# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
# ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ
# ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

## Πτυχιακή Εργασία

| | |
|---|---|
| Τίτλος Πτυχιακής Εργασίας | Χρήση του μοντέλου BERT για ταξινόμηση συναισθήματος<br><br>Sentiment classification using BERT Model |
| Ονοματεπώνυμο Φοιτητή | Αθανάσιος Παπούλιας |
| Πατρώνυμο | Μηνάς |
| Αριθμός Μητρώου | Π19135 |
| Επιβλέπων Καθηγητής | Διονύσης Σωτηρόπουλος, Αν. Καθηγητής |

Ημερομηνία Παράδοσης **Φεβρουάριος 2026**

# Copyright ©

# Ευχαριστίες

Με την ολοκλήρωση της πτυχιακής εργασίας, θέλω να πω ένα μεγάλο ευχαριστώ σε όλους όσοι με στήριξαν σε αυτή τη διαδρομή.
Ένα μεγάλο ευχαριστώ στον επιβλέποντα καθηγητή μου, τον κ. Διονύσιο Σωτηρόπουλο. Τον ευχαριστώ πολύ για την πολύτιμη καθοδήγησή του, τη βοήθεια και την υπομονή που μου έδειξε σε όλη τη διάρκεια της προσπάθειάς μου.
Φυσικά, το μεγαλύτερο ευχαριστώ ανήκει στην οικογένειά μου. Τους ευχαριστώ που ήταν πάντα εκεί για μένα, που με πίστεψαν και που με στήριξαν με κάθε τρόπο ώστε να φτάσω μέχρι εδώ. Χωρίς αυτούς, όλα θα ήταν πολύ πιο δύσκολα.
Τέλος, ένα μεγάλο ευχαριστώ στους φίλους μου. Τους ευχαριστώ για τις όμορφες στιγμές, τις συζητήσεις και την παρέα τους,

# Περίληψη

Η ανάλυση συναισθήματος σε δεδομένα κοινωνικής δικτύωσης αποτελεί ένα από τα πλέον ενεργά πεδία έρευνας στην Επεξεργασία Φυσικής Γλώσσας, με ευρείες εφαρμογές στην παρακολούθηση κοινής γνώμης, τη διαχείριση φήμης και την ανάλυση αγοράς. Στην εργασία παρουσιάζεται μια συγκριτική μελέτη μεθόδων αυτόματης ταξινόμησης συναισθήματος σε αναρτήσεις της πλατφόρμας Twitter, με χρήση του συνόλου δεδομένων Sentiment140, το οποίο αποτελείται από 1.6 εκατομμύρια tweets επισημασμένα μέσω απομακρυσμένης επίβλεψης. Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε στρωματοποιημένο υποσύνολο 200.000 tweets, ισορροπημένο ως προς τις δύο κλάσεις (θετικό/αρνητικό). Εξετάστηκαν συνολικά επτά μοντέλα, κατανεμημένα σε δύο κατηγορίες: πέντε κλασικά μοντέλα μηχανικής μάθησης με αναπαραστάσεις TF-IDF (Logistic Regression, Naive Bayes, Linear SVM, Random Forest, Gradient Boosting) και δύο μοντέλα βασισμένα σε αρχιτεκτονικές Transformer (DistilBERT, Twitter-RoBERTa). Τα αποτελέσματα καταδεικνύουν σαφή υπεροχή των μοντέλων Transformer, με το Twitter-RoBERTa να επιτυγχάνει την υψηλότερη απόδοση έναντι της καλύτερης κλασικής μεθόδου. Τα ευρήματα υπογραμμίζουν τη σημασία της εξειδικευμένης προεκπαίδευσης σε domain-specific δεδομένα και επιβεβαιώνουν την ανωτερότητα των μοντέλων βαθιάς μάθησης για το συγκεκριμένο έργο, ενώ παράλληλα αναδεικνύουν τα κλασικά μοντέλα ως βιώσιμη επιλογή υπό υπολογιστικούς περιορισμούς.


**Λέξεις κλειδιά**: Ανάλυση Συναισθήματος, Επεξεργασία Φυσικής Γλώσσας, Twitter, TF-IDF, Logistic Regression, Support Vector Machine, Transformer, BERT, DistilBERT, RoBERTa, Finetuning


# Abstract

Sentiment analysis in social networking data is one of the most active research fields in Natural Language Processing, with broad applications in public opinion monitoring, reputation management and market analysis. This paper presents a comparative study of automatic sentiment classification methods in Twitter platform posts, using the Sentiment140 dataset, which consists of 1.6 million tweets tagged through remote supervision. A stratified subset of 200,000 tweets, balanced in terms of the two classes (positive/negative), was used for training and evaluating the models. A total of seven models were examined, divided into two categories: five classic machine learning models with TF-IDF representations (Logistic Regression, Naive Bayes, Linear SVM, Random Forest, Gradient Boosting) and two models based on Transformer architectures (DistilBERT, Twitter-RoBERTa). The results demonstrate a clear superiority of the

Transformer models, with Twitter-RoBERTa achieving the highest performance (F1-score: 0.882, ROC-AUC: 0.953), compared to 0.780 for the best classical method. The findings highlight the importance of specialized pre-training on domain-specific data and confirm the superiority of deep learning models for this specific task, while also highlighting classical models as a viable option under computational constraints.

Keywords: Sentiment Analysis, Natural Language Processing, Twitter, TF-IDF, Logistic Regression, Support Vector Machine, Transformer, BERT, DistilBERT, RoBERTa, Fine-tuning.

## Πίνακας Περιεχομένων

## Πίνακας Εικόνων



## Πίνακας πινάκων

# Κεφάλαιο 1: Εισαγωγή

## 1.1 Γενική Εισαγωγή

Η ραγδαία ανάπτυξη των πλατφορμών κοινωνικής δικτύωσης κατά την τελευταία δεκαετία έχει μεταμορφώσει τον τρόπο με τον οποίο οι άνθρωποι εκφράζουν απόψεις, συναισθήματα και εμπειρίες. Πλατφόρμες όπως το Twitter, με εκατοντάδες εκατομμύρια ενεργούς χρήστες παγκοσμίως και πάνω από 500 εκατομμύρια αναρτήσεις ημερησίως , αποτελούν έναν ανεξάντλητο χώρο έκφρασης της ανθρώπινης γνώμης σε πραγματικό χρόνο. Το περιεχόμενο αυτό, αν και ανομοιόμορφο και θορυβώδες, ενέχει τεράστια αξία για οργανισμούς, επιχειρήσεις και ερευνητές που επιθυμούν να κατανοήσουν τη δημόσια γνώμη σε ένα ευρύ φάσμα θεμάτων.

Η αυτόματη εξαγωγή συναισθήματος από κείμενο, γνωστή ως **ανάλυση συναισθήματος** (sentiment analysis), αποτελεί έναν από τους πιο ενεργούς και εφαρμοσμένους τομείς της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Στόχος της είναι ο αυτόματος προσδιορισμός της συναισθηματικής πολικότητας ενός κειμένου, δηλαδή αν εκφράζει θετική, αρνητική ή ουδέτερη στάση, χωρίς την ανάγκη χειρωνακτικής επισκόπησης από ανθρώπους. Οι εφαρμογές της εκτείνονται από την παρακολούθηση της φήμης επιχειρήσεων και προϊόντων, έως την ανίχνευση κοινωνικών τάσεων, την πολιτική ανάλυση και την υποστήριξη αποφάσεων σε πραγματικό χρόνο.

Ωστόσο, η ανάλυση συναισθήματος σε δεδομένα Twitter παρουσιάζει μοναδικές προκλήσεις που την καθιστούν ιδιαίτερα απαιτητικό έργο. Ο περιορισμός των χαρακτήρων επιβάλλει συνοπτικές και συχνά ελλιπείς διατυπώσεις, ενώ η χρήση αργκό, συντομογραφιών, hashtags, emoji και ειρωνείας δυσκολεύει σημαντικά την αυτόματη κατανόηση. Παράλληλα, ο τεράστιος όγκος δεδομένων καθιστά απαγορευτική την ανθρώπινη επισήμανση σε μεγάλη κλίμακα, εντείνοντας την ανάγκη για αξιόπιστες αυτόματες μεθόδους.

Οι προσεγγίσεις για την αντιμετώπιση του προβλήματος έχουν εξελιχθεί σημαντικά με την πάροδο του χρόνου. Από τις πρώιμες λεξικογραφικές μεθόδους και τα κλασικά μοντέλα μηχανικής μάθησης με χειροποίητα χαρακτηριστικά, η έρευνα εξελίχθηκε προς νευρωνικά δίκτυα και τελικά στις αρχιτεκτονικές Transformer, που εισήγαγε ο [1], οι οποίες επανάστησαν το πεδίο του NLP. Μοντέλα όπως το BERT [2] και οι παραλλαγές του έδειξαν ότι η προεκπαίδευση

σε μεγάλα γλωσσικά σώματα, ακολουθούμενη από fine-tuning σε εξειδικευμένες εργασίες, οδηγεί σε επιδόσεις που υπερβαίνουν κατά πολύ τις παραδοσιακές μεθόδους.

Η υπόλοιπη εργασία δομείται ως εξής: το Κεφάλαιο 2 περιγράφει την βιβλιογραφική ανασκόπηση το Κεφάλαιο και παρουσιάζει το θεωρητικό υπόβαθρο, το Κεφάλαιο 3 περιγράφει το σύνολο δεδομένων και την διερευνητική ανάλυση, το Κεφάλαιο 4 αναλύει την προεπεξεργασία, το Κεφάλαιο 5 παρουσιάζει τα μοντέλα και τη μεθοδολογία, το Κεφάλαιο 6 αναφέρει τα αποτελέσματα, το Κεφάλαιο 7 τα συζητά κριτικά και το Κεφάλαιο 8 παρουσιάζει τα συμπεράσματα και τις μελλοντικές κατευθύνσεις.

## 1.2 Ορισμός Προβλήματος και Στόχοι

Αντικείμενο της εργασίας είναι η **δυαδική ταξινόμηση συναισθήματος** (binary sentiment classification) σε αναρτήσεις της πλατφόρμας Twitter, δηλαδή ο αυτόματος χαρακτηρισμός κάθε tweet ως θετικού ή αρνητικού. Το πρόβλημα διατυπώνεται επίσημα ως εξής: δοθέντος ενός tweet $t$, το σύστημα καλείται να προβλέψει μια ετικέτα $y \in \{0,1\}$, όπου $0$αντιστοιχεί σε αρνητικό και $1$σε θετικό συναίσθημα.

Για την αντιμετώπιση του προβλήματος αυτού χρησιμοποιείται το σύνολο δεδομένων **Sentiment140** [3], ένα ευρέως αναγνωρισμένο benchmark που αποτελείται από 1.6 εκατομμύρια tweets επισημασμένα μέσω απομακρυσμένης επίβλεψης. Η επιλογή του συνόλου αυτού επιτρέπει τη σύγκριση των αποτελεσμάτων με προηγούμενες μελέτες και εξασφαλίζει επαρκή όγκο δεδομένων για την εκπαίδευση μοντέλων βαθιάς μάθησης.

Οι επιμέρους στόχοι της εργασίας είναι οι εξής:

**Στόχος 1 — Σύγκριση κλασικών και σύγχρονων μεθόδων.** Αξιολόγηση και σύγκριση πέντε κλασικών μοντέλων μηχανικής μάθησης με TF-IDF αναπαραστάσεις έναντι δύο μοντέλων Transformer, εντοπίζοντας τις διαφορές απόδοσης και τους παράγοντες που τις ερμηνεύουν.

**Στόχος 2 — Αξιολόγηση domain-specific προεκπαίδευσης.** Διερεύνηση του κατά πόσο η προεκπαίδευση ενός μοντέλου Transformer σε δεδομένα του ίδιου domain (Twitter) προσφέρει μετρήσιμο πλεονέκτημα έναντι ενός γενικής χρήσης μοντέλου υπό ταυτόσημες συνθήκες finetuning.

**Στόχος 3 — Ανάπτυξη ολοκληρωμένης pipeline.** Σχεδιασμός και υλοποίηση μιας πλήρους και αναπαραγώγιμης pipeline προεπεξεργασίας, εκπαίδευσης και αξιολόγησης, η οποία να μπορεί να αποτελέσει βάση για μελλοντικές επεκτάσεις.

**Στόχος 4 — Τεκμηριωμένη επιλογή μοντέλου.** Παροχή τεκμηριωμένων συστάσεων για την επιλογή μοντέλου ανάλυσης συναισθήματος σε Twitter, λαμβάνοντας υπόψη τόσο την απόδοση όσο και το υπολογιστικό κόστος.

# 2. Βιβλιογραφική Ανασκόπηση και Ορισμοί Μηχανικής Μάθησης

## 2.1 Εισαγωγή στην Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis) ή opinion mining αποτελεί υποπεδίο της Επεξεργασίας Φυσικής Γλώσσας που ασχολείται με την αυτόματη αναγνώριση και εξαγωγή υποκειμενικής πληροφορίας από κείμενο. Ο πρωταρχικός στόχος είναι ο προσδιορισμός της συναισθηματικής πολικότητας ενός κειμένου, δηλαδή αν εκφράζει θετική, αρνητική ή ουδέτερη στάση απέναντι σε ένα θέμα, προϊόν ή γεγονός. Πέρα από την απλή πολικότητα, η έρευνα έχει επεκταθεί σε βαθύτερες μορφές ανάλυσης, όπως η αναγνώριση συναισθηματικής έντασης, η ανίχνευση ειρωνείας και σαρκασμού, η ανάλυση συναισθήματος σε επίπεδο πτυχής (aspectbased sentiment analysis) και η αναγνώριση συγκεκριμένων συναισθημάτων όπως χαρά, θυμός ή φόβος.

Οι [4] και [5] θεωρούνται από τους θεμελιώδεις συγγραφείς του πεδίου, με εργασίες τους να αποτελούν σημείο αναφοράς για δεκαετίες. Ο [4] ορίζει την ανάλυση συναισθήματος ως «την υπολογιστική μελέτη απόψεων, συναισθημάτων και υποκειμενικότητας στο κείμενο», διακρίνοντας τρία επίπεδα ανάλυσης: επίπεδο εγγράφου, επίπεδο πρότασης και επίπεδο πτυχής. Στην εργασία εστιάζουμε στο επίπεδο εγγράφου (tweet), όπου κάθε ανάρτηση αντιμετωπίζεται ως ενιαία μονάδα ανάλυσης και της αποδίδεται μία συνολική ετικέτα πολικότητας.

Από άποψη εφαρμογών, η ανάλυση συναισθήματος χρησιμοποιείται σήμερα σε ένα ευρύ φάσμα τομέων. Στον εμπορικό τομέα, εταιρείες αξιοποιούν συστήματα sentiment analysis για την παρακολούθηση της εικόνας των προϊόντων τους στα μέσα κοινωνικής δικτύωσης και την ανίχνευση αρνητικών σχολίων σε πραγματικό χρόνο. Στον πολιτικό τομέα, χρησιμοποιείται για την ανάλυση της κοινής γνώμης κατά τη διάρκεια εκλογικών αναμετρήσεων. Στον χρηματοοικονομικό τομέα, αξιοποιείται για την πρόβλεψη κινήσεων αγοράς βάσει δημόσιου

συναισθήματος. Η πολυδιάστατη αυτή χρησιμότητα εξηγεί την έντονη ερευνητική δραστηριότητα που παρατηρείται στο πεδίο εδώ και δύο δεκαετίες σε συνδυασμό με τις τελευταίες εξελίξεις στην μηχανική μάθηση, [30].

## 2.2 Κλασικές Μέθοδοι Ανάλυσης Συναισθήματος

Πριν από την κυριαρχία των νευρωνικών δικτύων, η ανάλυση συναισθήματος αντιμετωπιζόταν κυρίως μέσω δύο κατηγοριών προσεγγίσεων: λεξικογραφικών μεθόδων και μεθόδων μηχανικής μάθησης με χειροποίητα χαρακτηριστικά.

Οι **λεξικογραφικές μέθοδοι** βασίζονται σε προκατασκευασμένα λεξικά συναισθήματος, τα οποία αποδίδουν σε κάθε λέξη μια τιμή πολικότητας. Μεταξύ των πιο γνωστών λεξικών συγκαταλέγονται το SentiWordNet [7], το AFINN [8] και το VADER [9], το τελευταίο εκ των οποίων σχεδιάστηκε ειδικά για κείμενο κοινωνικής δικτύωσης. Οι λεξικογραφικές μέθοδοι έχουν το πλεονέκτημα ότι δεν απαιτούν εκπαίδευση, είναι ερμηνεύσιμες και εύκολα εφαρμόσιμες. Ωστόσο, αδυνατούν να χειριστούν ειρωνεία, αρνητικές εκφράσεις (π.χ. «not good») και νέες λέξεις που δεν περιλαμβάνονται στο λεξικό.

Οι μέθοδοι μηχανικής μάθησης [29] αντιμετωπίζουν την ανάλυση συναισθήματος ως πρόβλημα εποπτευόμενης ταξινόμησης. Οι [6] ήταν από τους πρώτους που εφάρμοσαν συστηματικά μεθόδους όπως Naive Bayes, Maximum Entropy και SVM σε κριτικές ταινιών, αποδεικνύοντας ότι τα μοντέλα μηχανικής μάθησης υπερέχουν έναντι των λεξικογραφικών προσεγγίσεων. Η αναπαράσταση κειμένου μέσω TF-IDF αποδείχθηκε ιδιαίτερα αποτελεσματική για αυτές τις μεθόδους, καθώς αναδεικνύει τις λέξεις που χαρακτηρίζουν κάθε κλάση. Ο [10] έδειξαν ότι ο συνδυασμός Naive Bayes features με SVM,γνωστός ως NBSVM, επιτυγχάνει εξαιρετικά αποτελέσματα σε πολλαπλά datasets ανάλυσης συναισθήματος, αποτελώντας ισχυρό baseline που παραμένει δύσκολο να νικηθεί από απλούστερες μεθόδους. Οι ensemble μέθοδοι, όπως Random Forest και Gradient Boosting, εφαρμόστηκαν επίσης στο πρόβλημα, αλλά με μικτά αποτελέσματα, συχνά υπολείπονται των γραμμικών μοντέλων όταν το feature space είναι υψηλοδιάστατο και αραιό, όπως συμβαίνει με TF-IDF αναπαραστάσεις.

## 2.3 Αρχιτεκτονικές Transformer και Προεκπαιδευμένα Γλωσσικά Μοντέλα

Η εισαγωγή της αρχιτεκτονικής Transformer από τους [1] σηματοδότησε τομή στο πεδίο του NLP. Ο μηχανισμός **self-attention** που εισήγαγε επιτρέπει στο μοντέλο να σταθμίζει δυναμικά τη σημασία κάθε token σε σχέση με όλα τα υπόλοιπα tokens της ακολουθίας, ανεξαρτήτως της μεταξύ τους απόστασης. Αυτό αντιμετωπίζει ένα από τα βασικά μειονεκτήματα των

παλαιότερων αρχιτεκτονικών RNN/LSTM, οι οποίες δυσκολεύονταν να διατηρήσουν μακρόχρονες εξαρτήσεις στο κείμενο.

Το **BERT** (Bidirectional Encoder Representations from Transformers), που παρουσίασαν οι [2], αξιοποίησε την αρχιτεκτονική Transformer για την κατασκευή ενός ισχυρού προεκπαιδευμένου γλωσσικού μοντέλου. Το BERT εκπαιδεύτηκε σε δύο αυτοεποπτευόμενες εργασίες, Masked Language Modeling (MLM) και Next Sentence Prediction (NSP), σε ένα τεράστιο σώμα κειμένου (Wikipedia και BookCorpus), αποκτώντας βαθιά γλωσσική γνώση. Η στρατηγική προεκπαίδευσης και στη συνέχεια fine-tuning σε εξειδικευμένες εργασίες απέδωσε state-of-the-art αποτελέσματα σε 11 διαφορετικά NLP benchmarks κατά την παρουσίασή του.

Το **DistilBERT** [12] αποτελεί συμπυκνωμένη έκδοση του BERT, εκπαιδευμένη μέσω knowledge distillation. Με 40% λιγότερες παραμέτρους και 60% ταχύτερο χρόνο συμπερασμού, διατηρεί το 97% της απόδοσης του BERT στο benchmark GLUE, καθιστώντας το ιδανικό για εφαρμογές που απαιτούν ισορροπία μεταξύ απόδοσης και υπολογιστικής αποδοτικότητας.

Το **RoBERTa** [13] βελτίωσε το BERT αφαιρώντας την εργασία NSP, αυξάνοντας το μέγεθος του batch και εκπαιδεύοντας σε μεγαλύτερο και ποικιλόμορφο σώμα κειμένου. Οι [13] έδειξαν ότι το BERT ήταν σημαντικά υποεκπαιδευμένο και ότι με κατάλληλη εκπαίδευση η απόδοση βελτιώνεται σημαντικά, με το RoBERTa να ξεπερνά το BERT σε σχεδόν όλα τα benchmarks.

## 2.4 Ανάλυση Συναισθήματος στο Twitter

Η ανάλυση συναισθήματος σε tweets παρουσιάζει ιδιαίτερες προκλήσεις που έχουν αποτελέσει αντικείμενο εκτεταμένης έρευνας. Οι [3] ήταν από τους πρώτους που αντιμετώπισαν συστηματικά το πρόβλημα, εισάγοντας τη μεθοδολογία της απομακρυσμένης επίβλεψης μέσω emoticons και δημιουργώντας το Sentiment140. Στη δουλειά τους, SVM και Naive Bayes με unigram και bigram χαρακτηριστικά επέτυχαν ακρίβεια περίπου 82–83%, θέτοντας ένα ισχυρό baseline για μελλοντικές έρευνες.

Οι [35] επέκτειναν αυτή την προσέγγιση, αναδεικνύοντας τις μοναδικές γλωσσικές ιδιαιτερότητες των tweets και την ανάγκη για εξειδικευμένη προεπεξεργασία. Οι [14] οργάνωσαν το SemEval Twitter Sentiment Analysis task επί πολλά χρόνια, παρέχοντας benchmark αξιολόγησης για δεκάδες ερευνητικές ομάδες παγκοσμίως και καταγράφοντας τη σταδιακή εξέλιξη των μεθόδων από κλασικά μοντέλα προς νευρωνικές αρχιτεκτονικές.

Ένα από τα σημαντικότερα βήματα για την εξειδίκευση των γλωσσικών μοντέλων στο Twitter ήταν η ανάπτυξη του **BERTweet** [16], ενός μοντέλου τύπου BERT εκπαιδευμένου αποκλειστικά σε 850 εκατομμύρια αγγλικά tweets. Το BERTweet επέδειξε σημαντική υπεροχή έναντι του κλασικού BERT σε τρία Twitter NLP tasks, επιβεβαιώνοντας ότι η προεκπαίδευση σε domain-specific δεδομένα είναι ουσιαστικής σημασίας. Παρόμοια συμπεράσματα εξήγαγαν οι [17] με το

**Twitter-RoBERTa**, δείχνοντας ότι ένα μοντέλο προεκπαιδευμένο σε 58 εκατομμύρια tweets υπερέχει συστηματικά έναντι γενικής χρήσης μοντέλων σε Twitter sentiment και άλλα Twitter-specific tasks, ακόμα και όταν τα γενικά μοντέλα έχουν μεγαλύτερο αριθμό παραμέτρων.

## 2.5 Σύγκριση Σχετικών Ερευνών

Ο επόμενος πίνακας συνοψίζει τις κυριότερες σχετικές έρευνες, παρουσιάζοντας τα σύνολα δεδομένων, τις μεθόδους και τα αποτελέσματα που αναφέρονται σε κάθε εργασία.

| **Έρευνα** | Dataset | Μέθοδος | **Καλύτερο Αποτέλεσμα** |
|---|---|---|---|
| [3] | Sentiment140 (1.6M) | SVM + Unigrams | Acc: 83% |
| [6] | Movie Reviews | SVM + BoW | Acc: 82.9% |
| [10] | Πολλαπλά | NBSVM | F1: ~0.89 |
| [14] | SemEval Twitter | Διάφορες | F1: 0.68–0.72 |
| [15] | SST | Fine-tuned BERT | Acc: 92.3% |
| [16] | Twitter NLP tasks | BERTweet | F1: ~0.89 |
| [17] | Twitter Sentiment | Twitter-RoBERTa | F1: ~0.90 |
| **Πτυχιακή εργασία** | **Sentiment140 (200K)** | Twitter-RoBERTa | **F1: 0.882** |

*Πίνακας 1: Σύγκριση σχετικών ερευνών στην ανάλυση συναισθήματος*

Από τη σύγκριση αυτή προκύπτουν χρήσιμα συμπεράσματα. Πρώτον, υπάρχει σαφής ανοδική τάση στις επιδόσεις με την πάροδο του χρόνου, αντανακλώντας την εξέλιξη των μεθόδων από κλασικά μοντέλα προς αρχιτεκτονικές Transformer. Δεύτερον, τα αποτελέσματα δεν είναι πάντα άμεσα συγκρίσιμα λόγω διαφορών στα σύνολα δεδομένων, τις μετρικές και τις συνθήκες αξιολόγησης, γεγονός που υπογραμμίζει την ανάγκη για τυποποιημένα benchmarks. Τρίτον, τα αποτελέσματα της εργασίας (F1: 0.882 για Twitter-RoBERTa) εντάσσονται στο άνω άκρο του εύρους που αναφέρεται στη βιβλιογραφία για το Sentiment140, επιβεβαιώνοντας την αποτελεσματικότητα του pipeline.

## 2.6 Μαθηματική Περιγραφή των Μεθόδων Μηχανικής Μάθησης

### 2.6.1 TF-IDF (Term Frequency – Inverse Document Frequency)

Πριν από την εφαρμογή οποιουδήποτε ταξινομητή, το κείμενο πρέπει να μετατραπεί σε αριθμητική αναπαράσταση. Το TF-IDF [18] είναι ένα στατιστικό μέτρο που αποδίδει σε κάθε λέξη $w$ενός εγγράφου $d$μια τιμή που εκφράζει τη σημασία της εντός του εγγράφου σε σχέση με το σύνολο των εγγράφων $D$. Ορίζεται ως:

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

όπου:

$$\text{TF}(w, d) = \frac{f_{w,d}}{\sum_{w' \in d} f_{w',d}}$$

είναι η κανονικοποιημένη συχνότητα εμφάνισης της λέξης $w$στο έγγραφο $d$, και:

$$\text{IDF}(w, D) = \log\left(\frac{|\,D\,|}{|\,\{d \in D : w \in d\}\,|}\right)$$

είναι ο αντίστροφος λογάριθμος του ποσοστού των εγγράφων που περιέχουν τη λέξη $w$. Λέξεις που εμφανίζονται σε πολλά έγγραφα (π.χ. stopwords) λαμβάνουν χαμηλή τιμή IDF, ενώ λέξεις χαρακτηριστικές συγκεκριμένων εγγράφων λαμβάνουν υψηλή. Στην εργασία εφαρμόστηκε λογαριθμική κλιμάκωση του TF ( sublinear_tf=True), δηλαδή $\text{TF}(w, d) = 1 + \log(f_{w,d})$, για την αντιμετώπιση της υπεραντιπροσώπευσης πολύ συχνών λέξεων.

### 2.6.2 Logistic Regression

Η Λογιστική Παλινδρόμηση [19] είναι γραμμικός ταξινομητής που εκτιμά την πιθανότητα ανήκειν στην κλάση $y = 1$μέσω της σιγμοειδούς συνάρτησης:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

όπου $\mathbf{x} \in \mathbb{R}^d$είναι το διάνυσμα TF-IDF του εγγράφου, $\mathbf{w} \in \mathbb{R}^d$το διάνυσμα βαρών και $b$η μεροληψία (bias). Η εκπαίδευση πραγματοποιείται με ελαχιστοποίηση της συνάρτησης απώλειας cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)]$$

Η απόφαση ταξινόμησης λαμβάνεται με κατώφλι $P(y = 1 \mid \mathbf{x}) \geq 0.5$.

### 2.6.3 Naive Bayes

Ο ταξινομητής Naive Bayes [20] βασίζεται στο θεώρημα Bayes:

$$P(y \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y) \cdot P(y)}{P(\mathbf{x})}$$

Υπό την υπόθεση **υπό συνθήκη ανεξαρτησίας** των χαρακτηριστικών (naive assumption), η πιθανότητα κλάσης απλοποιείται σε:

$$P(y \mid \mathbf{x}) \propto P(y) \prod_{i=1}^{d} P(x_i \mid y)$$

Για κείμενο με TF-IDF αναπαράσταση χρησιμοποιείται η Multinomial παραλλαγή, με εκτίμηση:

$$P(x_i \mid y) = \frac{N_{yi} + \alpha}{N_y + \alpha d}$$

όπου $N_{yi}$είναι η συχνότητα του χαρακτηριστικού $i$στην κλάση $y$, $N_y$το άθροισμα όλων των χαρακτηριστικών της κλάσης και $\alpha$η παράμετρος Laplace smoothing (στην εργασία $\alpha = 0.5$).

### 2.6.4 Linear SVM (Support Vector Machine)

Ο γραμμικός SVM [21], [29] αναζητά το υπερεπίπεδο $\mathbf{w}^T\mathbf{x} + b = 0$που **μεγιστοποιεί το περιθώριο** διαχωρισμού μεταξύ των δύο κλάσεων. Το πρόβλημα βελτιστοποίησης διατυπώνεται ως:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N} \xi_i$$

υπο΄ τους περιορισμου΄ς: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0$

όπου $\xi_i$είναι οι μεταβλητές χαλάρωσης (slack variables) που επιτρέπουν την ύπαρξη λανθασμένα ταξινομημένων σημείων, και $C$η παράμετρος ρύθμισης που ελέγχει τον συμβιβασμό μεταξύ μεγιστοποίησης περιθωρίου και ελαχιστοποίησης λαθών. Ο γραμμικός SVM είναι ιδιαίτερα αποδοτικός σε υψηλοδιάστατα αραιά feature spaces όπως το TF-IDF [28], καθώς δεν απαιτεί τον υπολογισμό kernel matrix.

### 2.6.5 Random Forest

Το Random Forest [22] είναι ensemble μέθοδος που συνδυάζει $T$αποφασιστικά δέντρα $\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\}$, κάθε ένα εκ των οποίων εκπαιδεύεται σε τυχαίο υποσύνολο των δεδομένων (bootstrap sampling) και τυχαίο υποσύνολο χαρακτηριστικών. Η τελική απόφαση λαμβάνεται με πλειοψηφική ψηφοφορία:

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\}$$

Η τυχαιότητα στην επιλογή χαρακτηριστικών εισάγει **αποσυσχέτιση** μεταξύ των δέντρων, μειώνοντας τη διασπορά (variance) του συνολικού εκτιμητή σε σχέση με ένα μεμονωμένο δέντρο. Κάθε κόμβος διαχωρισμού επιλέγει το βέλτιστο χαρακτηριστικό από υποσύνολο μεγέθους $\sqrt{d}$, όπου $d$ο συνολικός αριθμός χαρακτηριστικών. Στην εργασία χρησιμοποιήθηκαν $T = 100$δέντρα.

### 2.6.6 Gradient Boosting

Το Gradient Boosting [23] κατασκευάζει το τελικό μοντέλο **διαδοχικά**, με κάθε νέο δέντρο $h_t$να εκπαιδεύεται στα υπόλοιπα (residuals) του προηγούμενου συνολικού μοντέλου. Το μοντέλο στο βήμα $t$ορίζεται ως:

$$F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \eta \cdot h_t(\mathbf{x})$$

όπου $\eta$είναι ο ρυθμός μάθησης (learning rate). Κάθε $h_t$εκπαιδεύεται στα αρνητικά gradients της συνάρτησης απώλειας:

$$r_{it} = -\left[\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F=F_{t-1}}$$

Για δυαδική ταξινόμηση με cross-entropy loss, τα gradients αυτά αντιστοιχούν στα σφάλματα πρόβλεψης $r_{it} = y_i - \sigma(F_{t-1}(\mathbf{x}_i))$. Η διαδοχική φύση του αλγορίθμου οδηγεί σε χαμηλό bias αλλά υψηλό υπολογιστικό κόστος σε σχέση με τα παράλληλα ensemble μοντέλα όπως το Random Forest.

### 2.6.7 DistilBERT και Twitter-RoBERTa

Και τα δύο μοντέλα Transformer [24] βασίζονται στον μηχανισμό **selfattention**. Για μια ακολουθία tokens $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, το attention [25] υπολογίζεται ως:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

όπου $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$, $\mathbf{V} = \mathbf{X}\mathbf{W}^V$είναι τα διανύσματα ερωτήματος (query), κλειδιού (key) και τιμής (value) αντίστοιχα, και $d_k$η διάσταση των κλειδιών που χρησιμεύει για κανονικοποίηση. Το multi-head attention επεκτείνει αυτή την ιδέα εκτελώντας $h$παράλληλα attention heads:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$$

Για την ταξινόμηση, η αναπαράσταση του ειδικού token [CLS] από την τελευταία κρυφή στρώση τροφοδοτείται σε ένα γραμμικό επίπεδο ταξινόμησης:

$$\hat{y} = \text{softmax}(\mathbf{W}_c \cdot \mathbf{h}_{[CLS]} + \mathbf{b}_c)$$

Κατά το fine-tuning, ολόκληρο το δίκτυο, συμπεριλαμβανομένων των προεκπαιδευμένων βαρών, ενημερώνεται μέσω backpropagation με optimizer AdamW και linear warmup scheduler, όπως περιγράφεται στο Κεφάλαιο 5.

# Κεφάλαιο 3: Σύνολο Δεδομένων

## 3.1 Περιγραφή και Προέλευση

Για την εκπαίδευση και αξιολόγηση των μοντέλων που παρουσιάζονται στην εργασία χρησιμοποιήθηκε το **Sentiment140**, ένα ευρέως αναγνωρισμένο σύνολο δεδομένων στο πεδίο της ανάλυσης συναισθήματος σε κείμενα κοινωνικής δικτύωσης. Το σύνολο δεδομένων δημιουργήθηκε από τους [3] στο πλαίσιο ερευνητικής εργασίας του Πανεπιστημίου Stanford, με σκοπό τη μελέτη αυτόματης κατηγοριοποίησης συναισθήματος σε αναρτήσεις της πλατφόρμας Twitter.

Το Sentiment140 αποτελείται συνολικά από **1.600.000 tweets**, τα οποία συλλέχθηκαν μέσω του δημόσιου API του Twitter κατά τη διάρκεια του έτους 2009. Βασικό χαρακτηριστικό της μεθοδολογίας συλλογής είναι η χρήση distant supervision), δηλαδή αντί να γίνει χειροκίνητη επισήμανση από ανθρώπους, η πολικότητα κάθε tweet αντλήθηκε αυτόματα από τη χρήση emoticons. Συγκεκριμένα, tweets που περιείχαν θετικά emoticons (π.χ. :), :-)) ταξινομήθηκαν ως θετικά, ενώ tweets με αρνητικά emoticons (π.χ. :(, :-() ταξινομήθηκαν ως αρνητικά. Τα emoticons αφαιρέθηκαν στη συνέχεια από το κείμενο, ώστε να μην αποτελούν ενδείξεις για τα μοντέλα κατά την εκπαίδευση.

Η προσέγγιση αυτή επιτρέπει τη δημιουργία μεγάλης κλίμακας συνόλων δεδομένων χωρίς την ανάγκη δαπανηρής χειροκίνητης επισήμανσης, αν και εισάγει έναν βαθμό θορύβου, καθώς η χρήση emoticons δεν αντικατοπτρίζει πάντα με ακρίβεια το συναίσθημα ολόκληρου του κειμένου.

## 3.2 Δομή και Χαρακτηριστικά

Το αρχείο είναι σε μορφή CSV και περιλαμβάνει έξι χαρακτηριστικά για κάθε εγγραφή:

| Χαρακτηριστικό | Περιγραφή |
|---|---|
| polarity | Η πολικότητα του tweet: 0 (αρνητικό), 2 (ουδέτερο), 4 (θετικό) |
| id | Μοναδικό αναγνωριστικό του tweet |
| date | Ημερομηνία και ώρα δημοσίευσης |
| query | Το ερώτημα αναζήτησης που χρησιμοποιήθηκε για τη συλλογή |
| user | Το όνομα χρήστη του δημιουργού |
| text | Το πλήρες κείμενο του tweet |

*Πίνακας 2: Χαρακτηριστικά συνόλου δεδομένων*

Στο σύνολο δεδομένων απουσιάζουν παντελώς ουδέτερα tweets (polarity = 2), με αποτέλεσμα να πρόκειται ουσιαστικά για ένα δυαδικό πρόβλημα ταξινόμησης με δύο κλάσεις: αρνητική (polarity = 0) και θετική (polarity = 4). Κατά την προεπεξεργασία, η τιμή 4 αντιστοιχίστηκε στο 1, ώστε οι ετικέτες να ακολουθούν τη συμβατική δυαδική κωδικοποίηση {0, 1}.

Το σύνολο είναι ισορροπημένο: περιλαμβάνει ακριβώς 800.000 θετικά και 800.000 αρνητικά tweets, γεγονός που εξαλείφει την ανάγκη εφαρμογής τεχνικών αντιστάθμισης κλάσεων (class imbalance handling) κατά την εκπαίδευση [30].

## 3.3 Δειγματοληψία

Λόγω των υπολογιστικών περιορισμών που συνδέονται με την εκπαίδευση μοντέλων βαθιάς μάθησης, και ιδίως των αρχιτεκτονικών τύπου Transformer, δεν χρησιμοποιήθηκε το πλήρες σύνολο των 1,6 εκατομμυρίων tweets. Αντ' αυτού, εφαρμόστηκε **στρωματοποιημένη δειγματοληψία** (stratified sampling) [31] με σταθερό seed, η οποία έδωσε ισάριθμα δείγματα από κάθε κλάση, διατηρώντας έτσι την ισορροπία του αρχικού συνόλου.

Το τελικό υποσύνολο που χρησιμοποιήθηκε αποτελείται από **20.000 tweets** συνολικά, 10.000 θετικά και 10.000 αρνητικά. Η επιλογή αυτού του μεγέθους κρίθηκε ότι εξισορροπεί επαρκώς

την αντιπροσωπευτικότητα του δείγματος με την υπολογιστική αποδοτικότητα, ενώ παράλληλα είναι αρκετά μεγάλο ώστε να επιτρέπει αξιόπιστη σύγκριση μεταξύ κλασικών και νευρωνικών μεθόδων [38].

## 3.4 Διαχωρισμός Συνόλου Δεδομένων

Το σύνολο των 200.000 δειγμάτων διαχωρίστηκε σε τρία επιμέρους υποσύνολα, σύμφωνα με την κλασική πρακτική:

- **Σύνολο εκπαίδευσης (Training set):** 80% — 16.000 tweets
- **Σύνολο επικύρωσης (Validation set):** 10% — 2.000 tweets
- **Σύνολο ελέγχου (Test set):** 10% — 2.000 tweets

Ο διαχωρισμός πραγματοποιήθηκε με **στρωματοποίηση ως προς το label**, εξασφαλίζοντας ότι και τα τρία υποσύνολα διατηρούν την ίδια αναλογία θετικών και αρνητικών δειγμάτων (50%50%). Το σύνολο επικύρωσης χρησιμοποιήθηκε για τον έλεγχο της απόδοσης κατά τη διάρκεια της εκπαίδευσης των μοντέλων Transformer και για την έγκαιρη διακοπή (early stopping), ενώ το σύνολο ελέγχου χρησιμοποιήθηκε αποκλειστικά για την τελική αξιολόγηση όλων των μοντέλων, εξασφαλίζοντας αντικειμενική και αμερόληπτη σύγκριση.

## 3.5 Χαρακτηριστικά του Κειμένου

Τα tweets ως μορφή κειμένου παρουσιάζουν μια σειρά από ιδιαιτερότητες που τα διαφοροποιούν από τυπικά γραπτά κείμενα και επηρεάζουν σημαντικά τις επιλογές προεπεξεργασίας. Συγκεκριμένα:

**Περιορισμός χαρακτήρων.** Κατά την περίοδο συλλογής του Sentiment140, τα tweets περιορίζονταν σε 140 χαρακτήρες, επιβάλλοντας συνοπτικές και συχνά ελλειπτικές διατυπώσεις. Στο δείγμα που χρησιμοποιήθηκε, το μέσο μήκος tweet ανέρχεται περίπου στους **73 χαρακτήρες**, ενώ ο μέσος αριθμός λέξεων μετά την προεπεξεργασία είναι περίπου **7-8 λέξεις** ανά tweet.

**Ανεπίσημη γλώσσα.** Τα tweets περιέχουν συχνά αργκό, συντομογραφίες, ορθογραφικά λάθη και νεολογισμούς που δεν απαντώνται σε τυπικά γλωσσικά σώματα (corpora).

**Δομικά στοιχεία.** Τα tweets συχνά περιλαμβάνουν: αναφορές χρηστών (mentions, π.χ. @user), hashtags (π.χ. #topic), υπερσυνδέσμους (URLs) και emoticons. Όλα αυτά τα στοιχεία αντιμετωπίστηκαν ειδικά κατά τη φάση προεπεξεργασίας, όπως περιγράφεται στο επόμενο κεφάλαιο.

**Θόρυβος ετικετών.** Καθώς η επισήμανση βασίστηκε σε απομακρυσμένη επίβλεψη, ένα ποσοστό των tweets ενδέχεται να φέρει λανθασμένη ετικέτα, για παράδειγμα, ένα ironic tweet με θετικό emoticon που εκφράζει στην πραγματικότητα αρνητικό συναίσθημα. Αυτός ο θόρυβος είναι εγγενής στη μεθοδολογία συλλογής και αποτελεί γνωστό περιορισμό του συνόλου δεδομένων.

## 3.6 Επιλογή του Συνόλου Δεδομένων

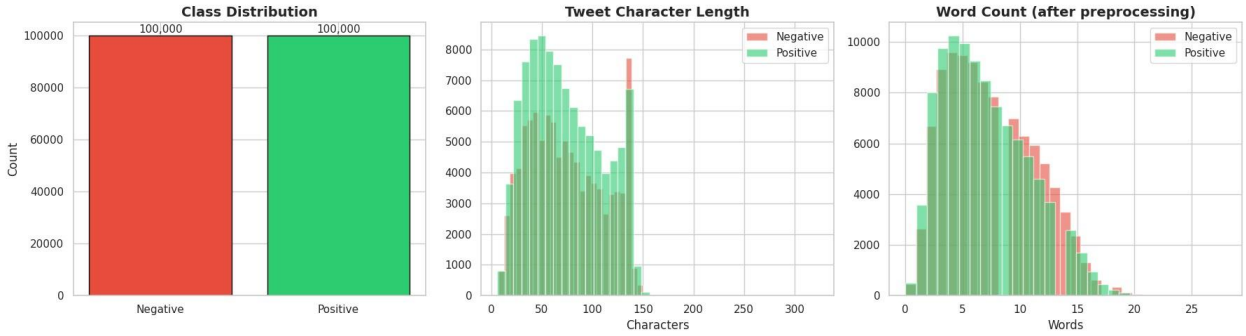
Η επιλογή του Sentiment140 ως βάση της εργασίας υπαγορεύτηκε από μια σειρά παραγόντων. Πρώτον, πρόκειται για ένα benchmark dataset που χρησιμοποιείται ευρέως στη βιβλιογραφία, επιτρέποντας την άμεση σύγκριση αποτελεσμάτων με προηγούμενες μελέτες. Δεύτερον, το μεγάλο του μέγεθος επιτρέπει την εκπαίδευση μοντέλων βαθιάς μάθησης υψηλής χωρητικότητας χωρίς κίνδυνο υπερπροσαρμογής λόγω έλλειψης δεδομένων. Τρίτον, η ισορροπία κλάσεων απλοποιεί τη διαδικασία αξιολόγησης, καθώς μετρικές όπως η ακρίβεια δεν επηρεάζονται από ανισορροπία. Τέλος, το γεγονός ότι τα δεδομένα προέρχονται από πραγματικούς χρήστες της πλατφόρμας Twitter προσδίδει στο σύνολο ρεαλιστικότητα και οικολογική εγκυρότητα.

## 3.7 Διερευνητική Ανάλυση Δεδομένων

Πριν από την εφαρμογή οποιασδήποτε τεχνικής προεπεξεργασίας ή εκπαίδευσης μοντέλου, πραγματοποιείται διερευνητική ανάλυση [37] του συνόλου δεδομένων με σκοπό την κατανόηση της κατανομής των κλάσεων, των στατιστικών χαρακτηριστικών του κειμένου και της λεξιλογικής σύνθεσης κάθε κατηγορίας συναισθήματος. Τα αποτελέσματα αυτής της ανάλυσης τεκμηριώνουν τις επιλογές προεπεξεργασίας που ακολουθούν και παρέχουν χρήσιμες ενδείξεις για τη φύση του προβλήματος.

### 3.7.1 Κατανομή Κλάσεων

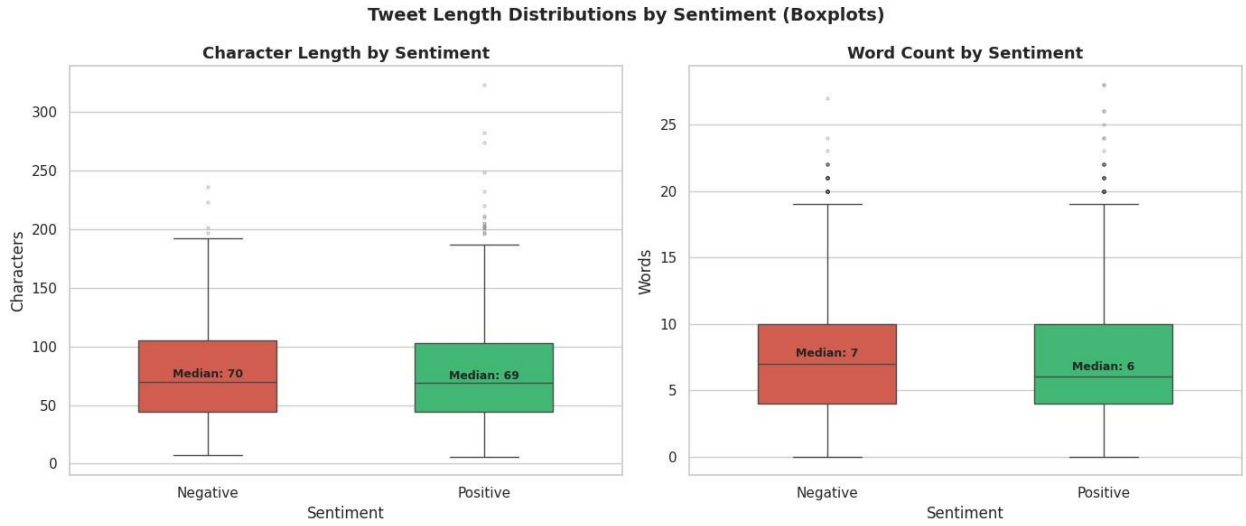
Όπως αναφέρθηκε στην ενότητα 3.3, το υποσύνολο που χρησιμοποιήθηκε αποτελείται από 100.000 θετικά και 100.000 αρνητικά tweets, με αποτέλεσμα απόλυτα ισορροπημένη κατανομή κλάσεων (50%-50%). Η ισορροπία αυτή επαληθεύτηκε οπτικά μέσω ραβδογράμματος κατανομής και αποτελεί εγγύηση ότι τα μοντέλα δεν θα αναπτύξουν μεροληψία υπέρ της πλειοψηφικής κλάσης.




*Εικόνα 1: Στατιστικά κλάσεων*

## 3.7.2 Στατιστικά Χαρακτηριστικά Μήκους Tweet

Εξετάστηκε το μήκος των tweets τόσο σε επίπεδο χαρακτήρων (πριν την προεπεξεργασία) όσο και σε επίπεδο αριθμού λέξεων (μετά την προεπεξεργασία). Από την ανάλυση προέκυψε ότι τα δύο σύνολα (θετικά και αρνητικά) παρουσιάζουν παρόμοιες κατανομές μήκους, γεγονός που υποδηλώνει ότι το μήκος του κειμένου από μόνο του δεν αποτελεί διακριτικό χαρακτηριστικό μεταξύ των δύο κλάσεων. Τα boxplots ανά κλάση επιβεβαιώνουν αυτή την παρατήρηση, αποκαλύπτοντας παρόμοιες διάμεσες τιμές και εύρος διασποράς και για τις δύο κατηγορίες. Η κατανομή του μήκους είναι ελαφρώς ασύμμετρη προς τα δεξιά και στις δύο κλάσεις, κάτι αναμενόμενο δεδομένου του ορίου των 140 χαρακτήρων.

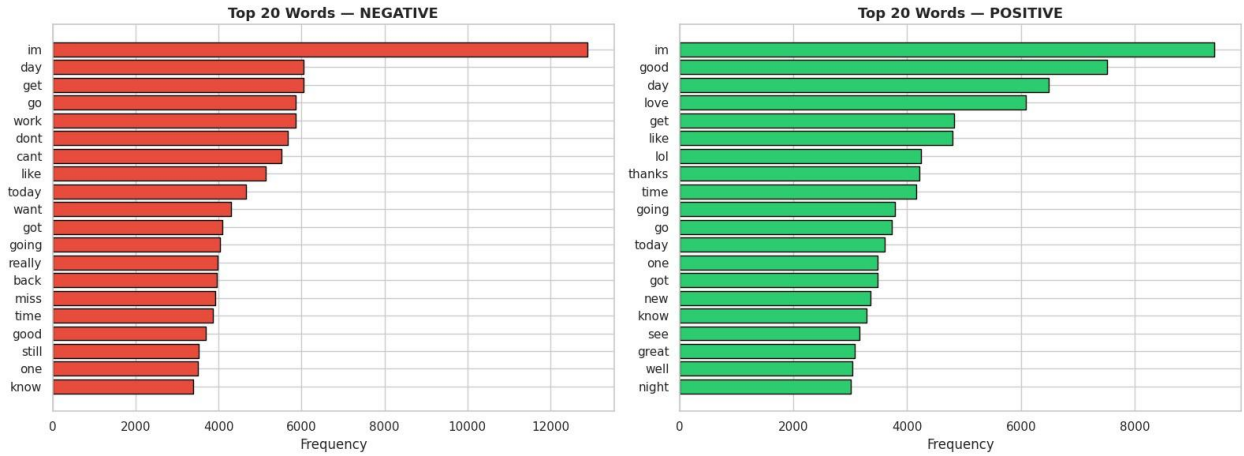



*Εικόνα 2: Μήκος χαρακτήρων και λέξεων ανα συναίσθημα*

## 3.7.3 Συχνότητα Λέξεων ανά Κλάση

Αναλύθηκαν οι 20 πιο συχνές λέξεις για κάθε κατηγορία συναισθήματος, μετά την εφαρμογή προεπεξεργασίας (αφαίρεση stopwords, lemmatization). Η ανάλυση αποκαλύπτει χαρακτηριστικά λεξιλογικά μοτίβα: τα αρνητικά tweets τείνουν να περιέχουν λέξεις που

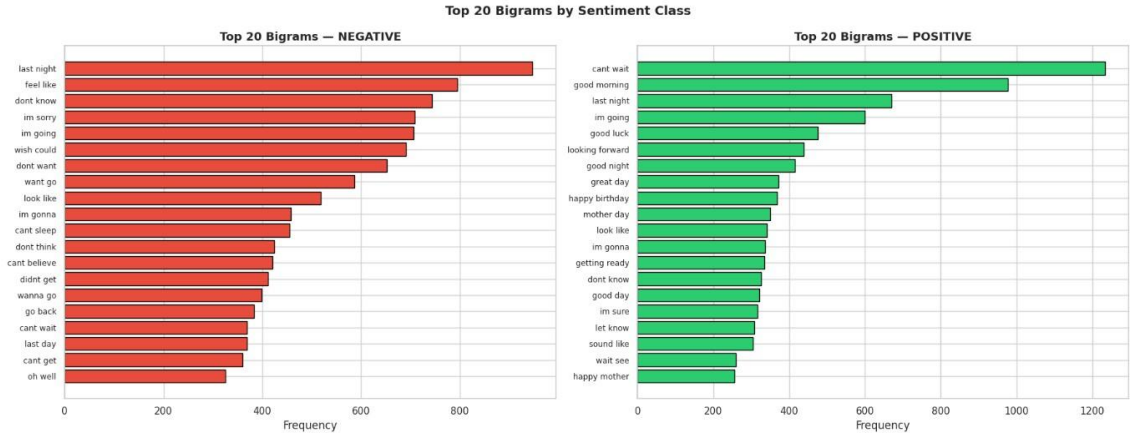
εκφράζουν αδυναμία, απογοήτευση ή δυσφορία, ενώ τα θετικά tweets χαρακτηρίζονται από λέξεις που σχετίζονται με χαρά, ευχαρίστηση και κοινωνική αλληλεπίδραση. Το εύρημα αυτό επιβεβαιώνει ότι το λεξιλόγιο φέρει σημαντική πληροφορία για το συναίσθημα και δικαιολογεί τη χρήση bag-of-words αναπαραστάσεων (TF-IDF) ως βάση για τα κλασικά μοντέλα.




*Εικόνα 3: Οι 20 συχνότερες λέξεις ανα συναίσθημα*

## 3.7.4 Ανάλυση Bigrams

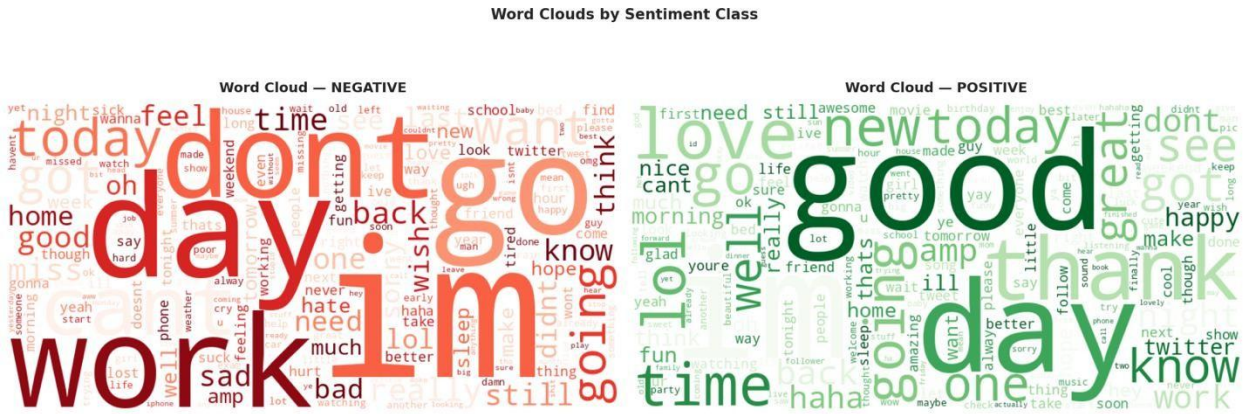
Επιπλέον της ανάλυσης μονόγραμμων (unigrams) [39], εξετάστηκαν τα 20 πιο συχνά **bigrams** (ζεύγη διαδοχικών λέξεων) [34] ανά κλάση. Η ανάλυση bigrams παρέχει πλουσιότερη σημασιολογική πληροφορία σε σχέση με μεμονωμένες λέξεις, καθώς αναδεικνύει φράσεις και συνδυασμούς που χαρακτηρίζουν το κάθε συναίσθημα. Τα αποτελέσματα υπογραμμίζουν ότι η συμπερίληψη bigrams στον TF-IDF vectorizer αποτελεί τεκμηριωμένη επιλογή για τη βελτίωση της αναπαραστατικής ικανότητας των κλασικών μοντέλων.




*Εικόνα 4: Οι 20 συχνότερες λέξεις-bigrams ανα συναίσθημα*

### 3.7.5 Word Clouds

Για την οπτική απεικόνιση της λεξιλογικής σύνθεσης κάθε κλάσης, δημιουργήθηκαν **word clouds** [32] για τα θετικά και αρνητικά tweets αντίστοιχα. Στα word clouds, το μέγεθος κάθε λέξης είναι ανάλογο της συχνότητάς της στην αντίστοιχη κλάση. Η οπτικοποίηση αυτή επιβεβαιώνει τα ευρήματα της ποσοτικής ανάλυσης συχνότητας λέξεων και παρέχει μια διαισθητική κατανόηση του γλωσσικού προφίλ κάθε κατηγορίας συναισθήματος.




*Εικόνα 5: Word clouds ανα συναίσθημα*

### 3.7.6 Συμπεράσματα Διερευνητικής Ανάλυσης

Συνολικά, η διερευνητική ανάλυση κατέδειξε ότι: (α) το σύνολο δεδομένων είναι πλήρως ισορροπημένο και κατάλληλο για απευθείας εφαρμογή χωρίς τεχνικές αντιστάθμισης, (β) τα δύο σύνολα παρουσιάζουν συγκρίσιμα στατιστικά χαρακτηριστικά μήκους, άρα το μήκος δεν αποτελεί από μόνο του χρήσιμο χαρακτηριστικό, και (γ) το λεξιλόγιο φέρει σαφή διακριτική πληροφορία μεταξύ των κλάσεων, δικαιολογώντας τόσο τη χρήση TF-IDF για τα κλασικά μοντέλα όσο και την ανάγκη για βαθύτερη σημασιολογική κατανόηση που παρέχουν τα μοντέλα Transformer.

# Κεφάλαιο 4: Προεπεξεργασία συνόλου δεδομένων

## 4.1 Εισαγωγή

Η προεπεξεργασία κειμένου αποτελεί κρίσιμο στάδιο σε οποιοδήποτε σύστημα επεξεργασίας φυσικής γλώσσας, και ιδιαίτερα σε εφαρμογές ανάλυσης κειμένου από κοινωνικά δίκτυα, όπου το κείμενο είναι εκ φύσεως θορυβώδες και ανομοιόμορφο.

Στην εργασία εφαρμόστηκε μια σειρά διαδοχικών βημάτων καθαρισμού και κανονικοποίησης, τα οποία υλοποιήθηκαν μέσω Python. Αξίζει να σημειωθεί ότι η προεπεξεργασία αυτή εφαρμόστηκε **αποκλειστικά για τα κλασικά μοντέλα μηχανικής μάθησης** (Logistic Regression, Naive Bayes, SVM, Random Forest, Gradient Boosting), τα οποία βασίζονται σε TF-IDF αναπαραστάσεις. Αντίθετα, τα μοντέλα Transformer (DistilBERT, Twitter-RoBERTa) εκπαιδεύτηκαν στο **ακατέργαστο κείμενο**, καθώς διαθέτουν τους δικούς τους tokenizers που έχουν σχεδιαστεί να χειρίζονται τέτοια στοιχεία εγγενώς.

## 4.2 Βήματα Προεπεξεργασίας

### 4.2.1 Μετατροπή σε Πεζά

Το πρώτο βήμα είναι η μετατροπή ολόκληρου του κειμένου σε πεζούς χαρακτήρες. Η κανονικοποίηση αυτή εξασφαλίζει ότι λέξεις όπως "Good", "GOOD" και "good" αντιμετωπίζονται ως η ίδια οντότητα. Χωρίς αυτό το βήμα, ένας TF-IDF vectorizer θα δημιουργούσε ξεχωριστές καταχωρήσεις για κάθε παραλλαγή, φουσκώνοντας τεχνητά το λεξιλόγιο.

### 4.2.2 Αφαίρεση URLs

Οι υπερσύνδεσμοι είναι εξαιρετικά συχνοί στα tweets και δεν φέρουν χρήσιμη σημασιολογική πληροφορία για το συναίσθημα. Η αφαίρεσή τους αποτρέπει την κατάληψη θέσεων στο λεξιλόγιο από αδιάφορες για το πρόβλημα συμβολοσειρές.

### 4.2.3 Αφαίρεση Αναφορών Χρηστών

Οι αναφορές σε χρήστες (π.χ. @nasa, @elonmusk) δεν σχετίζονται με το συναίσθημα του κειμένου και αφαιρούνται πλήρως. Η παρουσία τους θα εισήγαγε θόρυβο, καθώς το ίδιο όνομα χρήστη μπορεί να εμφανίζεται τόσο σε θετικά όσο και σε αρνητικά tweets.

### 4.2.4 Επεξεργασία Hashtags

Αντί να αφαιρεθούν εντελώς τα hashtags, επιλέχθηκε η διατήρηση της λέξης που τα ακολουθεί με αφαίρεση μόνο του συμβόλου #. Για παράδειγμα, το #happy μετατρέπεται σε happy. Αυτή η επιλογή είναι σκόπιμη: τα hashtags στο Twitter συχνά αποτελούν ισχυρούς δείκτες συναισθήματος και η πληροφορία τους αξίζει να διατηρηθεί.

### 4.2.5 Αφαίρεση μη-ASCII Χαρακτήρων και Emoji

Οι χαρακτήρες εκτός του ASCII εύρους, συμπεριλαμβανομένων των emoji, αντικαθίστανται με κενό διάστημα. Η απόφαση αυτή υπαγορεύεται από το γεγονός ότι το Sentiment140 έχει ήδη αφαιρέσει τα emoticons που χρησιμοποιήθηκαν για την επισήμανση, οπότε τα υπόλοιπα μηASCII σύμβολα αποτελούν θόρυβο. Παράλληλα, οι TF-IDF αναπαραστάσεις δεν μπορούν να αξιοποιήσουν emoji με τον ίδιο τρόπο που το κάνουν εξειδικευμένα μοντέλα.

### 4.2.6 Αφαίρεση Σημείων Στίξης

Τα σημεία στίξης αφαιρούνται στο σύνολό τους. Αν και σε ορισμένα προβλήματα NLP η στίξη μπορεί να φέρει πληροφορία (π.χ. θαυμαστικό για έντονο συναίσθημα), στο πλαίσιο των TF-IDF αναπαραστάσεων η συμβολή της είναι περιορισμένη και η παρουσία της αυξάνει την πολυπλοκότητα του λεξιλογίου.

### 4.2.7 Tokenization, Αφαίρεση Stopwords και Φιλτράρισμα

Το κείμενο διαχωρίζεται σε tokens [36] με απλό whitespace splitting. Στη συνέχεια εφαρμόζονται δύο φίλτρα: αφαιρούνται τα **stopwords** της αγγλικής γλώσσας (όπως "the", "is", "at") μέσω της λίστας του NLTK, και αποκλείονται λέξεις μήκους ενός χαρακτήρα (len(t) > 1), οι οποίες συνήθως δεν φέρουν σημασιολογική πληροφορία. Τα stopwords, αν και συχνά στο κείμενο, δεν διαφοροποιούνται μεταξύ των κλάσεων και η αφαίρεσή τους μειώνει σημαντικά τον όγκο του feature space.

### 4.2.8 Lemmatization

Το τελευταίο βήμα είναι η **λημματοποίηση** (lemmatization) [35] κάθε token μέσω του WordNetLemmatizer του NLTK, ο οποίος βασίζεται στο λεξικό WordNet. Η λημματοποίηση ανάγει κάθε λέξη στη βασική της μορφή (λήμμα): για παράδειγμα, "running" → "run", "better" → "good". Σε αντίθεση με το stemming, η λημματοποίηση παράγει πραγματικές λέξεις, διατηρώντας την αναγνωσιμότητα του κειμένου και αποφεύγοντας αποκομμένες ρίζες που μπορεί να μην αντιστοιχούν σε καμία πραγματική έννοια.

# Κεφάλαιο 5: Μοντέλα και Μεθοδολογία

## 5.1 Επισκόπηση

Για την αντιμετώπιση του προβλήματος ανάλυσης συναισθήματος εξετάστηκαν δύο κατηγορίες μοντέλων: κλασικά μοντέλα μηχανικής μάθησης που βασίζονται σε στατιστικές αναπαραστάσεις κειμένου, και μοντέλα βασισμένα σε αρχιτεκτονικές Transformer που αξιοποιούν βαθιά αναπαράσταση της γλώσσας. Η επιλογή δύο διαφορετικών κατηγοριών επιτρέπει άμεση σύγκριση της αποτελεσματικότητάς τους στο ίδιο σύνολο δεδομένων.

## 5.2 Κλασικά Μοντέλα Μηχανικής Μάθησης

### Αναπαράσταση Κειμένου — TF-IDF

Τα κλασικά μοντέλα δεν μπορούν να επεξεργαστούν κείμενο άμεσα και απαιτούν αριθμητική αναπαράσταση. Για τον σκοπό αυτό χρησιμοποιήθηκε ο **TF-IDF** (Term Frequency–Inverse Document Frequency) vectorizer, ο οποίος αποδίδει σε κάθε λέξη μια τιμή ανάλογη της συχνότητάς της στο έγγραφο, σταθμισμένη αντίστροφα με τη συχνότητά της στο σύνολο των εγγράφων. Με τον τρόπο αυτό, λέξεις που εμφανίζονται παντού (και άρα δεν φέρουν διακριτική πληροφορία) λαμβάνουν χαμηλές τιμές, ενώ λέξεις που είναι χαρακτηριστικές συγκεκριμένων εγγράφων λαμβάνουν υψηλές. Στην εργασία ο vectorizer διαμορφώθηκε ώστε να λαμβάνει υπόψη unigrams και bigrams (ngram_range=(1,2)), με μέγιστο αριθμό χαρακτηριστικών 50.000 (max_features=50_000) και εφαρμογή λογαριθμικής κλιμάκωσης (sublinear_tf=True).

Στη συνέχεια εκπαιδεύτηκαν πέντε κλασικά μοντέλα:

**Logistic Regression.** Γραμμικό μοντέλο ταξινόμησης που εκτιμά την πιθανότητα ανήκειν σε μία κλάση μέσω της λογιστικής συνάρτησης. Αποτελεί ισχυρό baseline για προβλήματα NLP λόγω της αποτελεσματικότητάς του σε υψηλοδιάστατα αραιά feature spaces όπως το TF-IDF.

**Naive Bayes.** Πιθανοτικός ταξινομητής που βασίζεται στο θεώρημα Bayes υπό την υπόθεση ανεξαρτησίας των χαρακτηριστικών. Παρά την απλοϊκή αυτή υπόθεση, αποδίδει ιδιαίτερα καλά σε προβλήματα ταξινόμησης κειμένου και είναι υπολογιστικά αποδοτικός.

**Linear SVM.** Ο γραμμικός Support Vector Machine αναζητά το υπερεπίπεδο που μεγιστοποιεί το περιθώριο διαχωρισμού μεταξύ των κλάσεων. Θεωρείται από τους πιο αξιόπιστους ταξινομητές για υψηλοδιάστατα προβλήματα κειμένου.

**Random Forest.** Σύνολο (ensemble) αποφασιστικών δέντρων που εκπαιδεύονται σε τυχαία υποσύνολα των δεδομένων και των χαρακτηριστικών. Η τελική απόφαση λαμβάνεται με πλειοψηφική ψηφοφορία, μειώνοντας την υπερπροσαρμογή σε σχέση με ένα μεμονωμένο δέντρο.

**Gradient Boosting.** Ένα ακόμη ensemble μοντέλο, στο οποίο τα δέντρα εκπαιδεύονται διαδοχικά, με κάθε νέο δέντρο να διορθώνει τα σφάλματα του προηγούμενου. Αποδίδει συνήθως υψηλή ακρίβεια αλλά απαιτεί σημαντικά μεγαλύτερο χρόνο εκπαίδευσης.

## 5.3 Μοντέλα Transformer

### DistilBERT

Το **DistilBERT** είναι μια συμπυκνωμένη έκδοση του BERT (Bidirectional Encoder Representations from Transformers) που διατηρεί το 97% της απόδοσής του με 40% λιγότερες παραμέτρους και 60% ταχύτερη εκπαίδευση. Χρησιμοποιήθηκε η προεκπαιδευμένη έκδοση distilbert-base-uncased, η οποία στη συνέχεια υποβλήθηκε σε fine-tuning στο σύνολο εκπαίδευσης για 2 εποχές με learning rate 2e-5 και μέγιστο μήκος ακολουθίας 128 tokens.

### Twitter-RoBERTa

Το **Twitter-RoBERTa** (cardiffnlp/twitter-roberta-base-sentiment) είναι ένα μοντέλο βασισμένο στην αρχιτεκτονική RoBERTa, το οποίο έχει προεκπαιδευτεί σε 58 εκατομμύρια tweets, καθιστώντας το ιδιαίτερα κατάλληλο για ανάλυση κειμένου από το Twitter. Αρχικά σχεδιασμένο για τριμερή ταξινόμηση (αρνητικό/ουδέτερο/θετικό), το classification head επανεκκινήθηκε για δυαδική ταξινόμηση (ignore_mismatched_sizes=True) και το μοντέλο υποβλήθηκε σε fine-tuning υπό τις ίδιες συνθήκες με το DistilBERT.

## 5.4 Αξιολόγηση Μοντέλων

Για την αξιολόγηση όλων των μοντέλων χρησιμοποιήθηκαν οι παρακάτω μετρικές, οι οποίες εφαρμόστηκαν αποκλειστικά στο σύνολο ελέγχου (test set): Accuracy, Precision, Recall, F1-score και ROC-AUC. Η πλήρης ανάλυση των αποτελεσμάτων παρουσιάζεται στο Κεφάλαιο 6.

# Κεφάλαιο 6: Αποτελέσματα

## 6.1 Επισκόπηση Αξιολόγησης

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της αξιολόγησης όλων των μοντέλων που εκπαιδεύτηκαν στο πλαίσιο της εργασίας. Η αξιολόγηση πραγματοποιήθηκε αποκλειστικά στο σύνολο ελέγχου (test set), το οποίο αποτελείται από 2.000 tweets (18.000 για εκπαίδευση) που δεν συμμετείχαν σε καμία φάση εκπαίδευσης ή επικύρωσης. Για κάθε μοντέλο υπολογίστηκαν πέντε μετρικές: Accuracy, Precision, Recall, F1-score και ROC-AUC. Η σύγκριση καλύπτει συνολικά επτά μοντέλα: πέντε κλασικά μοντέλα μηχανικής μάθησης (Logistic Regression, Naive Bayes, Linear SVM, Random Forest, Gradient Boosting) και δύο μοντέλα Transformer (DistilBERT, Twitter-RoBERTa).

## 6.2 Αποτελέσματα Κλασικών Μοντέλων

Ο επόμενος πίνακας παρουσιάζει τις επιδόσεις των πέντε κλασικών μοντέλων στο σύνολο ελέγχου.

| Μοντέλο | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| **Logistic Regression** | **0.776850** | **0.768552** | 0.792300 | **0.780245** | **0.858596** |

| | | | | | |
|---|---|---|---|---|---|
| **Naive Bayes** | 0.760550 | 0.765083 | 0.752000 | 0.758485 | 0.845420 |
| **Linear SVM** | 0.762800 | 0.758103 | 0.771900 | 0.764939 | 0.842145 |
| **Random Forest** | 0.764100 | 0.766069 | 0.760400 | 0.763224 | 0.839376 |
| **Gradient Boosting** | 0.684150 | 0.637047 | **0.856000** | 0.730469 | 0.754787 |

*Πίνακας 3: Επιδόσεις των πέντε κλασικών μοντέλων στο σύνολο ελέγχου*

Μεταξύ των κλασικών μοντέλων, η Logistic Regression επιτυγχάνει την υψηλότερη συνολική απόδοση με F1-score 0.780 και ROC-AUC 0.859, επιβεβαιώνοντας τη βιβλιογραφική παρατήρηση ότι τα γραμμικά μοντέλα αποδίδουν ιδιαίτερα καλά σε υψηλοδιάστατα αραιά feature spaces όπως το TF-IDF. Το Naive Bayes και το Linear SVM παρουσιάζουν παρόμοια, ελαφρώς χαμηλότερη απόδοση (F1: 0.759 και 0.765 αντίστοιχα), ενώ το Random Forest κινείται σε παρόμοια επίπεδα (F1: 0.763).

Αξιοσημείωτη εξαίρεση αποτελεί το Gradient Boosting, το οποίο εμφανίζει ιδιαίτερα υψηλό Recall (0.856) σε συνδυασμό με χαμηλό Precision (0.637), με αποτέλεσμα το χαμηλότερο F1score (0.731) και Accuracy (0.684) μεταξύ όλων των μοντέλων. Αυτή η ανισορροπία υποδηλώνει ότι το μοντέλο τείνει να ταξινομεί υπερβολικά μεγάλο ποσοστό των tweets ως θετικά, ελαχιστοποιώντας τα False Negatives εις βάρος των False Positives. Η συμπεριφορά αυτή πιθανώς οφείλεται στην ευαισθησία του Gradient Boosting στην κατανομή του feature space του TF-IDF σε συνδυασμό με τις προεπιλεγμένες υπερπαραμέτρους που χρησιμοποιήθηκαν.

## 6.3 Αποτελέσματα Μοντέλων Transformer

### 6.3.1 DistilBERT

Ο επόμενος πίνακας παρουσιάζει τα αποτελέσματα του DistilBERT στο σύνολο ελέγχου μετά από fine-tuning 2 εποχών.

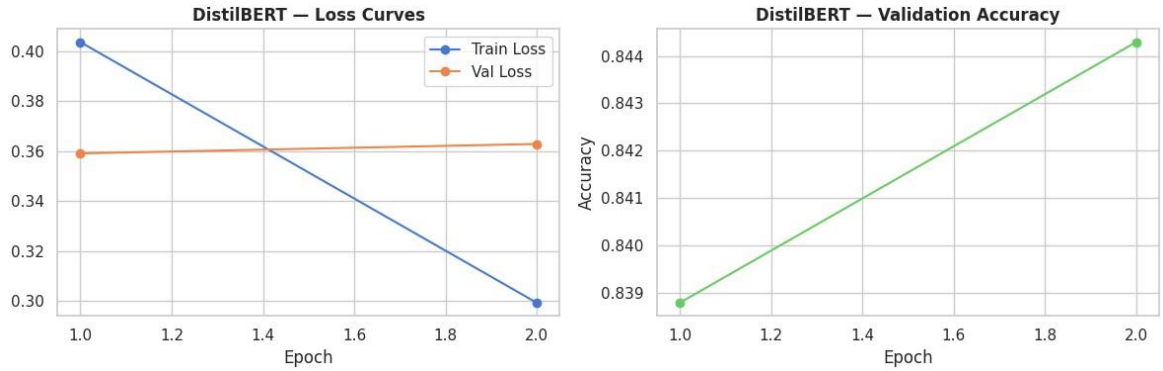
| Μετρική | Τιμή |
|---|---|
| **Accuracy** | 0.846150 |
| **Precision** | 0.849470 |
| **Recall** | 0.841400 |
| **F1-Score** | 0.845416 |
| **ROC-AUC** | 0.925107 |

*Πίνακας 4: Αποτελέσματα του DistilBERT στο σύνολο ελέγχου*

Το DistilBERT επιτυγχάνει Accuracy 84.6% και F1-score 0.845, υπερέχοντας σημαντικά έναντι όλων των κλασικών μοντέλων. Η βελτίωση σε F1-score σε σχέση με το καλύτερο κλασικό μοντέλο

(Logistic Regression: 0.780) ανέρχεται σε **+6.5 ποσοστιαίες μονάδες**, ενώ η βελτίωση σε ROC-AUC είναι ακόμη πιο εντυπωσιακή (+6.6 μονάδες, από 0.859 σε 0.925). Αξιοσημείωτη είναι επίσης η ισορροπία μεταξύ Precision (0.849) και Recall (0.841), που υποδηλώνει ότι το μοντέλο δεν παρουσιάζει συστηματική μεροληψία υπέρ κάποιας κλάσης.

Η επόμενη εικόνα απεικονίζει τις καμπύλες εκπαίδευσης (learning curves) του DistilBERT για τις δύο εποχές, τόσο για το training loss όσο και για το validation loss, καθώς και την εξέλιξη της ακρίβειας στο σύνολο επικύρωσης.




*Εικόνα 6: DistilBERT καμπύλες εκπαίδευσης*

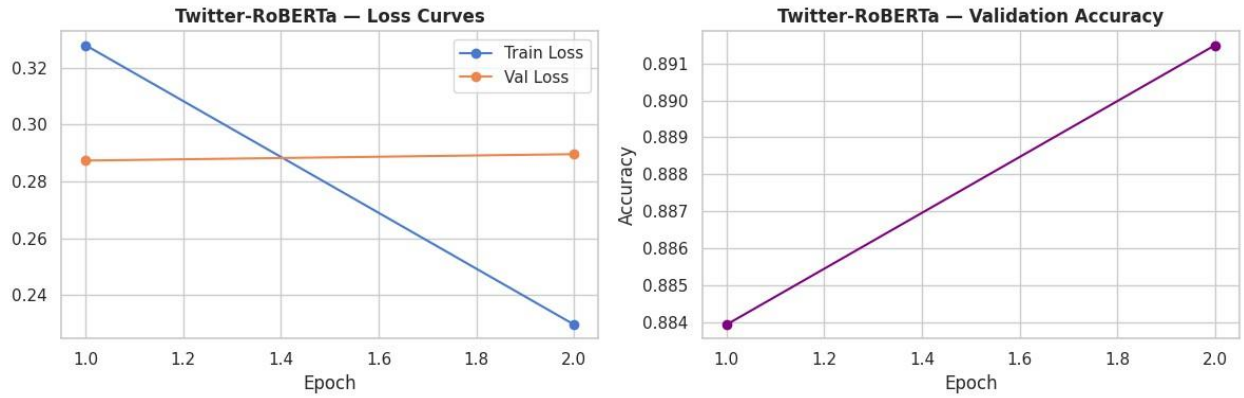
Παρατηρείται ότι το validation loss παρουσιάζει σημεία σταθεροποίησης]κατά τη διάρκεια της εκπαίδευσης, γεγονός που υποδηλώνει ότι οι 2 εποχές είναι επαρκείς για το συγκεκριμένο σύνολο δεδομένων. Η validation accuracy διαμορφώθηκε σε 0.844 στο τέλος της δεύτερης εποχής.

### 6.3.2 Twitter-RoBERTa

Ο επόμενος πίνακας παρουσιάζει τα αποτελέσματα του Twitter-RoBERTa στο σύνολο ελέγχου.

| Μετρική | Τιμή |
|---|---|
| **Accuracy** | **0.882400** |
| **Precision** | **0.887908** |
| **Recall** | **0.875300** |
| **F1-Score** | **0.881559** |
| **ROC-AUC** | **0.952800** |

*Πίνακας 5: Αποτελέσματα του Twitter-RoBERTa στο σύνολο ελέγχου*




*Εικόνα 7: Twitter-RoBERTa καμπύλες εκπαίδευσης*

Το Twitter-RoBERTa αναδεικνύεται ως το κορυφαίο μοντέλο της εργασίας, επιτυγχάνοντας Accuracy 88.2%, F1-score 0.882 και ROC-AUC 0.953 — την υψηλότερη τιμή σε κάθε μετρική. Η υπεροχή του έναντι του DistilBERT (+3.6 μονάδες F1, +2.8 μονάδες ROC-AUC) αποδίδεται στην εξειδικευμένη προεκπαίδευσή του σε 58 εκατομμύρια tweets, η οποία του επιτρέπει να κατανοεί εγγενώς τη γλώσσα του Twitter, συμπεριλαμβανομένων συντομογραφιών, slang και hashtags, χωρίς να απαιτείται εκτεταμένη προεπεξεργασία.

## 6.4 Συνολική Σύγκριση Μοντέλων

Ο επόμενος πίνακας συγκεντρώνει τα αποτελέσματα όλων των μοντέλων για άμεση σύγκριση. Με έντονη γραφή επισημαίνεται η καλύτερη τιμή ανά μετρική.

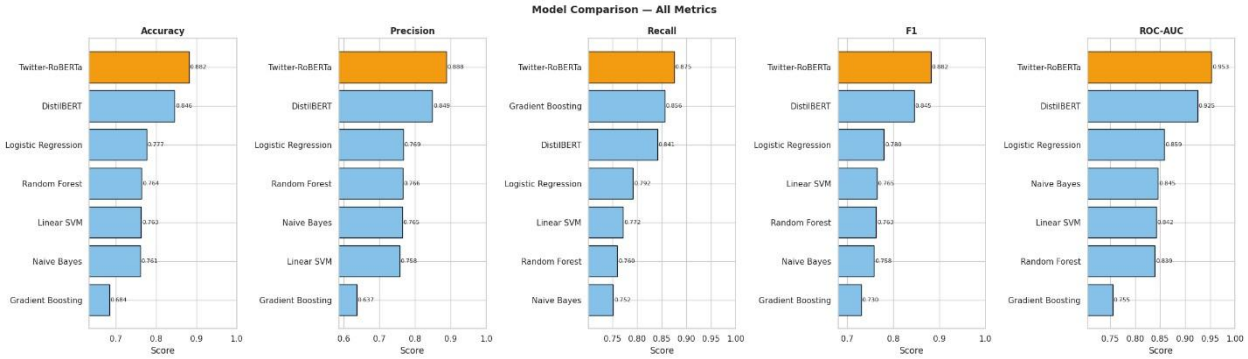
Στον επόμενο πίνακα γίνεται η σύγκριση όλων των μοντέλων:

| Μοντέλο | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.776850 | 0.768552 | 0.792300 | 0.780245 | 0.858596 |
| **Naive Bayes** | 0.760550 | 0.765083 | 0.752000 | 0.758485 | 0.845420 |
| **Linear SVM** | 0.762800 | 0.758103 | 0.771900 | 0.764939 | 0.842145 |
| **Random Forest** | 0.764100 | 0.766069 | 0.760400 | 0.763224 | 0.839376 |
| **Gradient Boosting** | 0.684150 | 0.637047 | 0.856000 | 0.730469 | 0.754787 |

| | | | | | |
|---|---|---|---|---|---|
| **DistilBERT** | 0.846150 | 0.849470 | 0.841400 | 0.845416 | 0.925107 |
| **Twitter-RoBERTa** | **0.882400** | **0.887908** | **0.875300** | **0.881559** | **0.952800** |

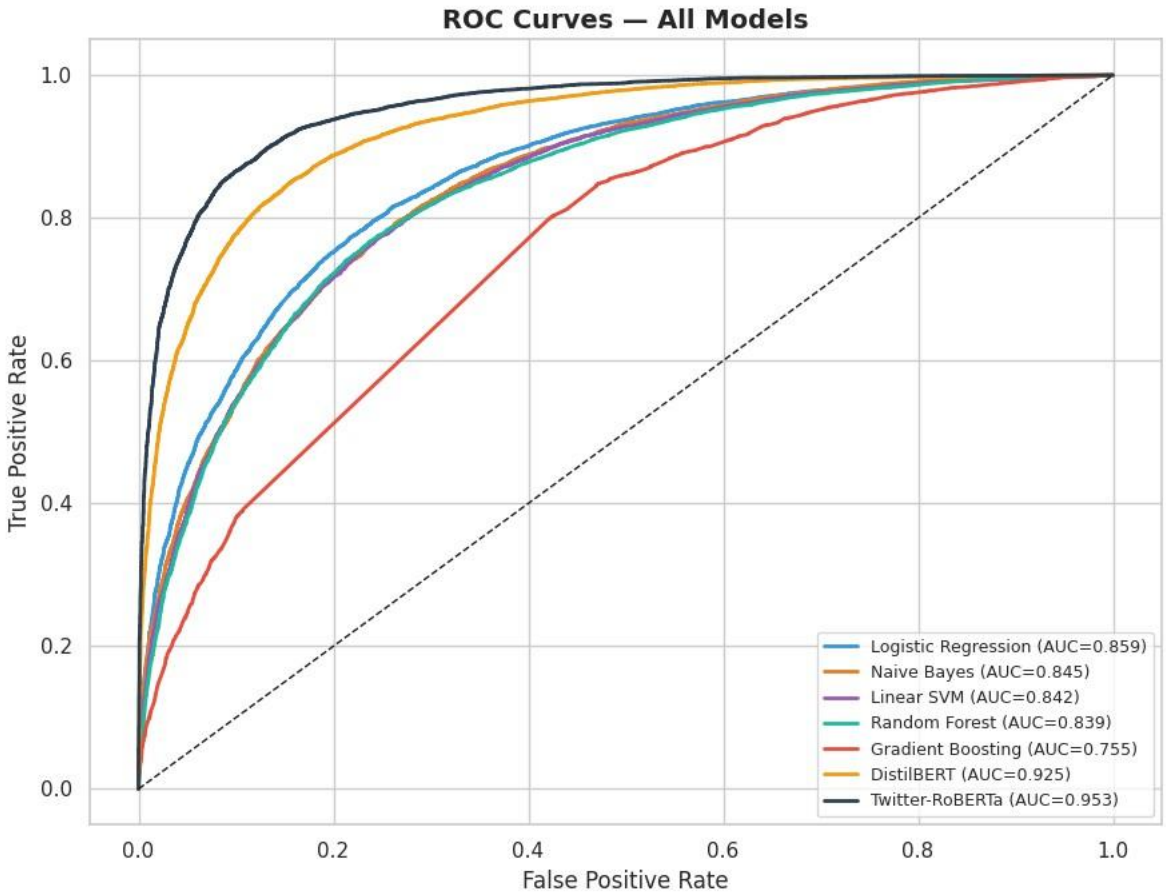
*Πίνακας 6: Τελικά αποτελέσματα ταξινομητών*

**Στην επόμενη εικόνα φαίνονται και αναλυτικά όλες οι μετρικές:**



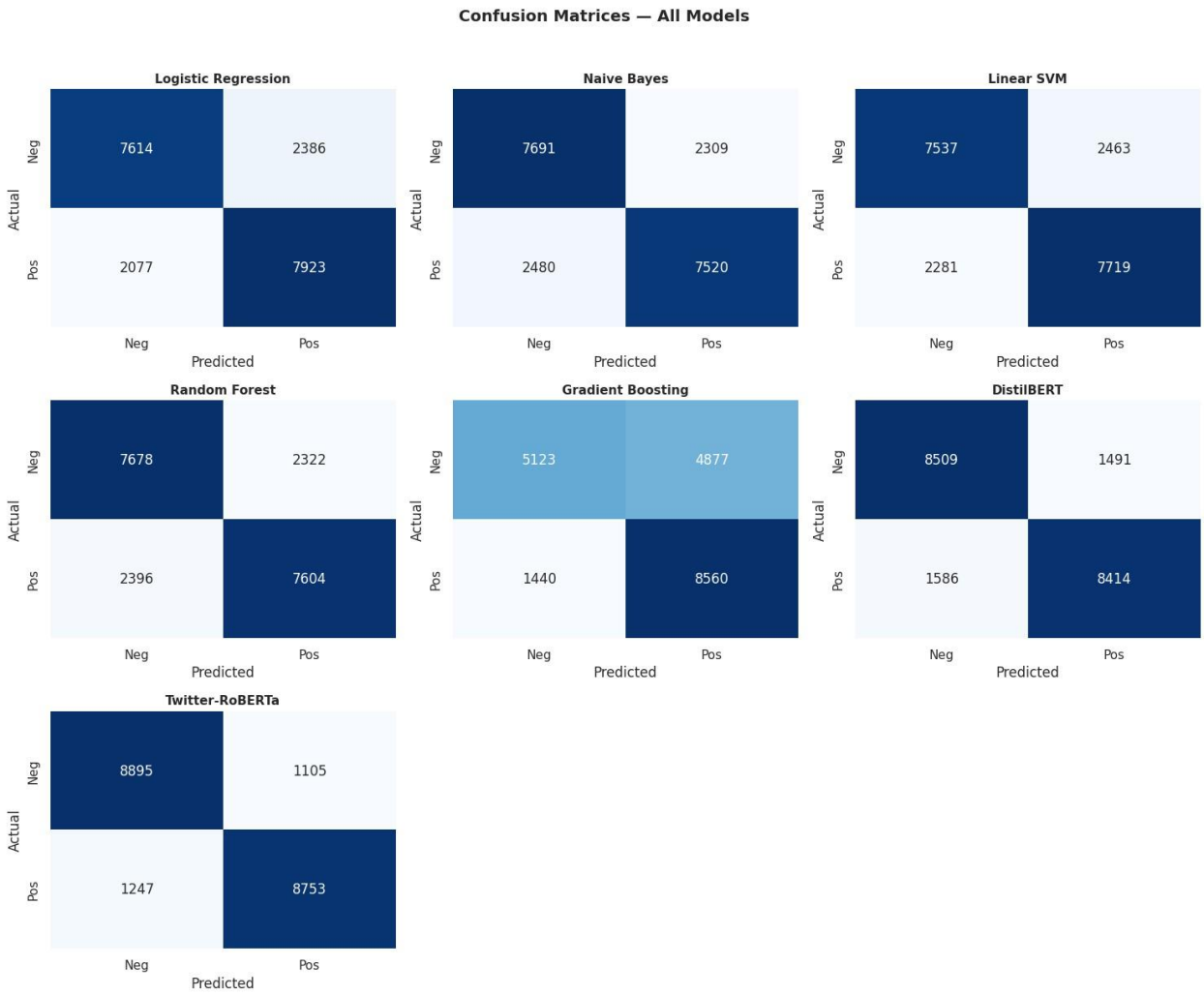

*Εικόνα 8: Τελικές μετρικές*

**Στην επόμενη εικόνα φαίνονται αναλυτικά όλα τα ROC curves:**



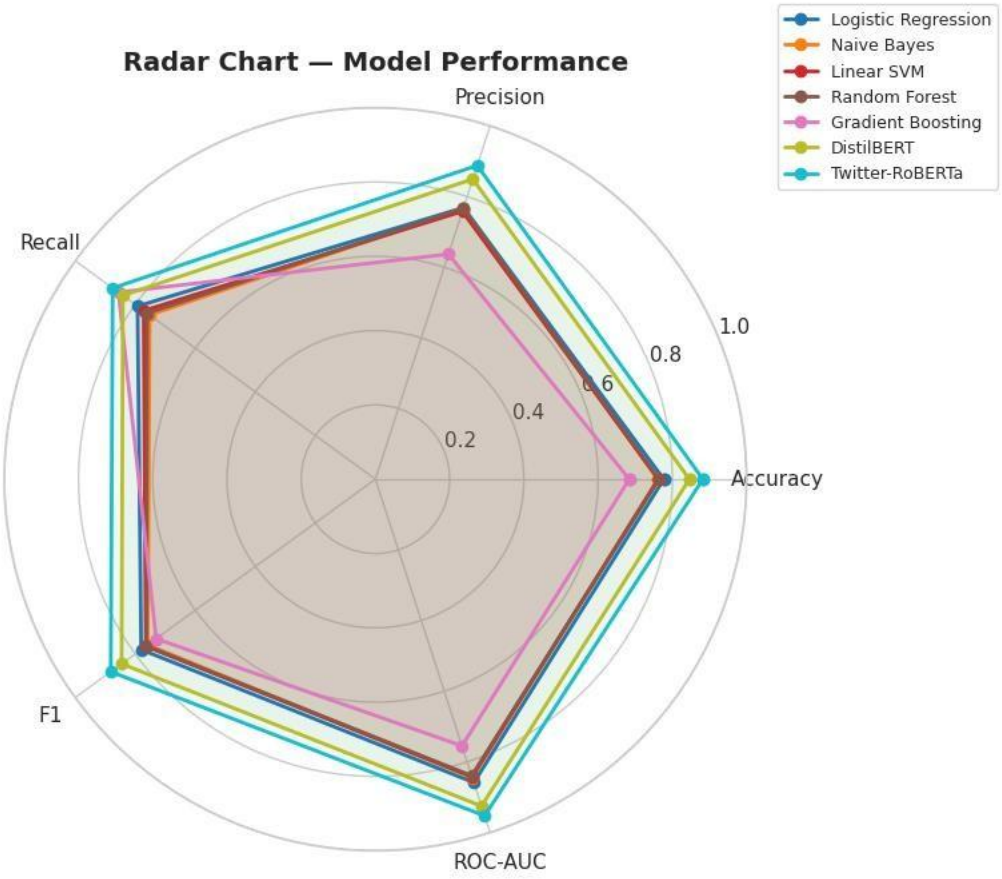

*Εικόνα 9: Τελικά ROC curves*

**Στην επόμενη εικόνα φαίνονται αναλυτικά όλα τα confusion matrices:**




*Εικόνα 10: Τελικά confusion matrices*

**Στην επόμενη εικόνα φαίνονται αναλυτικά όλα τα radar charts:**




*Εικόνα 11: Τελικά radar chart*

**Στην επόμενη εικόνα φαίνονται αναλυτικά όλα τα**

## 6.5 Συζήτηση Αποτελεσμάτων

Τα αποτελέσματα αναδεικνύουν τρία βασικά συμπεράσματα.

Πρώτον, υπάρχει σαφής και συστηματική υπεροχή των μοντέλων Transformer έναντι των κλασικών μοντέλων. Το χάσμα απόδοσης μεταξύ του καλύτερου κλασικού μοντέλου (Logistic Regression, F1: 0.780) και του χειρότερου Transformer (DistilBERT, F1: 0.845) ανέρχεται σε 6.5 ποσοστιαίες μονάδες, υποδηλώνοντας ότι η βαθιά κατανόηση του γλωσσικού πλαισίου που παρέχουν οι αρχιτεκτονικές Transformer αποτελεί ουσιαστικό πλεονέκτημα έναντι της bag-ofwords λογικής του TF-IDF.

Δεύτερον, μεταξύ των κλασικών μοντέλων, τα απλούστερα γραμμικά μοντέλα (Logistic Regression, Linear SVM) υπερτερούν έναντι των πιο σύνθετων ensemble μεθόδων (Random Forest, Gradient Boosting). Αυτό αποτελεί χαρακτηριστικό εύρημα στο πεδίο του NLP και οφείλεται στη φύση του TF-IDF feature space, το οποίο είναι εξαιρετικά αραιό και υψηλοδιάστατο, συνθήκες υπό τις οποίες τα γραμμικά μοντέλα γενικεύουν αποτελεσματικότερα.

Τρίτον, η εξειδικευμένη προεκπαίδευση του Twitter-RoBERTa σε δεδομένα του ίδιου domain αποδεικνύεται καθοριστική. Η διαφορά F1-score μεταξύ Twitter-RoBERTa (0.882) και DistilBERT (0.845), μοντέλα που εκπαιδεύτηκαν υπό τις ίδιες συνθήκες fine-tuning, αποδίδεται αποκλειστικά στην ποιότητα και τη συνάφεια της προεκπαίδευσης, καταδεικνύοντας τη σημασία επιλογής domain-specific μοντέλων για εξειδικευμένες εφαρμογές NLP.

# Κεφάλαιο 7: Συζήτηση

## 7.1 Ερμηνεία Αποτελεσμάτων

Τα αποτελέσματα της εργασίας επιβεβαιώνουν με συνέπεια την υπόθεση ότι τα μοντέλα βαθιάς μάθησης βασισμένα σε αρχιτεκτονικές Transformer υπερέχουν έναντι των κλασικών μεθόδων μηχανικής μάθησης στο έργο της ανάλυσης συναισθήματος σε tweets. Η υπεροχή αυτή δεν είναι τυχαία, αλλά αντανακλά θεμελιώδεις διαφορές στον τρόπο αναπαράστασης και κατανόησης του κειμένου.

Τα κλασικά μοντέλα, παρά τη χρήση ενός πλούσιου TF-IDF feature space με 50.000 χαρακτηριστικά και συμπερίληψη bigrams, περιορίζονται από τη bag-of-words λογική, η οποία αγνοεί την τάξη των λέξεων και το πλαίσιο εμφάνισής τους. Έτσι, προτάσεις όπως "not happy" και "happy" αντιμετωπίζονται ως παρόμοιες αναπαραστάσεις, εφόσον μοιράζονται το ίδιο unigram. Αντίθετα, τα μοντέλα Transformer επεξεργάζονται ολόκληρη την ακολουθία tokens μέσω μηχανισμών self-attention, επιτρέποντας την ανάδειξη σχέσεων εξάρτησης ανεξαρτήτως απόστασης στο κείμενο.

Η ανισορροπία Precision-Recall που εμφανίζει το Gradient Boosting (Precision: 0.637, Recall: 0.856) αξίζει ιδιαίτερης μνείας. Το μοντέλο αυτό τείνει να ταξινομεί υπερβολικά μεγάλο ποσοστό tweets ως θετικά, συμπεριφορά που πιθανώς οφείλεται στο συνδυασμό της διαδοχικής εκπαίδευσης του αλγορίθμου με τις προεπιλεγμένες υπερπαραμέτρους στο εξαιρετικά αραιό TF-IDF feature space. Η βελτιστοποίηση υπερπαραμέτρων μέσω crossvalidation θα μπορούσε πιθανώς να αντιμετωπίσει αυτή την ανισορροπία.

Μεταξύ των δύο Transformer μοντέλων, η διαφορά F1-score κατά 3.6 ποσοστιαίες μονάδες υπέρ του Twitter-RoBERTa (0.882 έναντι 0.845) αποδίδεται στην εξειδικευμένη προεκπαίδευσή του. Το γεγονός ότι και τα δύο μοντέλα εκπαιδεύτηκαν υπό ακριβώς τις ίδιες συνθήκες finetuning καθιστά τη σύγκριση αυτή ιδιαίτερα αξιόπιστη: η μόνη μεταβλητή που διαφέρει είναι η ποιότητα και η συνάφεια της προεκπαίδευσης.

## 7.2 Σύγκριση με τη Βιβλιογραφία

Τα αποτελέσματα της εργασίας εντάσσονται στο πλαίσιο ευρύτερων ερευνητικών ευρημάτων στο πεδίο της ανάλυσης συναισθήματος από δεδομένα βασισμένα στο Twitter.

Οι [3], οι δημιουργοί του Sentiment140, ανέφεραν ακρίβεια περίπου 83% με χρήση Maximum Entropy και SVM σε σύνολο 1.6 εκατομμυρίων tweets, αποτέλεσμα συγκρίσιμο με τα κλασικά μοντέλα της εργασίας, αν και οι διαφορές στο μέγεθος του δείγματος και τις παραμέτρους καθιστούν την άμεση σύγκριση ενδεικτική. Οι [14], στο πλαίσιο του διαγωνισμού SemEval, κατέγραψαν F1-scores της τάξης του 0.68–0.72 για κλασικά μοντέλα στο Twitter Sentiment Analysis task, επιβεβαιώνοντας ότι το εύρος 0.75–0.78 που επιτυγχάνεται εδώ αντιπροσωπεύει ανταγωνιστική απόδοση για την κατηγορία αυτή. Αναφορικά με τα μοντέλα Transformer, οι [17], οι δημιουργοί του TwitterRoBERTa, ανέφεραν υψηλές επιδόσεις σε διάφορα Twitter NLP benchmarks, με το μοντέλο να υπερέχει συστηματικά έναντι γενικής χρήσης μοντέλων όπως το BERT και το DistilBERT, εύρημα που επαληθεύεται πλήρως από τα αποτελέσματα της εργασίας. Τέλος, οι [15] απέδειξαν ότι το fine-tuned BERT επιτυγχάνει σημαντική βελτίωση έναντι κλασικών μοντέλων σε προβλήματα ανάλυσης συναισθήματος, με χάσμα απόδοσης της τάξης των 5–8 ποσοστιαίων μονάδων, αποτέλεσμα που συνάδει με τη διαφορά 6.5 μονάδων που παρατηρείται μεταξύ Logistic Regression και DistilBERT στην εργασία.

Συνολικά, τα αποτελέσματα επιβεβαιώνουν την υπάρχουσα βιβλιογραφία και τοποθετούν τα μοντέλα της εργασίας σε ανταγωνιστικό επίπεδο για το συγκεκριμένο έργο και σύνολο δεδομένων.

# Κεφάλαιο 8: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

## 8.1 Συμπεράσματα

Η εργασία πραγματεύτηκε το πρόβλημα της αυτόματης ανάλυσης συναισθήματος σε αναρτήσεις της πλατφόρμας Twitter, με έμφαση στη συγκριτική αξιολόγηση κλασικών μεθόδων μηχανικής μάθησης και σύγχρονων αρχιτεκτονικών Transformer. Για τον σκοπό αυτό σχεδιάστηκε και υλοποιήθηκε μια ολοκληρωμένη pipeline επεξεργασίας κειμένου, εκπαίδευσης και αξιολόγησης επτά διαφορετικών μοντέλων στο σύνολο δεδομένων Sentiment140.

Από τη συνολική ανάλυση προκύπτουν τα εξής βασικά συμπεράσματα:

**Τα μοντέλα Transformer υπερέχουν σαφώς έναντι των κλασικών μεθόδων.** Το DistilBERT και το Twitter-RoBERTa επέτυχαν F1-score 0.845 και 0.882 αντίστοιχα, έναντι 0.780 της καλύτερης κλασικής μεθόδου (Logistic Regression). Η βελτίωση αυτή οφείλεται στην ικανότητα των Transformer να κατανοούν το πλαίσιο και τις εξαρτήσεις μεταξύ λέξεων, κάτι που οι bag-ofwords αναπαραστάσεις TF-IDF αδυνατούν να συλλάβουν.

**Η εξειδικευμένη προεκπαίδευση σε domain-specific δεδομένα αποδίδει μετρήσιμο πλεονέκτημα.** Το Twitter-RoBERTa, προεκπαιδευμένο σε 58 εκατομμύρια tweets, υπερέχει του DistilBERT κατά 3.6 μονάδες F1 και 2.8 μονάδες ROC-AUC υπό ταυτόσημες συνθήκες fine-tuning, καταδεικνύοντας ότι η επιλογή του κατάλληλου προεκπαιδευμένου μοντέλου είναι εξίσου κρίσιμη με την αρχιτεκτονική του.

**Μεταξύ των κλασικών μοντέλων, τα γραμμικά μοντέλα υπερτερούν των ensemble μεθόδων.** Η Logistic Regression και το Linear SVM αποδίδουν καλύτερα από το Random Forest και το Gradient Boosting, επιβεβαιώνοντας ότι η αραιότητα και η υψηλή διαστατικότητα του TF-IDF feature space ευνοεί γραμμικούς ταξινομητές.

**Τα κλασικά μοντέλα παραμένουν βιώσιμη επιλογή υπό υπολογιστικούς περιορισμούς.** Παρά την κατωτερότητά τους έναντι των Transformer, επιτυγχάνουν Accuracy της τάξης του 76–78% με κλάσμα του υπολογιστικού κόστους, καθιστώντας τα κατάλληλα για εφαρμογές όπου οι πόροι είναι περιορισμένοι ή απαιτείται ταχύτατη εκπαίδευση και συμπερασμός.

## 8.2 Μελλοντικές Κατευθύνσεις

Η εργασία θέτει τις βάσεις για μια σειρά από επεκτάσεις και βελτιώσεις που θα μπορούσαν να αποτελέσουν αντικείμενο μελλοντικής έρευνας.

**Επέκταση σε πολυκατηγορική ταξινόμηση.** Η εργασία εστιάζει σε δυαδική ταξινόμηση (θετικό/αρνητικό). Μια φυσική επέκταση θα ήταν η συμπερίληψη ουδέτερης κλάσης ή ακόμη και η ανάλυση συναισθήματος σε κλίμακα έντασης (π.χ. πολύ αρνητικό, αρνητικό, ουδέτερο, θετικό, πολύ θετικό), προσεγγίζοντας πιο ρεαλιστικά τη φύση του ανθρώπινου συναισθήματος [39].

**Βελτιστοποίηση υπερπαραμέτρων.** Στην εργασία τα μοντέλα εκπαιδεύτηκαν με σε μεγάλο βαθμό προεπιλεγμένες λόγω της μεγάλη υπολογιστικής δύναμης που απαιτεί η αναζήτηση των βέλτιστων υπερπαραμέτρων. Η εφαρμογή συστηματικής αναζήτησης υπερπαραμέτρων (π.χ. grid search, Bayesian optimization) τόσο για τα κλασικά μοντέλα όσο και για τα Transformer (learning rate, batch size, αριθμός εποχών) θα μπορούσε να οδηγήσει σε περαιτέρω βελτίωση της απόδοσης [40].

**Αξιοποίηση μεγαλύτερου υποσυνόλου δεδομένων.** Λόγω υπολογιστικών περιορισμών χρησιμοποιήθηκαν 200.000 από τα 1.6 εκατομμύρια διαθέσιμα tweets. Η εκπαίδευση σε μεγαλύτερο υποσύνολο, ιδίως για τα μοντέλα Transformer, αναμένεται να βελτιώσει περαιτέρω την απόδοση, καθώς τα μοντέλα αυτά επωφελούνται σημαντικά από μεγαλύτερο όγκο δεδομένων fine-tuning [41].

**Εφαρμογή τεχνικών ερμηνευσιμότητας.** Μια σημαντική έλλειψη των μοντέλων Transformer είναι η περιορισμένη ερμηνευσιμότητά τους (black-box φύση). Η εφαρμογή τεχνικών όπως το SHAP (SHapley Additive exPlanations) [42] ή η ανάλυση attention weights θα επέτρεπε την κατανόηση των λέξεων και φράσεων που οδηγούν στις αποφάσεις του μοντέλου, αυξάνοντας την εμπιστοσύνη και τη διαφάνεια του συστήματος.

**Δοκιμή σε πραγματικά δεδομένα νέας εποχής.** Το Sentiment140 συλλέχθηκε το 2009, εποχή κατά την οποία η γλώσσα και οι συνήθειες των χρηστών του Twitter διέφεραν σημαντικά από τα σύγχρονα δεδομένα. Αξιολόγηση των μοντέλων σε πιο πρόσφατα tweets, τα οποία περιλαμβάνουν σύγχρονο slang, νέα emoji και αλλαγές στη χρήση της πλατφόρμας, θα αποκάλυπτε την πραγματική γενίκευση των μοντέλων σε σύγχρονες συνθήκες [43].

**Χρήση μεγαλύτερων μοντέλων.** Η εργασία χρησιμοποίησε το DistilBERT και το TwitterRoBERTa-base για λόγους υπολογιστικής αποδοτικότητας. Η αξιοποίηση μεγαλύτερων αρχιτεκτονικών [44], όπως το RoBERTa-large ή το DeBERTa, θα μπορούσε να οδηγήσει σε περαιτέρω βελτίωση της απόδοσης, εφόσον διατίθενται οι απαραίτητοι υπολογιστικοί πόροι.

# Βιβλιογραφικές αναφορές

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.*

[3] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford University, Stanford, CA, USA, Tech. Rep., 2009.

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, Jul. 2002, pp. 79–86.

[7] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. 5th Int. Conf. Language Resources and Evaluation (LREC)*, Genoa, Italy, May 2006, pp. 417–422.

[8] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proc. ESWC2011 Workshop Making Sense of Microposts*, Heraklion, Greece, May 2011, pp. 93–98.

[9] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs and Social Media (ICWSM)*, Ann Arbor, MI, USA, Jun. 2014, pp. 216–225.

[10] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju Island, Korea, Jul. 2012, pp. 90–94.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS)*, Vancouver, Canada, Dec. 2019, pp. 1–5.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, Jul. 2019.

[14] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Aug. 2017, pp. 502–518.

[15] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," in *Proc. 2019 Artificial Intelligence for Transforming Business and Society (AITB)*, Kathmandu, Nepal, Nov. 2019, pp. 1–5.

[16] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, Online, Nov. 2020, pp. 9–14.

[17] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Proc. Findings of the Association for Computational Linguistics (EMNLP 2020)*, Online, Nov. 2020, pp. 1644–1650.

[18] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[20] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *Proc. AAAI-98 Workshop on Learning for Text Categorization*, Madison, WI, USA, Jul. 1998, pp. 41–48.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp.

273–297, Sep. 1995.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–18.

[25] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018, June). Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 464-468).

[26] Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D. N., & Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices.

[27] Tsihrintzis, G. A., Sotiropoulos, D. N., & Jain, L. C. (2018). Machine learning paradigms: Advances in data analytics. In *Machine Learning Paradigms: Advances in Data Analytics* (pp. 14). Cham: Springer International Publishing.

[28] Panagoulias, D. P., Sotiropoulos, D. N., & Tsihrintzis, G. A. (2022). SVM-based blood exam classification for predicting defining factors in metabolic syndrome diagnosis. *Electronics*, *11*(6), 857.

[29] Sotiropoulos, D. N., Pournarakis, D. E., & Giaglis, G. M. (2017). SVM-based sentiment classification: a comparative study against state-of-the-art classifiers. *International Journal of Computational Intelligence Studies*, *6*(1), 52-67.

[30] Sotiropoulos, D. N., & Tsihrintzis, G. A. (2016). The class imbalance problem. In *Machine Learning Paradigms: Artificial Immune Systems and their Applications in Software Personalization* (pp. 51-78). Cham: Springer International Publishing.

[31] Singh, R., & Mangat, N. S. (1996). Stratified sampling. In *Elements of survey sampling* (pp. 102-144). Dordrecht: Springer Netherlands.

[32] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences* (pp. 1833-1842). IEEE.

[33] Mahmood, M., Jasem, F. M., Mukhlif, A. A., & Al-Khateeb, B. (2023). Classifying cuneiform symbols using machine learning algorithms with unigram features on a balanced dataset. *Journal of Intelligent Systems*, *32*(1), 20230087.

[34] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853-860.

[35] Akhmetov, I., Pak, A., Ualiyeva, I., & Gelbukh, A. (2020). Highly language-independent word lemmatization using a machine-learning classifier. *Computación y Sistemas*, *24*(3), 13531364.

[36] Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., ... & Tan, S. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

[37] Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human resource management review*, *27*(2), 265-276.

[38] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631–1642.

[39] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

[40] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

[41] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, No. 2, pp. 4171-4186).

[42] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

[43] O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010, May). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media* (Vol. 4, No. 1, pp. 122-129).

[44] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.