Ανάλυση Συναισθήματος - Memeability σε Παροιμίες με Χρήση Εργαλείων NLP

Ρέππας Παναγιώτης
f3662415

Εισαγωγή-Ερευνητικό Ερώτημα

Στην παρούσα εργασία μελετάται, αναφορικά με ένα σύνολο παροιμιών από την ελληνική λαογραφική παράδοση, η διαδικασία ανάλυσης δεδομένων για την ποιοτική κατηγοριοποίηση συναισθημάτων, την καταταξη τους ανάλογα με την φύση του συναισθήματος, και του memeability τους. Χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων και επεξεργασίας φυσικής γλώσσας διαφόρων τύπων. Η μεθοδολογία που ακολουθήθηκε μας επιτρέπει να απαντήσουμε σε ποικίλα ερευνητικά ερωτήματα, ανάλογα με τη φύση των πειραμάτων που θα διεξάγουμε. Πιο συγκεκριμένα, για τα πειράματα που έχουμε επιλέξει, το ερευνητικό ερώτημα μας διαμορφώνεται ως εξής: "Ποιες οι συναισθηματικές και μιμιδιακές διαστάσεις των ελληνικών παροιμιών και πώς η επεξεργασία φυσικής γλώσσας μπορεί να αποκαλύψει τις πιο συχνές και αντιπροσωπευτικές λέξεις-έννοιες σε σχέση με τα συναισθήματα;"

Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν είναι αποτελέσματα από διάφορους επισημειωτές. Κάθε παροιμία συνδέεται με επισημειώσεις σχετικά με το συναισθηματικό της περιεχόμενο (Emotion), το γενικό συναισθηματικό φορτίο (Sentiment) και το Memeability (καταλληλότητα για μετατροπή σε meme) της. Ο στόχος είναι να αναλυθούν και να κατηγοριοποιηθούν οι παροιμίες σε διάφορες ομάδες, με βάση τα χαρακτηριστικά αυτά. Χρησιμοποιήσαμε 12 από τα 13 διαθέσιμα αρχεία επισημείωσης. Ο επισημειωτής 9, παραλείφθηκε κατά την αξιολόγηση των δεδομένων καθώς το αρχείο δεν πληρούσε την τυπολογία που είχε οριστεί και τηρηθεί στα υπόλοιπα αρχεία xlsx.

Μεθοδολογία

Η μέθοδος που ακολουθήθηκε για την δημιουργία του κώδικα χωρίζεται σε κομμάτια για την καλύτερη κατανόηση και ευρετηρίαση, καθώς και για λόγους πιθανής επανάχρησης ή και για

τη διεξαγωγή διαφορετικών πειραμάτων. Παρακάτω περιγράφονται αναλυτικότερα τα βήματα που ακολουθήθηκαν. Για την επεξεργασία των δεδομένων, χρησιμοποιήθηκαν οι βιβλιοθήκες pandas, sklearn και itertools.

pandas: Η βιβλιοθήκη χρησιμοποιήθηκε για τη φόρτωση και οργάνωση των δεδομένων, καθώς και για την εκτέλεση υπολογισμών.
sklearn: Δείκτες όπως το Kappa, η ομοιότητα συνημιτόνου (cosine similarity) και το TF-IDF, χρησιμοποιήθηκαν για να μετατραπούν γλωσσικά δεδομένα σε αριθμητικές τιμές, οι οποίες χρησιμοποιούνται για την εξαγωγή συμπερασμάτων μέσω συγκρίσεων.
itertools: Χρησιμοποιήθηκε για να χειριστεί συνδυαστικά τα δεδομένα και να διευκολύνει τις διαδικασίες υπολογισμού.

Αρχικά, δημιουργήθηκαν τρία DataFrames για τα επισημειωμένα δεδομένα, ώστε να γίνει η σύγκριση των επισημειώσεων ανά επισημειωτή, πραγματοποιώντας ομαδοποίηση και ταξινόμηση των δεδομένων. Η συνάρτηση pivot στο pandas παραθέτει τις τιμές που έχουν οριστεί ανά επισημειωτή, όπου κάθε επισημειωτής γίνεται μια στήλη, διευκολύνοντας τη σύγκριση των απαντήσεών τους. Στη συνέχεια, κρίθηκε απαραίτητο να δημιουργηθεί ένας κώδικας που εξάγει μοναδικές παροιμίες, τις μετατρέπει σε διανύσματα με TF-IDF και υπολογίζει την ομοιότητά τους, προκειμένου να παραλειφθούν διπλότυπες ή παρεμφερείς παροιμίες και να δημιουργηθεί ένα DataFrame χωρίς επαναλήψεις. Ιδανικά, αυτή η διαδικασία θα γινόταν κατά τη δημιουργία του αρχείου επισημείωσης ή κατά την επεξεργασία των επισημειωμένων αρχείων, ώστε να αποφευχθεί η επιβάρυνση του υπολογιστικού συστήματος, ιδιαίτερα για μεγάλα datasets.

Η συμφωνία μεταξύ επισημειωτών όσον αφορά το αν μια παροιμία είναι ουδέτερη μετρήθηκε μέσω percentage agreement και Kappa. Το Kappa χρησιμοποιείται για να μετρηθεί η συμφωνία μεταξύ δύο επισημειωτών, λαμβάνοντας υπόψη την τυχαιότητα. Επειδή υπολογίζεται μόνο για δύο επισημειωτές, έγινε η απαραίτητη προσαρμογή ώστε να υπολογιστεί το Kappa για κάθε πιθανό ζευγάρι επισημειωτών. Στη συνέχεια, υπολογίστηκε ποιοι επισημειωτές συμφωνούν περισσότερο και ποιοι λιγότερο (ανά ζευγάρι) για το αν μια παροιμία είναι ουδέτερη. Τα αποτελέσματα απεικονίστηκαν σε heatmap, διευκολύνοντας την αναγνώριση του βαθμού συμφωνίας ή διαφωνίας, ενώ αποκρύφθηκαν οι διαγώνιες τιμές (συγκρίσεις μεταξύ του ίδιου επισημειωτή).

Εντοπίστηκαν οι 10 παροιμίες που οι επισημειωτές τις σημείωσαν λιγότερο ως ουδέτερες και αναφέρθηκαν ποιες από αυτές είχαν θετικά συναισθήματα. Υπολογίστηκε το μέσο memeability για κάθε μία και αναδείχθηκαν οι 10 πιο memeable παροιμίες. Επιπλέον, υπολογίστηκε η πιθανότητα συμφωνίας για το Emotion και αναδείχθηκαν οι 3 παροιμίες με τη μεγαλύτερη πιθανότητα συμφωνίας ανά επισημειωτή. Στην τελευταία φάση, χρησιμοποιήθηκε το TF-IDF (Term Frequency-Inverse Document Frequency) για να εξαχθούν οι πιο συχνές λέξεις σε κάθε συναισθηματική ομάδα, αποφεύγοντας τη χρήση γενικών λέξεων μέσω παραμετροποίησης της συνάρτησης TfidfVectorizer από τη βιβλιοθήκη sklearn.

Συμπεράσματα

Τα αποτελέσματα έδειξαν ότι η μέθοδος που ακολουθήθηκε ήταν επαρκής για την εξαγωγή συμπερασμάτων, τα οποία θα μπορούσαν να χρησιμεύσουν συμπληρωματικά σε άλλες έρευνες, κυρίως σε τομείς όπως η λαογραφία, η γλωσσολογία ή η ιστορία. Επίσης, θα μπορούσαν να αποτελέσουν αυτόνομη μελέτη, εφόσον υποστηριχθούν από το κατάλληλο θεωρητικό πλαίσιο, προκειμένου να μελετηθούν οι συνειρμοί και οι γλωσσικές τυπολογίες που σχετίζονται με συγκεκριμένες περιοχές, ιστορικές συνθήκες και μετακινήσεις. Η επιλογή δεδομένων από την προφορική παράδοση διασφαλίζει την αποφυγή αποστείρωσης του δείγματος, κάτι που είναι συχνό πρόβλημα στην επιλογή γραπτών κειμένων, και μας επιτρέπει να γενικεύσουμε τα αποτελέσματα μας εφόσον οι παροιμίες αποδίδονται σε συσσωρευμένη λαική σοφία. Το γεγονός οτι επικράτησαν σαν άγραφοι νόμοι, ή κανόνες ενδεικτικής συμπεριφοράς για μεγάλο χρονικό διάστημα, αποδεικνύει την καθολικότητα τους. Η επιτυχία εξαγωγής ουσιαστικών, ρημάτων, επιθέτων διαφόρων θεματικών και η μείωση της καταγραφής των συχνότερων λέξεων (όπως αντωνυμίες, άρθρα, σύνδεσμοι), πιστοποιεί την επιτυχία της μεθόδου ως προς την εκπλήρωση του ερευνητικού ερωτήματος, και φαίνεται να συνδέει τις φαινομενικά άσχετες αυτές λέξεις, με συναισθηματικές κατευθύνσεις.

Για την αποπεράτωση και επεξεργασία του κώδικα, καθώς και τη μορφοποίηση του τελικού κειμένου, χρησιμοποιήθηκαν τα ChatGPT o3-mini και GPT-4o.