



# **Anxiety / Stress Recognition in Texts**

A MACHINE LEARNING APPROACH

Αχιλλέας Οικονομόπουλος

# Dreaddit Dataset

**Reddit post corpus για stress recognition.**

- Pre-annotated
- Binary & Balanced

**~3500 Raw text samples**

- 2800 training - 700 testing

**Επιλεγμένα από συγκεκριμένα communities:**

- Abuse
- PTSD
- Finance
- Anxiety
- Relationships



# Tokens

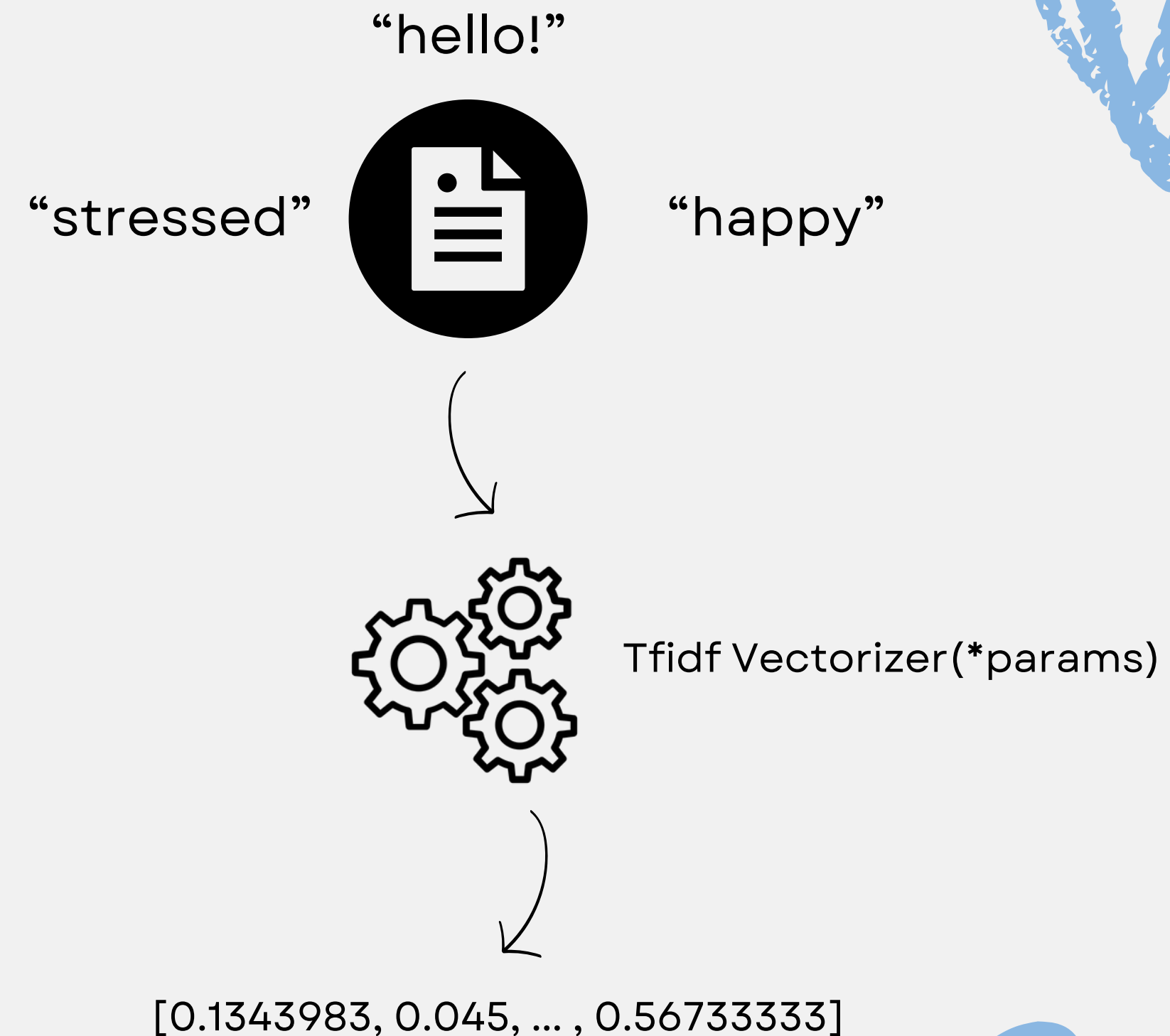
## Document -> tokens

- Σπάμε το κείμενο σε λέξεις.

## Tokenization rules

- Καθορίζουν πώς θα σπάσουμε το κείμενο
- Ευκαιρία για περαιτέρω επεξεργασία του κάθε token ξεχωριστά

**Features:** Tfidf scores (L2 Normalized)



# Η ιδέα του project

01

## Static Preprocessing

- Time Intensive διαδικασίες
- Υπολογίζουμε 1 φορά και αποθηκεύουμε τα αποτελέσματα

02

## Active Processing

- Επεξεργασία tokens πριν μετατραπούν σε features
- Στόχος η δημιουργία informative vocabulary

03

## Validation

- K-fold (5 folds)
- Logistic Regression, Multinomial Naive Bayes, XGBoost
- Παρακολούθηση των αποτελεσμάτων και tuning του Active Processing

04

## Evaluation

- Δοκιμή του feature set
- Παραγωγή μετρικών (macro f1) και γραφημάτων (Confusion Matrices, ROC curves)

# Static Preprocessing

## Ανάγκη για spell checking

- Τα social media posts δεν είναι και τα πιο καθαρογραμμένα
- “stressed” vs “stresed” -> ίδιο token
- ~10 λεπτά για όλο το corpus

## Worrywords Lexicon score (National Research Council of Canada)

- 44.000 λέξεις
- Λέξη -> βαθμός άγχους [-3,3]
- **feature:**  $\text{avg\_lexicon\_score}(\text{document}) = \frac{\text{sum}(\text{token lexicon scores})}{\text{count}(\text{lexicon words in document})}$
- ~5 λεπτά για το training set

# Active Preprocessing

## Tfidf Vectorizer Parameters

```
vectorizer_params = {  
    # "stop_words": "english", # SKLearn's default stopwords removal  
    "min_df": 3,                # min no. of documents allowed  
    "max_df": 0.2,              # max % of documents allowed  
    "max_features": 3000,        # retains X best performing tokens  
    "ngram_range": (1,3),        # 1 token can be 1-3 words  
    "tokenizer": custom_tokenizer # custom regex & stemming  
}
```

## Stopword Removal Tradeoff

- ✓ Πιο informative vocabulary & Καλύτερο generalization
- ✗ Ίσως κάποιες από αυτές (π.χ: can't) έχουν μεγάλη σημασία στο context μας, ειδικά με τη χρήση n-grams

# Active Processing

## Custom Tokenizer

### Default Sklearn Regex

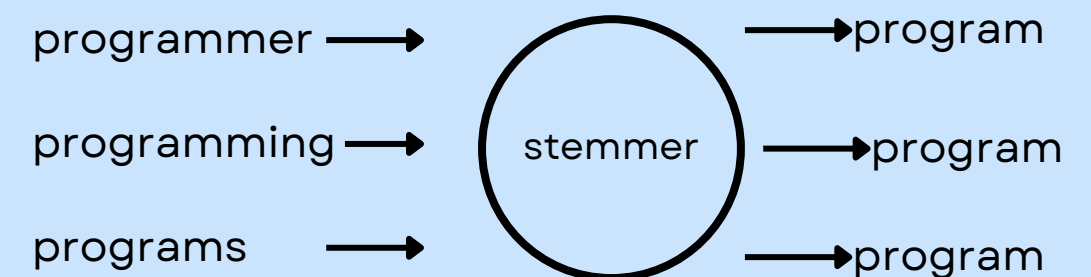
**$\geq 2$  alphanumeric στη σειρά**

- ✓ worry, happy, ve, 000zukor, α3
- ✗ i, i'm, i've, can't

### Προσθήκες

1. Αγνόησε tokens με αριθμό: (27f), 28m
2. Λάβε υπόψη αποστροφους: can't, won't
3. I | I'm | i'm | i

### Stemming






# Studied Models

## **Multinomial Naive Bayes & Logistic Regression**

- Η πιο απλοί και γρήγοροι classifiers
- Καλή απόδοση για textual δεδομένα

## **SVM (RBF Kernel) & XGBoost**

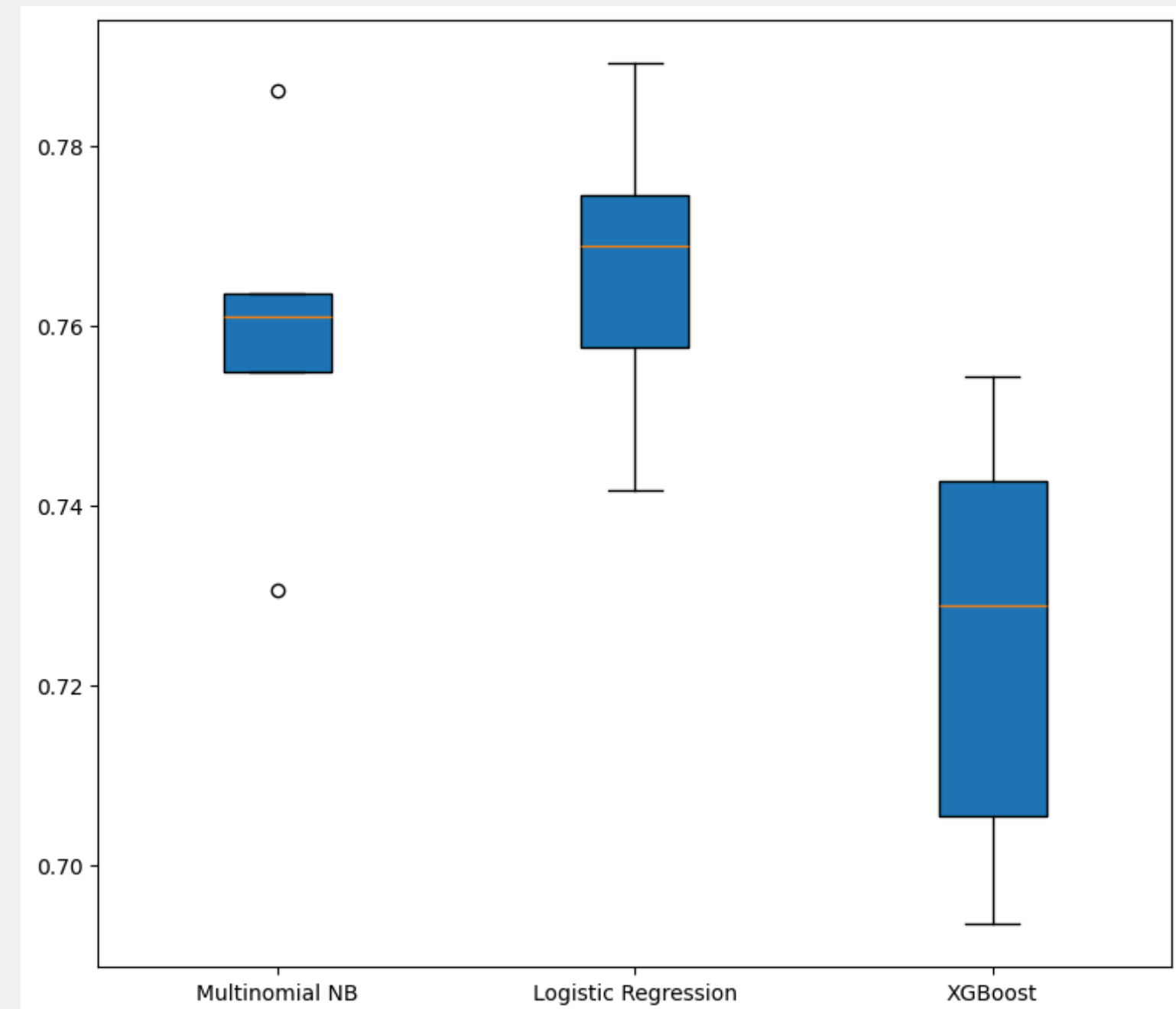
- Σημαντικά πιο αργοί αλλά,
  - Διαχείριση non-linear relationships των features
    - Το context παίζει μεγάλο ρόλο
    - “i’m stressed” vs “i’m stressed... not”
- 



# K-Fold Cross Validation

- 5 folds
- “Wrapper Methods”
  - Multinomial NB, LR, XGB
- Vocabulary Refinement
- Προσοχή για data leakage μεταξύ train / validation splits

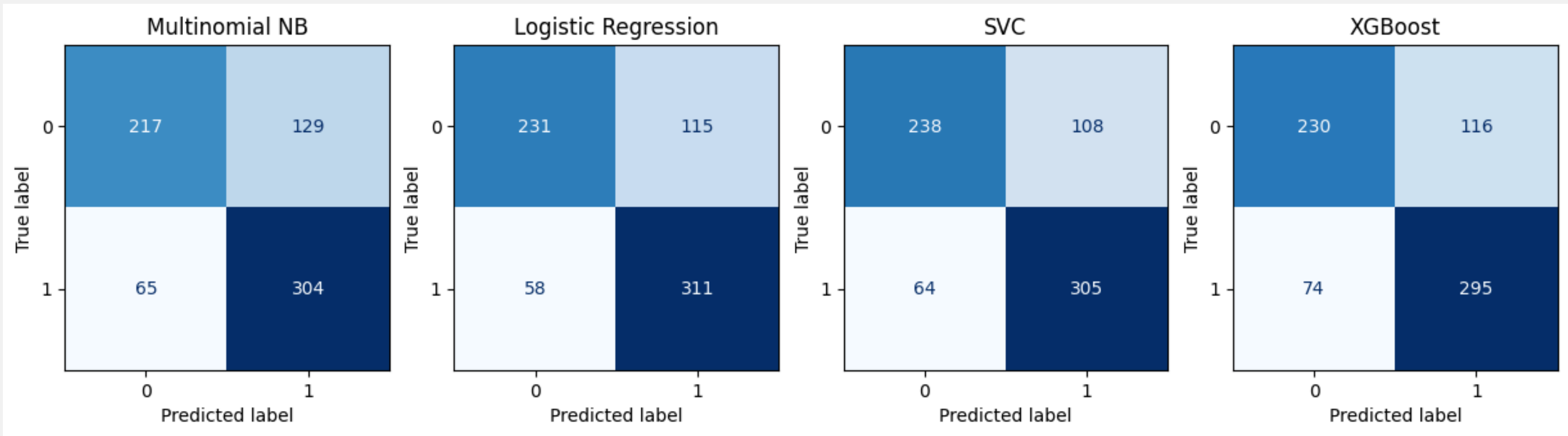
	Classifier	Mean	Variance
0	Multinomial NB	0.759195	0.000317
1	Logistic Regression	0.766354	0.000255
2	XGBoost	0.724950	0.000512



# Evaluation: Scores

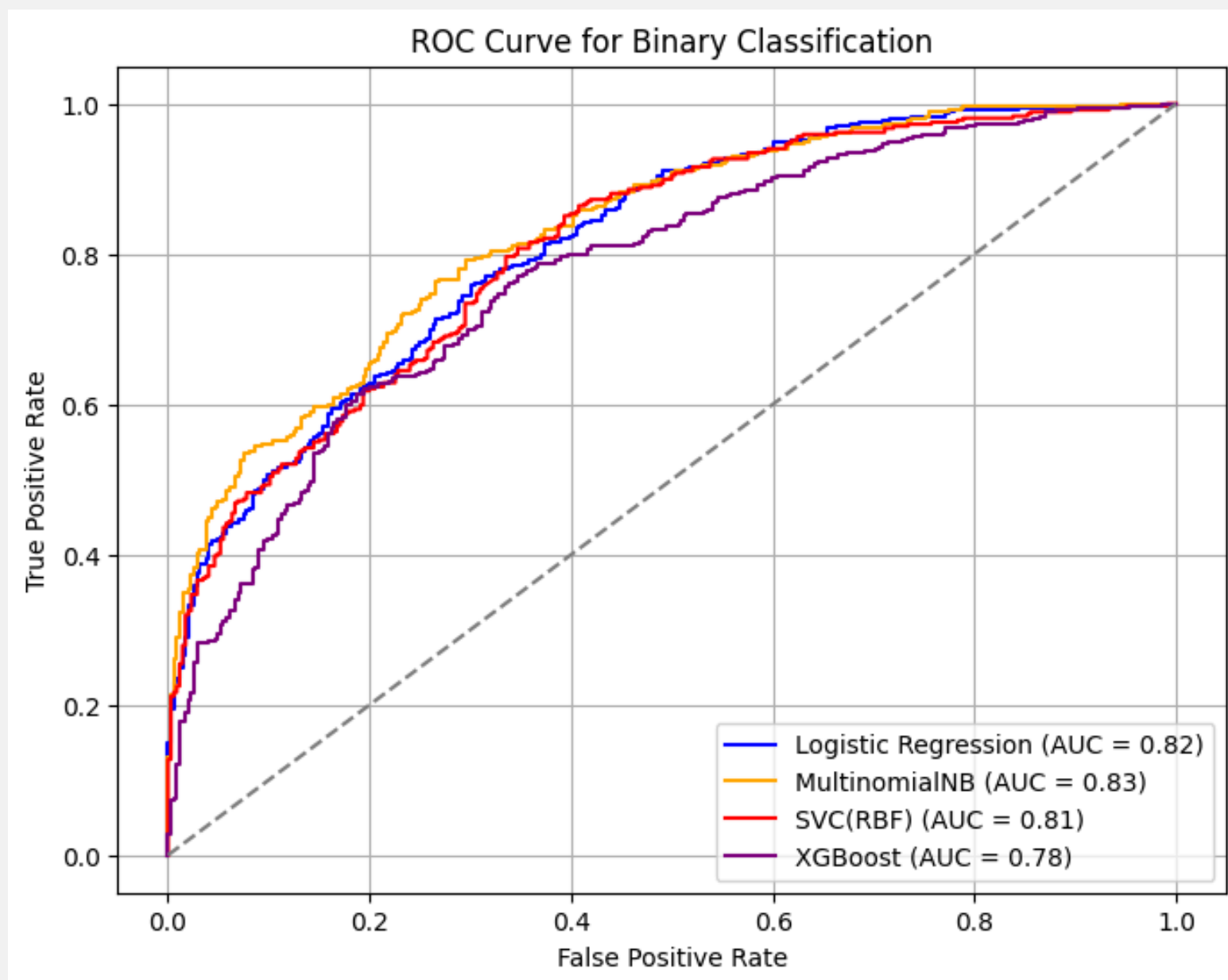
Classifier	Starting macro f1	Refined macro f1
Multinomial NB	<b>0.5960</b>	<b>0.7245</b>
Logistic Regression	0.7209	0.7549
SVM(RBF)	0.7203	0.7573
XGBoost	0.7118	0.7320

# Evaluation: Conf. Matrices

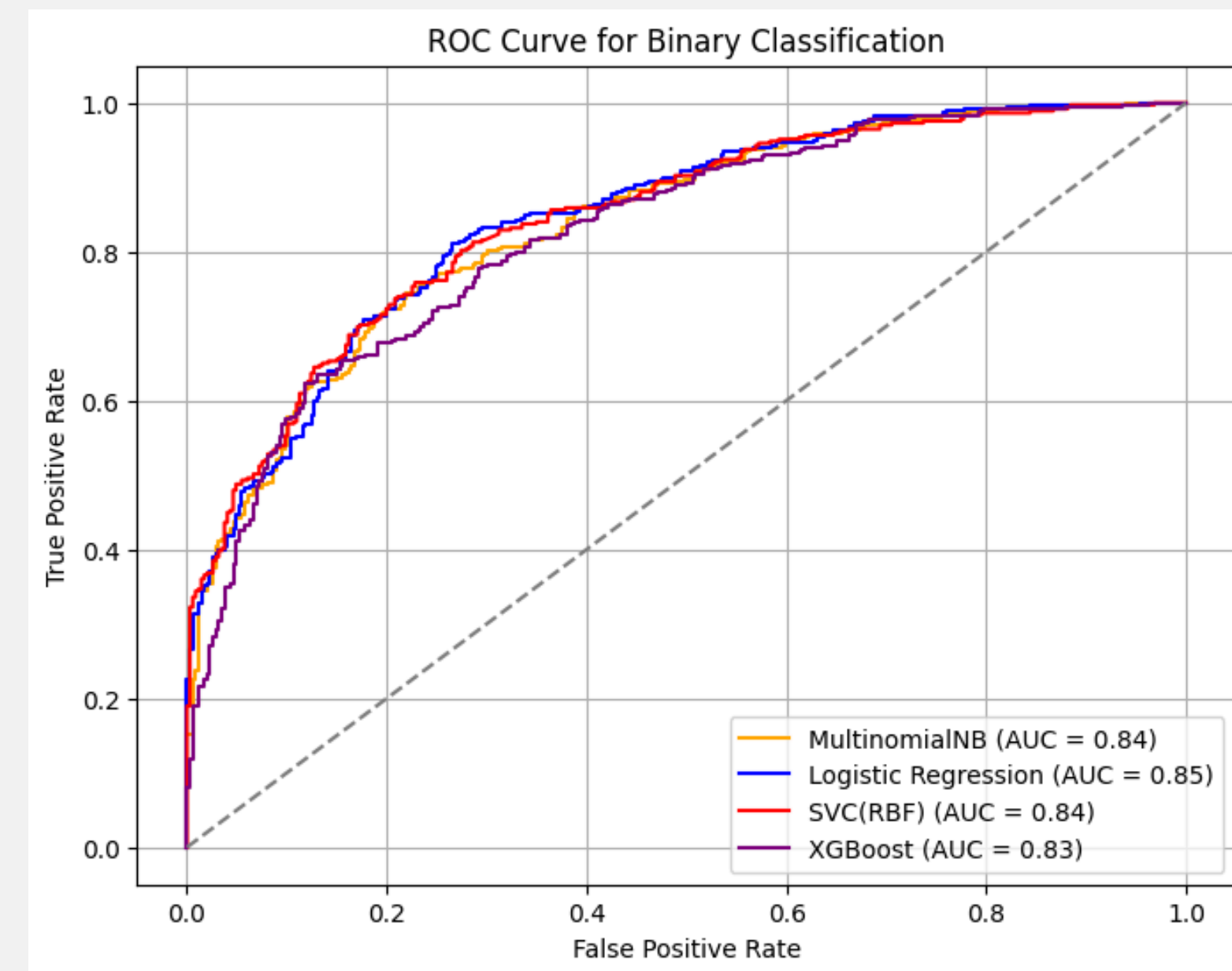


# Evaluation: ROC / AUC

Before



After






# Συμπεράσματα

**Το Logistic Regression μας καλύπτει**

**Πολλές παράμετροι που παίζουν ρόλο στο vocabulary refinement**

- Το lexicon feature μακράν το πιο αποδοτικό
- Πολύ σημαντική η γνώση του domain του προβλήματος

**Next steps?**

- **POS Tagging:** features βασισμένα στα % ρημάτων, επιθέτων, ουσιαστικών
  - Model hyperparameter tuning (SVM και XGBoost)
  - Vocabulary Refinement (v2) -> εισαγωγή domain expertise
- 

The background is a light gray color, decorated with various hand-drawn blue doodles. These include several overlapping circles and loops at the top, a series of concentric arcs at the bottom left, a wavy line at the bottom center, and several checkmarks at the bottom right. On the far right edge, there are some vertical blue strokes. The central text is in a bold, black, sans-serif font with a white drop shadow.

**Thank you  
very much!**