώνουμε ότι στο **Aylien η τιμή είναι υψηλότερη** σε σχέση με τους αλγορίθμους στο WEKA.

Επομένως τα αποτελέσματα που παίρνουμε στο Aylien μετά την μεταμάθηση είναι πιο αξιόπιστα στο Precision από αυτά των αλγορίθμων του WEKA.

Στη μετρική της **Ανάκλησης** τα αποτελέσματα των αλγορίθμων του WEKA έχουν υψηλότερο ποσοστό σε σχέση με τα αποτελέσματα του Aylien μετά την μεταμάθηση. Μάλιστα το μεγαλύτερο ποσοστό μας το δίνει ο **αλγόριθμος Bayes (0,75).**

Η τιμή του F-Measure που προέκυψε από τις τεχνικές και τους αλγορίθμους μας δείχνει την ισορροπία που δημιουργείται μεταξύ Recall και Precision και στο συγκεκριμένο πείραμα μας έδειξε ότι είναι παντού το ίδιο στο θετικό συναίσθημα, με πολύ μικρή απόκλιση στο αρνητικό συναίσθημα και καλύτερα αποτελέσματα στον αλγόριθμο Naïve Bayes του Weka ενώ στην ουδέτερη ταξινόμηση τα καλύτερα αποτελέσματα της μετρικής F-Measure μας τα δίνει ο αλγόριθμος S.V.M. του Weka.

Τα παραπάνω συμπεράσματα όσον αφορά το αρνητικό συναίσθημα μπορεί να οφείλονται λόγω του ότι η άρνηση σαν συναίσθημα είναι πιο δύσκολο να εκφραστεί γραπτώς. Επίσης, υπάρχει ανομοιογένεια λόγω κουλτούρας, γλώσσας κτλ. για το τι εννοούμε αρνητικό συναίσθημα. Το θετικό συναίσθημα φαίνεται να είναι πιο εύκολα αναγνωρίσιμο από τις πλατφόρμες που χρησιμοποιήσαμε. Είναι πιο εύκολο να ταξινομηθεί και από το Aylien και από το Weka (για τους αλγορίθμους bayes και SMO) και μας εμφάνισε σαφώς καλύτερα αποτελέσματα για το σετ δεδομένων και σχολίων που επιλέξαμε, σε αντίθεση με το αρνητικό συναίσθημα, το οποίο εμφανίζει δυσκολία στην σωστή ταξινόμηση. Το ουδέτερο συναίσθημα στην συγκεκριμένη περίπτωση δεν μας δίνει πληροφορίες, είναι πολύ μικρό το ποσοστό της ταξινόμησής του και θα μπορούσαμε να πούμε ότι δεν επηρεάζει τα αποτελέσματά μας.

5.2 Συζήτηση

Μετά την παρούσα εργασία θα έπρεπε να μας απασχολήσει το γεγονός του μικρού ποσοστού Ανάκλησης και Ακρίβειας που παρουσιάζουν οι εκφράσεις των αρνητικών συναισθημάτων και αφού έχουμε εφαρμόσει μεταμάθηση.

Τα παραπάνω συμπεράσματα όσον αφορά το αρνητικό συναίσθημα μπορεί να οφείλονται λόγω του ότι η άρνηση σαν συναίσθημα είναι πιο δύσκολο να εκφραστεί γραπτώς. Επίσης, υπάρχει ανομοιογένεια λόγω κουλτούρας και γλώσσας για το τι εννοούμε αρνητικό συναίσθημα. Το δείγμα μας είναι κοινό μεγάλου εύρους ηλικίας, διαφορετικών εθνοτήτων οι οποίοι εκφράζουν το συναίσθημά τους στην αγγλική γλώσσα που είναι πιθανό να μην είναι η μητρική τους και οι εκφράσεις τους να μην μεταφέρονται σωστά από τη μία γλώσσα στην άλλη.

Η περίπτωση της αποτύπωσης του αρνητικού συναισθήματος στο γραπτό λόγο είναι δυσκολότερη γιατί τις περισσότερες φορές προσπαθούμε να είμαστε ευγενικοί όταν εκφράζουμε κάτι αρνητικό. Στην Ανάλυση Συναισθήματος σε αυτήν την περίπτωση έρχεται σε σύγκρουση η ευγένεια που είναι κάτι θετικό και το αρνητικό συναίσθημα. Οι άνθρωποι τείνουν να χρησιμοποιούν πιο συχνά «θετικές» λέξεις συνήθως είναι περισσότερες σε πλήθος από τις «αρνητικές» και κατά συνέπεια ο προσανατολισμός ενός κειμένου δεν αντιστοιχεί στον πραγματικό. Αυτή μάλιστα είναι και μία από τις προκλήσεις της ευρύτερης περιοχή της Επε-

ξεργασίας Φυσικής Γλώσσας η οποία μαζί με την ειρωνεία και το σαρκασμό δυσκολεύουν την εξαγωγή ορθών αποτελεσμάτων.

Θα μπορούσαμε να σκεφτούμε πως τα χαρακτηριστικά που επιλέξαμε για τη μεταμάθηση δεν ήταν αυτά που τελικά θα βελτίωναν το αποτέλεσμα στα ποσοστά Ανάκλησης και Ακρίβειας του αρνητικού συναισθήματος.

Το ερώτημα που προκύπτει από το γεγονός αυτό αφορά τα αίτια αυτού του αποτελέσματος και με ποιον τρόπο θα πρέπει μελλοντικά να αντιμετωπιστούν.

Η φυσική γλώσσα είναι αδόμητη και δεν είναι εύκολο να μοντελοποιηθεί. Συνεπώς το νόημα που φέρει μια πρόταση είναι δύσκολο να γίνει κατανοητό από ένα μηχάνημα. Η παρούσα εργασία θα μπορούσε να επεκταθεί σε άλλη έρευνα σχετικά με την ορθότερη ταξινόμηση των αρνητικών συναισθημάτων και κατ' επέκταση να προταθούν τρόποι βελτίωσης των αλγορίθμων για να μπορεί να προκύπτουν μεγαλύτερη ανάκληση και ακρίβεια.

6

Βιβλιογραφία

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.

Aggarwal C., Zhai Cheng Xiang (2012). *A survey of text classification Algorithms, Mining Text Data Book,* pp 163-213 ISDN978-1-4614-3222-7 e-Isbn 978-1-4614-3223-4 Springer.

Aghaei, S., Nematbakhsh, M. A. & Farsani, H. K. (2012). Evolution of the world wide web: from Web 1.0 to Web 4.0. *International Journal of Web & Semantic Technology,* 3(1), 1-10.

Ahmed, S. R. (2004, April). Applications of data mining in retail business. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on* (Vol. 2, pp. 455-459). IEEE.

BBC Radio 4: The Bottom Line. (2014). "Interview with Stephen Kaufer".. October 16, 2014.

Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse processes, 11*(1), 1-34.

Biggs, J. B. (1985). The role of metalearning in study processes. *British journal of educational psychology*, *55*(3), 185-212.

Bishop, C. M. (2006). Graphical models. *Pattern recognition and machine learning*, *4*, 359-422.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems* (No. RR-150). YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.

Cardie, C., Wiebe, J., Wilson, T., & Litman, D. J. (2003). Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. In *New directions in question answering* (pp. 20-27).

Chafe, W. L., & Nichols, J. (1986). Evidentiality: The linguistic coding of epistemology.

Chelmis, C., & Prasanna, V. K. (2011, October). Social networking analysis: A state of the art and the effect of semantics. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 531-536). IEEE.

Conrad, S., & Biber, D. (2000). Adverbial marking of stance in speech and writing. *Evaluation in text: Authorial stance and the construction of discourse*, 56-73.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Custom Survey Research Engagement. (2015). Tripadvisor.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In*Proceedings of the 12th international conference on World Wide Web* (pp. 519-528). ACM.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *Icml* (Vol. 96, pp. 148-156).

Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the Stars: Improving Rating Predictions using Review Text Content. In *WebDB* (Vol. 9, pp. 1-6).

Georgiadis, C., & Γεωργιάδης, Χ. (2015). Τεχνολογίες παγκόσμιου ιστού και ηλεκτρονικού εμπορίου.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Harrington, P. (2012). Machine learning in action. *Shelter Island, NY: Manning Publications Co*.

Hearst, M. A. (1992). Direction-based text interpretation as an information access refinement. *Text-based intelligent systems: current research and practice in information extrac-*

tion and retrieval, 257-274.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Huettner, A., & Subasic, P. (2000). Fuzzy typing for document management.*ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, 26-27.

Indurkhya, N., & Damerau, F. J. (Eds.). (2010). *Handbook of natural language processing* (Vol. 2). CRC Press.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kantrowitz, M. (2003). *U.S. Patent No. 6,622,140*. Washington, DC: U.S. Patent and Trademark Office.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, *22*(2), 110-125.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.

Klösgen, W., & Zytkow, J. M. (2002, January). The knowledge discovery process. In *Handbook of data mining and knowledge discovery* (pp. 10-21). Oxford University Press, Inc.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Langacker, R. W. (1985). Observations and speculations on subjectivity. *Iconicity in syntax*, *1*(985), 109.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lyons, J. (1981). *Language and linguistics*. Cambridge University Press.

Marsden, P. V., & Campbell, K. E. (1984). Measuring tie strength. *Social forces*, *63*(2), 482-501.

Martin, J. R., & White, P. R. (2005). Appraisal in English.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes

text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

McLachlan, G., Do, K. A., & Ambroise, C. (2005). *Analyzing microarray gene expression data* (Vol. 422). John Wiley & Sons.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

Mitchell, T. M. (1997). Machine learning. WCB.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77). ACM.

Ogneva, M. (2010). How Companies Can Use Sentiment Analysis to Improve Their Business.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Parmanto, B., Munro, P. W., & Doyle, H. R. (1996). Reducing variance of committee prediction with resampling techniques. *Connection Science*, *8*(3-4), 405-426.

Pavlopoulos, J., & Androutsopoulos, I. (2014, April). Multi-granular aspect aggregation in aspect-based sentiment analysis. In *Proceedings of EACL* (pp. 78-87).

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications* (pp. 1-10). Springer, Dordrecht.

Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S. (2014, August). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 27-35).

Poria, S., Cambria, E., Gelbukh, A., Bisio, F., & Hussain, A. (2015). Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine*, *10*(4), 26-36.

Poria, S., Cambria, E., Winterstein, G., & Huang, G. B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*, 45-63.

Radovanović, M., & Ivanović, M. (2008). Text mining: Approaches and applications. *Novi Sad J. Math*, *38*(3), 227-234.

Sack, W. (1994, July). On the computation of point of view. In *AAAI* (p. 1488).

Sahami Mehran κ.α. (1998*) «A Bayesian Approach to Filtering Junk E-Mail»*. Στο:Learning for Text Categorization: Papers from the 1998 Workshop. Τομ. 62 σσ.98-105.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

Scheibman, J. (2002). *Point of view and grammar: Structural patterns of subjectivity in American English conversation* (Vol. 11). John Benjamins Publishing.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Suresh, V., Roohi, S., & Eirinaki, M. (2014, October). Aspect-based opinion mining and recommendationsystem for restaurant reviews. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 361-362). ACM.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tatemura, J. (2000, January). Virtual reviewers for collaborative exploration of movie reviews. In *Proceedings of the 5th international conference on Intelligent user interfaces* (pp. 272-275). ACM.

Terveen, L., Hill, W., Amento, B., McDonald, D., & Creter, J. (1997). PHOAKS: A system for sharing recommendations. *Communications of the ACM*, *40*(3), 59-62.

Tim, O. (2005). What is web 2.0? design patterns and business models for the next generation of software.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (Vol. 1, p. 6).

Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., & Hao, H. (2015). Semantic Clustering and Convolutional Neural Network for Short Text Categorization. Proceedings ACL 2015, 352–357.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20*(2), 233-287.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999, June). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting*

of the Association for Computational Linguistics on Computational Linguistics (pp. 246-253). Association for Computational Linguistics.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B.,.... & Maybury, M. T. (2003, March). Recognizing and Organizing Opinions Expressed in the World Press. In *New Directions in Question Answering* (pp. 12-19).

Wiebe, J. M., & Rapaport, W. J. (1988, June). A computational theory of perspective and reference in narrative. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics* (pp. 131-138). Association for Computational Linguistics.

Wilks, Y., & Bien, J. (1983). Beliefs, Points of View, and Multiple Environments*. *Cognitive Science*, *7*(2), 95-119.

YALSA (Young Adult Library Services Association), «Teens & social networking in school & public libraries: a toolkit for librarians & library workers: updated and expanded June 2009», Young adult library services association, 2009, σελ. 1-11.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *null* (p. 427). IEEE.

Yu, H., & Hatzivassiloglou, V. (2003, July). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 129-136). Association for Computational Linguistics.

Καρποδίνης, Κ. (2016) Ανάλυση συναισθημάτων σε δεδομένα από το Twitter χρησιμοποιώντας εργαλεία της R και μοντέλα μηχανικής μάθησης. Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών, Πανεπιστήμιο Πατρών.

Μαστρογιάννης, Ν.(2009) Μεθοδολογικό Πλαίσιο Υποστήριξης της Εξόρυξης Γνώσης από δεδομένα με την χρήση αρχών της πολυκριτήριας ανάλυσης αποφάσεων. Τμήμα Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Πατρών.

Χριστοπούλου, Κ. (2017) Ανάλυση συναισθήματος με χρήση τεχνικών μηχανικής μάθηση. Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

https://developer.aylien.com

http://jarche.com/.

https://www.tripadvisor.com/.

http://www.cs.waikato.ac.nz/ml/weka.

https://www.yelp.ie