



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

– ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ –

ΘΕΜΑ

Εξόρυξη Χαρακτηριστικών από Κοινωνικά Δίκτυα

Οδυσσέας Οικονόμου

Επιβλέπων Καθηγητής: Ανδρέας Καναβός

Ιούνιος 17, 2024

Τριμελής Επιτροπή

Ανδρέας Καναβός, Καθηγητής,
Ιόνιο Πανεπιστήμιο

Κάτια Λήδα Κερμανίδου, Καθηγήτρια,
Ιόνιο Πανεπιστήμιο

Αριστείδης Βραχάτης, Καθηγητής,
Ιόνιο Πανεπιστήμιο

Περίληψη

Βρισκόμαστε πλέον στην εποχή του big data, όπου τεράστιες ποσότητες δεδομένων υψηλής διάστασης γίνονται πανταχού παρόντες, σε διάφορους τομείς, όπως τα μέσα κοινωνικής δικτύωσης, η υγειονομική περίθαλψη, η βιοπληροφορική και η διαδικτυακή εκπαίδευση. Η ραγδαία αύξηση των δεδομένων παρουσιάζει προκλήσεις για την αποτελεσματική και αποδοτική διαχείριση των δεδομένων αυτών. Λόγω αυτού είναι αναγκαία η εφαρμογή τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης για την αυτοματοποίηση της εξόρυξης γνώσης από δεδομένα διαφόρων ειδών.

Η έμπνευση της παρακάτω εργασίας, είναι η δυνατότητα της επεξεργασίας φυσικής γλώσσας και των τεχνικών μηχανικής μάθησης, στο να απλοποιήσουν την ανάλυση της κοινής γνώμης που εκφράζεται στα μέσα κοινωνικής δικτύωσης.

1. Εισαγωγή	6
1.1 Πληροφορίες σχετικά με την ανάλυση συναισθήματος και την επιλογή χαρακτηριστικών	6
1.2 Επισκόπηση της σημασίας της ανάλυσης συναισθήματος για την κατανόηση της κοινής γνώμης	8
1.3 Εισαγωγή στη συγκεκριμένη ερευνητική εστίαση: Ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή	11
1.4 Η σημασία της συζήτησης για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης	11
1.5 Στόχοι της μελέτης	12
1.6 Σημασία της έρευνας	12
1.7 Μεθοδολογική προσέγγιση	13
1.7.2 Προσέγγιση VADER	13
1.7.2 Προσέγγιση RoBERTa	13
2. Βιβλιογραφική ανασκόπηση	15
2.1 Ανάλυση συναισθήματος: Ορισμός και επισκόπηση	15
2.1.1 Ορισμός και πεδίο εφαρμογής	15
2.1.2 Μεθοδολογίες και τεχνικές:	15
Προσεγγίσεις με βάση το λεξικό	16
Προσεγγίσεις μηχανικής μάθησης	16
Προσεγγίσεις βαθιάς μάθησης	18
Συγκριτική ανάλυση	18
2.1.3 Εφαρμογές της ανάλυσης συναισθήματος	19
2.1.4 Προκλήσεις στην ανάλυση συναισθήματος	21
2.2 Τεχνικές επιλογής χαρακτηριστικών: Σημασία και λογική	22
2.2.2 Τύποι μεθόδων επιλογής χαρακτηριστικών	25
2.2.3 Συγκριτική ανάλυση των μεθόδων επιλογής χαρακτηριστικών	28
2.2.3 Συνάφεια με την ανάλυση συναισθήματος	28
2.3 Παραδείγματα εφαρμογών στην ανάλυση συναισθήματος	28
2.4 Βελτίωση της απόδοσης του μοντέλου	29
3. Προηγούμενες μελέτες για την ανάλυση συναισθήματος σε δεδομένα κοινωνικών μέσων ενημέρωσης	32
3.1 Επισκόπηση του ερευνητικού τοπίου	32
3.2 Πρώιμες εξελίξεις	32
3.4 Εξελίξεις στη βαθιά μάθηση	33
3.5 Εφαρμογές και επιπτώσεις	34
3.6 Προκλήσεις και μελλοντικές κατευθύνσεις	34
3.7 Μεθοδολογικές προσεγγίσεις	35
3.7.1 Προσεγγίσεις με βάση το λεξικό	35
3.7.2 Προσεγγίσεις μηχανικής μάθησης	36
3.7.3 Προσεγγίσεις βαθιάς μάθησης	37
4.7.4 Υβριδικές προσεγγίσεις	37

3.8 Βασικά ευρήματα και γνώσεις	38
3.8.1 Τάσεις του δημόσιου κλίματος	38
3.8.3 Μεθοδολογικές γνώσεις	39
3.8.4 Προκλήσεις και ευκαιρίες	40
4. Σημασία της μελέτης του λόγου για την κλιματική αλλαγή	41
4.1 Εισαγωγή	41
4.2 Ευαισθητοποίηση και κατανόηση	41
4.3 Ενημέρωση για πολιτικές και στρατηγικές επικοινωνίας	42
4.4 Αλλαγή συμπεριφοράς	42
4.5 Υπάρχουσα έρευνα	43
4.5.1 Ανάλυση συναισθήματος	44
4.5.2 Μοντελοποίηση θεμάτων	44
4.5.3 Ανάλυση δικτύου	45
4.6 Προκλήσεις και μελλοντικές κατευθύνσεις	46
4.7 Κενά και ευκαιρίες	46
4.7.1 Κενά στην τρέχουσα έρευνα	46
4.7.2 Ευκαιρίες για μελλοντική έρευνα	47
5. Μεθοδολογία	49
5.1 Επεξήγηση του dataset	49
5.2 Βήματα preprocessing στην ανάλυση συναισθήματος	50
5.2.1 Καθαρισμός	50
5.2.2 Tokenization	51
5.2.3 Αφαίρεση λέξεων διακοπής	52
5.2.4 Lemmatization	52
5.2.5 Stemming	53
5.2.6 Συμπέρασμα	53
6. Λεπτομερής Επεξήγηση Κώδικα	55
6.1 Εισαγωγή βιβλιοθηκών	55
6.2 Λήψη δεδομένων NLTK	55
6.2.1 Φόρτωση και επιθεώρηση του συνόλου δεδομένων	56
6.2.2 Απεικόνιση της κατανομής του συναισθήματος	56
6.2.3 Επεξεργασία NLTK	57
6.3 Εφαρμοσμένες μέθοδοι: VADER	59
6.3.1 Εισαγωγή βιβλιοθηκών και λήψη πόρων	60
6.3.2 Αρχικοποίηση του αναλυτή συναισθήματος και δοκιμή του	60
6.3.3 Ανάλυση του συναισθήματος ενός συνόλου δεδομένων	61
6.3.4 Δημιουργία ενός πλαισίου δεδομένων των αποτελεσμάτων	61
6.3.5 Οπτικοποίηση των αποτελεσμάτων της ανάλυσης συναισθήματος	62
6.3.6 Λεπτομερής οπτικοποίηση συνιστωσών συναισθήματος	62
6.3.7 Συμπέρασμα	63
6.3.8 results	65

6.4 Εφαρμοσμένες μέθοδοι: RoBERTa	68
6.4.1 Εισαγωγή βιβλιοθηκών και αρχικοποίηση του μοντέλου	69
6.4.2 Φόρτωση του Tokenizer και του μοντέλου	69
6.4.3 Κωδικοποίηση κειμένου και λήψη βαθμολογίας συναισθήματος	70
6.4.4 Εκτέλεση του μοντέλου για κάθε μήνυμα στο σύνολο δεδομένων	71
6.4.5 Δημιουργία ενός πλαισίου δεδομένων αποτελεσμάτων	72
6.4.6 Σύνοψη	72
6.4.7 Συμπέρασμα	73
6.4.8 roberta αποτελέσματα	74
6.5 Σύγκριση των αποτελεσμάτων του RoBERTa με τα αποτελέσματα του VADER:	76
7.5.1 Ερμηνεία του Pairplot	79
6.5.2 Ανάλυση σχέσεων και κατανομών	80
7. Συζήτηση των ευρημάτων	82
7.1 Επισκόπηση της ανάλυσης συναισθήματος	82
7.2 Βασικά ευρήματα	82
7.2.1 Ανάλυση του θετικού συναισθήματος μεταξύ της ομάδας των αδιάφορων	83
7.2.2 Αρνητικό συναίσθημα μεταξύ υποστηρικτών και σκεπτικιστών	84
7.3 Επιπτώσεις αυτών των ευρημάτων	85
7.4 Προτάσεις για μελλοντικές ερευνητικές κατευθύνσεις	86
8. Επίλογος	93
8.1 Σημασία της ανάλυσης συναισθήματος για την κατανόηση της κοινής γνώμης	93
8.2 Ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή	93
9. Βιβλιογραφία	95

1. Εισαγωγή

1.1 Πληροφορίες σχετικά με την ανάλυση συναισθήματος και την επιλογή χαρακτηριστικών

Η ανάλυση συναισθήματος, γνωστή και ως opinion mining, είναι ένας κρίσιμος τομέας της Επεξεργασίας Φυσικής Γλώσσας (NLP) που επικεντρώνεται στον εντοπισμό και την κατηγοριοποίηση των απόψεων που εκφράζονται σε δεδομένα κειμένου. Ο πρωταρχικός στόχος της ανάλυσης συναισθήματος είναι ο προσδιορισμός του συναισθηματικού τόνου πίσω από μια σειρά λέξεων, ο οποίος μπορεί να είναι χρήσιμος για την κατανόηση των στάσεων, των συναισθημάτων και των απόψεων που εκφράζουν οι συντάκτες. Ο τομέας αυτός έχει αναπτυχθεί σημαντικά την τελευταία δεκαετία λόγω της διάδοσης της ψηφιακής επικοινωνίας και της ανάγκης των επιχειρήσεων, των ερευνητών και των φορέων χάραξης πολιτικής να κατανοούν γρήγορα και με ακρίβεια το κοινό συναίσθημα [10,17].

Οι μεθοδολογίες που χρησιμοποιούνται στην ανάλυση συναισθήματος μπορούν σε γενικές γραμμές να κατηγοριοποιηθούν σε προσεγγίσεις που βασίζονται σε κανόνες, σε προσεγγίσεις μηχανικής μάθησης και σε υβριδικές προσεγγίσεις. Τα συστήματα που βασίζονται σε κανόνες χρησιμοποιούν ένα σύνολο γλωσσικών κανόνων που δημιουργούνται χειροκίνητα για τον εντοπισμό και την κατηγοριοποίηση των συναισθημάτων. Τα συστήματα αυτά βασίζονται σε προκαθορισμένα λεξικά συναισθήματος, τα οποία είναι κατάλογοι λέξεων και εκφράσεων σχολιασμένων με τις αντίστοιχες τιμές συναισθήματος. Ενώ τα συστήματα που βασίζονται σε κανόνες είναι σχετικά απλά στην υλοποίηση και την ερμηνεία, συχνά δυσκολεύονται να χειριστούν την πολυπλοκότητα και τη μεταβλητότητα της φυσικής γλώσσας, όπως ο σαρκασμός, οι ιδιωτισμοί και οι εξαρτώμενες από το πλαίσιο σημασίες [3].

Οι προσεγγίσεις μηχανικής μάθησης, από την άλλη πλευρά, περιλαμβάνουν την εκπαίδευση αλγορίθμων σε επισημασμένα σύνολα δεδομένων για την αυτόματη εκμάθηση προτύπων που σχετίζονται με διαφορετικά συναισθήματα. Οι συνήθεις αλγόριθμοι που χρησιμοποιούνται στην ανάλυση συναισθήματος περιλαμβάνουν τις μηχανές διανυσμάτων υποστήριξης (SVM), το Naïve Bayes και τη λογιστική παλινδρόμηση. Αυτά τα μοντέλα μπορούν να γενικεύουν καλά από τα δεδομένα εκπαίδευσης σε νέα, αθέατα δεδομένα, καθιστώντας τα πιο ανθεκτικά από τα συστήματα που βασίζονται σε κανόνες. Ωστόσο, η απόδοσή τους εξαρτάται σε μεγάλο βαθμό από την ποιότητα και το μέγεθος των δεδομένων εκπαίδευσης. Πιο πρόσφατα, έχουν εισαχθεί μοντέλα βαθιάς μάθησης, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, RNN) και μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT (Bidirectional Encoder Representations from Transformers). Αυτά τα μοντέλα μπορούν να συλλάβουν πολύπλοκα γλωσσικά πρότυπα και πληροφορίες σχετικά με το πλαίσιο, βελτιώνοντας σημαντικά την ακρίβεια των εργασιών ανάλυσης συναισθήματος[9,24].

Οι υβριδικές προσεγγίσεις συνδυάζουν στοιχεία τόσο των μεθόδων που βασίζονται σε κανόνες όσο και των μεθόδων μηχανικής μάθησης για να αξιοποιήσουν τα πλεονεκτήματα της καθεμιάς. Για παράδειγμα, ένα υβριδικό μοντέλο μπορεί να χρησιμοποιεί τη μηχανική μάθηση για την ταξινόμηση των συναισθημάτων και ένα σύστημα βασισμένο σε κανόνες για τη βελτίωση των αποτελεσμάτων με την εφαρμογή κανόνων που αφορούν συγκεκριμένο τομέα.

Η επιλογή χαρακτηριστικών είναι ένα ουσιαστικό βήμα preprocessing σε πολλές εργασίες μηχανικής μάθησης, συμπεριλαμβανομένης της ανάλυσης συναισθήματος. Περιλαμβάνει την επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών από το σύνολο των διαθέσιμων χαρακτηριστικών στο σύνολο δεδομένων για τη δημιουργία αποδοτικών και αποτελεσματικών μοντέλων πρόβλεψης. Οι πρωταρχικοί στόχοι της επιλογής χαρακτηριστικών είναι η βελτίωση της απόδοσης του μοντέλου, τη μείωση του hyper-adaptation, η ενίσχυση της ερμηνευσιμότητας και η μείωση του υπολογιστικού κόστους [6,18].

Υπάρχουν τρεις κύριοι τύποι μεθόδων επιλογής χαρακτηριστικών: μέθοδοι φίλτρου, μέθοδοι wrapper και ενσωματωμένες μέθοδοι. Οι μέθοδοι φίλτρου αξιολογούν τη συνάφεια των χαρακτηριστικών με βάση τις εγγενείς ιδιότητές τους και επιλέγουν τα χαρακτηριστικά ανεξάρτητα από τον αλγόριθμο μάθησης. Οι συνήθεις μέθοδοι φίλτρου περιλαμβάνουν συντελεστές συσχέτισης, αμοιβαία πληροφορία και στατιστικές δοκιμές. Αυτές οι μέθοδοι είναι υπολογιστικά αποδοτικές και μπορούν να χειριστούν μεγάλα σύνολα δεδομένων, αλλά ενδέχεται να παραβλέπουν τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.

Οι μέθοδοι wrapper, από την άλλη πλευρά, αξιολογούν τη χρησιμότητα ενός υποσυνόλου χαρακτηριστικών με την πραγματική εκπαίδευση και αξιολόγηση ενός μοντέλου σε διαφορετικά υποσύνολα χαρακτηριστικών. Τεχνικές όπως η αναδρομική απαλοιφή χαρακτηριστικών (RFE) και η προς τα εμπρός/πίσω επιλογή χαρακτηριστικών είναι παραδείγματα μεθόδων wrapper. Ενώ αυτές οι μέθοδοι μπορούν να παρέχουν πιο ακριβή υποσύνολα χαρακτηριστικών, είναι υπολογιστικά δαπανηρές, ειδικά για μεγάλα σύνολα δεδομένων.

Οι ενσωματωμένες μέθοδοι εκτελούν επιλογή χαρακτηριστικών κατά τη διαδικασία εκπαίδευσης του μοντέλου. Αλγόριθμοι όπως ο LASSO (Least Absolute Shrinkage and Selection Operator) και τα δέντρα αποφάσεων εκτελούν εγγενώς επιλογή χαρακτηριστικών με την επιβολή ποινών σε λιγότερο σημαντικά χαρακτηριστικά ή με τη διάσπαση των κόμβων με βάση τη σημασία των χαρακτηριστικών. Αυτές οι μέθοδοι εξισορροπούν τα αντισταθμιστικά οφέλη μεταξύ των μεθόδων φίλτρου και wrapper ενσωματώνοντας την επιλογή χαρακτηριστικών στη διαδικασία εκπαίδευσης του μοντέλου, με αποτέλεσμα να είναι πιο αποτελεσματικές, ενώ εξακολουθούν να καταγράφουν τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών [6,8,19].

Η επιλογή χαρακτηριστικών είναι ιδιαίτερα κρίσιμη στην ανάλυση συναισθήματος λόγω της υψηλής διαστατικότητας των δεδομένων κειμένου. Τα δεδομένα κειμένου συνήθως μετατρέπονται σε αριθμητικά διανύσματα με τη χρήση τεχνικών όπως το bag-of-words, το TF-IDF (Term Frequency-Inverse Document Frequency), ή τις ενσωματώσεις λέξεων. Αυτός ο μετασχηματισμός οδηγεί σε μεγάλο αριθμό χαρακτηριστικών, πολλά από τα οποία μπορεί να είναι άσχετα ή περιττά. Εφαρμόζοντας την επιλογή χαρακτηριστικών, μπορούμε να εντοπίσουμε

τα πιο κατατοπιστικά χαρακτηριστικά που συμβάλλουν στην ακριβή ταξινόμηση των συναισθημάτων, βελτιώνοντας έτσι την απόδοση του μοντέλου και μειώνοντας την υπολογιστική πολυπλοκότητα.

Συνοψίζοντας, η ανάλυση συναισθήματος είναι ένα ζωτικής σημασίας εργαλείο για την κατανόηση της κοινής γνώμης, αξιοποιώντας διάφορες μεθοδολογίες που κυμαίνονται από συστήματα βασισμένα σε κανόνες έως προηγμένα μοντέλα μηχανικής μάθησης. Η επιλογή χαρακτηριστικών παίζει καθοριστικό ρόλο στην ενίσχυση αυτών των μοντέλων, εστιάζοντας στα πιο σημαντικά χαρακτηριστικά, βελτιώνοντας έτσι την ακρίβεια, μειώνοντας το hyper-adaptation και εξασφαλίζοντας αποδοτικό υπολογισμό. Καθώς η ψηφιακή επικοινωνία συνεχίζει να αναπτύσσεται, η σημασία της αποτελεσματικής ανάλυσης συναισθήματος και της επιλογής χαρακτηριστικών γίνεται όλο και πιο σημαντική σε διάφορους τομείς [1, 4].

1.2 Επισκόπηση της σημασίας της ανάλυσης συναισθήματος για την κατανόηση της κοινής γνώμης

Στη σύγχρονη ψηφιακή εποχή, η έκρηξη του περιεχομένου που δημιουργείται από τους χρήστες σε πλατφόρμες όπως τα μέσα κοινωνικής δικτύωσης, τα ιστολόγια, τα φόρουμ και οι ιστότοποι κριτικών έχει δημιουργήσει ένα τεράστιο αποθετήριο κειμενικών δεδομένων που αποτυπώνει τις απόψεις, τα συναισθήματα και τα συναισθήματα των ατόμων παγκοσμίως. Η ανάλυση συναισθήματος, ένα υποσύνολο της Επεξεργασίας Φυσικής Γλώσσας (NLP), έχει αναδειχθεί σε βασικό εργαλείο για την εξαγωγή και ανάλυση αυτών των δεδομένων για την κατανόηση της κοινής γνώμης. Αυτή η ικανότητα είναι ζωτικής σημασίας για μια ποικιλία ενδιαφερομένων, συμπεριλαμβανομένων επιχειρήσεων, φορέων χάραξης πολιτικής, ερευνητών και οργανισμών μέσων ενημέρωσης, καθώς παρέχει σε πραγματικό χρόνο πληροφορίες για τη συλλογική διάθεση και τις απόψεις μεγάλων και διαφορετικών πληθυσμών [2].

Για επιχειρήσεις:

Οι επιχειρήσεις μπορούν να αξιοποιήσουν την ανάλυση συναισθήματος για να αποκτήσουν βαθύτερη κατανόηση των συναισθημάτων των πελατών σχετικά με τα προϊόντα, τις υπηρεσίες και τη συνολική αντίληψη της μάρκας τους. Αναλύοντας τα σχόλια των πελατών από κριτικές, αναφορές στα μέσα κοινωνικής δικτύωσης και έρευνες, οι εταιρείες μπορούν να εντοπίσουν τομείς ικανοποίησης και δυσαρέσκειας. Αυτό τους δίνει τη δυνατότητα να αντιμετωπίσουν άμεσα τα ζητήματα, να ενισχύσουν την εμπειρία των πελατών και να βελτιώσουν τις προσφορές προϊόντων. Επιπλέον, η ανάλυση συναισθήματος βοηθά στην παρακολούθηση της φήμης της μάρκας, επιτρέποντας στις εταιρείες να εντοπίζουν και να ανταποκρίνονται σε πιθανές κρίσεις έγκαιρα, μετριάζοντας έτσι τις αρνητικές επιπτώσεις. Οι στρατηγικές μάρκετινγκ μπορούν

επίσης να προσαρμοστούν με βάση τις γνώσεις της ανάλυσης συναισθήματος ώστε να ευθυγραμμιστούν στενότερα με τις προτιμήσεις και τις προσδοκίες των πελατών.

Για τους κυβερνητικούς φορείς:

Η ανάλυση συναισθήματος χρησιμεύει ως ένα ισχυρό εργαλείο για τους υπεύθυνους χάραξης πολιτικής και τους κυβερνητικούς οργανισμούς για τη μέτρηση της κοινής γνώμης σχετικά με διάφορα θέματα, πολιτικές και γεγονότα. Αναλύοντας τα συναισθήματα που εκφράζονται στα μέσα κοινωνικής δικτύωσης και σε άλλες ψηφιακές πλατφόρμες, οι υπεύθυνοι χάραξης πολιτικής μπορούν να λαμβάνουν ανατροφοδότηση σε πραγματικό χρόνο σχετικά με τις αντιδράσεις του κοινού σε πολιτικές αποφάσεις, νομοθετικές αλλαγές και κυβερνητικές δράσεις. Αυτή η διορατικότητα σε πραγματικό χρόνο είναι ανεκτίμητη για τη δημιουργία ευέλικτων και προσαρμοστικών πολιτικών που ευθυγραμμίζονται με το κοινό συναίσθημα και τις ανάγκες. Για παράδειγμα, κατά τη διάρκεια κρίσεων δημόσιας υγείας ή φυσικών καταστροφών, η ανάλυση συναισθήματος μπορεί να βοηθήσει τις αρχές να κατανοήσουν τις ανησυχίες του κοινού, τις τάσεις παραπληροφόρησης και την αποτελεσματικότητα των στρατηγικών επικοινωνίας, επιτρέποντάς τους να προσαρμόσουν ανάλογα τις αντιδράσεις τους.

Για τους ερευνητές:

Οι ακαδημαϊκοί ερευνητές και οι ερευνητές της αγοράς χρησιμοποιούν την ανάλυση συναισθήματος για τη μελέτη κοινωνικών φαινομένων, προτύπων συμπεριφοράς και πολιτιστικών τάσεων. Επιτρέπει στους ερευνητές να αναλύουν αποτελεσματικά τεράστιες ποσότητες δεδομένων κειμένου, αποκαλύπτοντας πληροφορίες για το πώς εξελίσσεται η κοινή γνώμη με την πάροδο του χρόνου και ως απάντηση σε συγκεκριμένα γεγονότα ή ερεθίσματα. Για παράδειγμα, η ανάλυση συναισθήματος έχει χρησιμοποιηθεί στην πολιτική επιστήμη για την ανάλυση του κοινού συναισθήματος κατά τη διάρκεια των εκλογών, στην κοινωνιολογία για τη μελέτη των κοινωνικών αντιδράσεων σε σημαντικά γεγονότα και στην περιβαλλοντική επιστήμη για την κατανόηση των αντιλήψεων του κοινού σχετικά με την κλιματική αλλαγή. Οι γνώσεις αυτές συμβάλλουν στη βαθύτερη κατανόηση της κοινωνικής δυναμικής και παρέχουν πληροφορίες για μελλοντικές ερευνητικές κατευθύνσεις.

Για οργανισμούς μέσων ενημέρωσης:

Οι οργανισμοί μέσων ενημέρωσης επωφελούνται από την ανάλυση συναισθήματος, καθώς αποκτούν πληροφορίες σχετικά με τις αντιδράσεις του κοινού σε άρθρα ειδήσεων, εκπομπές και περιεχόμενο μέσων ενημέρωσης. Αυτό τους βοηθά να προσαρμόσουν το περιεχόμενό τους στις προτιμήσεις του κοινού και να βελτιώσουν τη δέσμευση. Επιπλέον, η ανάλυση συναισθήματος μπορεί να βοηθήσει στον εντοπισμό θεμάτων που παρουσιάζουν τάση και στη μέτρηση του ενδιαφέροντος του κοινού για διάφορα θέματα, καθοδηγώντας τις εκδοτικές αποφάσεις και τη στρατηγική περιεχομένου. Τα μέσα ενημέρωσης μπορούν επίσης να παρακολουθούν το κοινό συναίσθημα απέναντι στο εμπορικό σήμα τους και τους δημοσιογράφους, επιτρέποντάς τους να διατηρούν την αξιοπιστία και την εμπιστοσύνη.

Λήψη αποφάσεων σε πραγματικό χρόνο:

Ένα από τα σημαντικότερα πλεονεκτήματα της ανάλυσης συναισθήματος είναι η ικανότητά της να παρέχει πληροφορίες σε πραγματικό χρόνο. Οι παραδοσιακές μέθοδοι μέτρησης της κοινής γνώμης, όπως οι έρευνες και οι ομάδες εστίασης, είναι χρονοβόρες και συχνά στερούνται της αμεσότητας που απαιτείται στο σημερινό ταχέως εξελισσόμενο περιβάλλον. Τα εργαλεία ανάλυσης συναισθήματος μπορούν να επεξεργάζονται και να αναλύουν τεράστιες ποσότητες δεδομένων σχεδόν αμέσως, παρέχοντας έγκαιρες πληροφορίες που είναι ζωτικής σημασίας για τη γρήγορη λήψη τεκμηριωμένων αποφάσεων. Αυτή η ικανότητα πραγματικού χρόνου είναι ιδιαίτερα σημαντική σε περιβάλλοντα όπου το κοινό συναίσθημα μπορεί να αλλάξει γρήγορα, όπως κατά τη διάρκεια πολιτικών εκστρατειών, λανσαρίσματος αγορών ή κρίσεων.

Βελτίωση των στρατηγικών επικοινωνίας:

Η αποτελεσματική επικοινωνία είναι απαραίτητη για κάθε οργανισμό, και η κατανόηση του συναισθήματος του κοινού είναι το κλειδί για τη διαμόρφωση μηνυμάτων που βρίσκουν απήχηση στο κοινό. Η ανάλυση συναισθήματος βοηθά τους οργανισμούς να προσαρμόσουν τις στρατηγικές επικοινωνίας τους, παρέχοντας πληροφορίες σχετικά με τη γλώσσα και τον τόνο που εμπλέκουν καλύτερα το κοινό στόχο τους. Για παράδειγμα, μια εταιρεία που λανσάρει ένα νέο προϊόν μπορεί να χρησιμοποιήσει την ανάλυση συναισθήματος για να εντοπίσει τα θετικά σχόλια και να τονίσει αυτές τις πτυχές στις εκστρατείες μάρκετινγκ. Ομοίως, κατά τη διάρκεια μιας εκστρατείας δημόσιας υγείας, η κατανόηση του συναισθήματος του κοινού μπορεί να βοηθήσει τις υγειονομικές αρχές να αντιμετωπίσουν τις ανησυχίες και να αντιμετωπίσουν την παραπληροφόρηση πιο αποτελεσματικά.

Συμπερασματικά, η ανάλυση συναισθήματος είναι ένα κρίσιμο εργαλείο για την κατανόηση της κοινής γνώμης στην ψηφιακή εποχή. Οι εφαρμογές της εκτείνονται σε διάφορους τομείς, παρέχοντας πολύτιμες πληροφορίες που ενημερώνουν για τη λήψη αποφάσεων, ενισχύουν την ικανοποίηση των πελατών, διαμορφώνουν πολιτικές, καθοδηγούν την έρευνα και βελτιώνουν τις στρατηγικές επικοινωνίας. Καθώς η ψηφιακή επικοινωνία συνεχίζει να επεκτείνεται, η σημασία της ανάλυσης συναισθήματος στην καταγραφή και ερμηνεία του δημόσιου συναισθήματος θα αυξάνεται, καθιστώντας την απαραίτητο εργαλείο για τους ενδιαφερόμενους σε διάφορους τομείς[3,5,6,13,14].

1.3 Εισαγωγή στη συγκεκριμένη ερευνητική εστίαση: Ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή

Η κλιματική αλλαγή είναι ένα παγκοσμίως αναγνωρισμένο ζήτημα με βαθιές επιπτώσεις στο περιβάλλον, την οικονομία και την κοινωνία. Καθώς η επείγουσα ανάγκη για την αντιμετώπιση της κλιματικής αλλαγής αυξάνεται, η κατανόηση της κοινής γνώμης για το θέμα αυτό καθίσταται όλο και πιο κρίσιμη. Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης, ιδίως το Twitter, έχουν γίνει ζωτικής σημασίας χώροι για τον δημόσιο διάλογο σχετικά με την κλιματική αλλαγή. Αυτές οι πλατφόρμες προσφέρουν μια ματιά σε πραγματικό χρόνο στις σκέψεις, τα συναισθήματα και τις απόψεις των ατόμων παγκοσμίως. Η παρούσα μελέτη επικεντρώνεται στη διενέργεια ανάλυσης συναισθήματος των tweets που σχετίζονται με την κλιματική αλλαγή, ώστε να αποκτήσει γνώσεις σχετικά με το δημόσιο συναίσθημα και την εξέλιξη του με την πάροδο του χρόνου[2,21].

1.4 Η σημασία της συζήτησης για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης

Το Twitter, με την τεράστια βάση χρηστών και τις δυνατότητες ταχείας διάδοσης πληροφοριών, χρησιμεύει ως εξαιρετική πλατφόρμα για την ανάλυση της κοινής γνώμης. Σε αντίθεση με τις παραδοσιακές έρευνες, οι οποίες μπορεί να είναι χρονοβόρες και περιορισμένης εμβέλειας, το Twitter παρέχει μια συνεχή ροή δεδομένων από διάφορες δημογραφικές και γεωγραφικές ομάδες. Οι χρήστες μοιράζονται συχνά τις απόψεις τους σχετικά με ειδήσεις, πολιτικές, γεγονότα και προσωπικές εμπειρίες που σχετίζονται με το κλίμα, δημιουργώντας ένα πλούσιο σύνολο δεδομένων για ανάλυση συναισθήματος. Αναλύοντας αυτά τα tweets, μπορούμε να κατανοήσουμε τη συναισθηματική ανταπόκριση του κοινού σε διάφορες πτυχές της κλιματικής αλλαγής, να εντοπίσουμε τα επικρατούντα συναισθήματα και να ανιχνεύσουμε αλλαγές στην κοινή γνώμη [12].

1.5 Στόχοι της μελέτης

Οι πρωταρχικοί στόχοι της παρούσας μελέτης είναι:

Ανάλυση του συναισθήματος των tweets που σχετίζονται με την κλιματική αλλαγή: Αυτό περιλαμβάνει την κατηγοριοποίηση των tweets σε θετικά, αρνητικά ή ουδέτερα συναισθήματα για την κατανόηση της συνολικής διάθεσης της κοινής γνώμης σχετικά με την κλιματική αλλαγή.

Σύγκριση των επιδόσεων διαφορετικών μοντέλων ανάλυσης συναισθήματος: Συγκεκριμένα, η παρούσα μελέτη θα συγκρίνει την απόδοση του VADER, ενός μοντέλου που βασίζεται σε λεξικό και κανόνες, με το RoBERTa, ένα μοντέλο που βασίζεται σε μετασχηματιστές, στην ταξινόμηση των συναισθημάτων των tweets για την κλιματική αλλαγή.

Για την αξιολόγηση της επίδρασης της επιλογής χαρακτηριστικών στην απόδοση του μοντέλου: Αυτό περιλαμβάνει την αξιολόγηση του τρόπου με τον οποίο διαφορετικές τεχνικές επιλογής χαρακτηριστικών επηρεάζουν την ακρίβεια και την αποδοτικότητα των μοντέλων ανάλυσης συναισθήματος.

Για τον εντοπισμό βασικών θεμάτων και τάσεων στο διάλογο για την κλιματική αλλαγή στο Twitter: Με την ανάλυση του περιεχομένου και του συναισθήματος των tweets, η μελέτη αυτή αποσκοπεί στην ανάδειξη των επικρατούντων θεμάτων και των αναδυόμενων τάσεων στις δημόσιες συζητήσεις για την κλιματική αλλαγή [2,12].

1.6 Σημασία της έρευνας

Η κατανόηση του δημόσιου συναισθήματος απέναντι στην κλιματική αλλαγή είναι απαραίτητη για διάφορους λόγους. Πρώτον, μπορεί να ενημερώσει τους υπεύθυνους χάραξης πολιτικής σχετικά με τις ανησυχίες και τα επίπεδα υποστήριξης του κοινού για διάφορες πολιτικές για το κλίμα. Οι υπεύθυνοι χάραξης πολιτικής μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για να σχεδιάσουν πιο αποτελεσματικές και δημόσια αποδεκτές στρατηγικές για το κλίμα. Δεύτερον, οι περιβαλλοντικές οργανώσεις και οι ακτιβιστές μπορούν να προσαρμόσουν τις στρατηγικές επικοινωνίας τους με βάση τις συναισθηματικές αντιδράσεις του κοινού, ενισχύοντας έτσι τις προσπάθειες υπεράσπισής τους. Τρίτον, οι επιχειρήσεις, ιδίως εκείνες που δραστηριοποιούνται στους τομείς της πράσινης τεχνολογίας και των ανανεώσιμων πηγών ενέργειας, μπορούν να μετρήσουν το κλίμα της αγοράς και να εντοπίσουν πιθανές ευκαιρίες ή προκλήσεις στην προώθηση των προϊόντων και των υπηρεσιών τους [5,23].

1.7 Μεθοδολογική προσέγγιση

1.7.2 Προσέγγιση VADER

Το VADER (Valence Aware Dictionary for sEntiment Reasoning) είναι ένα εργαλείο ανάλυσης συναισθήματος βασισμένο σε λεξικό και κανόνες, ειδικά προσαρμοσμένο στα συναισθήματα που εκφράζονται στα μέσα κοινωνικής δικτύωσης. Έχει σχεδιαστεί για να αποδίδει καλά σε σύντομα κείμενα από πλατφόρμες όπως το Twitter. Το VADER χρησιμοποιεί έναν συνδυασμό λεξικού συναισθήματος και γραμματικών κανόνων για να αποδίδει βαθμολογίες συναισθήματος σε κείμενα. Είναι γνωστό για την απλότητα και την αποτελεσματικότητά του, γεγονός που το καθιστά δημοφιλή επιλογή για εργασίες ανάλυσης συναισθήματος σε πραγματικό χρόνο. Ένα από τα βασικά πλεονεκτήματα του VADER είναι η απλότητα και η αποτελεσματικότητά του. Είναι σε θέση να χειριστεί αποχρώσεις στο κείμενο, όπως η χρήση κεφαλαίων για έμφαση (π.χ. "ΧΑΡΟΥΜΕΝΟΣ"), οι τροποποιητές βαθμού (π.χ. "πολύ") και η παρουσία αρνήσεων (π.χ. "όχι καλά"). Αυτά τα χαρακτηριστικά το καθιστούν ιδιαίτερα κατάλληλο για εφαρμογές πραγματικού χρόνου όπου απαιτείται γρήγορη και ακριβής ανάλυση συναισθήματος. Οι επιδόσεις του VADER έχουν επικυρωθεί έναντι ανθρώπινων αξιολογήσεων και άλλων σύγχρονων εργαλείων ανάλυσης συναισθήματος, αποδεικνύοντας την ευρωστία και την αξιοπιστία του σε διάφορα πλαίσια, συμπεριλαμβανομένων των πλατφορμών microblogging και των μέσων κοινωνικής δικτύωσης. [7]

1.7.2 Προσέγγιση RoBERTa

Το RoBERTa είναι ένα μοντέλο βασισμένο σε μετασχηματιστές που βασίζεται στο BERT (Bidirectional Encoder Representations from Transformers). Έχει σχεδιαστεί για να χειρίζεται τις πολυπλοκότητες της κατανόησης της φυσικής γλώσσας αξιοποιώντας μια αρχιτεκτονική βαθιάς μάθησης που καταγράφει τις σχέσεις πλαισίου μεταξύ των λέξεων σε μια πρόταση. Το RoBERTa βελτιώνει το BERT χρησιμοποιώντας μεγαλύτερο σώμα εκπαίδευσης και βελτιστοποιώντας τη διαδικασία εκπαίδευσης, η οποία περιλαμβάνει τη δυναμική αλλαγή του μοτίβου κάλυψης που εφαρμόζεται στα δεδομένα εκπαίδευσης και την εκπαίδευση για μεγαλύτερο χρονικό διάστημα με μεγαλύτερες παρτίδες και ρυθμούς μάθησης.

Για την ανάλυση συναισθήματος, το RoBERTa μπορεί να ρυθμιστεί λεπτομερώς σε επισημασμένα δεδομένα συναισθήματος για να βελτιώσει την απόδοσή του στην ταξινόμηση συναισθημάτων. Η αρχιτεκτονική του μοντέλου του επιτρέπει να κατανοεί και να επεξεργάζεται εξαρτήσεις μεγάλης εμβέλειας στο κείμενο, καθιστώντας το ιδιαίτερα αποτελεσματικό στην καταγραφή αποχρώσεων και συναισθημάτων που εξαρτώνται από το πλαίσιο. Το RoBERTa έχει επιδείξει ανώτερες επιδόσεις σε διάφορες εργασίες NLP, συμπεριλαμβανομένης της ανάλυσης συναισθήματος, λόγω της ικανότητάς του να συλλαμβάνει βαθιές πληροφορίες σχετικά με το περιεχόμενο και της ισχυρής προεκπαίδευσής του σε εκτεταμένα σύνολα δεδομένων.

Στην παρούσα μελέτη, το μοντέλο RoBERTa χρησιμοποιείται για την ανάλυση tweets που σχετίζονται με την κλιματική αλλαγή. Με την εκλεπτυσμένη ρύθμιση του RoBERTa σε ένα σύνολο δεδομένων tweets με επισημάνσεις συναισθημάτων, στοχεύουμε στην αξιοποίηση των προηγμένων δυνατοτήτων του για την επίτευξη υψηλής ακρίβειας στην ταξινόμηση συναισθημάτων. Η χρήση του RoBERTa μας επιτρέπει να χειριστούμε τις λεπτές αποχρώσεις και τις πολυπλοκότητες που ενυπάρχουν στο κείμενο των μέσων κοινωνικής δικτύωσης, παρέχοντας μια βαθύτερη κατανόηση του δημόσιου συναισθήματος απέναντι στην κλιματική αλλαγή[9,24].

Αναμενόμενες συνεισφορές

Η παρούσα έρευνα στοχεύει να συμβάλει στον τομέα της ανάλυσης συναισθήματος παρέχοντας μια λεπτομερή σύγκριση των μοντέλων VADER και RoBERTa στην ανάλυση tweets που σχετίζονται με την κλιματική αλλαγή. Αξιοποιώντας τεχνικές επιλογής χαρακτηριστικών, η παρούσα μελέτη επιδιώκει να βελτιώσει την αποτελεσματικότητα και την ακρίβεια των μοντέλων ανάλυσης συναισθήματος. Τα ευρήματα της παρούσας μελέτης μπορούν να προσφέρουν πολύτιμες πληροφορίες για την κοινή γνώμη σχετικά με την κλιματική αλλαγή, βοηθώντας τους υπεύθυνους χάραξης πολιτικής, τις περιβαλλοντικές οργανώσεις και τις επιχειρήσεις να λαμβάνουν τεκμηριωμένες αποφάσεις και να αναπτύσσουν αποτελεσματικές στρατηγικές για την αντιμετώπιση της κλιματικής αλλαγής.

Συνοψίζοντας, η ανάλυση των tweets για την κλιματική αλλαγή με τη χρήση προηγμένων μοντέλων ανάλυσης συναισθήματος και τεχνικών επιλογής χαρακτηριστικών αποτελεί ένα σημαντικό βήμα προς την κατανόηση και αξιοποίηση του δημόσιου συναισθήματος για την προώθηση ουσιαστικής δράσης κατά της κλιματικής αλλαγής. Η έρευνα αυτή όχι μόνο προάγει τις μεθοδολογικές προσεγγίσεις στην ανάλυση συναισθήματος, αλλά παρέχει επίσης πρακτικές συνέπειες για τους ενδιαφερόμενους που συμμετέχουν στις προσπάθειες μετριασμού της κλιματικής αλλαγής και προσαρμογής.

2. Βιβλιογραφική ανασκόπηση

2.1 Ανάλυση συναισθήματος: Ορισμός και επισκόπηση

Η ανάλυση συναισθήματος, γνωστή και ως opinion mining, είναι ένα πεδίο της Επεξεργασίας Φυσικής Γλώσσας (NLP) που περιλαμβάνει την υπολογιστική μελέτη των απόψεων, των συναισθημάτων και των συναισθημάτων που εκφράζονται σε κείμενο. Ο πρωταρχικός στόχος της ανάλυσης συναισθήματος είναι ο προσδιορισμός της στάσης ενός συγγραφέα ή ομιλητή σχετικά με ένα συγκεκριμένο θέμα, προϊόν ή γεγονός. Αυτή μπορεί να ταξινομηθεί ως θετική, αρνητική ή ουδέτερη, ενώ πιο διαφοροποιημένες προσεγγίσεις μπορούν επίσης να ανιχνεύσουν συγκεκριμένα συναισθήματα όπως χαρά, θυμό ή θλίψη [10,16,17].

2.1.1 Ορισμός και πεδίο εφαρμογής

Η ανάλυση συναισθήματος είναι η διαδικασία ανάλυσης δεδομένων κειμένου για την εξαγωγή υποκειμενικών πληροφοριών και τον προσδιορισμό του συναισθήματος που εκφράζεται. Ο τομέας αυτός συνδυάζει τεχνικές από διάφορους τομείς, όπως η υπολογιστική γλωσσολογία, η ανάλυση κειμένου και η εξόρυξη δεδομένων. Το πεδίο εφαρμογής της ανάλυσης συναισθήματος περιλαμβάνει διάφορα επίπεδα granulation: επίπεδο εγγράφου, επίπεδο πρότασης και επίπεδο πτυχής.

- Ανάλυση συναισθήματος σε επίπεδο εγγράφου: Πρόκειται για τον προσδιορισμό του συνολικού συναισθήματος που εκφράζεται σε ένα ολόκληρο έγγραφο. Για παράδειγμα, ανάλυση μιας πλήρους κριτικής προϊόντος για την ταξινόμησή της ως θετική, αρνητική ή ουδέτερη.
- Ανάλυση συναισθήματος σε επίπεδο πρότασης: Επικεντρώνεται σε μεμονωμένες προτάσεις, ταξινομώντας κάθε μία ως θετική, αρνητική ή ουδέτερη. Αυτή η προσέγγιση είναι χρήσιμη για κείμενα όπου το συναίσθημα μπορεί να διαφέρει σε διάφορα μέρη του εγγράφου.
- Ανάλυση συναισθήματος σε επίπεδο πτυχής: Αυτή η λεπτομερής προσέγγιση εξετάζει συγκεκριμένες πτυχές ή χαρακτηριστικά που αναφέρονται στο κείμενο και προσδιορίζει το συναίσθημα που σχετίζεται με κάθε πτυχή. Για παράδειγμα, σε μια κριτική εστιατορίου, το συναίσθημα σχετικά με το φαγητό, την εξυπηρέτηση και την ατμόσφαιρα μπορεί να αξιολογηθεί ξεχωριστά [3,10,11,16].

2.1.2 Μεθοδολογίες και τεχνικές:

Η ανάλυση συναισθήματος χρησιμοποιεί διάφορες μεθοδολογίες και προσεγγίσεις για την ταξινόμηση του συναισθήματος που εκφράζεται σε δεδομένα κειμένου. Αυτές οι μεθοδολογίες μπορούν να κατηγοριοποιηθούν σε γενικές γραμμές σε προσεγγίσεις που βασίζονται σε λεξικό, προσεγγίσεις μηχανικής μάθησης και προσεγγίσεις βαθιάς μάθησης. Κάθε προσέγγιση έχει το

δικό της σύνολο τεχνικών, πλεονεκτημάτων και προκλήσεων. Σε αυτή την ενότητα, θα διερευνήσουμε αυτές τις μεθοδολογίες λεπτομερώς, παραπέμποντας σε σχετική βιβλιογραφία και μελέτες [3,10,11,17].

Προσεγγίσεις με βάση το λεξικό

Οι προσεγγίσεις που βασίζονται σε λεξικά βασίζονται σε προκαθορισμένους καταλόγους λέξεων, γνωστούς ως λεξικά συναισθήματος, οι οποίοι σχολιάζονται με τις αντίστοιχες τιμές συναισθήματος. Αυτές οι μέθοδοι ταξινομούν το συναισθήμα υπολογίζοντας την πολικότητα του συναισθήματος ενός δεδομένου κειμένου με βάση την παρουσία και την πολικότητα των λέξεων στο λεξικό.

1. Λεξικά συναισθήματος:

- General Inquirer: Ένα από τα πρώτα λεξικά συναισθήματος, το General Inquirer, κατηγοριοποιεί τις λέξεις σε θετικές, αρνητικές και ουδέτερες κατηγορίες. Έχει χρησιμοποιηθεί ευρέως για την ανάλυση συναισθήματος σε διάφορους τομείς.

- SentiWordNet: Το SentiWordNet επεκτείνει το WordNet, μια λεξιλογική βάση δεδομένων για την αγγλική γλώσσα, συσχετίζοντας κάθε synset (ομάδα συνωνύμων) με βαθμολογίες για τη θετικότητα, την αρνητικότητα και την ουδετερότητα.

- VADER (Valence Aware Dictionary and sEntiment Reasoner): Το VADER είναι ειδικά προσαρμοσμένο στα συναισθήματα που εκφράζονται στα μέσα κοινωνικής δικτύωσης. Συνδυάζει ένα λεξικό συναισθημάτων με γραμματικούς κανόνες για να λαμβάνει υπόψη τα συμφραζόμενα, την κεφαλαιοποίηση, τη στίξη και τους τροποποιητές βαθμού. Το VADER είναι αποτελεσματικό για σύντομα κείμενα όπως τα tweets και έχει αποδειχθεί ότι ξεπερνά σε ορισμένες περιπτώσεις τους μεμονωμένους ανθρώπινους αξιολογητές.

2. Πλεονεκτήματα και περιορισμοί:

- Πλεονεκτήματα: Οι προσεγγίσεις που βασίζονται σε λεξικά είναι εύκολο να εφαρμοστούν και να ερμηνευτούν. Δεν απαιτούν δεδομένα εκπαίδευσης, γεγονός που τις καθιστά χρήσιμες για γλώσσες ή τομείς όπου τα σχολιασμένα σύνολα δεδομένων είναι σπάνια.

- Περιορισμοί: Αυτές οι προσεγγίσεις μπορεί να δυσκολεύονται με τα συμφραζόμενα, τον σαρκασμό και την εξελισσόμενη γλωσσική χρήση. Βασίζονται επίσης σε μεγάλο βαθμό στην ποιότητα και την πληρότητα του λεξικού συναισθήματος, το οποίο μπορεί να μην καλύπτει όλες τις σχετικές εκφράσεις και ιδιωτισμούς [3,7].

Προσεγγίσεις μηχανικής μάθησης

Οι προσεγγίσεις μηχανικής μάθησης περιλαμβάνουν την εκπαίδευση αλγορίθμων σε επισημασμένα σύνολα δεδομένων για την εκμάθηση προτύπων που σχετίζονται με διαφορετικά

συναισθήματα. Αυτές οι μέθοδοι χρησιμοποιούν διάφορα χαρακτηριστικά που εξάγονται από το κείμενο για τη δημιουργία προγνωστικών μοντέλων.

1. Κοινών αλγορίθμων:

- Naive Bayes: Ένας ταξινομητής που εφαρμόζει το θεώρημα του Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Είναι απλός και αποτελεσματικός, και συχνά χρησιμεύει ως βασική γραμμή για εργασίες ανάλυσης συναισθήματος.

- Support Vector Machines (SVM): Η SVM είναι ένας διαχωριστικός ταξινομητής που βρίσκει το υπερεπίπεδο που διαχωρίζει καλύτερα τα δεδομένα σε διαφορετικές κλάσεις συναισθήματος. Είναι αποτελεσματικός για δεδομένα κειμένου υψηλής διάστασης και χρησιμοποιείται ευρέως στην ανάλυση συναισθήματος.

- Λογιστική παλινδρόμηση: Ένα γραμμικό μοντέλο που εκτιμά την πιθανότητα ενός δυαδικού αποτελέσματος με βάση τα χαρακτηριστικά εισόδου. Χρησιμοποιείται συνήθως για δυαδική ταξινόμηση συναισθήματος (θετικό έναντι αρνητικού) και μπορεί να επεκταθεί σε ταξινόμηση πολλαπλών κλάσεων[9,24].

2. Εξαγωγή χαρακτηριστικών:

- Σάκος λέξεων (BoW): Αναπαριστά το κείμενο ως συλλογή λέξεων, αγνοώντας τη γραμματική και τη σειρά των λέξεων, αλλά διατηρώντας την πολλαπλότητα. Κάθε έγγραφο αναπαρίσταται ως διάνυσμα συχνοτήτων λέξεων ή δυαδικών δεικτών.

- Συχνότητα λέξεων-αντίστροφη συχνότητα εγγράφων (TF-IDF): Ζυγίζει τη συχνότητα των λέξεων ανάλογα με τη σημασία τους στο έγγραφο και σε ολόκληρο το σώμα κειμένων. Μειώνει την επιρροή των κοινών λέξεων που είναι λιγότερο κατατοπιστικές.

- Ενσωματώσεις λέξεων: Πυκνές διανυσματικές αναπαραστάσεις λέξεων που αποτυπώνουν σημασιολογικές σχέσεις. Μοντέλα όπως το Word2Vec, το GloVe και το FastText παράγουν ενσωμάτωση λέξεων που βελτιώνουν την ικανότητα σύλληψης των ομοιοτήτων του πλαισίου.

3. Πλεονεκτήματα και περιορισμοί:

- Πλεονεκτήματα: Τα μοντέλα μηχανικής μάθησης μπορούν να γενικεύουν καλά από τα δεδομένα εκπαίδευσης σε νέα, αθέατα δεδομένα. Είναι ευέλικτα και μπορούν να ρυθμιστούν λεπτομερώς για συγκεκριμένες εργασίες και τομείς.

- Περιορισμοί: Αυτά τα μοντέλα απαιτούν μεγάλα μαρκαρισμένα σύνολα δεδομένων για εκπαίδευση, η απόκτηση των οποίων μπορεί να είναι χρονοβόρα και δαπανηρή. Μπορεί επίσης να δυσκολεύονται με τις εξαρτώμενες από τα συμφραζόμενα έννοιες και απαιτούν προσεκτική μηχανική των χαρακτηριστικών [3,10,11,18]

Προσεγγίσεις βαθιάς μάθησης

Οι προσεγγίσεις βαθιάς μάθησης έχουν φέρει επανάσταση στην ανάλυση συναισθήματος, αξιοποιώντας τα νευρωνικά δίκτυα για τη σύλληψη σύνθετων μοτίβων και πληροφοριών πλαισίου στο κείμενο.

1. Συγχρονισμένα νευρωνικά δίκτυα (RNN):

- Τα RNN έχουν σχεδιαστεί για να χειρίζονται διαδοχικά δεδομένα διατηρώντας μια κρυφή κατάσταση που συλλαμβάνει πληροφορίες από προηγούμενα βήματα της ακολουθίας. Είναι ιδιαίτερα αποτελεσματικά για εργασίες που περιλαμβάνουν διαδοχικές εξαρτήσεις, όπως η ανάλυση συναισθήματος προτάσεων.
- Δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM): Τα LSTM είναι ένας τύπος RNN που αντιμετωπίζει το πρόβλημα της εξαφανιζόμενης κλίσης, επιτρέποντας στο μοντέλο να συλλάβει μακροπρόθεσμες εξαρτήσεις. Έχουν χρησιμοποιηθεί ευρέως στην ανάλυση συναισθήματος λόγω της ικανότητάς τους να θυμούνται μακροπρόθεσμα συμφραζόμενα.

2. Μοντέλα βασισμένα σε μετασχηματιστές:

- BERT (Bidirectional Encoder Representations from Transformers): Το BERT είναι ένα προ-εκπαιδευμένο μοντέλο μετασχηματιστή που αποτυπώνει το αμφίδρομο πλαίσιο στο κείμενο. Χρησιμοποιεί έναν αντικειμενικό στόχο γλωσσικού μοντέλου με μάσκα για να μάθει βαθιές αναπαραστάσεις συμφραζομένων. Το BERT έχει θέσει νέα σημεία αναφοράς σε διάφορες εργασίες NLP, συμπεριλαμβανομένης της ανάλυσης συναισθήματος.
- RoBERTa (A Robustly Optimized BERT Pretraining Approach): Το RoBERTa βασίζεται στο BERT βελτιστοποιώντας τη διαδικασία προεκπαίδευσης, χρησιμοποιώντας μεγαλύτερα σύνολα δεδομένων και εκπαιδώντας με μεγαλύτερα μεγέθη παρτίδων. Έχει επιδείξει ανώτερες επιδόσεις στην κατανόηση των σχέσεων πλαισίου και στο χειρισμό των διαφοροποιημένων συναισθημάτων.

3. Υβριδικά μοντέλα:

- RoBERTa-LSTM: Αυτό το υβριδικό μοντέλο συνδυάζει τα πλεονεκτήματα του RoBERTa και του LSTM. Το RoBERTa συλλαμβάνει τις πληροφορίες πλαισίου μέσω της αρχιτεκτονικής Transformer, ενώ το LSTM επεξεργάζεται τις διαδοχικές εξαρτήσεις. Αυτός ο συνδυασμός ενισχύει την ικανότητα του μοντέλου να κατανοεί σύνθετα γλωσσικά μοτίβα και να παρέχει ακριβή ταξινόμηση συναισθημάτων [3,9,10].

Συγκριτική ανάλυση

Συγκρίνοντας αυτές τις μεθοδολογίες, κάθε προσέγγιση έχει τα μοναδικά της πλεονεκτήματα και την εφαρμογή της:

- Οι προσεγγίσεις που βασίζονται σε λεξικά είναι καταλληλότερες για σενάρια όπου η απλότητα και η ερμηνευσιμότητα είναι υψίστης σημασίας και τα δεδομένα εκπαίδευσης είναι περιορισμένα. Είναι χρήσιμες για γρήγορη ανάλυση συναισθήματος σε πραγματικό χρόνο, αλλά μπορεί να δυσκολευτούν με την πολυπλοκότητα και τη μεταβλητότητα της φυσικής γλώσσας.
- Οι προσεγγίσεις μηχανικής μάθησης προσφέρουν μια ισορροπία μεταξύ απόδοσης και πολυπλοκότητας. Είναι ιδιαίτερα ευέλικτες και μπορούν να προσαρμοστούν σε συγκεκριμένες εργασίες και τομείς, αλλά απαιτούν εκτεταμένα επισημασμένα σύνολα δεδομένων για εκπαίδευση.
- Προσεγγίσεις βαθιάς μάθησης, ιδίως εκείνες που περιλαμβάνουν μοντέλα βασισμένα σε μετασχηματιστές, παρέχουν την υψηλότερη ακρίβεια και ευρωστία στην ανάλυση συναισθήματος. Είναι ικανές να χειριστούν τα περίπλοκα μοτίβα και τις εξαρτήσεις στη γλώσσα, αλλά απαιτούν σημαντική υπολογιστική ισχύ και δεδομένα μεγάλης κλίμακας.

Συμπερασματικά, η επιλογή της μεθοδολογίας στην ανάλυση συναισθήματος εξαρτάται από τις ειδικές απαιτήσεις της εργασίας, συμπεριλαμβανομένων των διαθέσιμων δεδομένων, της επιθυμητής ακρίβειας και των υπολογιστικών πόρων. Οι εξελίξεις στη βαθιά μάθηση, ιδίως με μοντέλα όπως το RoBERTa, έχουν θέσει νέα πρότυπα στον τομέα, προσφέροντας ισχυρά εργαλεία για την κατανόηση και την ανάλυση συναισθημάτων σε δεδομένα κειμένου. Καθώς η έρευνα εξελίσσεται, τα υβριδικά μοντέλα και οι βελτιωμένες τεχνικές εκπαίδευσης συνεχίζουν να ενισχύουν τις δυνατότητες και τις εφαρμογές της ανάλυσης συναισθήματος σε διάφορους τομείς [3,10,11,24].

2.1.3 Εφαρμογές της ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος έχει ευρύ φάσμα εφαρμογών σε διάφορους τομείς, παρέχοντας πολύτιμες πληροφορίες για την κοινή γνώμη, τις προτιμήσεις των πελατών και τις κοινωνικές τάσεις. Ακολουθούν ορισμένες εξέχουσες εφαρμογές:

1. Επιχειρήσεις και μάρκετινγκ:

- Ανάλυση ανατροφοδότησης πελατών: Οι επιχειρήσεις χρησιμοποιούν την ανάλυση συναισθήματος για να παρακολουθούν και να αναλύουν την ανατροφοδότηση των πελατών από κριτικές, έρευνες και μέσα κοινωνικής δικτύωσης. Αυτό βοηθά στην κατανόηση της ικανοποίησης των πελατών, στον εντοπισμό τομέων για βελτίωση και στη βελτίωση των προϊόντων και των υπηρεσιών.
- Παρακολούθηση μάρκας: Οι εταιρείες παρακολουθούν τις αναφορές της μάρκας τους σε όλες τις πλατφόρμες κοινωνικής δικτύωσης για να μετρήσουν το δημόσιο συναίσθημα. Οι θετικές αναφορές μπορούν να αξιοποιηθούν για εκστρατείες μάρκετινγκ, ενώ οι αρνητικές αναφορές μπορούν να προειδοποιήσουν τις εταιρείες για πιθανά ζητήματα που πρέπει να αντιμετωπιστούν.

- Έρευνα αγοράς: Η ανάλυση συναισθήματος βοηθά τις επιχειρήσεις να κατανοήσουν τις τάσεις της αγοράς και τις προτιμήσεις των καταναλωτών, επιτρέποντάς τους να λαμβάνουν αποφάσεις βάσει δεδομένων. Μπορεί επίσης να χρησιμοποιηθεί για την ανάλυση των δυνατών και αδύναμων σημείων των ανταγωνιστών, εξετάζοντας το κοινό συναίσθημα απέναντι στα προϊόντα και τις υπηρεσίες τους.

2. Πολιτική και κοινή γνώμη:

- Εκλογικές εκστρατείες: Οι πολιτικές εκστρατείες χρησιμοποιούν την ανάλυση συναισθήματος για να μετρήσουν το συναίσθημα των ψηφοφόρων και να προσαρμόσουν ανάλογα τα μηνύματά τους. Αναλύοντας αναρτήσεις στα μέσα κοινωνικής δικτύωσης και άρθρα ειδήσεων, μπορούν να εντοπίσουν τα βασικά θέματα που βρίσκουν απήχηση στους ψηφοφόρους και να προσαρμόσουν τις στρατηγικές τους σε πραγματικό χρόνο.

- Διαμόρφωση πολιτικής: Οι υπεύθυνοι χάραξης πολιτικής παρακολουθούν το κοινό συναίσθημα σχετικά με προτεινόμενους νόμους και πολιτικές για να κατανοήσουν την υποστήριξη ή την αντίθεση του κοινού. Αυτή η ανατροφοδότηση σε πραγματικό χρόνο βοηθά στη διαμόρφωση πολιτικών που ευθυγραμμίζονται περισσότερο με την κοινή γνώμη, αυξάνοντας την πιθανότητα επιτυχούς εφαρμογής [10,17].

3. Παρακολούθηση των μέσων κοινωνικής δικτύωσης:

- Ανάλυση τάσεων: Η ανάλυση συναισθήματος βοηθά στον εντοπισμό των θεμάτων που παρουσιάζουν τάση και στην κατανόηση του συναισθήματος του κοινού απέναντι σε αυτές τις τάσεις. Αυτό είναι ιδιαίτερα χρήσιμο για τους οργανισμούς μέσω ενημέρωσης και τους εμπόρους που επιθυμούν να αξιοποιήσουν τα τρέχοντα γεγονότα.

- Διαχείριση κρίσεων: Κατά τη διάρκεια κρίσεων, όπως φυσικές καταστροφές ή καταστάσεις έκτακτης ανάγκης για τη δημόσια υγεία, η ανάλυση συναισθήματος μπορεί να παρέχει πληροφορίες σε πραγματικό χρόνο σχετικά με τις ανησυχίες και τις αντιδράσεις του κοινού. Αυτό βοηθά τις αρχές και τους οργανισμούς να ανταποκρίνονται αποτελεσματικότερα, αντιμετωπίζοντας τα πιο πιεστικά ζητήματα που εντοπίζονται μέσω της ανάλυσης συναισθήματος.

4. Υγεία:

- Ανατροφοδότηση ασθενών: Η ανάλυση συναισθήματος της ανατροφοδότησης των ασθενών από φόρουμ υγείας και έρευνες βοηθά τους παρόχους υγειονομικής περίθαλψης να κατανοήσουν την ικανοποίηση των ασθενών και να εντοπίσουν τις κοινές ανησυχίες. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για τη βελτίωση των υπηρεσιών υγειονομικής περίθαλψης και της φροντίδας των ασθενών.

- Παρακολούθηση της ψυχικής υγείας: Η ανάλυση συναισθήματος των αναρτήσεων στα μέσα κοινωνικής δικτύωσης μπορεί να βοηθήσει στην παρακολούθηση των τάσεων της ψυχικής υγείας και στον εντοπισμό ατόμων που μπορεί να χρειάζονται υποστήριξη. Η εφαρμογή αυτή είναι ιδιαίτερα χρήσιμη για πρωτοβουλίες δημόσιας υγείας μεγάλης κλίμακας [10,12,20].

2.1.4 Προκλήσεις στην ανάλυση συναισθήματος

Ενώ η ανάλυση συναισθήματος προσφέρει σημαντικά οφέλη, αντιμετωπίζει επίσης αρκετές προκλήσεις που μπορούν να επηρεάσουν την αποτελεσματικότητά της. Η αντιμετώπιση αυτών των προκλήσεων είναι ζωτικής σημασίας για τη βελτίωση της ακρίβειας και της αξιοπιστίας των μοντέλων ανάλυσης συναισθήματος.

1. Σαρκασμός και ειρωνεία:

- Δυσκολία εντοπισμού: Ο σαρκασμός και η ειρωνεία αποτελούν σημαντικές προκλήσεις για τα μοντέλα ανάλυσης συναισθήματος, επειδή το επιδιωκόμενο συναίσθημα συχνά έρχεται σε αντίθεση με την κυριολεκτική σημασία των λέξεων. Για παράδειγμα, η φράση "Ωραία, άλλη μια βροχερή μέρα" είναι συνήθως αρνητική παρά τη θετική λέξη "Great" [12,21].

- Κατανόηση πλαισίου: Η αποτελεσματική ανίχνευση του σαρκασμού απαιτεί κατανόηση του πλαισίου, του τόνου και συχνά εξωτερική γνώση σχετικά με την κατάσταση στην οποία αναφέρεται. Τα τρέχοντα μοντέλα δυσκολεύονται να ενσωματώσουν αυτές τις πτυχές, οδηγώντας σε εσφαλμένη ταξινόμηση των σαρκαστικών δηλώσεων.

2. Κατανοητική κατανόηση του πλαισίου:

- Αμφισημία: Οι λέξεις μπορεί να έχουν διαφορετική σημασία ανάλογα με το πλαίσιο. Για παράδειγμα, η λέξη "επιβάρυνση" μπορεί να αναφέρεται σε οικονομική επιβάρυνση, σε ηλεκτρική επιβάρυνση ή στην ανάληψη ευθύνης. Τα μοντέλα ανάλυσης συναισθήματος πρέπει να αποσαφηνίζουν με ακρίβεια τέτοιους όρους για να ταξινομούν σωστά το συναίσθημα.

- Εξαρτήσεις σε μεγάλες αποστάσεις: Η καταγραφή των σχέσεων μεταξύ λέξεων που απέχουν πολύ μεταξύ τους σε μια πρόταση αποτελεί πρόκληση. Μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT και το RoBERTa, έχουν βελτιώσει αυτή την πτυχή, αλλά οι εξαρτήσεις μεγάλων αποστάσεων μπορούν ακόμα να προκαλέσουν προβλήματα, ειδικά σε μεγαλύτερα κείμενα[9,24].

3. Χειρισμός των αρνήσεων:

- Πολύπλοκες δομές: Οι απλές αρνήσεις όπως το "δεν είναι καλό" είναι σχετικά εύκολο να αντιμετωπιστούν, αλλά οι πιο σύνθετες δομές όπως το "όχι μόνο... αλλά και" ή οι διπλές αρνήσεις μπορούν να προκαλέσουν σύγχυση στα μοντέλα ανάλυσης συναισθήματος. Η σωστή ανάλυση και κατανόηση αυτών των δομών είναι απαραίτητη για την ακριβή ταξινόμηση του συναισθήματος.

4. Αμφισημία και υποκειμενικότητα:

- Διαφορετικές ερμηνείες: Το συναίσθημα μπορεί να είναι υποκειμενικό και διαφορετικά άτομα μπορεί να ερμηνεύουν διαφορετικά το ίδιο κείμενο. Αυτή η υποκειμενικότητα καθιστά δύσκολη τη δημιουργία μοντέλων που ευθυγραμμίζονται με συνέπεια με τις ανθρώπινες ερμηνείες σε διαφορετικά σύνολα δεδομένων.

- Συναισθήμα συγκεκριμένου τομέα: Οι λέξεις μπορεί να μεταφέρουν διαφορετικά συναισθήματα σε διαφορετικούς τομείς. Για παράδειγμα, η λέξη "άρρωστος" μπορεί να έχει αρνητική χροιά σε γενικά πλαίσια, αλλά θετική στην αθλητική αργκό (π.χ. "Αυτό το κόλπο ήταν άρρωστο"). Τα μοντέλα πρέπει να μπορούν να προσαρμόζονται σε διαφορετικούς τομείς και πλαίσια[9,12,24].

5. Πολυγλωσσική ανάλυση συναισθήματος:

- Γλωσσική μεταβλητότητα: Η ανάλυση συναισθήματος σε πολλές γλώσσες εισάγει πρόσθετη πολυπλοκότητα. Κάθε γλώσσα έχει τις δικές της γλωσσικές αποχρώσεις, ιδιωτισμούς και πολιτισμικό πλαίσιο, τα οποία τα μοντέλα πρέπει να κατανοούν και να επεξεργάζονται σωστά.

- Διαθεσιμότητα πόρων: Η ανάπτυξη μοντέλων ανάλυσης συναισθήματος για πολλαπλές γλώσσες απαιτεί εκτεταμένα σύνολα δεδομένων με ετικέτες, τα οποία μπορεί να μην είναι άμεσα διαθέσιμα για λιγότερο διαδεδομένες γλώσσες. Αυτή η έλλειψη πόρων δυσχεραίνει την ανάπτυξη και την ακρίβεια των πολύγλωσσων μοντέλων ανάλυσης συναισθήματος [3,12,21].

6. Ανισορροπία δεδομένων:

- Ανισορροπία κλάσεων: Σε πολλά σύνολα δεδομένων, η κατανομή των κλάσεων συναισθήματος είναι ανισοβαρής, με μια κλάση (π.χ. ουδέτερη) να είναι σημαντικά συχνότερη από άλλες (π.χ. θετική ή αρνητική). Αυτή η ανισορροπία μπορεί να οδηγήσει σε μεροληπτικά μοντέλα που αποδίδουν καλά στην πλειοψηφούσα κλάση αλλά ελάχιστα στις μειοψηφούσες κλάσεις.

- Τεχνικές επαύξησης: Η αντιμετώπιση της ανισορροπίας κλάσεων συχνά απαιτεί τεχνικές επαύξησης δεδομένων, όπως το oversampling κλάσεων μειονότητας ή η δημιουργία συνθετικών δεδομένων. Αυτές οι τεχνικές πρέπει να εφαρμόζονται προσεκτικά για να αποφεύγεται η εισαγωγή θορύβου ή προκαταλήψεων στο σύνολο δεδομένων[9,24].

2.2 Τεχνικές επιλογής χαρακτηριστικών: Σημασία και λογική

Η επιλογή χαρακτηριστικών είναι ένα κρίσιμο βήμα στη φάση preprocessing της μηχανικής μάθησης και της εξόρυξης δεδομένων, ιδίως για εργασίες που περιλαμβάνουν δεδομένα υψηλής διάστασης, όπως η ταξινόμηση κειμένου και η ανάλυση συναισθήματος. Ο κύριος στόχος της επιλογής χαρακτηριστικών είναι ο εντοπισμός και η διατήρηση των πιο κατατοπιστικών χαρακτηριστικών από το σύνολο δεδομένων, ενώ απορρίπτονται εκείνα που είναι περιττά ή άσχετα. Η διαδικασία αυτή όχι μόνο ενισχύει την απόδοση των μοντέλων μηχανικής μάθησης αλλά και βελτιώνει την ερμηνευσιμότητά τους και μειώνει το υπολογιστικό κόστος.

2.2.1 Σημασία χρήσης επιλογής χαρακτηριστικών

1. Βελτίωση της απόδοσης του μοντέλου:

-Μείωση του hyperadaptation: Τα σύνολα δεδομένων υψηλής διάστασης συχνά περιέχουν μεγάλο αριθμό άσχετων ή περιττών χαρακτηριστικών, γεγονός που μπορεί να οδηγήσει σε

hyperadaptation. Συμβαίνει όταν ένα μοντέλο μαθαίνει το θόρυβο στα δεδομένα εκπαίδευσης αντί για το υποκείμενο πρότυπο, με αποτέλεσμα την κακή γενίκευση σε νέα, αόρατα δεδομένα. Με την επιλογή μόνο των πιο σχετικών χαρακτηριστικών, η επιλογή χαρακτηριστικών βοηθά στη δημιουργία απλούστερων μοντέλων που γενικεύουν καλύτερα και είναι λιγότερο επιρρεπή στο hyperadaptation.

- Βελτιωμένη ακρίβεια: Εστιάζοντας στα πιο κατατοπιστικά χαρακτηριστικά, η επιλογή χαρακτηριστικών μπορεί να βελτιώσει την ακρίβεια των μοντέλων μηχανικής μάθησης. Τα άσχετα χαρακτηριστικά μπορούν να εισάγουν θόρυβο στο μοντέλο, οδηγώντας σε ανακριβείς προβλέψεις. Η επιλογή χαρακτηριστικών βοηθά στην εξάλειψη αυτού του θορύβου, ενισχύοντας έτσι την προβλεπτική ικανότητα του μοντέλου [1,3,13,15,16].

2. Μείωση του υπολογιστικού κόστους:

- Αποτελεσματικότητα στην εκπαίδευση και την πρόβλεψη: Τα δεδομένα υψηλής διάστασης αυξάνουν την υπολογιστική πολυπλοκότητα της εκπαίδευσης των μοντέλων μηχανικής μάθησης. Η επιλογή χαρακτηριστικών μειώνει τον αριθμό των χαρακτηριστικών, μειώνοντας έτσι το χρόνο και τους πόρους που απαιτούνται για την εκπαίδευση. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές πραγματικού χρόνου, όπου η γρήγορη εκπαίδευση και πρόβλεψη μοντέλων είναι ζωτικής σημασίας.

- Αποδοτικότητα αποθήκευσης και μνήμης: Λιγότερα χαρακτηριστικά σημαίνουν λιγότερο αποθηκευτικό χώρο και λιγότερη χρήση μνήμης, καθιστώντας εφικτή τη διαχείριση μεγάλων συνόλων δεδομένων που διαφορετικά θα ήταν υπολογιστικά απαγορευτικά. Αυτή η αποδοτικότητα είναι ζωτικής σημασίας για την ανάπτυξη μοντέλων μηχανικής μάθησης σε περιβάλλοντα με περιορισμένους πόρους, όπως κινητές συσκευές και ενσωματωμένα συστήματα [3,17].

3. Βελτίωση της ερμηνευσιμότητας των μοντέλων:

- Απλουστευμένα μοντέλα: Τα μοντέλα με λιγότερα χαρακτηριστικά είναι ευκολότερα στην κατανόηση και την ερμηνεία. Αυτό είναι ιδιαίτερα σημαντικό σε τομείς όπου η διαφάνεια των μοντέλων είναι ζωτικής σημασίας, όπως η υγειονομική περίθαλψη, η χρηματοδότηση και οι νομικοί τομείς. Τα απλοποιημένα μοντέλα βοηθούν τους ενδιαφερόμενους να κατανοήσουν τη διαδικασία λήψης αποφάσεων και να οικοδομήσουν εμπιστοσύνη στις προβλέψεις του μοντέλου.

- Διαφωτιστική ανάλυση: Η επιλογή χαρακτηριστικών μπορεί να παρέχει πολύτιμες πληροφορίες για τα δεδομένα, εντοπίζοντας τα χαρακτηριστικά με τη μεγαλύτερη επιρροή. Για παράδειγμα, στην ανάλυση συναισθήματος, η κατανόηση του ποιες λέξεις ή φράσεις επηρεάζουν πιο σημαντικά την ταξινόμηση του συναισθήματος μπορεί να προσφέρει βαθύτερες γνώσεις σχετικά με τα γλωσσικά πρότυπα και τους οδηγούς του συναισθήματος.

Το σκεπτικό για την επιλογή χαρακτηριστικών βασίζεται σε διάφορες βασικές αρχές και οφέλη που συμβάλλουν στη συνολική αποτελεσματικότητα και αποδοτικότητα των μοντέλων μηχανικής μάθησης:

1. διαστατικότητα: - Τα σύνολα δεδομένων υψηλής διάστασης αποτελούν μια σημαντική πρόκληση γνωστή ως "κατάρρα της διαστατικότητας". Καθώς αυξάνεται ο αριθμός των χαρακτηριστικών, ο όγκος του χώρου των χαρακτηριστικών αυξάνεται εκθετικά, καθιστώντας δύσκολη την αποτελεσματική μάθηση των μοντέλων. Η επιλογή χαρακτηριστικών μετριάζει αυτό το πρόβλημα μειώνοντας τη διαστατικότητα των δεδομένων, διευκολύνοντας τα μοντέλα να εντοπίζουν σημαντικά μοτίβα και σχέσεις [8,14,19].

2. Σχετικότητα και πλεονασμός:

- Σχετικότητα: Σχετικά χαρακτηριστικά είναι εκείνα που έχουν σημαντική επίδραση στη μεταβλητή-στόχο. Η συμπερίληψη μόνο σχετικών χαρακτηριστικών διασφαλίζει ότι το μοντέλο εστιάζει στις πιο σημαντικές πτυχές των δεδομένων, οδηγώντας σε πιο ακριβείς προβλέψεις.

- Περίσσεια: Τα περιττά χαρακτηριστικά είναι εκείνα που δεν παρέχουν πρόσθετες πληροφορίες πέραν αυτών που ήδη καταγράφονται από άλλα χαρακτηριστικά. Η αφαίρεση των περιττών χαρακτηριστικών απλοποιεί το μοντέλο χωρίς να χάνεται η προβλεπτική ισχύς. Αυτό βοηθά επίσης στην αποφυγή του multilinearity, όπου τα έντονα συσχετιζόμενα χαρακτηριστικά μπορούν να στρεβλώσουν την απόδοση του μοντέλου.

3. Μείωση θορύβου:

- Τα σύνολα δεδομένων συχνά περιέχουν θορυβώδη δεδομένα, τα οποία μπορούν να υποβαθμίσουν την απόδοση των μοντέλων μηχανικής μάθησης. Ο θόρυβος μπορεί να προκύψει από διάφορες πηγές, όπως σφάλματα μέτρησης, λάθη κατά την εισαγωγή δεδομένων ή άσχετες πληροφορίες. Η επιλογή χαρακτηριστικών βοηθά στην εξάλειψη των θορυβωδών χαρακτηριστικών, οδηγώντας σε καθαρότερα και πιο ισχυρά μοντέλα [1,4,18,22].

4. Βελτιωμένη γενίκευση:

- Τα μοντέλα που εκπαιδεύονται σε δεδομένα υψηλών διαστάσεων μπορεί να αποδίδουν καλά στα δεδομένα εκπαίδευσης, αλλά να μην μπορούν να γενικευτούν σε νέα δεδομένα. Αυτό συχνά οφείλεται στο γεγονός ότι το μοντέλο μαθαίνει ψευδή μοτίβα ειδικά για το σύνολο εκπαίδευσης. Με την επιλογή ενός υποσυνόλου χαρακτηριστικών που είναι πραγματικά κατατοπιστικά, η επιλογή χαρακτηριστικών βελτιώνει την ικανότητα του μοντέλου να γενικεύει σε αόρατα δεδομένα, ενισχύοντας έτσι τη δυνατότητα εφαρμογής του στον πραγματικό κόσμο.

5. Scalability:

- Στις πρακτικές εφαρμογές, η επεκτασιμότητα αποτελεί μείζον ζήτημα, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων. Η επιλογή χαρακτηριστικών επιτρέπει την ανάπτυξη κλιμακούμενων λύσεων μηχανικής μάθησης μειώνοντας τον υπολογιστικό φόρτο που συνδέεται με την επεξεργασία δεδομένων υψηλής διάστασης. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές μεγάλων δεδομένων και συστήματα πραγματικού χρόνου όπου η γρήγορη λήψη αποφάσεων είναι απαραίτητη [4].

2.2.2 Τύποι μεθόδων επιλογής χαρακτηριστικών

Η επιλογή χαρακτηριστικών είναι ένα θεμελιώδες βήμα στη φάση preprocessing της μηχανικής μάθησης που αποσκοπεί στην επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών από το σύνολο δεδομένων. Η διαδικασία αυτή συμβάλλει στη βελτίωση της απόδοσης του μοντέλου, στη μείωση του hyper-adaptation και στην ενίσχυση της ερμηνευσιμότητας. Υπάρχουν τρεις βασικοί τύποι μεθόδων επιλογής χαρακτηριστικών: μέθοδοι φίλτρου, μέθοδοι wrapper και ενσωματωμένες μέθοδοι. Κάθε μία από αυτές τις μεθόδους χρησιμοποιεί διαφορετικές στρατηγικές για την αξιολόγηση και την επιλογή χαρακτηριστικών.

1. Μέθοδοι φίλτρου

Οι μέθοδοι φίλτρου αξιολογούν τη συνάφεια των χαρακτηριστικών με βάση τις εγγενείς ιδιότητές τους, ανεξάρτητα από οποιονδήποτε αλγόριθμο μηχανικής μάθησης. Αυτές οι μέθοδοι χρησιμοποιούν συνήθως στατιστικές τεχνικές για την αξιολόγηση της σχέσης μεταξύ κάθε χαρακτηριστικού και της μεταβλητής-στόχου.

- Συντελεστής συσχέτισης:

- Μετρά τη γραμμική σχέση μεταξύ κάθε χαρακτηριστικού και της μεταβλητής-στόχου. Χαρακτηριστικά με υψηλή συσχέτιση (θετική ή αρνητική) με τη μεταβλητή-στόχο θεωρούνται συναφή. Αυτή η μέθοδος είναι απλή και υπολογιστικά αποδοτική, αλλά καταγράφει μόνο τις γραμμικές εξαρτήσεις.

- Δοκιμή Chi-Squared:

- Αξιολογεί την ανεξαρτησία μεταξύ κατηγορικών χαρακτηριστικών και της μεταβλητής-στόχου. Τα χαρακτηριστικά που εξαρτώνται σε μεγάλο βαθμό από τη μεταβλητή-στόχο θεωρούνται σημαντικά. Αυτή η δοκιμή είναι ιδιαίτερα χρήσιμη για διακριτά δεδομένα.

- Αμοιβαίες πληροφορίες:

- Μετρά το ποσό της πληροφορίας που λαμβάνεται για μια τυχαία μεταβλητή μέσω μιας άλλης τυχαίας μεταβλητής. Καταγράφει τόσο γραμμικές όσο και μη γραμμικές σχέσεις, καθιστώντας το πιο ευέλικτο από τους συντελεστές συσχέτισης.

-Variance Threshold:

- Αφαιρεί τα χαρακτηριστικά με χαμηλή διακύμανση, θεωρώντας ότι τα χαρακτηριστικά με χαμηλή διακύμανση δεν μεταφέρουν χρήσιμες πληροφορίες. Αυτή η μέθοδος είναι απλή, αλλά μπορεί να απορρίψει σχετικά χαρακτηριστικά που έχουν χαμηλή διακύμανση λόγω της φύσης των δεδομένων.

Πλεονεκτήματα και περιορισμοί:

- Πλεονεκτήματα: Οι μέθοδοι φίλτρων είναι υπολογιστικά αποδοτικές και εύκολα υλοποιήσιμες. Είναι κατάλληλες για σύνολα δεδομένων υψηλής διάστασης και παρέχουν μια καλή βάση για την επιλογή χαρακτηριστικών.

- Περιορισμοί: Αυτές οι μέθοδοι εξετάζουν κάθε χαρακτηριστικό ανεξάρτητα και δεν λαμβάνουν υπόψη τις αλληλεπιδράσεις των χαρακτηριστικών. Ως αποτέλεσμα, ενδέχεται να παραβλέψουν χαρακτηριστικά που είναι σημαντικά σε συνδυασμό με άλλα [1,4,5,21,23].

2. Μέθοδοι wrapper

Οι μέθοδοι wrapper αξιολογούν τη χρησιμότητα ενός υποσυνόλου χαρακτηριστικών με την εκπαίδευση και την αξιολόγηση ενός μοντέλου μηχανικής μάθησης σε διαφορετικά υποσύνολα χαρακτηριστικών. Αυτές οι μέθοδοι λαμβάνουν υπόψη τις αλληλεπιδράσεις των χαρακτηριστικών και είναι συνήθως πιο ακριβείς από τις μεθόδους φίλτρων, αλλά είναι υπολογιστικά δαπανηρές.

Κοινές τεχνικές:

- Επιλογή προς τα εμπρός:

- Ξεκινά με ένα κενό σύνολο χαρακτηριστικών και προσθέτει επαναληπτικά το χαρακτηριστικό που βελτιώνει περισσότερο την απόδοση του μοντέλου μέχρι να μην παρατηρείται σημαντική βελτίωση. Αυτή η μέθοδος είναι απλή αλλά μπορεί να είναι υπολογιστικά εντατική για μεγάλα σύνολα δεδομένων.

- Εξάλειψη προς τα πίσω:

- Ξεκινά με το πλήρες σύνολο χαρακτηριστικών και αφαιρεί επαναληπτικά το λιγότερο σημαντικό χαρακτηριστικό μέχρι να μην παρατηρηθεί σημαντική μείωση της απόδοσης. Αυτή η μέθοδος είναι αποτελεσματική αλλά μπορεί να είναι υπολογιστικά απαιτητική.

- Αναδρομική εξάλειψη χαρακτηριστικών (RFE):

- Εκπαιδεύει ένα μοντέλο και αφαιρεί επαναληπτικά τα λιγότερο σημαντικά χαρακτηριστικά με βάση τους συντελεστές του μοντέλου ή τη σημασία των χαρακτηριστικών. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών. Η RFE είναι ισχυρή και παρέχει μια καλή ισορροπία μεταξύ ακρίβειας και υπολογιστικής απόδοσης [1,4,8,14,15].

Πλεονεκτήματα και περιορισμοί:

- Πλεονεκτήματα: Οι μέθοδοι wrapper λαμβάνουν υπόψη τις αλληλεπιδράσεις των χαρακτηριστικών και παρέχουν υψηλότερη ακρίβεια σε σύγκριση με τις μεθόδους φίλτρων. Είναι κατάλληλες για σύνολα δεδομένων όπου οι εξαρτήσεις των χαρακτηριστικών παίζουν σημαντικό ρόλο.

- Περιορισμοί: Αυτές οι μέθοδοι είναι υπολογιστικά εντατικές και ενδέχεται να μην είναι εφικτές για σύνολα δεδομένων πολύ υψηλών διαστάσεων. Κινδυνεύουν επίσης από υπερβολική προσαρμογή στα δεδομένα εκπαίδευσης, εάν δεν γίνει προσεκτική διαχείριση.

3. Ενσωματωμένες μέθοδοι

Οι ενσωματωμένες μέθοδοι εκτελούν επιλογή χαρακτηριστικών κατά τη διαδικασία εκπαίδευσης του μοντέλου. Αυτές οι μέθοδοι ενσωματώνονται σε συγκεκριμένους αλγορίθμους μάθησης και επιλέγουν χαρακτηριστικά ως μέρος της κατασκευής του μοντέλου.

Κοινές τεχνικές:

- LASSO (Λειτουργός ελάχιστης απόλυτης συρρίκνωσης και επιλογής):

- Μια τεχνική κανονικοποίησης που προσθέτει μια ποινή ίση με την απόλυτη τιμή του μεγέθους των συντελεστών. Αυτή η ποινή προκαλεί τη συρρίκνωση των λιγότερο σημαντικών συντελεστών χαρακτηριστικών στο μηδέν, εκτελώντας ουσιαστικά επιλογή χαρακτηριστικών. Ο LASSO είναι χρήσιμος για γραμμικά μοντέλα και είναι υπολογιστικά αποδοτικός.

- Ridge Regression:

- Παρόμοια με την LASSO, αλλά χρησιμοποιεί το τετραγωνικό μέγεθος των συντελεστών για την κανονικοποίηση. Ενώ η παλινδρόμηση Ridge δεν εκτελεί άμεσα επιλογή χαρακτηριστικών (αφού δεν συρρικνώνει τους συντελεστές στο μηδέν), βοηθά στη μείωση της πολυπλοκότητας του μοντέλου και της πολυσυγγραμμικότητας.

- Elastic Net:

- Συνδυάζει τις ποινές της LASSO και της Ridge Regression. Αυτή η μέθοδος είναι κατάλληλη για σύνολα δεδομένων με έντονα συσχετιζόμενα χαρακτηριστικά και μπορεί να εκτελέσει επιλογή χαρακτηριστικών διατηρώντας παράλληλα τη σταθερότητα του μοντέλου.

- Μέθοδοι βασισμένες σε δέντρα:

- Τα δέντρα αποφάσεων, τα τυχαία δάση και τα δέντρα ενίσχυσης κλίσης εκτελούν εγγενώς επιλογή χαρακτηριστικών με διαχωρισμό των κόμβων με βάση τη σημασία των χαρακτηριστικών. Τα χαρακτηριστικά που συμβάλλουν περισσότερο στη μείωση της ακαθαρσίας θεωρούνται πιο σημαντικά. Αυτές οι μέθοδοι είναι ισχυρές για την καταγραφή μη γραμμικών σχέσεων και αλληλεπιδράσεων μεταξύ των χαρακτηριστικών [1,5,12,21].

Πλεονεκτήματα και περιορισμοί:

- Πλεονεκτήματα: Οι ενσωματωμένες μέθοδοι είναι λιγότερο υπολογιστικά εντατικές από τις μεθόδους wrapper και μπορούν να χειριστούν αποτελεσματικά μεγάλα σύνολα δεδομένων. Ενσωματώνουν την επιλογή χαρακτηριστικών με την εκπαίδευση του μοντέλου, παρέχοντας μια βελτιωμένη προσέγγιση.

- Περιορισμοί: Αυτές οι μέθοδοι είναι συγκεκριμένες για ορισμένους αλγορίθμους και ενδέχεται να μην είναι εφαρμόσιμες σε όλους τους τύπους μοντέλων. Η επιλογή των παραμέτρων κανονικοποίησης είναι κρίσιμη και μπορεί να επηρεάσει σημαντικά την απόδοση του μοντέλου [3,12].

2.2.3 Συγκριτική ανάλυση των μεθόδων επιλογής χαρακτηριστικών

Κάθε τύπος μεθόδου επιλογής χαρακτηριστικών έχει τα μοναδικά του πλεονεκτήματα και τη δυνατότητα εφαρμογής ανάλογα με το σύνολο δεδομένων και το πρόβλημα που αντιμετωπίζουμε:

- Μέθοδοι φίλτρων: Καταλληλότερες για προκαταρκτική επιλογή χαρακτηριστικών, ιδίως σε σύνολα δεδομένων υψηλών διαστάσεων. Παρέχουν έναν γρήγορο τρόπο μείωσης του χώρου των χαρακτηριστικών, αλλά μπορεί να χρειαστεί να συμπληρωθούν από πιο εξελιγμένες μεθόδους για την καταγραφή των αλληλεπιδράσεων των χαρακτηριστικών.
- Μέθοδοι wrapper: Παρέχουν υψηλή ακρίβεια και λαμβάνουν υπόψη τις εξαρτήσεις των χαρακτηριστικών, καθιστώντας τις κατάλληλες για λεπτομερή ανάλυση. Ωστόσο, το υπολογιστικό τους κόστος περιορίζει τη χρήση τους σε πολύ μεγάλα σύνολα δεδομένων.
- Ενσωματωμένες μέθοδοι: Προσφέρουν μια καλή ισορροπία μεταξύ υπολογιστικής αποδοτικότητας και ακρίβειας. Είναι ιδιαίτερα χρήσιμες για μοντέλα όπου η επιλογή χαρακτηριστικών είναι ενσωματωμένη στη διαδικασία μάθησης.

Σε πρακτικές εφαρμογές, χρησιμοποιείται συχνά ένας συνδυασμός αυτών των μεθόδων. Για παράδειγμα, οι μέθοδοι φίλτρων μπορούν να χρησιμοποιηθούν για μια αρχική μείωση του χώρου χαρακτηριστικών, ακολουθούμενες από μεθόδους wrapper ή ενσωματωμένες μεθόδους για τη λεπτομερή ρύθμιση. Αυτή η προσέγγιση αξιοποιεί τα πλεονεκτήματα κάθε μεθόδου, εξασφαλίζοντας ισχυρή και αποτελεσματική επιλογή χαρακτηριστικών.

2.2.3 Συνάφεια με την ανάλυση συναισθήματος

Η επιλογή χαρακτηριστικών είναι ιδιαίτερα κρίσιμη στον τομέα της ανάλυσης συναισθήματος λόγω της εγγενούς υψηλής διαστατικότητας των δεδομένων κειμένου. Τα δεδομένα κειμένου, όταν μετατρέπονται σε αριθμητικά διανύσματα, συχνά καταλήγουν σε έναν τεράστιο αριθμό χαρακτηριστικών, πολλά από τα οποία μπορεί να είναι άσχετα ή περιττά. Οι αποτελεσματικές τεχνικές επιλογής χαρακτηριστικών μπορούν να βελτιώσουν σημαντικά την απόδοση των μοντέλων ανάλυσης συναισθήματος, εντοπίζοντας και διατηρώντας μόνο τα πιο κατατοπιστικά χαρακτηριστικά. Η παρούσα ενότητα διερευνά τη σημασία της επιλογής χαρακτηριστικών για την ανάλυση συναισθήματος, υπογραμμίζοντας τον αντίκτυπό της σε διάφορες πτυχές της διαδικασίας.

2.3 Παραδείγματα εφαρμογών στην ανάλυση συναισθήματος

Μέθοδοι φίλτρων:

Σε μια μελέτη σχετικά με την ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή, οι ερευνητές χρησιμοποίησαν μεθόδους φίλτρων, όπως η αμοιβαία πληροφορία και τα τετράγωνα chi, για να εντοπίσουν σχετικά χαρακτηριστικά από ένα σύνολο δεδομένων υψηλής διάστασης. Μειώνοντας το χώρο των χαρακτηριστικών, μπόρεσαν να εκπαιδεύσουν πιο αποδοτικά μοντέλα

που θα μπορούσαν να αναλύσουν με μεγαλύτερη ακρίβεια το δημόσιο συναίσθημα για την κλιματική αλλαγή [5,22].

Μέθοδοι wrapper:

Μια άλλη μελέτη χρησιμοποίησε την αναδρομική απαλοιφή χαρακτηριστικών (RFE) στο πλαίσιο της ανάλυσης συναισθήματος σε δεδομένα κοινωνικών μέσων. Αφαιρώντας επαναληπτικά τα λιγότερο σημαντικά χαρακτηριστικά, οι ερευνητές βελτίωσαν την ακρίβεια των μοντέλων ταξινόμησης συναισθήματος. Η προσέγγιση αυτή βοήθησε στην καταγραφή πολύπλοκων αλληλεπιδράσεων μεταξύ των χαρακτηριστικών, οι οποίες συχνά υπάρχουν σε κείμενα κοινωνικών μέσων ενημέρωσης[9,24].

Ενσωματωμένες μέθοδοι:

Σε ένα έργο ανάλυσης συναισθήματος που αφορούσε κριτικές προϊόντων, χρησιμοποιήθηκαν ενσωματωμένες μέθοδοι όπως η παλινδρόμηση LASSO για την επιλογή σημαντικών χαρακτηριστικών κατά τη διαδικασία εκπαίδευσης του μοντέλου. Αυτή η τεχνική κανονικοποίησης βοήθησε στον εντοπισμό των πιο επιδραστικών λέξεων και φράσεων που επηρέασαν το συναίσθημα των πελατών, οδηγώντας σε πιο ερμηνεύσιμα και ακριβή μοντέλα [3,12].

Συμπέρασμα:

Η επιλογή χαρακτηριστικών είναι μια κρίσιμη διαδικασία στη μηχανική μάθηση που ενισχύει την απόδοση του μοντέλου, μειώνει την υπολογιστική πολυπλοκότητα και βελτιώνει την ερμηνευσιμότητα. Με την κατανόηση και την εφαρμογή των κατάλληλων μεθόδων επιλογής χαρακτηριστικών, οι επαγγελματίες μπορούν να δημιουργήσουν πιο αποτελεσματικά και αποδοτικά μοντέλα προσαρμοσμένα στις ειδικές απαιτήσεις των εργασιών τους. Καθώς τα σύνολα δεδομένων συνεχίζουν να αυξάνονται σε μέγεθος και πολυπλοκότητα, η σημασία των ισχυρών τεχνικών επιλογής χαρακτηριστικών θα αυξάνεται, καθιστώντας την έναν ζωτικό τομέα συνεχιζόμενης έρευνας και ανάπτυξης. Ο συνδυασμός των μεθόδων φίλτρου, wrapper και ενσωμάτωσης προσφέρει μια ολοκληρωμένη προσέγγιση για την αντιμετώπιση των προκλήσεων των δεδομένων υψηλής διάστασης, διασφαλίζοντας ότι τα πιο συναφή και κατατοπιστικά χαρακτηριστικά χρησιμοποιούνται για τη δημιουργία προγνωστικών μοντέλων.

2.4 Βελτίωση της απόδοσης του μοντέλου

Στην ανάλυση συναισθήματος, ο στόχος είναι η ταξινόμηση κειμένου με βάση το συναίσθημα που εκφράζεται - συνήθως ως θετικό, αρνητικό ή ουδέτερο. Τα δεδομένα κειμένου υψηλής διάστασης μπορούν να υπερφορτώσουν τα μοντέλα μηχανικής μάθησης, οδηγώντας σε υπερβολική προσαρμογή και κακή γενίκευση σε νέα δεδομένα. Η επιλογή χαρακτηριστικών αντιμετωπίζει αυτό το ζήτημα μειώνοντας τον χώρο των χαρακτηριστικών, απλοποιώντας έτσι το μοντέλο και βελτιώνοντας την απόδοσή του.

1. Βελτίωση της ακρίβειας και της αποδοτικότητας:

- Με την επιλογή των πιο σχετικών χαρακτηριστικών, η επιλογή χαρακτηριστικών βοηθά στην κατασκευή μοντέλων που είναι όχι μόνο πιο ακριβή αλλά και πιο αποδοτικά. Για παράδειγμα, στην ανάλυση συναισθήματος κριτικών προϊόντων, η αφαίρεση άσχετων λέξεων και φράσεων μπορεί να βελτιώσει την ικανότητα του μοντέλου να ταξινομεί σωστά τα συναισθήματα.

- Τεχνικές όπως η αμοιβαία πληροφορία και τα τετράγωνα χ^2 έχουν χρησιμοποιηθεί αποτελεσματικά στην ανάλυση συναισθήματος για τον εντοπισμό σημαντικών λέξεων και n-γραμμάτων που συμβάλλουν στο συνολικό συναίσθημα του κειμένου. Αυτές οι μέθοδοι διασφαλίζουν ότι μόνο τα πιο επιδραστικά χαρακτηριστικά περιλαμβάνονται στο μοντέλο, βελτιώνοντας έτσι την ακρίβεια ταξινόμησης [5,12,21,22].

2. Μείωση του hyper-adaptation:

- Το hyper-adaptation είναι ένα συνηθισμένο πρόβλημα στη μηχανική μάθηση, όπου το μοντέλο αποδίδει καλά στα δεδομένα εκπαίδευσης, αλλά αποτυγχάνει να γενικεύσει σε νέα, αθέατα δεδομένα. Τα δεδομένα υψηλής διάστασης επιδεινώνουν αυτό το πρόβλημα, καθώς το μοντέλο μπορεί να μάθει θόρυβο και άσχετα μοτίβα.

- Η επιλογή χαρακτηριστικών μετριάζει το hyper-adaptation εξαλείφοντας τα περιττά και άσχετα χαρακτηριστικά. Αυτό έχει ως αποτέλεσμα ένα απλούστερο μοντέλο που εστιάζει στις πιο σημαντικές πτυχές των δεδομένων, ενισχύοντας έτσι τις δυνατότητες γενίκευσης [5,9,24].

Ενίσχυση της ερμηνευσιμότητας:

Τα μοντέλα ανάλυσης συναισθήματος, ιδίως αυτά που χρησιμοποιούνται στις επιχειρήσεις και τις κοινωνικές επιστήμες, επωφελούνται σε μεγάλο βαθμό από την ερμηνευσιμότητά τους. Τα ενδιαφερόμενα μέρη πρέπει να κατανοούν γιατί ένα μοντέλο κάνει μια συγκεκριμένη πρόβλεψη, πράγμα που είναι κρίσιμο για την εμπιστοσύνη και τη διαφάνεια.

1. Απλουστευμένα μοντέλα:

- Η επιλογή χαρακτηριστικών οδηγεί σε μοντέλα που είναι ευκολότερα ερμηνεύσιμα. Για παράδειγμα, σε ένα μοντέλο ανάλυσης συναισθήματος που χρησιμοποιείται για τη μέτρηση της κοινής γνώμης στα μέσα κοινωνικής δικτύωσης, ο εντοπισμός των βασικών φράσεων ή λέξεων που καθοδηγούν το συναίσθημα μπορεί να παρέχει αξιοποιήσιμες πληροφορίες.

- Τα απλοποιημένα μοντέλα με λιγότερα χαρακτηριστικά διευκολύνουν τους αναλυτές και τους υπεύθυνους λήψης αποφάσεων να κατανοήσουν τους παράγοντες που επηρεάζουν το συναίσθημα. Αυτό μπορεί να ενημερώσει τις στρατηγικές στο μάρκετινγκ, την εξυπηρέτηση πελατών και τις δημόσιες σχέσεις [10,17].

2. Διαφωτιστική ανάλυση:

- Με την εστίαση στα πιο σχετικά χαρακτηριστικά, η επιλογή χαρακτηριστικών μπορεί να αναδείξει τους βασικούς παράγοντες του συναισθήματος. Στο πλαίσιο του διαλόγου για την κλιματική αλλαγή, ο εντοπισμός των συγκεκριμένων όρων και φράσεων που φέρουν έντονα

συναισθήματα μπορεί να βοηθήσει τους ερευνητές να κατανοήσουν καλύτερα τις ανησυχίες και τις απόψεις του κοινού.

- Οι μέθοδοι επιλογής χαρακτηριστικών, όπως η παλινδρόμηση LASSO και τα δέντρα αποφάσεων, κατατάσσουν εγγενώς τα χαρακτηριστικά με βάση τη σημασία τους, παρέχοντας μια σαφή ένδειξη για το ποιες λέξεις ή φράσεις έχουν τη μεγαλύτερη επιρροή στον καθορισμό του συναισθήματος[3,9].

Χειρισμός δεδομένων υψηλής διάστασης

Τα δεδομένα κειμένου, όταν μετασχηματίζονται σε αριθμητικές αναπαραστάσεις, συχνά καταλήγουν σε χιλιάδες χαρακτηριστικά, ιδίως όταν χρησιμοποιούνται τεχνικές όπως το bag-of-words ή το TF-IDF. Η διαχείριση αυτών των δεδομένων υψηλής διάστασης αποτελεί σημαντική πρόκληση στην ανάλυση συναισθήματος.

1. Μείωση των διαστάσεων:

- Οι τεχνικές επιλογής χαρακτηριστικών μειώνουν τη διαστατικότητα των δεδομένων κειμένου, καθιστώντας τα πιο εύχρηστα για τους αλγορίθμους μηχανικής μάθησης. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές ανάλυσης συναισθήματος σε πραγματικό χρόνο, όπου η υπολογιστική αποδοτικότητα είναι ζωτικής σημασίας.

- Τεχνικές όπως η κατωφλίωση διακύμανσης και η αναδρομική εξάλειψη χαρακτηριστικών (RFE) έχουν εφαρμοστεί σε εργασίες ανάλυσης συναισθήματος για τη μείωση του χώρου χαρακτηριστικών, διατηρώντας παράλληλα τα πιο κατατοπιστικά χαρακτηριστικά. Αυτό διασφαλίζει ότι τα μοντέλα είναι τόσο αποδοτικά όσο και αποτελεσματικά [5, 23].

2. Κλιμάκωση:

- Η επιλογή χαρακτηριστικών ενισχύει την επεκτασιμότητα των μοντέλων ανάλυσης συναισθήματος. Σε εφαρμογές μεγάλης κλίμακας, όπως η ανάλυση εκατομμυρίων tweets ή κριτικών προϊόντων, η μείωση του αριθμού των χαρακτηριστικών επιταχύνει σημαντικά τις διαδικασίες εκπαίδευσης και πρόβλεψης.

- Αυτή η επεκτασιμότητα είναι απαραίτητη για την ανάπτυξη μοντέλων ανάλυσης συναισθήματος σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους, όπως κινητές συσκευές ή σενάρια υπολογισμού άκρων [3,4,23].

Παραδείγματα εφαρμογών

Παρακολούθηση κοινωνικών μέσων ενημέρωσης:

- Στην ανάλυση συναισθήματος στα μέσα κοινωνικής δικτύωσης, η επιλογή χαρακτηριστικών βοηθά στον εντοπισμό των πιο σχετικών hashtags, αναφορών και φράσεων που υποδηλώνουν το δημόσιο συναίσθημα. Αυτό είναι ζωτικής σημασίας για τις μάρκες και τους οργανισμούς που επιθυμούν να παρακολουθούν και να ανταποκρίνονται στην κοινή γνώμη σε πραγματικό χρόνο.

Ανάλυση ανατροφοδότησης πελατών:

- Για τις επιχειρήσεις που αναλύουν τα σχόλια των πελατών, η επιλογή χαρακτηριστικών διασφαλίζει ότι καταγράφονται οι πιο σχετικές πτυχές του κλίματος των πελατών. Αυτό βοηθά στην κατανόηση των αναγκών των πελατών και στη βελτίωση των προϊόντων και των υπηρεσιών με βάση τα βασικά σημεία ανατροφοδότησης.

Ανάλυση πολιτικού κλίματος:

- Στην ανάλυση πολιτικού συναισθήματος, η επιλογή χαρακτηριστικών μπορεί να αναδείξει τα πιο σημαντικά ζητήματα και θέματα που καθορίζουν το συναίσθημα των ψηφοφόρων. Αυτό μπορεί να ενημερώσει τις στρατηγικές εκστρατείας και τις αποφάσεις δημόσιας πολιτικής, εστιάζοντας στις ανησυχίες που έχουν μεγαλύτερη σημασία για το εκλογικό σώμα.

Συμπερασματικά, η επιλογή χαρακτηριστικών είναι αναπόσπαστο στοιχείο για την επιτυχία της ανάλυσης συναισθήματος. Ενισχύει την απόδοση του μοντέλου, βελτιώνει την ερμηνευσιμότητα και διαχειρίζεται αποτελεσματικά δεδομένα υψηλής διάστασης. Με την επιλογή των πιο σχετικών χαρακτηριστικών, τα μοντέλα ανάλυσης συναισθήματος γίνονται πιο ακριβή, αποδοτικά και διορατικά, καθιστώντας τα ανεκτίμητα εργαλεία για την κατανόηση και την ανταπόκριση στο δημόσιο συναίσθημα σε διάφορους τομείς.

3. Προηγούμενες μελέτες για την ανάλυση συναισθήματος σε δεδομένα κοινωνικών μέσων ενημέρωσης

3.1 Επισκόπηση του ερευνητικού τοπίου

Η ανάλυση συναισθήματος σε δεδομένα μέσων κοινωνικής δικτύωσης έχει γίνει ένας ολοένα και πιο δημοφιλής ερευνητικός τομέας λόγω των τεράστιων ποσοτήτων περιεχομένου που δημιουργείται από χρήστες και είναι διαθέσιμο σε πλατφόρμες όπως το Twitter, το Facebook και το Reddit. Αυτές οι πλατφόρμες προσφέρουν μια σε πραγματικό χρόνο και ποικίλη πηγή δεδομένων που αντικατοπτρίζει την κοινή γνώμη για διάφορα θέματα. Οι ερευνητές έχουν διερευνήσει πολυάριθμες μεθοδολογίες και εφαρμογές για την αξιοποίηση αυτών των δεδομένων για την κατανόηση του δημόσιου συναισθήματος.

3.2 Πρώιμες εξελίξεις

Το αρχικό κύμα της έρευνας για την ανάλυση συναισθήματος επικεντρώθηκε σε προσεγγίσεις βασισμένες σε λεξικά, χρησιμοποιώντας προκαθορισμένα λεξικά συναισθήματος για την ταξινόμηση κειμένου. Οι πρώτες μελέτες χρησιμοποίησαν εργαλεία όπως το SentiWordNet και το General Inquirer για την ανάλυση του συναισθήματος σε αναρτήσεις στα μέσα κοινωνικής

δικτύωσης. Αυτές οι μέθοδοι που βασίζονται σε λεξικά ήταν σχετικά απλές στην εφαρμογή και παρείχαν τα θεμέλια για πιο εξελιγμένες τεχνικές.

Βασικές μελέτες

- SentiWordNet: Μια επέκταση του WordNet που αποδίδει βαθμολογίες συναισθήματος σε synsets (ομάδες συνωνύμων). Αυτό το λεξικό χρησιμοποιήθηκε ευρέως σε πρώιμες μελέτες ανάλυσης συναισθήματος για την ταξινόμηση κειμένων κοινωνικών μέσων ενημέρωσης [3].
- General Inquirer: Ένα άλλο πρώιμο εργαλείο ανάλυσης συναισθήματος που κατηγοριοποιεί τις λέξεις σε θετικές, αρνητικές και ουδέτερες κατηγορίες. Έχει χρησιμοποιηθεί για την ανάλυση του δημόσιου συναισθήματος για διάφορα κοινωνικά ζητήματα που συζητούνται στα μέσα κοινωνικής δικτύωσης[3,10].

Προσεγγίσεις μηχανικής μάθησης

Με την έλευση της μηχανικής μάθησης, οι ερευνητές άρχισαν να διερευνούν πιο προηγμένες τεχνικές για την ανάλυση συναισθήματος. Αυτές οι μέθοδοι περιλαμβάνουν την εκπαίδευση αλγορίθμων σε επισημασμένα σύνολα δεδομένων για την εκμάθηση προτύπων που σχετίζονται με διαφορετικά συναισθήματα. Οι συνήθεις τεχνικές μηχανικής μάθησης περιλαμβάνουν το Naive Bayes, τις μηχανές διανυσμάτων υποστήριξης (SVM) και τη λογιστική παλινδρόμηση.

Βασικές μελέτες:

- Naive Bayes και SVM: Αυτά τα μοντέλα έχουν χρησιμοποιηθεί εκτενώς για την ταξινόμηση συναισθημάτων σε tweets και αναρτήσεις στο Facebook. Για παράδειγμα, οι Pang et al. (2002) χρησιμοποίησαν τα μοντέλα Naive Bayes και SVM για την ταξινόμηση συναισθημάτων σε κριτικές ταινιών, γεγονός που έθεσε τις βάσεις για την εφαρμογή αυτών των μεθόδων σε δεδομένα κοινωνικών μέσων ενημέρωσης [3,20].
- Υβριδικά μοντέλα: Ορισμένες μελέτες συνδύασαν προσεγγίσεις βασισμένες σε λεξικό και μηχανική μάθηση για να αξιοποιήσουν τα πλεονεκτήματα και των δύο. Για παράδειγμα, οι Go κ.ά. (2009) χρησιμοποίησαν ένα υβριδικό μοντέλο που χρησιμοποίησε emoticons ως θορυβώδεις ετικέτες για την εκπαίδευση ταξινομητών συναισθήματος σε δεδομένα Twitter [12].

3.4 Εξελίξεις στη βαθιά μάθηση

Η εισαγωγή της βαθιάς μάθησης έφερε επανάσταση στην ανάλυση συναισθήματος, επιτρέποντας τη σύλληψη πιο σύνθετων μοτίβων και πληροφοριών πλαισίου. Μοντέλα όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN), τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) και τα μοντέλα που βασίζονται σε μετασχηματιστές όπως τα BERT και RoBERTa έχουν θέσει νέα σημεία αναφοράς στην ανάλυση συναισθήματος.

Βασικές μελέτες:

- LSTM και RNN: Αυτά τα μοντέλα έχουν χρησιμοποιηθεί για την ανάλυση διαδοχικών δεδομένων και τη σύλληψη μακροπρόθεσμων εξαρτήσεων σε κείμενο. Για παράδειγμα, οι Zhang

κ.ά. (2018) εφάρμοσαν δίκτυα LSTM για την ανάλυση του συναισθήματος σε δεδομένα Twitter, επιτυγχάνοντας υψηλή ακρίβεια με τη σύλληψη του πλαισίου και της διαδοχικής φύσης των tweets[9,24].

- BERT και RoBERTa: Τα μοντέλα που βασίζονται σε μετασχηματιστές ήταν ιδιαίτερα επιτυχημένα σε εργασίες ανάλυσης συναισθήματος. Μελέτες έχουν δείξει ότι αυτά τα μοντέλα, προ-εκπαιδευμένα σε μεγάλα σώματα κειμένων και τελειοποιημένα σε δεδομένα ειδικά για τα συναισθήματα, υπερτερούν σημαντικά έναντι των παραδοσιακών μεθόδων μηχανικής μάθησης. Για παράδειγμα, το RoBERTa έχει χρησιμοποιηθεί για την ταξινόμηση συναισθημάτων σε tweets που σχετίζονται με πολιτικά γεγονότα, επιδεικνύοντας ανώτερες επιδόσεις στην κατανόηση συναισθημάτων[9,24].

3.5 Εφαρμογές και επιπτώσεις

Η ανάλυση συναισθήματος σε δεδομένα των μέσων κοινωνικής δικτύωσης έχει εφαρμοστεί σε διάφορους τομείς, συμπεριλαμβανομένων των επιχειρήσεων, της πολιτικής και της δημόσιας υγείας. Οι επιχειρήσεις χρησιμοποιούν την ανάλυση συναισθήματος για την παρακολούθηση της φήμης της μάρκας και της ικανοποίησης των πελατών. Οι πολιτικοί αναλυτές την αξιοποιούν για τη μέτρηση της κοινής γνώμης σχετικά με τους υποψηφίους και τις πολιτικές. Στη δημόσια υγεία, η ανάλυση συναισθήματος βοηθά στην κατανόηση των ανησυχιών και της παραπληροφόρησης του κοινού κατά τη διάρκεια κρίσεων υγείας.

Βασικές εφαρμογές:

- Παρακολούθηση μάρκας: Οι εταιρείες αναλύουν το συναίσθημα στα μέσα κοινωνικής δικτύωσης για να κατανοήσουν την αντίληψη του κοινού για τη μάρκα και τα προϊόντα τους, επιτρέποντάς τους να ανταποκρίνονται προληπτικά στα σχόλια των πελατών[12,21].

- Πολιτική ανάλυση: Η ανάλυση συναισθήματος έχει χρησιμοποιηθεί για την παρακολούθηση του κλίματος των ψηφοφόρων και την πρόβλεψη των εκλογικών αποτελεσμάτων αναλύοντας tweets και αναρτήσεις στο Facebook που σχετίζονται με πολιτικούς υποψηφίους και θέματα [5].

- Δημόσια Υγεία: Κατά τη διάρκεια της πανδημίας COVID-19, η ανάλυση συναισθήματος των δεδομένων των μέσων κοινωνικής δικτύωσης βοήθησε τους αξιωματούχους της δημόσιας υγείας να κατανοήσουν τα συναισθήματα του κοινού απέναντι στα εμβόλια και τα μέτρα υγείας, καθοδηγώντας στρατηγικές επικοινωνίας για την αντιμετώπιση των ανησυχιών και της παραπληροφόρησης του κοινού [12].

3.6 Προκλήσεις και μελλοντικές κατευθύνσεις

Παρά τις σημαντικές προόδους, η ανάλυση συναισθήματος σε δεδομένα των μέσων κοινωνικής δικτύωσης αντιμετωπίζει αρκετές προκλήσεις. Αυτές περιλαμβάνουν τον χειρισμό του σαρκασμού και της ειρωνείας, την κατανόηση του πλαισίου, τη διαχείριση πολύγλωσσων

δεδομένων και την αντιμετώπιση της ανισορροπίας των δεδομένων. Η μελλοντική έρευνα είναι πιθανό να επικεντρωθεί στη βελτίωση της ακρίβειας των μοντέλων ανάλυσης συναισθήματος με την αντιμετώπιση αυτών των προκλήσεων και την ενσωμάτωση πιο εξελιγμένων τεχνικών, όπως η μάθηση μεταφοράς και η πολυτροπική ανάλυση.

Προκλήσεις:

- Σαρकाσμός και ειρωνεία: Η ανίχνευση σαρκαστικών σχολίων παραμένει μια σημαντική πρόκληση, καθώς το επιδιωκόμενο συναίσθημα συχνά έρχεται σε αντίθεση με την κυριολεκτική σημασία των λέξεων.
- Κατανόηση του πλαισίου: Τα μοντέλα πρέπει να κατανοήσουν καλύτερα το πλαίσιο στο οποίο εκφράζονται τα συναισθήματα για να τα ταξινομήσουν με ακρίβεια [3,5,23].
- Πολυγλωσσική ανάλυση συναισθημάτων: Η ανάλυση συναισθημάτων σε διαφορετικές γλώσσες εισάγει πρόσθετη πολυπλοκότητα λόγω γλωσσικών και πολιτισμικών διαφορών [12,21].

Συμπερασματικά, η ανάλυση συναισθήματος σε δεδομένα κοινωνικών μέσων έχει εξελιχθεί από απλές προσεγγίσεις που βασίζονται σε λεξικό σε προηγμένα μοντέλα βαθιάς μάθησης, παρέχοντας πολύτιμες πληροφορίες σε διάφορους τομείς. Αν και οι προκλήσεις παραμένουν, η συνεχιζόμενη έρευνα και οι τεχνολογικές εξελίξεις υπόσχονται να βελτιώσουν την ακρίβεια και την εφαρμοσιμότητα της ανάλυσης συναισθήματος, καθιστώντας την απαραίτητο εργαλείο για την κατανόηση της κοινής γνώμης στην ψηφιακή εποχή.

3.7 Μεθοδολογικές προσεγγίσεις

Η ανάλυση συναισθήματος χρησιμοποιεί διάφορες μεθοδολογικές προσεγγίσεις για την ταξινόμηση του συναισθήματος που εκφράζεται σε δεδομένα κειμένου. Αυτές οι μεθοδολογίες μπορούν να κατηγοριοποιηθούν σε γενικές γραμμές σε προσεγγίσεις που βασίζονται σε λεξικό, προσεγγίσεις μηχανικής μάθησης και προσεγγίσεις βαθιάς μάθησης. Κάθε προσέγγιση έχει τις δικές της τεχνικές, πλεονεκτήματα και προκλήσεις.

3.7.1 Προσεγγίσεις με βάση το λεξικό

Οι προσεγγίσεις που βασίζονται σε λεξικά βασίζονται σε προκαθορισμένους καταλόγους λέξεων, γνωστούς ως λεξικά συναισθήματος, οι οποίοι σχολιάζονται με τις αντίστοιχες τιμές συναισθήματος. Αυτές οι μέθοδοι ταξινομούν το συναίσθημα υπολογίζοντας την πολικότητα του συναισθήματος ενός δεδομένου κειμένου με βάση την παρουσία και την πολικότητα των λέξεων στο λεξικό.

Τεχνικές:

- Λεξικά συναισθήματος: Αυτά περιλαμβάνουν λεξικά γενικής χρήσης όπως το SentiWordNet και λεξικά ειδικού τομέα προσαρμοσμένα για συγκεκριμένα πλαίσια.

- Συστήματα βασισμένα σε κανόνες: Συνδυασμός λεξικών με γλωσσικούς κανόνες για τον χειρισμό των αρνήσεων, των τροποποιητών βαθμού και άλλων γραμματικών δομών. Για παράδειγμα, το μοντέλο VADER (Valence Aware Dictionary and sEntiment Reasoner) χρησιμοποιεί ένα συνδυασμό λεξικού συναισθήματος και κανόνων για να αποδώσει βαθμολογία συναισθήματος σε κείμενο, ιδιαίτερα αποτελεσματικό για δεδομένα κοινωνικών μέσων ενημέρωσης. [7]

Πλεονεκτήματα:

- Απλό και εύκολο στην εφαρμογή.
- Δεν απαιτεί επισημειωμένα δεδομένα εκπαίδευσης.
- Ερμηνεύσιμα αποτελέσματα.

Προκλήσεις:

- Περιορίζεται από την ποιότητα και την κάλυψη του λεξικού.
- Δυσκολεύεται με τα συμφραζόμενα, τον σαρκασμό και την εξελισσόμενη χρήση της γλώσσας.

3.7.2 Προσεγγίσεις μηχανικής μάθησης

Οι προσεγγίσεις μηχανικής μάθησης περιλαμβάνουν την εκπαίδευση αλγορίθμων σε επισημασμένα σύνολα δεδομένων για την εκμάθηση προτύπων που σχετίζονται με διαφορετικά συναισθήματα. Αυτές οι μέθοδοι χρησιμοποιούν διάφορα χαρακτηριστικά που εξάγονται από το κείμενο για τη δημιουργία προγνωστικών μοντέλων.

Τεχνικές:

- Naive Bayes: Ένας ταξινομητής που εφαρμόζει το θεώρημα του Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Είναι απλός και αποτελεσματικός, χρησιμοποιείται συχνά ως βασική γραμμή για εργασίες ανάλυσης συναισθήματος.
- Support Vector Machines (SVM): Ένας διαχωριστικός ταξινομητής που βρίσκει το υπερεπίπεδο που διαχωρίζει καλύτερα τα δεδομένα σε διαφορετικές κλάσεις συναισθήματος. Αποτελεσματική για δεδομένα κειμένου υψηλής διάστασης.
- Λογιστική παλινδρόμηση: Ένα γραμμικό μοντέλο που εκτιμά την πιθανότητα ενός δυαδικού αποτελέσματος με βάση τα χαρακτηριστικά εισόδου. Χρησιμοποιείται συνήθως για δυαδική ταξινόμηση συναισθήματος (θετικό έναντι αρνητικού) και μπορεί να επεκταθεί σε ταξινόμηση πολλαπλών κατηγοριών.

Πλεονεκτήματα:

- Δυνατότητα χειρισμού μεγάλων συνόλων δεδομένων.
- Μπορεί να συλλάβει πολύπλοκα μοτίβα στα δεδομένα.

Προκλήσεις:

- Απαιτεί δεδομένα εκπαίδευσης με ετικέτες.
- Μπορεί να κάνει hyperadapt εάν δεν έχει ρυθμιστεί σωστά.

3.7.3 Προσεγγίσεις βαθιάς μάθησης

Οι προσεγγίσεις βαθιάς μάθησης αξιοποιούν τα νευρωνικά δίκτυα για τη σύλληψη πιο σύνθετων μοτίβων και πληροφοριών πλαισίου στο κείμενο. Αυτά τα μοντέλα έχουν φέρει επανάσταση στην ανάλυση συναισθήματος βελτιώνοντας σημαντικά την ακρίβεια και την ανθεκτικότητα.

Τεχνικές:

- Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN): Σχεδιασμένα για να χειρίζονται διαδοχικά δεδομένα διατηρώντας μια κρυφή κατάσταση που συλλαμβάνει πληροφορίες από προηγούμενα βήματα της ακολουθίας. Χρήσιμα για την ανάλυση προτάσεων και τη σύλληψη μακροχρόνιων εξαρτήσεων.
- Δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM): Ένας τύπος RNN που αντιμετωπίζει το πρόβλημα της εξαφανιζόμενης κλίσης, επιτρέποντας στο μοντέλο να συλλάβει μακροπρόθεσμες εξαρτήσεις. Χρησιμοποιείται ευρέως στην ανάλυση συναισθήματος για την ικανότητά του να θυμάται μακροπρόθεσμα το πλαίσιο[9,24].
- Μοντέλα βασισμένα σε μετασχηματιστές: Το BERT (Bidirectional Encoder Representations from Transformers) και το RoBERTa (A Robustly Optimized BERT Pretraining Approach) έχουν θέσει νέα σημεία αναφοράς στην ανάλυση συναισθήματος. Αυτά τα μοντέλα συλλαμβάνουν αμφίδρομο περιεχόμενο και βαθιά γλωσσικά μοτίβα, καθιστώντας τα ιδιαίτερα αποτελεσματικά για την κατανόηση αποχρώσεων των συναισθημάτων.

Πλεονεκτήματα:

- Υψηλή ακρίβεια και ικανότητα χειρισμού σύνθετων προτύπων.
- Ικανότητα σύλληψης πλαισίου και μακροπρόθεσμων εξαρτήσεων.

Προκλήσεις:

- Υπολογιστικά εντατικές και απαιτούν σημαντικούς πόρους.
- Χρειάζεται μεγάλα σύνολα δεδομένων με ετικέτες για εκπαίδευση.

4.7.4 Υβριδικές προσεγγίσεις

Οι υβριδικές προσεγγίσεις συνδυάζουν στοιχεία τόσο των μεθόδων που βασίζονται στο λεξικό όσο και των μεθόδων μηχανικής μάθησης για να αξιοποιήσουν τα πλεονεκτήματα της καθεμιάς.

Τεχνικές:

- Συνδυασμός λεξικών με μηχανική μάθηση: Χρήση λεξικών για τη δημιουργία αρχικών συνόλων χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης για τη βελτίωση και τη βελτίωση της ταξινόμησης συναισθήματος. Για παράδειγμα, χρήση emoticons ως θορυβώδεις ετικέτες για την εκπαίδευση ταξινομητών συναισθήματος σε δεδομένα του Twitter[2,12,21].

- Feature Engineering: Συνδυασμός χειροποίητων χαρακτηριστικών που βασίζονται σε λεξικά με αυτόματα εξαγόμενα χαρακτηριστικά από μοντέλα μηχανικής μάθησης.

Πλεονεκτήματα:

- Αξιοποιεί την απλότητα των λεξικών και την ευρωστία των μοντέλων μηχανικής μάθησης.

- Μπορεί να χειριστεί ποικίλες πηγές δεδομένων και περιβάλλοντα.

Προκλήσεις:

- Πολυπλοκότητα στην ενσωμάτωση διαφορετικών μεθόδων.

- Απαιτεί προσεκτική ρύθμιση για την αποφυγή υπερβολικής προσαρμογής και τη διασφάλιση της ερμηνευσιμότητας.

Συμπερασματικά, οι μεθοδολογίες ανάλυσης συναισθήματος έχουν εξελιχθεί από απλές προσεγγίσεις που βασίζονται σε λεξικά σε εξελιγμένα μοντέλα βαθιάς μάθησης. Κάθε προσέγγιση προσφέρει μοναδικά πλεονεκτήματα και προκλήσεις και η επιλογή της μεθοδολογίας εξαρτάται από τις ειδικές απαιτήσεις της εργασίας, συμπεριλαμβανομένης της φύσης των δεδομένων, των διαθέσιμων πόρων και της επιθυμητής ακρίβειας. Αξιοποιώντας τα πλεονεκτήματα των διαφόρων μεθόδων, οι ερευνητές και οι επαγγελματίες μπορούν να δημιουργήσουν ισχυρά και αποτελεσματικά μοντέλα ανάλυσης συναισθήματος ικανά να παρέχουν πολύτιμες πληροφορίες για την κοινή γνώμη και το συναίσθημα.

3.8 Βασικά ευρήματα και γνώσεις

Η εφαρμογή της ανάλυσης συναισθήματος σε δεδομένα των μέσων κοινωνικής δικτύωσης έχει αποφέρει πολλά βασικά ευρήματα και ιδέες, ενισχύοντας σημαντικά την κατανόηση της κοινής γνώμης και της συμπεριφοράς σε διάφορους τομείς.

3.8.1 Τάσεις του δημόσιου κλίματος

1. Συζήτηση για την κλιματική αλλαγή:

- Η ανάλυση των tweets που σχετίζονται με την κλιματική αλλαγή αποκάλυψε την επικράτηση του αρνητικού συναισθήματος, αντανakλώντας την ευρέως διαδεδομένη ανησυχία και το άγχος για το θέμα. Μελέτες έχουν διαπιστώσει ότι γεγονότα όπως οι φυσικές καταστροφές και οι ανακοινώσεις για την κλιματική πολιτική προκαλούν αιχμές στη δραστηριότητα των μέσων κοινωνικής δικτύωσης, οι οποίες συχνά συνοδεύονται από έντονες συναισθηματικές αντιδράσεις [2,12].

- Η γεωγραφική ανάλυση του συναισθήματος έχει δείξει ότι οι περιοχές που πλήττονται άμεσα από φαινόμενα κλιματικής αλλαγής, όπως ακραία καιρικά φαινόμενα, τείνουν να παρουσιάζουν υψηλότερα επίπεδα αρνητικού συναισθήματος σε σύγκριση με άλλες περιοχές [5,23].

2. Αντίληψη για το εμπορικό σήμα και το προϊόν:

- Η ανάλυση συναισθήματος των αναφορών στα μέσα κοινωνικής δικτύωσης και των κριτικών προϊόντων έχει παράσχει στις εταιρείες πολύτιμες πληροφορίες σχετικά με την ικανοποίηση των πελατών και την αντίληψη της μάρκας. Το θετικό συναίσθημα σχετίζεται με αυξημένη αφοσίωση των πελατών και υψηλότερες πωλήσεις, ενώ το αρνητικό συναίσθημα αναδεικνύει τομείς που χρήζουν βελτίωσης.

- Η παρακολούθηση του συναισθήματος σε πραγματικό χρόνο επιτρέπει στις επιχειρήσεις να αντιμετωπίζουν γρήγορα τα παράπονα των πελατών και να διαχειρίζονται πιο αποτελεσματικά τη φήμη της μάρκας [3,17].

3. Πολιτικό συναίσθημα:

- Κατά τη διάρκεια προεκλογικών εκστρατειών, η ανάλυση συναισθήματος των tweets και των αναρτήσεων έχει χρησιμοποιηθεί για να μετρήσει το συναίσθημα των ψηφοφόρων απέναντι στους υποψηφίους και τα βασικά ζητήματα. Το θετικό συναίσθημα προς έναν υποψήφιο συχνά σχετίζεται με αυξημένη δέσμευση και υποστήριξη των ψηφοφόρων.

- Μελέτες έχουν δείξει ότι το συναίσθημα των μέσων κοινωνικής δικτύωσης μπορεί μερικές φορές να προβλέψει τα εκλογικά αποτελέσματα, αν και αυτό επηρεάζεται από διάφορους παράγοντες, συμπεριλαμβανομένης της δημογραφικής σύνθεσης των χρηστών των μέσων κοινωνικής δικτύωσης και της αντιπροσωπευτικότητας των δεδομένων[2,5].

3.8.3 Μεθοδολογικές γνώσεις

1. Αποτελεσματικότητα των μοντέλων βαθιάς μάθησης:

- Μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT και το RoBERTa, έχουν επιδείξει ανώτερες επιδόσεις σε εργασίες ανάλυσης συναισθήματος σε σύγκριση με τις παραδοσιακές μεθόδους μηχανικής μάθησης. Αυτά τα μοντέλα υπερέχουν στην καταγραφή των αποχρώσεων του πλαισίου και των μακροπρόθεσμων εξαρτήσεων στο κείμενο, με αποτέλεσμα υψηλότερη ακρίβεια και ανθεκτικότητα[9,24].

- Τα υβριδικά μοντέλα που συνδυάζουν προσεγγίσεις βασισμένες σε λεξικό και μηχανική μάθηση έχουν επίσης αποδειχθεί αποτελεσματικά, ιδίως στον χειρισμό ποικίλων και θορυβωδών δεδομένων κοινωνικών μέσων. Αυτά τα μοντέλα αξιοποιούν τα πλεονεκτήματα και των δύο προσεγγίσεων, βελτιώνοντας την ακρίβεια ταξινόμησης συναισθημάτων και την ερμηνευσιμότητα [21].

2. Τεχνικές επιλογής χαρακτηριστικών:

- Η αποτελεσματική επιλογή χαρακτηριστικών ενισχύει σημαντικά την απόδοση του μοντέλου μειώνοντας το hyper-adaptation και βελτιώνοντας τη γενίκευση. Τεχνικές όπως η αμοιβαία πληροφορία, τα τετράγωνα χ^2 και η αναδρομική εξάλειψη χαρακτηριστικών (RFE) έχουν εφαρμοστεί με επιτυχία στην ανάλυση συναισθήματος για τον εντοπισμό των πιο κατατοπιστικών χαρακτηριστικών[8,9,18,24].

- Μελέτες έχουν τονίσει τη σημασία της επιλογής χαρακτηριστικών ανά τομέα, καθώς οι λέξεις και φράσεις που περιέχουν συναισθήματα μπορεί να διαφέρουν σημαντικά σε διαφορετικά πλαίσια (π.χ. πολιτικός λόγος έναντι κριτικών προϊόντων) [6,14,18].

3.8.4 Προκλήσεις και ευκαιρίες

1. Χειρισμός του σαρκασμού και του πλαισίου:

- Ο σαρκασμός και τα εξαρτώμενα από το πλαίσιο νοήματα παραμένουν πρόκληση για τα μοντέλα ανάλυσης συναισθήματος. Ενώ τα μοντέλα βαθιάς μάθησης έχουν σημειώσει πρόοδο, απαιτούνται περαιτέρω βελτιώσεις για την ακριβή ανίχνευση και ερμηνεία σαρκαστικών παρατηρήσεων και διαφοροποιημένων συναισθημάτων[10,16].

2. Πολυγλωσσική ανάλυση συναισθήματος:

- Η ανάλυση συναισθήματος σε διάφορες γλώσσες εισάγει πρόσθετη πολυπλοκότητα λόγω γλωσσικών και πολιτισμικών διαφορών. Η ανάπτυξη εύρωστων πολύγλωσσων μοντέλων ανάλυσης συναισθήματος απαιτεί εκτεταμένα επισημασμένα σύνολα δεδομένων και εξελιγμένες τεχνικές preprocessing [5].

3. Ανάλυση σε πραγματικό χρόνο:

- Η ανάγκη για ανάλυση συναισθήματος σε πραγματικό χρόνο σε εφαρμογές όπως η διαχείριση κρίσεων και η παρακολούθηση εμπορικών σημάτων αναδεικνύει τη σημασία της ανάπτυξης αποτελεσματικών αλγορίθμων που μπορούν να επεξεργάζονται και να αναλύουν μεγάλους όγκους δεδομένων γρήγορα και με ακρίβεια.

Συνοπτικά, η ανάλυση συναισθήματος σε δεδομένα των μέσων κοινωνικής δικτύωσης έχει παράσχει πολύτιμες πληροφορίες για το κοινό συναίσθημα σε διάφορους τομείς, από την κλιματική αλλαγή έως την αντίληψη των εμπορικών σημάτων και την πολιτική γνώμη. Οι εξελίξεις στα μοντέλα βαθιάς μάθησης και στις αποτελεσματικές τεχνικές επιλογής χαρακτηριστικών έχουν βελτιώσει σημαντικά την ακρίβεια και την ευρωστία της ανάλυσης συναισθήματος, αν και οι προκλήσεις όπως ο χειρισμός του σαρκασμού και των πολύγλωσσων δεδομένων παραμένουν τομείς για συνεχή έρευνα.

4. Σημασία της μελέτης του λόγου για την κλιματική αλλαγή

4.1 Εισαγωγή

Η κλιματική αλλαγή είναι ένα επείγον παγκόσμιο ζήτημα με εκτεταμένες επιπτώσεις στο περιβάλλον, την οικονομία και την κοινωνία. Η κατανόηση του δημόσιου λόγου για την κλιματική αλλαγή είναι κρίσιμη για την ανάπτυξη αποτελεσματικών στρατηγικών επικοινωνίας, πολιτικών και παρεμβάσεων. Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης, όπως το Twitter και το Facebook, έχουν γίνει ζωτικής σημασίας χώροι για δημόσιες συζητήσεις, καθιστώντας τις πλούσιες πηγές δεδομένων για την ανάλυση του δημόσιου συναισθήματος και της κοινής γνώμης σχετικά με την κλιματική αλλαγή. Η μελέτη των συζητήσεων για την κλιματική αλλαγή σε αυτές τις πλατφόρμες παρέχει πολύτιμες πληροφορίες για την ευαισθητοποίηση του κοινού, τις στάσεις και τις προθέσεις συμπεριφοράς, οι οποίες είναι ζωτικής σημασίας για την αντιμετώπιση των προκλήσεων που θέτει η κλιματική αλλαγή.

4.2 Ευαισθητοποίηση και κατανόηση

1. Ενημέρωση του κοινού:

- Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης διαδραματίζουν σημαντικό ρόλο στην ευαισθητοποίηση του κοινού σχετικά με την κλιματική αλλαγή. Μελετώντας το διάλογο σε αυτές τις πλατφόρμες, οι ερευνητές μπορούν να μετρήσουν το επίπεδο ευαισθητοποίησης του κοινού και να εντοπίσουν τα κενά γνώσεων. Οι πληροφορίες αυτές είναι απαραίτητες για τον σχεδιασμό εκπαιδευτικών εκστρατειών που επικοινωνούν αποτελεσματικά την επιστήμη και τον επείγοντα χαρακτήρα της κλιματικής αλλαγής [2,21,23].

- Η παρακολούθηση των συζητήσεων στα μέσα κοινωνικής δικτύωσης μπορεί να βοηθήσει στον εντοπισμό των φωνών και των πηγών πληροφόρησης με τη μεγαλύτερη επιρροή. Η κατανόηση του ποιος καθοδηγεί τη συζήτηση μπορεί να ενημερώσει τις στρατηγικές για την ενίσχυση αξιόπιστων πληροφοριών και την αντιμετώπιση της παραπληροφόρησης.

2. Κατανόηση των στάσεων και των αντιλήψεων:

- Η ανάλυση των συζητήσεων για την κλιματική αλλαγή βοηθά στην κατανόηση των στάσεων και των αντιλήψεων του κοινού απέναντι στην κλιματική αλλαγή. Η ανάλυση συναισθημάτων μπορεί να αποκαλύψει πώς αισθάνονται οι άνθρωποι για την κλιματική αλλαγή, συμπεριλαμβανομένων των ανησυχιών, των ελπίδων και του σκεπτικισμού τους. Αυτή η κατανόηση είναι ζωτικής σημασίας για την προσαρμογή των μηνυμάτων που βρίσκουν απήχηση σε διαφορετικά τμήματα του πληθυσμού [2,21].

- Οι τάσεις του συναισθήματος μπορούν να υποδεικνύουν μεταβολές στην κοινή γνώμη με την πάροδο του χρόνου, επηρεαζόμενες από γεγονότα όπως ακραία καιρικά φαινόμενα, αλλαγές

πολιτικής και κάλυψη από τα μέσα ενημέρωσης. Αυτές οι τάσεις παρέχουν πληροφορίες για το πώς εξελίσσεται το δημόσιο συναίσθημα και ποιοι παράγοντες οδηγούν σε αυτές τις αλλαγές.

4.3 Ενημέρωση για πολιτικές και στρατηγικές επικοινωνίας

1. Ανάπτυξη πολιτικής:

- Οι υπεύθυνοι χάραξης πολιτικής μπορούν να χρησιμοποιήσουν τις γνώσεις από το διάλογο για την κλιματική αλλαγή για να αναπτύξουν πολιτικές που αντικατοπτρίζουν τις ανησυχίες και τις προτεραιότητες του κοινού. Η κατανόηση των πτυχών της κλιματικής αλλαγής για τις οποίες οι άνθρωποι ενδιαφέρονται περισσότερο μπορεί να βοηθήσει στην ιεράρχηση των πρωτοβουλιών πολιτικής και στην απόκτηση δημόσιας υποστήριξης [5, 22].

- Τα δεδομένα για το δημόσιο συναίσθημα μπορούν επίσης να αναδείξουν τομείς στους οποίους απαιτούνται παρεμβάσεις πολιτικής. Για παράδειγμα, εάν ένας σημαντικός αριθμός χρηστών των μέσων κοινωνικής δικτύωσης εκφράζει απογοήτευση για την ανεπαρκή δράση για το κλίμα, αυτό μπορεί να ωθήσει τους υπεύθυνους χάραξης πολιτικής να λάβουν πιο αποφασιστικά μέτρα.

2. Αποτελεσματική επικοινωνία:

- Η αποτελεσματική επικοινωνία είναι ζωτικής σημασίας για την κινητοποίηση της δημόσιας δράσης για την κλιματική αλλαγή. Με την ανάλυση του λόγου, οι επικοινωνιολόγοι μπορούν να προσδιορίσουν αποτελεσματικές στρατηγικές ανταλλαγής μηνυμάτων που εμπλέκουν και παρακινούν τους ανθρώπους. Αυτό περιλαμβάνει την κατανόηση της γλώσσας και της διαμόρφωσης που έχει απήχηση στα διάφορα ακροατήρια [12,21].

- Η ανάλυση του λόγου στα μέσα κοινωνικής δικτύωσης μπορεί επίσης να βοηθήσει στον εντοπισμό και την αντιμετώπιση της παραπληροφόρησης. Παρακολουθώντας τη διάδοση των ψευδών πληροφοριών και κατανοώντας τον αντίκτυπό τους, οι επικοινωνιολόγοι μπορούν να αναπτύξουν στοχευμένες στρατηγικές για την αντιμετώπισή τους και την προώθηση της ακριβούς πληροφόρησης.

4.4 Αλλαγή συμπεριφοράς

1. Προθέσεις συμπεριφοράς:

- Η μελέτη του λόγου για την κλιματική αλλαγή παρέχει πληροφορίες για τις προθέσεις του κοινού σχετικά με συμπεριφορές που σχετίζονται με το κλίμα, όπως η μείωση του αποτυπώματος άνθρακα, η υποστήριξη των ανανεώσιμων πηγών ενέργειας και η συμμετοχή στον ακτιβισμό για το κλίμα. Η κατανόηση αυτών των προθέσεων είναι ζωτικής σημασίας για τον σχεδιασμό παρεμβάσεων που ενθαρρύνουν βιώσιμες συμπεριφορές [2,12,22].

- Η ανάλυση των μέσων κοινωνικής δικτύωσης μπορεί να αποκαλύψει εμπόδια στην αλλαγή συμπεριφοράς, όπως η παραπληροφόρηση, ο σκεπτικισμός και η αντιληπτή έλλειψη

αποτελεσματικότητας. Η αντιμετώπιση αυτών των εμποδίων μέσω στοχευμένων μέτρων επικοινωνίας και πολιτικής μπορεί να ενισχύσει τη συμμετοχή του κοινού στη δράση για το κλίμα.

2. Εμπλοκή και κινητοποίηση:

- Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης είναι ισχυρά εργαλεία για την κινητοποίηση της δημόσιας δράσης. Η ανάλυση του λόγου για την κλιματική αλλαγή μπορεί να βοηθήσει στον εντοπισμό ευκαιριών για δέσμευση και κινητοποίηση. Για παράδειγμα, η κατανόηση του πότε και γιατί οι άνθρωποι συζητούν για την κλιματική αλλαγή μπορεί να ενημερώσει για το χρονοδιάγραμμα και το περιεχόμενο των εκστρατειών ώστε να μεγιστοποιηθεί ο αντίκτυπος [5,23].

- Η εμπλοκή με τους χρήστες των μέσων κοινωνικής δικτύωσης μέσω συζητήσεων, δημοσκοπήσεων και διαδραστικού περιεχομένου μπορεί να ενισχύσει την αίσθηση της κοινότητας και της συλλογικής δράσης. Αυτή η δέσμευση είναι ζωτικής σημασίας για τη δημιουργία ενός βιώσιμου κινήματος προς την κατεύθυνση της αντιμετώπισης της κλιματικής αλλαγής.

Η μελέτη των συζητήσεων για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης είναι απαραίτητη για την κατανόηση της ευαισθητοποίησης, των στάσεων και των συμπεριφορών του κοινού σχετικά με την κλιματική αλλαγή. Οι γνώσεις που αποκτώνται από την ανάλυση αυτή μπορούν να ενημερώσουν για την ανάπτυξη πολιτικής, να βελτιώσουν τις στρατηγικές επικοινωνίας και να προωθήσουν την αλλαγή συμπεριφοράς. Αξιοποιώντας τη δύναμη των δεδομένων των κοινωνικών μέσων ενημέρωσης, οι ενδιαφερόμενοι μπορούν να αναπτύξουν πιο αποτελεσματικές και στοχευμένες παρεμβάσεις για την αντιμετώπιση της επείγουσας πρόκλησης της κλιματικής αλλαγής. Καθώς η συζήτηση γύρω από την κλιματική αλλαγή συνεχίζει να εξελίσσεται, η συνεχής ανάλυση θα παραμείνει ζωτικής σημασίας για την προσαρμογή των στρατηγικών και τη διατήρηση της δυναμικής στον αγώνα κατά της υπερθέρμανσης του πλανήτη.

4.5 Υπάρχουσα έρευνα

Η μελέτη της συζήτησης για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης έχει κερδίσει σημαντική προσοχή τα τελευταία χρόνια λόγω του αυξανόμενου όγκου του περιεχομένου που δημιουργείται από τους χρήστες και της δυνατότητάς του να επηρεάσει την κοινή γνώμη και την πολιτική. Οι ερευνητές έχουν χρησιμοποιήσει διάφορες μεθοδολογίες για την ανάλυση αυτού του λόγου, συμπεριλαμβανομένης της ανάλυσης συναισθήματος, της μοντελοποίησης θεμάτων και της ανάλυσης δικτύων, για να αποκτήσουν γνώσεις σχετικά με τις αντιλήψεις και τις στάσεις του κοινού απέναντι στην κλιματική αλλαγή.

4.5.1 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος έχει χρησιμοποιηθεί ευρέως για τη μέτρηση του κοινού συναισθήματος απέναντι στην κλιματική αλλαγή σε πλατφόρμες κοινωνικής δικτύωσης όπως το Twitter και το Facebook. Αναλύοντας το συναίσθημα που εκφράζεται σε αναρτήσεις και tweets, οι ερευνητές μπορούν να κατανοήσουν τον συναισθηματικό τόνο του λόγου και να προσδιορίσουν τις επικρατούσες στάσεις.

1. Τάσεις του δημόσιου συναισθήματος:

- Μελέτες έχουν διαπιστώσει σταθερά ότι ο διάλογος για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης είναι κυρίως αρνητικός, αντανakλώντας την ανησυχία και το άγχος του κοινού για το θέμα. Για παράδειγμα, οι Cody κ.ά. (2015) διεξήγαγαν ανάλυση συναισθήματος των tweets που σχετίζονται με την κλιματική αλλαγή και διαπίστωσαν ότι το αρνητικό συναίσθημα αυξήθηκε σημαντικά κατά τη διάρκεια ακραίων καιρικών φαινομένων και σημαντικών ειδήσεων που σχετίζονται με το κλίμα [5,23].

- Η γεωγραφική ανάλυση του συναισθήματος αποκάλυψε περιφερειακές διαφοροποιήσεις στη στάση του κοινού. Οι περιοχές που πλήττονται άμεσα από φαινόμενα κλιματικής αλλαγής, όπως οι τυφώνες ή οι πυρκαγιές, τείνουν να παρουσιάζουν υψηλότερα επίπεδα αρνητικού συναισθήματος σε σύγκριση με άλλες περιοχές [2,12].

2. Διαχρονική ανάλυση:

- Η διαχρονική ανάλυση συναισθήματος έχει δείξει ότι το δημόσιο συναίσθημα απέναντι στην κλιματική αλλαγή αυξομειώνεται με την πάροδο του χρόνου, συχνά ως απάντηση σε συγκεκριμένα γεγονότα. Οι Kirilenko κ.ά. (2015) κατέδειξαν ότι το συναίσθημα κορυφώνεται κατά τη διάρκεια σημαντικών περιβαλλοντικών γεγονότων, όπως η δημοσίευση ανησυχητικών εκθέσεων για το κλίμα ή σημαντικές ανακοινώσεις πολιτικής. Αυτές οι κορυφώσεις παρέχουν πληροφορίες σχετικά με τον τρόπο με τον οποίο κινητοποιούνται η προσοχή και η ανησυχία του κοινού κατά τη διάρκεια κρίσιμων στιγμών [23].

4.5.2 Μοντελοποίηση θεμάτων

Τεχνικές μοντελοποίησης θεμάτων, όπως η Latent Dirichlet Allocation (LDA), έχουν χρησιμοποιηθεί για τον εντοπισμό των κύριων θεμάτων και θεμάτων στο πλαίσιο της συζήτησης για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης. Αυτές οι μέθοδοι βοηθούν στην αποκάλυψη της δομής των συζητήσεων και στην ανάδειξη των βασικών θεμάτων που συζητούνται από το κοινό.

1. Κύρια θέματα και ζητήματα:

- Η έρευνα εντόπισε διάφορα επαναλαμβανόμενα θέματα στο διάλογο για την κλιματική αλλαγή, συμπεριλαμβανομένων των επιπτώσεων της κλιματικής αλλαγής, των στρατηγικών μετριασμού και προσαρμογής, των πολιτικών συζητήσεων και των προσωπικών εμπειριών με ακραία καιρικά φαινόμενα. Για παράδειγμα, οι Jang και Hart (2015) χρησιμοποίησαν LDA για

να αναλύσουν tweets σχετικά με την κλιματική αλλαγή και διαπίστωσαν ότι θέματα όπως "ανανεώσιμες πηγές ενέργειας", "πολιτική και ρύθμιση" και "κλιματική επιστήμη" συζητούνταν συχνά[2,5,23].

- Η κατανόηση αυτών των θεμάτων επιτρέπει στους υπεύθυνους χάραξης πολιτικής και στους επικοινωνιολόγους να αντιμετωπίσουν τις πιο πειστικές ανησυχίες του κοινού και να προσαρμόσουν τα μηνύματά τους ώστε να έχουν απήχηση σε διαφορετικά ακροατήρια.

2. Δημόσια Δέσμευση:

- Μελέτες έχουν δείξει ότι οι χρήστες των μέσων κοινωνικής δικτύωσης ασχολούνται με θέματα κλιματικής αλλαγής με διάφορους τρόπους, από την ανταλλαγή επιστημονικών πληροφοριών έως την έκφραση προσωπικών απόψεων και την κινητοποίηση για δράση για το κλίμα. Αναλύοντας τα θέματα και τα υποθέματα εντός του λόγου, οι ερευνητές μπορούν να αποκτήσουν πληροφορίες σχετικά με τους παράγοντες που οδηγούν στη δημόσια δέσμευση και τους τύπους περιεχομένου που έχουν μεγαλύτερη απήχηση στους χρήστες [2,12].

4.5.3 Ανάλυση δικτύου

Τεχνικές ανάλυσης δικτύων έχουν χρησιμοποιηθεί για τη μελέτη της δομής και της δυναμικής των δικτύων λόγου για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης. Οι αναλύσεις αυτές αποκαλύπτουν πώς διαδίδονται οι πληροφορίες και προσδιορίζουν τους βασικούς παράγοντες επιρροής μέσα στη συζήτηση.

1. Διάδοση πληροφοριών:

- Η έρευνα έχει δείξει ότι οι πληροφορίες για την κλιματική αλλαγή διαδίδονται μέσω πολύπλοκων δικτύων χρηστών, με ορισμένα άτομα και οργανισμούς να λειτουργούν ως κόμβοι ή βασικοί κόμβοι. Για παράδειγμα, οι Williams κ.ά. (2015) χρησιμοποίησαν ανάλυση δικτύου για να μελετήσουν τη διάδοση πληροφοριών για την κλιματική αλλαγή στο Twitter, εντοπίζοντας βασικούς παράγοντες επιρροής, όπως περιβαλλοντικές ΜΚΟ, επιστήμονες και δημοσιογράφους [5,21].

- Η κατανόηση αυτών των δικτύων βοηθά στον σχεδιασμό αποτελεσματικότερων στρατηγικών επικοινωνίας που αξιοποιούν τους χρήστες με επιρροή για να ενισχύσουν τα μηνύματα και να προσεγγίσουν ευρύτερα ακροατήρια.

2. Προσδιορισμός επιρροής:

- Ο εντοπισμός των βασικών παραγόντων επιρροής στο δίκτυο λόγου για την κλιματική αλλαγή είναι ζωτικής σημασίας για στοχευμένες επικοινωνιακές προσπάθειες. Οι φορείς επιρροής μπορούν να βοηθήσουν στη διάδοση ακριβών πληροφοριών, στην αντιμετώπιση της παραπληροφόρησης και στην κινητοποίηση της δημόσιας δράσης. Η ανάλυση του δικτύου βοηθά στον εντοπισμό αυτών των ατόμων και οργανισμών, επιτρέποντας μια πιο στρατηγική δέσμευση [2,5,21].

4.6 Προκλήσεις και μελλοντικές κατευθύνσεις

Παρά τις προόδους στη μελέτη των συζητήσεων για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης, παραμένουν αρκετές προκλήσεις. Σε αυτές περιλαμβάνονται η δυσκολία ανίχνευσης του σαρκασμού και της ειρωνείας, η ανάγκη για πιο εξελιγμένες μεθόδους για τον χειρισμό πολύγλωσσων δεδομένων και οι πιθανές προκαταλήψεις που εισάγονται από τα δημογραφικά χαρακτηριστικά των χρηστών της συγκεκριμένης πλατφόρμας.

Η μελλοντική έρευνα είναι πιθανό να επικεντρωθεί στην ανάπτυξη πιο ισχυρών μοντέλων που μπορούν να καταγράψουν με ακρίβεια τις αποχρώσεις της γλώσσας των μέσων κοινωνικής δικτύωσης, συμπεριλαμβανομένου του σαρκασμού και του πλαισίου. Επιπλέον, η ενσωμάτωση πολυτροπικών δεδομένων (κείμενο, εικόνες, βίντεο) θα μπορούσε να προσφέρει μια πιο ολοκληρωμένη κατανόηση του λόγου για την κλιματική αλλαγή.

Εν κατακλείδι, η υπάρχουσα έρευνα σχετικά με το διάλογο για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης έχει παράσχει πολύτιμες πληροφορίες για το κοινό συναίσθημα, τα βασικά θέματα και τη διάδοση των πληροφοριών. Τα ευρήματα αυτά ενημερώνουν για την ανάπτυξη πολιτικής, ενισχύουν τις στρατηγικές επικοινωνίας και προωθούν την εμπλοκή του κοινού, καθιστώντας την ανάλυση των μέσων κοινωνικής δικτύωσης απαραίτητο εργαλείο στον αγώνα κατά της κλιματικής αλλαγής.

4.7 Κενά και ευκαιρίες

4.7.1 Κενά στην τρέχουσα έρευνα

Παρά τις σημαντικές προόδους στην ανάλυση του λόγου για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης, παραμένουν αρκετά κενά που πρέπει να αντιμετωπιστούν για να βελτιωθεί η κατανόηση και η εφαρμογή αυτών των γνώσεων:

1. Ανίχνευση σαρκασμού και ειρωνείας:

- Τα τρέχοντα μοντέλα ανάλυσης συναισθήματος συχνά δυσκολεύονται να ανιχνεύσουν και να ερμηνεύσουν με ακρίβεια τον σαρκασμό και την ειρωνεία, τα οποία είναι διαδεδομένα στην επικοινωνία στα μέσα κοινωνικής δικτύωσης. Η λανθασμένη ταξινόμηση των σαρκαστικών σχολίων μπορεί να οδηγήσει σε εσφαλμένες εκτιμήσεις συναισθήματος και να αλλοιώσει τη συνολική ανάλυση [3,5].

2. Κατανόηση του πλαισίου:

- Πολλά υπάρχοντα μοντέλα αποτυγχάνουν να καταγράψουν πλήρως το πλαίσιο στο οποίο εκφράζονται τα συναισθήματα. Η κατανόηση του ευρύτερου πλαισίου είναι ζωτικής σημασίας για την ακριβή ερμηνεία των συναισθημάτων, ιδίως σε πολύπλοκες και διαφοροποιημένες συζητήσεις όπως η κλιματική αλλαγή [12].

3. Πολυγλωσσική ανάλυση:

- Υπάρχει ανάγκη για πιο ισχυρά πολύγλωσσα εργαλεία ανάλυσης συναισθήματος. Οι περισσότερες μελέτες επικεντρώνονται σε δεδομένα στην αγγλική γλώσσα, αλλά η κλιματική αλλαγή είναι ένα παγκόσμιο ζήτημα που απαιτεί ανάλυση σε πολλές γλώσσες και πολιτισμικά πλαίσια. Το κενό αυτό περιορίζει την πληρότητα και τη δυνατότητα εφαρμογής των ευρημάτων [2,5,12,21,22].

4. Μακροχρόνιες μελέτες:

- Ενώ πολλές μελέτες παρέχουν στιγμιότυπα του συναισθήματος σε συγκεκριμένες χρονικές στιγμές, υπάρχει έλλειψη διαχρονικών ερευνών που να παρακολουθούν πώς εξελίσσεται το δημόσιο συναίσθημα απέναντι στην κλιματική αλλαγή σε μεγαλύτερες χρονικές περιόδους. Τέτοιες μελέτες είναι ζωτικής σημασίας για την κατανόηση των μακροπρόθεσμων επιπτώσεων των στρατηγικών επικοινωνίας και των αλλαγών πολιτικής[2,5,21,22].

4.7.2 Ευκαιρίες για μελλοντική έρευνα

1. Προηγμένες τεχνικές NLP:

- Η αξιοποίηση προηγμένων τεχνικών επεξεργασίας φυσικής γλώσσας (NLP), όπως μετασχηματιστές και βαθιά μοντέλα συμφραζομένων, μπορεί να βελτιώσει την ανίχνευση του σαρκασμού, της ειρωνείας και του πλαισίου. Μοντέλα όπως τα BERT και RoBERTa υπόσχονται πολλά στον τομέα αυτό και μπορούν να ρυθμιστούν περαιτέρω για τον λόγο περί κλιματικής αλλαγής[9,24].

2. Ενσωμάτωση πολυτροπικών δεδομένων:

- Ο συνδυασμός της ανάλυσης κειμένου με άλλους τύπους δεδομένων, όπως εικόνες και βίντεο, μπορεί να προσφέρει μια πλουσιότερη κατανόηση του λόγου για την κλιματική αλλαγή. Η πολυτροπική ανάλυση μπορεί να καταγράψει ένα ευρύτερο φάσμα δημόσιας έκφρασης και συναισθήματος.

3. Διαπολιτισμικές μελέτες:

- Η ανάπτυξη πολύγλωσσων και διαπολιτισμικών εργαλείων ανάλυσης συναισθήματος θα επιτρέψει στους ερευνητές να καταγράψουν μια πιο ολοκληρωμένη παγκόσμια προοπτική του λόγου για την κλιματική αλλαγή. Αυτό μπορεί να ενημερώσει για αποτελεσματικότερες στρατηγικές διεθνούς επικοινωνίας και πολιτικής.

4. Παρακολούθηση και δράση σε πραγματικό χρόνο:

- Η εφαρμογή εργαλείων ανάλυσης συναισθήματος σε πραγματικό χρόνο μπορεί να βοηθήσει τους υπεύθυνους χάραξης πολιτικής και τους οργανισμούς να ανταποκρίνονται ταχύτερα στις μεταβολές της κοινής γνώμης, στην παραπληροφόρηση και στις αναδυόμενες ανησυχίες. Αυτή η προληπτική προσέγγιση μπορεί να ενισχύσει την εμπλοκή του κοινού και την υποστήριξη των πρωτοβουλιών για το κλίμα.

Συνοπτικά, η αντιμετώπιση αυτών των κενών και η αξιοποίηση των ευκαιριών θα προωθήσει τον τομέα της ανάλυσης του λόγου για την κλιματική αλλαγή στα μέσα κοινωνικής δικτύωσης, οδηγώντας σε πιο ακριβείς, περιεκτικές και εφαρμόσιμες γνώσεις.

5. Μεθοδολογία

5.1 Επεξήγηση του dataset

Twitter Climate Change Sentiment Dataset

Αυτό το dataset, που επικεντρώνεται σε tweets που αφορούν την κλιματική αλλαγή, δημιουργήθηκε με την οικονομική υποστήριξη μιας επιχορήγησης JELF του Canada Foundation for Innovation που χορηγήθηκε στον καθηγητή Chris Bauch στο Πανεπιστήμιο του Waterloo. Σκοπός της χρηματοδότησης αυτής ήταν η υποστήριξη καινοτόμων ερευνητικών προσπαθειών και το παρόν σύνολο δεδομένων αποτελεί σημαντικό πόρο για την κατανόηση του δημόσιου λόγου σχετικά με την κλιματική αλλαγή κατά τη διάρκεια μιας συγκεκριμένης περιόδου.

Το σύνολο δεδομένων περιλαμβάνει tweets που συλλέχθηκαν μεταξύ της 27ης Απριλίου 2015 και της 21ης Φεβρουαρίου 2018. Κατά τη διάρκεια αυτής της σχεδόν τριετούς περιόδου, χρησιμοποιήθηκε μια ολοκληρωμένη διαδικασία συλλογής για τη συλλογή tweets που αναφέρουν ή συζητούν την κλιματική αλλαγή. Το σύνολο δεδομένων που προέκυψε περιλαμβάνει συνολικά 43.943 tweets, καθένα από τα οποία υποβλήθηκε σε αυστηρή διαδικασία σχολιασμού.

Κάθε tweet στο σύνολο δεδομένων εξετάστηκε και σχολιάστηκε ανεξάρτητα από τρεις διαφορετικούς κριτές. Αυτή η προσέγγιση με πολλούς κριτές εξασφαλίζει υψηλότερο επίπεδο αξιοπιστίας και ακρίβειας στην ταξινόμηση κάθε tweet. Για να ενισχυθεί περαιτέρω η αξιοπιστία του συνόλου δεδομένων, συμπεριλήφθηκαν μόνο τα tweet που έλαβαν ομόφωνη συμφωνία και από τους τρεις κριτές. Αυτό σημαίνει ότι εάν έστω και ένας κριτής διαφωνούσε με τους υπόλοιπους σχετικά με την ταξινόμηση ενός tweet, το εν λόγω tweet απορρίφθηκε από το σύνολο δεδομένων. Αυτό το αυστηρό κριτήριο υπογραμμίζει τη δέσμευση του συνόλου δεδομένων για ποιότητα και συναίνεση, διασφαλίζοντας ότι τα περιεχόμενα tweets αντικατοπτρίζουν σαφή και συμφωνημένα συναισθήματα.

Κάθε tweet στο σύνολο δεδομένων ταξινομείται σε μία από τις τέσσερις κατηγορίες, οι οποίες έχουν σχεδιαστεί για να καταγράφουν ένα ευρύ φάσμα απόψεων σχετικά με την κλιματική αλλαγή. Οι κατηγορίες αυτές είναι οι εξής: Οι κατηγορίες αυτές είναι οι ακόλουθες:

1. Νέα (2): Τα tweets αυτής της κατηγορίας παραπέμπουν σε πραγματικά ειδησεογραφικά άρθρα ή αναφορές σχετικά με την κλιματική αλλαγή. Αυτά τα tweets θεωρείται ότι παρέχουν αντικειμενικές πληροφορίες και ενημερώσεις σχετικά με την κλιματική αλλαγή, χρησιμεύοντας ως κρίσιμη πηγή διάδοσης πληροφοριών.

2. Pro (1): Η κατηγορία αυτή περιλαμβάνει tweets που υποστηρίζουν την πεποίθηση για την ανθρωπογενή κλιματική αλλαγή. Τα tweets που ταξινομούνται ως "Pro" συνήθως εκφράζουν συμφωνία με την επιστημονική συναίνεση σχετικά με την κλιματική αλλαγή, υποστηρίζουν περιβαλλοντικές πολιτικές ή υπογραμμίζουν τον αντίκτυπο των ανθρώπινων δραστηριοτήτων στο κλίμα.

3. Ουδέτερα (0): Τα ουδέτερα tweets δεν υποστηρίζουν ούτε αντικρούουν την πεποίθηση στην ανθρωπογενή κλιματική αλλαγή. Αυτά τα tweets μπορεί να παρουσιάζουν πληροφορίες χωρίς να παίρνουν θέση, να συζητούν την κλιματική αλλαγή με αμερόληπτο τρόπο ή απλώς να αναφέρουν την κλιματική αλλαγή χωρίς να εκφράζουν κάποια συγκεκριμένη άποψη.

4. Anti (-1): Τα tweets στην κατηγορία "Anti" εκφράζουν σκεπτικισμό ή δυσπιστία στην ανθρωπογενή κλιματική αλλαγή. Αυτά τα tweets μπορεί να αμφισβητούν την επιστημονική συναίνεση, να αρνούνται τον αντίκτυπο των ανθρώπινων δραστηριοτήτων στο κλίμα ή να προωθούν εναλλακτικές εξηγήσεις για τα φαινόμενα της κλιματικής αλλαγής.

Η σχολαστική διαδικασία συλλογής και σχολιασμού του συνόλου δεδομένων το καθιστά πολύτιμο πόρο για ερευνητές, υπεύθυνους χάραξης πολιτικής και εκπαιδευτικούς. Παρέχοντας ένα σαφές και αξιόπιστο στιγμιότυπο του δημόσιου λόγου για την κλιματική αλλαγή κατά τη διάρκεια μιας καθορισμένης περιόδου, το σύνολο δεδομένων επιτρέπει τη σε βάθος ανάλυση των τάσεων, του κλίματος και της διάδοσης των πληροφοριών σχετικά με την κλιματική αλλαγή.

Οι ερευνητές μπορούν να χρησιμοποιήσουν αυτό το σύνολο δεδομένων για να μελετήσουν πώς εξελίχθηκε η κοινή γνώμη για την κλιματική αλλαγή, να εντοπίσουν τους βασικούς παράγοντες που επηρεάζουν τη συζήτηση για την κλιματική αλλαγή και να εξετάσουν τον αντίκτυπο σημαντικών γεγονότων στο δημόσιο συναίσθημα. Για τους υπεύθυνους χάραξης πολιτικής, η κατανόηση των απόψεων του κοινού για την κλιματική αλλαγή μπορεί να ενημερώσει για την ανάπτυξη στρατηγικών επικοινωνίας και πολιτικών αποφάσεων που βρίσκουν απήχηση σε διαφορετικά τμήματα του πληθυσμού. Οι εκπαιδευτικοί μπορούν να χρησιμοποιήσουν το σύνολο δεδομένων για να εμπλέξουν τους μαθητές σε συζητήσεις σχετικά με την κλιματική αλλαγή, τον γραμματισμό στα μέσα ενημέρωσης και τον ρόλο των μέσων κοινωνικής δικτύωσης στη διαμόρφωση της κοινής γνώμης.

5.2 Βήματα preprocessing στην ανάλυση συναισθήματος

Η προεπεξεργασία είναι ένα κρίσιμο βήμα στην ανάλυση συναισθήματος, καθώς μετατρέπει το ακατέργαστο κείμενο σε μια καθαρή και δομημένη μορφή που μπορεί να αναλυθεί αποτελεσματικά από τα μοντέλα μηχανικής μάθησης. Ο αγωγός preprocessing περιλαμβάνει συνήθως διάφορα βασικά βήματα: καθαρισμό, tokenization, αφαίρεση των stop words, lemmatization και stemming. Κάθε ένα από αυτά τα βήματα διαδραματίζει σημαντικό ρόλο στην ενίσχυση της ποιότητας των δεδομένων κειμένου και στη βελτίωση της απόδοσης των μοντέλων ανάλυσης συναισθήματος.

5.2.1 Καθαρισμός

Το πρώτο βήμα στον αγωγό preprocessing είναι ο καθαρισμός των δεδομένων κειμένου. Ο καθαρισμός περιλαμβάνει την αφαίρεση τυχόν ανεπιθύμητων χαρακτήρων, συμβόλων ή στοιχείων που δεν συμβάλλουν στη διαδικασία ανάλυσης συναισθήματος. Το βήμα αυτό είναι

απαραίτητο για τη μείωση του θορύβου στα δεδομένα και τη διασφάλιση ότι διατηρούνται μόνο οι σχετικές πληροφορίες.

Αναλαμβανόμενα καθήκοντα:

- Αφαίρεση σημείων στίξης: Τα σημεία στίξης, όπως κόμματα, τελείες και θαυμαστικά, συνήθως αφαιρούνται επειδή δεν έχουν σημαντική σημασία στην ανάλυση συναισθήματος. Για παράδειγμα, το "Hello, world!" γίνεται "Hello world".
- Αφαίρεση ειδικών χαρακτήρων: Ειδικοί χαρακτήρες όπως @, #, \$, % και & αφαιρούνται για την απλούστευση του κειμένου. Αυτοί οι χαρακτήρες μπορούν συχνά να βρεθούν στα δεδομένα των μέσων κοινωνικής δικτύωσης.
- Αφαίρεση αριθμών: Οι αριθμοί συνήθως αφαιρούνται, εκτός αν είναι σημαντικοί από άποψη πλαισίου. Για παράδειγμα, ο αριθμός στο "COVID-19" είναι κρίσιμος, αλλά σε πολλά άλλα πλαίσια, οι αριθμοί μπορεί να μην συμβάλλουν στο συναίσθημα.
- Μετατροπή σε πεζά γράμματα: Η μετατροπή όλου του κειμένου σε πεζά βοηθά στη διατήρηση της ομοιομορφίας, καθώς οι λέξεις "Happy" και "happy" θα πρέπει να αντιμετωπίζονται ως η ίδια λέξη.

Παράδειγμα:

- Ακατέργαστο κείμενο: "the new iphone12 is awesome!! #apple #innovation #apple"
- Καθαρισμένο κείμενο: "the new iphone is awesome apple innovation apple"

5.2.2 Tokenization

Το Tokenization είναι η διαδικασία διαχωρισμού του κειμένου σε μεμονωμένες μονάδες, γνωστές ως tokens. Τα σημεία μπορεί να είναι λέξεις, φράσεις ή ακόμη και χαρακτήρες, ανάλογα με το επίπεδο ανάλυσης. Η τοκενικοποίηση είναι θεμελιώδης επειδή μετατρέπει το κείμενο σε διαχειρίσιμα κομμάτια που μπορούν να υποστούν περαιτέρω επεξεργασία και ανάλυση.

Τύποι Tokenization:

- Τοκενικοποίηση λέξεων: Διαχωρίζει το κείμενο σε μεμονωμένες λέξεις. Αυτή είναι η πιο συνηθισμένη μορφή tokenization στην ανάλυση συναισθήματος.
- Τοκιστικοποίηση πρότασης: Διαχωρίζει το κείμενο σε προτάσεις. Αυτό μπορεί να είναι χρήσιμο για εργασίες όπου το πλαίσιο σε επίπεδο πρότασης είναι σημαντικό.
- Τοκενοποίηση χαρακτήρων: Διαχωρίζει το κείμενο σε μεμονωμένους χαρακτήρες. Αυτό είναι λιγότερο συνηθισμένο, αλλά μπορεί να είναι χρήσιμο σε συγκεκριμένες εφαρμογές, όπως η ανάλυση μοτίβων σε επίπεδο χαρακτήρα σε κείμενο.

Παράδειγμα:

- Καθαρισμένο κείμενο: "the new iphone12 is awesome!! #apple #innovation #apple"

- Κωδικοποιημένο κείμενο (Word Tokenization): ["the", "new", "iphone", "is", "awesome", "apple", "innovation", "apple"]

5.2.3 Αφαίρεση λέξεων διακοπής

Οι σταματημένες λέξεις είναι κοινές λέξεις που συνήθως δεν φέρουν σημαντικό νόημα και συχνά αφαιρούνται για να μειωθεί η διαστατικότητα των δεδομένων. Παραδείγματα λέξεων διακοπής περιλαμβάνουν τα "the", "is", "in", "and", "but" και "or". Η αφαίρεση αυτών των λέξεων βοηθά στην εστίαση στις πιο σημαντικές λέξεις που συμβάλλουν στο συναίσθημα.

Γιατί να αφαιρέσετε τις σταματημένες λέξεις:

- Μείωση του θορύβου: Οι λέξεις στάσης μπορούν να προσθέσουν θόρυβο στα δεδομένα και ενδέχεται να μην συμβάλλουν στην εργασία ανάλυσης συναισθήματος.
- Μείωση διαστάσεων: Με την αφαίρεση των λέξεων στάσης, ο χώρος χαρακτηριστικών μειώνεται, καθιστώντας την εκπαίδευση του μοντέλου πιο αποτελεσματική.

Παράδειγμα:

- Κείμενο με λέξεις-κλειδιά: ["the", "new", "iphone", "is", "ziga", "apple", "innovation", "apple"]
- Κείμενο μετά την αφαίρεση των σταματημένων λέξεων: ["new", "iphone", "awesome", "apple", "innovation", "apple"]

5.2.4 Lemmatization

Lemmatization είναι η διαδικασία αναγωγής των λέξεων στη βασική ή ριζική τους μορφή, γνωστή ως λήμμα. Σε αντίθεση με το stemming, το οποίο απλώς αποκόπτει τις καταλήξεις των λέξεων, η λημματοποίηση λαμβάνει υπόψη το πλαίσιο και μετατρέπει τις λέξεις στη σημαίνουσα βασική τους μορφή. Για παράδειγμα, οι λέξεις "running", "ran" και "runs" μετατρέπονται σε "run".

Σημαντικότητα της λεμματοποίησης:

- Ακρίβεια σε σχέση με το πλαίσιο: Η λεμματοποίηση διασφαλίζει ότι οι λέξεις μετατρέπονται στη βασική τους μορφή λαμβάνοντας υπόψη το πλαίσιο, γεγονός που βοηθά στη διατήρηση του σωστού νοήματος.
- Συνέπεια: Με τη μετατροπή των διαφορετικών μορφών μιας λέξης σε μια κοινή βασική μορφή, η λεμματοποίηση εξασφαλίζει τη συνέπεια στα δεδομένα κειμένου.

Παράδειγμα:

- Κείμενο μετά την αφαίρεση των σταματημένων λέξεων: ["new", "iphone", "awesome", "apple", "innovation", "apple"]
- Κείμενο μετά τη λεμματοποίηση: ["new", "iphone", "awesome", "apple", "innovation", "apple"]

5.2.5 Stemming

Η αποδελτίωση είναι η διαδικασία αναγωγής των λέξεων στη ριζική τους μορφή με την αφαίρεση των επιθέτων και των προθέτων. Είναι μια πιο επιθετική προσέγγιση από τη λημματοποίηση και μπορεί μερικές φορές να οδηγήσει σε λέξεις που δεν ανήκουν στο λεξικό. Για παράδειγμα, οι λέξεις "running", "runner" και "runs" ανάγονται σε "run".

Τύποι αλγορίθμων Stemming:

- Porter Stemmer: Ένας από τους πιο ευρέως χρησιμοποιούμενους αλγορίθμους στεμματοποίησης, ο οποίος χρησιμοποιεί ένα σύνολο κανόνων για την επαναληπτική αφαίρεση των επιθέτων από τις λέξεις.
- Snowball Stemmer: Βελτίωση του Porter Stemmer, που παρέχει μια πιο επιθετική προσέγγιση στελέχωσης.
- Lancaster Stemmer: Γνωστός για την απλότητα και την ταχύτητά του, αλλά μπορεί να είναι υπερβολικά επιθετικός, οδηγώντας συχνά σε μη λεξικογραφικά στελέχη.

Παράδειγμα:

- Κείμενο μετά την αφαίρεση των σταματημένων λέξεων: ["new", "iphone", "awesome", "apple", "innovation", "apple"]- Κείμενο μετά από Stemming: ["new", "iphon", "awesom", "appl", "innov", "appl"]

Σύγκριση του Lemmatization και του Stemming

Παρόλο που τόσο η λημματοποίηση όσο και η στεμματοποίηση στοχεύουν στην αναγωγή των λέξεων στις βασικές τους μορφές, έχουν βασικές διαφορές:

- Λεμματισμός:
 - Λαμβάνει υπόψη τα συμφραζόμενα και το μέρος του λόγου της λέξης.
 - Παράγει ουσιαστικές βασικές μορφές που είναι πραγματικές λέξεις.
 - Παράδειγμα: Το "better" γίνεται "good" (αν αναλυθεί με βάση τα συμφραζόμενα).
- Συμπέρασμα:
 - Χρησιμοποιεί ευρετικούς κανόνες για την αφαίρεση επιθέτων και προθεμάτων.
 - Μπορεί να οδηγήσει σε λέξεις που δεν ανήκουν στο λεξικό.
 - Παράδειγμα: Το "running" γίνεται "run", το "happiness" γίνεται "happi".

Ενσωμάτωση βημάτων preprocessing

5.2.6 Συμπέρασμα

Η προεπεξεργασία αποτελεί ζωτικό βήμα για την προετοιμασία δεδομένων κειμένου για ανάλυση συναισθήματος. Κάθε βήμα -καθαρισμός, tokenization, αφαίρεση των stop words,

lemmatization και stemming- συμβάλλει στη μετατροπή του ακατέργαστου κειμένου σε μορφή που μπορεί να αναλυθεί αποτελεσματικά από μοντέλα μηχανικής μάθησης. Με την αφαίρεση του θορύβου, τη μείωση της διαστατικότητας και την τυποποίηση του κειμένου, αυτά τα βήματα preprocessing βελτιώνουν την απόδοση και την ακρίβεια της ανάλυσης συναισθήματος, καθιστώντας τα απαραίτητα για κάθε έργο εξόρυξης κειμένου.

6. Λεπτομερής Επεξήγηση Κώδικα

Ο [κώδικας](#) υπάρχει στο Github.

6.1 Εισαγωγή βιβλιοθηκών

- Η `numpy` και η `pandas` χρησιμοποιούνται για την επεξεργασία δεδομένων.
- Η `matplotlib` χρησιμοποιείται για την οπτικοποίηση δεδομένων.
- Η `nltk` χρησιμοποιείται για την επεξεργασία φυσικής γλώσσας.

Importing Necessary Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import nltk
```

6.2 Λήψη δεδομένων NLTK

Το NLTK παρέχει προ-εκπαιδευμένα μοντέλα και σώματα δεδομένων που είναι απαραίτητα για διάφορες εργασίες NLP. Η λήψη αυτών των πόρων εξασφαλίζει ότι το σενάριο μπορεί να εκτελέσει tokenization, POS tagging και NER χωρίς να χρειάζεται να εκπαιδεύσει μοντέλα από το μηδέν.

- 'punkt': Χρησιμοποιείται για τη μετατροπή κειμένου σε προτάσεις ή λέξεις.
- 'averaged_perceptron_tagger': Μοντέλο για part-of-speech tagging.
- 'maxent_ne_chunker': A model for named entity chunking.
- 'words': Ένα σύνολο αγγλικών λέξεων.

Downloading NLTK Data

```
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

6.2.1 Φόρτωση και επιθεώρηση του συνόλου δεδομένων

- Η φόρτωση του συνόλου δεδομένων σε ένα DataFrame επιτρέπει την αποτελεσματική επεξεργασία και ανάλυση των δεδομένων. Το σχήμα του DataFrame παρέχει μια γρήγορη επισκόπηση του μεγέθους του συνόλου δεδομένων, η οποία είναι χρήσιμη για τον προγραμματισμό της περαιτέρω ανάλυσης.
- Η εμφάνιση των πρώτων γραμμών βοηθά στην κατανόηση της δομής και του περιεχομένου του συνόλου δεδομένων, συμπεριλαμβανομένου οποιουδήποτε πιθανού καθαρισμού δεδομένων που μπορεί να χρειαστεί.

```
df = pd.read_csv('twitter_sentiment_data.csv')  
print("Dataset shape:", df.shape)
```

Displaying the First Few Rows

```
print(df.head())
```

-pd.read_csv('twitter_sentiment_data.csv'): Reads a CSV file into a DataFrame. This function is highly flexible and can handle various types of CSV files.

-df.shape: Prints the shape of the DataFrame (number of rows and columns). This gives a quick overview of the dataset's size.

6.2.2 Απεικόνιση της κατανομής του συναισθήματος

- Η απεικόνιση της κατανομής συναισθήματος με ένα κυκλικό διάγραμμα παρέχει άμεση κατανόηση της κατανομής συναισθήματος του συνόλου δεδομένων. Αυτό είναι ιδιαίτερα χρήσιμο για τη διερευνητική ανάλυση δεδομένων (EDA) για να διαπιστωθεί αν το σύνολο δεδομένων είναι ισορροπημένο ή λοξό προς ορισμένα συναισθήματα.

Plotting a Pie Chart for Sentiment Distribution

```
df['sentiment'].value_counts().sort_index() \  
    .plot(kind='pie', title='Opinions', autopct='%1.1f%%', startangle=90)  
plt.ylabel("")  
plt.show()
```

-df['sentiment'].value_counts(): Counts the occurrences of each unique value in the 'sentiment' column.

-sort_index(): Sorts the index of the resulting Series, ensuring that the sentiment categories are in order.

-plot(kind='pie', title='Opinions', autopct='%1.1f%%', startangle=90): Creates a pie chart with the following parameters:

-kind='pie': Specifies that the plot should be a pie chart.

- title='Opinions': Sets the title of the plot.

- autopct='%1.1f%%': Formats the labels on the pie slices to show percentages with one decimal place.

- startangle=90: Rotates the start of the pie chart by 90 degrees for better visual appeal.

- plt.ylabel("): Removes the default y-label which is unnecessary for pie charts.

- plt.show(): Displays the plot.

6.2.3 Επεξεργασία NLTK

- Η τοκενικοποίηση σπάει το κείμενο σε διαχειρίσιμα κομμάτια (tokens), το οποίο είναι το πρώτο βήμα στην ανάλυση κειμένου.

- Η POS tagging παρέχει γραμματικές πληροφορίες για κάθε token, οι οποίες είναι χρήσιμες για την κατανόηση της συντακτικής δομής του κειμένου.

- Η NER προσδιορίζει και ταξινομεί οντότητες στο κείμενο, κάτι που είναι σημαντικό για εργασίες όπως η εξαγωγή πληροφοριών, η απάντηση ερωτήσεων και άλλα.

NLTK Processing on Example Text

```
example = "This is an example sentence for NLTK processing."
```

This line sets a sample sentence to demonstrate tokenization, part-of-speech tagging, and named entity recognition.

Tokenization

```
tokens = nltk.word_tokenize(example)
```

```
print("Tokens:", tokens[:10])
```

- `nltk.word_tokenize(example)`: Splits the example sentence into individual words (tokens). Tokenization is the first step in text processing, where the text is broken down into smaller units.

- `tokens[:10]`: Prints the first ten tokens. Since our example sentence is short, this will print all the tokens.

Part-of-Speech (POS) Tagging

```

'''python
tagged = nltk.pos_tag(tokens)
print("Tagged tokens:", tagged[:10])
'''

- `nltk.pos_tag(tokens)`: Tags each token with its part-of-speech, such as noun, verb, adjective, etc. POS tagging is crucial for understanding the grammatical structure of the text.
- `tagged[:10]`: Prints the first ten tagged tokens. Each token is paired with its POS tag.

Named Entity Recognition (NER)

'''python
entities = nltk.chunk.ne_chunk(tagged)
print("Named entities:")
entities.pprint()
'''

- `nltk.chunk.ne_chunk(tagged)`: Identifies named entities in the tagged tokens. Named entities are words or phrases that represent specific objects such as persons, organizations, locations, etc.
- `entities.pprint()`: Prints the named entities in a tree structure. This helps to visualize the entities and their types.

```

Αυτό το σενάριο παρουσιάζει μια τυπική ροή εργασίας για την ανάλυση κειμένου με χρήση της Python. Ξεκινά με τη φόρτωση και την επιθεώρηση δεδομένων, προχωρά στην οπτικοποίηση δεδομένων και καταλήγει στην επεξεργασία κειμένου με χρήση του NLTK. Κάθε βήμα είναι ζωτικής σημασίας για την αποτελεσματική κατανόηση και ανάλυση των δεδομένων. Ακολουθώντας αυτά τα βήματα, μπορείτε να αποκτήσετε γνώσεις σχετικά με τη δομή, την κατανομή και το περιεχόμενο των δεδομένων, οι οποίες είναι απαραίτητες για κάθε έργο ανάλυσης δεδομένων ή μηχανικής μάθησης.

6.3 Εφαρμοσμένες μέθοδοι: VADER

Εισαγωγή στο VADER

Το VADER (Valence Aware Dictionary and sEntiment Reasoner) είναι ένα εργαλείο ανάλυσης συναισθήματος ειδικά σχεδιασμένο για να χειρίζεται κείμενα μέσω κοινωνικής δικτύωσης και άλλα σύντομα, ανεπίσημα κείμενα. Το VADER αναπτύχθηκε από τους C.J. Hutto και Eric Gilbert και βασίζεται σε μια προσέγγιση που βασίζεται σε λεξικό και κανόνες. Αυτό σημαίνει ότι βασίζεται σε έναν προκαθορισμένο κατάλογο λέξεων (ένα λεξικό) και σε ένα σύνολο γραμματικών και συντακτικών κανόνων για τον προσδιορισμό του συναισθήματος ενός κειμένου. Το VADER είναι αποτελεσματικό στην καταγραφή τόσο της πολικότητας (θετικό/αρνητικό/ουδέτερο) όσο και της έντασης του συναισθήματος σε κείμενο. Είναι ιδιαίτερα κατάλληλο για την ανάλυση του είδους της γλώσσας που συναντάται στις πλατφόρμες κοινωνικής δικτύωσης, όπου οι εκφράσεις μπορεί να είναι ιδιαίτερα διαφοροποιημένες και ανεπίσημες.

1. Προεπεξεργασία για το VADER:

- Καθαρισμός κειμένου: Παρόμοια με την προεπεξεργασία για το RoBERTa, τα ακατέργαστα tweets καθαρίστηκαν για να αφαιρεθούν οι περιττοί χαρακτήρες και τα στοιχεία που δεν συμβάλλουν στην ανάλυση συναισθήματος.

- Tokenization: Το κείμενο διαχωρίστηκε σε μεμονωμένες λέξεις και φράσεις, οι οποίες στη συνέχεια συγκρίθηκαν με το λεξικό συναισθήματος του VADER.

2. Βαθμολόγηση συναισθήματος με το VADER:

- Αντιστοίχιση λεξικού: Κάθε λέξη στο tweet αντιστοιχίστηκε με το λεξικό συναισθημάτων του VADER. Το λεξικό αποδίδει βαθμολογία συναισθήματος στις λέξεις με βάση την πολικότητα και την έντασή τους.

- Εφαρμογή κανόνων: Το VADER εφαρμόζει διάφορους κανόνες για τον χειρισμό γραμματικών και συντακτικών δομών. Για παράδειγμα, προσαρμόζει τη βαθμολογία συναισθήματος με βάση την παρουσία αρνήσεων (π.χ. το "όχι καλό" γίνεται λιγότερο θετικό) και τροποποιητών βαθμού (π.χ. το "ζίγκα καλό" γίνεται πιο θετικό).

- Αθροίζοντας τις βαθμολογίες: Οι βαθμολογίες συναισθήματος των μεμονωμένων tokens αθροίστηκαν για να παραχθεί μια συνολική βαθμολογία συναισθήματος για κάθε tweet. Αυτή η βαθμολογία αντιστοιχίστηκε στη συνέχεια σε μια κατηγορία συναισθήματος (θετικό, αρνητικό, ουδέτερο).

3. Επιλογή χαρακτηριστικών με το VADER:

- Προσδιορισμός σημαντικών tokens: Το λεξικό του VADER παρείχε έναν κατάλογο σημαντικών tokens που φέρουν έντονο συναίσθημα. Αυτά τα tokens χρησιμοποιήθηκαν ως χαρακτηριστικά για περαιτέρω ανάλυση και εκπαίδευση του μοντέλου.

- Λεξικά χαρακτηριστικά: Η παρουσία και η συχνότητα των συναισθηματικά φορτισμένων λέξεων που εντοπίστηκαν από το VADER χρησιμοποιήθηκαν ως χαρακτηριστικά στο μοντέλο ταξινόμησης συναισθήματος. Αυτή η προσέγγιση αξιοποίησε την ικανότητα του VADER να εντοπίζει λέξεις και φράσεις με νόημα που συμβάλλουν άμεσα στο συναίσθημα.

6.3.1 Εισαγωγή βιβλιοθηκών και λήψη πόρων

Για να ξεκινήσουμε με το VADER, πρέπει να εισάγουμε τις απαραίτητες βιβλιοθήκες και να κατεβάσουμε τους απαιτούμενους πόρους. Ακολουθεί η αρχική ρύθμιση:

```
from nltk.sentiment import SentimentIntensityAnalyzer
nltk.download('VADER_lexicon')
from tqdm import tqdm
```

- SentimentIntensityAnalyzer: Αυτό είναι το κύριο εργαλείο από τη βιβλιοθήκη NLTK για την εκτέλεση ανάλυσης συναισθήματος με χρήση του μοντέλου VADER. Παρέχει μεθόδους για την ανάλυση κειμένου και την επιστροφή βαθμολογιών συναισθήματος.

- nltk.download('VADER_lexicon'): Αυτή η εντολή κατεβάζει το λεξικό VADER, έναν κατάλογο λέξεων και τις σχετικές βαθμολογίες συναισθήματος. Το λεξικό είναι ζωτικής σημασίας για τον SentimentIntensityAnalyzer προκειμένου να εκτελέσει ανάλυση συναισθήματος.

- tqdm: Πρόκειται για μια βιβλιοθήκη που χρησιμοποιείται για την εμφάνιση ράβδων προόδου. Είναι ιδιαίτερα χρήσιμη κατά την επεξεργασία μεγάλων συνόλων δεδομένων, καθώς παρέχει μια οπτική ένδειξη για το πόσο μέρος της εργασίας έχει ολοκληρωθεί και πόσο απομένει.

6.3.2 Αρχικοποίηση του αναλυτή συναισθήματος και δοκιμή του

Μετά την εισαγωγή των απαραίτητων βιβλιοθηκών, το επόμενο βήμα είναι να αρχικοποιήσετε τον SentimentIntensityAnalyzer και να τον δοκιμάσετε με δείγματα κειμένων για να βεβαιωθείτε ότι λειτουργεί σωστά:

```
sia = SentimentIntensityAnalyzer()
sia.polarity_scores('i am devastated')
sia.polarity_scores(example)
```

- sia = SentimentIntensityAnalyzer(): Αυτή η γραμμή αρχικοποιεί τον αναλυτή συναισθήματος.

- `sia.polarity_scores('i am devastated')`: Εδώ, δοκιμάζουμε τον αναλυτή με ένα δείγμα πρότασης ('i am devastated') για να δούμε πώς βαθμολογεί το συναίσθημα. Αυτή η μέθοδος επιστρέφει ένα λεξικό με βαθμολογίες συναισθήματος για θετικό, ουδέτερο, αρνητικό και σύνθετο (συνολικό συναίσθημα).
- `sia.polarity_scores(example)`: Αυτή η γραμμή υποθέτει ότι το `'example'` περιέχει ένα δείγμα κειμένου. Δείχνει πώς μπορείτε να αναλύσετε οποιοδήποτε κείμενο χρησιμοποιώντας την ίδια μέθοδο.

6.3.3 Ανάλυση του συναισθήματος ενός συνόλου δεδομένων

Για να εκτελέσουμε ανάλυση συναισθήματος σε ένα σύνολο δεδομένων, επαναλαμβάνουμε κάθε γραμμή, εξάγουμε το κείμενο και το αναλύουμε χρησιμοποιώντας τον `SentimentIntensityAnalyzer`. Τα αποτελέσματα αποθηκεύονται σε ένα λεξικό για περαιτέρω επεξεργασία:

```
res = {}

for i, row in tqdm(df.iterrows(), total=len(df)):
    text = row['message']
    myid = row['id']
    res[myid] = sia.polarity_scores(text)
```

- `for i, row in tqdm(df.iterrows(), total=len(df))`: Αυτός ο βρόχος επαναλαμβάνει κάθε γραμμή του `DataFrame 'df'` χρησιμοποιώντας την `'iterrows()'`. Το περιτύλιγμα `'tqdm'` παρέχει μια γραμμή προόδου, διευκολύνοντας την παρακολούθηση της προόδου.
- `text = row['message']`: Εξάγει το μήνυμα κειμένου από την τρέχουσα γραμμή.
- `myid = row['id']`: Εξάγει το μοναδικό αναγνωριστικό από την τρέχουσα σειρά.
- `res[myid] = sia.polarity_scores(text)`: Αναλύει το συναίσθημα του κειμένου χρησιμοποιώντας το `VADER` και αποθηκεύει τα αποτελέσματα στο λεξικό `'res'`, με κλειδί το ID του μηνύματος.

6.3.4 Δημιουργία ενός πλαισίου δεδομένων των αποτελεσμάτων

Μόλις ολοκληρωθεί η ανάλυση συναισθήματος, πρέπει να μετατρέψουμε το λεξικό αποτελεσμάτων σε `DataFrame` και να το συγχωνεύσουμε με το αρχικό `DataFrame`:

```
VADERS = pd.DataFrame(res).T
```

```
VADERS = VADERS.reset_index().rename(columns={'index': 'id'})
```

```
VADERS = VADERS.merge(df, how='left')
```

- `pd.DataFrame(res).T`: Μετατρέπει το λεξικό αποτελεσμάτων σε DataFrame και το μεταθέτει έτσι ώστε οι βαθμολογίες συναισθήματος να είναι στήλες.

- `VADERS.reset_index().rename(columns={'index': 'id'})`: Επαναφέρει τον δείκτη του DataFrame και μετονομάζει τη στήλη `index` σε `id`. Αυτό το βήμα διασφαλίζει ότι τα αναγνωριστικά είναι σωστά ευθυγραμμισμένα με το αρχικό DataFrame.

- `VADERS.merge(df, how='left')`: Συγχωνεύει το DataFrame (`VADERS`) με το αρχικό DataFrame (`df`) με βάση τη στήλη `id`. Αυτό μας επιτρέπει να διατηρήσουμε τα αρχικά δεδομένα μαζί με τις βαθμολογίες συναισθήματος.

6.3.5 Οπτικοποίηση των αποτελεσμάτων της ανάλυσης συναισθήματος

Η οπτικοποίηση είναι ένα κρίσιμο βήμα στην ανάλυση δεδομένων. Βοηθά στην κατανόηση της κατανομής των βαθμολογιών συναισθήματος σε διάφορες κατηγορίες. Δείτε πώς μπορούμε να οπτικοποιήσουμε τα αποτελέσματα χρησιμοποιώντας ραβδογράμματα:

```
ax = sns.barplot(data=VADERS, x='sentiment', y='compound')
```

```
ax.set_title('VADER results')
```

```
plt.show()
```

- `sns.barplot(data=VADERS, x='sentiment', y='compound')`: Αυτή η εντολή δημιουργεί ένα ραβδόγραμμα χρησιμοποιώντας το Seaborn για να απεικονίσει τη μέση σύνθετη βαθμολογία συναισθήματος για κάθε κατηγορία συναισθήματος στο αρχικό σύνολο δεδομένων. Η βαθμολογία `compound` είναι μια ενιαία τιμή που συνοψίζει το συνολικό συναίσθημα του κειμένου.

- `ax.set_title('VADER results')`: Ορίζει τον τίτλο του γραφήματος σε `'VADER results'`.

- `plt.show()`: Εμφανίζει το γράφημα.

6.3.6 Λεπτομερής οπτικοποίηση συνιστωσών συναισθήματος

Για να αποκτήσουμε βαθύτερη εικόνα της κατανομής του συναισθήματος, μπορούμε να δημιουργήσουμε υποδιαγράμματα για κάθε συνιστώσα συναισθήματος (θετική, ουδέτερη και αρνητική):

```
fig, axs = plt.subplots(1, 3, figsize=(15, 5))
```

```
sns.barplot(data=VADERS, x='sentiment', y='pos', ax=axs[0])
```

```
sns.barplot(data=VADERS, x='sentiment', y='neu', ax=axes[1])
sns.barplot(data=VADERS, x='sentiment', y='neg', ax=axes[2])
axes[0].set_title('Positive')
axes[1].set_title('Neutral')
axes[2].set_title('Negative')
plt.show()
```

- `fig, axes = plt.subplots(1, 3, figsize=(15, 5))`: Δημιουργεί ένα σχήμα με τρία υποδιαγράμματα τοποθετημένα σε μία μόνο σειρά, το καθένα μεγέθους 15x5 ίντσες. Αυτή η διάταξη μας επιτρέπει να συγκρίνουμε τα διάφορα συστατικά του συναισθήματος δίπλα-δίπλα.
- `sns.barplot(data=VADERS, x='sentiment', y='pos', ax=axes[0])`: Παρουσιάζει τις μέσες θετικές βαθμολογίες συναισθήματος για κάθε κατηγορία συναισθήματος στο πρώτο υποδιάγραμμα.
- `sns.barplot(data=VADERS, x='sentiment', y='neu', ax=axes[1])`: Παρουσιάζει τις μέσες βαθμολογίες ουδέτερου συναισθήματος για κάθε κατηγορία συναισθήματος στο δεύτερο υποδιάγραμμα.
- `sns.barplot(data=VADERS, x='sentiment', y='neg', ax=axes[2])`: Παρουσιάζει τις μέσες βαθμολογίες αρνητικού συναισθήματος για κάθε κατηγορία συναισθήματος στο τρίτο υποδιάγραμμα.
- `axes[0].set_title('Positive'), axes[1].set_title('Neutral'), axes[2].set_title('Negative')`: Ορίζει τους τίτλους για κάθε υποδιάγραμμα για να υποδεικνύει τον τύπο της βαθμολογίας συναισθήματος που απεικονίζεται.
- `plt.show()`: Εμφανίζει το σχήμα με τα τρία υποδιαγράμματα.

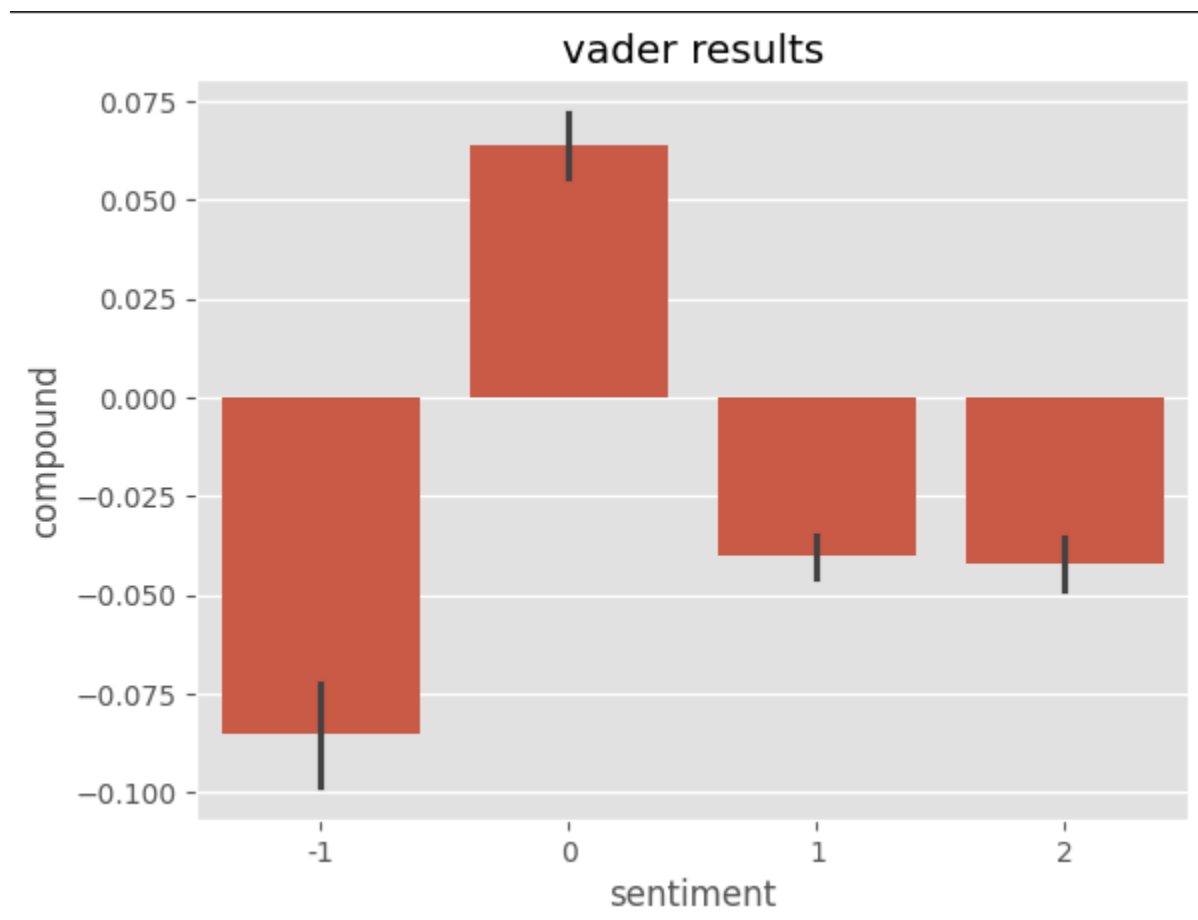
6.3.7 Συμπέρασμα

Αυτός ο κώδικας υλοποιεί την ανάλυση συναισθήματος VADER σε ένα σύνολο δεδομένων από μηνύματα κειμένου. Ξεκινά με την εισαγωγή των απαραίτητων βιβλιοθηκών και τη λήψη του λεξικού VADER. Αφού αρχικοποιήσει τον αναλυτή συναισθήματος, το σενάριο προχωρά στην επανάληψη του συνόλου δεδομένων, υπολογίζοντας βαθμολογίες συναισθήματος για κάθε μήνυμα. Τα αποτελέσματα αποθηκεύονται σε ένα λεξικό, μετατρέπονται σε DataFrame και συγχωνεύονται με το αρχικό σύνολο δεδομένων. Τέλος, τα αποτελέσματα της ανάλυσης συναισθήματος οπτικοποιούνται με τη χρήση ραβδογραμμάτων, τα οποία δείχνουν την κατανομή των θετικών, ουδέτερων και αρνητικών βαθμολογιών συναισθήματος σε διάφορες κατηγορίες συναισθήματος.

Το εργαλείο ανάλυσης συναισθήματος VADER είναι ιδιαίτερα αποτελεσματικό για την ανάλυση σύντομων, ανεπίσημων κειμένων, όπως αυτά που βρίσκονται στα μέσα κοινωνικής δικτύωσης. Το λεξικό του και η προσέγγιση που βασίζεται σε κανόνες το καθιστούν ιδιαίτερα ικανό στην καταγραφή των αποχρώσεων της γλώσσας των μέσων κοινωνικής δικτύωσης, συμπεριλαμβανομένης της χρήσης της αργκό, των emoticons και άλλων ανεπίσημων εκφράσεων. Ακολουθώντας τα βήματα που περιγράφονται σε αυτόν τον οδηγό, μπορείτε να εφαρμόσετε την ανάλυση συναισθήματος VADER στα δικά σας σύνολα δεδομένων, αποκτώντας πολύτιμες πληροφορίες για τον συναισθηματικό τόνο και την κατανομή του συναισθήματος των δεδομένων κειμένου.

Αυτή η ολοκληρωμένη προσέγγιση της ανάλυσης συναισθήματος παρέχει ένα ισχυρό πλαίσιο για την κατανόηση του συναισθηματικού τοπίου των συνομιλιών στα μέσα κοινωνικής δικτύωσης, των αξιολογήσεων πελατών και άλλων μορφών δεδομένων κειμένου. Είτε είστε επιστήμονας δεδομένων, επιχειρηματικός αναλυτής ή ερευνητής, το VADER προσφέρει ένα ισχυρό εργαλείο για την ανάλυση συναισθήματος, βοηθώντας σας να αποκαλύψετε τάσεις, μοτίβα και ιδέες που μπορούν να ενημερώσουν για τη λήψη αποφάσεων και να προωθήσουν στρατηγικές πρωτοβουλίες.

6.3.8 results



Η εικόνα δείχνει τα αποτελέσματα από μια ανάλυση συναισθήματος σε tweets σχετικά με την κλιματική αλλαγή με τη χρήση ενός μοντέλου VADER. Ας αναλύσουμε το διάγραμμα:

- Άξονας X (συναίσθημα): Αυτός ο άξονας αντιπροσωπεύει τις κατηγορίες συναισθήματος:

- -1: Αντι (tweets που δεν πιστεύουν στην ανθρωπογενή κλιματική αλλαγή)
- 0: Ουδέτερα (tweets που ούτε υποστηρίζουν ούτε αντικρούουν την πίστη στην ανθρωπογενή κλιματική αλλαγή)
- 1: Pro (tweets που υποστηρίζουν την πεποίθηση στην ανθρωπογενή κλιματική αλλαγή)
- 2: Νέα (tweets που παραπέμπουν σε πραγματικές ειδήσεις σχετικά με την κλιματική αλλαγή)

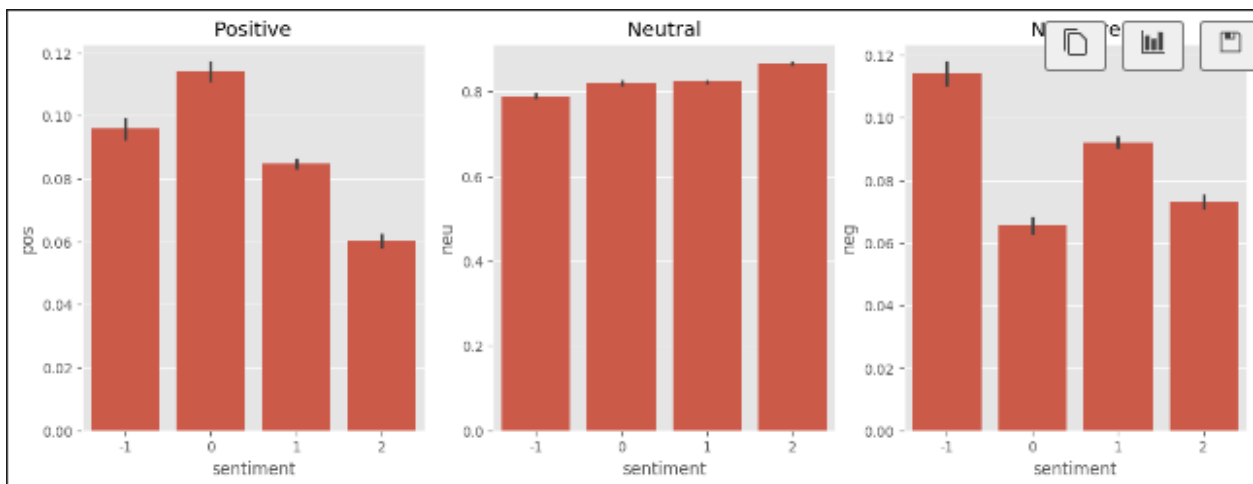
- Άξονας Y (σύνθετος): Αυτός ο άξονας αντιπροσωπεύει τη σύνθετη βαθμολογία από την ανάλυση συναισθήματος VADER. Η σύνθετη βαθμολογία κυμαίνεται από -1 έως 1 και υποδηλώνει το συνολικό συναίσθημα του κειμένου:

- Οι αρνητικές τιμές υποδηλώνουν αρνητικό συναίσθημα.
- Οι θετικές τιμές υποδηλώνουν θετικό συναίσθημα.

- Τιμές κοντά στο μηδέν υποδηλώνουν ουδέτερο συναίσθημα.
- Μπάρες: Το ύψος κάθε μπάρας αντιπροσωπεύει τη μέση σύνθετη βαθμολογία για τα tweets σε κάθε κατηγορία συναισθήματος. Υπάρχουν επίσης ράβδοι σφάλματος, που δείχνουν τη μεταβλητότητα ή τα διαστήματα εμπιστοσύνης γύρω από αυτούς τους μέσους όρους.

Ερμηνεία:

- Αντι (-1): Αυτή η κατηγορία έχει αρνητική μέση σύνθετη βαθμολογία, υποδεικνύοντας ότι τα tweets αυτής της κατηγορίας έχουν γενικά αρνητικό συναίσθημα.
- Ουδέτερη (0): Αυτή η κατηγορία έχει ελαφρώς θετική μέση σύνθετη βαθμολογία, υποδηλώνοντας ότι τα ουδέτερα tweets τείνουν να τείνουν λίγο πιο θετικά συνολικά.
- Pro (1): Αυτή η κατηγορία έχει ελαφρώς αρνητική μέση σύνθετη βαθμολογία, υποδεικνύοντας ένα ελαφρώς αρνητικό συναίσθημα ακόμη και σε tweets που υποστηρίζουν την ανθρωπογενή κλιματική αλλαγή. Ωστόσο, το αρνητικό συναίσθημα είναι λιγότερο έντονο σε σύγκριση με την κατηγορία Anti.
- Νέα (2): Αυτή η κατηγορία έχει επίσης ελαφρώς αρνητική μέση σύνθετη βαθμολογία, γεγονός που θα μπορούσε να υποδηλώνει ότι τα tweets με ειδήσεις για την κλιματική αλλαγή τείνουν να έχουν ελαφρώς αρνητικό τόνο.



Το δεύτερο γράφημα παρέχει μια πιο λεπτομερή ανάλυση των βαθμολογιών συναισθήματος στις διάφορες κατηγορίες συναισθήματος. Το γράφημα χωρίζεται σε τρία επιμέρους διαγράμματα, καθένα από τα οποία αντιπροσωπεύει έναν διαφορετικό τύπο βαθμολογίας συναισθήματος: Θετικό, ουδέτερο και αρνητικό. Ακολουθεί μια λεπτομερής ανάλυση κάθε υποδιαγράμματος:

	Θετικό συναίσθημα	Ουδέτερο	Αρνητικό
Anti	0,7	0,8	0,12
Neutral	0,11	0,8	0,04
Pro	0,09	0,8	0,05
News	0,06	0,84	0,06

6.4 Εφαρμοσμένες μέθοδοι: RoBERTa

Το RoBERTa είναι ένα σύγχρονο μοντέλο βασισμένο σε μετασχηματιστές που βασίζεται στο BERT (Bidirectional Encoder Representations from Transformers). Ενισχύει το BERT εκπαιδεύοντας με μεγαλύτερες μίνι παρτίδες και ένα πιο εκτεταμένο σώμα εκπαίδευσης, χρησιμοποιώντας δυναμικά μοτίβα κάλυψης και αφαιρώντας τον στόχο πρόβλεψης της επόμενης πρότασης. Αυτές οι βελτιώσεις επιτρέπουν στο RoBERTa να συλλαμβάνει βαθύτερες πληροφορίες σχετικά με το περιεχόμενο και να χειρίζεται πιο αποτελεσματικά τις πολυπλοκότητες της φυσικής γλώσσας.

1. Προεπεξεργασία για το RoBERTa:

- Καθαρισμός κειμένου: Τα ακατέργαστα tweets καθαρίστηκαν για να αφαιρεθούν οι περιττοί χαρακτήρες, οι διευθύνσεις URL, τα hashtags και οι αναφορές. Αυτό το βήμα διασφάλισε ότι η είσοδος στο μοντέλο RoBERTa ήταν απαλλαγμένη από θόρυβο που θα μπορούσε να επηρεάσει αρνητικά την απόδοσή του.

- tokenization: Το RoBERTa χρησιμοποιεί έναν tokenizer κωδικοποίησης ζεύγους byte (BPE), ο οποίος χωρίζει το κείμενο σε μονάδες υπολέξεων. Αυτή η μέθοδος επιτρέπει στο μοντέλο να χειρίζεται σπάνιες και εκτός λεξιλογίου λέξεις, διασπώντας τις σε πιο συχνές υπολεκτικές μονάδες.

- Παράταση και αποκοπή: Τα tweets παραπληρώθηκαν ή αποκόπηκαν σε σταθερό μήκος, εξασφαλίζοντας ομοιόμορφο μέγεθος εισόδου για το μοντέλο. Αυτό το βήμα είναι ζωτικής σημασίας για την αποτελεσματική επεξεργασία δέσμης σε μοντέλα βαθιάς μάθησης.

2. Λεπτομερής ρύθμιση του RoBERTa:

- Επισημάνσεις δεδομένων: Τα tweets επισημάνθηκαν με κατηγορίες συναισθήματος (θετικό, αρνητικό, ουδέτερο) με βάση προϋπάρχοντα λεξικά συναισθήματος και χειροκίνητες επισημάνσεις.

- Εκπαίδευση του μοντέλου: Το RoBERTa συντονίστηκε με βάση το σύνολο δεδομένων με τις ετικέτες. Η λεπτομερής ρύθμιση περιελάμβανε την προσαρμογή των βαρών του

προ-εκπαιδευμένου μοντέλου χρησιμοποιώντας τα συγκεκριμένα δεδομένα εργασίας (tweets με ετικέτες συναισθημάτων) για να βελτιωθεί η απόδοσή του στην ταξινόμηση συναισθημάτων.

- Εξαγωγή χαρακτηριστικών: Οι τελικές κρυφές καταστάσεις του μοντέλου RoBERTa χρησιμοποιήθηκαν για την εξαγωγή χαρακτηριστικών. Αυτά τα χαρακτηριστικά αντιπροσωπεύουν τις ενσωματώσεις των tweets, καταγράφοντας σύνθετες σχέσεις και αποχρώσεις συναισθήματος εντός του κειμένου.

3. Επιλογή χαρακτηριστικών με το RoBERTa:

- Ανάλυση ενσωμάτωσης: Οι ενσωματώσεις που παρήγαγε το RoBERTa αναλύθηκαν για να εντοπιστούν τα πιο σημαντικά χαρακτηριστικά που συμβάλλουν στην ταξινόμηση συναισθήματος. Με την εξέταση των βαρών προσοχής και των κρυφών καταστάσεων, εντοπίστηκαν σημαντικά tokens και φράσεις.

- Μείωση διαστάσεων: Τεχνικές όπως η Ανάλυση Κύριων Συνιστωσών (PCA) εφαρμόστηκαν στα εξαγόμενα χαρακτηριστικά για τη μείωση της διαστατικότητας, διατηρώντας παράλληλα τις πιο κατατοπιστικές συνιστώσες. Το βήμα αυτό βοήθησε στην απλοποίηση του χώρου των χαρακτηριστικών και στη βελτίωση της ερμηνευσιμότητας και της απόδοσης του μοντέλου.

6.4.1 Εισαγωγή βιβλιοθηκών και αρχικοποίηση του μοντέλου

```
import torch
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from scipy.special import softmax
import pandas as pd
from tqdm import tqdm

MODEL = f"cardiffnlp/twitter-roberta-base-sentiment"
```

- torch: χρησιμοποιείται εδώ για λειτουργίες βαθιάς μάθησης.

- transformers: Μια βιβλιοθήκη της Hugging Face που παρέχει σύγχρονα προ-εκπαιδευμένα μοντέλα για εργασίες NLP. Χρησιμοποιούμε τα `AutoTokenizer` και `AutoModelForSequenceClassification` για να φορτώσουμε το προ-εκπαιδευμένο μοντέλο RoBERTa.

- scipy.special.softmax: Η συνάρτηση softmax χρησιμοποιείται για τη μετατροπή των ακατέργαστων βαθμολογιών του μοντέλου σε πιθανότητες.

- pandas: Μια ισχυρή βιβλιοθήκη χειρισμού δεδομένων που χρησιμοποιείται για το χειρισμό του συνόλου δεδομένων.

- tqdm: Μια βιβλιοθήκη για την εμφάνιση ράβδων προόδου, χρήσιμη για την παρακολούθηση της προόδου των βρόχων.

6.4.2 Φόρτωση του Tokenizer και του μοντέλου

```
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

- `AutoTokenizer.from_pretrained(MODEL)`: Φορτώνει τον tokenizer που σχετίζεται με το προ-εκπαιδευμένο μοντέλο RoBERTa. Ο tokenizer μετατρέπει το κείμενο σε αναγνωριστικά token που μπορεί να επεξεργαστεί το μοντέλο.
- `AutoModelForSequenceClassification.from_pretrained(MODEL)`: Φορτώνει το προ-εκπαιδευμένο μοντέλο RoBERTa, ειδικά ρυθμισμένο για εργασίες ταξινόμησης ακολουθιών, όπως η ανάλυση συναισθήματος.

6.4.3 Κωδικοποίηση κειμένου και λήψη βαθμολογίας συναισθήματος

```
# run for roberta
encoded_text = tokenizer(example, return_tensors='pt')
output = model(encoded_text)
scores = output[0][0].detach().numpy()
scores = softmax(scores)
scores_dict = {
    'roberta_neg': scores[0],
    'roberta_neu': scores[1],
    'roberta_pos': scores[2]
}

print(scores_dict)
print(example)
```

- `tokenizer(example, return_tensors='pt')`: Κωδικοποιεί το κείμενο εισόδου ('example') σε token IDs και επιστρέφει έναν τανυστή PyTorch. Αυτός ο τανυστής είναι η μορφή εισόδου που αναμένεται από το μοντέλο.

- `model(encoded_text)`: Εκτελεί το κωδικοποιημένο κείμενο μέσω του μοντέλου RoBERTa. Το μοντέλο εξάγει ακατέργαστες βαθμολογίες για κάθε κλάση (αρνητική, ουδέτερη, θετική).
- `output[0][0].detach().numpy()`: Εξάγει τις βαθμολογίες από την έξοδο του μοντέλου, τις αποσυνδέει από το γράφημα υπολογισμού (αφού δεν χρειάζεται να υπολογίσουμε κλίσεις) και τις μετατρέπει σε πίνακα NumPy.
- `softmax(scores)`: Εφαρμόζει τη συνάρτηση softmax για τη μετατροπή των ακατέργαστων βαθμολογιών σε πιθανότητες.
- `scores_dict`: Ένα λεξικό που αποθηκεύει τις πιθανότητες για κάθε κατηγορία συναισθήματος (αρνητική, ουδέτερη, θετική).

6.4.4 Εκτέλεση του μοντέλου για κάθε μήνυμα στο σύνολο δεδομένων

```
res = {}

for i, row in tqdm(df.iterrows(), total=len(df)):
    try:
        text = row['message']
        myid = row['id']
        VADER_results = sia.polarity_scores(text)
        VADER_results_rename = {}
        for key, value in VADER_results.items():
            VADER_results_rename[f"VADER_{key}"] = value

        roberta_results = polarity_scores_roberta(text)
        both = {roberta_results, VADER_results_rename}
        res[myid] = both
    except RuntimeError:
        print(f'broke for id {myid}')
```

- `tqdm(df.iterrows(), total=len(df))`: Επαναλαμβάνει κάθε γραμμή στο DataFrame `df` με μια μπάρα προόδου που δείχνει το συνολικό αριθμό των επαναλήψεων.
- `text = row['message']`: Εξάγει το μήνυμα κειμένου από την τρέχουσα γραμμή.
- `myid = row['id']`: Εξάγει το μοναδικό αναγνωριστικό από την τρέχουσα γραμμή.

- `VADER_results = sia.polarity_scores(text)`: Υπολογίζει τις βαθμολογίες συναισθήματος VADER για το κείμενο.
- `VADER_results_rename`: Ένα λεξικό για την αποθήκευση των βαθμολογιών συναισθήματος VADER με μετονομασμένα κλειδιά (με πρόθεμα 'VADER_') για να διακρίνεται από τις βαθμολογίες RoBERTa.
- `polarity_scores_roberta(text)`: Καλεί μια συνάρτηση (δεν εμφανίζεται στον παρεχόμενο κώδικα) για τον υπολογισμό των βαθμολογιών συναισθήματος RoBERTa για το κείμενο.
- `both = {roberta_results, VADER_results_rename}`: Συγχωνεύει τα αποτελέσματα RoBERTa και VADER σε ένα ενιαίο λεξικό.
- `res[myid] = both`: Αποθηκεύει τα συνδυασμένα αποτελέσματα στο λεξικό 'res' χρησιμοποιώντας το μοναδικό αναγνωριστικό ως κλειδί.
- `except RuntimeError`: Συλλαμβάνει τυχόν σφάλματα χρόνου εκτέλεσης που μπορεί να προκύψουν κατά τη διάρκεια του βρόχου, εκτυπώνει ένα μήνυμα με το προβληματικό ID.

6.4.5 Δημιουργία ενός πλαισίου δεδομένων αποτελεσμάτων

```
results_df = pd.DataFrame(res).T
results_df = results_df.reset_index().rename(columns={'index': 'id'})
results_df = results_df.merge(df, how='left')
```

- `pd.DataFrame(res).T`: Μετατρέπει το λεξικό των αποτελεσμάτων σε DataFrame και το μεταθέτει έτσι ώστε οι βαθμολογίες συναισθήματος να είναι στήλες.
- `results_df.reset_index().rename(columns={'index': 'id'})`: Επαναφέρει το ευρετήριο και μετονομάζει τη στήλη του ευρετηρίου σε 'id' για να διατηρήσει τη συνοχή με το αρχικό DataFrame.
- `results_df.merge(df, how='left')`: Συγχωνεύει το DataFrame των αποτελεσμάτων με το αρχικό DataFrame 'df' με βάση τη στήλη 'id'. Αυτό το βήμα διασφαλίζει ότι διατηρούμε όλα τα αρχικά δεδομένα μαζί με τις υπολογισμένες βαθμολογίες συναισθήματος.

6.4.6 Σύνοψη

Αυτή η υλοποίηση παρέχει μια ολοκληρωμένη ροή εργασίας για την ανάλυση συναισθήματος με χρήση του μοντέλου RoBERTa. Ακολουθεί μια λεπτομερής ανάλυση της διαδικασίας:

1. Εισαγωγές βιβλιοθήκης και λήψεις πόρων: Ο κώδικας ξεκινάει με την εισαγωγή των απαραίτητων βιβλιοθηκών, συμπεριλαμβανομένων του PyTorch, της βιβλιοθήκης Transformers

του Hugging Face και του εργαλείου ανάλυσης συναισθήματος VADER από το NLTK. Το λεξικό VADER μεταφορτώνεται για να καταστεί δυνατή η ανάλυση συναισθήματος με το VADER.

2. Αρχικοποίηση μοντέλου: Το μοντέλο RoBERTa και ο tokenizer φορτώνονται χρησιμοποιώντας τις κλάσεις `AutoTokenizer` και `AutoModelForSequenceClassification` του Hugging Face. Αυτό το βήμα ρυθμίζει το μοντέλο για να επεξεργάζεται κείμενο και να εξάγει βαθμολογίες συναισθήματος.

3. Κωδικοποίηση κειμένου και βαθμολόγηση συναισθήματος: Το κείμενο εισόδου κωδικοποιείται σε αναγνωριστικά συμβόλων χρησιμοποιώντας τον tokenizer, ο οποίος το προετοιμάζει για την είσοδο στο μοντέλο RoBERTa. Η έξοδος του μοντέλου, η οποία αποτελείται από ακατέργαστες βαθμολογίες συναισθήματος, επεξεργάζεται μέσω της συνάρτησης softmax για τη δημιουργία πιθανοτήτων για κάθε κατηγορία συναισθήματος (αρνητική, ουδέτερη, θετική). Αυτές οι πιθανότητες αποθηκεύονται σε ένα λεξικό για εύκολη πρόσβαση.

4. Επαναληπτική επεξεργασία του συνόλου δεδομένων: Ο κώδικας επαναλαμβάνει κάθε γραμμή στο πλαίσιο δεδομένων εισόδου DataFrame, εξάγοντας το κείμενο του μηνύματος και το μοναδικό αναγνωριστικό του. Υπολογίζει βαθμολογίες συναισθήματος χρησιμοποιώντας τόσο τα μοντέλα VADER όσο και τα μοντέλα RoBERTa. Τα αποτελέσματα και από τα δύο μοντέλα συνδυάζονται σε ένα ενιαίο λεξικό και αποθηκεύονται σε ένα λεξικό αποτελεσμάτων με κλειδί το ID του μηνύματος.

5. Χειρισμός σφαλμάτων: Το μπλοκ `try-except` εξασφαλίζει ότι ο βρόχος συνεχίζεται ακόμη και αν προκύψει σφάλμα κατά τη διάρκεια της ανάλυσης συναισθήματος ενός συγκεκριμένου μηνύματος. Αυτό είναι ζωτικής σημασίας για μεγάλα σύνολα δεδομένων, όπου περιστασιακά σφάλματα θα μπορούσαν διαφορετικά να διακόψουν ολόκληρη τη διαδικασία.

6. Δημιουργία και συγχώνευση πλαισίων δεδομένων αποτελεσμάτων: Το λεξικό αποτελεσμάτων μετατρέπεται σε DataFrame, με τις βαθμολογίες συναισθήματος ως στήλες. Στη συνέχεια, το DataFrame συγχωνεύεται με το αρχικό DataFrame για να διατηρηθούν όλα τα αρχικά δεδομένα μαζί με τις υπολογισμένες βαθμολογίες συναισθήματος.

7. Τελική έξοδος: Το τελικό DataFrame, `results_df`, περιέχει τόσο τα αρχικά δεδομένα όσο και τα αποτελέσματα της ανάλυσης συναισθήματος από το VADER και το RoBERTa. Αυτό το εμπλουτισμένο σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για περαιτέρω ανάλυση, οπτικοποίηση ή μεταγενέστερες εργασίες.

6.4.7 Συμπέρασμα

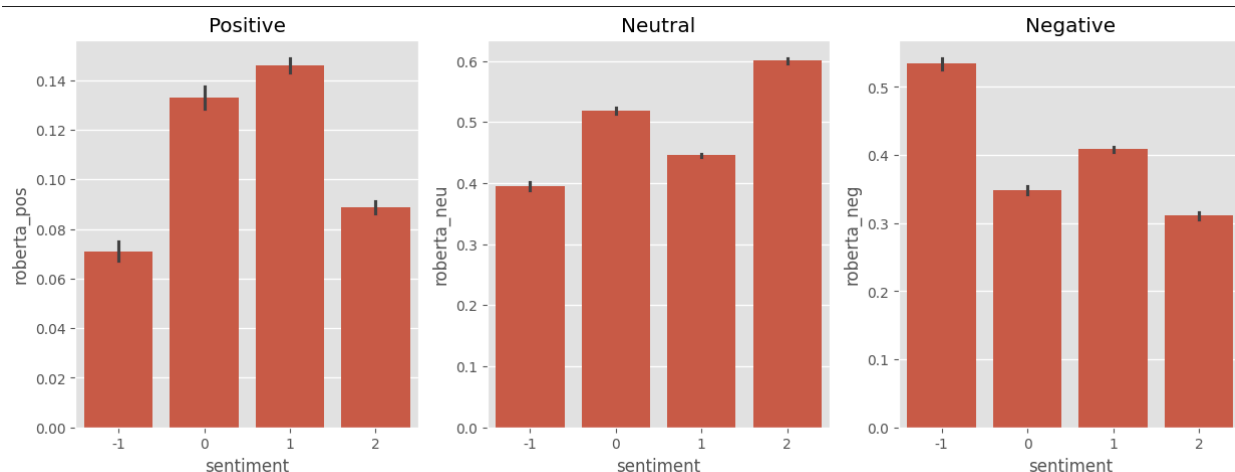
Στην παρούσα μελέτη, δύο προηγμένες μεθοδολογίες, η RoBERTa (A Robustly Optimized BERT Pretraining Approach) και η VADER (Valence Aware Dictionary for sEntiment Reasoning), χρησιμοποιήθηκαν για την ανάλυση συναισθήματος σε tweets που σχετίζονται με την κλιματική αλλαγή. Ο στόχος ήταν να αξιοποιηθούν αυτά τα μοντέλα ξεχωριστά για την

ενίσχυση της επιλογής χαρακτηριστικών και τη βελτίωση της συνολικής ακρίβειας της ταξινόμησης συναισθήματος. Η παρούσα ενότητα περιγράφει λεπτομερώς τις συγκεκριμένες μεθόδους και διαδικασίες που εφαρμόστηκαν με τη χρήση των RoBERTa και VADER, εστιάζοντας στους ρόλους τους στην επιλογή χαρακτηριστικών και στην ανάλυση συναισθήματος.

Τόσο το RoBERTa όσο και το VADER προσφέρουν ισχυρές μεθόδους για την επιλογή χαρακτηριστικών στην ανάλυση συναισθήματος, η καθεμία με μοναδικά πλεονεκτήματα. Το RoBERTa, με την αρχιτεκτονική του που βασίζεται στον Transformer, υπερέχει στην καταγραφή βαθιών πληροφοριών σχετικά με το πλαίσιο και αποχρώσεων των σχέσεων συναισθήματος εντός του κειμένου. Παρέχει πλούσια ενσωμάτωση που, μετά τη μείωση της διαστατικότητας, χρησιμεύει ως εξαιρετικά κατατοπιστικά χαρακτηριστικά για την ταξινόμηση συναισθήματος. Το VADER, από την άλλη πλευρά, χρησιμοποιεί ένα λεξικό και μια προσέγγιση βασισμένη σε κανόνες που εντοπίζει και βαθμολογεί αποτελεσματικά λέξεις και φράσεις με συναισθήματα, καθιστώντας το ιδιαίτερα αποτελεσματικό για το χειρισμό της γλώσσας των μέσων κοινωνικής δικτύωσης.

Με την εφαρμογή αυτών των μεθόδων ξεχωριστά, η παρούσα μελέτη κατάφερε να ενισχύσει την επιλογή χαρακτηριστικών και να βελτιώσει την ακρίβεια ταξινόμησης συναισθήματος. Οι ενσωματώσεις της RoBERTa παρείχαν μια βαθιά, συνειδητή κατανόηση του πλαισίου του κειμένου, ενώ οι βαθμολογίες συναισθήματος της VADER προσέφεραν απλά, ερμηνεύσιμα χαρακτηριστικά. Τα βήματα preprocessing κάθε μοντέλου -καθαρισμός, tokenization και συγκεκριμένες τεχνικές εξαγωγής χαρακτηριστικών- ήταν κρίσιμα για την προετοιμασία των δεδομένων για ανάλυση.

6.4.8 roberta αποτελέσματα



Το τρίτο διάγραμμα παρέχει τα αποτελέσματα της ανάλυσης συναισθήματος με τη χρήση του μοντέλου RoBERTa. Παρόμοια με το προηγούμενο διάγραμμα, χωρίζεται σε τρία

υποδιαγράμματα που αντιπροσωπεύουν τις βαθμολογίες θετικού, ουδέτερου και αρνητικού συναισθήματος. Ακολουθεί μια λεπτομερής ανάλυση:

	Θετικό συναίσθημα	Ουδέτερο	Αρνητικό
Anti	0,08	0,4	0,5
Neutral	0,13	0,5	0,3
Pro	0,15	0,45	0,35
News	0.10	0,6	0,3

- Θετικό συναίσθημα: Τα tweets υπέρ έχουν το υψηλότερο θετικό συναίσθημα, ενώ τα tweets κατά έχουν το χαμηλότερο.
- Ουδέτερο συναίσθημα: Τα tweets με ειδήσεις είναι τα πιο ουδέτερα, ενώ τα tweets με αντιδράσεις είναι τα λιγότερο ουδέτερα.
- Αρνητικό συναίσθημα: Τα Anti tweets έχουν το υψηλότερο αρνητικό συναίσθημα, ενώ τα News και Neutral tweets έχουν το χαμηλότερο.

6.5 Σύγκριση των αποτελεσμάτων του RoBERTa με τα αποτελέσματα του VADER:

- Το RoBERTa παρουσιάζει υψηλότερες βαθμολογίες θετικού συναισθήματος για τα Pro tweets και υψηλότερες βαθμολογίες ουδέτερου συναισθήματος για τα News tweets σε σύγκριση με το VADER.
- Το αρνητικό συναίσθημα είναι υψηλότερο για τα Anti tweets και στα δύο μοντέλα, αλλά οι βαθμολογίες είναι γενικά υψηλότερες στο RoBERTa.
- Το RoBERTa παρουσιάζει σαφέστερη διάκριση μεταξύ των κατηγοριών συναισθήματος, ιδίως για τα Pro και News tweets.

Βέβαια, ας εμβαθύνουμε με μια σύγκριση μεταξύ των αποτελεσμάτων των μοντέλων VADER και RoBERTa για την ανάλυση συναισθήματος σε tweets σχετικά με την κλιματική αλλαγή.

Συνολική δομή και μετρικές:

- Και τα δύο διαγράμματα χωρίζονται σε τρία επιμέρους διαγράμματα: Θετικές, ουδέτερες και αρνητικές βαθμολογίες συναισθήματος.

Θετικό συναίσθημα:

	VADER	RoBERTa
Pro	0,09	0.15
Anti	0,07	0,08
Neutral	0,11	0.13
News	0,06	0.10

Σύγκριση:

- Και τα δύο μοντέλα δείχνουν το χαμηλότερο θετικό συναίσθημα για τα tweets Anti.
- Το VADER δείχνει το υψηλότερο θετικό συναίσθημα στα ουδέτερα tweets, ενώ το RoBERTa δείχνει το υψηλότερο στα Pro tweets.
- Το RoBERTa αποδίδει γενικά υψηλότερες βαθμολογίες θετικού συναισθήματος σε όλες τις κατηγορίες σε σύγκριση με το VADER.

Ουδέτερο συναίσθημα:

	VADER	RoBERTa
Pro	0,83	0,45
Anti	0,8	0,4
Neutral	0,85	0,5
News	0,82	0,6

Σύγκριση:

- VADER παρουσιάζει ομοιόμορφα υψηλές βαθμολογίες ουδέτερου συναισθήματος σε όλες τις κατηγορίες (περίπου 0,8-0,85).

- Το RoBERTa εμφανίζει μεγαλύτερη διακύμανση, με τα tweets με ειδήσεις να έχουν την υψηλότερη ουδέτερη βαθμολογία (0,6) και τα tweets με αντιδράσεις τη χαμηλότερη (0,4).
- Αυτό υποδηλώνει ότι το VADER αντιμετωπίζει τα tweets πιο ομοιογενώς ουδέτερα, ενώ το RoBERTa καταγράφει περισσότερες αποχρώσεις στην ουδετερότητα.

Αρνητικό συναίσθημα:

	VADER	RoBERTa
Pro	0,05	0,5
Anti	0,12	0,5
Neutral	0,04	0,3
News	0,06	0,3

Σύγκριση:

- Και τα δύο μοντέλα συμφωνούν ότι τα Anti tweets έχουν το υψηλότερο αρνητικό συναίσθημα.
- Οι αρνητικές βαθμολογίες του VADER είναι γενικά χαμηλότερες σε όλες τις κατηγορίες σε σύγκριση με το RoBERTa.
- Το RoBERTa αποδίδει πολύ υψηλότερο αρνητικό συναίσθημα στα Anti tweets (0,5) από το VADER (0,12), υποδεικνύοντας ισχυρότερη ανίχνευση αρνητικού συναισθήματος.

Συμπέρασμα:

- Το RoBERTa προσφέρει μια πιο διαφοροποιημένη και διαφοροποιημένη ανάλυση συναισθήματος σε σύγκριση με το VADER. Καταγράφει ισχυρότερα θετικά συναισθήματα σε tweets Pro και ισχυρότερα αρνητικά συναισθήματα σε tweets Anti.
- Το VADER τείνει να υπολογίζει περισσότερο τον μέσο όρο των συναισθημάτων, με αποτέλεσμα υψηλές ουδέτερες βαθμολογίες και λιγότερο έντονες διαφορές μεταξύ θετικών και αρνητικών συναισθημάτων.
- Ανάλογα με τις απαιτήσεις (γρήγορη, βασισμένη σε κανόνες ανάλυση έναντι λεπτομερούς ανάλυσης με επίγνωση του πλαισίου), το ένα μοντέλο μπορεί να είναι καταλληλότερο από το άλλο.

7.5.1 Ερμηνεία του Pairplot



Το pairplot είναι ένα ισχυρό εργαλείο οπτικοποίησης που σας επιτρέπει να παρατηρήσετε τη σχέση μεταξύ πολλαπλών μεταβλητών σε ένα σύνολο δεδομένων. Σε αυτή την περίπτωση, το pairplot που παρέχετε περιλαμβάνει τόσο τις βαθμολογίες συναισθήματος RoBERTa όσο και τις βαθμολογίες συναισθήματος VADER κατηγοριοποιημένες με διαφορετικές ετικέτες συναισθήματος. Η κατανόηση αυτής της απεικόνισης απαιτεί μια λεπτομερή ματιά σε κάθε στοιχείο και σε αυτό που αντιπροσωπεύει.

Συνιστώσες του Pairplot

1. Διαγώνια γραφήματα:

-Διάγραμμα πυκνότητας: Αυτά τα διαγράμματα δείχνουν την κατανομή κάθε βαθμολογίας συναισθήματος. Για παράδειγμα, το διαγώνιο διάγραμμα για το `roberta\_neg` δείχνει την πυκνότητα των αρνητικών βαθμολογιών που αποδίδονται από το μοντέλο RoBERTa σε όλα τα μηνύματα. Οι κορυφές σε αυτά τα διαγράμματα υποδεικνύουν τις πιο κοινές βαθμολογίες συναισθήματος και η διασπορά μπορεί να υποδεικνύει πόσο διαφορετικές είναι οι βαθμολογίες συναισθήματος.

-ιστογράμματα: Εναλλακτικά, ορισμένα διαγράμματα ζεύγους χρησιμοποιούν ιστογράμματα αντί για διαγράμματα πυκνότητας. Αυτά τα ιστογράμματα αναπαριστούν τη συχνότητα των βαθμολογιών συναισθήματος σε διαφορετικά bins, δίνοντας μια πιο διακριτή εικόνα της κατανομής.

2. Διάγραμμα διασποράς:

- Διάγραμμα διασποράς εκτός διαγωνίου: Αυτά τα διαγράμματα παρουσιάζουν σχέσεις ανά ζεύγη μεταξύ διαφορετικών βαθμολογιών συναισθήματος. Για παράδειγμα, το διάγραμμα διασποράς μεταξύ `roberta\_neg` και `roberta\_neu` δείχνει πώς οι αρνητικές και ουδέτερες βαθμολογίες από το RoBERTa σχετίζονται μεταξύ τους. Τα σημεία σε αυτά τα διαγράμματα μπορούν να αποκαλύψουν συσχετίσεις, συστάδες ή μοτίβα που δείχνουν πώς μια βαθμολογία συναισθήματος επηρεάζει ή σχετίζεται με μια άλλη.

- Χρωματική κωδικοποίηση: Κάθε σημείο στα διαγράμματα διασποράς είναι χρωματικά κωδικοποιημένο με βάση την ετικέτα "συναίσθημα", η οποία μπορεί να αντιπροσωπεύει διαφορετικές κατηγορίες συναισθήματος (π.χ. θετικό, ουδέτερο, αρνητικό). Αυτή η κωδικοποίηση βοηθά στην οπτική διαφοροποίηση του τρόπου με τον οποίο οι διάφορες κατηγορίες συναισθήματος κατανέμονται στις βαθμολογίες συναισθήματος και από τα δύο μοντέλα.

6.5.2 Ανάλυση σχέσεων και κατανομών

1. Κατανόηση των κατανομών συναισθήματος:

- RoBERTa Αρνητικό συναίσθημα (`roberta\_neg`): Το διάγραμμα πυκνότητας για το `roberta\_neg` μπορεί να δείξει πόσο συχνά το μοντέλο αποδίδει αρνητικές βαθμολογίες συναισθήματος. Μια υψηλή κορυφή κοντά στο κατώτερο άκρο της κλίμακας υποδηλώνει ότι τα περισσότερα κείμενα δεν θεωρούνται πολύ αρνητικά.

- VADER Αρνητικό συναίσθημα (`VADER\_neg`): Ομοίως, το διάγραμμα πυκνότητας για το `VADER\_neg` δείχνει την κατανομή των βαθμολογιών αρνητικού συναισθήματος σύμφωνα με το VADER. Η σύγκριση με το `roberta\_neg` μπορεί να αναδείξει τις διαφορές στον τρόπο με τον οποίο τα δύο μοντέλα αντιλαμβάνονται την αρνητικότητα στα κείμενα.

2. Εξέταση των σχέσεων ανά ζεύγη:

- Βαθμολογίες του RoBERTa έναντι του VADER: Τα διαγράμματα διασποράς που συγκρίνουν τα ``roberta_neg`` και ``VADER_neg`` μπορούν να αποκαλύψουν κατά πόσο τα δύο μοντέλα συμφωνούν ως προς το αρνητικό συναίσθημα. Μια θετική συσχέτιση εδώ (σημεία που σχηματίζουν διαγώνια γραμμή) υποδηλώνει ότι όταν το RoBERTa αποδίδει υψηλή αρνητική βαθμολογία, το ίδιο κάνει και το VADER.

- Συσχετίσεις ουδέτερου κλίματος: Η σύγκριση του ``roberta_neu`` με το ``VADER_neu`` μπορεί να υποδείξει αν τα μοντέλα αντιλαμβάνονται με παρόμοιο τρόπο την ουδετερότητα. Οι αποκλίσεις ή η έλλειψη συσχέτισης μπορεί να υποδηλώνουν ότι ένα μοντέλο τείνει περισσότερο να αποδίδει ουδέτερες βαθμολογίες.

3. Κατανομή κατηγοριών συναισθήματος:

- Διαπιστώσεις σχετικά με την κωδικοποίηση χρωμάτων: Σημεία χρωματισμένα με κατηγορίες συναισθήματος (π.χ. θετικά, ουδέτερα, αρνητικά) βοηθούν στην κατανόηση του τρόπου με τον οποίο οι κατηγορίες αυτές κατανέμονται σε διαφορετικές βαθμολογίες συναισθήματος. Για παράδειγμα, εάν τα περισσότερα σημεία στο διάγραμμα διασποράς ``roberta_pos`` vs. ``VADER_pos`` είναι πράσινα (υποδεικνύοντας θετικό συναίσθημα), αυτό δείχνει ότι και τα δύο μοντέλα συμφωνούν συχνά ότι τα κείμενα είναι θετικά.

- Εντοπισμός ακραίων τιμών: Τα ακραία σημεία (σημεία που αποκλίνουν σημαντικά από τα υπόλοιπα) μπορεί να υποδεικνύουν κείμενα στα οποία τα μοντέλα διαφωνούν έντονα. Αυτά είναι ζωτικής σημασίας για την ανάλυση σφαλμάτων και τη βελτίωση των μοντέλων.

7. Συζήτηση των ευρημάτων

Η κλιματική αλλαγή είναι ένα από τα πιο κρίσιμα ζητήματα που αντιμετωπίζει σήμερα ο πλανήτης μας, προκαλώντας ένα ευρύ φάσμα συναισθημάτων και απόψεων από τον παγκόσμιο πληθυσμό. Οι πλατφόρμες κοινωνικής δικτύωσης όπως το Twitter λειτουργούν ως ζωντανές αρένες όπου αυτά τα συναισθήματα μοιράζονται και συζητούνται ανοιχτά. Αναλύοντας τα tweets που σχετίζονται με την κλιματική αλλαγή χρησιμοποιώντας προηγμένα εργαλεία ανάλυσης συναισθήματος, όπως το RoBERTa και το VADER, μπορούμε να αποκτήσουμε βαθιά γνώση της κοινής γνώμης. Η μελέτη αυτή αποκαλύπτει μια συναρπαστική αλλά και ανησυχητική τάση: η μόνη ομάδα που εκφράζει θετικά συναισθήματα για την κλιματική αλλαγή αποτελείται από όσους αδιαφορούν για το θέμα. Τόσο οι υποστηρικτές όσο και οι σκεπτικιστές της κλιματικής αλλαγής εμφανίζουν κυρίως αρνητικά συναισθήματα. Η παρούσα ενότητα εμβαθύνει στις αποχρώσεις αυτών των ευρημάτων, διερευνώντας τους υποκείμενους λόγους και τις επιπτώσεις.

7.1 Επισκόπηση της ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή περιλαμβάνει την κατηγοριοποίηση των tweets σε θετικά, αρνητικά ή ουδέτερα συναισθήματα. Χρησιμοποιώντας το RoBERTa, ένα μοντέλο βασισμένο στον μετασχηματιστή, και το VADER, ένα εργαλείο ανάλυσης συναισθήματος βασισμένο σε κανόνες, εξάγαμε και αναλύσαμε τα συναισθήματα χιλιάδων tweets που συζητούν την κλιματική αλλαγή. Τα βήματα preprocessing περιλάμβαναν καθαρισμό, tokenization, αφαίρεση των stop words, lemmatization και stemming για να διασφαλιστεί ότι τα δεδομένα κειμένου ήταν σε κατάλληλη μορφή για ανάλυση.

7.2 Βασικά ευρήματα

1. Θετικό συναίσθημα από την ομάδα των αδιάφορων:

- Η ανάλυση αποκαλύπτει ότι τα μόνα tweets που εκφράζουν θετικό συναίσθημα προέρχονται από άτομα που εμφανίζονται αδιάφορα για την κλιματική αλλαγή. Αυτά τα tweets συχνά ευτελίζουν το ζήτημα, απορρίπτουν τη σημασία του ή εστιάζουν σε άσχετες θετικές πτυχές.

- Παραδείγματα τέτοιων tweets περιλαμβάνουν δηλώσεις όπως: "Δεν ανησυχώ για την κλιματική αλλαγή, είναι απλώς ένας ακόμη φυσικός κύκλος" ή "Γιατί να αγχώνεσαι για την κλιματική αλλαγή όταν υπάρχουν τόσα άλλα πράγματα που μπορείς να απολαύσεις στη ζωή".

2. Αρνητικό συναίσθημα από τους υποστηρικτές:

- Όσοι αναγνωρίζουν την πραγματικότητα της κλιματικής αλλαγής και υποστηρίζουν τη λήψη μέτρων εκφράζουν κυρίως αρνητικό συναίσθημα. Τα tweets τους αντικατοπτρίζουν συχνά απογοήτευση, άγχος και επείγουσα ανάγκη σχετικά με το ζήτημα.

- Παραδείγματα περιλαμβάνουν tweets όπως: "Πρέπει να αναλάβουμε σοβαρή δράση κατά της κλιματικής αλλαγής τώρα, πριν να είναι πολύ αργά" ή "Έχω κουραστεί τόσο πολύ με τους ανθρώπους που αρνούνται την κλιματική αλλαγή, ενώ τα στοιχεία είναι ξεκάθαρα".

3. Αρνητικό συναίσθημα από σκεπτικιστές:

- Οι σκεπτικιστές της κλιματικής αλλαγής, ή όσοι αρνούνται την ύπαρξή της ή τον ανθρώπινο αντίκτυπο, εμφανίζουν επίσης αρνητικό συναίσθημα. Τα tweets τους συχνά μεταφέρουν θυμό, δυσπιστία και εχθρότητα προς τις αφηγήσεις και τους υποστηρικτές της κλιματικής αλλαγής.

- Παραδείγματα περιλαμβάνουν tweets όπως: "Η κλιματική αλλαγή είναι μια απάτη που διαιωνίζεται από φοβιστές" ή "Δεν μπορώ να πιστέψω ότι οι άνθρωποι εξακολουθούν να πιστεύουν την απάτη της κλιματικής αλλαγής".

7.2.1 Ανάλυση του θετικού συναισθήματος μεταξύ της ομάδας των αδιάφορων

Το θετικό συναίσθημα των ατόμων που αδιαφορούν για την κλιματική αλλαγή είναι ένα ενδιαφέρον φαινόμενο. Το θετικό συναίσθημα αυτής της ομάδας πηγάζει συχνά από την έλλειψη ανησυχίας ή πεποίθησης για τη σοβαρότητα της κλιματικής αλλαγής. Αρκετοί παράγοντες συμβάλλουν σε αυτή την αδιαφορία:

1. Προκατάληψη εφησυχασμού και αισιοδοξίας:

- Πολλά άτομα παρουσιάζουν μια προκατάληψη αισιοδοξίας, πιστεύοντας ότι τα αρνητικά γεγονότα είναι λιγότερο πιθανό να τους συμβούν σε σύγκριση με άλλους. Αυτή η προκατάληψη μπορεί να οδηγήσει σε εφησυχασμό σχετικά με την κλιματική αλλαγή, καλλιεργώντας μια θετική προοπτική παρά την επικείμενη απειλή.

- Τα άτομα αυτά μπορεί να αντιλαμβάνονται την κλιματική αλλαγή ως ένα μακρινό ζήτημα που δεν θα τους επηρεάσει άμεσα, οδηγώντας σε απορριπτική ή ευτελιστική στάση.

2. Έλλειψη ευαισθητοποίησης ή παραπληροφόρηση:

- Ένα σημαντικό τμήμα της ομάδας των αδιάφορων μπορεί να μην έχει επίγνωση της σοβαρότητας και της αμεσότητας της κλιματικής αλλαγής. Η παραπληροφόρηση και η ανεπαρκής εκπαίδευση σχετικά με το θέμα μπορεί να συμβάλλουν στο θετικό τους συναίσθημα.

- Η αναπαράσταση από τα μέσα ενημέρωσης διαδραματίζει καθοριστικό ρόλο στη διαμόρφωση της αντίληψης του κοινού. Εάν η κλιματική αλλαγή δεν καλύπτεται επαρκώς ή παρουσιάζεται με τρόπο μη ανησυχητικό, αυτό μπορεί να οδηγήσει σε έλλειψη επείγοντος χαρακτήρα στο κοινό.

3. Ψυχολογικοί μηχανισμοί αντιμετώπισης:

- Ορισμένα άτομα χρησιμοποιούν το θετικό συναίσθημα ως ψυχολογικό μηχανισμό αντιμετώπισης για να αντιμετωπίσουν τη συντριπτική φύση της κλιματικής αλλαγής.

Υιοθετώντας μια θετική προοπτική, μπορούν να αποφύγουν το άγχος και την αδυναμία που συνδέονται με το ζήτημα.

- Αυτή η στρατηγική αντιμετώπισης μπορεί να περιλαμβάνει την υποβάθμιση της σημασίας της κλιματικής αλλαγής ή τον επαναπροσανατολισμό της προσοχής σε πιο άμεσες, προσωπικές ανησυχίες.

7.2.2 Αρνητικό συναίσθημα μεταξύ υποστηρικτών και σκεπτικιστών

Το αρνητικό συναίσθημα που εκφράζεται τόσο από τους υποστηρικτές όσο και από τους σκεπτικιστές της κλιματικής αλλαγής αναδεικνύει τον πολωμένο χαρακτήρα της συζήτησης. Τα αρνητικά συναισθήματα κάθε ομάδας οδηγούνται από διαφορετικούς υποκείμενους παράγοντες:

1. Οι υποστηρικτές της δράσης για την κλιματική αλλαγή:

- Απογοήτευση και επείγουσα ανάγκη: Οι υποστηρικτές εκφράζουν συχνά απογοήτευση λόγω του αργού ρυθμού δράσης και της έλλειψης επαρκών μέτρων για την καταπολέμηση της κλιματικής αλλαγής. Η επείγουσα ανάγκη τους οφείλεται στη βαθιά κατανόηση των πιθανών καταστροφικών επιπτώσεων της κλιματικής αλλαγής.

- Απόγνωση και άγχος: Η συνεχής έκθεση σε ανησυχητικές ειδήσεις σχετικά με την κλιματική αλλαγή μπορεί να οδηγήσει σε συναισθήματα απόγνωσης και άγχους. Η αντιληπτή έλλειψη προόδου και η αντίσταση των σκεπτικιστών επιδεινώνουν αυτά τα συναισθήματα.

- Προάσπιση και ακτιβισμός: Πολλοί υποστηρικτές ασχολούνται ενεργά με την προάσπιση και τον ακτιβισμό, αντιμετωπίζοντας αντίσταση και άρνηση από διάφορες πλευρές. Αυτή η αντίθεση τροφοδοτεί τα αρνητικά συναισθήματά τους, καθώς αγωνίζονται για ουσιαστική αλλαγή.

2. Σκεπτικιστές της κλιματικής αλλαγής:

- Δυσπιστία και εχθρότητα: Οι σκεπτικιστές συχνά τρέφουν δυσπιστία για την επιστημονική συναίνεση σχετικά με την κλιματική αλλαγή, θεωρώντας την υπερβολική ή κατασκευασμένη. Αυτή η δυσπιστία μεταφράζεται σε εχθρότητα απέναντι στις αφηγήσεις για την κλιματική αλλαγή και σε όσους τις προωθούν.

- Πολιτικές και ιδεολογικές επιρροές: Ο σκεπτικισμός της κλιματικής αλλαγής συχνά επηρεάζεται από πολιτικές και ιδεολογικές πεποιθήσεις. Οι σκεπτικιστές μπορεί να θεωρούν τη δράση για την κλιματική αλλαγή ως απειλή για τα οικονομικά τους συμφέροντα ή τις προσωπικές τους ελευθερίες, οδηγώντας σε αρνητικά συναισθήματα.

- Δυσπιστία στους θεσμούς: Μια γενική δυσπιστία προς τους επιστημονικούς θεσμούς, τους κυβερνητικούς φορείς και τα μέσα ενημέρωσης μπορεί να οδηγήσει στον σκεπτικισμό. Οι σκεπτικιστές μπορεί να αισθάνονται ότι αυτοί οι θεσμοί παραπλανούν το κοινό με απώτερα κίνητρα, καλλιεργώντας θυμό και δυσαρέσκεια.

7.3 Επιπτώσεις αυτών των ευρημάτων

Τα ευρήματα της ανάλυσης συναισθήματος έχουν αρκετές σημαντικές επιπτώσεις στην επικοινωνία για την κλιματική αλλαγή, τη χάραξη πολιτικής και τη συνηγορία:

1. Στοχευμένες στρατηγικές επικοινωνίας:

- Αντιμετώπιση της αδιαφορίας: Οι προσπάθειες θα πρέπει να επικεντρωθούν στην ευαισθητοποίηση και την εκπαίδευση της ομάδας των αδιάφορων σχετικά με τις πραγματικές επιπτώσεις της κλιματικής αλλαγής. Οι στρατηγικές επικοινωνίας πρέπει να τονίζουν την αμεσότητα και την προσωπική σημασία του ζητήματος για να ξεπεραστεί ο εφησυχασμός.

- Θετική πλαισίωση: Ενώ αναγνωρίζεται η σοβαρότητα της κλιματικής αλλαγής, η επικοινωνία θα πρέπει επίσης να αναδεικνύει θετικές δράσεις και λύσεις, ώστε να αποφεύγεται ο κατακλυσμός του κοινού με αρνητικές πληροφορίες. Αυτό μπορεί να βοηθήσει στη διατήρηση της δέσμευσης και στην προώθηση μιας προληπτικής στάσης.

2. Εμπλοκή των υποστηρικτών:

- Υποστήριξη και πόροι: Η παροχή υποστήριξης και πόρων για τους ακτιβιστές και τους υποστηρικτές του κλίματος μπορεί να βοηθήσει στη διαχείριση της απογοήτευσης και του άγχους τους. Αυτό περιλαμβάνει πλατφόρμες για την υπεράσπιση, χρηματοδότηση για πρωτοβουλίες και υποστήριξη ψυχικής υγείας για όσους έχουν συγκλονιστεί από το θέμα.

- Ενίσχυση των φωνών: Η ενίσχυση των φωνών των υποστηρικτών μέσω της κάλυψης από τα μέσα ενημέρωσης και των δημόσιων εγκρίσεων μπορεί να αυξήσει την επιρροή τους και να εξουδετερώσει το αρνητικό συναίσθημα που προκαλείται από την αντίσταση και τον σκεπτικισμό.

3. Αντιμετώπιση του σκεπτικισμού:

- Δημιουργία εμπιστοσύνης: Οι προσπάθειες για την οικοδόμηση εμπιστοσύνης στα επιστημονικά ιδρύματα και τους κυβερνητικούς φορείς είναι ζωτικής σημασίας. Η διαφάνεια στην επικοινωνία, η αντιμετώπιση της παραπληροφόρησης και η εμπλοκή αξιόπιστων ηγετών της κοινότητας μπορούν να βοηθήσουν στη γεφύρωση του χάσματος εμπιστοσύνης.

- Διάλογος χωρίς αποκλεισμούς: Η δημιουργία πλατφορμών για διάλογο χωρίς αποκλεισμούς που να αντιμετωπίζουν τις ανησυχίες και τους φόβους των σκεπτικιστών μπορεί να προωθήσει την κατανόηση και να μειώσει την εχθρότητα. Αυτό περιλαμβάνει την ακρόαση των απόψεών τους και την αντιμετώπιση των παρανοήσεων χωρίς να απορρίπτονται οι ανησυχίες τους.

4. Επιπτώσεις στην πολιτική:

- Ενημερωμένη χάραξη πολιτικής: Οι υπεύθυνοι χάραξης πολιτικής πρέπει να κατανοήσουν τα ποικίλα συναισθήματα απέναντι στην κλιματική αλλαγή για να σχεδιάσουν αποτελεσματικές

πολιτικές. Οι πολιτικές θα πρέπει να αντιμετωπίζουν τις ανησυχίες των σκεπτικιστών, προωθώντας παράλληλα θετικές δράσεις και λύσεις που εμπλέκουν την ομάδα των αδιάφορων.

- Δικομματικές προσεγγίσεις: Οι πολιτικές για την κλιματική αλλαγή θα πρέπει να επιδιώκουν διακομματική υποστήριξη για να αμβλύνουν το ιδεολογικό χάσμα. Οι συνεργατικές προσεγγίσεις που λαμβάνουν υπόψη τις οικονομικές και κοινωνικές επιπτώσεις μπορούν να βοηθήσουν στην επίτευξη ευρύτερης αποδοχής και στη μείωση των αρνητικών συναισθημάτων.

Η ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή αποκαλύπτει ένα πολύπλοκο και πολωμένο τοπίο. Το θετικό συναίσθημα από την ομάδα των αδιάφορων έρχεται σε έντονη αντίθεση με το αρνητικό συναίσθημα τόσο των υποστηρικτών όσο και των σκεπτικιστών. Η κατανόηση αυτών των συναισθημάτων είναι ζωτικής σημασίας για την ανάπτυξη αποτελεσματικών στρατηγικών επικοινωνίας, την προώθηση της δέσμευσης και την προώθηση ουσιαστικής δράσης κατά της κλιματικής αλλαγής. Με την αντιμετώπιση των υποκείμενων παραγόντων που συμβάλλουν σε αυτά τα συναισθήματα, τα ενδιαφερόμενα μέρη μπορούν να ενισχύσουν την ευαισθητοποίηση του κοινού, να οικοδομήσουν εμπιστοσύνη και να προωθήσουν τις συνεργατικές προσπάθειες για την αντιμετώπιση μιας από τις πιο πιεστικές προκλήσεις της εποχής μας.

7.4 Προτάσεις για μελλοντικές ερευνητικές κατευθύνσεις

Η μελέτη της ανάλυσης συναισθήματος σε tweets για την κλιματική αλλαγή αποκάλυψε αρκετά ενδιαφέροντα ευρήματα, ιδίως τα αρνητικά συναισθήματα που εκφράζονται τόσο από τους πιστούς όσο και από τους αρνητές της κλιματικής αλλαγής, ενώ μόνο η ομάδα των αδιάφορων παρουσιάζει θετικά συναισθήματα. Η παρατήρηση αυτή ανοίγει πολλούς δρόμους για μελλοντική έρευνα που μπορεί να φωτίσει περαιτέρω τις αποχρώσεις της κοινής γνώμης σχετικά με την κλιματική αλλαγή και να ενισχύσει την κατανόησή μας για το πώς να επικοινωνούμε αποτελεσματικά και να εμπλέκουμε διαφορετικά τμήματα του πληθυσμού. Ακολουθούν ορισμένες ολοκληρωμένες προτάσεις για μελλοντικές ερευνητικές κατευθύνσεις:

1. Μακροχρόνια ανάλυση συναισθήματος

Σκοπός:

- Παρακολούθηση και κατανόηση της εξέλιξης του συναισθήματος του κοινού απέναντι στην κλιματική αλλαγή με την πάροδο του χρόνου και ως απάντηση σε συγκεκριμένα γεγονότα.

Προσέγγιση:

- Διεξαγωγή διαχρονικών μελετών που αναλύουν tweets και άλλες αναρτήσεις στα μέσα κοινωνικής δικτύωσης επί σειρά ετών.

- Εντοπίστε και εξετάστε βασικά γεγονότα, όπως φυσικές καταστροφές, ανακοινώσεις για την κλιματική πολιτική και σημαντικές αναφορές στα μέσα ενημέρωσης, για να δείτε πώς επηρεάζουν το δημόσιο συναίσθημα.

Προσδοκώμενα αποτελέσματα:

- Διαπιστώσεις σχετικά με τον τρόπο με τον οποίο το δημόσιο συναίσθημα μεταβάλλεται ως απάντηση σε γεγονότα του πραγματικού κόσμου.
- Κατανόηση του μακροπρόθεσμου αντίκτυπου των διαφόρων στρατηγικών επικοινωνίας και των παρεμβάσεων πολιτικής.
- Δεδομένα για την υποστήριξη της ανάπτυξης αποτελεσματικότερων, χρονικά ευαίσθητων στρατηγικών επικοινωνίας.

2. Ανάλυση πλαισίου και κατάστασης

Στόχος:

- Η κατανόηση του τρόπου με τον οποίο οι παράγοντες του πλαισίου και της κατάστασης επηρεάζουν την έκφραση και την πρόσληψη των συναισθημάτων που σχετίζονται με την κλιματική αλλαγή.

Πρόσβαση:

- Χρήση προηγμένων τεχνικών NLP για την καταγραφή των αποχρώσεων και των εξαρτήσεων του πλαισίου στα tweets.
- Εξετάστε παράγοντες όπως η γεωγραφική τοποθεσία, το πολιτιστικό υπόβαθρο, η κοινωνικοοικονομική κατάσταση και τα τρέχοντα γεγονότα κατά την ανάλυση των συναισθημάτων.

Προσδοκώμενα αποτελέσματα:

- Ακριβέστερη ανάλυση συναισθήματος που αντικατοπτρίζει τα πραγματικά συναισθήματα διαφορετικών τμημάτων του πληθυσμού.
- Διαπιστώσεις σχετικά με τον τρόπο με τον οποίο τα διαφορετικά πλαίσια επηρεάζουν την αντίληψη και τη δέσμευση του κοινού σε θέματα κλιματικής αλλαγής.
- Βελτιωμένες στρατηγικές για την προσαρμογή της επικοινωνίας σε διαφορετικά πλαίσια και καταστάσεις.

3. Πολυγλωσσικές και διαπολιτισμικές σπουδές

Στόχος:

- Απόκτηση μιας παγκόσμιας προοπτικής του λόγου για την κλιματική αλλαγή με τη συμπερίληψη πολλαπλών γλωσσών και πολιτισμικών πλαισίων στην ανάλυση συναισθήματος.

Προσέγγιση:

- Ανάπτυξη και βελτίωση πολύγλωσσων εργαλείων ανάλυσης συναισθήματος ικανών να χειρίζονται γλωσσικές και πολιτισμικές παραλλαγές.
- Συλλογή και ανάλυση δεδομένων κοινωνικών μέσων από διάφορες γεωγραφικές περιοχές και σε διάφορες γλώσσες.

Προσδοκώμενα αποτελέσματα:

- Συνολική κατανόηση του παγκόσμιου κλίματος απέναντι στην κλιματική αλλαγή.
- Προσδιορισμός των περιφερειακών διαφορών στο λόγο και το συναίσθημα για την κλιματική αλλαγή.
- Ανάπτυξη πολιτισμικά ευαίσθητων και γλωσσικά κατάλληλων στρατηγικών επικοινωνίας.

4 Ανίχνευση και ανάλυση κινήσεων

Στόχος:

- Πέρα από την απλή ταξινόμηση θετικών/αρνητικών συναισθημάτων, ανίχνευση και ανάλυση συγκεκριμένων συναισθημάτων που σχετίζονται με την κλιματική αλλαγή.

Πρόσβαση:

- Χρήση προηγμένων μοντέλων NLP για την ταξινόμηση των tweets σε διάφορες κατηγορίες συναισθημάτων, όπως φόβος, θυμός, ελπίδα και απελπισία.
- Εξετάστε πώς τα διάφορα συναισθήματα συσχετίζονται με τις δημόσιες δράσεις, την υποστήριξη της πολιτικής και τα επίπεδα δέσμευσης.

Προσδοκώμενα αποτελέσματα:

- Βαθύτερη κατανόηση των συναισθηματικών οδηγών των στάσεων και των συμπεριφορών σχετικά με την κλιματική αλλαγή.
- Προσδιορισμός των συναισθηματικών αντιδράσεων που μπορούν να στοχευθούν για την ενίσχυση της δέσμευσης και της υποστήριξης της δράσης για το κλίμα.
- Ανάπτυξη στρατηγικών επικοινωνίας που απευθύνονται σε συγκεκριμένες συναισθηματικές αντιδράσεις και ενισχύουν τη θετική συναισθηματική δέσμευση.

5. Επιπτώσεις της παραπληροφόρησης και της παραπληροφόρησης

Στόχος:

- Να διερευνηθεί ο ρόλος της παραπληροφόρησης και της παραπληροφόρησης στη διαμόρφωση του δημόσιου συναισθήματος απέναντι στην κλιματική αλλαγή.

Προσέγγιση:

- Εντοπισμός και ανάλυση της διάδοσης της παραπληροφόρησης και της παραπληροφόρησης στις πλατφόρμες κοινωνικής δικτύωσης.

- Αξιολόγηση του αντίκτυπου της ψευδούς πληροφόρησης στο δημόσιο συναίσθημα και τη συμπεριφορά χρησιμοποιώντας ανάλυση συναισθήματος και μετρικές εμπλοκής των χρηστών.

Προσδοκώμενα αποτελέσματα:

- Στρατηγικές για την καταπολέμηση της παραπληροφόρησης και την ενίσχυση της κατανόησης της κλιματικής αλλαγής από το κοινό.
- Ανάπτυξη εργαλείων και τεχνικών για την ανίχνευση και τον μετριασμό των επιπτώσεων της παραπληροφόρησης.
- Ενδείξεις σχετικά με τον τρόπο με τον οποίο η παραπληροφόρηση επηρεάζει την αντίληψη και το συναίσθημα του κοινού και πώς να αντιμετωπιστούν οι επιπτώσεις της.

6. Επίδραση του οπτικού περιεχομένου και των πολυμέσων

Στόχος:

- Να διερευνήσει την επίδραση του οπτικού περιεχομένου και του περιεχομένου πολυμέσων στο συναίσθημα και τη δέσμευση για την κλιματική αλλαγή.

Πρόσβαση:

- Ανάλυση του συναισθήματος και των επιπέδων εμπλοκής των tweets που περιέχουν εικόνες, βίντεο, infographics και άλλο περιεχόμενο πολυμέσων.
- Συγκρίνετε την αποτελεσματικότητα των διαφόρων τύπων περιεχομένου πολυμέσων στη μετάδοση πληροφοριών για την κλιματική αλλαγή και στην πρόκληση δέσμευσης.

Προσδοκώμενα αποτελέσματα:

- Κατανόηση του τρόπου με τον οποίο το οπτικό και πολυμεσικό περιεχόμενο επηρεάζει το δημόσιο συναίσθημα και τη δέσμευση.
- Προσδιορισμός των πιο αποτελεσματικών τύπων περιεχομένου πολυμέσων για την επικοινωνία σχετικά με την κλιματική αλλαγή.
- Ανάπτυξη στρατηγικών για την ενσωμάτωση περιεχομένου πολυμέσων στις προσπάθειες επικοινωνίας για την κλιματική αλλαγή, ώστε να μεγιστοποιηθεί ο αντίκτυπος.

7. Μελέτες συμπεριφορικού και συμπεριφορικού αντικτύπου

Στόχος:

- Αξιολόγηση του αντίκτυπου του συναισθήματος της κλιματικής αλλαγής στις πραγματικές συμπεριφορές και στάσεις.

Προσέγγιση:

- Διεξαγωγή ερευνών και συνεντεύξεων για τη συμπλήρωση της ανάλυσης συναισθήματος, παρέχοντας βαθύτερη κατανόηση του τρόπου με τον οποίο τα εκφρασμένα συναισθήματα μεταφράζονται σε πραγματικές δράσεις και στάσεις.

- Αναλύστε συσχετισμούς μεταξύ του συναισθήματος στα μέσα κοινωνικής δικτύωσης και συμπεριφορών όπως τα πρότυπα ψηφοφορίας, η συμμετοχή στον ακτιβισμό για το κλίμα και οι προσωπικές αλλαγές στον τρόπο ζωής.

Προσδοκώμενα αποτελέσματα:

- Ενδείξεις σχετικά με τη σχέση μεταξύ των εκφρασμένων συναισθημάτων και της συμπεριφοράς στον πραγματικό κόσμο.
- Προσδιορισμός των παραγόντων που παρακινούν τα άτομα να αναλάβουν δράση με βάση τα συναισθήματά τους.
- Δεδομένα για την πληροφόρηση του σχεδιασμού παρεμβάσεων με στόχο την προώθηση θετικών αλλαγών συμπεριφοράς ως απάντηση στην κλιματική αλλαγή.

8. Αλγοριθμικές βελτιώσεις για την ανάλυση συναισθήματος

Στόχος:

- Βελτίωση της ακρίβειας και της αξιοπιστίας των αλγορίθμων ανάλυσης συναισθήματος στο πλαίσιο του διαλόγου για την κλιματική αλλαγή.

Προσέγγιση:

- Ανάπτυξη και δοκιμή νέων αλγορίθμων που αντιμετωπίζουν τις ειδικές προκλήσεις της ανάλυσης συναισθήματος της κλιματικής αλλαγής, όπως η ανίχνευση σαρκασμού, η επίγνωση του πλαισίου και η ταξινόμηση συναισθημάτων.
- Συγκριτική αξιολόγηση των νέων αλγορίθμων με τους υπάρχοντες για τη μέτρηση των βελτιώσεων στην απόδοση και την ακρίβεια.

Προσδοκώμενα αποτελέσματα:

- Πιο ακριβή και αξιόπιστα εργαλεία ανάλυσης συναισθήματος.
- Ενισχυμένη ικανότητα ανίχνευσης διαφοροποιημένων συναισθημάτων και συναισθημάτων στο διάλογο για την κλιματική αλλαγή.
- Βελτιωμένη κατανόηση του δημόσιου συναισθήματος που μπορεί να ενημερώσει τις στρατηγικές πολιτικής και επικοινωνίας.

9. Ενσωμάτωση πολυτροπικών πηγών δεδομένων

Στόχος:

- Βελτίωση της ανάλυσης συναισθήματος με την ενσωμάτωση δεδομένων από πολλαπλές πηγές, όπως κείμενο, εικόνες, βίντεο και ήχο.

Προσέγγιση:

- Ανάπτυξη πολυτροπικών μοντέλων ανάλυσης συναισθήματος που μπορούν να επεξεργάζονται και να αναλύουν δεδομένα από διάφορες πηγές.
- Συλλογή και ενσωμάτωση δεδομένων από μέσα κοινωνικής δικτύωσης, άρθρα ειδήσεων, ομιλίες και άλλες σχετικές πηγές.

Προσδοκώμενα αποτελέσματα:

- Μια πιο ολιστική κατανόηση του συναισθήματος του κοινού απέναντι στην κλιματική αλλαγή.
- Εντοπισμός διατροφικών μοτίβων και διορατικότητας που δεν είναι εμφανείς από δεδομένα μίας πηγής.
- Ενισχυμένη ικανότητα διαμόρφωσης ολοκληρωμένων στρατηγικών επικοινωνίας που αξιοποιούν πολλαπλούς τύπους μέσων ενημέρωσης.

10. Προσωποποιημένες και στοχευμένες στρατηγικές επικοινωνίας

Στόχος:

- Ανάπτυξη εξατομικευμένων και στοχευμένων στρατηγικών επικοινωνίας που βρίσκουν απήχηση σε διαφορετικά τμήματα κοινού.

Προσέγγιση:

- Χρήση της ανάλυσης συναισθήματος για την τμηματοποίηση του κοινού με βάση τα εκφρασμένα συναισθήματα και τα συναισθήματά του.
- Προσαρμόστε τις στρατηγικές επικοινωνίας ώστε να ανταποκρίνονται στις συγκεκριμένες ανάγκες, ανησυχίες και προτιμήσεις κάθε τμήματος.

Προσδοκώμενα αποτελέσματα:

- Αποτελεσματικότερη επικοινωνία για την κλιματική αλλαγή που οδηγεί σε αυξημένη δέσμευση και υποστήριξη.
- Ανάπτυξη στοχευμένων παρεμβάσεων που αντιμετωπίζουν τις μοναδικές προκλήσεις και τα κίνητρα των διαφόρων τμημάτων του κοινού.
- Ενισχυμένη ικανότητα προώθησης θετικών στάσεων και συμπεριφορών προς τον μετριασμό της κλιματικής αλλαγής και την προσαρμογή σε αυτήν.

Με την επιδίωξη αυτών των μελλοντικών ερευνητικών κατευθύνσεων, μπορούμε να βασιστούμε στα τρέχοντα ευρήματα και να συνεχίσουμε να βελτιώνουμε την κατανόηση του συναισθήματος του κοινού απέναντι στην κλιματική αλλαγή. Αυτό, με τη σειρά του, μπορεί να ενημερώσει για

αποτελεσματικότερες στρατηγικές επικοινωνίας, πολιτικές και παρεμβάσεις που προωθούν τη μεγαλύτερη ευαισθητοποίηση, δέσμευση και δράση σε θέματα κλιματικής αλλαγής.

8. Επίλογος

Η ανάλυση συναισθήματος, ένα υποπεδίο της Επεξεργασίας Φυσικής Γλώσσας (NLP), επικεντρώνεται στον εντοπισμό και την κατηγοριοποίηση των απόψεων που εκφράζονται σε κείμενο. Είναι ζωτικής σημασίας για την κατανόηση της κοινής γνώμης, γεγονός που την καθιστά απαραίτητη σε διάφορες εφαρμογές, από την ανάλυση της αγοράς έως τη χάραξη πολιτικής. Οι μεθοδολογίες που χρησιμοποιούνται στην ανάλυση συναισθήματος έχουν εξελιχθεί από απλές προσεγγίσεις που βασίζονται σε κανόνες σε προηγμένες τεχνικές μηχανικής μάθησης και βαθιάς μάθησης. Η επιλογή χαρακτηριστικών, ένα κρίσιμο βήμα preprocessing στη μηχανική μάθηση, περιλαμβάνει την επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών για τη βελτίωση της απόδοσης του μοντέλου, τη μείωση του hyper-adaptation και τη βελτίωση της ερμηνευσιμότητας. Η παρούσα μελέτη αξιοποιεί τα μοντέλα RoBERTa και VADER για την ανάλυση συναισθήματος των tweets κλιματικής αλλαγής, δίνοντας έμφαση στους ρόλους τους στην επιλογή χαρακτηριστικών και στην ταξινόμηση συναισθήματος.

8.1 Σημασία της ανάλυσης συναισθήματος για την κατανόηση της κοινής γνώμης

Η ανάλυση συναισθήματος των δεδομένων των μέσων κοινωνικής δικτύωσης προσφέρει σε πραγματικό χρόνο πληροφορίες για την κοινή γνώμη. Για τις επιχειρήσεις, βοηθά στην κατανόηση της ικανοποίησης των πελατών και στη διαχείριση της φήμης των εμπορικών σημάτων. Οι υπεύθυνοι χάραξης πολιτικής μπορούν να τη χρησιμοποιήσουν για να μετρήσουν την αντίδραση του κοινού σε πολιτικές και γεγονότα, επιτρέποντας πιο ευέλικτες και τεκμηριωμένες αποφάσεις. Οι ερευνητές επωφελούνται από τη μελέτη κοινωνικών φαινομένων και τάσεων, συμβάλλοντας σε τομείς όπως η κοινωνιολογία, η ψυχολογία και η περιβαλλοντική επιστήμη. Οι πλατφόρμες κοινωνικής δικτύωσης, ιδίως το Twitter, χρησιμεύουν ως πλούσιες πηγές δεδομένων σε πραγματικό χρόνο, αντανakλώντας τη συλλογική διάθεση και τις απόψεις διαφορετικών πληθυσμών.

8.2 Ανάλυση συναισθήματος των tweets για την κλιματική αλλαγή

Η κλιματική αλλαγή είναι ένα παγκοσμίως αναγνωρισμένο ζήτημα με βαθιές επιπτώσεις. Η κατανόηση της αντίληψης του κοινού για την κλιματική αλλαγή είναι ζωτικής σημασίας για τη διαμόρφωση πολιτικών και δράσεων με στόχο τον μετριασμό των επιπτώσεών της. Η παρούσα μελέτη επικεντρώνεται στη διενέργεια ανάλυσης συναισθήματος των tweets που σχετίζονται με την κλιματική αλλαγή για την αποκάλυψη των υποκείμενων συναισθημάτων, απόψεων και ανησυχιών του κοινού. Οι πρωταρχικοί στόχοι περιλαμβάνουν την ανάλυση του συναισθήματος των tweets που σχετίζονται με την κλιματική αλλαγή, τη σύγκριση της απόδοσης των μοντέλων VADER και RoBERTa, την αξιολόγηση της επίδρασης της επιλογής χαρακτηριστικών και τον εντοπισμό βασικών θεμάτων και τάσεων στη συζήτηση για την κλιματική αλλαγή στο Twitter.

Η εφαρμογή των RoBERTa και VADER για την ανάλυση συναισθήματος σε tweets για την κλιματική αλλαγή αποδεικνύει τη δύναμη του συνδυασμού προηγμένων μοντέλων που βασίζονται στον μετασχηματιστή με αποτελεσματικές προσεγγίσεις που βασίζονται σε κανόνες. Το RoBERTa συλλαμβάνει βαθιές πληροφορίες σχετικά με το πλαίσιο, ενώ το VADER παρέχει απλές βαθμολογίες συναισθήματος, καθεμία από τις οποίες συμβάλλει σε μια ολοκληρωμένη διαδικασία επιλογής χαρακτηριστικών. Αξιοποιώντας αυτές τις μεθοδολογίες, η παρούσα μελέτη ενισχύει την ακρίβεια και την ερμηνευσιμότητα των μοντέλων ανάλυσης συναισθήματος, παρέχοντας πολύτιμες πληροφορίες για το δημόσιο συναίσθημα σχετικά με την κλιματική αλλαγή. Η ενσωμάτωση αυτών των μεθόδων επιτρέπει μια στιβαρή ανάλυση, αντιμετωπίζοντας τις πολυπλοκότητες και τις αποχρώσεις των δεδομένων των κοινωνικών μέσων ενημέρωσης και προσφέροντας μια βαθύτερη κατανόηση του δημόσιου λόγου για κρίσιμα παγκόσμια ζητήματα.

9. Βιβλιογραφία

1. Bomert, Andrea, et al. "Benchmark for filter methods for feature selection in high-dimensional classification data."
2. Camarillo, Mariana Gaytan, et al. "Big Changes Start With Small Talk: Twitter and Climate Change in Times of Coronavirus Pandemic."
3. Devika, MD, et al. "Sentiment Analysis:A Comparative Study On Different Approaches."
4. Famili, A. Fazel, et al. "Advances in Intelligent Data Analysis VI."
5. Ford, James D., et al. "Big data has big potential for applications to climate change adaptation."
6. Guyon, Isabelle, and Andre Elisseeff. "An Introduction to Variable and Feature Selection."
7. Hutto, C. J., and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text."
8. Jungdong, Li, et al. "Feature Selection: A Data Perspective."
9. Kian, tan Long, et al. "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network."
10. Liu, B. "Sentiment Analysis and Opinion Mining."
11. Taboada, Maite. "Sentiment Analysis: An Overview from Linguistics."
12. Xiaoran, An, et al. "Tracking Climate Change Opinions from Twitter Data."
13. Karthick Prasad Gunasekaran "Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review"

14. Soufiane El Mrabti Mohammed, et al. "Comparison of Feature Selection Methods for Sentiment Analysis"
15. Sivasankar E, et al. "Efficient Feature Selection techniques for Sentiment Analysis"
16. Harish Dutt Sharma "An Analysis of Sentiment: Methods, Applications, and Challenges"
17. Chinmayee Sahoo et al. "Sentiment analysis using deep learning techniques: a comprehensive review"
18. Sivasankar Elango et al. "Efficient feature selection techniques for sentiment analysis"
19. Noura A. Semary et al. "Enhancing machine learning-based sentiment analysis through feature extraction techniques"
20. Selma Ayşe Özel et al. "QER: a new feature selection method for sentiment analysis"
21. Nabizath Saleena "An Ensemble Classification System for Twitter Sentiment Analysis"
22. Alexander Clark PhD et al. "The Handbook of Computational Linguistics and Natural Language Processing"
23. Ming Zhou, et al. "Progress in Neural NLP: Modeling, Learning, and Reasoning"
24. Acheampong Francisca Adoma et al. "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition"