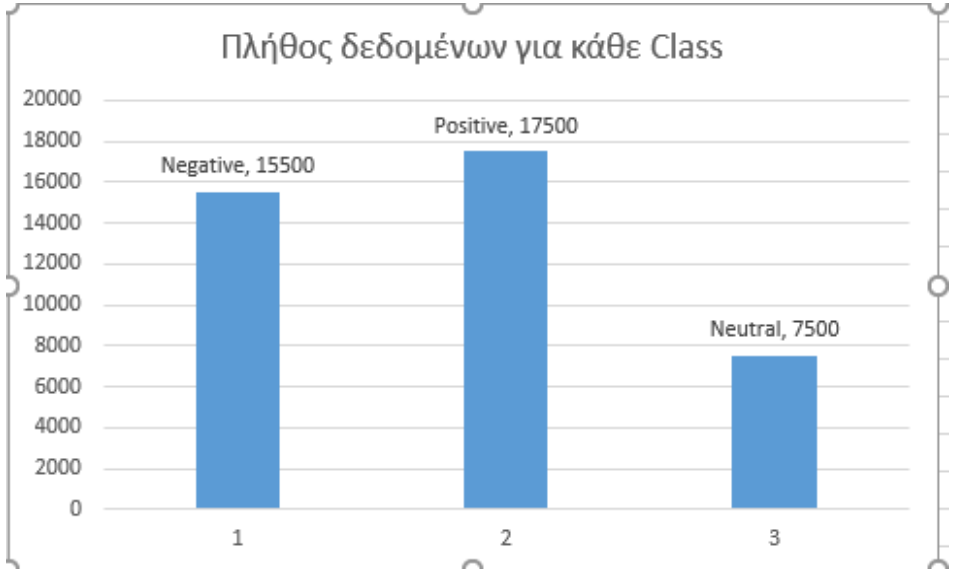
Sentiment Analysis on Covid19 Tweets
Multi-Classification Task

Πετρόπουλος Παναγιώτης

panos.petr1@gmail.com

Δεδομένα

- Τα δεδομένα βρέθηκαν στο Kaggle σε μορφή CSV.
- Πρόκειται για tweets, άρα texts/Documents.
- Ένα train CSV με 41157 samples & ένα τεστ CSV το οποίο χρησιμοποιήθηκε μόνο στο testing on Unseen Data με 3798 samples.
- Έχουμε 3 πιθανά class labels : Negative, Neutral & Positive.



Επεξεργασία Δεδομένων

Για να προκύψουν τα τελικά Features με τα οποία θα εκπαιδευτεί ο Classifier έγιναν τα παρακάτω βήματα:

- Lookup σε custom dictionaries για την αντικατάσταση ή διαγραφή λέξεων/emojis/emoticons
- Διαγραφή θορύβου, όπως urls, hashtags, mentions, numbers etc. Στην περίπτωση των hashtags αφαιρέθηκε απλά το σύμβολο #. Αυτό διότι μπορεί κάποιο από τα hashtags να προσφέρει κάποιο νόημα ως προς το συναίσθημα. Π.χ. #love -> love
- Tokenization & αφαίρεση stopwords.
- Part of speech tagging για να το χρησιμοποιήσουμε στην διαδικασία του Lemmatization.
- Lemmatization.
- Αποθήκευση των Cleaned/processed data σε DataFrame.
- Παραγωγή uni-grams & bi-grams:
    - Για παράδειγμα: Πρόταση: 'I don't like'
        - unigrams: I, do, not, like
        - bi-grams: I do, do not, not like Όπως βλέπουμε η λέξη Like απο μόνη της δείχνει συνήθως θετικό συναίσθημα ένω με την λέξη Not μπροστά δείχνει αρνητικό.
- Μετατροπή των καθαρών πλέον κειμένων σε Vectors , με την διαδικασία του TF-IDF.

$$\text{Score}(d,q) = \sum_{t \in q \cap d} \text{tf}_{t,d} \times \text{idf}_t$$

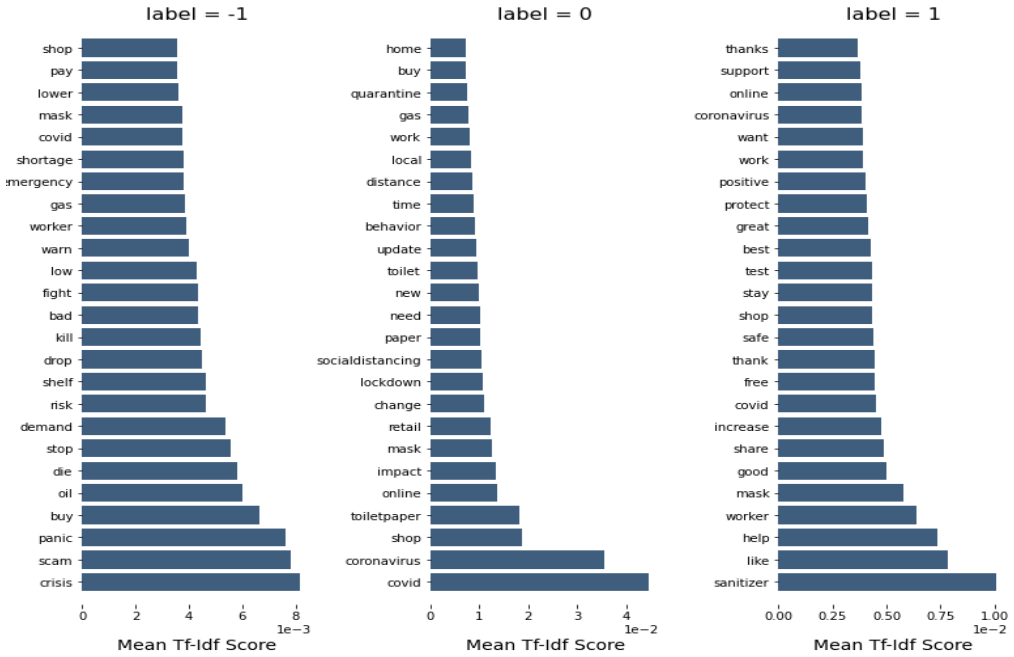
Μείωση Διαστάσεων

Μετά την εφαρμογή της διαδικασίας TF-IDF οι διαστάσεις του Vector που αποτελεί την είσοδο στους Classifiers είναι (41157,20000). Δηλαδή περιέχει 20000 features.

Για την μείωση εφαρμόστηκε η X^2 test στατιστική μέθοδος στην οποία ως threshold για την απόρριψη της H0 ορίστηκε 1-p = 5% οτιδήποτε μεγαλύτερο του 1-p το κρατάμε.

Όπου p = score που αντιστοιχεί σε πιθανότητα ένα feature να βρίσκεται σε κάποια κλάση.

Επομένως, έγινε ανάκτηση των πιο συνηθισμένων feature όπου είναι πιο σχετικά με την κάθε κλάση. Οι τελικές διαστάσεις πλέον είναι: (41157, 1638)



Όσο πιο μεγάλο tf-idf score τόσο πιο μεγάλη διακριτική δύναμη έχει το Feature. Δηλαδή εμφανίστηκαν πολλές φορές σε λίγα documents.

```
# -1:
 . selected features: 900
 . top features: bad,buy,crisis,fear,oil,panic,panic buy,sanitizer,scam,stop,covid crisis,kill,help,low,die,warn,selfish,co
virus crisis,shortage,stop panic

# 0:
 . selected features: 646
 . top features: toiletpaper,panic,like,demand,help,crisis,panic buy,coronavirus,stop,safe,thank,coronavirus toiletpaper,su
t,behavior,great,impact,retail closure,care,low,free

# 1:
 . selected features: 798
 . top features: best,crisis,free,great,help,like,panic,panic buy,safe,sanitizer,support,thank,love,friend,scam,thanks,care
y safe,hero,share
```

Classifiers

- Οι παράμετροι των αλγορίθμων επιλέχθηκαν με την διαδικασία του GridSearchCV το οποίο κάνει και ταυτόχρονα Kfold Cross-Validation με K=5.
- Για τα πειράματα εκπαιδεύτηκαν:
  - SVM : RBF kernel, C=1.5, gamma=scale=$\frac{1}{n \times \text{Διακύμανση } (\sigma^2)}$
  - Logistic Regression: solver=newton-cg, C=10, max_iter=200
  - Naïve Bayes: alpha=0.70001

Ορισμοί

- **Recall:** Πρακτικά είναι το ποσοστό των σωστών προβλέψεων από όλα τα actual της κάθε κλάσης. P(Predicted|Actual)
- **Precision:** Πρακτικά είναι το ποσοστό των σωστών προβλέψεων από τα predicted της κάθε κλάσης. P(Actual|Predicted)

Παράδειγμα: Περίπτωση test PCR Covid-19. Μας ενδιαφέρει το Recall, διότι δεν θέλουμε το μοντέλο να κάνει λάθος και να μας «υποδείξει» ένα θετικό κρούσμα ως ένα αρνητικό (αυτό μπορεί να φανεί από το precision). Αυτό που θέλουμε πρακτικά είναι να μας δώσει για το αποτέλεσμα του τεστ έστω και λάθος απάντηση. Προτιμάμε δηλαδή ένα False Alarm (False Negative).

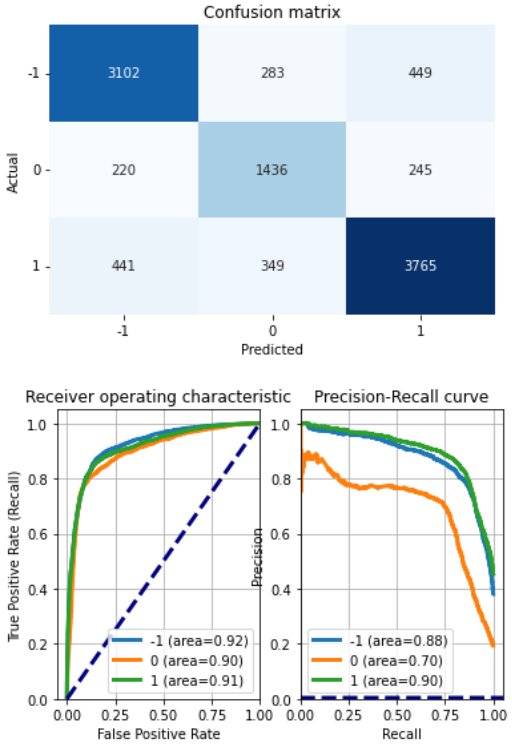
- **F1-score:** Εδώ θα χρησιμοποιήσουμε το macro F1. Πρακτικά το χρησιμοποιούμαι για να «τιμωρήσουμε» την κακή επίδοση (αν υπάρχει) σε ένα από τα 2 precision & recall. Έτσι αναζητούμε με αυτό μια ισορροπία μεταξύ precision & recall, δηλαδή το F1 θα βρίσκεται πιο κοντά στην μικρότερη από τις 2 τιμές (precision, recall). Για παράδειγμα πολλές φορές εστιάζουμε στα False Positives & False Negatives. Εδώ έρχεται να μας εξηγήσει πιο πολλά πράγματα το F1 σε σχέση με την Accuracy.
- **ROC:** Υπολογίζεται με βάση το TPR & FPR.

  TPR: πόσα από τα θετικά (TP + FN) ταξινομεί σωστά. Δηλαδή το Recall.

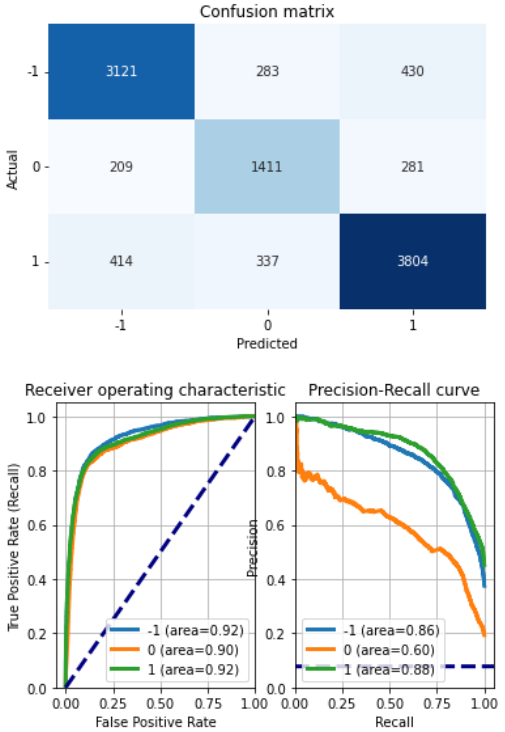
  FPR: πόσα από τα αρνητικά (TN + FP) ταξινομεί λάθος.

Metrics(1/2)

SVM

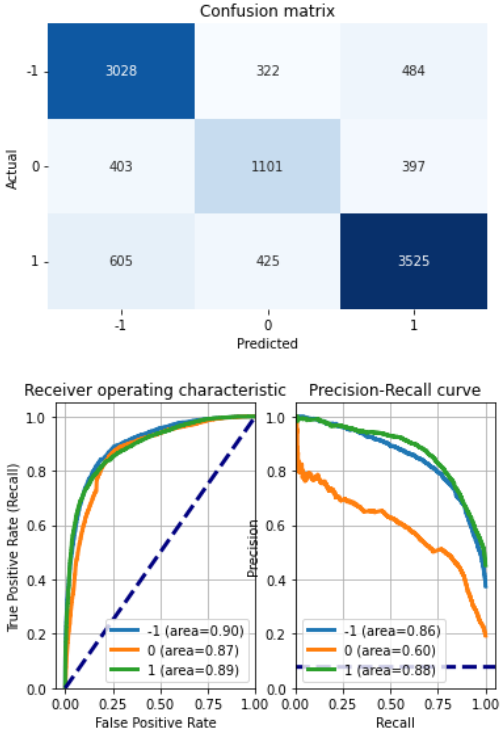


LR




```
Accuracy: 0.802 (+/- 0.003) [SVM]
Accuracy: 0.741 (+/- 0.003) [Naive Bayes]
Accuracy: 0.803 (+/- 0.001) [Logistic Regression]
```

Naïve Bayes



Metrics(2/2)

SVM

	precision	recall	f1-score	support
-1	0.82	0.82	0.82	3803
0	0.73	0.70	0.72	1971
1	0.83	0.84	0.83	4516
accuracy			0.81	10290
macro avg	0.79	0.79	0.79	10290
weighted avg	0.81	0.81	0.81	10290

LR

	precision	recall	f1-score	support
-1	0.81	0.83	0.82	3737
0	0.74	0.70	0.72	2009
1	0.84	0.84	0.84	4544
accuracy			0.81	10290
macro avg	0.80	0.79	0.79	10290
weighted avg	0.81	0.81	0.81	10290

Naïve Bayes

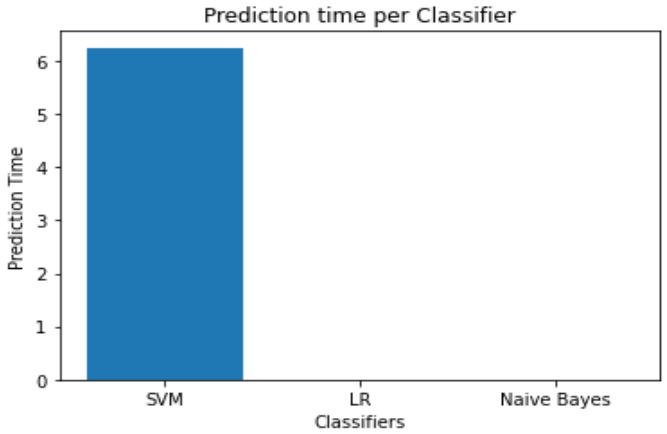
	precision	recall	f1-score	support
-1	0.79	0.76	0.77	4017
0	0.58	0.60	0.59	1829
1	0.78	0.80	0.79	4444
accuracy			0.75	10290
macro avg	0.72	0.72	0.72	10290
weighted avg	0.75	0.75	0.75	10290

	Models	macro-F1 scores	avg Accuracies	Precision	Recall
0	SVM RBF	0.791871	0.8069	0.787698	0.797011
1	Naive Bayes	0.714535	0.743829	0.715357	0.714273
2	Logistic Regression	0.793403	0.810107	0.790286	0.797133

Χρόνοι κατά το Prediction

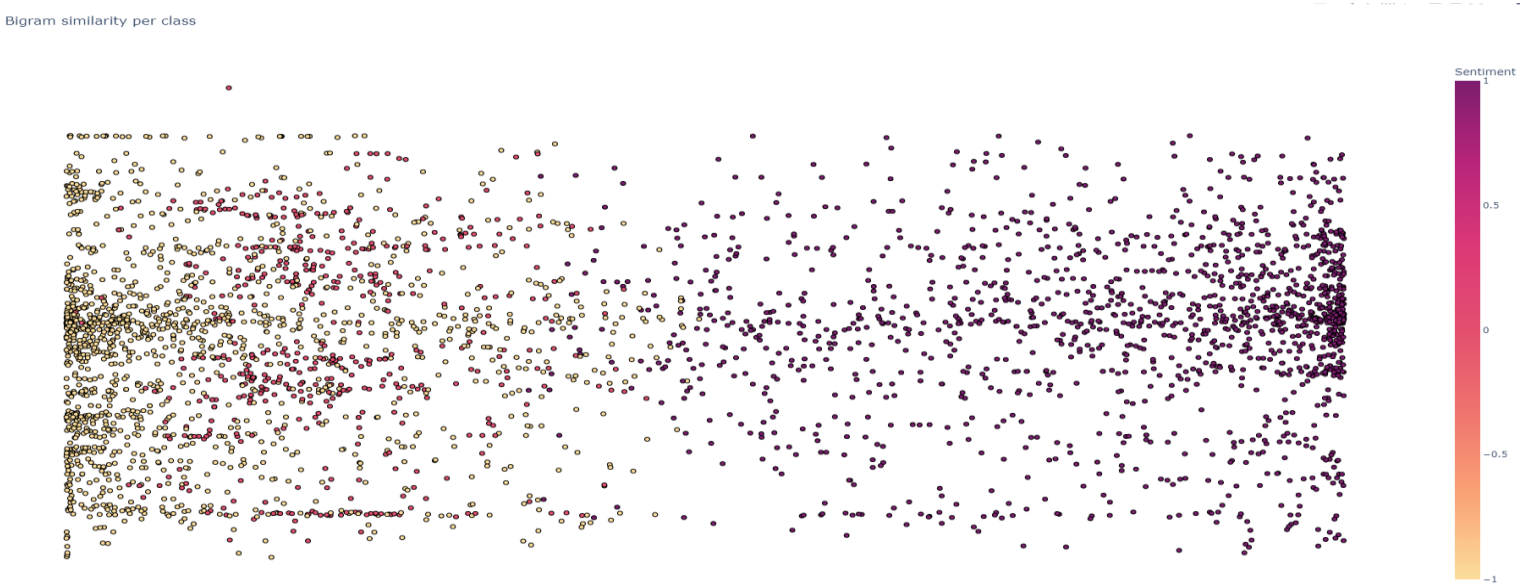
Πραγματοποιήθηκε διαδικασία prediction στα unknown Data.

- SVM Duration: 6.2 seconds
- LR Duration: 0.00092 seconds
- Naïve Bayes Duration: 0.0010 seconds



Prediction

- T-SNE:
  1. Χρησιμοποιεί ευκλείδεια απόσταση σημείων.
  2. Διατηρεί τις συσχετίσεις των σημείων. Δηλαδή αν το A συσχετίζεται με το B στο high Dimensional space το ίδιο θα ισχύει και στο low Dimensional.
  3. Κάνει και για μη-γραμμικά χωρισμένα σημεία.
  4. Αυτή η μέθοδος βρίσκει την πιθανότητα ένα Feature να διαλέξει κάποιο άλλο σαν γείτονα.



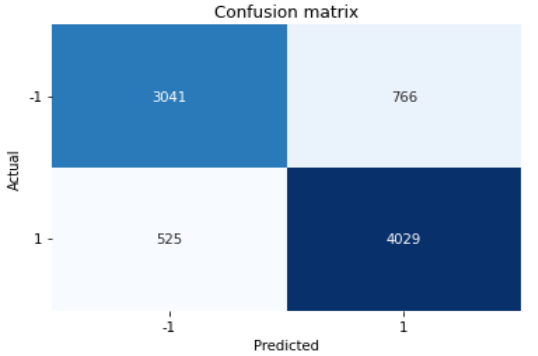
Binary Classification Task

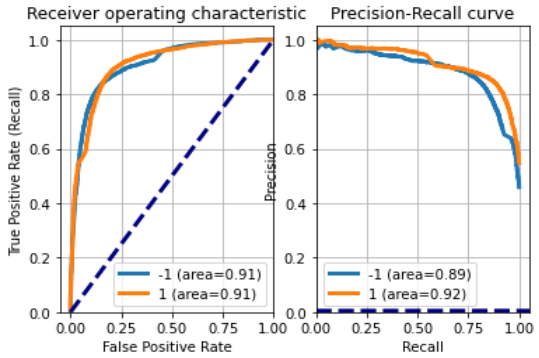
- Πραγματοποιήθηκε στα ίδια δεδομένα, ώστε να μπορεί να προβλέπει μόνο το Negative & Positive.
- Ίδια επεξεργασία.

	Models	macro-F1 scores	avg Accuracies	Precision	Recall
0	SVM RBF	0.843406	0.845593	0.846513	0.841754
1	Naive Bayes	0.815847	0.816529	0.815413	0.817616
2	Logistic Regression	0.849679	0.851812	0.852987	0.847937

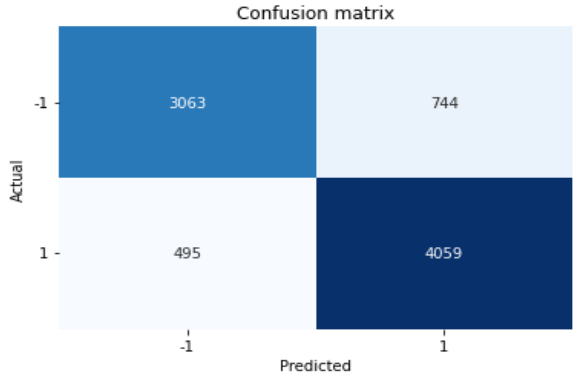
Metrics

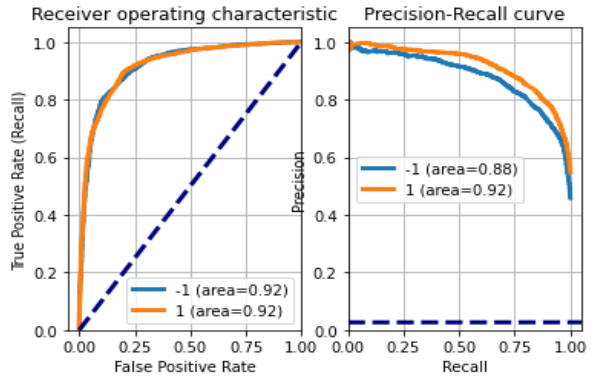
SVM



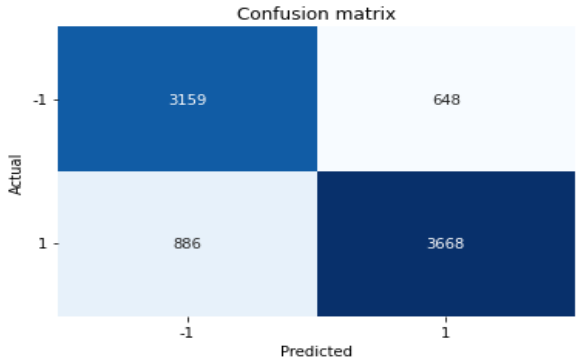


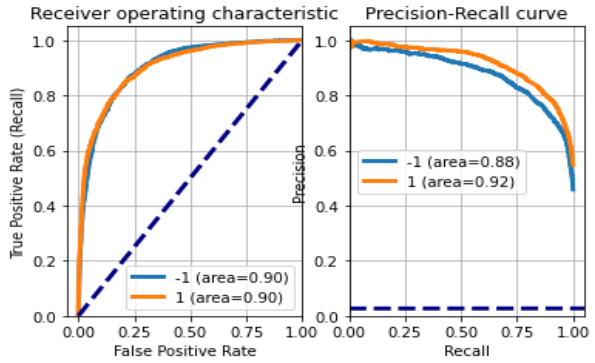
LR





Naïve Bayes






```
Accuracy: 0.844 (+/- 0.004) [SVM]
Accuracy: 0.818 (+/- 0.005) [Naive Bayes]
Accuracy: 0.846 (+/- 0.004) [Logistic Regression]
```

T-SNE for Binary Task