ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Σχολή Τεχνολογίας

Τμήμα Ψηφιακών Συστημάτων

Ανάλυση Συναισθήματος με Χρήση LDA και Οπτικοποίηση Word Cloud στη Γλώσσα R

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μελισσοπούλου Ειρήνη Μυροφόρα (ΑΜ: 3121193)

Χρυσανθακοπούλου Ζωή (ΑΜ:3121053)

Επιβλέπων: Όμηρος Ιατρέλλης, Επίκουρος Καθηγητής

ΛΑΡΙΣΑ, Μάιος 2025

UNIVERSITY OF THESSALY

School of Technology

Digital Systems Department

Sentiment Analysis Using LDA and Word Cloud Visualization in R Programming Language

BACHELOR THESIS

Melissopoulou Eirini Myrofora (RN: 3121193)

Chrysanthakopoulou Zoi (RN: 3121053)

Supervisor: ***Omiros Iatrellis, Assistant Professor***

LARISSA, May 2025

Υπεύθυνη Δήλωση περί Ακαδημαϊκής Δεοντολογίας και Πνευματικών Δικαιωμάτων

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα πτυχιακή εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον, και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο/Η Δηλών/ούσα

(Υπογραφή)

Ονοματεπώνυμο Φοιτητή/-ήτριας

Μελισσοπούλου Ειρήνη Μυροφόρα

Χρυσανθακοπούλου Ζωή

Ημερομηνία 25/09/2025

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων/πουσα **Όμηρος Ιατρέλλης**
Επίκουρος Καθηγητής,
Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Θεσσαλίας

Μέλος **Ξενάκης Απόστολος**
Επίκουρος Καθηγητής, Τμήμα Ψηφιακών Συστημάτων/
Πανεπιστήμιο Θεσσαλίας

Μέλος **Χαικάλης Κωνσταντίνος**
Αναπληρωτής Καθηγητής ,Τμήμα Ψηφιακών Συστημάτων/
Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: dd-mm-yyyy

Περίληψη

Η ανάλυση συναισθήματος αποτελεί έναν από τους βασικότερους τομείς της Επεξεργασίας Φυσικής Γλώσσας (NLP), με αυξανόμενη σημασία σε περιβάλλοντα όπου παράγονται μαζικά κειμενικά δεδομένα, όπως τα κοινωνικά δίκτυα, οι επιχειρήσεις και η πολιτική ανάλυση. Η παρούσα εργασία εξετάζει τη χρήση του αλγορίθμου Latent Dirichlet Allocation (LDA) για την εξαγωγή θεμάτων από κείμενα και τη σύνδεσή τους με την ανάλυση συναισθήματος. Η μεθοδολογία αναπτύχθηκε στη γλώσσα R, χρησιμοποιώντας βιβλιοθήκες για την προεπεξεργασία δεδομένων, την εφαρμογή του LDA και την αντιστοίχιση συναισθημάτων με τα παραγόμενα θέματα. Για την καλύτερη κατανόηση των αποτελεσμάτων δημιουργήθηκαν wordclouds που αναδεικνύουν τις λέξεις-κλειδιά ανά συναίσθημα. Τα αποτελέσματα δείχνουν ότι η θεματική μοντελοποίηση μπορεί να αποκαλύψει κρυμμένα μοτίβα συναισθημάτων, παρέχοντας πολύτιμες πληροφορίες για την ερμηνεία μεγάλων όγκων κειμένου. Η εργασία προτείνει μια αποδοτική και επεκτάσιμη προσέγγιση ανάλυσης συναισθήματος και δημιουργεί βάση για μελλοντικές έρευνες με πιο προηγμένες τεχνικές, όπως υβριδικά ή νευρωνικά μοντέλα. Συνολικά, η προσέγγιση συμβάλλει στην κατανόηση της συναισθηματικής διάστασης των δεδομένων, ενισχύοντας την εφαρμοσιμότητα των ευρημάτων σε ποικίλα πεδία.

Abstract

Sentiment analysis is one of the core fields of Natural Language Processing (NLP), with increasing importance in environments where large volumes of textual data are generated, such as social media, business applications, and political analysis. This study explores the use of the Latent Dirichlet Allocation (LDA) algorithm for extracting topics from text and linking them to sentiment analysis. The methodology was developed using the R programming language, employing libraries for data preprocessing, LDA implementation, and associating emotions with the generated topics. To enhance result interpretation, wordclouds were created to highlight the key words related to each sentiment. The results indicate that topic modeling can reveal hidden sentiment patterns, offering valuable insights into the emotional tone of large text corpora. This work proposes an efficient and scalable approach to sentiment analysis and lays the groundwork for future research employing more advanced techniques, such as hybrid or neural models. Overall, the proposed approach contributes to a deeper understanding of the emotional dimension of textual data, enhancing the applicability of the findings across various fields.

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή μας, κ. Όμηρο Ιατρέλλη, για την καθοδήγηση, τη συνεργασία και τη στήριξη κατά την εκπόνηση της παρούσας εργασίας. Η συμβολή του υπήρξε καθοριστική για την ολοκλήρωση και την ποιότητα του έργου μας.

Ευχαριστούμε επίσης τις οικογένειες μας για την αμέριστη υποστήριξη, την κατανόηση και την ενθάρρυνση που μας προσέφεραν καθ'όλη τη διάρκεια των σπουδών μας.

Μελισσοπούλου Ειρήνη Μυροφόρα

Χρυσανθακοπούλου Ζωή

25/09/02025

Περιεχόμενα

1 Εισαγωγή

Η ανάλυση συναισθήματος, γνωστή και ως opinion mining, αποτελεί έναν κλάδο της Επεξεργασίας Φυσικής Γλώσσας που αποσκοπεί στον προσδιορισμό του συναισθηματικού τόνου ενός κειμένου μέσω υπολογιστικών τεχνικών. Πρόκειται για μια διαδικασία που αναλύει μεγάλα σύνολα δεδομένων, ταξινομώντας τα ως θετικά, αρνητικά ή ουδέτερα, με στόχο την εξαγωγή πολύτιμων πληροφοριών για τις απόψεις και τη συναισθηματική κατάσταση των ανθρώπων. Τα τελευταία χρόνια, η κατανόηση και ανάλυση των απόψεων των χρηστών έχει καταστεί ζωτικής σημασίας για επιχειρήσεις, οργανισμούς και ερευνητές. Η ανάπτυξη του διαδικτύου και των κοινωνικών δικτύων έχει οδηγήσει στη δημιουργία τεράστιων όγκων δεδομένων, παρέχοντας νέες ευκαιρίες αλλά και προκλήσεις για την εξαγωγή χρήσιμων συμπερασμάτων. Η ανάλυση συναισθήματος έχει συμβάλλει σημαντικά στην αυτοματοποίηση της επεξεργασίας αυτών των δεδομένων, επιτρέποντας τη λήψη τεκμηριωμένων αποφάσεων σε τομείς όπως το μάρκετινγκ, η πολιτική ανάλυση, η εξυπηρέτηση πελατών και η ψυχολογία. Μία από τις μεθόδους που χρησιμοποιούνται στην ανάλυση συναισθήματος είναι η Latent Dirichlet Allocation (LDA), μια στατιστική τεχνική που εφαρμόζεται στην εξαγωγή θεμάτων από μεγάλα σύνολα κειμένων. Το LDA αποτελεί μια μορφή μη επιβλεπόμενης μηχανικής μάθησης που αναγνωρίζει τις υποκείμενες θεματικές δομές ενός κειμένου και βοηθά στην κατηγοριοποίησή του. Στο πλαίσιο της ανάλυσης συναισθήματος, το LDA μπορεί να αξιοποιηθεί για την αναγνώριση των κυρίαρχων συναισθημάτων που εκφράζονται σε ένα σύνολο δεδομένων, προσφέροντας μια πιο λεπτομερή και ερμηνεύσιμη εικόνα της συναισθηματικής κατάστασης των χρηστών. Για την υλοποίηση της παρούσας μελέτης, χρησιμοποιείται η R, ένα ισχυρό περιβάλλον λογισμικού και γλώσσα προγραμματισμού που έχει δημιουργηθεί για στατιστικούς υπολογισμούς και οπτικοποίηση δεδομένων. Χάρη στις βιβλιοθήκες της, η R χρησιμοποιείται ευρέως στην επιστήμη δεδομένων και στην Ανάλυση Συναισθήματος για εξαγωγή πολύτιμων συμπερασμάτων από μεγάλα σύνολα δεδομένων. Παράλληλα, για την παρουσίαση των αποτελεσμάτων της ανάλυσης συναισθήματος θα χρησιμοποιηθεί το Word Cloud το οποίο είναι μια τεχνική που αναπαριστά τις λέξεις του κειμένου με μέγεθος που αντανακλά τη συχνότητά τους.

Χρησιμοποιείται για να αναδείξει τα πιο συχνά ή σημαντικά θέματα σε ένα σύνολο δεδομένων, προσφέροντας μια οπτικά ελκυστική και εύκολη στην κατανόηση αναπαράσταση του περιεχομένου. Το Word Cloud είναι ιδιαίτερα χρήσιμο για την ανάλυση δεδομένων από κοινωνικά δίκτυα ή άρθρα, επιτρέποντας την εύκολη ταυτοποίηση των βασικών θεμάτων που απασχολούν τους χρήστες. Για την υλοποίηση των πειραμάτων, η R θα χρησιμοποιηθεί ως βασικό εργαλείο ανάλυσης και οπτικοποίησης των δεδομένων, αξιοποιώντας τις διαθέσιμες βιβλιοθήκες για στατιστική επεξεργασία. Στο Κεφάλαιο 1 παρουσιάζεται το θέμα της εργασίας και η ερευνητική του σημασία, με έμφαση στην ανάλυση συναισθήματος ως πεδίο με αυξανόμενο ενδιαφέρον. Περιγράφονται οι στόχοι και η μεθοδολογία της εργασίας, ενώ δίνεται μια επισκόπηση των τεχνολογιών που θα χρησιμοποιηθούν. Στο Κεφάλαιο 2 γίνεται εκτενής ανάλυση της έννοιας και της λειτουργίας της ανάλυσης συναισθήματος. Παρουσιάζονται οι τρόποι με τους οποίους εφαρμόζεται, οι βασικές τεχνικές και τα εργαλεία που χρησιμοποιούνται, καθώς και τα βασικά πλεονεκτήματα και μειονεκτήματά της. Το Κεφάλαιο 3 εστιάζει στην τεχνική της εξόρυξης θεμάτων (Topic Modeling), παρουσιάζοντας μεθόδους όπως το LDA και το NMF. Αναλύονται τα πλεονεκτήματα και τα μειονεκτήματα αυτών των μεθόδων. Έπειτα, στο Κεφάλαιο 4 περιγράφεται η εφαρμογή της ανάλυσης συναισθήματος μέσω φυσικής γλώσσας (NLP) και η χρήση της μεθόδου LDA για την εξαγωγή θεμάτων από δεδομένα. Αναλύεται η διαδικασία εφαρμογής των μεθόδων, με έμφαση στη συσχέτισή τους με τα συναισθήματα των χρηστών. Το Κεφάλαιο 5 εξετάζει την τεχνική του Word Cloud ως μέσο οπτικοποίησης λέξεων, παρουσιάζοντας την ιστορία, τη λειτουργία, τα πλεονεκτήματα και τα μειονεκτήματά του. Επιπλέον, αναδεικνύεται ο ρόλος του Word Cloud στην ερμηνεία των αποτελεσμάτων της ανάλυσης συναισθήματος. Το Κεφάλαιο 6 επικεντρώνεται στη γλώσσα προγραμματισμού R, η οποία χρησιμοποιείται στην παρούσα εργασία για την ανάλυση και οπτικοποίηση των δεδομένων. Παρουσιάζονται βασικά χαρακτηριστικά της γλώσσας, οι λόγοι για τους οποίους είναι κατάλληλη για ανάλυση δεδομένων, καθώς και παραδείγματα εφαρμογών. Στο Κεφάλαιο 7 αναφέρονται η αναλυτική υλοποίηση και εφαρμογή της παρούσας εργασίας, ενώ στο Κεφάλαιο 8 παρουσιάζονται τα συμπεράσματα και οι μελλοντικές προοπτικές που προκύπτουν από την μελέτη. Στο τέλος του κειμένου παρατίθεται η βιβλιογραφία, η οποία περιλαμβάνει όλες τις πηγές που χρησιμοποιήθηκαν κατά τη συγγραφή της εργασίας αυτής, καθώς και τα παραρτήματα, τα οποία περιλαμβάνουν τον κώδικα, τις οπτικοποιήσεις Word Cloud και τα αποτελέσματα που προέκυψαν από την ανάλυση [1].

2 Ανάλυση Συναισθήματος (Sentiment Analysis)

2.1 Ορισμός και Λειτουργία της Ανάλυσης Συναισθήματος

Η ανάλυση συναισθήματος συνιστά μία υπολογιστική διαδικασία εντοπισμού, αναγνώρισης και κατηγοριοποίησης των συναισθηματικών φορτίσεων που εκφράζονται σε κείμενο, ταξινομώντας τις ως θετικές, αρνητικές ή ουδέτερες. Η εν λόγω ανάλυση εφαρμόζεται σε κείμενα που συσχετίζονται με υπηρεσίες, προϊόντα, γεγονότα ή άλλα χαρακτηριστικά ενδιαφέροντος. Οι βασικές πηγές των δεδομένων προέρχονται από ψηφιακά μέσα κοινωνικής δικτύωσης και επικοινωνίας, όπως είναι οι ιστότοποι αξιολογήσεων, τα διαδικτυακά φόρουμ, τα ιστολόγια, και πλατφόρμες όπως το Twitter. Αυτός ο κλάδος έρευνας είναι πολύ διαδεδομένος στις μέρες μας λόγω των δεδομένων με απόψεις που παρέχουν, στα οποία οι χρήστες μπορούν να βρουν κριτικές για ποικίλες υπηρεσίες που είναι απαραίτητες στην καθημερινότητά τους. Ένας σημαντικός όγκος δεδομένων παράγεται και αποθηκεύεται σε ψηφιακή μορφή, κυρίως μέσω διαδικτυακών πλατφορμών. Η ανάλυση συναισθήματος εφαρμόζεται σε τέτοιου είδους δεδομένα, με σκοπό την εξαγωγή πληροφορίας σχετικά με τη συναισθηματική στάση των χρηστών απέναντι σε ένα συγκεκριμένο θέμα ή αντικείμενο ενδιαφέροντος, αποδίδοντας αποτελέσματα που αντικατοπτρίζουν τη συνολική τάση ή αντίληψη του κοινού [2]. Η ανάλυση συναισθήματος, ως επιστημονικός τομέας, εντοπίζεται στα μέσα του 20ού αιώνα, όταν οι ερευνητές ανέλυαν και συνέκριναν γραπτά κείμενα προκειμένου να κατανοήσουν την πρόθεση των συγγραφέων. Ωστόσο, η ανάλυση συναισθήματος απέκτησε επιχειρηματική αξία και βιωσιμότητα μόνο με την εμφάνιση της ψηφιακής επικοινωνίας και της εξόρυξης μεγάλων δεδομένων. Σήμερα, οι τεχνολογικές εξελίξεις στον τομέα της Τεχνητής Νοημοσύνης, της βαθιάς μάθησης και της Επεξεργασίας Φυσικής Γλώσσας επιτρέπουν στους οργανισμούς να αντλούν και να αναλύουν τεράστιες ποσότητες δεδομένων πελατών, προκειμένου να αξιολογούν την κοινή γνώμη, να διεξάγουν έρευνες αγοράς, να παρακολουθούν τη φήμη τους και να αποκομίζουν βαθύτερη κατανόηση της εμπειρίας των πελατών. Η ανάλυση συναισθήματος

απαρτίζεται από συγκεκριμένα βήματα ώστε να υλοποιηθεί. Αρχικά, πραγματοποιείται η εξόρυξη των δεδομένων, όπου κατά την ανάλυση του κειμένου χρησιμοποιούνται συχνά εργαλεία απόξεσης ιστού (web scraping) ή διεπαφές προγραμματισμού εφαρμογών (APIs) για την απόκτηση των δεδομένων. Στην συνέχεια γίνεται καθαρισμός και επεξεργασία των δεδομένων, όπου αφαιρούνται τυχόν μέρη που δεν συνδέονται άμεσα με σχετικό με το συναίσθημα του κειμένου όπως άρθρα, ειδικοί χαρακτήρες και σημεία στίξης. Η διαδικασία αυτή ονομάζεται τυποποίηση. Έπειτα, πραγματοποιείται η εξαγωγή χαρακτηριστικών, όπου ένας αλγόριθμος μηχανικής μάθησης αναλύει το κείμενο και εξάγει χαρακτηριστικά για να εντοπίσει θετικά ή αρνητικά συναισθήματα. Στη συνέχεια, οι αναλυτές δεδομένων επιλέγουν το κατάλληλο μοντέλο μηχανικής μάθησης. Το εργαλείο ανάλυσης συναισθήματος χρησιμοποιεί το μοντέλο για να αξιολογήσει το κείμενο, είτε μέσω κανόνων, είτε αυτόματα, είτε με υβριδικό σύστημα. Στο τελικό βήμα, το κείμενο ταξινομείται σύμφωνα με το συναίσθημα που εκφράζει, και αποδίδεται μια βαθμολογία συναισθήματος. Η συγκεκριμένη μέθοδος χρησιμοποιείται για την ανάλυση διαφόρων τμημάτων του κειμένου, όπως ένα ολόκληρο έγγραφο ή μια μόνο πρόταση. Η ανάλυση συναισθήματος μπορεί να πραγματοποιηθεί με δυο τρόπους οι οποίοι είναι : η μηχανική μάθηση και η βασισμένη στο λεξικό. Η μέθοδος μηχανικής εκμάθησης χρησιμοποιεί δεδομένα με ανθρώπινες ετικέτες για την εκπαίδευση ενός ταξινομητή κειμένου, καθιστώντας την εποπτευόμενη μέθοδο εκμάθησης. Από την άλλη, η προσέγγιση που βασίζεται στο λεξικό αναλύει μια πρόταση σε λέξεις και αποδίδει βαθμολογία στον σημασιολογικό προσανατολισμό κάθε λέξης, χρησιμοποιώντας ένα λεξικό αναφοράς. Τέλος, αθροίζει τις διάφορες βαθμολογίες για να βρεθεί ένα συμπέρασμα [3].

2.2 Σημασία και Χρησιμότητα της Ανάλυσης Συναισθήματος

Οι κριτικές των καταναλωτών σε πλατφόρμα γνώμης και μέσα κοινωνικής δικτύωσης μπορούν να έχουν τόσο θετικό όσο και αρνητικό αντίκτυπο σε μια επιχείρηση. Γι’ αυτό, η παρακολούθηση των διαδικτυακών αξιολογήσεων των πελατών είναι εξαιρετικά απαραίτητη. Τα εργαλεία ανάλυσης συναισθήματος στο διαδίκτυο απλοποιούν τη διαδικασία παρακολούθησης της αντίληψης του κοινού για κάθε επωνυμία. Έτσι, μπορούν οι επιχειρήσεις να εντοπίζουν ποιες πηγές επηρεάζουν περισσότερο την εικόνα

της επιχείρησης είτε θετικά είτε αρνητικά. Η ανάλυση συναισθήματος ορίζεται ως ένα απαραίτητο εργαλείο επιχειρηματικής ευφυΐας που διευκολύνει τις επιχειρήσεις να αναβαθμίσουν τα προϊόντα και τις υπηρεσίες τους. Παρακάτω παρουσιάζονται κάποια οφέλη από την ανάλυση συναισθήματος. Αρχικά παρέχονται αντικειμενικές γνώσεις. Οι επιχειρήσεις μπορούν να μειώσουν την ανθρώπινη μεροληψία στην αξιολόγηση χρησιμοποιώντας εργαλεία ανάλυσης συναισθήματος με Τεχνητή Νοημοσύνη. Έτσι, διασφαλίζουν ότι η ανάλυση των απόψεων των πελατών γίνεται με συνέπεια και αντικειμενικότητα. Οι υπεύθυνοι μάρκετινγκ ενδέχεται να αγνοήσουν τις αρνητικές αξιολογήσεις και να έχουν θετική προκατάληψη σχετικά με την απόδοση του επεξεργαστή. Ωστόσο, τα ακριβή εργαλεία ανάλυσης συναισθήματος αναλύουν και ταξινομούν το κείμενο, καταγράφοντας τα συναισθήματα με αντικειμενικό τρόπο. Επιπλέον ένα σύστημα ανάλυσης συναισθήματος ενισχύει τις εταιρείες να τροποποιήσουν τα προϊόντα και τις υπηρεσίες τους με βάση γνήσια και συγκεκριμένες κριτικές των πελατών [4]. Οι επιχειρήσεις αντλούν διαρκώς πληροφορίες από έναν τεράστιο όγκο μη δομημένων δεδομένων, όπως email, μεταγραφές chatbot, έρευνες, αρχεία διαχείρισης πελατών και σχόλια προϊόντων. Τα cloud εργαλεία ανάλυσης συναισθήματος δίνουν τη δυνατότητα στις επιχειρήσεις να επεκτείνουν τη διαδικασία ανίχνευσης των συναισθημάτων των πελατών σε δεδομένα κειμένου με οικονομικά προσιτό τρόπο. Τέλος ,οι επιχειρήσεις πρέπει να ανταποκρίνονται άμεσα σε πιθανές τάσεις της αγοράς στο σημερινό τοπίο που αλλάζει πολύ γρήγορα. Οι έμποροι χρησιμοποιούν λογισμικό ανάλυσης συναισθήματος για να κατανοήσουν σε πραγματικό χρόνο τα συναισθήματα των πελατών σχετικά με την επωνυμία, τα προϊόντα και τις υπηρεσίες της εταιρείας, λαμβάνοντας άμεσες δράσεις βάσει των ευρημάτων τους. Μπορούν, επίσης, να προσαρμόσουν το λογισμικό ώστε να αποστέλλει ειδοποιήσεις όταν ανιχνεύονται αρνητικά συναισθήματα για συγκεκριμένες λέξεις-κλειδιά [5].

2.3 Εφαρμογές και Εργαλεία της Ανάλυσης Συναισθήματος

2.3.1 Εφαρμογές της Ανάλυσης Συναισθήματος

Η ανάλυση συναισθήματος αποτελεί έναν από τους σημαντικότερους τομείς της Επεξεργασίας Φυσικής Γλώσσας, με εφαρμογές σε διάφορους κλάδους, όπως τα μέσα κοινωνικής δικτύωσης, η εξυπηρέτηση πελατών, η πολιτική ανάλυση και η ψυχολογία. Υπάρχουν πολλές εφαρμογές και εργαλεία που χρησιμοποιούνται για την ανάλυση

συναισθήματος, τα οποία βασίζονται είτε σε λεξικογραφικές μεθόδους, είτε σε τεχνικές μηχανικής μάθησης, είτε σε υβριδικές προσεγγίσεις. Στη σύγχρονη εποχή, οι επιχειρήσεις καλούνται να διαχειριστούν μεγάλους όγκους δεδομένων, όπως σχόλια στα social media, κριτικές και emails. Η ανάλυση συναισθήματος έχει καταστεί απαραίτητη, ιδίως στον τομέα του marketing, βοηθώντας στον εντοπισμό των συναισθημάτων των καταναλωτών, τη βελτίωση της φήμης των διαφημίσεων και τη λήψη στρατηγικών αποφάσεων [6]. Μέσω αυτής της μεθόδου, οι εταιρείες μπορούν να κατανοούν βαθύτερα τους πελάτες τους, να ενισχύουν την αφοσίωσή τους και να ανταποκρίνονται άμεσα σε κρίσιμες καταστάσεις. Επιπλέον, εξοικονομούν χρόνο και πόρους, ενώ αξιοποιούν τις διαθέσιμες πληροφορίες από διάφορες πηγές, όπως email, tweets, έρευνες και συνομιλίες εξυπηρέτησης. Τα συστήματα ανάλυσης συναισθήματος επιτρέπουν στις επιχειρήσεις να μετατρέπουν τα δεδομένα σε χρήσιμες πληροφορίες, βελτιώνοντας τόσο τις εμπειρίες των πελατών όσο και την ανταγωνιστικότητά τους [7].

2.3.2 Εργαλεία της Ανάλυσης Συνασθήματος

1. SentiStrength

Το SentiStrength είναι ένα εργαλείο που χρησιμοποιείται κυρίως για την ανάλυση του συναισθήματος σε κοινωνικά δίκτυα, όπως το Twitter και τα σχόλια στο YouTube. Πιο συγκεκριμένα, είναι ένα πρόγραμμα που συγκρίνει το κείμενο των μέσων κοινωνικής δικτύωσης με έναν ταξινομητή συναισθημάτων που βασίζεται σε λεξικό. Το SentiStrength μετρά τη δύναμη του συναισθήματος εκχωρώντας βαθμολογίες που κυμαίνονται από -5 έως +5. Οι θετικοί αριθμοί δηλώνουν θετικές στάσεις ενώ οι αρνητικοί αριθμοί δηλώνουν αρνητικές στάσεις [8].

2. Sentiment140

Το Sentiment140 είναι ένα εργαλείο που αναλύει συναισθήματα από tweets, επιτρέποντας στους ερευνητές και τους επιχειρηματίες να κατανοήσουν τη στάση των χρηστών απέναντι σε ένα θέμα, ένα προϊόν ή ένα εμπορικό σήμα. Χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την ταξινόμηση των συναισθημάτων σε θετικά και αρνητικά επιτρέποντας σημαντικές προόδους στον τομέα της επεξεργασίας φυσικής γλώσσας [9].

3. LIWC (Linguistic Inquiry and Word Count)

Το LIWC είναι ένα εργαλείο που χρησιμοποιείται στην ψυχολογία και την κοινωνική επιστήμη για την ανάλυση συναισθήματος σε κείμενα. Το LIWC συνοδεύεται από

περισσότερα από 100 ενσωματωμένα λεξικά που έχουν δημιουργηθεί για να καταγράφουν τις κοινωνικές και ψυχολογικές καταστάσεις των ανθρώπων. Κάθε λεξικό περιέχει από μια λίστα λέξεων και άλλες συγκεκριμένες λεκτικές κατασκευές που έχουν αναγνωριστεί ως αντιπροσωπευτικά μιας συγκεκριμένης ψυχολογικής κατηγορίας ενδιαφέροντος.

4. SenticNet

Το SenticNet παρέχει ένα σύνολο σημασιολογικών, συναισθηματικών και πολικότητας χαρακτηριστικών που συνδέονται με έννοιες της φυσικής γλώσσας. Ειδικά, η σημασιολογία δηλώνει τις δηλωτικές πληροφορίες που σχετίζονται με λέξεις και εκφράσεις πολλαπλών λέξεων όπου οι αισθητικοί ορίζουν τις υποδηλωτικές πληροφορίες που σχετίζονται με έννοιες φυσικής γλώσσας (δηλαδή, τιμές κατηγοριοποίησης συναισθημάτων) και η πολικότητα είναι κυμαινόμενος αριθμός μεταξύ -1 και +1[10].

5. VADER (Valence Aware Dictionary and sEntiment Reasoner)

Το VADER είναι προσαρμοσμένο για να εντοπίζει το συναίσθημα από σύντομα κομμάτια κειμένου, όπως tweets, κριτικές προϊόντων ή οποιοδήποτε περιεχόμενο που δημιουργείται από χρήστες που μπορεί να περιέχει αργκό και συντομογραφίες. Χρησιμοποιεί ένα προκατασκευασμένο λεξικό λέξεων που σχετίζονται με τις συναισθηματικές αξίες και χρησιμοποιεί ένα σύνολο κανόνων για τον υπολογισμό των βαθμολογιών συναισθήματος [11]. Η ανάλυση συναισθήματος αποτελεί ένα ισχυρό εργαλείο για την αντίληψη των ανθρώπινων συναισθημάτων σε κείμενα, με πληθώρα εφαρμογών στην έρευνα και την επιχειρηματικότητα. Η επιλογή της ιδανικής μεθόδου καθορίζεται από τον τύπο των δεδομένων και την απαιτούμενη ακρίβεια. Τα σύγχρονα εργαλεία, όπως το SentiStrength, το SenticNet και το VADER, επιτρέπουν την αυτοματοποιημένη εξόρυξη συναισθημάτων, ενώ οι εξελιγμένες τεχνικές μηχανικής μάθησης οδηγούν στην ανάλυση και ερμηνεία συναισθημάτων με μεγαλύτερη ακρίβεια [12].

2.4 Πλεονεκτήματα και Μειονεκτήματα της Ανάλυσης Συναισθήματος

2.4.1 Πλεονεκτήματα της Ανάλυσης Συναισθήματος

Τα οφέλη της ανάλυσης συναισθήματος είναι διαδοχικά, καθώς παρέχουν μια σαφή εικόνα των μεταβαλλόμενων τάσεων της αγοράς και των προτιμήσεων των πελατών, ανεξάρτητα από τον κλάδο στον οποίο δραστηριοποιείται κάποιος. Η ανάλυση συναισθημάτων από δεδομένα εμπειρίας κοινού που προέρχονται από διάφορες πηγές, όπως μέσα κοινωνικής δικτύωσης, ιστότοποι κριτικών, ειδησεογραφικά άρθρα και έρευνες, προσφέρει πολύτιμες πληροφορίες για τη διαμόρφωση μιας αποτελεσματικής στρατηγικής ανάπτυξης, η οποία είναι καθοριστική για τη βιωσιμότητα της επιχείρησης. Ας εξετάσουμε, λοιπόν, μερικά από τα σημαντικότερα πλεονεκτήματα της ανάλυσης συναισθήματος. Αρχικά, εφαρμόζεται τακτοποίηση δεδομένων σε κλίμακα. Πιο συγκεκριμένα, η ανάλυση συναισθήματος συμβάλλει ώστε οι οργανισμοί να χειρίζονται τεράστιες μετρήσεις αδόμητων πληροφοριών με παραγωγικό και έξυπνο τρόπο, καθώς υπάρχουν απλώς πολλές επιχειρηματικές πληροφορίες για φυσική επεξεργασία. Είναι σημαντικό να αναφερθεί πως η ανάλυση συναισθήματος μπορεί να εντοπίσει σταδιακά τα κύρια προβλήματα, για παράδειγμα, είναι μια έκτακτη ανάγκη δημοσίων σχέσεων μέσω της εικονικής ψυχαγωγίας [13]. Επίσης, η ανάλυση συναισθήματος δίνει την δυνατότητα στις επιχειρήσεις να ανιχνεύουν τα αρνητικά συναισθήματα που εκφράζονται από τους πελάτες και να τα διαχειρίζονται άμεσα. Μέσω της αποτελεσματικής επίλυσης των προβλημάτων των πελατών, οι επιχειρήσεις έχουν τη δυνατότητα να εξελίξουν τα συνολικά επίπεδα ικανοποίησης και να διασφαλίσουν την αξιοπιστία των πελατών τους. Γενικά όταν οι επιχειρήσεις λαμβάνουν ενεργά τα συναισθήματα των πελατών τους και προχωρούν σε κατάλληλες ενέργειες, ενδυναμώνεται η εμπιστοσύνη και η αφοσίωση της πελατειακής τους βάσης. Αυτό μπορεί να οδηγήσει σε επαναλαμβανόμενες συναλλαγές και σε θετικές συστάσεις, ενισχύοντας τη φήμη και τη βιωσιμότητα της επιχείρησης. Οι επιχειρήσεις όταν χρησιμοποιούν την ανάλυση συναισθήματος αποκτούν ανταγωνιστικό πλεονέκτημα μένοντας μπροστά από τις τάσεις της αγοράς και τις προτιμήσεις των πελατών σε σχέση με άλλες επιχειρήσεις. Αυτό τους δίνει την ευκαιρία να προσαρμόζουν τις στρατηγικές και τις προσφορές τους για να ανταποκρίνονται στις εξελισσόμενες απαιτήσεις σε πραγματικό χρόνο. Τέλος, η ανάλυση συναισθήματος μπορεί να βοηθήσει στην ανίχνευση ύποπτων δραστηριοτήτων ή αρνητικών συναισθημάτων που συνδέονται με

δόλιες ενέργειες. Εντοπίζοντας έγκαιρα ενδείξεις πιθανής απάτης, οι επιχειρήσεις έχουν τη δυνατότητα να περιορίσουν τις οικονομικές απώλειες και να προστατεύσουν τη φήμη τους, διασφαλίζοντας έτσι την αξιοπιστία τους στην αγορά [14].

2.4.2 Μειονεκτήματα της Ανάλυσης Συναισθήματος

Παρά τα θετικά στοιχεία που αναφέρθηκαν, υπάρχουν και κάποια αρνητικά σημεία που πρέπει να ληφθούν υπόψη. Είναι γεγονός ότι οι άνθρωποι εκφράζουν αρνητικά συναισθήματα χρησιμοποιώντας θετικές λέξεις, γεγονός που καθιστά τον σαρκασμό ικανό να παραπλανήσει εύκολα τα μοντέλα ανάλυσης συναισθημάτων, εκτός αν είναι ειδικά σχεδιασμένα για να τον αναγνωρίζουν. Ο σαρκασμός συχνά εμφανίζεται σε περιεχόμενο που δημιουργούν οι χρήστες, όπως σχόλια στο Facebook και tweets. Η ανίχνευση του σαρκασμού στην ανάλυση συναισθημάτων είναι εξαιρετικά δύσκολη χωρίς μια βαθιά κατανόηση του πλαισίου, του συγκεκριμένου θέματος και του περιβάλλοντος. Αυτή η πρόκληση καθιστά δύσκολη την ανάλυση, όχι μόνο για τις μηχανές, αλλά και για τους ανθρώπους. Η συνεχής παραλλαγή των λέξεων που χρησιμοποιούνται σε σαρκαστικές προτάσεις καθιστά την εκπαίδευση μοντέλων ανάλυσης συναισθημάτων ακόμη πιο απαιτητική [15]. Μία από τις βασικές προκλήσεις στην ανάλυση συναισθημάτων είναι η κατανόηση του σαρκασμού, της ειρωνείας ή του χιούμορ. Επιπλέον, οι πολιτιστικές αναφορές και η εξειδικευμένη ορολογία ενός τομέα μπορούν να προκαλέσουν παρερμηνείες. Οι ερευνητές και οι προγραμματιστές συνεχώς εξελίσσουν αλγόριθμους για να βελτιώσουν την κατανόηση του πλαισίου, μειώνοντας έτσι τα περιθώρια λάθους στην ανάλυση. Μαζί με τα παραπάνω, ένα ακόμη ζήτημα που προκύπτει είναι ότι η ακρίβεια της ανάλυσης συναισθήματος βασίζεται σε μεγάλο βαθμό στην ποιότητα των δεδομένων στα οποία έχει εκπαιδευτεί. Τα αποτελέσματα ενδέχεται να παραμορφωθούν εάν τα δεδομένα εκπαίδευσης είναι μεροληπτικά ή ελλιπή. Παρόλο που η ανάλυση συναισθήματος προσφέρει αμερόληπτα αποτελέσματα, καθώς δεν συμμετέχουν άνθρωποι στην ανάλυση, μπορεί να εμφανιστεί προκατάληψη αν το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του αλγορίθμου είναι ήδη προκατειλημμένο [16]. Ένα τελευταίο, αλλά εξαιρετικά σημαντικό μειονέκτημα είναι πως συνήθως, το συναίσθημα κατηγοριοποιείται σε θετικό, αρνητικό και ουδέτερο. Στο τυπικό μοντέλο ανάλυσης συναισθήματος, η έξοδος για ένα ολόκληρο κείμενο εκφράζεται συνήθως ως ένας αριθμός μεταξύ -1 (αρνητικό) και 1 (θετικό), με το 0 να δηλώνει ουδέτερο ή καθόλου συναίσθημα. Αυτή η προσέγγιση δεν λαμβάνει υπόψη τις αποχρώσεις του συναισθήματος. Για παράδειγμα, η κατανόηση της διαφοράς μεταξύ του

«δεν μου αρέσει» και του «μισώ» είναι σημαντική, αλλά σε ένα απλό μοντέλο ανάλυσης συναισθήματος, και τα δύο θα μπορούσαν να αποδοθούν με την ίδια βαθμολογία -1. Η ακριβής μέτρηση των αποχρώσεων του συναισθήματος είναι δύσκολη, ειδικά επειδή οι αναλυτές δεν συμφωνούν πάντα μεταξύ τους. Μοντέλα με περισσότερες αποχρώσεις (όπως 5 ή 7, π.χ. πολύ αρνητικό, αρνητικό, ουδέτερο, θετικό, πολύ θετικό) προσφέρουν περισσότερες ευκαιρίες για υποκειμενικές διαφορές μεταξύ των σχολιαστών σε σύγκριση με τα μοντέλα που χρησιμοποιούν μόνο 3 αποχρώσεις. Η κατανόηση της κλίμακας συναισθημάτων και της ακρίβειας της μέτρησης είναι σημαντική για να αξιολογήσουμε εάν μια διαφορά στο συναίσθημα είναι ουσιαστική ή όχι. Εκτός από τον αριθμό των αποχρώσεων, είναι επίσης σημαντικό να γίνει διάκριση μεταξύ «ουδέτερου» και «χωρίς συναίσθημα». Για παράδειγμα, η πρόταση «Το προϊόν είναι εντάξει» εκφράζει συναισθηματική κρίση, ενώ η πρόταση «Αγόρασα το προϊόν» δεν εκφράζει συναίσθημα. Ένα απλό μοντέλο ανάλυσης συναισθήματος θα επιστρέψει μια βαθμολογία 0 και στις δύο προτάσεις, ενώ ένα πιο εξελιγμένο μοντέλο θα αναγνωρίσει τη διαφορά [17].

3 Θεματική Μοντελοποίηση (Topic Modeling) και Εξόρυξη Θεμάτων

3.1 Εισαγωγή στην έννοια του Topic Modeling

Η ανάπτυξη του Topic Modeling ξεκίνησε στις αρχές της δεκαετίας του 1980 και αναπτύχθηκε σε επιμέρους τομείς από το αντικείμενο μελέτης της «γενετικής πιθανοτικής μοντελοποίησης». Η συγκεκριμένη μέθοδος βασίζεται στην υπόθεση ότι οι παρατηρούμενες μεταβλητές αλληλεπιδρούν με μη παρατηρούμενες ή λανθάνουσες παραμέτρους, σχηματίζοντας έτσι μια πιθανοτική σχέση που οδηγεί στη δημιουργία των δεδομένων εντός ενός συνόλου δεδομένων. Η ανάπτυξη αυτών των διαδικασιών αναδύθηκε από την ανάγκη συνοπτικής περιγραφής πληροφοριών μέσα σε συνεχώς αυξανόμενες και μεγάλες συλλογές δεδομένων χωρίς να διακυβεύονται οι στατιστικές σχέσεις που απαιτούνται για την περάτωση πιο απλών αναλύσεων. Η πρώτη μέθοδος που εφαρμόστηκε για το έργο αυτό προσδιορίστηκε ως σχέδιο μείωσης tf-idf και προτάθηκε από δύο ερευνητές στον κλάδο της ανάκτησης πληροφορίας, τους Salton και McGill, το 1983. Σε αυτή τη τεχνική, κάθε μεμονωμένο έγγραφο σε ένα σύνολο κειμένων θεωρείται

από το λεξιλόγιό του και ο αριθμός των εμφανίσεων κάθε λέξης υπολογίζεται ώστε να προκύψει μια τιμή συχνότητας εμφάνισης (term frequency - tf) για τη συγκεκριμένη λέξη μέσα στο εκάστοτε έγγραφο. Επιπλέον, υπολογίζεται το σύνολο των εμφανίσεων μιας λέξης σε ολόκληρο το σύνολο το οποίο ονομάζεται αντίστροφος αριθμός συχνότητας εγγράφων (inverse document frequency- idf). Αφού προηγηθεί η κανονικοποίηση των τιμών στο σύνολο δεδομένων, εφόσον απαιτείται, οι δύο τιμές συγκρίνονται, δημιουργώντας έναν πίνακα όρων-εγγράφων, στον οποίο περιλαμβάνονται οι τιμές tf-idf για κάθε όρο σε όλα τα έγγραφα ανά στήλη. Η διαδικασία αυτή, αν και αποτελεσματική στην εύρεση συνόλων λέξεων που διακρίνουν τα έγγραφα, παρουσίασε περιορισμούς διότι η μείωση του μήκους της περιγραφής ήταν σχετικά ασήμαντη και απέδωσε λίγες σημαντικές πληροφορίες σχετικά με τις στατιστικές σχέσεις εντός ή μεταξύ των εγγράφων. Για το λόγο αυτό, προτάθηκε μια άλλη μέθοδος, η λανθάνουσα σημασιολογική ευρετηρίαση (LSI/LSA), η οποία παρουσιάστηκε από τους Deerwester et al. (μια ομάδα ερευνητών) το 1990 και αποσκούσε στην αποτελεσματικότερη μείωση των διαστάσεων και στην επίτευξη μεγαλύτερης συμπίεσης δεδομένων.Σε συνέχεια, το 1999 προτάθηκε από τον Hofman το πιθανοτικό μοντέλο LSI ή LSA (pLSI/pLSA). Βασικός σκοπός του ήταν να αποτελεί μια ακόμα πιο άμεση μέθοδο ανάλυσης που θα μπορούσε να εφαρμοστεί σχηματίζοντας ένα παραγωγικό μοντέλο και προσαρμόζοντάς το χρησιμοποιώντας πιθανοτικές μεθόδους [46]. Ωστόσο, το 2003, με την εισαγωγή του Latent Dirichlet Allocation (LDA) από τους Blei, Ng και Jordan, ο τομέας του Topic Modeling πήρε μια σημαντική στροφή. Το LDA επαναστατεί στο πεδίο της ανάλυσης θεμάτων, εστιάζοντας στην πιθανοτική αναπαράσταση των θεμάτων. Στο μοντέλο αυτό, κάθε έγγραφο θεωρείται ως συνδυασμός θεμάτων, με κάθε θέμα να αποτελεί συνδυασμό λέξεων, προσφέροντας έτσι μια πιο φυσική και βελτιωμένη αναπαράσταση της σχέσης λέξεων και θεμάτων στα κείμενα. Η μέθοδος αυτή βελτίωσε την ακρίβεια των αναλύσεων και τη δυνατότητα ανίχνευσης πιο σύνθετων και αφηρημένων σχέσεων μεταξύ των εγγράφων. Η εξέλιξη του Topic Modeling έχει εφαρμοστεί σε πολλούς τομείς, όπως στην ανάλυση συναισθήματος, την ανάκτηση πληροφορίας και την κατηγοριοποίηση κειμένων. Η δυνατότητα του LDA και άλλων μεθόδων του Topic Modeling να εξάγουν λανθάνοντα θέματα που συμπεριέχονται σε μεγάλα σύνολα δεδομένων καθιστά τη μέθοδο εξαιρετικά χρήσιμη στην ανάλυση μεγάλων όγκων κειμένων, κάτι που είναι αναγκαίο για την επεξεργασία δεδομένων από κοινωνικά δίκτυα, επιστημονικά έγγραφα και άλλες μεγάλες συλλογές πληροφορίας. Επιπλέον, με την ανάπτυξη πιο προηγμένων τεχνικών και εργαλείων, το Topic Modeling συνεχίζει να προσφέρει δυναμικές λύσεις

και να βελτιώνει τη δυνατότητα της μηχανικής μάθησης να κατανοεί τη δομή του λόγου και τις σχέσεις μέσα σε μεγάλες ποσότητες δεδομένων [18]. Η μοντελοποίηση θεμάτων είναι ένα δημοφιλές αναλυτικό εργαλείο για την αξιολόγηση δεδομένων. Η μοντελοποίηση θεμάτων είναι μια τεχνική εκμάθησης συστήματος που ανακαλύπτει τα κύρια θέματα που αντιπροσωπεύουν έναν μεγάλο όγκο συλλογής εγγράφων. Βασικός σκοπός της θεματικής μοντελοποίησης είναι ο εντοπισμός των κρυμμένων σημασιολογικά συστημάτων μέσα σε «γεγονότα» κειμενικού περιεχομένου, επιτρέποντας στους χρήστες να κατανοήσουν και να συνοψίσουν τα δεδομένα με τρόπο άμεσο και εύκολο. Ένα «θέμα» ή «topic» ορίζεται ως ένα επαναλαμβανόμενο μοτίβο λέξεων που αντιπροσωπεύει με περισσότερη ακρίβεια ένα θέμα μέσα στο σύνολο κειμένου. Τα μοντέλα θεμάτων είναι αλγόριθμοι που σαρώνουν τη συλλογή εγγράφων για να αναγνωρίσουν αυτά τα θέματα [19]. Στην ανάλυση δεδομένων ενός κειμένου, παρατηρείται συχνά ο προσδιορισμός των γεγονότων ή των εννοιών που υπάρχει σε ένα έγγραφο. Αυτές οι πληροφορίες είναι σαφείς σε έναν άνθρωπο που διαβάζει ένα κείμενο, όμως σε ένα πρόγραμμα δίνεται μόνο το κείμενο όπως είναι γραμμένο και όχι το θέμα κάθε εγγράφου. Προκειμένου να ολοκληρωθεί αυτό το έργο σε ένα πρόγραμμα, οι επιστήμονες δεδομένων χρησιμοποιούν την μέθοδο της μοντελοποίησης θεμάτων. Η μοντελοποίηση θεμάτων αποτελεί στατιστικό εργαλείο για την εξαγωγή λανθανουσών μεταβλητών από μεγάλα σύνολα δεδομένων. Θεωρείται κατάλληλη τόσο για χρήση σε δεδομένα κειμένου αλλά όχι μόνο, καθώς έχει επίσης χρησιμοποιηθεί για την ανάλυση δεδομένων βιοπληροφορικής, κοινωνικών δεδομένων και περιβαλλοντικών δεδομένων. Όλες οι αναλύσεις είναι ικανές να βοηθήσουν στην οργάνωση συνόλων δεδομένων για πιο εύκολη πρόσβαση. Παρά τη δημοτικότητά της, η μοντελοποίηση θεμάτων είναι υποκείμενη σε σημαντικά ζητήματα βελτιστοποίησης τα οποία είναι πιθανό να οδηγήσουν σε αναξιόπιστα αποτελέσματα. Ορισμένες τεχνικές δεν αντικατοπτρίζουν σωστά τις πραγματικές σχέσεις στα δεδομένα. Αυτό συμβαίνει λόγω υποθέσεων που γίνονται για βασικές παραμέτρους στη διαδικασία υπολογισμού και λόγω της αναποτελεσματικότητας πολλών μεθόδων βελτιστοποίησης, οι οποίες προσπαθούν να ξεπεράσουν την αβεβαιότητα εκτελώντας πολλές χρονοβόρες επαναλήψεις για να βρουν την καλύτερη τιμή για μια παράμετρο. Υπάρχουν επίσης λίγες ακαδημαϊκές μελέτες που αναλύουν τις διάφορες μεθόδους μοντελοποίησης θεμάτων και τις στρατηγικές που μπορούν να εφαρμοστούν για τη βελτιστοποίηση αυτής της διαδικασίας σε ένα σύνολο δεδομένων. Η θεματική μοντελοποίηση είναι μια πολλά υποσχόμενη μέθοδος ανάλυσης κειμένου που ενδιαφέρει ένα ευρύ φάσμα μελετητών σε ποικίλες επιστήμες. Οι

αλγόριθμοι είναι ικανοί να το επιτύχουν με ελάχιστη ανθρώπινη παρέμβαση, καθιστώντας τη μέθοδο πιο επαγωγική σε σύγκριση με τις συνηθισμένες προσεγγίσεις στην ανάλυση κειμένου. Στον ευρύτερο τομέα της επιστήμης των δεδομένων, η μοντελοποίηση θεμάτων αποτελεί παράδειγμα πιθανοτικής μοντελοποίησης. Μεταξύ των αποτελεσμάτων αυτής της διαδικασίας, προκύπτει ένα σύνολο κατανομών λέξεων ανά θέμα που αντιστοιχούν σε πιθανότητες για κάθε συνδυασμό θέματος-λέξης, καθώς και μια αντίστοιχη κατανομή εγγράφων ανά θέμα, που περιγράφει την πιθανότητα να επιλεγεί ένα συγκεκριμένο θέμα για κάθε έγγραφο στο σύνολο. Ωστόσο, είναι σημαντικό να αναφερθεί ότι η δομή που προκύπτει είναι λανθάνουσα, πράγμα που σημαίνει ότι οι κατανομές θεμάτων ανά λέξη δεν συνδέονται με μια ξεκάθαρη ονομασία για το θέμα [20]. Συμπερασματικά, η μοντελοποίηση θεμάτων είναι μια μεθοδολογία ανάλυσης μεγάλων δεδομένων που χρησιμοποιείται για την ανακάλυψη αφηρημένων θεμάτων που εμφανίζονται επανειλημμένα σε μια σειρά εγγράφων. Κατά τη συγγραφή ενός άρθρου, ο συγγραφέας έχει κατά νου συγκεκριμένες λέξεις-κλειδιά, οι οποίες επαναλαμβάνονται καθ' όλη τη διάρκεια του κειμένου. Αυτές οι λέξεις-κλειδιά αναπαρίστανται ως ένα πεπερασμένο μείγμα βασισμένο σε ένα υποκείμενο σύνολο πιθανοτήτων θεμάτων, και στη συνέχεια επιστρέφεται ένα λανθάνον θέμα για κάθε συγκεκριμένο έγγραφο [21].

3.2 Μέθοδοι εξόρυξης θεμάτων: LDA & NMF

Τα θεματικά μοντέλα λειτουργούν ως γέφυρα μεταξύ διαφορετικών τομέων ανάλυσης, συνδυάζοντας επιστήμες, δομημένες και μη δομημένες προσεγγίσεις και μεθόδους συλλογισμού για την επεξεργασία μεγάλων δεδομένων. Τα μέσα κοινωνικής δικτύωσης προσφέρουν νέες δυνατότητες για την έρευνα, ειδικά στην αλληλεπίδραση ανθρώπινων σχέσεων και τεχνολογίας. Το περιεχόμενο που δημιουργείται από χρήστες αναγνωρίζεται για την αξία του, καθώς προσφέρει πλούσιες και υποκειμενικές πληροφορίες που επιτρέπουν βαθύτερη υπολογιστική ανάλυση. Έρευνες έχουν εξετάσει την κοινωνική δυναμική αθλητικών εκδηλώσεων μέσω Facebook, την πολιτισμική ποικιλία μέσω Instagram και τις κοινωνικές αντιδράσεις στην πανδημία μέσω Twitter. Αυτές οι προσεγγίσεις ανοίγουν νέες προοπτικές και ενισχύουν την κατανόηση των κοινωνικών φαινομένων, ανανεώνοντας τις γνώσεις σε πολλούς τομείς [22]. Στη συγκεκριμένη μελέτη γίνεται χρήση της μεθόδου LDA, η οποία είναι από τις πιο συχνά χρησιμοποιούμενες μεθόδους μοντελοποίησης. Γενικότερα, η τεχνική αυτή δομεί τα δεδομένα σε τρία επίπεδα, λέξη, θέμα και έγγραφο. Η προσέγγιση της LDA επέδειξε

ιδιαίτερη βελτίωση σε σχέση με προηγούμενες μεθόδους όσον αφορά την εξέταση της εξαγωγής συμπερασμάτων από έγγραφα και την κατανόηση των μη δομημένων δεδομένων. Παρόλα αυτά, το πρόβλημα είναι ότι τα τελικά αποτελέσματα μπορεί να διαφέρουν σημαντικά για τα ίδια δεδομένα χρησιμοποιώντας τον ίδιο αλγόριθμο, γεγονός που οδηγεί σε σημαντικά προβλήματα. Γι' αυτό έχουν προταθεί διάφορα παράγωγα της LDA για προσδιορισμένους ρόλους και επιδιορθώσεις. Μερικά παραδείγματα αυτών των μοντέλων είναι το LDMM, το fLDA και το hLDA. Κάθε μία από αυτές τις μεθόδους συνδυάζει τα βασικά χαρακτηριστικά της LDA με μια άλλη πιθανοτική ή στατιστική διαδικασία, προκειμένου να βελτιώσει την αποτελεσματικότητα και να ενσωματώσει την εξαγωγή συμπερασμάτων ή διατηρώντας ταυτόχρονα την αρχική προσέγγιση LDA [23]. Σε αντίθεση με το LDA, το NMF (Non-negative Matrix Factorization) είναι ένας αποσυνθετικός, μη πιθανοτικός αλγόριθμος ο οποίος που αναπτύχθηκε το 1999 από τους Lee και Seung. και χρησιμοποιεί factorization matrix ενώ ανήκει στην ομάδα των γραμμικών-αλγεβρικών αλγορίθμων (Egger, 2022). Δηλαδή, ανακαλύπτει τις κρυφές δομές ή τα μοτίβα μέσα σε δεδομένα. Αυτή η μέθοδος έχει αποδειχθεί ότι λειτουργεί αποτελεσματικά για την εξαγωγή θεμάτων από σύνολα κειμένων. Το NMF είναι μη εποπτευόμενο, σε σύγκριση με το LDA που μπορεί να είναι είτε εποπτευόμενο είτε μη εποπτευόμενο [24]. Η λειτουργία του NMF είναι η ανάλυση ενός δεδομένου πίνακα VV σε δύο πίνακες W και H έτσι ώστε: $V \approx W \times H$

Όπου:

- V είναι ο αρχικός πίνακας δεδομένων.
- W είναι ένας πίνακας που αντιπροσωπεύει τις "συνιστώσες" ή τις "χαρακτηριστικές" του δεδομένου πίνακα.
- H είναι ο πίνακας των συντελεστών που δείχνει τη συσχέτιση κάθε συνιστώσας με τα δεδομένα.

Παρά τη σημαντική πρόοδο στη χρήση της μεθόδου NMF για τη μοντελοποίηση θεμάτων, παραμένουν συγκεκριμένες κρίσιμες προκλήσεις. Το NMF περιορίζεται στη διαχωρισιμότητα, δηλαδή συχνά βασίζεται στην υπόθεση ότι κάθε θέμα έχει μία τουλάχιστον "μοναδική" λέξη (λέξη-άγκυρα) που δεν εμφανίζεται σε άλλα θέματα. Αντίθετα, οι λέξεις είναι πολύσημες και μπορούν να χρησιμοποιούνται σε διαφορετικά συμφραζόμενα, άρα αυτή η υπόθεση δεν είναι ρεαλιστική. Επιπλέον, έχει μεγάλη απαίτηση σε δεδομένα, όπως η χρήση tensor factorization, που απαιτούν πολύ περισσότερα δεδομένα για να δώσουν αξιόπιστα αποτελέσματα, κάτι που δεν είναι πάντα

εφικτό ειδικά σε μικρότερες συλλογές κειμένων. Τέλος, είναι ένα μοντέλο ρηχής μάθησης καθώς περιλαμβάνει μόνο ένα επίπεδο αναπαράστασης. Αυτό σημαίνει πως δεν μπορεί να αποτυπώσει εύκολα τη μη γραμμική ή ιεραρχική φύση της γλώσσας (π.χ. έννοιες που περιέχουν υποέννοιες ή διαφορετικά επίπεδα σημασιολογίας). Από την άλλη, το LDA αποτελεί μία από τις πιο θεμελιώδεις μεθόδους μοντελοποίησης θεμάτων, προσφέροντας ένα σαφές πιθανοτικό πλαίσιο για την αναγνώριση κρυμμένων θεμάτων μέσα σε μεγάλα σύνολα κειμένων. Η βασική του δύναμη έγκειται στην δυνατότητα του να κατανέμει κάθε έγγραφο σε πολλαπλά θέματα και να αποδίδει πιθανότητες λέξεων σε κάθε θέμα, επιτρέποντας την αναλυτική και ερμηνεύσιμη εξερεύνηση του περιεχομένου. Παρότι η απόδοσή του επηρεάζεται από παραμέτρους που απαιτούν ρύθμιση, και η κατανομή Dirichlet δεν είναι πάντα η βέλτιστη προσέγγιση για όλα τα δεδομένα, το LDA παραμένει ένα αξιόπιστο σημείο αναφοράς. Η θεματική μοντελοποίηση μέσω LDA και NMF συνιστά δύο συμπληρωματικές αλλά διαφοροποιημένες προσεγγίσεις στη θεματική μοντελοποίηση, προσφέροντας πολύτιμη υποστήριξη σε αναλύσεις που απαιτούν εννοιολογική ερμηνεία και πληροφορική εξόρυξη γνώσης. Παρά τις μεθοδολογικές προκλήσεις, συνεχίζουν να αποτελούν αντικείμενο ερευνητικού ενδιαφέροντος και να εφαρμόζονται ευρέως σε πληθώρα επιστημονικών και τεχνολογικών πεδίων [25].

3.3 Πλεονεκτήματα και Μειονεκτήματα Topic Modeling

3.3.1 Πλεονεκτήματα Topic Modeling

Τα πλεονεκτήματα του Topic Modeling έχουν αναγνωριστεί ευρέως στη βιβλιογραφία, καθώς η τεχνική αυτή παρέχει αποδοτικούς τρόπους οργάνωσης και ερμηνείας μεγάλου όγκου κειμενικών δεδομένων. Στη συνέχεια, παρουσιάζονται τα βασικά οφέλη που προσφέρει η χρήση της. Αρχικά, οι σχετικές πληροφορίες μπορούν να ανακαταταχθούν και να οργανωθούν σε μεγάλη κλίμακα, καθώς η μοντελοποίηση θεμάτων επιτρέπει την αυτοματοποίηση της διαδικασίας εξόρυξης κειμένου. Αντί να απαιτείται χειροκίνητη επεξεργασία μεγάλου όγκου δεδομένων όπως αναρτήσεις σε μέσα κοινωνικής δικτύωσης ή μηνύματα εξυπηρέτησης πελατών, οι αλγόριθμοι Topic Modeling διευκολύνουν την ανάλυση και εντοπίζουν τα κυρίαρχα θέματα με αποδοτικό τρόπο. Επιπλέον, συνδυάζει τη μοντελοποίηση θεμάτων με άλλους τύπους NLP, όπως η ανάλυση συναισθημάτων. Το πιο σημαντικό είναι ότι οι χρήστες μπορούν να μετατρέψουν τη διορατικότητα σε δράση αφού όλη διαδικασία είναι αυτοματοποιημένη. Ένα ακόμη σημαντικό πλεονέκτημα είναι ότι δίνεται η δυνατότητα εξαγωγής νοήματος από μεγάλα σύνολα

δεδομένων, καθιστώντας εφικτή την κατανόηση των αλληλεπιδράσεων σε κάθε σημείο επαφής, από ερωτηματολόγια έως τα σχόλια των χρηστών. Οι πληροφορίες που προκύπτουν μπορούν να αξιοποιηθούν για τη βελτίωση της εμπειρίας των χρηστών και τη λήψη στοχευμένων αποφάσεων. Τελικά, η ανάλυση μοντελοποίησης θεμάτων αναλύει τεράστια τμήματα δεδομένων κειμένου με σκοπό να εξάγονται σχετικές πληροφορίες σε πραγματικό χρόνο. Η μοντελοποίηση θεμάτων βρίσκει ευρεία εφαρμογή σε ποικίλα επιχειρηματικά περιβάλλοντα, όπου απαιτείται η κατανόηση των μεγάλων ποσοτήτων μη δομημένων δεδομένων. Παρακάτω, παρουσιάζονται ορισμένα χαρακτηριστικά παραδείγματα επιχειρηματικής αξιοποίησης της τεχνικής. Συχνά, για παράδειγμα, ομάδες ανθρώπινου δυναμικού δαπανούν πολλές ώρες για τη συλλογή, ταξινόμηση και ερμηνεία μεγάλου όγκου δεδομένων, με στόχο τη μετατροπή τους σε αξιοποιήσιμη πληροφορία. Επιπλέον, αντί να απαιτείται η χειροκίνητη αναγνώριση του αρμόδιου τμήματος που πρέπει να απαντήσει σε ένα αίτημα, η μοντελοποίηση θεμάτων μπορεί να χρησιμοποιηθεί για την αυτόματη επισήμανση των συνομιλιών και, στη συνέχεια, με τη χρήση κατάλληλων ροών εργασίας, να γίνεται η προώθησή τους στην κατάλληλη ομάδα. Ενδεικτικά αναφέρεται μία συνομιλία που περιλαμβάνει λέξεις-κλειδιά όπως «τιμολόγηση», «συνδρομές», «ανανέωση» ή «προβλήματα χρέωσης», μπορεί να προωθηθεί αυτόματα στο τμήμα λογιστηρίου ή χρεώσεων για επίλυση. Εξάλλου, η ανάλυση κειμένου, σε συνδυασμό με τεχνικές ανάλυσης συναισθήματος, μπορεί να συμβάλει στον εντοπισμό της συναισθηματικής κατάστασης των πελατών και στην εκτίμηση του βαθμού επείγοντος των αιτημάτων τους. Ανάμεσα στα παραδείγματα είναι ένα αίτημα το οποίο περιλαμβάνει λέξεις όπως «άμεσα», «επείγον» ή «το συντομότερο δυνατόν», με σκοπό να ενεργοποιηθούν οι κατάλληλοι υπεύθυνοι για την άμεση ανταπόκριση. Η ταχεία διαχείριση τέτοιων καταστάσεων μπορεί να μετατρέψει μία αρνητική εμπειρία σε θετική, ενώ παράλληλα μειώνει τον κίνδυνο κρίσεων. Τέλος, η χρήση τεχνικών μοντελοποίησης θεμάτων σε συνδυασμό με ανάλυση συναισθήματος επιτρέπει την εις βάθος κατανόηση των συνομιλιών με τους πελάτες, τόσο ως προς τα θέματα που συζητούνται όσο και ως προς τη συναισθηματική τους φόρτιση. Έτσι, καθίσταται δυνατός ο εντοπισμός των σημαντικότερων θεμάτων ή προκλήσεων που αντιμετωπίζουν οι χρήστες, καθώς και η αξιολόγηση του τρόπου με τον οποίο βιώνουν αυτά τα ζητήματα. Οι πληροφορίες αυτές μπορούν να αξιοποιηθούν για την προώθηση αλλαγών σε ευρεία κλίμακα [26].

3.3.1 Μειονεκτήματα Topic Modeling

Η ανάλυση περιεχομένου έρχεται να συμπληρώσει τα κενά που συχνά αφήνουν οι παραδοσιακές ποιοτικές ή ποσοτικές μέθοδοι έρευνας, προσφέροντας μια συστηματική και οργανωμένη προσέγγιση στην ερμηνεία επικοινωνιακού υλικού. Ο ερευνητής που επιλέγει να εφαρμόσει την ανάλυση περιεχομένου, πριν προχωρήσει στη διαδικασία της κατηγοριοποίησης και της μέτρησης, είναι απαραίτητο να έχει μελετήσει τη γλώσσα, τη μορφή, τη δομή και το ευρύτερο πλαίσιο μέσα στο οποίο παράγεται το περιεχόμενο. Μόνο μέσα από αυτή τη βαθιά κατανόηση μπορεί να διαμορφώσει τεκμηριωμένα ερευνητικά ερωτήματα, υποθέσεις και αναλυτικά σχήματα. Εκεί που μια παραδοσιακή μελέτη πιθανόν να ολοκληρωνόταν, η ανάλυση περιεχομένου ξεκινά τη συστηματική της πορεία, οργανώνοντας και δοκιμάζοντας τις αναλυτικές της κατηγορίες πριν την εφαρμογή τους σε μεγαλύτερη κλίμακα [27]. Η μοντελοποίηση θεμάτων, αν και ισχυρό εργαλείο ανάλυσης μεγάλων ποσοτήτων κειμενικών δεδομένων, παρουσιάζει αρκετά μειονεκτήματα που πρέπει να ληφθούν υπόψη κατά την εφαρμογή της. Κατ' αρχάς, ενώ η μοντελοποίηση θεμάτων έχει την ικανότητα να προσδιορίζει τα θέματα ,τα ίδια τα θέματα συνήθως αντιπροσωπεύονται ως λίστες και η σημασία τους μπορεί να χρειάζεται ανθρώπινη κρίση και γνώσεις στοχευμένες στο τομέα αυτό .Ένα ακόμη σημαντικό μειονέκτημα της μοντελοποίησης θεμάτων αφορά τη δυσκολία καθορισμού του αριθμού θεμάτων που πρέπει να εξάγονται από το μοντέλο. Η επιλογή του σωστού αριθμού θεμάτων αποτελεί συχνά μια υποκειμενική διαδικασία, η οποία βασίζεται σε εμπειρικές δοκιμές, χωρίς σαφή κριτήρια ή κατευθυντήριες γραμμές, κάτι που μπορεί να οδηγήσει σε αναξιόπιστα ή παραπλανητικά αποτελέσματα. Επιπλέον, ένα συχνό πρόβλημα που προκύπτει από τη μοντελοποίηση θεμάτων είναι η έλλειψη πλαισίου. Δεν καταγράφει τα συμφραζόμενα μεταξύ λέξεων ή φράσεων μέσα σε ένα θέμα με αποτέλεσμα να δυσκολεύεται να κατανοήσει διαφοροποιημένες έννοιες. Αξιοσημείωτο είναι επίσης ότι η μοντελοποίηση θεμάτων βασίζεται κυρίως σε στατιστικές συσχετίσεις μεταξύ λέξεων και όχι σε βαθύτερη εννοιολογική κατανόηση του περιεχομένου. Αυτό σημαίνει ότι τα εξαγόμενα θέματα ενδέχεται να μην αντανακλούν πλήρως το πραγματικό νόημα ή τις έννοιες πίσω από τις λέξεις, κάτι που μπορεί να οδηγήσει σε λιγότερο συνεκτικά και αποδοτικά αποτελέσματα. Τέλος, η μεροληψία δειγματοληψίας αποτελεί έναν ουσιαστικό κίνδυνο για την εγκυρότητα οποιουδήποτε πειράματος, καθώς τα ευρήματα που εξάγονται ενδέχεται να μην είναι γενικεύσιμα σε άλλα συμφραζόμενα. Συγκεκριμένα, τα σύνολα δεδομένων που χρησιμοποιούνται σε οποιαδήποτε μελέτη έχουν υποστεί διάφορα στάδια προεπεξεργασίας, γεγονός που ενδέχεται να επηρεάσει τα

τελικά αποτελέσματα. Είναι πιθανό ότι, υπό διαφορετικές διαδικασίες προεπεξεργασίας, τα δεδομένα θα οδηγούσαν σε διαφορετικά συμπεράσματα [28].

3.4 Σχέση Topic Modeling και Ανάλυση Συναισθήματος

Σε αυτή την ενότητα, συζητούνται οι τεχνικές που χρησιμοποιούν το Topic Modeling για την εξαγωγή των δεδομένων. Υπάρχουν δύο βασικές μεθοδολογίες για το Topic Modeling: το pLSA και το LDA. Ωστόσο, η έρευνα της εργασίας μας επικεντρώθηκε σε τεχνικές βασισμένες στο LDA, και συνεπώς, σε αυτή την ενότητα, εξετάζονται μόνο οι τεχνικές που χρησιμοποίησαν μεθοδολογίες βασισμένες στο LDA για την εξαγωγή άκρων [29]. Με απλούς όρους, η ανάλυση συναισθήματος εντοπίζει ιδέες από πληροφορίες κειμένου που βασίζονται σε θετικές και αρνητικές λέξεις. Υπάρχουν αρκετές περιπτώσεις χρήσης, όπως κριτικές πελατών για προϊόντα καθώς και εξόρυξη γνώμης αναρτήσεων στα μέσα κοινωνικής δικτύωσης. Η ανάλυση συναισθήματος παίζει σημαντικό ρόλο στην ανάπτυξη της επιχείρησης και στην οικοδόμηση μιας ισχυρότερης σχέσης με τους πελάτες με τα προϊόντα. Η μοντελοποίηση θεμάτων είναι μια στατιστική μέθοδος για την εξαγωγή θεμάτων από ένα σώμα κειμένου ή εγγράφων. Οι συνήθεις περιπτώσεις επιχειρηματικής χρήσης θα περιλαμβάνουν την εξαγωγή θεμάτων άρθρων εφημερίδων, ερευνητικών άρθρων ή ακόμα και διαδικτυακών αναρτήσεων [30].

4 Εφαρμογή NLP και LDA για Ανάλυση Συναισθήματος και Εξόρυξη Θεμάτων

4.1 Χρήση Μοντέλων NLP για Ανάλυση Συναισθήματος

Τα μοντέλα Επεξεργασίας Φυσικής Γλώσσας (NLP) αποτελούν έναν κλάδο της επιστήμης των υπολογιστών και της Τεχνητής Νοημοσύνης. Συγκεκριμένα, γίνεται χρήση μηχανικής μάθησης με στόχο οι υπολογιστές να κατανοούν, να ερμηνεύουν, να δημιουργούν και να επικοινωνούν την ανθρώπινη γλώσσα. Τα μοντέλα αυτά είναι σχεδιασμένα για να χειρίζονται την πολυπλοκότητα της φυσικής γλώσσας, δίνοντας τη

δυνατότητα στις μηχανές να εκτελούν εργασίες όπως ανάλυση συναισθήματος, μετάφραση γλώσσας, σύνοψη, απάντηση ερωτήσεων και άλλα [31]. Το NLP μπορεί μέσω τον υπολογιστών, να κατανοεί, να δημιουργεί, κείμενο και ομιλία σε συνδυασμό της υπολογιστικής γλωσσολογίας (τη μοντελοποίηση της ανθρώπινης γλώσσας) με τη στατιστική μοντελοποίηση, τη μηχανική και τη βαθιά μάθηση. Τα τελευταία χρόνια, τα μοντέλα NLP έχουν σημειώσει σημαντική εξέλιξη. Αυτό οφείλεται στην πρόοδο της βαθιάς μάθησης και πρόσβασης σε μεγάλα σύνολα δεδομένων. Το NLP περιλαμβάνει ένα μεγάλο φάσμα εφαρμογών όπως προαναφέρθηκε για παράδειγμα, η ανάλυση συναισθήματος. Στην εφαρμογή αυτή γίνεται προσδιορισμός και ταξινόμηση του συναισθήματος που εκφράζεται σε κάθε κείμενο ως «θετικό», «αρνητικό» και «ουδέτερο» [32]. Την τελευταία δεκαετία, η ανάλυση συναισθήματος συναντάται ευρέως σε διάφορους και ποικίλους τομείς όπως κυρίως η παρακολούθηση των μέσων κοινωνικής δικτύωσης, η διαχείριση εταιρικής φήμης και η ανάλυση σχολίων πελατών. Η ανάλυση συναισθήματος αποτελεί μια μορφή ταξινόμησης κειμένου που εστιάζει στην επεξεργασία και κατηγοριοποίηση υποκειμενικών δηλώσεων ενώ παράλληλα αναλύει απόψεις ώστε να κατανοήσει την αντίληψη του κοινού. Η ανάλυση συναισθήματος και η εξόρυξη γνώμης είναι όροι που έχουν την ίδια σημασία και βασίζονται στην NLP για τη συλλογή και ανάλυση λέξεων που εκφράζουν απόψεις ή συναισθήματα. Με πιο κατανοητό τρόπο, η ανάλυση συναισθήματος περιγράφεται ως η διαδικασία αναγνώρισης των συναισθημάτων των ανθρώπων σχετικά με ένα θέμα και τα χαρακτηριστικά του [33]. Η χρήση NLP για την ανάλυση συναισθήματος, γενικά προσδιορίζει το συναίσθημα πίσω από μία κατάσταση [34]. Η γλώσσα λειτουργεί ως μεσολαβητής στην ανθρώπινη επικοινωνία ενώ κάθε δήλωση μπορεί να φέρει ένα από τα τρία συναισθήματα που αναφέρθηκαν. Σήμερα, η ανάλυση συναισθήματος είναι ένα ισχυρό εργαλείο για την αποκρυπτογράφηση ή αποκωδικοποίηση σύνθετων ανθρώπινων συναισθημάτων που αποτυπώνονται σε δεδομένα κειμένου. Κάνοντας χρήση τεχνικών και μεθοδολογιών, όπως η ανάλυση κειμένου, οι αναλυτές μπορούν να εξαγάγουν πολύτιμες πληροφορίες, οι οποίες μπορούν να αφορούν για παράδειγμα από τις προτιμήσεις των καταναλωτών έως και τις πολιτικές τάσεις. Κατά αυτόν τον τρόπο διευκολύνεται η διαδικασία λήψης αποφάσεων σε διάφορους τομείς [35].

4.2 Εφαρμογή του Αλγορίθμου LDA στην Εξόρυξη Θεμάτων

Η μέθοδος Latent Dirichlet Allocation, η οποία διατυπώθηκε από τους Bley, Ng και Jordan το 2003, είναι μία από τις πιο γνωστές μεθόδους για την ανάλυση θεμάτων, λόγω της ευκολίας και της αποδοτικότητας της. Ο αλγόριθμος υποθέτει ότι κάθε έγγραφο σε ένα σύνολο κειμένων αποτελεί ένα μείγμα πολλών θεμάτων, με κάθε θέμα να αντιστοιχεί σε ένα συνολικό αριθμό λέξεων. Ο LDA αναθέτει επαναληπτικά λέξεις σε θέματα και έγγραφα σε διάφορα θέματα, για να αυξήσει την πιθανότητα εμφάνισης των δεδομένων κειμένου. Ως συνέπεια, ο αλγόριθμος προσδιορίζει την κατανομή των θεμάτων σε ένα σύνολο κειμένων και υπολογίζει την πιθανότητα κάθε λέξης να σχετίζεται με ένα συγκεκριμένο θέμα. Η ευχέρεια και η ικανότητα επέκτασης του LDA το έχουν καταστήσει δημοφιλές τόσο σε ακαδημαϊκά όσο και σε βιομηχανικά πλαίσια. Συγκεκριμένα, η εφαρμογή του LDA μπορεί να γίνει σε ποικιλία κειμένων και να διαχειρίζεται διαφορετικούς τύπους δεδομένων χωρίς την ανάγκη να απαιτούνται βασικές αλλαγές στη διαδικασία. Αυτό τον καθιστά ωφέλιμο σε πολλές διαφορετικές περιπτώσεις και περιοχές εφαρμογής. Η αποτελεσματικότητά του του επιτρέπει να επεξεργάζεται ένα εκτεταμένο εύρος από δεδομένα κειμένου. Ενδεικτικά, η μοντελοποίηση θεμάτων που στηρίζεται σε αναλύσεις μέσων κοινωνικής δικτύωσης κάνει εύκολη την κατανόηση συνομιλιών και αντιδράσεων μεταξύ μέλη διαδικτυακών κοινοτήτων. Τα θεματικά μοντέλα είναι ιδιαίτερα χρήσιμα για την απεικόνιση δεδομένων παρέχοντας μια αποτελεσματική μέθοδο για την εύρεση λανθάνουσας σημασιολογικής δομής σε μεγάλα σύνολα δεδομένων [36].

4.3 Περιγραφή της Μεθοδολογικής Διαδικασίας

1. Συλλογή δεδομένων

Αρχικά, γίνεται εξαγωγή των απαραίτητων δεδομένων από τα άρθρα, όπως η ημερομηνία δημοσίευσης, ο τίτλος, τα στοιχεία των συγγραφέων, η περίληψη και ενδεχομένως keywords. Τα δεδομένα είναι κρίσιμα για την ανάλυση και κατηγοριοποίηση των άρθρων, καθώς προσφέρουν περαιτέρω εμβάθυνση στις θεματικές τους. Για την παρούσα εργασία χρησιμοποιήθηκε έτοιμο σύνολο δεδομένων από την πλατφόρμα Kaggle. Το dataset αυτό αποτέλεσε τη βάση για την εφαρμογή τεχνικών ανάλυσης συναισθήματος με τη μέθοδο

LDA. Στη συνέχεια, τα δεδομένα αποθηκεύονται σε αρχείο CSV (Comma-Separated Values) για επιπλέον επεξεργασία και ανάλυση (τιμές διαχωρισμένες με κόμμα). Αυτή η τεχνική διασφαλίζει την ακριβή συλλογή δεδομένων από ένα μεγάλο όγκο άρθρων, καθιστώντας την κατάλληλη για τη διεξαγωγή μεγάλων ερευνών ή για την ανάπτυξη συστημάτων που προϋποθέτουν εκτεταμένη ανάλυση κειμένων.

2. Προ-επεξεργασία

Τα ακατέργαστα δεδομένα τις περισσότερες φορές περιλαμβάνουν επιπλέον πληροφορίες οι οποίες απαιτούνται να αφαιρεθούν πριν δοθούν για ανάλυση. Η προ-επεξεργασία δεδομένων αποτελείται από μία σειρά βημάτων όπως lowercase, κατάργηση σημείων στίξης (punctuation removal), tokenization, αφαίρεση λέξεων διακοπής (stop word removal) και lemmatization. Ειδικότερα, η διαδικασία lowercase μετατρέπει τα δεδομένα σε ομοιόμορφα γράμματα, δηλαδή πεζά. Για παράδειγμα, με την μετατροπή του κειμένου σε μικρά γράμματα, οι λέξεις "Apple" και "apple" αντιμετωπίζονται ως η ίδια λέξη. Στην περίπτωση που αυτό παραληφθεί θα υπήρχαν διπλές καταχωρήσεις και θα έπρεπε να γίνει διαχείριση κάθε παραλλαγής μεμονωμένα. Η κατάργηση των σημείων στίξης είναι αναγκαία διότι ορισμένες λέξεις που ενσωματώνουν τα μοντέλα δεν τα υποστηρίζουν. Ταυτόχρονα, τα σημεία στίξης όπως οι άνω τελείες, δεν προσφέρουν πάντα χρήσιμες πληροφορίες για αλγορίθμους που αναλύουν κείμενα. Η αφαίρεσή τους καθιστά το κείμενο πιο απλό στην επεξεργασία. Στη συνέχεια γίνεται tokenization, δηλαδή αποξενώνονται «μέρη» ή «κομμάτια» του κειμένου σε μικρότερα τα οποία αναφέρονται και ως «tokens». Ο βασικός τους σκοπός είναι ότι κατανομή του κειμένου σε αυτά επιτρέπει την εκτέλεση λειτουργιών, όπως η ανάλυση της συχνότητας των λέξεων και η κατηγοριοποίηση τους. Ουσιαστικά, η αφαίρεση τους συμβάλλει στην βελτίωση της απόδοσης καθώς ο χώρος περιορίζεται. Η lemmatization αφορά μια διαδικασία που μετατρέπει τις παραλλαγές μιας λέξης στη βασική ή «αρχική» της μορφή και προσπαθεί να αναγνωρίσει τη λέξη στο λεξικό της γλώσσας και να τη μειώσει σε μία μορφή που δεν περιλαμβάνει διαφοροποιήσεις ή παραλλαγές, όπως πληθυντικός αριθμός, κλίσεις και πτώσεις. Ως παράδειγμα αναφέρονται οι παρακάτω προτάσεις.

- Η λέξη "τρέχω" παραμένει "τρέχω"
- Η λέξη "έτρεχα" μετατρέπεται σε "τρέχω"

3. Δημιουργία Document-Term Matrix (DTM)

Το Document-Term Matrix, ή αλλιώς Πίνακας Εγγράφων-Όρων, αποτελεί μια θεμελιώδη δομή δεδομένων στον χώρο της Επεξεργασίας Φυσικής Γλώσσας και της εξόρυξης κειμένου. Η κύρια λειτουργία του είναι να μετασχηματίζει τα μη δομημένα δεδομένα κειμένου (μια συλλογή εγγράφων) σε μια δομημένη, αριθμητική αναπαράσταση, καθιστώντας τα κατάλληλα για επεξεργασία από αλγορίθμους μηχανικής μάθησης και στατιστικά μοντέλα. Όσον αφορά τη δομή και την αναπαράσταση του, ο DTM είναι ένας δισδιάστατος πίνακας με γραμμές, στήλες και τιμές. Ειδικότερα, κάθε γραμμή του πίνακα αντιπροσωπεύει ένα μοναδικό έγγραφο από τη συλλογή κειμένων. Αυτό μπορεί να είναι οτιδήποτε, από ένα άρθρο και ένα tweet μέχρι ένα σχόλιο χρήστη ή μια παράγραφος. Αντίθετα, κάθε στήλη αντιστοιχεί σε έναν μοναδικό όρο (λέξη ή token) που έχει εμφανιστεί τουλάχιστον μία φορά σε οποιοδήποτε έγγραφο της συλλογής. Το σύνολο αυτών των μοναδικών όρων αποτελεί το λεξιλόγιο της συλλογής. Τέλος, Η τιμή σε κάθε κελί (i,j) του πίνακα (δηλαδή, στη γραμμή i και στη στήλη j) υποδεικνύει τη συχνότητα εμφάνισης του όρου j εντός του εγγράφου i. Αυτή η συχνότητα μπορεί να είναι απλή καταμέτρηση (raw count), δυαδική παρουσία (binary, 0 ή 1), ή κάποια σταθμισμένη τιμή, όπως η συχνότητα όρου-αντίστροφη συχνότητα εγγράφου (TF-IDF) [72]. Το DTM είναι ζωτικής σημασίας για την εφαρμογή του αλγορίθμου LDA, καθώς παρέχει την απαραίτητη αριθμητική είσοδο που επιτρέπει στον αλγόριθμο να αναλύσει τα μοτίβα εμφάνισης των λέξεων και να ανακαλύψει τα υποκείμενα, "κρυμμένα" θέματα σε μια συλλογή κειμένων. Κάθε θέμα που προκύπτει από το LDA χαρακτηρίζεται από ένα σύνολο λέξεων που τείνουν να εμφανίζονται συχνά μαζί, και το DTM προσφέρει τα δεδομένα για αυτή την ανίχνευση. Πέρα από τη θεματική μοντελοποίηση, το DTM χρησιμοποιείται και σε άλλες εφαρμογές, όπως η ανάλυση συναισθήματος και η δημιουργία οπτικοποιήσεων δεδομένων κειμένου, όπως τα Word Clouds, όπου η συχνότητα των λέξεων από το DTM καθορίζει το μέγεθος της κάθε λέξης στην οπτική αναπαράσταση. Με λίγα λόγια, το DTM είναι ο σύνδεσμος που μετατρέπει το αδόμητο κείμενο σε ένα κατανοητό αριθμητικό σύνολο, απαραίτητο για κάθε προηγμένη ανάλυση κειμένων [73].

4. Δημιουργία θεμάτων

Ο αλγόριθμος LDA χρησιμοποιείται για να δημιουργήσει τα θέματα που υπάρχουν σε ένα σύνολο χωρίς τον εκ των προτέρων καθορισμό τους. Έτσι, αναλύονται οι λέξεις που εμφανίζονται σε κάθε κείμενο αλλά και η συχνότητα τους. Το LDA βρίσκει κάποια κοινά

μοτίβα από τις λέξεις που περιέχονται μέσα σε ένα κείμενο. Το κάθε θέμα είναι και μία ομάδα λέξεων οι οποίες εμφανίζονται πιο συχνά μαζί. Για παράδειγμα, το θέμα «υγεία» μπορεί να έχει λέξεις όπως ασθένεια θεραπεία γιατρός. Επιπλέον, ο αλγόριθμος ερευνά τις λέξεις που παρουσιάζονται συχνά στο ίδιο κείμενο αλλά και κοντά η μία στην άλλη. Στην περίπτωση που δύο λέξεις εμφανίζονται συχνά μαζί το LDA υποθέτει πως ανήκουν στο ίδιο θέμα. Οι λέξεις που εμφανίζονται ως «co-occurrences» σημαίνει πως θεωρούνται ως συναφείς και είναι πιθανό να ανήκουν στο ίδιο θέμα αντίστοιχα. Για παράδειγμα, αν σε πολλά άρθρα εμφανίζονται οι λέξεις διατροφή και υγεία μαζί, ο αλγόριθμος θα τις κατατάσσει στο ίδιο θέμα που ανήκουν δηλαδή «υγιεινή ζωή». Κατά αυτόν τον τρόπο, ο αλγόριθμος καταλαβαίνει ποια είναι και τα «κρυμμένα» θέματα στα κείμενα τα οποία βασίζονται στο μεγαλύτερο τους μέρος στην συχνότητα των λέξεων και την σύνδεση που έχουν μεταξύ τους [37] [38] [39].

5. Φόρμουλα Gibbs (Gibbs Sampling)

$$P(z_i = t \mid z^{-i}, w) = \frac{n_{m,t}^{-i} + \alpha \qquad n_{t,w_i}^{-i} + \beta}{\sum_{t'=1}^{T} (n_{m,t'}^{-i} + \alpha) \sum_{v'=1}^{V} (n_{t,v'}^{-i} + \beta)}$$

Ο αλγόριθμος LDA χρησιμοποιεί την δειγματοληψία Gibbs κατά την ανάθεση θεμάτων για να εκτιμήσει τα θέματα και τις κατανομές των λέξεων που τα συνδέουν. Συγκεκριμένα, ανακαλύπτει τα «κρυμμένα» θέματα μέσω του Topic Modeling και η Gibbs sampling βοηθάει να εκτιμηθούν οι παράμετροι κάτι το οποίο είναι δύσκολο να υπολογιστούν ακριβώς με άλλες μεθόδους. Ο τύπος δειγματοληψίας Gibbs αποτελείται από τον πρώτο παράγοντα ο οποίος εκφράζει την πιθανότητα να ανήκει το θέμα t στο κείμενο d. Δηλαδή, υπολογίζει αυτή την πιθανότητα βασισμένο στον αριθμό των λέξεων στο κείμενο d που σχετίζονται με το θέμα t. Αυτό ουσιαστικά προσπαθεί να απαντήσει στο ερώτημα:

Πόσο συχνά εμφανίζεται το θέμα t στο κείμενο d;

Ο δεύτερος παράγοντας εκφράζει την πιθανότητα μια λέξη w να ανήκει στο θέμα t. Ο αλγόριθμος υπολογίζει αυτή την πιθανότητα μετρώντας πόσες φορές η λέξη w εμφανίζεται στο θέμα t σε όλα τα έγγραφα που περιέχουν το θέμα t. Αυτό ουσιαστικά δημιουργεί το ερώτημα:

Πόσο συχνά εμφανίζεται η λέξη w σε σχέση με το θέμα t στο συνολικό σύνολο των εγγράφων;

Η δειγματοληψία Gibbs είναι μια διαδικασία επανάληψης. Αυτό σημαίνει ότι μια λέξη δεν ανατίθεται σε ένα θέμα μια μόνο φορά και στη συνέχεια ξεχνιέται. Αντιθέτως, η δειγματοληψία Gibbs περνά κάθε λέξη μέσω αρκετών επαναλήψεων, συνεχώς ενημερώνοντας τις πιθανότητες σύνδεσης λέξης και θέματος με βάση τις προηγούμενες εκτιμήσεις.

6. Ταξινόμηση κειμένων

Ο αλγόριθμος LDA είναι ικανός για την οργάνωση και ταξινόμηση των εγγράφων σε κατηγορίες (ή θέματα) χωρίς ανθρώπινη ενέργεια και καταφεύγοντας μόνο στις λέξεις και στη συχνότητά τους. Με αυτόν τον τρόπο, μπορεί να γίνει μία σχετική εκτίμηση (δηλαδή να εντοπίσει ποια θέματα κυριαρχούν) σε μια συλλογή εγγράφων. Έτσι, το LDA συμβάλλει στην οργάνωση των εγγράφων με βάση τα θέματα τους. Αυτές οι εκτιμήσεις χρησιμοποιούνται για διάφορους σκοπούς, όπως η ταξινόμηση των εγγράφων σε κατηγορίες ή η ενίσχυση της αναζήτησης δεδομένων σε εκτενείς βάσεις πληροφοριών

7. Βελτιστοποίηση

Όπως συμβαίνει με πολλές τεχνικές εξόρυξης κειμένου στην επιστήμη των δεδομένων, η προ-επεξεργασία, όπως η αφαίρεση λέξεων διακοπής και η lemanization βοηθούν στη βελτίωση των αποτελεσμάτων του μοντέλου LDA, αφαιρώντας συνηθισμένες λέξεις που δεν προσφέρουν ιδιαίτερη σημασία και ενοποιώντας διάφορες μορφές λέξεων για να γίνει το μοντέλο πιο ακριβές. Ωστόσο, η βελτιστοποίηση των μοντέλων είναι πιθανό να είναι μία δύσκολη διαδικασία καθώς δεν υπάρχει συγκεκριμένος αριθμός θεμάτων που εξάγει τα καλύτερα αποτελέσματα. Αυτό ορισμένες φορές οδηγεί σε δοκιμές και λάθη μέσω της προσαρμογής του μοντέλου [40].

8. Ανάλυση συχνότητας λέξεων (tf-idf)

Η ανάλυση συχνότητας λέξεων με tf-idf είναι ένα βασικό εργαλείο στην εξόρυξη κειμένου και την επεξεργασία φυσικής γλώσσας και στοχεύει στην «quantification of the importance of words in a document» του περιεχομένου ενός εγγράφου, δηλαδή η

εκτίμηση της βαρύτητας των λέξεων εντός ενός εγγράφου. Με τον όρο αυτό εννοείται η διαδικασία μετατροπής του περιεχομένου του εγγράφου σε αριθμητικά δεδομένα ώστε να γίνει η ανάλυση με υπολογιστικούς ή στατιστικούς τρόπους. Με το tf-idf, υπολογίζεται η μέτρηση της «σημασίας» κάθε λέξης μέσα σε ένα κείμενο, συγκριτικά με τη συνολική συλλογή εγγράφων. Παρόλα αυτά, ορισμένες λέξεις εμφανίζονται περισσότερο όμως δεν είναι απαραίτητο να έχουν ιδιαίτερη σημασία. Στην αγγλική γλώσσα, παραδείγματα αυτών των λέξεων είναι τα "the", "is", "of'. Το tf (team frequency) δηλαδή η συχνότητα του όρου μετρά πόσο συχνά εμφανίζεται μια λέξη σε ένα έγγραφο. Η βασική ιδέα είναι ότι όσο πιο συχνά εμφανίζεται μια λέξη σε ένα έγγραφο, τόσο πιο σημαντική είναι για το συγκεκριμένο. Ο υπολογισμός του tf είναι συνήθως:

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{N_d}$$

Όπου:

- t = ο όρος (λέξη)
- d = το έγγραφο
- $f_{t,d}$ = το πλήθος εμφανίσεων του όρου t στο έγγραφο d
- N_d = ο συνολικός αριθμός λέξεων στο έγγραφο ddd

Ωστόσο, κάποιες από αυτές τις λέξεις μπορεί να είναι πιο σημαντικές σε ορισμένα έγγραφα. Μία άλλη προσέγγιση για τον υπολογισμό της σημασία των λέξεων είναι η χρήση της idf (inverse document frequency) δηλαδή η αντίστροφη συχνότητα του εγγράφο, χρησιμοποιείται για να μετρήσει τη «σπανιότητα» ενός όρου στο σύνολο των εγγράφων. Σε αυτήν την συχνότητα μειώνεται η αξία των λέξεων που εμφανίζονται συχνά και έχουν μεγαλύτερο ρόλο οι λέξεις που εμφανίζονται πολύ σπάνια σε μια συλλογή δεδομένων. Πιο απλά, εάν μια λέξη εμφανίζεται σε πολλά έγγραφα, τότε θεωρείται λιγότερο σημαντική, ενώ αν εμφανίζεται μόνο σε λίγα έγγραφα, θεωρείται πιο σημαντική. Ο υπολογισμός του idf είναι:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{n_t}\right)$$

Όπου:

- t = ο όρος (λέξη)
- N = συνολικός αριθμός εγγράφων στο σύνολο δεδομένων
- n_t = αριθμός των εγγράφων που περιέχουν τον όρο t

Τελικά, ο υπολογισμός του tf-idf προκύπτει από τον συνδυασμό αυτών των δύο πολλαπλασιάζοντας τους μεταξύ τους:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

Ο υπολογισμός βοηθά στο να αποδοθεί μεγαλύτερη σημασία σε λέξεις που είναι συχνές σε ένα έγγραφο αλλά σπάνιες στη συλλογή εγγράφων, και να δίνεται μικρότερη σημασία σε λέξεις που είναι συχνές σε πολλά έγγραφα. Παρόλο που αυτή η μέθοδος είναι εμπειρική και χρησιμοποιείται ευρέως σε εφαρμογές όπως η εξόρυξη κειμένου και οι μηχανές αναζήτησης, η θεωρητική της βάση θεωρείται λιγότερο σταθερή από τους ειδικούς της θεωρίας της πληροφορίας [41] [42].

5 Χρήση Word Cloud στην Ανάλυση Συναισθήματος

5.1 Ιστορική Εξέλιξη και Εισαγωγή στο Word Cloud

Σύμφωνα με το βιβλίο Introduction to Text Visualisation, η πρώτη οπτικοποίηση τύπου Word Cloud δημιουργήθηκε το 1976 από τον κοινωνικό ψυχολόγο Stanley Milgram. Ο Milgram ζήτησε από ανθρώπους να καταγράψουν γνωστά ορόσημα του Παρισιού και κατόπιν δημιούργησε έναν χάρτη στον οποίο τα ονόματα των ορόσημων εμφανίζονταν σε κείμενο. Συγχρόνως, το μέγεθος κάθε λέξης καθοριζόταν από τον αριθμό των απαντήσεων που είχε λάβει. Αυτό σημαίνει ότι όσο περισσότερες φορές αναφερόταν μία λέξη, τόσο μεγαλύτερη ήταν η γραμματοσειρά της. Το πρώτο έντυπο παράδειγμα μιας «σταθμισμένης λίστας» λέξεων-κλειδιών εμφανίστηκε το 1995 στο βιβλίο Microserfs του Douglas Coupland, όπου οι λέξεις σε γεωγραφικούς χάρτες προσαρμόζονταν σε μέγεθος για να απεικονίζουν τη σχετική σημασία των πόλεων. Εντούτοις, το Word Cloud έγινε γνωστό σε πολλούς μόλις το 2006, τη στιγμή που ο όρος εμφανίστηκε στο Flickr, έναν δημοφιλή ιστότοπο κοινής χρήσης φωτογραφιών [43]. Σε λίγο χρονικό διάστημα, η χρήση του έγινε ευρύτερη στην κοινότητα του UX Design (User Experience Design) η οποία αναφέρεται στη διαδικασία σχεδιασμού προϊόντων ή υπηρεσιών με στόχο την καλύτερη εμπειρία του χρήστη κατά την αλληλεπίδρασή του με το προϊόν ή την υπηρεσία αντίστοιχα. Σκοπός του UX Design είναι να βελτιώσει την ικανοποίηση του χρήστη, δημιουργώντας την αίσθηση ότι το προϊόν είναι εύχρηστο και ικανοποιητικό [44]. Η

εμφάνιση του ξεκίνησε σε διάφορες πλατφόρμες στο διαδίκτυο ενώ την ίδια περίοδο, το Word Cloud χρησιμοποιούταν και από υπηρεσίες όπως το Del.icio.us και το Technorati (το οποίο πλέον δεν υπάρχει). Καθ 'όλη τη δεκαετία του 2010, τα Wordclouds έγιναν όλο και πιο κοινά στις επιχειρηματικές παρουσιάσεις, την εκπαίδευση, τη δημοσιογραφία δεδομένων και την ανάλυση κοινωνικών μέσων. Στο σύγχρονο ψηφιακό κόσμο, το Word Cloud έχει εξελιχθεί σε ευέλικτα εργαλεία για ποικίλες εφαρμογές. Ενδεικτικά, όπως στην παρούσα μελέτη, οι τεχνικές οπτικοποίησης χρησιμοποιούνται για την ανάλυση συναισθήματος με ακαδημαϊκό σκοπό [45].

5.2 Τι είναι και πώς λειτουργεί το Word Cloud

Το Word Cloud είναι ένας τύπος τεχνικής οπτικοποίησης δεδομένων κειμένου που αποδίδει μία έννοια για ένα θέμα με τη χρήση της οπτικής αναπαράστασης των λέξεων και είναι μια λειτουργική μέθοδος για την ανάλυση, την έρευνα και τη συλλογή κειμενικών απόψεων. Αναφορικά με άλλες μεθόδους οπτικής απεικόνισης δεδομένων κειμένου, το Word Cloud είναι περισσότερο επωφελής και απλή διαδικασία καθώς όχι μόνο αντιπροσωπεύει μια εικόνα συγκεκριμένων λέξεων, αλλά αντιπροσωπεύει επίσης κάτι διαφορετικό από τη λέξη. Ευρύτερα, ο όρος «cloud» έχει χρησιμοποιηθεί για να αναφερθεί στην αρχική φάση της ανάλυσης ενός εγγράφου κειμένου. Ως εκ τούτου, η συχνότητα των λέξεων είναι ο πρωτεύον παράγοντας κατά την κατηγοριοποίηση ή την οπτικοποίηση ενός αποτελέσματος [46]. Ταυτόχρονα, χαρακτηρίζεται και ως μια συλλογή λέξεων που απεικονίζονται σε διαφορετικά μεγέθη. Επιπλέον, η έρευνα δείχνει ότι όσο πιο έντονη και ιδιαίτερη είναι η λέξη, τόσο πιο συχνά εμφανίζεται στο κείμενο και τόσο μεγαλύτερη είναι και η σημασία της. Η δυνατότητα άντλησης πολλών ιδεών με ταχύτητα παρέχει προσαρμοστικότητα στην ερμηνεία των οπτικοποιήσεων. Το Word Cloud είναι ουσιαστικά ένα περιγραφικό εργαλείο. Κατά συνέπεια, θα πρέπει να αξιοποιείται μόνο για την καταγραφή συγκεκριμένων ποιοτικών γνώσεων. Στην παρούσα εργασία έγινε χρήση του Word Cloud με την γλώσσα R. Σημασιολογικά, στην R, ο όρος "πακέτο" (package) αναφέρεται σε ένα σύνολο λειτουργιών και δεδομένων που μπορούν να εγκατασταθούν και να χρησιμοποιηθούν, ενώ η "βιβλιοθήκη" (library) είναι η συλλογή των εγκατεστημένων πακέτων.

Α. Πακέτο (package): Ένα σύνολο συναρτήσεων, δεδομένων και τεκμηρίωσης που μπορεί να εγκατασταθεί μέσω install.packages("package_name").

Β. Βιβλιοθήκη (library): Ο χώρος όπου αποθηκεύονται τα εγκατεστημένα πακέτα. Ένα πακέτο γίνεται διαθέσιμο για χρήση μέσω library(package_name).

Στην τρέχουσα εργασία για την προ επεξεργασία των δεδομένων μετά την εισαγωγή των κειμένων, χρησιμοποιήθηκαν τα εξής κύρια πακέτα R:

a. Tidyverse: Χρησιμοποιείται για το χειρισμό δεδομένων (π.χ., dplyr για επιλογή και μετασχηματισμό δεδομένων).
b. Tidytext: Χρησιμοποιείται για την ανάλυση κειμένου, όπως η unnest_tokens() για tokenization (διαχωρισμός λέξεων).
c. Stopwords: Χρησιμοποιείται για την αφαίρεση των stopwords από το κείμενο.
d. Tm: Χρησιμοποιείται για την επεξεργασία κειμένου και τη δημιουργία του Document-Term Matrix (DTM).
e. Topicmodels: Χρησιμοποιείται για τη μοντελοποίηση θεμάτων με τον αλγόριθμο LDA .

Οι βασικές βιβλιοθήκες που χρησιμοποιήθηκαν για την προ επεξεργασία των δεδομένων κειμένου είναι οι εξής:

1. Tidytext: Για tokenization(unnest_tokens()).
2. Stopwords: Για την αφαίρεση των stopwords.
3. Tm: Για τη δημιουργία του Document-Term Matrix (DTM) [42] [47] [48].

5.3 Πλεονεκτήματα και Μειονεκτήματα του Word Cloud

5.3.1 Πλεονεκτήματα του Word Cloud

Οι λόγοι για τους οποίους το Word Cloud είναι χρήσιμο ποικίλουν και σχετίζονται τόσο με την οπτικοποίηση όσο και με την ανάλυση δεδομένων κειμένου. Αποτελεί με βεβαιότητα ένα ισχυρό εργαλείο το οποίο επιτρέπει την απεικόνιση λέξεων με τέτοιο τρόπο ώστε να γίνεται άμεσα αντιληπτή η συχνότητα εμφάνισής τους [49]. Οι λέξεις που όπως προαναφέρθηκε εμφανίζονται πιο συχνά σε ένα κείμενο αποδίδονται με μεγαλύτερο και εντονότερο μέγεθος, δίνοντας έτσι μια ξεκάθαρη και εύκολα κατανοητή εικόνα του περιεχομένου. Αυτή η προσέγγιση καθιστά το Word Cloud οπτικά ευχάριστο, απλό στην κατανόηση και αποτελεσματικό στη μεταφορά πληροφορίας. Παρέχει δηλαδή, μια συνολική εικόνα του περιεχομένου με άμεσο και εύληπτο τρόπο,

καθιστώντας τα ιδιαίτερα χρήσιμα για ερευνητές, αναλυτές δεδομένων και επαγγελματίες του μάρκετινγκ [50] [51]. Ένας ακόμη σημαντικός λόγος για τη χρήση του Word Cloud είναι η δυνατότητα του να ενισχύει τη διαδικασία λήψης αποφάσεων. Μέσω της οπτικοποίησης των πιο συχνών όρων, δίνεται η δυνατότητα στους χρήστες να εντοπίσουν τάσεις και μοτίβα που δεν θα ήταν άμεσα εμφανή σε μια απλή ανάγνωση ενός κειμένου. Σε συνδυασμό με αυτό, επιτρέπουν μια ευέλικτη ερμηνεία των δεδομένων, καθώς η γραφική τους απεικόνιση προάγει τη δημιουργική σκέψη και τη διαμόρφωση νέων ιδεών. Ιδιαίτερη σημασία έχει το ότι ένα Word Cloud βοηθά να επισημανθούν οι περιοχές ή τα πεδία δημοτικότητας και αναγνώρισης από την άποψη και την προοπτική του κοινού. Ως εκ τούτου, βοηθά στην ανταλλαγή ιδεών και στην ανάπτυξη νέων εννοιών και ιδεών. Πιο συγκεκριμένα, επιτρέπει στο θέμα, το ενδιαφέρον ή τα σχόλια των συμμετεχόντων να προσδιοριστούν και να επισημανθούν [52]. Καταλήγοντας, η χρήση του Word Cloud για την παρουσίαση δεδομένων κειμένου προσφέρει σημαντικά πλεονεκτήματα αφού διακρίνονται για την απλότητα και τη σαφήνειά τους και διευκολύνουν την προβολή των πιο συχνά χρησιμοποιούμενων λέξεων. Με αυτόν τον τρόπο, οι λέξεις-κλειδιά γίνονται πιο ευδιάκριτες και ευανάγνωστες, επιτρέποντας στον χρήστη να αποκομίσει τις βασικές πληροφορίες. Εκτός αυτού, αποτελεί ένα εξαιρετικά αποτελεσματικό μέσο επικοινωνίας [53] [54]. Κάνουν τη μετάδοση πληροφοριών με σαφή και άμεσο τρόπο, ενώ παράλληλα καθιστούν το περιεχόμενο προσαρμοστικό για το ευρύ κοινό. Αξίζει τέλος να σημειωθεί ότι σε αντίθεση με τις παραδοσιακές μορφές παρουσίασης δεδομένων, όπως οι πίνακες, το Word Cloud ενισχύει την προσοχή και διατηρεί το ενδιαφέρον του αναγνώστη. Η δυναμική του μορφή συμβάλλει στην αρτιότερη αφομοίωση της πληροφορίας, καθιστώντας τα ένα κατεξοχήν αποτελεσματικό εργαλείο οπτικής απεικόνισης [55].

5.3.2 Μειονεκτήματα του Word Cloud

Παρόλα αυτά, είναι σημαντικό να αναφερθούν κάποιοι περιορισμοί οι οποίοι αφορούν κυρίως τον χαρακτηρισμό του ως περιγραφικό εργαλείο, το οποίο προσφέρει μια συνοπτική εικόνα. Ειδικότερα, η χρήση τους είναι ιδανική για την αρχική διερεύνηση και παρουσίαση δεδομένων, όμως δεν επαρκεί για εξειδικευμένες αναλύσεις. Η αρχική παρατήρηση δείχνει ότι το Word Cloud προσφέρει ένα περιορισμένο πλαίσιο, διότι απεικονίζεται μόνο η συχνότητα εμφάνισης των λέξεων, ενώ δεν παρουσιάζεται πώς χρησιμοποιούνται ή ποιες λέξεις συνδέονται μεταξύ τους. Επιπλέον, μπορεί να καταλήξουν σε σύγχυση της σημασίας, καθώς το γεγονός ότι μια λέξη εμφανίζεται με μεγαλύτερη γραμματοσειρά δεν δηλώνει απαραίτητα ότι είναι πιο σημαντική. Για

παράδειγμα, λέξεις όπως «το» ή «και» ενδέχεται να εμφανίζονται μεγαλύτερες λόγω της συχνής χρήσης, χωρίς να έχουν ιδιαίτερη σημασία. Μία ακόμη δυσκολία που καλείται να αντιμετωπίσει το Word Cloud είναι η αποτυχία αντίληψης των λεπτομερειών της γλώσσας, αφού δεν λαμβάνει υπόψη συνώνυμα ή διαφορετικές γραμματικές μορφές μιας λέξης, σχηματίζοντας με αυτόν τον τρόπο μια λανθασμένη εντύπωση για το περιεχόμενο του κειμένου [56]. Επιπλέον, ο περιορισμός στην απεικόνιση σύνθετων δεδομένων, λόγω των πολύπλοκων συνόλων με μεγάλο λεξιλόγιο συχνά δυσκολεύει την ακρίβεια. Τελικά, το αποτέλεσμα κρίνεται περίπλοκο και μη κατατοπιστικό. Λειτουργούν καλύτερα όταν εφαρμόζονται σε μικρότερα και πιο απλά σύνολα δεδομένων. Παρόλο που το Word Cloud είναι βοηθητικό για μια πρώτη, βασική εξερεύνηση του κειμένου, δεν καταφέρνει να αποτυπώσουν τις λεπτομερώς και με βαθύτερη σημασία τα πολύπλοκα δεδομένα που μπορεί να υπάρχουν. Ωστόσο, η απλότητά τους και η ικανότητά τους να μετατρέπουν σύνθετα δεδομένα σε μια απλή και κατανοητή οπτική απεικόνιση το καθιστούν χρήσιμο σε διάφορους τομείς, από την έρευνα έως τη διαφήμιση και την ανάλυση περιεχομένου [57].

5.4 Εφαρμογή Word Cloud στην Ανάλυση Συναισθήματος

Η χρήση του Word Cloud είναι συχνά σχετική με την ανάλυση συναισθήματος. Αποκτά ολοένα και μεγαλύτερη σημασία λόγω της αύξησης της διαδικτυακής δραστηριότητας. Η εφαρμογή τους παρατηρείται σε διάφορους τύπους εργασιών και στρατηγικών. Ο λόγος που η ανάλυση συναισθήματος είναι σημαντική έγκειται στο ότι προσφέρει μία αποδοτική μέθοδο αντίληψης και εξήγησης των αυθεντικών συναισθημάτων των ατόμων. Δεν είναι τυχαίο ότι η ανάλυση συναισθήματος αποκτά ολοένα και μεγαλύτερη δημοτικότητα στον τομέα της επεξεργασίας φυσικής γλώσσας, προσελκύοντας τα τελευταία χρόνια το ενδιαφέρον πλήθους ερευνητών [55]. Πολλές μέθοδοι για την ανάλυση συναισθήματος σε κείμενα έχουν προταθεί στη βιβλιογραφία και χρησιμοποιούνται σε ποικίλες εφαρμογές, όπως η ανάλυση των συναισθημάτων στα tweets του Twitter. Εκτός από την εφαρμογή της στα μέσα κοινωνικής δικτύωσης, η ανάλυση συναισθήματος βρίσκει εκτεταμένη χρήση και σε άλλους τομείς. Στον τομέα του μάρκετινγκ, οι εταιρίες τη χρησιμοποιούν για να αναπτύξουν στρατηγικές, να κατανοήσουν τις αντιδράσεις των χρηστών απέναντι στα προϊόντα ή τις επωνυμίες τους, να αξιολογήσουν την αποδοχή ή απόρριψη των καμπανιών και των νέων προϊόντων,

καθώς και να εντοπίσουν τους λόγους για τους οποίους οι καταναλωτές δεν πραγματοποιούν αγορές [58]. Καθώς επίσης και στον πολιτικό τομέα στον οποίο η ανάλυση συναισθήματος χρησιμοποιείται για την παρακολούθηση των πολιτικών απόψεων του κοινού, τον εντοπισμό αντιφάσεων και ασυνέπειας μεταξύ δημόσιων δηλώσεων και δράσεων των πολιτικών προσώπων, ενώ μπορεί επίσης να συμβάλλει στην εκτίμηση των εκλογικών αποτελεσμάτων με βάση τις τάσεις των ψηφοφόρων. Επιπλέον, η ανάλυση συναισθήματος χρησιμοποιείται για την παρακολούθηση και αξιολόγηση κοινωνικών φαινομένων, αναγνωρίζοντας καταστάσεις που μπορεί να προκαλέσουν κινδύνους ή αναταραχές, και βοηθά στην καταγραφή της γενικής στάσης του κοινού ή των bloggers σε διάφορα θέματα. Οι εν λόγω εφαρμογές επιβεβαιώνουν την έκταση και τη χρησιμότητα της ανάλυσης συναισθήματος σε ποικιλόμορφους τομείς, καθιστώντας την ένα σημαντικό εργαλείο. Για παράδειγμα, σε μια μελέτη περίπτωσης μάρκετινγκ θα ήταν ενδιαφέρον για μια ομάδα εξυπηρέτησης πελατών να απεικονίσει ποιες λέξεις χρησιμοποιούνται συχνότερα σε θετικά και αρνητικά σχόλια για προϊόντα που διατίθενται στην αγορά. Αυτό τους επιτρέπει να επικοινωνούν πιο αποτελεσματικά και να ανταποκρίνονται καλύτερα σε διαφορετικά ερωτήματα. Αφού πραγματοποιήσουν μια ανάλυση συναισθήματος σχετικά με τα σχόλια που έχουν στη διάθεσή τους, το σύννεφο λέξεων θα υποστηρίξει αυτήν την αρχική ανάλυση, επιτρέποντας στις ομάδες να δουν ποιες λέξεις είναι οι πιο συχνές και έχουν οδηγήσει στην ταξινόμησή τους. Δηλαδή, εάν οι ομάδες πελατών παρατηρήσουν ότι οι λέξεις που εμφανίζονται περισσότερο στα αρνητικά σχόλια έχουν ισχυρή σχέση με την υπηρεσία παράδοσης: "καθυστερήσεις", "κόστος", "ταχυδρομικά τέλη". Αυτός είναι ένας πολύ απλός τρόπος προσανατολισμού της καθοδήγησης και της λήψης αποφάσεων, όπως η απόφαση μείωσης του κόστους αποστολής ή αλλαγής μεταφορέα. Με τη βοήθεια σύννεφων λέξεων, οι ομάδες θα μπορούν εύκολα να δικαιολογήσουν τις επιλογές τους σε άλλες ομάδες χωρίς πρόσθετη εργασία [59].

6 Η Γλώσσα Προγραμματισμού R στην Ανάλυση Δεδομένων

6.1 Εισαγωγή στη Γλώσσα R

Η R είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου και ένα περιβάλλον για ανάλυση δεδομένων και γραφικά. Ο σχεδιασμός της R επηρεάστηκε έντονα από δύο υπάρχουσες γλώσσες: τις S των Becker, Chambers και Wilks και τη Scheme του Sussman. Η προκύπτουσα γλώσσα μοιάζει πολύ στην εμφάνιση με την S, αλλά η υποκείμενη υλοποίηση και σημασιολογία προέρχεται από τη Scheme [60]. Με το ενδιαφέρον για τη γλώσσα να αυξάνεται, όπως φαίνεται στους δείκτες δημοτικότητας γλωσσών, η R εμφανίστηκε για πρώτη φορά τη δεκαετία του 1990 και υπηρέτησε ως υλοποίηση της γλώσσας προγραμματισμού S για στατιστικά. Η R έγινε πιο γρήγορη με την πάροδο του χρόνου και λειτουργεί ως μια γλώσσα σύνδεσης για τη σύνθεση διαφόρων συνόλων δεδομένων, εργαλείων ή πακέτων λογισμικού [61]. Η γλώσσα προγραμματισμού R είναι ένα σημαντικό εργαλείο για την ανάπτυξη στον τομέα της αριθμητικής ανάλυσης και της μηχανικής μάθησης. Καθώς οι μηχανές γίνονται ολοένα και πιο σημαντικές ως παραγωγοί δεδομένων, αναμένεται η δημοτικότητα της γλώσσας να αυξάνεται. Η R είναι ένα σύστημα στατιστικών υπολογιστών και γραφικών. Αυτό το σύστημα αποτελείται από δυο διακριτά μέρη: την ίδια τη γλώσσα R (που είναι αυτό που εννοούν οι περισσότεροι όταν μιλούν για R) και ένα περιβάλλον χρόνου εκτέλεσης. Η R είναι μια διερμηνευμένη γλώσσα, που σημαίνει ότι οι χρήστες μπορούν να εκτελούν εντολές και να χρησιμοποιούν τις λειτουργίες της άμεσα μέσω ενός διερμηνέα γραμμής εντολών. Σε αντίθεση με γλώσσες όπως η Python και η Java, η R δεν είναι μια γλώσσα προγραμματισμού γενικής χρήσης. Αντίθετα, θεωρείται γλώσσα συγκεκριμένης περιοχής (DSL), που σημαίνει ότι οι λειτουργίες είναι προσαρμοσμένες σε έναν συγκεκριμένο τομέα ή πεδίο εφαρμογής. Στην περίπτωση της R, αυτό είναι στατιστικός υπολογισμός και ανάλυση. Ως εκ τούτου, η R χρησιμοποιείται κυρίως για διάφορες εργασίες στην επιστήμη δεδομένων. Επιπρόσθετα, είναι εξοπλισμένη με ένα μεγάλο σύνολο λειτουργιών που επιτρέπουν οπτικοποιήσεις δεδομένων, ώστε οι χρήστες να έχουν τη δυνατότητα να αναλύουν δεδομένα, να τα μοντελοποιούν όπως απαιτείται και στη συνέχεια να δημιουργούν γραφήματα πίτας ή ιστογράμματα [62]. Εν κατακλείδι, παρέχει ποικίλα πακέτα (βιβλιοθήκες συναρτήσεων) που δίνουν τη δυνατότητα να

χρησιμοποιούνται για την επίλυση διαφόρων προβλημάτων. Εκτός από τις ενσωματωμένες γραφικές λειτουργίες της γλώσσας, υπάρχουν πολλά πρόσθετα ή ενότητες που διευκολύνουν αυτό[63].

6.2 Ο Ρόλος της R στην Ανάλυση Δεδομένων

Η R είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου που χρησιμοποιείται ευρέως για ανάλυση δεδομένων και στατιστικούς υπολογισμούς. Ο βασικός λόγος για την ευρεία χρήση της μεταξύ των αναλυτών δεδομένων είναι η ικανότητά της να διαχειρίζεται και να επεξεργάζεται μεγάλες ποσότητες δεδομένων με ταχύτητα και αποτελεσματικότητα. Αυτό επιτυγχάνεται μέσω αποτελεσματικών αλγορίθμων και ενσωματωμένων λειτουργιών που δίνουν την πρόσβαση στους χρήστες να χειρίζονται γρήγορα, να αναλύουν και να οπτικοποιούν μεγάλα σύνολα δεδομένων .Επιπλέον, η R είναι διαδεδομένη για την ικανότητά της να δημιουργεί οπτικοποιήσεις υψηλής ποιότητας, κάτι που είναι σημαντικό για την αποτελεσματική μετάδοση πληροφοριών από την ανάλυση δεδομένων. Οι απεικονίσεις που παράγονται μπορούν να προσαρμοστούν με ευκολία ώστε να ανταποκρίνονται στις συγκεκριμένες ανάγκες του χρήστη, ενώ τα παραγόμενα αποτελέσματα μπορούν να διαμοιραστούν εύκολα με άλλους [2].

6.3 Παραδείγματα εφαρμογών

6.3.1 1ο Παράδειγμα

Βήμα 1: Εγκατάσταση και φόρτωση απαραίτητων βιβλιοθηκών.

Πραγματοποιείται εγκατάσταση και φόρτωση όλων των απαραίτητων πακέτων για επεξεργασία φυσικής γλώσσας, ανάλυση συναισθήματος και θεματική μοντελοποίηση.

```
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "cld2", "stopwords", "topicmodels", "tm", "LDAvis"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(cld2)
```

```
library(stopwords)
library(topicmodels)
library(tm)
library(LDAvis)
```

Βήμα 2: Φόρτωση του συνόλου δεδομένων

Εισάγεται αρχείο .csv με σχόλια χρηστών σχετικά με το Candy Crush Saga.

```
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/Candy Crush Saga.csv")
```

Βήμα 3: Επιλογή στήλης κειμένου και προσθήκη μοναδικού αναγνωριστικού

Επιλέγεται μόνο η στήλη content και προστίθεται αριθμητικό id σε κάθε γραμμή.

```
text_data <- df %>% select(content) %>% mutate(id = row_number())
```

Βήμα 4: Αφαίρεση ελλιπών εγγραφών

Αφαιρούνται οι εγγραφές με κενές τιμές (NA).

```
text_data <- text_data %>% drop_na()
```

Βήμα 5: Καθαρισμός κειμένου

Όλα τα γράμματα μετατρέπονται σε πεζά και αφαιρείται η στίξη.

```
text_data <- text_data %>% mutate(clean_text = tolower(content))
text_data <- text_data %>% mutate(clean_text = gsub("[[:punct:]]", "", clean_text))
```

Βήμα 6: Tokenization (διαχωρισμός λέξεων)

Το καθαρό κείμενο διασπάται σε μεμονωμένες λέξεις.

```
text_tokens <- text_data %>% unnest_tokens(word, clean_text)
```

Βήμα 7: Αφαίρεση άχρηστων λέξεων (custom stopwords)

Αφαιρούνται λέξεις που δεν προσφέρουν πληροφορία στην ανάλυση (π.χ. “app”, “use”).

```
custom_stopwords <- stopwords("app", "follow", "op", "posts", "social", "media", "time", "use", "h")
```

```
text_tokens <- text_tokens %>% filter(!word %in% custom_stopwords)
```

Βήμα 8: Ανάλυση Συναισθήματος με Bing Lexicon

Συνδυάζονται οι λέξεις του dataset με το λεξικό Bing που χαρακτηρίζει τις λέξεις ως "positive" ή "negative".

```
bing <- get_sentiments("bing")
sentiment_scores <- text_tokens %>%
inner_join(bing, by = "word") %>%
count(word, sentiment, sort = TRUE)
```

Βήμα 9: Οπτικοποίηση κατανομής συναισθήματος

Δημιουργείται γράφημα που δείχνει πόσες λέξεις έχουν θετικό ή αρνητικό πρόσημο.

```
sentiment_plot <- sentiment_scores %>%
  count(sentiment) %>%
  ggplot(aes(x = sentiment, y = n, fill = sentiment)) +
  geom_col() +
  theme_minimal() +
  labs(title = "Sentiment Analysis of Reviews", x = "Sentiment", y = "Frequency")
print(sentiment_plot)
```

Βήμα 10: Word Cloud

Οπτική αναπαράσταση λέξεων με διαφορετικά χρώματα για θετικά και αρνητικά συναισθήματα.

```
set.seed(1234)
wordcloud(
  words = sentiment_scores$word,
  freq = sentiment_scores$n,
  min.freq = 2,
  max.words = 100,
  colors = ifelse(sentiment_scores$sentiment == "positive", "blue", "red"),
  random.order = FALSE
)
```

-

Βήμα 11: Δημιουργία Document-Term Matrix (DTM)

Κατασκευάζεται πίνακας όπου κάθε γραμμή είναι ένα έγγραφο και κάθε στήλη μία λέξη.

```
dtm <- text_tokens %>% count(id, word) %>% cast_dtm(document = id, term = word, value = n)
```

Βήμα 12: Εκπαίδευση μοντέλου LDA με 5 θέματα

Γίνεται εκπαίδευση θεματικού μοντέλου (Latent Dirichlet Allocation) για εξαγωγή θεμάτων.

```
k <- 5
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
```

Βήμα 13: Εξαγωγή κορυφαίων λέξεων ανά θέμα

Ανάλυση των πιο χαρακτηριστικών λέξεων για κάθε θέμα.

```
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
group_by(topic) %>%
 top_n(10, beta) %>%
 ungroup()
```

Βήμα 14: Οπτικοποίηση θεμάτων
Γράφημα με τις 10 πιο σημαντικές λέξεις για κάθε θέμα που αναγνωρίστηκε από το LDA.

```
lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
 geom_col(show.legend = FALSE) +
 facet_wrap(~ topic, scales = "free") +
 coord_flip() +
 labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
print(lda_plot)
```

Βήμα 15: Διαδραστική απεικόνιση θεμάτων με LDAvis
Δημιουργείται διαδραστικό γράφημα για να εξερευνήσεις την κατανομή των θεμάτων.

```
json_lda <- LDAvis::createJSON(
 phi = posterior(lda_model)$terms,
 theta = posterior(lda_model)$topics,
 doc.length = rowSums(as.matrix(dtm)),
 vocab = colnames(as.matrix(dtm)),
 term.frequency = colSums(as.matrix(dtm))
)
LDAvis::serVis(json_lda)
```

6.3.2 2ο Παράδειγμα

Βήμα 1: Εγκατάσταση και φόρτωση απαραίτητων βιβλιοθηκών

Πραγματοποιείται εγκατάσταση και φόρτωση όλων των απαραίτητων πακέτων για προεπεξεργασία κειμένου, ανάλυση συναισθήματος και θεματική μοντελοποίηση.

```
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "topicmodels", "tm", "LDAvis", "stringi"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(topicmodels)
library(tm)
library(LDAvis)
library(stringi)
```

Βήμα 2: Φόρτωση του συνόλου δεδομένων με διόρθωση κωδικοποίησης

Εισάγεται το αρχείο .csv με tweets από τη Σρι Λάνκα και διορθώνεται η κωδικοποίηση χαρακτήρων.

```
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/SriLankaTweets.csv",
         locale = locale(encoding = "ISO-8859-1"))
```

Βήμα 3: Αφαίρεση ελλιπών εγγραφών

Αφαιρούνται οι εγγραφές με κενές τιμές στο πεδίο tweet.

```
df <- df %>% drop_na(tweet)
```

Βήμα 4: Καθαρισμός κειμένου

Μετατροπή κειμένου σε UTF-8, πεζά γράμματα, αφαίρεση σημείων στίξης και κενών.

```
df <- df %>% mutate(
  clean_text = stri_encode(tweet, "", "UTF-8"),
  clean_text = tolower(clean_text),
  clean_text = str_replace_all(clean_text, "[[:punct:]]", ""),
```

```
  clean_text = str_squish(clean_text)
)
```

Βήμα 5: Tokenization (διαχωρισμός λέξεων)

Το καθαρό κείμενο διασπάται σε μεμονωμένες λέξεις.

```
tokens <- df %>%
  filter(nchar(clean_text) > 1) %>%
  unnest_tokens(word, clean_text, token = "words")
```

Βήμα 6: Αφαίρεση λέξεων χωρίς σημασιολογική πληροφορία

Αφαιρούνται κοινές αγγλικές stopwords που δεν προσφέρουν χρήσιμη πληροφορία.

```
custom_stopwords <- tibble(word = stopwords("en"))
tokens <- tokens %>% anti_join(custom_stopwords, by = "word")
```

Βήμα 7: Ανάλυση Συναισθήματος με Bing Lexicon

Οι λέξεις συσχετίζονται με το λεξικό Bing για χαρακτηρισμό θετικού ή αρνητικού συναισθήματος.

```
bing <- get_sentiments("bing") %>% rename(sentiment_category = sentiment)
tokens <- tokens %>% inner_join(bing, by = "word")
```

Βήμα 8: Οπτικοποίηση Συναισθηματικής Συχνότητας με Wordcloud

Δημιουργείται word cloud με λέξεις ανάλογα με τη συναισθηματική τους κατηγορία.

```
sentiment_scores <- tokens %>% count(sentiment_category, word, sort = TRUE)
set.seed(1234)
wordcloud(
  words = sentiment_scores$word,
  freq = sentiment_scores$n,
  min.freq = 2,
  max.words = 100,
  colors = ifelse(sentiment_scores$sentiment_category == "positive", "blue",
                  ifelse(sentiment_scores$sentiment_category == "negative", "red", "gray")),
  random.order = FALSE
```

```
)
```

Βήμα 9: Δημιουργία Document-Term Matrix (DTM)

Κατασκευάζεται πίνακας συχνοτήτων λέξεων για κάθε tweet ή ID.

```
if ("id" %in% colnames(df)) {
  dtm <- tokens %>% count(id, word) %>% cast_dtm(document = id, term = word, value = n)
} else {
  dtm <- tokens %>% count(tweet, word) %>% cast_dtm(document = tweet, term = word, value = n)
}
```

Βήμα 10: Εκπαίδευση μοντέλου LDA με 3 θέματα

Γίνεται εκπαίδευση του θεματικού μοντέλου LDA για εξαγωγή θεμάτων.

```
k <- 3
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
```

Βήμα 11: Εξαγωγή κορυφαίων λέξεων ανά θέμα

Ανακτώνται οι 10 πιο χαρακτηριστικές λέξεις για κάθε θέμα.

```
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()
```

Βήμα 12: Οπτικοποίηση θεμάτων

Γράφημα με τις σημαντικότερες λέξεις κάθε θέματος.

```
lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
print(lda_plot)
```

Βήμα 13: Διαδραστική απεικόνιση θεμάτων με LDAvis

Δημιουργείται διαδραστικό γράφημα για εξερεύνηση θεμάτων.

```
tryCatch({
 json_lda <- LDAvis::createJSON(
  phi = posterior(lda_model)$terms,
  theta = posterior(lda_model)$topics,
  doc.length = rowSums(as.matrix(dtm)),
  vocab = colnames(as.matrix(dtm)),
  term.frequency = colSums(as.matrix(dtm))
 )
 LDAvis::serVis(json_lda)
}, error = function(e) {
 print("Error in LDA visualization:")
 print(e$message)
})
```

6.4 Πλεονεκτήματα και Μειονεκτήματα της R

6.4.1 Πλεονεκτήματα της R

Στον σημερινή εποχή, το μεγαλύτερο ποσοστό τοπικών και νεοσύστατων επιχειρήσεων και εταιριών, χρησιμοποιούν ορισμένα δεδομένα για την ανάλυση των προτύπων του παρελθόντος και για την εκτίμηση των μελλοντικών τάσεων. Γι' αυτό λοιπόν, καθώς το σύνολο δεδομένων αυξάνεται και ο όγκος των πληροφοριών επεκτείνεται, οι εταιρείες και οι επιχειρήσεις είναι αναγκαίο να έχουν ένα εργαλείο που τους βοηθά στην κατανόηση των αριθμών. Η γλώσσα R είναι ικανή να προσφέρει στον τομέα της ανάλυσης δεδομένων και της στατιστικής επεξεργασίας με την προσαρμοστικότητα και την ευκολία στη χρήση της να την καθιστούν ιδιαίτερα ελκυστική για ένα μεγάλο πλαίσιο εφαρμογών. Στη συνέχεια, θα παρουσιαστούν τα βασικά πλεονεκτήματα της, τα οποία την αναφέρουν ως ένα από τα πιο δημοφιλή εργαλεία επεξεργασίας και ανάλυσης δεδομένων. Ένα από τα πιο σημαντικά πλεονεκτήματα είναι το πλούσιο οικοσύστημα πακέτων της, το οποίο περιλαμβάνει μια πληθώρα συλλογή εργαλείων διαθέσιμων σε CRAN, Bioconductor και GitHub (πλατφόρμες που σχετίζονται με την ανάπτυξη και διαχείριση πακέτων για τη γλώσσα R, προσφέροντας εργαλεία και κοινότητες για ανάλυση δεδομένων, βιοπληροφορική και ανοιχτό λογισμικό αντίστοιχα) [64],[65],[66].

Αυτά τα πακέτα καλύπτουν ένα μεγάλο φάσμα ανάλυσης δεδομένων και στατιστικών εργασιών, παρέχοντας έτοιμες προς χρήση λειτουργίες και εργαλεία, εξοικονομώντας χρόνο και προσπάθεια. Επιπρόσθετα, η R ξεχωρίζει για την δυναμική της βάση στη στατιστική ανάλυση, καθιστώντας τη ιδανική για την εκτέλεση πολλαπλών στατιστικών αναλύσεων, όπως γραμμική και μη γραμμική μοντελοποίηση, ανάλυση χρονοσειρών και δοκιμές υποθέσεων. Η γλώσσα R προσφέρει ενσωματωμένες βιβλιοθήκες και λειτουργίες που επιτρέπουν την εύκολη πρόσβαση σε σύνθετες στατιστικές διαδικασίες. Ένα ακόμα πλεονέκτημα αλλά και η πιο σημαντική λειτουργία για την παρούσα εργασία αποτελεί η απεικόνιση του σύνολο δεδομένων με σαφή και κατανοητό τρόπο [67]. Ουσιαστικά, η R έχει δυνατότητες οπτικοποίησης δεδομένων που παρέχονται από βιβλιοθήκες όπως το ggplot2, οι οποίες επιτρέπουν τη δημιουργία ευέλικτων και ευπαρουσίαστων γραφημάτων και άλλων απεικονίσεων. Έτσι, διευκολύνει την κοινοποίηση των αποτελεσμάτων στους ενδιαφερόμενους. Επιπλέον, η R επωφελείται από μια ενεργή και συνεχώς αναπτυσσόμενη κοινότητα χρηστών και προγραμματιστών, που προάγουν τη συνεργασία και την ανταλλαγή γνώσεων και ιδεών. Αυτή η κοινότητα βοηθά στην λειτουργία νέων πακέτων και λειτουργιών και παρέχει συνεχή υποστήριξη στους χρήστες της. Αναφορικά με την ερευνητική διαδικασία, η R επιτρέπει την αναπαραγώγιμη έρευνα μέσω εργαλείων όπως το RMarkdown και το knitr, τα οποία συνδυάζουν κώδικα, δεδομένα και αποτελέσματα σε ένα ενιαίο έγγραφο, διευκολύνοντας τη διασφάλιση ότι τα αποτελέσματα μπορούν εύκολα να αναπαραχθούν και να κοινοποιηθούν. Η ευελιξία της R, σε συνδυασμό με την ικανότητά της να διασυνδέεται με άλλες γλώσσες προγραμματισμού όπως C, Python και Java, την καθιστά ιδανική για προσαρμοσμένες λειτουργίες και αναλύσεις που καλύπτουν τις συγκεκριμένες ανάγκες των χρηστών. Αξίζει να τονιστεί η χρήση της R στον ακαδημαϊκό και ερευνητικό τομέα, όπου έχει γίνει η κύρια γλώσσα επιλογής λόγω των ισχυρών στατιστικών της δυνατοτήτων και της φύσης ανοιχτού κώδικα. Η R παρέχει επίσης δυνατότητες χειρισμού δεδομένων, επιτρέποντας στους χρήστες να εισάγουν, να καθαρίζουν και να επεξεργάζονται δεδομένα με αποτελεσματικότητα, κάτι που είναι ιδιαίτερα χρήσιμο για σύνολα δεδομένων πραγματικού κόσμου που απαιτούν προ επεξεργασία και τροποποιήσεις. Συγκεκριμένα, μέσα από φόρουμ και σεμινάρια που κάνουν οι χρήστες ενισχύεται η εμπειρία μάθησης. Υπάρχουν ιστότοποι όπως το Stack Overflow και το R-bloggers που αποτελούν πολύτιμους πόρους για να λάβει κάποιος βοήθεια και να παραμένει ενημερωμένος [68]. Για αρχάριους, η ύπαρξη γραφικών διεπαφών χρήστη (GUIs) όπως τα RStudio και R Commander καθιστούν τη γλώσσα πιο εύκολα

προσβάσιμη. Εξίσου σημαντικό πλεονέκτημα είναι ότι η R είναι μια οικονομικά αποδοτική λύση, αφού είναι ανοιχτού κώδικα, πράγμα που σημαίνει ότι είναι ελεύθερη για χρήση σε όλους χωρίς υψηλό κόστος. Τέλος, η γλώσσα R έχει εκπαιδευτικό χαρακτήρα. Με άλλα λόγια, μπορεί να διδάξει και να μάθει κανείς επιστήμη δεδομένων καθώς ο τρόπος που λειτουργεί δηλαδή με εντολές που δίνονται γραπτώς και η ποικιλία πακέτων που διαθέτει, ενθαρρύνουν τους μαθητές να μαθαίνουν με την πράξη και να πειραματίζονται [69]. Η τελευταία παρατήρηση που αποτελεί παράλληλα όφελος για την R με τους Matt Adams και Roger D. Peng, είναι ότι η γλώσσα αυτή δεν απευθύνεται αποκλειστικά σε προγραμματιστές αλλά σε οποιονδήποτε επιθυμεί να αναλύσει δεδομένα. Ο Matt Adams, επιστήμονας δεδομένων στην Code School, και ο Roger D. Peng, καθηγητής Βιοστατιστικής στο Πανεπιστήμιο Johns Hopkins και συγγραφέας του βιβλίου "R Programming for Data Science", υποστηρίζουν ότι η R είναι μια προσιτή γλώσσα για αναλυτές δεδομένων ανεξαρτήτως προγραμματιστικής εμπειρίας. Στο άρθρο "Why R? The pros and cons of the R language" του InfoWorld, αναφέρεται ότι «Η R δεν είναι μόνο για προχωρημένους προγραμματιστές». Ωστόσο, ο Adams και ο Peng βλέπουν και οι δύο την R ως προσιτή γλώσσα. «Δεν προέρχομαι από την επιστήμη των υπολογιστών και ποτέ δεν είχα φιλοδοξίες να γίνω προγραμματιστής. Η γνώση των βασικών αρχών προγραμματισμού σίγουρα βοηθά όταν προσθέτετε η R στην εργαλειοθήκη σας, αλλά δεν θα έλεγα ότι απαιτείται για να ξεκινήσετε», λέει ο Adams. Ο Roger D. Peng προσθέτει: «Δεν θα έλεγα καν ότι η R είναι για προγραμματιστές. Ταιριάζει καλύτερα σε άτομα που έχουν προβλήματα προσανατολισμένα στα δεδομένα που προσπαθούν να λύσουν, ανεξάρτητα από την ικανότητα προγραμματισμού τους». Συμπερασματικά, η γλώσσα R είναι ένα ισχυρό εργαλείο για στατιστική ανάλυση και επεξεργασία δεδομένων, κατάλληλο όχι μόνο για προγραμματιστές αλλά και για επιστήμονες, αναλυτές και άλλους επαγγελματίες που επικεντρώνονται στην ανάλυση δεδομένων [70].

6.4.2 Μειονεκτήματα της R

Δεν αποτελεί έκπληξη ότι όπως κάθε γλώσσα προγραμματισμού, η R δεν είναι απαλλαγμένη από αδυναμίες. Παρόλο που προσφέρει ισχυρά εργαλεία για στατιστική ανάλυση και επεξεργασία δεδομένων, παρουσιάζει ορισμένες προκλήσεις που αξίζει να διερευνηθούν. Η κατανόηση αυτών των ζητημάτων είναι απαραίτητη για τη σωστή αξιοποίηση των δυνατοτήτων της. Στα επόμενα σημεία, θα εξεταστούν τα βασικά μειονεκτήματα της R και οι λόγοι για τους οποίους μπορούν να επηρεάσουν την εμπειρία

χρήσης της. Μέσα από αυτή την ανάλυση, θα αποσαφηνιστούν οι βασικές πτυχές που πρέπει να λαμβάνονται υπόψη. Αρχικά, ένα σχόλιο που κινεί αρνητικά το ενδιαφέρον των χρηστών είναι πως η γλώσσα R έχει απότομη «καμπύλη μάθησης», δηλαδή ότι είναι δύσκολο να μάθει κάποιος στην αρχή, αλλά μετά γίνεται ευκολότερο. Ουσιαστικά, δείχνει τη σχέση μεταξύ χρόνου και προσπάθειας που επενδύει κάποιος και του επιπέδου δεξιοτήτων που αποκτά. Λόγω της περίπλοκης σύνταξης και των μοναδικών τύπων δεδομένων που διαθέτει, για παράδειγμα τα data frames, τα διανύσματα και οι λίστες, είναι πιθανό να δυσκολέψουν τους νέους χρήστες. Παρόλο που η R μπορεί να θεωρείται σχετικά απλή, η γνώση προηγμένων λειτουργιών και πακέτων απαιτεί ιδιαίτερη προσπάθεια. Κάτι τέτοιο μπορεί να αποτελέσει εμπόδιο για πολλούς. Ακόμη, η R συγκριτικά με άλλες γλώσσες όπως η C ή η C++ είναι πιο αργή. Αυτό αποτελεί πρόβλημα για κάποιον που δουλεύει με μεγάλο όγκο δεδομένων ή στην περίπτωση εκτέλεσης απαιτητικών εργασιών. Το παραπάνω μειονέκτημα αφορούσε γενικά την επίδοση της R. Στη συνέχεια, το σύστημα διαχείρισης μνήμης της R ενδέχεται να μην είναι πάντα αποτελεσματικό καθώς τα μεγάλα σύνολα δεδομένων και τα ενδιάμεσα αντικείμενα που δημιουργούνται κατά τη διάρκεια της ανάλυσης μπορεί να καταναλώνουν μεγάλες ποσότητες μνήμης, κάτι που μπορεί να προκαλέσει προβλήματα στη μνήμη, όπως επιβραδύνσεις και σφάλματα. Η σύνταξη της R κρίνεται περίπλοκη και δεν φαίνεται «φυσική» για εξοικειωμένους χρήστες με γλώσσες προγραμματισμού, όπως η Python και η Java. Ωστόσο, όταν οι χρήστες εξοικειωθούν με αυτήν, συχνά εκτιμούν την εκφραστικότητα και τη συνοπτικότητά της για εργασίες στατιστικής ανάλυσης και δεδομένων. Ο βασικός σχεδιασμός της R είναι η ανάλυση δεδομένων και η οπτικοποίηση, και όχι η γενική ανάπτυξη λογισμικού. Γι' αυτό δεν προσφέρει τις ίδιες δυνατότητες και την ευελιξία που προσφέρουν γλώσσες όπως η Python ή η Java για την ανάπτυξη ευρύτερων εφαρμογών. Για παράδειγμα, η δημιουργία εφαρμογών ιστού ή κινητών συσκευών είναι πιο εύκολη σε άλλες γλώσσες. Όσον αφορά τα περιβαλλοντικά θέματα και τα θέματα συμβατότητας, ο κώδικας στην R μπορεί κάποιες φορές να εμφανίζει διαφορετική συμπεριφορά σε διάφορα λειτουργικά συστήματα, προκαλώντας σφάλματα που σχετίζονται με το συγκεκριμένο περιβάλλον, τα οποία μπορεί να είναι δύσκολο να λυθούν. Είναι ιδιαίτερα σημαντική η εξασφάλιση της συμβατότητας καθώς ορισμένες φορές απαιτεί λεπτομερείς δοκιμές. Οι νέες εκδόσεις της R και των πακέτων της είναι πιθανό να προκαλέσουν ασυμβατότητες με παλαιότερους κώδικες, οδηγώντας σε αλλαγές. Αυτό μπορεί να αποτελέσει σοβαρό πρόβλημα σε έργα μακροπρόθεσμης διάρκειας, όπου η σταθερότητα του κώδικα είναι βασική [71]. Συμπληρωματικά, η R, ως

γλώσσα ανοιχτού κώδικα, ενέχεται να προκαλέσει ανησυχίες όσον αφορά την ασφάλεια των δεδομένων, ειδικά όταν χρησιμοποιείται για την επεξεργασία κρίσιμων πληροφοριών. Η προστασία του απορρήτου απαιτεί ακριβή διαχείριση και αναγνώριση των πιθανών απειλών. Καταλήγοντας, ένα ακόμη ελάττωμα της R είναι ότι δεν υπάρχει μια επίσημη εμπορική υποστήριξη που να παρέχεται από κάποιον κεντρικό φορέα. Αυτό μπορεί να δημιουργήσει πρόβλημα για οργανισμούς που εξαρτώνται σε μεγάλο βαθμό από αυτή τη γλώσσα, καθώς δεν είναι πάντα εύκολο να εντοπίσουν εξειδικευμένη υποστήριξη. Παρά τα πλεονεκτήματα της γλώσσας R αποτελεί μια από τις πιο απαιτητικές γλώσσες στον τομέα της επιστήμης των δεδομένων και υπάρχουν ορισμένα αρνητικά που είναι καλό να αναφέρονται. Ωστόσο, το μέλλον της R παραμένει θετικό, καθώς συνεχίζει να είναι εξαιρετικά αποτελεσματική για στατιστικές αναλύσεις και διαχείριση μεγάλων δεδομένων, με την αυξανόμενη ζήτηση και την αδιάκοπη εξέλιξή της στον τομέα της επιστήμης των δεδομένων [69].

7 Εφαρμογή και Υλοποίηση

Η ραγδαία αύξηση του όγκου των ψηφιακών δεδομένων καθιστά αναγκαία τη χρήση προηγμένων τεχνικών επεξεργασίας και ανάλυσης, προκειμένου να εξάγονται ουσιαστικές πληροφορίες και τάσεις από το πλήθος των διαθέσιμων κειμένων. Στο πλαίσιο της παρούσας πτυχιακής εργασίας, αξιοποιείται η τεχνική του Topic Modeling, και συγκεκριμένα ο αλγόριθμος LDA, με στόχο την ανάδειξη θεμάτων που αναδύονται μέσα από ένα μεγάλο σύνολο δεδομένων κειμένου. Παράλληλα, για την ενίσχυση της ερμηνείας των αποτελεσμάτων, εφαρμόζεται η τεχνική οπτικοποίησης Word Cloud, χρησιμοποιώντας τη γλώσσα προγραμματισμού R. Ο σκοπός της εργασίας επικεντρώνεται στην εφαρμογή του αλγορίθμου LDA στην ανάλυση κειμένων που αντλήθηκαν από την πλατφόρμα Kaggle, καθώς και στη δημιουργία οπτικοποιημένων αποτελεσμάτων μέσω συννεφιών λέξεων. Η μεθοδολογία περιλαμβάνει τη συλλογή και καθαρισμό των δεδομένων, την προεπεξεργασία των κειμένων ώστε να καταστούν κατάλληλα για ανάλυση, την εφαρμογή του LDA και, τέλος, την οπτικοποίηση των αποτελεσμάτων. Η συλλογή των δεδομένων πραγματοποιήθηκε μέσω της πλατφόρμας Kaggle, η οποία παρέχει πλούσιο υλικό ανοικτών δεδομένων για αναλύσεις. Αφού τα δεδομένα εισήχθησαν στο περιβάλλον R, ακολούθησε εκτενής διαδικασία

προεπεξεργασίας: αφαιρέθηκαν τα κενά και μη απαραίτητα πεδία με τη χρήση της συνάρτησης drop_na(), απομακρύνθηκαν τα HTML tags μέσω της str_replace_all(), ενώ όλα τα κείμενα μετατράπηκαν σε πεζά γράμματα για να διασφαλιστεί η ομοιομορφία. Επιπλέον, αφαιρέθηκαν οι μη αλφαριθμητικοί χαρακτήρες, όπως σημεία στίξης, καθώς και οι λεγόμενες "σταματημένες λέξεις" (stopwords), δηλαδή λέξεις χωρίς σημασιολογική βαρύτητα, όπως "the", "and", "in", κ.ά., μέσω της βιβλιοθήκης stopwords. Στη συνέχεια, τα προεπεξεργασμένα δεδομένα μετατράπηκαν σε Ματρίδα Όρων-Εγγράφων (Document-Term Matrix - DTM), ώστε να γίνουν κατανοητά από τον αλγόριθμο LDA. Ο LDA, μέσω του πακέτου topicmodels, εφαρμόστηκε πάνω στη ματρίδα με στόχο την εξαγωγή των θεμάτων. Ο αλγόριθμος στηρίζεται στην παραδοχή ότι κάθε έγγραφο αποτελείται από ένα μίγμα θεμάτων και ότι κάθε θέμα αποτελείται από ένα σύνολο λέξεων που σχετίζονται μεταξύ τους με διαφορετικές πιθανότητες εμφάνισης. Η επιλογή του αριθμού των θεμάτων (k) έγινε κατόπιν ανάλυσης και δοκιμών με στόχο τη βέλτιστη απόδοση και την ερμηνευσιμότητα των αποτελεσμάτων. Η κατανόηση των αποτελεσμάτων ενισχύθηκε μέσω της τεχνικής Word Cloud, η οποία προσφέρει μια εύληπτη απεικόνιση των λέξεων που σχετίζονται με κάθε θέμα. Συγκεκριμένα, για κάθε θέμα δημιουργήθηκε ένα σύννεφο λέξεων όπου οι συχνότερες λέξεις εμφανίζονται με μεγαλύτερο μέγεθος, επιτρέποντας την άμεση οπτική αναγνώριση των πιο σημαντικών όρων ανά θέμα. Η τεχνική αυτή αναδείχθηκε ιδιαιτέρως χρήσιμη για την ερμηνεία των αποτελεσμάτων που παρήγαγε ο αλγόριθμος LDA. Η εφαρμογή του Topic Modeling σε συνδυασμό με την οπτικοποίηση των αποτελεσμάτων προσέφερε ουσιαστικά ευρήματα, συμβάλλοντας στην αποκάλυψη κρυφών θεμάτων που ενυπάρχουν στα δεδομένα. Η μεθοδολογία που ακολουθήθηκε αναδεικνύεται ως ένα ισχυρό εργαλείο για την ανάλυση μεγάλου όγκου κειμενικών δεδομένων, με σημαντική εφαρμογή σε ποικίλα επιστημονικά και επαγγελματικά πεδία, όπως τα κοινωνικά δίκτυα, η πληροφορική υγείας, η επιχειρηματική ανάλυση και η δημοσιογραφία. Στο επόμενο κεφάλαιο, θα παρουσιαστούν αναλυτικά τα αποτελέσματα της ανάλυσης, συνοδευόμενα από ερμηνεία των θεμάτων που αναδείχθηκαν, καθώς και προτάσεις για πιθανές μελλοντικές επεκτάσεις και εφαρμογές της παρούσας μεθοδολογίας.

8 Συμπεράσματα και Μελλοντικές Προοπτικές

8.1 Συμπεράσματα

Η παρούσα εργασία εξετάζει τη συνδυαστική αξιοποίηση της μεθόδου LDA για την εξόρυξη θεμάτων και της ανάλυσης συναισθήματος σε κειμενικά δεδομένα, με τη συνδρομή οπτικοποιήσεων τύπου Wordclouds για την απεικόνιση των αποτελεσμάτων. Η υλοποίηση πραγματοποιήθηκε με τη γλώσσα προγραμματισμού R, αξιοποιώντας τα πλεονεκτήματα που προσφέρει στους τομείς της NLP, της στατιστικής ανάλυσης και της γραφικής απεικόνισης δεδομένων. Μέσω της ανάλυσης, η μέθοδος κατόρθωσε να ομαδοποιήσει σχετικούς όρους σε σαφώς διακριτές θεματικές ενότητες, προσφέροντας μια οργανωμένη επισκόπηση του περιεχομένου. Ο περαιτέρω συνδυασμός της LDA με την ανάλυση συναισθήματος, επέτρεψε τη σύνδεση θεματικών περιοχών με συναισθηματικούς τόνους, επιτρέποντας βαθύτερη κατανόηση των στάσεων και συναισθημάτων που εκφράζονται στο κείμενο. Τα Wordclouds αποτέλεσαν ένα ιδιαίτερα χρήσιμο εργαλείο απεικόνισης, προσφέροντας άμεση και κατανοητή παρουσίαση των συχνότερων όρων που σχετίζονται με κάθε συναίσθημα ή θεματική ενότητα. Η χρήση τους διευκόλυνε σημαντικά την ερμηνεία των αποτελεσμάτων, ενισχύοντας τη διαύγεια της ανάλυσης. Επιπλέον, κρίσιμο ρόλο διαδραμάτισε η διαδικασία προεπεξεργασίας των δεδομένων όπως η μετατροπή των χαρακτήρων σε πεζά, η αφαίρεση σημείων στίξης, η κανονικοποίηση λέξεων (tokenization) και η αφαίρεση κοινών λέξεων (stopwords) καθώς επηρέασε ουσιαστικά την ποιότητα και αξιοπιστία των αποτελεσμάτων. Η γλώσσα R, μέσω των εξειδικευμένων βιβλιοθηκών της, παρείχε ένα ευέλικτο και ολοκληρωμένο περιβάλλον για την υλοποίηση όλων των σταδίων της ανάλυσης, από τον καθαρισμό των δεδομένων μέχρι την τελική οπτικοποίηση. Συνοψίζοντας, η εργασία ανέδειξε ότι ο συνδυασμός της θεματικής μοντελοποίησης μέσω LDA με την ανάλυση συναισθήματος και την απεικόνιση μέσω Wordclouds συνιστά μια ισχυρή μεθοδολογική προσέγγιση για την εξαγωγή γνώσης και συναισθηματικών τάσεων από εκτενή σύνολα κειμενικών δεδομένων. Η προσέγγιση αυτή παρείχε πολύτιμες ευκαιρίες για την κατανόηση της κοινής γνώμης, την ανάλυση σχολίων πελατών και άλλες εφαρμογές όπου η ανάλυση κειμένου είναι απαραίτητη.

8.2 Μελλοντικές Προοπτικές

Παρά τα θετικά αποτελέσματα, η παρούσα έρευνα ανοίγει δρόμους για περαιτέρω βελτιώσεις και μελλοντικές διερευνήσεις. Αρχικά, μελλοντικές εργασίες θα μπορούσαν να ενσωματώσουν πιο εξελιγμένα μοντέλα NLP, όπως μοντέλα βασισμένα σε νευρωνικά δίκτυα και βαθιά μάθηση για την ανάλυση συναισθήματος. Τα εν λόγω μοντέλα επιτυγχάνουν συχνά ανώτερα ποσοστά ακρίβειας, καθώς διαθέτουν αυξημένη ικανότητα κατανόησης του συμφραζομένου των λέξεων και ανταποκρίνονται πιο αποτελεσματικά σε σύνθετα φαινόμενα, όπως ο σαρκασμός και η ειρωνεία, τα οποία συνιστούν σημαντική πρόκληση για τις συμβατικές μεθόδους. Σε μεταγενέστερο στάδιο θα μπορούσαν να διερευνηθούν περαιτέρω, αντί να υπάρχει η τριπολική ταξινόμηση (θετικό, αρνητικό, ουδέτερο), θα μπορούσε να διερευνηθεί η ανάλυση συναισθήματος βάσει διαστάσεων (π.χ., χαρά, λύπη, θυμός, φόβος) ή η ανάλυση συναισθήματος σε επίπεδο όψης (Aspect-Based Sentiment Analysis), η οποία εντοπίζει το συναίσθημα για συγκεκριμένα χαρακτηριστικά ή όψεις ενός προϊόντος ή μιας υπηρεσίας. Αυτό θα παρείχε πιο λεπτομερή και αξιοποιήσιμη πληροφόρηση. Ένα επόμενο βήμα θα μπορούσε να είναι η εφαρμογή των μεθόδων σε δεδομένα από διαφορετικές γλώσσες θα διεύρυνε το πεδίο εφαρμογής και θα επέτρεπε συγκριτικές αναλύσεις. Αυτό απαιτεί τη χρήση πολυγλωσσικών λεξικών συναισθήματος και μοντέλων NLP. Τέλος, στο μέλλον θα μπορούσε να εξεταστεί, πέρα από τα Wordclouds, να υπάρχουν και άλλες τεχνικές οπτικοποίησης για την αναπαράσταση των θεμάτων, των συναισθημάτων και των σχέσεων μεταξύ τους, όπως διαδραστικά γραφήματα δικτύων. Η συνεχής βελτίωση στον τομέα της επεξεργασίας φυσικής γλώσσας και της μηχανικής μάθησης υπόσχεται περαιτέρω βελτιώσεις στις τεχνικές ανάλυσης συναισθήματος και εξόρυξης θεμάτων, καθιστώντας τις ολοένα και πιο ισχυρά εργαλεία για την κατανόηση του τεράστιου όγκου κειμενικών δεδομένων που παράγονται καθημερινά.

Βιβλιογραφία

[1] *Python or R for Data Analysis: Which One Should You Learn?* (2025, February 5). Coursera. https://www.coursera.org/articles/python-or-r-for-data-analysis

[2] Mehta, P., & Pandya, D. S. (2020). *A review on sentiment analysis methodologies, practices and applications*. A-Review-On-Sentiment-Analysis-Methodologies-Practices-And-Applications.pdf

[3] Qualtrics *Sentiment Analysis and How to Use It.* https://www.qualtrics.com/experience-management/research/sentiment-analysis/

[4] *What is Sentiment Analysis.* InMoment. https://inmoment.com/blog/sentiment-analysis/

[5] *What is Sentiment Analysis? – Explained.* Amazon Web Services. https://aws.amazon.com/what-is/sentiment-analysis/

[6] *Data Science Training Programs.* Big Blue Data Academy. https://bigblue.academy/gr

[7] *What is Sentiment Analysis?* (2024, August 23). IBM. https://www.ibm.com/think/topics/sentiment-analysis

[8] Khan, L. (2019, March 29). *Sentiment Analysis with SentiStrength.* https://professorkhan.com/2019/03/29/sentiment-analysis-with-sentistrength/

[9] Mahmood, M. (2024, June 21). *The Sentiment140 Dataset: A Benchmark for Sentiment Classification*. Lexiconia-Medium. https://medium.com/lexiconia/the-sentiment140-dataset-a-benchmark-for-sentiment-classification-7f37313bf757

[10] *SenticNet.* https://sentic.net/

[11] *Sentiment Analysis Using VADER in Python.* GeeksforGeeks. https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/

[12] 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), 26–33. https://doi.org/10.5120/ijca2015905866.

[13] *14 Benefits of Sentiment Analysis.* (2022, July 27). Repustate Inc. https://www.repustate.com/blog/sentiment-analysis-benefits/

[14] *What is Sentiment Analysis?* (2024. August 1). InMoment https://inmoment.com/blog/sentiment-analysis/

[15] *4 Sentiment Analysis Accuracy Traps.* Toptal. https://www.toptal.com/deep-learning/4-sentiment-analysis-accuracy-traps

[16] *What is Sentiment Analysis? Techniques, Benefits, and Implementation.* Bright Data. https://brightdata.com/blog/web-data/sentiment-analysis-explained

[17] 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), 26–33. https://doi.org/10.5120/ijca2015905866.

[18] Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). *A comprehensive overview of topic modeling: Techniques, applications and challenges*. Neurocomputing, 628, 129638. https://doi.org/10.1016/j.neucom.2025.129638

[19] *Topic Modeling-Types, Working, Applications*. (2024, June 14). Geeksfor-Geeks. https://www.geeksforgeeks.org/what-is-topic-modeling/

[20] Mohr, J. W., & Bogdanov, P. (2013). *Introduction-Topic models: What they are and why they matter.* Poetics, 41(6), 545–569. https://doi.org/10.1016/j.po-etic.2013.10.001

[21] Cho, H.-W. (2019). *Topic Modeling. Osong Public Health and Research Perspectives*, 10(3), 115–116. https://doi.org/10.24171/j.phrp.2019.10.3.01

[22] Egger, R., & Yu, J. (2022). *A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts.* Frontiers in Sociology, 7. https://doi.org/10.3389/fsoc.2022.886498

[23] Agrawal, A., Fu, W., & Menzies, T. (2018*). What is wrong with topic modeling? And how to fix it using search-based software engineering. Information and Software Technology*, 98, 74–88. https://doi.org/10.1016/j.infsof.2018.02.005

[24] Vayansky, I., & Kumar, S. A. P. (2020). *A review of topic modeling methods. Information Systems*, 94, 101582. https://doi.org/10.1016/j.is.2020.101582

[25] Wang, J., & Zhang, X.-L. (2023*). Deep NMF topic modeling. Neurocomputing*, 515, 157–173.
https://doi.org/10.1016/j.neucom.2022.10.002

[26] *Topic Modeling: Definition, Benefits & Use Cases.* Qualtrics.
https://www.qualtrics.com/experience-management/research/topic-modeling/

[27] Mohr, J. W., & Bogdanov, P. (2013). *Introduction-Topic models: What they are and why they matter*. Poetics, 41(6), 545–569.
https://doi.org/10.1016/j.poetic.2013.10.001

[28] *What is Topic Modeling? Discuss core algorithms, work, applications, pros and cons.* (2024, October 17). AIML.com. https://aiml.com/what-is-topic-modeling/

[29] *Topic Modeling in Sentiment Analysis*.
https://d1wqtxts1xzle7.cloudfront.net/101413362/Topic_Modeling_in_Sentiment_1442-9685-3-PB-libre.pdf

[30] Budiman, A. (2023, July 7). *Sentiment Analysis and Topic Modelling of Reddit Headlines using Vader and gensim*. Data and Beyond.
https://medium.com/data-and-beyond/sentiment-analysis-and-topic-modelling-of-reddit-headlines-using-vader-and-gensim-a3ecf2062815

[31] Stryker, C., & Holdsworth, J. (2024, August 11). *What is NLP (natural language processing)?* IBM. https://www.ibm.com/think/topics/natural-language-processing

[32] Naydenova, B. (2024). *Ανάπτυξη Συστήματος Ανάλυσης Συναισθήματος με Χρήση Επεξεργασίας Φυσικής Γλώσσας*. Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών, Πανεπιστήμιο Δυτικής Αττικής.
https://polynoe.lib.uniwa.gr/xmlui/bitstream/handle/11400/7124/Naydenova_18390186.pdf?sequence=3&isAllowed=y

[33] Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). *Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. Applied Sciences*, 11(9), 3986.
https://doi.org/10.3390/app11093986

[34] De La Cruz, R. (2023, December 18). *Sentiment analysis using natural language processing (NLP).* Medium.
https://medium.com/@robdelacruz/sentiment-analysis-using-natural-language-processing-nlp-3c12b77a73ec

[35] Nikhil. (2021, June 15). *Guide to sentiment analysis using natural language processing.* Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/

[36] Μοσχολιός, Α.-Ε. (2023). *Μοντελοποίηση Θεμάτων σε Τεχνολογικές Μάρκες.* Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών, Πανεπιστήμιο Πειραιώς.
https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16408/ME2118.pdf?sequence=3

[37] Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. S. K. (2022). *Prediction of research trends using LDA based topic modeling. Global Transitions Proceedings*, 3(1), 298–304.
https://doi.org/10.1016/j.gltp.2022.03.015

[38] Anupriya, P., Karpagavalli, S. (2015). *LDA based topic modeling of journal abstracts.* IEEE. 10.1109/ICACCS.2015.7324058

[39] Murel J., Kavlakoglu E. (2024, April 22). *What is Latent Dirichlet allocation?* IBM https://www.ibm.com/think/topics/latent-dirichlet-allocation

[40] Kibe K. (2025, April 15*). Topic Modeling Using Latent Dirichlet Allocation (LDA).* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/02/topic-modeling-using-latent-dirichlet-allocation-lda/

[41] Silge, J., & Robinson, D. (2017). *Chapter 3: The importance of being earnest: Tfidf. In Text mining with R: A tidy approach.* O'Reilly Media. https://www.tidytextmining.com/tfidf

[42] Hossain, A., Karimuzzaman, M., Hossain, M. M., Rahman, A. (2021). *Text Mining and Sentiment Analysis of Newspaper Headlines*. MDPI. 12(10), 414. https://doi.org/10.3390/info12100414

[43] Mukherjee, A. (2023, August 21). *Why Are Word Clouds Inefficient and How to Improve Them?* Olvy. https://olvy.co/blog/word-clouds/

[44] Hartson, R., & Pyla, P. (2012). *The UX Book: Agile UX Design for a Quality User Experience.* Morgan Kaufmann Publishers. https://books.google.gr/books?hl=el&lr=&id=RHIGCwAAQBAJ&oi=fnd&pg=PP1&dq=ux+design&ots=EOGhHmiWxQ&sig=0FgzGZ3TH2dpBWhYjgBNmYq6uCo&redir_esc=y#v=onepage&q&f=false

[45] emperinter. (2024, July 2). *Exploring the Visual Revolution: Word Clouds – Past, Present, and Future Trends*. WordCloudMaster. https://wordcloudmaster.com/exploring-the-visual-revolution-word-clouds-past-present-and-future-trends/

[46] Nagadia, M., (2021, October 28). *Word cloud in R*. Kaagle.

https://www.kaggle.com/code/meetnagadia/word-cloud-in-r/notebook

[47] Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). *The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R.* Informatica Economică, 22(1), 25-38. 10.12948/issn14531305/22.1.2018.03

[48] Maceli, M. (2016, July 18*). Introduction to text mining with R for information professionals.* The Code4Lib Journal. https://journal.code4lib.org/articles/11626

[49] Alida. (2022, September 22). *The pros and cons of word clouds as visualizations.* The Alida Journal. https://www.alida.com/the-alida-journal/the-pros-and-cons-of-word-clouds-as-visualizations

[50] Vdr, C. (2019, October 15). Create a word cloud with R. Medium. https://medium.com/towards-data-science/create-a-word-cloud-with-r-bde3e7422e8a

[51] Kane. (2021, October 11). *What is a word cloud and what are the benefits?* Chart Attack. https://www.chartsattack.com/what-is-word-cloud/

[52] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). *Word Cloud Explorer: Text Analytics Based on Word Clouds.* IEEE. 10.1109/HICSS.2014.231

[53] WordCloud.app. (2023, July 20). *Using Word Clouds for Powerful Data Visualization*. WordCloud.App Blog. https://wordcloud.app/blog/using-word-clouds-for-powerful-data-visualization/

[54] *Word Cloud-Learn about this chart and tools to create it*. https://datavizcatalogue.com/methods/wordcloud.html

[55] Melanie. (2023, October 30). *Wordcloud: What's it all about?* DataScientest. https://datascientest.com/en/wordcloud-whats-it-all-about

[56] smith, J. (2024, May 27). *What is Word Cloud and What are the Disadvantages of Word Cloud?* Medium. https://jenny-smith.medium.com/what-is-word-cloud-and-what-are-the-disadvantages-of-word-cloud-59fb85b0d5b8

[57] Jayashankar, S., & Sridaran, R. (2016). *Moving word cloud from visual towards text analysis to endow eLearning.* 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 3493–3498. https://ieeexplore.ieee.org/abstract/document/7724914

[58] Saini, S., Punhani, R., Bathla, R., & Shukla, V. K. (2019*). Sentiment Analysis on Twitter Data using R.* International Conference on Automation, Computational and Technology Management (ICACTM), (68–72).
https://doi.org/10.1109/ICACTM.2019.8776685

[59] Tubishat, M., Al-Obeidat, F., & Shuhaiber, A. (2023). *Sentiment Analysis of Using ChatGPT in Education*. International Conference on Smart Applications, Communications and Networking (SmartNets), (1–7). https://doi.org/10.1109/SmartNets58706.2023.10215977

[60] Crawley, M. J. (2012). *The R Book*. John Wiley & Sons.
The R Book | Wiley Online Books

[61] Krill, P. (2015, June 30). *Why R? The pros and cons of the R language*. InfoWorld.
https://www.infoworld.com/article/2245310/r-programming-language-statistical-data-analysis-2.html

[62] Worsley, S. (2023, October 13). *What is R? – Introduction to Statistical Computing Power.* DataCamp. https://www.datacamp.com/blog/all-about-r

[63] *What is R?* The R Project. https://www.r-project.org/about.html

[64] *What Is CRAN In R Language?* (2024, May 10). GeeksforGeeks. https://www.geeksforgeeks.org/what-is-cran-in-r-language/

[65] *Bioconductor.* (2025). Wikipedia. https://en.wikipedia.org/w/index.php?title=Bioconductor&oldid=1285946087

[66] *What is GitHub and How to Use It?* (2024, June 30). GeeksforGeeks. https://www.geeksforgeeks.org/what-is-github-and-how-to-use-it/

[67] Daisy. *Why Data Scientists Use R: The Pros, Cons & Alternatives.* Data Science Nerd. https://datasciencenerd.com/why-data-scientists-use-r/

[68] *Pros and Cons of R Programming Language*. GeeksforGeeks. https://www.geeksforgeeks.org/pros-and-cons-of-r-programming-language/

[69] *Top Advantages and Disadvantages of R programming.* TheKnowledgeAcademy.https://www.theknowledgeacademy.com/blog/advantages-and-disadvantages-of-r-programming/

[70] Krill, P. *Why R? The pros and cons of the R language*. InfoWorld. https://www.infoworld.com/article/2245310/r-programming-language-statistical-data-analysis-2.html

[71] *Pros and Cons of R Programming Language.* (2024, May 23). GeeksforGeeks. https://www.geeksforgeeks.org/pros-and-cons-of-r-programming-language/

[72] Torkkola, K. *Linear Discriminant Analysis in Document Classification*. https://www.cse.unr.edu/~buchmann/doc_class/papers/LDA_doc_classification.pdf

[73] Yang, P., Yao, Y. & Zhou, H. (2020). *Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering.* IEEE Xplore. 10.1109/ACCESS.2020.2969525

Παράρτημα Α- Κώδικας Υλοποίησης

Στο παρόν παράρτημα παρατίθεται το σύνολο των κώδικων που αναπτύχθηκαν στο πλαίσιο της παρούσας πτυχιακής εργασίας και υλοποιήθηκαν στη γλώσσα προγραμματισμού R. Οι κώδικες χρησιμοποιήθηκαν για την προεπεξεργασία των δεδομένων, την ανάλυση και οπτικοποίησή τους, καθώς και για την εφαρμογή των μεθόδων και τεχνικών που περιγράφονται στο κυρίως σώμα της εργασίας. Η αναφορά των κωδίκων γίνεται με σκοπό τη διασφάλιση της παραγωγής των αποτελεσμάτων και την υποστήριξη της διαφάνειας της μεθοδολογίας που ακολουθήθηκε.

Candy Crush Saga

```
# Install and Load Required Libraries
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "cld2", "stopwords", "topicmodels", "tm", "LDAvis"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(cld2)
library(stopwords)
library(topicmodels)
library(tm)
library(LDAvis)
# Load the Dataset
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/Candy Crush Saga.csv")
# Select Relevant Text Column
text_data <- df %>% select(content) %>% mutate(id = row_number())
# Remove Missing Values
text_data <- text_data %>% drop_na()
# Convert to Lowercase and Remove Punctuation
text_data <- text_data %>% mutate(clean_text = tolower(content))
```

```
text_data <- text_data %>% mutate(clean_text = gsub("[[:punct:]]", "", clean_text))
# Tokenization
text_tokens <- text_data %>% unnest_tokens(word, clean_text)
# Remove English Stopwords
custom_stopwords <- stopwords("app", "follow", "op", "posts", "social", "media", "time", "use", "h" )
text_tokens <- text_tokens %>% filter(!word %in% custom_stopwords)
# Sentiment Analysis using Bing Lexicon
bing <- get_sentiments("bing")
sentiment_scores <- text_tokens %>%
inner_join(bing, by = "word") %>%
count(word, sentiment, sort = TRUE)
# Plot Sentiment Distribution
sentiment_plot <- sentiment_scores %>%
count(sentiment) %>%
 ggplot(aes(x = sentiment, y = n, fill = sentiment)) +
 geom_col() +
 theme_minimal() +
 labs(title = "Sentiment Analysis of Reviews", x = "Sentiment", y = "Frequency")
print(sentiment_plot)
# Word Cloud for Sentiment Words
set.seed(1234)
wordcloud(
 words = sentiment_scores$word,
 freq = sentiment_scores$n,
 min.freq = 2,
 max.words = 100,
 colors = ifelse(sentiment_scores$sentiment == "positive", "blue", "red"),
 random.order = FALSE
)
# Prepare Document-Term Matrix (DTM)
dtm <- text_tokens %>% count(id, word) %>% cast_dtm(document = id, term = word, value = n)
# Apply LDA Topic Modeling with 5 Topics
k <- 5
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
```

```
# Extract Topics
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
group_by(topic) %>%
 top_n(10, beta) %>%
 ungroup()
# Visualize Topics
lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
 geom_col(show.legend = FALSE) +
 facet_wrap(~ topic, scales = "free") +
 coord_flip() +
 labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
print(lda_plot)
# Interactive LDA Visualization
json_lda <- LDAvis::createJSON(
phi = posterior(lda_model)$terms,
 theta = posterior(lda_model)$topics,
 doc.length = rowSums(as.matrix(dtm)),
 vocab = colnames(as.matrix(dtm)),
 term.frequency = colSums(as.matrix(dtm))
)
LDAvis::serVis(json_lda)
```

Sri Lanka

```
# Install and Load Required Libraries
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "topicmodels", "tm", "LDAvis", "stringi"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(topicmodels)
library(tm)
library(LDAvis)
```

```
library(stringi)
# Load the Dataset with Encoding Fix
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/SriLankaTweets.csv",
         locale = locale(encoding = "ISO-8859-1"))
# Remove Missing Values
df <- df %>% drop_na(tweet)
# Preprocess Text: Encoding Fix, Lowercase, Remove Punctuation
df <- df %>% mutate(
  clean_text = stri_encode(tweet, "", "UTF-8"),
  clean_text = tolower(clean_text),
  clean_text = str_replace_all(clean_text, "[[:punct:]]", ""),
  clean_text = str_squish(clean_text)
)
# Tokenization
tokens <- df %>%
  filter(nchar(clean_text) > 1) %>%
unnest_tokens(word, clean_text, token = "words")
# Remove Stopwords
custom_stopwords <- tibble(word = stopwords("en"))
tokens <- tokens %>% anti_join(custom_stopwords, by = "word")
#  Keep only semantically meaningful words (from the Bing lexicon)
bing <- get_sentiments("bing") %>% rename(sentiment_category = sentiment)
tokens <- tokens %>% inner_join(bing, by = "word")
# Word frequencies for visualization
sentiment_scores <- tokens %>% count(sentiment_category, word, sort = TRUE)
# Wordcloud
set.seed(1234)
wordcloud(
  words = sentiment_scores$word,
  freq = sentiment_scores$n,
  min.freq = 2,
  max.words = 100,
  colors = ifelse(sentiment_scores$sentiment_category == "positive", "blue",
           ifelse(sentiment_scores$sentiment_category == "negative", "red", "gray")),
```

```
  random.order = FALSE
)
# Prepare Document-Term Matrix (DTM) for LDA
if ("id" %in% colnames(df)) {
  dtm <- tokens %>% count(id, word) %>% cast_dtm(document = id, term = word, value = n)
} else {
  dtm <- tokens %>% count(tweet, word) %>% cast_dtm(document = tweet, term = word, value = n)
}
# Apply LDA Topic Modeling with 3 Topics
k <- 3
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
# Extract Topics
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()
# Visualize Topics
lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
print(lda_plot)
# Interactive LDA Visualization
tryCatch({
  json_lda <- LDAvis::createJSON(
    phi = posterior(lda_model)$terms,
    theta = posterior(lda_model)$topics,
    doc.length = rowSums(as.matrix(dtm)),
    vocab = colnames(as.matrix(dtm)),
    term.frequency = colSums(as.matrix(dtm))
  )
  LDAvis::serVis(json_lda)
```

```
}, error = function(e) {
  print(" Error in LDA visualization: ")
  print(e$message)
})
```

Cyberbullying tweets

```
# Install and Load Required Libraries
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "topicmodels",
"tm", "LDAvis"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(topicmodels)
library(tm)
library(LDAvis)
# Load the Dataset
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/cyberbullying_tweets.csv",
          locale = locale(encoding = "ISO-8859-1"))
# Remove Missing Values
df <- df %>% drop_na()
# Convert Text Column to Lowercase and Remove Punctuation
df <- df %>%
  mutate(clean_text = iconv(tweet_text, "ISO-8859-1", "UTF-8")) %>%  # Fix encoding
  mutate(clean_text = tolower(clean_text)) %>%
  mutate(clean_text = str_remove_all(clean_text, "[[:punct:]]"))  # Remove punctuation
# Tokenization
tokens <- df %>% unnest_tokens(word, clean_text)
# Remove Stopwords
custom_stopwords <- stopwords("en")
tokens <- tokens %>% filter(!word %in% custom_stopwords)
bing <- get_sentiments("bing") %>% rename(sentiment_category = sentiment)
tokens <- tokens %>% inner_join(bing, by = "word")
# Filter out words that appear <5 times
```

```
tokens <- tokens %>%
 group_by(word) %>%
 filter(n() >= 5) %>%
 ungroup()
# Data sampling (optional)
set.seed(1234)
df_sampled <- df %>% sample_n(min(10000, nrow(df)))
# Calculation of Frequencies of Emotional Words
sentiment_scores <- tokens %>%
count(sentiment_category, word, sort = TRUE)
# If there are words, display wordcloud
if (nrow(sentiment_scores) == 0) {
 print(" No matching sentiment words found in the dataset!")
} else {
 print(" Sentiment words found:")
 print(head(sentiment_scores, 10))
 # Word Cloud
 set.seed(1234)
 wordcloud(
  words = sentiment_scores$word,
  freq = sentiment_scores$n,
  min.freq = 2,
  max.words = 100,
  colors = ifelse(sentiment_scores$sentiment_category == "positive", "blue",
          ifelse(sentiment_scores$sentiment_category == "negative", "red", "gray")),
  random.order = FALSE
 )}
# Prepare Document-Term Matrix (DTM)
if ("id" %in% colnames(df_sampled)) {
 dtm <- tokens %>% count(id, word) %>% cast_dtm(document = id, term = word, value = n)
} else {
 dtm <- tokens %>% count(tweet_text, word) %>% cast_dtm(document = tweet_text, term = word,
value = n)}
# Check if DTM is valid
```

```
if (dim(dtm)[1] == 0 | dim(dtm)[2] == 0) {
  stop(" DTM is empty! Ensure tokens contain valid words.")
}
# LDA Topic Modeling
k <- max(2, min(5, dim(dtm)[1] - 1))
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
# Extract Topics
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()
# Display Topics
if (nrow(top_terms) == 0) {
  print(" No topics extracted! Check document-term matrix.")
} else {
  print(" Topics extracted:")
  print(head(top_terms))
  # Plot Topics
  lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~ topic, scales = "free") +
    coord_flip() +
    labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
  print(lda_plot)
}
# Interactive LDA Visualization
tryCatch({
  json_lda <- LDAvis::createJSON(
    phi = posterior(lda_model)$terms,
    theta = posterior(lda_model)$topics,
    doc.length = rowSums(as.matrix(dtm)),
    vocab = colnames(as.matrix(dtm)),
    term.frequency = colSums(as.matrix(dtm))
```

```
 )
 LDAvis::serVis(json_lda)
}, error = function(e) {
 print(" Error in LDA visualization: ")
 print(e$message)
})
```

All Data

```
# Install and Load Required Libraries
install.packages(c("tidyverse", "tidytext", "ggplot2", "wordcloud", "RColorBrewer", "topicmodels", "tm", "LDAvis"))
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(topicmodels)
library(tm)
library(LDAvis)
# Load the Dataset with Encoding Fix
df <- read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy/all-data.csv",
           col_names = c("sentiment", "text"), skip = 0, locale = locale(encoding = "ISO-8859-1"))
# Remove Missing Values
df <- df %>% drop_na()
# Convert Sentiment Labels to Factors
df$sentiment <- as.factor(df$sentiment)
# Clean text: lowercase & remove punctuation
df <- df %>%
 mutate(clean_text = tolower(text),
      clean_text = gsub("[[:punct:]]", "", clean_text),
      doc_id = row_number())  # Add document ID
# Tokenization
tokens <- df %>%
 unnest_tokens(word, clean_text)
# Remove Stopwords
```

```
custom_stopwords <- stopwords("en")
tokens <- tokens %>% filter(!word %in% custom_stopwords)
# Load Bing Lexicon & Semantic Filtering
bing <- get_sentiments("bing")
tokens_sentiment <- tokens %>%
inner_join(bing, by = "word") %>%
  inner_join(df %>% select(doc_id, text), by = c("doc_id"))
# Word Cloud Data
sentiment_scores <- tokens_sentiment %>%
  count(word, sentiment, sort = TRUE)
# Word Cloud with Sentiment Colors
set.seed(1234)
wordcloud(
  words = sentiment_scores$word,
  freq = sentiment_scores$n,
  min.freq = 2,
  max.words = 100,
  colors = ifelse(sentiment_scores$sentiment == "positive", "blue", "red"),
  random.order = FALSE
)
# Document-Term Matrix using only sentiment words
dtm <- tokens_sentiment %>%
  count(doc_id, word) %>%
  cast_dtm(document = doc_id, term = word, value = n)
# LDA Topic Modeling
k <- 3
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
# Extract Topics
topics_terms <- tidy(lda_model, matrix = "beta")
top_terms <- topics_terms %>%
group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()
# Visualize Topics
```

```
lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
 geom_col(show.legend = FALSE) +
 facet_wrap(~ topic, scales = "free") +
 coord_flip() +
 labs(title = "Top Words in Each Topic (Filtered by Sentiment)", x = "Words", y = "Probability")
print(lda_plot)
# Interactive LDA Visualization
json_lda <- LDAvis::createJSON(
phi = posterior(lda_model)$terms,
 theta = posterior(lda_model)$topics,
 doc.length = rowSums(as.matrix(dtm)),
 vocab = colnames(as.matrix(dtm)),
 term.frequency = colSums(as.matrix(dtm))
)
LDAvis::serVis(json_lda)
```

Tweets

```
#  Load Required Libraries
library(tidyverse)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(topicmodels)
library(tm)
library(LDAvis)
#  Load the Dataset
df <- tryCatch({
 read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy /tweets.csv",
        locale = locale(encoding = "UTF-8"),
        guess_max = 10000)
}, error = function(e) {
 print(" Error loading CSV. Retrying with ISO-8859-1 encoding...")
 read_csv("C:/Users/rafai/Downloads/ptyxiaki/datasets - Copy /tweets.csv",
        locale = locale(encoding = "ISO-8859-1"),
```

```
        guess_max = 10000)
})
#  Display Column Names for Debugging
print(" Column names in dataset:")
print(colnames(df))
#  Ensure "text" Column Exists (Handles Variations in Column Names)
if (!"text" %in% colnames(df)) {
  stop(" Error: Column 'text' not found in dataset. Please check column names!")
}
#  Select & Rename Column for Processing
df <- df %>% select(text)
#  Remove Missing Values
df <- df %>% drop_na(text)
#  Ensure Text Column is Character Type
df <- df %>% mutate(text = as.character(text))
#  Create a Unique Row ID (for LDA document identification)
df <- df %>% mutate(row_id = row_number())
#  Fix Encoding & Text Cleaning
df <- df %>%
  mutate(clean_text = text) %>%
  mutate(clean_text = iconv(clean_text, from = "UTF-8", to = "ASCII", sub = "")) %>%
  mutate(clean_text = tolower(clean_text)) %>%
  mutate(clean_text = str_replace_all(clean_text, "<.*?>", "")) %>%
  mutate(clean_text = str_replace_all(clean_text, "[^a-zA-Z\\s#]", "")) %>%
  mutate(clean_text = str_squish(clean_text)) %>%
  mutate(clean_text = str_replace_all(clean_text, "", ""))

#  Tokenization
tokens <- df %>% unnest_tokens(word, clean_text)
#  Ensure Tokens Are Not Empty
if (nrow(tokens) == 0) {
  stop(" Tokenization failed. No valid words found!")
}
#  Remove Stopwords
```

```
custom_stopwords <- stopwords("en")
tokens <- tokens %>% filter(!word %in% custom_stopwords)
#  Keep Only Frequent Words (Lower Threshold for More Data)
tokens <- tokens %>% count(row_id, word, sort = TRUE) %>% filter(n >= 2)
#  Ensure Tokens Are Still Present
if (nrow(tokens) == 0) {
  stop(" After filtering, no words remain. Consider lowering frequency threshold.")
}
#  Sentiment Analysis using Bing Lexicon
bing <- get_sentiments("bing") %>% rename(sentiment_category = sentiment)
#  Semantic Filtering Using Bing
tokens <- tokens %>% inner_join(bing, by = "word")
#  Compute Sentiment Scores
sentiment_scores <- tokens %>%
count(sentiment_category, word, sort = TRUE)
#  Generate Sentiment-Based Word Cloud
wordcloud_words <- c()
if (nrow(sentiment_scores) > 0) {
  print(" Sentiment words found:")
  print(head(sentiment_scores, 10))
  set.seed(1234)
  wordcloud_words <- sentiment_scores$word
  wordcloud(
    words = sentiment_scores$word,
    freq = sentiment_scores$n,
    min.freq = 1,
    max.words = 100,
    colors = ifelse(sentiment_scores$sentiment_category == "positive", "blue",
                    ifelse(sentiment_scores$sentiment_category == "negative", "red", "gray")),
    random.order = FALSE
  )
} else {
  print(" No matching sentiment words found in the dataset!")
}
```

```
#  Prepare Document-Term Matrix (DTM)
dtm <- tokens %>% select(row_id, word, n) %>% cast_dtm(document = row_id, term = word, value
= n)
#  Ensure No Empty Documents Before LDA
if (sum(rowSums(as.matrix(dtm)) == 0) > 0) {
 dtm <- dtm[rowSums(as.matrix(dtm)) > 0, ]
}
#  Debug Step: Check if DTM is Valid
if (dim(dtm)[1] == 0 | dim(dtm)[2] == 0) {
 stop(" DTM is empty! Ensure tokens contain valid words.")
}
#  Choose Optimal `k` for LDA (Dynamic Selection)
k <- max(2, min(5, dim(dtm)[1] - 1))
#  Run LDA Model
lda_model <- LDA(dtm, k = k, control = list(seed = 1234))
#  Extract Topics
topics_terms <- tidy(lda_model, matrix = "beta")
# Filter topics to only include words from wordcloud
if (length(wordcloud_words) > 0) {
 topics_terms <- topics_terms %>% filter(term %in% wordcloud_words)
}
top_terms <- topics_terms %>%
 group_by(topic) %>%
 top_n(10, beta) %>%
 ungroup()
#  Generate Topic Modeling Visualization
if (nrow(top_terms) > 0) {
 print(" Topics extracted:")
 print(head(top_terms))
 lda_plot <- ggplot(top_terms, aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "Top Words in Each Topic", x = "Words", y = "Probability")
```

```
 print(lda_plot)
} else {
 print(" No topics extracted! Check document-term matrix.")
}
#  Optimized LDA Visualization (Memory-Safe)
tryCatch({
 #  Reduce Memory Usage for LDAvis by Keeping Only Top 300 Words
 term_freq <- colSums(as.matrix(dtm))
 correct_vocab <- terms(lda_model, dim(dtm)[2])
 vocab <- intersect(correct_vocab, names(term_freq))
 vocab <- gsub("\uFEFF", "", vocab)
 dtm_filtered <- dtm[, vocab, drop=FALSE]
 #  Ensure No NA Values in LDAvis Input
 phi <- posterior(lda_model)$terms[, vocab, drop=FALSE]
 theta <- posterior(lda_model)$topics
 doc_length <- rowSums(as.matrix(dtm_filtered))
 term_frequency <- term_freq[vocab]
 #  Remove Any NA & Infinite Values
 phi[is.na(phi) | is.infinite(phi)] <- 0
 theta[is.na(theta) | is.infinite(theta)] <- 0
 doc_length[is.na(doc_length) | is.infinite(doc_length)] <- 0
 term_frequency[is.na(term_frequency) | is.infinite(term_frequency)] <- 0
 json_lda <- LDAvis::createJSON(
  phi = phi,
  theta = theta,
  doc.length = doc_length,
  vocab = vocab,
  term.frequency = term_frequency
 )
 LDAvis::serVis(json_lda)
}, error = function(e) {
 print(" Error in LDA visualization: ")
 print(e$message)
})
```

Παράρτημα Β- Οπτικοποιήσεις Word Cloud

Στο παρόν παράρτημα παρουσιάζονται οι γραφικές απεικονίσεις τύπου word cloud που δημιουργήθηκαν στο πλαίσιο της ανάλυσης κειμένου, με στόχο την αποτύπωση της συχνότητας και της σχετικής σημαντικότητας των λέξεων που εντοπίστηκαν στο σύνολο των δεδομένων. Οι word clouds υλοποιήθηκαν με τη χρήση της γλώσσας R, αξιοποιώντας κατάλληλες βιβλιοθήκες για την επεξεργασία κειμένου και τη δημιουργία γραφικών. Οι οπτικοποιήσεις αυτές συμβάλλουν στην κατανόηση των κύριων θεματικών εννοιών και προσφέρουν μια συνοπτική και εύληπτη εικόνα των πιο συχνών όρων στο κείμενο.

Candy Crush Saga

Sri Lanka

Cyberbullying Tweets

All-Data

Tweets

Social Media

Elon Musk Tweets

Amazon

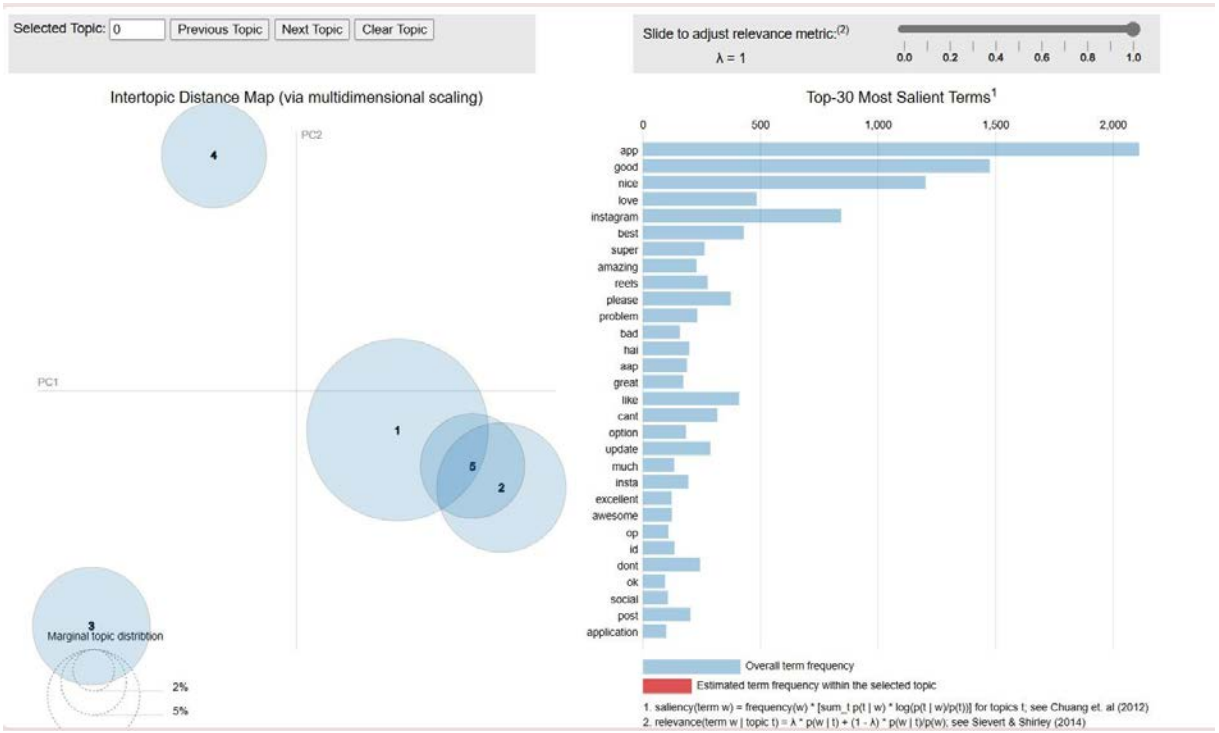
Tweet-3

IMDB

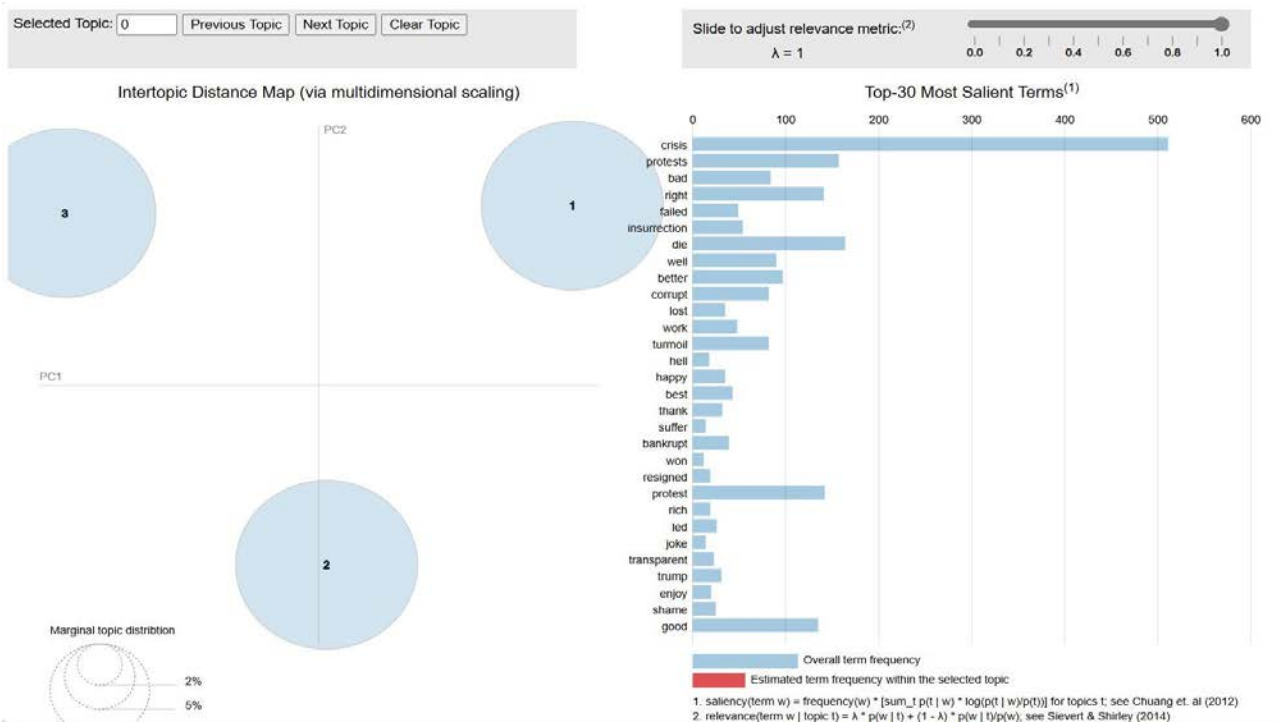
Παράρτημα Γ- Οπτικοποιήσεις LDA

Στο παρόν παράρτημα παρουσιάζονται οι οπτικοποιήσεις που προέκυψαν από την εφαρμογή της μεθόδου LDA για τη θεματική μοντελοποίηση του συνόλου κειμένων. Οι γραφικές αναπαραστάσεις δημιουργήθηκαν με χρήση της γλώσσας R, αξιοποιώντας εξειδικευμένες βιβλιοθήκες, όπως οι topicmodels, LDAvis, tidytext και ggplot2. Μέσω αυτών των οπτικοποιήσεων, αποτυπώνονται τα κυρίαρχα θέματα που εντοπίστηκαν στο σώμα των δεδομένων, καθώς και η κατανομή των λέξεων εντός κάθε θέματος. Οι απεικονίσεις αυτές προσφέρουν χρήσιμη εποπτική πληροφόρηση για τη δομή των θεμάτων και διευκολύνουν την ερμηνεία των αποτελεσμάτων της θεματικής ανάλυσης.

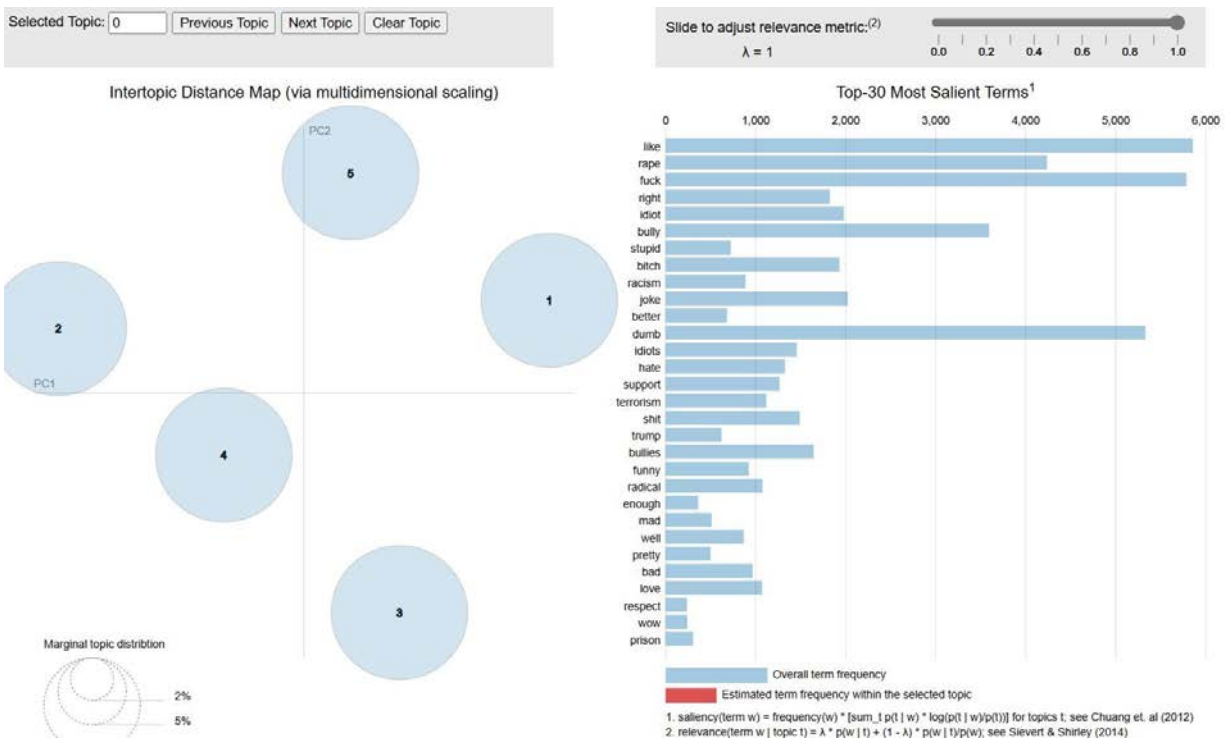
Candy Crush Saga



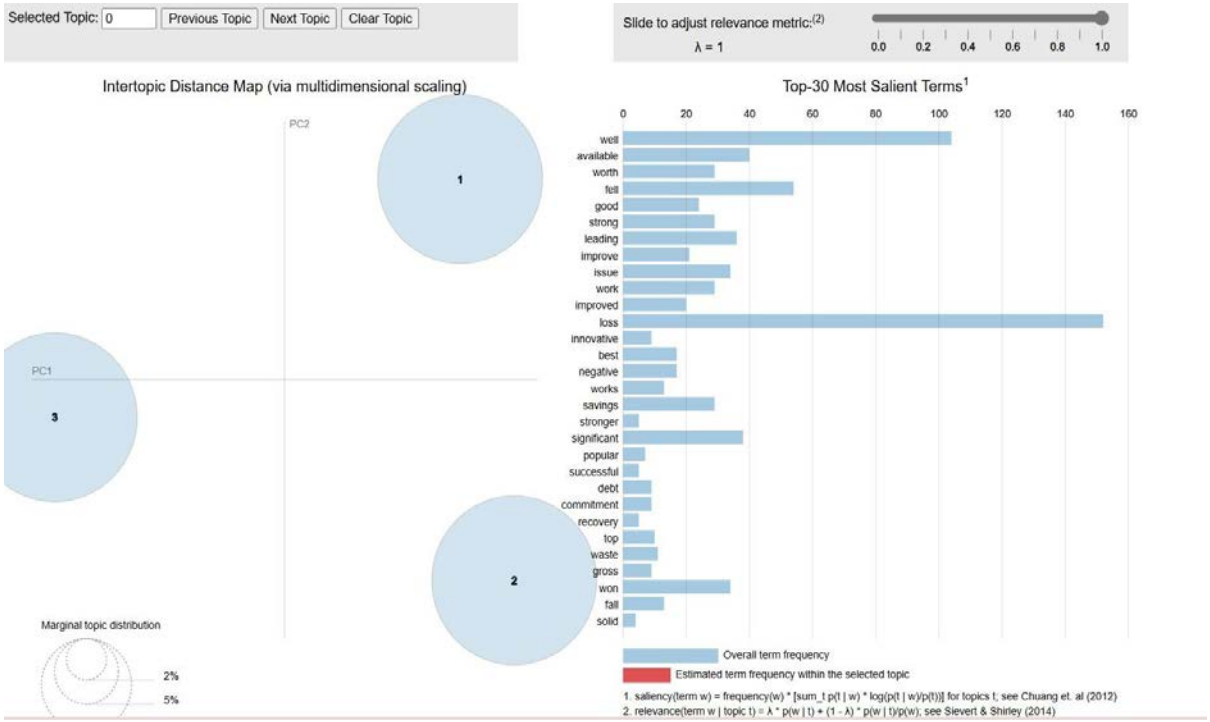
Sri Lanka



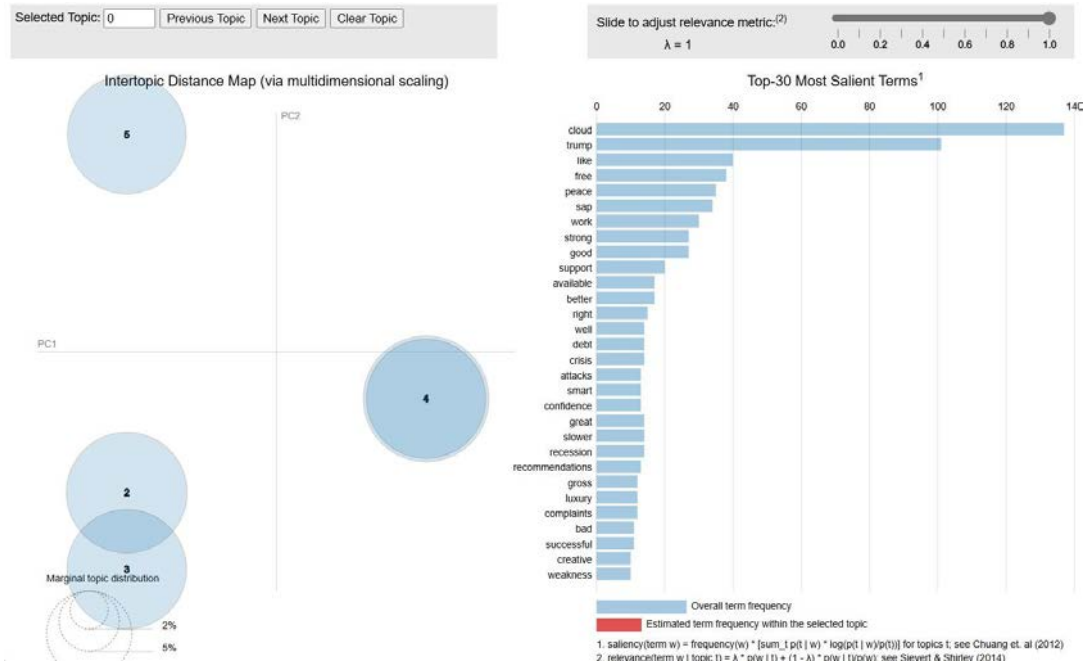
Cyberbullying Tweets



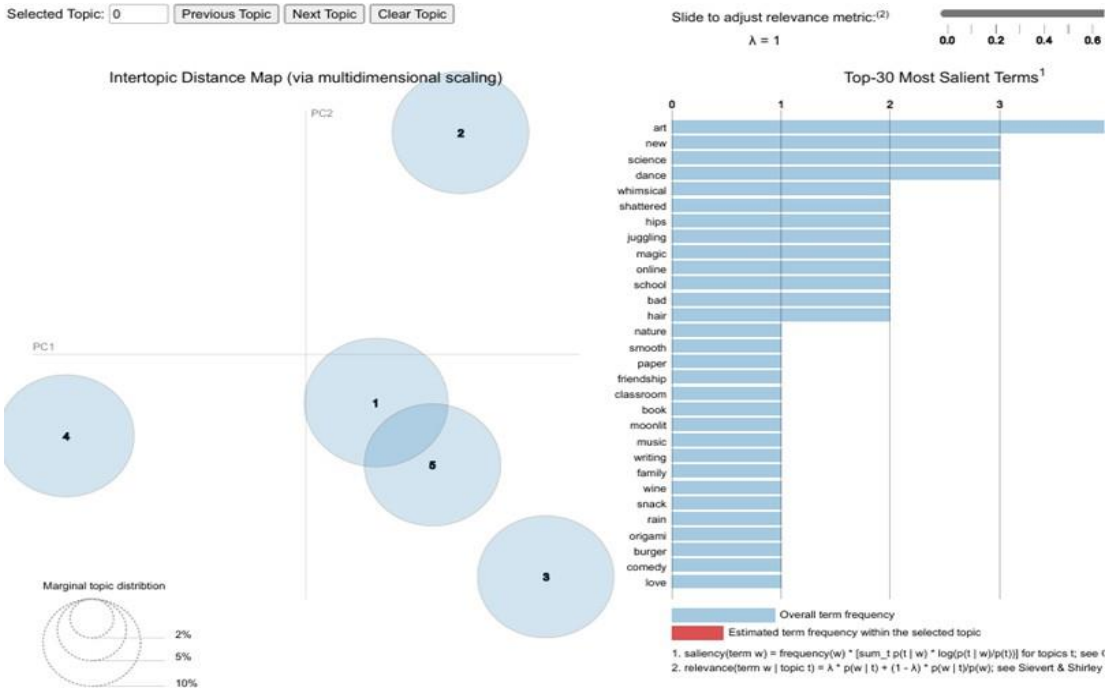
All Data



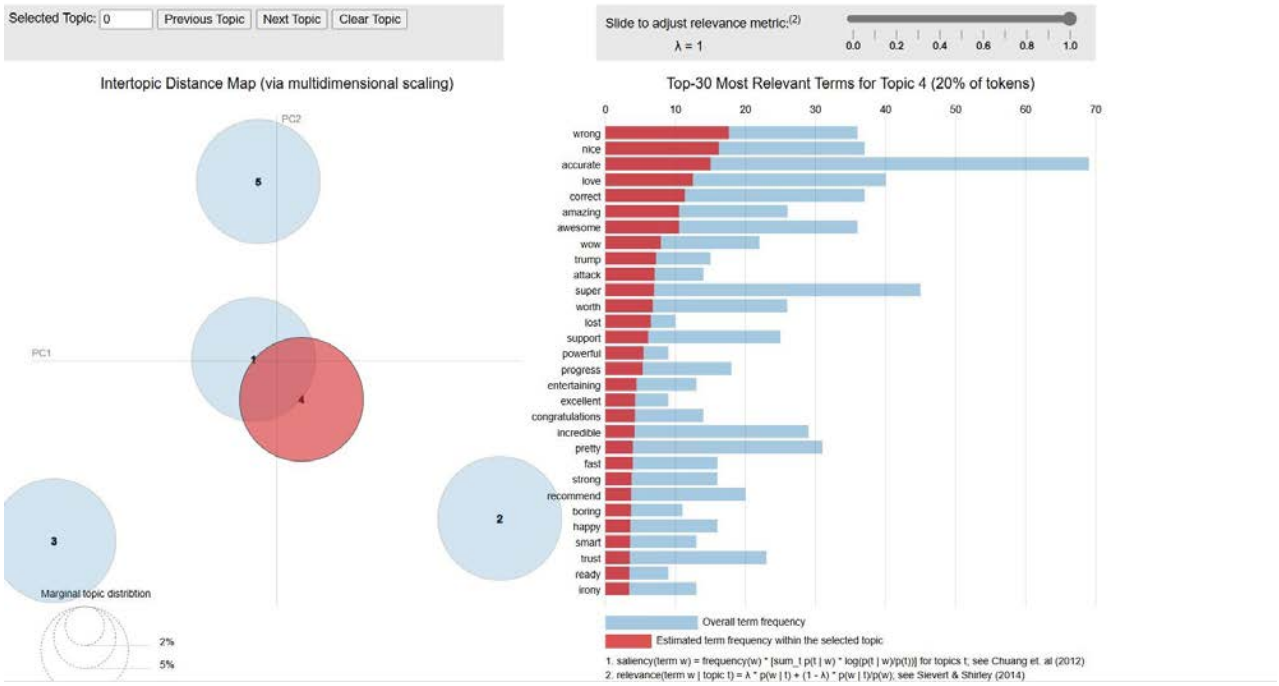
Tweets



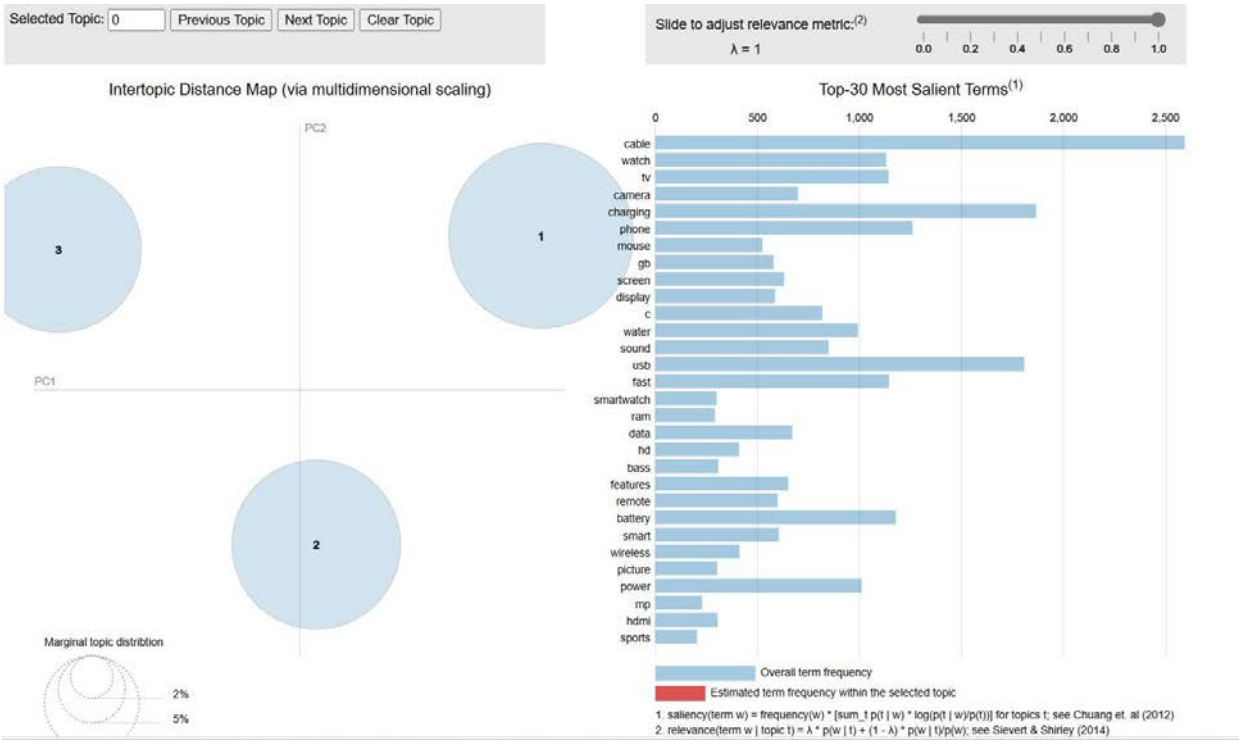
Social Media



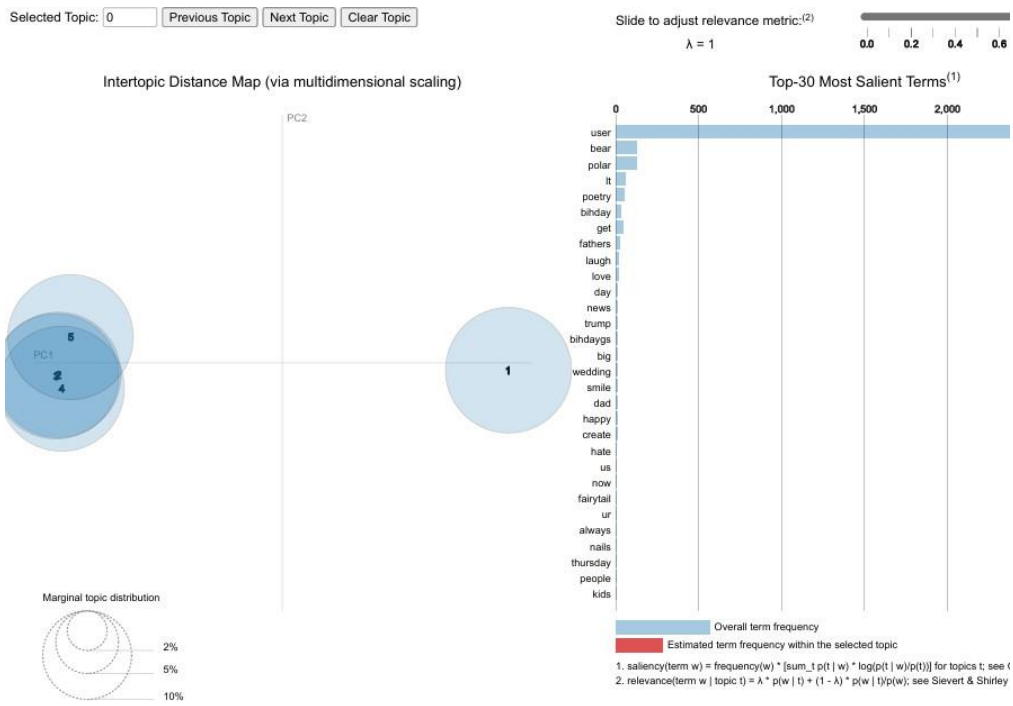
Elon Musk Tweets



Amazon



Tweet3



IMDB