Υπολογιστική Γλωσσολογία και Επεξεργασία Φυσικής Γλώσσας και Επεξεργασία και Διαχείριση Γλωσσικών Πόρων

Demented speech classification (via text)

Καθηγητής: Χάρης Παπαγεωργίου

Φοιτήτριες:

Φωτεινή Ηλιάδου ΑΜ: 7115182200006

Άρτεμις Χαρδούβελη ΑΜ: 7115182200026

Νίκη Αποστολοπούλου ΑΜ: 7115182200002

Περίληψη - Εισαγωγή:

Στόχος της παρούσας εργασίας είναι η δημιουργία ενός binary classifier, που θα διακρίνει τα data μας σε δύο κατηγορίες. Το dataset που χρησιμοποιήσαμε είναι λεκτικές περιγραφές της εικόνας "η κλοπή του μπισκότου" από υγιή πληθυσμό και από πληθυσμό πασχόντων με άνοια. Ακριβώς αυτές είναι και δύο κατηγορίες, όπου ταξινομήσαμε τα data: σε dementia και non dementia. Για αυτόν τον σκοπό, ξεκινήσαμε κάνοντας text preprocessing, όπου προχωρήσαμε σε δύο μονοπάτια, με και χωρίς lemmatization. Για κάθε μονοπάτι αφού χωρίσαμε τα δεδομένα μας σε train και test set προχωρήσαμε σε feature extraction χρησιμοποιώντας δύο διαφορετικούς τρόπους. Το επόμενο βήμα ήταν να εκπαιδεύσουμε πέντε επιλεγμένα μοντέλα από τα οποία θα καταλήξουμε στα αποτελέσματα με βάση τα accuracy scores και τα confusion matrices τους. Η παραπάνω διαδικασία ακολουθείται και για τα δύο προαναφερθέντα μονοπάτια.

Προς επιβεβαίωση των παραπάνω αποτελεσμάτων θα χρησιμοποιήσουμε εφαρμογή text to image για να δούμε, αφού δημιουργηθεί εικόνα από το εκάστοτε δοσμένο text, αν αυτή ταιριάζει/ μοιάζει με την αρχική εικόνα πάνω στην οποία βασίστηκε η διαλογή των data.

Subject/ Task	Classification in Dementia/Non-Dementia
Type of Data	Textual
Models	Naive-Bayes, Logistic Regression (LR), Support Vector Machines (SVM), Random Forest, Artificial Neural Network (ANN)

Μεθοδολογία:

1. Περιγραφή Dataset

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των classifiers συλλέχθηκαν δια ζώσης και παραχωρήθηκαν για την εκπόνηση αυτής της εργασίας από την εταιρεία *Langaware Inc*, εφόσον τα δεδομένα δεν πρόκειται να τεθούν στην διάθεση του κοινού. Πρόκειται για περιγραφές δοκιμασίας εκμαίευσης λόγου και συγκεκριμένα της νέας εκδοχής της εικόνας "η κλοπή του μπισκότου" από υγιή πληθυσμό και από πληθυσμό πασχόντων με άνοια. Στην αρχική τους μορφή τα δεδομένα ήταν προφορικά. Στη συνέχεια, μεταγράφηκαν πιστά και αφαιρέθηκαν παρεμβολές από τον λόγο του ερευνητή. Στη διάθεσή μας είχαμε το κείμενο των προφορικών περιγραφών και

την ετικέτα της διάγνωσης των συμμετεχόντων σε ασθενείς και υγιείς (ομάδα ελέγχου). Στο σύνολο πρόκειται για 126 κείμενα εκ των οποίων τα 96 αποτελούν δείγματα συμμετεχόντων, που πάσχουν με άνοια και τα 73 δείγματα συμμετεχόντων χωρίς γνωστική έκπτωση.

Εικόνα_1: Η κλοπή του μπισκότου (Stealing Cookies)

Για την υλοποίηση του στόχου μας, έπρεπε να τροποποιήσουμε τα δεδομένα μας με τέτοιο τρόπο, ώστε να έχουμε ίσα δείγματα ανοϊκών και υγιών συμμετεχόντων. Για την επίτευξη αυτού, επιλέξαμε downsampling αλλά δεδομένου ότι δεν έχουμε μεγάλο όγκο δεδομένων, καταλήξαμε να κάνουμε oversampling. Επίσης, επειδή, ακριβώς δεν είναι τόσο μεγάλος ο αριθμός των data αναμένουμε να μην έχουμε φαινόμενα overfitting.

2. Data / Text processing

Stopwords

Κατά το text preprocessing, όσον αφορά τον συνήθη διαχωρισμό των λέξεων ενός κειμένου σε stopwords και content words, και την συνακόλουθη απομάκρυνση των stopwords από το dataset, κατόπιν συζήτησης για την φύση των δεδομένων μας, αποφασίστηκε να μην ακολουθηθεί τέτοια διαδικασία. Τα stopwords αυτά αποτελούνται από τις λεγόμενες λειτουργικές λέξεις (function words), δηλαδή τις προθέσεις, τα άρθρα και γενικά όλες εκείνες τις βοηθητικές λέξεις, που όμως δεν περιλαμβάνουν το νόημα του εκάστοτε κειμένου. Στην περίπτωσή μας, λόγω της φύσης του προβλήματος θεωρήθηκε ότι θα ήταν χρήσιμα. Ουσιαστικά αποτελούν κομμάτι του *προβλήματος* άρα και της επίλυσής του. Στον διαχωρισμό μεταξύ σημαντικών και μη σημαντικών λέξεων, πάντα αναφερόμενες στο συγκεκριμένο *πρόβλημα*, θα κατατάσσαμε τα stopwords στις σημαντικές λέξεις, για το λόγο ότι αποτελούν λέξεις ικανές να μας δώσουν πολλά συμπεράσματα με βάση τη

χρήση ή τη μη-χρήση τους, καθώς και με την ποσότητα της χρήσης τους, την επανάληψη αυτών κ.ο.κ.. Γλωσσολογικά τα stopwords ισοδυναμούν με τις λειτουργικές λέξεις (έναντι των λέξεων περιεχομένου), ο οποίες μπορούν να προσφέρουν πολύτιμες πληροφορίες στη μελέτη του παθολογικού λόγου.

Stemming

Stemming είναι μια διαδικασία, που συντελείται στο preprocessing και αναφέρεται σε αφαίρεση χαρακτήρων από τις λέξεις -αφαίρεση των επιθημάτων/καταλήξεων- στην λέξη-ρίζα, ακόμη κι αν το ίδιο το δημιουργηθέν stem δεν αποτελεί από μόνο του λέξη. Απόρροια αυτού είναι μια μικρότερη μορφή, η βάσης της λέξης. Επιλέξαμε, λοιπόν, να μην προχωρήσουμε σε stemming λόγω της ιδιομορφίας που παρατήσαμε στο dataset μας. Συγκεκριμένα, στον λόγο των ανοϊκών ασθενών συχνά εμφανίζονται λέξεις παραφθαρμένες φωνητικά. Η μεταγραφή του ηχητικού αποσπάσματος είναι πιστή. Συνεπώς, είναι αναμενόμενο κατά το stemming αυτές οι λέξεις να μην λάβουν ρίζα, που να ανταποκρίνεται στην λέξη στόχο του συμμετέχοντα. Επιπροσθέτως, δεν είναι αναμενόμενο να χρησιμοποιούν όλοι οι πάσχοντες από άνοια τις ίδιες παραφθαρμένες λέξεις, οπότε, έστω ότι κάναμε stemming, θα έχουμε διαφορετικές ρίζες στον πληθυσμό των ασθενών.

Lemmatization

Άλλη μια διαδικασία που συντελείται κατά το preprocessing είναι αυτή του lemmatization. Το lemmatization μειώνει το μέγεθος της λέξης, κρατώντας το λήμμα, εξασφαλίζοντας κατάλληλα ότι η λέξη-ρίζα ανήκει στη γλώσσα, σε αντίθεση με το stemming. Η επιλογή που πάρθηκε ήταν να πειραματιστούμε σε αυτό το κομμάτι. Επιλέξαμε να αναλύσουμε το dataset με τα αντίστοιχα μοντέλα, μία χωρίς lemmatization (semi-preprocessing) και μία με lemmatization (preprocessing).

Punctuation - Μορφή χαρακτήρων - Tokenization

Και στα δύο μονοπάτια, που προήλθαν από την προηγούμενη απόφασή μας (με ή χωρις lemmatization), προβαίνουμε σε κάποιες ενέργειες εξομάλυνσης του κειμένου. Αφαιρούμε προηγουμένως τα σημεία στίξης (punctuation), ώστε να μην μετρηθούν ως ξεχωριστές λέξεις και επηρεάσουν τα αποτελέσματα. Μετατρέπουμε, επίσης, το σύνολο των text σε πεζά γράμματα (lowercase), για την μείωση των διαφορετικών λέξεων ενώ δεν είναι πραγματικά διαφορετικές, πόσω μάλλον λόγω του ότι επιλέξαμε να μην αφαιρέσουμε τα stopwords. Τέλος, προχωράμε σε word segmentation (tokenization).

3. Data split

Επιλέξαμε να χωρίσουμε τα data μας σε δύο set, train και test set, έναντι των τριών set, train - validation - test set, θεωρώντας ότι αν και πρόκειται για μια λιγότερο σωστή διαδικασία χωρισμού των δεδομένων, στην υλοποίηση προχωρήσαμε μόνο σε train και test set.

4. Feature extraction

Κατόπιν συζήτησης, για το πως θα μπορούσε να τρέχει καλύτερα ο εκάστοτε classifier μας, επιλέξαμε να προσθέσουμε περισσότερα στοιχεία ως features.

Αρχικά, ήμασταν ανάμεσα στις ακόλουθες μετρικές *rouge score, blue score* και *big5 personality traits*. Καταλήξαμε να χρησιμοποιήσουμε 2 εξ αυτών ως features, με το *big5 personality traits* να μην αποτελεί τελική επιλογή μας λόγω του ότι η εξαγωγή *big5* δεν είναι δυνατή στα ελληνικά, καθώς δεν υπάρχει Ελληνικό Λεξικό Συναισθημάτων (GSL) επισημειωμένο σύμφωνα με τις ετικέτες από το *big5*. Επιπλέον, δεν υπάρχουν ελληνικά δεδομένα επισημειωμένα σύμφωνα με τα *big5* που να αφορούν στο λεξιλόγιο που περιλαμβάνεται και παράγεται κατά την δοκιμασία εκμαίευσης της εικόνας της κλοπής του μπισκότου. Ο υπολογισμός αυτών των features έγινε πριν κάνουμε split τα data.

Bleu score (Bilingual Evaluation Understudy)

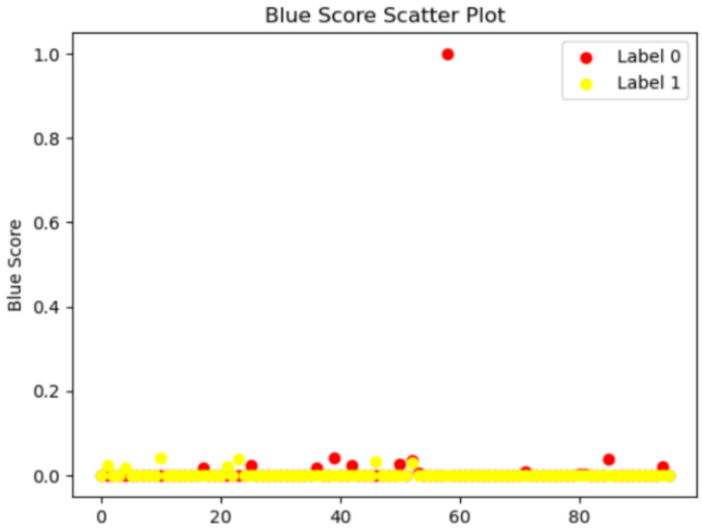
Αποτελεί μετρική για την ποιότητα των μοντέλων μηχανικής μετάφρασης. Αν και αρχικά είχε σχεδιαστεί μόνο για μοντέλα μετάφρασης, πλέον χρησιμοποιείται και για άλλες εφαρμογές επεξεργασίας φυσικής γλώσσας και έτσι το χρησιμοποιήσαμε και εμείς. Για τον υπολογισμό του blue score χρειαζόμαστε προτάσεις αναφοράς (reference) και υποψήφιες προς εξέταση προτάσεις (candidate). Στο δικό μας *πρόβλημα* θα θεωρήσουμε ως προτάσεις αναφοράς τις προτάσεις των υγιών, και ως υποψήφια πρόταση την εκάστοτε πρόταση από αυτές των ασθενών. Το bleu score συγκρίνει μια πρόταση με μία ή περισσότερες προτάσεις αναφοράς και δείχνει την ομοιότητα, το πόσο καλά δηλαδή ταιριάζει η υποψήφια πρόταση με την πρόταση αναφοράς. Δίνει μια βαθμολογία εξόδου μεταξύ 0 και 1. Όσο πιο μεγάλη είναι αυτή η βαθμολογία τόσο περισσότερη ομοιότητα υπάρχει. Γενικά, διαιρείται ο πληθικός αριθμός των προτάσεων που κάνουν matching με τον πληθικό αριθμό των προτάσεων που ειπώθηκαν από τον candidate, προσαρμοσμένο, όμως, στη συνθήκη του αν είναι πολύ μικρό το εκφώνημα του candidate “τιμωρείται” (brevity penalty). Αξίζει να αναφερθεί ότι τα συνώνυμα δεν θεωρούνται όμοια.

Συλλογιστική κατεύθυνση:

Κατά την προσπάθεια να εντάξουμε το blue score στα features ήρθαμε αντιμέτωπες με πολλούς σκοπέλους. Αρχικά, υπολογίσαμε το blue score για όλο το dataset. Προβήκαμε σε διάφορες συγκρίσεις: υγιείς (reference) VS υγιείς, υγιείς VS ασθενείς, ασθενείς VS ασθενείς. Όμως το matching γινόταν one-to-one στις περιγραφές με την σειρά, οπότε ανάμεσα στις ίδιες κλάσεις πχ. υγιείς (reference) VS υγιείς συνέκρινε το ίδιο κείμενο με τον εαυτό του και είχαμε απόλυτη ταύτιση (BLEU score: 1.0).

```
Blue score for healthy vs demented
BLEU score: 0.1304157290371472
<class 'float'>
Blue score for healthy vs healthy
BLEU score: 1.0
Blue score for demented vs demented
BLEU score: 1.0
```

Οπότε και η προσέγγιση αυτή εγκαταλείφθηκε. Και αποφασίσαμε να ορίσουμε να κάνει randomly τις συγκρίσεις. Κατά την ένταξη όμως του feature συνειδητοποιήσαμε ότι, αφού οι τιμές ήταν για όλο το dataset δεν προσφέρουν κάτι. Για αυτό και προσπαθήσαμε στην συνέχεια να εξαγάγουμε την μετρική αυτή για κάθε περιγραφή ξεχωριστά. Αυτή συγκρίνεται με ένα τυχαία επιλεγμένο κείμενο υγιούς συμμετέχοντα (reference).
Σε αυτό το σημείο υπήρξε η σκέψη να δημιουργήσουμε ενα threshold. Για όποια μέτρηση βρίσκεται πάνω από αυτό, θα κατατάσσεται η περιγραφή που της αντιστοιχεί σε αυτή των υγιών και αν είναι χαμηλότερη σε αυτή των ανοϊκων. Πήραμε, λοιπόν, την minimum μέτρηση των υγιών και την maximum των ασθενών (και τις μέσες τιμές). Όμως, κάνοντας τα variation plots παρατηρήσαμε ότι η διασπορά δεν ήταν η αναμενόμενη (σχετικά διακριτά) αλλά ήταν πολύ ανακατεμένα μεταξύ τους. Για αυτό και εγκαταλείφθηκε και αυτή η προσέγγιση και τα αποτελέσματα τα καταχωρήσαμε απλά ως features σε κάθε data point.



Βέβαια, σε αυτό το σημείο, πήραμε warning message ότι πρέπει να κάνουμε smoothing γιατί οι τιμές που παίρνουμε είναι πολύ μικρές και ετσι το υιοθετήσαμε.

```
C:\Users\artem\anaconda3\lib\site-packages\nltk\translate\bleu_score.py:552: UserWarning
The hypothesis contains 0 counts of 3-gram overlaps.
Therefore the BLEU score evaluates to 0, independently of
how many N-gram overlaps of lower order it contains.
Consider using lower n-gram order or use SmoothingFunction()
  warnings.warn(_msg)
C:\Users\artem\anaconda3\lib\site-packages\nltk\translate\bleu_score.py:552: UserWarning
The hypothesis contains 0 counts of 4-gram overlaps.
Therefore the BLEU score evaluates to 0, independently of
how many N-gram overlaps of lower order it contains.
Consider using lower n-gram order or use SmoothingFunction()
  warnings.warn(_msg)
C:\Users\artem\anaconda3\lib\site-packages\nltk\translate\bleu_score.py:552: UserWarning
The hypothesis contains 0 counts of 2-gram overlaps.
Therefore the BLEU score evaluates to 0, independently of
how many N-gram overlaps of lower order it contains.
Consider using lower n-gram order or use SmoothingFunction()
  warnings.warn(_msg)
```

	Label	BLEU Score
66	0	0.0
67	0	0.0
68	0	0.0
69	0	0.0
70	0	0.0
..	...	...
187	0	0.0
188	0	0.0
189	0	0.0
190	0	0.021178
191	0	0.0

[96 rows x 2 columns]

Rouge score (Recall - Oriented Understudy for Gisting Evaluation)

Χρησιμοποιείται συνήθως για την αξιολόγηση της ποιότητας των περιλήψεων ή των κειμένων που παράγονται από μηχανές. Αποτελείται από τρεις κύριες βαθμολογίες:

- *ROUGE-1* (unigram overlap): Αυτό το score μετρά την επικάλυψη των unigrams (μεμονωμένες λέξεις) μεταξύ των υποψηφίων κειμένων και των κειμένων αναφοράς.

Επικεντρώνεται στο precision και το recall των unigrams. Τα unigrams, όπως ξέρουμε, αποτελούν n-gram ενός πράγματος, άρα ενός token. Όσα μοντέλα βασίζονται σε unigrams αγνοούν πλήρως την σειρά των λέξεων, γι'αυτό και τα bigrams, trigrams κ.ο.κ θεωρούνται περισσότερο χρήσιμα.

- *ROUGE-2* (bigram overlap): Αυτό το score μετρά την επικάλυψη bigrams (ζεύγη διαδοχικών λέξεων) μεταξύ των υποψηφίων κειμένων και των κειμένων αναφοράς. Παρέχει ένα πιο συγκεκριμένο μέτρο επικάλυψης (overlap) σε επίπεδο φράσεων.
- *ROUGE-L* (longest common subsequence): Αυτό το score μετρά τη μεγαλύτερη κοινή υπο-ακολουθία μεταξύ των υποψηφίων κειμένων και των κειμένων αναφοράς. Λαμβάνει υπόψη τη μεγαλύτερη αλληλουχία λέξεων που ταιριάζει, ανεξάρτητα από τη σειρά των λέξεων.

Κάθε ένα από αυτά τα score αντιπροσωπεύεται από μια τιμή f, η οποία συνδυάζει το precision και το recall. Όσο μεγαλύτερο είναι το rouge score, τόσο καλύτερη είναι η ποιότητα του παραγόμενου κειμένου σε σύγκριση με το κείμενο αναφοράς. Αξίζει να αναφερθεί ότι τα συνώνυμα δεν θεωρούνται όμοια.

Συλλογιστική κατεύθυνση:

Κατά την προσπάθεια να εντάξουμε το rouge score στα features ήρθαμε αντιμέτωπες με λιγότερες προκλήσεις, καθώς οι προσεγγίσεις που είχαν ήδη εγκαταλειφθεί για το blue score δεν επανα χρησιμοποιήθηκαν και για αυτή την μετρική. Αρχικά, υπολογίσαμε το rouge score για όλο το dataset. Ομοίως με το blue score, κάναμε διάφορες συγκρίσεις: υγιείς (reference) VS υγιείς, υγιείς VS ασθενείς, ασθενείς VS ασθενείς. Όμως, το matching γινόταν one-to-one στις περιγραφές με την σειρά, οπότε ανάμεσα στις ίδιες κλάσεις πχ. υγιείς (reference) VS υγιείς συνέκρινε το ίδιο κείμενο με τον εαυτό του και είχαμε σχεδόν απόλυτη ταύτιση.

```
Rouge scores for healthy vs demented:
ROUGE-1 score: 0.19512194743319458
ROUGE-2 score: 0.017291062893970383
ROUGE-L score: 0.08780487426246299
Rouge scores for healthy vs healthy:
ROUGE-1 score: 0.999999995
ROUGE-2 score: 0.999999995
ROUGE-L score: 0.999999995
Rouge scores for demented vs demented:
ROUGE-1 score: 0.999999995
ROUGE-2 score: 0.999999995
ROUGE-L score: 0.999999995
```

Στην συνέχεια εξαγάγαμε την μετρική αυτή για κάθε περιγραφή ξεχωριστά. Αυτή συγκρίνεται με ένα randomly επιλεγμένο κείμενο υγιούς συμμετέχοντα (reference). Καθώς έγινε σύγκριση με όλα τα εκφωνήματα, καταχωρήθηκαν εν τέλει τρεις στήλες - feature με τις τιμές για τα rouge score: rouge 1, rouge2 , rouge L για κάθε εγγραφή στο dataset.

Τεχνικές αναπαράστασης

Έγινε δοκιμή με *Bow (bag of words)* και *Word Embeddings (Word2Vec)*. Και οι δύο τεχνικές χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας (NLP) για αναπαράσταση κειμένου, αλλά διαφέρουν ως προς τις προσεγγίσεις τους και τις πληροφορίες που συλλαμβάνουν.

- Bow: είναι μια απλή και βασική τεχνική για αναπαράσταση κειμένου. Αντιμετωπίζει κάθε έγγραφο ως αταξινόμητη συλλογή ή "σακούλα" λέξεων, αγνοώντας τη γραμματική, τη σειρά των λέξεων και τα συμφραζόμενα. Δηλαδή, δεν αποτυπώνει σημασιολογικές ή συντακτικές σχέσεις μεταξύ των λέξεων. Πιο συγκεκριμένα, αναπαριστά ένα έγγραφο δημιουργώντας ένα διάνυσμα, που μετρά τη συχνότητα κάθε λέξης στο έγγραφο. Το διάνυσμα που προκύπτει είναι συνήθως υψηλής διάστασης, με κάθε διάσταση να αντιπροσωπεύει μια μοναδική λέξη του λεξιλογίου.
- Word embeddings: είναι πυκνές διανυσματικές αναπαραστάσεις, που αποτυπώνουν τις σημασιολογικές και συντακτικές σχέσεις μεταξύ των λέξεων. Στόχος τους είναι να συλλάβουν το νόημα των λέξεων σε έναν συνεχή διανυσματικό χώρο. Παράγονται συνήθως με τη χρήση μοντέλων νευρωνικών δικτύων, όπως το Word2Vec. Πιο συγκεκριμένα, αυτά τα μοντέλα εκπαιδεύονται σε μεγάλα σώματα κειμένων για να μάθουν διανυσματικές αναπαραστάσεις για τις λέξεις με βάση τα μοτίβα συνύπαρξής τους. Συνεπώς, αποτυπώνουν τις σημασιολογικές σχέσεις και το νόημα μεταξύ των λέξεων, επιτρέποντας έτσι πιο διαφοροποιημένες αναπαραστάσεις. Παράγουν διανύσματα ικανά να αποτυπώσουν καλύτερα τις ομοιότητες και τις διαφορές μεταξύ των λέξεων. Ωστόσο, η παραγωγή των word embeddings απαιτεί μεγάλες ποσότητες training data και υπολογιστικών πόρων συγκριτικά με το Bow.

Συνοπτικά, το Bow αναπαριστά έγγραφα κειμένου μετρώντας τη συχνότητα των λέξεων, ενώ τα Word embeddings αποτυπώνουν σημασιολογικές σχέσεις με την αναπαράσταση των λέξεων σε ένα συνεχή διανυσματικό χώρο. Το Bow είναι απλό και υπολογιστικά αποδοτικό, αλλά αγνοεί τη

σειρά των λέξεων και το πλαίσιο, ενώ τα Word embeddings λαμβάνουν υπόψη το πλαίσιο και τη σειρά των λέξεων, αποτυπώνουν πιο διαφοροποιημένο νόημα, αλλά απαιτούν περισσότερους πόρους για το training.

Στο *πρόβλημά* μας, χρησιμοποιήσαμε Bow και Word Embeddings, και συγκεκριμένα Word2Vec, για κάθε ένα από τα μοντέλα που θα αναπτύξουμε παρακάτω, καθώς και για κάθε ένα από τα δύο μονοπάτια που ακολουθήσαμε (semi-preprocessed και preprocessed), ώστε να εξάγουμε συμπεράσματα για το ποιος τρόπος εξυπηρετεί καλύτερα το dataset μας καθώς και το *πρόβλημά* μας. Δεν χρησιμοποιήσαμε Word Embeddings μόνο για το μοντέλο Naive Bayes, λόγω αρνητικών τιμών.

5. Μοντέλα

Για κάθε μονοπάτι (semi-preprocessed/tokens και preprocessed/lemmas) είτε με Bow είτε με Embeddings, αρχικά, τρέξαμε τα εξής 5 διαφορετικά μοντέλα, ώστε να καταλήξουμε στα δύο, που θα φέρουν την καλύτερη απόδοση (accuracy):

- *Naive Bayes:* είναι ένας πιθανοτικός αλγόριθμος που βασίζεται στο θεώρημα του Bayes. Υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών και υπολογίζει την πιθανότητα ένα δείγμα να ανήκει σε μια κλάση με βάση τις πιθανότητες των χαρακτηριστικών. Γενικά, είναι κατάλληλος αλγόριθμος για ταξινόμηση, ιδίως όταν πρόκειται για δεδομένα υψηλής διάστασης. Είναι απλός, γρήγορος και μπορεί να αποδώσει καλά όταν η υπόθεση ανεξαρτησίας είναι λογική.
- *Logistic Regression:* είναι ένα γραμμικό μοντέλο που χρησιμοποιεί τη λογιστική συνάρτηση για να μοντελοποιήσει την πιθανότητα ενός δυαδικού αποτελέσματος. Εκτιμά τους συντελεστές των χαρακτηριστικών, για να κάνει προβλέψεις. Γενικά, χρησιμοποιείται ευρέως για προβλήματα δυαδικής ταξινόμησης. Αποδίδει καλά όταν υπάρχει γραμμική σχέση μεταξύ των χαρακτηριστικών. Είναι ερμηνεύσιμο και αποτελεσματικό μοντέλο για μεγάλα σύνολα δεδομένων.
- *SVM Linear:* είναι ένα γραμμικό μοντέλο που διαχωρίζει τις κλάσεις με την εύρεση του υπερ-επιπέδου, που τις διαχωρίζει στο μέγιστο βαθμό σε έναν χώρο χαρακτηριστικών υψηλής διάστασης. Γενικά, είναι αποτελεσματικό για προβλήματα δυαδικής ταξινόμησης,

ιδίως όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα. Μπορεί να χειριστεί καλά δεδομένα υψηλών διαστάσεων και χρησιμοποιείται συχνά όταν υπάρχουν σαφή decision boundaries.

- *Random Forest:* είναι ένας αλγόριθμος που συνδυάζει πολλαπλά δέντρα αποφάσεων για να κάνει προβλέψεις. Δημιουργεί τυχαία υποσύνολα δεδομένων και χαρακτηριστικών για την εκπαίδευση κάθε δέντρου και συνδυάζει τις προβλέψεις τους. Γενικά, είναι ένας ευέλικτος αλγόριθμος κατάλληλος για ευρύ φάσμα προβλημάτων, συμπεριλαμβανομένης της δυαδικής ταξινόμησης. Χειρίζεται μη γραμμικές σχέσεις, αλληλεπιδράσεις χαρακτηριστικών και μπορεί να χειριστεί δεδομένα υψηλής διάστασης.
- *Artificial Neural Networks (ANN):* είναι μια κατηγορία μοντέλων που αποτελούνται από νευρώνες, δηλαδή διασυνδεδεμένους κόμβους, οργανωμένους σε επίπεδα και χρησιμοποιούν activation functions για την επεξεργασία πληροφοριών και την πραγματοποίηση προβλέψεων. Γενικά, είναι ένα ισχυρό και ευέλικτο μοντέλο, που μπορεί να χειριστεί πολύπλοκες σχέσεις μεταξύ των δεδομένων. Είναι κατάλληλο για προβλήματα δυαδικής ταξινόμησης, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων και πολύπλοκα πρότυπα. Ωστόσο, συχνά απαιτούν περισσότερους υπολογιστικούς πόρους συγκριτικά με άλλους αλγορίθμους.

Στη συνέχεια, τρέξαμε τον Dummy αλγόριθμο, για να εξετάσουμε αν έχει κάποιο νόημα το μοντέλο μας. Τα αποτελέσματα του Dummy μας κάλυψαν, άρα έχει νόημα το μοντέλο μας και συνεχίζουμε.

6. Text to Image

Μετά από έρευνα για δωρεάν διαθέσιμες προς χρήση εφαρμογές καθώς και για εφαρμογές που λαμβάνουν input text στα ελληνικά, δεν καταφέραμε να βρούμε εφαρμογή για τα ελληνικά. Καταλήξαμε στην χρήση της εφαρμογής text-to-image της Adobe, Firefly. Όπως αναφέραμε, το dataset μας είναι στα ελληνικά, άρα αναγκαστικά κάναμε χρήση machine translation (google translation) και έπειτα, το μεταφρασμένο κείμενο το περάσαμε ως input text στην εφαρμογή για text to image. Δεδομένων των ζητημάτων που υπάρχουν και αποτελούν κομμάτι θεωρητικής και πρακτικής έρευνας, στο κομμάτι της *μετάφρασης από γλώσσα σε γλώσσα*, θεωρήσαμε κάποιες συμβάσεις ώστε να προχωρήσουμε περαιτέρω το συγκεκριμένο κομμάτι υλοποίησης.

Επομένως, δώσαμε την περιγραφή της εικόνας, δηλαδή τυχαία δύο από τα text της απομαγνητοφώνησης (ένα υγιούς και ένα ανοϊκού), και για κάθε text input πήραμε ως output μια εικόνα που αντιστοιχεί στο δοθέν text. Αυτή την εικόνα (output) την επεξεργαστήκαμε οπτικά ώστε να καταλάβουμε αν μοιάζει (και κατά πόσο) ή όχι στην αρχική εικόνα (Εικόνα_1), πάνω στην οποία βασίστηκε το κείμενο, που δώσαμε ως input.

Η γενική ιδέα συνοψίζεται στην παραδοχή να συγκρίνουμε τυχαία δύο εικόνες - output της text-to-image εφαρμογής της Adobe, Firefly, με την αρχική εικόνα (Εικόνα_1). Όποια εικόνα μοιάζει με την αρχική αναμένεται να προέρχεται από input text συμμετέχοντα που ανήκει στην ομάδα των υγιών. Ενώ όποια εικόνα δεν μοιάζει με την αρχική, αναμένεται να προέρχεται από input text συμμετέχοντα που ανήκει στην ομάδα των ανοϊκών.

Αποτελέσματα:

2.Data / Text processing

Lemmatization

Με lemmatization: 'ο', 'βασικός', 'μερη', 'ο', 'εγώ', 'ο', 'βασικος', 'μου', 'μέριμνα', 'είμαι', 'ο', 'φροντίδα', 'ο', 'εαυτός', 'μου', 'ο', 'ένας', 'τηλεφωνώ', 'ο', 'άλλος', 'θαυμάζω', 'ο', 'γυναίκα', 'μου', 'που', 'κάνω', 'ο', 'δουλειά', 'που', 'γλίτωσε', 'ο', 'κούρεμα'
Χωρίς lemmatization: 'το', 'βασικό', 'μερη', 'η', 'τους', 'η', 'βασικη', 'τους', 'μέριμνα', 'είναι', 'η', 'φροντίδα', 'του', 'εαυτού', 'τους', 'ο', 'ένας', 'τηλεφωνάει', 'ο', 'άλλος', 'θαυμάζει', 'τη', 'γυναίκα', 'του', 'που', 'κάνει', 'τη', 'δουλειά', 'που', 'γλίτωσε', 'το', 'κούρεμα'

Semi-preprocessing είναι ο όρος που χρησιμοποιήσαμε (καταχρηστικά) για να περιγράψουμε οτι συντελέστηκαν τα ακόλουθα: removing punctuation, converting text to lowercase, tokenization. Ενώ preprocessing είναι ο όρος που χρησιμοποιήσαμε για τις άνωθεν διαδικασίες με ειδοποιό διαφορά ότι κάναμε lemmatization.

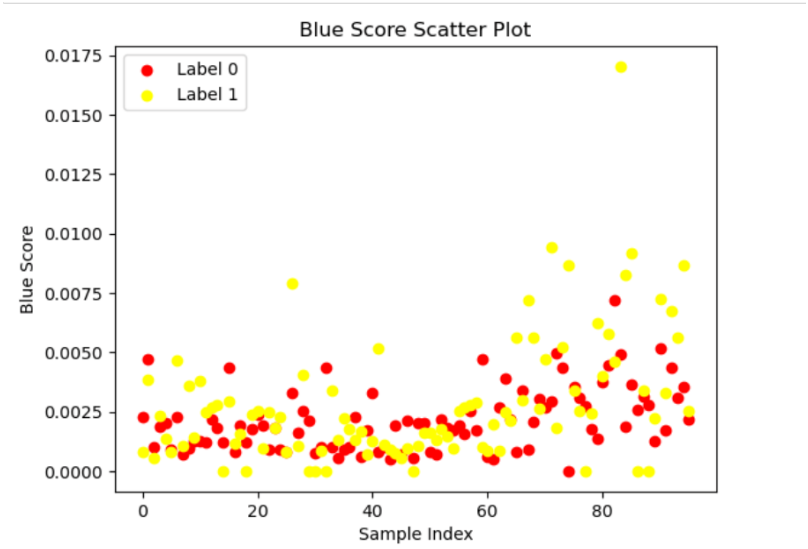
4.Feature extraction

```
[192 rows x 2 columns]
     Label bleu_score   rouge_1   rouge_2   rouge_L
0        1     0.0008  0.193878   0.01875  0.112245
1        1   0.003825  0.181818  0.021505  0.103896
2        1    0.00056   0.14902  0.012848  0.109804
3        1   0.002318  0.235294  0.013333  0.156863
4        1   0.001363  0.243902      0.01  0.195122
..     ...        ...       ...       ...       ...
187      0   0.001712  0.296875  0.036145  0.140625
188      0   0.004361  0.309524  0.082474  0.214286
189      0   0.003093  0.247191       0.0   0.11236
190      0    0.00357  0.252632   0.05042  0.126316
191      0   0.002156  0.309278  0.015748  0.123711
```

Blue score

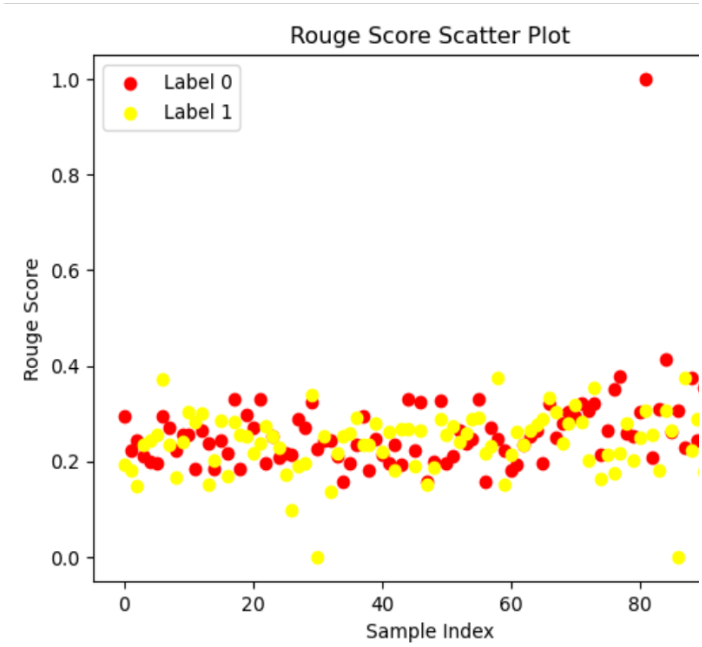
Όπως αναλύσαμε στο αντίστοιχο σημείο της μεθοδολογίας παραπάνω, καταλήξαμε να έχουμε μια τιμή blue score για κάθε εγγραφή (και υγιούς και ανοϊκού συμμετέχοντα) συγκρινόμενο με ένα τυχαίο κείμενο υγιούς (reference). Οι τιμές, οι οποίες πήραμε ως εξαγόμενο δεν ανταποκρίθηκαν στις προσδοκώμενες, καθώς δεν αποτελούν χαρακτηριστικό που μπορεί να σταθεί επαρκές για να κατηγοριοποιήσουμε τα κείμενα στις binary κλάσεις μας (demented, healthy). Καταχωρήθηκαν, όμως, ως feature σε κάθε data point.

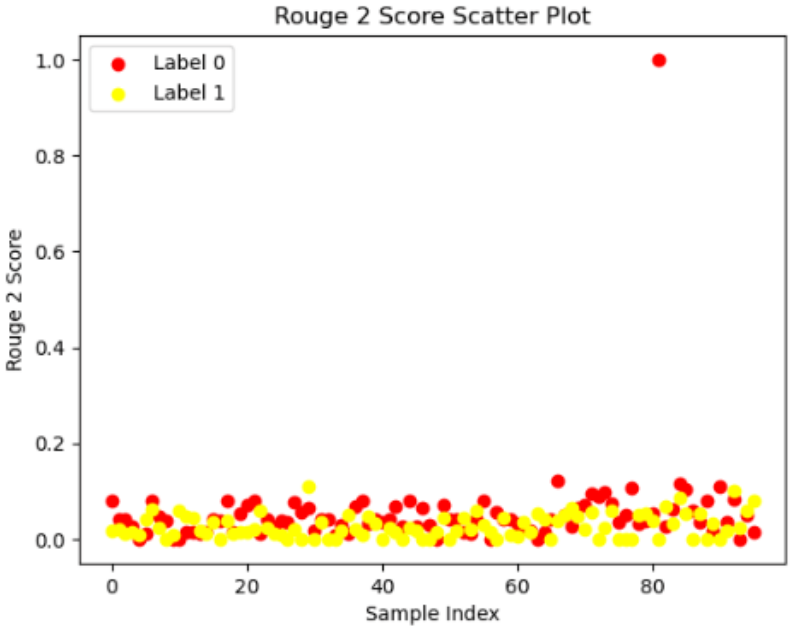
Η ανεπάρκεια αυτή τονίστηκε από το scatter plot που κάναμε στο οποίο, όπως διακρίνεται, η διασπορά των δεδομένων δεν ευνοεί την κατηγοριοποίηση.

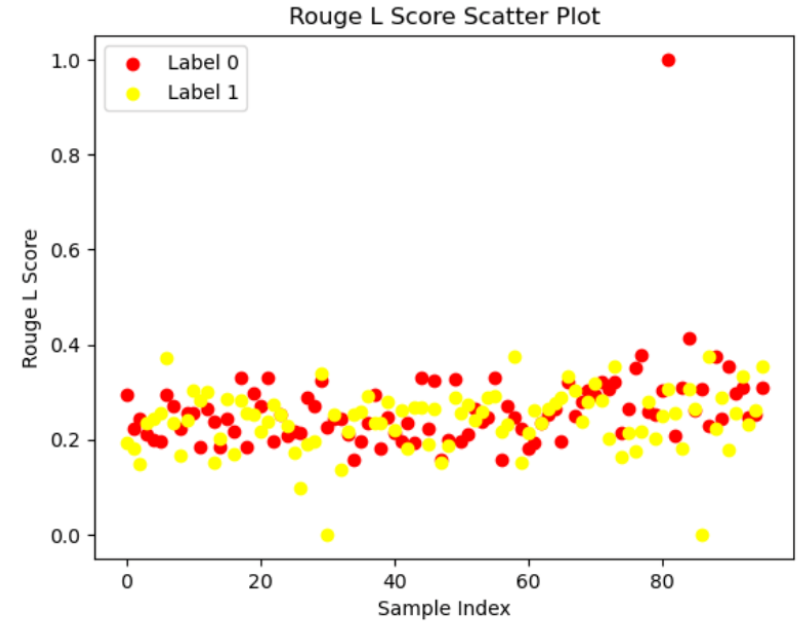


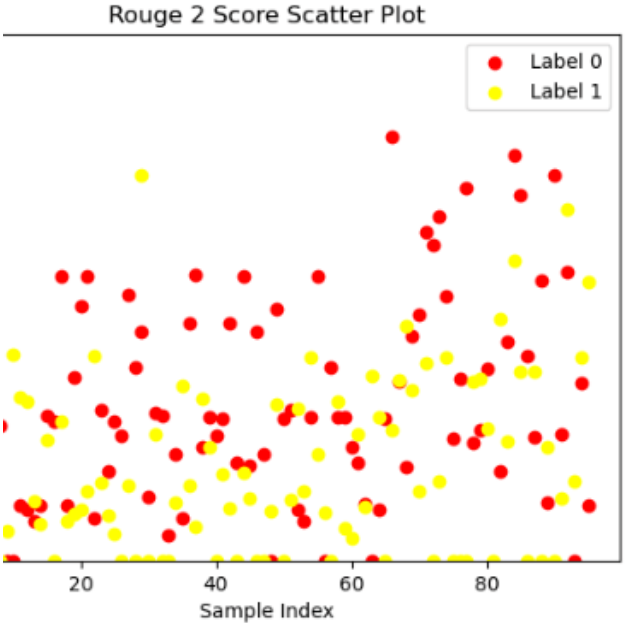
Rouge score

Στην ίδια κατεύθυνση με το blue score, για την μετρική αυτή καταχωρήθηκαν ως feature σε κάθε data point οι τρεις κύριες βαθμολογίες, που παίρνουμε για το rouge score. Παλι, όπως θα γίνει αντιληπτό από τα παρακάτω διαγράμματα, η διασπορά δεν συνηγορεί υπέρ της θεωρίας μας οτι μπορούμε να κατηγοριοποιήσουμε τα data μας με βάση τις τιμές που δέχονται.

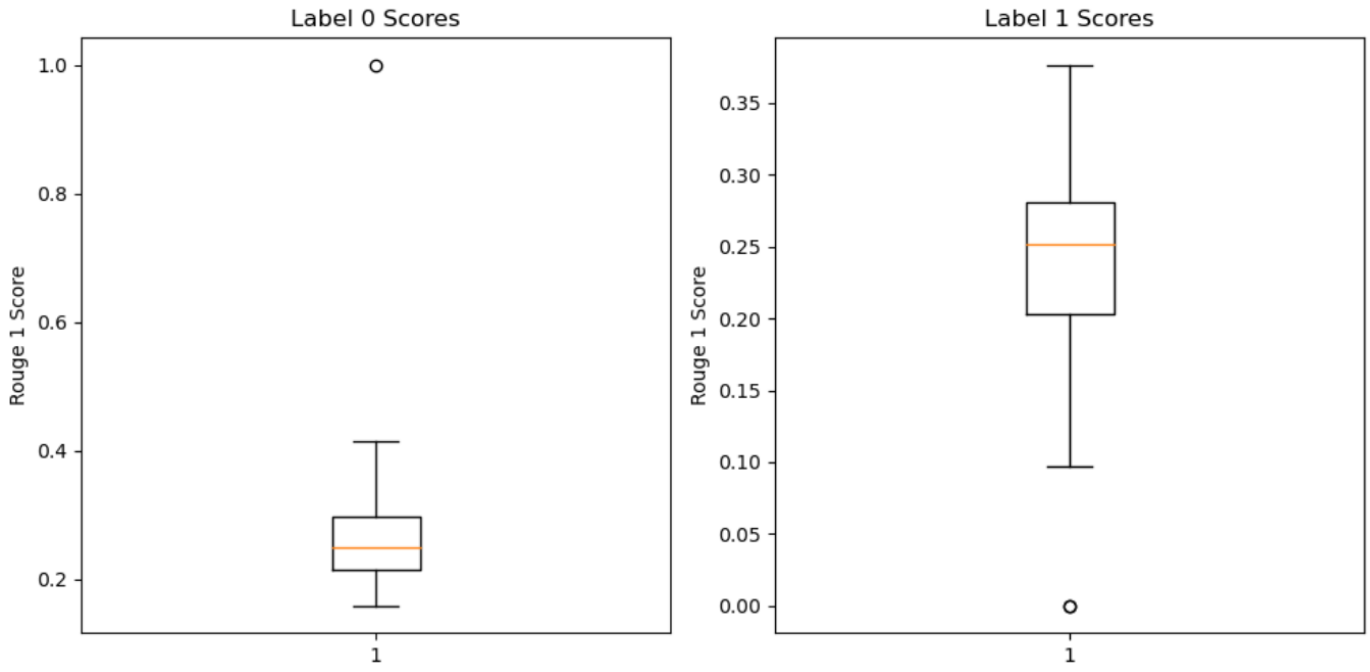








Ενδεικτικά, παραθέτουμε και ένα boxplot για το Rouge 1 Score:



5. Μοντέλα

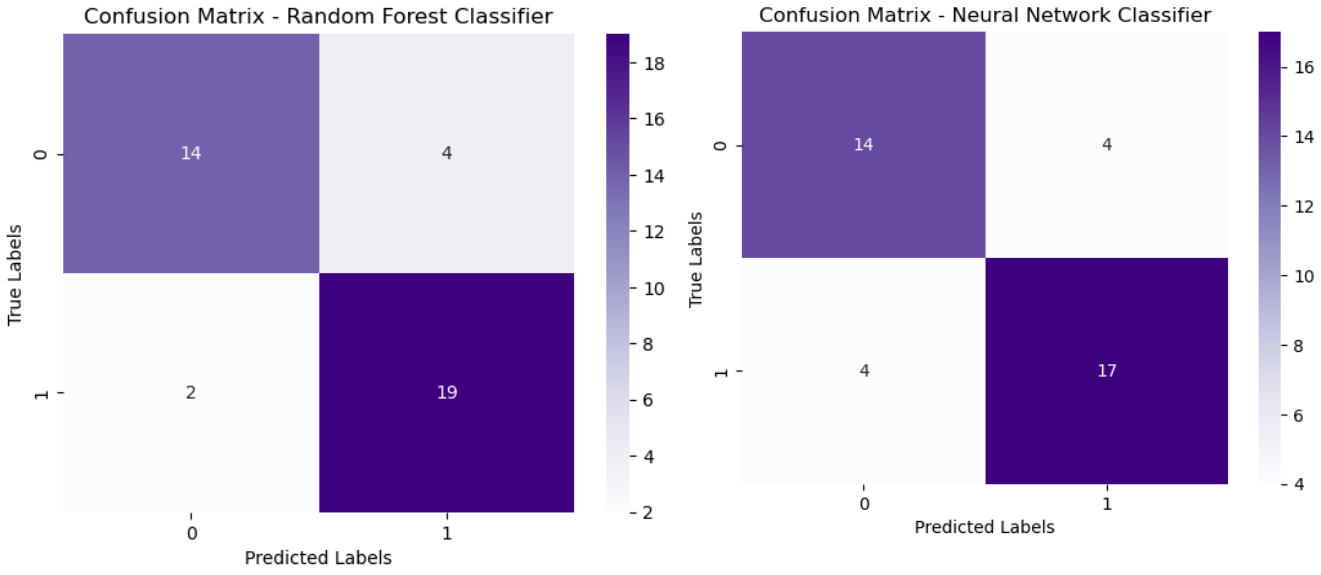
Ακολουθεί screenshot με τα αποτελέσματα των accuracy του κάθε μοντέλου, τα οποία τα κατατάξαμε σε πίνακα για να είναι ευανάγνωστα:

```
[('nn_semi_bow', 0.8205128312110901), ('nn_pre_bow', 0.8205128312110901), ('rf_semi_bow', 0.8205128205128205), ('rf_pre_bow',
0.8205128205128205), ('rf_semi_w2v', 0.7948717948717948), ('rf_pre_w2v', 0.7692307692307693), ('logreg_semi_bow', 0.74358974358
97436), ('logreg_pre_bow', 0.7435897435897436), ('svm_pre_w2v', 0.7435897435897436), ('svm_semi_bow', 0.717948717948718), ('svm
_pre_bow', 0.717948717948718), ('nb_semi_bow', 0.6923076923076923), ('nb_pre_bow', 0.6923076923076923), ('nn_semi_w2v', 0.66666
66865348816), ('logreg_pre_w2v', 0.6666666666666666), ('svm_semi_w2v', 0.6153846153846154), ('nn_pre_w2v', 0.5897436141967773),
('logreg_semi_w2v', 0.5897435897435898)]
```

Feature Extraction	***Bow***	***Word Embeddings (Word2Vec)***
Μοντέλα		
Naive Bayes_semi	0.6923076923076923	-
Naive Bayes_pre	0.6923076923076923	-
Logistic Regression_semi	0.7435897435897436	0.5897435897435898
Logistic Regression_pre	0.7435897435897436	0.6666666666666666
SVM Linear_semi	0.717948717948718	0.6153846153846154
SVM Linear_pre	0.717948717948718	0.7435897435897436
Random Forest_semi	0.8205128205128205	0.7948717948717948
Random Forest_pre	0.8205128205128205	0.7692307692307693
Artificial_Neural_Networks_ANN_semi	0.8205128312110901	0.6666666865348816
Artificial_Neural_Networks_ANN_pre	0.8205128312110901	0.5897436141967773

Από τα παραπάνω βλέπουμε ότι τα δύο μοντέλα με το καλύτερο accuracy είναι τα ANN και Random Forest, γεγονός αναμενόμενο βάση του ότι ο RF είναι ένας ευέλικτος αλγόριθμος κατάλληλος για το χειρισμό μη γραμμικών σχέσεων και το ANN είναι ένα ευέλικτο μοντέλο, που μπορεί να χειριστεί πολύπλοκες σχέσεις μεταξύ των δεδομένων, όπως είδαμε και στη θεωρία. Όπως φαίνεται και στα παραπάνω plot δεν έχουμε γραμμικές σχέσεις, άρα σίγουρα δεν αναμέναμε να πάρουμε καλό accuracy από μοντέλα όπως Logistic Regression, SVM linear.

Ακολουθούν τα confusion matrices των δύο καλύτερων μοντέλων (τυχαία ένα semi και ένα pre αφού δεν αποτελούν διαφοροποιητικό παράγοντα) με Bow προσέγγιση:



Όσον αφορά στην διάκριση μεταξύ Bow και Word Embeddings (Word2Vec) παρουσιάζεται σημαντικά υψηλότερο accuracy σε όλα τα μοντέλα με την Bow προσέγγιση. Εξαίρεση αποτελεί μόνο το μοντέλο SVM linear pre, δηλαδή με lemmatization.

Παρατηρείται ότι για κάθε μοντέλο με Bow προσέγγιση δεν παρατηρείται καμία αλλαγή στα accuracy είτε αν κάνουμε lemmatization είτε όχι. Ενώ η Word2Vec προσέγγιση λαμβάνει υπ' όψιν το lemmatization δεδομένου ότι για κάθε μοντέλο δίνει διαφορετικό accuracy με lemmatization και διαφορετικό χωρίς lemmatization (semi-preprocessed και preprocessed). Πιο συγκεκριμένα, στα μοντέλα Logistic Regression (LR) και SVM_linear το accuracy αυξάνεται στην preprocessed προσέγγιση (δηλαδή με lemmatization) ενώ στα μοντέλα Random Forest και ANN το accuracy μειώνεται στην preprocessed προσέγγιση, όταν κάνουμε lemmatization.

6. Text to Image

Original text από φάκελο AD το 2ο:

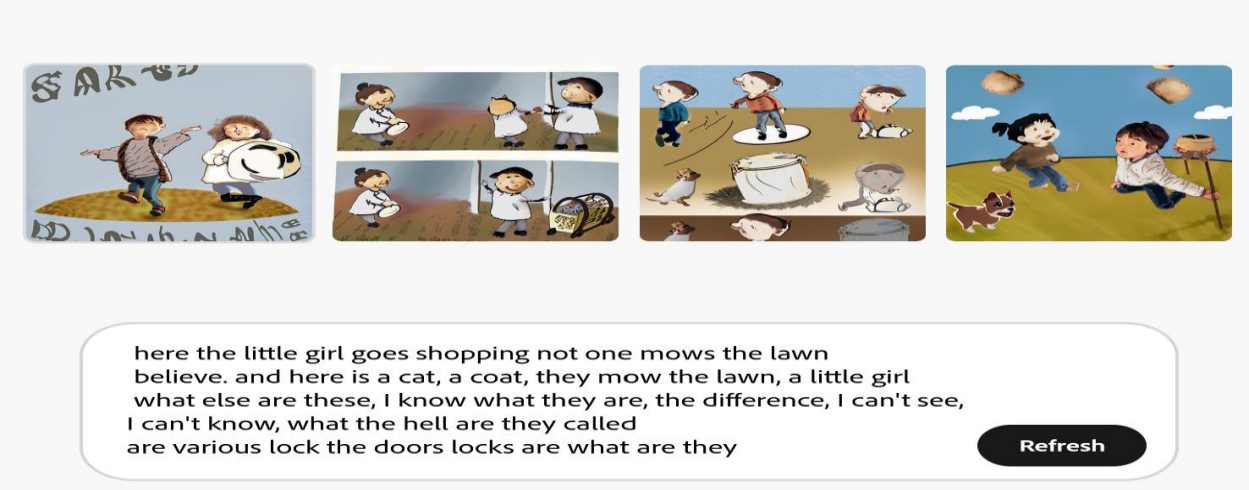
αφου παιδάκια βλεπώ πέφτουν απάνω άλλο ανεβαίνει άλλο κατεβαίνει το σκυλάκι από κάτω ε τι να πω εδώ; έναν νεαρό και αυτός Κάτι κρατά εκεί περα τι είναι αυτό δεν ξέρω τύμπανο; σοκολάτα; τρώει εδώ το κοριτσάκι πάει για ψώνια όχι ένα κουρεύει το γκαζόν πιστεύω. και εδώ είναι μια η

γάτα η κατσούλα κουρεύουνε το γκαζόν το κοριτσάκι ε άλλο τι είναι αυτά ξέρω 'γω τι είναι αυτά διαφορά δεν μπορώ να βλέπω δεν μπορώ να ξέρω τι διάλο πως τα λένε αυτά είναι διάφορες κλειδώνουνε τις πόρτες κλειδαριές είναι τι είναι αυτά

Google translate text:

since I see little children falling up, one goes up, another goes down, the dog from below what can i say here; a young man and he Something is holding there but what it is I don't know drum; chocolate; he eats here the little girl goes shopping not one mows the lawn believe. and here is a cat, a coat, they mow the lawn, a little girl what else are these, I know what they are, the difference, I can't see, I can't know, what the hell are they called are various lock the doors locks are what are they

Adobe firefly image:



Original text από φάκελο Healthy το 5ο:

και μία οικογένεια Η μαμά μιλώντας στο κινητό κόβει το γκαζόν στον κήπο Παρέα με το γάτο που κυνηγάει τα πουλάκια στην κουζίνα Προφανώς εκείνη Μάλλον έχει αφήσει τα πιάτα και τρέχει το νερό για να τα καθαρίσει και να τα σκουπίσει ο μπαμπάς ο οποίος έχει αφήσει το νερό και τρέχει και σαπουνάδα έχει πλημμυρίσει τον κόσμο ανδρικές δουλειές τα παιδιά είναι μέσα στη σκανταλιά ο πιτσιρίκος ανεβαίνει στο ντουλάπι και παίρνει κούκις να φάει πέφτουνε κάτω παίρνει και η μικρή χαρούμενη και τρώει και χαρούμενος και ο σκύλος γιατί και εκείνος έχει να φάει ώσπου ο πιτσιρίκος χάνει την ισορροπία του και είναι έτοιμος να πέσει από το σκαμπό αυτά

Google translate text:

and a family Mom talking on the cell phone mowing the lawn in the garden Hanging out with the cat chasing the birds in the kitchen Apparently she has probably left the dishes and is running the water to clean and wipe them Dad who has left the water and she runs and soap has flooded the world men's jobs the children are in trouble the little one goes up to the cupboard and takes cookies to eat they fall down the little girl gets them happy and eats and the dog is happy too because he has to eat too until the little one loses his balance and he is ready to fall off the stool

Adobe firefly image:

Από τα δύο παραπάνω πειράματα επιβεβαιώνουμε την αρχική πρόβλεψη, ότι το πρώτο κείμενο ανήκει σε ανοϊκό συμμετέχοντα και επομένως η εικόνα που παράχθηκε είναι πιο αφαιρετική και δεν προσεγγίζει τόσο την αρχική (*Εικόνα_1.*), ενώ το δεύτερο κείμενο ανήκει σε υγιή συμμετέχοντα, συνεπώς παρατηρείται μεγαλύτερη ομοιότητα με την πρότυπη εικόνα.

7. Περαιτέρω Συζήτηση

Θα είχε ενδιαφέρον να δούμε ένα αντίστοιχο πείραμα αλλά χωρίς τα stopwords. Θα είχε άραγε λόγω ύπαρξης ένα τέτοιο πείραμα για ένα τέτοιας φύσεως πρόβλημα; Σε ποιο βαθμό ανταποκρίνεται στην πραγματικότητα;

Επίσης, ανάμεσα στις τόσες συζητήσεις για τις γλωσσικές ποικιλίες που χάνονται ή παραγκωνίζονται, θα πρέπει να φτιαχτούν και να υπάρξει μέριμνα και ενδιαφέρον για εφαρμογές που να μην ανταποκρίνονται αποκλειστικά και μόνο στην Αγγλική γλώσσα. Προτείνουμε τη δημιουργία μιας text to image εφαρμογής που να λαμβάνει ως input (και) greek text.

Βιβλιογραφία:

1. Berube S, Nonnemacher J, Demsky C, Glenn S, Saxena S, Wright A, Tippett DC, Hillis AE. Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy Versus Speakers With Aphasia. Am J Speech Lang Pathol. 2019 Mar 11;28(1S):321-329. doi: 10.1044/2018_AJSLP-17-0131. PMID: 30242341; PMCID: PMC6437702.
2. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc. 2020 Nov 1;27(11):1784-1797. doi: 10.1093/jamia/ocaa174. PMID: 32929494; PMCID: PMC7671617.
3. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine. 2020 Oct 22;28:100583. doi: 10.1016/j.eclinm.2020.100583. PMID: 33294808; PMCID: PMC7700896.
4. V. Rentoumi *et al.*, "Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis," *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen, Hungary, 2017, pp. 000033-000038, doi: 10.1109/CogInfoCom.2017.8268212.
5. Rentoumi V, Raoufian L, Ahmed S, de Jager CA, Garrard P. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. J Alzheimers Dis. 2014;42 Suppl 3:S3-17. doi: 10.3233/JAD-140555. PMID: 25061045.
6. Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.