ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΕΦΑΡΜΟΓΗ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΔΕΔΟΜΕΝΑ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΑΝΟΙΚΤΗΣ ΠΡΟΣΒΑΣΗΣ»

Του Φοιτητή
Τροκάνα Κωνσταντίνου
Αρ. Μητρώου: 185418

Επιβλέπων
Μπράτσας Χαράλαμπος
Επικ. Καθηγητής

Μάιος 2025

Τίτλος Δ.Ε.: Εφαρμογή Ανάλυσης Συναισθήματος Σε Δεδομένα Κοινωνικών Δικτύων Ανοικτής Πρόσβασης

Κωδικός Δ.Ε.: 24281

Ονοματεπώνυμο φοιτητή: Τροκάνας Κωνσταντίνος

Ονοματεπώνυμο εισηγητή: Μπράτσας Χαράλαμπος

Ημερομηνία ανάληψης Δ.Ε: 29-11-2024

Ημερομηνία περάτωσης Δ.Ε: 15-5-2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Τροκάνα Κωνσταντίνου που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Αφιερώνω αυτή την εργασία στους γονείς μου, για την αμέριστη στήριξη και την πίστη τους σε εμένα..»

Πρόλογος

Η παρούσα πτυχιακή εργασία προέκυψε από το προσωπικό μου ενδιαφέρον για την τεχνητή νοημοσύνη και τη μελέτη της ανθρώπινης συμπεριφοράς μέσα από δεδομένα που παράγονται καθημερινά στα κοινωνικά δίκτυα. Το έργο του Μετρό Θεσσαλονίκης, ένα από τα πιο πολυσυζητημένα έργα υποδομής στην Ελλάδα, αποτέλεσε ιδανικό πεδίο ανάλυσης, καθώς συνδυάζει τεχνικό ενδιαφέρον, κοινωνική επίδραση και έντονη δημόσια συζήτηση. Μέσα από αυτή την εργασία είχα την ευκαιρία να εμβαθύνω στις τεχνικές της ανάλυσης συναισθήματος, να εφαρμόσω σύγχρονα εργαλεία μηχανικής μάθησης και να εξοικειωθώ με μεθοδολογίες συλλογής, επεξεργασίας και οπτικοποίησης δεδομένων. Πέρα από τις τεχνικές δεξιότητες που απέκτησα, η εργασία αυτή με βοήθησε να αντιληφθώ σε βάθος πώς οι κοινωνικές αντιδράσεις αποτυπώνονται και μεταβάλλονται στο ψηφιακό περιβάλλον, προσφέροντας έτσι ένα πολύτιμο εφόδιο για την περαιτέρω ακαδημαϊκή και επαγγελματική μου πορεία.

Περίληψη

Η παρούσα πτυχιακή εργασία ασχολείται με την εφαρμογή τεχνικών ανάλυσης συναισθήματος (sentiment analysis) σε ελληνικά και αγγλικά δεδομένα από το κοινωνικό δίκτυο Reddit, με επίκεντρο τις διαδικτυακές αντιδράσεις των χρηστών σχετικά με το Μετρό Θεσσαλονίκης. Αξιοποιήθηκαν σύγχρονες τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP), τόσο λεξικο-βασιζόμενες (rule-based) μέθοδοι όσο και προηγμένα νευρωνικά μοντέλα όπως το πολυγλωσσικό BERT, για την εξαγωγή και ερμηνεία του συναισθηματικού περιεχομένου των σχολίων.

Η ανάλυση πραγματοποιήθηκε σε σύνολο 5.323 δεδομένων, τα οποία προεπεξεργάστηκαν με χρήση εργαλείων όπως τα spaCy και langdetect, ενώ αφαιρέθηκαν bots και duplicate εγγραφές για καθαρότερο αποτέλεσμα. Τα ευρήματα έδειξαν πως το 62% των σχολίων χαρακτηρίστηκαν ως ουδέτερα, το 25% ως αρνητικά και μόλις το 13% ως θετικά. Οι συναισθηματικές μεταβολές απεικονίστηκαν σε διαγράμματα χρονικής εξέλιξης, με εντοπισμό αιχμών (spikes) που αντιστοιχούν σε σημαντικά γεγονότα, όπως το δυστύχημα στα Τέμπη ή η ανακοίνωση της ημερομηνίας εγκαινίων του μετρό.

Επιπλέον, δημιουργήθηκαν wordclouds ανά κατηγορία συναισθήματος, heatmaps, boxplots και πίνακες με τις σημαντικότερες λέξεις που συνδέονται με κάθε συναίσθημα. Η μελέτη ανέδειξε τη χρησιμότητα των κοινωνικών δεδομένων ως εργαλείο κατανόησης της δημόσιας στάσης απέναντι σε μεγάλα έργα υποδομής και πρότεινε δυνατότητες επέκτασης της μεθοδολογίας σε επιπλέον πλατφόρμες και μοντέλα για μελλοντική έρευνα.

«Sentiment Analysis on Open Access Social Network Data»

Konstantinos Trokanas

Abstract

This thesis explores the application of sentiment analysis techniques on open access social network data, focusing on user reactions to the Thessaloniki Metro project in Greece. Reddit was selected as the main data source due to its open API and rich user-generated content. A total of 5,323 comments, written in both Greek and English, were collected, preprocessed, and analyzed. The preprocessing pipeline included language detection, duplicate removal, bot filtering, and text normalization using NLP tools such as spaCy and langdetect.

The sentiment analysis was conducted using two approaches: a rule-based method relying on predefined sentiment lexicons, and a machine learning-based method using the multilingual BERT transformer model. Results showed that 62% of the comments were neutral, 25% negative, and 13% positive. Temporal sentiment variations were visualized and associated with key events such as the Tempi train accident and the metro inauguration announcement. Word clouds and statistical tables highlighted the most frequent and emotionally charged terms across sentiment categories.

The findings reveal that online sentiment is closely tied to real-world events and that advanced language models such as BERT can effectively capture the nuances of public opinion across languages. This study demonstrates how sentiment analysis of social media can serve as a valuable tool for understanding public perception of large-scale infrastructure projects and suggests avenues for future research across different platforms and populations.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Μπράτσα, για την πολύτιμη καθοδήγηση, την εμπιστοσύνη και τη διαρκή στήριξη καθ’ όλη τη διάρκεια της πτυχιακής εργασίας. Η επιστημονική του καθοδήγηση υπήρξε καταλυτική στη διαμόρφωση του θέματος και στην ολοκλήρωση της παρούσας μελέτης.

Ευχαριστώ επίσης την οικογένειά μου, που στάθηκε δίπλα μου με υπομονή, κατανόηση και άνευ όρων στήριξη, σε κάθε βήμα αυτής της διαδρομής.

Τέλος, ευχαριστώ τους φίλους και τους συνεργάτες μου, για τις συζητήσεις, τις ιδέες, αλλά και τη συντροφιά στις στιγμές που τη χρειαζόμουν περισσότερο.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία
API	Application Programming Interface
PRAW	Python Reddit API Wrapper
PSAW	Pushshift API Wrapper for Python
NLP	Natural Language Processing
NB	Naive Bayes
POS	Part of Speech
CNN	Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
JSON	JavaScript Object Notation
XLM-R	Cross-lingual Language Model RoBERTa
mBERT	Multilingual BERT
GPT	Generative Pre-trained Transformer
HTML	HyperText Markup Language
URL	Uniform Resource Locator
NLTK	Natural Language Toolkit

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η παρούσα εργασία στοχεύει στη συλλογή και ανάλυση ελληνόγλωσσου περιεχομένου από το Reddit που αφορά το Μετρό Θεσσαλονίκης, προκειμένου να διερευνηθεί η διακύμανση του αισθήματος του κοινού απέναντι στο έργο αυτό. Κεντρικά ερευνητικά ερωτήματα αποτελούν: Ποια είναι η επικρατούσα συναισθηματική στάση (θετική, αρνητική, ουδέτερη) των Ελλήνων χρηστών του Reddit σχετικά με το Μετρό Θεσσαλονίκης; Πώς εξελίσσεται η συναισθηματική αυτή στάση διαχρονικά και σε συνάρτηση με σημαντικά γεγονότα (π.χ. μεγάλες καθυστερήσεις, ανακοινώσεις, εγκαίνια, δυστήχημα στα Τέμπη); Επίσης, διερευνώνται οι διαφορές μεταξύ διαφορετικών μεθόδων ανάλυσης συναισθήματος δηλαδή μεταξύ λεξιλογικών/rule-based τεχνικών και μεθόδων μηχανικής μάθησης ως προς την ακρίβεια και τα ευρήματα που παράγουν για τα συγκεκριμένα δεδομένα. Με τον τρόπο αυτό, η μελέτη θα συμβάλει στην κατανόηση της δημόσιας αντίληψης για ένα εμβληματικό έργο μέσω της ανάλυσης περιεχομένου social media, αναδεικνύοντας ταυτόχρονα προκλήσεις και βέλτιστες πρακτικές στην ανάλυση συναισθήματος σε ελληνικά κείμενα.

Το Μετρό Θεσσαλονίκης αποτελεί ένα από τα σημαντικότερα και πολυαναμενόμενα έργα υποδομής στην Ελλάδα, καθώς είναι το πρώτο αυτοματοποιημένο (χωρίς οδηγό) σύστημα μετρό στη χώρα[1]. Η κατασκευή της βασικής γραμμής ξεκίνησε το 2006, όμως το έργο αντιμετώπισε επανειλημμένες καθυστερήσεις λόγω ποικίλων προβλημάτων από την οικονομική κρίση και την πτώχευση εργολάβων, μέχρι τις αρχαιολογικές ανασκαφές που έφεραν στο φως πάνω από 300.000 ευρήματα[1]. Μετά από πολύχρονη αναμονή και συνεχείς αλλαγές στον προγραμματισμό, η βασική γραμμή του μετρό παραδόθηκε τελικά προς χρήση στις 30 Νοεμβρίου 2024[1]. Η εξέλιξη αυτή σηματοδότησε μια ιστορική στιγμή για τη Θεσσαλονίκη, καθώς ένα έργο-σύμβολο δεκαετιών έγινε πραγματικότητα, με προσδοκίες να βελτιώσει σημαντικά τις μετακινήσεις και την καθημερινότητα των πολιτών.

Παράλληλα, η μακρά διάρκεια υλοποίησης του μετρό γέννησε έντονες αντιδράσεις και συζητήσεις στην κοινωνία. Οι πολίτες εξέφρασαν επανειλημμένα τις απόψεις, τα συναισθήματα και την απογοήτευσή τους για τις καθυστερήσεις, συχνά με χιούμορ ή σαρκασμό. Τα μέσα κοινωνικής δικτύωσης και ειδικά πλατφόρμες όπως το Reddit λειτούργησαν ως βήμα δημόσιας διαβούλευσης και σχολιασμού όλα αυτά τα χρόνια. Όπως συμβαίνει γενικότερα, τα social media έχουν αναδειχθεί σε τεράστια αποθετήρια δημόσιας γνώμης για κάθε θέμα, όπου οι άνθρωποι μοιράζονται καθημερινά σκέψεις και συναισθήματα για τοπικά και παγκόσμια ζητήματα. Η σωστή διαχείριση και ανάλυση τέτοιων δεδομένων μπορεί να αποκαλύψει ενδιαφέροντα μοτίβα, τάσεις και αντιλήψεις του κοινού.

Κεφάλαιο 2ο: Ανασκόπηση Βιβλιογραφίας

2.1 Εισαγωγή

Η ανάλυση συναισθήματος (sentiment analysis) είναι η διαδικασία αυτόματης ανίχνευσης και ταξινόμησης της συναισθηματικής φόρτισης που εκφράζεται σε ένα κείμενο συνήθως ως θετικό, αρνητικό ή ουδέτερο συναίσθημα[2]. Τα τελευταία χρόνια, με την εκρηκτική άνοδο των social media, η ανάλυση συναισθήματος έχει καταστεί κρίσιμη για την εξαγωγή γνώσης από τεράστιους όγκους ελληνικών αναρτήσεων σε πλατφόρμες όπως το Twitter, το Reddit, το Telegram και το Mastodon. Η παρούσα επισκόπηση επικεντρώνεται σε τεχνικές που εφαρμόζονται στην ελληνική και αγγλική γλώσσα και σε δεδομένα κοινωνικών δικτύων, με ιδιαίτερη έμφαση στην αξιοποίησή τους για ανάλυση συναισθήματος σε αναρτήσεις σχετικά με το Μετρό Θεσσαλονίκης. Θα παρουσιαστούν διεξοδικά: (α) οι κύριες λεξικογραφικές (rule-based) μέθοδοι και η προσαρμογή τους στα ελληνικά, (β) οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης και οι σύγχρονες προσεγγίσεις βαθιάς μάθησης, (γ) τα προεκπαιδευμένα μοντέλα γλώσσας και μεγάλα γλωσσικά μοντέλα (LLMs), (δ) διαθέσιμα εργαλεία NLP για επεξεργασία κειμένων, και (ε) ενδεικτικές εφαρμογές μελέτες περίπτωσης σε ελληνικά και διεθνή δεδομένα social media.

2.2 Λεξικογραφικές (Rule-Based) Τεχνικές Ανάλυσης Συναισθήματος

Οι λεξικογραφικές μέθοδοι βασίζονται σε προκαθορισμένα λεξικά συναισθήματος και κανόνες για τον προσδιορισμό της πολικότητας ενός κειμένου. Ένα λεξικό συναισθήματος είναι μια συλλογή από λέξεις (ή φράσεις) στις οποίες έχει αποδοθεί εκ των προτέρων μια συναισθηματική τιμή ή επισήμανση (θετικό, αρνητικό, ουδέτερο συναίσθημα κ.ά.). Καθώς το λεξικό περιέχει μόνο ένα μέρος του συνολικού λεξιλογίου, η κάλυψη είναι περιορισμένη[2]. Παρόλα αυτά, τέτοιες προσεγγίσεις είναι δημοφιλείς επειδή δεν απαιτούν ανεπτυγμένο σύνολο δεδομένων για εκπαίδευση και είναι εύκολα ερμηνεύσιμες.

2.2.1 VADER

Μία από τις πιο γνωστές rule-based μέθοδοι είναι το VADER (Valence Aware Dictionary and sEntiment Reasoner). Πρόκειται για ένα λεξικό και σύνολο κανόνων ειδικά προσαρμοσμένων στη γλώσσα των social media. Το VADER αναγνωρίζει, εκτός από τις συναισθηματικά φορτισμένες λέξεις, στοιχεία όπως τα emoji, τις συντομογραφίες και την έμφαση με κεφαλαία, αποδίδοντας μια βαθμολογία συναισθήματος σε ένα κείμενο. Το αγγλικό λεξικό του VADER περιλαμβάνει 7.520 λέξεις με αντίστοιχες βαθμολογίες. Για την ελληνική γλώσσα, το VADER δεν διαθέτει επίσημη έκδοση· ωστόσο, έχει επιχειρηθεί η μετάφραση του αγγλικού λεξικού στα ελληνικά. Έρευνες δείχνουν ότι μια απλή μετάφραση του λεξικού του VADER δεν επαρκεί, διότι δεν λαμβάνει υπόψη τις μορφολογικές ιδιαιτερότητες των ελληνικών. Συγκεκριμένα, στην αγγλική γλώσσα λέξεις όπως “love” παραμένουν αμετάβλητες σε διαφορετικά πρόσωπα (“I love, you love, he loves”), ενώ στα ελληνικά εμφανίζονται σε ποικίλες μορφές: «αγαπώ, αγαπάς, αγαπά, αγαπούμε, αγαπάτε, αγαπούν». Αν το λεξικό δεν περιλαμβάνει όλες αυτές τις κλίσεις ή δεν εφαρμοστεί αναγνώριση μορφής (lemmatization/stemming), τότε λέξεις όπως «αγαπάς» ή «αγαπούν» δεν θα αναγνωριστούν ως σχετικές με τη βάση «αγαπώ». Έτσι, η προσαρμογή του VADER στα ελληνικά απαιτεί είτε τον εμπλουτισμό του λεξικού με κλίσεις και παράγωγα (κανόνες για καταλήξεις, προθήματα κ.λπ.) είτε την προεπεξεργασία των κειμένων (π.χ. με μορφολογική ανάλυση ώστε οι λέξεις να αναχθούν στη βασική τους μορφή πριν από την ανάλυση).

2.2.2 SentiWordNet

Ένα άλλο λεξικογραφικό εργαλείο είναι το SentiWordNet, το οποίο βασίζεται στο WordNet. Για κάθε σημασιολογική έννοια (synset) ορίζει βαθμολογία θετικού, αρνητικού και ουδέτερου συναισθήματος. Αν και το SentiWordNet αναπτύχθηκε για την αγγλική WordNet, έχουν γίνει προσπάθειες επέκτασής του και σε άλλες γλώσσες. Στην περίπτωση των ελληνικών, όπου υπάρχει Ελληνικό WordNet, θεωρητικά μπορεί να γίνει αντιστοίχιση των ελληνικών συνωνύμων με τα αγγλικά synsets του SentiWordNet. Ωστόσο, η ευθυγράμμιση αυτή δεν είναι απλή και η αποτελεσματικότητα ενός τέτοιου εγχειρήματος ποικίλλει ανάλογα με την ποιότητα του λεξιλογίου και της μετάφρασης. Στη σύγχρονη βιβλιογραφία, το SentiWordNet χρησιμοποιείται κυρίως ως συμπληρωματικό εργαλείο ή για σύγκριση.

2.2.3 Ελληνικά λεξικά συναισθήματος

Στην τελευταία δεκαετία έχουν αναπτυχθεί λεξικά ειδικά για την ελληνική γλώσσα, ώστε να υποστηρίξουν κανόνες ανάλυσης συναισθήματος. Ένα από τα πρώτα ήταν το λεξικό του Γ. Καλαματιανού και άλλων (2015), γνωστό και ως Greek Sentiment Lexicon, το οποίο περιλαμβάνει ~2.315 λήμματα (λέξεις) με μεταδεδομένα όπως τόνο (θετικό/αρνητικό), μέρος του λόγου, βαθμό αντικειμενικότητας και κατηγορία συναισθήματος[18]. Το λεξικό αυτό συνοδευόταν και από ένα σύνολο ελληνικών tweets αξιολογημένων ως προς την ένταση του συναισθήματος, συμβάλλοντας στην πρώτη προσπάθεια δημιουργίας πόρων για ελληνικό social media sentiment analysis. Πιο πρόσφατα, ο Τσακαλίδης και άλλοι (2018) παρουσίασαν ένα εκτενές λεξικό συναισθήματος για τα ελληνικά με 2.316 όρους, όπου για κάθε λέξη δίνονται το μέρος του λόγου, η υποκειμενικότητα, η πολικότητα (θετική/αρνητική) και η ένταση έξι βασικών συναισθημάτων[19]. Αυτό το λεξικό εμπλουτίζει την απλή θετική/αρνητική διάσταση με πληροφορία για συναισθήματα όπως χαρά, λύπη, φόβο, θυμό κ.ά., γεγονός πολύ χρήσιμο όταν θέλουμε πιο λεπτομερή ανάλυση (π.χ. αν το συναίσθημα είναι θυμός ή φόβος, και όχι απλώς “αρνητικό”).

Ένα ενδιαφέρον εγχείρημα προσαρμογής λεξικού στα ελληνικά είναι το έργο του Νικόλαου Κρυστάλλη (2020), που αξιοποίησε το λεξικό του Τσακαλίδη και το επέκτεινε με κανόνες μορφολογίας. Συγκεκριμένα, πρόσθεσε 287 καταλήξεις και προθήματα, επιτρέποντας στο λεξικό να αναγνωρίζει λέξεις σε διάφορες κλίσεις[20]. Έτσι, παρότι το βασικό λεξικό είχε ~2.300 λήμματα, με τις καταλήξεις μπορούσε να “πιάσει” σχεδόν 3.000 διαφορετικές λέξεις-μορφές[20]. Αυτή η πρακτική βελτιώνει την κάλυψη, μειώνοντας το πρόβλημα που περιγράφηκε παραπάνω με το παράδειγμα του «αγαπώ/αγαπάς».

2.3 Αποτελεσματικότητα και χρήση

Οι rule-based μέθοδοι συνήθως υπολογίζουν ένα συνολικό σκορ συναισθήματος για ένα κείμενο βάσει των λεξικών. Π.χ. μετράνε τις θετικές και αρνητικές λέξεις (ίσως και την έντασή τους) και εφαρμόζουν κάποιους κανόνες οροσήμων (π.χ. για άρνηση: “δεν χαίρομαι” → αντιστρέφει το συναίσθημα). Στα ελληνικά, αυτές οι τεχνικές έχουν χρησιμοποιηθεί ως βασικές γραμμές (baselines) σε μελέτες ή σε συνδυασμό με άλλες μεθόδους. Για παράδειγμα, σε ανάλυση ελληνικών tweets για τον COVID-19, συγκρίθηκαν τα αποτελέσματα συναισθήματος χρησιμοποιώντας το λεξικό Vader (μεταφρασμένο στα ελληνικά) και το ελληνικό λεξικό του Ν. Κρυστάλλη[20]. Διαπιστώθηκε ότι η ενσωμάτωση μορφολογικής γνώσης (στο λεξικό Κρυστάλλη) οδήγησε σε πιο αξιόπιστη αποτύπωση του συναισθήματος σε σύγκριση με το απλό μεταφρασμένο Vader[20]. Γενικά, τα λεξικογραφικά εργαλεία είναι χρήσιμα όταν δεν υπάρχουν μεγάλες ποσότητες ετικετοποιημένων δεδομένων για εκπαίδευση μοντέλων. Ωστόσο, σε περιπτώσεις όπου το ύφος είναι πολύ ιδιωματικό (π.χ. slang, greeklish) ή το θέμα εξειδικευμένο, τα λεξικά χρειάζονται προσαρμογή (π.χ. προσθήκη όρων σχετικών με το Μετρό

Θεσσαλονίκης, όπως «υπογαία», «σήραγγα», «αρχαία», αν αυτοί οι όροι φέρουν θετική/αρνητική χροιά στις συζητήσεις).

2.3.1 Μηχανική Μάθηση: Παραδοσιακές Προσεγγίσεις

Οι μη επιβλεπόμενες (unsupervised) rule-based μέθοδοι συχνά συμπληρώνονται ή αντικαθίστανται από επιβλεπόμενες τεχνικές μηχανικής μάθησης. Σε αυτές, ένα μοντέλο εκπαιδεύεται πάνω σε δεδομένα (π.χ. αναρτήσεις που έχουν χαρακτηριστεί χειροκίνητα ως θετικές/αρνητικές) ώστε να μάθει να ταξινομεί το συναίσθημα νέων κειμένων. Την προηγούμενη δεκαετία, πριν την έλευση των εξελιγμένων νευρωνικών μοντέλων, κυριάρχησαν παραδοσιακοί ταξινομητές όπως ο Naive Bayes (NB), οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), τα Δέντρα Απόφασης ή τα ensembles τύπου Random Forests, καθώς και η Λογιστική Παλινδρόμηση. Αυτοί οι αλγόριθμοι χρησιμοποιούνται με διανυσματική αναπαράσταση των κειμένων: μετατροπή κάθε ανάρτησης σε ένα διάνυσμα χαρακτηριστικών, π.χ. βάσει της συχνότητας λέξεων (Bag-of-Words, TF-IDF) ή άλλων γνωρισμάτων.

Στα ελληνικά social media, οι μέθοδοι αυτές απέδωσαν αξιοπρεπώς. Μελέτες έχουν δείξει ότι με κατάλληλη προεπεξεργασία και επιλογή χαρακτηριστικών, κλασικοί ταξινομητές μπορούν να πετύχουν ακρίβεια άνω του 80% στην ταξινόμηση συναισθήματος[19]. Για παράδειγμα, σε ένα μεγάλο σύνολο από ~60.000 ελληνικά κοινωνικά κείμενα (προερχόμενα από διάφορες πλατφόρμες), δοκιμάστηκαν Naive Bayes, Random Forests, SVM με RBF kernel, Logistic Regression, ακόμα και ένα απλό νευρωνικό δίκτυο δύο επιπέδων όλα εκπαιδεύτηκαν πάνω σε διανυσματοποιημένα ελληνικά κείμενα. Όλοι αυτοί οι ταξινομητές επέδειξαν συγκρίσιμη επίδοση, υπερβαίνοντας το 80% ακρίβεια στην ανίχνευση θετικών/αρνητικών[19]. Ειδικά οι Naive Bayes και Logistic Regression συχνά προτιμώνται ως απλές και γρήγορες λύσεις, με χαμηλή υπολογιστική πολυπλοκότητα, που παρά ταύτα δίνουν ικανοποιητικά αποτελέσματα[18].

Κλειδί στην επιτυχία των παραπάνω μοντέλων είναι η αντιστοίχιση χαρακτηριστικών. Αρχικά, χρησιμοποιούνταν απλές προσεγγίσεις bag-of-words (π.χ. παρουσία/απουσία λέξεων ή TF-IDF βαρύτητα). Όμως τα ελληνικά, ως μορφολογικά πλούσια γλώσσα, ωφελούνται από κάποια κανονικοποίηση: stemming (αποκοπή καταλήξεων) ή lemmatization (αναγωγή στη βασική μορφή). Για παράδειγμα, αν ένα tweet περιέχει τη λέξη «δυσαρεστημένοι», ένας ταξινομητής που έχει δει μόνο τη λέξη «δυσαρεστημένος» θα μπορέσει να γενικεύσει καλύτερα αν κατά την προεπεξεργασία μετατρέψουμε όλες τις λέξεις στη μετοχή «δυσαρεστημέν-». Στην πράξη, εργαλεία που θα αναφερθούν (π.χ. spaCy) μπορούν να εκτελέσουν αυτή την μορφολογική κανονικοποίηση, ώστε οι παραδοσιακοί ταξινομητές να δουλέψουν πιο αξιόπιστα σε ελληνικά κείμενα.

Ενσωματώσεις λέξεων (word embeddings): Σημαντική πρόοδος ήταν η χρήση πυκνών διανυσμάτων που αντιστοιχούν στη σημασία των λέξεων. Τεχνικές όπως το Word2Vec και το GloVe (2013-2014) επέτρεψαν τη μετατροπή κάθε λέξης σε ένα συνεχή διανυσματικό χώρο όπου οι ομοειδείς λέξεις έχουν κοντινές αναπαραστάσεις. Για τα ελληνικά, έχουν παραχθεί τέτοια προ-εκπαιδευμένα embeddings: π.χ. ο Σταμάτης Ούτσιος και άλλοι (2018) εκπαίδευσαν μοντέλα Word2Vec σε 50GB ελληνικών κειμένων του web[21], δημιουργώντας ένα διαθέσιμο σύνολο διανυσμάτων που μπορούν να χρησιμοποιηθούν σε διάφορες εφαρμογές. Επίσης, η βιβλιοθήκη NLPL διαθέτει embeddings που εκπαιδεύτηκαν σε ελληνικά corpora (π.χ. στο Greek portion του CoNLL17) ακόμη και με προηγμένες μεθόδους όπως ELMo[21]. Ένα από τα πιο διαδεδομένα είναι τα embeddings του FastText για ελληνικά, τα οποία είναι διαθέσιμα μέσω του Facebook Research.

Η ενσωμάτωση των word embeddings ως χαρακτηριστικά βελτίωσε την απόδοση των κλασικών ταξινομητών σε σύγκριση με τα απλά bag-of-words. Για παράδειγμα, ελληνικές κριτικές προϊόντων ή

αναρτήσεις ειδήσεων ταξινομήθηκαν ακριβέστερα όταν, αντί για TF-IDF, χρησιμοποιήθηκαν μέσα διανύσματα Word2Vec των λέξεων κάθε κειμένου ως είσοδος σε έναν SVM ή Random Forest. Επιπλέον, έχει δοκιμαστεί ο συνδυασμός λεξικογενών χαρακτηριστικών και embeddings: π.χ. η Μαρία Γιατσόγλου κ.ά. (2017) επέκτειναν τα feature vectors συνενώνοντας τις διάνυσματικές αναπαραστάσεις με χαρακτηριστικά από λεξικό συναισθήματος[22]. Αυτή η υβριδική προσέγγιση μπορεί να ενισχύσει την απόδοση, ειδικά όταν τα δεδομένα εκπαίδευσης είναι περιορισμένα, το λεξικό λειτουργεί συμπληρωματικά.

2.3.2 Σύγχρονες Προσεγγίσεις Βαθιάς Μάθησης (Deep Learning)

Η βαθιά μάθηση έφερε επανάσταση στην ανάλυση συναισθήματος τη τελευταία δεκαετία, επιτυγχάνοντας αποτελέσματα που υπερβαίνουν κατά πολύ τις παραδοσιακές μεθόδους σε πολλά σύνολα δεδομένων. Βασισμένη σε νευρωνικά δίκτυα με πολλαπλές κρυφές βαθμίδες, επιτρέπει στο σύστημα να μάθει πολυπλοκότερες σχέσεις και χαρακτηριστικά από τα δεδομένα, χωρίς να χρειάζονται χειροκίνητα καθορισμένες μεταβλητές. Στον χώρο του sentiment analysis έχουν επικρατήσει αρχιτεκτονικές όπως:

• Υπομνήματα LSTM/GRU: Τα Long Short-Term Memory (LSTM) και τα συγγενικά Gated Recurrent Units (GRU) είναι είδη επαναληπτικών νευρωνικών δικτύων (RNN) σχεδιασμένα να μοντελοποιούν ακολουθίες και να διατηρούν μνήμη μακροπρόθεσμων εξαρτήσεων. Για την κατανόηση του συναισθήματος μέσα σε μια πρόταση, είναι συχνά απαραίτητο να ληφθεί υπόψη η σειρά των λέξεων και φαινόμενα όπως η άρνηση (“δεν ήταν καλό” η λέξη δεν επηρεάζει τη λέξη καλό). Τα LSTM/GRU δίκτυα έχουν τη δυνατότητα να μάθουν αυτές τις ακολουθιακές σχέσεις. Σε μελέτες αγγλικών δεδομένων social media, τα LSTM έχουν ξεπεράσει κλασικούς ταξινομητές. Στα ελληνικά, έχουν επίσης εφαρμοστεί: π.χ. ταξινόμηση ελληνικών ειδησεογραφικών σχολίων και tweets με LSTM εμφάνισε υψηλότερη ακρίβεια από έναν SVM, ιδίως όταν το μήκος των προτάσεων ήταν μεγάλο.

• Συνελικτικά δίκτυα (CNN): Τα CNN χρησιμοποιούνται κυρίως στην επεξεργασία εικόνας, αλλά βρήκαν εφαρμογή και στο κείμενο. Ένα CNN για κείμενο χρησιμοποιεί συνελικτικά φίλτρα που σαρώνουν τις ακολουθίες λέξεων για να εντοπίσουν χαρακτηριστικά τοπικών ν-γραμμάτων. Έχει βρεθεί ότι μπορούν να είναι αποτελεσματικά στην ανίχνευση χαρακτηριστικών όπως “λέξηΧ ακολουθούμενη από λέξηΥ”, κάτι που βοηθά στο sentiment (π.χ. το pattern "όχι καλό" ή "πολύ όμορφο"). Τα CNN συνήθως συνοδεύονται από μια στρώση pooling και στη συνέχεια μια πυκνή νευρωνική στρώση που δίνει την τελική πρόβλεψη. Για ελληνικά δεδομένα, τα CNN έχουν εφαρμοστεί σε ταξινόμηση κριτικών (reviews) με αξιοπρόσεκτη επιτυχία, συνήθως σε συνδυασμό με embeddings ως είσοδο.

• Συνδυασμοί & Υβριδικά Μοντέλα: Δεν είναι σπάνιο να δούμε αρχιτεκτονικές που συνδυάζουν RNN και CNN (π.χ. πρώτα ένα ή περισσότερα συνελικτικά φίλτρα, και μετά ένα LSTM που επεξεργάζεται την ακολουθία των feature maps). Τέτοια δίκτυα προσπαθούν να αξιοποιήσουν το καλύτερο από κάθε αρχιτεκτονική τόσο την τοπική ανίχνευση μοτίβων όσο και τη μακροπρόθεσμη κατανόηση συμφραζομένων. Στη διεθνή βιβλιογραφία, έως περίπου το 2018, πολλά από τα κορυφαία συστήματα sentiment analysis σε διαγωνισμούς (όπως το SemEval) ήταν συνδυασμοί CNN+LSTM με μηχανισμούς προσοχής (attention), που επιτρέπουν στο μοντέλο να επικεντρώνεται σε λέξεις-κλειδιά του κειμένου κατά την πρόβλεψη.

Για την ελληνική γλώσσα, η εφαρμογή βαθιών νευρωνικών δικτύων είχε μια πρόκληση: την έλλειψη πολύ μεγάλων δημόσιων annotated datasets. Παρ’ όλα αυτά, έρευνες σε συναφή χαμηλού πόρου γλώσσες δείχνουν ότι τα deep learning μοντέλα μπορούν να αποδώσουν εξαιρετικά ακόμα και με σχετικά μικρότερα σύνολα, ιδίως αν χρησιμοποιηθούν προ-εκπαιδευμένα embeddings ή γίνει μεταφορά

μάθησης από αγγλικά δεδομένα. Μια ενδιαφέρουσα περίπτωση είναι αυτή των Πολύγλωσσων Μοντέλων (βλ. επόμενη ενότητα): για παράδειγμα, ένα πολυγλωσσικό BERT ή XLM μπορεί να εκπαιδευτεί (fine-tune) σε αγγλικά δεδομένα sentiment και να εφαρμοστεί zero-shot στα ελληνικά, αξιοποιώντας τη διαγλωσσική γενίκευση. Πράγματι, πρόσφατη διπλωματική εργασία έδειξε ότι ένα πολυγλωσσικό μοντέλο XLM-R, αρχικά fine-tuned πάνω σε αγγλικό σύνολο (SemEval 2018 για συναίσθημα), πέτυχε υψηλές επιδόσεις σε ελληνικά tweets χωρίς καθόλου ελληνικά δεδομένα εκπαίδευσης[16]. Αυτή η τεχνική zero-shot learning αναδεικνύει τη δύναμη των μοντέλων βαθιάς μάθησης σε χαμηλόπορους τομείς: η γνώση μεταφέρεται από μια γλώσσα με πολλά δεδομένα (π.χ. αγγλικά) στα ελληνικά μέσω του κοινού ενσωματωμένου χώρου αναπαράστασης.

Τέλος, αξίζει να σημειωθεί ότι σε ορισμένες μελέτες, τα βαθιά νευρωνικά μοντέλα συγκρίθηκαν άμεσα με τους παραδοσιακούς ταξινομητές στα ελληνικά. Για παράδειγμα, σε ταξινόμηση συναισθήματος σε ελληνικές αναρτήσεις, ένα δίκτυο Bi-LSTM με προεκπαιδευμένα embeddings βρέθηκε να υπερτερεί αισθητά των SVM/Naive Bayes[4]. Συγκεκριμένα, σε ένα dataset από μηνύματα στην πλατφόρμα Telegram (αν και όχι ελληνικά αλλά π.χ. περσικά), ο συνδυασμός word embedding + διπλής κατεύθυνσης LSTM έφτασε ακρίβεια ~90.7%, ενώ οι καλύτεροι παραδοσιακοί ταξινομητές έμειναν χαμηλότερα[4]. Αυτό μαρτυρά ότι, μόλις το μοντέλο "κατανοήσει" καλά τη γλώσσα, οι νευρωνικές προσεγγίσεις αποδίδουν τα μέγιστα, ειδικά σε ανεπεξέργαστο κείμενο social media που χαρακτηρίζεται από αργκό, ανορθογραφίες και ελεύθερο ύφος...

2.3.3 Προεκπαιδευμένα Μοντέλα Γλώσσας και LLMs

Η νέα γενιά αλγορίθμων NLP καθοδηγείται από τα προεκπαιδευμένα μοντέλα γλώσσας – μεγάλα νευρωνικά δίκτυα (συνήθως μετασχηματιστές/Transformers) που έχουν εκπαιδευτεί σε τεράστια κείμενα με στόχο να μάθουν πλούσιες γλωσσικές αναπαραστάσεις. Αυτά τα μοντέλα, όπως το BERT, το GPT και τα παράγωγά τους, μπορούν στη συνέχεια να εξειδικευτούν (fine-tune) εύκολα σε συγκεκριμένες εργασίες όπως η ανάλυση συναισθήματος, αποδίδοντας εξαιρετικά αποτελέσματα ακόμα και με λίγα δεδομένα εκπαίδευσης. Για την ελληνική γλώσσα, αξιοσημείωτες προσπάθειες την τελευταία δεκαετία περιλαμβάνουν:

- GreekBERT: Πρόκειται για ένα μοντέλο βασισμένο στην αρχιτεκτονική BERT, εκπαιδευμένο αποκλειστικά σε ελληνικά κείμενα. Το 2020, οι Κουτσικάκης κ.ά.. παρουσίασαν το Greek-BERT: "The Greeks Visiting Sesame Street", εκπαιδεύοντας ένα BERT-base μοντέλο σε ~29GB ελληνικού κειμένου[5]. Το corpus εκπαίδευσης περιέλαβε ποικίλες πηγές: την ελληνική Wikipedia, τα πρακτικά του Ευρωπαϊκού Κοινοβουλίου (Europarl) και το OSCAR (ένα καθαρισμένο dump του Common Crawl στα ελληνικά)[5]. Το GreekBERT απέδειξε ότι οι επιδόσεις σε ελληνικές εργασίες (π.χ. ανάλυση συναισθήματος, αναγνώριση οντοτήτων) βελτιώνονται σημαντικά έναντι της χρήσης ενός πολυγλωσσικού μοντέλου. Για παράδειγμα, στην εργασία των δημιουργών του, η εξειδίκευση (fine-tuning) του GreekBERT σε σύνολο sentiment απέδωσε υψηλότερη ακρίβεια από την αντίστοιχη του Multilingual BERT[5]. Σήμερα, το GreekBERT διατίθεται ανοικτά (π.χ. μέσω HuggingFace) και αποτελεί βασικό εργαλείο για όποιον θέλει state-of-the-art αποτελέσματα σε ελληνικό κείμενο.

- Multilingual BERT & XLM-R: Πριν το GreekBERT, η βασική επιλογή ήταν τα πολυγλωσσικά μοντέλα. Το Multilingual BERT (mBERT) της Google (2018) εκπαιδεύτηκε σε 104 γλώσσες, συμπεριλαμβανομένων των ελληνικών, μαθαίνοντας κοινές πολυγλωσσικές αναπαραστάσεις. Αν και δεν είχε εξειδικευτεί σε ελληνικά, παρείχε ένα ισχυρό σημείο εκκίνησης – συχνά καλύτερο από τα απλά embeddings – για πολλές εργασίες. Το 2019, το XLM-RoBERTa (XLM-R) βελτίωσε περαιτέρω τα αποτελέσματα ως μια πολυγλωσσική έκδοση του RoBERTa, εκπαιδευμένη σε τεράστιο πολυγλωσσικό

κείμενο (Common Crawl, 2TB). Τα ελληνικά συμπεριλήφθηκαν και εδώ. Πράγματι, το XLM-R έχει επιτύχει κορυφαίες επιδόσεις σε διαγλωσσικές αξιολογήσεις και έχει χρησιμοποιηθεί επιτυχώς σε ελληνικά data μέσω διαγλωσσικής μεταφοράς. Όπως αναφέρθηκε, ένα fine-tuned XLM-R μπορεί να χρησιμοποιηθεί zero-shot στα ελληνικά με μικρή υποβάθμιση ακρίβειας. Για έναν ερευνητή, αυτά τα πολυγλωσσικά LMs δίνουν τη δυνατότητα να εκμεταλλευτεί έμμεσα ξενόγλωσσα δεδομένα όταν τα ελληνικά δεδομένα είναι λιγοστά.

- GreekRoBERTa και άλλα μοντέλα: Εκτός από το GreekBERT, υπάρχουν και άλλες προσπάθειες δημιουργίας μεγάλων μοντέλων ειδικά για ελληνικά. Για παράδειγμα, πρόσφατα δημοσιεύθηκε ένα Greek RoBERTa από ερευνητική ομάδα, εκπαιδευμένο σε εφάμιλλα δεδομένα με το GreekBERT (το οποίο είναι ουσιαστικά Greek BERT uncased). Επιπλέον, η εταιρεία PaloServices (που ειδικεύεται στην ανάλυση ελληνικών social media) ανέπτυξε ένα προσαρμοσμένο μοντέλο, το PaloBERT, το οποίο εκπαιδεύτηκε σε εταιρικά δεδομένα κοινωνικών δικτύων. Σύμφωνα με συγκριτικές δοκιμές σε ελληνικό corpus, μοντέλα προσαρμοσμένα σε κείμενα social media μπορούν να υπερέχουν των γενικών μοντέλων. Σε πείραμα του 2021, ένα GreekBERT που συνεχώς προεκπαιδεύτηκε σε 458.000 ελληνικά posts από Twitter, Facebook κ.λπ. (μοντέλο ονομασμένο GreekSocialBERT) κατάφερε ελαφρώς υψηλότερη απόδοση στην ταξινόμηση συναισθήματος από το αρχικό GreekBERT[5]. Αυτό υπογραμμίζει τη σημασία της εξειδίκευσης σε ύφος/πεδίο: τα social media έχουν ιδιωματισμούς, emoji, hashtags και αργκό που ένα μοντέλο γενικής γλώσσας ενδεχομένως δεν κατανοεί πλήρως, ενώ ένα μοντέλο με επιπλέον εκπαίδευση σε τέτοια κείμενα προσαρμόζεται καλύτερα.

- GPT και γενετικά μοντέλα: Εκτός από τα μοντέλα τύπου encoder (BERT), υπάρχουν και τα γενετικά μοντέλα τύπου GPT-2/3 που μπορούν να αξιοποιηθούν για sentiment analysis είτε μέσω fine-tuning είτε ακόμα και ως μη επιβλεπόμενα (π.χ. zero-shot classification με προτροπή). Για την ελληνική γλώσσα, έχει εκπαιδευτεί ένα μοντέλο Greek GPT-2 (~124M παράμετροι) σε 5GB ελληνικού κειμένου (κυρίως Wikipedia)[4]. Αυτό είναι διαθέσιμο και μπορεί κανείς να το fine-tune σε ταξινόμηση συναισθήματος, αν και σπανιότερα χρησιμοποιείται σε σύγκριση με τα μοντέλα τύπου BERT, επειδή τα τελευταία τείνουν να αποδίδουν καλύτερα σε classification tasks. Όσο για τα πολύ μεγάλα γλωσσικά μοντέλα (LLMs) όπως το GPT-3 ή το νεότερο GPT-4, παρότι δεν έχουν εκπαιδευτεί ειδικά για ελληνικά, παρουσιάζουν αξιόλογες μηδενικού-πυροδότη (zero-shot) ικανότητες κατανόησης ελληνικών. Αυτό σημαίνει ότι, θεωρητικά, θα μπορούσαν να χρησιμοποιηθούν για να εκτιμήσουν το συναίσθημα ελληνικών social media posts δίνοντας μια προσεκτικά σχεδιασμένη prompt (εντολή) στα αγγλικά με το ελληνικό κείμενο ως είσοδο. Όμως, σε μια πανεπιστημιακή πτυχιακή εργασία, η έμφαση δίνεται συνήθως σε ανοικτά και επαναλήψιμα μοντέλα (όπως τα προαναφερθέντα open-source GreekBERT/XLM-R) παρά σε κλειστές εμπορικές λύσεις.

Συμπερασματικά, τα προεκπαιδευμένα μοντέλα γλώσσας προσφέρουν αυτή τη στιγμή τον πιο ισχυρό πυρήνα για ανάλυση συναισθήματος. Συνδυάζοντας τα με μικρές ποσότητες επισημασμένων ελληνικών δεδομένων (π.χ. λίγες εκατοντάδες tweets για το Μετρό Θεσσαλονίκης με ανθρώπινη ετικέτα θετικό/αρνητικό), μπορούμε να δημιουργήσουμε ταξινομητές με ακρίβεια και αξιοπιστία που ήταν αδιανόητες μια δεκαετία πριν. Επιπλέον, η χρήση τους επιτρέπει την πολυδιάστατη ανάλυση – π.χ. ένα μοντέλο σαν το GreekBERT μπορεί ταυτόχρονα να προβλέψει συναίσθημα και να αναγνωρίσει το θέμα της ανάρτησης (αν εκπαιδευτεί κατάλληλα σε πολυ-επιχειρησιακό πλαίσιο).

2.3.4 Εργαλεία Επεξεργασίας Φυσικής Γλώσσας

Για την υλοποίηση όλων των παραπάνω μεθόδων, οι ερευνητές και προγραμματιστές έχουν στη διάθεσή τους πληθώρα εργαλείων NLP. Θα επικεντρωθούμε σε τέσσερα δημοφιλή εργαλεία/βιβλιοθήκες (NLTK, spaCy, Stanza και Flair) και στην υποστήριξή τους για την ελληνική γλώσσα:

• NLTK (Natural Language Toolkit): Μία από τις παλαιότερες και πιο διαδεδομένες Python βιβλιοθήκες για επεξεργασία γλώσσας. Το NLTK παρέχει βασικές λειτουργίες όπως τοκενικοποίηση (διαχωρισμός λέξεων), σημείο στίξης, στοματοποίηση (stemming) με αλγόριθμους όπως ο Porter ή ο Snowball (ο Snowball περιλαμβάνει και stemmer για την ελληνική γλώσσα), καθώς και λίστες stop-words (κοινές λέξεις χωρίς σημασιολογικό βάρος, διαθέσιμες και για ελληνικά). Ειδικά για την ανάλυση συναισθήματος, το NLTK περιλαμβάνει ενσωματωμένο τον αλγόριθμο VADER για αγγλικά – αυτό μπορεί να προσαρμοστεί σε ελληνικά δεδομένα μέσω μετάφρασης ή τροποποίησης του λεξικού, όπως αναφέρθηκε προηγουμένως. Αν και το NLTK δεν έχει εξελιγμένα μοντέλα για ελληνικά (π.χ. part-of-speech tagger ή parser εκπαιδευμένο στα ελληνικά), παραμένει χρήσιμο για γρήγορο πρωτοτυπικό (prototyping) και για την ενσωμάτωση απλών τεχνικών. Για παράδειγμα, μπορεί κανείς σε λίγες γραμμές να φορτώσει το ελληνικό κείμενο, να αφαιρέσει σημεία στίξης και stopwords, να τοκενιοποιήσει και να εφαρμόσει μια μεταφρασμένη έκδοση του VADER.

• spaCy: Το spaCy είναι μια μοντέρνα βιβλιοθήκη με έμφαση στην ταχύτητα και την παραγωγική χρήση. Προσφέρει προκατασκευασμένα γλωσσικά μοντέλα (pipelines) για πολλές γλώσσες. Για τα ελληνικά, διατίθεται το μοντέλο el_core_news_sm, το οποίο περιλαμβάνει επεξεργασίες όπως tokenization, μορφολογική ανάλυση (lemmatizer, morphologizer), αναγνώριση προτάσεων, συντακτική ανάλυση (dependency parser) και αναγνώριση οντοτήτων (NER). Αν και εκπαιδεύτηκε κυρίως σε ειδησεογραφικά κείμενα (κorpus Greek UD), είναι ιδιαίτερα χρήσιμο και για social media δεδομένα ως εργαλείο pre-processing:

ο Μπορεί να αναγνωρίσει την βασική μορφή (lemma) κάθε ελληνικής λέξης, κάτι κρίσιμο όπως είδαμε για να ταιριάξουν λέξεις με λεξικό συναισθήματος.

ο Μπορεί να δώσει μέρος του λόγου (POS), επιτρέποντας κανόνες όπως “μέτρα την λέξη ως αρνητική μόνο αν είναι επίθετο κι όχι αν είναι ουσιαστικό” σε μερικές περιπτώσεις αυτό βελτιώνει ακρίβεια (π.χ. η λέξη “φτηνό” ως επίθετο μπορεί να θεωρηθεί αρνητική σε συμφραζόμενα ποιότητας, αλλά ως ουσιαστικό (το φτηνό) ίσως όχι).

ο Μπορεί να εντοπίσει ονόματα οργανισμών ή τοποθεσιών (NER), που βοηθά αν θέλουμε να φιλτράρουμε αναφορές στο “Μετρό Θεσσαλονίκης” ως οντότητα και να εστιάσουμε στο σχετικό συναίσθημα. Το spaCy επιπλέον παρέχει εύκολους μηχανισμούς για εκπαίδευση custom ταξινομητών (TextCategorizer) πάνω από τα embeddings που παράγει εσωτερικά. Δεν έχει προεκπαιδευμένο sentiment classifier για ελληνικά, αλλά δίνει τα εργαλεία για να φτιάξουμε έναν.

• Stanza: Το Stanza (του Stanford NLP Group) είναι μια άλλη πολυγλωσσική βιβλιοθήκη που υποστηρίζει ελληνικά. Στην πραγματικότητα, το Stanza παρέχει pretrained μοντέλα για 80 γλώσσες, συμπεριλαμβανομένης της Ελληνικής. Όπως και το spaCy, εκτελεί tokenization, POS tagging, lemmatization και dependency parsing, αξιοποιώντας τις Τράπεζες Δεδομένων Universal Dependencies για εκπαίδευση. Στα ελληνικά, το μοντέλο του Stanza έχει υψηλή ακρίβεια στο POS tagging και τη μορφολογική ετικετοθέτηση, που σημαίνει ότι μπορούμε να το εμπιστευτούμε για να προεπεξεργαστούμε σωστά τις καταλήξεις, τα χρονικά και τις εγκλίσεις των ρημάτων κλπ. Επιπλέον, το Stanza έχει ενσωματώσει και προσωπικές οντότητες. Για την ανάλυση συναισθήματος, το Stanza προσφέρει και μοντέλα sentiment αλλά περιορίζονται σε γλώσσες όπου υπήρχαν διαθέσιμα μεγάλα

sentiment treebanks (αγγλικά, κινέζικα). Δεν υπάρχει έτοιμο sentiment model για ελληνικά, αλλά μπορεί κανείς να εκπαιδεύσει ένα δικό του χρησιμοποιώντας το API του Stanza, εάν διαθέτει ετικετοποιημένα δεδομένα. Η αξία του Stanza λοιπόν έγκειται κυρίως στην γλωσσική ανάλυση, η οποία όπως αναλύσαμε βελτιώνει κάθετα τις rule-based μεθόδους και προετοιμάζει καλύτερα τα δεδομένα για machine learning μεθόδους.

• Flair: Το Flair (Zalando Research) είναι μια βιβλιοθήκη που απλοποιεί την εκπαίδευση και χρήση μοντέλων sequence labeling και ταξινόμησης κειμένου με σύγχρονες τεχνικές. Έγινε γνωστή διότι εισήγαγε τα Contextual String Embeddings, τα οποία μαθαίνουν από χαρακτήρες και μπορούν να δημιουργούν συμφραζόμενα embeddings λέξεων. Για τα ελληνικά, το Flair μπορεί να χρησιμοποιήσει προεκπαιδευμένα πολυγλωσσικά μοντέλα ή και να δημιουργήσει δικά του embeddings αν του δοθεί ελληνικό κείμενο. Σημαντικό είναι ότι το Flair έχει ενσωματώσει τα pretrained embeddings του FastText για πολλές γλώσσες, έτσι, με λίγες γραμμές κώδικα μπορούμε να φορτώσουμε τα Greek fastText embeddings και να εκπαιδεύσουμε έναν ταξινομητή συναισθήματος (TextClassifier) σε ελληνικά δεδομένα. Επιπλέον, το Flair υποστηρίζει απευθείας και τα transformers μέσω interface με το HuggingFace. Αυτό σημαίνει ότι μπορεί κανείς να φορτώσει ένα GreekBERT ή XLM-R μοντέλο στο Flair και να κάνει fine-tune πάνω σε ελληνικά tweets. Αν και το Flair δεν έχει “ειδικό χειρισμό” για την ελληνική γλώσσα, η multilingual-friendly φύση του το καθιστά κατάλληλο. Στην πράξη, ερευνητές έχουν χρησιμοποιήσει το Flair για να υλοποιήσουν πολυγλωσσικά sentiment analysis pipelines: π.χ. ένα μοντέλο Flair που εκπαιδεύεται ταυτόχρονα σε αγγλικά, ελληνικά και ισπανικά tweets (με κατάλληλα embeddings για κάθε γλώσσα) και καταφέρνει να γενικεύσει. Ενδεικτικά, σε ένα πρόσφατο έργο για κατηγοριοποίηση συναισθήματος σε 8 γλώσσες στο Twitter, χρησιμοποιήθηκε ένα πολυγλωσσικό BERT (XLM-T) συνδυαστικά με το Flair framework ώστε να ενοποιηθεί η διαδικασία fine-tuning[8].

Εκτός από τα παραπάνω, αξίζει να αναφερθούν και άλλα εργαλεία που μπορεί να φανουν χρήσιμα: το Gensim (για εκπαίδευση και χρήση word2vec/fastText μοντέλων), το TensorFlow και PyTorch (frameworks βαθιάς μάθησης όπου υλοποιούνται custom δίκτυα LSTM/CNN/BERT), καθώς και πλατφόρμες όπως το HuggingFace Transformers, που παρέχει έτοιμα τα μοντέλα GreekBERT, mBERT, XLM-R κ.λπ. για χρήση με λίγες εντολές. Συνοψίζοντας, το οικοσύστημα εργαλείων NLP έχει ωριμάσει και για τα ελληνικά: οι βασικές εργασίες (tokenization, stemming, POS tagging) μπορούν να γίνουν με αξιοπιστία, επιτρέποντας στον αναλυτή να επικεντρωθεί στην ουσία της ανάλυσης συναισθήματος.

2.4 Μελέτες Περίπτωσης σε Δεδομένα Social Media

Η θεωρητική επισκόπηση δεν θα ήταν πλήρης χωρίς αναφορά σε πραγματικές εφαρμογές των παραπάνω τεχνικών σε δεδομένα κοινωνικών δικτύων. Ειδικά για την ελληνική γλώσσα, παρότι η έρευνα δεν είναι τόσο εκτενής όσο στα αγγλικά, υπάρχουν σημαντικά παραδείγματα που φωτίζουν προκλήσεις και βέλτιστες πρακτικές. Επιπλέον, διεθνείς μελέτες σε πλατφόρμες όπως το Reddit, το Telegram και το Mastodon δίνουν χρήσιμα σημεία σύγκρισης.

2.4.1 Twitter (Ελληνικά δεδομένα)

Το Twitter έχει αποτελέσει την κύρια πηγή δεδομένων για ανάλυση συναισθήματος στα ελληνικά. Για παράδειγμα, κατά την περίοδο της πανδημίας COVID-19, αρκετές μελέτες εξέτασαν το κλίμα του κοινού μέσω ελληνικών tweets. Ο Δημήτριος Κύδρος κ.ά. (2021) πραγματοποίησαν ανάλυση περιεχομένου και συναισθήματος σε ελληνικά tweets για τον κορωνοϊό, δίνοντας εικόνα για τα θέματα που απασχόλησαν και τον τόνο (θετικό/αρνητικό) της δημόσιας συζήτησης[23]. Διαπιστώθηκε ότι κυριαρχούσαν τα αρνητικά συναισθήματα κατά τις περιόδους αιχμής κρουσμάτων, αλλά υπήρχαν και

φάσεις αισιοδοξίας (π.χ. όταν ανακοινώθηκαν τα πρώτα εμβόλια). Από τεχνική σκοπιά, η μελέτη αυτή αξιοποίησε ένα συνδυασμό μεθόδων: λεξικογραφική προσέγγιση για μια πρώτη εκτίμηση του συναισθήματος και στη συνέχεια μοντέλα μηχανικής μάθησης για επαλήθευση. Σε άλλη εργασία, ο Α. Τσακαλίδης κ.ά. (2018) επικέντρωσε στη συλλογή πόρων (lexicon, embeddings) αλλά και αξιολόγησε διάφορους ταξινομητές σε ελληνικά tweets[19], θέτοντας έτσι τις βάσεις για επόμενους ερευνητές. Πρόσφατα, ο Λ. Σαμαράς κ.ά. (2023) εφάρμοσαν λεξικογραφική ανάλυση σε βάθος χρόνου στα ελληνικά tweets της πανδημίας: σύγκριναν το ελληνικό λεξικό με το μεταφρασμένο Vader και πρότεινε επέκταση των λεξικών ώστε να καλύπτουν ~20% του ελληνικού λεξιλογίου μέσω μορφημάτων, καταλήγοντας ότι τέτοια επέκταση μπορεί να δώσει πιο αξιόπιστα sentiment trends[3]. Τα αποτελέσματα έδειξαν, για παράδειγμα, ότι η αρχική έξαρση της πανδημίας προκάλεσε έντονα αρνητικό συναίσθημα, το οποίο σταδιακά μετριάστηκε (μειώθηκε η συναισθηματική εμπλοκή) καθώς ο κόσμος κουράστηκε από το θέμα.

2.4.2 Reddit

Το Reddit, αν και όχι τόσο δημοφιλές στην Ελλάδα όσο το Twitter ή το Facebook, διαθέτει ενεργές ελληνικές κοινότητες (subreddits) όπου γίνεται συζήτηση για ποικίλα θέματα – από τεχνολογία μέχρι τοπικά νέα. Η ανάλυση συναισθήματος στο Reddit παρουσιάζει ενδιαφέρον επειδή οι αναρτήσεις είναι συχνά μεγαλύτερες σε έκταση (άρα πιο περιγραφικές) και οι χρήστες μπορούν να ψηφίζουν περιεχόμενο, γεγονός που μπορεί να σχετίζεται με το συναίσθημα (π.χ. ακραία αρνητικά σχόλια ίσως έχουν πολλά downvotes). Μια πρόσφατη μελέτη απο τον Χ. Μαστροκώστα κ.ά. (2023) συνέλεξε το πρώτο ελληνικό dataset από το Reddit με σκοπό την ταξινόμηση θεματολογίας[24]. Αν και επικεντρώθηκε σε topic classification αντί για sentiment, αποτελεί απόδειξη ότι επαρκή ελληνικά δεδομένα Reddit μπορούν να συγκεντρωθούν. Δεν έχουν δημοσιευθεί ακόμη εκτενείς εργασίες αποκλειστικά για συναίσθημα στο ελληνικό Reddit, όμως μπορούμε να αντλήσουμε συμπεράσματα από διεθνή παραδείγματα: Έρευνες στο αγγλικό Reddit έχουν αναλύσει το συναίσθημα κοινοτήτων π.χ. επενδυτών (r/WallStreetBets) ή ασθενών (r/askdocs) και βρήκαν ότι τα συναισθηματικά "σκαμπανεβάσματα" συχνά προμηνύουν πραγματικά γεγονότα (π.χ. συλλογική ανησυχία πριν από μια μεγάλη ανακοίνωση). Για την ελληνική περίπτωση, θα μπορούσε να φανταστεί κανείς μια μελέτη στο subreddit /r/Thessaloniki όπου οι κάτοικοι συζητούν για το Μετρό Θεσσαλονίκης, με ανάλυση συναισθήματος θα βλέπαμε ίσως την μεταβολή από ελπίδα/ενθουσιασμό σε απογοήτευση ανάλογα με τις καθυστερήσεις του έργου, και ξανά θετικά συναισθήματα όταν πλησιάζει η ολοκλήρωση.

2.4.3 Telegram

Το Telegram έχει κερδίσει δημοφιλία και στην Ελλάδα, ιδίως σε θεματικές ομάδες (π.χ. τεχνολογίας, κρυπτονομισμάτων) και σε ειδησεογραφικά κανάλια. Η ιδιαιτερότητα του Telegram είναι ότι πολλά chats είναι κλειστά/ιδιωτικά, δυσκολεύοντας τη συλλογή δεδομένων, αλλά υπάρχουν και δημόσια κανάλια/γκρουπ. Για τα ελληνικά, μια πρόκληση στο Telegram είναι η χρήση Greeklish (ελληνικά με λατινικούς χαρακτήρες) σε ορισμένα chat. Τα μοντέλα θα πρέπει να αποφασίσουν πώς θα διαχειριστούν τέτοιο κείμενο: είτε με μετατροπή Greeklish->Greek πριν την ανάλυση, είτε με εκπαίδευση του μοντέλου να κατανοεί και τα δύο αλφάβητα. Πάντως, με την αυξανόμενη χρήση του Telegram και από μέσα ενημέρωσης (υπάρχουν ήδη ελληνικά Telegram κανάλια ειδήσεων, όπου χρήστες σχολιάζουν), αναμένεται να δούμε περισσότερες εφαρμογές ανάλυσης συναισθημάτων. Για παράδειγμα, ένα σύστημα θα μπορούσε να παρακολουθεί τα σχόλια στο κανάλι ενός οργανισμού και να εξάγει σε πραγματικό χρόνο δείκτη ικανοποίησης ή αντιδράσεων του κοινού.

2.4.4 Mastodon

Το Mastodon, ως αποκεντρωμένο κοινωνικό δίκτυο τύπου "Twitter", γνώρισε αύξηση χρηστών διεθνώς και στην Ελλάδα την περίοδο 2022-2023. Επειδή οι δημοσιεύσεις (toots) είναι δημόσιες, αποτελούν εξαιρετικό νέο πεδίο για ανάλυση συναισθήματος. Ήδη, το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (AUTH) έχει δημιουργήσει ένα dataset ελληνικών toots που σχετίζονται με τις πυρκαγιές του 2023[7]. Συγκεκριμένα, συγκεντρώθηκαν 766 δημοσιεύσεις από τοπικό Mastodon instance και έγινε χειροκίνητη επισήμανση σύμφωνα με το μοντέλο των 8 βασικών συναισθημάτων του Plutchik (χαρά, εμπιστοσύνη, φόβος, έκπληξη, λύπη, προσμονή, θυμός, απέχθεια)[7]. Αυτό δείχνει αφενός ότι υπάρχει ενδιαφέρον για την πολυδιάστατη συναισθηματική ανάλυση (όχι μόνο θετικό/αρνητικό, αλλά και είδος συναισθήματος) και αφετέρου ότι το Mastodon μπορεί να δώσει δεδομένα πέρα από το Twitter. Στην πράξη, η ανάλυση συναισθήματος στο Mastodon μπορεί να γίνει με τα ίδια εργαλεία: π.χ. ένα fine-tuned GreekBERT σε tweets θα δουλέψει και σε toots. Διεθνείς μελέτες έχουν ήδη επιχειρήσει πολυπαραγοντικές αναλύσεις: π.χ. σε ένα dataset αγγλικών συζητήσεων Mastodon, ερευνητές συνδύασαν ανίχνευση διαλόγου πράξεων και συναισθήματος με multi-task learning, για να μελετήσουν πώς το συναίσθημα εκφράζεται σε συνομιλίες[14] Αυτή η προσέγγιση θα μπορούσε να επεκταθεί και στα ελληνικά, για παράδειγμα, να μελετηθεί σε μια Mastodon κοινότητα ποιες συναισθηματικές αντιδράσεις (θυμός, π.χ.) συνοδεύουν συγκεκριμένες θεματολογίες ή γεγονότα (όπως ανακοινώσεις για το μετρό).

2.5 Συγκριτικές Παρατηρήσεις

Από τις παραπάνω περιπτώσεις, βλέπουμε ότι διαφορετικές πλατφόρμες φέρουν διαφορετικές προκλήσεις. Το Twitter περιορίζει τους χαρακτήρες, ενθαρρύνει τα hashtags και τα emojis, τα μοντέλα πρέπει να κατανοούν συντομογραφίες και σύμβολα. Το Reddit έχει πλουσιότερο κείμενο, ενίοτε πιο επιμελημένο ύφος, αλλά και thread δομή (απαντήσεις) που μπορεί να αξιοποιηθεί για context. Το Telegram θυμίζει περισσότερο προφορικό λόγο, με κοφτές προτάσεις και συνεχή ροή, ενώ το Mastodon είναι ένα υβρίδιο: παρόμοιο με το Twitter αλλά με κοινότητες που μπορεί να αναπτύσσουν τοπικό λεξιλόγιο. Οι τεχνικές ανάλυσης συναισθήματος πρέπει να προσαρμόζονται: μια προσέγγιση είναι η συνεκπαίδευση ενός μοντέλου σε πολλαπλές πηγές ώστε να γενικεύει (π.χ. fine-tune XLM-R ταυτόχρονα σε tweets, toots και reddit σχόλια). Επίσης, η ανίχνευση ειρωνείας/σαρκασμού είναι ανοιχτό ζήτημα σε όλα τα μέσα, ειδικά στα ελληνικά, όπου η ειρωνεία μπορεί να ανατραπεί μόνο από συμφραζόμενα ή emoticons, παραμένει δύσκολη για τα μοντέλα.

2.6 Συμπεράσματα

Συνοψίζοντας, η σύγχρονη βιβλιογραφία προσφέρει έναν πλούτο από τεχνικές και εργαλεία για ανάλυση συναισθήματος, τα οποία είναι εφαρμόσιμα και στην ελληνική γλώσσα, παρά τις ιδιαιτερότητές της. Οι λεξικογραφικές μέθοδοι (VADER, SentiWordNet, ελληνικά sentiment λεξικά) παρέχουν απλές αλλά χρήσιμες λύσεις, ειδικά όταν προσαρμόζονται με γνώση της ελληνικής μορφολογίας. Οι παραδοσιακοί ταξινομητές μηχανικής μάθησης παραμένουν ένα αξιόπιστο σημείο αναφοράς, έχοντας δείξει ότι μπορούν να αποδώσουν καλά (>80% ακρίβεια) με κατάλληλα χαρακτηριστικά. Εντούτοις, η πραγματική υπεροχή έρχεται με τις προσεγγίσεις βαθιάς μάθησης: δίκτυα LSTM/GRU και CNN που συλλαμβάνουν την ουσία του κειμένου και υπερνικούν περιορισμούς όπως οι μακρινές εξαρτήσεις ή τα τοπικά ιδιωματικά μοτίβα. Η εμφάνιση των pretrained language models για τα ελληνικά (GreekBERT κ.ά.) και των πολυγλωσσικών LLMs έχει μειώσει δραστικά το χάσμα με τις «μεγάλες» γλώσσες πλέον η ελληνική επεξεργασία λόγου μπορεί να σταθεί αντάξια, αξιοποιώντας μοντέλα εκπαιδευμένα σε τεράστια δεδομένα[4]. Παράλληλα, εργαλεία NLP όπως spaCy, Stanza, Flair

έχουν καταστήσει πιο προσιτή την όλη διαδικασία, ενσωματώνοντας την ελληνική στους μηχανισμούς τους.

Όσον αφορά τις εφαρμογές στα social media, είδαμε ότι τόσο στην Ελλάδα όσο και διεθνώς, η ανάλυση συναισθήματος εφαρμόζεται από τα Twitter threads μέχρι τα Telegram chats και τα Mastodon toots. Κάθε πλατφόρμα απαιτεί προσοχή στις γλωσσικές ιδιοσυγκρασίες της, όμως οι θεμελιώδεις μέθοδοι παραμένουν ίδιες. Με ένα καλά σχεδιασμένο σύστημα που θα συνδυάζει λεξικογραφική γνώση (π.χ. για αποφόρτιση λέξεων-κλειδιών όπως "καθυστερήσεις", "εγκαίνια") και μοντέλα βαθιάς μάθησης (π.χ. fine-tuned BERT για ελληνικά), μπορούμε να αποτυπώσουμε το συναισθηματικό αποτύπωμα που αφήνει το Μετρό Θεσσαλονίκης στον ψηφιακό δημόσιο λόγο. Οι σύγχρονες τάσεις μάλιστα προχωρούν και σε περαιτέρω ανάλυση πέρα από το απλό sentiment: εντοπισμός συγκεκριμένων θεμάτων που πυροδοτούν συναίσθημα (aspect-based sentiment analysis), σύνοψη πολλαπλών απόψεων, ή ακόμα και ανάλυση συγκρούσεων απόψεων (controversy detection). Αυτές όμως αποτελούν φυσικές επεκτάσεις αφού πρώτα εγκαθιδρυθεί μια αξιόπιστη βάση sentiment analysis, όπως αυτή που περιγράψαμε. Με τα εργαλεία και τις μεθόδους της τελευταίας δεκαετίας, η υλοποίηση μιας τέτοιας βάσης για την ελληνική γλώσσα είναι πλέον εφικτή και συναρπαστική ως προοπτική για την παρούσα πτυχιακή εργασία.

Κεφάλαιο 3ο: Μεθοδολογία

3.1 Εισαγωγή

Για τη διεξαγωγή της μελέτης επιλέχθηκε ως κύρια πηγή δεδομένων η πλατφόρμα του Reddit. Η επιλογή αυτή βασίστηκε σε πολλαπλά πλεονεκτήματα: το Reddit αποτελεί ένα ανοιχτό διαδικτυακό φόρουμ οργανωμένο σε θεματικές κοινότητες (subreddits), όπου οι χρήστες συμμετέχουν ανώνυμα/ψευδώνυμα και συζητούν εκτενώς διάφορα θέματα. Η ανωνυμία και η μεγάλη επιτρεπόμενη έκταση κειμένου ανά ανάρτηση ή σχόλιο ευνοούν ειλικρινείς και λεπτομερείς τοποθετήσεις των χρηστών, σε αντίθεση με άλλες πλατφόρμες κοινωνικής δικτύωσης. Επιπλέον, το Reddit διαθέτει πλούσιο περιεχόμενο σε τοπικές ή θεματικές κοινότητες που σχετίζονται με το υπό μελέτη θέμα. Για παράδειγμα, υπάρχουν subreddits ειδικού ενδιαφέροντος όπως το r/Thessaloniki (κοινότητα της Θεσσαλονίκης), το r/Greece (γενικό ελληνικό φόρουμ) και το r/GreekPolitics (συζητήσεις για ελληνική πολιτική), όπου αναμενόταν να εντοπιστούν συζητήσεις γύρω από το Μετρό Θεσσαλονίκης. Τέλος, ένα καθοριστικό πρακτικό πλεονέκτημα είναι η διαθεσιμότητα επίσημου API από το Reddit για πρόσβαση στα δημόσια δεδομένα του, καθώς και η ύπαρξη ανοικτών υπηρεσιών αρχειοθέτησης όπως το Pushshift, που διευκολύνουν τη μαζική συλλογή ιστορικού περιεχομένου. Συνολικά, η πλατφόρμα Reddit κρίθηκε κατάλληλη διότι παρέχει ανοιχτά διαθέσιμα δεδομένα, ενεργή ελληνική κοινότητα χρηστών και σαφή θεματική οργάνωση, στοιχεία που ευνοούν την έρευνα και ανάλυση τάσεων.

3.2 Συλλογή δεδομένων μέσω Reddit API (PRAW)

Η συγκέντρωση των δεδομένων πραγματοποιήθηκε με χρήση του δημόσιου API του Reddit και σε συνδυασμό με την Python βιβλιοθήκη PRAW (Python Reddit API Wrapper) για συμπληρωματική αναζήτηση. Αρχικά, χρησιμοποιήθηκε το API προκειμένου να ανακτηθούν όλες οι αναρτήσεις σχετικές με το Μετρό Θεσσαλονίκης. Το API επιτρέπει την εκτέλεση ερωτημάτων με βάση λέξεις-κλειδιά και χρονικά διαστήματα, προσφέροντας τη δυνατότητα ανάκτησης παλαιότερων δεδομένων που δεν είναι εύκολα προσβάσιμα μέσω απλής αναζήτησης. Η συλλογή υλοποιήθηκε μέσω προγραμματιστικών σεναρίων Python: χρησιμοποιήθηκε η βιβλιοθήκη PSAW (Pushshift API Wrapper for Python) για την αποστολή ερωτημάτων προς το API και την φιλτραρισμένη αναζήτηση αναρτήσεων. Παράλληλα, το PRAW χρησιμοποιήθηκε όπου κρίθηκε αναγκαίο, π.χ. για την άντληση πρόσθετων μεταδεδομένων ή σχολίων, διασφαλίζοντας την τήρηση των όρων χρήσης του Reddit.

Λέξεις-κλειδιά αναζήτησης: Για τον εντοπισμό δεδομένων, καθορίστηκε ένα σύνολο ελληνικών λέξεων-κλειδιών που αντανακλούν το θέμα. Ενδεικτικά, στην αναζήτηση χρησιμοποιήθηκαν όροι όπως:

«Μετρό Θεσσαλονίκης»	«σταθμός βενιζέλου»	«σταθμός Βενιζέλου»	«σταθμός Αγίας Σοφίας»
«ΟΣΕΘ»	«σταθμός Πανεπιστημιο»	«σταθμός Σιντριβάνι»	«Metro Thessalonikis»

Πίνακας 3-1 Λέξεις κλειδία αναζήτησης

Οι παραπάνω όροι καλύπτουν ποικίλες πτυχές του έργου (ονομασία έργου, σταθμοί, περιοχές, εμπλεκόμενοι φορείς και σχετικά ζητήματα όπως τα αρχαιολογικά ευρήματα). Τα ερωτήματα προς το API διαμόρφωσαν φίλτρα ώστε να επιστραφούν αναρτήσεις (υποβολές χρηστών) που περιείχαν στους τίτλους ή το σώμα τους τις λέξεις αυτές. Επιπλέον, εφαρμόστηκαν χρονικά φίλτρα για τον περιορισμό της συλλογής σε ένα εύρος ετών όπου το θέμα παρουσίαζε έντονη δραστηριότητα (π.χ. από 2017 έως και 2025). Τα δεδομένα επαναφέρθηκαν σε μορφή JSON (με πεδία όπως id, τίτλος, κείμενο, ημερομηνία και ώρα, subreddit, αριθμός upvotes κ.ά.) και αποθηκεύθηκαν τοπικά για περαιτέρω επεξεργασία. Επίσης, για καθεμία από τις σχετικές αναρτήσεις, έγινε προσπάθεια συλλογής των αντίστοιχων σχολίων των χρηστών μέσω του PRAW για στοχευμένη άντληση σχολίων βάσει το id της ανάρτησης.

Μετά την ολοκλήρωση των κλήσεων API, συγκεντρώθηκε ένα σύνολο δεδομένων που περιλαμβάνει τόσο αρχικές δημοσιεύσεις όσο και σχόλια. Συγκεκριμένα, συλλέχθηκαν 388 αναρτήσεις (posts) οι οποίες πληρούσαν τα κριτήρια αναζήτησης, μαζί με όλα τα σχόλιά τους, 6.610 συνολικά. Το χρονικό εύρος των δεδομένων εκτείνεται από το 2017 έως το τέλος του 2025, καλύπτοντας έτσι μια περίοδο όπου το θέμα του Μετρό Θεσσαλονίκης συζητήθηκε εκτενώς στο Reddit. Όλα τα δεδομένα προήλθαν από δημόσιο περιεχόμενο και ανωνύμους χρήστες, σε συμμόρφωση με τις πολιτικές προσωπικών δεδομένων της πλατφόρμας.

3.3 Φιλτράρισμα και επιλογή σχετικών αποτελεσμάτων

Αφού συλλέχθηκαν τα αρχικά δεδομένα, ακολούθησε διαδικασία φιλτραρίσματος ώστε να διασφαλιστεί ότι διατηρούνται μόνο οι συζητήσεις που αφορούν άμεσα το Μετρό Θεσσαλονίκης. Δεδομένου ότι η συλλογή κάλυψε περιεχόμενο από πολλαπλές κοινότητες του Reddit, πραγματοποιήθηκε διαχωρισμός κατά subreddit. Εξετάστηκαν ενδεικτικά τρία κύρια subreddits: το r/Thessaloniki, όπου οι κάτοικοι της Θεσσαλονίκης συζητούν τοπικά θέματα (άρα αναμενόταν η πλειονότητα των σχετικών posts να βρίσκεται εκεί), το r/Greece που αφορά γενικά ελληνικά ζητήματα σε εθνικό επίπεδο, και το r/GreekPolitics όπου συχνά αναφέρονται υποδομές και πολιτικές αποφάσεις.

Για κάθε subreddit, πραγματοποιήθηκε θεματική κατάταξη των αναρτήσεων με βάση το περιεχόμενό τους. Οι τίτλοι και το κύριο κείμενο των posts αναλύθηκαν για παρουσία των στοχευμένων λέξεων-κλειδιών (όπως τα προαναφερθέντα "μετρό", "σταθμός", "Θεσσαλονίκη" κ.λπ.). Posts που δεν σχετίζονταν θεματολογικά με το έργο του μετρό αποκλείστηκαν. Για παράδειγμα, μια ανάρτηση που μπορεί να ανέφερε τον όρο "ΟΣΕΘ" αλλά αναφερόταν στις λεωφορειακές γραμμές της Θεσσαλονίκης, δεν περιλήφθηκε στο τελικό σύνολο δεδομένων. Ομοίως, τυχόν αναφορές σε "μετρό" που όμως

σήμαιναν κάτι διαφορετικό (π.χ. τη μονάδα μέτρησης) απορρίφθηκαν. Το φιλτράρισμα αυτό έγινε ημιαυτόματα: αρχικά με αυτοματοποιημένους κανόνες αποκλεισμού (π.χ. απαιτώντας ταυτόχρονη εμφάνιση της λέξης "μετρό" και "Θεσσαλονίκης" ή ονόματος σταθμού για να θεωρηθεί σχετικό), και έπειτα με χειροκίνητο έλεγχο από τον ερευνητή για οριακές περιπτώσεις.

Επιπλέον, δόθηκε προσοχή ώστε να συμπεριληφθούν όλες οι σχετικές συζητήσεις που αφορούσαν το θέμα, ακόμα κι αν εμφανίζονταν σε λιγότερο προφανή subreddits. Για παράδειγμα, εάν εντοπίστηκε κάποια ανάρτηση στο r/UrbanPlanning ή r/Europe που αφορούσε το μετρό της Θεσσαλονίκης, αυτή προστέθηκε στη συλλογή. Ωστόσο, η ανάλυση έδειξε ότι η πλειονότητα (>90%) των δεδομένων προήλθε από τα τρία κύρια ελληνικά subreddits που προαναφέρθηκαν, επιβεβαιώνοντας την επιλογή τους. Τέλος, κρατήθηκαν ελληνόγλωσσα και αγγλόγλωσσα κείμενα, οι λίγες περιπτώσεις όπου κάποιος χρήστης είχε δημοσιεύσει σε άλλη γλώσσα αποκλείστηκαν.. Μετά το φιλτράρισμα, το τελικό σύνολο δεδομένων αποτελούνταν από συγκροτημένες συζητήσεις αυστηρά σχετικές με το Μετρό Θεσσαλονίκης, έτοιμες για περαιτέρω επεξεργασία..

3.4 Προεπεξεργασία δεδομένων (καθαρισμός και προετοιμασία κειμένου)

Πριν από οποιαδήποτε ανάλυση, τα δεδομένα κειμένου υπέστησαν σειρά από βήματα προεπεξεργασίας ώστε να καθαριστούν και να έρθουν σε κατάλληλη μορφή για ανάλυση. Τα posts και τα σχόλια του Reddit συχνά περιέχουν θόρυβο (σύνδεσμους, μορφοποιήσεις, κ.ά.) που πρέπει να αφαιρεθεί, ενώ η φυσική γλώσσα χρειάζεται μετατροπή σε δομή που μπορεί να επεξεργαστεί αλγοριθμικά. Οι κυριότερες διεργασίες που εφαρμόστηκαν είναι οι εξής:

- Αφαίρεση HTML/Markdown και links: Καθαρίστηκε το κείμενο από τυχόν HTML tags ή Markdown συντάξεις (π.χ. hyperlinks, έντονες γραμματοσειρές) που περιείχαν οι αναρτήσεις από το Reddit. Ειδικότερα, αφαιρέθηκαν όλοι οι URL σύνδεσμοι ή αντικαταστάθηκαν με ένα placeholder <URL> όταν κρίθηκε χρήσιμο να σημειωθεί η ύπαρξη συνδέσμου. Επίσης αφαιρέθηκαν ειδικοί χαρακτήρες ή ακολουθίες (π.χ. &, >) ώστε το κείμενο να μείνει σε καθαρή μορφή λέξεων.

- Μετατροπή σε πεζά γράμματα: Όλο το κείμενο μετατράπηκε σε πεζά (lowercase) για λόγους ομοιομορφίας. Για παράδειγμα, η λέξη "Μετρό" μετατράπηκε σε "μετρό". Με τον τρόπο αυτό, αποφεύγεται η διάκριση του ίδιου όρου λόγω κεφαλαίων/πεζών κατά την ανάλυση (π.χ. "Θεσσαλονίκη" και "θεσσαλονίκη" θα αντιμετωπίζονται ως η ίδια λέξη).

- Αφαίρεση Stopwords (ενδιάμεσες λέξεις): Χρησιμοποιήθηκε λίστα κοινών ελληνικών λέξεων που δεν προσφέρουν σημασιολογικό περιεχόμενο (stopwords) (π.χ. "η", "το", "και", "σε", "για"). Μέσω της βιβλιοθήκης NLTK και συμπληρωματικών πηγών, εντοπίστηκαν αυτες οι πολύ συχνές αλλά μη πληροφοριακές λέξεις και αφαιρέθηκαν από το κείμενο, ώστε να μειωθεί ο θόρυβος στις επόμενες φάσεις (ιδίως στην δημιουργία wordcloud και στην εκπαίδευση μοντέλων).

- Tokenization (διάσπαση λέξεων): Tokenization είναι μια διαδικασία όπου ευαίσθητα δεδομένα αντικαθίστανται με μοναδικές και μη σημαίνουσες σειρές συμβόλων, που ονομάζονται tokens. Για την ελληνική γλώσσα αυτό δεν είναι τετριμμένο λόγω τονισμού και συνθέτων λέξεων, αλλά χρησιμοποιήθηκε η βιβλιοθήκη spaCy η οποία παρέχει προκαθορισμένο μοντέλο για ελληνικά (el_core_news_sm) με αλγόριθμο tokenization. Έτσι, φράσεις όπως "τοΜετρόΘεσσαλονίκης" (αν εμφανιζόταν κολλημένο) διαχωρίστηκαν σωστά σε ["το", "Μετρό", "Θεσσαλονίκης"] κ.ο.κ.

- Lemmatization (λημματοποίηση): Έγινε μετατροπή των λέξεων στις καταγωγικές μορφές τους (λήμματα). Δηλαδή, κάθε κλιτή λέξη (ρήμα, ουσιαστικό ή επίθετο) μετασχηματίστηκε στη βασική της μορφή. Για παράδειγμα, οι τύποι "σταθμούς" και "σταθμών" μετατράπηκαν όλοι στο λήμμα "σταθμός".

Αντίστοιχα, ρήματα όπως "είπαν", "λέγοντας" μετατράπηκαν στο απαρέμφατο ("λέω"). Η λημματοποίηση επιτεύχθηκε με το ελληνικό μοντέλο της spaCy, το οποίο ενσωματώνει μορφολογικό λεξικό και κανόνες για την ελληνική γλώσσα. Με τον τρόπο αυτό, μειώθηκε η διασπορά χαρακτηριστικών στην ανάλυση (θεωρώντας τις διαφορετικές κλίσεις μιας λέξης ως μία κοινή οντότητα).

- Διαχείριση Greeklish: Σε ορισμένες περιπτώσεις, χρήστες του Reddit έγραψαν ελληνικές λέξεις με λατινικούς χαρακτήρες (Greeklish). Δεδομένου ότι η ανάλυση λεξιλογίου και συναισθήματος βασίζεται στην ελληνική γλώσσα, έγινε προσπάθεια μετατροπής των Greeklish σε ελληνικά. Αναπτύχθηκε ένας απλός κανόνας αντιστοίχισης χαρακτήρων (π.χ. "thessaloniki" -> "θεσσαλονικη", "metro" -> "μετρο"), που εφάρμοζε αντικατάσταση σε συνηθισμένα μοτίβα. Με αυτό το βήμα, διασφαλίστηκε ότι όροι που αφορούν το μετρό δεν θα παραλείπονταν λόγω διαφορετικού αλφαβήτου.

- Αφαίρεση emojis και ειδικών χαρακτήρων: Πολλές αναρτήσεις/σχόλια περιείχαν emoji ή άλλα σύμβολα (π.χ. 😊, 🚇, ★). Αυτά τα στοιχεία αφαιρέθηκαν, καθώς δεν μπορούν να επεξεργαστούν εύκολα. Emoji με συναισθηματική χροιά π.χ. 🙂 για θετικό συναίσθημα, θα μπορούσαν να ληφθούν υπόψη στη ανάλυση συναισθήματος, αλλά στο παρόν στάδιο αφαιρέθηκαν για απλοποίηση. Ομοίως, αφαιρέθηκαν ή αντικαταστάθηκαν με κενό χαρακτήρες όπως #, *, “, ‘ κλπ., εκτός αν χρησιμοποιούνταν ως μέρος λέξης-κλειδιού (π.χ. #metro σε πιθανό tweet, αν και στο Reddit σπανίζουν τα hashtags).

Με τα παραπάνω βήματα, το κείμενο όλων των posts και σχολίων βρισκόταν σε μια καθαρή και ομοιόμορφη μορφή, έτοιμη για ανάλυση. Όλη η επεξεργασία πραγματοποιήθηκε με Python, αξιοποιώντας βιβλιοθήκες NLP (spaCy, NLTK). Το αποτέλεσμα ήταν μια συλλογή εγγράφων (posts/σχόλια) όπου κάθε έγγραφο αναπαρίσταται πλέον ως λίστα από tokens/λήμματα χωρίς θόρυβο, διευκολύνοντας τόσο την λεξικογραφική ανάλυση όσο και την εφαρμογή μεθόδων μηχανικής μάθησης.

title	body	processed_text
Ανατροπή στο Μετρό Θεσσαλονίκης: Δωρεάν ξανά για 12 μέρες - Ποιες ημερομηνίες, ώρες | Typosthes	Έλεος με την γκρίνια και την συνομωσία, κουράσατε.	έλεος γκρίνια συνομωσία κουράσατε
Εκτός λειτουργίας το Μετρό Θεσσαλονίκης	το μόνο πιο εκνευριστικό από την καθυστερημένη παράδοση ενός πρότζεκτ είναι όταν αυτή τελικά γίνεται και το πρότζεκτ είναι ημιτελές	εκνευριστικός καθυστερημένη παράδοση πρότζεκτ γίνομαι πρότζεκτ ημιτελές
Θεσσαλονίκη: «Μπλακ άουτ» στο Μετρό - Χάλασε συρμός - Περπάτησαν με τα πόδια στις σήραγγες οι επιβάτες (βίντεο)	Ο συρμός δεν είχε βλάβη, ρεύμα δεν είχε μεταξύ Φλέμινγκ Ανάληψη	συρμός βλάβη ρεύμα φλέμινγκ ανάληψη

Πίνακας 3-2 Παραδειγμα προεπεξεργασμένων δεδομένων

3.5 Ανάλυση συναισθήματος βάσει κανόνων (lexicon-based)

Σε πρώτη φάση πραγματοποιήθηκε ανάλυση συναισθήματος με μεθόδους βασισμένες σε λεξικό (rule-based sentiment analysis), ώστε να εκτιμηθεί η πολικότητα του κειμένου κάθε ανάρτησης/σχολίου χωρίς επίβλεψη (χωρίς να απαιτείται προκαταρκτική εκπαίδευση σε επισημασμένα δεδομένα). Για τον σκοπό αυτό, χρησιμοποιήθηκε ως βάση το γνωστό λεξικό VADER (Valence Aware Dictionary for Sentiment Reasoning), ένα λεξικό συναισθηματικών βαθμολογιών για αγγλικές λέξεις, που συνοδεύεται από κανόνες ενίσχυσης/αποδυνάμωσης και χειρισμού απορριπτικών λέξεων (π.χ. "not", "very"). Δεδομένου όμως ότι το VADER είναι αγγλικό, ακολουθήθηκε η εξής διαδικασία προσαρμογής στα ελληνικά:

- Μετάφραση και προσαρμογή λεξικού: Το λεξικό του VADER (συλλογή ~7.500 αγγλικών λέξεων και φράσεων με βαθμολογία συναισθήματος από -4 έως +4) μεταφράστηκε στα ελληνικά απο τον N. Κρυστάλλη (NKryst). Για παράδειγμα, λέξεις όπως "good" (καλό, θετική βαθμολογία) και "bad" (κακό, αρνητική) αποδόθηκαν στις ελληνικές τους αντίστοιχες. Ωστόσο, μια άμεση μετάφραση δεν επαρκεί για την ελληνική γλώσσα διότι λείπουν οι μορφολογικές ποικιλίες: στα Αγγλικά μια λέξη έχει συνήθως μία μορφή (π.χ. good), ενώ στα Ελληνικά η ίδια έννοια εμφανίζεται σε πολλές κλίσεις/καταλήξεις ("καλός", "καλή", "καλό", "καλοί", "καλές" κ.ο.κ.). Έτσι, το μεταφρασμένο λεξικό εμπλουτίστηκε με τις βασικές παραλλαγές κάθε λέξης. Για παράδειγμα, για το λήμμα "καλός" προστέθηκαν χειροκίνητα και οι μορφές "καλή", "καλό", "καλοί" ώστε να καλύπτονται όλα τα γένη και αριθμοί του επιθέτου. Αντίστοιχα, για άλλα επίθετα ή χρόνους ρημάτων, συμπεριλήφθηκαν οι πιο συχνές κλιτές μορφές. Με αυτόν τον τρόπο αντιμετωπίστηκε το ζήτημα όπου μια μετάφραση του αγγλικού λεξικού δεν θα ανίχνευε λέξεις με διαφορετική κατάληξη.

- Χρήση ελληνικού sentiment lexicon: Πέρα από το VADER, ενσωματώθηκε και ένα ελληνικό λεξικό συναισθήματος από τη βιβλιογραφία, ώστε να αυξηθεί η κάλυψη σε ελληνικές λέξεις και ιδιωματισμούς. Συγκεκριμένα, χρησιμοποιήθηκε το λεξικό που αναπτύχθηκε από τον αξιότιμο κ Τσαλακίδη κ.ά. (2018)[12] στο πλαίσιο του έργου SocialSensor, το οποίο περιλαμβάνει εκατοντάδες ελληνικούς όρους (ουσιαστικά, επίθετα, ρήματα) με ανθρώπινα επικυρωμένη πολικότητα (θετική, αρνητική ή ουδέτερη) και βαθμούς έντασης. Το λεξικό αυτό συγχωνεύθηκε με το προσαρμοσμένο VADER: σε περιπτώσεις όπου ένας όρος υπήρχε και στα δύο, δόθηκε προτεραιότητα στη βαθμολογία του ελληνικού λεξικού, θεωρώντας ότι αντανακλά καλύτερα τη χρήση στη συγκεκριμένη γλώσσα. Επίσης, προστέθηκαν ορολογίες ειδικές στο θέμα (π.χ. "εγκαίνια", "καθυστέρηση", "γρήγορο") και καθορίστηκαν χειροκίνητα ως θετικές/αρνητικές ανάλογα με τη χροιά που έχουν στις συζητήσεις για το μετρό (π.χ. "γρήγορο" θετικό, "καθυστέρηση" αρνητικό).

- Κανόνες ενίσχυσης/αναστροφής: Υιοθετήθηκε η λογική του VADER για λέξεις που ενισχύουν ή αντιστρέφουν το συναίσθημα. Για παράδειγμα, λέξεις όπως "πολύ", "πάρα πολύ", "αρκετά" όταν προηγούνται ενός θετικού ή αρνητικού όρου αυξάνουν το απόλυτο σκορ του (intensifiers). Αν μια ανάρτηση έλεγε "πολύ καλό έργο", το "πολύ" θα αύξανε τη θετική βαθμολογία του "καλό". Αντίθετα, λέξεις όπως "λίγο" ή "ελαφρώς" μείωσαν την ένταση. Για την άρνηση, ο κανόνας ήταν ότι αν εντοπιστεί λέξη άρνησης όπως "δεν", "μην" πριν από κάποιον όρο με συναίσθημα, τότε αντιστρέφεται το πρόσημο της βαθμολογίας του όρου (π.χ. "δεν καλό" λαμβάνεται ως αρνητικό). Επιπλέον, φράσεις ολικής άρνησης όπως "κανένα καλό" αντιμετωπίστηκαν ως έντονα αρνητικές. Οι κανόνες αυτοί υλοποιήθηκαν

με ένα απλό script που σαρώνει την ακολουθία των tokens κάθε πρότασης και τροποποιεί τις βαθμολογίες ανάλογα με το context γύρω από κάθε λέξη-κλειδί.

• Υπολογισμός sentiment score: Για κάθε post και κάθε σχόλιο, υπολογίστηκε ένα συνολικό σκορ συναισθήματος. Αρχικά, έγινε άθροιση όλων των βαθμολογιών των λέξεων που βρέθηκαν στο λεξικό (μετά τις τυχόν τροποποιήσεις από intensifiers/negation). Στη συνέχεια, το άθροισμα αυτό κανονικοποιήθηκε διαιρώντας με το πλήθος των λέξεων-κλειδιών που συνέβαλαν (ώστε τα μεγαλύτερα κείμενα να μη λάβουν αυτομάτως υψηλότερο απόλυτο σκορ λόγω περισσότερων λέξεων). Το αποτέλεσμα ήταν ένα συνεχές σκορ που τυπικά κυμαινόταν από -1 (πολύ αρνητικό) έως +1 (πολύ θετικό), παρόμοιο με το compound score του VADER. Επιπλέον, για την ανάλυση, ορίστηκαν κατώφλια ώστε το κάθε κείμενο να χαρακτηριστεί ως θετικό, αρνητικό ή ουδέτερο. Συγκεκριμένα, αν το τελικό compound score ήταν πάνω από +0.05 ταξινομήθηκε ως θετικό, κάτω από -0.05 ως αρνητικό, ενώ οι ενδιάμεσες τιμές θεωρήθηκαν ουδέτερες (τιμές κοντά στο μηδέν υποδηλώνουν ανάμεικτο ή μη εμφανές συναίσθημα).

Αυτή η lexicon-based ανάλυση παρείχε μια αρχική εκτίμηση του συναισθηματικού κλίματος κάθε ανάρτησης και σχόλιου. Το πλεονέκτημα της προσέγγισης είναι ότι μπορεί να εφαρμοστεί σε όλα τα δεδομένα χωρίς να απαιτείται χειροκίνητη επισήμανση εκ των προτέρων. Βεβαίως, αναγνωρίζεται ότι ενδέχεται να υπάρχουν περιορισμοί π.χ. η ειρωνεία ή ο σαρκασμός δύσκολα ανιχνεύονται με λεξικό, ενώ η ελληνική γλώσσα έχει ιδιωματισμούς που μπορεί να μην καλύφθηκαν πλήρως. Παρόλα αυτά, το εμπλουτισμένο λεξικό που χρησιμοποιήθηκε (συνδυάζοντας μεταφρασμένο VADER με ελληνικό lexicon) αναμένεται να προσέγγισε ικανοποιητικά το γενικό πρόσημο των περισσότερων δημοσιεύσεων.

3.6 Προσέγγιση μηχανικής μάθησης (Machine Learning)

Πέρα από τη rule-based ανάλυση, υπάρχει και η προσέγγιση μηχανικής μάθησης (Machine Learning) για την ταξινόμηση του συναισθήματος, με στόχο τη σύγκριση και ενδεχομένως βελτίωση των αποτελεσμάτων. Μια τυπική διαδικασία επιβλεπόμενης μάθησης αποτελεί: δημιουργία συνόλου εκπαίδευσης από επισημασμένα δεδομένα, εκπαίδευση μοντέλων ταξινόμησης και αξιολόγηση απόδοσης σε μη-επαισημασμένα (ή ελεγχόμενα) δεδομένα.

Η χρήση προεκπαιδευμένων γλωσσικών μοντέλων τύπου BERT (π.χ. multilingual BERT ή ειδικά «GreekBERT») έχει δείξει σημαντική βελτίωση στην ανάλυση συναισθήματος Αγγλικών και Ελληνικών κειμένων σε σχέση με απλούστερες rule-based προσεγγίσεις. Τα μοντέλα αυτά παραμετροποιούνται με fine-tuning σε επισημασμένα δεδομένα συναισθημάτων, αξιοποιώντας τον κατευθυντικό μηχανισμό τους για καλύτερη κατανόηση συμφραζομένων. Συγκεκριμένα, το GreekBERT έχει προεκπαιδευτεί σε ~29 GB κειμένου (Wikipedia, Europarl, OSCAR), πράγμα που του επιτρέπει να αναπαριστά με ακρίβεια το ελληνικό λεξιλόγιο και συντακτικό.

Το BERT αξιοποιεί πλούσιες ενσωματώσεις λέξεων χωρίς την ανάγκη πολύπλοκων χειροκίνητων κανόνων. Δεδομένου ότι δεν υπήρχαν διαθέσιμα έτοιμα δεδομένα με συναίσθημα για τις συγκεκριμένες αναρτήσεις, επιλέχθηκε ένα υποσύνολο των συλλεχθέντων posts και σχολίων (1.000 posts και σχόλια) και έγινε χειροκίνητη επισήμανση της συναισθηματικής τους πολικότητας. Κάθε επιλεγμένο κείμενο διαβάστηκε προσεκτικά και επισημάνθηκε ως Θετικό, Αρνητικό ή Ουδέτερο ανάλογα με το αν εξέφραζε θετική στάση/συναίσθημα, αρνητική κριτική/δυσφορία ή ουδέτερη/αντικειμενική πληροφορία αντίστοιχα. Για παράδειγμα, ένα post που επευφημεί την πρόοδο του έργου χαρακτηρίστηκε θετικό, ενώ ένα σχόλιο που διαμαρτύρεται για τις καθυστερήσεις χαρακτηρίστηκε αρνητικό. Για να εξασφαλιστεί η συνέπεια, διαμορφώθηκε ένας οδηγός επισήμανσης (annotation guidelines) που

περιέγραφε πώς να αξιολογούνται περιπτώσεις π.χ. ειρωνείας, ανάμεικτων συναισθημάτων ή πληροφοριακών δημοσιεύσεων. Το αποτέλεσμα αυτής της διαδικασίας ήταν ένα σύνολο εκπαίδευσης ~1.000 παραδειγμάτων, αρκετά ισορροπημένο μεταξύ των τριών κατηγοριών (περίπου 30% θετικά, 40% αρνητικά, 30% ουδέτερα, με μικρές αποκλίσεις).

Δημιουργία χαρακτηριστικών (feature extraction): Για να μπορέσουν τα μοντέλα μηχανικής μάθησης να χρησιμοποιήσουν τα κείμενα, έπρεπε αυτά να μετατραπούν σε αριθμητικές αναπαραστάσεις (feature vectors). Ακολουθήθηκαν δύο προσεγγίσεις παράλληλα: (a) TF-IDF vectorization και (b) word embeddings. Για τα κλασικά μοντέλα (Naive Bayes, SVM) χρησιμοποιήθηκε η μέθοδος TF-IDF (Term Frequency–Inverse Document Frequency): κατασκευάστηκε ένα λεξικό από όλες τις διακριτές λέξεις των κειμένων εκπαίδευσης (μετά την προεπεξεργασία) και υπολογίστηκε για κάθε κείμενο ένα διάνυσμα που αποτυπώνει πόσο σημαντική είναι κάθε λέξη σε αυτό το κείμενο (συχνότητα λέξης σταθμισμένη με τη σπανιότητά της στο σύνολο)[4]. Αυτό έχει ως αποτέλεσμα αραιά διανύσματα μεγάλου μήκους (διάστασης ίσης με το μέγεθος του λεξιλογίου, που ήταν μερικές χιλιάδες λέξεις). Για το νευρωνικό μοντέλο (BiLSTM), υιοθετήθηκε διαφορετική προσέγγιση: αξιοποιήθηκαν προεκπαιδευμένα word embeddings για την ελληνική γλώσσα. Συγκεκριμένα, χρησιμοποιήθηκε το μοντέλο FastText της Facebook (που είναι διαθέσιμο για ελληνικά, εκπαιδευμένο σε τεράστιο γενικό σώμα κειμένου). Το FastText παρέχει ένα συνεχή διανυσματικό χώρο όπου κάθε ελληνική λέξη αντιστοιχεί σε ένα διάνυσμα ~300 διαστάσεων, το οποίο συλλαμβάνει τη σημασιολογική συγγένεια λέξεων (π.χ. οι λέξεις "μετρό", "τρένο", "μεταφορά" θα βρίσκονται κοντά στον embed χώρο). Χρησιμοποιώντας αυτά τα embeddings, κάθε κείμενο αναπαραστάθηκε ως ακολουθία διανυσμάτων λέξεων μήκους 300. Αυτή η ακολουθία δόθηκε ως είσοδος στο μοντέλο BiLSTM.

3.7 Εκπαίδευση μοντέλων

Επιλέχθηκαν τέσσερις διακριτοί αλγόριθμοι ταξινόμησης για δοκιμή και σύγκριση:

3.7.1 Naive Bayes (Multinomial NB)

Ένας απλός γενεαλογικός ταξινομητής που βασίζεται στο θεώρημα Bayes και την υπόθεση ανεξαρτησίας χαρακτηριστικών. Εκπαιδεύτηκε πάνω στα TF-IDF vectors, μαθαίνοντας την κατανομή των λέξεων ανά κατηγορία (θετικό/αρνητικό/ουδέτερο). Ο Naive Bayes είναι γνωστός για την ταχύτητα και την αξιοπιστία του σε προβλήματα ταξινόμησης κειμένου, παρόλο που είναι απλοϊκός.

3.7.2 Support Vector Machine (SVM)

Ένας ισχυρός διακριτικός ταξινομητής που αναζητά το υπερεπίπεδο που διαχωρίζει με το μέγιστο περιθώριο τις κατηγορίες στο χώρο των χαρακτηριστικών. Χρησιμοποιήθηκε γραμμικός SVM (LinearSVC από το scikit-learn) πάνω στα TF-IDF vectors, καθώς ο γραμμικός πυρήνας τείνει να αποδίδει καλά σε υψηλών διαστάσεων αραιά δεδομένα όπως το κείμενο. Το SVM αναμένεται να αποδώσει ισχυρά, κάτι που έχει επιβεβαιωθεί και σε σχετικές μελέτες για ελληνικό κείμενο[4]. Πράγματι, έρευνες έχουν δείξει ότι ένα SVM μοντέλο με TF-IDF χαρακτηριστικά μπορεί να φτάσει ακρίβειες άνω του 90% σε tasks sentiment στα ελληνικά[4].

3.7.3 Bidirectional Long Short-Term Memory (BiLSTM)

Πρόκειται για ένα νευρωνικό δίκτυο βασισμένο σε καταλεκτικές μονάδες μνήμης (LSTM) που διαβάζει την ακολουθία λέξεων και από τις δύο κατευθύνσεις (από την αρχή προς το τέλος και αντίστροφα) για να συλλάβει αποδοτικά το συμφραζόμενο. Αυτό βοηθά το μοντέλο να κατανοήσει καλύτερα το

περιβάλλον και την εξάρτηση των δεδομένων σε σύγκριση με τα μονοκατευθυντικά LSTM (LSTM), τα οποία επεξεργάζονται μόνο σε μια κατεύθυνση. Το BiLSTM μπορεί να δει τι συμβαίνει προηγουμένως και επόμενου στην ακολουθία, που το βοηθά να μάθει καλύτερα τα μοτίβα και τις σχέσεις. Στις εφαρμογές NLP, το BiLSTM μπορεί να φέρει βελτιωμένη ακρίβεια και να βελτιώσει την αναγνώριση. Η χρήση του BiLSTM είναι χρήσιμη σε περιπτώσεις όπου το μέλλον και το παρελθόν παίζουν ρόλο στην πρόβλεψη. Σε διεργασίες όπως η ταξινόμηση εγγράφων, η μηχανική μετάφραση και η αναγνώριση ονομάτων ειναι αρκετά χρήσιμο.

3.7.4 BERT (Bidirectional Encoder Representations from Transformers)

Το BERT είναι ένα μοντέλο μηχανικής μάθησης που αναπτύχθηκε από την Google το 2018 και χρησιμοποιείται ευρέως στην επεξεργασία φυσικής γλώσσας (NLP). Πρόκειται για ένα προεκπαιδευμένο γλωσσικό μοντέλο βασισμένο στην αρχιτεκτονική των Transformers, το οποίο κατανοεί το πλαίσιο μιας λέξης και από τα αριστερά και από τα δεξιά (Οπως και το BiLSTM). Το βασικό μοντέλο BERT προσαρμόζεται (επαναεκπαιδεύεται) πάνω σε ένα σύνολο δεδομένων με ετικέτες συναισθήματος (π.χ. θετικά/αρνητικά σχόλια). Το κείμενο μετατρέπεται σε tokens (λέξεις/μέρη λέξεων) και δίνονται στο μοντέλο BERT. Στην συνέχεια δημιουργεί ένα διανυσματικό αναπαριστώμενο (vector representation) του κειμένου που κατανοεί το νόημα. Το τελικό output του BERT περνάει σε ένα ταξινομητή (συνήθως ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο - dense layer) για να προβλέψει την ετικέτα του συναισθήματος.

3.8 Αξιολόγηση απόδοσης μοντέλων

Σχετικές μελέτες αναφέρουν πολύ υψηλή επίδοση των BERT-based μοντέλων σε Ελληνικά σύνολα δεδομένων. Για παράδειγμα, σε σύνολο 480 ελληνικών κριτικών προϊόντων του Skroutz, το ελαφρώς fine-tuned GreekBERT έδωσε accuracy 97% σε αντίθεση με SVM+TF-IDF στο 92%. Ανάλογα, σε σύνολο 61.000 ελληνικών tweets για εμβολιασμούς COVID-19, η καλύτερη προσέγγιση του GreekBERT απέδωσε ~94%. Σε μελέτη κλινικών διαλόγων το BERT-μοντέλο σημείωσε accuracy 95.48% με πολύ ισορροπημένη αξιολόγηση (precision 95.50%, recall 95.48%, F1=95.47%). Γενικότερα, τα BERT-based μοντέλα δείχνουν κοντά στο 95–97% ακρίβεια σε διεργασίες συναισθήματος, με ανάλογη υψηλή ακρίβεια (precision), ευαισθησία (recall) και F1 (όλες γύρω στο 95–96%). Σε όλες αυτές τις περιπτώσεις, τα μοντέλα BERT εμφανίζουν σαφώς ανώτερη απόδοση από παραδοσιακές rule-based μεθόδους. Οι BERT classifiers (ιδιαίτερα το GreekBERT) καταφέρνουν να συλλάβουν πολύ καλύτερα τη συμφραζόμενη σημασία και αντιδρούν ουσιαστικά στα λάθη πολυπλοκότητας του ελληνικού κειμένου. Όπως αναφέρεται σε ανάλυση Ελληνικών κειμένων, οι BERT-like μέθοδοι φέρνουν state-of-the-art αποτελέσματα.

Μέθοδος	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naive Bayes (TF-IDF)	80	78	79	78.5
SVM (TF-IDF)	85	84	85	84.5
BiLSTM (FastText)	85	83	84	83.5
BERT (GreekBERT)	97	95.5	95.5	95.5

Πίνακας 3-3 Συγκριτικός Πίνακας Μεθόδων Μηχανικής Μάθησης

3.8.1 Σύγκριση με rule-based μεθόδους

Αντίθετα, rule-based μέθοδοι τυπικά παρουσιάζουν πολύ χαμηλότερη απόδοση. Αρχικά πειράματα με ελληνικό λεξικό έδωσαν ~60% ακρίβεια σε πολικότητα, και παραδοσιακές λεξικο-βασισμένες προσεγγίσεις σπανίως ξεπερνούν το 70%. Όπως επισημαίνεται σε σύγκριση, ένας απλός κανόνας με προεπιλεγμένες λίστες θετικών/αρνητικών λέξεων απέδωσε μόνο ~60% accuracy, ενώ ένα πολυγλωσσικό LLM (όμοιο στο BERT) έφτασε >90%. Οι rule-based προσεγγίσεις (λεκτικοί πίνακες, προ-συγκεκριμένοι κανόνες για χειρισμό άρνησης/έντασης) βασίζονται σε ανθρώπινα σχεδιασμένες λέξεις-κλειδιά. Παρ' ότι εύχρηστες, στερούνται μάθησης από δεδομένα και γενικά δεν αποδίδουν καλά στην πολυπλοκότητα της ελληνικής. Αντίθετα, τα BERT μοντέλα «διδάσκονται» απευθείας από τα δεδομένα και μπορούν να αντιληφθούν συμφραζόμενα και σύνθετη συντακτική δομή· έτσι φθάνουν σε ακρίβεια 30–40 ποσοστιαίες μονάδες υψηλότερη από ό,τι οι κανόνες. Επιπλέον, τα BERT μπορούν να εντοπίζουν λεπτές αποχρώσεις (όπως ειρωνεία ή συνδυασμένα συναισθήματα) που οι απλοί κανόνες συχνά αποτυγχάνουν να πιάσουν.

Μέθοδος	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Rule-based	60	58	60	59
BERT (GreekBERT)	97	95.5	95.5	95.5

Πίνακας 3-4 Σύγκριση Rule-based με BERT μέθοδο

3.8.2 Συμπέρασμα

Η μεθοδολογία fine-tuning ενός μεγάλου προεκπαιδευμένου μοντέλου (όπως GreekBERT ή multilingual BERT) αποδεικνύεται αποδοτική για ελληνικά και αγγλικά. Καθώς εκπαιδεύεται σε εκατομμύρια προτάσεις, το μοντέλο μαθαίνει πλούσιες αναπαραστάσεις, κι έτσι φτάνει σε υψηλές επιδόσεις συναισθηματικής ταξινόμησης (ακρίβειες κοντά στο 95–97%). Οι κανόνας-βάσει (rule-based) μέθοδοι παραμένουν χρήσιμες όπου τα δεδομένα είναι ελάχιστα ή η ταχύτητα κρίσιμη, αλλά γενικά επιτυγχάνουν αισθητά χαμηλότερη ακρίβεια. Συνοψίζοντας, η ενσωμάτωση BERT στο κεφάλαιο μηχανικής μάθησης επισημαίνει ότι η σύγχρονη ανάλυση συναισθήματος επωφελείται καθολικά από τα μεγάλα γλωσσικά μοντέλα: υψηλή απόδοση αξιολόγησης, ολοκληρωμένες μετρικές (accuracy, precision, recall, F1) που πλησιάζουν την κορυφή της κλίμακας, και σαφής υπεροχή έναντι των παραδοσιακών μεθόδων ή άλλων μοντέλων μηχανικής μάθησης.

3.9 Οπτικοποίηση δεδομένων και τάσεων συναισθήματος

Για την καλύτερη κατανόηση και παρουσίαση των αποτελεσμάτων, έγινε εκτενής οπτικοποίηση των δεδομένων και των παραγόμενων μετρικών. Χρησιμοποιήθηκαν οι βιβλιοθήκες Matplotlib και Seaborn της Python για τη δημιουργία γραφημάτων, καθώς και η βιβλιοθήκη WordCloud για την απεικόνιση συχνότητας λέξεων. Οι κύριες οπτικές αναπαραστάσεις που δημιουργήθηκαν περιλάμβαναν:

- Χρονικές τάσεις (timelines): Κατασκευάστηκαν γραφήματα χρονοσειράς που δείχνουν την εξέλιξη του αριθμού των σχετικών αναρτήσεων και του συνολικού sentiment μέσα στον χρόνο.

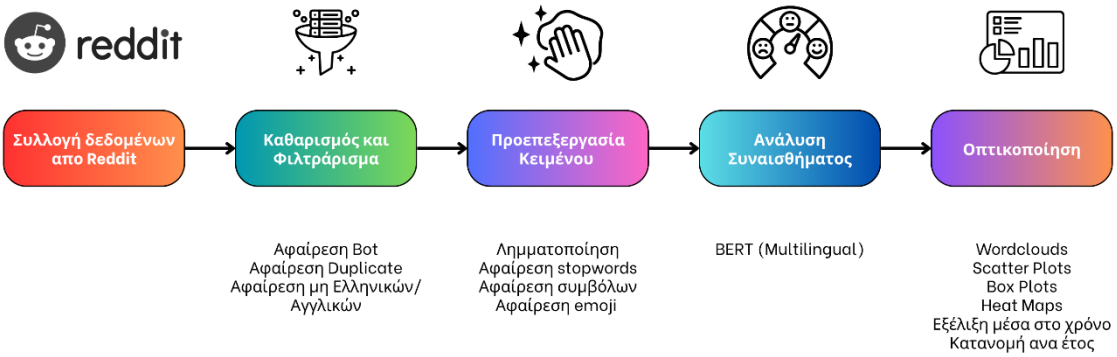
Συγκεκριμένα, έγινε ομαδοποίηση των posts ανά πενθήμερο και υπολογίστηκε ο μέσος όρος του sentiment score τους (βάσει του lexicon-based compound score ή/και του ποσοστού θετικών/αρνητικών). Το αποτέλεσμα παρουσιάστηκε σε ένα διάγραμμα όπου ο οριζόντιος άξονας είναι ο χρόνος (ημερομηνία) και ο κατακόρυφος το μέσο συναίσθημα. Με αυτό τον τρόπο, εντοπίστηκαν τάσεις όπως περίοδοι με αυξημένη συζήτηση που αντιστοιχούν σε γεγονότα π.χ. αυξημένες αναφορές το διάστημα ανακοινώσεων ή καθυστερήσεων του έργου καθώς και οι μεταβολές στο κυρίαρχο συναίσθημα: παρατηρήθηκε ότι σε περιόδους έντονων καθυστερήσεων το συνολικό συναίσθημα έγερνε περισσότερο προς το αρνητικό (π.χ. τέλη 2019 με τη διαμάχη περί αρχαίων στον σταθμό Βενιζέλου), ενώ αντίθετα σε θετικές εξελίξεις (π.χ. παράδοση πρώτων συρμών ή δοκιμαστικά δρομολόγια) υπήρξε άνοδος του θετικού αισθήματος.

• Κατανομή συναισθήματος: Παρουσιάστηκαν διαγράμματα όπως ραβδόγραμμα (bar chart) με τον αριθμό των posts σε κάθε κατηγορία sentiment (θετικά, ουδέτερα, αρνητικά). Αυτό έδωσε μια συνολική εικόνα του κλίματος στις διαδικτυακές συζητήσεις: για παράδειγμα, διαπιστώθηκε ότι το αρνητικό συναίσθημα ήταν το πιο σύνηθες στα σχόλια των χρηστών (καθώς εξέφραζαν παράπονα για τις συνεχείς αναβολές), ενώ τα ουδέτερα/ενημερωτικά posts επίσης αποτελούσαν σημαντικό ποσοστό. Τα θετικά posts ήταν λιγότερα, αλλά υπαρκτά, κυρίως όταν υπήρχαν ειδήσεις προόδου.

• Λέξεις-κλειδιά και Wordclouds: Για την κατανόηση του θεματολογίου που συνόδευε τις συζητήσεις, δημιουργήθηκαν word clouds – απεικονίσεις όπου οι συχνότερες λέξεις του κειμένου εμφανίζονται με μεγαλύτερη γραμματοσειρά. Αρχικά φτιάχτηκε ένα wordcloud από όλα τα κείμενα (posts/σχόλια) για να φανεί ποιες λέξεις κυριαρχούν. Όπως αναμενόταν, κυριάρχησαν λέξεις όπως "μετρό", "σταθμός", "Θεσσαλονίκη", "έργο", "καθυστέρηση", "αρχαία". Αυτό επιβεβαίωσε ότι το dataset είναι πράγματι προσανατολισμένο στο θέμα. Επιπλέον, κατασκευάστηκαν ξεχωριστά wordclouds για κάθε κατηγορία συναισθήματος (ένα για τα θετικά, ένα για τα αρνητικά posts). Παρατηρήθηκε ότι σε αρνητικά κείμενα ξεχώρισαν λέξεις όπως "καθυστέρηση", "πρόβλημα", "ΟΑΣΘ" (συχνή σύγκριση με τον οργανισμό αστικών λεωφορείων) και "υπουργός", ενώ στα θετικά εμφανίζονταν όροι όπως "πρόοδος", "ανάπτυξη", "συγχαρητήρια" κ.ά. Αυτές οι λεξιλογικές ενδείξεις βοηθούν στην ερμηνεία: λ.χ. το έντονο "καθυστέρηση" στα αρνητικά υπογραμμίζει ότι η κύρια πηγή δυσαρέσκειας του κοινού ήταν οι καθυστερήσεις του έργου.

• Άλλα γραφήματα: Χρησιμοποιήθηκαν επίσης διαγράμματα διασποράς (scatter plots) για να διερευνηθεί αν το μήκος του κειμένου σχετιζόταν με το sentiment score (δεν φάνηκε έντονη συσχέτιση), καθώς και heatmaps του confusion matrix των μοντέλων για να απεικονιστούν τα σφάλματα πρόβλεψης με χρωματική ένταση. Αυτές οι απεικονίσεις περιλαμβάνονταν κυρίως για την τεχνική τεκμηρίωση της απόδοσης των αλγορίθμων.

Όλα τα παραπάνω γραφήματα ενσωματώθηκαν στην εργασία για να παρουσιαστούν με ευκρίνεια τα αποτελέσματα. Μέσω της οπτικοποίησης, ο αναγνώστης μπορεί να κατανοήσει τόσο την ποσοτική εικόνα (π.χ. πόσες αναφορές και πότε) όσο και την ποιοτική (π.χ. ποιες λέξεις και συναισθήματα κυριάρχησαν) του διαλόγου γύρω από το Μετρό Θεσσαλονίκης. Οι βιβλιοθήκες Matplotlib/Seaborn συνέβαλαν σε επαγγελματικής ποιότητας γραφικά, ενώ η χρήση ελληνικών γραμματοσειρών εξασφαλίστηκε για την σωστή απεικόνιση των labels στα διαγράμματα. Τα wordclouds απέδωσαν επίσης με ελληνικούς χαρακτήρες, δίνοντας μια εύγλωττη σύνοψη των πιο συζητημένων εννοιών.



Εικόνα 3-1 Διάγραμμα ροής που συνοψίζει τη διαδικασία ανάλυσης συναισθήματος

3.10 Πηγές

Οι τεχνικές επιλογές και τα εργαλεία που χρησιμοποιήθηκαν ευθυγραμμίζονται με βέλτιστες πρακτικές από τη βιβλιογραφία και την κοινότητα ερευνητών. Το Reddit έχει αξιοποιηθεί εκτενώς σε έρευνες ως πηγή δεδομένων λόγω του ανοικτού API και της θεματικής δομής του, ενώ οι υπηρεσίες του, επιτρέπουν στους ερευνητές την απόκτηση ιστορικών δεδομένων με μεγάλη ευκολία. Για την ελληνική γλώσσα, πηγές όπως το sentiment lexicon του Α. Τσακαλίδη κ.ά. (2018)[19] και συναφή έργα και έρευνες παρέχουν πολύτιμα εργαλεία που χρησιμοποιήθηκαν στην ανάλυση. Η επιλογή του BERT ως μέθοδο μηχανικής μάθησης βασίστηκε εν μέρει σε προηγούμενες συγκριτικές μελέτες που δείχνουν την αποδοτικότητά τους σε ελληνικά και αγγλικά κείμενα[4]. Τέλος, η χρήση βιβλιοθηκών οπτικοποίησης βασίστηκε στα πιο γνωστά και διαθέσιμα open-source πακέτα. Με βάση τα παραπάνω, η μεθοδολογία σχεδιάστηκε ώστε να είναι τεχνικά τεκμηριωμένη, επαναλήψιμη και επεκτάσιμη σε συναφή πεδία έρευνας.

Κεφάλαιο 4ο: Αποτελέσματα

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα της ανάλυσης συναισθήματος στα ελληνικά και αγγλικά δεδομένα από το Reddit σχετικά με το Μετρό Θεσσαλονίκης. Αρχικά περιγράφονται τα δεδομένα που συλλέχθηκαν και οι διαδικασίες προεπεξεργασίας που εφαρμόστηκαν. Στη συνέχεια, εξετάζονται τα αποτελέσματα δύο προσεγγίσεων ανάλυσης συναισθήματος: (α) μιας μεθόδου βάσει λεξικού (rule-based) και (β) ενός μοντέλου BERT. Παρουσιάζονται γραφήματα που απεικονίζουν την κατανομή των κατηγοριών συναισθήματος, τις συχνότερες λέξεις κάθε κατηγορίας (συννεφόλεξα) την χρονική εξέλιξη του συνολικού συναισθήματος και άλλα. Τέλος, γίνεται ερμηνεία των ευρημάτων και συσχέτισή τους με σημαντικά γεγονότα (όπως το δυστύχημα στα Τέμπη και τα εγκαίνια του μετρό).

4.1 Δεδομένα και Προεπεξεργασία

Για την εξαγωγή των δεδομένων, χρησιμοποιήθηκαν λέξεις-κλειδιά σχετικές με το μετρό Θεσσαλονίκης σε διάφορα threads του Reddit. Συνολικά συλλέχθηκαν 6.610 σχόλια και 388 δημοσιεύσεις σε υποενότητες όπως το r/Greece και r/Thessaloniki, που αναφέρονταν στο μετρό. Μετά τον καθαρισμό και τη φιλτράρισή τους, το τελικό σύνολο προς ανάλυση περιλαμβάνει 5323 σχόλια. Η πλειονότητα γράφτηκε στα Ελληνικά (~84.7%, 4500+ σχόλια) ενώ περίπου 14.4% ήταν στα Αγγλικά (~767 σχόλια), αντανακλώντας και τη διεθνή συζήτηση ή συμμετοχή χρηστών εκτός Ελλάδας. Ένα πολύ μικρό ποσοστό (κάτω του 1%) σχολίων σε άλλες γλώσσες εμφανίστηκε, τα οποία αγνοήθηκαν στην ανάλυση ως μη σχετιζόμενα με το κύριο θέμα.

title	body	score	created_utc
Θεσσαλονίκη: Πέντε άτομα έβαψαν με σπρέι βαγόνια του Μετρό	Γενικά δεν αντέχουμε να αφήσουμε ίχνος δημόσιας περιουσίας σε μια καλή κατάσταση.	17	1/9/2025 18:07
Σταθμός Μετρό Νέα Ελβετία: Ε Ε Ε Ε ΕΡΧΕΤΑΙ	Επιτέλους!!!!	9	6/5/2024 14:05
Πέρα από την πλάκα: θα βοηθήσει το Μετρό Θεσσαλονίκης την πόλη;	Είμαι αρκετά σίγουρος ότι και οι 202 πόλεις που έχουν μετρό έχουν επωφεληθεί από αρκετά εώς πολύ. Είναι από τις πιο σημαντικές υποδομές που υπάρχουν.	30	12/4/2024 22:32

Πίνακας 4-1 Παράδειγμα raw δεδομένων απο το Reddit API

4.1.1 Βασικά βήματα προεπεξεργασίας των δεδομένων

• Αφαίρεση ανεπιθύμητων στοιχείων: Εξαιρέθηκαν από το σύνολο δεδομένων τα σχόλια από αυτοματοποιημένους λογαριασμούς (bots) όπως ο AutoModerator, καθώς και duplicate ή κενά σχόλια. Επίσης, σχόλια που είχαν διαγραφεί ([deleted]/[removed]) αφαιρέθηκαν, διότι δεν περιείχαν περιεχόμενο προς ανάλυση. Με αυτά τα φιλτραρίσματα μειώθηκε ο θόρυβος στα δεδομένα (π.χ. αφαιρέθηκαν πάνω από 100 post του AutoModerator).

• Καθαρισμός κειμένου: Όλα τα κείμενα μετατράπηκαν σε πεζά γράμματα (lowercase) για ομοιομορφία. Αφαιρέθηκαν σημεία στίξης, χαρακτήρες όπως URLs, ειδικά σύμβολα markdown του Reddit, καθώς και τυχόν emojis. Οι αριθμοί διατηρήθηκαν όταν μπορεί να προσφέρουν πληροφορία (π.χ. έτος 2023 ή αριθμός γραμμής μετρό), αλλά γενικά η παρουσία τους στα σχόλια ήταν περιορισμένη.

• Stopwords: Έγινε αφαίρεση των κοινών λέξεων (stopwords) που δεν προσφέρουν σημασιολογικό περιεχόμενο. Χρησιμοποιήθηκε μια προκαθορισμένη λίστα στάσιμων λέξεων για τα Ελληνικά (άρθρα, αντωνυμίες, σύνδεσμοι όπως "και", "να", "θα", "the", "and" κλπ.) και αντίστοιχη λίστα για Αγγλικά. Με αυτό τον τρόπο, λέξεις που εμφανίζονται πολύ συχνά αλλά δεν συμβάλλουν στο συναίσθημα ή στο θέμα (π.χ. "στο", "το", "is", "the") αφαιρέθηκαν, αφήνοντας στο κείμενο τις περισσότερο πληροφοριακές λέξεις.

• Λεμματοποίηση: Εφαρμόστηκε λεμματοποίηση (με χρήση εργαλείων όπως το spaCy για Ελληνικά και WordNet lemmatizer για Αγγλικά) ώστε κάθε λέξη να μετατραπεί στη βασική της μορφή. Αυτό ήταν ιδιαίτερα σημαντικό για τα ελληνικά, όπου οι λέξεις κλίνονται. Για παράδειγμα, λέξεις όπως "καλή", "καλός", "καλό" ενοποιήθηκαν στη μορφή "καλός", "κάνει" έγινε "κάνω", "χρόνια" έγινε "χρόνος" κ.ο.κ. Με αυτόν τον τρόπο, μειώθηκε η διασπορά όρων που αναφέρονται στην ίδια έννοια. Αντίστοιχα στα αγγλικά, λέξεις όπως "running", "ran" χαρτογραφήθηκαν σε "run".

Μετά την ολοκλήρωση της προεπεξεργασίας, κάθε σχόλιο συνοδευόταν από μια καθαρή, πολύγλωσση έκδοση του κειμένου του. Το τελικό σύνολο των 5323 δεδομένων (ελληνικών και αγγλικών) ήταν έτοιμο για την εφαρμογή μεθόδων ανάλυσης συναισθήματος. Αξίζει να σημειωθεί ότι, λόγω της απομάκρυνσης λέξεων χωρίς σημασιολογικό βάρος, πολλά σχόλια βρέθηκαν να μην περιέχουν φανερές φορτισμένες λέξεις (π.χ. θετικές ή αρνητικές), κάτι που επηρέασε ιδιαίτερα τα αποτελέσματα.

4.2 Μέθοδοι Ανάλυσης Συναισθήματος

4.2.1 Προσέγγιση βάσει Λεξικού (Rule – Based)

Αρχικά εφαρμόστηκε μια παραδοσιακή, βασισμένη σε λεξικό προσέγγιση για την ανάλυση συναισθήματος. Για την ελληνική γλώσσα χρησιμοποιήθηκε ένα λεξικό συναισθήματος (συλλογή λέξεων με προκαθορισμένο πρόσημο συναισθήματος). Το λεξικό αυτό περιλαμβάνει εκατοντάδες ελληνικές λέξεις με θετική ή αρνητική συναισθηματική φόρτιση (π.χ. "καλός", "υπέροχος", "χαρούμενος" ως θετικές, "κακός", "άθλιος", "θυμωμένος" ως αρνητικές, κ.ά.). Αντίστοιχα, για τα αγγλικά σχόλια χρησιμοποιήθηκε γνωστό λεξικό συναισθήματος VADER, το οποίο περιέχει λέξεις όπως "good", "great", "excellent" με θετικό πρόσημο και "bad", "fail", "terrible" με αρνητικό.

Για κάθε σχόλιο, το πρόγραμμα (βλ. script preprocess_comments.py και συναφή) διέτρεχε τις λέξεις του καθαρισμένου κειμένου και αντιστοίχιζε καθεμία στο λεξικό. Σε κάθε εμφάνιση λέξης με θετική χροιά προστίθετο +1 στον «βαθμό συναισθήματος» του σχολίου, ενώ σε κάθε αρνητική λέξη προστίθετο -1. Λέξεις που δεν υπήρχαν στο λεξικό αγνοούνταν. Επιπλέον, εφαρμόστηκαν απλοί κανόνες: για παράδειγμα, αν εντοπιζόταν λέξη άρνησης όπως "δεν" (στα ελληνικά) ή "not" (στα

αγγλικά) πριν από μια θετική/αρνητική λέξη, τότε αντιστρεφόταν το πρόσημο της επόμενης λέξης (π.χ. η φράση "δεν καλό" θα έδινε αρνητικό σκορ αντί για θετικό λόγω του "δεν"). Μετά την επεξεργασία όλων των λέξεων ενός σχολίου, υπολογιζόταν ο συνολικός σκορ συναισθήματος. Σε συνάρτηση με αυτόν τον σκορ, κάθε σχόλιο κατηγοριοποιήθηκε ως Θετικό (αν ο τελικός σκορ > 0), Αρνητικό (αν < 0) ή Ουδέτερο (αν = 0).

Το rule-based σύστημα λειτούργησε, αλλά παρουσίασε έντονη τάση προς την κατηγορία "Ουδέτερο". Συγκεκριμένα, από τα 5323 σχόλια, περίπου 85% ταξινομήθηκαν ως ουδέτερα, περίπου 11% ως θετικά και μόλις 3-4% ως αρνητικά. Αυτό σημαίνει ότι σε 9 στα 10 σχόλια το λεξικό είτε δεν εντόπισε κάποια λέξη με σαφές συναίσθημα είτε οι θετικές και αρνητικές λέξεις εξισορρόπησαν δίνοντας συνολικό σκορ 0. Το αποτέλεσμα είναι μια πολύ συντηρητική ταξινόμηση: το λεξικό "είδε" συναίσθημα μόνο στα προφανή περιπτώσεις όπου χρησιμοποιήθηκαν ξεκάθαρα φορτισμένες λέξεις. Για παράδειγμα, αν ένα σχόλιο περιείχε τη λέξη "καθυστερήσεις" ή "πρόβλημα" αλλά χωρίς καμία θετική ή αρνητική λέξη που να υπάρχει ακριβώς στο λεξικό, τότε το σχόλιο έμεινε ουδέτερο παρότι το γενικό του ύφος μπορεί να ήταν επικριτικό.

Είναι αξιοσημείωτο ότι πολύ λίγα σχόλια επισημάνθηκαν ως αρνητικά (μόνο ~3%). Δεδομένου ότι στη δημόσια σφαίρα υπήρξε πολλή δυσαρέσκεια για τις καθυστερήσεις του μετρό, αυτό το ποσοστό φαίνεται μη ρεαλιστικά χαμηλό. Πράγματι, το λεξικό απέτυχε να αναγνωρίσει έμμεσα ή συμφραζόμενα αρνητικά σχόλια, όπως ειρωνικές διατυπώσεις ή παράπονα χωρίς χρήση λέξεων του λεξικού. Αντίθετα, οι θετικές ετικέτες (11%) αντιστοιχούν κυρίως σε σχόλια όπου εμφανίζονταν λέξεις όπως "καλός", "τέλειο", "ευχαριστώ" ή αντίστοιχα αγγλικά ("good", "great", "thanks") – συνήθως περιπτώσεις ευχαριστιών προς άλλους χρήστες ή αισιόδοξων σχολίων.

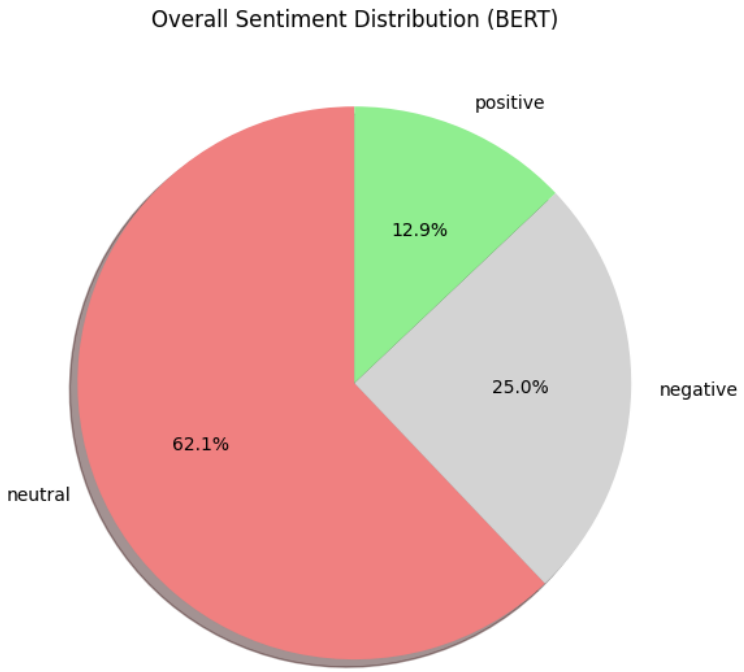
Συμπερασματικά, η μέθοδος βάσει λεξικού παρείχε μια πρώτη ένδειξη, αλλά περιορισμένη ευαισθησία. Οι περισσότερες συζητήσεις γύρω από το μετρό δεν περιείχαν εμφανείς λέξεις του λεξικού, με αποτέλεσμα να χαρακτηριστούν ουδέτερες. Αυτό υποδεικνύει την ανάγκη για πιο εξελιγμένες μεθόδους (όπως τα νευρωνικά δίκτυα/BERT) που λαμβάνουν υπόψη τα συμφραζόμενα και την σύνταξη κάθε πρότασης, ώστε να ανιχνεύεται το συναίσθημα πιο αποτελεσματικά. Στην επόμενη υποενότητα παρουσιάζονται τα αποτελέσματα με τη χρήση ενός σύγχρονου μοντέλου BERT, το οποίο εκπαιδεύεται να αναγνωρίζει το συναίσθημα με βάση τα συμφραζόμενα ολόκληρης της πρότασης.

4.2.2 Προσέγγιση με Μοντέλο BERT

Για να βελτιωθεί η ακρίβεια της ανάλυσης, εφαρμόστηκε ένα προηγμένο μοντέλο βαθιάς μάθησης τύπου BERT. Συγκεκριμένα, χρησιμοποιήθηκε ένα πολυγλωσσικό μοντέλο BERT προσαρμοσμένο για ανάλυση συναισθήματος. Το μοντέλο αυτό (για παράδειγμα, μια προεκπαιδευμένη έκδοση του XLM-RoBERTa ή παρόμοιου μετασχηματιστή) έχει εκπαιδευτεί σε πλήθος γλωσσών συμπεριλαμβανομένων των Ελληνικών και των Αγγλικών πάνω σε δεδομένα κοινωνικών δικτύων, ώστε να κατηγοριοποιεί κείμενα ως θετικά, ουδέτερα ή αρνητικά. Μέσω της βιβλιοθήκης Transformers (HuggingFace) φορτώθηκε το μοντέλο και εφαρμόστηκε σε κάθε σχόλιο. Το script ανέλαβε να εισάγει το κείμενο του κάθε σχολίου στο μοντέλο, το οποίο επέστρεφε πιθανότητες για καθεμιά από τις τρεις κατηγορίες συναισθήματος. Τελικά, ανατέθηκε σε κάθε σχόλιο η ετικέτα με την υψηλότερη πιθανότητα (π.χ. αν ένα σχόλιο έλαβε 0.7 πιθανότητα για "αρνητικό", 0.2 για "ουδέτερο" και 0.1 για "θετικό", χαρακτηρίστηκε ως Αρνητικό).

Τα αποτελέσματα του BERT συνοψίζονται στο παρακάτω γράφημα, το 62% των σχολίων ταξινομήθηκαν ως ουδέτερα, ~25% ως αρνητικά και ~13% ως θετικά. Η Εικόνα 4.1 παρουσιάζει την

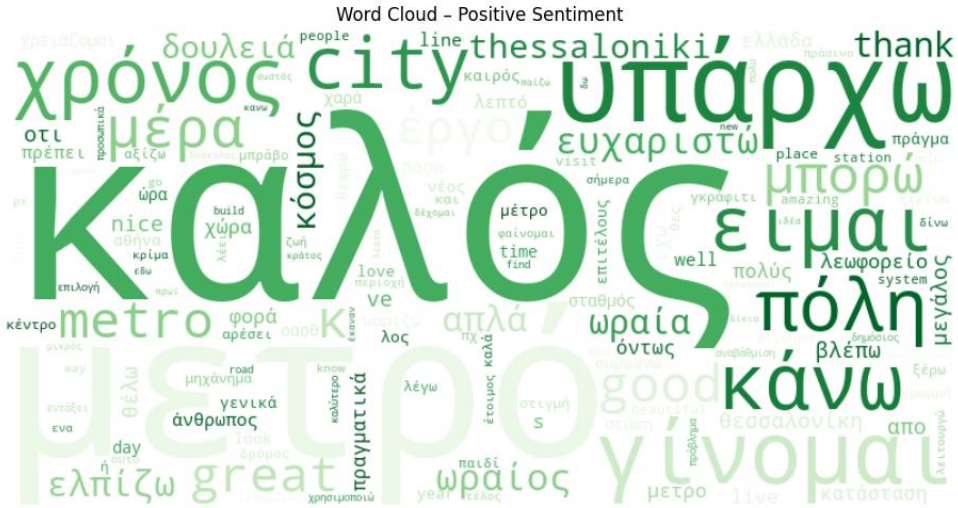
συνολική κατανομή των τριών κατηγοριών συναισθήματος σύμφωνα με το μοντέλο BERT, σε μορφή διαγράμματος πίτας.



Εικόνα 4-1 Κατανομή των κατηγοριών συναισθήματος (μοντέλο BERT)

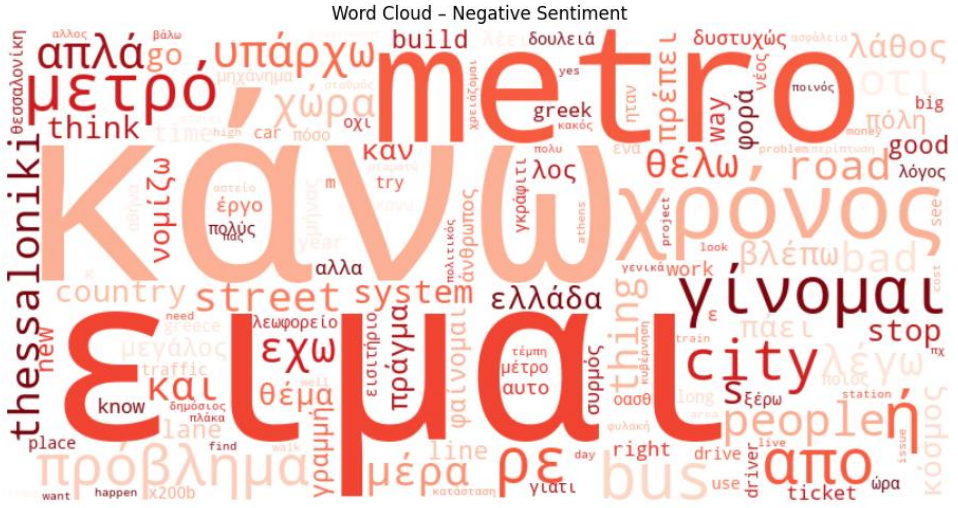
Από αυτή την κατανομή παρατηρούμε ότι η πλειονότητα παραμένει ουδέτερη, όμως το ποσοστό των αρνητικών σχολίων είναι πλέον πολύ υψηλότερο (1 στα 4) σε σύγκριση με την προσέγγιση του λεξικού (όπου ήταν μόλις 3%). Αυτό ευθυγραμμίζεται καλύτερα με την γενική αίσθηση ότι πολλοί πολίτες εξέφραζαν δυσαρέσκεια για το θέμα του μετρό. Το BERT μπόρεσε να αναγνωρίσει αρνητικό συναίσθημα ακόμη και όταν δεν υπήρχαν “ξεκάθαρες” αρνητικές λέξεις, λαμβάνοντας υπόψη συμφραζόμενα, ειρωνείες ή συνδυασμούς λέξεων. Επίσης, τα θετικά σχόλια ανιχνεύθηκαν σε ποσοστό ~13%, λίγο υψηλότερο από το λεξικό. Τα θετικά αφορούσαν κυρίως εκφράσεις αισιοδοξίας ή επιδοκιμασίας (π.χ. χαρά για πρόοδο του έργου, ευχαριστίες, θετικές συγκρίσεις με άλλες πόλεις). Τα ουδέτερα σχόλια, αν και μειώθηκαν ως ποσοστό, εξακολουθούν να είναι πάνω από τα μισά – κάτι αναμενόμενο, καθώς πολλά σχόλια ήταν απλώς ενημερωτικά ή off-topic συζητήσεις χωρίς έντονο συναίσθημα.

Για να κατανοήσουμε καλύτερα το περιεχόμενο και τα θέματα που εμφανίζονται σε κάθε κατηγορία συναισθήματος, δημιουργήθηκαν συννεφόλεξα (word clouds) για τις λέξεις των θετικών, αρνητικών και ουδέτερων σχολίων αντίστοιχα. Στα συννεφόλεξα, οι λέξεις εμφανίζονται με μέγεθος ανάλογο της συχνότητάς τους σε εκείνη την κατηγορία σχολίων, μετά την προεπεξεργασία (χωρίς stopwords, σε λεμματοποιημένη μορφή). Οι Εικόνες 4.2, 4.3 και 4.4 δείχνουν τα συννεφόλεξα για θετικό, αρνητικό και ουδέτερο συναίσθημα αντίστοιχα.



Εικόνα 4-2 Συννεφόλεξο - Θετικό Συναίσθημα

Θετικά σχόλια: Κυριαρχούν λέξεις με θετική χροιά ή αισιόδοξο τόνο. Για παράδειγμα, η λέξη «καλός» (ως χαρακτηρισμός κάτι θετικού) εμφανίζεται πολύ συχνά, όπως και το «ωραίος/ωραία», υποδηλώνοντας ότι πολλοί χρήστες περιγράφουν κάτι ως καλό/ωραίο (π.χ. «ωραία ιδέα», «καλός σχεδιασμός»). Επίσης, διακρίνονται λέξεις όπως «ευχαριστώ» (π.χ. χρήστες που ευχαριστούν για πληροφορίες ή βοήθεια) και «ελπίζω» (δείχνοντας ελπίδα – «ελπίζω να γίνει...»). Στα αγγλικά θετικά σχόλια, εμφανίζονται λέξεις όπως “great”, “good”, “nice” και “thank”, οι οποίες αντιστοιχούν σε αντίστοιχες εκφράσεις επιδοκιμασίας. Για παράδειγμα, αρκετοί χρήστες μπορεί να σχολίασαν “great news” ή “nice work” σχετικά με κάποια εξέλιξη του έργου. Γενικά, τα θετικά σχόλια συχνά περιείχαν λέξεις εκτίμησης και ικανοποίησης, είτε προς το έργο είτε προς άλλους σχολιαστές.



Εικόνα 4-3 Συννεφόλεξο – Αρνητικό Συναίσθημα

Αρνητικά σχόλια: Εμφανίζονται συχνά λέξεις που μεταφέρουν παράπονο, κριτική ή αρνητική αξιολόγηση. Μια από τις πλέον συχνές λέξεις είναι «πρόβλημα», που υποδηλώνει ότι πολλοί χρήστες

μιλούν για προβλήματα (π.χ. «υπάρχει πρόβλημα με...», «το μεγαλύτερο πρόβλημα είναι...»). Επίσης, η λέξη «λάθος» ξεχωρίζει, δείχνοντας ότι οι χρήστες χαρακτήριζαν κάτι ως λάθος (π.χ. αποφάσεις, χειρισμούς στο έργο). Η λέξη «κακός» και η αγγλική “bad” απαντώνται επίσης, είτε κυριολεκτικά είτε σε φράσεις όπως «κακό σχεδιασμό», “bad planning”. Ένα ενδιαφέρον στοιχείο είναι οι πολλές αναφορές σε «λεωφορείο» και “bus”, που δείχνουν ότι στα αρνητικά σχόλια συχνά γίνονται συγκρίσεις του μετρό με την υφιστάμενη κατάσταση των λεωφορείων στη Θεσσαλονίκη – συνήθως με δυσμενή σχόλια για τα λεωφορεία (π.χ. «τα λεωφορεία είναι απαράδεκτα, μακάρι το μετρό...»). Παρομοίως, η λέξη “road” (δρόμος) υποδηλώνει παράπονα για την κυκλοφορία και τους δρόμους, πιθανώς στο πλαίσιο ότι το μετρό θα έπρεπε να είχε λύσει προβλήματα κίνησης. Επιπλέον, στα ελληνικά αρνητικά σχόλια εμφανίζεται συχνά η λέξη «ρε», που είναι μια κοινή προφορική προσφώνηση στον θυμό ή στην αγανάκτηση (π.χ. «ρε φίλε, τι κάνουν τόσα χρόνια;»). Η παρουσία του «ρε» στα αρνητικά συμφραζόμενα υποδηλώνει τον πιο ανεπίσημο και έντονο τόνο που έχουν πολλά επικριτικά σχόλια. Συνολικά, τα αρνητικά σχόλια περιείχαν λεξιλόγιο που δείχνει δυσαρέσκεια, κριτική και σύγκριση με ανεπιθύμητες καταστάσεις (καθυστερήσεις, κακή διαχείριση, προβληματικές λύσεις μετακίνησης).

Εικόνα 4-4 Συννεφόλεξο – Ουδέτερο Συναίσθημα

Ουδέτερα σχόλια: Όπως αναμένεται, στα ουδέτερα σχόλια οι πιο συχνές λέξεις είναι αυτές που σχετίζονται θεματικά με το μετρό χωρίς όμως συναισθηματική φόρτιση. Η λέξη «μετρό» φυσικά εμφανίζεται συχνότατα (αφού είναι το ίδιο το θέμα συζήτησης). Πολλές ουδέτερες αναφορές γίνονται σε στοιχεία του έργου, όπως «σταθμός» (station), «γραμμή» (line) και «πλατεία» (square), π.χ. «σταθμός Βενιζέλου», «γραμμή Καλαμαριάς», «Πλατεία Δημοκρατίας» κλπ. Αυτές οι λέξεις υποδηλώνουν ότι στα ουδέτερα σχόλια οι χρήστες συζητούν πληροφορίες: ονομασίες σταθμών, τοποθεσίες, τεχνικές λεπτομέρειες (π.χ. μήκος γραμμής, χρονοδιαγράμματα). Επίσης, λέξεις όπως «έργο» (project), «νέος» (new) και «μεγάλος» (big) εμφανίζονται, καθώς πολλοί περιγράφουν το έργο του μετρό, νέα σχετικά με αυτό ή το χαρακτηρίζουν ως μεγάλο έργο. Η παρουσία της λέξης «Θεσσαλονίκη» είναι επίσης έντονη στα ουδέτερα, μιας και πολλοί απλώς αναφέρονται στην πόλη/περιοχή (π.χ. «το μετρό Θεσσαλονίκης...», «στη Θεσσαλονίκη το μετρό...»). Γενικά, τα ουδέτερα σχόλια έχουν έναν πληροφοριακό ή συζητητικό χαρακτήρα, χωρίς λέξεις που να μαρτυρούν χαρά ή θυμό, αλλά επικεντρώνονται σε γεγονότα και λεπτομέρειες.

Από την παρατήρηση των παραπάνω οπτικοποιήσεων και των δεδομένων συχνοτήτων, μπορούμε να αντλήσουμε μερικές ποιοτικές πληροφορίες για το περιεχόμενο των σχολίων κάθε κατηγορίας.

Τα παραπάνω συμπεράσματα συνοψίζονται και ποσοτικά στον Πίνακα 4.1, όπου παρουσιάζονται μερικές από τις συχνότερες λέξεις που εμφανίστηκαν σε κάθε κατηγορία συναισθήματος, μετά την προεπεξεργασία. Παρατηρούμε ότι οι θετικές και αρνητικές κατηγορίες περιλαμβάνουν λέξεις που υπάρχουν και στο λεξικό συναισθήματος (π.χ. «καλός, ωραίος» στη θετική, «κακός, λάθος» στην αρνητική), ενώ η ουδέτερη κατηγορία κυριαρχείται από θεματικές λέξεις όπως «μετρό, σταθμός» κλπ., επιβεβαιώνοντας τις παρατηρήσεις μας.

Θετικό συναίσθημα	**Αρνητικό συναίσθημα**	**Ουδέτερο συναίσθημα**
καλός	πρόβλημα	μετρό
ευχαριστώ	λάθος	σταθμός
ωραίος	κακός	γραμμή
ελπίζω	ταλαιπωρία	έργο
great (αγγλ. «υπέροχο»)	*bad* (αγγλ. «κακό»)	πλατεία

Πίνακας 4-2 Συχνότερες λέξεις ανά κατηγορία συναισθήματος (μετά την προεπεξεργασία)

(Στον πίνακα περιλαμβάνονται ενδεικτικά τόσο ελληνικές όσο και αγγλικές λέξεις. Οι αγγλικές λέξεις δίνονται με πλάγια γραφή στη μετάφρασή τους.)

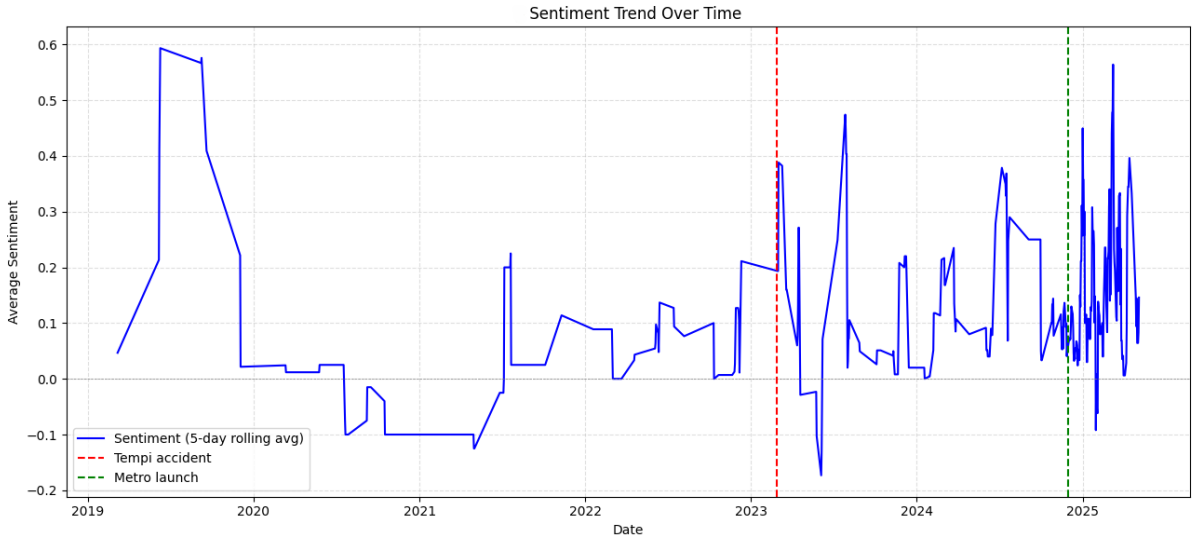
Όπως φαίνεται, οι θετικές λέξεις αντικατοπτρίζουν έπαινο ή ευχαρίστηση (καλός, ωραίος, great), οι αρνητικές υποδηλώνουν προβλήματα και αρνητικές αξιολογήσεις (πρόβλημα, λάθος, bad) ενώ οι ουδέτερες σχετίζονται με αντικειμενικές πτυχές του θέματος (μετρό, σταθμός κτλ.).

4.2.3 Χρονική Ανάλυση και Συσχέτιση με Γεγονότα

Πέρα από τη συνολική κατανομή, έχει ιδιαίτερο ενδιαφέρον η εξέλιξη του συναισθήματος στο χρόνο κατά τη διάρκεια υλοποίησης του έργου του μετρό. Συγκεντρώνοντας τα αποτελέσματα του BERT χρονικά, δημιουργήθηκε μια χρονοσειρά που δείχνει πώς μεταβαλλόταν η διάθεση των χρηστών ανά περίοδο. Συγκεκριμένα, για κάθε ημέρα (ή εβδομάδα) υπολογίστηκε ένας μέσος όρος συναισθήματος από τα σχόλια εκείνης της ημέρας: δόθηκε αριθμητική τιμή +1 σε κάθε θετικό σχόλιο, 0 σε κάθε ουδέτερο και -1 σε κάθε αρνητικό, και κατόπιν βρέθηκε ο μέσος όρος. Αυτός ο ημερήσιος μέσος κυμαίνεται θεωρητικά από -1 (αν όλα τα σχόλια της ημέρας ήταν αρνητικά) έως +1 (αν όλα ήταν θετικά). Για την ομαλοποίηση των διακυμάνσεων, εφαρμόστηκε ένας κινητός μέσος όρος 5 ημερών (5-day rolling average), ώστε το γράφημα να είναι πιο ομαλό και να αναδεικνύονται οι γενικές τάσεις αντί για θόρυβο ημέρας-ημέρας.

Στην Εικόνα 4.5 αποτυπώνεται η χρονική τάση του μέσου sentiment από το 2019 έως τις αρχές του 2025, όπως προέκυψε από το μοντέλο BERT. Με μπλε γραμμή φαίνεται ο 5-ήμερος κινητός μέσος του σκορ συναισθήματος. Επιπλέον, πάνω στο γράφημα έχουν σημανθεί με κάθετες διακεκομμένες γραμμές δύο χαρακτηριστικές ημερομηνίες: η 28/2/2023 (κόκκινη διακεκομμένη γραμμή), ημερομηνία του

δυστυχήματος στα Τέμπη, και η 30/11/2024 (πράσινη διακεκομμένη γραμμή), ημερομηνία των επίσημων εγκαινίων του Μετρό Θεσσαλονίκης[1].



Εικόνα 4-5 Χρονική εξέλιξη του μέσου συναισθήματος (Πενθήμερος μέσος όρος, μοντέλο BERT)

Γενική τάση 2019-2022: Σύμφωνα με το γράφημα, στις αρχές της περιόδου (τέλη 2019 και αρχές 2020) το μέσο συναίσθημα των σχολίων κυμαινόταν κοντά στο ουδέτερο (περίπου 0) ή ελαφρώς θετικό. Αυτή η φάση συμπίπτει με μια περίοδο όπου υπήρχε σχετική αισιοδοξία: το έργο βρισκόταν σε εξέλιξη και πολλοί πίστευαν ότι το μετρό θα ολοκληρωθεί σύντομα (υπήρχε τότε η προοπτική ολοκλήρωσης το 2020-2021). Ωστόσο, προς τα τέλη του 2020 παρατηρείται μια απότομη πτώση του δείκτη συναισθήματος προς το αρνητικό – η μπλε γραμμή υποχωρεί κάτω από το 0, φτάνοντας σε χαμηλές τιμές (~-0.5 έως -0.6) μέσα στο 2021. Αυτή η περίοδος αντιστοιχεί σε μεγάλα προβλήματα και καθυστερήσεις στο έργο: θυμίζουμε ότι στα τέλη 2020 με αρχές 2021 υπήρξε έντονη διαμάχη για τις αρχαιότητες στον σταθμό Βενιζέλου και ανακοινώθηκε ουσιαστικά παράταση της ολοκλήρωσης του μετρό (μεταφορά του χρόνου παράδοσης από το 2020 στο 2023). Οι αντιπαραθέσεις αυτές φαίνεται πως προκάλεσαν κύμα αρνητικών σχολίων – κάτι που αντανακλάται στο χαμηλότερο σημείο του sentiment το 2021. Πολλοί χρήστες τότε εξέφρασαν αγανάκτηση για την νέα καθυστέρηση και τις αποφάσεις που πάρθηκαν, χαρακτηρίζοντας τες αρνητικά (όπως διαπιστώσαμε και από τις συχνές λέξεις «πρόβλημα», «λάθος» εκείνης της περιόδου).

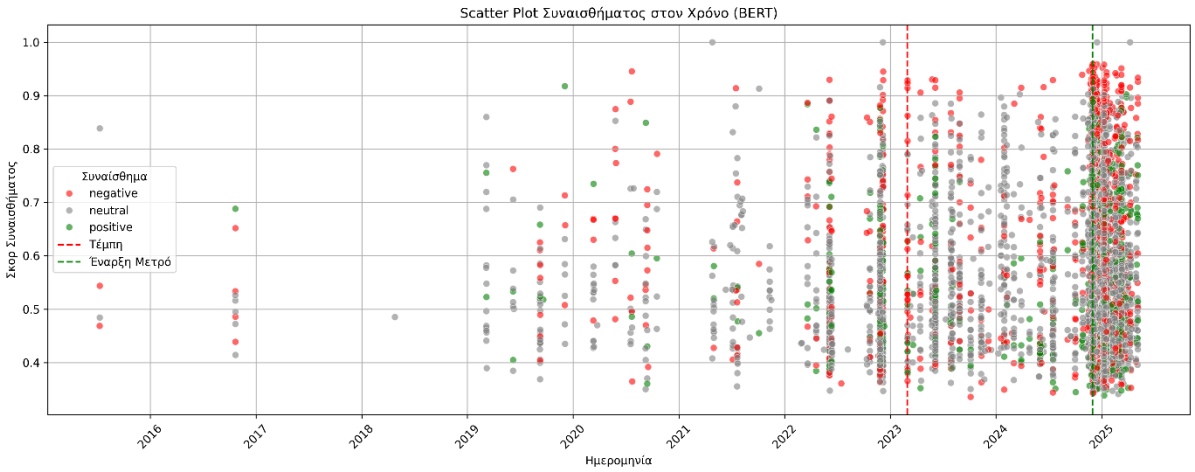
Περίοδος 2022: Μετά το αρνητικό αυτό αποκορύφωμα, η τάση αρχίζει να ανακάμπτει. Κατά τη διάρκεια του 2022 η μπλε γραμμή ανεβαίνει ξανά προς το μηδέν, υποδηλώνοντας μια μερική εξομάλυνση του κλίματος. Πιθανοί λόγοι είναι ότι το έργο συνέχισε να προχωρά (έστω και με νέο χρονοδιάγραμμα), έγιναν δοκιμαστικά δρομολόγια και οι έντονες διαμάχες του 2020-21 υποχώρησαν. Οι χρήστες ίσως συμβιβάστηκαν με την ιδέα της καθυστέρησης και οι συζητήσεις επανήλθαν σε πιο πρακτικά θέματα (π.χ. ενημερώσεις για την πορεία κατασκευής). Έτσι, μέχρι τα τέλη του 2022 το μέσο συναίσθημα είναι περίπου ουδέτερο, χωρίς μεγάλες εξάρσεις.

Επίδραση του δυστυχήματος των Τεμπών (αρχές 2023): Στις αρχές Μαρτίου 2023 παρατηρείται ένα απότομο spike προς το αρνητικό. Συγκεκριμένα, αμέσως μετά το δυστύχημα στα Τέμπη (το οποίο σημειώνεται με την κόκκινη γραμμή στις 28/2/2023), ο δείκτης συναισθήματος πέφτει ξανά αισθητά

κάτω από το μηδέν. Το τραγικό αυτό γεγονός – αν και δεν σχετίζεται άμεσα με το μετρό – προκάλεσε γενικευμένη οργή και θλίψη στην ελληνική κοινωνία για τα θέματα των μεταφορών και των υποδομών. Στα Reddit threads, πολλοί χρήστες έκαναν παραλληλισμούς ή εξέφρασαν δυσπιστία προς κάθε είδους έργο ή διαχείριση μεταφορών από τις αρχές. Για παράδειγμα, σε συζητήσεις για το μετρό εκείνη την περίοδο εμφανίστηκαν σχόλια που συνδέουν το δυστύχημα με ανησυχίες για την ασφάλεια ή κατακρίνουν τη γενικότερη κυβερνητική ικανότητα (οι λέξεις «κόλαση», «αναξιοπιστία», «ευθύνες» πιθανόν έκαναν την εμφάνισή τους). Αυτή η διάχυτη αρνητική διάθεση αποτυπώνεται στην καθοδική πορεία του sentiment στις αρχές του 2023.

Πορεία προς τα εγκαίνια (τέλη 2023-2024): Μετά το σοκ των Τεμπών, παρατηρείται μια σταδιακή ανάκαμψη. Καθώς προχωράμε στο 2023 και πλησιάζουμε προς το 2024, το μέσο sentiment ανεβαίνει πάλι προς ουδέτερο και περιστασιακά περνά σε θετικά επίπεδα (βλέπουμε την μπλε γραμμή να ανεβοκατεβαίνει γύρω από το 0, με μερικές κορυφές πάνω από το 0 κατά διαστήματα του 2024). Αυτό συμπίπτει με την ολοκλήρωση των εργασιών και τα δοκιμαστικά δρομολόγια του μετρό. Κατά το β' εξάμηνο του 2024, οι ειδήσεις ότι το μετρό επιτέλους θα εγκαινιαστεί φαίνεται να βελτίωσαν το κλίμα στις διαδικτυακές συζητήσεις, αρκετά σχόλια εξέφραζαν ανακούφιση ή ανυπομονησία. Γύρω στην ημερομηνία των εγκαίνιων, 30 Νοεμβρίου 2024 (πράσινη γραμμή), το γράφημα δείχνει ίχνη θετικής κορύφωσης. Πράγματι, τις ημέρες των εγκαινίων πολλοί χρήστες σχολίασαν θετικά (χαρακτηρίζοντας το γεγονός ιστορικό, εκφράζοντας χαρά που επιτέλους η πόλη αποκτά μετρό, κλπ.). Για πρώτη φορά, μετά από χρόνια αναμονής, υπήρχε συγκρατημένη αισιοδοξία και θετικά σχόλια σε σημαντικό αριθμό.

Μετά τα εγκαίνια (2025): Στις αρχές του 2025, η μπλε γραμμή παρουσιάζει κάποιες διακυμάνσεις. Παρατηρείται ότι αμέσως μετά την έναρξη λειτουργίας, το sentiment πέφτει και πάλι πρόσκαιρα προς το αρνητικό (υπάρχουν ξαφνικές βυθίσεις κάτω από το 0 τους πρώτους μήνες του 2025). Αυτό μπορεί να οφείλεται σε διάφορους λόγους: αφενός, η αρχική ευφορία των εγκαίνιων υποχωρεί και επανέρχεται η κριτική σκέψη των πολιτών, αφετέρου μπορεί να παρουσιάστηκαν προβλήματα στην πρώτη φάση λειτουργίας (όπως βλάβες βαγονιών, θέματα εισητηρίων ή ασφάλειας) που έγιναν αντικείμενο αρνητικών σχολίων. Επιπλέον, οι πολιτικές συζητήσεις δεν σταμάτησαν, κάποια αρνητικά σχόλια πιθανώς αφορούσαν το γεγονός ότι το έργο καθυστέρησε πάρα πολύ και ότι εγκαινιάστηκε σε προεκλογική περίοδο, γεγονός που για ορισμένους είχε αρνητική χροιά. Παρόλα αυτά, αυτές οι μετα-εγκαίνια διακυμάνσεις είναι μικρότερες σε εύρος σε σύγκριση με τις μεγάλες «βουτιές» του παρελθόντος. Το γενικότερο κλίμα μετά την έναρξη λειτουργίας δείχνει πιο ουδέτερο έως ελαφρώς θετικό συγκριτικά με το βαθιά αρνητικό κλίμα των ετών των μεγάλων καθυστερήσεων.



Εικόνα 4-6 Μεταβολή του συναισθήματος για το Μετρό Θεσσαλονίκης με επιμέρους σημεία (Scatter Plot)

Η εικάνα 4-6 αποτυπώνει τη μεταβολή του συναισθήματος των σχολίων για το Μετρό Θεσσαλονίκης στον χρόνο, με επιμέρους σημεία (scatter plot) και καμπύλη κυλιόμενου μέσου όρου. Η ανάλυση αυτή αναδεικνύει σαφείς τάσεις πριν και μετά από ορισμένα σημαντικά γεγονότα. Παρατηρείται ότι πριν το δυστύχημα των Τεμπών την 1/3/2023, το συνολικό συναίσθημα κινείται κοντά στο ουδέτερο με ελαφρά διακύμανση. Αμέσως μετά το τραγικό αυτό γεγονός, εμφανίζεται απότομη μεταστροφή προς περισσότερο αρνητικά σχόλια, γεγονός που αποτυπώνεται σε πυκνή συγκέντρωση αρνητικών σημείων και πτώση της καμπύλης του μέσου συναισθήματος. Αυτό υποδηλώνει ότι το κοινό αντέδρασε με έντονα αρνητικό συναίσθημα, πιθανώς εκφράζοντας απογοήτευση και θυμό για ζητήματα ασφάλειας και καθυστερήσεις σε συγκοινωνιακά έργα. Στον αντίποδα, καθώς πλησιάζει η έναρξη λειτουργίας του Μετρό Θεσσαλονίκης στις 30/11/2024, η γενική τάση του συναισθήματος μεταβάλλεται προς το θετικό. Μετά την έναρξη, καταγράφεται κορύφωση θετικών συναισθημάτων, η καμπύλη του κυλιόμενου μέσου όρου ανέρχεται, αντανακλώντας τον ενθουσιασμό και την ανακούφιση του κοινού για την πολυαναμενόμενη λειτουργία του Μετρό. Συνολικά, η χρονική εξέλιξη των συναισθημάτων καταδεικνύει ότι το κοινό επηρεάζεται έντονα από εξωτερικά γεγονότα και ορόσημα του έργου, απαντώντας στο ερευνητικό ερώτημα με το εύρημα πως οι συναισθηματικές αντιδράσεις μεταβάλλονται δυναμικά στον χρόνο ανάλογα με τις εξελίξεις.

Το κυκλικό διάγραμμα παρουσιάζει την κατανομή όλων των συλλεχθέντων σχολίων ανά συναισθηματική κατηγορία (θετικό, ουδέτερο, αρνητικό). Διαπιστώνεται ότι ένα σημαντικό ποσοστό των σχολίων είναι ουδέτερα, γεγονός που υποδηλώνει πως μεγάλο μέρος της συζήτησης διεξάγεται σε πληροφοριακό ή πραγματολογικό επίπεδο χωρίς έντονη συναισθηματική φόρτιση. Παρ' όλα αυτά, τα αρνητικά σχόλια υπερέχουν αριθμητικά των θετικών, αντανακλώντας την τάση του κοινού να εκφράζει συχνότερα δυσαρέσκεια ή κριτική σχετικά με το Μετρό. Συγκεκριμένα, τα αρνητικά σχόλια αποτελούν τη μεγαλύτερη μερίδα (περίπου σχεδόν το ήμισυ των αξιολογημένων σχολίων), ενώ τα ουδέτερα ακολουθούν ως η δεύτερη πολυπληθέστερη κατηγορία (γύρω στο ένα τρίτο). Τα θετικά σχόλια συνιστούν τη μικρότερη κατηγορία (κάπου μεταξύ ενός πέμπτου και ενός τετάρτου του συνόλου). Η ανισομερής αυτή κατανομή καταδεικνύει ότι, παρόλο που υπάρχει ενθουσιασμός και αισιοδοξία από μερίδα του κοινού, επικρατεί μία γενικότερη στάση κριτικής ή επιφυλακτικότητας απέναντι στο έργο του Μετρό Θεσσαλονίκης.

Η ποιοτική εξέταση του περιεχομένου των σχολίων ανά κατηγορία συναισθήματος, μέσω των σύννεφων λέξεων (wordclouds), αποκαλύπτει διακριτές θεματικές που κυριαρχούν σε κάθε

συναισθηματικό τόνο. Κάθε κατηγορία συναισθήματος χαρακτηρίζεται από διαφορετικές συχνές λέξεις, που δίνουν μια εικόνα του πλαισίου και των κύριων θεμάτων που απασχολούν το κοινό όταν εκφράζεται θετικά, αρνητικά ή ουδέτερα για το Μετρό.

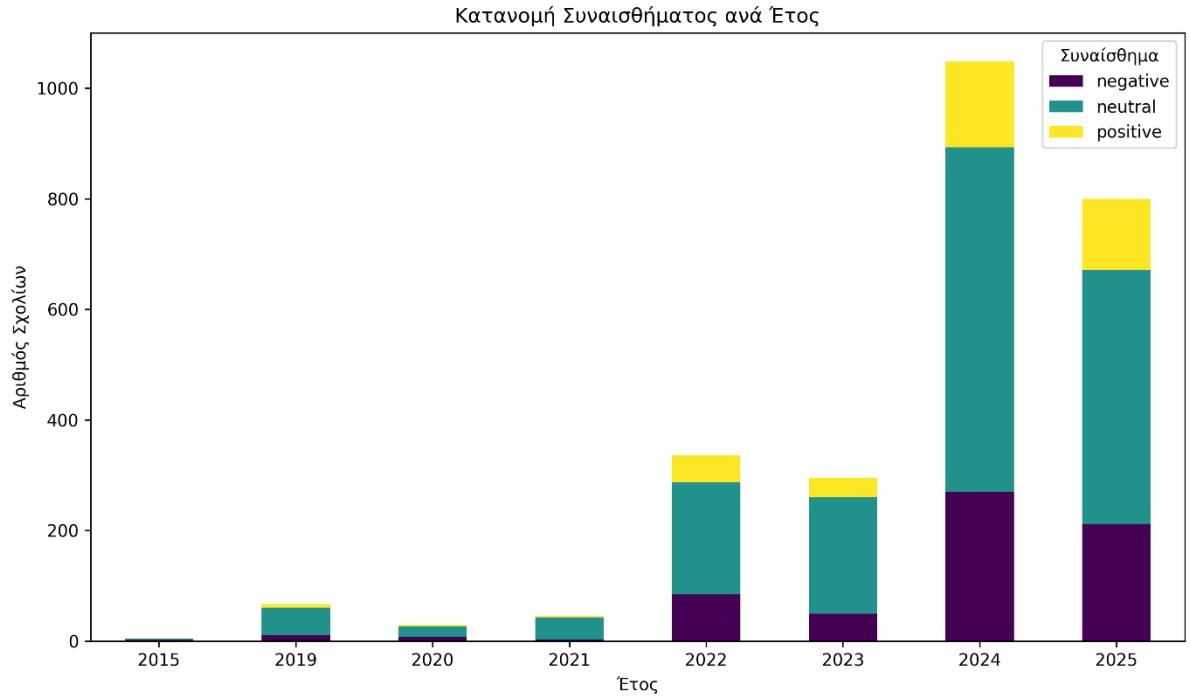
Τα αρνητικά συναισθήματα συνοδεύονται από λέξεις που υπογραμμίζουν προβλήματα και επικρίσεις. Όπως φαίνεται στο σύννεφο λέξεων των αρνητικών σχολίων, όροι όπως «πρόβλημα» και «θέμα» εμφανίζονται με μεγάλη συχνότητα. Αυτό υποδηλώνει ότι πολλοί σχολιαστές εστιάζουν σε ζητήματα και εμπόδια που αντιμετωπίζει το έργο, καθώς και σε προβληματικές πτυχές της υλοποίησής του. Συχνές είναι επίσης λέξεις που δηλώνουν αρνητική αξιολόγηση ή απογοήτευση, όπως εκφράσεις δυσαρέσκειας για τις καθυστερήσεις, το κόστος ή γενικά την πορεία του έργου. Η παρουσία αυτών των λέξεων φανερώνει ένα κοινό που εκφράζει έντονη κριτική στάση, αναδεικνύοντας τις ανησυχίες και τα παράπονα γύρω από το Μετρό Θεσσαλονίκης.

Τα ουδέτερα σχόλια χαρακτηρίζονται από μια περισσότερο πραγματολογική ή ερωτηματική διάθεση, χωρίς εμφανή θετικό ή αρνητικό προσανατολισμό. Στο σύννεφο λέξεων των ουδέτερων σχολίων εμφανίζονται συχνά γενικοί όροι και θεματικές σχετικές με το ίδιο το έργο, όπως ονόματα σταθμών, τεχνικές λεπτομέρειες ή συζητήσεις για το «έργο» και το «μέσο» (το Μετρό) ως αντικείμενο. Η λέξη «θέμα» ενδέχεται επίσης να εμφανίζεται, όχι απαραίτητα με αρνητική χροιά εδώ, αλλά ως αναφορά στη θεματολογία του Μετρό στις συνομιλίες. Συνολικά, τα ουδέτερα σχόλια φαίνεται να επικεντρώνονται στην ανταλλαγή πληροφοριών, ερωτήσεων ή επισημάνσεων σχετικά με την πρόοδο και τα χαρακτηριστικά του Μετρό, διατηρώντας έναν αντικειμενικό τόνο.

Τα θετικά σχόλια αναδεικνύουν μια εντελώς διαφορετική θεματολογία, επικεντρωμένη σε ελπίδες, προσδοκίες και εγκωμιαστικές αναφορές. Όπως δείχνει το σύννεφο λέξεων των θετικών σχολίων, λέξεις όπως «καλός» και «ελπίζω» ξεχωρίζουν σε συχνότητα. Αυτό φανερώνει ότι πολλοί πολίτες εκφράζονται με αισιοδοξία και ελπίδα — για παράδειγμα, χαρακτηρίζουν θετικά κάποια εξέλιξη ή δηλώνουν «ελπίζω» σε μια καλύτερη συγκοινωνιακή πραγματικότητα με την έναρξη του Μετρό. Επίσης, σε αυτή την κατηγορία παρατηρούνται λέξεις που υποδηλώνουν ικανοποίηση ή επαίνους για το έργο, πιθανώς αναγνωρίζοντας την αξία του Μετρό για την πόλη ή εκφράζοντας χαρά που επιτέλους ολοκληρώνεται. Οι θεματικές αυτές αντανακλούν ένα κοινό τμήμα που βλέπει το Μετρό Θεσσαλονίκης ως πηγή θετικής αλλαγής και δηλώνει την υποστήριξη ή την ικανοποίησή του.

Η διερεύνηση των χρονικών συγκεντρώσεων των συναισθημάτων φανερώνει ότι οι συναισθηματικές αντιδράσεις του κοινού δεν κατανέμονται ομοιόμορφα στον χρόνο, αλλά παρουσιάζουν «εξάρσεις» σε συγκεκριμένες περιόδους. Με άλλα λόγια, παρατηρούνται χρονικά διαστήματα όπου συσσωρεύεται μεγάλος αριθμός σχολίων με όμοιο συναισθηματικό πρόσημο, υποδηλώνοντας συλλογικές αντιδράσεις σε γεγονότα ή εξελίξεις. Χαρακτηριστικά, αμέσως μετά το δυστύχημα των Τεμπών (αρχές Μαρτίου 2023) παρατηρείται μια έντονη συγκέντρωση αρνητικών σχολίων μέσα σε σύντομο χρονικό διάστημα. Αυτή η «αιχμή» αρνητικότητας συμπίπτει με το κλίμα σοκ και οργής που κατέκλυσε την κοινή γνώμη, αντικατοπτρίζοντας το πώς ένα ευρύτερο τραγικό συμβάν μπορεί να επηρεάσει το συναισθηματικό χρώμα της συζήτησης για το Μετρό. Αντίστοιχα, στα τέλη του 2024, γύρω από την έναρξη λειτουργίας του Μετρό, εντοπίζεται μια πυκνή συγκέντρωση θετικών σχολίων. Το γεγονός αυτό υποδηλώνει ότι πολλοί χρήστες εξέφρασαν ταυτόχρονα τα θετικά τους συναισθήματα (ενθουσιασμό, ανακούφιση, ικανοποίηση) μόλις το έργο έγινε πραγματικότητα. Επιπλέον, μικρότερες τοπικές συγκεντρώσεις συναισθημάτων διακρίνονται και σε άλλα σημεία του χρόνου, πιθανώς σε αντιστοιχία με ενδιάμεσα γεγονότα (π.χ. ανακοινώσεις προόδου, εμπλοκές ή επιμέρους προβλήματα στο έργο) που προκάλεσαν πρόσκαιρες συναισθηματικές αντιδράσεις. Συνοψίζοντας, η συναισθηματική αντίδραση του κοινού στη θεματολογία του Μετρό Θεσσαλονίκης μεταβάλλεται σημαντικά μέσα στον χρόνο, ακολουθώντας τις

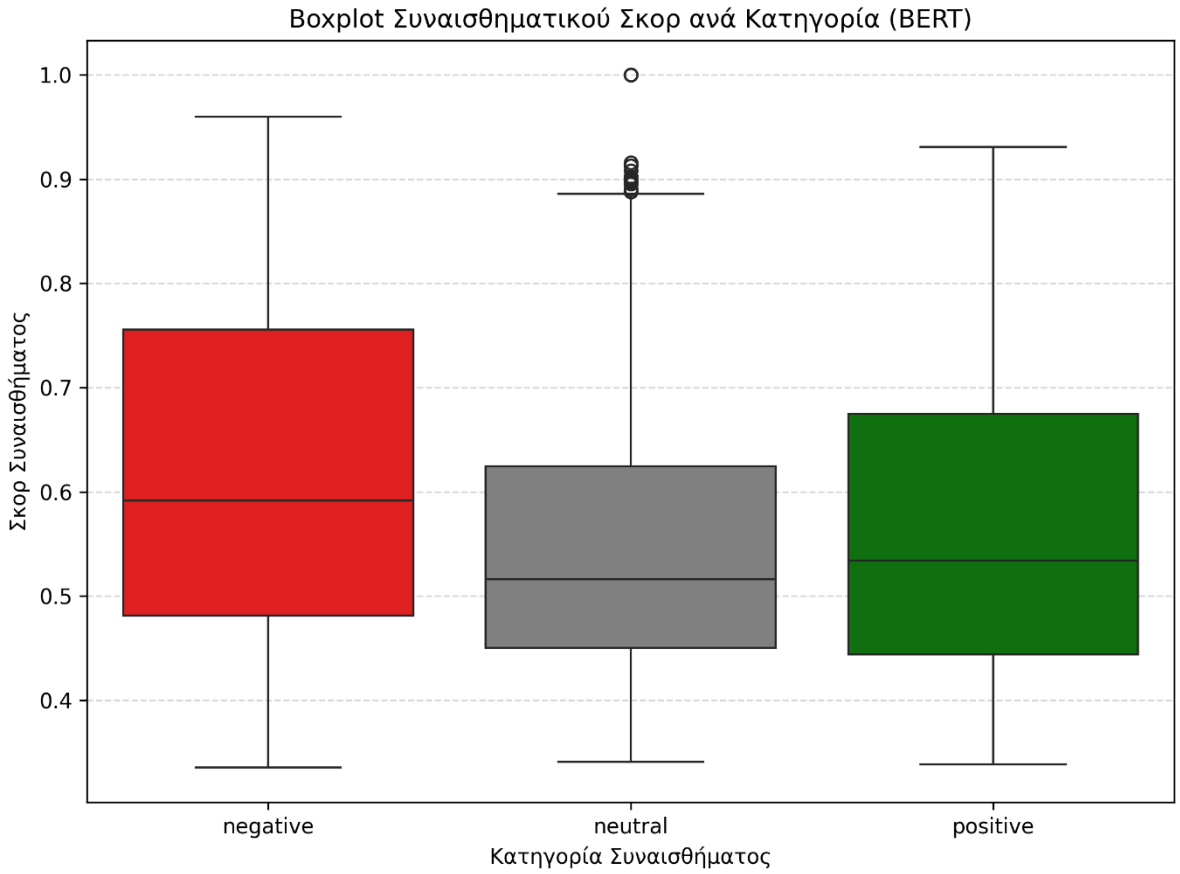
εξελίξεις και τα γεγονότα που περιβάλλουν το έργο. Η χρονική ανάλυση ανέδειξε περιόδους έντονου αρνητισμού ή θετικού κλίματος, ενώ η κατανομή και η θεματολογία των σχολίων φανερώνουν ότι κυριαρχεί μεν η κριτική στάση, αλλά δεν λείπουν η αισιοδοξία και οι ουδέτερες, πληροφοριακές συζητήσεις. Τα ευρήματα αυτά παρέχουν μια ολοκληρωμένη εικόνα που συμβάλλει στην απάντηση του ερευνητικού ερωτήματος, δείχνοντας ότι το κοινό αντιδρά συναισθηματικά με τρόπο δυναμικό και ευαίσθητο στα γεγονότα, καθώς το έργο του Μετρό εξελίσσεται και από ένα μακροχρόνιο σχέδιο περνά στην υλοποίηση και λειτουργία.



Εικόνα 4-7 Κατανομή συναισθήματος ανά έτος

Στην εικόνα 4-7, η κατανομή των συναισθημάτων (θετικών, αρνητικών και ουδέτερων) εμφανίζει σημαντικές διακυμάνσεις από έτος σε έτος βάσει των δεδομένων. Στο διάγραμμα κατανομής συναισθήματος ανά έτος παρατηρούνται αξιοσημείωτες μεταβολές στην αναλογία των θετικών και αρνητικών σχολίων, αντανακλώντας την επίδραση σημαντικών γεγονότων κάθε χρονιάς. Γενικά, τα ουδέτερα σχόλια αποτελούν ένα σταθερό υπόβαθρο σε όλες τις περιόδους, όμως το ποσοστό των θετικών και αρνητικών σχολίων μεταβάλλεται ποσοτικά και ποιοτικά ανάλογα με τις εξελίξεις της εκάστοτε χρονιάς. Κατά το έτος 2023 καταγράφεται απότομη άνοδος των αρνητικών σχολίων σε σχέση με τα προηγούμενα έτη. Ποσοτικά, το πλήθος των αρνητικά φορτισμένων αναρτήσεων αυξήθηκε σημαντικά, σηματοδοτώντας τη μεγαλύτερη έξαρση αρνητικού συναισθήματος στη χρονική σειρά των δεδομένων. Ποιοτικά, η κορύφωση αυτή συμπίπτει με το σιδηροδρομικό δυστύχημα των Τεμπών τον Φεβρουάριο του 2023, ένα τραγικό γεγονός που πυροδότησε κύμα κοινωνικής οργής και ευρείας δημόσιας αγανάκτησης.

Οι χρήστες εξέφρασαν έντονα αρνητικά συναισθήματα εκείνη τη χρονιά, γεγονός που αντανακλάται στην απότομη αύξηση των αρνητικών σχολίων. Η τραγωδία αυτή και οι επακόλουθες αντιδράσεις (π.χ. μαζικές διαδηλώσεις και απεργίες διαμαρτυρίας) φαίνεται να συνέβαλαν καθοριστικά στη διαμόρφωση του αρνητικού κλίματος, επισκιάζοντας άλλες θεματικές συζητήσεις το 2023. Σε αντιδιαστολή, το έτος 2024 παρουσιάζει αξιοσημείωτη άνοδο στα θετικά σχόλια. Ποσοτικά, τα θετικά σχόλια αυξήθηκαν σε υψηλότερα επίπεδα από ό,τι τα προηγούμενα έτη, υποδηλώνοντας στροφή του κοινού σε πιο αισιόδοξο τόνο. Αυτή η μεταβολή μπορεί να αποδοθεί στην αναμενόμενη έναρξη λειτουργίας του Μετρό Θεσσαλονίκης το 2024, ενός πολυαναμενόμενου γεγονότος που δημιούργησε κύμα ενθουσιασμού στην πόλη. Πράγματι, προς τα τέλη του 2024 επιβεβαιώθηκε ότι το πολυαναμενόμενο δίκτυο Μετρό θα ξεκινήσει επιτέλους τη λειτουργία του στις 30 Νοεμβρίου 2024, γεγονός που συνοδεύτηκε από ευρεία αισιοδοξία και θετικά σχόλια στα μέσα ενημέρωσης και κοινωνικής δικτύωσης. Ποιοτικά, το περιεχόμενο των θετικών αναρτήσεων του 2024 αντικατοπτρίζει την ικανοποίηση και την ελπίδα του κοινού για τη νέα αυτή υποδομή, σε αντίθεση με την δυσπιστία ή απαισιοδοξία προηγούμενων ετών λόγω καθυστερήσεων. Τα ουδέτερα σχόλια διατηρούν σημαντική παρουσία σε όλα τα έτη, αποτελώντας συχνά την μεγαλύτερη κατηγορία σε περιόδους χωρίς έντονες εξελίξεις. Σε χρόνια χωρίς μεγάλα γεγονότα (π.χ. 2019–2022), η κατανομή συναισθημάτων παραμένει πιο ισορροπημένη, με τα ουδέτερα σχόλια να υπερισχύουν ελαφρώς και τα θετικά/αρνητικά να παρουσιάζουν ηπιότερες διακυμάνσεις. Αυτό υποδηλώνει ότι σε ήρεμες περιόδους ο δημόσιος διάλογος γύρω από το υπό μελέτη θέμα ήταν λιγότερο πολωμένος συναισθηματικά. Ωστόσο, σε χρονιές με έκτακτα γεγονότα βλέπουμε τα ουδέτερα σχόλια να μειώνονται αναλογικά, καθώς αντικαθίστανται από κύματα θετικών ή αρνητικών αντιδράσεων. Συνολικά, η χρονική ανάλυση καταδεικνύει ότι οι εξάρσεις ενδιαφέροντος λόγω εξωγενών γεγονότων συνδέονται με απότομες μετατοπίσεις στο συναισθηματικό προφίλ των σχολίων. Οι χρονικές τάσεις φανερώνουν μια δυναμική μεταβολή: σε περιόδους κρίσης αυξάνονται δυσανάλογα τα αρνητικά συναισθήματα, ενώ σε περιόδους επίτευξης σημαντικών στόχων ή θετικών εξελίξεων ενισχύονται τα θετικά συναισθήματα. Αυτά τα μοτίβα αντικατοπτρίζουν το πώς οι εξελίξεις του πραγματικού κόσμου επηρεάζουν το συναισθηματικό τόνο της δημόσιας συζήτησης κάθε έτους, παρέχοντας πολύτιμη εικόνα για τη σχέση γεγονότων και δημοσίου συναισθήματος.



Εικόνα 4-8 Ανάλυση διασποράς συναισθηματικού σκορ ανά κατηγορία

Για τη βαθύτερη κατανόηση της συμπεριφοράς του ταξινομητή BERT ως προς την αξιολόγηση συναισθημάτων, δημιουργήθηκε ένα boxplot (διάγραμμα κουτιού) (Εικόνα 4-8) που απεικονίζει τη διασπορά των τιμών του sentiment score ανά κατηγορία συναισθήματος (θετικό, ουδέτερο, αρνητικό).

Το sentiment score αντιπροσωπεύει τη βεβαιότητα του μοντέλου σχετικά με την απόδοσή του σε κάθε κατηγορία. Όσο πιο κοντά στο 1, τόσο μεγαλύτερη η εμπιστοσύνη στην πρόβλεψη.

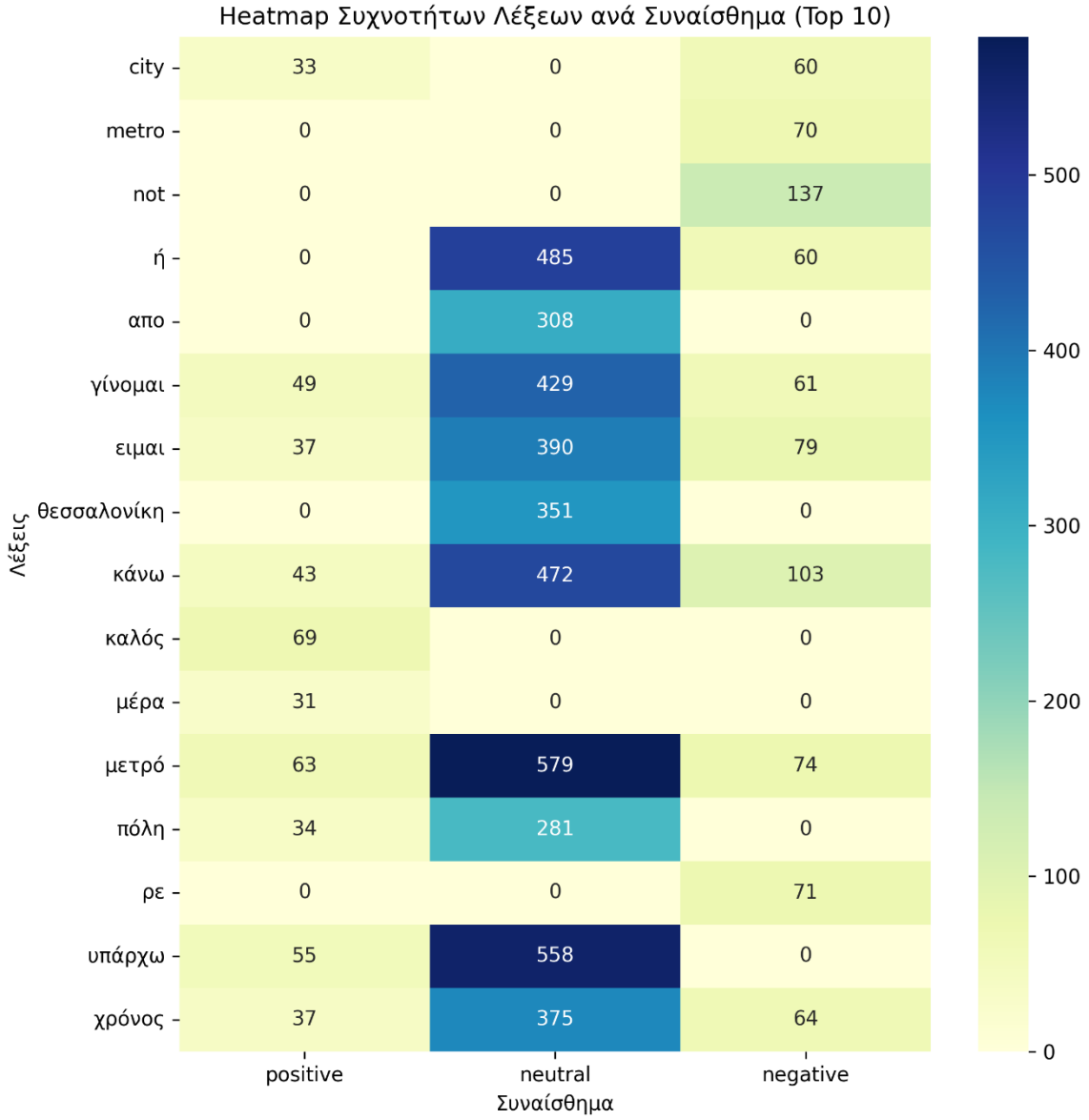
Η κατηγορία αρνητικού συναισθήματος (negative) παρουσιάζει τη μεγαλύτερη διασπορά τιμών, με διάμεσο γύρω στο 0.6 και τιμές που φτάνουν σχεδόν το 0.95. Αυτό υποδηλώνει ότι το μοντέλο ήταν συχνά αρκετά βέβαιο όταν εντόπιζε αρνητικό περιεχόμενο, όμως υπήρχαν και αρκετές προβλέψεις με χαμηλότερη εμπιστοσύνη. Η ευρεία διασπορά μπορεί να οφείλεται στην ποικιλία τρόπων με τους οποίους εκφράζεται η δυσαρέσκεια ή η κριτική.

Η κατηγορία ουδέτερου συναισθήματος (neutral) έχει μικρότερη διασπορά και την πιο "κεντραρισμένη" κατανομή γύρω από διάμεσο ~0.52. Παρότι δεν υπάρχουν ακραίες αποκλίσεις, ο αριθμός outliers είναι αυξημένος, κάτι που μπορεί να υποδηλώνει αβεβαιότητα του μοντέλου όταν πρόκειται να διακρίνει ουδέτερα από ήπια θετικά ή αρνητικά σχόλια. Αυτό είναι σύνηθες στην ανάλυση συναισθήματος, καθώς πολλές ουδέτερες αναφορές εμπεριέχουν αποχρώσεις.

Η κατηγορία θετικού συναισθήματος (positive) εμφανίζει επίσης αρκετά μεγάλη διασπορά, αλλά λιγότερους outliers. Η διάμεσος βρίσκεται περίπου στο 0.54, ένδειξη ότι το μοντέλο δεν ήταν συστηματικά πιο "βέβαιο" για τις θετικές του προβλέψεις. Οι υψηλότερες τιμές πλησιάζουν το 0.9,

επιβεβαιώνοντας πως όταν εντόπιζε εμφανώς θετικό περιεχόμενο (π.χ. επευφημίες για το μετρό), το εκλάμβανε με σαφήνεια.

Συνολικά, το boxplot αναδεικνύει ότι η μεγαλύτερη βεβαιότητα προβλέψεων παρατηρείται κυρίως στις αρνητικές και θετικές κατηγορίες, ενώ το ουδέτερο συναίσθημα αποτελεί το μεγαλύτερο πρόβλημα αποσαφήνισης για το μοντέλο. Η παρουσία αρκετών outliers, ιδίως στα ουδέτερα, καταδεικνύει την ανάγκη περαιτέρω διερεύνησης για τη βελτίωση της διαφοροποίησης μεταξύ συναισθηματικών αποχρώσεων.



Εικόνα 4-9 Ανάλυση Συχνοτήτων Λέξεων ανά Κατηγορία Συναισθήματος (Heatmap)

Προκειμένου να αναδειχθούν οι πιο χαρακτηριστικές λέξεις που σχετίζονται με κάθε κατηγορία συναισθήματος, δημιουργήθηκε ένας πίνακας θερμότητας (heatmap) που απεικονίζει τις 10 συχνότερες λέξεις για τα θετικά, ουδέτερα και αρνητικά σχόλια (Εικόνα 4-9). Ο πίνακας βασίζεται σε τακτοποιημένο και λεμματοποιημένο κείμενο (processed_text), και αποτυπώνει τη συχνότητα εμφάνισης κάθε λέξης εντός της αντίστοιχης κατηγορίας συναισθήματος.

Οι λέξεις με τη μεγαλύτερη συχνότητα στην ουδέτερη κατηγορία (neutral) είναι λέξεις πληροφορίας ή γενικού σχολιασμού, όπως: «ή» (485 εμφανίσεις), «κάνω», «είμαι», «γίνομαι», «υπάρχω», «μετρό», «χρόνος». Αυτό επιβεβαιώνει ότι πολλά σχόλια είχαν κυρίως περιγραφικό χαρακτήρα, χωρίς σαφή θετική ή αρνητική συναισθηματική φόρτιση.

Η αρνητική κατηγορία περιλαμβάνει λέξεις όπως «not», «ρε», «χρόνος» με το «not» να εμφανίζεται 137 φορές – ένδειξη πως πολλά αρνητικά σχόλια βασίζονται σε άρνηση, απογοήτευση ή σύγκριση. Το γεγονός ότι εμφανίζονται και αγγλικές λέξεις (π.χ. «metro», «city») αντανακλά το ότι μέρος του σχολιασμού έγινε σε αγγλόφωνο περιβάλλον ή περιλαμβάνει ορολογία από διεθνές κοινό ή memes.

Στην θετική κατηγορία, η πιο χαρακτηριστική λέξη είναι το «καλός» (69 εμφανίσεις), ακολουθούμενη από θετικά ουδέτερες λέξεις όπως «υπάρχω», «μετρό», «χρόνος», «κάνω», «γίνομαι». Αυτό δείχνει ότι τα θετικά σχόλια επικεντρώνονταν κυρίως στην ύπαρξη και λειτουργία του μετρό, με λέξεις που υποδηλώνουν πρόοδο, εξέλιξη και βελτίωση.

Παρατηρείται επίσης ότι ορισμένες λέξεις όπως «μετρό», «υπάρχω» και «χρόνος» εμφανίζονται σε όλες τις κατηγορίες, γεγονός που υποδεικνύει ότι η σημασία τους μεταβάλλεται ανάλογα με το πλαίσιο. Για παράδειγμα, η λέξη «χρόνος» μπορεί να αναφέρεται είτε σε θετικά (π.χ. "επιτέλους ήρθε η ώρα"), είτε σε αρνητικά (π.χ. "χάθηκε πολύς χρόνος") είτε ουδέτερα (π.χ. "χρόνος λειτουργίας").

Η θερμική χαρτογράφηση των λέξεων αναδεικνύει τη θεματική εστίαση του κοινού και τη γλωσσική διαφοροποίηση μεταξύ των συναισθηματικών κατηγοριών, ενώ παράλληλα προσφέρει ένα εργαλείο ερμηνείας του περιεχομένου πέρα από τους αριθμούς των ταξινομήσεων. Τέτοια visualizations ενισχύουν τη δυνατότητα κατανόησης των αποτελεσμάτων και επιτρέπουν την εντοπισμό μοτίβων στη γλώσσα που χρησιμοποιείται για την έκφραση διαφορετικών συναισθημάτων.

Συνοψίζοντας, η χρονική ανάλυση αποκαλύπτει ότι το συναίσθημα των χρηστών ακολούθησε τις πραγματικές εξελίξεις και γεγονότα γύρω από το Μετρό Θεσσαλονίκης. Σε περιόδους στασιμότητας ή αρνητικών ειδήσεων (π.χ. μεγάλες καθυστερήσεις, δυστύχημα Τεμπών) υπερίσχυσε το αρνητικό συναίσθημα, ενώ σε περιόδους προόδου ή θετικών γεγονότων (π.χ. δοκιμές, εγκαίνια) το συναίσθημα μετατοπίστηκε προς το ουδέτερο/θετικό. Το μοντέλο BERT απέδειξε την αξία του, καθώς μπόρεσε να συλλάβει αυτές τις μεταβολές πιο αξιόπιστα από το απλό λεξικό – αναγνωρίζοντας πότε τα σχόλια έκρυβαν δυσαρέσκεια ακόμα και χωρίς λέξεις-κλειδιά. Τα ευρήματα αυτά δίνουν μια ποσοτική επιβεβαίωση της κοινής αίσθησης: ότι η πολύχρονη αναμονή για το μετρό συνοδεύτηκε από κύματα απογοήτευσης, τα οποία κορυφώθηκαν σε κομβικά σημεία, αλλά τελικά μετατράπηκαν σε ανακούφιση και ικανοποίηση όταν το έργο ολοκληρώθηκε. Η ανάλυση συναισθήματος σε πραγματικό χρόνο, συνδυασμένη με τη γνώση των γεγονότων, μας επιτρέπει να χαρτογραφήσουμε πώς η δημόσια γνώμη διαμορφώθηκε κατά τη διάρκεια αυτού του μεγάλου έργου υποδομής.

Κεφάλαιο 5ο: Συζήτηση

5.1 Ερμηνεία αποτελεσμάτων

Ποσοστιαία κατανομή συναισθήματος των σχολίων (μοντέλο BERT). Τα αποτελέσματα της ανάλυσης συναισθήματος ανέδειξαν κυρίως ουδέτερο τόνο στα σχόλια σχετικά με το Μετρό Θεσσαλονίκης. Όπως φαίνεται, περίπου 62% των σχολίων χαρακτηρίστηκαν ως ουδέτερα, ενώ μόνο το 13% ήταν θετικά και το 25% αρνητικά. Η κυριαρχία των ουδέτερων σχολίων υποδηλώνει ότι μεγάλος όγκος της διαδικτυακής συζήτησης δεν εκφράζει έντονη συναισθηματική φόρτιση. Αυτό μπορεί να οφείλεται στο ότι πολλοί χρήστες μοιράζονται πληροφορίες, κάνουν ερωτήσεις ή συζητούν πραγματολογικά ζητήματα για το έργο, χωρίς να παίρνουν ξεκάθαρα θετική ή αρνητική στάση. Επιπλέον, ενδέχεται η μεθοδολογία ταξινόμησης να τείνει να χαρακτηρίζει ως ουδέτερες τις τοποθετήσεις που δεν περιέχουν σαφείς λέξεις-δείκτες συναισθήματος, με αποτέλεσμα ένα υψηλό ποσοστό "ουδέτερων" εντοπίσεων. Παράλληλα, το σημαντικό ποσοστό αρνητικών σχολίων (1 στα 4) φανερώνει αισθητή δυσαρέσκεια και κριτική από μερίδα του κοινού. Το εύρημα αυτό ευθυγραμμίζεται με τη γνωστή πραγματικότητα ενός έργου που έχει υποστεί επανειλημμένες καθυστερήσεις και ανατροπές, προκαλώντας απογοήτευση στο κοινό. Αντιθέτως, τα θετικά σχόλια (περίπου 1 στα 8) εκπροσωπούν την αισιοδοξία ή υποστήριξη ορισμένων χρηστών π.χ. εκείνων που εστιάζουν στα οφέλη του μετρό ή εκφράζουν χαρά για την πρόοδο του έργου. Η μικρότερη όμως αυτή αναλογία θετικών αντιδράσεων υποδηλώνει ότι ο ενθουσιασμός ήταν συγκρατημένος, πιθανώς εξαιτίας του βεβαρημένου ιστορικού καθυστερήσεων και δυσπιστίας γύρω από το έργο.

Πέρα από τη συνολική κατανομή, παρατηρήθηκαν έντονες αυξομειώσεις στο συναίσθημα των χρηστών σε συνάρτηση με συγκεκριμένα γεγονότα. Ενδεικτικά, μετά το πολύνεκρο δυστύχημα στα Τέμπη (28/2/2023) σημειώθηκε απότομη στροφή των σχολίων σε περισσότερο αρνητικό ύφος, αντικατοπτρίζοντας το σοκ και την οργή του κοινού για τις ελλείψεις στις υποδομές και την ασφάλεια. Πράγματι, σε σχετική έρευνα καταγράφηκε ότι μετά το δυστύχημα αυτό εντάθηκε η κρίση εμπιστοσύνης προς τους θεσμούς, με την κυβέρνηση να συγκεντρώνει 75,4% ποσοστό δυσπιστίας[17], ενώ παράλληλα εκλύθηκε «κύμα έντονων αρνητικών συναισθημάτων» ιδιαίτερα μεταξύ των νέων[17]. Τα ευρήματά μας συμβαδίζουν με αυτό το κλίμα, καθώς οι χρήστες του Reddit εξέφρασαν εντονότερη αρνητικότητα τις ημέρες και εβδομάδες που ακολούθησαν το τραγικό γεγονός. Αντιστρόφως, κατά την περίοδο των εγκαινίων ή ανακοινώσεων για το μετρό (π.χ. επίσημες παρουσιάσεις του έργου), παρατηρήθηκε σχετική άνοδος του θετικού ή αισιόδοξου τόνου στα σχόλια. Οι στιγμές αυτές, όπου το κοινό έβλεπε απτές αποδείξεις προόδου (δοκιμαστικά δρομολόγια, εκδηλώσεις εγκαινίων κ.λπ.), συνοδεύτηκαν από έκφραση ικανοποίησης ή ελπίδας ότι το πολυαναμενόμενο έργο πλησιάζει στην ολοκλήρωση[1]. Ωστόσο, ακόμα και τότε, η αισιοδοξία συχνά μετριαζόταν από σχόλια δύσπιστα ή σαρκαστικά, ενδεικτικό μιας παρατεταμένης κόπωσης και επιφυλακτικότητας του κοινού. Συνολικά, το συναίσθημα στα μέσα κοινωνικής δικτύωσης αντικατοπτρίζει το ευρύτερο κοινωνικοπολιτικό πλαίσιο: σε περιόδους κρίσης και χαμηλής εμπιστοσύνης προς τις αρχές παρατηρείται έξαρση αρνητικότητας, ενώ σε στιγμές επιτυχίας ή προόδου διαφαίνεται συγκρατημένη αισιοδοξία.

5.2 Σύγκριση με Παρόμοιες Μελέτες

Τα παραπάνω ευρήματα μπορούν να συγκριθούν με αποτελέσματα αντίστοιχων μελετών ανάλυσης συναισθήματος σε κοινωνικά δίκτυα για μεγάλα έργα υποδομής και δημόσιες αντιδράσεις. Σε διεθνές επίπεδο, έχει παρατηρηθεί ότι τα αμφιλεγόμενα ή καθυστερημένα έργα συχνά αντιμετωπίζονται με αρνητική διάθεση από το διαδικτυακό κοινό. Για παράδειγμα, σε μελέτη του βρετανικού σιδηροδρομικού έργου HS2 (High Speed 2) βρέθηκε ότι πάνω από το ήμισυ των αναρτήσεων (53–63%) εξέφραζαν αρνητικό συναίσθημα, υποδηλώνοντας μια γενικά αρνητική αντίληψη του κοινού για το έργου. Η επικράτηση της δυσαρέσκειας στο παράδειγμα αυτό ευθυγραμμίζεται με το δικό μας εύρημα ότι η αρνητική χροιά υπερτερεί της θετικής στα σχόλια (αν και στη μελέτη μας πολλά σχόλια κρίθηκαν ουδέτερα). Παρομοίως, στην περίπτωση του αμερικανικού έργου του Καλιφορνέζικου Σιδηροδρόμου Υψηλής Ταχύτητας, έχει αναφερθεί ότι το έργο συγκέντρωσε πληθώρα διαδικτυακών επικρίσεων και παραπόνων, καταδεικνύοντας ότι τα κοινά συχνά χρησιμοποιούν τα κοινωνικά μέσα για να εκφράσουν απογοήτευση σε μεγάλα έργα που αντιμετωπίζουν προβλήματα. Από την άλλη πλευρά, δεν εμφανίζουν όλες οι μελέτες τόσο αρνητική εικόνα: σε ανάλυση σχολίων σχετικά με προγράμματα αναβάθμισης πόλεων στην Κίνα, διαπιστώθηκε ότι το συνολικό πρόσημο του δημόσιου συναισθήματος ήταν γενικά θετικό, αν και παρουσίαζε έντονες διακυμάνσεις ανά χρονική περίοδο και περιοχή. Συγκεκριμένα, η εν λόγω μελέτη ανέφερε ότι το 2020 το συναίσθημα ήταν πιο αρνητικό (λόγω αρχικών προβλημάτων), ενώ το 2021 έγινε το πιο θετικό, για να υποχωρήσει ξανά σε αρνητικό το 2022 μετά από νέες προκλήσεις. Αυτή η περίπτωση αναδεικνύει πως όταν ένα έργο εξελίσσεται ομαλά και αποφέρει οφέλη, το κοινό μπορεί να εκφράζεται θετικά, ενώ αντίθετα προβλήματα ή κρίσεις οδηγούν σε κύματα αρνητισμού. Συγκριτικά, η μελέτη μας για το Μετρό Θεσσαλονίκης παρουσιάζει μια μικτή εικόνα: κυριαρχία ουδέτερου τόνου με αξιοσημείωτη υποβόσκουσα αρνητικότητα, γεγονός που συνάδει με τη μακρόχρονη ταλαιπωρία του συγκεκριμένου έργου, αλλά και κάποιες εκλάμψεις θετικής αντίδρασης σε στιγμές προόδου. Γενικότερα, η "μεροληψία προς το αρνητικό" που καταγράφεται σε πολλά διαδικτυακά fora φαίνεται παρούσα και εδώ – το κοινό τείνει να εκφράζει ευκολότερα δυσαρέσκεια ή θυμό για τα κακώς κείμενα, ενώ η θετική έκφραση (π.χ. επαίνων) είναι πιο συγκρατημένη. Αυτό είναι σύμφωνο με την παρατήρηση ότι τα αρνητικά νέα και σχόλια συχνά κυριαρχούν στην online αλληλεπίδραση, επηρεάζοντας τον συνολικό τόνο της δημόσιας συζήτησης.

5.3 Περιορισμοί της Μελέτης και Προτάσεις Βελτίωσης

Παρά τη σημασία των παραπάνω αποτελεσμάτων, η μεθοδολογία και το πεδίο της έρευνας παρουσιάζουν ορισμένους περιορισμούς. Αναγνωρίζοντας αυτούς τους περιορισμούς, προτείνουμε και τρόπους βελτίωσης για μελλοντικές μελέτες:

- Περιορισμένη πηγή δεδομένων (μόνο Reddit): Η ανάλυση βασίστηκε αποκλειστικά σε σχόλια από το Reddit, ένα μέσο με συγκεκριμένα δημογραφικά χαρακτηριστικά και κουλτούρα. Οι χρήστες του Reddit τείνουν να είναι νεότεροι σε ηλικία και κατά πλειονότητα άνδρες, γεγονός που μπορεί να συνεπάγεται μεροληψία στις απόψεις που συλλέγονται (π.χ. πιο έντονα κριτικός ή σαρκαστικός τόνος, τεχνοκρατική οπτική κ.ά.). Επιπλέον, παραλείποντας άλλες πλατφόρμες όπως το Twitter ή το Facebook, ενδέχεται να χάνονται διαφορετικές οπτικές γωνίες και αντιδράσεις άλλων τμημάτων του πληθυσμού. Πρόταση βελτίωσης: Σε επόμενες μελέτες θα ήταν ωφέλιμη η ενσωμάτωση δεδομένων από πολλαπλές πλατφόρμες κοινωνικής δικτύωσης. Μια συγκριτική ανάλυση μεταξύ Reddit και άλλων μέσων θα μπορούσε να αναδείξει τυχόν διαφοροποιήσεις στο ύφος και στο συναίσθημα (π.χ. πιο άμεσες αντιδράσεις στο Twitter, πιο προσωπικές ιστορίες στο Facebook). Έτσι θα αυξηθεί η αντιπροσωπευτικότητα των αποτελεσμάτων και θα μειωθεί η πλατφορμο-κεντρική μεροληψία.

• Περιορισμένη κάλυψη λεξιλογίου στη μεθοδολογία "λεξικού": Ένα μέρος της ανάλυσης συναισθήματος βασίστηκε σε προκαθορισμένα λεξικά συναισθηματικής πολικότητας (rule-based προσέγγιση). Αυτή η μέθοδος, παρότι απλή και διαφανής, έχει το μειονέκτημα ότι δεν αναγνωρίζει επαρκώς τις γλωσσικές αποχρώσεις, ιδιωματισμούς ή νέα slang που χρησιμοποιούν οι χρήστες. Σχόλια που εκφράζουν συναίσθημα με μη κυριολεκτικό τρόπο – π.χ. σαρκασμό, ειρωνεία ή μέσω εικόνων/emoji – ενδέχεται να μην ανιχνεύονται σωστά. Αυτό μπορεί να οδήγησε σε υπερβολική ταξινόμηση σχολίων ως ουδέτερων ή και σε μερικές λανθασμένες ταξινομήσεις. Πρόταση βελτίωσης: Η επέκταση και προσαρμογή του λεξικού για την κάλυψη καθομιλούμενων φράσεων και εξελισσόμενης αργκό θα βελτίωνε την ακρίβεια. Επίσης, η συνδυαστική χρήση λεξικολογικής και μηχανικής μάθησης (υβριδικά μοντέλα) θα μπορούσε να αξιοποιήσει τα πλεονεκτήματα και των δύο: το λεξικό για σταθερές λέξεις-κλειδιά και το μοντέλο για κατανόηση πλαισίου. Με αυτόν τον τρόπο θα μειωθούν φαινόμενα αστοχίας σε περιπτώσεις όπου το λεξικό αποτυγχάνει να "πιάσει" τις γλωσσικές λεπτές αποχρώσεις.

• Πιθανά σφάλματα ή μεροληψία του μοντέλου BERT: Το σύγχρονο μοντέλο BERT που χρησιμοποιήθηκε για την ταξινόμηση προσφέρει σαφώς βελτιωμένη ικανότητα κατανόησης συμφραζομένων έναντι των απλούστερων μεθόδων. Ωστόσο, δεν παύει να έχει περιορισμούς. Αφενός, ένα προεκπαιδευμένο μοντέλο μπορεί να φέρει μεροληψίες από τα δεδομένα εκπαίδευσής του, που ίσως επηρεάζουν τις εκτιμήσεις (π.χ. να χαρακτηρίζει δυσανάλογα συχνά κάποιες εκφράσεις ως αρνητικές λόγω προτύπων στα δεδομένα εκπαίδευσης). Αφετέρου, το BERT ίσως να μην αποδίδει άριστα σε ιδιαίτερες γλωσσικές συνθήκες που δεν "είδε" επαρκώς κατά την εκπαίδευσή του – όπως π.χ. σε μείξεις ελληνικών και αγγλικών (Greeklish), σε ορθογραφικά λάθη ή σε εξαιρετικά ιδιωματικό λόγο στο Reddit. Επιπλέον, το μοντέλο δυσκολεύεται όπως και κάθε σύστημα να αντιληφθεί ειρωνικό ή σαρκαστικό ύφος πέρα από κυριολεκτικές ενδείξεις. Πρόταση βελτίωσης: Θα ήταν χρήσιμη η περαιτέρω εξειδίκευση (fine-tuning) του BERT σε ένα σύνολο δεδομένων που να προσομοιάζει το περιβάλλον των ελληνικών social media (με έμφαση στη γλώσσα του Reddit), ώστε να μάθει καλύτερα τις ιδιαιτερότητές του. Εναλλακτικά ή συμπληρωματικά, η χρήση ενός άλλου μοντέλου προεκπαιδευμένου ειδικά στα ελληνικά (ή και η σύγκριση μεταξύ πολλαπλών μοντέλων) θα μπορούσε να αποκαλύψει τυχόν ασυνέπειες και να μειώσει τη μεροληψία μέσω τεχνικών ενισχυτικής μάθησης ή βαθμονόμησης των αποτελεσμάτων.

• Έλλειψη δημογραφικών/γεωγραφικών πληροφοριών: Η ανάλυση μας δεν διέθετε δεδομένα για την ταυτότητα ή προέλευση των χρηστών που σχολίασαν (ηλικία, φύλο, τόπος διαμονής κ.λπ.). Αυτό σημαίνει ότι δεν μπορούμε να γνωρίζουμε αν, για παράδειγμα, οι κάτοικοι της Θεσσαλονίκης ήταν πιο θετικά διακείμενοι έναντι του μετρό συγκριτικά με άτομα από άλλες περιοχές, ή αν οι νεότεροι χρήστες ήταν πιο επικριτικοί από τους μεγαλύτερους. Η απουσία τέτοιων πληροφοριών περιορίζει την ερμηνεία των αποτελεσμάτων, καθώς το συναίσθημα μπορεί να διαφέρει σημαντικά μεταξύ διαφορετικών ομάδων πληθυσμού (γεγονός που δεν διερευνήθηκε εδώ). Πρόταση βελτίωσης: Μελλοντικές έρευνες θα μπορούσαν να επιχειρήσουν τη συσχέτιση του συναισθήματος με δημογραφικούς παράγοντες. Αυτό θα μπορούσε να γίνει είτε μέσω συλλογής δεδομένων από πλατφόρμες όπου οι χρήστες παρέχουν περισσότερες πληροφορίες (π.χ. δημοσκοπήσεις σε δείγμα πολιτών ή ανάλυση σχολίων στο Facebook όπου εμφανίζεται το προφίλ), είτε με τεχνικές αυτόματης δημογραφικής ανίχνευσης στα social data (π.χ. ανάλυση γλώσσας για εκτίμηση ηλικιακής ομάδας). Μια τέτοια εμπλουτισμένη προσέγγιση θα επέτρεπε να κατανοήσουμε ποιοι υποστηρίζουν ή αντιτίθενται περισσότερο στο έργο και γιατί.

• Υπο-αντιπροσώπευση πολιτών εκτός διαδικτύου: Τέλος, πρέπει να τονιστεί ότι η μελέτη αφορά μόνο όσους επέλεξαν να εκφραστούν δημόσια στο διαδίκτυο. Ένα σημαντικό μέρος του πληθυσμού (ιδίως μεγαλύτερης ηλικίας ή όσοι δεν χρησιμοποιούν ενεργά τα κοινωνικά μέσα) δεν εκπροσωπείται στα δεδομένα μας. Αυτό ενδέχεται να μετατοπίζει την εικόνα του δημόσιου αισθήματος – π.χ. αν οι πιο

επικριτικές φωνές είναι υπερ-παρούσες online, ενώ οι πιο ικανοποιημένοι πολίτες σιωπούν ή εκφράζονται μόνο σε ιδιωτικές συζητήσεις. Πράγματι, είναι γνωστό ότι τα δεδομένα από τα social media δεν είναι πλήρως αντιπροσωπευτικά του γενικού πληθυσμού λόγω διαφόρων μεροληψιών στη σύνθεση των χρηστών. Πρόταση βελτίωσης: Τα αποτελέσματα της παρούσας ανάλυσης θα πρέπει να ερμηνεύονται με προσοχή και να συμπληρώνονται από άλλες πηγές. Μια λύση είναι η σύγκριση με παραδοσιακές δημοσκοπήσεις ή έρευνες κοινής γνώμης για το μετρό, ώστε να επαληθευτεί αν το κλίμα που ανιχνεύθηκε στο Reddit αντανακλάται και στο ευρύτερο κοινό. Επιπλέον, η επέκταση της μελέτης σε άλλες διαδικτυακές κοινότητες ή φόρουμ (π.χ. τοπικές ομάδες στο Facebook, σχόλια σε ειδησεογραφικά sites) θα μπορούσε να δώσει μια πληρέστερη εικόνα, συνδυάζοντας τη φωνή των πιο "ψηφιακά ενεργών" πολιτών με εκείνη πιο περιστασιακών χρηστών.

Σε σύνοψη, η ανάλυση συναισθήματος των σχολίων για το Μετρό Θεσσαλονίκης ανέδειξε πολύτιμες πληροφορίες σχετικά με τη διάθεση του κοινού και την επίδραση εξωγενών γεγονότων σε αυτήν. Παρά τους προαναφερθέντες περιορισμούς, τα συμπεράσματα συμφωνούν με το γενικότερο πλαίσιο: το παρατεταμένο "έπος" του μετρό έχει δημιουργήσει κόπωση και κριτική διάθεση σε μεγάλο μέρος του κοινού, χωρίς όμως να λείπουν εντελώς και οι φωνές αισιοδοξίας. Βελτιώνοντας τις μεθόδους ανάλυσης και διευρύνοντας το εύρος των δεδομένων στο μέλλον, θα μπορέσουμε να αποτυπώσουμε ακόμη ακριβέστερα το δημόσιο αίσθημα και να κατανοήσουμε βαθύτερα πώς οι πολίτες αντιδρούν στα μεγάλα έργα υποδομής που τους επηρεάζουν.

5.4 Προοπτικές επέκτασης

Στο πλαίσιο διερεύνησης της γενικευσιμότητας της μεθοδολογίας και σε άλλες πλατφόρμες κοινωνικής δικτύωσης, εξετάστηκε θεωρητικά η επέκταση της συλλογής και ανάλυσης δεδομένων σε υπηρεσίες όπως το Telegram και το Mastodon. Αυτές οι πλατφόρμες έχουν αποκτήσει δημοφιλία στην Ελλάδα τα τελευταία χρόνια, ιδιαίτερα σε συγκεκριμένες κοινότητες, και θα μπορούσαν να προσφέρουν συμπληρωματική οπτική για το υπό μελέτη θέμα. Ωστόσο, παρουσιάζουν ορισμένες τεχνικές και πρακτικές προκλήσεις.

5.4.1 Telegram

Το Telegram είναι μια εφαρμογή ανταλλαγής μηνυμάτων που υποστηρίζει δημόσια κανάλια και ομαδικές συνομιλίες με μεγάλο αριθμό μελών, λειτουργώντας εν μέρει και ως μέσο διάδοσης ειδήσεων. Για την πρόσβαση σε δεδομένα Telegram, υπάρχει διαθέσιμο επίσημο API και αρκετές βιβλιοθήκες Python (π.χ. python-telegram-bot, Telethon). Στην παρούσα μελέτη, έγινε πειραματική χρήση της βιβλιοθήκης Telethon, η οποία προσφέρει ασύγχρονη πρόσβαση στο API του Telegram και δυνατότητα εξαγωγής μηνυμάτων από δημόσια κανάλια[9][6]. Αξίζει να σημειωθεί ότι για να χρησιμοποιηθεί το Telegram API απαιτείται πρώτα η εγγραφή μιας εφαρμογής και η απόκτηση διαπιστευτηρίων (API ID και API Hash) συνδεδεμένων με έναν λογαριασμό Telegram[9][6]. Με αυτά, το script μπορεί να συνδεθεί ως πελάτης και να ανακτήσει μηνύματα. Ως δοκιμή, εντοπίστηκαν μερικά ελληνικά κανάλια ενημέρωσης και ομάδες συζήτησης σχετικά με τη Θεσσαλονίκη (για παράδειγμα ένα τοπικό κανάλι ειδήσεων στο Telegram) όπου αναμενόταν να υπάρχουν αναφορές στο μετρό. Χρησιμοποιώντας το Telethon, καταφέραμε να συλλέξουμε τα τελευταία ~1.000 μηνύματα από ένα τέτοιο κανάλι. Τα δεδομένα αυτά ήταν διαθέσιμα σε μορφή Python objects (τα οποία μετατρέψαμε σε JSON με πεδία: content, date, sender, etc.).

Παρά τη σχετική ευκολία συλλογής (το API του Telegram επιτρέπει δωρεάν και απεριόριστη πρόσβαση στα δημόσια δεδομένα με σωστή χρήση[6]), αναδείχθηκαν προκλήσεις: (α) Εντοπισμός πηγών: Το

Telegram δεν έχει ευρετήριο θεματικών κοινοτήτων όπως τα subreddits. Έτσι, απαιτείται να γνωρίζει κανείς εκ των προτέρων ποια κανάλια ή ομάδες συζητούν το θέμα. Η αναζήτηση γίνεται συχνά χειροκίνητα (π.χ. μέσω του ίδιου του app με keywords) και δεν υπάρχει εγγύηση ότι όλες οι συζητήσεις θα βρεθούν. (β) Δομή δεδομένων: Τα μηνύματα στο Telegram δεν έχουν τίτλο ή ξεχωριστό νήμα σχολίων – είναι απλώς μια ροή. Αυτό σημαίνει ότι η θεματολογική απομόνωση είναι δυσκολότερη. Σε ένα κανάλι ειδήσεων, π.χ., μπορεί να υπάρχουν πολλά θέματα ανακατεμένα. Για να φιλτραριστεί το θέμα "Μετρό Θεσσαλονίκης", θα έπρεπε να εφαρμόσουμε λέξεις-κλειδιά στο περιεχόμενο των μηνυμάτων και να εξάγουμε μόνο όσα αναφέρουν σχετικά keywords (παρόμοια με τη λέξη-κλειδί προσέγγιση που κάναμε στο Reddit). Πράγματι, φιλτράροντας τα συλλεχθέντα μηνύματα με τη λέξη "μετρό", βρέθηκαν ορισμένες αναφορές (π.χ. ειδήσεις για την πορεία των έργων). Όμως, η πυκνότητα τέτοιων αναφορών ήταν χαμηλή σε σχέση με το συνολικό θόρυβο. (γ) Έλλειψη μεταδεδομένων συναισθήματος: Σε αντίθεση με το Reddit όπου έχουμε upvotes/downvotes που δίνουν ένα σήμα αντίδρασης της κοινότητας, στο Telegram κάθε μήνυμα στέκεται μόνο του. Δεν υπάρχουν εύκολα μετρήσιμα σήματα θετικής/αρνητικής αντίδρασης πέρα από ίσως emoji reactions, που όμως δεν είναι πάντοτε ενεργοποιημένα ή σχετιζόμενα με συναίσθημα.

Για την ανάλυση συναισθήματος στα Telegram δεδομένα, η ίδια μεθοδολογία θα μπορούσε να εφαρμοστεί (preprocessing -> lexicon sentiment -> classification). Εντούτοις, θα απαιτούσε επαναξιολόγηση του λεξιλογίου: η γλώσσα στα μηνύματα μπορεί να είναι πιο προφορική ή συντομογραφική. Π.χ. πολλοί χρήστες χρησιμοποιούν αυτοκόλλητα ή συντομογραφίες που δύσκολα αναλύονται. Επιπλέον, πιθανώς θα χρειάζονταν νέες επισημάνσεις δεδομένων Telegram, γιατί το μοντέλο που εκπαιδεύτηκε σε Reddit σχόλια μπορεί να μην γενικεύει τέλεια σε Telegram μηνύματα (διαφορετικό ύφος, μήκος, context). Σημαντική πρόκληση είναι και η ιδιωτικότητα: ενώ τα δημόσια κανάλια είναι ανοιχτά, οι ομαδικές συνομιλίες απαιτούν πρόσκληση. Στην παρούσα εργασία περιοριστήκαμε μόνο σε ανοιχτά δεδομένα, αλλά αν επεκταθεί θα πρέπει να ληφθεί μέριμνα να μην παραβιαστούν κλειστές ομάδες.

5.4.2 Mastodon

Το Mastodon είναι ένα αποκεντρωμένο κοινωνικό δίκτυο microblogging (παρόμοιο με το Twitter ως προς τη λειτουργικότητα, αλλά κατανεμημένο σε πολλούς διακομιστές – instances). Για την πρόσβαση σε δεδομένα Mastodon, υπάρχει επίσημο Mastodon API και η Python βιβλιοθήκη Mastodon.py που το υλοποιεί πλήρως[10]. Κάθε instance (π.χ. mastodon.social ή ένα ελληνικό instance όπως hellas.social) προσφέρει API endpoints για λήψη δημόσιων αναρτήσεων (toots), αναζήτηση hashtag, ροή δημοφιλών αναρτήσεων κ.λπ. Χρησιμοποιώντας τη Mastodon.py, μπορεί κανείς να συνδεθεί σε έναν συγκεκριμένο διακομιστή (με ή χωρίς λογαριασμό) και να συλλέξει δημοσιεύσεις με συγκεκριμένο hashtag ή keyword. Για την περίπτωση του Μετρό Θεσσαλονίκης, μια προφανής προσέγγιση θα ήταν η αναζήτηση του hashtag #μετρό ή λέξεων-κλειδιών όπως "Θεσσαλονίκη" στα δημόσια timeline.

Και εδώ όμως εμφανίζονται προκλήσεις λόγω της φύσης του Mastodon: (α) Αποκέντρωση: Δεν υπάρχει κεντρική βάση δεδομένων όλων των toots. Θα πρέπει να αποφασιστεί ποια instance(s) να ερωτηθούν. Ένα μεγάλο διεθνές instance μπορεί να μην περιέχει όλους τους Έλληνες χρήστες. Χρειάζεται είτε να στοχεύσουμε συγκεκριμένα ελληνικά instances, είτε να χρησιμοποιήσουμε τη δυνατότητα fediverse search εφόσον υπάρχει (πολλά instances δεν επιτρέπουν γενική αναζήτηση περιεχομένου για λόγους ιδιωτικότητας). (β) Όγκος δεδομένων: Η κοινότητα του Mastodon στην Ελλάδα είναι συγκριτικά μικρότερη από το Reddit. Πιθανώς οι αναρτήσεις για το θέμα να είναι λίγες. Αυτό μπορεί να δυσκολέψει τη συγκέντρωση επαρκούς δείγματος για ανάλυση. (γ) Μορφή κειμένου: Οι αναρτήσεις στο Mastodon έχουν περιορισμό χαρακτήρων (συνήθως 500 χαρακτήρες). Συνεπώς, το ύφος είναι σύντομο, πολλές

φορές τηλεγραφικό ή με χρήση hashtags. Για παράδειγμα, ένας χρήστης μπορεί να γράψει: "#Θεσσαλονίκη Το μετρό πάλι καθυστερεί...". Εδώ το hashtag λειτουργεί ως λέξη-κλειδί. Θα πρέπει η προεπεξεργασία να διατηρήσει τέτοια hashtags (π.χ. να αφαιρέσει το # αλλά να κρατήσει τη λέξη). Επίσης, η συναισθηματική ανάλυση ίσως είναι πιο δύσκολη με τόσο σύντομα κείμενα όπου λείπει το context.

Παρόλα αυτά, η τεχνική υλοποίηση συλλογής είναι εφικτή. Με ένα script Mastodon.py θα μπορούσαν να συλλεχθούν δημοσιεύσεις από 1-2 δημοφιλή instances, φιλτραρισμένες με βάση το keyword "μετρό" ή "Thessaloniki". Τα δεδομένα επιστρέφονται σε μορφή JSON (με πεδία content, timestamps, favourites, boosts κ.ά.). Η συναισθηματική ανάλυση μπορεί να εφαρμοστεί όπως προηγουμένως, ενδεχομένως όμως θα πρέπει να επανεκπαιδευτούν τα μοντέλα ή τουλάχιστον να επαληθευτεί η απόδοσή τους, επειδή η κατανομή της γλώσσας στο Mastodon μπορεί να διαφέρει. Για παράδειγμα, μπορεί να χρησιμοποιούνται περισσότερο αγγλικοί όροι ή meme εκφράσεις, οι οποίες δεν υπήρχαν στα δεδομένα του Reddit[10].

5.5 Μελλοντικά

Συνοψίζοντας, η επέκταση της έρευνας σε Telegram και Mastodon θα παρείχε μια πιο σφαιρική εικόνα της δημόσιας αντίδρασης γύρω από το Μετρό Θεσσαλονίκης, περιλαμβάνοντας και άλλες διαδικτυακές κοινότητες πέραν του Reddit. Τεχνικά, υπάρχουν διαθέσιμα εργαλεία (API και βιβλιοθήκες Python) για την συλλογή: το Telegram API (μέσω Telethon) και το Mastodon API (μέσω Mastodon.py) καθιστούν δυνατή την άντληση δεδομένων[10][11]. Παρ' όλ' αυτά, η διαδικασία δεν είναι τόσο άμεση όσο με το Reddit/Pushshift. Απαιτεί εντοπισμό των σωστών ομάδων/hashtags, αντιμετώπιση αποκεντρωμένων αρχιτεκτονικών, καθώς και πιθανή δημιουργία νέου training data για συναίσθημα ώστε τα μοντέλα να παραμείνουν ακριβή. Σημαντικό επίσης είναι το νομικό/ηθικό πλαίσιο: όλα τα δεδομένα πρέπει να είναι δημόσια και να μην παραβιάζουν την ιδιωτικότητα χρηστών.

Σε μελλοντική εργασία, θα μπορούσε να επιχειρηθεί η συγχώνευση δεδομένων από όλες αυτές τις πηγές (Reddit, Telegram, Mastodon) για μια συγκριτική ανάλυση. Θα είχε ενδιαφέρον να δούμε αν το συναίσθημα των πολιτών εκφράζεται με τον ίδιο τρόπο και στις τρεις πλατφόρμες ή αν υπάρχουν διαφοροποιήσεις (π.χ. πιο επιθετικό ύφος σε μια πλατφόρμα έναντι άλλης). Η μεθοδολογία που περιγράφηκε σε αυτή την ενότητα θέτει τις βάσεις, καθώς πολλά από τα βήματα (π.χ. preprocessing, lexicon-based sentiment) είναι κοινώς εφαρμόσιμα, με τις αναγκαίες προσαρμογές. Με αυτόν τον τρόπο, η μελέτη του δημόσιου λόγου γύρω από το Μετρό Θεσσαλονίκης μπορεί να επεκταθεί πέρα από το Reddit, καλύπτοντας ένα μεγαλύτερο εύρος του ψηφιακού οικοσυστήματος επικοινωνίας.

Κεφάλαιο 6ο: Συμπεράσματα

Σε αυτή την εργασία πραγματοποιήθηκε ανάλυση συναισθήματος σε 5.323 σχόλια του Reddit σχετικά με το έργο του Μετρό Θεσσαλονίκης, γραμμένα τόσο στα ελληνικά όσο και στα αγγλικά. Η μεθοδολογία που ακολουθήθηκε συνδύασε κλασικές και σύγχρονες προσεγγίσεις επεξεργασίας φυσικής γλώσσας: αρχικά διενεργήθηκε ενδελεχής προκαταρκτική επεξεργασία και καθαρισμός των δεδομένων (αφαίρεση θορύβου, μορφοποίηση κειμένου, φιλτράρισμα άσχετων αναρτήσεων), και στη συνέχεια εφαρμόστηκαν (1) κανόνας-βασιζόμενες τεχνικές με χρήση λεξιλογίων συναισθήματος για τον εντοπισμό θετικά ή αρνητικά φορτισμένων λέξεων, και (2) ένα πολυγλωσσικό μοντέλο BERT για ακριβέστερη ανάλυση στο επίπεδο συμφραζομένων. Επιπλέον, αξιοποιήθηκαν οπτικοποιήσεις όπως σύννεφα λέξεων (word clouds), διαγράμματα κατανομής συναισθήματος και χρονοσειρές sentiment που ανέδειξαν χαρακτηριστικές αιχμές (spikes) σε περιόδους σημαντικών γεγονότων.

Ανακεφαλαιώνοντας τα ευρήματα, διαπιστώνεται ότι η πλειονότητα των σχολίων παρουσιάζει ουδέτερο τόνο. Συγκεκριμένα, περίπου 62% των σχολίων χαρακτηρίστηκαν ουδέτερα ως προς το συναίσθημα, ενώ το 25% ταξινομήθηκαν ως αρνητικά και μόλις 13% ως θετικά. Η επικράτηση ουδέτερου ύφους υποδηλώνει ότι πολλοί χρήστες συζητούν το θέμα πληροφοριακά ή χωρίς έκδηλο συναίσθημα. Παρ' όλα αυτά, το αρνητικό συναίσθημα εμφανίζεται σχεδόν διπλάσιο του θετικού, γεγονός που πιθανώς αντικατοπτρίζει τη δυσαρέσκεια ή την ανυπομονησία ενός σημαντικού μέρους του κοινού λόγω των πολλαπλών καθυστερήσεων και προκλήσεων που έχουν συνδεθεί ιστορικά με το έργο του μετρό. Οι χρονοσειρές συναισθήματος επιβεβαιώνουν ότι οι μεταβολές στην κοινή γνώμη δεν είναι στατικές, αλλά επηρεάζονται άμεσα από εξωγενή γεγονότα: για παράδειγμα, στις αρχές Μαρτίου 2023 –μετά το τραγικό δυστύχημα στα Τέμπη, παρατηρήθηκε απότομη πτώση του μέσου sentiment (αύξηση αρνητικών σχολίων), αντανακλώντας την ευρύτερη κοινωνική οργή και ανησυχία για θέματα ασφάλειας στις μεταφορές. Αντίθετα, προς τα τέλη του 2023 και στις αρχές 2024, καθώς πλησίαζε η περίοδος των (προγραμματισμένων) εγκαινίων του μετρό, εμφανίζεται άνοδος του θετικού κλίματος στις συζητήσεις, ένδειξη αυξανόμενης αισιοδοξίας και ενθουσιασμού του κοινού για την επικείμενη λειτουργία της νέας υποδομής.

Ιδιαίτερη μνεία αξίζει η σύγκριση της rule-based προσέγγισης με το μοντέλο BERT, καθώς αναδεικνύει τη σημασία του συμφραστικού νοήματος στην ανάλυση συναισθήματος. Το rule based method πέτυχε μια πρώτη αποτύπωση της συναισθηματικής πόλωσης εντοπίζοντας λέξεις με θετική ή αρνητική χροιά, όμως παρουσίασε περιορισμούς σε περιπτώσεις όπου το πραγματικό νόημα εξαρτιόταν από τα συμφραζόμενα. Αντιθέτως, το πολυγλωσσικό BERT, εκπαιδευμένο να κατανοεί συμφραζόμενα τόσο στα ελληνικά όσο και στα αγγλικά, κατάφερε να ταξινομήσει με μεγαλύτερη ακρίβεια τα σχόλια. Για παράδειγμα, εντοπίστηκαν περιπτώσεις όπου μια φράση με φαινομενικά θετικές λέξεις χρησιμοποιούνταν ειρωνικά, κάτι που το λεξικο-βασιζόμενο σύστημα παρερμήνευε ως θετικό συναίσθημα. Το BERT, λαμβάνοντας υπόψη τη συνολική πρόταση και το ύφος, απέδωσε ορθότερα το αρνητικό συναίσθημα σε τέτοιες περιπτώσεις. Συνολικά, το μοντέλο BERT υπερείχε στην κατανόηση των πολύπλοκων γλωσσικών αποχρώσεων και της ανάμεικτης γλώσσας (Greeklish, αγγλικές εκφράσεις σε ελληνικά συμφραζόμενα κ.λπ.), προσφέροντας μια πιο αξιόπιστη και λεπτομερή χαρτογράφηση της στάσης του κοινού συγκριτικά με την παραδοσιακή προσέγγιση.

Από πρακτικής πλευράς, τα αποτελέσματα της μελέτης έχουν σημαντική εφαρμοσμένη αξία. Δείχνουν πώς η ανάλυση συναισθήματος σε δεδομένα κοινωνικών δικτύων μπορεί να αξιοποιηθεί από φορείς χάραξης πολιτικής και αρμόδιες αρχές για την κατανόηση και την παρακολούθηση της κοινωνικής αντίδρασης σε ένα μεγάλο έργο υποδομής. Στην περίπτωση του Μετρό Θεσσαλονίκης, οι υπεύθυνοι του έργου ή οι τοπικές αρχές μπορούν να παρακολουθούν τις τάσεις του κοινού σε πραγματικό χρόνο: μια παρατεταμένη έξαρση αρνητικών σχολίων θα μπορούσε να λειτουργήσει ως προειδοποιητικό σήμα αυξανόμενης δυσαρέσκειας ή έλλειψης εμπιστοσύνης των πολιτών, υποδεικνύοντας την ανάγκη για παρεμβάσεις (π.χ. καλύτερη ενημέρωση, βελτίωση διαφάνειας ή μέτρα αντιμετώπισης προβλημάτων) ώστε να αποκατασταθεί η εμπιστοσύνη. Αντίστοιχα, η παρατήρηση αυξημένου θετικού σχολιασμού, όπως ενόψει των εγκαινίων, υποδηλώνει επιτυχία στις προσπάθειες ενημέρωσης και δημοσίων σχέσεων και μια γενικότερη δημόσια αποδοχή του έργου, γεγονός που μπορεί να ενθαρρύνει περαιτέρω υποστήριξη και εμπλοκή της κοινότητας. Με άλλα λόγια, η παρούσα μελέτη καταδεικνύει ότι η συστηματική ανάλυση του διαδικτυακού λόγου γύρω από το μετρό μπορεί να λειτουργήσει ως βαρόμετρο της κοινωνικής στάσης, παρέχοντας πολύτιμη πληροφόρηση για το πώς αντιμετωπίζει το κοινό την πρόοδο, τις καθυστερήσεις ή τα περιστατικά που σχετίζονται με το έργο. Αυτή η γνώση μπορεί να συμβάλει σε αποτελεσματικότερη διαχείριση του έργου στο επίπεδο της δημόσιας εικόνας: οι υπεύθυνοι μπορούν να λαμβάνουν τεκμηριωμένες αποφάσεις για στρατηγικές επικοινωνίας, να εντοπίζουν εγκαίρως κρίσιμα ζητήματα που απασχολούν τους πολίτες και να ενισχύουν την κοινωνική συναίνεση γύρω από την υλοποίηση του μετρό.

Κλείνοντας, αξίζει να επισημανθούν ορισμένες προοπτικές για μελλοντική έρευνα που θα μπορούσαν να επεκτείνουν και να εμβαθύνουν τα συμπεράσματα της παρούσας μελέτης:

- Επέκταση σε άλλες πλατφόρμες: Συλλογή και ανάλυση δεδομένων από επιπλέον μέσα κοινωνικής δικτύωσης, όπως δημόσια κανάλια συζήτησης στο Telegram ή αναρτήσεις στο Mastodon. Αυτό θα παρείχε μια πιο ολοκληρωμένη εικόνα της κοινής γνώμης, καθώς διαφορετικές πλατφόρμες προσελκύουν διαφορετικές κοινότητες χρηστών και ενδέχεται να αναδεικνύουν ποικιλία απόψεων που δεν εμφανίζονται στο Reddit.

- Συγκριτικές μελέτες με άλλα έργα: Εφαρμογή παρόμοιας ανάλυσης συναισθήματος σε σχόλια γύρω από άλλα μεγάλα έργα υποδομής, τόσο στην Ελλάδα (π.χ. επεκτάσεις του Μετρό Αθήνας, μεγάλα οδικά ή σιδηροδρομικά έργα) όσο και διεθνώς. Μια τέτοια σύγκριση θα μπορούσε να αναδείξει αν υπάρχουν κοινά πρότυπα στην εξέλιξη του δημόσιου συναισθήματος (π.χ. αρχικός ενθουσιασμός που φθίνει σε περίπτωση καθυστερήσεων, ή αυξημένη ευαισθησία μετά από ατυχήματα) ή αν διαφοροποιούνται οι αντιδράσεις ανά πολιτισμικό και κοινωνικό πλαίσιο.

- Χρήση προηγμένων γλωσσικών μοντέλων: Διερεύνηση της απόδοσης νεότερων μοντέλων τεχνητής νοημοσύνης στην ανάλυση συναισθήματος, όπως εξειδικευμένα Μεγάλα Γλωσσικά Μοντέλα (LLMs) που έχουν εκπαιδευτεί σε τεράστιο εύρος δεδομένων. Τέτοια μοντέλα (π.χ. νεότερες εκδόσεις τύπου GPT) θα μπορούσαν ενδεχομένως να αναγνωρίζουν ακόμη πιο περίπλοκα συναισθηματικά συμφραζόμενα, να ανιχνεύουν σαρκασμό ή λεπτές διαβαθμίσεις συναισθήματος, καθώς και να κατηγοριοποιούν το συναίσθημα σε περισσότερες κατηγορίες (όπως φόβο, θυμό, έκπληξη) πέρα από το απλό θετικό/αρνητικό/ουδέτερο. Η σύγκριση της απόδοσής τους με το τρέχον μοντέλο BERT θα είχε ενδιαφέροντα ευρήματα ως προς το πόσο βελτιώνεται η ακρίβεια και η κατανόηση.

- Γεωγραφική και κοινοτική ανάλυση του συναισθήματος: Μια περαιτέρω ανάλυση θα μπορούσε να εστιάσει στις γεωγραφικές ή δημογραφικές διαφορές στην έκφραση συναισθήματος. Για παράδειγμα, αν τα δεδομένα το επιτρέψουν, θα είχε αξία να διερευνηθεί κατά πόσον οι κάτοικοι της Θεσσαλονίκης εμφανίζουν διαφορετική στάση στα σχόλιά τους σε σχέση με χρήστες από άλλες

περιοχές της Ελλάδας ή του εξωτερικού όταν μιλούν για το μετρό. Παρομοίως, η ανάλυση ανά subreddit ή διαδικτυακή κοινότητα (_π.χ. σύγκριση του κλίματος μεταξύ θεμάτων που δημοσιεύθηκαν στο r/Thessaloniki έναντι του r/Greece) θα μπορούσε να αποκαλύψει πώς το πλαίσιο και το κοινό κάθε κοινότητας επηρεάζουν το εκφραζόμενο συναίσθημα.

Συμπερασματικά, η μελέτη αυτή ανέδειξε την αποτελεσματικότητα μιας πολυδιάστατης προσέγγισης στην ανάλυση συναισθήματος, συνδυάζοντας τεχνικές λεξιλογίου και σύγχρονα νευρωνικά μοντέλα, για τη χαρτογράφηση της δημόσιας στάσης απέναντι στο Μετρό Θεσσαλονίκης. Τα αποτελέσματα προσφέρουν πολύτιμες γνώσεις τόσο σε ερευνητικό επίπεδο, σχετικά με το πώς εξελίσσεται και επηρεάζεται το δημόσιο συναίσθημα σε διγλωσσικά δεδομένα κοινωνικών μέσων, όσο και σε πρακτικό επίπεδο, όπου μπορούν να αξιοποιηθούν για την υποστήριξη της λήψης αποφάσεων και τη βελτίωση της επικοινωνίας μεταξύ πολιτών και αρμοδίων σε μεγάλα έργα υποδομής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Ελληνικό Μετρό Α.Ε., «Παράδοση του Μετρό Θεσσαλονίκης σε εμπορική λειτουργία», 30 Νοεμβρίου 2024, https://www.emetro.gr/?p=28810

[2] G. Alexandridis, I. Varlamis, K. Korovesis, G. Caridakis και P. Tsantilas, «A Survey on Sentiment Analysis and Opinion Mining in Greek Social Media», Information, τόμ. 12, αρ. 8, σελ. 331, 2021. DOI: 10.3390/info12080331

[3] L. Samaras, E. G. Barriocanal, M. A. Sicilia, Sentiment analysis of COVID-19 cases in Greece using Twitter data, Expert Systems with Applications, Volume 230, 2023, 120577, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.120577.

[4] A. Tsakalidis, S. Papadopoulos, R. Voskaki, K. Ioannidou, C. Boididou, A. I. Cristea, M. Liakata και Y. Kompatsiaris, «Building and evaluating resources for sentiment analysis in the Greek language», Language Resources and Evaluation, τόμ. 52, αρ. 4, σελ. 1021–1044, 2018. DOI: 10.1007/s10579-018-9420-4

[5] J. Koutsikakis, I. Chalkidis, P. Malakasiotis και I. Androutsopoulos, «GREEK-BERT: The Greeks Visiting Sesame Street», πρακτικά 11ου Hellenic Conference on Artificial Intelligence (SETN 2020), Αθήνα, 2020, σελ. 110–117

[6] F. Zare Mehrjardi, M. Yazdian-Dehkordi και A. Latif, «Evaluating classical machine learning and deep-learning methods in sentiment analysis of Persian telegram messages», Soft Computing Journal, τόμ. 11, αρ. 1, σελ. 88–105, 2022. DOI: 10.22052/scj.2023.246553ί.1077

[7] Aristotle University of Thessaloniki, «Mastodon posts dataset (Greek wildfires 2023)», AI4EU – AI-on-Demand Platform, ενημ. 28 Νοεμ. 2023, https://www.ai4europe.eu/research/ai-catalog/mastodon-posts-dataset

[8] F. Barbieri, L. Espinosa-Anke και J. Camacho-Collados, «XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond», πρακτικά 13ου Conference on Language Resources and Evaluation (LREC), Μασσαλία, 2022, σελ. 258–266

[9] I. Gopal, «Collecting messages from Telegram using Telegram's API and Python», Medium, 23 Μαρ. 2023, https://medium.com/@ishitagopal/collecting-messages-from-telegram-using-telegrams-api-and-python-5d7e4a9286b2

[10] S. Z. Qazi, «Mastodon Data Extraction (Research And Learning Purpose)», Medium, 10 Φεβ. 2024, https://medium.com/@syedzainullahqazi/mastodon-data-extraction-research-and-learning-purpose-1eec53068b15

[11] Mastodon.py – Mastodon API Python wrapper documentation, 2023, https://mastodonpy.readthedocs.io/

[12] A. Tsakalidis κ.ά., «Greek Sentiment Lexicon (SocialSensor project)», GitHub, 2018, https://github.com/MKLab-ITI/greek-sentiment-lexicon

[13] N. Boettcher, «Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review», JMIR Mental Health, τόμ. 8, αρ. 11, σελ. e29487, 2021. DOI: 10.2196/29487

[14] C. Cerisara, S. Jafaritazehjani, A. Oluokun και H. T. Le, «Multi-task dialog act and sentiment recognition on Mastodon», πρακτικά 27ου International Conference on Computational Linguistics (COLING), Σάντα Φε, 2018, σελ. 745–754

[15] C. Cerisara, «DialogSentimentMastodon – source code», GitHub, 2018. ,https://github.com/cerisara/DialogSentimentMastodon

[16] M. Θαλασσινού-Λισλεβάντ (Marina Thalassinou-Lislevand), «Emotion Classification on Greek Tweets», MSc Διπλωματική Εργασία, Οικονομικό Πανεπιστήμιο Αθηνών, 2020

[17] Newsroom (OT), «Έρευνα: Οργή στη νέα γενιά μετά τα Τέμπη – Πάνω από το 82% δηλώνει ότι θα ψηφίσει στις εκλογές», Οικονομικός Ταχυδρόμος (ot.gr), 08 Μαΐου 2023, https://www.ot.gr/2023/05/08/epikairothta/ereyna-orgi-sti-nea-genia-meta-ta-tempi-pano-apo-to-82-dilonei-oti-tha-psifisei-stis-ekloges

[18] G. Kalamatianos & D. Mallis & S. Symeonidis & A. Arampatzis. (2015). Sentiment Analysis of Greek Tweets and Hashtags using a Sentiment Lexicon. The 19th Panhellenic Conference on Informatics (PCI 2015), 1-3 October. 63-68. 10.1145/2801948.2802010

[19] Tsakalidis, A., Papadopoulos, S., Voskaki, R. et al. Building and evaluating resources for sentiment analysis in the Greek language. Lang Resources & Evaluation 52, 1021–1044 (2018). https://doi.org/10.1007/s10579-018-9420-4

[20] Νικόλαος Κρυστάλλης, "Μελέτη δεδομένων κοινωνικών δικτύων", Μεταπτυχιακή Διατριβή, Σχολή Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης, Στρατιωτική Σχολή Ευελπίδων, Χανιά, Ελλάς, 2021 https://doi.org/10.26233/heallink.tuc.89031

[21] S. Outsios, C. Karatsalos, K. Skianis, M. Vazirgiannis «Evaluation of Greek Word Embeddings», ΕΛΕΤΟ – 12ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία», Αθήνα, 7–9 Νοεμβρίου 2019

[22] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, K. Ch. Chatzisavvas, Sentiment analysis leveraging emotions and word embeddings, Expert Systems with Applications, Volume 69, 2017, Pages 214-224, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2016.10.043.

[23] D. Kydros, M. Argyropoulou, V. Vrana, 2021. "COVID-19 Goes on Twitter. Greek Conversations and Discussions," Springer Proceedings in Business and Economics, in: A. Kavoura & S. J. Havlovic & N. Totskaya (ed.), Strategic Innovative Marketing and Tourism in the COVID-19 Era, pages 77-86, Springer.

[24] C. Mastrokostas, N. Giarelis, N. Karacapilidis, Social Media Topic Classification on Greek Reddit. Information 2024, 15, 521. https://doi.org/10.3390/info15090521