ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΑΚΑΔΗΜΑΪΚΟΥ ΕΤΟΥΣ 2022-2023

ΣΤΡΑΤΙΩΤΙΚΗ ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ

Τμήμα Στρατιωτικών Επιστημών

ΕΥΦΥΗ ΣΥΣΤΗΜΑΤΑ

MASTER OF SCIENCE IN INTELLIGENT SYSTEMS

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μη επιβλεπόμενη προσέγγιση Ανάλυσης Συναισθήματος: η περίπτωση του Παλιού Λιμανιού των Χανίων

Μεταπτυχιακή φοιτήτρια *:* Παπαδοπούλου Ελένη

Α.Μ. *:* 2021018008

ΜΑΡΤΙΟΣ 2023

Η Μεταπτυχιακή Διατριβή της Παπαδοπούλου Ελένης εγκρίνεται:

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Αναπληρωτής Καθηγητής ΣΤΕΛΙΟΣ ΤΣΑΦΑΡΑΚΗΣ (Επιβλέπων) Stelios Tsafarakis Digitally signed by Stelios Tsafarakis DN: c=GR, l=Chania, o=Technical University of Crete, ou=Class B - Private Key created and stored in software CSP, sn=Tsafarakis, givenName=Stelios, serialNumber=4644645843, cn=Stelios Tsafarakis, email=tsafarakis@dpem.tuc.gr Date: 2023.06.16 19:30:01 +03'00'

Καθηγητής ΝΙΚΟΛΑΟΣ ΜΑΤΣΑΤΣΙΝΗΣ Nikolaos Matsatsinis Digitally signed by Nikolaos Matsatsinis Date: 2023.06.17 08:49:37 +03'00'

Καθηγητής ΝΙΚΟΛΑΟΣ ΠΑΠΑΔΑΚΗΣ Nikolaos Papadakis Signed by: Nikolaos Papadakis Date: 2023-06-19 08:47:46 EEST

ΣΕΛΙΔΑ ΣΚΟΠΙΜΑ ΚΕΝΗ

Έτος 2023

ΕΥΧΑΡΙΣΤΙΕΣ

Πρώτα, θα ήθελα να ευχαριστήσω θερμά τον κ. Τσαφαράκη Στέλιο, Αναπληρωτή Καθηγητή της Σχολής Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου το παρόν θέμα, τη βοήθεια και διάθεσή του να προσφέρει τις γνώσεις και τις συμβουλές του καθ' όλη την εξέλιξη της εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω στον κ. Κυριακίδη Αναστάσιο, Υποψήφιο Διδάκτορα της Σχολής Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης που μου παρείχε τα δεδομένα προς διερεύνηση από το Tripadvisor. Με καθοδήγησε όσον αφορά το πλαίσιο μελέτης του αντικειμένου και ήταν πρόθυμος ανά πάσα χρονική στιγμή για να προσφέρει τις γνώσεις του και να λύσει τυχόν απορίες μου.

Τέλος, θέλω να ευχαριστήσω τους δικούς μου ανθρώπους και ιδιαίτερα τους γονείς μου, τα αδέρφια μου και τη νονά μου που με στηρίζουν και πιστεύουν σε μένα. Η εργασία και όλη μου η προσπάθεια είναι αφιερωμένη στους παππούδες και τις γιαγιάδες μου.

ΣΕΛΙΔΑ ΣΚΟΠΙΜΑ ΚΕΝΗ

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Αναντίρρητα, οι ανθρώπινες γνώμες κατέχουν θέση-κλειδί στη διαμόρφωση της ατομικής και συλλογικής συμπεριφοράς, καθώς οι πεποιθήσεις, οι αντιλήψεις, οι επιλογές και οι αποφάσεις εξαρτώνται σε αρκετά μεγάλο βαθμό από το πώς οι υπόλοιποι αντιλαμβάνονται και αξιολογούν τα γεγονότα που συμβαίνουν. Σημαντικό ρόλο στη διάχυσή τους έπαιξαν και συνεχίζουν να παίζουν οι τεχνολογίες του Διαδικτύου οι οποίες δίνουν τη δυνατότητα σε χρήστες από κάθε μέρος του πλανήτη να εκφράζονται ελεύθερα σχετικά με ένα προϊόν ή υπηρεσία. Η τεράστια διαθεσιμότητα και χρήση κειμενικών βάσεων δεδομένων κατέστησε εξαιρετικά χρήσιμη την αυτοματοποιημένη ανάλυση συναισθήματος, υπό την έννοια ότι αυτή μπορεί να ανιχνεύσει κάθε είδους απόψεις.

Στόχος, λοιπόν, της παρούσας διπλωματικής είναι να αξιολογηθεί ως προς την απόδοσή της, μία μη επιβλεπόμενη μεθοδολογία της βιβλιογραφίας για την ταξινόμηση της βαθμολογίας διαδικτυακών κριτικών του Tripadvisor που αφορούν το παλιό λιμάνι των Χανίων Κρήτης. Αυτή επιχειρείται να υλοποιηθεί διαμέσου της ανάλυσης συναισθήματος βάσει των χαρακτηριστικών στα οποία γίνεται αναφορά. Η απόδοσή της συγκρίνεται με άλλες εναλλακτικές προσεγγίσεις που διαφοροποιούνται από τη βασική, στις τεχνικές στάθμισης και τον τρόπο κατασκευής του λεξικού συναισθήματος.

Η εργασία δομείται από 10 βασικά κεφάλαια. Τα τρία πρώτα παρουσιάζουν το ευρύτερο πλαίσιο ανάλυσης συναισθήματος και ειδικότερα την ανάλυση συναισθήματος βάσει χαρακτηριστικών. Επίσης, αναλύεται η μεθοδολογία που έχει ακολουθηθεί από διάφορους ερευνητές, στο πέρασμα του χρόνου, με σκοπό την παραγωγή αξιόπιστων και ακριβών μεθόδων ταξινόμησης δεδομένων σε μορφή κειμένου. Τα κεφάλαια 4, 5 και 6 πραγματεύονται το θεωρητικό υπόβαθρο που αφορά την εξαγωγή των χαρακτηριστικών από μία συλλογή κειμένου και την αναπαράσταση των λέξεων με το Word2vec μοντέλο, παραθέτοντας, τελικά, τον τρόπο δημιουργίας ορισμένων λεξικών συναισθηματικής πολικότητας. Ακολουθεί το κεφάλαιο 7 το οποίο αφιερώνεται στην πειραματική εφαρμογή, από τη διερεύνηση των δεδομένων αναφοράς και τον καθαρισμό του κειμένου από περιττές πληροφορίες, έως την αποτίμηση των αποτελεσμάτων ταξινόμησης. Είναι σημαντικό να επισημανθεί ότι τα συμπεράσματα όποιας μεθοδολογίας εφαρμόζεται, επεξηγούνται στο Κεφάλαιο 8, όπου και γίνεται λόγος για προτάσεις προς μελλοντική μελέτη. Για την περάτωση της εργασίας κατέστη πολύτιμη η χρήση του περιβάλλοντος της Python στο Google Colab όπως άλλωστε γίνεται σαφές στο Κεφάλαιο 9. Τέλος, το Κεφάλαιο 10 περιλαμβάνει όλες τις βιβλιογραφικές αναφορές στις οποίες γίνεται λόγος καθ’ όλη την έκταση της εργασίας.

1. ΕΙΣΑΓΩΓΗ

Η εμφάνιση και ανάπτυξη του Διαδικτύου, εκτός του ότι άλλαξε τον τρόπο επικοινωνίας μεταξύ των ανθρώπων, δημιούργησε ένα αυξανόμενο πλήθος ιστοσελίδων, ιστοτόπων ηλεκτρονικού εμπορίου (π.χ. Amazon), εφαρμογών ανταλλαγής μηνυμάτων (π.χ. WhatsApp) και μέσων κοινωνικής δικτύωσης (π.χ. Instagram και Facebook). Μέσω αυτών, άνθρωποι από οποιοδήποτε μέρος του πλανήτη έχουν τη δυνατότητα να αλληλεπιδρούν μεταξύ τους διαμοιράζοντας πολυμεσικό περιεχόμενο, αλλά και να έχουν πρόσβαση σε ολοένα και περισσότερες πηγές πληροφοριών παγκόσμιας εμβέλειας. Το πιο σημαντικό, όμως, γεγονός, είναι ότι οι άνθρωποι, πλέον, εμπιστεύονται ολοένα και περισσότερο σε σχέση με προηγούμενες χρονικές περιόδους, τις πληροφορίες που προσφέρονται στον Παγκόσμιο Ιστό από άλλους ανθρώπους, μη επαγγελματίες στον εκάστοτε χώρο εργασίας. Άλλωστε, λαμβάνουν υπόψιν ότι η γνώμη ενός επαγγελματία υποκινείται και από την επιθυμία του να προωθήσει το προϊόν του και κατά συνέπεια να αυξήσει τα κέρδη του (Shinde, Pawar, Ahirrao, & Phansalkar, 2019).

Ο όγκος περιεχομένου στο Διαδίκτυο, παραγόμενου οικειοθελώς από τον ίδιο τον άνθρωπο, δομημένου και μη, αυξάνεται με ταχύτατο ρυθμό, ιδιαίτερα εκείνου που είναι σε μορφή κειμένου. Με αυτόν τον τρόπο, όπως γίνεται αντιληπτό, ο τομέας της ανάλυσης μεγάλων δεδομένων που ασχολείται με την εξέταση μεγάλων βάσεων δεδομένων και την εξόρυξη χρήσιμης πληροφορίας βρίσκει άμεση εφαρμογή. Πλέον, στον Παγκόσμιο Ιστό διατίθεται ελεύθερα τεράστιος αριθμός σχολίων και κριτικών για οποιοδήποτε προϊόν ή υπηρεσία. Τα τελευταία είναι ιδιαίτερα χρήσιμα, αφού συγκεντρώνουν μεγάλο όγκο απόψεων, κυρίως υποκειμενικών, οι οποίες με τη σειρά τους είναι ικανές να επηρεάσουν τις ιδεολογικές επιλογές και την καταναλωτική συμπεριφορά τεράστιας μερίδας ανθρώπων παγκοσμίως. Επίσης, βοηθούν τους ίδιους τους παρόχους υπηρεσιών, τους πωλητές, τους κατασκευαστές προϊόντων, ακόμα και τους πολιτικούς, παρακολουθώντας τη διαμόρφωση της γνώμης των χρηστών, να προβαίνουν σε βελτιωτικές αλλαγές των υπό συζήτηση χαρακτηριστικών και να λαμβάνουν κατάλληλες αποφάσεις. Τόσο εταιρείες όσο και άτομα μπορούν να ελέγχουν άμεσα και σε πραγματικό χρόνο τη φήμη τους και να ανταποκρίνονται έγκαιρα στη θετική ή αρνητική ανταπόκριση των χρηστών (Feldman, 2013).

Από την άλλη πλευρά, αυτή η υπερπληθώρα δεδομένων σε μορφή κειμένου εμποδίζει τον μέσο χρήστη του Διαδικτύου από την εύρεση όλων των σχετικών με ένα επιθυμητό θέμα, ιστοχώρων, την ολοκληρωμένη μελέτη τους και την εξαγωγή περιληπτικής εικόνας για την κυριαρχούμενη άποψη (Liu & Zhang, A Survey of Opinion Mining and Sentiment Analysis, 2012). Υπό το πλαίσιο αυτό, καθίσταται πολύτιμη η παρουσία κάποιας αυτοματοποιημένης προσέγγισης για την εξαγωγή και ανάλυση γνώμης και συναισθήματος στις διαδικτυακές κριτικές.

2. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS)

Η ανάλυση συναισθήματος (sentiment analysis) αποτελεί το πεδίο έρευνας που περιλαμβάνει τη συγκέντρωση και μελέτη των γνώμεων και αξιολογήσεων σχετικά με διάφορες οντότητες (Nigam & Hurst, 2006), όπως επίσης και την ανάλυση και συνάθροιση του συναισθήματος το οποίο υποκρύπτεται γι'αυτές σε μία μεγάλη συλλογή κειμένου. Παραδείγματα τέτοιων οντοτήτων είναι προϊόντα, υπηρεσίες, οργανισμοί, άτομα, εκδηλώσεις και χαρακτηριστικά τους (Liu, Gao, Liu, & Zhang, 2016). Η ανάλυση συναισθήματος συναντάται στη βιβλιογραφία και με τον όρο εξόρυξη γνώμης (opinion mining), υπό την έννοια ότι επιχειρείται να αναγνωριστεί εάν η γνώμη για ένα θέμα είναι θετική ή αρνητική. Συνήθως, η εργασία της ανάλυσης συναισθήματος ενός κειμένου μετατίθεται σε πρόβλημα ταξινόμησης, είτε ανάμεσα σε δύο κλάσεις, είτε σε τρεις, στοχεύοντας στη διάκριση της πόλωσης ανάμεσα σε θετική / αρνητική ή θετική / αρνητική / ουδέτερη αντίστοιχα. Ο βασικός λόγος πίσω από τόσο λίγες κατηγορίες εντοπίζεται στην αδυναμία αποτελεσματικής ανίχνευσης όλων των συναισθημάτων του ανθρώπου σε δεδομένο κείμενο, εξαιτίας του μεγάλου αριθμού τους.

Σε αυτό το σημείο, κρίνεται σημαντικό να αναφερθεί ότι η γνώμη ορίζεται ως η θετική ή αρνητική στάση, αξιολόγηση ή συναίσθημα ενός ανθρώπου, που ονομάζεται κάτοχος γνώμης, προς μία οντότητα ή χαρακτηριστικό αυτής (Liu B. , Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2011). Οι οντότητες μπορεί να είναι από προϊόντα, υπηρεσίες και πρόσωπα, μέχρι γεγονότα, θέματα γενικού ενδιαφέροντος και χαρακτηριστικά τους (Liu B. , Sentiment Analysis and Opinion Mining, 2012). Οι Munezero, Montero et al. (Munezero, Montero, Sutinen, & Pajunen, 2014) διαφοροποίησαν την έννοια της γνώμης από το συναίσθημα. Όπως αναφέρουν, η γνώμη ενός ανθρώπου είναι η προσωπική ερμηνεία διαφόρων πληροφοριών που ο ίδιος λαμβάνει, φορτισμένη ενδεχομένως συναισθηματικά. Στον αντίποδα, το συναίσθημα που εκφράζεται για μία οντότητα παράγεται βάσει των κοινωνικών επιταγών και αναπτύσσεται σε βάθος χρόνου. Παρά τη διαφορά τους, η παγκόσμια βιβλιογραφία συχνά χρησιμοποιεί αδιακρίτως αυτούς τους δύο όρους.

Η ανάλυση συναισθήματος σε ένα σύνολο εγγράφων αποτελεί ένα από τα πιο δύσκολα ερευνητικά προβλήματα, καθώς συσχετίζεται άμεσα με την υποκειμενική παρατήρηση και σκέψη κάθε ανθρώπου. Η πολυπλοκότητα στη δομή κάθε φυσικής γλώσσας και στη λειτουργία του υπολογιστή δυσκολεύει το έργο της «μετάφρασης» του ψηφιακού κειμένου. Ταυτόχρονα, όμως, η ικανότητα εντοπισμού της υποκειμενικής διάστασης σε αυτό είναι ιδιαίτερα πολύτιμη και μπορεί να εφαρμοστεί σε διάφορους τομείς (D'Andrea, Ferri, Grifoni, & Guzzo, 2015). Χαρακτηριστικό παράδειγμα αποτελεί ο κοινωνιολογικός τομέας, όπου η διερεύνηση των κειμένων των ιστολογίων μπορεί να βοηθήσει στον εντοπισμό και την πρόληψη επικίνδυνων φαινομένων, όπως ο διαδικτυακός εκφοβισμός. Επίσης, μπορεί να εφαρμοστεί στο χρηματοοικονομικό πεδίο, αναγνωρίζοντας τόσο τη διάθεση των ανθρώπων απέναντι σε προϊόντα,

υπηρεσίες ή επιχειρήσεις όσο και τους λόγους ενδεχόμενης δυσαρέσκειάς τους, προβλέποντας την οικονομική απόδοση, π.χ. μετοχών (Cristescu, Nerisanu, Mara, & Oprea, 2022), καθορίζοντας την τιμολόγηση των προϊόντων (Archak, Ghose, & Ipeirotis, 2007) αλλά και ανιχνεύοντας το ρίσκο στο τραπεζικό σύστημα (Nopp & Hanbury, 2015). Η ανάλυση συναισθήματος μπορεί να εφαρμοστεί και στον χώρο της ψυχικής υγείας, υπό την έννοια ότι η μελέτη όσων γράφονται στο Διαδίκτυο από τους ανθρώπους βοηθάει τους ειδικούς να αντιληφθούν τη συναισθηματική κατάσταση ενός ατόμου και ενδεχομένως τις συναισθηματικές του διακυμάνσεις κατά τη διάρκεια της ημέρας απέναντι σε ορισμένα θέματα (Gaind, Syal, & Padgalwar, 2019).

Ο όρος της συναισθηματικής ανάλυσης πρωτοεμφανίστηκε το 2003, όπου αποσπάστηκε η συναισθηματική πόλωση (θετική ή αρνητική) συγκεκριμένων θεμάτων ενός εγγράφου, μέσω εκφράσεων, τόσο υπαρκτών όσο και έμμεσα εννοούμενων (Nasukawa & Yi, 2003). Βέβαια, ήδη από το 2000, υπήρχαν μελέτες χρήσης των διαδικτυακών κριτικών για την κατηγοριοποίηση του συναισθήματος που υποκρύπτονταν πίσω από αυτές (Wiebe J. , 2000), (Pang, Lee, & Vaithyanathan, 2002). Ύστερα από το μεγάλο ενδιαφέρον των ερευνητών γύρω από αυτόν τον τομέα, προέκυψαν κυρίως τρεις υπό-περιοχές, καθεμία από τις οποίες αφορά διαφορετικό επίπεδο ανάλυσης (Schouten & Frasincar, 2016).

Το πρώτο αφορά την ταξινόμηση ανάμεσα σε θετική και αρνητική της συνολικής άποψης που εκφράζεται σε δεδομένο έγγραφο το οποίο περιλαμβάνει πολλές προτάσεις. Είναι γνωστή ως ταξινόμηση συναισθήματος σε επίπεδο εγγράφου (document-level sentiment classification) και εφαρμόζεται ευρέως σε περιπτώσεις κριτικών, όπου η καθεμία αφορά μοναδική οντότητα. Οι βασικές προσεγγίσεις που ακολουθούνται για αυτόν τον σκοπό ανήκουν στο πεδίο της επιβλεπόμενης μάθησης και της μη επιβλεπόμενης μάθησης (Feldman, 2013). Στην πρώτη περίπτωση, με τη βοήθεια ενός συνόλου εγγράφων που αποτελούν το σύνολο εκπαίδευσης, το επιλεγμένο μοντέλο ταξινόμησης μαθαίνει, βάσει της αναπαράστασής του, να κατηγοριοποιεί κάθε έγγραφο σε μία ποιοτική κλάση συναισθηματικής πολικότητας. Αντίθετα, στη δεύτερη περίπτωση, επιλέγονται, αρχικά, σε κάθε έγγραφο, συγκεκριμένες λέξεις και φράσεις, είτε βάσει προκαθορισμένων μοτίβων που αφορούν τα μέρη του λόγου είτε με τη βοήθεια κάποιου λεξικού συναισθημάτων. Ύστερα, καθορίζεται ο σημασιολογικός τους προσανατολισμός με τη χρήση της κατά σημεία αμοιβαίας πληροφορίας (Pointwise Mutual Information / PMI) η οποία συσχετίζει καθεμία από αυτές με ένα πρότυπο σύνολο λέξεων, τόσο θετικού όσο και αρνητικού συναισθήματος.

Σε αυτό το σημείο, κρίνεται χρήσιμο να αποσαφηνιστούν η επιβλεπόμενη και η μη επιβλεπόμενη μάθηση. Η επιβλεπόμενη μάθηση είναι μια μέθοδος εκπαίδευσης στην οποία το μοντέλο εξασκείται σε ένα δοσμένο «ερέθισμα», δίνοντάς του την επιθυμητή «αντίδραση» σε αυτό το ερέθισμα. Αυτή η επιθυμητή αντίδραση χρησιμοποιείται για την παραγωγή ενός σήματος σφάλματος στην έξοδο, λόγω της διαφοράς που υπάρχει ανάμεσα σε αυτή και την πραγματική έξοδο του μοντέλου. Αντίθετα, στη μη

επιβλεπόμενη κατηγορία μάθησης δεν δίνονται οι επιθυμητές τιμές-στόχοι (labels), όπως στην εκπαίδευση με επίβλεψη. Αντίθετα, το εκάστοτε μοντέλο ψάχνει να βρει τάσεις ή κανονικότητα στα σήματα εισόδου, ώστε τα διανύσματα εξόδου να εμφανίζουν τα ίδια χαρακτηριστικά με τα διανύσματα εισόδου.

Στο δεύτερο επίπεδο ανάλυσης συναισθήματος, ο στόχος είναι παρόμοιος με αυτόν του πρώτου επιπέδου, μόνο που σε αυτή την περίπτωση, έγκειται στον καθορισμό του εάν μία δεδομένη πρόταση εκφράζει μία θετική, αρνητική ή ουδέτερη γνώμη (sentence-level sentiment classification) (Kim & Hovy, 2004). Σχετίζεται άμεσα, χωρίς, ωστόσο, να ταυτίζεται με την ταξινόμηση των προτάσεων σε αντικειμενικές ή υποκειμενικές, ανάλογα με το αν εκφράζονται σε αυτές, αληθινά γεγονότα ή αντίθετα υποκειμενικές απόψεις και θέσεις (Riloff & Wiebe, Learning Extraction Patterns for Subjective Expressions, 2003). Αυτού του είδους η προσέγγιση υποθέτει ότι σε κάθε πρόταση εκφράζεται γνώμη μόνο για ένα θέμα, αδυνατώντας στην κατανόηση ποικίλων παρατιθέμενων γνώμεων, ενδεχομένως και διαφορετικής συναισθηματικής κατεύθυνσης, για περισσότερες του ενός οντότητες ή χαρακτηριστικά τους. Ο σκοπός της ταξινόμησης του συναισθήματος των προτάσεων επιτυγχάνεται με παρόμοιες μεθόδους όπως στο επίπεδο ανάλυσης συναισθήματος ολόκληρων εγγράφων. Το είδος των μελετώμενων προτάσεων παίζει πολύ σημαντικό ρόλο στην ανάπτυξη μεθοδολογίας για την επιτυχημένη ανάλυσή τους, με τους ερευνητές να καταλήγουν ότι απαιτείται, σε κάθε περίπτωση, διαφορετική προσέγγιση. Για παράδειγμα, οι σαρκαστικές προτάσεις (Wang, et al., 2022) (Kumar Bhadra, Shaila, & Banga, 2022), όπως επίσης και οι υποθετικές προτάσεις (Liu B. , Dealing with Conditional Sentences, 2015) απαιτούν ειδικό χειρισμό από τις υπόλοιπες προτάσεις.

Το τρίτο επίπεδο ανάλυσης συναισθήματος που είναι και αυτό που διερευνά μία γνώμη με τη μεγαλύτερη δυνατή λεπτομέρεια, αναλύεται εκτενώς σε επόμενο κεφάλαιο. Επιλέγεται αρκετά συχνά σε πρακτικές εφαρμογές, καθώς προσφέρει σε έναν χρήστη λεπτομερή γνώση των απόψεων και των συναισθημάτων αναφορικά με συγκεκριμένα μόνο χαρακτηριστικά. Για παράδειγμα στην περίπτωση των φορητών υπολογιστών, ένας υποψήφιος πελάτης μπορεί να ενδιαφέρεται αποκλειστικά για τη διάρκεια μπαταρίας ή την ανάλυση της οθόνης.

Οι τεχνικές που χρησιμοποιούνται για την ανάλυση συναισθήματος διακρίνονται σε δύο μεγάλες κατηγορίες. Μπορούν είτε να ανήκουν στο πεδίο της μηχανικής μάθησης, είτε να βασίζονται στην ύπαρξη κάποιου λεξικού (lexicon-based). Υπάρχουν, βέβαια, και οι υβριδικές τεχνικές που συνδυάζουν τις δύο βασικές προσεγγίσεις .

Η πρώτη κατηγορία χωρίζεται κατά βάση σε επιβλεπόμενη και μη επιβλεπόμενη μάθηση και εφαρμόζεται συνήθως για την εξόρυξη γνώμης σε επίπεδο προτάσεων και χαρακτηριστικών. Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), τα Μπεϋζιανά μοντέλα (Wang & Manning, 2012; Jeyapriya & Selvi Kanimozhi, 2015), τα μοντέλα μέγιστης εντροπίας και τα νευρωνικά δίκτυα (Chen, Xu, He, & Wang, 2017; Kim Y. , 2014) είναι μερικά από τα μοντέλα που ανήκουν σε αυτήν την κατηγορία μεθοδολογίας, οπότε και εκπαιδεύονται σε επισημειωμένα δεδομένα με στόχο την

αναγνώριση της κατεύθυνσης συναισθηματικού προσανατολισμού. Παρά την υψηλή τους απόδοση, κρίνεται σημαντικό να τονιστεί ότι ιδίως τα μοντέλα επιβλεπόμενης μάθησης παρουσιάζουν αδυναμία αποτελεσματικής ταξινόμησης ενός κειμένου που ανήκει σε διαφορετικό εννοιολογικό πεδίο από αυτό στο οποίο εκπαιδεύτηκαν. Οι θετικές και αρνητικές λέξεις που αποσπώνται από έναν ταξινομητή μετά από την εκπαίδευσή του ενδέχεται να είναι επιζήμιες στην εξαγωγή αντιπροσωπευτικής τιμής συναισθηματικού προσανατολισμού από ένα άλλο κείμενο που πραγματεύεται άλλο αντικείμενο (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Τα χαρακτηριστικά που χρησιμοποιούνται για την ταξινόμηση ποικίλουν και μπορεί να είναι από απλά μονογράμματα και διγράμματα μέχρι και αυτά που φέρουν και τις ετικέτες των μερών του λόγου (Salvetti & Reichenbach, 2006).

Κατά τη δεύτερη προσέγγιση, η συνολική πολικότητα του κειμένου εξαρτάται σε μεγάλο βαθμό από την πολικότητα των λέξεων και φράσεων που υπάρχουν σε αυτό, οτιδήποτε και αν είναι αυτές (ουσιαστικά, επίθετα, επιρρήματα και ρήματα) (Benamara, Cesarano, Picariello, Recupero, & Subrahmanian, 2007; Vermeij). Οπότε, συνδυάζει τις τιμές συναισθηματικής πολικότητας προερχόμενες από ένα λεξικό συναισθημάτων, για όσες λέξεις εντοπίζονται σε δεδομένο κείμενο, παράγοντας, εν τέλει μία μοναδική τιμή (Muhammad, Wiratunga, & Lothian, 2016). Αυτό το λεξικό, παρ' όλο που μπορεί να κατασκευαστεί από τον ίδιο τον ερευνητή, μπορεί και να δημιουργηθεί με αυτόματο τρόπο, επεκτείνοντας μία υπάρχουσα λίστα λέξεων με τη βοήθεια ορισμένων μέτρων που αποτυπώνουν την ισχύ σύνδεσης μεταξύ των λέξεων (όπως είναι η κατά σημεία αμοιβαία πληροφορία και η Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis)) (Turney & Littman, Measuring praise and criticism: Inference of semantic orientation from association, 2003). Εμπλέκει αρκετές φορές τον αλγόριθμο k κοντινότερων γειτόνων (k Nearest Neighbors / k-NN), υπό συνθήκη τυχαία πεδία (Conditional Random Fields / CRFs) και κρυφά μαρκοβιανά μοντέλα (Hidden Markov Models / HMMs).

3. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΑΣΙΣΜΕΝΗ ΣΕ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (ASPECT-BASED SENTIMENT ANALYSIS)

Η Ανάλυση Συναισθήματος βασισμένη σε Χαρακτηριστικά (Aspect-based Sentiment Analysis) είναι η εργασία που αποσκοπεί στην ανάλυση και τον καθορισμό του συναισθηματικού προσανατολισμού μίας σειράς οντοτήτων ή/και χαρακτηριστικών τους (aspects), εντοπισμένων σε μία πρόταση (Gu, Zhao, He, Li, & Ying, 2023). Είναι αρκετά συχνό το φαινόμενο, σε διαδικτυακές κριτικές προϊόντων ή υπηρεσιών αλλά και σε διαδικτυακά φόρουμ συζητήσεων, να γίνεται αναφορά σε πολλές οντότητες και να γίνεται αξιολόγηση ή να εκφράζονται απόψεις, ίδιας ή διαφορετικής συναισθηματικής πολικότητας, για πολλά χαρακτηριστικά τους. Για να επιτευχθεί αυτό, είναι απαραίτητες τόσο η κατανόηση της διάρθρωσης και του εννοιολογικού πλαισίου του κειμένου όσο και η εξαγωγή των εξαρτήσεων μεταξύ των λέξεων που απαρτίζουν την πρόταση και της συνακόλουθης σημασιολογίας των χαρακτηριστικών.

Δεδομένου ότι σε μία πρόταση, μπορεί να εκφράζονται γνώμες διαφορετικής συναισθηματικής πόλωσης για δεδομένα χαρακτηριστικά, η ανάλυση συναισθήματος επιχειρεί, αρχικά, να ανιχνεύσει και να εξάγει εκείνο το κείμενο που αφορά αποκλειστικά τα συγκεκριμένα χαρακτηριστικά, απαλλαγμένο από λέξεις, σχετικές με άλλες οντότητες που αποτελούν θόρυβο. Η εργασία αυτή είναι γνωστή στην αγγλική βιβλιογραφία ως aspect term extraction και αποτελεί το πρώτο υπό-πρόβλημα που καλείται να λυθεί στο πλαίσιο της Ανάλυσης Συναισθήματος βάσει Χαρακτηριστικών. Μέσω αυτής, αποκτάται χρήσιμη και σημαντική πληροφορία για το περιεχόμενο της συλλογής κειμένου, αναφορικά με τα ζητούμενα χαρακτηριστικά. Στην εικόνα που ακολουθεί, φαίνεται το κείμενο μίας κριτικής εστιατορίου, όπου γίνεται αναφορά σε δύο χαρακτηριστικά, τη θέα και το φαγητό, στα οποία αποδίδεται διαφορετικός συναισθηματικός προσανατολισμός, θετικός για τον πρώτο και αρνητικός για τον δεύτερο.

Εικόνα 1. Παράδειγμα μίας κριτικής εστιατορίου.

Οι προσεγγίσεις που μπορούν να ακολουθηθούν για τη συγκεκριμένη εργασία, διαφέρουν ανάλογα με τη φύση των δεδομένων. Στη βιβλιογραφία έχουν διερευνηθεί μέθοδοι επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης, μέθοδοι που στηρίζονται στη συχνότητα εμφάνισης των λέξεων (frequency-based) και τη

συντακτική τους θέση μέσα στις προτάσεις του κειμένου (syntax-based), όπως επίσης και υβριδικές προσεγγίσεις που συνδυάζουν περισσότερες της μίας μεθόδους.

Οι μέθοδοι που λαμβάνουν υπόψιν τη συχνότητα για την ανίχνευση των χαρακτηριστικών ενός γενικού θέματος, στηρίζονται στην παρατήρηση των ερευνητών ότι οι συχνότερες λέξεις σε μία συλλογή κειμένου (συνήθως ουσιαστικά), αποτελούν , τις περισσότερες φορές, τα ζητούμενα χαρακτηριστικά. Προκειμένου να αποφευχθεί ο κίνδυνος απόρριψης λέξεων εξαιτίας της χαμηλής τους συχνότητας, συνηθίζεται οι προσεγγίσεις αυτού του είδους να συνδυάζονται με άλλες μεθοδολογίες.

Οι Hu & Liu (2004a,2004b), ασχολούμενοι με κριτικές πωλητών για διαδικτυακές αγορές προϊόντων, αναγνώριζοντας, αρχικά, ως υποψήφια χαρακτηριστικά, εκείνες τις φράσεις που είχαν μεγάλη συχνότητα εμφάνισης στους κανόνες συσχέτισης των προτάσεων, κατέληξαν στα τελικά χαρακτηριστικά, αφαιρώντας τα μη συμπαγή και πλεονάζοντα. Στη συνέχεια, συμπεριέλαβαν σε αυτά και άλλα ουσιαστικά ή / και φράσεις ουσιαστικών, σπανίως εμφανιζόμενα, εκμεταλλευόμενοι την εγγύτητά τους σε επίθετα που εξέφραζαν κάποια άποψη ή συναίσθημα. Παρόμοια προσέγγιση ακολούθησαν και οι Popescu & Etzioni (2005), θέτοντας, σε πρώτη φάση, ως χαρακτηριστικά των κριτικών, τις πιο συχνές ονοματικές φράσεις και έπειτα, βάσει των λέξεων που εξέφραζαν γνώμες (opinion words), εντόπιζαν τα μη συχνά χαρακτηριστικά. Σε αυτό το σημείο, καθίσταται σημαντικό να επισημανθεί ότι σε γενικές γραμμές τα χαρακτηριστικά μπορεί να μην αναφέρονται αποκλειστικά ξεκάθαρα μέσα σε κομμάτια κειμένου, όπως συνέβη στις προαναφερθείσες μελέτες, αλλά μπορεί και να εννοούνται έμμεσα. Η ανίχνευση των χαρακτηριστικών, άμεσων και έμμεσων, όχι απαραίτητα ουσιαστικών, επιχειρήθηκε σε κριτικές οι οποίες ανέφεραν χωριστά τα πλεονεκτήματα και τα μειονεκτήματα των προϊόντων, εφαρμόζοντας μία επιβλεπομένη μορφή εξόρυξης των κανόνων συσχέτισης (Liu, Hu, & Cheng, Opinion observer: analyzing and comparing opinions on the Web, 2005). Οι Long, Zhang, & Zhu (2010), απέσπασαν από τις κριτικές ξενοδοχείων, με χρήση στατιστικών μεθόδων και αξιοποιώντας τη συχνότητα εμφάνισης των λέξεων, τα βασικά χαρακτηριστικά (core feature words), και έπειτα, συμπλήρωσαν με επιπρόσθετες, λιγότερο συχνά χρησιμοποιούμενες λέξεις βάσει τόσο της πληροφοριακής τους απόστασης από τα βασικά χαρακτηριστικά όσο και των γραμματικών εξαρτήσεών τους.

Οι μεθοδολογίες που βασίζονται στη συντακτική ανάλυση των προτάσεων, κατά βάση, αποσκοπούν στη μοντελοποίηση των υπαρκτών συντακτικών δομών και εξαρτήσεων (Xu, Pang, Wu, Cai, & Peng, 2023).

Το 2010, οι Zhao et al. τόνισαν ότι η απόσπαση μόνο των ονοματικών φράσεων, κατά την εργασία εξαγωγής των χαρακτηριστικών, προκαλεί τη μετρική της ανάκλησης να παραμένει σε χαμηλά επίπεδα. Πρότειναν ως λύσεις, αφενός, τη γενίκευση των συντακτικών δομών των προτάσεων των κριτικών με δύο ευρετικές συναρτήσεις και αφετέρου την εξόρυξή τους μέσω της διάσπασής σε υποδομές (substructures) αλλά και την χρήση μίας συνάρτησης πυρήνα για τον υπολογισμό της ομοιότητας μεταξύ τους.

Η ανακάλυψη υποψήφιων χαρακτηριστικών στο σύνολο ελέγχου έγινε, βρίσκοντας παρόμοια συντακτικά μοτίβα με αυτά των σημειωμένων χαρακτηριστικών στο σύνολο εκπαίδευσης. Στη μελέτη των Zhang et al. (2010), η λήψη των χαρακτηριστικών προτάθηκε να γίνει μέσα από μία διαδικασία δύο σταδίων που περιελάμβανε την απόσπαση υποψήφιων χαρακτηριστικών και την κατάταξη βάσει της σημαντικότητάς τους. Κατά το πρώτο στάδιο, εφαρμόζοντας τη μεθοδολογία της διπλής μετάδοσης (double propagation), ορισμένες βελτιωμένες σχέσεις μέρους-όλου (part-whole) και το μοτίβο «no», μετά από κάθε αναγνώριση χαρακτηριστικού, προσθέτονται καινούριες λέξεις στο λεξικό των λέξεων που εκφράζουν κάποιο συναίσθημα ή γνώμη. Αλλά συμβαίνει και το αντίστροφο, αφού με κάθε καινούρια λέξη που εκφράζει συναίσθημα, σημειώνονται καινούρια χαρακτηριστικά, εκμεταλλευόμενοι τις γραμμικές εξαρτήσεις μεταξύ τους. Όπως αναφέρεται στη συγκεκριμένη έρευνα, το μεγαλύτερο πλεονέκτημά της είναι ότι πρόκειται για μία μη επιβλεπόμενη μέθοδο που το μόνο που απαιτεί είναι ένα λεξικό με κάποιες βασικές λέξεις που εκφράζουν συναίσθημα. Οι Gu, Zhao et al. (2023) κατασκεύασαν ένα συνελικτικό νευρωνικό δίκτυο γράφων (graph convolutional neural network) βασισμένο στα συντακτικά δέντρα εξάρτησης το οποίο λάμβανε υπόψιν τόσο την πληροφορία που προερχόταν από την ανάλυση σε μέρη του λόγου (part-of-speech) όσο και τα σκορ συναισθήματος των λέξεων από ένα λεξικό συναισθημάτων (sentiment lexicon).

Οι μέθοδοι επιβλεπόμενης μάθησης στο πεδίο της απόσπασης των χαρακτηριστικών διάφορων οντοτήτων σε μία πρόταση έχουν αποτελέσει αντικείμενο έρευνας πολλών επιστημόνων. Το 2006 (Zhuang, Jing, & Zhu, 2006), παρουσιάστηκε αλγόριθμος επιβλεπόμενης μάθησης ο οποίος εξήγαγε ζεύγη χαρακτηριστικών – απόψεων, τόσο ξεκάθαρα αναφερόμενα όσο και υποδηλούμενα μέσα σε κάθε πρόταση. Για την υλοποίηση της πρώτης εργασίας, κατά το στάδιο της εκπαίδευσης, δημιουργούνταν τα γραφήματα συντακτικών εξαρτήσεων μεταξύ των λέξεων κάθε πρότασης και αναλύονταν οι λέξεις σε ποια μέρη του λόγου ανήκουν. Ύστερα, ανιχνευόταν και καταγραφόταν το συντομότερο μονοπάτι μεταξύ του χαρακτηριστικού και της εκφραζόμενης γνώμης ώστε η παραγόμενη ακολουθία μερών του λόγου και συντακτικής σχέσης να χρησιμοποιηθεί αυτούσια για τον εντοπισμό ζευγών χαρακτηριστικών-απόψεων στο σύνολο ελέγχου. Οι Jacob και Gurevych (Jakob & Gurevych, 2010) μοντελοποίησαν την εργασία εξόρυξης γνώμης με τη βοήθεια των Υπό Συνθήκη Τυχαίων Πεδίων (Conditional Random Fields / CRF) που ανήκουν στην κατηγορία της επιβλεπόμενης μάθησης. Η είσοδος σε αυτό το μοντέλο αποτελούνταν, αρχικά, από το αλφαριθμητικό της μικρότερης μονάδας κατάτμησης και την ετικέτα με το μέρος του λόγου στο οποίο αυτή ανήκε. Επίσης, είχε ως χαρακτηριστικά, τη συντακτική εξάρτηση όσων μονάδων κατάτμησης συσχετίζονταν άμεσα και απευθείας με την έκφραση που αντιπροσώπευε μία γνώμη μέσα σε μία πρόταση καθώς επίσης και την επισήμανση της παρουσίας ή μη των μονάδων κατάτμησης σε πρόταση που εξέφραζε κάποιου είδους γνώμη. Η ταξινόμηση κάθε μονάδας-όρου γινόταν με βάση το εάν αυτή ήταν ή δεν ήταν χαρακτηριστικό της οντότητας για την οποία δηλωνόταν

μία άποψη και σε περίπτωση που ήταν, εάν ήταν μόνη της ή συνοδευόταν και από την επόμενη σε σειρά μονάδα.

Το γεγονός ότι οι συλλογές κειμένου με επισημειωμένα τα χαρακτηριστικά συγκεκριμένων ζητούμενων οντοτήτων είναι δύσκολο να αποκτηθούν από τους ερευνητές, οδήγησε στη στροφή προς τη μελέτη και εύρεση μεθόδων μη επιβλεπόμενης μάθησης για την εξόρυξη γνώμης. Η λανθάνουσα κατανομή του Dirichlet (Latent Dirichlet Allocation) είναι η κατεξοχήν μη επιβλεπόμενη προσέγγιση απόσπασης χαρακτηριστικών, μοντελοποιώντας κάθε διαθέσιμο έγγραφο ως μείγμα θεμάτων-χαρακτηριστικών και κάθε θέμα ως ξεχωριστή κατανομή λέξεων (Brody & Elhadad, 2010). Ωστόσο, το γεγονός ότι δεν εκμεταλλεύεται τη συνύπαρξη των λέξεων, θεωρώντας ότι η εμφάνιση κάθε λέξης είναι ανεξάρτητη από τις υπόλοιπες σε συνδυασμό με την εξάρτηση από το μέγεθος κάθε εγγράφου για την εκτίμηση των θεμάτων οδήγησε σε στροφή προς τα νευρωνικά μοντέλα προσοχής (He, Lee, Ng, & Dahlmeier, 2017). Αυτά, αποδίδοντας τις λέξεις ως νευρωνικές αναπαραστάσεις, κωδικοποιούν την ‘εγγύτητά’ τους και εισάγοντας έναν μηχανισμό προσοχής, πετυχαίνουν την αφαίρεση άσχετων και μη συναφών εντοπισμένων χαρακτηριστικών. Οι Yuan et al. (2020) διερεύνησαν για την εξαγωγή των χαρακτηριστικών δύο μοντέλα που ανήκουν στην λανθάνουσα σημασιολογική ανάλυση: ένα μοντέλο κατανομής του Dirichlet που στηρίζεται στα λεξικά WordNet και SentiWordNet (SLDA) και μία βελτιωμένη παραλλαγή του LDA που συνδυάζει το SLDA με το MaxEnt-LDA (HME-LDA). Το ζήτημα της μεταφοράς χρήσιμων πληροφοριών και της αποτελεσματικής γενίκευσης σε άλλον εννοιολογικό τομέα (target domain) από αυτόν στον οποίο πραγματοποιείται η εκπαίδευση για την εξαγωγή χαρακτηριστικών (source domain) αποτέλεσε αντικείμενο έρευνας των Chen & Wan (Chen & Wan, 2022). Πρότειναν μία παραλλαγή της τεχνικής μεγιστοποίησης της αμοιβαίας πληροφορίας, επονομαζόμενη ως ‘FMIM’ (Fine-grained Mutual Information Maximization), η οποία βελτιστοποιεί το μέτρο της αμοιβαίας πληροφορίας και στους δύο τομείς.

Η εξαγωγή των χαρακτηριστικών μπορεί να επιτευχθεί και με υβριδικό τρόπο, συνδυάζοντας δύο ή περισσότερες από τις παραπάνω μεθόδους. Χαρακτηριστικό παράδειγμα αποτελεί η εργασία των Blair-Goldensohn, Hannan et al. (2008), στην οποία απέσπασαν τα χαρακτηριστικά από κριτικές εστιατορίων και ξενοδοχείων με τη βοήθεια ενός δυναμικού τρόπου και ενός στατικού μηχανισμού. Ο πρώτος αναγνώριζε ως χαρακτηριστικά τα πιο συχνά ουσιαστικά και σύνθετα ουσιαστικά, ενώ ο δεύτερος περιλάμβανε την ιδιόχειρη επισήμανση των προτάσεων που αφορούσαν και τους δύο τομείς, με συγκεκριμένα χαρακτηριστικά που είχαν προσδιοριστεί από τους ίδιους τους ερευνητές. Ύστερα, εκπαίδευαν έναν ταξινομητή μέγιστης εντροπίας ώστε να μαθαίνει το εάν σε κάποια πρόταση γινόταν λόγος για δεδομένο χαρακτηριστικό.

Μετά την ολοκλήρωση του εντοπισμού και της αναγνώρισης των χαρακτηριστικών στη δεδομένη συλλογή κειμένου, ακολουθεί η ανάθεση συναισθηματικής πολικότητας σε καθέναν από αυτούς τους όρους, δηλαδή η ταξινόμησή τους σε μία από τις προκαθορισμένες κατηγορίες. Αυτές είναι συνήθως η θετική και αρνητική πολικότητα.

Μπορεί, όμως να προτιμηθεί από τον εκάστοτε ερευνητή η κατηγοριοποίηση να γίνει πιο λεπτομερής, δηλαδή μεταξύ θετικής, αρνητικής, ουδέτερης και αντιφατικής συναισθηματικής πολικότητας (Pontiki, et al., 2014). Η ανάθεση κάποιου είδους συναισθήματος επιτυγχάνεται μέσω εκφράσεων που αναφέρονται και χαρακτηρίζουν τα χαρακτηριστικά (Pontiki, Galanis, Papageorgiou, Manandhar, & Androutsopoulos, 2015).

Στο τελευταίο στάδιο πραγματοποιείται η ταξινόμηση κάθε μελετώμενης πρότασης ή εγγράφου σε μία από τις προκαθορισμένες κατηγορίες, λαμβάνοντας υπόψιν όλες τις τιμές συναισθηματικού προσανατολισμού που έχουν υπολογιστεί προηγουμένως. Αυτή η διαδικασία κάθε άλλο παρά αυτόματη και προφανής είναι, καθώς η τελική κατηγοριοποίηση της πρότασης δεν σημαίνει ότι εκφράζονται σε αυτήν αποκλειστικά γνώμες ίδιας συναισθηματικής πολικότητας. Ας μην ξεχνούμε ότι μπορεί μία συλλογή κειμένου να αξιολογείται, για παράδειγμα, με βαθμολογία ίση με 4 (με μέγιστη τιμή το 5) από τον χρήστη, αλλά να εκφράζονται σε αυτήν απόψεις που εμπερικλείουν και αρνητική χροιά.

Τα βασικά βήματα της μεθοδολογίας που ακολουθείται με σκοπό την ανάλυση συναισθήματος είναι εν συντομία τα παρακάτω:

1. Αρχικά, δίνεται η είσοδος στο σύστημα ανάλυσης συναισθήματος. Αυτή είναι μία συλλογή εγγράφων που διατίθεται σε ηλεκτρονική μορφή, όπως HTML, XML, Excel και Word.
2. Βασικό χαρακτηριστικό αυτών των εγγράφων, δεδομένου ότι πρόκειται για δημιουργήματα ανθρώπινης σκέψης και γραφής, είναι η ανακολουθία στην οργάνωση και η διαφορετικότητα στην έκφραση. Συνεπώς, η προετοιμασία και προεπεξεργασία του διαθέσιμου κειμένου, με τη βοήθεια ποικίλων γλωσσικών εργαλείων όπως η κατάτμηση σε λέξεις και η λημματοποίηση, κρίνονται απαραίτητες προκειμένου να αποσπαστεί χρήσιμο περιεχόμενο για την επιδιωκόμενη εργασία. Παράλληλα, απομακρύνεται οτιδήποτε θεωρείται άχρηστο για την επακόλουθη ανάλυση, όντας είτε σε μορφή αριθμητικών ψηφίων, είτε σε μορφή χαρακτήρων, μεμονωμένων ή ομαδοποιημένων.
3. Ακολουθεί η ανάλυση του καθαρισμένου κειμένου, δηλαδή εξαγωγή των χαρακτηριστικών, συνήθως, βάσει παρουσίας ή συχνότητας. Από αυτά, μπορεί να αποφασιστεί από τον ερευνητή να επιλεγούν μόνο όσα ανήκουν σε συγκεκριμένα μέρη του λόγου (π.χ. ουσιαστικά).
4. Το επόμενο βήμα είναι η ανίχνευση του συναισθήματος που υποκρύπτεται στις υποκειμενικές προτάσεις των εγγράφων οι οποίες παραθέτουν γνώμες, απόψεις και πεποιθήσεις. Για το σκοπό αυτό, συχνά χρησιμοποιούνται λεξικά συναισθημάτων ή άλλες γλωσσικές πηγές που εντοπίζουν τα σημεία εκείνα που καθορίζουν την κατηγορία συναισθηματικής πολικότητας στην οποία ανήκει το εκάστοτε έγγραφο, η πρόταση ή το χαρακτηριστικό. Αυτό προϋποθέτει η κατασκευή του λεξικού συναισθημάτων (στην περίπτωση που αυτό κατασκευάζεται) να έχει γίνει με τέτοιο τρόπο ώστε κάθε όρος που

συμπεριλαμβάνεται σε αυτό, να συνοδεύεται από ένα μέτρο που αναπαριστά το βαθμό θετικότητας ή αρνητικότητας με έγκυρο τρόπο.

5. Ακολουθεί η ταξινόμηση συναισθήματος ανάμεσα σε κατηγορίες, των οποίων το πλήθος και το είδος έχει προκαθορίσει ο χρήστης.
6. Η έξοδος του συστήματος είναι οι επισημάνσεις με το είδος συναισθήματος που ενυπάρχει στα έγγραφα, τις προτάσεις ή τα χαρακτηριστικά οντοτήτων (ανάλογα με το επίπεδο ανάλυσης συναισθήματος). Τα αποτελέσματα μπορούν να οπτικοποιηθούν με ποικίλους τρόπους, διευκολύνοντας την κατανόησή τους.

4. ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΚΕΙΜΕΝΟΥ

Πρώτο βήμα για την επεξεργασία μίας μεγάλης συλλογής κειμένου είναι η κωδικοποίηση και μετατροπή των δεδομένων κειμένου σε τέτοια μορφή ώστε να μπορούν να δοθούν ως είσοδο σε κάποιον αλγόριθμο ταξινόμησης. Το αποτέλεσμα της διαδικασίας αυτής είναι η δημιουργία διανυσμάτων με συνιστώσες πραγματικούς αριθμούς. Υπάρχει εκτεταμένη έρευνα γύρω από αυτό το αντικείμενο, με αποτέλεσμα να υπάρχουν πολλοί διαθέσιμοι τρόποι, καθένας από τους οποίους επιλέγεται ανάλογα με την προσέγγιση και τον αλγόριθμο που πρόκειται να εφαρμοστεί σε επόμενο στάδιο. Οι κυριότεροι αναφέρονται παρακάτω.

4.1. BAG OF WORDS (BoW)

Η τεχνική Bag of Words παράγει την αναπαράσταση ενός κειμένου ως ένα διάνυσμα τόσου μήκους όσο είναι το πλήθος των μοναδικών λέξεων που το αποτελούν. Κάθε συνιστώσα του περιγράφει την εμφάνιση μοναδικής λέξης από ένα προκαθορισμένο λεξιλόγιο μέσα στο έγγραφο.

Το μέτρο παρουσίας των λέξεων αποφασίζεται από τον ερευνητή. Αυτό μπορεί να είναι η απόλυτη συχνότητα των εμφανίσεων μίας λέξης σε ένα έγγραφο ή το εάν μία λέξη εμφανίζεται ή όχι σε ένα έγγραφο. Στην πρώτη περίπτωση, κάθε συνιστώσα λαμβάνει μία ακέραια τιμή μεγαλύτερη ή ίση του μηδενός, ενώ στη δεύτερη, παίρνει είτε την τιμή 1, είτε την τιμή 0, όταν παρουσιάζεται ή αντίστοιχα απουσιάζει από το εκάστοτε έγγραφο.

Ο συγκεκριμένος αλγόριθμος φέρει αυτό το όνομα γιατί όπως δηλώνει και η αγγλική λέξη ‘bag’, δεν λαμβάνει καθόλου υπόψιν τη σειρά των λέξεων μέσα στην εκάστοτε πρόταση, παρά μόνο την ύπαρξή τους. Έτσι, ίδιες λέξεις, παρ’ όλο που είναι διαφορετικά διατεταγμένες μέσα σε μία πρόταση, μοντελοποιούνται σαν να είναι ίδιες. Αυτό το γεγονός είναι και το βασικό του μειονέκτημα, που σε συνδυασμό με την αδυναμία του να κατανοήσει το εννοιολογικό περιεχόμενο, τη γραμματική και τη συντακτική δομή του κειμένου συντελούν στη χρήση άλλων πιο εξελιγμένων μεθόδων αναπαράστασης (Gogula, Rahouti, Gogula, Jalamuri, & Jagatheesaperumal, 2023; Xia, Zong, & Li, 2011).

4.2. N-GRAMS

Ως N-gram ορίζεται μία ακολουθία N σε πλήθος βασικών μονάδων κατάτμησης του κειμένου που είναι διαδοχικά τοποθετημένες σε αυτό. Η μεταβλητή N μπορεί να πάρει

οποιονδήποτε φυσικό αριθμό. Έτσι, για N=1, δημιουργούνται τα λεγόμενα μονογράμματα των μονάδων κατάτμησης (unigrams), για N=2, τα λεγόμενα διγράμματα (bigrams), για N=3, τα τριγράμματα (trigrams) κ.ο.κ.. Σε αυτού του είδους την αναπαράσταση, το λεξιλόγιο πλέον δημιουργείται από όλες τις μοναδικές N-άδες διπλανών λέξεων. Για παράδειγμα, από την αγγλική φράση 'the weather was rainy today', σχηματίζονται τα εξής διγράμματα: 'the weather', 'weather was', 'was rainy', 'rainy today'.

Γίνεται εύκολα κατανοητό ότι μέσω αυτής της αναπαράστασης, αποτυπώνεται η διάταξη των λέξεων και τα γραμματικά μοτίβα που υπάρχουν μέσα σε κάθε πρόταση. Η συμπερίληψη των N-grams, τόσο για χαμηλές τάξεις του N (δηλαδή N=1 ή N=2), όσο και για υψηλές τάξεις, δεν κλιμακώνει απαραίτητα την απόδοση σε μία εργασία επεξεργασίας φυσικής γλώσσας. Αυτό συμβαίνει ιδιαίτερα στην περίπτωση που το βάρος των χαρακτηριστικών χαμηλών τάξεων είναι υψηλότερο συγκριτικά με αυτό των χαρακτηριστικών υψηλής τάξης (Dave, Lawrence, & Pennock, 2003). Αν και η έννοια του βάρους των λέξεων αναφέρεται σε επόμενη σχετική υπό-ενότητα, μπορεί να σημειωθεί ότι πρόκειται για μία τιμή που φανερώνει το πόσο σημαντική είναι η εκάστοτε λέξη για την εύρεση του συναισθηματικού προσανατολισμού ολόκληρου του εγγράφου.

4.3. ΛΕΞΕΙΣ ΩΣ ΜΕΡΗ ΤΟΥ ΛΟΓΟΥ

Σχετική έρευνα στο πεδίο εύρεσης αντιπροσωπευτικών όρων για τις εκφράσεις συναισθήματος σε μία συλλογή κειμένου έχει δείξει ότι οι φράσεις που περιέχουν επίθετα έχουν μεγάλη επίδραση στην αποτελεσματικότητα της εργασίας ανίχνευσης της υποκειμενικότητας (Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, 2002; Wiebe J. , 2000). Χρησιμοποιώντας τα μαζί με τα ουσιαστικά (Riloff, Wiebe, & Wilson, Learning subjective nouns using extraction pattern bootstrapping, 2003), τα ρήματα (Wiebe, Wilson, & Bell, 2001) και τα επιρρήματα, μπορεί να δημιουργηθεί μία διαφορετική αναπαράσταση ενός κειμένου συγκριτικά με τις προηγούμενες μεθόδους. Η βασική διαφορά με τις προηγούμενες μεθόδους έγκειται στην αξιοποίηση πληροφορίας αναφορικά με τα μέρη του λόγου στα οποία ανήκουν οι επιμέρους λέξεις, η οποία δεν παρέχεται στον ερευνητή με άμεσο τρόπο μέσω της ακολουθιακής δομής του κειμένου, όπως πριν.

4.4. ΣΥΝΤΑΚΤΙΚΗ ΕΞΑΡΤΗΣΗ ΜΕΤΑΞΥ ΤΩΝ ΛΕΞΕΩΝ

Η μελέτη χρήσης των γλωσσικών χαρακτηριστικών που απορρέουν από συντακτικά δέντρα, που είναι είτε σε μορφή υπό-δέντρων, είτε σε μορφή φράσεων, έχει δώσει ανάμεικτα συμπεράσματα όσον αφορά τη χρησιμότητα τους. Από τη μία πλευρά, έχει

βρεθεί ότι μπορούν να βελτιώσουν την αποσκοπούμενη εργασία της πρόβλεψης των βαθμολογιών και της ταξινόμησης κριτικών πελατών (Gamon, 2004; Matsumoto, Takamura, & Okumura, 2005). Από την άλλη πλευρά, η απόσπαση μόνο των σχέσεων της μορφής ‘επίθετο-ουσιαστικό’ έχει αποδειχτεί αρκετά περιοριστική, αδυνατώντας να πετύχει αποτελεσματικότερη ταξινόμηση πολικότητας από την απλή μέθοδο Bag of Words. Κατά συνέπεια, οι ερευνητές οδηγήθηκαν στην επέκταση του συνόλου των σχέσεων εξάρτησης που λάμβαναν υπόψιν. Ξεκίνησαν να εμφανίζονται ως χαρακτηριστικά, εξαρτήσεις του τύπου ‘υποκείμενο-ρήμα’ και ‘ρήμα-αντικείμενο’ αλλά και μετασχηματισμοί των βασικών σχέσεων συντακτικής εξάρτησης, ώστε να αποκτώνται μοτίβα που βοηθούν τον εκάστοτε ταξινομητή να γενικεύει καλύτερα (Ng, Dasgupta, & Niaz Arifin, 2006; Joshi & Penstein-Rosé, 2009).

4.5. ΣΤΑΘΜΙΣΗ ΟΡΩΝ

Μέσω της στάθμισης, οι όροι που απαρτίζουν μία μεγάλη συλλογή κειμένου, οργανωμένου σε έγγραφα, αποκτούν μία τιμή βάρους που φανερώνει τη σημασία και τη βαρύτητά τους για κάθε έγγραφο.

4.5.1. Tf-idf

Η tf-idf είναι μία ιδιαίτερα δημοφιλής προσέγγιση στάθμισης των όρων καθώς λαμβάνει υπόψιν τόσο τη σημασία που έχει κάθε όρος σε κάθε έγγραφο μίας μεγάλης συλλογής κειμένου, όσο και την αξία του σε όλα τα έγγραφα που συνιστούν αυτή τη συλλογή. Ορίζεται ως το γινόμενο δύο ποσοτήτων, συνδυάζοντας, με αυτόν τον τρόπο, δύο διαφορετικές έννοιες:

$$tf\_idf_{t.d} = tf_{t,d} \cdot idf_t \,. \tag{1}$$

Η πρώτη είναι η συχνότητα εμφάνισης κάθε όρου σε κάθε έγγραφο, η οποία για τον όρο t σε δεδομένο έγγραφο d, δίνεται από τον τύπο:

$$tf_{t,d} = count(t, d) \,. \tag{2}$$

Δεν είναι λίγες οι φορές που η συχνότητα του τύπου (2) αντικαθίσταται από αυτή του τύπου (3), με τη λογική ότι το πλήθος των φορών που εμφανίζεται κάποιος όρος σε δεδομένο έγγραφο δεν μπορεί να είναι ανάλογο με το πόσο σχετικός είναι αυτός με τη σημασία που υποδηλώνεται από το συγκεκριμένο έγγραφο. Αυτή η διαπίστωση μπορεί

να γίνει περισσότερο κατανοητή, αν αναλογιστούμε τα άρθρα και τις αντωνυμίες που παρ' όλο που έχουν πολύ υψηλή συχνότητα εμφάνισης σε οποιοδήποτε κείμενο, δεν προσφέρουν κάποια χρήσιμη πληροφορία για την κατανόηση της συναισθηματικής χροιάς του. Χρησιμοποιώντας τη λογαριθμική συνάρτηση, επιτυγχάνεται η συμπύκνωση των τιμών απόλυτης συχνότητας σε μικρότερο εύρος τιμών. Ο καινούριος τύπος είναι

$$tf_{t,d} = log(count(t,d) + 1)\,. \tag{3}$$

Επιπλέον, συναντάται, αρκετά συχνά, η πρώτη ποσότητα του γινομένου της μελετώμενης στάθμισης να αντιπροσωπεύεται από τη σχετική συχνότητα εμφάνισης ενός όρου σε δεδομένο έγγραφο. Οπότε, ο τύπος υπολογισμού της γίνεται:

$$tf_{t,d} = \frac{count(t,d)}{\sum_{t' \epsilon d} count(t',d)}\,. \tag{4}$$

Οριζόμενη ως το πλήθος των φορών που εμφανίζεται ένας συγκεκριμένος όρος σε συγκεκριμένο έγγραφο προς το συνολικό πλήθος των όρων που περιέχονται σε αυτό το έγγραφο, αποτυπώνει το πόσο σημαντικός είναι ο συγκεκριμένος όρος για αυτό το έγγραφο.

Η δεύτερη έννοια σε αυτού του είδους τη στάθμιση αφορά τη συχνότητα των εγγράφων στα οποία εμφανίζεται ο εκάστοτε όρος, γνωστή στην αγγλική βιβλιογραφία ως document frequency (ή εν συντομία 'df'). Είναι απαραίτητο να επισημανθεί ότι αυτή διαφέρει από την απόλυτη συχνότητα εμφάνισης κάθε όρου στη συλλογή εγγράφων, καθώς στη δεύτερη περίπτωση πρόκειται για το πλήθος των φορών που ο εκάστοτε όρος εμφανίζεται σε όλα τα έγγραφα. Υπό αυτό το πλαίσιο, ο δεύτερος παράγοντας του γινομένου που υπολογίζει την tf-idf στάθμιση, γνωστός ως αντίστροφη συχνότητα εγγράφων (στα αγγλικά ως inverse document frequency), δίνεται, για τον όρο t, από τον τύπο:

$$idf_t = \frac{N}{df_t}, \tag{5}$$

όπου N είναι το συνολικό πλήθος των εγγράφων που απαρτίζουν τη μελετώμενη συλλογή κειμένου και df_t είναι το πλήθος των εγγράφων στα οποία εμφανίζεται ο όρος t. Από τον τύπο (5), γίνεται αντιληπτό ότι η αντίστροφη συχνότητα εγγράφων για έναν όρο και κατά συνέπεια το βάρος του λαμβάνουν υψηλότερες τιμές όταν αυτός ο όρος εμφανίζεται σε ελάχιστα μόνο έγγραφα της συλλογής. Αυτό σημαίνει ότι οι όροι που περιέχονται λίγες φορές στα έγγραφα, είναι αυτοί που μπορούν πιο εύκολα να ξεχωρίσουν τον συναισθηματικό προσανατολισμό των εγγράφων.

Η συχνή μεταβολή των τιμών της αντίστροφης συχνότητας εγγράφων σε λογαριθμική κλίμακα εξυπηρετεί στον περιορισμό του μεγάλου εύρους τιμών της και τη συνακόλουθη εξομάλυνση της επίδρασής της στον υπολογισμό των τιμών στάθμισης. Με αυτόν τον τρόπο, ο πολλαπλασιασμός της με τη συχνότητα εμφάνισης του εκάστοτε όρου σε δεδομένο έγγραφο (που με τον τύπο (4) ανήκει στο διάστημα [0,1]) παράγει πιο λογικά αποτελέσματα.

4.5.2. Τροποποιημένες εκδοχές της Tf-idf

Οι Zhu, Wang & Zou (Zhu, Wang, & Zou, 2016) πρότειναν μία βελτιωμένη εναλλακτική της κλασικής στάθμισης tf-idf. Ο μαθηματικός τύπος που την υπολογίζει, χρησιμοποιεί τις ποσότητες $tf_{t,d}$ και idf_t της τεχνικής tf-idf, εισάγοντας έναν επιπλέον παράγοντα e που καλείται παράγοντας επίδρασης:

$$tf\_idf_{t.d}{}' = tf_{t,d} - e_t \cdot idf_t \,. \qquad (6)$$

Αυτός επινοήθηκε με σκοπό να ληφθεί υπόψιν η κατανομή των όρων στα έγγραφα διαφορετικών κλάσεων. Πιο συγκεκριμένα, η σπάνια εμφάνιση κάποιας λέξης στα έγγραφα κάποιων κλάσεων σε αντιδιαστολή με την συχνή εμφάνιση στα έγγραφα άλλων κλάσεων μπορεί να αποτελέσει καθοριστικό παράγοντα για τη διάκριση των εγγράφων που κατηγοριοποιούνται σε διαφορετικές κλάσεις. Οι Al-Ghuribi, Mohd Noah & Tiun (Al-Ghuribi, Mohd Noah, & Tiun, 2020) τροποποίησαν τη μεταβλητή e για τον τυχαίο όρο t, ώστε να διαφοροποιείται ανάλογα με το αν ο όρος αυτός είναι κύριο χαρακτηριστικό ή βασικός όρος που συμπληρώνει και περιγράφει κάποιο κύριο χαρακτηριστικό. Περισσότερες λεπτομέρειες γι' αυτούς τους όρους παρατίθενται σε ακόλουθο κεφάλαιο. Τελικά, ο τύπος είναι:

$$e_t = \begin{cases} \left(\frac{1}{Cl}\sum_{i=1}^{Cl}\left(df_{t,i} - \frac{1}{Cl}\right)^2\right)^{1/2} + 3, & \text{εάν } t\text{: κύριο χαρακτηριστικό} \\ \left(\frac{1}{Cl}\sum_{i=1}^{Cl}\left(df_{t,i} - \frac{1}{Cl}\right)^2\right)^{1/2} + 1, & \text{εάν } t\text{: βασικός όρος} \end{cases}, \qquad (7)$$

όπου Cl είναι το συνολικό πλήθος των δυνατών κλάσεων ταξινόμησης των εγγράφων και $df_{t,i}$ είναι το πλήθος των εγγράφων που περιέχουν τον όρο t και είναι γνωστό ότι ανήκουν στην κλάση i.

Μία άλλη προσέγγιση στάθμισης μετρά τον βαθμό σημαντικότητας ενός χαρακτηριστικού σε ένα έγγραφο εξετάζοντας και το πλήθος των προτάσεων που το συγκροτούν (Nguyen Thi Ngoc, Nguyen Thi Thu, & Nguyen, 2019). Έτσι, το μόνο που αλλάζει σε σύγκριση με τον τύπο (4) είναι ο υπολογισμός της σχετικής συχνότητας ενός όρου σε δεδομένο έγγραφο. Σε αυτήν την περίπτωση υπολογίζεται ως:

$$tf_{t,d} = \frac{count(t,d)}{count(s,d)}, \tag{8}$$

με τον παρονομαστή να δηλώνει το πλήθος των προτάσεων s από τις οποίες αποτελείται το έγγραφο d.

5. ΚΑΤΑΝΕΜΗΜΕΝΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΛΕΞΕΩΝ

Ένα είδος αναπαράστασης των λέξεων είναι η λεγόμενη κατανεμημένη αναπαράσταση (distributional representation), η οποία στηρίζεται στην ύπαρξη ενός πίνακα F (co-occurrence matrix), με μετρήσεις συνυπάρξεων των λέξεων μίας μεγάλης συλλογής κειμένου που το απαρτίζουν (Sahlgren, 2006). Αυτός ο πίνακας έχει διάσταση ίση με $|V| \times C$, όπου $|V|$ είναι το μέγεθος του λεξιλογίου του κειμένου (δηλαδή το πλήθος των μοναδικών λέξεων που το αποτελούν) και το C ισούται είτε με το $|V|$, είτε με το πλήθος των εγγράφων (documents) στα οποία οργανώνεται η συλλογή κειμένου.

Το στοιχείο της γραμμής i και της στήλης j, έστω $F_{i,j}$, στην πρώτη περίπτωση, αντιπροσωπεύει το πλήθος των φορών που η λέξη i έχει στο σύνολο των περιβαλλουσών λέξεών της τη λέξη j, ενώ στη δεύτερη περίπτωση, είναι ίσο με το βάρος της λέξης i στο έγγραφο j. Όσον αφορά το σύνολο των περιβαλλουσών λέξεων δεδομένης λέξης, πρόκειται για το 'παράθυρο' (context) λέξεων προκαθορισμένου μεγέθους τοποθετημένο στην αριστερή ή τη δεξιά πλευρά της λέξης. Με αυτόν τον τρόπο, κάθε λέξη, έστω η i, αναπαρίσταται από το διάνυσμα $\boldsymbol{F_{i,\cdot}}$ της γραμμής i του πίνακα F το οποία αποτελείται από C στοιχεία.

Ωστόσο, η υψηλή διάσταση της διανυσματικής αναπαράστασης κάθε λέξης αυξάνει εκθετικά τις απαιτήσεις σε μνήμη και πολυπλοκότητα, οδηγώντας στη δημιουργία μίας απεικόνισης g του πίνακα F σε έναν άλλο πίνακα f (Turlan, Ratlnov, & Benglo, 2010). Ο νέος πίνακας έχει διάσταση $|V| \times d$, με το πλήθος των στηλών του, d, να είναι πολύ μικρότερο του C. Η απεικόνιση g ορίζεται ως εξής:

$$g: \ |V| \times C \to |V| \times d$$
$$F \longmapsto g(F) = f \ . \qquad (9)$$

Η επιλογή της g κρίνει την τελική αναπαράσταση κάθε λέξης ως διάνυσμα συγκεκριμένης σταθερής διάστασης, ανεξάρτητης από το μήκος της λέξης (δηλαδή το πλήθος των χαρακτήρων της).

5.1. WORD2VEC EMBEDDINGS

Το Word2vec είναι ένα μοντέλο εκμάθησης κατανεμημένων αναπαραστάσεων λέξεων (Mikolov, Chen, Corrado, & Dean, 2013), ικανό να μετατρέψει λέξεις σε αριθμητικά διανύσματα, γνωστά στην αγγλική βιβλιογραφία ως word embeddings. Η κατασκευή των διανυσματικών αυτών αναπαραστάσεων στηρίζεται στη σημασιολογική σχέση μεταξύ των λέξεων του κειμένου, με αποτέλεσμα η παρόμοια σημασία μεταξύ

διαφορετικών λέξεων σε δεδομένο κείμενο να υποδηλώνεται από την εγγύτητα των αντίστοιχων διανυσμάτων στον διανυσματικό χώρο. Η εκπαίδευση ενός τέτοιου μοντέλου παρέχει τις ζητούμενες διανυσματικές αναπαραστάσεις των λέξεων, λειτουργώντας οι ίδιες, ως κάποιες από τις παραμέτρους του υπό εκμάθηση μοντέλου (Mikolov, Yih, & Zweig, Linguistic Regularities in Continuous Space Word Representations, 2013).

Το Word2vec εμφανίζεται με δύο διαφορετικές αρχιτεκτονικές μοντέλων που ανήκουν στη κατηγορία των νευρωνικών δικτύων: το Continuous Bag-of-Words (CBOW) και το Continuous Skip-gram. Καθεμία μπορεί να εκπαιδευτεί με διάφορες μεθόδους όπως η αρνητική δειγματοληψία και η χρήση της ιεραρχικής softmax (hierarchical softmax) αντί της softmax ως συνάρτηση ενεργοποίησης του επιπέδου εξόδου.

5.1.1. Continuous Bag-of-Words

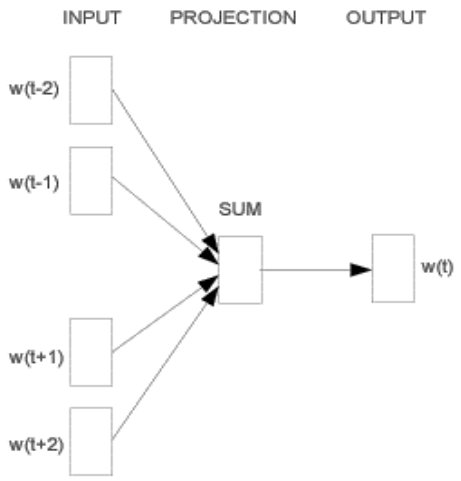
Το πρώτο μοντέλο (Continuous Bag-of-Words) έχει παρόμοια αρχιτεκτονική με αυτή ενός γλωσσικού μοντέλου βασισμένου σε εμπρός τροφοδότησης νευρωνικό δίκτυο (neural net language model / NNLM). Αποτελείται από 3 επίπεδα, το επίπεδο εισόδου (input layer), το επίπεδο προβολής (projection layer) και το επίπεδο εξόδου (output layer) (Mikolov, Chen, Corrado, & Dean, 2013).

Στο επίπεδο εισόδου εισάγονται οι λέξεις που εντοπίζονται στο παράθυρο περιβαλλουσών λέξεων προκαθορισμένου μεγέθους, τοποθετημένο τόσο από τα αριστερά όσο και από τα δεξιά εκείνης της λέξης που επιθυμούμε να προβλεφθεί. Καθεμία από τις λέξεις έχει προηγουμένως κωδικοποιηθεί σε one-hot διάνυσμα, διάστασης όσης και το μέγεθος του λεξιλογίου του κειμένου. Πιο συγκεκριμένα, αυτό αποτελείται από 0 και 1, με το μοναδικό στοιχείο 1 να εντοπίζεται στη θέση εκείνη που βρίσκεται και η αντίστοιχη λέξη μέσα στο λεξιλόγιο. Ακολουθεί η μετατροπή κάθε τέτοιου διανύσματος σε διάνυσμα χαμηλότερης διάστασης, ίσης με την επιθυμητή διάσταση των κατανεμημένων αναπαραστάσεων των λέξεων. Αυτή επιτυγχάνεται μέσω του πολλαπλασιασμού του εκάστοτε διανύσματος με έναν πίνακα παραμέτρων E, γνωστό ως Embedding matrix.

Στη συνέχεια, λαμβάνεται ο μέσος όρος ανά συνιστώσα όλων αυτών των διανυσμάτων που εισάγονται στο μοντέλο αποσκοπώντας στην πρόβλεψη της λέξης-στόχου και το παραγόμενο διάνυσμα φτάνει στο επίπεδο εξόδου το οποίο είναι ένα πλήρως συνδεδεμένο επίπεδο με τόσες ενώσεις όσο είναι το μέγεθος του λεξιλογίου. Κάθε συνιστώσα του διανύσματος εξόδου αντιπροσωπεύει τη δεσμευμένη πιθανότητα η αντίστοιχη λέξη να είναι εντός καθορισμένου παραθύρου λέξεων (Mai, Galke, & Scherp, 2019). Το μοντέλο ονομάστηκε σκόπιμα Bag-of-Words. Ο λόγος έγκειται στο ότι ενδεχόμενη αλλαγή σειράς εισαγωγής των λέξεων κάθε δείγματος στο επίπεδο

εισόδου δεν αλλάζει την προβλεπόμενη λέξη, καθώς οι μέσοι όροι που υπολογίζονται σε ενδιάμεσο επίπεδο δεν λαμβάνουν υπόψιν τη σειρά των τιμών.

Στην εικόνα 2, φαίνονται τα επίπεδα της αρχιτεκτονικής με την υπερπαράμετρο του μεγέθους του παραθύρου περιβαλλουσών λέξεων να ισούται με 2. Γι' αυτό το λόγο, για την πρόβλεψη της $\boldsymbol{w}(t)$, εισάγονται οι 2 προηγούμενες λέξεις $(\boldsymbol{w}(t-2), \boldsymbol{w}(t-1))$ και οι 2 επόμενες λέξεις $(\boldsymbol{w}(t+1), \boldsymbol{w}(t+2))$ κατά μήκος της εκάστοτε πρότασης. Έτσι, το συγκεκριμένο δείγμα εκπαίδευσης είναι το ζευγάρι $([\boldsymbol{w}(t-2), \boldsymbol{w}(t-1), \boldsymbol{w}(t+1), \boldsymbol{w}(t+2)], \boldsymbol{w}(t))$. Ο συμβολισμός $\boldsymbol{w}(t)$ αντιπροσωπεύει το one-hot διάνυσμα εκείνης της λέξης που βρίσκεται στη σειρά t μέσα στη συλλογή κειμένου. Έτσι, δεδομένου ότι η αρίθμηση των θέσεων ενός διανύσματος ξεκινάει από το 0 και τελειώνει στον αριθμό που ταυτίζεται με τη διάσταση του διανύσματος μειωμένο κατά 1, η συνιστώσα με τιμή 1 εντοπίζεται στη θέση με δείκτη το $t-1$.



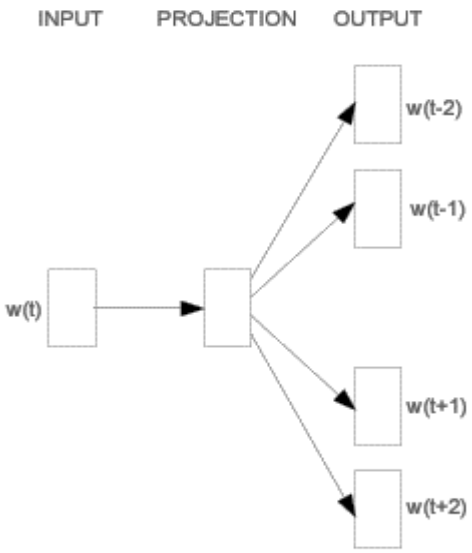
Εικόνα 2. Αρχιτεκτονική του Continuous Bag-of-Words μοντέλου.

Οι παράμετροι του νευρωνικού δικτύου που αναπροσαρμόζονται κατά την εκπαίδευσή του είναι τόσο τα στοιχεία του πίνακα Ε όσο και οι πίνακες βαρών και πόλωσης του επιπέδου εξόδου. Με την ολοκλήρωση της εκπαίδευσης του μοντέλου, ο πίνακας Ε που προκύπτει είναι αυτός που περιέχει σε κάθε γραμμή του τη ζητούμενη συνεχή κατανεμημένη αναπαράσταση καθεμίας λέξης (Sarkar, 2018).

5.1.2. Continuous Skip-gram

Ένα άλλο νευρωνικό δίκτυο που δημιουργεί κατανεμημένες αναπαραστάσεις λέξεων είναι το μοντέλο Continuous Skip-gram. Η αρχιτεκτονική του στην εικόνα 3 δείχνει ότι επιχειρεί να πετύχει το αντίστροφο από το Continuous Bag-of-Words μοντέλο. Δηλαδή

δεδομένης μίας λέξης-στόχου, επιδιώκει την πρόβλεψη καθεμίας λέξης που την περιβάλλει αριστερά και δεξιά εντός προκαθορισμένης εμβέλειας.



Εικόνα 3. Αρχιτεκτονική του Continuous Skip-gram μοντέλου.

Από τη μία πλευρά, στο επίπεδο εισόδου δίνεται το one-hot διάνυσμα της λέξης-στόχου, με διάσταση τόση όσο το πλήθος των διαφορετικών λέξεων του κειμένου. Από την άλλη πλευρά, η έξοδος του μοντέλου είναι επίσης ένα διάνυσμα ίδιας διάστασης με το μέγεθος του λεξιλογίου που έχει όμως ως συνιστώσες, τις πιθανότητες καθεμία λέξη του λεξιλογίου να είναι η τυχαία επιλεγμένη λέξη εντός της επιτρεπόμενης εμβέλειας. Η κατανομή πιθανοτήτων που δίνει το επίπεδο εξόδου επιτυγχάνεται μέσω της εφαρμογής της συνάρτησης ενεργοποίησης softmax.

Το διάνυσμα εξόδου του ενδιάμεσου κρυφού επιπέδου προβολής προκύπτει από τον πολλαπλασιασμό του one-hot διανύσματος εισόδου με έναν πίνακα βαρών E (Embedding matrix). Τα στοιχεία του είναι παράμετροι του νευρωνικού δικτύου, με τιμές που μαθαίνονται κατά την εκπαίδευση. Το πλήθος των ενώσεων-νευρώνων του κρυφού στρώματος αποτελεί μία υπερπαράμετρο του νευρωνικού δικτύου, η τιμή της οποίας αποφασίζεται από τον χρήστη και συμπίπτει με τη διάσταση των αναπαραστάσεων των λέξεων.

Συγκρίνοντας τα μοντέλα Continuous Bag-of-Words και Continuous Skip-gram, το πρώτο, σε αρκετές περιπτώσεις, είναι γρηγορότερο από το δεύτερο και πετυχαίνει μεγαλύτερη ακρίβεια για τις συχνά εμφανιζόμενες λέξεις (Google Code Archive-word2vec, χ.χ.). Αντίθετα, το νευρωνικό δίκτυο Continuous Skip-gram έχει μεγαλύτερη αποτελεσματικότητα σε λέξεις και φράσεις που εμφανίζονται σπάνια στο

υπό διερεύνηση κείμενο και τείνει να αποδίδει καλύτερα όταν οι συλλογές κειμένου που χρησιμοποιούνται για την εκπαίδευσή του είναι μικρού και μεσαίου μεγέθους.

5.1.3. Αρνητική δειγματοληψία

Δεδομένης μίας ακολουθίας λέξεων στο σύνολο εκπαίδευσης: $w_1, w_2, \dots, w_T$, ο σκοπός ενός Skip-gram μοντέλου, με μέγεθος παραθύρου περιβαλλουσών λέξεων ίσο με c, είναι η μεγιστοποίηση της μέσης λογαριθμικής πιθανότητας που δίνεται από τον ακόλουθο τύπο (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013):

$$\frac{1}{T} \cdot \sum_{t=1}^{T} \sum_{\substack{-c \le j \le c \\ j \neq 0}} log\left(p\left(w_{t+j} \middle| w_t\right)\right). \tag{10}$$

Εφόσον εφαρμοστεί η συνάρτηση softmax στο επίπεδο εξόδου του νευρωνικού δικτύου, η δεσμευμένη πιθανότητα του παραπάνω τύπου, συμβολίζοντας την πιθανότητα, δεδομένης της λέξης εισόδου w_t, η κοντινή λέξη $w_{t+j}, -c \le j \le c$, $j \neq 0$, να ανήκει σε γειτονική περιοχή της w_t, υπολογίζεται με τη βοήθεια του ακόλουθου τύπου:

$$p\left(w_{t+j} \middle| w_t\right) = y(t + j - 1). \tag{11}$$

Με $y(t + j - 1)$ δηλώνεται η συνιστώσα, με δείκτη θέσης $t + j - 2$, του διανύσματος εξόδου $\boldsymbol{y}$ του δείγματος εκπαίδευσης που έχει ως είσοδο το one-hot διάνυσμα της λέξης w_t. Το συγκεκριμένο διάνυσμα εξόδου δίνεται από τον παρακάτω τύπο:

$$\boldsymbol{y} = softmax(\boldsymbol{\Theta}^T \cdot \boldsymbol{e}_t) = \left[\frac{exp\left(\boldsymbol{\Theta}_{1,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}{\sum_{i=1}^{|V|} exp\left(\boldsymbol{\Theta}_{i,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}, \frac{exp\left(\boldsymbol{\Theta}_{2,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}{\sum_{i=1}^{|V|} exp\left(\boldsymbol{\Theta}_{i,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}, \dots, \frac{exp\left(\boldsymbol{\Theta}_{|V|,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}{\sum_{i=1}^{|V|} exp\left(\boldsymbol{\Theta}_{i,\cdot}{}^T \cdot \boldsymbol{e}_t\right)}\right], \tag{12}$$

όπου $\boldsymbol{\Theta}_{i,\cdot}{}^T$ είναι το διάνυσμα-γραμμή που αντιστοιχεί στην i-οστή γραμμή του πίνακα βαρών του επιπέδου εξόδου $\boldsymbol{\Theta}$ και $\boldsymbol{e}_t$ είναι το λεγόμενο word embedding της λέξης $\boldsymbol{w}_t$.

Γίνεται κατανοητό από τους τύπους (11) και (12) ότι ο υπολογισμός κάθε δεσμευμένης πιθανότητας απαιτεί προηγουμένως τον υπολογισμό ενός αθροίσματος με τόσους προσθετέους όσο και το μέγεθος του λεξιλογίου. Δεδομένου ότι το μέγεθος αυτό μπορεί να φτάσει να είναι χιλιάδες, ακόμα και εκατομμύρια λέξεις, το υπολογιστικό κόστος αυξάνεται κατά πολύ.

Ένας τρόπος επίλυσης αυτού του προβλήματος είναι η εφαρμογή αρνητικής δειγματοληψίας. Το σκεπτικό πίσω από αυτού του είδους τη δειγματοληψία είναι αφενός ο στόχος να μεγιστοποιηθεί η πιθανότητα οι όντως κοντινές λέξεις στις εκάστοτε λέξεις εισόδου να ανήκουν στο παράθυρο περιβαλλουσών λέξεων προκαθορισμένου μεγέθους. Αφετέρου η ιδανική κατάσταση περιλαμβάνει και την ελαχιστοποίηση της πιθανότητας οι λέξεις εκτός των ορίων του παραθύρου να είναι κάθε φορά αυτές που προβλέπονται από το μοντέλο.

Αυτές οι διαπιστώσεις οδήγησαν για κάθε δείγμα εκπαίδευσης που περιέχει ζεύγος λέξεων εντός καθορισμένου παραθύρου (οπότε και έχει ετικέτα 1), να επιλέγονται, επίσης, k στο πλήθος αρνητικά δείγματα με ετικέτα 0 που είναι ζεύγη λέξεων οι οποίες δεν εμφανίζονται όσο κοντά απαιτεί το μέγεθος του παραθύρου. Η ετικέτα 1 ή 0 σχετίζεται άμεσα με το εάν οι λέξεις του εκάστοτε δείγματος εκπαίδευσης ανήκουν στην ίδια γειτονιά λέξεων κατά μήκος του κειμένου ή όχι αντίστοιχα. Η επιλογή των επιθυμητών λέξεων εξόδου στα αρνητικά δείγματα γίνεται βάσει της λεγόμενης κατανομής θορύβου (noise distribution) P_n. Σύμφωνα με αυτή, η πιθανότητα μία λέξη του λεξιλογίου να είναι λέξη θορύβου (δηλαδή να μην ανήκει στο παράθυρο περιβαλλουσών λέξεων της κεντρικής λέξης που εισάγεται στο μοντέλο) ορίζεται ως ακολούθως (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013):

$$P(\text{επιλογής λέξης } w_i \text{ ως λέξη θορύβου}) = \frac{U(w_i)^{3/4}}{Z}\ ,\ w_i \in V, \quad (13)$$

όπου Z είναι μία σταθερά κανονικοποίησης και $U(w_i)$ είναι η κατανομή unigram για τη λέξη w_i.

Η ταξινόμηση για κάθε δείγμα εκπαίδευσης μετατρέπεται πλέον σε δυαδική μεταξύ των τιμών 0 ή 1, με τη λογιστική συνάρτηση να αντικαθιστά τη softmax στη θέση της επιλεχθείσας συνάρτησης ενεργοποίησης του επιπέδου εξόδου. Τότε, σε περίπτωση που η λέξη εξόδου ενός δείγματος εκπαίδευσης ανήκει στην προκαθορισμένη 'γειτονιά' μεγέθους c από την λέξη εισόδου, η πιθανότητα $P(w_{t+j}|w_t)$ ισούται με $P(D = 1|w_{t+j}, w_t)$ για $j \in \{-c, \dots, c\}\backslash\{0\}$, ενώ σε αντίθετη περίπτωση ισούται με $P(D = 0|w_{t+j}, w_t)$ για $j \notin \{-c, \dots, c\}$. Επιπρόσθετα, ισχύει:

$$P(D = 1|w_{t+j}, w_t) = \sigma(\boldsymbol{\Theta}_{t+j,\cdot}{}^T \cdot \boldsymbol{e}_t), \text{για } j \in \{-c, \dots, c\}\backslash\{0\} \quad (14)$$

και

$$P(D = 0|w_{t+j}, w_t) = 1 - P(D = 1|w_{t+j}, w_t) = 1 - \sigma(\boldsymbol{\Theta}_{t+j,\cdot}{}^T \cdot \boldsymbol{e}_t), \text{για } j \notin \{-c, \dots, c\} \quad (15)$$

Δεδομένου ότι ο τύπος λογιστικής συνάρτησης είναι:

$$f(x) = \frac{1}{1 + e^{-x}}\,, \tag{16}$$

οι παραπάνω τύποι (14) και (15) γίνονται αντίστοιχα:

$$P\big(D = 1|w_{t+j}, w_t\big) = \frac{1}{1 + exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big)} \tag{17}$$

και

$$P\big(D = 0|w_{t+j}, w_t\big) = 1 - \frac{1}{1 + exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big)} \Leftrightarrow P\big(D = 0|w_{t+j}, w_t\big) = \frac{1 + exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big) - 1}{1 + exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big)} \Leftrightarrow$$

$$\Leftrightarrow P\big(D = 0|w_{t+j}, w_t\big) = \frac{exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big)}{1 + exp\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big)} \xLeftrightarrow{Απλοποίηση\ με\ exp\left(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\right)}$$

$$\Leftrightarrow P\big(D = 0|w_{t+j}, w_t\big) = \frac{1}{exp\big(\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big) + 1} \overset{(16)}{\Longleftrightarrow}$$

$$\overset{(16)}{\Longleftrightarrow} P\big(D = 0|w_{t+j}, w_t\big) = \sigma\big(-\boldsymbol{\Theta}_{t+j,\cdot}{}^{T} \cdot \boldsymbol{e}_t\big), \text{για } j \notin \{-c, \dots, c\}\,. \tag{18}$$

Ο παραπάνω τύπος δείχνει ότι κατά την οπισθοδιάδοση του σφάλματος σε κάθε επανάληψη δεν ανανεώνονται όλα τα στοιχεία του πίνακα βαρών του επιπέδου εξόδου, παρά μόνο αυτά που αντιστοιχούν στις k λέξεις θορύβου και την λέξη εξόδου του θετικού δείγματος (McCormick, 2017).

Ένα Continuous Skip-Gram μοντέλο, με εφαρμογή αρνητικής δειγματοληψίας για την εκπαίδευση ενός θετικού δείγματος, έστω (w_ℓ, w_o), και k αρνητικών δειγμάτων της μορφής $(w_\ell, w_i), i = 1,2, \dots, k$, έχει στόχο την εύρεση των παραμέτρων που μεγιστοποιούν την ακόλουθη αντικειμενική συνάρτηση:

$$log\big(p(w_o|w_\ell)\big) + \sum_{i=1}^{k} E_{w_i \sim P_n}\big[log\big(p(w_i|w_\ell)\big)\big]. \tag{19}$$

Με τη βοήθεια των τύπων (14) και (18) η συνάρτηση αυτή γράφεται εναλλακτικά (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013):

$$log\left(\sigma\left(\boldsymbol{\Theta}_{o,\cdot}{}^{T}\cdot\boldsymbol{e}_{\ell}\right)\right)+\sum_{i=1}^{k}E_{w_i\sim P_n}\left[log\left(\sigma\left(-\boldsymbol{\Theta}_{i,\cdot}{}^{T}\cdot\boldsymbol{e}_{\ell}\right)\right)\right]. \qquad (20)$$

Όσον αφορά το πλήθος k των αρνητικών δειγμάτων που επιλέγονται για κάθε θετικό δείγμα, αυτό εξαρτάται από το μέγεθος του συνόλου εκπαίδευσης. Πιο συγκεκριμένα, προτείνεται από τη βιβλιογραφία να καθορίζεται με μία τιμή μεταξύ 5 και 20 όταν το σύνολο εκπαίδευσης είναι μικρό, ελαττώνοντάς την μεταξύ 2 και 5 όταν είναι μεγάλο.

6. ΛΕΞΙΚΟ ΣΥΝΑΙΣΘΗΜΑΤΩΝ

Το λεξικό συναισθημάτων που χρησιμοποιείται στις εργασίες ανάλυσης συναισθήματος είναι ένα θεμελιώδες και πολύ χρήσιμο εργαλείο το οποίο μπορεί να παραχθεί με διάφορους τρόπους.

Ένας από αυτούς είναι η δημιουργία του από τον ίδιο τον χρήστη ανάλογα με τον τομέα στον οποίο ανήκουν τα υπό μελέτη έγγραφα. Ωστόσο, όπως γίνεται αντιληπτό, πρόκειται για μία ιδιαίτερα επίπονη και χρονοβόρα διαδικασία, γι' αυτό και πολύ συχνά οι ερευνητές καταφεύγουν σε κάποια άλλη διαδικασία απόκτησης του λεξικού.

Μία άλλη προσέγγιση περιλαμβάνει τη χρήση ενός συνόλου πρότυπων όρων που αφορούν το εννοιολογικό πεδίο της υπό μελέτη συλλογής κειμένου ώστε αυτό, ύστερα, να επεκταθεί με συμπληρωματικούς όρους, με τη βοήθεια κάποιας λεξιλογικής πηγής και την εύρεση συνωνύμων και αντωνύμων. Το μειονέκτημα σε αυτού του είδους τη μεθοδολογία εντοπίζεται στην αδυναμία επαρκούς συνυπολογισμού του εννοιολογικού τομέα στον οποίο ανήκουν τα έγγραφα που εξετάζονται. Οπότε, το παραγόμενο λεξικό καταλήγει να συσχετίζεται ελάχιστα με το περιεχόμενο της συλλογής κειμένου και κατά συνέπεια να μην είναι αποτελεσματικό για τον επιδιωκόμενο σκοπό αιχμαλώτισης της συναισθηματικής πολικότητας των λέξεων. Η προσαρμογή και συσχέτιση του λεξικού με τον τομέα των κειμένων είναι εξαιρετικά σημαντική για την αποτελεσματικότητα της εργασίας ανάλυσης συναισθήματος καθώς οι λέξεις λαμβάνουν διαφορετικό νόημα και κατ' επέκταση διαφορετική πολικότητα σε διαφορετικά εννοιολογικά πλαίσια.

Προκειμένου να επιτευχθεί σύνδεση με τα μελετώμενα έγγραφα, ο χρήστης, δεδομένων κάποιων πρότυπων όρων με γνωστή συναισθηματική πολικότητα, ορίζει μία σειρά από επιθυμητούς κανόνες εξάρτησης μεταξύ των λέξεων ώστε να φανερωθεί η γλωσσική συνοχή μεταξύ τους, αν αυτή υπάρχει, και να εμπλουτιστεί ύστερα το λεξικό με νέες λέξεις κάθε είδους συναισθηματικού προσανατολισμού. Τα λεξικά που χρησιμοποιούνται για την απόδοση συναισθηματικής πολικότητας στην πρακτική εφαρμογή της συγκεκριμένης διπλωματικής είναι 5 σε πλήθος και ο τρόπος κατασκευής τους αναλύεται παρακάτω. Πρόκειται ονομαστικά για τα λεξικά 'SentiWordNet', 'SPLM', 'SentiDomain', 'SentiPosNeg' και 'SentiDraw'.

6.1. SentiWordNet

Το SentiWordNet 3.0 ή εν συντομία SWN είναι ένα ευρέως γνωστό λεξικό γενικού σκοπού για την εξαγωγή του συναισθηματικού προσανατολισμού των λέξεων. Αποτελεί μία βελτιωμένη εκδοχή του SentiWordNet 1.0. Και τα δύο συνιστούν επεκτάσεις διαφορετικών εκδοχών του WordNet (2.0 και 3.0 αντίστοιχα), παραγόμενα

αποδίδοντας σε κάθε ομάδα συνωνύμων (synset), τρεις αριθμητικές τιμές που αφορούν τον βαθμό θετικότητας, αρνητικότητας και ουδετερότητας.

Το WordNet είναι μία ελεύθερα διαθέσιμη λεξιλογική βάση δεδομένων στα αγγλικά η οποία ξεκίνησε να αναπτύσσεται το 1985 σε εργαστήριο του Πανεπιστημίου Princeton και ακόμη συνεχίζει να ανανεώνεται. Αποτελείται από ουσιαστικά, ρήματα, επίθετα και επιρρήματα, ομαδοποιημένα σε ομάδες συνωνύμων οι οποίες συνδέονται μεταξύ τους μέσω εννοιολογικών σχέσεων (Miller, WordNet: A Lexical Database for English, 1995), (Miller & Fellbaum, Semantic networks of English, 1991). Κάθε ομάδα συνωνύμων συνοδεύεται από έναν σύντομο ορισμό και μερικές φορές και από μία ή περισσότερες προτάσεις που περιέχουν τις λέξεις της ομάδας. Διαφορετικός συντακτικός ρόλος δεδομένης λέξης στο γλωσσικό συγκείμενο έχει ως αποτέλεσμα επιπρόσθετη καταχώρισή της στη βάση δεδομένων, συνδυασμένη με τον εκάστοτε συντακτικό ρόλο.

Στο WordNet τα ουσιαστικά μπορούν να συνδέονται με τα συνώνυμά τους, τα αντώνυμά τους και τις σχέσεις υπωνυμίας-υπερνυμίας και μερωνυμίας-ολονυμίας. Στη σημασιολογική σχέση υπωνυμίας-υπερνυμίας, η σημασία μίας λέξης συμπεριλαμβάνεται στη σημασία μίας άλλης λέξης-γενικότερης έννοιας, ενώ στη σχέση μερωνυμίας-ολονυμίας, μία λέξη είναι μέρος μίας άλλης (π.χ. πόρτα με κτίριο). Η κατεύθυνση της εκάστοτε σχέσης καθορίζει τον όρο που χρησιμοποιείται για τη δήλωσή της, δηλαδή το εάν πρόκειται για υπωνυμία ή υπερνυμία στην πρώτη περίπτωση και αν έχουμε μερωνυμία ή ολονυμία στη δεύτερη. Τα ρήματα, από την άλλη πλευρά, μπορούν να συνδέονται με τα συνώνυμά τους, τα αντώνυμά τους, με σχέση τροπωνυμίας ή συνεπαγωγής. Όπως δηλώνεται και από την ονομασία της, μία σχέση τροπωνυμίας αφορά δύο ρήματα, εκ των οποίων το πρώτο εκφράζει τον τρόπο με τον οποίο γίνεται η ενέργεια του δεύτερου ρήματος. Τα επίθετα και τα επιρρήματα νοηματοδοτούνται στο WordNet είτε μέσω των συνωνύμων τους, είτε μέσω των αντωνύμων τους.

Το SentiWordnet κατασκευάστηκε συνδυάζοντας την ενισχυτική μάθηση με χρήση ενός μικρού αρχικού συνόλου με ομάδες συνωνύμων από το WordNet και τον αλγόριθμο τυχαίου περιπάτου. Οι καινούριες τιμές Pos(s), Neg(s) και Obj(s) που προστίθενται στο SentiWordNet φανερώνουν το πόσο θετικοί, αρνητικοί και ουδέτεροι (αντικειμενικοί) αντίστοιχα είναι οι όροι που συμπεριλαμβάνονται σε κάθε ομάδα συνωνύμων s (Baccianella, Esuli, & Sebastiani, 2010). Καθεμία τιμή ανήκει στο διάστημα [0,1] και για δεδομένη ομάδα συνωνύμων λέξεων, το άθροισμα των τριών τιμών ισούται με 1. Σε περίπτωση που μία ομάδα όρων μπορεί να λάβει διαφορετικά νοήματα, εμφανίζεται στο λεξικό τόσες φορές όσα και τα διαφορετικά νοήματα, χαρακτηριζόμενη από διαφορετική τριάδα τιμών σε κάθε ξεχωριστή εγγραφή. Χαρακτηριστικό παράδειγμα καταχωρίσεων στο SentiWordNet απεικονίζεται παρακάτω, όπου φαίνεται να λείπει η στήλη με τις τιμές αντικειμενικότητας. Παρ' όλα αυτά, υπολογίζεται εύκολα ως:

$$ObjScore = 1 - PosScore - NegScore. \quad (21)$$

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00754107	0.375	0.125	diligent#2	characterized by care and perseverance in carrying out tasks; "a diligent detective investigates all clues"; "a diligent search of the files"
a	00754393	0.125	0.25	sedulous#1 assiduous#1	marked by care and persistent effort; "her assiduous attempts to learn French"; "assiduous research"; "sedulous pursuit of legal and moral principles"

Εικόνα 4. Δείγμα 2 καταχωρίσεων στο SentiWordNet.

Η ελεύθερα προσβάσιμη λεξιλογική πηγή SentiWordNet αποτέλεσε ένα από τα λεξικά που αποφασίστηκε να χρησιμοποιηθούν κατά την πειραματική μελέτη. Δεδομένου ότι οι τιμές πολικότητας σε αυτό το λεξικό ακολουθούν μία κλασματική και όχι συνεχή επισήμανση, όπως ιδανικά θα ήταν επιθυμητό (Labille, Gauch, & Alfarhood, 2017), ακολουθείται η προσέγγιση του Petter Tonberg (ideone.com, n.d.) για να γίνει ο απαραίτητος μετασχηματισμός.

Οι επιτρεπτές ετικέτες μερών του λόγου στις οποίες μπορούν να ανήκουν οι λέξεις είναι οι ‘n’, ‘v’, ‘a’ και ‘r’, δηλαδή ‘ουσιαστικό’, ‘ρήμα’, ‘επίθετο’ και ‘επίρρημα’ αντίστοιχα. Ο αλγόριθμος βρίσκει στο αρχείο του λεξικού μορφής txt, για κάθε εγγραφή, το μέρος του λόγου (‘POS’) των όρων που ανήκουν σε αυτή καθώς και τον μοναδικό αριθμό ταυτοποίησής της (‘ID’). Για να επιτευχθεί αυτό, εκμεταλλεύεται την οργάνωση του λεξικού σε στήλες που έχουν την ακόλουθη σειρά: ‘POS’, ‘ID’, ‘PosScore’, ‘NegScore’, ‘SynsetTerms’, ‘Gloss’. Έπειτα, υπολογίζει τη διαφορά της θετικής τιμής (‘PosScore’) από την αρνητική (‘NegScore’) προκειμένου να αποκτηθεί μία συνολική τιμή s που θα φανερώνει το πόσο θετικοί ή αρνητικοί είναι οι όροι που συμπεριλαμβάνονται σε κάθε ομάδα συνωνύμων. Ταυτόχρονα, λαμβάνεται για κάθε εγγραφή του λεξικού, κάθε συνώνυμος όρος, η κατάταξη r που του έχει αποδοθεί και το μέρος του λόγου στο οποίο αυτός ανήκει. Η τιμή κατάταξης προκύπτει από τη στήλη ‘SynsetTerms’, ως ο αριθμός μετά τη δίεση #, σε αντίθεση με το μέρος του λόγου που βρίσκεται από τη στήλη ‘POS’. Έτσι, φτιάχνεται ένα νέο λεξικό που αντιστοιχίζει κάθε μοναδικό συνδυασμό λέξης-μέρους του λόγου με την κατάταξη που έχει και την υπολογισμένη διαφορά. Εξαιτίας του γεγονότος ότι μπορεί μία λέξη με το μέρος του λόγου στο οποίο ανήκει, να εμφανίζεται πολλές φορές στο αρχικό λεξικό, ενδέχεται να υπάρχουν γι’ αυτήν πολλά ζεύγη κατάταξης-τιμής συναισθηματικής πολικότητας. Γι’ αυτό και για κάθε καταχωρημένο συνδυασμό λέξης-μέρους του λόγου, υπολογίζεται ένα συνολικό σκορ συναισθηματικού προσανατολισμού βάσει του ακόλουθου τύπου:

$$score_j = \frac{\sum_{i=1}^{p_j} \frac{s_i}{r_i}}{\sum_{i=1}^{p_j} \frac{1}{r_i}} = \sum_{i=1}^{p_j} \frac{s_i}{r_i} \frac{1}{r_i} = \sum_{i=1}^{p_j} s_i \ . \quad (22)$$

Η μεταβλητή $score_j$ συμβολίζει τη συνολική τιμή συναισθηματικής πολικότητας της j καταχώρησης στο νέο λεξικό και s_i είναι η υπολογισμένη τιμή πολικότητας του i ζεύγους τιμών που ανατίθεται στον j συνδυασμό λέξης-μέρους του λόγου. Επίσης, με r_i αντιπροσωπεύεται η κατάταξη που έχει ο j συνδυασμός λέξης-μέρους του λόγου στο i ζεύγος, ενώ με p_j το πλήθος των ζευγών κατάταξης-τιμής συναισθήματος για τον j συνδυασμό λέξης-μέρους του λόγου.

Το παραγόμενο λεξικό μετατρέπεται σε πλαίσιο δεδομένων και αποθηκεύεται σε μορφή csv, με την ονομασία 'SENTIWORDNET.csv'. Οι τιμές συναισθηματικού προσανατολισμού για όλες τις εγγραφές ανήκουν στο κλειστό διάστημα [-1, 1]. Οπότε, γίνεται κατανοητό ότι δεν χρειάζεται να πραγματοποιηθεί περαιτέρω κάποιου είδους κανονικοποίηση.

6.2. SPLM

Το λεξικό SPLM είναι ένα λεξικό που προτείνουν οι Almatarneh & Gamallo (Almatarneh & Gamallo, 2017) και κατασκευάζεται σύμφωνα με τους παρακάτω υπολογισμούς. Αρχικά, υπολογίζεται η σχετική συχνότητα εμφάνισης κάθε λέξης στα έγγραφα κάθε δυνατής βαθμολογίας. Έτσι, ο όρος t έχει σχετική συχνότητα εμφάνισης στα έγγραφα βαθμολογίας c, ίση με:

$$tf_{t,d_c} = \frac{count(t, d_c)}{\sum_{t' \epsilon d_c} count(t', d_c)} , \quad (23)$$

όπου $count(t, d_c)$ είναι το πλήθος των φορών που εμφανίζεται ο όρος t στα έγγραφα με βαθμολογία c. Ο παρονομαστής του κλάσματος δηλώνει το συνολικό πλήθος των όρων που εμφανίζονται στα έγγραφα με βαθμολογία c, όσες φορές κι αν αυτοί παρουσιάζονται. Αν θεωρήσουμε ότι οι κριτικές βαθμολογίας 3 που εκλαμβάνονται ως ουδέτερες, αγνοούνται, η οριακή τιμή B που ξεχωρίζει τις αρνητικές από τις θετικές κριτικές τίθεται να είναι το 2. Επεξηγηματικά, μέχρι και τη βαθμολογία 2, οι κριτικές είναι αρνητικές, ενώ μετά τις παραλειπόμενες ουδέτερες κριτικές, ακολουθούν οι θετικές κριτικές με δύο δυνατές βαθμολογίες (4 και 5).

Τότε, οι μέσοι όροι σχετικών συχνοτήτων εμφάνισης κάθε όρου t στις αρνητικές και τις θετικές κριτικές, συμβολισμένοι, αντίστοιχα, ως Avn και Avp, υπολογίζονται ως εξής:

$$Avn_t = \frac{\sum_{c=1}^{2} tf_{t,d_c}}{2}, \tag{24}$$

$$Avp_t = \frac{\sum_{c=4}^{5} tf_{t,d_c}}{2}. \tag{25}$$

Χρησιμοποιώντας τις ποσότητες Avn_t και Avp_t, η συναισθηματική πολικότητα κάθε λέξης δίνεται από τη διαφορά τους. Δηλαδή, ισχύει:

$$D_t = Avp_t - Avn_t. \tag{26}$$

Βάσει του αν η τιμή αυτή είναι θετική ή αρνητική για έναν όρο t, αποφασίζεται αν ο συγκεκριμένος όρος είναι θετικά ή αρνητικά προσκείμενος. Προκειμένου να αποκτηθούν τιμές στο διάστημα [-1,1], πραγματοποιείται μία min-max κανονικοποίηση, παράγοντας, τελικά, στη θέση της D_t, την $Norm\_D_t$.

Η min-max κανονικοποίηση, όντας μία από τις πιο συνηθισμένες μεθόδους κανονικοποίησης των δεδομένων, πετυχαίνει τη συμπίεση όλων των αριθμητικών τιμών σε ένα περιορισμένο εύρος τιμών. Εξαιτίας του γεγονότος ότι οι τιμές πολικότητας μπορούν να λάβουν και αρνητικές τιμές, το καινούριο εύρος τιμών αποφασίστηκε να είναι το [-1, 1]. Ο μαθηματικός τύπος μετασχηματισμού κάθε τιμής x_{old} σε μία νέα τιμή x_{new}, ώστε αυτή να ανήκει πλέον στο διάστημα $[min_{new}, max_{new}]$ αντί για το διάστημα $[min_{old}, max_{old}]$, δίνεται από τον τύπο (27) :

$$x_{new} = (max_{new} - min_{new}) \cdot \frac{x - min_{old}}{(max_{old} - min_{old})} + min_{new}. \tag{27}$$

Η εκτέλεση κατάλληλου κώδικα παράγει ένα αρχείο τύπου csv, το ‘SPLM1.csv’ που συμπεριλαμβάνει τις τιμές πολικότητας για όλες τις λέξεις που συμπεριλαμβάνονται στις κριτικές του συνόλου εκπαίδευσης. Στην πραγματικότητα, αν μελετηθεί ο εκτελούμενος κώδικας, μπορεί να διαπιστωθεί ότι ο αρχικός διαχωρισμός των κριτικών, απομακρύνοντας όσες ήταν ουδέτερες, ήταν διαφορετικός. Είχε εφαρμοστεί στρωματοποιημένη 5-fold cross-validation, παράγοντας 5 διαφορετικά σύνολα εκπαίδευσης και σύνολα ελέγχου, προκειμένου να ελαττωθεί η επίδραση της τυχαιότητας διαμερισμού του συνόλου κριτικών.

Ο αλγόριθμος της στρωματοποιημένης k-fold cross-validation είναι ίδιος με αυτόν της k-fold cross-validation με τη μόνη διαφορά ότι ο αρχικός διαχωρισμός σε k folds υλοποιείται με τέτοιον τρόπο ώστε κάθε fold να περιλαμβάνει προσεγγιστικά το ίδιο ποσοστό δειγμάτων σε καθεμία κλάση με αυτό που είχε το αρχικό σύνολο δεδομένων, πριν τον διαχωρισμό. Ωστόσο, ο υψηλός χρόνος εκτέλεσης που απαιτούνταν από τον

υπολογιστή σε επόμενο βήμα προκειμένου να χρησιμοποιηθούν τα αποτελέσματα και των 5 λεξικών, οδήγησε στην εγκατάλειψη αυτής της προσέγγισης. Χρησιμοποιήθηκε, τελικά, μόνο το σύνολο εκπαίδευσης και ελέγχου του 1ου fold. Η στρωματοποίηση πραγματοποιήθηκε βάσει των δύο κλάσεων βαθμολογίας των κριτικών προκειμένου να ληφθεί υπόψιν η μεγάλη ανισορροπία κλάσεων.

6.3. SentiDomain

Μία άλλη προσέγγιση υπολογισμού της πολικότητας των λέξεων ενσωματώνει υπό συνθήκη πιθανότητες που περικλείουν την πληροφορία του είδους πολικότητας των κριτικών (δηλαδή θετικές / αρνητικές) αντί για την ακριβή αριθμητική τιμή αξιολόγησής τους (δηλαδή 1 ή 2 ή 4 ή 5). Πιο συγκεκριμένα, πρώτα, βρίσκονται οι τιμές των εκ των υστέρων πιθανοτήτων κάθε όρος t να έχει θετική πολικότητα και αρνητική πολικότητα, $p(pos|t)$ και $p(neg|t)$ αντίστοιχα σύμφωνα με τους τύπους:

$$p(pos|t) = \frac{p(pos) \cdot p(t|pos)}{p(t)} \tag{28}$$

και

$$p(neg|t) = \frac{p(neg) \cdot p(t|neg)}{p(t)} . \tag{29}$$

Ως $p(pos)$ ορίζεται ο λόγος του πλήθους των όρων που ανήκουν στο κείμενο των θετικά αξιολογούμενων κριτικών προς το πλήθος των όρων που εμφανίζονται σε όλες τις κριτικές (δηλαδή το συνολικό πλήθος φορών εμφάνισης όλων των όρων). Με παρόμοιο τρόπο, ορίζεται και η μεταβλητή $p(neg)$, με την εξαίρεση ότι αυτή αφορά τη συλλογή κειμένου των κριτικών αρνητικής βαθμολογίας. Και οι δύο συναντώνται στη βιβλιογραφία ως εκ των προτέρων πιθανότητες των δύο κλάσεων. Οι μεταβλητές $p(t|pos)$ και $p(t|neg)$ αντιπροσωπεύουν τις πιθανότητες δεδομένης μίας θετικής κριτικής ή αρνητικής κριτικής αντίστοιχα, να βρεθεί ο όρος t. Δίνονται από τους ακόλουθους τύπους, με τις $count(t, d_{pos})$ και $count(t, d_{neg})$ να συμβολίζουν την απόλυτη συχνότητα εμφάνισης του όρου t στα έγγραφα-κριτικές θετικής και αρνητικής αξιολόγησης αντίστοιχα:

$$p(t|pos) = \frac{count(t, d_{pos})}{\sum_{t' \epsilon d_{pos}} count(t', d_{pos})} \tag{30}$$

και

$$p(t|neg) = \frac{count(t, d_{neg})}{\sum_{t' \epsilon d_{neg}} count(t', d_{neg})}. \quad (31)$$

Τέλος, η μεταβλητή $p(t)$ υποδηλώνει την πιθανότητα εμφάνισης του όρου t σε κείμενο οποιουδήποτε είδους κριτικής και δίνεται με τη βοήθεια του Θεωρήματος Ολικής Πιθανότητας από τον τύπο:

$$p(t) = p(pos) \cdot p(t|pos) + p(neg) \cdot p(t|neg) \quad (32)$$

Για τον παραπάνω τύπο λαμβάνεται υπόψιν ότι τα κείμενα των κριτικών διαχωρίζονται μεταξύ των θετικά και αρνητικά βαθμολογημένων. Τελικά, η πιθανολογική τιμή συναισθηματικού προσανατολισμού που ανατίθεται στον όρο t είναι η ακόλουθη:

$$Scoreprob(t) = p(pos|t) - p(neg|t). \quad (33)$$

Η εκτέλεση ανάλογου κώδικα έχει ως αποτέλεσμα ένα αρχείο τύπου csv με όνομα ‘Sentidomain1.csv’.

6.4. SentiPosNeg

Η προσέγγιση SentiPosNeg θέτει ως τιμή πολικότητας για έναν όρο το πόσο πιθανόν είναι αυτός ο όρος να είναι θετικά φορτισμένος. Για το σκοπό αυτό, υπολογίζει το πηλίκο του πλήθους των εμφανίσεων του συγκεκριμένου όρου στις θετικές κριτικές προς το πλήθος των εμφανίσεων σε όλων των ειδών τις κριτικές. Επεξηγηματικά, ο τύπος είναι ο (34):

$$posprob(t) = \frac{count(t, d_{pos})}{count(t, d_{pos}) + count(t, d_{neg})}. \quad (34)$$

Επειδή η μεταβλητή $posprob$ για κάθε όρο t, λόγω του ορισμού της ως πιθανότητα, λαμβάνει τιμές στο διάστημα [0, 1], ενώ τα υπόλοιπα μέτρα υπολογισμού τιμών πολικότητας παίρνουν και αρνητικές τιμές, εφαρμόζεται η min-max κανονικοποίηση στη μεταβλητή. Στόχος είναι η επέκταση των τιμών της στο κλειστό διάστημα [-1, 1].

Ο κατασκευασμένος κώδικας για αυτήν την τεχνική δίνει το αρχείο τύπου csv ‘Sentiposneg1.csv’.Η τελική τιμή συναισθηματικής πολικότητας για τους συμπεριλαμβανόμενους όρους είναι καταγεγραμμένη στη στήλη ‘Norm_Pos_prob_w’.

6.5. SentiDraw

Σε αυτήν την τεχνική δημιουργίας του λεξικού συναισθημάτων, βρίσκεται, σε πρώτη φάση, η τιμή της μεταβλητής P_{t,d_c} που προσδιορίζεται ως εξής:

$$P_{t,d_c} = \frac{count(t, d_c)}{\sum_{c' \epsilon \{1,2,4,5\}} count(t, d_{c'})}, c = 1,2,4,5. \quad (35)$$

Ο ορισμός της μεταβλητής $count(t, d_c)$ έχει γραφεί στην περίπτωση της μεθόδου SPLM. Ακολουθεί η πρόσδοση μίας ακέραιας τιμής σε κάθε αριθμητική ετικέτα βαθμολογίας μεταξύ του 1 και του 5 (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Sharma & Dutta, 2021). Ειδικότερα, οι βαθμολογίες 1, 2, 3, 4 και 5 αντιστοιχίζονται με τις τιμές -5, -3, 0, +3 και +5 αντίστοιχα. Τελικά, η τιμή που εκφράζει το συναίσθημα πίσω από έναν όρο t δίνεται ως:

$$Ws(t) = (-5) \cdot P_{t,d_1} + (-3) \cdot P_{t,d_2} + 3 \cdot P_{t,d_4} + 5 \cdot P_{t,d_5} \quad (36)$$

και καλείται σταθμισμένη τιμή συναισθήματος για αυτόν τον όρο. Όλες οι παραγόμενες σταθμισμένες τιμές κανονικοποιούνται με min-max κανονικοποίηση ώστε το καινούριο διάστημα να είναι και εδώ το [-1, 1].

Το αρχείο τύπου csv με αυτές τις τιμές είναι το 'Sentidraw1.csv' που τις συγκεντρώνει στη στήλη 'Norm_W_s_t'.

7. ΠΕΙΡΑΜΑΤΙΚΗ ΕΦΑΡΜΟΓΗ - ΔΕΔΟΜΕΝΑ, ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

7.1. ΔΕΔΟΜΕΝΑ ΑΝΑΦΟΡΑΣ

Στην παρούσα εργασία, αξιολογήθηκε ως προς τη συναισθηματική πόλωση ένα σύνολο 3.270 online κριτικών, αντλημένων από το Tripadvisor, που περιλαμβάνουν γνώμες ανθρώπων αναφορικά με το παλιό λιμάνι των Χανίων. Το κείμενο όλων των κριτικών είναι γραμμένο με αγγλικούς χαρακτήρες, παρ' όλο που οι χρήστες μπορεί να προέρχονται από οποιοδήποτε μέρος του πλανήτη, κάτι που αποδεικνύεται από τα ονόματά τους (π.χ. 'ειρηνη κ' ή 'Владимир П'). Οι διαφορετικοί χρήστες αριθμούν 3.142 σε πλήθος, που σημαίνει ότι μερικοί έχουν γράψει περισσότερες της μίας κριτικής. Δεδομένου ότι κάθε καινούρια καταχώριση ενός χρήστη στο σύνολο κριτικών σημαίνει και διαφορετική αξιολόγηση του εκάστοτε χρήστη με ή χωρίς την ίδια βαθμολογία, συνεπάγεται ότι χρειάζεται να μελετηθούν και οι 3.270. Χαρακτηριστικά παραδείγματα τέτοιων χρηστών και κριτικών φαίνονται στην ακόλουθη εικόνα. Υπό αυτό το πλαίσιο, κάθε εγγραφή του πίνακα διαθέσιμων δεδομένων, όποια και αν είναι η έκτασή της σε προτάσεις ή λέξεις, αντιμετωπίζεται ως ένα ξεχωριστό έγγραφο που χρήζει περαιτέρω ανάλυσης.

Αναγνωριστικός αριθμός κριτικής	Όνομα χρήστη (User)	Τίτλος (Title)	Σχόλια (Comments)	Βαθμολογία (Rating)
2179	BostonProf	Stunning	Loved it here, so beautiful and such a welcoming feel both during the day and late into the evening.Easy to navigate around on foot (bit tricky by car due to one way system and crazy driving style).The walk out to the lighthouse gives great views and the cooling sea breeze from high on the wall meant we didnt melt from the heat.Stunning Venetian buildings, good place to buy silver jewellery, good priced places to eat too (esp a few streets back from the waterfront).Enjoyed by all the family.	5
3253	BostonProf	The hawkers have ruined it!	This was our second trip to Chania and we were excited to go back to the Old Venetian Harbor where we spent part of the afternoon on our wedding day in 2006. We we pretty disappointed with the cheesy restaurant hawkers who actually block you from walking, get right in your face with their B.S. about Mama cooking Cretan food in their kitchen and then get nasty when you turn them down. They have ruined such a beautiful place. It has turned into a real tourist trap.	2

2062	Carol R	Gorgeous, but a tourist trap	The Old Venetian Harbor in Chania is indeed lovely and a wonderful area to explore. But... it's sadly cluttered with ultra cheesy souvenir shops and mediocre tourist trap restaurants, and endless mounds of tourists and tour groups. This can all be avoided by walking it early or late. Also, the streets just behind the harbor are lovely-with interesting shops and beautiful old Venetian and Ottoman architecture. It's definitely worth a visit, especially for the elegant light house and old walls. Just be sure to avoid the crowds if you can.	4
3265	Carol R	Beautiful harbor!	We loved strolling around the harbor - enjoying the views, the beautiful weather, and the colors of all the old buildings!	4

Εικόνα 5. Καταχώριση κριτικών από τον χρήστη 'BostonProf' και του 'Carol R'. Ο πρώτος δίνει στην πρώτη κριτική του βαθμολογία ίση με 5, ενώ στη δεύτερη δίνει 2. Αντίθετα, ο χρήστης 'Carol R' δίνει και στις 2 κριτικές που κάνει, βαθμολογία ίση με 4.

Το αρχείο τύπου .csv που περιλαμβάνει τις μελετώμενες κριτικές αποτελείται από τις στήλες:

- ✓ User

  Πρόκειται για το όνομα του χρήστη που δημοσίευσε μία κριτική. Θεωρείται ως το μοναδικό αναγνωριστικό για τους χρήστες.

- ✓ Title

  Ο τίτλος κάθε κριτικής προσφέρει μία σύντομη περιγραφή της κριτικής.

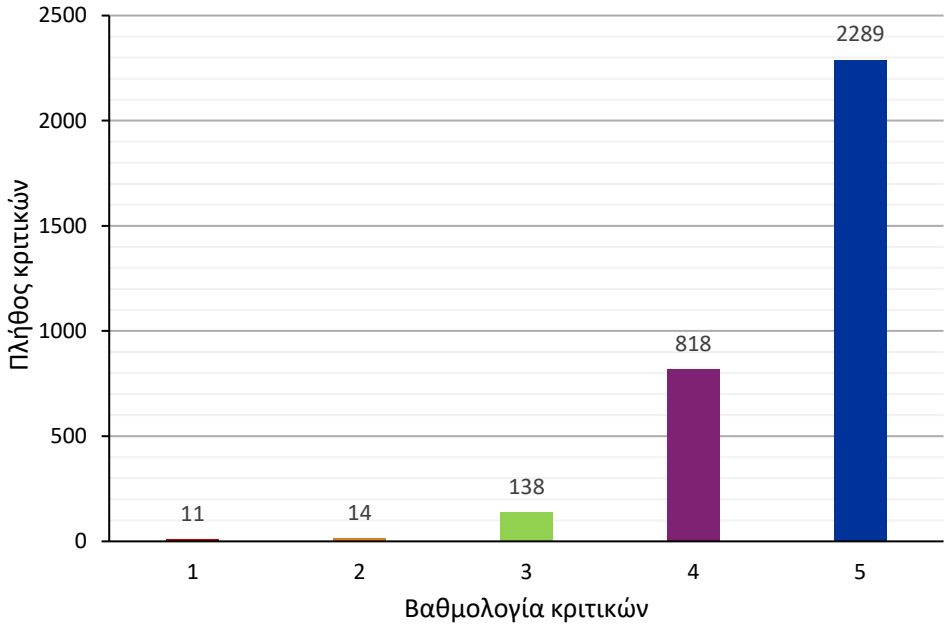
- ✓ Comments

  Αυτή η στήλη περιλαμβάνει το βασικό κείμενο κάθε κριτικής.

- ✓ Rating

  Κάθε γραμμή αυτής της στήλης περιέχει τη βαθμολογία που δίνει κάθε χρήστης βάσει της κριτικής του. Οι δυνατές τιμές είναι οι ακέραιοι μεταξύ του 1 και του 5.

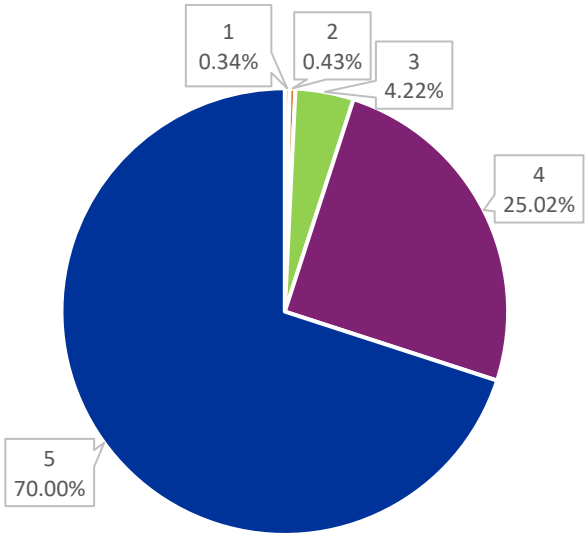
Στο Σχήμα 1 φαίνεται η κατανομή των βαθμολογιών σε όλες τις διαθέσιμες καταχωρίσεις κριτικών. Οι κλάσεις 1 και 2 συναντώνται αθροιστικά σε κάτω από 50 κριτικές, ενώ η κλάση 5 εμφανίζεται σε 2.289 κριτικές. Έτσι, γίνεται αντιληπτό ότι η ανισορροπία μεταξύ των κλάσεων είναι πολύ μεγάλη, με τις κριτικές βαθμολογίας 1 και 2 να αποτελούν λιγότερο από το 1% των κριτικών (Σχήμα 2).



Σχήμα 1. Ραβδόγραμμα απόλυτων συχνοτήτων των βαθμολογιών των κριτικών.

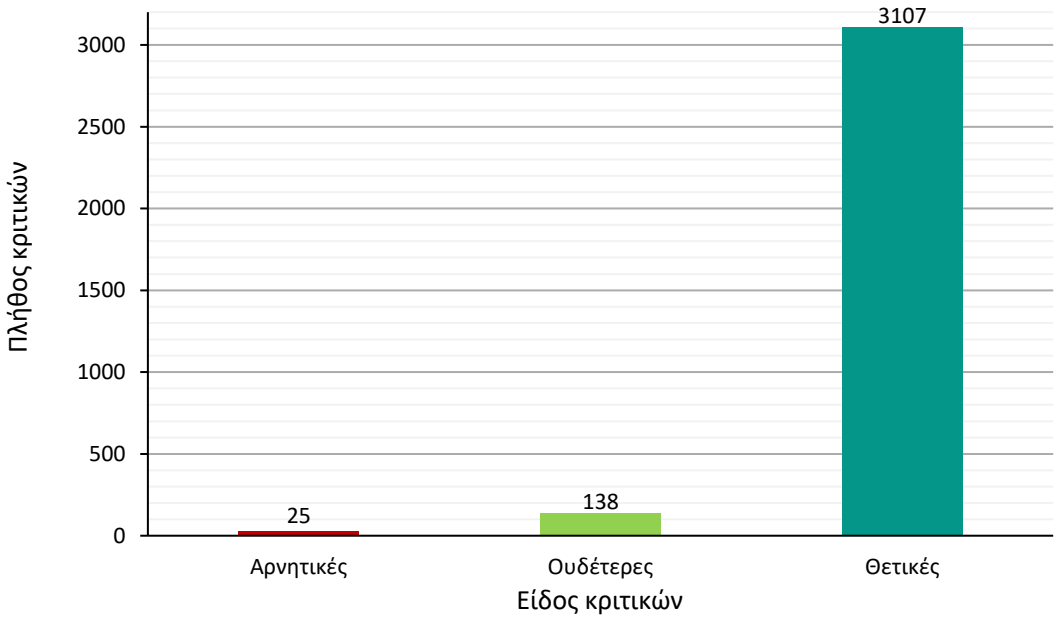
Πίνακας 1. Απόλυτες συχνότητες εμφάνισης κριτικών κάθε δυνατής βαθμολογίας μεταξύ του 1 και του 5.

Βαθμολογία κριτικών	Πλήθος κριτικών
1	11
2	14
3	138
4	818
5	2289



Σχήμα 2. Διάγραμμα πίτας με τις σχετικές συχνότητες των βαθμολογιών των κριτικών σε μορφή ποσοστού.

Οι κριτικές με βαθμολογία 1 και 2, όπως επίσης και αυτές με βαθμολογία 4 και 5 συνήθως είναι δύσκολα διαχωρίσιμες μεταξύ τους βάσει του περιεχομένου τους (Wu & Ji, 2016). Γι' αυτόν τον λόγο, όπως συνηθίζεται, οι υποψήφιες κλάσεις ταξινόμησης κάθε κριτικής μετασχηματίστηκαν, ύστερα από ομαδοποίηση, από πέντε σε τρεις: τις θετικές, τις ουδέτερες και τις αρνητικές. Ως θετικές κριτικές εκλαμβάνονται όσες έχουν βαθμολογηθεί με 4 ή 5, ως ουδέτερες όσες έχουν βαθμολογηθεί με 3 και ως αρνητικές, όλες οι υπόλοιπες, δηλαδή όσες έχουν βαθμολογηθεί με 1 ή 2. Εκτός από αυτό, εύκολα μπορεί να παρατηρηθεί πως μία τέτοιου είδους συγχώνευση των κλάσεων καθίσταται αναγκαία, καθώς οι καταχωρίσεις κριτικών με βαθμολογία 1 και 2 δεν μπορούν να θεωρηθούν επαρκείς σε πλήθος για ικανοποιητική ανάλυση της σημασιολογικής δομής του κειμένου που περιέχουν.



Σχήμα 3. Ραβδόγραμμα απόλυτων συχνοτήτων των ομαδοποιημένων βαθμολογιών των κριτικών.

(α) (β) (γ)

Σχήμα 4. Γράφημα πίτας της ποσοστιαίας κατανομής (α) των βαθμολογιών 1 και 2 στο σύνολο των αρνητικών κριτικών, (β) της βαθμολογίας 3 στο σύνολο των ουδέτερων κριτικών και (γ) των βαθμολογιών 4 και 5 στο σύνολο των θετικών κριτικών.

Έπειτα από διερεύνηση, δεν υπάρχει καμία καταχώριση κριτικής στην οποία δεν έχει δοθεί βαθμολογία από τον χρήστη. Σε πέντε καταχωρίσεις δεν είναι διαθέσιμο το όνομα του χρήστη, χωρίς, ωστόσο, αυτό να επηρεάζει, οδηγώντας τυχόν σε μη συμπερίληψή τους στο σύνολο των υπό μελέτη κριτικών.

7.2. TRIPADVISOR

Το Tripadvisor είναι η μεγαλύτερη ταξιδιωτική διαδικτυακή πλατφόρμα στον κόσμο. Μέσω αυτής, ταξιδιώτες από όλο τον κόσμο μπορούν να επιλέγουν τους καταλληλότερους γι' αυτούς προορισμούς, καταλύματα, εστιατόρια, πτήσεις, τουριστικές περιηγήσεις, αξιοθέατα, κρουαζιέρες και αυτοκίνητα προς ενοικίαση, συγκρίνοντας τις τιμές μεταξύ των εκατομμύρια εναλλακτικών επιλογών. Η λειτουργία της στηρίζεται στα σχόλια και τις αξιολογήσεις των χρηστών που έχουν ήδη ταξιδέψει ή επισκεφτεί διάφορα μέρη και έχουν μοιραστεί τις εμπειρίες τους. Η απόφαση ενός τουρίστα, για παράδειγμα, για το πού θα ταξιδέψει ή σε ποιο εστιατόριο θα φάει, επηρεάζεται σε μεγάλο βαθμό, από την προδιάθεση που του δημιουργούν οι υπάρχουσες κριτικές άλλων ανθρώπων. Η επιτυχία του Tripadvisor έγκειται στο γεγονός ότι συγκεντρώνει πάνω από 859 εκατομμύρια κριτικές για 43 αγορές οι οποίες είναι διαθέσιμες σε 22 γλώσσες (About Tripadvisor, χ.χ.).

Για το 2022, τόσο το πλήθος, όσο και το είδος των αξιολογήσεων για το νησί της Κρήτης συνέβαλαν στο να διακριθεί η Κρήτη από το Tripadvisor ως ο δεύτερος πιο δημοφιλής προορισμός της Ευρώπης, κατακτώντας παράλληλα και την 5^{η} θέση σε παγκόσμιο επίπεδο. Επίσης, αποφασίστηκε από τους αναγνώστες της ίδιας τουριστικής πλατφόρμας, για το 2023, ο ίδιος προορισμός να ανακηρυχτεί δεύτερος καλύτερος παγκοσμίως από πλευράς φαγητού, αποτελώντας πλέον ένα από τα πιο αγαπητά και δημοφιλή μέρη του πλανήτη.

7.3. ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ

Η διαδικασία αυτή πραγματοποιείται πριν από την ανάλυση του κειμένου και χρησιμεύει στον καθαρισμό των ακατέργαστων αρχικών δεδομένων, δηλαδή την αναγνώριση και την αφαίρεση άχρηστου περιεχομένου. Έτσι, μπορούν, έπειτα, να αποσπαστούν χαρακτηριστικά (features) που θα αξιοποιηθούν αποτελεσματικά σε αλγορίθμους μηχανικής μάθησης ή κάποια ευρύτερη μεθοδολογία εξόρυξης γνώμης. Τα άχρηστα δεδομένα, όπως τα σημεία στίξης, τα άρθρα, οι ημερομηνίες και γενικά οι αριθμοί, αποτελούν θόρυβο ο οποίος δεν προσφέρει κάποια πολύτιμη πληροφορία, αλλά αντίθετα καθιστά τον όγκο του κειμένου μη διαχειρίσιμο. Υπό το πλαίσιο αυτό, η εφαρμογή της προεπεξεργασίας κρίνεται πολύτιμη, καθώς, μέσω αυτής, τα δεδομένα που διατίθενται σε μορφή κειμένου μετατρέπονται σε θεμελιώδεις μονάδες κατάλληλης μορφής για περαιτέρω ανάλυση.

Παρακάτω, αναφέρονται τα κυριότερα στάδια προεπεξεργασίας που ακολουθούνται στη βιβλιογραφία, είτε αποσπασματικά, είτε σε συνδυασμό μεταξύ τους. Ταυτόχρονα, είναι και αυτά που ακολουθήθηκαν στην πειραματική εφαρμογή της παρούσας διπλωματικής.

1. Μετατροπή των emoticons στις αντίστοιχες λεξιλογικές ετικέτες

Για το σκοπό αυτό, χρησιμοποιείται η βιβλιοθήκη της Python «emot» (Shah & Rohilla, χ.χ.) η οποία, σε πρώτο στάδιο, ανιχνεύει σε δεδομένο κείμενο όλα τα emoticons και ύστερα με τη βοήθεια λεξικού τα αντικαθιστά με τις ερμηνείες τους, γραμμένες με αγγλικά πεζά γράμματα. Τα emoticons είναι σύμβολα που δημιουργούνται με χρήση σημείων στίξης, αριθμών και χαρακτήρων και χρησιμοποιούνται στην επικοινωνία μέσω ηλεκτρονικών μηνυμάτων για να αναπαραστήσουν ανθρώπινα συναισθήματα και εκφράσεις (List of emoticons, χ.χ.). Για παράδειγμα, το emoticon :-) αντικαθίσταται από την ετικέτα 'happy face smiley' και το emoticon :-(από τις λέξεις 'frown, sad, andry or pouting'. Το βήμα αυτό είναι αναγκαίο να προηγείται της αφαίρεσης των σημείων στίξης, καθώς σε διαφορετική περίπτωση, παρερμηνεύονται τα emoticons και μειώνεται η απόδοση της προεπεξεργασίας.

2. Μετατροπή των emojis στις αντίστοιχες λεξιλογικές ετικέτες

Τα emojis είναι μικρά εικονίδια που συναντώνται κυρίως σε μηνύματα κειμένου με τη μορφή κίτρινων φατσών και παριστάνουν ένα σύνολο αντικειμένων, από φαγητά και ζώα μέχρι άτομα και δραστηριότητες. Η παραμονή τους μέσα στο κείμενο μπορεί να δημιουργήσει προβλήματα στην μετέπειτα επεξεργασία του κειμένου. Γι’ αυτό και με τη βοήθεια της ίδιας βιβλιοθήκης με αυτήν του προηγούμενου βήματος, όλα τα emojis δεδομένου κειμένου μετατρέπονται σε αγγλικές λέξεις με πεζούς χαρακτήρες. Έτσι, για παράδειγμα το emoji 🙂 μεταφράζεται, με το συγκεκριμένο βήμα, ως ':slightly_smiling_face:'.

3. Κατακερμάτιση σε προτάσεις (sentence tokenization)

Η κατακερμάτιση ενός κειμένου λειτουργεί ως μία πολύ βασική διαδικασία στο πεδίο επεξεργασίας φυσικής γλώσσας και εξαρτάται από τη γλώσσα στην οποία είναι γραμμένο το κείμενο (Webster & Kit, 1992). Χρησιμοποιείται για το διαχωρισμό ενός κειμένου σε μικρότερες μονάδες, τις λεγόμενες στην αγγλική βιβλιογραφία ως tokens. Αυτές μπορεί να είναι προτάσεις, λέξεις, όροι, αριθμοί ή σημεία στίξης, με την πιο κλασική επιλογή μονάδων διαχωρισμού να είναι οι λέξεις (Kumhar, et al., 2023). Έτσι, για παράδειγμα, ο ερευνητής μπορεί να αποφασίσει για δεδομένο έγγραφο, αυτό να χωριστεί, σε πρώτη φάση, σε προτάσεις και έπειτα κάθε πρόταση να χωριστεί σε επιμέρους λέξεις.

Στην πειραματική εφαρμογή της παρούσας διπλωματικής, ο διαχωρισμός του κειμένου κάθε κριτικής σε επιμέρους προτάσεις υλοποιήθηκε με τη βοήθεια της σειράς βιβλιοθηκών της Python, NLTK. Η τεχνική της κατακερμάτισης εφαρμόζεται και σε επόμενο στάδιο, απλώς, τη δεύτερη φορά, με διαφορετικές μονάδες διαχωρισμού.

Η NLTK (Natural Language Toolkit) είναι μία πλατφόρμα που χρησιμοποιείται στην κατασκευή προγραμμάτων στη γλώσσα Python ώστε αυτά να χειρίζονται δεδομένα κειμένου. Προσφέρει μία τεράστια ποικιλία από βιβλιοθήκες επεξεργασίας κειμένου, με σκοπό να φέρνει σε πέρας εργασίες όπως η κατάτμηση του κειμένου σε μικρότερες μονάδες, η αναγωγή των όρων στη ρίζα τους μέσω της αφαίρεσης των καταλήξεων (stemming) και η χρήση περιτυλιγμάτων (wrappers) για βιβλιοθήκες NLP βιομηχανικής ισχύος (Natural Language Toolkit, 2022). Συμπληρωματικά, παρέχει διεπαφές με λεξιλογικές βάσεις δεδομένων και μεγάλες συλλογές κειμένου όπως το Open Multilingual Wordnet, το Brown Corpus και το Stopwords Corpus (Bird, Klein, & Loper, 2019).

4. Αφαίρεση των διευθύνσεων του Διαδικτύου

Όπως γίνεται κατανοητό, οι διευθύνσεις του Διαδικτύου (URLs) που παρουσιάζονται εντός του κειμένου δεν είναι χρήσιμες στην εξαγωγή συναισθήματος, γι' αυτό και χρειάζεται να αφαιρεθούν. Μετά από κατάλληλη διερεύνηση στις κριτικές του διαθέσιμου συνόλου δεδομένων, εντοπίστηκαν κάποιες ηλεκτρονικές διευθύνσεις στο τέλος των προτάσεων των κριτικών. Οπότε, με τη βοήθεια της βιβλιοθήκης «re» (re-Regular expression operations, χ.χ.), αναζητείται μέσα στο κείμενο κάθε κριτικής, κάθε μοτίβο κανονικής έκφρασης (regular expression) που αντιπροσωπεύει μία διεύθυνση URL και αφού ανιχνευτεί, αντικαθίσταται από έναν κενό χαρακτήρα.

Αξίζει, σε αυτό το σημείο, να γίνει αναφορά στο ότι οι κανονικές εκφράσεις είναι μία γλώσσα, όπου κάθε ακολουθία συμβόλων και χαρακτήρων εκφράζει διαφορετικό αλφαριθμητικό κείμενο. Ουσιαστικά, χρησιμοποιούνται για την αναζήτηση και την επιστροφή όσων κομματιών από μεγάλες συλλογές κειμένου ταιριάζουν με το δηλωμένο μοτίβο.

5. Μετάφραση του κειμένου στην αγγλική γλώσσα

Σε αυτό το σημείο, λαμβάνεται υπόψιν ότι οι χρήστες του Διαδικτύου μπορεί να μιλούν σε οποιαδήποτε άλλη γλώσσα εκτός από τα αγγλικά και συνεπώς να γράφουν και να εκφράζονται με λέξεις της τοπικής τους γλώσσας γραμμένες με χαρακτήρες του αγγλικού αλφαβήτου. Υπό αυτό το πλαίσιο, προκειμένου όλες οι λέξεις του συνόλου των κριτικών να ανήκουν στο αγγλικό λεξιλόγιο, απαιτείται, σε πρώτο στάδιο, η αναγνώριση των διαφορετικών γλωσσών που ενδεχομένως εμφανίζονται σε δεδομένη πρόταση κάποιας κριτικής. Έπειτα, με τη βοήθεια ενός αντικειμένου της κλάσης Translator που περιέχεται στο πακέτο «googletrans» της Python (googletrans 4.0.0rc1, χ.χ.), γίνεται η μετάφραση όλης της πρότασης στα αγγλικά. Εξαιτίας του περιορισμού του μέγιστου επιτρεπόμενου πλήθους χαρακτήρων προς μετάφραση στους 5.000, η εκάστοτε πρόταση, όποτε είναι μεγαλύτερη, διαχωρίζεται σε κομμάτια (chunks) των

2.000 χαρακτήρων και έπειτα πραγματοποιείται η μετάφραση ανά κομμάτι με χρονοκαθυστέρηση 0,5 δευτερολέπτων ανάμεσα σε διαδοχικές μεταφράσεις.

6. Μετατροπή όλων των χαρακτήρων σε πεζούς

Η τροποποίηση των κεφαλαίων γραμμάτων σε πεζά συνηθίζεται να συμπεριλαμβάνεται στα βήματα προεπεξεργασίας μίας συλλογής κειμένου, αποσκοπώντας στη μη απόδοση διαφορετικής σημασίας σε λέξεις εξαιτίας του ότι αυτές περιλαμβάνουν κεφαλαία γράμματα. Παραδείγματα τέτοιων λέξεων είναι τα κύρια ονόματα, οι λέξεις στην αρχή μίας πρότασης και τα ακρωνύμια. Το συγκεκριμένο βήμα έχει αποδειχτεί ότι βελτιώνει την ακρίβεια της ταξινόμησης κειμένου ανάμεσα σε ένα προκαθορισμένο σύνολο κλάσεων (Uysal & Gunal, 2014), (HaCohen-Kerner, Miller, & Yigal, 2020).

7. Διόρθωση των συγκοπών

Η συγκοπή είναι η συντόμευση μίας λέξης ή φράσης μέσω της παράλειψης μερικών γραμμάτων ή της αντικατάστασής τους με απόστροφο. Στην αγγλική γλώσσα, τα παραλειπόμενα γράμματα συνήθως είναι φωνήεντα.

Ένας βασικός λόγος διόρθωσης των συγκοπών, κατά τον καθαρισμό των δεδομένων κειμένου, είναι η ανάγκη θεώρησης ίδιας σημασίας σε λέξεις όπου η μία αποτελεί συγκοπή της άλλης. Επιπρόσθετα, είναι χρήσιμο να αποφεύγεται η προσθήκη συγκοπών ήδη υπαρχουσών λέξεων στο λεξιλόγιο, καθώς αυξάνεται η υπολογιστική πολυπλοκότητα της επακόλουθης εργασίας.

Το πακέτο της Python «contractions» είναι αυτό που χρησιμοποιείται στην πειραματική εφαρμογή της διπλωματικής έτσι ώστε να επεκτείνει τις συγκοπές, όπου αυτό τις ανιχνεύει. Αρχικά, η ανάλογη εντολή εκτελείται για όλες της λέξεις κάθε πρότασης, και μετά από τα βήματα 8 και 10, ξαναεκτελείται, ώστε να διορθωθούν τυχόν συγκοπές που, είτε παράγονται από τη μετάφραση λέξεων της αργκό, είτε συμπεριλαμβάνονται στη λίστα των λέξεων που υποδηλώνουν άρνηση.

8. Μετάφραση λέξεων αργκό και ακρωνυμίων

Η καθημερινή και ανεπίσημη προφορική γλώσσα, η λεγόμενη αργκό, όπως επίσης και τα ακρωνύμια συναντώνται σε μεγάλο βαθμό σε διαδικτυακά φόρουμ συζητήσεων και ιστολόγια. Ιδιαίτερα συχνά εμφανίζονται και όροι που παράγονται από τη συντόμευση αγγλικών λέξεων και φράσεων, βάσει της φωνητικής των αριθμών και των γραμμάτων που περιέχουν. Αυτοί συγκροτούν το λεγόμενο μικροκείμενο (microtext), ένα σύντομο και ημιδομημένο κείμενο, γραμμένο σε μία πρόταση ή λιγότερο (Gunawan, Saniyah, & Hizriadi, 2019). Η κανονικοποίησή του, δηλαδή η μετατροπή του σε μία πιο βολική και τυπική μορφή, έχει αποδειχτεί σε διάφορες εφαρμογές ότι βελτιώνει την ακρίβεια

της ανίχνευσης της πόλωσης ενός κειμένου (Satapathy, Guerreiro, Chaturvedi, & Cambria, 2017). Η πιο ουσιώδης αιτία είναι το γεγονός ότι πίσω από την εμφάνιση λέξεων αργκό, υποκρύπτονται σημαντικές λέξεις για την εξαγωγή του συναισθηματικού περιεχομένου ενός κειμένου. Ωστόσο, η μετάφρασή τους είναι δύσκολη τόσο για τους ανθρώπους όσο και για τις μηχανές (Mattiello, 2009), με αποτέλεσμα τα τυπικά εργαλεία της επεξεργασίας φυσικής γλώσσας να αδυνατούν να χειριστούν τις συγκεκριμένες λέξεις.

Στην παρούσα εργασία, χρησιμοποιείται ένα ηλεκτρονικό λεξικό επιλεγμένης ιστοσελίδας (Internet & Text Slang Dictionary & Translator, χ.χ.) για τον χειρισμό των λέξεων αργκό και των ακρωνυμίων που μπορεί να υπάρχουν στο κείμενο των κριτικών. Η συλλογή του μικροκειμένου και της μετάφρασής του από το λεξικό πραγματοποιούνται μέσω της διαδικασίας της απόξεσης του ιστού (web scraping) η οποία επεξεργάζεται και αναλύει τα περιεχόμενα της ιστοσελίδας, όντας σε μορφή κώδικα HTML.

Πιο συγκεκριμένα, η απόξεση ιστού είναι μία τεχνική που εξάγει τις ακατέργαστες πληροφορίες και τα δεδομένα που περιλαμβάνει ένας ιστότοπος ώστε να αποθηκευτούν για μελλοντική επεξεργασία από τον χρήστη. Για την πραγμάτωσή της, απαιτείται μία γλώσσα προγραμματισμού εκτελούμενη από την πλευρά του διακομιστή (server-side), όπως είναι η Python. Στην πειραματική εφαρμογή, με τη βοήθεια των πακέτων της Python «requests» και «BeautifulSoup», εξήχθη, αρχικά, το περιεχόμενο 27 ιστοτόπων, καθένας από τους οποίους ερμήνευε όρους διαφορετικού αρχικού γράμματος και έπειτα αυτό οργανώθηκε σε ένα ενιαίο λεξικό μεγέθους 5.828 εγγραφών.

9. Αφαίρεση αριθμών

Η αφαίρεση των χαρακτήρων των λέξεων που είναι σε μορφή αριθμού κρίνεται απαραίτητο βήμα προεπεξεργασίας του διαθέσιμου κειμένου, καθώς δεν εκφράζουν κάποιους είδους πολικότητα.

10. Αφαίρεση ειδικών χαρακτήρων και σημείων στίξης

Όσοι χαρακτήρες δεν ανήκουν στο αγγλικό αλφάβητο ή είναι ο κενός χαρακτήρας απομακρύνονται από το κείμενο της εκάστοτε πρότασης. Ο λόγος είναι ότι δεν φέρουν αξία στην κατανόηση του συναισθηματικού περιεχομένου του κειμένου και η ενδεχόμενη διατήρησή τους θα πρόσθετε θόρυβο στους αλγορίθμους, όποιοι και αν ήταν αυτοί.

11. Αφαίρεση πολλαπλών κενών

Μέσω αυτού του βήματος, εξαλείφεται πιθανή εμφάνιση πολλών συγκεντρωμένων κενών χαρακτήρων ενδιάμεσα σε κάθε πρόταση η οποία πιθανόν οφείλεται σε κάποιο

προηγούμενο στάδιο της διαδικασίας της προεπεξεργασίας ή σε λάθος του χρήστη κατά τη συγγραφή του κειμένου.

12. Κατακερμάτιση σε λέξεις (word tokenization)

Σε αυτό το στάδιο προεπεξεργασίας, κάθε πρόταση δεδομένης κριτικής από το σύνολο κριτικών της πειραματικής εφαρμογής διαχωρίζεται σε λέξεις με τη βοήθεια της σειράς βιβλιοθηκών της Python NLTK και της εργαλειοθήκης Stanford CoreNLP.

Η Stanford CoreNLP (Manning, et al., 2014) είναι μία Java εργαλειοθήκη που υποστηρίζει τις περισσότερες βασικές διεργασίες επεξεργασίας φυσικής γλώσσας (NLP), όπως την επισήμανση μερών του λόγου (POS tagging), την αναγνώριση ονοματικών οντοτήτων (Named Entity Recognition) και τη συντακτική ανάλυση. Η πρόσβαση σε αυτήν είναι δυνατή και μέσω άλλων γλωσσών προγραμματισμού, όπως η Python, η Perl, η Scala και η C. Η βασική διαφορά της με την ανοιχτής πηγής βιβλιοθήκη της Python NLTK εντοπίζεται στο πλήθος των υποστηριζόμενων φυσικών γλωσσών (Sun, Luo, & Chen, 2017).

Ο προτεινόμενος τρόπος χρήσης της βιβλιοθήκης Stanford CoreNLP στην Python είναι μέσω του πακέτου ‘stanza’ της Python. Αυτό παρέχει μία Διεπαφή Προγραμματισμού Εφαρμογών (API) η οποία έχει τη δυνατότητα να αλληλεπιδράει, όποτε ζητείται, με έναν εξυπηρετητή της Stanford CoreNLP (Stanford NLP Group, 2020).

13. Αφαίρεση των συχνά εμφανιζόμενων λέξεων (stop words)

Οι stop words αποτελούν λέξεις συχνά εμφανιζόμενες σε όλα τα έγγραφα δεδομένης συλλογής κειμένου που σχεδόν ποτέ δεν φέρουν κάποια χρήσιμη πληροφορία για τη συναισθηματική πολικότητα (Ghag & Shah, 2015). Ο προσδιορισμός αυτών των λέξεων και η συνακόλουθη εύρεσή τους επηρεάζεται από τη γλώσσα στην οποία είναι γραμμένο το κείμενο που μελετάται.

Ο πρώτος τύπος περιλαμβάνει όσες λέξεις είναι γενικού χαρακτήρα. Σε αυτόν ανήκουν κοινές λέξεις όπως άρθρα (π.χ. a, an), αντωνυμίες (π.χ. she, ours) και προθέσεις (π.χ. to, of) οι οποίες δεν έχουν σημαντική συνεισφορά σε εφαρμογές επεξεργασίας φυσικής γλώσσας, γι’ αυτό και απομακρύνονται. Η σειρά βιβλιοθηκών της Python NLTK προσφέρει μία έτοιμη λίστα από stop words στην αγγλική γλώσσα, βοηθώντας, με αυτόν τον τρόπο, στην αυτόματη εύρεση και αφαίρεση αυτού του είδους λέξεων από τα δεδομένα κειμένου.

Υπάρχουν και οι συχνά εμφανιζόμενες λέξεις που αλλάζουν κάθε φορά ανάλογα με τον τομέα στον οποίο αναφέρεται η συλλογή του κειμένου που αποτελεί το αντικείμενο έρευνας (Maree, Eleyat, Rabayah, & Belkhatir, 2023). Μία μέθοδος που συνηθίζεται να εφαρμόζεται με σκοπό την εύρεση τέτοιων λέξεων είναι το tf-idf μοντέλο, όπου βάσει του βάρους κάθε λέξης (υπολογισμένου σύμφωνα με την tf-idf στάθμιση),

αφαιρούνται όσες λέξεις έχουν τιμή μικρότερη ή ίση από ένα προεπιλεγμένο κατώφλι (Maree, Kmail, & Belkhatir, Analysis and shortcomings of e-recruitment systems: Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage, 2019). Στην πρακτική εφαρμογή, εκλαμβάνονται ως stop words μόνο οι συχνές λέξεις γενικού χαρακτήρα ανεξαρτήτως του περιεχομένου του κειμένου.

14. Διατήρηση των λέξεων που εκφράζουν άρνηση

Η παρουσία λέξεων σε μία πρόταση που εκφράζουν άρνηση παίζει αναμφίβολα σημαντικό ρόλο καθώς επηρεάζει τον συναισθηματικό προσανατολισμό μίας σειράς άλλων λέξεων της ίδιας πρότασης. Συνεπώς, απαιτείται να παραμείνουν στην πρόταση μετά την ολοκλήρωση της διαδικασίας της προεπεξεργασίας. Για να αποφευχθεί η αφαίρεσή τους εξαιτίας της ενδεχόμενης συμπερίληψής τους στις συχνά εμφανιζόμενες λέξεις, πραγματοποιείται προηγούμενη καταγραφή τους σε μορφή λίστας και εντοπισμός τους στην εκάστοτε πρόταση με τη βοήθεια κατάλληλων εντολών. Το σύνολο των λέξεων της αγγλικής γλώσσας οι οποίες θεωρείται ότι δηλώνουν κάποιου είδους άρνηση είναι το ακόλουθο (Farooq, Mansoor, Nongaillard, Ouzrout, & Qadir, 2017):

$$A = \left\{ \begin{array}{c} \text{no, not, rather, couldn't, wasn't, wasn, didn't, didn, wouldn't, wouldn, shouldn't,} \\ \text{shouldn, weren't, weren, don't, doesn't, doesn, haven't, haven, hasn't, hasn, won't,} \\ \text{wont, hadn't, hadn, never, none, nobody, nothing, neither, nor, nowhere, isn't, isn,} \\ \text{can't, cannot, mustn't, mustn, mightn't, mightn, shan't, shan, without, needn't, needn,} \\ \text{hardly, less, little, rarely, scarcely, seldom, λέξεις με πρόθεμα: de−, dis−, il−, im−,} \\ \text{in−, ir−, mis−, non−, un − και κατάληξη − less.} \end{array} \right\}$$

15. Αφαίρεση των σύντομων λέξεων

Όσες λέξεις αποτελούνται από λιγότερο από 3 χαρακτήρες αφαιρούνται, αφού λόγω του πολύ μικρού μήκους τους, αδυνατούν να συνεισφέρουν στην ανάλυση συναισθήματος. Επιπλέον, λέξεις τόσο μικρού μήκους είναι αρκετές φορές τυπογραφικά λάθη των χρηστών, με την απομάκρυνσή τους σε αυτή την περίπτωση να μετατρέπεται σε αναγκαία.

Εκτός από τα παραπάνω, στο πλαίσιο της προεπεξεργασίας του κειμένου κάθε κριτικής εφαρμόζονται και μία σειρά από διαδικασίες, ικανές να εξάγουν χρήσιμη πληροφορία σχετικά με τη γραμματική και τη συντακτική δομή κάθε πρότασης. Αυτές αναλύονται ακολούθως:

- Αναγνώριση λέξεων ως μέρη του λόγου

Η επισήμανση των λέξεων μίας πρότασης ως μέρη του λόγου, γνωστή στην αγγλική βιβλιογραφία ως Part-Of-Speech (POS) tagging, αποσκοπεί στην αναγνώριση της λεξιλογικής κατηγορίας στην οποία ανήκει κάθε λέξη από σημασιολογικής, γραμματικής και συντακτικής πλευράς. Οι βάσεις της λεξιλογικής κατηγοριοποίησης είναι τα μέρη του λόγου: ουσιαστικό, ρήμα και επίθετο, τα οποία συμπληρώνονται από επιπρόσθετες κατηγορίες δευτερεύουσας σημασίας και υποκατηγορίες τόσο των πρωταρχικών όσο και των δευτερευουσών κατηγοριών (Güngör, 2010).

Το σύνολο που επιλέγεται για τη συγκεκριμένη διεργασία χρειάζεται να περιλαμβάνει ικανοποιητικό αριθμό ετικετών με τις διαφορές ανάμεσά τους να είναι σαφείς και ευδιάκριτες. Ένα από τα πιο συχνά χρησιμοποιούμενα και πλήρη σύνολα είναι το UPenn Treebank II το οποίο περιέχει λίστα 36 ετικετών που παρατίθεται ακολούθως.

Πίνακας 2. Λίστα ετικετών UPenn Treebank II.

Ετικέτα	Περιγραφή
CC	Coordinating conjunction / Συντονιστικός σύνδεσμος
CD	Cardinal number / Απόλυτος αριθμός
DT	Determiner / Προσδιοριστής
EX	Existential *there* / Υπαρξιακό *there*
FW	Foreign word / Ξένη λέξη
IN	Preposition or subordinating conjunction / Πρόθεση ή δευτερεύων σύνδεσμος
JJ	Adjective / Επίθετο
JJR	Adjective, comparative / Επίθετο σε συγκριτικό βαθμό
JJS	Adjective, superlative / Επίθετο σε υπερθετικό βαθμό
LS	List item marker / Δείκτης αντικειμένου λίστας
MD	Modal / Τροπικό βοηθητικό ρήμα
NN	Noun, singular or mass / Ουσιαστικό σε ενικό αριθμό ή περιληπτικό ουσιαστικό
NNS	Noun, plural / Ουσιαστικό σε πληθυντικό αριθμό
NNP	Proper noun, singular / Κύριο όνομα σε ενικό αριθμό
NNPS	Proper noun, plural / Κύριο όνομα σε πληθυντικό αριθμό
PDT	Predeterminer / Προπροσδιοριστής
POS	Possessive ending / Κτητική κατάληξη
PRP	Personal pronoun / Προσωπική αντωνυμία
PRP$	Possessive pronoun / Κτητική αντωνυμία
RB	Adverb / Επίρρημα
RBR	Adverb, comparative / Επίρρημα σε συγκριτικό βαθμό
RBS	Adverb, superlative / Επίρρημα σε υπερθετικό βαθμό
RP	Particle / Μόριο
SYM	Symbol / Σύμβολο
TO	*To* / Λέξη *to*
UH	Interjection / Επιφώνημα
VB	Verb, base form / Ρήμα σε βασική μορφή
VBD	Verb, past tense / Ρήμα σε παρελθοντικό χρόνο
VBG	Verb, gerund or present participle / Ρήμα ως γερούνδιο ή μετοχή ενεστώτα
VBN	Verb, past participle / Ρήμα ως μετοχή αορίστου
VBP	Verb, non-3rd person singular present / Ρήμα σε μη γ’ ενικό πρόσωπο ενεστώτα
VBZ	Verb, 3rd person singular present / Ρήμα σε γ’ ενικό πρόσωπο ενεστώτα

WDT	Wh-determiner / Wh-προσδιοριστής
WP	Wh-pronoun / Wh-αντωνυμία
WP$	Possessive wh-pronoun / Κτητική wh-αντωνυμία
WRB	Wh-adverb / Wh-επίρρημα

- Λημματοποίηση

Η λημματοποίηση είναι μία διαδικασία μορφολογικής ανάλυσης κατά την οποία οι λέξεις μετατρέπονται στη βασική τους μορφή, μειώνοντας με αυτόν τον τρόπο, το πλήθος των διαφορετικών μονάδων κατάτμησης-λέξεων. Αυτή επιτυγχάνεται μέσω της περικοπής των κλιτικών καταλήξεων και προθεμάτων, ώστε οι υπάρχουσες παράγωγες μορφές κάθε λέξης να αναχθούν και να αναλυθούν τελικά ως μία μοναδική λέξη που είναι το λήμμα.

Ο καθορισμός του λήμματος της εκάστοτε λέξης προαπαιτεί τη σωστή αναγνώριση του μέρους του λόγου στο οποίο ανήκει η συγκεκριμένη λέξη. Έτσι, η αγγλική λέξη ‘saw’ επιστρέφεται, μετά από τη λημματοποίηση, ως ‘see’ ή ‘saw’ ανάλογα με το αν αποφασίζεται να χρησιμοποιηθεί ως ρήμα ή ως ουσιαστικό αντίστοιχα (Ghosh, et al., 2023). Η διαδικασία της λημματοποίησης είναι σημαντικό να λαμβάνει χώρα πριν από τον υπολογισμό των απόλυτων συχνοτήτων εμφάνισης των λέξεων σε δεδομένη συλλογή κειμένου.

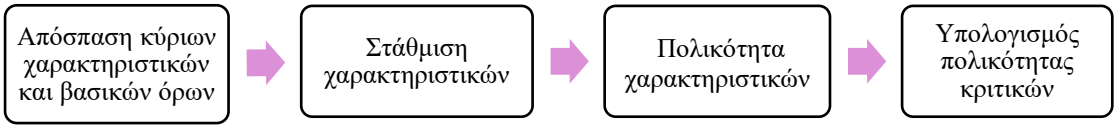
- Ανάλυση συντακτικών εξαρτήσεων

Η διερεύνηση των σχέσεων και κατ’ επέκταση ο καθορισμός των συντακτικών εξαρτήσεων μεταξύ των όρων μίας πρότασης, στηρίζεται στην υπόθεση ότι όλες οι γλωσσικές μονάδες συνδέονται με κάποιο τρόπο μεταξύ τους. Εφαρμόζεται με σκοπό να αξιοποιηθεί η πληροφορία της συντακτικής δομής των προτάσεων η οποία εξυπηρετεί σε σημαντικό βαθμό την ανάλυση των προτάσεων. Η διαδικασία αυτή συναντάται στη βιβλιογραφία ως dependency parsing και παράγει ως αποτέλεσμα, τριάδες, σε καθεμία από τις οποίες αναφέρεται το είδος της σχέσης, η λέξη που αποτελεί την κεφαλή (head) ή αλλιώς τον κυβερνήτη (governor) της σχέσης και η εξαρτώμενη λέξη (dependent).

7.4. ΒΗΜΑΤΑ ΜΕΘΟΔΟΛΟΓΙΑΣ

Η διαδικασία που ακολουθείται, στηρίζεται κατά βάση σε αυτή των Al-Ghuribi, Mohd Noah & Tiun (2020) περιλαμβάνοντας, εν συντομία, τα ακόλουθα βήματα. Σε πρώτο στάδιο, πραγματοποιείται μέσω της αναλυθείσας προεπεξεργασίας, ο μετασχηματισμός του κειμένου κάθε κριτικής, έτσι ώστε να μετατραπεί σε μία κανονική μορφή, κατάλληλη για περαιτέρω ανάλυση. Ύστερα, ακολουθεί η εξαγωγή των χαρακτηριστικών, χρησιμοποιώντας μία υβριδική μέθοδο που συνδυάζει τη

συχνότητα των λέξεων και τη συντακτική τους θέση μέσα σε κάθε κριτική, αφαιρώντας κάποιες λέξεις βάσει της σημασιολογικής τους ομοιότητας. Σε επόμενο βήμα, αποκτάται η στάθμιση των λέξεων με τρεις διαφορετικούς τρόπους και ακολουθεί η εύρεση των ζευγών 'χαρακτηριστικό-λέξη εκφραζόμενου συναισθήματος'. Στη συνέχεια, πραγματοποιείται η βαθμολόγηση της πολικότητας των λέξεων με τη βοήθεια των λεξικών, στην πλειονότητά τους, πλήρως προσαρμοσμένων στο σημασιολογικό πεδίο των κριτικών. Τέλος, υπολογίζεται μία συνολική τιμή για τον συναισθηματικό προσανατολισμό κάθε κριτικής, οπότε, ύστερα, βάσει αυτής, υλοποιείται και η ζητούμενη εργασία ταξινόμησής της.



Εικόνα 6. Βασικό διάγραμμα ροής της εφαρμοσμένης μεθοδολογίας.

7.4.1. Απόσπαση χαρακτηριστικών

Σε αυτό το στάδιο, η μεθοδολογία αποσκοπεί στην παράλληλη συλλογή, σε δύο διαφορετικά λεξικά, των ουσιαστικών και όσων ουσιαστικών προσδιορίζονται από επίθετα. Σχετική έρευνα έχει δείξει ότι το 60%-70% των χαρακτηριστικών είναι ουσιαστικά ή φράσεις ουσιαστικών που έχουν υψηλή συχνότητα εμφάνισης στα μελετώμενα δεδομένα κειμένου (Liu B. , Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2011; Moghaddam & Ester, 2010). Σε κάθε περίπτωση δημιουργούνται 26 συστάδες (ή ομάδες), καθεμία από τις οποίες περιέχει όλες τις ζητούμενες λέξεις που ξεκινούν με πρώτο γράμμα, ένα από τα 26 γράμματα του αγγλικού αλφαβήτου A έως Z (Thabit & Al-Ghuribi, 2013). Κρίνεται απαραίτητο, προηγουμένως, να έχει βρεθεί το πρώτο γράμμα κάθε ουσιαστικού, να μετατραπεί σε κεφαλαίο και να αντιστοιχηθεί με τον αριθμό κατά το πρότυπο ASCII. Δεδομένου ότι οι ομάδες αριθμούν από το 0 έως το 25, χρειάζεται η αφαίρεση του αριθμού 65 από τους δεκαδικούς αριθμούς ASCII, με την ακριβή αντιστοίχιση να φαίνεται στην Εικόνα 6.

A	B	C	D	E	F	G	H	I	J	K	L	M
↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕
65	66	67	68	69	70	71	72	73	74	75	76	77
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕
78	79	80	81	82	83	84	85	86	87	88	89	90

Εικόνα 7. ASCII κωδικοποίηση των κεφαλαίων αγγλικών χαρακτήρων.

Με αυτόν τον τρόπο, η επεξεργασία που αφορά κάθε ουσιαστικό περιορίζεται σε μία μόνο συστάδα και η πολυπλοκότητα εκτέλεσης του αλγορίθμου ελαττώνεται σημαντικά. Κατά τη φάση αυτή, κάθε ουσιαστικό συνοδεύεται και από τη συχνότητα εμφάνισής του στο κείμενο κάθε κριτικής. Ο αλγόριθμος έχει κατασκευαστεί έτσι ώστε να ελέγχεται αν η εκάστοτε λέξη υπάρχει ήδη στο λεξικό (αφού προηγουμένως έχει αναγνωριστεί ότι πρόκειται για ουσιαστικό), ώστε στη συνέχεια να προστίθεται στην αντίστοιχη λίστα συχνοτήτων ο αριθμός 1. Το παραγόμενο λεξικό είναι το ‘Noun_Block_Dictionary’.

Με παρόμοιο τρόπο, παράγεται και το δεύτερο λεξικό ‘Noun_Adj_Block_Dictionary’ με τα ουσιαστικά των οποίων η σημασία προσδιορίζεται από κάποια επιθετική φράση (De Marneffe & D. Manning, 2008). Η αναζήτησή τους υλοποιείται μέσω του αναλυτή εξαρτήσεων του StanfordCoreNLP και πιο συγκεκριμένα της συντακτικής σχέσης ‘amod’. Στη συνέχεια, διερευνάται αν τα συνοδευόμενα επίθετα βρίσκονται σε συγκριτικό/υπερθετικό ή θετικό βαθμό βάσει των ετικετών των μερών του λόγου στα οποία ανήκουν (‘JJR’ για συγκριτικό βαθμό και ‘JJS’ για υπερθετικό βαθμό). Έτσι, σε αυτό το λεξικό, οι λίστες συχνοτήτων είναι δύο σε πλήθος: ‘Adj_Frequency’ και ‘Comp_Sup_Frequency’, περιλαμβάνοντας το πλήθος των ουσιαστικών που συνοδεύονται από επίθετα οποιουδήποτε βαθμού (στην πρώτη λίστα συχνοτήτων) και από επίθετα μόνο συγκριτικού ή υπερθετικού βαθμού (στη δεύτερη λίστα).

Έπειτα, τα δύο λεξικά συμπτύσσονται σε ένα ενιαίο λεξικό ‘Main_Dictionary’, το οποίο εμπεριέχει όλα τα ουσιαστικά του λεξικού ‘Noun_Block_Dictionary’ σε μορφή συστάδων. Η στήλη με τη συχνότητα εμφάνισής τους ‘Total_Frequency’ παραμένει αυτούσια με την αντίστοιχη του πρώτου λεξικού (‘Noun_Block_Dictionary’). Στον αντίποδα, το περιεχόμενο των δύο άλλων στηλών καθορίζεται από το εάν το εκάστοτε ουσιαστικό του πρώτου λεξικού ανήκει ή όχι και στο δεύτερο λεξικό. Εάν δεν ανήκει, τότε οι τιμές των στηλών ‘Adj_Frequency’ και ‘Comp_Sup_Frequency’ προσδιορίζονται να είναι μηδενικές, ενώ εάν ανήκει, λαμβάνονται οι τιμές των αντίστοιχων στηλών του δεύτερου λεξικού (‘Noun_Adj_Block_Dictionary’). Τελικά, το λεξικό μετατρέπεται σε πλαίσιο δεδομένων (‘df_Main_Dictionary’) με 3.003 λέξεις-γραμμές και 3 στήλες.

Ύστερα λαμβάνονται τρεις λίστες ίδιου πλήθους ουσιαστικών, σε καθεμία από τις οποίες περιλαμβάνονται τα ουσιαστικά εκείνα που παρουσιάζουν τις 200 μεγαλύτερες συχνότητες σε καθεμία από τις στήλες συχνοτήτων αντίστοιχα. Η τιμή 200 αποφασίστηκε μετά από διάφορες δοκιμές και τέθηκε λαμβάνοντας υπόψιν τον χαμηλό αριθμό διαθέσιμων κριτικών συγκριτικά με αυτόν της έρευνας των Al-Ghuribi, Mohd Noah, & Tiun (2020). Με τη συγχώνευσή των λιστών, δημιουργείται μία καινούρια λίστα με 150 ουσιαστικά η οποία παρατίθεται στον παρακάτω πίνακα:

Πίνακας 3. Λίστα ουσιαστικών με τις πιο υψηλές συχνότητες τριών τύπων.

harbour	person	trip	part	hotel	spot	cruise	week	lane
restaurant	bar	way	stroll	dinner	couple	parking	setting	mountain
place	boat	wall	light	photo	experience	venice	turtle	nothing
chania	tourist	taverna	front	thing	afternoon	charm	tavern	quality
harbor	cafe	port	end	year	something	island	holiday	character
shop	night	coffee	house	tour	sight	ship	Greek	promenade
view	building	atmosphere	architecture	picture	location	beer	number	path
town	food	plenty	shopping	life	bus	site	staff	style
lot	water	city	meal	waterfront	season	face	crowd	alleyway
day	sunset	history	morning	local	choice	store	value	culture
lighthouse	evening	museum	price	lunch	glass	art	scenery	point
time	drink	side	sun	fish	opportunity	variety	wave	road
street	visit	bit	souvenir	ride	beach	attraction	wine	yacht
walk	sea	horse	carriage	alley	music	fortress	weather	eaterie
area	crete	hour	mosque	market	one	trap	child	fort
selection	gem	boutique	seafood	option	right	exhibition	church	fun
walkway	walking	service	stay	goods	scene			

Τα φαινόμενα εμφάνισης τόσο λέξεων που, ενώ δεν σχετίζονται με το σημασιολογικό πεδίο του κειμένου, είναι συχνές οπότε και εξάγονται από την προηγούμενη διαδικασία, όσο και λέξεων που παρ' όλο που είναι συνώνυμες των χαρακτηριστικών, δεν εμφανίζονται τόσο συχνά στο μελετώμενο κείμενο και κατά συνέπεια απορρίπτονται, λύνονται ως εξής.

Για το πρώτο ζήτημα, έχει προταθεί η εξέταση της τιμής ομοιότητας κάθε ουσιαστικού που εξήχθη από την προηγούμενη διαδικασία με μία ευρύτερη λέξη που αντιπροσωπεύει τον τομέα ενδιαφέροντος. Αυτή η λέξη αποφασίστηκε να είναι η αγγλική λέξη 'tourism', μιας και το παρόν σύνολο δεδομένων αντλήθηκε από το Tripadvisor και συνίσταται από κριτικές επισκεπτών για το λιμάνι των Χανίων. Η ομοιότητα υπολογίζεται με τη βοήθεια ενός προεκπαιδευμένου με αρνητική δειγματοληψία Word2vec μοντέλου σε μέρος του συνόλου δεδομένων Google News που περιλαμβάνει 3 εκατομμύρια λέξεις και φράσεις. Οι κατανεμημένες αναπαραστάσεις των λέξεων έχουν διάσταση ίση με 300, δηλαδή κάθε διαφορετική λέξη αντιπροσωπεύεται από ένα μοναδικό διάνυσμα 300 συνιστωσών.

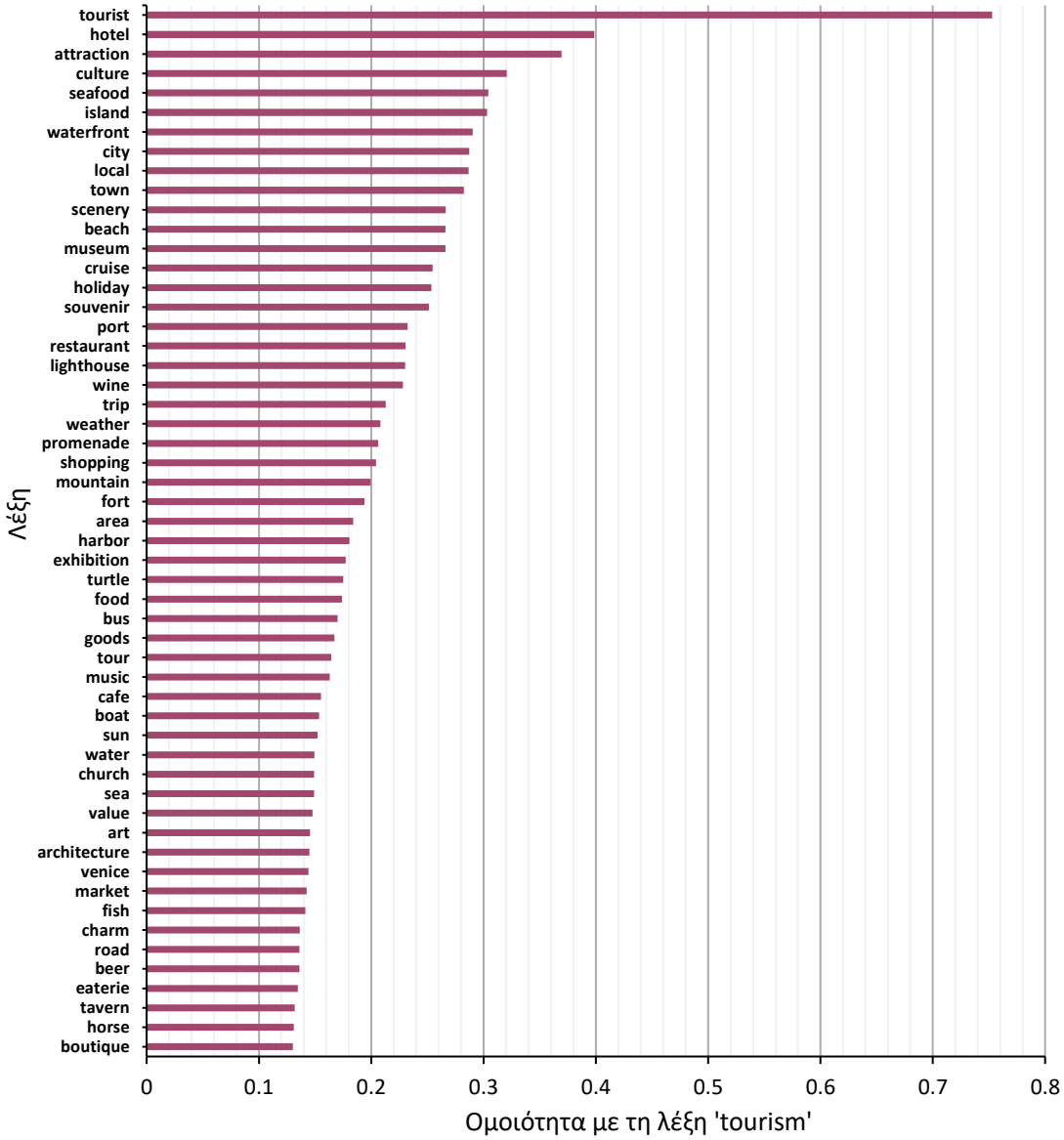
Ο αλγόριθμος κατασκευάζεται με τέτοιο τρόπο έτσι ώστε για τις λέξεις του πίνακα 3 που δεν ανήκουν στο σύνολο των 3 εκατομμυρίων λέξεων, να αφήνεται κενή (nan) η τιμή ομοιότητας. Τα αποτελέσματα φανερώνουν ότι αποκτώνται ελλείπουσες τιμές ομοιότητας για τις λέξεις 'harbour' και 'chania'. Όμως, εύκολα παρατηρείται ότι παρ' όλο που η λέξη 'harbor' έχει την ίδια ακριβώς σημασία με την προηγούμενη λέξη 'harbour', η τιμή ομοιότητας της'harbour' με τη λέξη 'tourism' είναι κενή, ενώ της 'harbor' με τη λέξη 'tourism' υπολογίζεται ίση με 0,18. Η μόνη διαφορά μεταξύ των λέξεων είναι ότι η λέξη 'harbor' είναι γραμμένη σε Αμερικάνικα Αγγλικά ενώ η. 'harbour' σε Βρετανικά Αγγλικά.

Προκειμένου να θεωρηθεί ότι δύο λέξεις είναι σχετικές μεταξύ τους, αποφασίζεται ότι πρέπει να έχουν ομοιότητα με τον τομέα κατ' ελάχιστον ίση με 0,13. Η συγκεκριμένη τιμή καθορίζεται ύστερα από διερεύνηση μεταξύ των τιμών του διαστήματος [0,10, 0,20] και την εύρεση του πλήθους και της σημασίας των λέξεων που λαμβάνονται ως

σχετικές με τον τομέα των κριτικών. Τα ουσιαστικά που τελικά παραμένουν, αποτελούν τα κύρια χαρακτηριστικά (main aspects) των κριτικών και είναι 54 σε πλήθος.

Πίνακας 4. Τιμές ομοιότητας των κύριων χαρακτηριστικών με τη λέξη 'tourism'.

Λέξη	Ομοιότητα με τον τομέα	Λέξη	Ομοιότητα με τον τομέα
restaurant	0.230312	beach	0.266164
harbor	0.180686	music	0.163135
town	0.282573	cruise	0.254764
lighthouse	0.230275	venice	0.144026
area	0.183970	charm	0.136446
boat	0.153453	island	0.303155
tourist	0.752909	beer	0.135906
cafe	0.155119	art	0.145540
food	0.173800	attraction	0.369380
water	0.149468	turtle	0.174938
sea	0.149054	tavern	0.131885
trip	0.212945	holiday	0.253480
port	0.232373	value	0.147603
city	0.287286	scenery	0.266417
museum	0.266120	wine	0.228228
horse	0.131155	weather	0.208195
architecture	0.144954	mountain	0.199408
shopping	0.204351	promenade	0.206275
sun	0.152291	culture	0.320693
souvenir	0.251389	road	0.136162
hotel	0.398608	eaterie	0.134772
tour	0.164455	fort	0.194014
waterfront	0.290188	goods	0.167321
local	0.286684	church	0.149152
fish	0.141377	exhibition	0.177111
market	0.142336	seafood	0.304165
bus	0.169888	boutique	0.130196



Σχήμα 5. Ραβδόγραμμα τιμών ομοιότητας κάθε κύριου χαρακτηριστικού με τον τομέα των κριτικών.

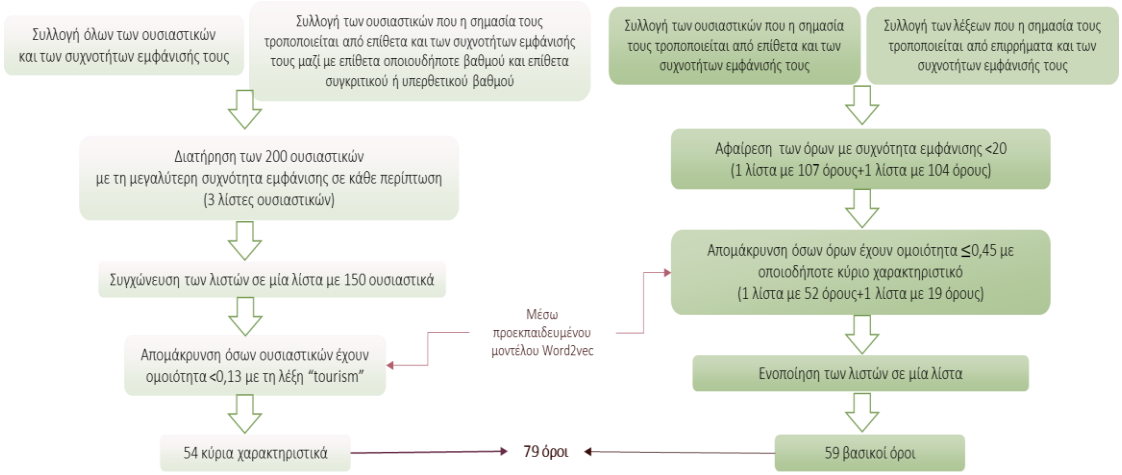
Για την εξάλειψη του ζητήματος αδυναμίας του αλγορίθμου αιχμαλώτισης σπάνιων, όμως σημαντικών λέξεων για την ταξινόμηση, αποσπώνται, αρχικά, από το κείμενο όλων των κριτικών τα ουσιαστικά που προσδιορίζονται από επίθετα και οι λέξεις που προσδιορίζονται από επιρρήματα, με τη βοήθεια των συντακτικών σχέσεων ‘amod’ και ‘advmod’ του StanfordCoreNLP αντίστοιχα. Από τα σύνολα των ουσιαστικών και των λέξεων που εξάγονται, αφαιρούνται όσοι όροι έχουν συχνότητα μικρότερη από 20. Η τιμή αυτή προέκυψε ύστερα από τη μελέτη των λέξεων που περιέχονταν σε κάθε παραγόμενο σύνολο. Έτσι, τελικά το πρώτο σύνολο με τα ουσιαστικά που τροποποιούνται από επίθετα αποτελείται από 107 όρους, ενώ το δεύτερο με τις λέξεις

των οποίων η σημασία μεταβάλλεται από επιρρήματα περιέχει 104 όρους. Στη συνέχεια, από τις εναπομείνασες λέξεις, διατηρούνται μόνο αυτές που έχουν τιμή ομοιότητας με οποιοδήποτε κύριο χαρακτηριστικό, μεγαλύτερη του 0,45. Τελικά, προκύπτουν 52 όροι από το πρώτο σύνολο και 19 από το δεύτερο που μαζί αποτελούν εκείνους τους βασικούς όρους που περιγράφουν και συμπληρώνουν τα κύρια χαρακτηριστικά (γνωστούς στη βιβλιογραφία ως core terms). Ενώνοντας τους σε ένα ενιαίο σύνολο και απομακρύνοντας τις διπλότυπες λέξεις, συγκροτούν το σύνολο του Πίνακα 5 με 59 λέξεις. Τελικά, οι μοναδικές λέξεις που αποτελούν τα κύρια χαρακτηριστικά και τους βασικούς όρους είναι 79 σε πλήθος.

Πίνακας 5. Βασικοί όροι στο μελετώμενο κείμενο κριτικών.

architecture	area	attraction	bar	beach	beautiful
beer	boat	cafe	charming	city	coffee
cuisine	culture	drink	eat	fish	food
fortress	goods	harbor	hotel	island	lane
lighthouse	location	market	meal	mosque	museum
music	picturesque	port	price	restaurant	road
scenery	sea	ship	shop	shopping	souvenir
store	street	stroll	taverna	tour	tourist
touristic	touristy	town	trip	turtle	value
vibe	visit	water	wine	worth	

Όλη η προαναφερθείσα διαδικασία απόκτησης των κύριων χαρακτηριστικών και των βασικών όρων παρατίθεται σχηματικά στο ακόλουθο διάγραμμα ροής.



Εικόνα 8. Διάγραμμα ροής για την απόσπαση των κύριων χαρακτηριστικών και των βασικών όρων.

Μετά τη λήψη των κύριων χαρακτηριστικών και των βασικών όρων, επαναλαμβάνεται όλη η προαναφερθείσα μεθοδολογία, αυτή τη φορά, για το κείμενο στους τίτλους των καταχωρίσεων του αρχικού συνόλου δεδομένων. Ο τίτλος σε μία εγγραφή του πίνακα

δεδομένων αντανακλά μία σύντομη και περιληπτική έκφραση γνώμης γύρω από ένα θέμα, στη συγκεκριμένη περίπτωση αναφορικά με την επίσκεψη στο παλιό λιμάνι των Χανίων. Οπότε, καθίσταται ζωτικής σημασίας να ελεγχθεί αν οι όροι που προκύπτουν από την ανάλυση των τίτλων θα διαφέρουν ή όχι σε σχέση με αυτούς του πίνακα 5. Η εκτέλεση του κώδικα επιβεβαιώνει την πληρότητα των ήδη υπαρχουσών λιστών με τα κύρια χαρακτηριστικά και τους βασικούς όρους.

7.4.2. Στάθμιση των χαρακτηριστικών

Σε αυτό το βήμα, οι όροι που απαρτίζουν το βασικό κείμενο των κριτικών σταθμίζονται με τρεις διαφορετικές μεθόδους και είναι αυτές που περιγράφονται αναλυτικά στην ενότητα 4.5: η κλασική τεχνική tf-idf και οι δύο εναλλακτικές προσεγγίσεις υπολογισμού που προτείνονται από τους Zhu, Wang, & Zou και Nguyen Thi Ngoc, Nguyen Thi Thu, & Nguyen.

Όσον αφορά τις εντολές του κώδικα που τρέχουν για τον συγκεκριμένο σκοπό, εισάγεται σε ένα καινούριο σημειωματάριο του Google Colab (‘weighting.ipynb’), το αρχικό διαθέσιμο σύνολο δεδομένων με τροποποιημένο το κείμενο που αφορά τα σχόλια των χρηστών. Η αλλαγή γίνεται ώστε να ξεπεραστεί η αδυναμία εισαγωγής της λέξης ‘harbour’ στο σύνολο των κύριων χαρακτηριστικών εξαιτίας της διαφορετικής γραφής. Υπό το πλαίσιο αυτό, οι λέξεις ‘harbour’ και ‘Harbour’ αντικαθίστανται από τις λέξεις ‘harbor’ και ‘Harbor’ αντίστοιχα που έχουν ακριβώς την ίδια σημασία. Πραγματοποιείται η προεπεξεργασία του κειμένου όπως και πριν και στη συνέχεια, βάσει των τύπων των υποενοτήτων 4.5.1. και 4.5.2. υπολογίζεται ένα μέτρο σημαντικότητας για κάθε όρο-λήμμα σε καθεμία καταχώριση κριτικής. Κάθε όρος συμπληρώνεται με την ετικέτα του μέρους του λόγου στο οποίο ανήκει. Η ανάθεση των ετικετών και η εύρεση των λημμάτων των όρων κάθε πρότασης και κατ’ επέκταση εγγράφου είναι δυνατή μέσω των υπομνηματιστών του Stanford CoreNLP: ‘pos’ και ‘lemma’.

Στον τύπο (7), το πλήθος Cl των κλάσεων ισούται με 5, όσες είναι και οι δυνατές βαθμολογίες που έχουν βάλει οι χρήστες στο σύνολο των κριτικών τους. Η προσέγγιση των Zhu, Wang, & Zou παρουσιάζει μία βασική διαφορά σε σχέση με τις υπόλοιπες υπό διερεύνηση μεθόδους, υπό την έννοια ότι υπολογίζει τα βάρη μόνο όσων λέξεων αποτελούν κύρια χαρακτηριστικά ή βασικοί όροι, σε οποιοδήποτε μέρος του λόγου και αν ανήκουν. Τα βάρη των υπόλοιπων λέξεων λαμβάνουν ελλείπουσες τιμές.

7.4.3. Βαθμολόγηση της πολικότητας των χαρακτηριστικών

Σε πρώτη φάση, οι κριτικές διαχωρίζονται σε θετικές και αρνητικές, αγνοώντας τις ουδέτερες κριτικές (με βαθμολογία ίση με 3). Ως θετικές εκλαμβάνονται όσες έχουν

βαθμολογία 4 και 5, ενώ ως αρνητικές αυτές που έδωσαν βαθμολογία ίση με 1 ή 2. Το 20% του συνόλου των εγγράφων-κριτικών, στρωματοποιημένο ως προς το ποσοστό των κριτικών ανά βαθμολογία, ανατίθεται ως το σύνολο των δειγμάτων ελέγχου των αλγορίθμων ταξινόμησης. Το υπόλοιπο 80% συγκροτεί το σύνολο εκπαίδευσης, απ' όπου λαμβάνεται το βασικό κείμενο με τη γνώμη των χρηστών. Ύστερα, αποκτάται η κανονική του μορφή και ο αλγόριθμος εντοπίζει τη γενική κατηγορία μέρους του λόγου στην οποίο ανήκει κάθε όρος. Οι κατηγορίες αυτές είναι πέντε σε πλήθος: 'ουσιαστικό', 'επίθετο', 'επίρρημα', 'ρήμα' και οτιδήποτε άλλο. Σε περίπτωση που ένας όρος θεωρηθεί από τον υπομνηματιστή 'pos' ότι ανήκει σε κάποια από τις τέσσερις πρώτες κατηγορίες ή ανήκει στις αποδεκτές λέξεις άρνησης, τότε και μόνο τότε, συμπεριλαμβάνεται στη μελέτη για την εύρεση της τιμής συναισθηματικού προσανατολισμού.

Για την επίτευξη ανάθεσης μίας τιμής συναισθηματικού προσανατολισμού σε κάθε όρο που εξήχθη κατά το βήμα απόσπασης χαρακτηριστικών, είναι απαραίτητο να βρεθούν πρώτα οι λέξεις συναισθήματος που τον χαρακτηρίζουν. Χωρίς περιορισμό στο είδος των λέξεων συναισθήματος και στο πλήθος τους για κάθε κύριο χαρακτηριστικό ή βασικό όρο, η μεθοδολογία χρησιμοποιεί κατά βάση τον αναλυτή συντακτικών εξαρτήσεων και την αναγνώριση ονοματικών οντοτήτων (Stanford NLP Group, Named Entity Recognition, n.d.) του Stanford CoreNLP.

Οι ονοματικές οντότητες που εντοπίζονται είναι οι ονοματικές φράσεις και όροι που ανήκουν στις κατηγορίες: πρόσωπα ('PERSON'), περιοχές ('LOCATION'), τίτλους ('TITLE') ή σε άλλες ανάμεικτες κατηγορίες ('MISC'). Οι συντακτικές εξαρτήσεις μεταξύ των λέξεων που χρειάζεται να εντοπιστούν, δίνονται και ερμηνεύονται ακολούθως (De Marneffe & D. Manning, 2008; Stanford NLP Group, Class UniversalEnglishGrammaticalRelations, n.d.):

- amod: Μία επιθετική φράση που τροποποιεί τη σημασία μίας ονοματικής φράσης.
- advmod: Ένα επίρρημα ή μία φράση που έχει ως κεφαλή ένα επίρρημα που τροποποιεί τη σημασία μίας λέξης.
- nmod: Μία ονοματική φράση που εξαρτάται από άλλο ουσιαστικό ή ονοματική φράση και αντιστοιχεί σε επιθετικό προσδιορισμό ή συμπλήρωμα σε γενική κτητική.
- agent: Το συμπλήρωμα ενός ρήματος σε παθητική φωνή το οποίο εισάγεται από την πρόθεση 'by'.
- dobj: Το άμεσο αντικείμενο μίας ρηματικής φράσης, δηλαδή μία ονοματική φράση που αποτελεί το αντικείμενο ενός ρήματος.
- nsubj: Μία ονοματική φράση που αποτελεί το υποκείμενο μίας υποπρότασης.
- nsubjpass: Μία ονοματική φράση που αποτελεί το υποκείμενο μίας υποπρότασης σε παθητική σύνταξη.
- pobj: Το αντικείμενο μίας πρόθεσης, δηλαδή η κεφαλή μίας ονοματικής φράσης που βρίσκεται μετά από την πρόθεση ή τα επιρρήματα 'here' και 'there'.

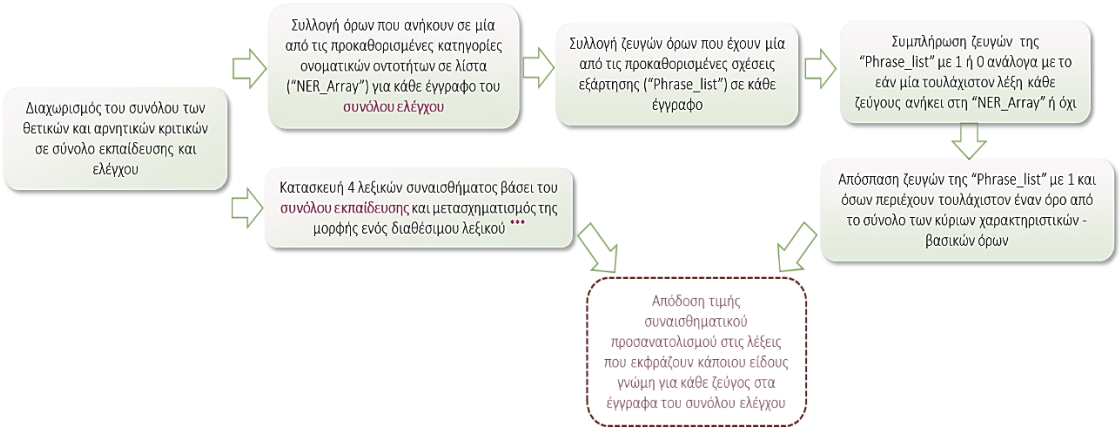
- xcomp: Ένα κατηγορηματικό ή προτασιακό συμπλήρωμα χωρίς το υποκείμενό του.
- neg: Πρόκειται για τη σχέση ανάμεσα σε μία λέξη άρνησης και τη λέξη που τροποποιεί.
- case: Μία προθετική φράση ενός ρήματος, επιθέτου ή ουσιαστικού που τροποποιεί τη σημασία του ρήματος, του επιθέτου ή του ουσιαστικού.
- dep: Πρόκειται για κάποια σχέση εξάρτησης μεταξύ δύο λέξεων που δεν μπορεί να αποσαφηνιστεί το ποια είναι, είτε λόγω ανικανότητας του λογισμικού, είτε λόγω ύπαρξης περίεργης γραμματικής δομής.

Ο αλγόριθμος περιλαμβάνει τον εντοπισμό στο κείμενο κάθε κριτικής του συνόλου ελέγχου των ζευγών όρων που συσχετίζονται μεταξύ τους με μία από τις παραπάνω σχέσεις και αποθηκεύονται σε μία νέα λίστα (‘Phrase_list’). Το ίδιο συμβαίνει και με τα λήμματα των λέξεων που ανήκουν σε μία από τις προκαθορισμένες κατηγορίες ονοματικών οντοτήτων, δημιουργώντας τη λίστα ‘NER_Array’. Έπειτα, παράγεται η λίστα με το όνομα ‘Phrase_list1’ έτσι ώστε να περιέχει, αρχικά, κάθε ζευγάρι λέξεων της Phrase_list, με μία τουλάχιστον λέξη στην ‘NER_Array’ ή χωρίς καμία λέξη στη ‘NER_Array’. Ανάλογα, με το πού ανήκουν οι λέξεις κάθε ζευγαριού, επιτελείται, έπειτα, η συγχώνευσή του με την τιμή 1 ή την τιμή 0 αντίστοιχα. Τέλος, δημιουργείται η λίστα ‘AspectSentimentWordsPair’ με τα ζεύγη της ‘Phrase_list1’ που φέρουν την τιμή 1 ή περιέχουν τουλάχιστον μία λέξη που ανήκει στο σύνολο των κύριων χαρακτηριστικών και των βασικών όρων. Έτσι, αποκτάται για κάθε κριτική του συνόλου ελέγχου, ένα σύνολο ζευγών, καθένα από τα οποία περιέχει τα κύρια χαρακτηριστικά (ή τους βασικούς όρους) καθώς και τις λέξεις που αποτυπώνουν κάποιου είδους γνώμη γι’ αυτά.

Ταυτόχρονα με την προηγούμενη διαδικασία, κατασκευάζονται από την αρχή ή μετασχηματίζονται τα λεξικά συναισθήματος που αναφέρονται στην ενότητα 6. Τα τέσσερα από αυτά (‘SPLM’, ‘SentiDomain’, ‘SentiPosNeg’, ‘SentiDraw’), με τον τρόπο δημιουργίας τους, χρησιμοποιώντας τις κριτικές του συνόλου εκπαίδευσης (Labille, Gauch, & Alfarhood, 2017), είναι πλήρως προσαρμοσμένα στον τομέα στον οποίο ενσωματώνονται οι διαθέσιμες κριτικές. Το μόνο που εξαιρείται, χωρίς να συνδέεται με αυτόν, είναι το λεξικό SentiWordNet, για το οποίο μόνο πραγματοποιείται κάποιου είδους αναδόμηση, όπως περιγράφεται στην ενότητα 6.1. Σε καθένα από αυτά, κάθε μοναδικός συνδυασμός λέξης-μέρους του λόγου (εντοπισμένος στις κριτικές του συνόλου εκπαίδευσης) χαρακτηρίζεται από μία τιμή που φανερώνει τη συναισθηματική του χροιά.

Σε επόμενο βήμα, διερευνάται η αποτελεσματικότητα των λεξικών με τη βοήθεια επισημασμένων δεδομένων, δηλαδή δεδομένων με γνωστή βαθμολογία. Υπό το πλαίσιο αυτό, σε κάθε κριτική του συνόλου ελέγχου, αποδίδεται σε κάθε αναφερόμενο χαρακτηριστικό μία τιμή συναισθηματικού προσανατολισμού, μέσω αντιστοίχισης της συνοδευόμενης λέξης που αποτυπώνει γνώμη γι’ αυτό με την αριθμητική τιμή που δίνει το εκάστοτε λεξικό.

Η ακόλουθη εικόνα παρουσιάζει σχηματικά τα επιμέρους βήματα που οδηγούν στην επίτευξη της επιδιωκόμενης πρόσδοσης τιμής πολικότητας στα αναφερόμενα χαρακτηριστικά.



Εικόνα 9. Διάγραμμα ροής για τη βαθμολόγηση της πολικότητας των χαρακτηριστικών.

7.4.4. Αποτύπωση υποδηλούμενου συναισθήματος των κριτικών

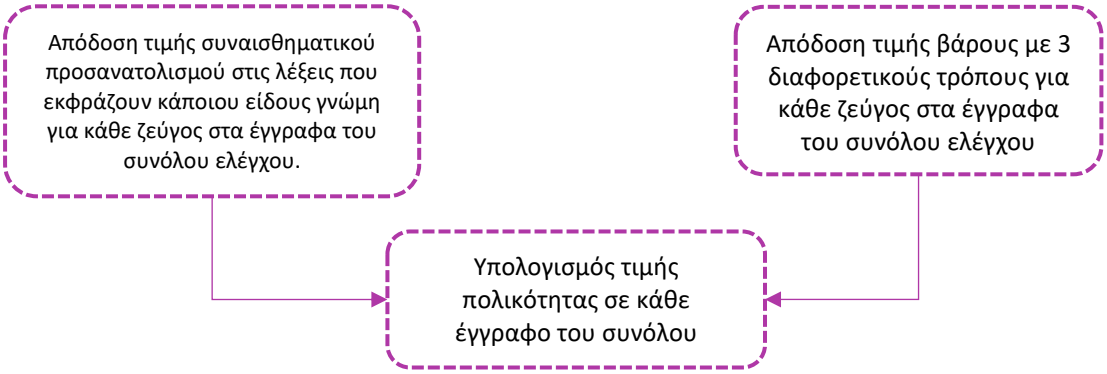
Ο σκοπός καταγραφής μίας ενιαίας τιμής, ικανής να συνοψίζει τη γνώμη ενός χρήστη, επιτυγχάνεται σε αυτό το βήμα. Για κάθε κριτική του συνόλου ελέγχου, η τιμή πολικότητας που έχει ανατεθεί σε κάθε εντοπισμένο ζευγάρι (κατά το προηγούμενο στάδιο) πολλαπλασιάζεται με το βάρος που έχει το χαρακτηριστικό του ζευγαριού (σύμφωνα με επιλεγμένη προσέγγιση στάθμισης). Έτσι, παράγεται, τελικά, μία μοναδική τιμή, έστω $P_{i,k}$ για το ζευγάρι i στην κριτική k του συνόλου ελέγχου. Δηλαδή, ισχύει:

$$P_{i,k} = S_i \cdot W_{i,k} \,, \tag{37}$$

όπου S_i είναι η τιμή πολικότητας των λέξεων στο ζευγάρι i που υποδηλώνουν κάποια γνώμη αναφορικά με ένα κύριο χαρακτηριστικό (ή έναν βασικό όρο) και $W_{i,k}$ είναι το βάρος που έχει το κύριο χαρακτηριστικό (ή ο βασικός όρος) του ζευγαριού i στην κριτική k.
Αφού επαναληφθεί η ίδια διαδικασία για όλα τα ζευγάρια κάθε κριτικής, αθροίζονται οι επιμέρους τιμές και προκύπτει η εκτιμώμενη αριθμητική τιμή συναισθηματικού προσανατολισμού ολόκληρης της κριτικής. Έτσι, το συνολικό σκορ της κριτικής k υπολογίζεται ως

$$R_k = \sum_{i=1}^{L_k} P_{i,k} = \sum_{i=1}^{L_k} S_i \cdot W_{i,k} \ , \qquad (38)$$

όπου L_k είναι το πλήθος των ζευγαριών που εντοπίζει ο αλγόριθμος στην κριτική k. Η ακόλουθη εικόνα παριστάνει σχηματικά τη συγκεκριμένη διαδικασία για την καλύτερη κατανόησή της.



Εικόνα 10. Διάγραμμα ροής για την αποτύπωση του υποδηλούμενου συναισθήματος των κριτικών.

Η συνάρτηση στον κώδικα που πραγματοποιεί τη ζητούμενη εργασία είναι η 'ReviewScore_AspectBased' η οποία καλεί τις συναρτήσεις 'Aspect_SentimentWords_Pair' και 'Sentiment_Score'.

7.5. ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΣΗΣ

7.5.1. Μετρικές ταξινόμησης

Μετρικές ταξινόμησης ονομάζονται οι συναρτήσεις που χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός κατασκευασμένου μοντέλου όσον αφορά την επιδιωκόμενη ταξινόμηση, συνηθέστερα, στο σύνολο επικύρωσης αλλά ενδέχεται και στο σύνολο ελέγχου.

- ***Ακρίβεια ταξινόμησης (Classification Accuracy)***

Η ακρίβεια ταξινόμησης θεωρείται ως η απλούστερη μετρική μιας εργασίας ταξινόμησης. Ορίζεται ως το κλάσμα των ορθά ταξινομημένων δειγμάτων προς το συνολικό πλήθος των δειγμάτων που ανήκουν σε εκείνο το σύνολο για το οποίο

μελετάται η απόδοσή του στην επιθυμητή ταξινόμηση. Πιο συγκεκριμένα, η ακρίβεια ταξινόμησης υπολογίζεται ως

$$accuracy = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\left\{d^{(k)}=\hat{d}^{(k)}\right\}}, \quad (39)$$

όπου $d^{(k)}$: η ετικέτα-στόχος που αντιστοιχεί στην είσοδο του δείγματος k, εκφρασμένη πλέον, όχι ως διάνυσμα διάστασης Cl, αλλά $d^{(k)} \in \{1,2,\dots,Cl\}$. Ως $\hat{d}^{(k)}$ συμβολίζεται η ετικέτα που προβλέπεται από τον εκάστοτε ταξινομητή για την είσοδο του δείγματος k.

- ***Πίνακας σύγχυσης (Confusion Matrix)***

Σε μια εργασία ταξινόμησης Cl, σε πλήθος, κλάσεων, ο πίνακας σύγχυσης για δεδομένο σύνολο, είναι ένας τετραγωνικός πίνακας C, διάστασης $ClxCl$, για τα στοιχεία $C_{i,j}$ ($i,j \in \{1,2,\dots,Cl\}$) του οποίου, ισχύει η ιδιότητα ότι $C_{i,j}$= πλήθος δειγμάτων του συνόλου των οποίων η επιθυμητή κλάση είναι η i και η προβλεπόμενη, από το μοντέλο, κλάση είναι η j. Έτσι, γίνεται αντιληπτό ότι η κύρια διαγώνιος του πίνακα σύγχυσης περιλαμβάνει το συνολικό αριθμό των δειγμάτων που ταξινομούνται κατά το δοκούν σε καθεμία κλάση. Κατά συνέπεια, όσα περισσότερα στοιχεία του πίνακα, εκτός αυτών της κύριας διαγωνίου είναι μηδενικά, τόσο πιο σωστή είναι η ταξινόμηση.

Στην περίπτωση της δυαδικής ταξινόμησης, με 2 δυνατές κλάσεις (συνήθως 0/1), ο πίνακας σύγχυσης έχει διάσταση 2 x 2 και τα στοιχεία του καλούνται με τα ονόματα που φαίνονται στην ακόλουθη εικόνα.

		Predicted	
		Positive (1)	Negative (0)
Real	Positive (1)	True Positive (TP)	False Negative (FN)
	Negative (0)	False Positive (FP)	True Negative (TN)

Εικόνα 11. Πίνακας σύγχυσης στην περίπτωση δυαδικής ταξινόμησης.

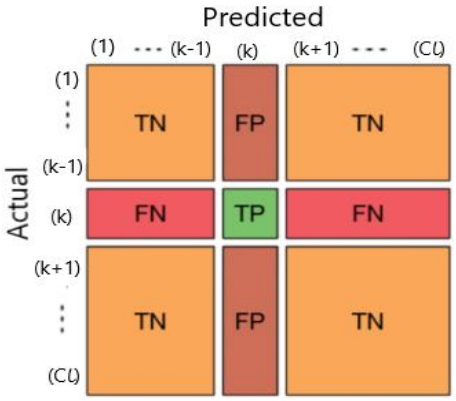
Ως αληθώς θετικά (True-Positive/TP) χαρακτηρίζονται τα δείγματα για τα οποία τόσο η επιθυμητή κλάση όσο και η προβλεπόμενη κλάση είναι αυτή που συμβολίζεται από τον χρήστη ως 1. Έτσι, το στοιχείο $C_{1,1}$ αντιπροσωπεύει το πλήθος των αληθώς θετικών

δειγμάτων στο σύνολο των δειγμάτων που γίνεται η αξιολόγηση της απόδοσης. Στην ίδια γραμμή του πίνακα, το στοιχείο $C_{1,2}$ υπολογίζεται ως το πλήθος των ψευδώς αρνητικών (False-Negative/FN**)** δειγμάτων, δηλαδή των δειγμάτων για τα οποία, παρ' ότι η πραγματική τους κλάση είναι η 1, το μοντέλο τα 'τοποθετεί' στην κλάση 0. Προχωρώντας στη δεύτερη και τελευταία γραμμή του πίνακα C, γίνεται αντιληπτό ότι αυτή αφορά το πλήθος των δειγμάτων που ανήκουν στην κλάση 0 και είτε έχουν ταξινομηθεί εσφαλμένα από το μοντέλο (ψευδώς θετικά ή False-Positive/FP), είτε έχουν ταξινομηθεί στην επιθυμητή κατηγορία (αληθώς αρνητικά ή True-Negative/TN).

Με τη βοήθεια του πίνακα σύγχυσης, μπορεί να βρεθεί και η τιμή της ακρίβειας ταξινόμησης που επιτυγχάνεται στο υπό μελέτη σύνολο. Ειδικότερα, υπολογίζεται με τη βοήθεια του ακόλουθου τύπου:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (40)$$

Όταν κάθε διάνυσμα εισόδου έχει περισσότερες από δύο υποψήφιες κατηγορίες, έστω Cl σε πλήθος, στις οποίες μπορεί να ανήκει η αντίστοιχη ετικέτα του (multi-class classification), τότε οι χαρακτηρισμοί των στοιχείων του πίνακα σύγχυσης ακολουθούν το μοτίβο που παριστάνεται στην παρακάτω εικόνα για την τυχαία κλάση k.



Εικόνα 12. Πίνακας σύγχυσης στην περίπτωση ταξινόμησης ανάμεσα σε περισσότερες από 2 κλάσεις.

Όπως δείχνει η εικόνα 8, στην περίπτωση της ταξινόμησης σε πολλές κλάσεις, εξακολουθεί να υφίσταται η διάκριση σε ψευδώς αρνητικά, ψευδώς θετικά, αληθώς θετικά και αληθώς αρνητικά δείγματα. Η διαφορά με τη δυαδική ταξινόμηση είναι ότι ο χαρακτηρισμός ενός δείγματος με ένα από τα 4 δυνατά ονόματα εξαρτάται από την υπό μελέτη κλάση. Έτσι, αν ορίσουμε $F = \{1,2, \dots Cl\}$ να είναι το σύνολο των δυνατών κλάσεων, το συνολικό πλήθος των ψευδώς αρνητικών δειγμάτων για την κλάση k ($k \in F$), είναι

$$FN_k = \sum_{\substack{j \in F \\ j \neq k}} C_{k,j}, \tag{41}$$

ενώ το πλήθος των ψευδώς θετικών και των αληθώς αρνητικών δειγμάτων είναι αντίστοιχα

$$FP_k = \sum_{\substack{i \in F \\ i \neq k}} C_{i,k} \tag{42}$$

και

$$TN_k = \sum_{i,j \in {}^{F}/_{\{k\}}} C_{i,j}. \tag{43}$$

Τότε, η συνολική ακρίβεια ταξινόμησης σε όλες τις κλάσεις δίνεται από τον τύπο:

$$accuracy = \frac{\sum_{i \in F} C_{i,i}}{\sum_{i \in F} \sum_{j \in F} C_{i,j}}. \tag{44}$$

- ***Precision***

Στην περίπτωση της ταξινόμησης δειγμάτων ενός συνόλου ανάμεσα σε δύο κλάσεις 0 ή 1, η μετρική ταξινόμησης γνωστή ως precision (ή user's accuracy), ορίζεται ως το κλάσμα του πλήθους των δειγμάτων που το μοντέλο σωστά 'τοποθέτησε' στην κλάση 1 προς το συνολικό αριθμό των δειγμάτων για τα οποία το μοντέλο προέβλεψε ότι ανήκουν στην κλάση 1. Δηλαδή, ισχύει:

$$precision = \frac{TP}{TP + FP}. \tag{45}$$

Ο ορισμός αυτός επεκτείνεται για καθεμία κλάση, όταν το αποτέλεσμα της ταξινόμησης είναι μια τιμή που ανήκει σε ένα σύνολο πληθικού αριθμού μεγαλύτερου του 2. Η σχέση για την τιμή της precision της κλάσης k ($k \in F$), δίνεται από τον τύπο (46)

$$precision_k = \frac{C_{k,k}}{C_{k,k} + \sum_{\substack{i \in F \\ i \neq k}} C_{i,k}} = \frac{C_{k,k}}{\sum_{i \in F} C_{i,k}}, \tag{46}$$

με τη συνολική τιμή της precision της ταξινόμησης να προκύπτει αθροίζοντας τις επιμέρους τιμές precision για κάθε κλάση. Επομένως, ισχύει:

$$precision = \sum_{k \in F} precision_k = \sum_{k \in F} \frac{C_{k,k}}{\sum_{i \in F} C_{i,k}}. \quad (47)$$

Επιπλέον, όταν οι κατηγορίες ταξινόμησης είναι πολλές, δύναται να χρησιμοποιηθεί και η μετρική macro-precision, η οποία υπολογίζεται ως ο μέσος όρος των τιμών της precision όλων των κλάσεων. Δίνεται από τον τύπο:

$$macro\_average\_precision = \frac{\sum_{k \in F} precision_k}{Cl}. \quad (48)$$

Όπως γίνεται αντιληπτό, κάθε κλάση έχει την ίδια στάθμιση στον υπολογισμό του μέσου όρου, οπότε δεν λαμβάνεται υπόψιν αν μια κλάση έχει υψηλή ή χαμηλή συχνότητα εμφάνισης στο εκάστοτε υποσύνολο δειγμάτων.

Η μετρική weighted average precision υπολογίζεται ως ο σταθμισμένος μέσος όρος των τιμών της precision όλων των κλάσεων. Όσον αφορά αυτή τη στάθμιση, πρόκειται για τον λόγο του συνολικού πλήθους εμφανίσεων της εκάστοτε κλάσης στο σύνολο δεδομένων προς το πλήθος εμφανίσεων όλων των κλάσεων.

- ***Recall***

Η μετρική recall (ή αλλιώς producer's accuracy) ορίζεται ως ο λόγος του πλήθους των σωστά ταξινομημένων δειγμάτων σε δεδομένη κλάση, προς το πλήθος των δειγμάτων που ήταν γνωστό εξαρχής ότι ανήκουν στη συγκεκριμένη κλάση. Με άλλα λόγια, η recall, με τη βοήθεια των στοιχείων του πίνακα σύγχυσης, προκύπτει, για τη δυαδική ταξινόμηση, να είναι ίση με

$$recall = \frac{TP}{TP + FN}\ , \quad (49)$$

ενώ για την ταξινόμηση ανάμεσα σε πολλές πιθανές κλάσεις, ορίζεται για την κλάση k ($k \in F$) ως:

$$recall_k = \frac{C_{k,k}}{C_{k,k} + \sum_{\substack{j \in F \\ j \neq k}} C_{k,j}} = \frac{C_{k,k}}{\sum_{j \in F} C_{k,j}}. \quad (50)$$

Σε αντιστοιχία με την περίπτωση της precision, η συνολική recall ορίζεται από τον ακόλουθο τύπο:

$$recall = \sum_{k \in F} recall_k = \sum_{k \in F} \frac{C_{k,k}}{\sum_{j \in F} C_{k,j}}. \quad (51)$$

Εύκολα μπορεί να συναχθεί το συμπέρασμα, ιδιαίτερα από τους τύπους (46) και (50) για τις $precision_k$ και $recall_k$ αντίστοιχα ότι όσο πιο κοντά στο 1 είναι οι τιμές precision και recall για κάθε κλάση, τόσο πιο ακριβής είναι η συνολική ταξινόμηση.

Η μετρική macro-recall ορίζεται παρόμοια με την macro-precision, δηλαδή ως ο αριθμητικός μέσος των μετρικών recall των επιμέρους κατηγοριών ταξινόμησης. Πιο συγκεκριμένα, η σχέση, με τη βοήθεια της οποίας υπολογίζεται η macro-recall είναι η:

$$macro\_average\_recall = \frac{\sum_{k \in F} recall_k}{Cl}. \quad (52)$$

Όπως μπορεί να γίνει κατανοητό, η weighted average recall υπολογίζεται με την ίδια λογική όπως η weighted average precision. Επεξηγηματικά, ορίζεται ως ο σταθμισμένος μέσος όρος των τιμών της recall όλων των κλάσεων.

- ***F1 Score***

Το F1 score είναι μια μετρική που χρησιμοποιείται με σκοπό να αντλήσει και να συνδυάσει την πληροφορία που πηγάζει από τις δύο προαναφερθείσες μετρικές ταξινόμησης. Ορίζεται ως ο αρμονικός μέσος των τιμών precision και recall, οπότε

$$F1\ score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \overset{(39),(40)}{\Longleftrightarrow} F1\ score = \frac{2}{\frac{TP + FP}{TP} + \frac{TP + FN}{TP}} \Leftrightarrow$$

$$\Leftrightarrow F1\ score = \frac{2}{\frac{2 \cdot TP + FP + FN}{TP}} \Leftrightarrow$$

$$\Leftrightarrow F1\ score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \quad (53)$$

Όταν οι διαθέσιμες κατηγορίες ταξινόμησης είναι περισσότερες από δύο, το συνολικό F1 score δίνεται από το άθροισμα των F1 score που πετυχαίνει κάθε κατηγορία ξεχωριστά. Δηλαδή, ισχύει ότι

$$F1\ score = \sum_{k \in F} (F1\ score)_k \Leftrightarrow F1\ score = \sum_{k \in F} \frac{2 \cdot C_{k,k}}{2 \cdot C_{k,k} + \sum_{\substack{i \in F \\ i \neq k}} C_{i,k} + \sum_{\substack{j \in F \\ j \neq k}} C_{k,j}} \Leftrightarrow$$

$$\Leftrightarrow F1\ score = \sum_{k \in F} \frac{2 \cdot C_{k,k}}{\sum_{i \in F} C_{i,k} + \sum_{j \in F} C_{k,j}}. \quad (54)$$

Το εύρος τιμών του F1 score, όταν αυτό υπολογίζεται για κάθε κλάση, είναι το διάστημα [0,1], με τη μέγιστη τιμή, όταν αυτή επιτυγχάνεται, να δηλώνει τέλεια ταξινόμηση των δειγμάτων στη συγκεκριμένη κλάση.

Καθίσταται χρήσιμο, σε αυτό το σημείο, να επισημανθεί ότι η μετρική macro-F1 score, εφόσον αποφασιστεί να υπολογιστεί για δεδομένη εργασία ταξινόμησης, παράγεται ως ο μέσος όρος των F1 scores που προκύπτουν για κάθε κλάση. Συνεπώς, ισούται με:

$$macro\_average\_F1\_score = \frac{\sum_{k \in F}(F1\ score)_k}{Cl}. \quad (55)$$

Το weighted average F1 score ορίζεται ως ο σταθμισμένος αριθμητικός μέσος της μετρικής F1 score υπολογισμένης για όλες τις κλάσεις ταξινόμησης. Η στάθμιση κάθε κλάσης, και σε αυτήν την περίπτωση, είναι ίση με την αναλογία εμφανίσεων της συγκεκριμένης κλάσης προς το συνολικό πλήθος των δεδομένων.

7.5.2. Αποτελέσματα μετρικών ταξινόμησης των δειγμάτων ελέγχου

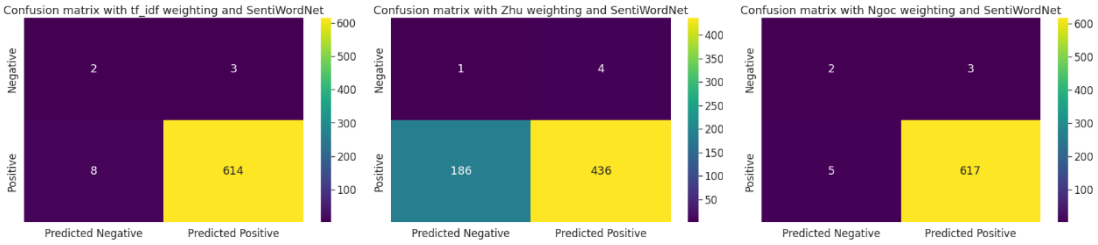
Στην περίπτωση χρήσης, για παράδειγμα, της tf-idf στάθμισης όρων και του λεξικού συναισθήματος SentiWordNet, όταν το εκτιμώμενο συνολικό σκορ των κριτικών (υπολογισμένο σύμφωνα με τον τύπο (38)) είναι μεγαλύτερο του -0,04, τότε το αντίστοιχο έγγραφο-κριτική εκλαμβάνεται ότι ανήκει στη θετική κλάση. Σε διαφορετική περίπτωση, θεωρείται ότι είναι επιφορτισμένο με αρνητικό συναίσθημα. Παρόμοια εξήγηση δίνεται και για τους υπόλοιπους συνδυασμούς τεχνικών.

Γενικά, όπως φαίνεται και παρακάτω, προτιμώνται για τη σύγκριση, οι μετρικές που είναι στη μορφή macro, καθώς η ανισορροπία μεταξύ των κλάσεων είναι αρκετά μεγάλη (622:5 υπέρ των θετικών εγγράφων). Καμία από τις δύο κλάσεις δεν είναι πιο σημαντική για τον ερευνητή ώστε να καταφύγουμε στους σταθμισμένους μέσους όρους των αντίστοιχων κριτικών. Τα κρίσιμα σημεία στα οποία αλλάζει το είδος συναισθήματος, αποφασίστηκαν σε κάθε περίπτωση, ώστε βάσει καθεμίας κατανομής εκτιμώμενων σκορ πολικότητας, να λαμβάνεται το μέγιστο δυνατό macro average F1 score.

Δεν είναι απόλυτα σωστό να βασιστούμε μόνο στη μετρική της συνολικής ακρίβειας ώστε να εξάγουμε τον καλύτερο αλγόριθμο για τους σκοπούς της ταξινόμησης, ιδιαίτερα τώρα, που η κατανομή των κριτικών στις 2 κλάσεις δεν είναι ισορροπημένη. Η απόφαση περί βέλτιστης επιλογής αλγορίθμου, συνήθως εξαρτάται κάθε φορά από το τι επιθυμεί να πετύχει ο χρήστης των αποτελεσμάτων και τι είναι αναγκαίο να αποφύγει. Συνήθως, όταν υπάρχει άνιση κατανομή κλάσεων, το F1 score είναι πιο χρήσιμο και πιο αποτελεσματικό από τη συνολική ακρίβεια ταξινόμησης, καθώς συνδυάζει τις τιμές precision και recall.

Πίνακας 6. Μετρικές απόδοσης ταξινόμησης στο σύνολο ελέγχου με χρήση του λεξικού συναισθήματος SentiWordNet και ένα από τα τρία μελετώμενα είδη στάθμισης.

	Tf_idf+SentiWordNet (-0,04)	Στάθμιση Zhu +SentiWordNet (-2,3)	Στάθμιση Ngoc +SentiWordNet (-0,2)
Overall accuracy	0,98	0,7	0,99
Misclassification error	0,02	0,3	0,01
Macro average precision	0,6	0,5	0,64
Macro average recall	0,69	0,45	0,7
Macro average F1 score	0,63	0,42	0,66

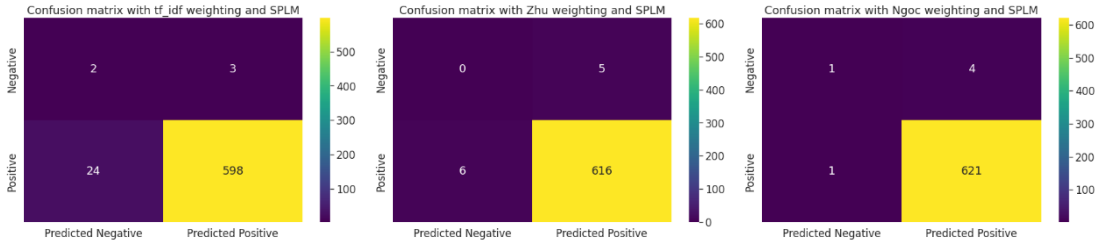


(α) (β) (γ)

Σχήμα 6. Πίνακες σύγχυσης για ταξινόμηση κριτικών του συνόλου ελέγχου με χρήση του λεξικού συναισθήματος SentiWordNet και (α) της κλασικής tf-idf στάθμισης, (β) της παραλλαγής της tf-idf κατά τον Zhu και (γ) της παραλλαγής της tf-idf κατά τον Ngoc.

Πίνακας 7. Μετρικές απόδοσης ταξινόμησης στο σύνολο ελέγχου με χρήση του λεξικού συναισθήματος SPLM και ένα από τα τρία μελετώμενα είδη στάθμισης.

	Tf_idf+SPLM (-0,1)	Στάθμιση Zhu +SPLM (-4,5)	Στάθμιση Ngoc +SPLM (-1,8)
Overall accuracy	0,96	0,98	0,99
Misclassification error	0,04	0,02	0,01
Macro average precision	0,54	0,5	0,75
Macro average recall	0,68	0,5	0,6
Macro average F1 score	0,55	0,5	0,64

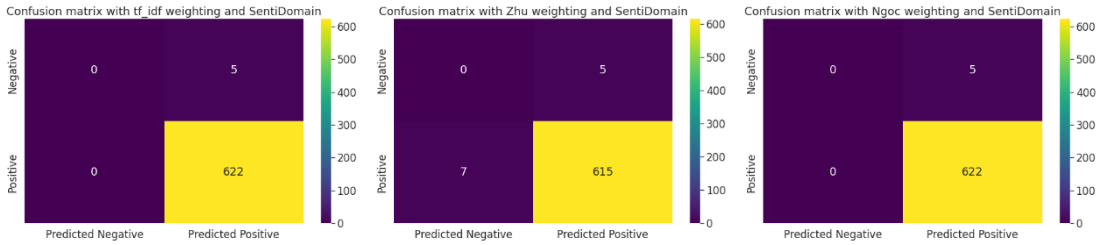


(α) (β) (γ)

Σχήμα 7. Πίνακες σύγχυσης για ταξινόμηση κριτικών του συνόλου ελέγχου με χρήση του λεξικού συναισθήματος SPLM και (α) της κλασικής tf-idf στάθμισης, (β) της παραλλαγής της tf-idf κατά τον Zhu και (γ) της παραλλαγής της tf-idf κατά τον Ngoc.

Πίνακας 8. Μετρικές απόδοσης ταξινόμησης στο σύνολο ελέγχου με χρήση του λεξικού συναισθήματος SentiDomain και ένα από τα τρία μελετώμενα είδη στάθμισης.

	Tf_idf+SentiDomain (-0,1)	Στάθμιση Zhu +SentiDomain (-80)	Στάθμιση Ngoc +SentiDomain (-0,01)
Overall accuracy	0,99	0,98	0,99
Misclassification error	0,01	0,02	0,01
Macro average precision	0,5	0,5	0,5
Macro average recall	0,5	0,49	0,5
Macro average F1 score	0,5	0,5	0,5



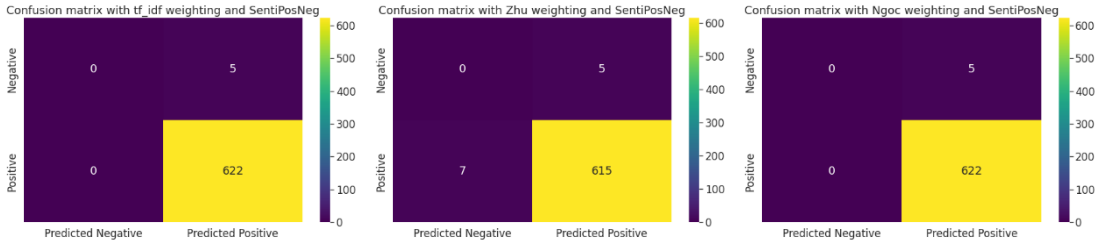
(α) (β) (γ)

Σχήμα 8. Πίνακες σύγχυσης για ταξινόμηση κριτικών του συνόλου ελέγχου με χρήση του λεξικού συναισθήματος SentiDomain και (α) της κλασικής tf-idf στάθμισης, (β) της παραλλαγής της tf-idf κατά τον Zhu και (γ) της παραλλαγής της tf-idf κατά τον Ngoc.

Πίνακας 9. Μετρικές απόδοσης ταξινόμησης στο σύνολο ελέγχου με χρήση του λεξικού συναισθήματος SentiPosNeg και ένα από τα τρία μελετώμενα είδη στάθμισης.

	Tf_idf+SentiPosNeg (-0,01)	Στάθμιση Zhu +SentiPosNeg (-80)	Στάθμιση Ngoc +SentiPosNeg (-0,01)
Overall accuracy	0,99	0,98	0,99
Misclassification error	0,01	0,02	0,01

Macro average precision	0,5	0,5	0,5
Macro average recall	0,5	0,49	0,5
Macro average F1 score	0,5	0,5	0,5

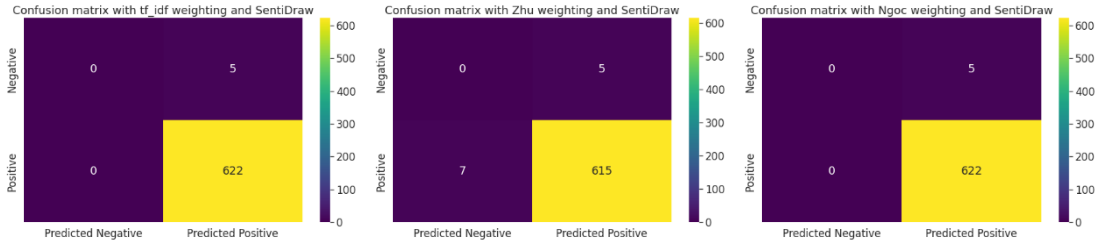


(α) (β) (γ)

Σχήμα 9. Πίνακες σύγχυσης για ταξινόμηση κριτικών του συνόλου ελέγχου με χρήση του λεξικού συναισθήματος SentiPosNeg και (α) της κλασικής tf-idf στάθμισης, (β) της παραλλαγής της tf-idf κατά τον Zhu και (γ) της παραλλαγής της tf-idf κατά τον Ngoc.

Πίνακας 10. Μετρικές απόδοσης ταξινόμησης στο σύνολο ελέγχου με χρήση του λεξικού συναισθήματος SentiDraw και ένα από τα τρία μελετώμενα είδη στάθμισης.

	Tf_idf+SentiDraw (-0,01)	Στάθμιση Zhu +SentiDraw(-70)	Στάθμιση Ngoc +SentiDraw(-0,01)
Overall accuracy	0,99	0,98	0,99
Misclassification error	0,01	0,02	0,01
Macro average precision	0,5	0,5	0,5
Macro average recall	0,5	0,49	0,5
Macro average F1 score	0,5	0,5	0,5

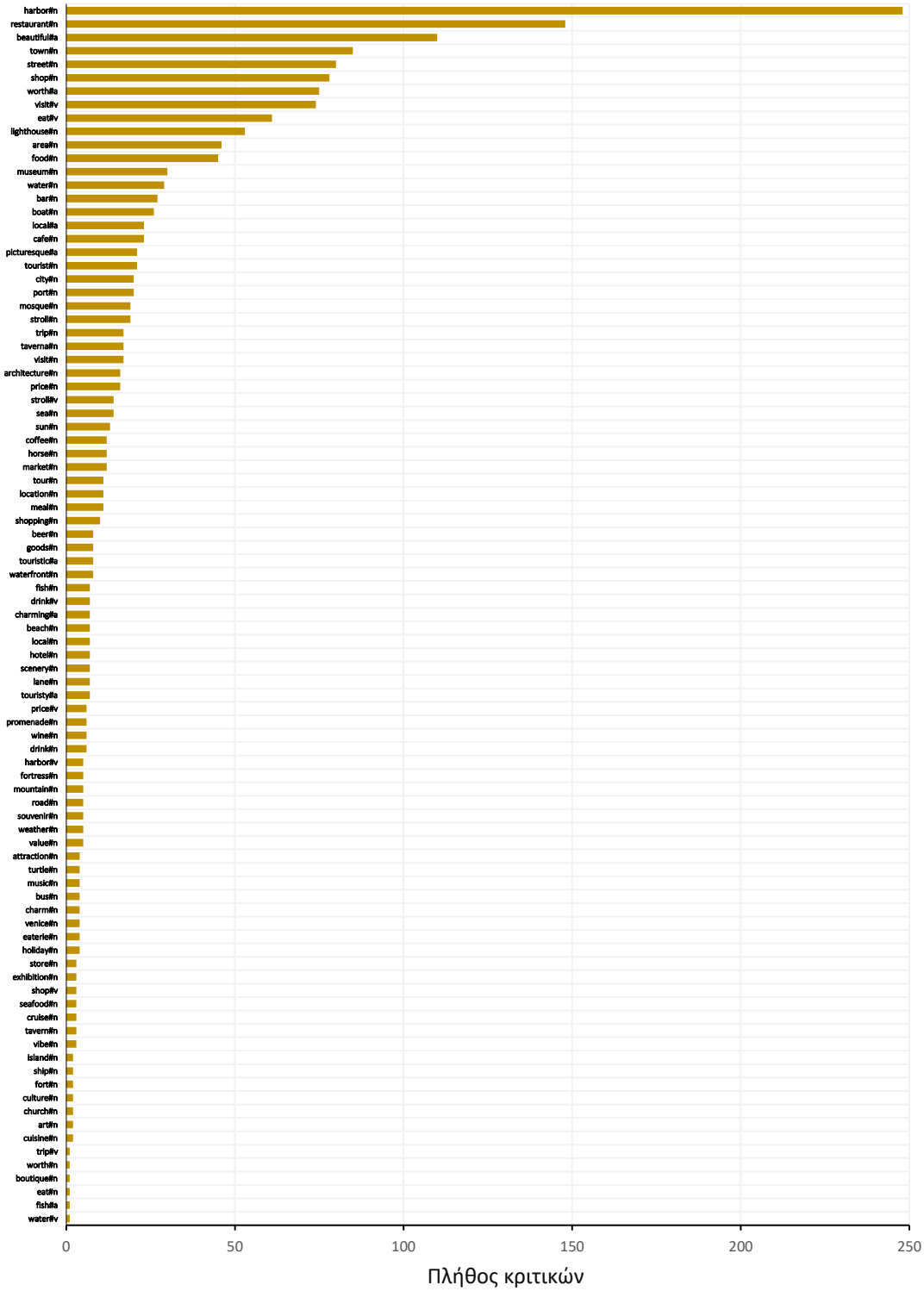


(α) (β) (γ)

Σχήμα 10. Πίνακες σύγχυσης για ταξινόμηση κριτικών του συνόλου ελέγχου με χρήση του λεξικού συναισθήματος SentiDraw και (α) της κλασικής tf-idf στάθμισης, (β) της παραλλαγής της tf-idf κατά τον Zhu και (γ) της παραλλαγής της tf-idf κατά τον Ngoc.

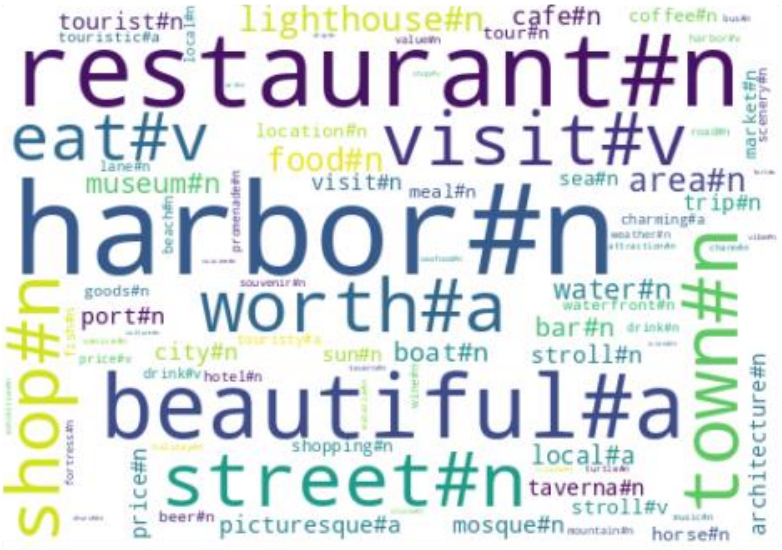
Βάσει των παραπάνω αποτελεσμάτων, η καλύτερη ταξινόμηση ανάμεσα σε θετικές και αρνητικές κριτικές πραγματοποιείται με τη βοήθεια του λεξικού συναισθήματος SentiWordNet και της στάθμισης λέξεων που προτείνουν οι Nguyen Thi Ngoc, Nguyen

Thi Thu, & Nguyen. Οπότε, προχωρώντας τη διερεύνηση ένα βήμα παρακάτω, μπορούμε να αντλήσουμε σε ποια χαρακτηριστικά αναφέρονται κυρίως οι χρήστες στο κείμενο των κριτικών αλλά και να υπολογίσουμε μία μέση τιμή συναισθηματικής πολικότητας για καθένα από αυτά σε όλες τις κριτικές του συνόλου ελέγχου.

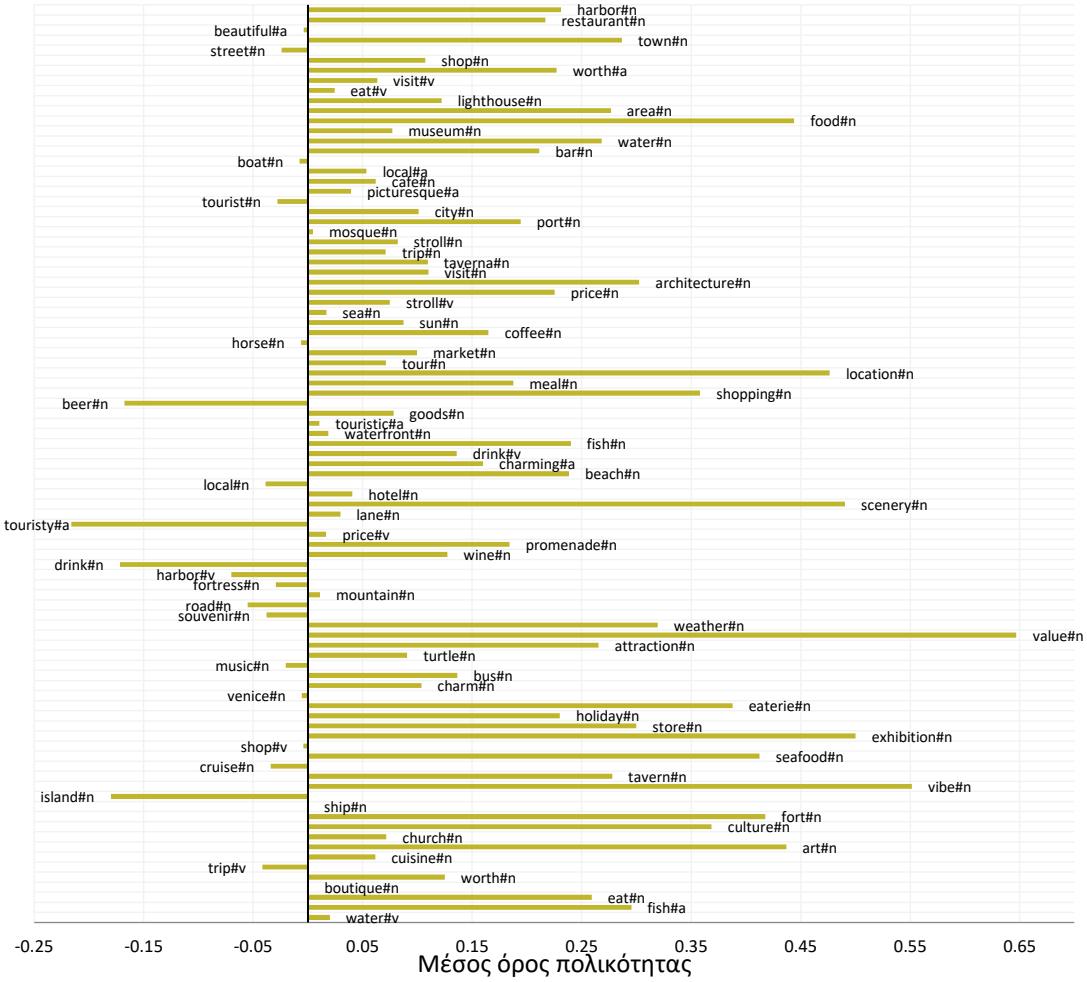


Σχήμα 11. Πλήθος κριτικών του συνόλου ελέγχου στις οποίες εμφανίζεται κάθε μοναδικός συνδυασμός χαρακτηριστικού-μέρους του λόγου.

Το νέφος λέξεων που ακολουθεί είναι μία γραφική αναπαράσταση της συχνότητας παρουσίας των λέξεων, με αυτήν να εκφράζεται μέσω του μεγέθους των λέξεων. Δηλαδή, όσο πιο μεγάλη απεικονίζεται μία λέξη, τόσο πιο μεγάλο είναι το πλήθος των εμφανίσεών της στην υπό μελέτη συλλογή κειμένου.



Εικόνα 13. Νέφος κύριων χαρακτηριστικών και βασικών όρων (ενωμένα με το μέρος του λόγου στο οποίο ανήκουν) βάσει της συχνότητας παρουσίας τους στις κριτικές του συνόλου ελέγχου.



Εικόνα 14. Μέση πολικότητα των κύριων χαρακτηριστικών και των βασικών όρων στις κριτικές του συνόλου ελέγχου.

Πίνακας 11. Συγκεντρωτικός πίνακας με το πλήθος των κριτικών του συνόλου ελέγχου στις οποίες εμφανίζεται κάθε συνδυασμός χαρακτηριστικού-μέρους του λόγου και μέση πολικότητα σε αυτές τις κριτικές.

Χαρακτηριστικό # μέρος του λόγου	Πλήθος κριτικών	Μέση πολικότητα	Χαρακτηριστικό # μέρος του λόγου	Πλήθος κριτικών	Μέση πολικότητα
harbor#n	248	0.23111	hotel#n	7	0.040482
restaurant#n	148	0.216679	local#n	7	-0.03863
beautiful#a	110	-0.00381	beach#n	7	0.238485
town#n	85	0.286743	charming#a	7	0.159889
street#n	80	-0.0242	drink#v	7	0.135909
shop#n	78	0.107176	fish#n	7	0.240141
worth#a	75	0.227137	drink#n	6	-0.17179
visit#v	74	0.063342	wine#n	6	0.127428
eat#v	61	0.024344	promenade#n	6	0.183968
lighthouse#n	53	0.122199	price#v	6	0.016643
area#n	46	0.276657	value#n	5	0.646999
food#n	45	0.443964	weather#n	5	0.319513

museum#n	30	0.076941	souvenir#n	5	-0.03794
water#n	29	0.268334	road#n	5	-0.05499
bar#n	27	0.211129	mountain#n	5	0.010914
boat#n	26	-0.00766	fortress#n	5	-0.02933
cafe#n	23	0.061951	harbor#v	5	-0.0699
local#a	23	0.053356	holiday#n	4	0.229851
tourist#n	21	-0.02788	eaterie#n	4	0.387839
picturesque#a	21	0.039292	venice#n	4	-0.00568
port#n	20	0.194374	charm#n	4	0.1037
city#n	20	0.10119	bus#n	4	0.13653
stroll#n	19	0.082049	music#n	4	-0.02041
mosque#n	19	0.00454	turtle#n	4	0.090619
visit#n	17	0.110102	attraction#n	4	0.265413
taverna#n	17	0.109434	vibe#n	3	0.5515
trip#n	17	0.071005	tavern#n	3	0.277981
price#n	16	0.225236	cruise#n	3	-0.03409
architecture#n	16	0.302601	seafood#n	3	0.412511
sea#n	14	0.01678	shop#v	3	-0.00418
stroll#v	14	0.074788	exhibition#n	3	0.5
sun#n	13	0.087358	store#n	3	0.299989
market#n	12	0.09971	cuisine#n	2	0.061505
horse#n	12	-0.00638	art#n	2	0.436882
coffee#n	12	0.164891	church#n	2	0.071511
meal#n	11	0.187661	culture#n	2	0.368613
location#n	11	0.47653	fort#n	2	0.41771
tour#n	11	0.071343	ship#n	2	0
shopping#n	10	0.358165	island#n	2	-0.18
waterfront#n	8	0.018684	water#v	1	0.02
touristic#a	8	0.010392	fish#a	1	0.295455
goods#n	8	0.07817	eat#n	1	0.259354
beer#n	8	-0.16763	boutique#n	1	0
touristy#a	7	-0.2161	worth#n	1	0.125
lane#n	7	0.029588	trip#v	1	-0.04167
scenery#n	7	0.49053			

8. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εξόρυξη γνώμης των χρηστών του Διαδικτύου είναι μία εξαιρετικά πολύτιμη διαδικασία, με όλο και περισσότερες επιχειρήσεις και ιδιώτες να ξεκινούν, σταδιακά, να ασχολούνται με το πεδίο ανάλυσης συναισθήματος. Θεμελιώδης στόχος είναι η απόκτηση τόσο μίας εποπτικής εικόνας, όσο και μίας λεπτομερέστερης ανάλυσης για την απόδοση των προϊόντων ή υπηρεσιών. Το ίδιο ισχύει και για τους ερευνητές, οι οποίοι χρησιμοποιούν αυτήν την πληροφορία, προκειμένου να παρακολουθούν τις τάσεις της αγοράς.

Είναι ξεκάθαρο, από τα αποτελέσματα που προέκυψαν στην προηγούμενη ενότητα, ότι η χρήση του λεξικού συναισθήματος γενικού σκοπού SentiWordNet το οποίο δεν προσαρμόστηκε στον τομέα των κριτικών, παρά μόνο τροποποιήθηκε ώστε να αποκτηθούν τιμές πολικότητας για κάθε μοναδικό συνδυασμό λέξης-μέρους του λόγου, ήταν η πιο αποδοτική. Η ταυτόχρονη στάθμιση των όρων με τη βοήθεια μίας τροποποιημένης τεχνικής tf-idf που εκμεταλλεύεται το πλήθος των προτάσεων παρήγαγε την υψηλότερη τιμή macro average F1 score σε σύγκριση με όλες τις υπόλοιπες προσεγγίσεις. Αυτή είναι ίση με 0,66, η οποία σε γενικές γραμμές είναι σε χαμηλά επίπεδα, οπότε και δεν είναι τόσο ικανοποιητική ώστε να ισχυριστούμε ότι η ταξινόμηση με όλη την προαναφερθείσα μεθοδολογία είναι πετυχημένη. Σε κάθε περίπτωση, όμως, αποδεικνύει ότι τα λεξικά με λέξεις που λαμβάνουν μία τιμή συναισθηματικού προσανατολισμού βάσει των διαθέσιμων εγγράφων δεν σημαίνει απαραίτητα ότι θα κατορθώσουν να αιχμαλωτίσουν αποτελεσματικά την εκφραζόμενη συναισθηματική χροιά. Σε όλες τις συγκρινόμενες προσεγγίσεις, αξιοποιήθηκε η ίδια λίστα κύριων χαρακτηριστικών και βασικών όρων, η οποία παρ' όλο που δεν μπορεί να αξιολογηθεί, λόγω του ότι δεν υπάρχει κάποια άλλη διαθέσιμη λίστα από παρόμοια μελέτη, κρίνεται ότι περιλαμβάνει αρκετά αντιπροσωπευτικούς όρους και έννοιες που απασχολούν έναν επισκέπτη του παλιού λιμανιού των Χανίων.

Το σημαντικό πρόβλημα των διαθέσιμων δεδομένων είναι η μεγάλη ανισορροπία των κλάσεων. Αν βασιστούμε μόνο στο γεγονός ότι το υποσύνολο ελέγχου το οποίο σχηματίστηκε με στρωματοποιημένο τρόπο από το αρχικό σύνολο, περιέχει μόνο 5 έγγραφα-κριτικές της αρνητικής κλάσης, ενώ το σύνολο στο οποίο βασίστηκε η κατασκευή των λεξικών αποτελείται από 20 κριτικές αρνητικού προσήμου, γίνεται κατανοητό ότι οι χαμηλές τιμές μετρικών μπορούν να δικαιολογηθούν. Το μικρό πλήθος εγγράφων δεν βοηθάει κανέναν αλγόριθμο να αντλήσει όλη την αναγκαία πληροφορία.

Ενδεχόμενες λύσεις του συγκεκριμένου προβλήματος, είναι η υπερδειγματοληψία ώστε να αυξηθεί το πλήθος των εγγράφων που ανήκουν στην αρνητική κλάση (έχοντας βαθμολογία 1 ή 2) ή η υποδειγματοληψία ώστε να ελαττωθούν οι θετικές κριτικές (με βαθμολογία 4 ή 5). Βέβαια, η δεύτερη λύση χρειάζεται προσοχή ώστε να μην καταλήξει το καινούριο σύνολο δεδομένων να είναι πολύ μικρό σε μέγεθος γιατί προφανώς τότε

τα αποτελέσματα με κανέναν αλγόριθμο δεν πρόκειται να είναι αξιόπιστα. Σε κάθε περίπτωση, συνιστάται μεγάλη προσοχή, ώστε, όταν μεταβάλλεται το πλήθος των δειγμάτων, η ανταλλαγή (tradeoff) που επιτελείται μεταξύ μεροληψίας και διασποράς, να μην επιφέρει ούτε υποπροσαρμογή (underfitting) ούτε υπερπροσαρμογή (overfitting). Ένας άλλος τρόπος αντιμετώπισης της μεγάλης ανισορροπίας των κλάσεων μπορεί να είναι η κατασκευή λεξικού με τέτοιες τιμές συναισθηματικού προσανατολισμού έτσι ώστε σε αυτές να υπεισέρχεται κάποιος κατάλληλος παράγοντας. Θα ήταν χρήσιμο, αυτός να σταθμίζει την επίδραση του είδους πολικότητας και να εξισορροπεί τις μεγάλες διαφορές στις επιμέρους ποσότητες που οφείλονται κατά βάση στο άνισο πλήθος εγγράφων κάθε κατηγορίας.

9. ΠΑΡΑΡΤΗΜΑ 1: ΚΩΔΙΚΑΣ ΣΕ PYTHON

Ο κώδικας που συνοδεύει όλα τα βήματα της μεθοδολογίας δημιουργήθηκε σε σημειωματάρια του Google Colab το οποία τρέχουν σε Python. Η βασική ιδέα πίσω από τον στόχο και την υλοποίηση κάθε βήματος των αλγορίθμων δόθηκε από την εργασία των Al-Ghuribi, Mohd Noah, & Tiun (2020), συμπληρώνοντας ή τροποποιώντας την με κατάλληλες εντολές σε ορισμένα σημεία που κρίθηκε ότι χρειαζόταν. Τα βασικά αρχεία τύπου .ipynb είναι τα εξής:

- Basic ABSA.ipynb

Περιλαμβάνει τις εντολές της βασικής μεθοδολογίας, από την είσοδο των δεδομένων, την προεπεξεργασία του κειμένου, την εξαγωγή κύριων χαρακτηριστικών και βασικών όρων και τον υπολογισμό των εκτιμώμενων τιμών συναισθηματικού προσανατολισμού των κριτικών.

- weighting.ipynb

Περιλαμβάνει τις εντολές για την παραγωγή πλαισίων δεδομένων που δίνουν το βάρος κάθε όρου (με τον τρόπο που δημιουργείται αυτός) για κάθε κριτική, εφαρμόζοντας και τις 3 τεχνικές στάθμισης όρων.

- sentiment lexicon.ipynb

Αφορά τη δημιουργία των λεξικών συναισθήματος που είναι προσαρμοσμένα στο σημασιολογικό πεδίο των κριτικών.

- sentiwordnet.ipynb

Αφορά κυρίως το μετασχηματισμό του λεξικού SentiWordNet βάσει της λογικής του Petter Tonberg αλλά και την προσθήκη αλλαγών ώστε το παραγόμενο πλαίσιο δεδομένων να συμβαδίζει στη μορφή του με τα υπόλοιπα των άλλων προσεγγίσεων.

- aspect_rating.ipynb

Περιλαμβάνει τον κώδικα που παράγει τους μέσους όρους συναισθηματικού προσανατολισμού για κάθε χαρακτηριστικό στις κριτικές του συνόλου ελέγχου καθώς επίσης και το νέφος με το πλήθος των κριτικών στις οποίες παρουσιάζεται το εκάστοτε χαρακτηριστικό ανήκοντας σε συγκεκριμένο μέρος του λόγου.

Στην πραγματικότητα, τα αρχεία είναι πολύ περισσότερα από αυτά τα τέσσερα, καθώς η αδυναμία εκτέλεσης ορισμένων συναρτήσεων για όλο το μέγεθος της εισόδου τους (εξαιτίας ελλιπούς διαθέσιμης μνήμης RAM) οδήγησε στη δημιουργία ίδιων σημειωματάριων ώστε αυτά να εκτελούν τις ίδιες συναρτήσεις ανά τμήματα της αρχικής εισόδου. Στο τέλος, τα επιμέρους αρχεία εξόδων της ίδιας συνάρτησης

ενοποιούνται σε ένα μοναδικό αρχείο και συνεχίζεται ο κώδικας με τις υπόλοιπες εντολές.

Όλα τα αρχεία που παρήχθησαν κατά την εκτέλεση του κώδικα ή δημιουργήθηκαν ώστε να εξυπηρετήσουν κάποιο στάδιο των αλγορίθμων είναι διαθέσιμα στον επισυναπτόμενο φάκελο.

10. ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

- *About Tripadvisor*. (χ.χ.). Ανάκτηση Δεκέμβριος 29, 2022, από https://tripadvisor.mediaroom.com/US-about-us
- Al-Ghuribi, S., Mohd Noah, S., & Tiun, S. (2020). Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews. *IEEE Access, 8*, pp. 218592-218613. doi:10.1109/ACCESS.2020.3042312
- Almatarneh, S., & Gamallo, P. (2017). Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification. *International Conference on Practical Applications of Agents and Multi-Agent Systems. 619*, pp. 175-182. Springer, Cham. doi:10.1007/978-3-319-61578-3_17
- Archak, N., Ghose, A., & Ipeirotis, P. (2007). Show me the money!: deriving the pricing power of product features by mining consumer reviews. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, (σσ. 56-65). doi:10.1145/1281192.1281202
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).
- Benamara, F., Cesarano, C., Picariello, A., Recupero, D., & Subrahmanian, V. (2007). Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. *International Conference on Web and Social Media.*
- Bird, S., Klein, E., & Loper, E. (2019). Accessing Text Corpora and Lexical Resources. Στο S. Bird, E. Klein, & E. Loper, *Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit.* Ανάκτηση από https://www.nltk.org/book/ch02.html
- Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., & Reynar, J. (2008). Building a Sentiment Summarizer for Local Service Reviews. *Proceedings of the WWW2008 Workshop: NLP in the Information Explosion Era (NLPIX 2008).*
- Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 804-812).
- Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications, 72*, pp. 221-230. doi:10.1016/j.eswa.2016.10.065

- Chen, X., & Wan, X. (2022). A Simple Information-Based Approach to Unsupervised Domain-Adaptive Aspect-Based Sentiment Analysis. *arXiv:2201.12549v1*. doi:10.48550/arXiv.2201.12549
- Cristescu, M. P., Nerisanu, R. A., Mara, D. A., & Oprea, S.-V. (2022). Using Market News Sentiment Analysis for Stock Market Prediction. *Mathematics, 10*(22). doi:10.3390/math10224255
- D'Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications, 125*(3), pp. 26-33. doi:10.5120/ijca2015905866
- Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, (pp. 519-528). doi:10.1145/775152.775226
- De Marneffe, M.-C., & D. Manning, C. (2008). *Stanford typed dependencies manual.* Retrieved Φεβρουάριος 18, 2023, from https://nlp.stanford.edu/software/dependencies_manual.pdf
- Farooq, U., Mansoor, H., Nongaillard, A., Ouzrout, Y., & Qadir, M. (2017). Negation Handling in Sentiment Analysis at Sentence Level. *Journal of Computers, 12*(5), pp. 470-478. doi:10.17706/jcp.12.5.470-478
- Feldman, R. (2013). Techniques and Applications For Sentiment Analysis. *Communications of the ACM, 56*(4). doi:10.1145/2436256.2436274
- Gaind, B., Syal, V., & Padgalwar, S. (2019). Emotion Detection and Analysis on Social Media. *arXiv:1901.08458*. Retrieved from https://arxiv.org/abs/1901.08458
- Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. *Proceedings of the 20th international conference on Computational Linguistics.* doi:10.3115/1220355.1220476
- Ghag, K., & Shah, K. (2015). Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification. *IEEE International Conference on Computer, Communication and Control (IC4-2015)*, (pp. 1-6).
- Ghosh, S., Bhaduri, S., Kumar, S., Verma, J., Katyal, Y., & Saraswat, A. (2023). A Semi-supervised Vulnerability Management System. In K. Arai, *Lecture Notes in Networks and Systems-Intelligent Systems and Applications* (Vol. 1, pp. 97-113). Springer International Publishing. doi:https://doi.org/10.1007/978-3-031-16072-1_7
- Gogula, S., Rahouti, M., Gogula, S., Jalamuri, A., & Jagatheesaperumal, S. (2023). An Emotion-Based Rating System for Books Using Sentiment Analysis and Machine Learning in the Cloud. *Applied Sciences, 13*(2). doi:10.3390/app13020773

- *Google Code Archive-word2vec*. (χ.χ.). Ανάκτηση Ιανουάριος 17, 2023, από https://code.google.com/archive/p/word2vec/
- *googletrans 4.0.0rc1*. (χ.χ.). Ανάκτηση Νοέμβριος 27, 2022, από https://pypi.org/project/googletrans/4.0.0rc1/
- Gu, T., Zhao, H., He, Z., Li, M., & Ying, D. (2023). Integrating external knowledge into aspect-based sentiment analysis using graph neural network. *Knowledge-Based Systems, 259*. doi:10.1016/j.knosys.2022.110025.
- Gunawan, D., Saniyah, Z., & Hizriadi, A. (2019). Normalization of Abbreviation and Acronym on Microtext in Bahasa Indonesia by Using Dictionary-Based and Longest Common Subsequence (LCS). *Procedia Computer Science, 161*, pp. 553-559. doi:10.1016/j.procs.2019.11.155
- Güngör, T. (2010). Part-of-Speech Tagging. Στο N. Indurkhya, & F. J. Damerau, *Handbook of Natural Language Processing-Second Edition.* Chapman & Hall.
- HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One, 15*(5). doi:10.1371/journal.pone.0232525
- He, R., Lee, W., Ng, H., & Dahlmeier, D. (2017). An Unsupervised Neural Attention Model for Aspect Extraction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, (pp. 388-397). doi:10.18653/v1/P17-1036
- Hu, M., & Liu, B. (2004a). Mining opinion features in customer reviews. *Proceedings of the 19th national conference on Artifical intelligence*, (pp. 755-760).
- Hu, M., & Liu, B. (2004b). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 168-177).
- *ideone.com*. (n.d.). Retrieved Φεβρουάριος 21, 2023, from https://ideone.com/fork/M0G455
- *Internet & Text Slang Dictionary & Translator*. (χ.χ.). Ανάκτηση Δεκέμβριος 4, 2022, από https://www.noslang.com/dictionary/
- Jakob, N., & Gurevych, I. (2010). Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (pp. 1035-1045).
- Jeyapriya, A., & Selvi Kanimozhi, C. (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, (pp. 548-552). doi:10.1109/ECS.2015.7124967

- Joshi, M., & Penstein-Rosé, C. (2009). Generalizing Dependency Features for Opinion Mining. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 313-316). Association for Computational Linguistics.
- Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics*, (pp. 1367-1373). doi:10.3115/1220355.1220555
- Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751). Association for Computational Linguistics. doi:10.3115/v1/D14-1181
- Kumar Bhadra, A., Shaila, S., & Banga, M. (2022). Review on Sentiment Analysis and Polarity Classification of Sarcastic Sentences using Deep Learning in Social Media. Στο *Data Engineering and Intelligent Computing. Lecture Notes in Networks and Systems* (Τόμ. 446). Bhateja, Vikrant; Khin Wee, Lai; Lin, Jerry Chun-Wei; Satapathy, Suresh Chandra; Rajesh, T.M. doi:10.1007/978-981-19-1559-8_24
- Kumhar, S., Ansarullah, S., Gardezi, A., Ahmad, S., Sayed, A., & Shafiq, M. (2023). Translation of English Language into Urdu Language Using LSTM Model. *Computers, Materials and Continua, 74*(2), pp. 3899-3912. doi:10.32604/cmc.2023.032290
- Labille, K., Gauch, S., & Alfarhood, S. (2017). Creating Domain-Specific Sentiment Lexicons via Text Mining. *WISDOM 2017 : 6th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining.*
- *List of emoticons*. (χ.χ.). Ανάκτηση Νοέμβριος 25, 2022, από https://en.wikipedia.org/wiki/List_of_emoticons
- Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer Berlin, Heidelberg. doi:10.1007/978-3-642-19460-3
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.
- Liu, B. (2015). Dealing with Conditional Sentences. In B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.*
- Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In *Mining Text Data* (pp. 415-463). Aggarwal, Charu C.; Zhai, ChengXiang;. doi:10.1007/978-1-4614-3223-4_13
- Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the Web. *Proceedings of the 14th international conference on World Wide Web*, (pp. 342-351).
- Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2016). Automated rule selection for opinion target extraction. *Knowledge-Based Systems, 104*, pp. 74-88. doi:10.1016/j.knosys.2016.04.010

- Long, C., Zhang, J., & Zhu, X. (2010). A Review Selection Approach for Accurate Feature Rating Estimation. *Proceedings of the 23rd International Conference on Computational Linguistics*, (pp. 766-774).
- Mai, F., Galke, L., & Scherp, A. (2019). CBOW Is Not All You Need: Combining CBOW with the Compositional Matrix Space Model. *ICLR (Poster).*
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). Association for Computational Linguistics. doi:10.3115/v1/P14-5010
- Maree, M., Eleyat, M., Rabayah, S., & Belkhatir, M. (2023). A hybrid composite features based sentence level sentiment analyzer. *IAES International Journal of Artificial Intelligence (IJ-AI), 12*(1), pp. 284-294. doi:10.11591/ijai.v12.i1.pp284-294
- Maree, M., Kmail, A., & Belkhatir, M. (2019). Analysis and shortcomings of e-recruitment systems: Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage. *Journal of Information Science, 45*(6), pp. 713-735. doi:10.1177/0165551518811449
- Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. *Proceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, (pp. 301-311). doi:10.1007/11430919_37
- Mattiello, E. (2009). Difficulty of Slang Translation. Στο *Translation Practices* (σσ. 65-83). Brill. doi:10.1163/9789042029040_007
- McCormick, C. (2017). *Word2Vec Tutorial Part 2 - Negative Sampling.* Ανάκτηση από http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*. Ανάκτηση από https://arxiv.org/abs/1301.3781
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546*. Ανάκτηση από https://arxiv.org/abs/1310.4546
- Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751). Atlanta,Georgia: Association for Computational Linguistics.
- Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), pp. 39-41. doi:10.1145/219717.219748

- Miller, G., & Fellbaum, C. (1991). Semantic networks of English. *Cognition, 41*, pp. 197-229. doi:10.1016/0010-0277(91)90036-4
- Moghaddam, S., & Ester, M. (2010). Opinion digger: an unsupervised opinion miner from unstructured product reviews. *Proceedings of the 19th ACM international conference on Information and knowledge management*, (pp. 1825-1828). doi:10.1145/1871437.1871739
- Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems, 108*, pp. 92-101.
- Munezero, M., Montero, C., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing, 5*(2), pp. 101-111. doi:10.1109/TAFFC.2014.2317187
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003).*
- *Natural Language Toolkit*. (2022, Δεκέμβριος 12). Ανάκτηση Δεκέμβριος 29, 2022, από https://www.nltk.org/
- Ng, V., Dasgupta, S., & Niaz Arifin, S. M. (2006). Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, (pp. 611-618).
- Nguyen Thi Ngoc, T., Nguyen Thi Thu, H., & Nguyen, V. A. (2019). Mining aspects of customer's review on the social network. *Journal of Big Data, 6*(22). doi:10.1186/s40537-019-0184-5
- Nigam, K., & Hurst, M. (2006). Towards a Robust Metric of Polarity. Στο *Computing Attitude and Affect in Text: Theory and Applications* (Τόμ. 20, σσ. 265-279). Shanahan, James G.; Qu, Yan; Wiebe, Janyce;. doi:10.1007/1-4020-4102-0_20
- Nopp, C., & Hanbury, A. (2015). Detecting Risks in the Banking System by Sentiment Analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 591-600). Association for Computational Linguistics. doi:10.18653/v1/D15-1071
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, (pp. 79-86).
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, (pp. 486-495).
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), (pp. 27-35). doi:10.3115/v1/S14-2004

- Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (pp. 339-346).
- *re-Regular expression operations*. (χ.χ.). Ανάκτηση Νοέμβριος 26, 2022, από https://docs.python.org/3/library/re.html
- Riloff, E., & Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, (σσ. 105-112).
- Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, (pp. 25-32).
- Sahlgren, M. (2006). *Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces (PhD dissertation, Stockholm University).* Ανάκτηση από http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-1037.
- Salvetti, F., & Reichenbach, C. (2006). Opinion Polarity Identification of Movie Reviews. In J. G. Shanahan, Y. Qu, & J. Wiebe, *Computing Attitude and Affect in Text: Theory and Applications* (Vol. 20, pp. 303-316). Springer. doi:10.1007/1-4020-4102-0_23
- Sarkar, D. (2018). *Implementing Deep Learning Methods and Feature Engineering for Text Data: The Continuous Bag of Words (CBOW).* Ανάκτηση από https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html
- Satapathy, R., Guerreiro, C., Chaturvedi, I., & Cambria, E. (2017). Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, (pp. 407-413). doi:10.1109/ICDMW.2017.59
- Schouten, K., & Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 28*(3), pp. 813-830.
- Shah, N., & Rohilla, S. (χ.χ.). *Open source Emoticons and Emoji detection library: emot*. Ανάκτηση από https://github.com/NeelShah18/emot
- Sharma, S., & Dutta, G. (2021). SentiDraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Information Processing & Management, 58*(1). doi:10.1016/j.ipm.2020.102412

- Shinde, V., Pawar, A., Ahirrao, S., & Phansalkar, S. (2019). Emotions Identification by Using Unsupervised Aspect Category Based Sentiment Classification. *International Journal of Engineering and Advanced Technology (IJEAT), 8*(6), pp. 4224-4230. doi:10.35940/ijeat.f8902.088619
- Stanford NLP Group. (2020). *Using CoreNLP within other programming languages and packages*. Ανάκτηση Δεκέμβριος 29, 2022, από https://stanfordnlp.github.io/CoreNLP/other-languages.html
- Stanford NLP Group. (n.d.). *Class UniversalEnglishGrammaticalRelations*. Retrieved Φεβρουάριος 19, 2023, from https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/UniversalEnglishGrammaticalRelations.html
- Stanford NLP Group. (n.d.). *Named Entity Recognition*. Retrieved Φεβρουάριος 19, 2023, from https://stanfordnlp.github.io/CoreNLP/ner.html
- Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, pp. 10-25. doi:10.1016/j.inffus.2016.10.004
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics, 37*(2), pp. 267-307. doi:10.1162/COLI_a_00049
- Thabit, K., & Al-Ghuribi, S. (2013). A new search algorithm for documents using blocks and words prefixes. *Scientific Research and Essays, 8*(16), pp. 640-648. doi:10.5897/SRE2013. 5373
- Turlan, J., Ratlnov, L., & Benglo, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384-394). Uppsala, Sweden: Association for Computational Linguistics.
- Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 417-424).
- Turney, P., & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), pp. 315-346. doi:10.1145/944012.944013
- Uysal, A., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management, 50*(1), pp. 104-112.
- Vermeij, M. (n.d.). *The Orientation of User Opinions through Adverbs, Verbs and Nouns.* Retrieved Φεβρουάριος 14, 2023, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.4909&rep1&pdf

- Wang, R., Wang, Q., Liang, B., Chen, Y., Wen, Z., Qin, B., & Xu, R. (2022). Masking and Generation: An Unsupervised Method for Sarcasm Detection. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 2172-2177). doi:10.1145/3477495.3531825
- Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 90-94). Association for Computational Linguistics.
- Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *Proceedings of the 14th conference on Computational linguistics*, *4*, pp. 1106-1110. doi:10.3115/992424.992434
- Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. *Proceedings of National Conf. on Artificial Intelligence (AAAI-2000).*
- Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying Collocations for Recognizing Opinions. *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction,Analysis, and Exploitation*, (pp. 24-31).
- Wu, J., & Ji, T. (2016). *Deep Learning for Amazon Food Review Sentiment Analysis.* Retrieved from http://cs224d.stanford.edu/reports/WuJi.pdf
- Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*(6), pp. 1138-1152. doi:10.1016/j.ins.2010.11.023
- Xu, L., Pang, X., Wu, J., Cai, M., & Peng, J. (2023). Learn from structural scope: Improving aspect-level sentiment analysis with hybrid graph convolutional networks. *Neurocomputing, 518*, pp. 373-383. doi:10.1016/j.neucom.2022.10.071
- Yuan, L., Bin, J., Wei, Y., Huang, F., Hu, X., & Tan, M. (2020). Big Data Aspect-Based Opinion Mining Using the SLDA and HME-LDA Models. *Wireless Communications and Mobile Computing*. doi:10.1155/2020/8869385
- Zhang, L., Liu, B., Lim, S., & O'Brien-Strain, E. (2010). Extracting and Ranking Product Features in Opinion Documents. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, (pp. 1462-1470).
- Zhao, Y., Qin, B., Hu, S., & Liu, T. (2010). Generalizing Syntactic Structures for Product Attribute Candidate Extraction. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 377-380).
- Zhu, L., Wang, G., & Zou, X. (2016). A Study of Chinese Document Representation and Classification with Word2vec. *2016 9th International Symposium on Computational Intelligence and Design (ISCID)* (pp. 298-302). IEEE. doi:10.1109/ISCID.2016.1075

- Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management*, (pp. 43-50). doi:10.1145/1183614.1183625