Ετήσιο Ελληνόφωνο Επιστημονικό Συνέδριο Εργαστηρίων Επικοινωνίας

Τόμ. 3, Αρ. 1 (2024)

3ο Ετήσιο Ελληνόφωνο Συνέδριο Εργαστηρίων Επικοινωνίας: Δημοσιογραφία, Μέσα και Επικοινωνία: Σύγχρονες προκλήσεις στην εποχή της Τεχνητής Νοημοσύνης

Τεχνικές μοντελοποίησης θεμάτων σε κείμενα και ομιλίες ειδησεογραφικού περιεχομένου

Παναγιώτης Κοσμάς , Μαρίνα-Ειρήνη Σταματιάδου , Χαράλαμπος Δημούλας

doi: 10.12681/cclabs.8066

Βιβλιογραφική αναφορά:

Κοσμάς Π., Σταματιάδου Μ.-Ε., & Δημούλας Χ. (2025). Τεχνικές μοντελοποίησης θεμάτων σε κείμενα και ομιλίες ειδησεογραφικού περιεχομένου. *Ετήσιο Ελληνόφωνο Επιστημονικό Συνέδριο Εργαστηρίων Επικοινωνίας*, *3*(1), 119–133. https://doi.org/10.12681/cclabs.8066

Τεχνικές μοντελοποίησης θεμάτων σε κείμενα και ομιλίες ειδησεογραφικού περιεχομένου

Κοσμάς Παναγιώτης
Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
pkosmas@jour.auth.gr

Σταματιάδου Μαρίνα-Ειρήνη
Υπ. Διδάκτωρ, Τμήμα Δημοσιογραφίας & ΜΜΕ,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
mstamat@jour.auth.gr

Δρ. Δημούλας Χαράλαμπος
Καθηγητής, Τμήμα Δημοσιογραφίας & ΜΜΕ,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
babis@eng.auth.gr

Περίληψη

Η επεξεργασία φυσικής γλώσσας βρίσκει εφαρμογή και στη δημοσιογραφία, αναδιαμορφώνοντας τον τρόπο ανάλυσης μεγάλων δεδομένων κειμένων, με στόχο την καλύτερη διαχείρισή τους. Η μοντελοποίηση θεμάτων είναι μια τεχνική ανάλυσης δεδομένων που χρησιμοποιείται για την ανάδειξη κυρίαρχων εννοιών που διατρέχουν μεγάλα σύνολα κειμένων, με στόχο την αναγνώριση των σημαντικότερων θεμάτων που υπάρχουν μέσα σε αυτά και την κατάλληλη ομαδοποίησή τους. Λύσεις σε τέτοια προβλήματα προσφέρουν τα Γλωσσικά Μοντέλα Μετασχηματιστών (Transformers). Στην παρούσα εργασία προτείνεται ένα υβριδικό σύστημα επεξεργασίας φυσικής γλώσσας, που συνδυάζει γλωσσικά μοντέλα BERTopic, για άρθρα και πολιτικές ομιλίες, καθώς και την ελληνική έκδοση του BERT για ταξινόμηση θεμάτων. Τα κείμενα υποβάλλονται σε επεξεργασία για την εξαγωγή θεμάτων και λέξεων-κλειδιών, επιτρέποντας την οπτικοποίηση των αποτελεσμάτων και διευκολύνοντας την αναζήτηση και ομαδοποίηση δεδομένων. Το προτεινόμενο σύστημα επιτρέπει σε δημοσιογράφους, επαγγελματίες και ερασιτέχνες να συντάσσουν ειδησεογραφικά άρθρα, να υπαγορεύουν κείμενα και να καταγράφουν ομιλίες, με στόχο την εξαγωγή των κύριων θεμάτων που αναγνωρίζονται στο περιεχόμενο τους και την ταξινόμηση τους σε 12 δημοσιογραφικές κατηγορίες.

Λέξεις-κλειδιά: Μοντελοποίηση Θεμάτων, Επεξεργασία Φυσικής Γλώσσας, Ταξινόμηση, BERT, Φορητή Δημοσιογραφία

1. Εισαγωγή

Η ραγδαία πρόοδος στις Τεχνολογίες Πληροφορίας και Επικοινωνιών (ΤΠΕ) έχει αναμορφώσει τον τομέα της δημοσιογραφίας, επηρεάζοντας δραστικά τη δημιουργία,

διανομή και κατανάλωση ειδησεογραφικού περιεχομένου. Η εξάπλωση των ψηφιακών μέσων και φορητών συσκευών έχει αυξήσει τις απαιτήσεις για επεξεργασία μεγάλου όγκου των δεδομένων. Στο πλαίσιο αυτό, η μοντελοποίηση θεμάτων (Topic Modeling - TM) και η ταξινόμηση κειμένων με τη χρήση προηγμένων τεχνολογιών έχουν καταστεί κρίσιμα εργαλεία για την ανάλυση του ειδησεογραφικού περιεχομένου.

Η μοντελοποίηση θεμάτων και η ταξινόμηση κειμένων είναι κρίσιμες τεχνικές επεξεργασίας φυσικής γλώσσας για την ανάλυση μεγάλων όγκων κειμενικών δεδομένων και την εξαγωγή χρήσιμων πληροφοριών. Τα σύγχρονα γλωσσικά μοντέλα μετασχηματιστών, όπως το BERT (Bidirectional Encoder Representations from Transformers), έχουν αποδειχθεί ιδιαίτερα χρήσιμα στην κατεύθυνση αυτή, επιτρέποντας την εξαγωγή των κύριων θεμάτων από κείμενα. Η εφαρμογή αυτών των τεχνολογιών στη δημοσιογραφία διευκολύνει τη σημασιολογική ανάλυση των δεδομένων, γεγονός που επιτρέπει την αποτελεσματική διαχείριση και ανάκτηση πληροφοριών.

1.1 Αντικείμενο της Εργασίας

Η παρούσα εργασία επικεντρώνεται στην ανάπτυξη ενός συστήματος που συνδυάζει παραδοσιακές και σύγχρονες μεθόδους Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) για την ανάλυση δημοσιογραφικού περιεχομένου. Το σύστημα έχει σχεδιαστεί για να εξυπηρετεί επαγγελματίες και ερασιτέχνες δημοσιογράφους, επιτρέποντας την ανάλυση κειμένων και ομιλιών στα Νέα Ελληνικά, καθώς και την εξαγωγή και κατηγοριοποίηση των κύριων θεμάτων που προκύπτουν από αυτά τα κείμενα. Η υλοποίηση περιλαμβάνει τη χρήση δύο μοντέλων BERTopic, ένα για την ανάλυση άρθρων και ένα για την ανάλυση πολιτικών ομιλιών, αξιοποιώντας έναν ελληνικό μετασχηματιστή προτάσεων για την εξαγωγή και ταξινόμηση των θεμάτων. Το σύστημα επιτρέπει την επεξεργασία αυτούσιων κειμένων αλλά και όσων προκύπτουν μέσω φωνητικής εισόδου, με χρήση μετατροπέα ομιλίας σε κείμενο (Speech to Text - STT).

1.2. Στόχος της Εργασίας

Ο κύριος στόχος της εργασίας είναι η δημιουργία ενός ευέλικτου συστήματος που να επιτρέπει στους χρήστες να καταχωρούν περιεχόμενο είτε ως απλό κείμενο είτε μέσω φωνητικής εισόδου και να το επεξεργάζονται για την εξαγωγή και ταξινόμηση θεμάτων. Το σύστημα επιδιώκει να εντοπίζει τα κύρια θέματα που περιλαμβάνονται στο περιεχόμενο και να τα ταξινομεί σε ευρείες δημοσιογραφικές κατηγορίες. Αυτή η ταξινόμηση διευκολύνει τη διαχείριση, την αναζήτηση και την ανάκτηση των πληροφοριών, καθιστώντας τα αρχεία πιο εύκολα διαχειρίσιμα και συνδεδεμένα με άλλα αρχεία παρόμοιου σημασιολογικού περιεχομένου. Επιπλέον, το σύστημα έχει σχεδιαστεί για να είναι δυναμικό και να ενημερώνεται συνεχώς με νέα δεδομένα, επιτρέποντας την εκ νέου εκπαίδευση των μοντέλων του. Αυτό εξασφαλίζει την αξιοπιστία του συστήματος, διευκολύνοντας την παρακολούθηση των αλλαγών στα θέματα με την πάροδο του χρόνου.

1.3. Διάρθρωση της Εργασίας

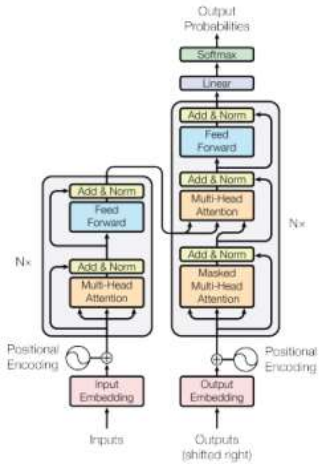
Η εργασία είναι δομημένη σε πέντε κεφάλαια. Στο πρώτο κεφάλαιο παρουσιάζεται το αντικείμενο και ο στόχος της εργασίας. Στο δεύτερο δίνεται όλο το σχετικό θεωρητικό υπόβαθρο που αφορά βασικές έννοιες NLP, με ιδιαίτερη έμφαση στα γλωσσικά μοντέλα και τις αναπαραστάσεις κειμένου, δίνοντας ταυτόχρονα βαρύτητα και σε αλγόριθμους και υβριδικά μοντέλα μοντελοποίησης θεμάτων. Στο τρίτο κεφάλαιο περιγράφεται η μεθοδολογία και ο σχεδιασμός του συστήματος. Στο τέταρτο κεφάλαιο, καταγράφονται τα αποτελέσματα των πειραμάτων και οι αξιολογήσεις τους. Τέλος, στο πέμπτο κεφάλαιο συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται ιδέες για μελλοντικές επεκτάσεις.

2. Επεξεργασία Φυσικής Γλώσσας: θεωρητικό υπόβαθρο και σχετική βιβλιογραφία

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι ένας κλάδος της Τεχνητής Νοημοσύνης που επιδιώκει την κατανόηση και ανάλυση της ανθρώπινης γλώσσας από τους υπολογιστές. Στόχος της είναι η ανάπτυξη αλγορίθμων και μοντέλων που μπορούν να κατανοούν, να επεξεργάζονται και να εξάγουν νόημα από κείμενα και ομιλία. Το NLP συνδυάζει γλωσσικές και υπολογιστικές τεχνικές για να επιλύσει προβλήματα όπως η σύνταξη, η σημασιολογία και η μορφολογία της γλώσσας. Στη βάση του NLP βρίσκονται τα τεχνητά νευρωνικά δίκτυα (ANNs), τα οποία μιμούνται τη λειτουργία του ανθρώπινου εγκεφάλου, αποτελούμενα από νευρώνες συνδεδεμένους σε στρώσεις.

2.1. Γλωσσικά Μοντέλα

Τα γλωσσικά μοντέλα είναι θεμελιώδη για την πρόβλεψη της επόμενης λέξης σε μια πρόταση βασισμένα σε προηγούμενες λέξεις. Το πιο παραδοσιακό μοντέλο είναι το μοντέλο n-gram, ενώ στην συνέχεια προτάθηκαν τα Νευρωνικά Πιθανοτικά Γλωσσικά Μοντέλα για την καλύτερη αναπαράσταση της σημασιολογικής εγγύτητας των λέξεων. Η επανάσταση στον τομέα ήρθε με τα μοντέλα Μετασχηματιστών (Transformers), τα οποία χρησιμοποιούν μηχανισμούς προσοχής (attention mechanisms) για τη βελτίωση της αποτελεσματικότητας και της ακρίβειας στην επεξεργασία γλώσσας. Το μοντέλο Μετασχηματιστή (Εικόνα 1) βασίζεται σε επίπεδα Αυτο-Προσοχής (Self-Attention), επιτρέποντας τη σύνθεση πληροφοριών από ολόκληρη την ακολουθία λέξεων με βελτιωμένη ακρίβεια.



Εικόνα 1: Αρχιτεκτονική του Μετασχηματιστή. Αριστερά είναι ο Κωδικοποιητής και δεξιά ο Αποκωδικοποιητής.

2.2. Διανυσματοποίηση Κειμένου

Η διανυσματοποίηση κειμένου επιτρέπει στους υπολογιστές να κατανοούν κείμενο μετατρέποντάς το σε αριθμητικά διανύσματα. Αρχικά, χρησιμοποιήθηκαν μέθοδοι όπως η One-hot encoding (Kuuluvainen, E, 2023) και το Bag-of-Words (Galke et al., 2021), οι οποίες παρουσιάζουν περιορισμούς λόγω της αραιότητας και της έλλειψης σημασιολογικής αναπαράστασης. Σύγχρονες μέθοδοι περιλαμβάνουν τις ενσωματώσεις λέξεων (word embeddings), όπως τα μοντέλα Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), και FastText (Bojanowski et al., 2017), τα οποία δημιουργούν διανύσματα λέξεων με βάση τη σημασιολογική εγγύτητα και τα συμφραζόμενα. Στην επόμενη γενιά μοντέλων, περιλαμβάνονται οι συμφραζόμενες ενσωματώσεις (contextual embeddings) όπως το ELMo και το BERT. Το BERT είναι ένα προηγμένο μοντέλο βασισμένο στην αρχιτεκτονική των Transformers, η οποία χρησιμοποιεί μηχανισμούς προσοχής (attention mechanisms) για να κατανοήσει τις σχέσεις μεταξύ των λέξεων σε ένα κείμενο. Εισάγει την αμφίδρομη εκπαίδευση, επιτρέποντας την εκτίμηση των συμφραζόμενων από αριστερά προς δεξιά και αντίστροφα, επιτυγχάνοντας κορυφαία αποτελέσματα σε ποικιλία εργασιών.

2.3. Διανυσματικές Αναπαραστάσεις Προτάσεων

Η διανυσματοποίηση κειμένου αναφέρεται στην διαδικασία κατά την οποία λέξεις ή/και φράσεις ενός κειμένου μετατρέπονται σε αριθμητικές αναπαραστάσεις-διανύσματα τα οποία στη συνέχεια μπορούν να υποστούν επεξεργασία από αλγορίθμους μηχανικής μάθησης. Οι κωδικοποιητές προτάσεων, δημιουργούν διανύσματα για προτάσεις, διατηρώντας σημασιολογικές σχέσεις. Το Sentence-BERT, που χρησιμοποιήθηκε στην εργασία, είναι μια τροποποίηση του προ-εκπαιδευμένου BERT που χρησιμοποιεί σιαμαίες και τριπλές δομές δικτύων για την παραγωγή σημασιολογικά ισχυρών ενσωματώσεων

προτάσεων. Επιτρέπει τη σύγκριση προτάσεων και είναι κατάλληλο για εργασίες όπως η αναζήτηση σημασιολογικής ομοιότητας, η ομαδοποίηση και η ανάκτηση πληροφοριών.

Η διανυσματοποίηση για τα Νέα Ελληνικά έχει προχωρήσει σημαντικά, με τη χρήση τεχνολογιών όπως το fastText, το Greek-BERT και τα προεκπαιδευμένα μοντέλα από το HuggingFace.

2.4. Μοντελοποίηση Θεμάτων

Η Μοντελοποίηση Θεμάτων επικεντρώνεται στην ανακάλυψη και οργάνωση των αφηρημένων θεμάτων σε μεγάλα σύνολα κειμένων. Ο κύριος στόχος της είναι η αναγνώριση αυτών των θεμάτων ως λανθάνουσες μεταβλητές και η συσχέτισή τους με συγκεκριμένα έγγραφα, χρησιμοποιώντας ποσοτικούς δείκτες. Τα παραδοσιακά μοντέλα θεμάτων απαιτούσαν από τον χρήστη να καθορίσει εκ των προτέρων τον αριθμό των θεμάτων. Αυτή η προσέγγιση έχει καταστεί ουσιαστικό εργαλείο για την οργάνωση μεγάλων κειμενικών συνόλων, διευκολύνοντας τη δημιουργία βάσεων δεδομένων με κοινά θέματα.

Η Μοντελοποίηση Θεμάτων ξεκίνησε τη δεκαετία του 1980 με την τεχνική TF-IDF, η οποία παρίστανε τα έγγραφα ως διανύσματα σταθερού μήκους, χωρίς να αποκαλύπτει σχέσεις μεταξύ των λέξεων. Στη συνέχεια, αναπτύχθηκαν πιο προηγμένα μοντέλα, όπως η Λανθάνουσα Σημειολογική Ανάλυση (LSA), η Μη Αρνητική Παραγοντοποίηση Πίνακα (NNMF) και η Λανθάνουσα Κατανομή Dirichlet (LDA). Τέλος, διάφορες παραλλαγές του LDA, όπως τα CTM και PAM, προσφέρουν πιο εξελιγμένα χαρακτηριστικά και αποτυπώνουν συσχετίσεις μεταξύ των θεμάτων.

Στη σύγχρονη ανάλυση θεμάτων, δύο σημαντικές μέθοδοι που χρησιμοποιήθηκαν στην εργασία μας είναι το Top2Vec και το BERTopic, λόγω των εξελιγμένων τεχνικών τους για την ανακάλυψη και αναπαράσταση θεμάτων. Εργασίες έχουν δείξει ότι το BERTopic επιτυγχάνει υψηλότερη συνοχή θεμάτων σε σύγκριση με άλλα μοντέλα, όπως το LDA και το NMF (Grootendorst, 2022). Από την άλλη, το Top2Vec ενσωματώνει αυτόματα τα έγγραφα και τις λέξεις στον ίδιο διανυσματικό χώρο, χρησιμοποιώντας το UMAP για μείωση διαστάσεων και το HDBSCAN για συσταδοποίηση, αποφεύγοντας τον καθορισμό προκαθορισμένου αριθμού θεμάτων.

Το Top2Vec (Angelov, 2020) ενσωματώνει αυτόματα έγγραφα και λέξεις στον ίδιο διανυσματικό χώρο και εντοπίζει θεματικές περιοχές χωρίς την ανάγκη προκαθορισμένου αριθμού θεμάτων, χρησιμοποιώντας τεχνικές όπως το UMAP (McInnes et al., 2018) και το HDBSCAN (Campello et al., 2015). Από την άλλη πλευρά, το BERTopic συνδυάζει σύγχρονα γλωσσικά μοντέλα όπως το Sentence-BERT με παραδοσιακές μεθόδους, όπως το TF-IDF, για την αναπαράσταση και εξαγωγή θεμάτων μετά τη συσταδοποίηση. Οι βασικές διαφορές μεταξύ των δύο μοντέλων έγκεινται στην προσέγγιση της αναπαράστασης και εξαγωγής θεμάτων, με το Top2Vec να επικεντρώνεται στην αυτοματοποιημένη ανακάλυψη, ενώ το BERTopic συνδυάζει σύγχρονες και παραδοσιακές τεχνικές για μεγαλύτερη ακρίβεια.

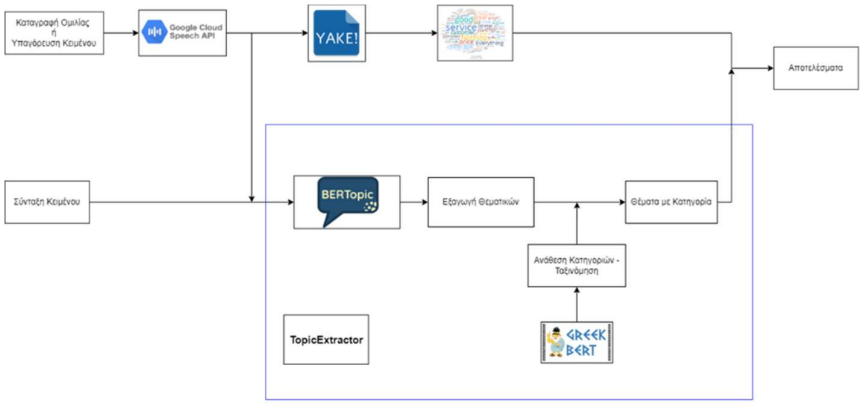
Το Topic Modeling έχει χρησιμοποιηθεί ευρέως σε επιστημονικές μελέτες, ειδικά για την ανάλυση μεγάλων όγκων δεδομένων. Σε αντίθεση με τα παραδοσιακά στατιστικά μοντέλα, η προσέγγιση με BERT επιτρέπει την καλύτερη ενσωμάτωση θεμάτων με βάση τη σημασιολογική εγγύτητα των λέξεων. Εφαρμογές τέτοιων μοντέλων έχουν γίνει κυρίως σε αγγλόφωνα δεδομένα, όπως το σύνολο ειδήσεων 20 NewsGroups και το BBC News (George

L., 2023; Grootendorst, 2020), ενώ οι εφαρμογές στα ελληνικά δεδομένα είναι περιορισμένες, γεγονός που υπογραμμίζει το πρωτοποριακό χαρακτήρα της έρευνάς μας για τα Νέα Ελληνικά.

Συνοψίζοντας, η σύγχρονη Επεξεργασία Φυσικής Γλώσσας συνδυάζει προηγμένες τεχνικές νευρωνικών δικτύων και γλωσσικών μοντέλων για τη βελτίωση της κατανόησης και ανάλυσης κειμένων. Η εφαρμογή αυτών των τεχνικών για την εξαγωγή και ταξινόμηση θεμάτων σε ειδησεογραφικό περιεχόμενο αποτελεί ένα σημαντικό βήμα προς την αυτοματοποίηση και βελτίωση της διαχείρισης πληροφορίας στον τομέα της δημοσιογραφίας και της έρευνας.

3. Μεθοδολογία

Ο βασικός στόχος της εργασίας είναι η ανάπτυξη ενός καινοτόμου συστήματος για την ανάλυση και κατηγοριοποίηση ειδησεογραφικών κειμένων και πολιτικών ομιλιών στα Νέα Ελληνικά. Το σύστημα αυτό, το οποίο ονομάζεται TopicExtractor, στοχεύει στην αυτοματοποίηση της εξαγωγής κύριων θεμάτων από κείμενα και ομιλίες, καθώς και στην ταξινόμηση αυτών των θεμάτων σε προκαθορισμένες δημοσιογραφικές κατηγορίες.



Εικόνα 2: Γενικό Σενάριο Χρήσης και Αρχιτεκτονική Συστήματος TopicExtractor.

Συγκεκριμένα, οι επιμέρους στόχοι περιλαμβάνουν:

- Ανάπτυξη Συστήματος Εξαγωγής Θέματων: Δημιουργία ενός μοντέλου που αναγνωρίζει και εξάγει κύρια θέματα από ειδησεογραφικά άρθρα και πολιτικές ομιλίες.
- Ταξινόμηση Θέματων: Κατηγοριοποίηση των εξαγόμενων θεμάτων σε 12 προκαθορισμένες δημοσιογραφικές κατηγορίες, με εκπαίδευση του Μοντέλου Greek-BERT (Martin et al., 2020) για μεγαλύτερη ακρίβεια στην ταξινόμηση θεμάτων.
- Αξιολόγηση Απόδοσης: Μέτρηση της ακρίβειας και της αποδοτικότητας του συστήματος χρησιμοποιώντας μετρικές όπως η ακρίβεια, η ανάκληση και η τιμή F1.
- Εκτίμηση Ευαισθησίας Αναγνώρισης Ομιλίας: Ανάλυση της αποτελεσματικότητας του συστήματος υπό διάφορες συνθήκες, όπως θόρυβος.

3.1. Κατασκευή Συνόλου Δεδομένων

Η συλλογή δεδομένων πραγματοποιήθηκε από διάφορες ειδησεογραφικές πηγές λόγω της έλλειψης διαθέσιμων συνόλων στα Νέα Ελληνικά. Για την αποτελεσματική συλλογή των δεδομένων αναπτύχθηκε ειδικός αλγόριθμος ανίχνευσης δεδομένων, αξιοποιώντας τη βιβλιοθήκη BeautifulSoup στην Python. Η επιλογή των 8 ειδησεογραφικών ιστοσελίδων εξασφάλισε μια ευρεία κάλυψη θεμάτων και ελαχιστοποίησε την προκατάληψη. Από τα αρχικά 66.999 έγγραφα, διατηρήθηκαν 59.984 μετά τη διαδικασία καθαρισμού, η οποία περιλάμβανε την αφαίρεση διπλοεγγραφών και κενών κειμένων.

Εικόνα 3: Στοιχεία άντλησης και οι ιστοσελίδες κατά την Ιστοσυγκομιδή.

3.2. Μοντελοποίηση Θεμάτων

Η μοντελοποίηση θεμάτων πραγματοποιήθηκε με προηγμένες τεχνικές ανάλυσης. Αρχικά, για την ενσωμάτωση εγγράφων χρησιμοποιήθηκε ο μετασχηματιστής "lighteternal/stsb-xlm-r-greek-transfer," δημιουργώντας διανύσματα 768 διαστάσεων, που στη συνέχεια μειώθηκαν σε 5 μέσω της μεθόδου UMAP, διατηρώντας τη δομή των δεδομένων. Η ομαδοποίηση έγινε με την εφαρμογή του Ιεραρχικού DBSCAN, ο οποίος επέτρεψε την ανίχνευση εξωκείμενων τιμών και απέφυγε την αναγκαστική κατάταξη σε ακατάλληλες συστάδες. Τέλος, η αναπαράσταση των θεμάτων επιτεύχθηκε με τη μέθοδο c-TF-IDF για τον υπολογισμό της σημασίας των λέξεων σε κάθε θέμα, ενώ η μέθοδος MMR χρησιμοποιήθηκε για την αφαίρεση επαναλαμβανόμενων ή παρόμοιων λέξεων.

3.3. Προεπεξεργασία Εισόδου και Παραμετροποίηση

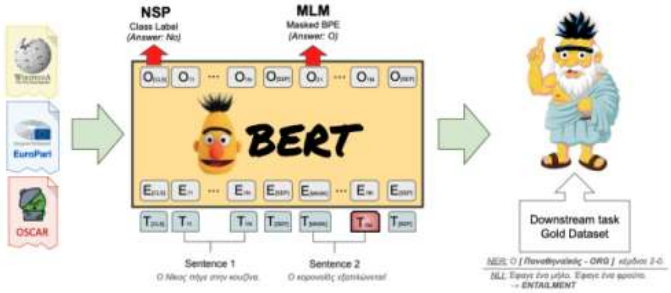
Στην προεπεξεργασία των δεδομένων, έγινε καθαρισμός των κειμένων από ειδικούς χαρακτήρες και περιττά στοιχεία, ενώ τα κείμενα οργανώθηκαν σε παραγράφους για βελτιωμένη ανάλυση. Επιπλέον, χρησιμοποιήθηκε το εργαλείο CountVectorizer με εξειδικευμένες ρυθμίσεις, όπως n-grams και λέξεις τερματισμού, για τη βελτίωση της

θεματικής αναπαράστασης. Η παραμετροποίηση επικεντρώθηκε στη μείωση διαστάσεων με το UMAP, όπου ρυθμίστηκαν βασικές παράμετροι όπως το n_neighbors και το min_dist, και στην ομαδοποίηση με το HDBSCAN, με έμφαση στο min_cluster_size για τον καθορισμό του ελάχιστου μεγέθους συστάδων. Η μέθοδος k-Means εξετάστηκε, αλλά απορρίφθηκε λόγω της φύσης των μη δομημένων δεδομένων.

3.4. Ταξινόμηση Θέματων

Η ταξινόμηση των θεμάτων με χρήση του Μετασχηματιστή BERT είναι συνήθης σε μεγάλη κλίμακα και καταδεικνύει την σημασία της παραμετροποίησης και χρήσης πολλών τεχνικών NLP για τη μείωση του χρόνου εκπαίδευσης χωρίς να θυσιάζεται η ακρίβεια (Nugroho et al., 2021).

Εν προκειμένω, έγινε χρήση του μοντέλου Greek-BERT. Το Greek-BERT, το πρώτο μονογλωσσικό μοντέλο BERT αποκλειστικά για τα Νέα Ελληνικά, εκπαιδεύτηκε σε 29 GB ελληνικών κειμένων και προσφέρει 5-10% βελτίωση σε σχέση με πολυγλωσσικά μοντέλα. Επιπλέον, το HuggingFace παρέχει προεκπαιδευμένους μετασχηματιστές προτάσεων για τα Νέα Ελληνικά, με το "lighteternal/stsb-xlm-r-greek-transfer" να είναι το πρώτο μονογλωσσικό μοντέλο μετασχηματιστή προτάσεων για την ελληνική γλώσσα. Το παραπάνω έχει χρησιμοποιηθεί και σε άλλες εργασίες για ταξινόμηση σε πλατφόρμες κοινωνικής δικτύωσης όπως το Reddit (Mastrokostas C., 2024).



Εικόνα 4: Το προ-εκπαιδευμένο μονογλωσσικό μοντέλο Greek-BERT.

Συνολικά, οι εξελίξεις αυτές, με το Greek-BERT να ξεχωρίζει, βελτιώνουν την απόδοση σε εφαρμογές μηχανικής μάθησης για τις ελληνικές γλωσσικές προκλήσεις. Η διαδικασία περιλάμβανε το fine-tuning του Greek-BERT σε σύνολο δεδομένων με κατανομή 70% για εκπαίδευση και 30% για επικύρωση, ενώ η ανισορροπία κατηγοριών αντιμετωπίστηκε μέσω αντιγραφής των μειοψηφικών δειγμάτων. Το μοντέλο εκπαιδεύτηκε για 10 epochs με συνεχή παρακολούθηση της συνάρτησης κόστους και της ακρίβειας.

3.5. Παρουσίαση Εξαγόμενων Αποτελεσμάτων

Η παρουσίαση των αποτελεσμάτων περιλαμβάνει ανάλυση κειμένων για την εξαγωγή σημείων ενδιαφέροντος, με χρήση του Google Speech-to-Text για την μετατροπή ομιλιών σε κείμενο. Εφαρμόζονται εργαλεία όπως το YAKE! και το SpaCy για την εξαγωγή λέξεων-κλειδιών και την αναγνώριση ονομαστικών οντοτήτων, ενώ το BERTopic API μετατρέπει την

είσοδο σε διανυσματική αναπαράσταση και αξιολογεί τη σημασιολογική ομοιότητα με τα θέματα του μοντέλου, παρουσιάζοντας τα αποτελέσματα μέσω σύννεφων λέξεων.

3.6. Διαχείριση Ανισορροπίας στο Σύνολο Δεδομένων

Η διαχείριση της ανισορροπίας ήταν κρίσιμη για την αποφυγή προκατάληψης του μοντέλου. Εφαρμόστηκαν τεχνικές εξομάλυνσης μέσω απλής αντιγραφής των μειοψηφικών δειγμάτων, αποφεύγοντας την τεχνική SMOTE για την αποφυγή εισαγωγής πλασματικών δεδομένων. Αυτή η προσέγγιση διασφάλισε την ακριβή αξιολόγηση του μοντέλου στο αμετάβλητο σύνολο ελέγχου.

Με αυτές τις τεχνικές και προσεγγίσεις, το TopicExtractor αποδεικνύεται ως ένα ισχυρό εργαλείο για την ανάλυση και κατηγοριοποίηση θεμάτων, συμβάλλοντας ουσιαστικά στη βελτίωση της διαχείρισης και ανάλυσης ειδησεογραφικού περιεχομένου.

4. Αποτελέσματα και Αξιολόγηση Συστήματος

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν με σκοπό την αξιολόγηση και βελτίωση των μοντέλων ενσωμάτωσης κειμένων. Τα αποτελέσματα αξιολογήθηκαν με διάφορες μετρικές και συγκρίθηκαν με άλλες μεθόδους για να αναδειχθούν τα πλεονεκτήματα και μειονεκτήματα.

4.1. Πειράματα και Αποτελέσματα στα Μοντέλα Θεμάτων

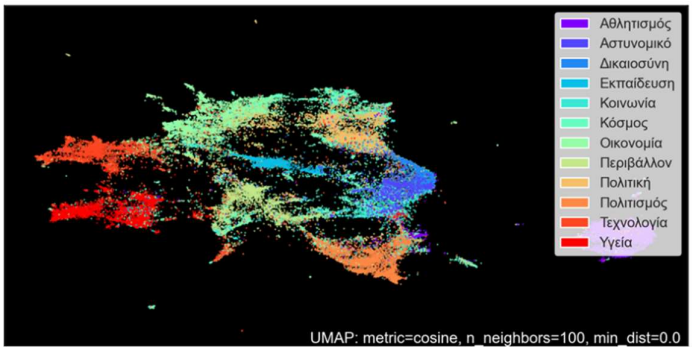
Οι πειραματικές διαδικασίες εκτελέστηκαν κυρίως μέσω της πλατφόρμας Google Colab+ Pro και τοπικά σε υπολογιστές με επεξεργαστές Nvidia A100-SXM4-40GB και Nvidia GTX 1660 Super. Η υπολογιστική ισχύς της Nvidia A100 χρησιμοποιήθηκε κυρίως για τις πιο απαιτητικές διαδικασίες, ενώ οι τοπικοί υπολογιστές παρουσίασαν περιορισμούς όσον αφορά τη μνήμη και την ταχύτητα εκτέλεσης. Για την μελέτη της ομαδοποίησης και της συνοχής των θεμάτων, αξιολογήθηκαν δύο μοντέλα: το πολυγλωσσικό "paraphrasemultilingual-mpnet-base-v2" και το μονογλωσσικό “lighteternal/stsb-xlm-r-greek-transfer”.

Οι μετρικές που αξιολογήθηκαν περιλάμβαναν το DBCV και το Silhouette για την ομαδοποίηση, καθώς και τα NPMI, UMass, και UCI για τη συνοχή των θεμάτων. Η χρήση του CountVectorizer αποδείχθηκε καθοριστική και για τα δύο μοντέλα. Χωρίς αυτό, τα παραγόμενα θέματα περιείχαν λέξεις με περιορισμένη πληροφοριακή αξία. Με την ενσωμάτωση του CountVectorizer, οι μετρικές βελτιώθηκαν σημαντικά.

Πίνακας 1: Σύγκριση Πολυγλωσσικού και Μονογλωσσικού Μοντέλου - Επίδραση CountVectorizer.

Μετρική	Πολυγλωσσικό Μοντέλο		Μονογλωσσικό Μοντέλο	
Χρήση CountVectorizer	Όχι	Ναι	Όχι	Ναι
DBCV	0.2233	0.2439	0.2829	0.2639
Ποικιλία Θεμάτων	0.4488	0.6535	0.2560	0.4041
NPMI (Άρθρα Ειδήσεων)	0.0928	0.2412	0.1480	0.2208
NPMI (Ομιλίες)	0.1105	0.3161	0.1076	0.3152
Συμπέρασμα	Σημαντική βελτίωση στη συνοχή και ποικιλία θεμάτων		Βελτίωση συνοχής με χρήση του CountVectorizer	

Στο πολυγλωσσικό μοντέλο, η χρήση του CountVectorizer βελτίωσε την συνοχή και την ποικιλία των θεμάτων, ενώ στο μονογλωσσικό μοντέλο παρατηρήθηκε αύξηση του NPMI τόσο στα άρθρα ειδήσεων όσο και στις ομιλίες, υποδεικνύοντας επίσης καλύτερη συνοχή.



Εικόνα 5: Οπτικοποίηση κειμένων στον δισδιάστατο χώρο ανά κατηγορία ειδήσεων.

Και στα δύο μοντέλα, η αύξηση των διαστάσεων (n_compontents) δεν βελτίωσε τις μετρικές συνοχής, ενώ επηρέασε αρνητικά τον χρόνο επεξεργασίας. Αντίθετα, η παράμετρος n_neighbors βελτίωσε την ποιότητα των θεμάτων, ειδικά όταν αυξήθηκε σε 25, τόσο στο πολυγλωσσικό όσο και στο μονογλωσσικό μοντέλο.

Πίνακας 2: Επίδραση των n_components και n_neighbors.

Παράμετρος	Πολυγλωσσικό Μοντέλο	Μονογλωσσικό Μοντέλο
n_components (Διαστάσεις)	Η αύξηση από 5 σε 10 δεν βελτίωσε τις μετρικές συνοχής, ενώ επιβράδυνε τον χρόνο εκτέλεσης.	Παρόμοιο αποτέλεσμα: Η αύξηση των διαστάσεων δεν έφερε βελτίωση και επηρέασε την ταχύτητα.
n_neighbors (Γείτονες)	Η καλύτερη απόδοση παρατηρήθηκε με n_neighbors=25, βελτιώνοντας συνοχή και ποικιλία.	Βελτίωση στην συνοχή και ποικιλία με n_neighbors=25 σε σχέση με την προκαθορισμένη τιμή των 15.

Τα αποτελέσματα δείχνουν ότι η αύξηση του n_neighbors σε 25 και για τα δυο μοντέλα βελτίωσε σημαντικά την συνοχή και την ποικιλία των θεμάτων. Από την άλλη, η αύξηση των διαστάσεων (n_components) πέρα από τις 5 δεν έδωσε θετικά αποτελέσματα και επιβάρυνε τον χρόνο επεξεργασίας.

4.2. Ταξινόμηση Ειδήσεων - Πειράματα

Η αξιολόγηση των αλγορίθμων ταξινόμησης έγινε με βάση τις μετρικές: ακρίβεια ταξινόμησης (accuracy), η ακρίβεια (precision), ανάκληση (recall) και F1-score. Δοκιμάστηκαν διάφορα μοντέλα, συμπεριλαμβανομένων της Λογιστικής Παλινδρόμησης (Regression), των Δένδρων Απόφασης (Decision Trees), Random Forest και Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM). Η κωδικοποίηση των δεδομένων έγινε με δύο μεθόδους: Tf-Idf και doc2vec.

Πίνακας 3: Απόδοση μοντέλων με κωδικοποίηση Tf-Idf.

Μοντέλο	Ακρίβεια	F1-score (Αθλητισμός)	F1-score (Πολιτισμός)	F1-score (Κοινωνία)	F1-score (Μέσος Όρος)
Λογιστική Παλινδρόμηση	71%	0.83	0.81	0.58	0.69
Δένδρα Απόφασης	58%	0.72	0.68	0.41	0.55
Random Forest	72%	0.79	0.81	0.70	0.69
SVMs	77%	0.84	0.84	0.64	0.74

Σύμφωνα με την κωδικοποίηση Tf-Idf, το μοντέλο SVM πέτυχε την καλύτερη απόδοση με ακρίβεια 77%, παρουσιάζοντας εξαιρετικά αποτελέσματα στις κατηγορίες "Αθλητισμός" και "Πολιτισμός". Αντίθετα, τα Δένδρα Απόφασης είχαν τη χαμηλότερη απόδοση με 58% ακρίβεια.

Πίνακας 4: Απόδοση μοντέλων με κωδικοποίηση doc2vec.

Μοντέλο	Ακρίβεια	F1-score (Αθλητισμός)	F1-score (Πολιτισμός)	F1-score (Κοινωνία)	F1-score (Μέσος Όρος)
Λογιστική Παλινδρόμηση	81%	0.94	0.85	0.67	0.79
Δένδρα Απόφασης	55%	0.67	0.84	0.47	0.46
Random Forest	65%	0.79	0.72	0.50	0.49
SVMs	84%	0.95	0.87	0.73	0.82

Με την κωδικοποίηση doc2vec, το μοντέλο Λογιστικής Παλινδρόμησης είχε σημαντική βελτίωση στην ακρίβεια (81%), όμως το SVM απέδωσε καλύτερα με 84% ακρίβεια.

Τέλος, το μοντέλο Greek-BERT είχε την καλύτερη απόδοση με ακρίβεια 88% στο σύνολο επικύρωσης, ξεπερνώντας τα άλλα μοντέλα, καθιστώντας το το πιο αποδοτικό για την ταξινόμηση ειδήσεων στα Νέα Ελληνικά.

Πίνακας 5: Απόδοση μοντέλου Greek-BERT.

Μοντέλο	Ακρίβεια	F1-score (Αθλητισμός)	F1-score (Πολιτισμός)	F1-score (Κοινωνία)	F1-score (Μέσος Όρος)
Greek-BERT	88%	0.96	0.89	0.81	0.85

Εικόνα 6: Πίνακας Σύγχυσης του μοντέλου Greek-BERT για την διεργασία της ταξινόμησης.

4.3. Πειράματα Απώλειας λόγω Θορύβου

Ο λόγος σήματος προς θόρυβο (SNR) είναι ένα κρίσιμο μέτρο που καθορίζει την ποιότητα του σήματος σε σχέση με τον θόρυβο. Ένα υψηλό SNR σημαίνει καθαρότερο σήμα και βελτιωμένη ακρίβεια αναγνώρισης ομιλίας. Η σχέση μεταξύ SNR και χωρητικότητας καναλιού, όπως περιγράφεται από το θεώρημα Shannon-Hartley, αποδεικνύει ότι η αύξηση του SNR βελτιώνει την απόδοση του καναλιού επικοινωνίας.

Με την βοήθεια της υπηρεσίας Google Cloud Speech-To-Text πραγματοποιήθηκαν δυο πειράματα. Στο πρώτο πείραμα, πραγματοποιήθηκε μέτρηση του Word Error Rate (WER) σε ομιλία του Πρωθυπουργού στην Διεθνή Έκθεση Θεσσαλονίκης, με διάρκεια 30 λεπτών, υπό συνθήκες SNR = 20 dB. Το πρωτότυπο κείμενο της ομιλίας συγκρίθηκε με την καταγραφή που παρήγαγε το σύστημα. Τα αποτελέσματα της μέτρησης έδειξαν ότι υπήρχαν 19 εισαγωγές (I), 89 αντικαταστάσεις (S), και 133 διαγραφές (D), με συνολικό αριθμό λέξεων (N) 2509. Ο τελικός συντελεστής WER υπολογίστηκε στο 9.6%.

Πίνακας 6: Αλλαγές στα θέματα και τις κατηγορίες σε περίπτωση απώλειας λόγω θορύβου.

	Αλλαγή Θέματος	Αλλαγή Κατηγορίας	Ποσοστό Λάθος (%)
5 %	16	1	0.016
10 %	54	7	0.054
15 %	186	24	0.186

Στο δεύτερο πείραμα (Πίνακας 6), εξετάστηκε η επίδραση του θορύβου στις επιδόσεις του συστήματος, χρησιμοποιώντας 10.000 παράγραφους δεδομένων. Για κάθε παράγραφο, συγκρίθηκαν οι ground-truth ετικέτες με τις προβλέψεις του μοντέλου σε τρία επίπεδα

απώλειας: 5%, 10%, και 15%. Η απώλεια εισάγεται κατά 60% ως διαγραφή (D), 30% ως αντικατάσταση (S), και 10% ως εισαγωγή (I). Ωστόσο, όπως φαίνεται στην Εικόνα 7, τα αποτελέσματα του συστήματος δεν επηρεάζονται στην έξοδό του.

5. Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η παρούσα εργασία ξεχωρίζει για την πρωτοτυπία της, καθώς συνδυάζει σύγχρονες μεθόδους ταξινόμησης κειμένων και μοντελοποίησης θεμάτων με την εφαρμογή των γλωσσικών μοντέλων μετασχηματιστών για τα Νέα Ελληνικά. Ειδικότερα, η καινοτομία έγκειται στη συνδυασμένη χρήση κειμενικών και φωνητικών εισόδων, μέσω της ενσωμάτωσης ενός συστήματος αναγνώρισης φωνής για την επεξεργασία της ομιλίας. Η υλοποίηση αυτής της προσέγγισης επιτρέπει την αξιολόγηση και την επεξεργασία της ομιλίας σε πραγματικό χρόνο, κάτι που δεν είχε εξεταστεί σε παρόμοιες έρευνες για τα Νέα Ελληνικά. Η συνεχής ανανέωση των συνόλων δεδομένων και η επανεκπαίδευση των μοντέλων, που επιτρέπει στο σύστημα να προσαρμόζεται δυναμικά στις νέες εξελίξεις στην ειδησεογραφία, αποτελεί στοχευμένη καινοτομία της εργασίας και διαφοροποιεί το σύστημα TopicExtractor από τα παραδοσιακά συστήματα που βασίζονται σε στατικά δεδομένα.

Η συνεισφορά της εργασίας είναι διπλή: πρώτον, η ανάπτυξη ενός συστήματος που ενσωματώνει προηγμένα γλωσσικά μοντέλα και αναγνώριση φωνής, και δεύτερον, η ανακάλυψη σημαντικών ευρημάτων μέσω των πειραμάτων που επιβεβαιώνουν την αξιοπιστία των μετρικών αξιολόγησης της μοντελοποίησης θεμάτων.

5.1. Συμπεράσματα και Σχολιασμός Αποτελεσμάτων

Από την ανάλυση των πειραματικών αποτελεσμάτων προκύπτουν αρκετά σημαντικά συμπεράσματα. Η μετρική DBCV αποδείχθηκε ότι υπερέχει της silhouette ως κριτήριο για την καλή συσταδοποίηση με την μέθοδο του Ιεραρχικού DBSCAN. Αν και οι δύο μετρικές δεν συμφωνούν πάντα μεταξύ τους, η DBCV παρέχει πιο αξιόπιστες ενδείξεις για την αποτελεσματικότητα της συσταδοποίησης.

Όσον αφορά τα μέτρα συνοχής, οι μετρικές NPMI και Umass αποτυπώνουν καλύτερα την αποτελεσματικότητα της μοντελοποίησης θεμάτων, υποδεικνύοντας ότι η μέθοδος BERTopic επιτυγχάνει υψηλής ποιότητας θεματικές αναπαραστάσεις χωρίς ανάγκη εκτενούς παραμετροποίησης. Η συγκριτική ανάλυση των μονόγλωσσων και πολυγλωσσικών μετασχηματιστών έδειξε ότι ο μονόγλωσσος μετασχηματιστής 'lighteternal/stsb-xlm-r-greek-transfer' προσφέρει καλύτερα αποτελέσματα για τα Νέα Ελληνικά.

Η εργασία επανεξετάζει επίσης τη σημασία της παραμετροποίησης, επισημαίνοντας ότι η ελαχιστοποίηση του θορύβου δεν εγγυάται απαραίτητα ένα καλύτερο μοντέλο και ότι οι επιλογές παραμέτρων όπως οι min_samples και min_cluster_size μπορούν να επηρεάσουν δραστικά τα αποτελέσματα.

5.2. Μελλοντικές Επεκτάσεις

Για την περαιτέρω εξέλιξη της εργασίας, προτείνονται αρκετές κατευθύνσεις για μελλοντική έρευνα:

- Νέες Μέθοδοι Συσταδοποίησης: Εξετάστε την εφαρμογή εναλλακτικών μεθόδων συσταδοποίησης, όπως οι μέθοδοι πυκνότητας (π.χ. BIRCH, OPTICS), για την πιθανή βελτίωση της απόδοσης στην μοντελοποίηση θεμάτων.
- Ενίσχυση Συστήματος Αναγνώρισης Φωνής: Διερευνήστε την εφαρμογή των γλωσσικών μοντέλων για την πρόβλεψη χαμένων λέξεων ή τη διόρθωση λέξεων που έχουν τοποθετηθεί εσφαλμένα στις προτάσεις μέσω της τεχνικής Fill-Mask. Αυτό μπορεί να βελτιώσει την αξιοπιστία των συστημάτων αναγνώρισης φωνής.
- Ενσωμάτωση Εικόνας και Συναισθηματικής Ανάλυσης: Εξετάστε την αξιοποίηση εικόνων για σημασιολογική ανάλυση κειμένων και ομιλίας, όπως η παραγωγή λεζάντας εικόνων για κατηγοριοποίηση άρθρων και ανάλυση εικόνας ανά καρέ σε βίντεο. Επίσης, η συνδυασμένη ανάλυση συναισθήματος στην ομιλία και διάκριση ομιλητών μπορεί να προσφέρει νέες δυνατότητες στην ανάλυση συνεντεύξεων.
- Εφαρμογή σε Κινητές Συσκευές: Η τελική ενσωμάτωση της υλοποίησης σε εφαρμογές κινητών μπορεί να ενισχύσει την πρακτική χρησιμότητα του συστήματος, παρέχοντας στους χρήστες (π.χ., δημοσιογράφους) πιο προηγμένα εργαλεία ανάλυσης και κατηγοριοποίησης περιεχομένου.

Συμπερασματικά, η εργασία αυτή ανοίγει το δρόμο για περαιτέρω έρευνα και ανάπτυξη στον τομέα της μοντελοποίησης θεμάτων και αναγνώρισης φωνής, προτείνοντας νέες προσεγγίσεις και τεχνολογίες που θα μπορούσαν να βελτιώσουν σημαντικά την ακρίβεια και την αποτελεσματικότητα τέτοιων συστημάτων.

Αναφορές

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *10*(1), 1-51.

Galke, L., & Scherp, A. (2021). Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. *arXiv preprint arXiv:2109.03777*.

George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology*, *15*(4), 2187-2195.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020, September). Greek-bert: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence* (pp. 110-117).

Kuuluvainen, E. (2023). Classifying news articles based on user needs using transfer learning and deep neural networks: a multi-class approach combining BERT with non-textual features.

Mastrokostas, C., Giarelis, N., & Karacapilidis, N. (2024). Social Media Topic Classification on Greek Reddit. *Information*, *15*(9), 521.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Nugroho, K. S., Sukmadewa, A. Y., & Yudistira, N. (2021, September). Large-scale news classification using BERT language model: Spark NLP approach. In *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology* (pp. 240-246).

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).