Detecting Indirect Racism and Sexism through Speech and Text Analysis

Machine Learning Project

Αγόρη Φωτεινή - MTN2401
Τόλια Άννα - MTN2418

Δ.Π.Μ.Σ. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΕΡΕΥΝΑΣ
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ «ΔΗΜΟΚΡΙΤΟΣ»

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Περιεχόμενα

Εισαγωγή

Η συγκεκριμένη εργασία, στοχεύει στην ανάπτυξη ενός μοντέλου μηχανικής μάθησης που θα μπορεί να ανιχνεύει έμμεσες μορφές ρητορικής μίσους, βασιζόμενο τόσο στα γλωσσικά χαρακτηριστικά όσο και στα φωνητικά χαρακτηριστικά της ομιλίας.

Η δυσκολία στην αναγνώριση έμμεσου λόγου σχετίζεται με τη χρήση ειρωνείας, τόνου φωνής και συναισθηματικής φόρτισης, κάτι που δεν αποτυπώνεται εύκολα από απλές τεχνικές ανάλυσης κειμένου.

Συλλογή δεδομένων

Τα δεδομένα που συλλέξαμε είναι youtube videos και προέρχονται από ελληνικές τηλεοπτικές εκπομπές, δελτία ειδήσεων και δημόσιες συζητήσεις. Το dataset, περιλαμβάνει αποσπάσματα ομιλίας πολλαπλών ομιλητών.

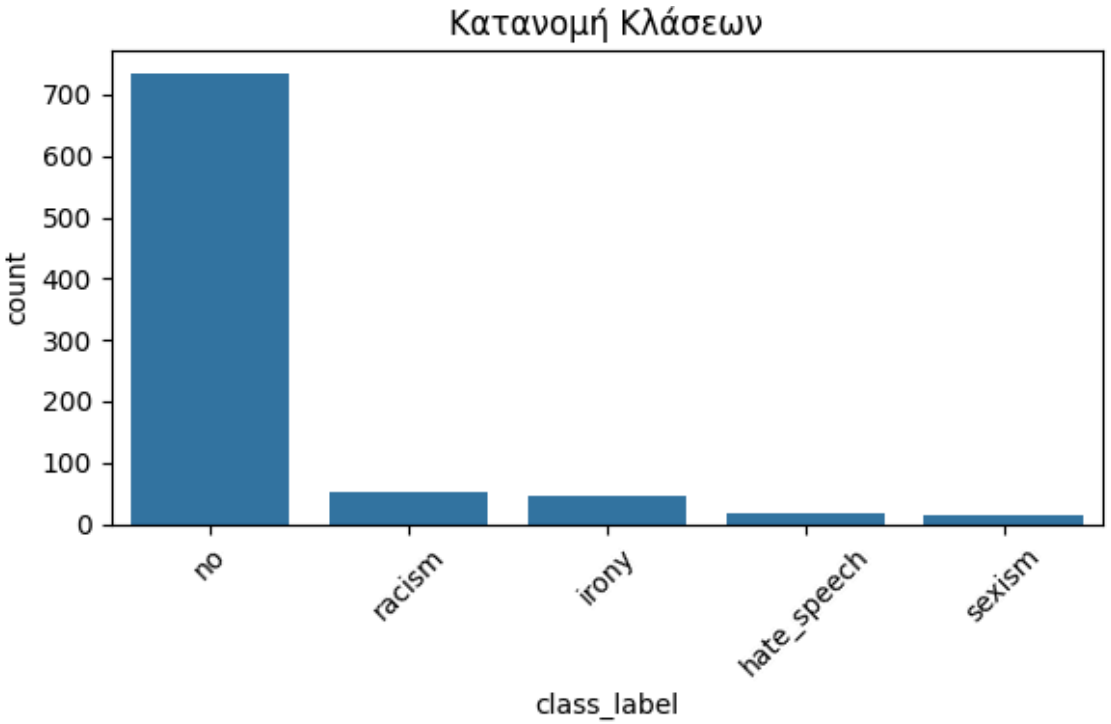
Αρχικά μετατρέψαμε τα βίντεο σε αρχεία ήχου mp3 και στη συνέχεια προχωρήσαμε σε speech-to-text χρησιμοποιώντας το WhisperX, (Medium Model). Κάναμε μία μικρή διόρθωση στα κείμενα καιπροχωρήσαμε στην επισημείωση (annotation) των δεδομένων. Για να επισημειώσουμε τα δεδομένα σαν εύκολη μέθοδο χρησιμοποιούμε το excel όπου κάθε πρόταση (utterance) από το transcription αντιστοιχούσε σε μία γραμμή και στην δίπλα στήλη δώσαμε το αντίστοιχο annotation από τις κατηγορίες: **hate speech, racism, sexism, irony, no**.

Στην συνέχεια εξάγαμε το αρχείο σε csv, και προχωρήσαμε με την προετοιμασία των δεδομένων και την εκπαίδευση του μοντέλου.

Exploratory Data Analysis

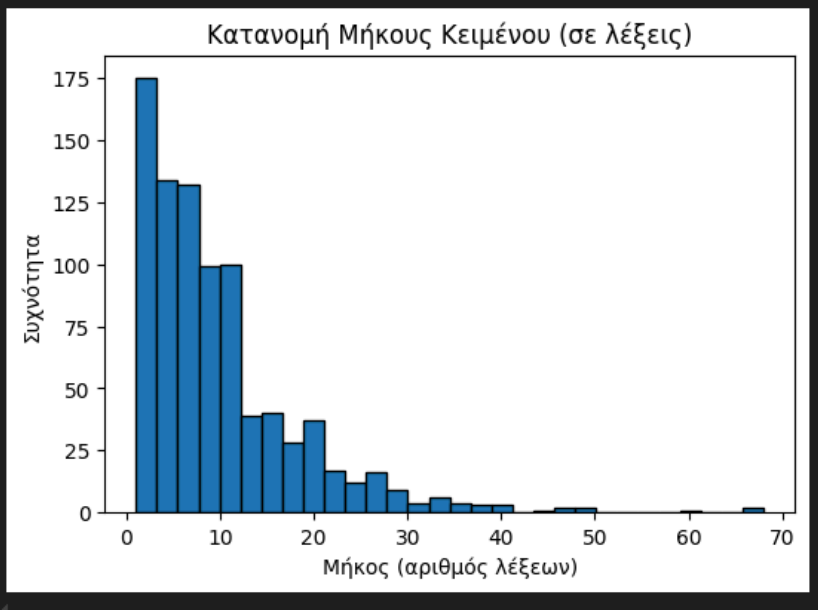
Κατά την εξερεύνηση των δεδομένων πραγματοποιήθηκαν οι εξής αναλύσεις:

- **Κατανομή κλάσεων**: Παρατηρήθηκε μεγάλη ανισορροπία μεταξύ των κλάσεων, με την κατηγορία "no" να κυριαρχεί.



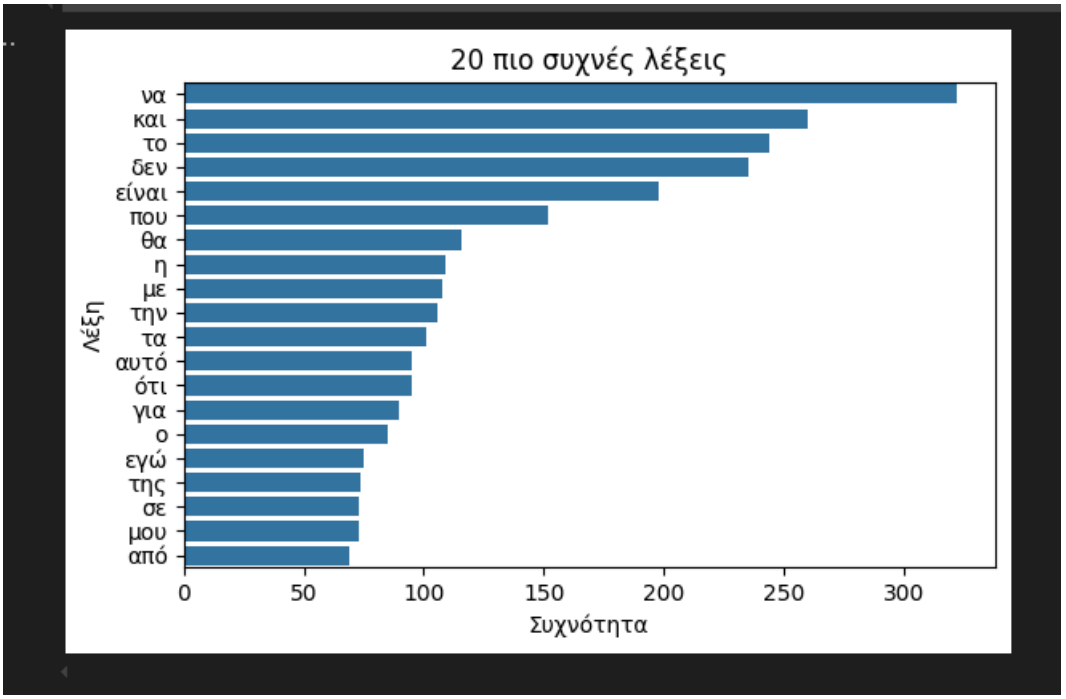
- **Μέσος αριθμός λέξεων ανά πρόταση**: Αναλύθηκε η έκταση των προτάσεων, παρέχοντας πληροφορίες για τη δομή του dataset.

```
Στατιστικά μήκους κειμένου:
count    866.000000
mean       9.942263
std        8.717739
min        1.000000
25%        4.000000
50%        7.000000
75%       13.000000
max       68.000000
Name: text_length, dtype: float64
```



-

- **Συχνότητα λέξεων**: Έγινε μελέτη των πιο κοινών λέξεων ανά κλάση για την καλύτερη κατανόηση των μοτίβων στη γλώσσα του dataset.



Συμπεράσματα

Η ανάλυση των δεδομένων αποκάλυψε ορισμένες κρίσιμες παρατηρήσεις που επηρέασαν τις αποφάσεις μας όσον αφορά την προεπεξεργασία και την εκπαίδευση των μοντέλων.

Μία από τις σημαντικότερες προκλήσεις που αντιμετωπίσαμε ήταν η ανισορροπία μεταξύ των κλάσεων, με την κλάση "no" να υπερισχύει σημαντικά στο dataset. Αυτή η υπερεκπροσώπηση οδήγησε σε προκατάληψη του μοντέλου (bias), καθώς στις αρχικές δοκιμές, το μοντέλο έτεινε να προβλέπει σχεδόν αποκλειστικά αυτή την κατηγορία, αγνοώντας τις υπόλοιπες. Για την αντιμετώπιση αυτού του προβλήματος, εφαρμόσαμε τεχνικές εξισορρόπησης, δοκιμάζοντας τόσο downsampling όσο και SMOTE (Synthetic Minority Over-sampling Technique).

Feature Extraction

Στη διαδικασία εξαγωγής χαρακτηριστικών, χρησιμοποιήθηκαν δεδομένα κειμένου σε μορφή **utterances** (μικρές προτάσεις ή φράσεις). Για τη μετατροπή των δεδομένων σε αριθμητικά χαρακτηριστικά, εφαρμόστηκαν τα παρακάτω βήματα:

Προεπεξεργασία Κειμένου

Πριν την εξαγωγή των χαρακτηριστικών, εφαρμόστηκαν οι εξής τεχνικές προεπεξεργασίας:

- **Αφαίρεση Stopwords:** Απομακρύνθηκαν κοινές λέξεις που δεν προσφέρουν πληροφορία, χρησιμοποιώντας τη βιβλιοθήκη NLTK.Αρχικά, χρησιμοποιήθηκαν οι προεπιλεγμένες λίστες stopwords από τη βιβλιοθήκη NLTK, αλλά παρατηρήθηκε ότι αυτές περιείχαν λέξεις από αρχαία ελληνικά, οι οποίες δεν ήταν σχετικές με το σύγχρονο κείμενο των δεδομένων μας. Για τον λόγο αυτό, δημιουργήσαμε custom stopword list, ενσωματώνοντας τις πιο συχνές αλλά μη πληροφοριακές λέξεις από το dataset.
- **Tokenization:** Το κείμενο διασπάστηκε σε λέξεις (tokens) με τη χρήση NLTK για να επιτρέψει την ανάλυση των λέξεων ξεχωριστά.

Εξαγωγή Χαρακτηριστικών

- TF-IDF (Term Frequency-Inverse Document Frequency): Χρησιμοποιήθηκε το TfidfVectorizer από τη βιβλιοθήκη Scikit-learn για την αναπαράσταση του κειμένου ως αριθμητικά χαρακτηριστικά. Το TF-IDF επιλέχθηκε επειδή μειώνει τη σημασία των πολύ συχνών λέξεων και ενισχύει τη σημασία των λιγότερο συχνών αλλά σημαντικών όρων.
- PCA (Principal Component Analysis): Χρησιμοποιήθηκε για τη μείωση της διάστασης των χαρακτηριστικών και την αφαίρεση της πλεοναστικής πληροφορίας, βελτιώνοντας έτσι την απόδοση των μοντέλων και μειώνοντας την πολυπλοκότητα.

Βιβλιοθήκες και Εργαλεία

Για την εξαγωγή των χαρακτηριστικών, χρησιμοποιήθηκαν οι εξής βιβλιοθήκες:

- Scikit-learn: Για την εφαρμογή του TF-IDF Vectorization και του PCA.
- Pandas: Για τη διαχείριση και επεξεργασία των δεδομένων.
- NLTK: Για την προεπεξεργασία κειμένου, όπως stopword removal και tokenization.

Επιλογή Παραμέτρων

- TF-IDF: Χρησιμοποιήθηκε max_features=5000 για να περιορίσουμε τον αριθμό των χαρακτηριστικών και να αποφύγουμε το overfitting.
- PCA: Επιλέχθηκαν n_components=100, διατηρώντας το μεγαλύτερο ποσοστό της πληροφορίας αλλά μειώνοντας τη διάσταση.

Downsampling / Διαχωρισμός Δεδομένων

Αρχικά, το μοντέλο εκπαιδεύτηκε χωρίς να εφαρμοστεί κάποια τεχνική εξισορρόπησης των δεδομένων. Ωστόσο, παρατηρήθηκε ότι λόγω του μεγάλου ανισοζυγίου μεταξύ των κλάσεων, το μοντέλο είχε έντονη προκατάληψη προς την επικρατέστερη κλάση ("no"), καταλήγοντας να προβλέπει σχεδόν αποκλειστικά αυτή την κατηγορία.

Downsampling της κλάσης "No"

Για την αντιμετώπιση του προβλήματος της ανισορροπίας των δεδομένων, εφαρμόστηκε downsampling στην κλάση "no". Συγκεκριμένα, μειώσαμε τα δείγματα αυτής της κατηγορίας στα 60 δείγματα, ώστε να πλησιάζει το μέγεθος των υπόλοιπων κλάσεων.

Για να αντιμετωπίσουμε την ανισορροπία των δεδομένων, εκτός από downsampling, δοκιμάσαμε επίσης SMOTE (Synthetic Minority Over-sampling Technique). Η τεχνική αυτή δημιουργεί συνθετικά δείγματα για την υπο εκπροσωπούμενη κλάση, αυξάνοντας την ποικιλομορφία των δεδομένων.

Ως τεχνική, λειτουργεί ακολουθώντας τα εξής βήματα:

1. Επιλέγει τυχαία ένα δείγμα από τη μειονοτική κλάση.
2. Υπολογίζει τους k πλησιέστερους γείτονες του.
3. Δημιουργεί ένα νέο δείγμα στο διάστημα μεταξύ του αρχικού δείγματος και ενός γείτονα.
4. Η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί το επιθυμητό επίπεδο ισορροπίας.

Αρχικά χρησιμοποιήθηκε εσφαλμένα και εφαρμόστηκε σε ολόκληρο το dataset, πριν από την εκτέλεση του (LOO-CV), προκαλώντας data leakage, και υπερβολικά καλές αποδόσεις σε επίπεδο metrics. Παρατηρήθηκε επίσης και παραμόρφωση στο συνολικό support στο classification report, που εμφανίζεται το πλήθος των test instances που χρησιμοποιήθηκαν, καθώς σε όλες τις κλάσεις εμφανίζοταν ο αριθμός 60, κάτι που δεν ίσχυε στην πραγματικότητα βάσει του dataset μας.

Όταν επιδιορθώθηκε η λανθασμένη εφαρμογή του και χρησιμοποιήθηκε μόνο στο training set σε κάθε iteration του LOO-CV, τα metrics μειώθηκαν εμφανώς, αλλά και πάλι ήταν αρκετά βελτιωμένα, συγκριτικά με τα metrics που είχαν προκύψει σε όλες της τεχνικές classification που δοκιμάσαμε, χωρίς την χρήση της τεχνικής smote. Το test-set αυτήν τη φορά, είχε διατηρήσει την αρχική του μορφή, αφού στο support του classification report, εμφανίζονταν, 60 δείγματα no, 52 δείγματα racism, 46 δείγματα irony, 17 δείγματα hate speech και 15 δείγματα sexism.

Διαχωρισμός Δεδομένων & Εκπαίδευση με Leave-One-Out (LOO)

Λόγω του πολύ μικρού μεγέθους του dataset, δεν εφαρμόστηκε η κλασική προσέγγιση train-test split ή cross-validation. Αντί αυτών, χρησιμοποιήθηκε η τεχνική Leave-One-Out Cross Validation (LOO-CV).

- Κάθε φορά, ένα μόνο δείγμα αφαιρούνταν ως test set, ενώ το υπόλοιπο dataset χρησιμοποιούνταν για εκπαίδευση.
- Η διαδικασία αυτή επαναλαμβανόταν για κάθε δείγμα στο dataset, διασφαλίζοντας ότι κάθε σημείο χρησιμοποιείται τόσο για εκπαίδευση όσο και για αξιολόγηση.
- Η μέθοδος αυτή επιλέχθηκε επειδή επιτρέπει βέλτιστη χρήση των δεδομένων, ειδικά όταν το δείγμα είναι μικρό, όπως στην περίπτωσή μας.

Αποτελέσματα & Συμπεράσματα

Η χρήση downsampling βελτίωσε σημαντικά την απόδοση του μοντέλου, καθώς μειώθηκε η προκατάληψη προς την κλάση "no" και οι προβλέψεις έγιναν πιο ισορροπημένες. Επιπλέον, η τεχνική Leave-One-Out διασφάλισε ότι όλα τα δεδομένα συνέβαλαν τόσο στην εκπαίδευση όσο και στην αξιολόγηση του μοντέλου, μεγιστοποιώντας τη χρήση των διαθέσιμων δειγμάτων.

Παράλληλα, η εφαρμογή του SMOTE στο training set κάθε iteration του LOO-CV, συνέβαλε περαιτέρω στη βελτίωση της απόδοσης του μοντέλου, ιδίως ως προς τις μειονοτικές κλάσεις. Αν και η αρχική εσφαλμένη χρήση του σε ολόκληρο το dataset προκάλεσε, υπερβολικά αισιόδοξες επιδόσεις, η σωστή εφαρμογή του βελτίωσε την ικανότητα του μοντέλου να προβλέπει τις υποεκπροσωπούμενες κατηγορίες. Ωστόσο, παρατηρήθηκε μια ελαφρώς αυξημένη τάση προς overfitting, γεγονός που υποδηλώνει ότι η προσεκτική ρύθμιση των υπερπαραμέτρων και η χρήση συμπληρωματικών τεχνικών μπορεί να βελτιώσει περαιτέρω τη γενίκευση του μοντέλου.

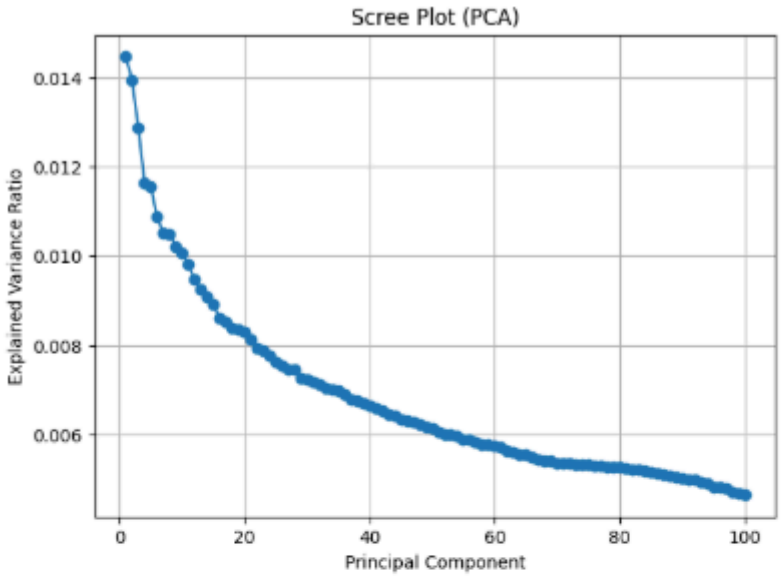
Hyperparameter Tuning - Εκπαίδευση Μοντέλου

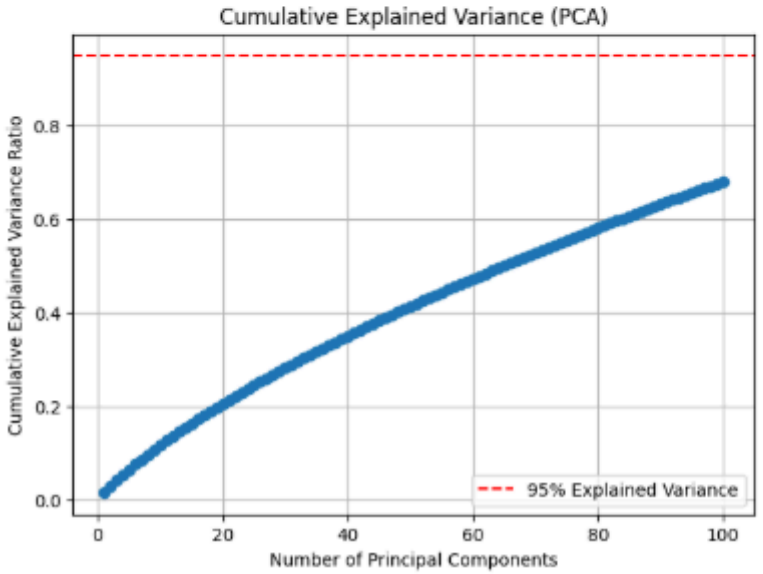
Στη διαδικασία εκπαίδευσης του μοντέλου, δοκιμάστηκαν τρεις διαφορετικοί ταξινομητές: Logistic Regression, Naive Bayes και SVM. Για την αξιολόγηση της απόδοσής τους, πραγματοποιήθηκαν δύο πειράματα:

1. Εκπαίδευση των μοντέλων χωρίς PCA.
2. Εκπαίδευση των μοντέλων με PCA, μειώνοντας τη διάσταση των χαρακτηριστικών.

Επιλογή του αριθμού των Principal Components

Για να προσδιορίσουμε τον ιδανικό αριθμό χαρακτηριστικών μετά το PCA, αρχικά δημιουργήθηκε ένα Scree Plot (εικόνα 1).





Ωστόσο, δεν υπήρχε κάποιο εμφανές elbow point, γεγονός που δυσκόλεψε την επιλογή του βέλτιστου αριθμού components. Βοηθητικά, προστέθηκε ένα Cumulative Explained Variance Plot (εικόνα 2), για να οπτικοποιηθεί η σωρευτική διακύμανση και να μας δώσει μια ποσοτική μέτρηση, του πόσο χρήσιμη ήταν η κάθε επιπλέον προσθήκη συνιστωσών. Με αυτόν τον τρόπο, είχαμε μια πιο ξεκάθαρη εικόνα επιλογής αριθμού principal components. Έτσι, η επιλογή έγινε χειροκίνητα, δοκιμάζοντας τιμές από 50 έως 200 components και αξιολογώντας την απόδοση των μοντέλων σε κάθε περίπτωση.

Αξιολόγηση & Αποτελέσματα

- Δεν τροποποιήθηκαν οι υπερπαράμετροι σε κανένα από τα μοντέλα. και χρησιμοποιήθηκαν οι προεπιλεγμένες τιμές της Scikit-learn, οι οποίες είναι επιλεγμένες με βάση γενικές βέλτιστες πρακτικές και αποδίδουν

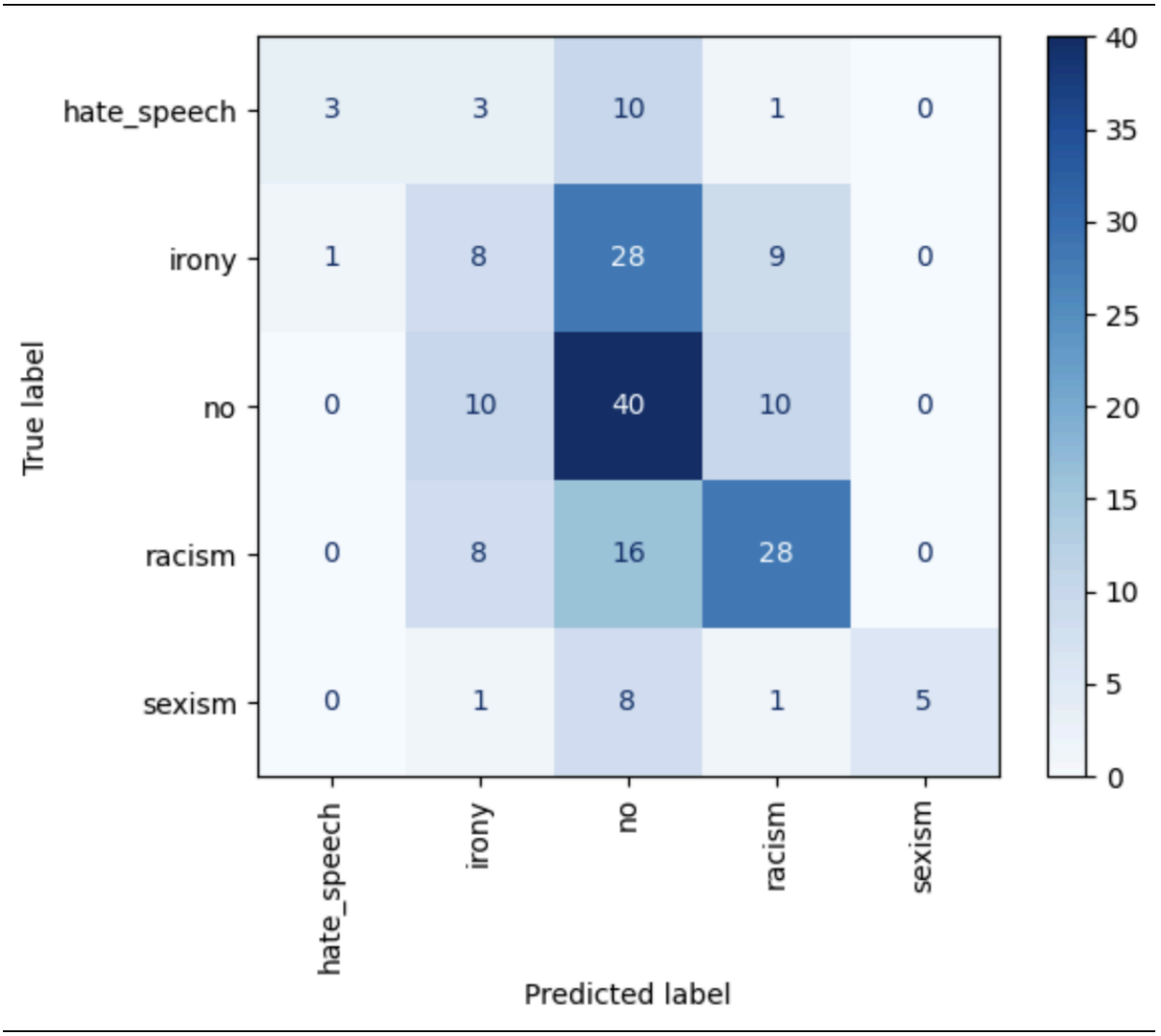
ικανοποιητικά σε ένα μεγάλο εύρος προβλημάτων. Αποφασίστηκε πως έτσι, θα παρεχόταν μια baseline επίδοση για κάθε μοντέλο, ώστε να υπάρχει ένα ουδέτερο σημείο σύγκρισης. Επιπλέον, λόγω του μικρού εύρους του dataset, που ήταν ήδη επιρρεπές στην πιθανότητα overfitting, αποφύγαμε την την ρύθμιση των υπερπαραμέτρων, καθώς είναι γνωστό, πως η λανθασμένη χρήση τους μπορεί να οδηγήσει σε περαιτέρω υπερπροσαρμογή.

- Η αξιολόγηση πραγματοποιήθηκε με F1-score, σε συνδυασμό με την ανάλυση του Confusion Matrix για την κατανόηση των λαθών των ταξινομητών.
- Δεν παρατηρήθηκε σημαντική διαφορά μεταξύ των δύο πειραμάτων (με και χωρίς PCA). Αυτό υποδηλώνει ότι η μείωση της διάστασης δεν βελτίωσε την απόδοση των ταξινομητών στο συγκεκριμένο dataset, απεναντίας την μείωσε σε κάποιες περιπτώσεις.

Μετρήσεις

Όλες οι μετρικές μετρήθηκαν, αλλά η κύρια εστίαση ήταν στο F1-score, το οποίο αποτελεί ισορροπημένο μέτρο μεταξύ Precision και Recall. Από το Confusion Matrix, φαίνεται ότι η κατηγορία "no" είχε την καλύτερη απόδοση, με 40 σωστές ταξινομήσεις, ενώ υπήρχε σύγχυση μεταξύ κατηγοριών όπως hate_speech, irony και racism, γεγονός που υποδηλώνει ότι το μοντέλο δυσκολεύεται να διαχωρίσει αυτές τις έννοιες. Επιπλέον, η κατηγορία sexism παρουσίασε πολύ χαμηλά ποσοστά σωστής ταξινόμησης, δείχνοντας ότι το μοντέλο δεν έχει μάθει καλά να τη διακρίνει. Η χρήση του SMOTE (Synthetic Minority Over-sampling Technique) βελτίωσε σημαντικά την απόδοση όλων των ταξινομητών, με το SVM να επιτυγχάνει την καλύτερη απόδοση, φτάνοντας F1-score 0.41 όταν χρησιμοποιήθηκε χωρίς PCA αλλά με SMOTE. Ο Naive Bayes είχε τη χαμηλότερη απόδοση στις περισσότερες περιπτώσεις, ενώ το Logistic Regression είχε σταθερή απόδοση, η οποία όμως βελτιώθηκε σημαντικά με την εφαρμογή του SMOTE. Συμπερασματικά, το SVM φαίνεται να είναι το καλύτερο μοντέλο, καθώς επιτυγχάνει το υψηλότερο F1-score, πιθανότατα λόγω της ικανότητάς του να διαχειρίζεται καλύτερα μη γραμμικά δεδομένα και πολύπλοκα σύνολα δεδομένων. Επομένως, η χρήση του SMOTE συνέβαλε σημαντικά στη βελτίωση της ταξινόμησης, ειδικά για το SVM, το οποίο πέτυχε την υψηλότερη απόδοση σε σύγκριση με τους υπόλοιπους ταξινομητές.

f1-score	Logistic Regression	Naive Base	SVM
PCA	0.21	0.05	0.28
no PCA	0.21	0.22	0.32
no PCA & smote	0.41	0.37	0.41



Επιπλέον πειράματα

Fusion Voting

Δοκιμάστηκε η τεχνική Fusion Voting, συνδυάζοντας τα αποτελέσματα των καλύτερων μοντέλων (SVM & Logistic Regression) με στόχο τη βελτίωση της συνολικής απόδοσης.

Η διαδικασία έχει ως εξής:

Τα μοντέλα εκπαιδεύτηκαν ξεχωριστά και στη συνέχεια συνδυάστηκαν με έναν κανόνα πλειοψηφίας (majority voting). Αν οι δύο classifiers παρήγαγαν διαφορετική πρόβλεψη, προτιμήθηκε η απόφαση του SVM, καθώς είχε υψηλότερη ακρίβεια στις περισσότερες κατηγορίες.

Με την ορθή εφαρμογή του SMOTE, το Fusion Voting έδειξε διαφορετικά αποτελέσματα:
Χωρίς SMOTE, η απόδοση ήταν macro-F1 = 0.30 και accuracy = 0.43.

- Πολύ χαμηλή επίδοση στην κατηγορία "racism" και "sexism", με recall 0.00.
- Η "irony" έχει recall 0.78, κάτι που δείχνει ότι το μοντέλο καταφέρνει να την αναγνωρίσει αρκετά καλά.
- Η κατηγορία "hate speech" έχει recall 0.58 και f1-score 0.57, που είναι αρκετά υψηλό.
- Η κατηγορία "no" έχει recall 0.27, κάτι που σημαίνει ότι η ικανότητα του μοντέλου να αναγνωρίζει μη τοξικά σχόλια είναι χαμηλή.

Με σωστό SMOTE, η απόδοση αυξήθηκε σε macro-F1 = 0.37 και accuracy = 0.44.

- Το Precision βελτιώθηκε σημαντικά (0.66 έναντι 0.38 πριν).
- Το recall δεν αυξήθηκε αισθητά (0.37 έναντι 0.33).
- Το f1-score αυξήθηκε ελαφρώς (0.37 έναντι 0.30).
- Παρατηρείται σημαντική βελτίωση στην κατηγορία "racism" (precision 0.67 από 0.00), αλλά το recall παραμένει χαμηλό (0.12).
- Το "sexism" έχει precision 0.43 (από 0.00), αλλά recall μόλις 0.07.
- Η "irony" έχει recall 0.98, το οποίο είναι πολύ υψηλό.
- Η "hate speech" έχει recall 0.27, που είναι μικρότερο από πριν, αλλά το precision είναι σημαντικά βελτιωμένο (0.82).

Τα αποτελέσματα λοιπόν, έδειξαν:

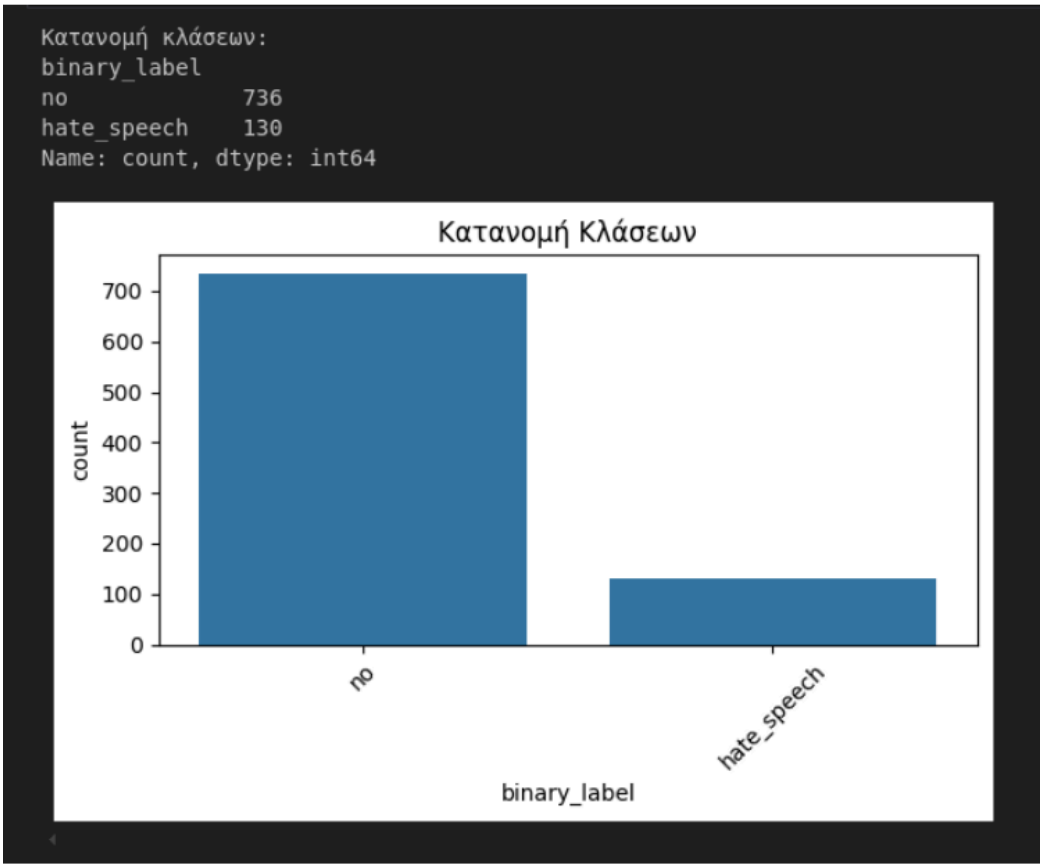
- Το SMOTE αύξησε την ακρίβεια του Fusion Voting από 0.4263 σε 0.4421, κάτι που σημαίνει ότι υπήρξε βελτίωση, αλλά όχι ριζική.
- Το Precision βελτιώθηκε σημαντικά, ειδικά στις κατηγορίες που πριν είχαν πολύ χαμηλή απόδοση, όπως το "racism" (από 0.00 σε 0.67) και το "sexism" (από 0.00 σε 0.43).
- Η κατηγορία "irony" είχε ήδη υψηλό recall (0.78 πριν, 0.98 μετά), και με το SMOTE το precision της βελτιώθηκε, κάνοντας την κατηγορία πιο ακριβή.
- Η κατηγορία "hate speech" είχε καλές επιδόσεις και πριν, αλλά μετά το SMOTE το precision αυξήθηκε, ενώ το recall μειώθηκε ελαφρώς.

Το Fusion Voting παρείχε μικρή βελτίωση στη σταθερότητα των προβλέψεων, αλλά δεν παρουσίασε σημαντική αύξηση της συνολικής απόδοσης σε σχέση με το απλό SVM.

Binary Classification

Παρατηρήθηκε σημαντική ανισορροπία μεταξύ των κλάσεων, με την κλάση **"No Hate Speech"** να υπερισχύει στο dataset. Η υπερεκπροσώπηση αυτής της κλάσης οδήγησε σε **προκατάληψη του μοντέλου** (bias), καθώς στις αρχικές δοκιμές, το μοντέλο έτεινε να προβλέπει σχεδόν αποκλειστικά αυτή την κατηγορία, αγνοώντας τις περιπτώσεις **Hate Speech**.

Για την αντιμετώπιση αυτού του προβλήματος, εφαρμόστηκε downsampling, όπου μειώθηκε ο αριθμός των δειγμάτων της κλάσης "No Hate Speech" ώστε να πλησιάζει το μέγεθος της κατηγορίας "Hate Speech" (130).

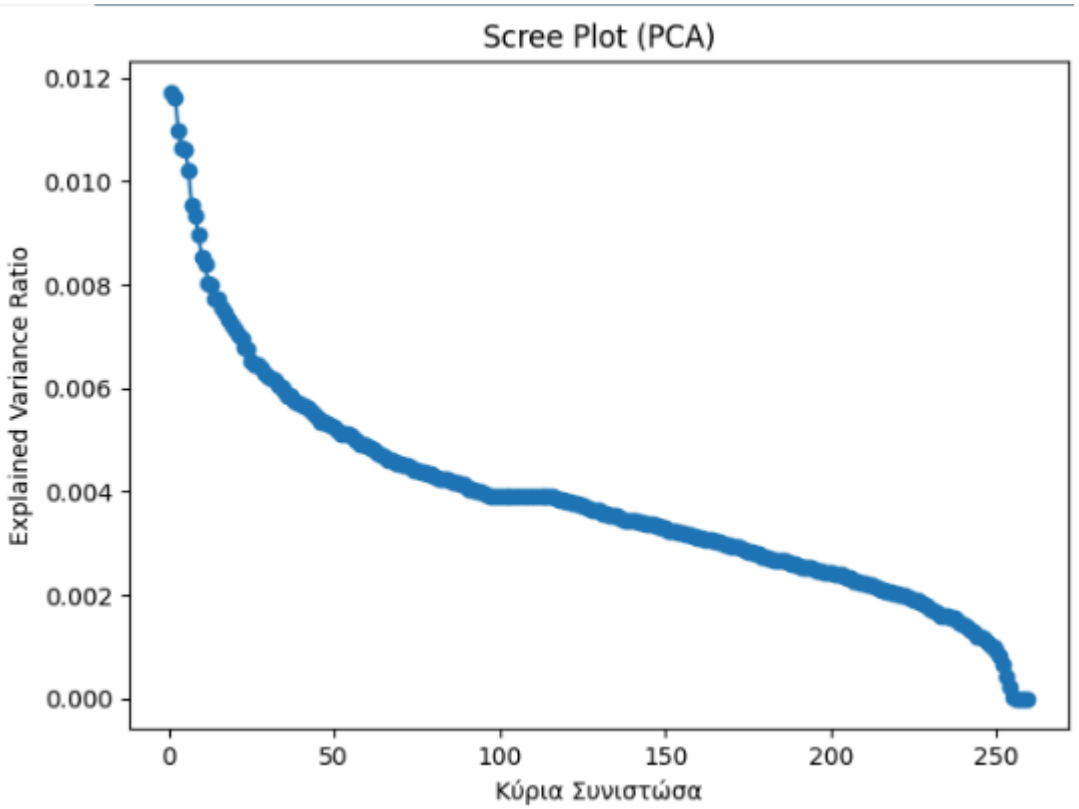


Στη διαδικασία εκπαίδευσης του μοντέλου, δοκιμάστηκαν τρεις διαφορετικοί ταξινομητές: **Logistic Regression, Naive Bayes και SVM**. Για την αξιολόγηση της απόδοσής τους, πραγματοποιήθηκαν **δύο πειράματα**:

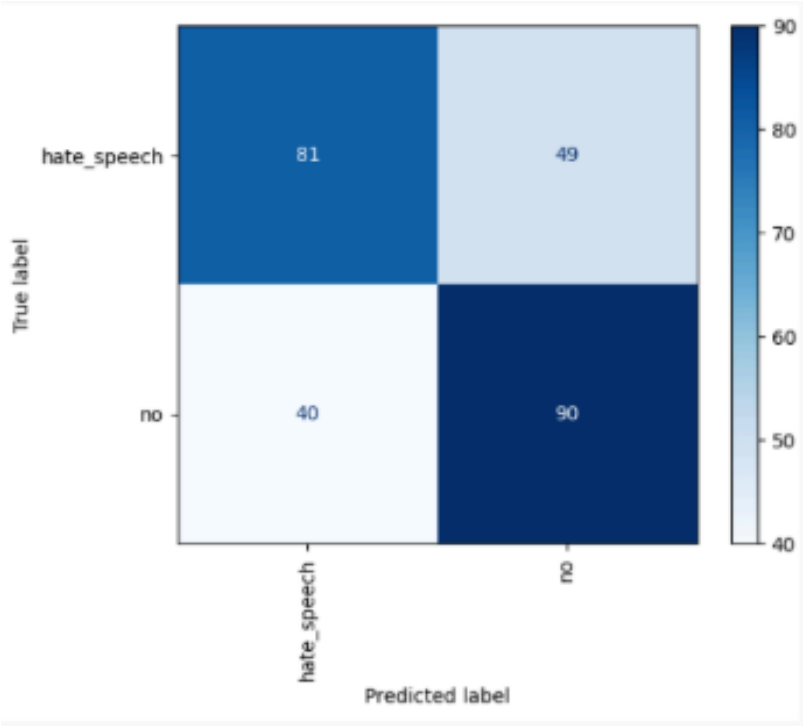
3. **Εκπαίδευση των μοντέλων χωρίς PCA**.
4. **Εκπαίδευση των μοντέλων με PCA**, μειώνοντας τη διάσταση των χαρακτηριστικών.

Επιλογή του αριθμού των Principal Components

Για να προσδιορίσουμε τον ιδανικό αριθμό χαρακτηριστικών μετά το PCA, αρχικά δημιουργήθηκε ένα Scree Plot (εικόνα 1).



Το μεγαλύτερο f1 score= 0.65 σημειώθηκε στις 250 συνιστώσες, το οποίο όμως δεν ξεπέρασε το f1 score του μοντέλου χωρίς PCA 0.66



Επιπλέον Βελτιώσεις

1. Επέκταση του Dataset

Παρόλο που το dataset περιλαμβάνει ελληνικές τηλεοπτικές εκπομπές και δημόσιες συζητήσεις, το μέγεθός του παραμένει περιορισμένο. Η συλλογή περισσότερων δεδομένων μπορεί να βελτιώσει την απόδοση του μοντέλου και να μειώσει την πιθανότητα overfitting.

2. Χρήση Ήχου για Ανάλυση Φωνητικών Χαρακτηριστικών

Η έμμεση ρητορική μίσους συχνά εκφράζεται μέσω ειρωνείας, σαρκασμού, τόνου φωνής και συναισθηματικής φόρτισης, που δεν είναι εμφανή στο απλό κείμενο.

- Χρήση του PyAudioAnalysis για εξαγωγή χαρακτηριστικών ήχου:
    - Pitch & Intonation: Εντοπισμός ειρωνείας μέσω αλλαγών στον τόνο φωνής.
    - Speech Rate: Ανάλυση της ταχύτητας ομιλίας για ανίχνευση συναισθηματικής φόρτισης.
    - Energy & Amplitude Variability: Εντοπισμός έντασης και stress στην ομιλία.
- Fusion μεταξύ text και speech classifiers:
    - Συνδυασμός των αποτελεσμάτων του μοντέλου ανάλυσης κειμένου και του μοντέλου ανάλυσης ήχου μέσω voting για πιο ακριβή ταξινόμηση.

3. Βελτίωση της Χρήσης του SMOTE

Το SMOTE βελτίωσε τις επιδόσεις του μοντέλου, αλλά η εφαρμογή του χρειάζεται προσοχή για να αποφευχθεί η εισαγωγή θορύβου. Αντ'αυτού, μελλοντικά, μπορεί να χρησιμοποιηθεί το Adaptive Synthetic Sampling (ADASYN), που δημιουργεί συνθετικά δείγματα πιο κοντά στη φυσική κατανομή των δεδομένων. Επίσης, θα μπορούσαν να δοκιμαστούν και διαφορετικές τεχνικές, όπως το Cluster-Based SMOTE, που δημιουργει συνθετικά δεδομένα με βάση centroids, αντί για τυχαία επιλογή γειτόνων, ή το KNN-based undersampling, που διαγράφει περιττά σημεία της υπερέχουσας κλάσης, χωρίς να χαθεί πληροφορία.

4. Βελτίωση των Κατηγοριών και Επανασχεδιασμός του Annotation

Η υπάρχουσα ταξινόμηση περιλαμβάνει τις κατηγορίες: Hate Speech, Racism, Sexism, Irony και No (ουδέτερο κείμενο).Ωστόσο, η ειρωνεία δεν αποτελεί από μόνη της μια μορφή ρητορικής μίσους, αλλά μπορεί να συνυπάρχει με άλλες κατηγορίες (π.χ., "Sexism + Irony"). Μια πιθανή βελτίωση είναι η εφαρμογή Multi-Label Classification. Αντί να ταξινομείται κάθε utterance σε μία μόνο κατηγορία, να επιτρέπεται να έχει πολλαπλές ετικέτες (π.χ. ένα σχόλιο μπορεί να είναι και "racist" και "ironic"). Κατ'επέκταση και η χρήση multi-label classifiers όπως BERT με sigmoid activation αντί για softmax, θα ήταν αρκετά βοηθητική.