ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Πληροφοριακά Συστήματα στη Διοίκηση Επιχειρήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Δεδομένων από Ελληνικά Κείμενα με Στόχο τη Κατηγοριοποίηση με τη Χρήση του Greek BERT

Γεώργιος Ν. Γκολφόπουλος

Αθήνα, 2022

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

HAROKOPIO UNIVERSITY

SCHOOL OF DIGITAL TECHNOLOGY

INFORMATICS AND TELEMATICS

Information Systems in Business

Master Thesis

Data Mining from Greek Articles for Text Classification with Greek Bert

Georgios N. Gkolfopoulos

Athens, 2022

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Πληροφοριακά Συστήματα στη Διοίκηση Επιχειρήσεων

Τριμελής Εξεταστική Επιτροπή

Βαρλάμης Ηρακλής (Επιβλέπων)

Αναπληρωτής Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής, Χαροκόπειο Πανεπιστήμιο

Δίου Χρήστος

Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής, Χαροκόπειο Πανεπιστήμιο

Μιχαήλ Δημήτριος

Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής, Χαροκόπειο Πανεπιστήμιο

Ο Γκολφόπουλος Γεώργιος

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

3) Όπου υφίστανται δικαιώματα άλλων δημιουργών έχουν διασφαλιστεί όλες οι αναγκαίες άδειες χρήσης ενώ το αντίστοιχο υλικό είναι ευδιάκριτο στην υποβληθείσα εργασία.

ΕΥΧΑΡΙΣΤΙΕΣ

Η συγκεκριμένη διπλωματική εργασία, πραγματοποιήθηκε στο τμήμα Πληροφορικής και Τηλεματικής του Χαροκόπειου Πανεπιστημίου.

Αρχικά θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα καθηγητή μου, κύριο Βαρλάμη Ηρακλή, καθηγητή του τμήματος Πληροφορικής και Τηλεματικής, για την βοήθεια του σε όλη την διάρκεια της διπλωματικής εργασίας και για τις συμβουλές που μου έδωσε όλο αυτό το διάστημα που εργαστήκαμε μαζί.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για την στήριξη που μου παρείχαν όλο αυτό τον καιρό αλλά ιδιαιτέρως θέλω να ευχαριστήσω τους φίλους μου για ότι έκαναν για μένα σε όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας.

Περιεχόμενα

Περίληψη

Η κατηγοριοποίηση κειμένων είναι μια σημαντική μελέτη στον τομέα της εξαγωγής πληροφορίας από κείμενα (Text Mining), έχοντας ένα μεγάλο εύρος εφαρμογής. Τα τελευταία χρόνια, μέσω της εξέλιξης αλγορίθμων νευρωνικών δικτύων (Neural Networks), έχουν αναπτυχθεί πολλές τεχνικές εξαγωγής γλωσσικών μοντέλων από μεγάλες συλλογές κειμένων γνωστά ως προ-εκπαιδευμένα γλωσσικά μοντέλα (Pre-Trained Language Models), οι οποίες βρίσκουν εφαρμογή σε ποικίλες εργασίες επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP). Την συγκεκριμένη χρονική στιγμή, η βέλτιστη πρακτική για ταξινόμηση κειμένων είναι η εφαρμογή των Pre-Trained Language Models με την κατάλληλη προσαρμογή τους (Fine-Tuning).

Στόχος της παρούσας διπλωματικής εργασίας ήταν η δημιουργία ενός κατηγοριοποιητή κειμένων για εφαρμογή πάνω σε Ελληνικά άρθρα, κείμενα και ειδήσεις. Οι βασικοί πυλώνες αυτού του εγχειρήματος είναι το καινοτόμο μοντέλο BERT της Google και η εξελληνισμένη του εκδοχή (GreekBERT). Ο απώτερος σκοπός της συγκεκριμένης διπλωματικής ήταν, με την χρήση της ήδη υπάρχουσας γνώσης και τεχνογνωσίας στον τομέα της Επεξεργασίας Φυσικής Γλώσσας, η ανάπτυξη και η εκπαίδευση ενός μοντέλου κατηγοριοποίησης Ελληνικών ειδήσεων που μέσω συνεχής εκπαίδευσης και αλληλεπίδρασης με τους χρήστες θα καταφέρει να επιτύχει μέγιστα αποτελέσματα.

Αρχικά γίνεται αναφορά στο θεωρητικό κομμάτι των τεχνικών του text mining, των μοντέλων κατηγοριοποίησης και του μοντέλου BERT. Στη συνέχεια και αφού έχει γίνει η κατανόηση του μοντέλου BERT παρουσιάζονται τα βασικά σημεία του μοντέλου που αναπτύχθηκε και εκπαιδεύτηκε για τον Κατηγοριοποιητή Ελληνικών Κειμένων με τη προγραμματιστική γλώσσα python καθώς και της web σελίδας με τη χρήση της προγραμματιστικής γλώσσας HTML. Κατά την ολοκλήρωση της εργασίας παρουσιάζονται τα αποτελέσματα της ακρίβειας και της αποτελεσματικότητας των κατηγοριοποιήσεων του μοντέλου (accuracy) κατά την πρώτη εκπαίδευση και χρήση , με το dataset των 4000 κειμένων που είχαμε στην διάθεσή μας, καθώς και ο έλεγχος της ακρίβειας και της αποτελεσματικότητας των κατηγοριοποιήσεων μετά από αρκετές επανεκπαιδεύσεις και χρήσης της εφαρμογής από πραγματικούς χρήστες.

Λέξεις κλειδιά: [Εξόρυξη Δεδομένων, Greek Bert, Εξόρυξη Κειμένων, Αυτόματη Κατηγοριοποίηση, Μηχανική Μάθηση]

Abstract

The categorization of texts is an important study in the field of text mining, having a wide range of application. In recent years, through the evolution of Neural Networks, many techniques have been developed to extract language models from large collections of texts known as Pre-Trained Language Models, which find application in a variety of Natural Language Processing (NLP) processes. At this point in time, the best practice for text classification is the application of Pre-Trained Language Models with their proper adaptation (Fine-Tuning).

The aim of this diploma thesis was to create a categorizer of texts for application on Greek articles, texts and news. The main pillars of this project are Google's innovative BERT model and its Hellenized version (GreekBERT). The ultimate goal of this diploma thesis was, with the use of the already existing knowledge and know-how in the field of Natural Language Processing, the development and education of a model of categorization of Greek news that through continuous training and interaction with users will manage to achieve maximum results.

Initially, reference is made to the theoretical part of the techniques of text mining, the categorization models and the BERT model. Then, after the understanding of the BERT model, the main points of the model developed and trained for the Greek Text Categorizer with the programming python language as well as the web page with the use of html programming language are presented. During the completion of the project, the results of the accuracy and effectiveness of the classifications of the model (accuracy) during the first training and use are presented, with the dataset of 4000 texts that we had at our disposal, as well as the control of the accuracy and effectiveness of the categorizations after several retrainings and use of the application by real users.

Keywords: [Data Mining, Greek Bert, Text Mining, NLP, A.I.]

Κατάλογος εικόνων

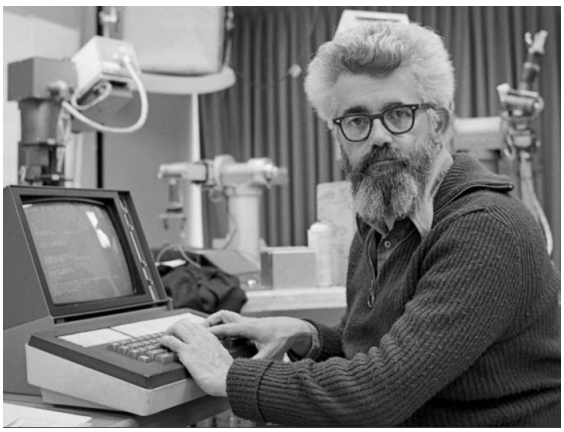
1) Εισαγωγή

Στην σύγχρονη εποχή που ζούμε παράγεται καθημερινά μεγάλος αριθμός ψηφιακών ειδήσεων, άρθρων και κειμένων από διάφερες πηγές. Όλα αυτά αποτελούν ένα σύνολο δεδομένων που είναι δύσκολο και χρονοβόρο να αντιστοιχηθεί σε κατηγορίες που θα βοηθήσουν την αρχειοθέτηση και καταγραφή τους. Συνεπώς με την πάροδο των χρόνων έχει καταστεί απαραίτητη η δημιουργία αυτόματων μηχανισμών επεξεργασίας και κατηγοριοποίησης αρχείων και κειμένων. Οι έρευνες σε θέματα Πληροφορικής, Στατιστικής και Μαθηματικών, έχουν επιφέρει σημαντικές αναβαθμίσεις σε τομείς όπως είναι η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση. Μέσω αυτής της εξέλιξης, έχει επιτευχθεί η δημιουργία νέων μεθόδων δημιουργίας αξιόπιστων μοντέλων μηχανικής μάθησης, τα οποία βρίσκουν εφαρμογή σε ένα μεγάλο πλήθος προβλημάτων. Ένα από αυτά τα προβλήματα είναι και η αυτόματη Κατηγοριοποίηση Κειμένου.

1.1) Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη αναφέρεται στην ικανότητα μιας μηχανής να αναπαράγει τις γνωστικές λειτουργίες ενός ανθρώπου, όπως είναι η μάθηση, ο σχεδιασμός και η δημιουργικότητα.

Η τεχνητή νοημοσύνη καθιστά τις μηχανές ικανές να μπορούν να κατανοήσουν το περιβάλλον τους, να επιλύσουν προβλήματα και να δρουν προς την επίτευξη ενός συγκεκριμένου στόχου. Ο υπολογιστής λαμβάνει δεδομένα, τα επεξεργάζεται και ανταποκρίνεται βάσει αυτών. Τα συστήματα τεχνητής νοημοσύνης είναι ικανά να προσαρμόζουν τη συμπεριφορά τους, αναλύοντας τις συνέπειες προηγούμενων δράσεων και επιλύοντας προβλήματα με αυτονομία. Την τεχνητή νοημοσύνη τη συναντάμε καθημερινά σε διαδικτυακές αγορές και διαφημίσεις, σε έξυπνα σπίτια, πόλεις και υποδομές, σε αυτοκίνητα, στον κλάδο της υγείας, στις μεταφορές και σε άλλα. Ο όρος της τεχνητής νοημοσύνης χρησιμοποιήθηκε για πρώτη φορά το 1956 από τον John McCarthy για να περιγράψει ένα θερινό εργαστήριο με την ονομασία “The Dartmouth Summer Research Project on Artificial Intelligence”, στο οποίο συζητήθηκε το βασικό θέμα, «οι μηχανές που σκέπτονται». Αυτό θεωρήθηκε σαν το πρώτο Συνέδριο της A.I. και υπήρξε καθοριστικό στη γέννηση και υλοποίησή της. Δόθηκε επίσης η εγκυκλοπαιδική ερμηνεία ως η «ικανότητα της μηχανής να μπορεί να σκέπτεται και να μιμείται την ανθρώπινη συμπεριφορά και ευφυΐα, αλλά να μην την αντικαθιστά» (A.I).



Εικόνα 1: John McCarthy

1.2) Μηχανική Μάθηση

Η μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί" (Wikipedia). Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Έχει ισχυρούς δεσμούς με την μαθηματική βελτιστοποίηση, η οποία παρέχει τις μεθόδους, τη θεωρία και τους τομείς εφαρμογής. Η μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση.

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

1.2.1) Τεχνικές Μηχανικής Μάθησης

Υπάρχει μεγάλο πλήθος αλγορίθμων μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν για την επίλυση προβλημάτων. Αντίστοιχα, υπάρχει μεγάλη ποικιλία προβλημάτων με διαφορετικές απαιτήσεις και επιθυμητές εξόδους. Ανάλογα με τη φύση και τις ανάγκες του προβλήματος καθορίζεται και ένα σύνολο τεχνικών μηχανικής μάθησης που πρέπει να εφαρμοστούν. Οι εργασίες μηχανικής μάθησης ταξινομούνται σε δυο βασικές κατηγορίες: την επιβλεπόμενη μάθηση και τη μη επιβλεπόμενη μάθηση.

1.2.2) Επιβλεπόμενη μηχανική μάθηση

Στην επιβλεπόμενη μάθηση ένας χρήστης εισάγει παραδειγματικές εισόδους και τις αντίστοιχες εξόδους σε μια υπολογιστική μηχανή προκειμένου να δημιουργηθεί ένας γενικός κανόνας ο οποίος θα αντιστοιχίζει τις επόμενες εισόδους στις αντίστοιχες εξόδους τους. Ο αλγόριθμος δημιουργεί μια συνάρτηση για να καταφέρει να απεικονίσει τα δεδομένα από το σύνολο εκπαίδευσης στις εξόδους τους που είναι ήδη γνωστές. Μετέπειτα, γενικεύοντας τη συνάρτηση αυτή, έχει τη δυνατότητα να εκτελεί προβλέψεις για δεδομένα με άγνωστη εκ των προτέρων έξοδο.

Οι κυριότερες τεχνικές επιβλεπόμενης μηχανικής μάθησης είναι οι εξής:

- Μάθηση Εννοιών (Concept Learning)
- Δένδρα Απόφασης (Decision Trees)
- Μάθηση Κανόνων (Rule Learning)
- Μάθηση κατά Περίπτωση (Instance Based Learning)
- Μάθηση κατά Bayes
- Γραμμική Παρεμβολή (Linear Regression)
- Νευρωνικά Δίκτυα (Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vectors Machines)

1.2.3) Με επιβλεπόμενη μηχανική μάθηση

Στα προβλήματα Μη επιβλεπόμενης Μάθησης ο υπολογιστής προσπαθεί μέσα από τη δομή των δεδομένων να βρει συσχετίσεις και ομάδες που υπάρχουν χωρίς να έχει καμία πρότερη εμπειρία. Σαν αποτέλεσμα προκύπτουν πρότυπα καθένα εκ των οποίων περιγράφει ένα μέρος των δεδομένων που σχετίζονται μέσω κάποιας ιδιότητας. Τα δεδομένα που έχουμε στη διάθεσή μας στην περίπτωση αυτή, δεν έχουν κάποια γνωστή ετικέτα και επομένως δεν υπάρχει η δυνατότητα εκτίμησης πιθανού λάθους και αξιολόγησης της αποδοτικότητας του μοντέλου.

1.3) Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)

Ως Επεξεργασία Φυσικής Γλώσσας (NLP), ορίζουμε το κοινό πεδίο ανάμεσα στην επιστήμη της Γλωσσολογίας, της Πληροφορικής και της Τεχνητής Νοημοσύνης το οποίο μελετά τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων φυσικών γλωσσών, με απώτερο σκοπό την πλήρη κατανόηση της φυσικής γλώσσας από έναν υπολογιστή, ώστε να μπορεί να εξάγει νοήματα αλλά και φυσική γλώσσα από γλωσσικά δεδομένα. Με άλλα λόγια, ο κύριος στόχος της Επεξεργασίας Φυσικής Γλώσσας είναι ένας υπολογιστής να μπορεί να

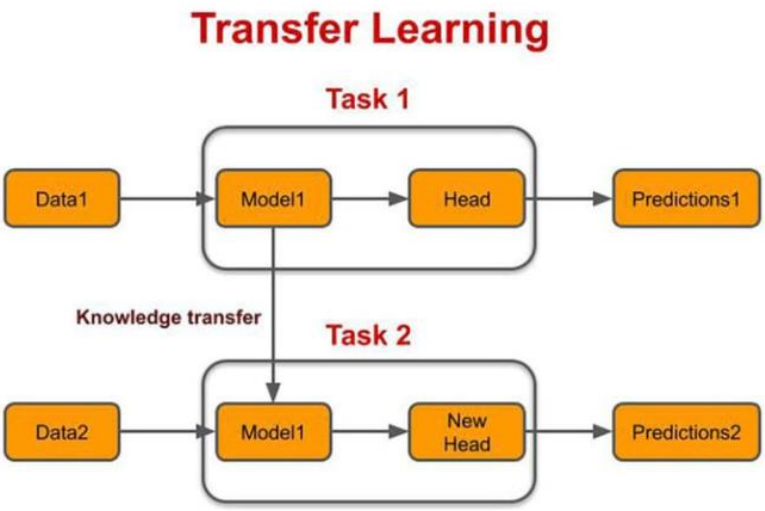
αποκρυπτογραφήσει κάθε έννοια της ανθρώπινης γλώσσας, όπως ο ίδιος ο άνθρωπος αντιλαμβάνεται, ώστε να μπορεί να την χρησιμοποιήσει για οποιοδήποτε όφελος. Λόγω του μεγάλου εύρους της, ο τομέας της Επεξεργασίας Φυσικής Γλώσσας περιλαμβάνει πολλούς υποκλάδους οι οποίοι με την σειρά τους έχουν εξελιχθεί σε αυτόνομες επιστήμες με πολλές έρευνες και έργα. Έννοιες όπως η Γλωσσική Μοντελοποίηση (Language Modeling), Ομαδοποίηση και Ενσωμάτωση Λέξεων (Clustering and Word Embeddings) και Ανάκτηση Πληροφορίας (Information Retrieval) στελεχώνουν το ευρύ πεδίο της NLP. Η διφορούμενη φύση της φυσικής γλώσσας την καθιστά πολύ δύσκολη στην κατανόηση και την ερμηνεία καθώς δημιουργεί σημαντικές ασάφειες.

Οι ασάφειες αυτές εντοπίζονται σε πολλά επίπεδα. Ενδεικτικά αναφέρονται τα εξής:

- Συντακτικές ασάφειες: Μια συντακτικά ορθή πρόταση επιδέχεται πολλές πιθανές ερμηνείες όπως για παράδειγμα η πρόταση: «Χτύπησα τον κλέφτη με το τσεκούρι». Δεν είναι σαφές αν χρησιμοποίησα το τσεκούρι ως όπλο ή αν ο κλέφτης κρατούσε τσεκούρι.
- Λεξιλογικές ασάφειες: Η πρόταση «Το πρώτο γράμμα του Γιώργου» μπορεί να αναφέρεται είτε στο γράμμα που γράφει ο Γιώργος είτε στο σύμβολο του αλφαβήτου με το οποίο αρχίζει το όνομα Γιώργος.
- Αναφορικές ασάφειες: Τέτοιου είδος ασάφεια παρατηρείται όταν δεν είναι ξεκάθαρο σε ποιόν αναφέρεται το καθετί. Για παράδειγμα στην πρόταση: «Ο Γιάννης χτύπησε το Γιώργο γιατί του αρέσει η Μαίρη». Δεν είναι σαφές αν η Μαίρη αρέσει στο Γιάννη ή στο Γιώργο.
- Σημασιολογικές ασάφειες: Μια πρόταση με την ίδια σύνταξη επιδέχεται δύο ή περισσότερες διαφορετικές ερμηνείες.
- Πραγματολογικές ασάφειες: Για τη διερμηνεία μιας πρότασης πολλές φορές είναι απαραίτητο να λάβουμε υπόψη και το κείμενο το οποίο την περιέχει. Για παράδειγμα στη φράση: «Οι δεινόσαυροι εξαφανίστηκαν πριν πολλά χρόνια». Δεν είναι σαφές πόσα είναι τα χρόνια εξαφάνισης των δεινοσαύρων. Οι ασάφειες που υπάρχουν στις προτάσεις της φυσικής γλώσσας αποτελούν σημαντικό πρόβλημα στην εξέλιξη της αυτόματης αναγνώρισης όρων και επομένως στην επεξεργασία της φυσικής γλώσσας. Για το λόγο αυτό απασχολούν ιδιαίτερα τους επιστήμονες που ασχολούνται με την αυτόματη αναγνώριση όρων.

1.4) Transfer Learning & Pretrained Models

Με τον όρο Transfer Learning προσδιορίζουμε την μέθοδο της μηχανικής μάθησης με την οποία ένα μοντέλο το οποίο είναι κατασκευασμένο για ένα συγκεκριμένο πρόβλημα, χρησιμοποιείται ως αρχικό μοντέλο ενός άλλου παρόμοιου προβλήματος. Τα τελευταία χρόνια η χρήση του Transfer Learning έχει γνωρίσει μεγάλη άνθηση μιας και λύνει δύο σημαντικά προβλήματα της μηχανικής μάθησης, την ταχύτητα εκμάθησης του μοντέλου και το μέγεθος των πληροφοριών που απαιτείται για την εκπαίδευση του. Χρησιμοποιώντας μοντέλα τα οποία έχουν εκπαιδευτεί ήδη σε κάποιο παρόμοιο πρόβλημα με αυτό που καλούμαστε να λύσουμε (Pretrained Models), μειώνουμε σημαντικά τον χρόνο εκπαίδευσης του μοντέλου, καθώς η παραμετροποίηση του δεν ξεκινά από το μηδέν αλλά έχει ήδη παραμετροποιηθεί με την χρήση ενός άλλου Dataset σε παρόμοιο πρόβλημα. Με αυτό τον τρόπο, μπορούμε να χρησιμοποιήσουμε ένα μικρό πλήθος δεδομένων πάνω σε ένα προ εκπαιδευμένο μοντέλο και να έχουμε ένα σημαντικό πλεονέκτημα χρόνου και κόστους στην εκπαίδευση του. Η εφαρμογή του Transfer Learning θεωρείται η μεγάλη εξέλιξη στα προβλήματα μηχανικής μάθησης και γενικότερα Τεχνητής Νοημοσύνης καθώς αν παρατηρήσουμε την εκμάθηση ενός ανθρώπινου εγκεφάλου, θα διαπιστώσουμε ότι ο ανθρώπινος εγκέφαλος δεν μαθαίνει τα πάντα από την αρχή, αλλά μεταδίδει γνώση την οποία κατέχει ήδη, σε ένα άλλο πρόβλημα το οποίο καλείται να λύσει.



Εικόνα 2: Transfer Learning

1.5) Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων (Data Mining) είναι ένας ερευνητικός τομέας που προσπαθεί να επιλύσει το πρόβλημα του μεγάλου όγκου πληροφοριών με τη χρήση διάφορων τεχνικών. Αποτελεί την προσπάθεια να αποκτήσουμε τις πληροφορίες που μας ενδιαφέρουν από πολλές πηγές πληροφοριών, όπως είναι οι βάσεις δεδομένων, εξωτερικά αρχεία, ρεύματα δεδομένων κ.ά. Επιπλέον, είναι σημαντικό τα δεδομένα που θα αποκτήσουμε να είναι κατανοητά για να μπορέσουμε να τα χειριστούμε σωστά.

Βασικός στόχος του Data Mining είναι η ανάλυση μεγάλων όγκων δεδομένων ώστε να κατηγοριοποιηθούν, να εντοπιστούν τυχόν ανωμαλίες και διαφοροποιημένες εγγραφές αλλά και η εύρεση προτύπων. Σαν απώτερος σκοπός του συγκεκριμένου κλάδου είναι η αυτόματη ή ημιαυτόματη ανάλυση πολλών δεδομένων για την εξαγωγή κάποιου νέου προτύπου. Οι πολλές πηγές και πληροφορίες που μπορούμε να αποκτήσουμε δεν αποτελεί πάντα θετικό στοιχείο μιας και ο όγκος που συλλέγουμε από αυτές μπορεί να μας οδηγήσει σε σύγχυση με αρκετά ασήμαντα στοιχεία. Για αυτό από τα μεγάλα σύνολα δεδομένων παίρνουμε μόνο τις λίγες και σημαντικές πληροφορίες.

Υπάρχουν πολλές χρήσεις και παραδείγματα που χρησιμοποιείται η εξόρυξη δεδομένων. Για παράδειγμα στα διαδικτυακά μηνύματα αλληλογραφίας με ανεπιθύμητο περιεχόμενο (spam email) χρησιμοποιείται φίλτρο που υλοποιείται με κανόνες ή κάποιο άλλο μοντέλο που έχει εκπαιδευτεί με αλγόριθμους Data Mining. Εκεί έχει μάθει από την εξέταση εκατομμυρίων μηνυμάτων, ποια έχουν χαρακτηριστεί ως ανεπιθύμητα (spam) και ποια τα βασικά χαρακτηριστικά τους. Ένας άλλος κλάδος όπου χρησιμοποιείται είναι στις αγορές και στο «ψάρεμα» υποψήφιων πελατών ή στις τράπεζες και στην έγκριση τραπεζικών προϊόντων η οποία είναι βασισμένη σε στοιχεία των αιτούντων. Τέλος είναι πολύ χρήσιμη στις φορολογικές αρχές και στην εντόπιση φορολογικών δηλώσεων που είναι πιθανόν να είναι ψευδείς. Οι τομείς που χρησιμοποιούν πλέον εξόρυξη πληροφοριών είναι πολλοί. Ενδεικτικά η ιατρική, οι τηλεπικοινωνίες όπως και το χρηματιστήριο και γενικά η οικονομία εφαρμόζουν το data mining για τη συλλογή των σημαντικών δεδομένων μέσα από μεγάλο όγκο στοιχείων.

1.6) Εξόρυξη Δεδομένων από Κείμενα

Το Text Mining είναι παρόμοιος τομέας με το Data Mining. Ενώ το Data Mining είναι μία διαδικασία βασισμένη σε αλγορίθμους για εξόρυξη και ανάλυση χρήσιμων στοιχείων από διάφορης μορφής δεδομένα, το Text Mining είναι το σύνολο των διαδικασιών που απαιτούνται για τη μετατροπή αδόμητων εγγράφων, που μπορούν να θεωρηθούν δεδομένα σε γραπτή μορφή, σε πολύτιμες και δομημένες πληροφορίες.

Τα συστήματα στο Data Mining ασχολούνται με πηγές που θεωρούνται ομογενείς και είναι εύκολες γενικά προς την κατανόησή τους, ενώ στο Text Mining έχουμε μία νέα και πιο δύσκολη πρόκληση. Και αυτό γιατί η πρόβλεψη των διαφόρων χρήσιμων αποτελεσμάτων από τις μεγάλες βάσεις γραπτών δεδομένων περιλαμβάνει το δύσκολο πρόβλημα της ετερογενούς μορφής των έγγραφων και πηγών. Για παράδειγμα μία πηγή μπορεί να είναι σε μορφή email, μίας δημοσίευσης σε κοινωνικό δίκτυο ή ακόμα και ενός κλασσικού μηνύματος SMS.

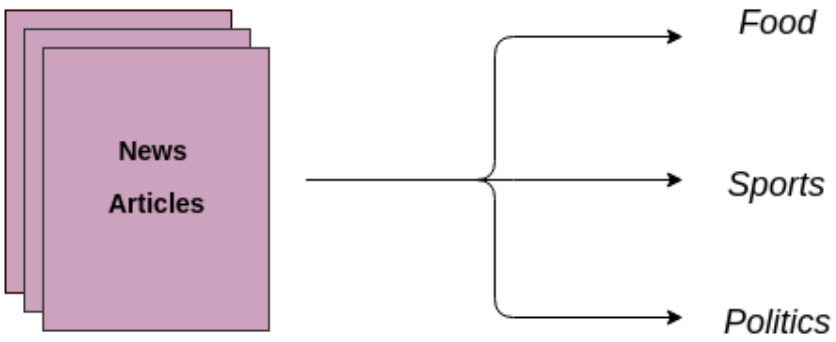
Η γενική εξόρυξη δεδομένων είναι αποδεδειγμένη, ισχυρή, κατανοητή και γρήγορη τεχνολογία για πολλές δεκαετίες. Αλλά με τη πάροδο του χρόνου η συνεχής χρησιμοποίηση του διαδικτύου και οι διαφορετικές μορφές κειμένων που περιέχει κάνουν αναγκαία πλέον και την εξόρυξη δεδομένων από κείμενα. Οι περισσότερες πληροφορίες είναι σε κάποια μορφή κειμένου και σίγουρα η απόκτηση μόνο των σημαντικών πληροφοριών από τεράστια κείμενα και δεδομένα είναι ένα μεγάλο επίτευγμά σε κέρδος κόστους και χρόνου. Η εξόρυξη κειμένου έχει καταστεί πρακτικότερη για τους ειδικούς αλλά και τους απλούς χρήστες λόγω της ανάπτυξης μεγάλων πλατφορμών δεδομένων και αλγορίθμων βαθιάς μάθησης Deep Learning που μπορούν να αναλύσουν μαζικά σύνολα μη δομημένων δεδομένων.

2) Κατηγοριοποίηση Κειμένων

2.1) Τύποι Κατηγοριοποίησης Κειμένων

Ως Κατηγοριοποίηση Κειμένων (Text Classification) ορίζουμε τη διαδικασία κατά την οποία διάφορα αντικείμενα ή έννοιες αναγνωρίζονται και ανατίθενται σε μία συγκεκριμένη κατηγορία ή σε ένα προκαθορισμένο σύνολο κατηγοριών.

Κατά την διαδικασία αυτή, αποδίδεται στο εκάστοτε κείμενο μια ή περισσότερες ετικέτες (labels) ώστε να προσδιοριστεί σε ποιες κατηγορίες ανήκει σε σχέση με τα υπόλοιπα κείμενα της συλλογής. Για παράδειγμα για ένα σύνολο ειδήσεων οι οποίες έχουν αποσπαστεί από ιστοσελίδες μπορούμε να αποδώσουμε ετικέτες ανάλογα με το είδος στο οποίο ανήκουν. Με αυτό τον τρόπο, καταφέρνουμε να οργανώσουμε σε κατηγορίες τα κείμενα μιας συλλογής κειμένων, ώστε να αξιοποιήσουμε την πληροφορία αυτή σε διάφορες άλλες εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας.



Εικόνα 3: Κατηγοριοποίηση Κειμένων

2.1.1) Τύποι Κατηγοριοποίησης Κειμένων

Ένα κείμενο μπορεί να ανήκει σε παραπάνω από μια ετικέτες κατηγορίας. Πολλά συγγράμματά περιέχουν πλέον πάνω από ένα χαρακτηριστικό και δεν υπάρχει κανένας περιορισμός ως προς τον αριθμό των κατηγοριών στις οποίες μπορεί να αναγνωριστεί το

έγγραφο. Ενώ στη περίπτωση κατηγοριοποίησης μονής ετικέτας (Single-Label Classification) έχουμε την συνθήκη ότι για κάθε έγγραφο υπάρχει αντιστοίχιση σε μία κατηγορία και μόνο.

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

Εικόνα 4: Πίνακας επίλυσης κατηγοριοποίησης κειμένων με πολλές ετικέτες

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Εικόνα 5: Πίνακας επίλυσης κατηγοριοποίησης κειμένων με μία ετικέτα

2.1.2) Αυτόματη Κατηγοριοποίηση Κειμένου

Ένας επιστημονικός κλάδος, ο οποίος επιχειρεί να αντιμετωπίσει την ραγδαία αύξηση των πληροφοριών, διευκολύνοντας την πρόσβαση και αναζήτηση στην πληθώρα των πηγών πληροφόρησης που παρέχονται σε ηλεκτρονική μορφή, είναι εκείνος της Αυτόματης Κατηγοριοποίησης Κειμένου – Α.Κ.Κ. (Automated Text Categorization), δηλαδή της αυτόματης ανάθεσης κειμένων, γραμμένων σε φυσική γλώσσα, σε ένα σύνολο προκαθορισμένων κατηγοριών βάσει του περιεχομένου τους. Οι πρώτες προσεγγίσεις στην κατηγοριοποίηση κειμένου περιλάμβαναν την κατασκευή κανόνων από επιστήμονες της τεχνολογίας γνώσεων και από επαγγελματίες, εξειδικευμένους στο γνωστικό αντικείμενο των υπό κατηγοριοποίηση κειμένων. Με την πρόοδο που σημειώθηκε ωστόσο τα τελευταία χρόνια στο επιστημονικό πεδίο της μηχανικής μάθησης, το κέντρο βάρους της Α.Κ.Κ. άρχισε να μετατοπίζεται προς την κατασκευή ταξινομητών, ικανών να κατηγοριοποιήσουν ηλεκτρονικά κείμενα αυτόματα, μέσω της εκμάθησης των χαρακτηριστικών των κατηγοριών αυτών, από ένα ήδη ταξινομημένο σώμα

κειμένων. Η νέα αυτή αντιμετώπιση του προβλήματος προσέφερε στην Α.Κ.Κ. συγκρίσιμη ακρίβεια με εκείνη που επιτύγχαναν οι κανόνες των επιστημόνων της τεχνολογίας γνώσεων, ανεξαρτησία από τη θεματολογία των υπό κατηγοριοποίηση κειμένων και ελαχιστοποίηση της ανθρώπινης παρέμβασης στην όλη διαδικασία.

Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων. Η κατηγοριοποίηση μπορεί να περιγράφει ως μία διαδικασία δύο βημάτων:

1. Εκμάθηση (Learning): Στο πρώτο βήμα της διαδικασίας δημιουργείται/προσδιορίζεται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων παραδειγμάτων, που ονομάζονται δεδομένα εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η κατηγοριοποίηση αποτελεί μέθοδο εποπτευομένης μάθησης (supervised learning). Το μοντέλο, που λέγεται και αλλιώς κατηγοριοποιητής (classifier), αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων απόφασης (decision trees) ή μαθηματικών τύπων.

2. Κατηγοριοποίηση (Classification): Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα (test data) για να αξιολογήσουμε την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο. Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη. Παραδείγματα μιας τέτοιας τεχνικής αποτελούν η ανίχνευση ανεπιθύμητων μηνυμάτων με βάση την επικεφαλίδα τους ή το περιεχόμενό τους, η πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη ή η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστοληπτική τους ικανότητα.

2.1.3) Ορισμοί του προβλήματος της Κατηγοριοποίησης Κειμένου

Η Αυτόματη Κατηγοριοποίηση Κειμένου μπορεί να οριστεί ως η διαδικασία ανάθεσης μιας τιμής 0 ή 1 σε κάθε κελί a_{ij} του πίνακα απόφασης:

	d_1	...	d_j	...	d_n
c_1	a_{11}	...	a_{1j}	...	a_{1n}
...	...	...	...	...	...
c_i	a_{i1}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...
c_m	a_{m1}	...	a_{mj}	...	a_{mn}

Εικόνα 6: Πίνακας Απόφασης της Κατηγοριοποίησης

όπου: C = {c1, c2, …, cm} ένα σύνολο m προκαθορισμένων κατηγοριών και D = {d1, d2, …, dn} το σύνολο των προς κατηγοριοποίηση εγγράφων. Η τιμή 1 ενός πεδίου aij δηλώνει την απόφαση να καταταχθεί το κείμενο dj στην κατηγορία ci, ενώ η τιμή 0 να μην καταταχθεί στην κατηγορία ci. Πιο τυπικά, σκοπός είναι η προσέγγιση της άγνωστης συνάρτησης fDC : 0,1 × →{ } , η οποία αντιστοιχίζει τα έγγραφα στις κατηγορίες στις οποίες ανήκουν, μέσω της συνάρτησης f DC ' : 0,1 × →{ }, η οποία καλείται ταξινομητής ή υπόθεση, έτσι ώστε οι f και ' f να συμπίπτουν (coincide) όσο το δυνατό περισσότερο, με βάση κάποια μέτρα αξιολόγησης του βαθμού σύμπτωσης, δηλ. της αποτελεσματικότητας (effectiveness) του αλγόριθμου.

Η ταξινόμηση βασίζεται γενικά στη σημασιολογία (semantics) των κειμένων και όχι σε διαθέσιμα μεταδεδομένα. Αυτό σημαίνει ότι η απόφαση του ταξινομητή στηρίζεται στην ενδογενή γνώση των κειμένων, δηλαδή αυτή που μπορεί να εξαχθεί από το περιεχόμενό τους, χωρίς τη χρήση γνώσης που μπορεί να παράσχει μια εξωτερική πηγή. Δεδομένου ότι η σημασιολογία ενός κειμένου είναι μια ενδογενώς υποκειμενική έννοια, η συνάφεια ενός κειμένου με μια κατηγορία δεν μπορεί να καθοριστεί μονοσήμαντα. Δεν είναι άλλωστε σπάνια η περίπτωση διαφωνίας μεταξύ δύο ανθρώπων που καλούνται να κατατάξουν ένα κείμενο σε μια κατηγορία, π.χ. να αποφασίσουν κατά πόσο ένα άρθρο για την ελληνική συμμετοχή στη Eurovision ανήκει στην κατηγορία Μουσική, Εθνικά Θέματα ή σε καμία από τις δύο. Το

πρόβλημα της ΑΚΚ μπορεί να προσεγγιστεί είτε με τεχνικές που βασίζονται στη γνώση (knowledge-based) είτε με τεχνικές μηχανικής μάθησης. Στην πρώτη περίπτωση σκοπός είναι η ανάπτυξη ενός έμπειρου προγράμματος που συνίσταται από ένα σύνολο κανόνων της μορφής "if then else" που έχουν οριστεί με το χέρι από ειδικούς. Εναλλακτικά, μπορούν να αξιοποιηθούν γλωσσολογικοί θησαυροί (λεξικά, οντολογίες, κλπ). Οι αλγόριθμοι μηχανικής μάθησης αντιμετωπίζουν το πρόβλημα της ΑΚΚ αναπτύσσοντας έναν ταξινομητή που μαθαίνει από τα χαρακτηριστικά ενός συνόλου κειμένων που έχουν κατηγοριοποιηθεί από κάποιον ειδικό (το σώμα εκπαίδευσης). Ακολουθώντας μια επαγωγική διαδικασία, ο ταξινομητής αποφασίζει την κατηγορία στην οποία πρέπει να ανήκει ένα νέο κείμενο από το σώμα ελέγχου. Η προσέγγιση της μηχανικής μάθησης πλεονεκτεί έναντι των συστημάτων γνώσης, καθώς όταν αλλάζει το σύνολο των κατηγοριών ή όταν το σύστημα εφαρμόζεται σε έναν διαφορετικό τομέα το μόνο που χρειάζεται είναι η επανεκπαίδευση του ταξινομητή με βάση τα νέα δεδομένα, χωρίς να απαιτείται η εμπλοκή των ειδικών για την επαναδιατύπωση των κανόνων.

Ένας από τους περιορισμούς που επιβάλλεται από ορισμένες εφαρμογές εντοπίζεται στον αριθμό των κατηγοριών που είναι δυνατόν να ανατεθούν σε ένα κείμενο. Για παράδειγμα, μπορεί να υπάρχει η απαίτηση κάθε κείμενο να αντιστοιχηθεί σε k ακριβώς (ή λιγότερες ή περισσότερες) κατηγορίες από το σύνολο C, ή αντίστοιχα κάθε κατηγορία να αντιστοιχισθεί σε l ακριβώς (ή λιγότερα ή περισσότερα) κείμενο, στην περίπτωση που μας ενδιαφέρει η ισοκατανομή τους στις υπάρχουσες κατηγορίες. Η σημαντικότερη από τις δύο περιπτώσεις είναι εκείνη της αντιστοίχισης ακριβώς k κατηγοριών (ή λιγότερων ή περισσότερων) ανά κείμενο. Όταν ο αριθμός $k = 1$, τότε κάνουμε λόγο για κατηγοριοποίηση μονής ετικέτας (single-label categorization), ενώ όταν ισχύει ότι $k > 1$, για κατηγοριοποίηση πολλαπλής ετικέτας (multi-label categorization), περίπτωση ειδικότερη της προηγούμενης, καθώς αποδεικνύεται ότι ανάγεται σε k διαφορετικά προβλήματα κατηγοριοποίησης μονής ετικέτας ($k \neq 1$).

2.2) Ερευνητικές Εργασίες πάνω στην Κατηγοριοποίηση Κειμένων

Τα τελευταία χρόνια έχουν αναπτυχθεί αρκετές εργασίες γύρω από την κατηγοριοποίηση κειμένων. Παρακάτω παρουσιάζονται μερικές από αυτές:

- **Experiments in Text Classification: Analyzing the Sentiment of Electronic Product Reviews in Greek**

  Η συγκεκριμένη εργασία ασχολήθηκε με την ανάλυση συναισθημάτων των ηλεκτρονικών κριτικών για προϊόντα που είναι γραμμένες στα ελληνικά. Για το σκοπό αυτό, χρησιμοποιήθηκε ένα μικρό σύνολο δεδομένων με 480 θετικές και αρνητικές κριτικές, το οποίο λήφθηκε από τη δημοφιλή ελληνική ιστοσελίδα ηλεκτρονικού εμπορίου, www.skroutz.gr. Αξιολογήθηκαν διαφορετικά υπολογιστικά μοντέλα για την εκπαίδευση και τη δοκιμή του συνόλου δεδομένων. Τα αποτελέσματα φάνηκαν πολύ ελπιδοφόρα για ένα τόσο μικρό όγκο δεδομένων.

- **A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets**

  Το παρόν έργο περιγράφει ένα σχήμα ταξινόμησης για την ανίχνευση ειρωνείας στα ελληνικά πολιτικά tweets. Η υπόθεση αναφέρει ότι τα χιουμοριστικά πολιτικά tweets θα μπορούσαν να προβλέψουν πραγματικά εκλογικά αποτελέσματα. Η προτεινόμενη προσέγγιση βασίζεται σε περιορισμένα δεδομένα κατάρτισης με ετικέτες, επομένως ακολουθείται μια ημι-εποπτευόμενη προσέγγιση, όπου οι αλγόριθμοι συλλογικής μάθησης λαμβάνουν υπόψη τόσο τα δεδομένα με ετικέτα όσο και τα δεδομένα χωρίς ετικέτα. Συγκρίθηκαν τα ημι-εποπτευόμενα αποτελέσματα με τα εποπτευόμενα από προηγούμενη έρευνά. Η υπόθεση αξιολογείται μέσω μελέτης συσχέτισης μεταξύ της ειρωνείας που λαμβάνει ένα κόμμα στο Twitter, των αντίστοιχων πραγματικών εκλογικών αποτελεσμάτων του κατά τις ελληνικές βουλευτικές εκλογές του Μαΐου 2012 και της διαφοράς μεταξύ αυτών των αποτελεσμάτων και αυτών των προηγούμενων εκλογών του 2009.

- **Multilabel Text Classification for Automated Tag Suggestion**

  Το Bibsonomy είναι ένα σύστημα κοινωνικού σελιδοδείκτη και ανταλλαγής εκδόσεων. Ένας χρήστης μπορεί να αποθηκεύσει και να οργανώσει σελιδοδείκτες (ιστοσελίδες). Το κύριο εργαλείο που παρέχεται για τη διαχείριση περιεχομένου στο BibSonomy είναι η προσθήκη ετικετών. Οι χρήστες μπορούν να εκχωρούν ελεύθερα tags τα υποβάλλουν στο σύστημα. Στην συγκεκριμένη εργασία δημιουργήθηκε εργαλείο το οποίο πρότεινε στον χρήστη ένα σχετικό σύνολο ετικετών.

2.2.1) Εφαρμογές Κατηγοριοποίησης

Παρακάτω παρουσιάζονται κάποιες από τις εφαρμογές κατηγοριοποίησης:

• Αυτόματη Ευρετηριοποίηση Συστημάτων Ανάκτησης Πληροφορίας (Information Retrieval ή IR Systems): Στην περίπτωση των συστημάτων ανάκτησης πληροφορίας, η χρήση της Α.Κ.Κ. συνίσταται στη δημιουργία ευρετηρίων από κείμενα, με βάση ένα ελεγχόμενο λεξικό. Πιο συγκεκριμένα, σε κάθε κείμενο ανατίθενται μια σειρά από λέξεις ή φράσεις κλειδιά που εννοιολογικά ταιριάζουν με το περιεχόμενό του, και οι οποίες συστήνουν το προαναφερθέν λεξικό. Υπό το παραπάνω πρίσμα, οι λέξεις και οι φράσεις κλειδιά του λεξικού αντιστοιχούν στις κατηγορίες ενός συστήματος Α.Κ.Κ.

• Αυτόματη Παραγωγή Μεταδεδομένων: Η εφαρμογή αυτή, η οποία σχετίζεται πολύ με την προηγούμενη, αποσκοπεί στη δημιουργία βιβλιογραφικών στοιχείων (μεταδεδομένων), όπως ημερομηνία συγγραφής, όνομα συγγραφέα, τύπος κειμένου, κ.α. τα οποία χρησιμοποιούνται από ψηφιακές βιβλιοθήκες. Καθώς πολλά από τα στοιχεία αυτά έχουν θεματικό περιεχόμενο, η εφαρμογή θα μπορούσε να αντιμετωπιστεί ως ειδική περίπτωση της αυτόματης ευρετηριοποίησης κειμένων, οδηγούμενης από ένα κατευθυνόμενο λεξικό (τα θεματικά χαρακτηριστικά που προαναφέρθηκαν), αλλά και συντακτικά πρότυπα, η οποία παρουσιάστηκε προηγουμένως. Παράδειγμα αποτελεί το σύστημα CLARITY (http://www.topic.com.au/products/clarity.html).

• Οργάνωση Εγγράφων: Η εφαρμογή αυτή αναφέρεται στην αυτόματη κατάταξη εγγράφων που λαμβάνονται / δημιουργούνται σε πραγματικό χρόνο, σε κατηγορίες, προς διευκόλυνση της διαχείρισής τους, όπως για παράδειγμα η αυτόματη κατηγοριοποίηση των ειδήσεων που καταφθάνουν στα γραφεία κάποιου ειδησεογραφικού πρακτορείου σε θεματικές περιοχές (π.χ. Πολιτιστικά νέα, Διεθνή, κλπ.).

• Επίλυση προβλημάτων που απασχολούν την επεξεργασία φυσικής γλώσσας: Στην ενότητα αυτή εντάσσονται οι επιμέρους εφαρμογές της εννοιολογικής αποσαφήνισης λέξεων (Word Sense Disambiguation – WSD), της εύρεσης δηλαδή του νοηματικού περιεχομένου μιας λέξης σε ένα κείμενο , του ορθογράφου βασισμένου στα συμφραζόμενα (context-sensitive spelling correction), της αναγνώρισης μέρους του λόγου (part of speech tagging), καθώς και της κατάλληλης επιλογής λέξης (word choice selection) που συναντάται στη μηχανική μετάφραση .

• Κατηγοριοποίηση δικτυακών τόπων: Τα αποτελέσματα της εφαρμογής αυτής απαντώνται συχνά σε διάφορες μηχανές αναζήτησης στο διαδίκτυο (π.χ. Yahoo!, INFOSEEK, κ.ά.). Πρόκειται για ιεραρχικούς καταλόγους οι οποίοι περιλαμβάνουν ιστοσελίδες ή ακόμα και ολόκληρους δικτυακούς τόπους με σχετική θεματολογία, διευκολύνοντας έτσι την περιήγηση των χρηστών σε αυτές, καθώς και την αναζήτηση πληροφοριών.

• Κατηγοριοποίηση ομιλίας: Εφαρμογή η οποία κάνει παράλληλη χρήση της αναγνώρισης ομιλίας με την Α.Κ.Κ. .

• Κατηγοριοποίηση εγγράφων πολυμέσων: Εφαρμογή η οποία επικεντρώνεται στην κατηγοριοποίηση εγγράφων με βάση τους υπότιτλους ή τις λεζάντες που συνοδεύουν ένα multimedia έγγραφο (π.χ. φωτογραφία, video clip, κ.α.).

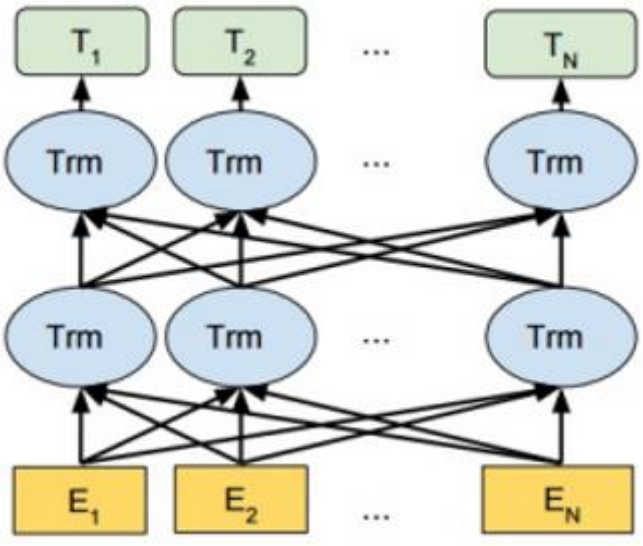
• Αναγνώριση του συγγραφέα κειμένων αμφισβητούμενης ή άγνωστης προέλευσης.

• Φιλτράρισα Εγγράφων: Μια από τις σημαντικότερες εφαρμογές της Α.Κ.Κ., η οποία αποτελεί και το τελικό προϊόν της εργασίας αυτής, είναι το φιλτράρισα εγγράφων. Πρόκειται για τη διαδικασία ταξινόμησης μιας συλλογής εγγράφων που τροφοδοτείται δυναμικά στο σύστημα από κάποια πηγή πληροφορίας, η οποία προσφέρει τις υπηρεσίες της στον λεγόμενο καταναλωτή της πληροφορίας. Τέτοια συστήματα μπορούν να εγκατασταθούν τόσο στο άκρο του καταναλωτή, φιλτράροντας τα κείμενα που απευθύνονται σ’ αυτόν, όσο και στο άκρο της πηγής, περίπτωση κατά την οποία απαιτείται η δημιουργία ενός προφίλ για κάθε καταναλωτή του συστήματος, το οποίο θα καθοδηγεί το σύστημα ταξινόμησης ανάλογα με τις προτιμήσεις του τελευταίου. Ως παράδειγμα μιας τέτοιας εφαρμογής, θα μπορούσε να θεωρηθεί ένα

σύστημα φιλτραρίσματος διαφημιστικής αλληλογραφίας, εγκατεστημένο στο Mail Server (πηγή) και ικανό να διακρίνει και να χαρακτηρίζει αυτόματα τα διαφημιστικά μηνύματα που απευθύνονται μαζικά στους χρήστες που εξυπηρετούνται από αυτόν (καταναλωτές).

3) BERT

Το μοντέλο BERT (Bidirectional Encoder Representations Form Transformers) είναι ένα γλωσσικό μοντέλο το οποίο στηρίζεται στην αρχιτεκτονική του Μετασχηματιστή. Πρόκειται για ένα αμφίδρομο (bidirectional) μοντέλο, το οποίο παράγει βαθιές ενσωματώσεις συμφραζομένων (deep contextualized embeddings). Οι ενσωματώσεις αυτές χρειάζονται πολύ μικρή προσαρμογή προκειμένου να επιτύχουν εξαίρετα αποτελέσματα σε σύνθετα προβλήματα του τομέα επεξεργασίας της φυσικής γλώσσας - πρόβλημα συνεπαγωγής (entailment), πρόβλημα απάντησης σε ερωτήματα (question-answering).



Εικόνα 7: Bert

Ο Μετασχηματιστής περιλαμβάνει δύο διαφορετικούς μηχανισμούς, έναν κωδικοποιητή και έναν αποκωδικοποιητή. Ο κωδικοποιητής διαβάζει το κείμενο εισόδου και ο αποκωδικοποιητής παράγει με τη σειρά του μία πρόβλεψη για το συγκεκριμένο πρόβλημα κάθε φορά. Καθώς ο στόχος του μοντέλου BERT είναι να παράξει ένα γλωσσικό μοντέλο, κάνει

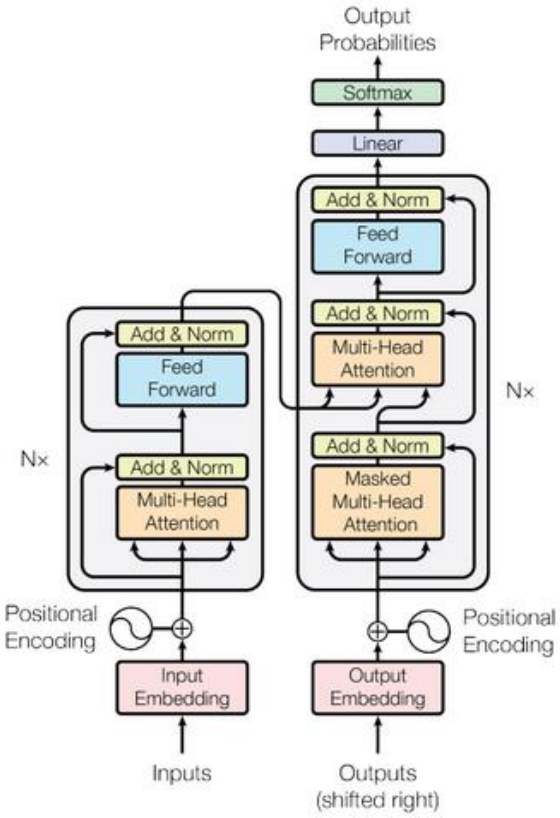
χρήση μόνο της συνιστώσας του κωδικοποιητή, διαφοροποιώντας έτσι σε ένα βαθμό την αρχιτεκτονική του από αυτή του Μετασχηματιστή. Ωστόσο, κληρονομεί την ιδιότητα του τελευταίου να διαβάζει ολόκληρη την ακολουθία των λέξεων απευθείας και όχι ακολουθιακά, με το χαρακτηριστικό αυτό να το καθιστά αμφίδρομο και να του επιτρέπει να μαθαίνει το περιεχόμενο της κάθε λέξης βασιζόμενο στις λέξεις που βρίσκονται αριστερά και δεξιά αυτής.

Το πλαίσιο πάνω στο οποίο στηρίζεται ένα μοντέλο BERT αποτελείται από δύο βήματα. Το πρώτο βήμα αφορά στην προ-εκπαίδευση του μοντέλου (pre-training) και το δεύτερο στην προσαρμογή αυτού (fine-tuning).

3.1) Transformer

Η αρχιτεκτονική του Transformer βρίσκεται στο κέντρο σχεδόν όλων των πρόσφατων σημαντικών εξελίξεων στο τομέα του NLP και επηρεάζει και άλλα γλωσσικά μοντέλα, όπως θα αναφερθεί και παρακάτω. Εισήχθη πρόσφατα, το 2017 από την Google (Uszkoreit, 2017). Ως τότε, χρησιμοποιούνταν διάφορα επαναλαμβανόμενα νευρωνικά δίκτυα για γλωσσικές εργασίες, όπως η μηχανική μετάφραση και τα συστήματα απάντησης ερωτήσεων. Είχε λοιπόν καλύτερα αποτελέσματα από RNN και CNN νευρωνικά δίκτυα και επιπρόσθετα απαιτούσε λιγότερους πόρους. Τα RNN (αναδρομικά νευρωνικά δίκτυα) από τη μία είναι μία τάξη τεχνητών νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ των κόμβων σχηματίζουν ένα κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής αλληλουχίας και μπορούν να χρησιμοποιήσουν την εσωτερική τους κατάσταση (μνήμη) για να επεξεργαστούν ακολουθίες εισόδων. Δεν δίνει μια συγκεκριμένη σειρά για τις λέξεις και το πως θα τοποθετηθούν, πέρα από το να δοθεί σε κάθε λέξη η ακριβής της θέση. Η Google κυκλοφόρησε πέρυσι μια βελτιωμένη έκδοση του Transformer, γνωστή ως Universal Transformer. Υπάρχει μια ακόμη νεότερη και πιο έξυπνη έκδοση, που ονομάζεται Transformer-XL. Το Transformer βασίζεται αποκλειστικά σε self-attention θεωρία. Η συγκεκριμένη θεωρία έχει ως στόχο να μάθουμε τη σχέση των λέξεων σε 1 φράση. Σε αυτό μπορεί να βοηθήσει και η επόμενη πρόταση ή η προηγούμενη. Για παράδειγμα αν η μία πρόταση "Οι Transformers" είναι μια ιαπωνική μπάντα. Το συγκρότημα δημιουργήθηκε το 1968". Εδώ μπορούμε να καταλάβουμε ότι η λέξη συγκρότημα της δεύτερης πρότασης αναφέρεται στους "Transformers" από τη πρώτη πρόταση. Άρα η προηγούμενη φράση βοηθάει στο να μάθουμε την αναφορά και έννοια της λέξης "συγκρότημα" της δεύτερης φράσης. Και γενικά το να εξετάζουμε έναν όρο μέσω των

γειτόνων του είναι ένα μεγάλο πλεονέκτημα. Άρα εν γένει, εξετάζεται μία πρόταση μέσω πολλών επαναλήψεων, στις οποίες επαναλήψεις εξετάζονται οι σχέσεις που έχει ο κάθε όρος με τους γείτονές του.



Εικόνα 8: Transformers in NLP

3.2) Προ-εκπαίδευση του Μοντέλου

Για την προ-εκπαίδευση του μοντέλου χρησιμοποιούνται δύο στρατηγικές χωρίς επίβλεψη, το Γλωσσικό Μοντέλο Απόκρυψης (Masked Language Model) και η Πρόβλεψη της Επόμενης Πρότασης (Next Sentence Prediction). Το σώμα που χρησιμοποιήθηκε για την προ-εκπαίδευση του αγγλικού μοντέλου BERT ήταν το BooksCorpus και η αγγλική Βικιπαίδεια.

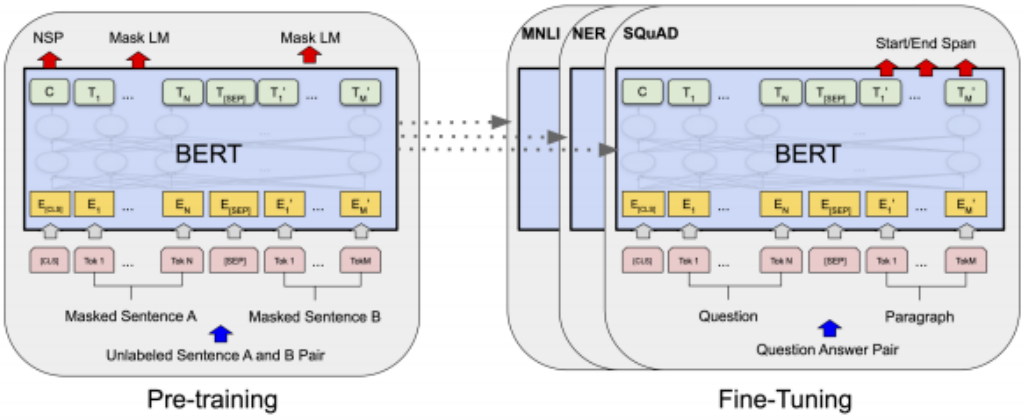
1. Γλωσσικό Μοντέλο Απόκρυψης

Πρόκειται για ένα μοντέλο το οποίο αντικαθιστά το 15% των λέξεων της ακολουθίας εισόδου με τα σύμβολα [MASK]. Η αντικατάσταση των λέξεων γίνεται με τυχαίο τρόπο κάθε φορά και ο αντικειμενικός σκοπός είναι το μοντέλο να προβλέψει όλες εκείνες τις λέξεις που

αντικαταστάθηκαν, βασιζόμενο σε όλες τις υπόλοιπες λέξεις που βρίσκονται μέσα στην ακολουθία. Σε αντίθεση με τους αυτόματους κωδικοποιητές εξομάλυνσης (denoising auto-encoders), η πρόβλεψη αφορά μόνο τις λέξεις οι οποίες αποκρύφτηκαν και δεν γίνεται ανακατασκευή ολόκληρης της ακολουθίας εισόδου.

2. Πρόβλεψη της Επόμενης Πρότασης

Ο αντικειμενικός σκοπός είναι το μοντέλο να προβλέψει την ύπαρξη σχέσης ή μη ανάμεσα σε δύο προτάσεις. Για την εκπαίδευση του μοντέλου χρησιμοποιείται ένα σύνολο δεδομένων αποτελούμενο από ζεύγη προτάσεων. Στο 50% των περιπτώσεων η δεύτερη πρόταση του ζεύγους είναι ακριβώς εκείνη η πρόταση η οποία διαδέχεται την πρώτη σύμφωνα με το πρωτότυπο κείμενο, ενώ στο υπόλοιπο 50% των περιπτώσεων η επιλογή της γίνεται με τυχαίο τρόπο. Πρόκειται λοιπόν για ένα πρόβλημα ταξινόμησης με ετικέτες ‘Είναι η Επόμενη’ (IsNext) και ‘Δεν Είναι η Επόμενη’ (NotNext). Κατά τη διαδικασία της προσαρμογής (fine-tuning) , το μοντέλο αρχικοποιείται με τις παραμέτρους που έχουν προκύψει από την προ-εκπαίδευση και στη συνέχεια ακολουθεί εκ νέου εκπαίδευση, με τη διαφορά ότι τα δεδομένα πλέον είναι επισημειωμένα και συγκεκριμένου τύπου, ανάλογα με το πρόβλημα που τίθεται κάθε φορά. Κάθε πρόβλημα έχει ξεχωριστά μοντέλα προσαρμογής, παρόλο που αυτά αρχικοποιούνται με τις ίδιες προ-εκπαιδευμένες παραμέτρους. Ένα ξεχωριστό χαρακτηριστικό του BERT είναι η ενοποιημένη αρχιτεκτονική σε διαφορετικού τύπου προβλήματα, καθώς η διαφοροποίηση ανάμεσα στην αρχιτεκτονική της προ-εκπαίδευσης και στην αρχιτεκτονική της προσαρμογής του μοντέλου είναι πολύ μικρή. Παρακάτω παρουσιάζονται οι διαδικασίες για το μοντέλο BERT στο πρόβλημα της απάντησης ερωτήσεων. Συγκριτικά με την προ-εκπαίδευση, η διαδικασία της προσαρμογής είναι σχετικά ‘ανέξοδη’ καθώς οι παράμετροι που πρέπει να ‘μάθει’ το μοντέλο είναι πολύ λίγες.



Εικόνα 9: Pre-training and Fine-Tuning

Οι διαδικασίες της προ-εκπαίδευσης και της προσαρμογής για το μοντέλο Bert στο πρόβλημα της απάντησης ερωτήσεων (SQuAD)[1].
Για την προ-εκπαίδευση χρησιμοποιούνται δύο στρατηγικές χωρίς επίβλεψη - το Γλωσσικό Μοντέλο Απόκρυψης (Masked Language Model, Mask LM) και η Πρόβλεψη της Επόμενης Πρότασης (Next Sentence Prediction, NSP). Κατά τη διαδικασία της προσαρμογής, χρησιμοποιούνται οι ίδιες προ-εκπαιδευμένες παράμετροι για να αρχικοποιήσουν μοντέλα για διαφορετικού τύπου προβλήματα - αναγνώριση ονοματικών οντοτήτων (Named-entity recognition, NER), πρόβλημα συνεπαγωγής (Multi-Genre Natural Language Inference, MNLI). Εκτός από τα επίπεδα εξόδου, η αρχιτεκτονική που χρησιμοποιείται και στις δύο διαδικασίες είναι η ίδια.

3.3) Προσαρμογή μοντέλου (fine-tuning)

Κατά τη διαδικασία της προσαρμογής (fine-tuning), το μοντέλο αρχικοποιείται με τις παραμέτρους που έχουν προκύψει από την προ-εκπαίδευση και στη συνέχεια ακολουθεί εκ νέου εκπαίδευση, με τη διαφορά ότι τα δεδομένα πλέον είναι επισημειωμένα και συγκεκριμένου τύπου, ανάλογα με το πρόβλημα που τίθεται κάθε φορά. Κάθε πρόβλημα έχει ξεχωριστά μοντέλα προσαρμογής, παρόλο που αυτά αρχικοποιούνται με τις ίδιες προ-εκπαιδευμένες παραμέτρους. Παρόλης της μεγάλης ποικιλίας των δεδομένων των οποίων έχει προ-εκπαιδευτεί το Γλωσσικό Μοντέλο, το πιθανότερο είναι ότι τα δεδομένα της εργασίας της

[1] https://rajpurkar.github.io/SQuAD-explorer/

οποίας θέλουμε να εφαρμόσουμε, διαφέρουν. Για αυτό το λόγο εφαρμόζουμε Fine-Tuning πάνω στο Γλωσσικό Μοντέλο με τα δεδομένα της εργασίας-στόχου. Η διαδικασία αυτή είναι αρκετά ταχύτερη από την εκπαίδευση του Γλωσσικού Μοντέλου σε γενικού τομέα δεδομένα, καθώς το μόνο που χρειάζεται είναι το μοντέλο να υιοθετήσει τις ιδιαιτερότητες των δεδομένων δίνοντας την δυνατότητα της δημιουργίας ενός ισχυρού Γλωσσικού Μοντέλου ακόμα και από μικρές συλλογές δεδομένων.

3.4) Προ-εκπαιδευμένα μοντέλα BERT

Για τα προ-εκπαιδευμένα μοντέλα BERT (Devlin et al, 2018) υπάρχουν δύο εκδοχές

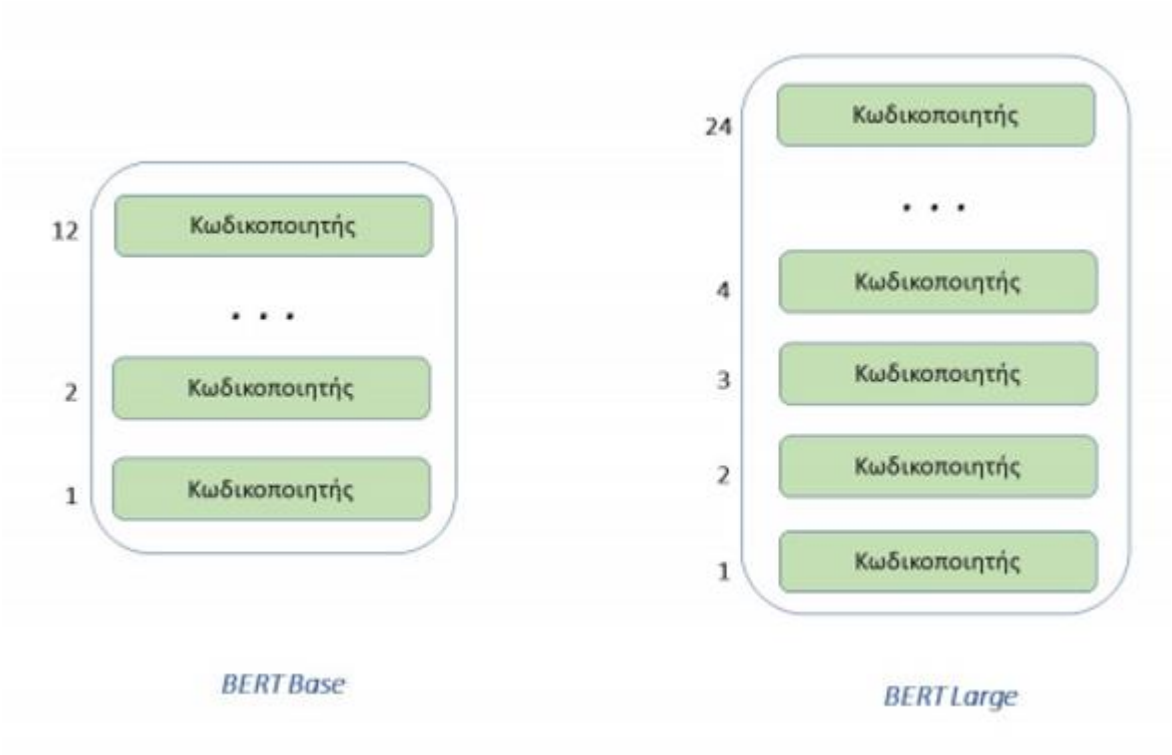
• BERTBase (βασικό μοντέλο)

• BERTLarge (μεγάλο μοντέλο)

Και στα δύο μοντέλα υπάρχει προς μεγάλος αριθμός από επίπεδα κωδικοποιητών. Τα επίπεδα αυτά αποτελούνται από μηχανισμούς αυτό-προσοχής και καθένας από προς ακολουθείται από ένα προς τα εμπρός τροφοδοτούμενο νευρωνικό δίκτυο. Ωστόσο, σε αντίθεση με την προκαθορισμένη διαμόρφωση του Μετασχηματιστή (L = 6 επίπεδα κωδικοποιητών, H = 512 κρυμμένες μονάδες και A = 8 κεφαλές), τα προς τα εμπρός τροφοδοτούμενα νευρωνικά δίκτυα των μοντέλων BERT είναι μεγαλύτερα και οι κεφαλές του μηχανισμού αυτό-προσοχής είναι περισσότερες. Συγκεκριμένα:

• Βασικό μοντέλο : 12 επίπεδα κωδικοποιητών, 768 κρυμμένες μονάδες, 12 κεφαλές, 110M παράμετροι.

• Μεγάλο μοντέλο : 24 επίπεδα κωδικοποιητών, 1024 κρυμμένες μονάδες, 16 κεφαλές, 340M παράμετροι.

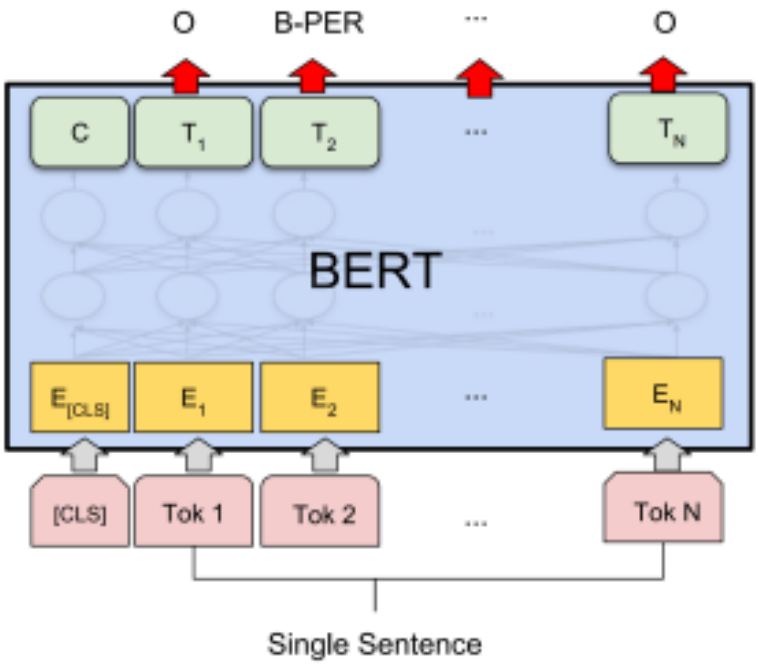


Εικόνα 10: BERT base vs BERT large

3.5) Είσοδος και Έξοδος μοντέλου Bert

Προκειμένου το μοντέλο BERT να μπορεί να χειρίζεται πολλά διαφορετικά προβλήματα, η είσοδος αναπαράστασης έχει σχεδιαστεί με τέτοιο τρόπο, ώστε να είναι εφικτή και ξεκάθαρη η αναπαράσταση όχι μίας μόνο πρότασης, αλλά και ενός ζεύγους προτάσεων (π.χ σε προβλήματα απάντησης ερωτήσεων) σε μία ακολουθία συμβόλων. Το πρώτο σύμβολο σε κάθε ακολουθία εισόδου είναι πάντα το ειδικό σύμβολο [CLS]. Πρόκειται για ένα σύμβολο ταξινόμησης, του οποίου οι τελευταίες κρυμμένες καταστάσεις χρησιμοποιούνται ως την αθροιστική ακολουθιακή αναπαράσταση σε προβλήματα ταξινόμησης. Όσον αφορά τα ζεύγη προτάσεων, τοποθετούνται μαζί σε μία ακολουθία και ο διαχωρισμός τους προκύπτει με δύο τρόπους. Πρώτα, διαχωρίζονται με το ειδικό σύμβολο [SEP] και έπειτα σε κάθε ένα σύμβολο προστίθεται ένα εκπαιδευμένο διάνυσμα ενσωμάτωσης (learned embedding) που υποδεικνύει αν αυτό το σύμβολο ανήκει στην πρόταση A ή στην πρόταση B.

Όπως φαίνεται και στο σχήμα παρακάτω, το διάνυσμα ενσωμάτωσης εισόδου δηλώνεται με το γράμμα E, το τελικό κρυμμένο διάνυσμα του ειδικού συμβόλου [CLS] με το διάνυσμα $C \in R^H$ και το τελικό κρυμμένο διάνυσμα για το i-οστό σύμβολο εισόδου με το διάνυσμα $Ti \in R^H$.



Εικόνα 11: Είσοδος και Έξοδος μοντέλου Bert

4) Δημιουργία και εκπαίδευση κατηγοριοποιητή και υλοποίηση Web εφαρμογής για χρήση του εκπαιδευμένου μοντέλου

Στόχος της συγκεκριμένης υλοποίησης ήταν η δημιουργία ενός κατηγοριοποιητή με τη χρήση του προ-εκπαιδευμένου μοντέλου Greek BERT. Ο κατηγοριοποιητής θα έπρεπε να εκπαιδευτεί και να μπορεί να κατηγοριοποιήσει ελληνικές ειδήσεις και άρθρα σε μια ή και περισσότερες από τις διαθέσιμες 17 κατηγορίες που είχαμε στην διάθεσή μας. Στην συνέχεια και αφού τα αποτελέσματα των test ήταν ικανοποιητικά, θα έπρεπε να υλοποιηθεί και μια web εφαρμογή ώστε να δοθεί η δυνατότητα σε απλούς χρήστες να μπορούν να κατηγοριοποιήσουν ελληνικές ειδήσεις ή άρθρα, καθώς επίσης και να αλληλοεπιδρούν με τον κατηγοριοποιητή συμφωνώντας ή όχι με την κατηγοριοποίηση. Τέλος ο κατηγοριοποιητής θα έπρεπε να εκπαιδεύεται εκ νέου με τις διορθώσεις ώστε να φτάσει στο μέγιστο βαθμό ακρίβειας των επόμενων αποτελεσμάτων.

4.1) Δημιουργία και Εκπαίδευση μοντέλου

4.1.1) Επεξεργασία Dataset

Για τη δημιουργία και εκπαίδευση ενός μοντέλου είναι απαραίτητη η ύπαρξη ενός dataset που να περιέχει πληθώρα δεδομένων και στοιχείων. Για την επίτευξη της συγκεκριμένης εργασίας είχαμε στην διάθεσή μας ένα json αρχείο με ελληνικές ειδήσεις και τις κατηγορίες στις οποίες ανήκει το καθένα.


```
{
  "category":[
    {
      "id":1,
      "name":"Ελλάδα:Πολιτική",
      "short_name":"Ελλάδα:Πολιτική"
    },
    {
      "id":2,
      "name":"Ελλάδα:Κοινωνία",
      "short_name":"Ελλάδα:Κοινωνία"
    },
    {
      "id":3,
      "name":"Ελλάδα:Οικονομία",
      "short_name":"Ελλάδα:Οικονομία"
    },
    {
      "id":4,
      "name":"Διεθνή:Πολιτική",
      "short_name":"Διεθνή:Πολιτική"
    },
    {
      "id":5,
      "name":"Διεθνή:Κοινωνία",
      "short_name":"Διεθνή:Κοινωνία"
    },
    {
      "id":6,
      "name":"Διεθνή:Οικονομία",
      "short_name":"Διεθνή:Οικονομία"
    },
    {
      "id":7,
      "name":"Επιστήμες",
      "short_name":"Επιστήμες"
    },
    {
      "id":8,
      "name":"Πολιτισμός",
```

```
      "short_name":"Πολιτισμός"
    },
    {
      "id":9,
      "name":"Αθλητισμός",
      "short_name":"Αθλητισμός"
    },
    {
      "id":10,
      "name":"Βιοτροπία:Δραστηριότητες Ελεύθερου Χρόνου",
      "short_name":"Βιοτροπία:Δραστηριότητες_Ελεύθερου_Χρόνου"
    },
    {
      "id":11,
      "name":"Βιοτροπία:Διατροφή-Σώμα",
      "short_name":"Βιοτροπία:Διατροφή-Σώμα"
    },
    {
      "id":12,
      "name":"Βιοτροπία:Κοινωνικά-Κοσμικά-Τηλεοπτικά Νέα / Διαπροσωπικές Σχέσεις",
      "short_name":"Βιοτροπία:Κοινωνικά-Κοσμικά-Τηλεοπτικά_Νέα-Διαπροσωπικές_Σχέσεις"
    },
    {
      "id":13,
      "name":"Κατανάλωση:Σπίτι-Οικογένεια-Εμφάνιση",
      "short_name":"Κατανάλωση:Σπίτι-Οικογένεια-Εμφάνιση"
    },
    {
      "id":14,
      "name":"Κατανάλωση:Εμπορική Τεχνολογία",
      "short_name":"Κατανάλωση:Εμπορική_Τεχνολογία"
    },
    {
      "id":15,
      "name":"Κατανάλωση:Αυτοκίνηση",
      "short_name":"Κατανάλωση:Αυτοκίνηση"
    },
    {
      "id":16,
      "name":"Ξενόγλωσσα",
      "short_name":"Ξενόγλωσσα"
    },
    {
      "id":17,
      "name":"Άλλο",
      "short_name":"Άλλο"
    }
  ],
"corpus":[
    {
      "id":1,
      "content":"Μητσοτάκης: Τα πρόσθετα μέτρα θα βουλιάξουν την οικονομία\r\n  \r\n  Στη
Δυτική Αττική περιόδευσε ο πρόεδρος της ΝΔ Κυριάκος Μητσοτάκης και στις δηλώσεις\r\n  του
αναφέρθηκε στα μέτρα που ετοιμάζεται να ψηφίσει η κυβέρνηση. «Θα βουλιάξουν την
ικονομία\r\n  πιο βαθιά στην ύφεση και θα υπονομεύσουν κάθε ουσιαστική προσπάθεια για την
ανάκαμψή της»\r\n  τόνισε. «Δυστυχώς, και εδώ, στο Περιστέρι, επιβεβαιώνεται, για άλλη μια
```

φορά, ότι το μεγάλο\r\n πρόβλημα της ελληνικής οικονομίας είναι η αδυναμία της να πιστρέψει σε μια αναπτυξιακή\r\n δυναμική. » Η πολιτική της κυβέρνησης, όπως δρομολογείται με τα πρόσθετα μέτρα, τα οποία θα\r\n ψηφιστούν στη Βουλή, θα έχει τελικά ακριβώς το ανάποδο αποτέλεσμα. Θα βουλιάξει την οικονομία\r\n πιο βαθιά στην ύφεση και θα υπονομεύσει κάθε ουσιαστική προσπάθεια για την ανάκαμψή της»\r\n τόνισε. Ο ίδιος είπε ότι η ΝΔ έχει χρέος να είναι κοντά «στους ανθρώπους, οι οποίοι βάλλονται\r\n περισσότερο από την πολιτική του ΣΥΡΙΖΑ. Και να αποδείξουμε στην πράξη ότι η δήθεν κοινωνική\r\n ευαισθησία του ΣΥΡΙΖΑ για τις πιο αδύναμες κοινωνικές ομάδες είναι απλά ένας μύθος».",
"primary_category":1,
"creation_date":{
"date":"2021-01-01 00:00:00.000000",
"timezone_type":3,
"timezone":"Europe/Helsinki"
},
"modification_date":{
"date":"2021-01-01 00:00:00.000000",
"timezone_type":3,
"timezone":"Europe/Helsinki"
}
}
{
"id":2,
"content":"Κάμερον: «Εκανα λάθη στην υπόθεση των Panama Papers»\r\n \r\n «Θα έπρεπε να είχα διαχειριστεί καλύτερα την υπόθεση των Panama Papers» παραδέχτηκε ο\r\n πρωθυπουργός της Βρετανίας, Ντέιβιντ Κάμερον, δηλώνοντας ότι «το λάθος είναι δικό του», δύο\r\n ημέρες μετά την ομολογία του ότι τελικά είχε μετοχές της υπεράκτιας εταιρίας του πατέρα του, ο\r\n οποίος πέθανε το 2010. «Λοιπόν, δεν ήταν μια σπουδαία εβδομάδα αυτή. Ξέρω ότι θα έπρεπε να\r\n είχα διαχειριστεί καλύτερα αυτήν την υπόθεση, μην κατηγορείτε τους συμβούλους μου, το λάθος\r\n είναι δικό μου, πήρα το μάθημά μου» δήλωσε ο Κάμερον κατά τη διάρκεια του εαρινού συνεδρίου\r\n του Συντηρητικού Κόμματος στο Λονδίνο. Μετά τις αποκαλύψεις των Panama Papers την περασμένη\r\n Κυριακή, χρειάστηκε να εκδοθούν τέσσερις ανακοινώσεις με μπερδεμένες διατυπώσεις από τις\r\n υπηρεσίες του πρωθυπουργού πριν ο ίδιος αποφασίσει τελικά να παραδεχθεί το βράδυ της Πέμπτης\r\n ότι είχε μετοχές σε αυτή την εταιρία που εδρεύει στις Μπαχάμες. Εκφράζοντας παράλληλα τη λύπη\r\n του για την άσκηση μιας καταστροφικής επικοινωνιακής πολιτικής, ο ηγέτης των Συντηρητικών είχε\r\n διαβεβαιώσει ότι δεν έκανε τίποτε παράνομο και ότι πάντοτε πλήρωνε τους φόρους του. Ο Κάμερον\r\n επανέλαβε την υπόσχεσή του να δώσει στη δημοσιότητα «προσεχώς» τις φορολογικές του δηλώσεις\r\n των τελευταίων ετών, κάτι πρωτοφανές για πρωθυπουργό στην ιστορία της Βρετανίας. Την ίδια ώρα,\r\n και μόλις δύο χιλιόμετρα πιο μακριά, αρκετές εκατοντάδες διαδηλωτές είχαν συγκεντρωθεί στη\r\n Ντάουνινγκ Στριτ ζητώντας την παραίτησή του. «Κάμερον πρέπει να φύγεις» φώναζαν οι\r\n συγκεντρωμένοι, ορισμένοι φορώντας καπέλα Παναμά και άλλοι, πιο τολμηροί, χαβανέζικα\r\n πουκάμισα. «Ο πρωθυπουργός έχασε την εμπιστοσύνη των Βρετανών» σχολίασε την Παρασκευή ο ηγέτης\r\n της εργατικής αντιπολίτευσης Τζέρεμι Κόρμπιν, χωρίς ωστόσο να ζητήσει την παραίτησή\r\n του.",
"primary_category":4,
"secondary_category":5,
"creation_date":{
"date":"2021-01-01 00:00:00.000000",
"timezone_type":3,
"timezone":"Europe/Helsinki"
},
"modification_date":{
"date":"2021-01-01 00:00:00.000000",
"timezone_type":3,
"timezone":"Europe/Helsinki"

```
  }
}
```

Αφού επεξεργαστήκαμε το αρχείο, είδαμε πως υπήρχαν συνολικά 3992 ειδήσεις. Παρακάτω παρουσιάζονται οι κατηγορίες και οι συνολικές ειδήσεις που ανήκαν σε κάθε κατηγορία. Να σημειωθεί ότι κάθε είδηση μπορεί να ανήκει σε παραπάνω από 1 κατηγορίες.

ID	Κατηγορία	Συνολικές Εγγραφές
1	Ελλάδα:Πολιτική	261
2	Ελλάδα:Κοινωνία	252
3	Ελλάδα:Οικονομία	252
4	Διεθνή:Πολιτική	299
5	Διεθνή:Κοινωνία	244
6	Διεθνή:Οικονομία	222
7	Επιστήμες – Υγεία	248
8	Πολιτισμός	240
9	Αθλητισμός	348
10	Βιοτροπία:Δραστηριότητες Ελεύθερου Χρόνου	277
11	Βιοτροπία:Διατροφή-Σώμα	302
12	Βιοτροπία:Κοινωνικά-Κοσμικά-Τηλεοπτικά Νέα-Διαπροσωπικές Σχέσεις	295
13	Κατανάλωση:Σπίτι-Οικογένεια-Εμφάνιση	227
14	Κατανάλωση:Εμπορική Τεχνολογία	277
15	Κατανάλωση:Αυτοκίνηση	253
16	Ξενόγλωσσα	343
17	Άλλο	0

Εικόνα 13: Συνολικά κείμενα ανά κατηγορία στο dataset

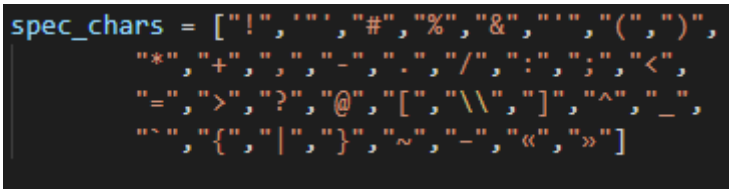
Στη συνέχεια δημιουργήσαμε ένα excel αρχείο (.xslx) με 18 στήλες. Η πρώτη στήλη είχε την ονομασία News όπου αποθηκεύσαμε τα κείμενα και στις άλλες 17 δώσαμε το όνομα της κάθε κατηγορίας. Όταν ένα κείμενο ανήκει σε κάποια κατηγορία, τότε στην συγκεκριμένη στήλη βάλαμε «1» ενώ στις άλλες «0» όπως φαίνεται στην παρακάτω εικόνα.

	News	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	Αύριο Τρίτη θα υπάρξουν οι σχετικές ανακοινώσεις και σύμφωνα με πληροφορίες από το φθινόπωρο στους κλειστού	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Πηγές από την Πυροσβεστική αναφέρουν ότι κατά την φετινή αντιπυρική περίοδο θυμούνται πολλές ακόμα εστίες φ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Την προσφιλή στα τουρκικά μέσα ενημέρωσης προπαγάνδα για τον Έβρο ξεκίνησε η φιλοκυβερνητική εφημερίδα Sc	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
5	Όπως αναφέρεται σε ανακοίνωση του υπουργείου Εργασίας και Κοινωνικών Υποθέσεων, το νομοσχέδιο αναπτύσσετ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	«Νεκροταφείο των αυτοκρατοριών»: Ο διάσημος τίτλος του αφιλόξενου, αιματοβαμμένου Αφγανιστάν έχει επανέλ	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	Οι εκδρομείς επιστρέφουν από τους τουριστικούς προορισμούς, ενώ ταυτόχρονα τους επόμενους μήνες θα αρχίσου	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Τη μάχη με το οριοθετημένο πλέον μέτωπο της φωτιάς στην Κάρυστο συνεχίζουν να δίνουν οι πυροσβέστες καθώς οι	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Για ένα «μεγάλο δράμα» που ζει ο ίδιος και η οικογένειά του μίλησε ο κύριος Γιάννης, ο πατέρας του 9χρονου αγορι	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	Οι «καταστροφικές» πλημμύρες στην πολιτεία Τενεσί στοίχισαν τη ζωή σε τουλάχιστον 21 ανθρώπους, ανέφεραν χθε	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
11	Ολυμπιακός: Παρουσία-έκπληξη με Σλόβαν - Μεγάλη ανατροπή στον Πειραιά (photos)Ο Ολυμπιακός προετοιμάζετα	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
12	Ολυμπιακός: Παρουσία-έκπληξη με Σλόβαν - Μεγάλη ανατροπή στον Πειραιά (photos)Ο Ολυμπιακός προετοιμάζετα	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
13	Ο κ. Οικονόμου αρχικά είπε ότι «υπάρχει κόσμος που επιμένει να μην εμβολιάζεται με επιχειρήματα που δεν γίνοντα	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Με ένα story του στο Instagram, ο Γιάννης Αντετοκούνμπο έκανε γνωστή ακόμη μια πολύ σπουδαία κίνηση του ίδιου	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0

Εικόνα 14: Μετασχηματισμός json σε excel

4.1.2) Επεξεργασία Κειμένων

Αρχικά, αφαιρέσαμε τα σημεία στίξης, ειδικούς χαρακτήρες, spaces και new lines από όλα τα κείμενα.



Εικόνα 15: Special Characters για διαγραφή

Στην συνέχεια μετατρέψαμε όλα τα κεφαλαία σε πεζά και αφαιρέσαμε τα stopwords.

(Η παρουσία ορισμένων λέξεων όπως είσαι, είμαι, στην, το, αλλά, όσα, κλπ., οι οποίες αποκαλούνται stopwords, στα δεδομένα δεν προσδίδει καμία χρήσιμη πληροφορία, αντιθέτως, εισάγει σημαντικό θόρυβο στα δεδομένα και αυξάνει κατά πολύ τον όγκο τους. Είναι εμφανές ότι τα φιλτραρισμένα δεδομένα έχουν μικρότερο όγκο και περιέχουν την ίδια σημαντική πληροφορία με τα αρχικά.)

4.1.3) Tokenization – Build dataloader

Με τη βοήθεια της βιβλιοθήκης pandas φορτώσαμε σαν dataframe το excel και χωρίσαμε το dataset σε 3 κατηγορίες,

- 60% (2396 εγγραφές) για εκπαίδευση του μοντέλου (training dataset)
- 20% (798 εγγραφές) για έλεγχο του μοντέλου (test dataset)
- 20% (798 εγγραφές) για επικύρωση του μοντέλου (validation dataset)

Το BERT απαιτεί συγκεκριμένη μορφή των δεδομένων εκπαίδευσης και για αυτό χρειάστηκε να επεξεργαστούμε τα δεδομένα εισόδου σύμφωνα με τον τρόπο που αυτό εκπαιδεύτηκε. Κάθε πρόταση που πρόκειται να εισαχθεί στο μοντέλο, αποτελεί μία ακολουθία λέξεων. Η ακολουθία αυτή πρέπει πρώτα να περάσει από τη διαδικασία του δειγματισμού (tokenization), δηλαδή να χωριστεί σε μέρη (λέξεις) και έπειτα κάθε λέξη να μετατραπεί σε έναν αριθμό (word id). Το προ-εκπαιδευμένο μοντέλο βασίζεται στον δικό του δειγματιστή (tokenizer), ο οποίος περιλαμβάνει επίσης και τη μετατροπή των λέξεων σε αριθμούς. Αναλυτικότερα, ο Δειγματιστής BERT (BERT-Tokenizer) βασίζεται στον αλγόριθμο WordPiece.

Φορτώσαμε τον Greek Bert από το huggingface (https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1) για να μετασχηματίσουμε τα δεδομένα μας λαμβάνοντας τα input_ids, attention_mask. Έπειτα φτιάξαμε dictionaries που αποτελούνται από τα text, τα input_ids, τα attention_mask, και τα labels.

Με την βοήθεια της βιβλιοθήκης torch δημιουργήσαμε dataloaders τα οποία αποτελούνται από τα dataset που δημιουργήσαμε προηγουμένως χωρισμένα σε 10 batch, με αποτέλεσμα να έχουμε 3 dataloaders (train_dataloader , test_dataloader , val_dataloader)

	TEXT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
136	πρεμιερα στην αγορα για το νεο samsung galaxy ...	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
35	απριλιου επιχειρησεις η κωτσοβολος φιλοξενει ...	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
273	το πρωτο βημα πως θα αντιμετωπισεις την υπερφα...	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
492	ισχυρος σεισμος ριχτερ στα συνορα πακιστα αφγ...	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
201	ο συγγραφεας του game of thrones αποκαλυπτει γ...	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
..	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
303	η ρωσια βουλιαξε την εθνικη γυναικων κατωτερη ...	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
127	ενα εφιαλτικο σεναριο το οποιο προκυπτει απο ...	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
403	looks για να αντιγραψουμε με οπλο τη μπεζ γκα...	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
149	δοκιμαζουμε τη νεα τετρακινητη bmw x συμπληρωσ...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
457	για να αποσυμφορηθει το σωμα η νηστεια μπορει ...	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Εικόνα 16: Παράδειγμα dataloader

4.1.4) Fine-Tuning στο Μοντέλο Ταξινομητή

Εφόσον εφαρμόσαμε Fine-Tuning στο Γλωσσικό Μοντέλο, χρησιμοποιήσαμε Fine-Tuning στο Μοντέλο Ταξινομητή (Classifier) μέσω του αποθηκεμένου encoder και στη συνέχεια προχωρήσαμε στην εκπαίδευση του ταξινομητή για 3 epochs.

Φορτώσαμε το BertModel.from_pretrained(nlpaueb/bert-base-greek-uncased-v1) περνώντας σαν είσοδο τα input_ids και attention_mask. Δώσαμε τα αποτελέσματα στον ταξινομητή που με την βοήθεια του γραμμικού μετασχηματισμού απόδωσε για κάθε άρθρο του train dataloader έναν αριθμό σε κάθε μια από τις 17 κατηγορίες. Με την βοήθεια του sigmoid φέρνουμε τα αποτελέσματα κάθε κειμένου για κάθε κατηγορία μεταξύ των αριθμών 0 – 1 όπως φαίνεται παρακάτω.

```
[0.3339, 0.5345, 0.5369, 0.3440, 0.4877, 0.4765, 0.4646, 0.4246, 0.6032,
 0.5410, 0.5611, 0.4661, 0.4727, 0.4133, 0.3811, 0.5114, 0.6254],
```

Εικόνα 17: Sigmoid – Αποτελέσματα μεταξύ 0-1

Όπως είναι φανερό στην αρχή και πριν εκπαιδευτεί το μοντέλο αποδίδει σε κάθε κατηγορία ποσοστό που δεν αποτελεί σωστή πρόβλεψη (σωστή επιλογή >0.5) :

```
[0.4094, 0.3963, 0.3266, 0.3087, 0.3196, 0.3595, 0.2611, 0.2667, 0.3849,
 0.3659, 0.4660, 0.2554, 0.3738, 0.3416, 0.3150, 0.3554, 0.5521],
[0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
```

Εικόνα 18: Αποτελέσματα στην αρχή της εκπαίδευσης

Ενώ όσο προχωράει η εκπαίδευση του μοντέλου έχουμε πολύ καλύτερα αποτελέσματα:

```
[0.0232, 0.0404, 0.0127, 0.9110, 0.9055, 0.0365, 0.0304, 0.0209, 0.0461,
 0.0296, 0.0212, 0.0483, 0.0270, 0.0282, 0.0134, 0.0204, 0.0185],
[0., 0., 0., 1., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
```

Εικόνα 19: Αποτελέσματα μετά τα μισά της εκπαίδευσης

Στη συνέχεια και αφού ολοκληρώθηκαν όλα τα epochs και η εκπαίδευση του μοντέλου, προχωρήσαμε σε αξιολόγηση ώστε να γίνει αποθήκευση του μοντέλου με το epoch με τα καλύτερα αποτελέσματα σε αρχείο με όνομα model.pt .

Έπειτα, προχωρήσαμε σε test του αποθηκευμένου μοντέλου στο test dataloader. Φορτώσαμε λοιπόν το αποθηκευμένο μοντέλο και προχωρήσαμε σε prediction για κάθε κείμενο.

Τέλος, συγκρίναμε τα αποτελέσματα και με την βοήθεια του classification_report από τη βιβλιοθήκη sklearn.metrics μετρήσαμε τις παρακάτω μετρικές ώστε να βγάλουμε συμπεράσματα για το μοντέλο.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Εικόνα 20: Μετρικές

Όπου:

- True Positive = όσα παραδείγματα ανήκουν στην κλάση X και ταξινομήθηκαν στην X
- False Negative = όσα παραδείγματα ανήκουν στην κλάση X, αλλά ταξινομήθηκαν στην Y
- False Positive = όσα παραδείγματα ανήκουν στην κλάση Y, αλλά ταξινομήθηκαν στην X
- True Negative = όσα παραδείγματα ανήκουν στην κλάση Y και ταξινομήθηκαν στην Y

Αναλυτικότερα, με το metric Precision μπορέσαμε να βρούμε πόσα από τα παραδείγματα που ο ταξινομητής ταξινόμησε ως θετικά ήταν πραγματικά θετικά. Με το metric Recall μπορέσαμε να βρούμε πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει και με το F1 score χρησιμοποιείται όταν θέλουμε να έχουμε μια ισορροπία ανάμεσα σε precission και recall.

Ένας άλλος διαχωρισμός που αξίζει να σημειώσουμε για τις μετρικές που αναφέρουμε είναι το averaging. Μία macro averaged μετρική υπολογίζεται παίρνοντας τον μέσο όρο των μετρικών που αφορούν κάθε κλάση ξεχωριστά, συνεπώς αντιμετωπίζει την κάθε κλάση ισότιμα. Μια micro averages μετρική προσθέτει τις συνεισφορές όλων των κλάσεων και υπολογίζει την μετρική στο συνολικό αποτέλεσμα.

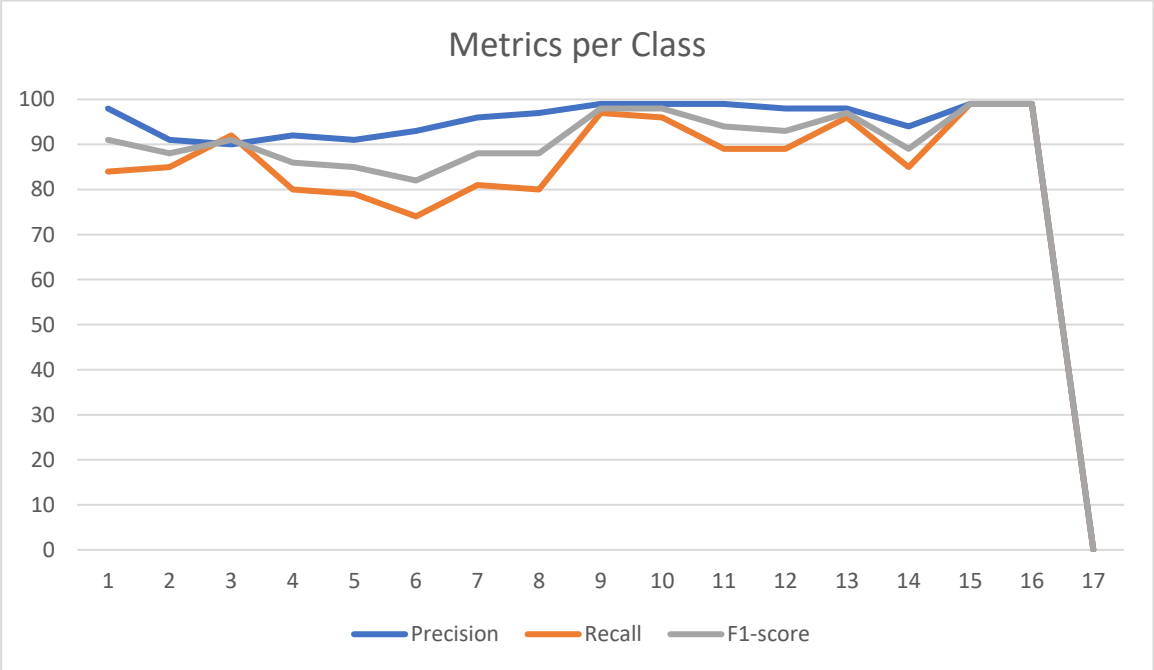
Όπως φαίνεται και στις παρακάτω εικόνες τα αποτελέσματα των μετρικών ήταν πολύ ικανοποιητικά με το accuracy να φτάνει το 93%, το precision, το recall και το f1-score για κάθε κλάση να είναι συνολικά >90% για όλες τις κλάσεις (με εξαίρεση την τελευταία που δεν είχαμε δεδομένα από το αρχικό dataset).

```
accuracy: 0.9371827465734617
```

Εικόνα 21: Accuracy

	precision	recall	f1-score	support
1	0.98	0.84	0.91	51
2	0.91	0.85	0.88	46
3	0.90	0.92	0.91	48
4	0.92	0.80	0.86	61
5	0.91	0.79	0.85	53
6	0.93	0.74	0.82	38
7	0.96	0.81	0.88	59
8	0.97	0.80	0.88	46
9	1.00	0.97	0.98	61
10	1.00	0.96	0.98	54
11	1.00	0.89	0.94	54
12	0.98	0.89	0.93	63
13	0.98	0.96	0.97	50
14	0.94	0.85	0.89	59
15	1.00	1.00	1.00	50
16	1.00	1.00	1.00	74
17	0.00	0.00	0.00	0
micro avg	0.96	0.88	0.92	867
macro avg	0.91	0.83	0.86	867
weighted avg	0.96	0.88	0.92	867
samples avg	0.94	0.91	0.92	867

Εικόνα 22: Αποτελέσματα Test



Εικόνα 23: Μετρικές ανα κλάση

4.1.5) Επανεκπαίδευση μοντέλου με νέα δεδομένα

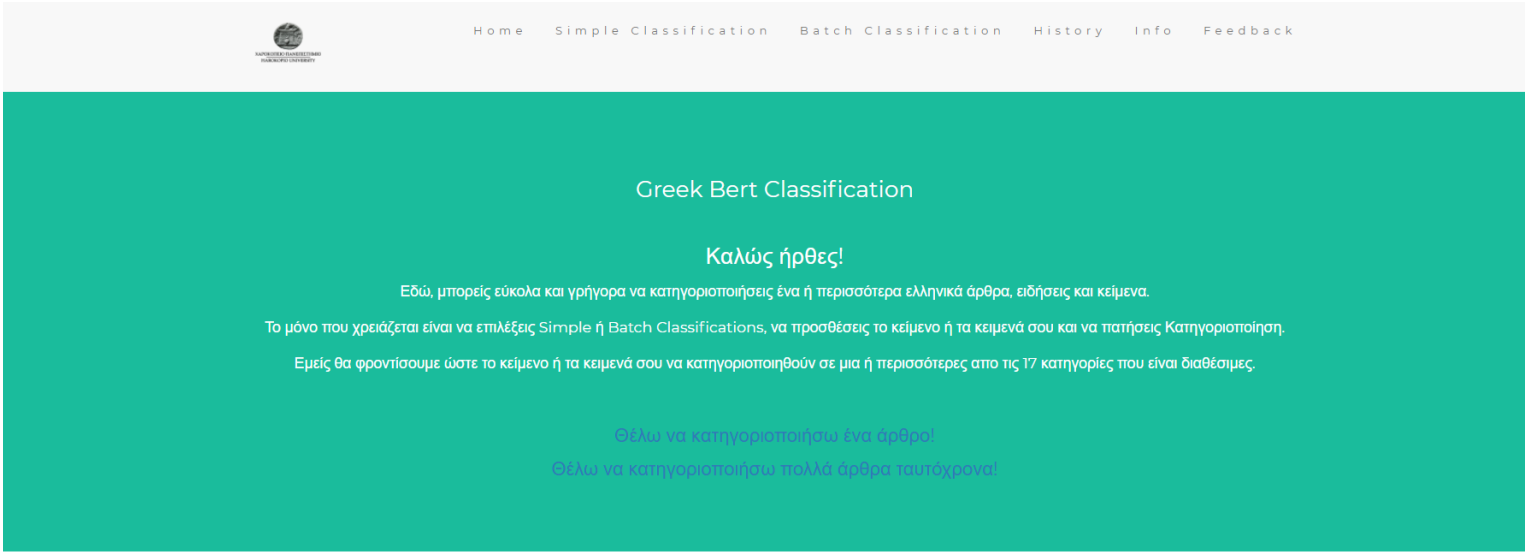
Πριν προχωρήσουμε στην δημιουργία της Web εφαρμογής υλοποιήσαμε ένα process με το οποίο μπορούμε να επανεκπαιδεύσουμε το μοντέλο με νέα δεδομένα. Για παράδειγμα, δημιουργήσαμε ένα νέο dataset με 10 εγγραφές στο οποίο προσθέσαμε 4 κείμενα που αφορούσαν τον καιρό και τα εντάξαμε στην κατηγορία 17 (Άλλο). Προχωρήσαμε σε test για το συγκεκριμένο dataset με accuracy 60%, αφού το μοντέλο δεν πρόβλεψε σωστά την κατηγορία αυτών των 4 κειμένων. Στη συνέχεια προχωρήσαμε σε επανεκπαίδευση του μοντέλου για 20 epochs και έπειτα σε εκ νέου test. Με βάση τα αποτελέσματα, υπήρξε accuracy 100% , αφού το μοντέλο κατάφερε μετά την επανεκπαίδευση να προβλέψει σωστά τις κατηγορίες των 4 νέων κειμένων.

4.2) Δημιουργία Web εφαρμογής και επεξήγηση δυνατοτήτων

Με την βοήθεια του web framework της python, flask ,της γλώσσας προγραμματισμού html και την συλλογή εργαλείων bootstrap δημιουργήθηκε το web application με το οποίο υπάρχει αλληλεπίδραση του εκπαιδευμένου μοντέλου με τους χρήστες. Παρακάτω παρουσιάζονται και περιγράφονται οι σελίδες που έχουν δημιουργηθεί για την λειτουργία αυτή.

4.2.1) Επεξήγηση Αρχική Σελίδας

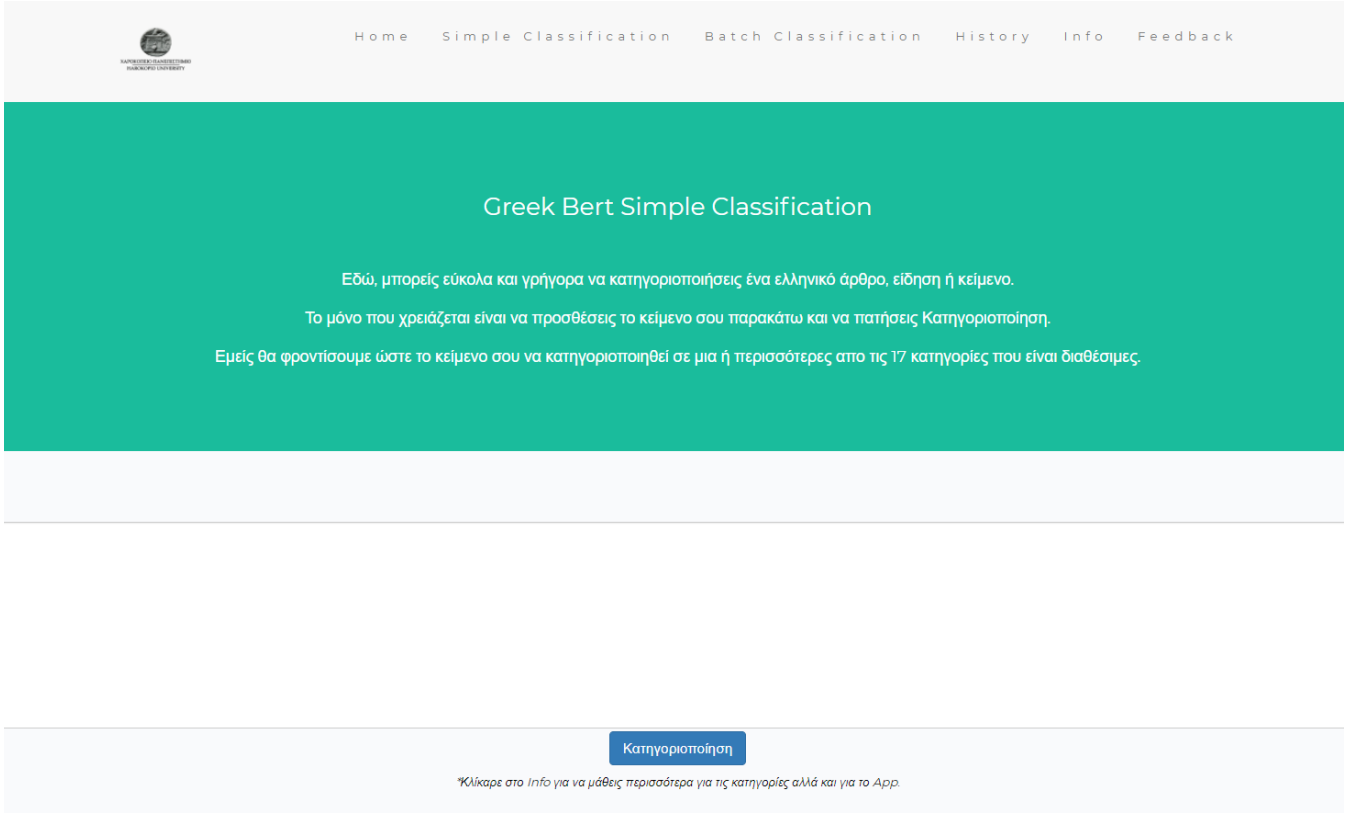
Στην αρχική σελίδα ο χρήστης μπορεί να ενημερωθεί σχετικά με τις δυνατότητες που του προσφέρει το συγκεκριμένο application. Έχει δυνατότητα να επιλέξει από το menu Simple Classification για να κατηγοριοποιήσει ένα κείμενο κάθε φορά ή Batch Classification για να κατηγοριοποιήσει παραπάνω από ένα κείμενα ταυτόχρονα. Επίσης έχει την επιλογή History με την οποία μπορεί να δει τα 10 τελευταία κείμενα που κατηγοριοποιήθηκαν και την επιλογή Info στην οποία παρουσιάζονται διαφάνειες που περιγράφουν τη χρήση του application, τις 17 κατηγορίες που είναι διαθέσιμες καθώς και πληροφορίες για το μοντέλο BERT. Τέλος μπορεί να αξιολογήσει την συγκεκριμένη εφαρμογή και την εμπειρία που είχε κατά τη χρήση της, πατώντας την επιλογή Feedback.



Εικόνα 24: Αρχική σελίδα εφαρμογής

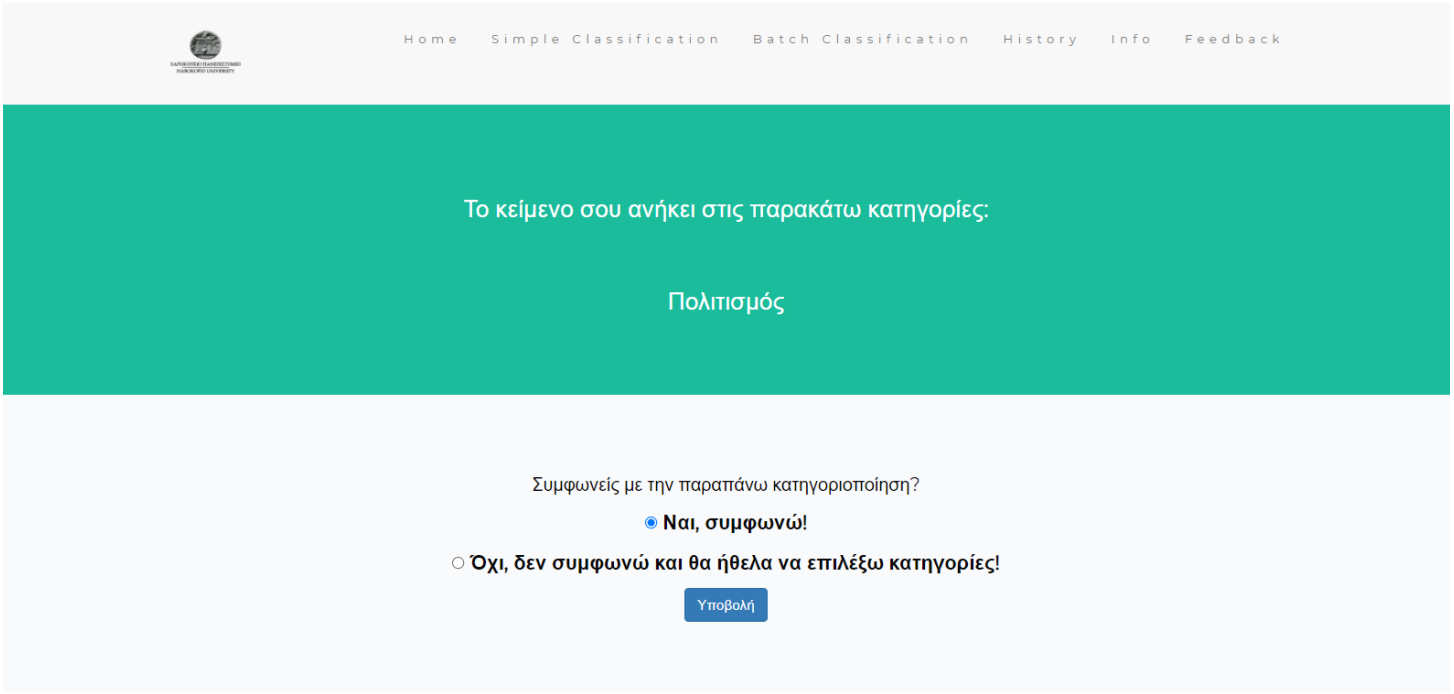
4.2.2) Επεξήγηση Simple Classification

Στην περίπτωση που ο χρήστης επιλέξει να κατηγοριοποιήσει ένα άρθρο τη φορά, εμφανίζονται στην οθόνη κάποιες λεπτομέρειες για τα βήματα που θα πρέπει να ακολουθήσει και το πεδίο που θα πρέπει να επικολλήσει το κείμενο του πριν πατήσει το κουμπί Κατηγοριοποίηση.



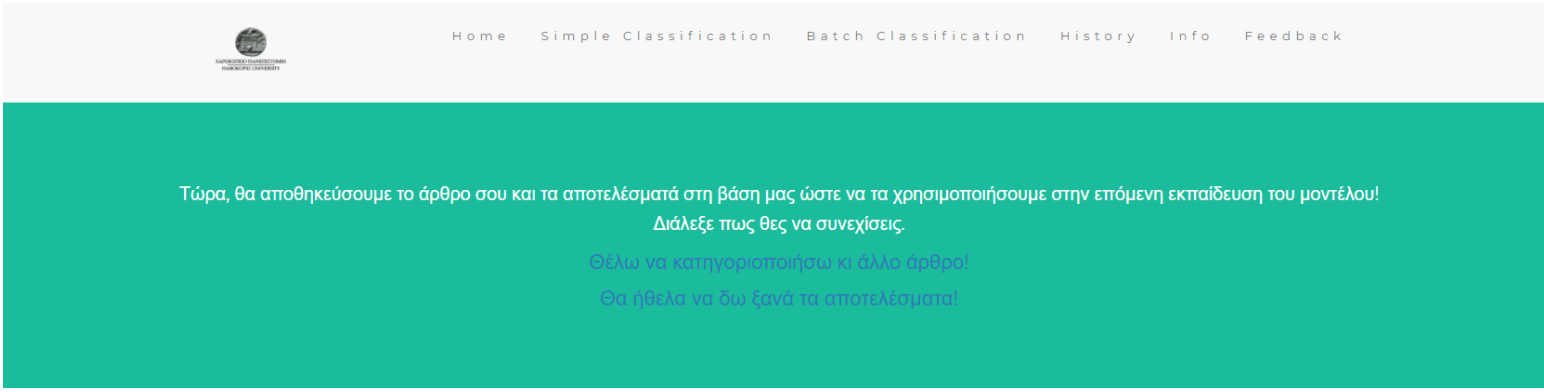
Εικόνα 25: Σελίδα Simple Classification

Σε λιγότερο από 1 δευτερόλεπτο από την επιλογή του χρήστη να Κατηγοριοποιήσει το κείμενο, θα εμφανιστεί η οθόνη με το αποτέλεσμα της κατηγοριοποίησης που αναφέρει σε ποια ή ποιες κατηγορίες ανήκει το κείμενο. Στην συνέχεια δίνεται η δυνατότητα στο χρήστη να επιλέξει αν συμφωνεί με την συγκεκριμένη κατηγοριοποίηση ή όχι.



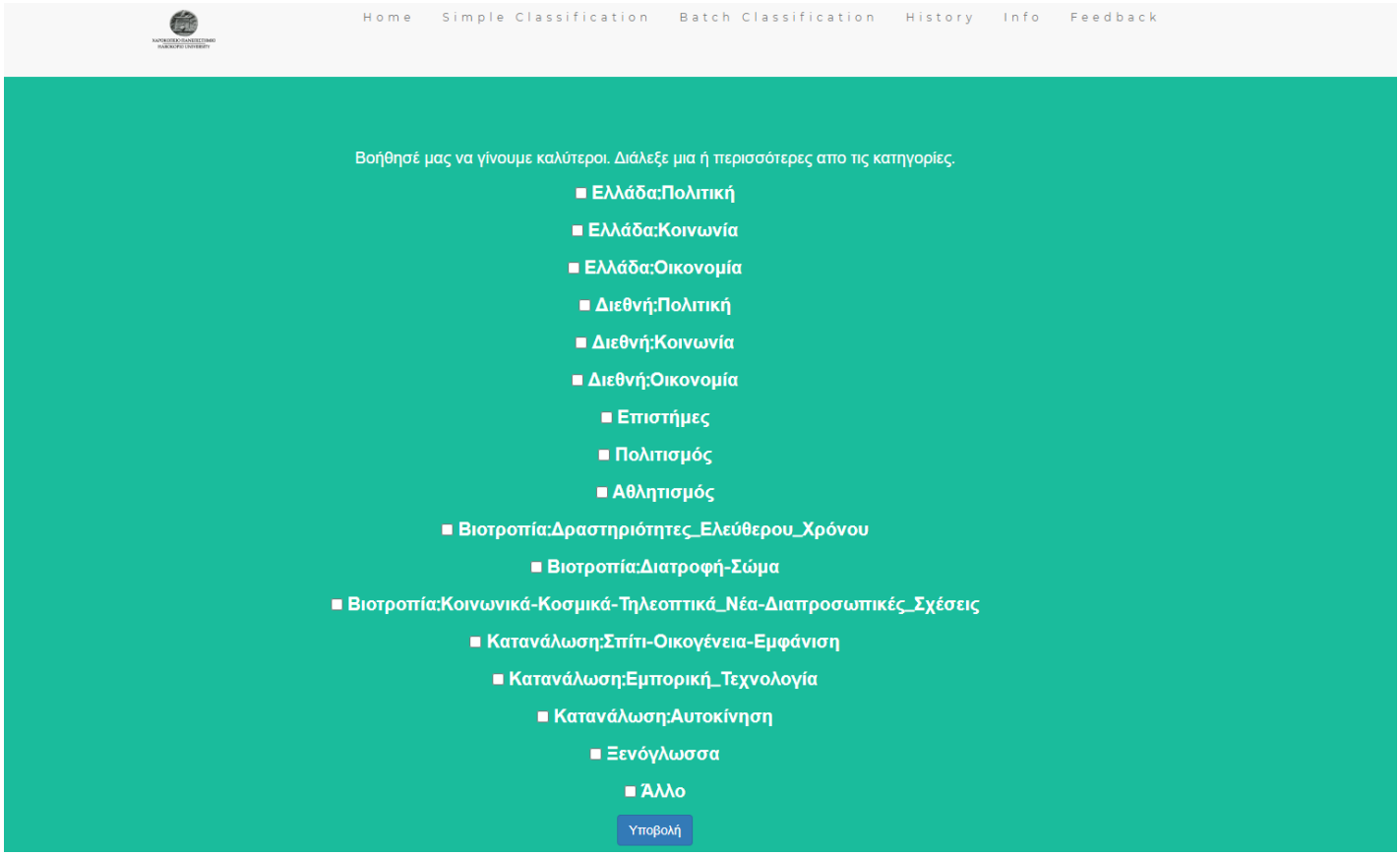
Εικόνα 26: Σελίδα αποτελεσμάτων Simple Classification

Στην περίπτωση που συμφωνεί με την κατηγοριοποίηση γίνεται αποθήκευση του κειμένου και τον κατηγοριών σε αρχείο excel. Επιπλέον εμφανίζεται η παρακάτω οθόνη δίνοντας τη δυνατότητα στο χρήστη να κατηγοριοποιήσει εκ νέου άλλο κείμενο ή να δει ξανά τα αποτελέσματα της προηγούμενης κατηγοριοποίησης. Σε αυτό το σημείο θα πρέπει να σημειωθεί πως σε περίπτωση που στο αρχείο excel αποθήκευσης των κειμένων υπάρχουν 10 εγγραφές, ενεργοποιείται στο background ένα process μέσω του οποίου γίνεται εκπαίδευση του μοντέλου από το τελευταίο checkpoint με τα νέα δεδομένα χωρίς αυτό να εμποδίζει τους χρήστες να προχωρήσουν σε εκ νέου κατηγοριοποίηση κειμένων.



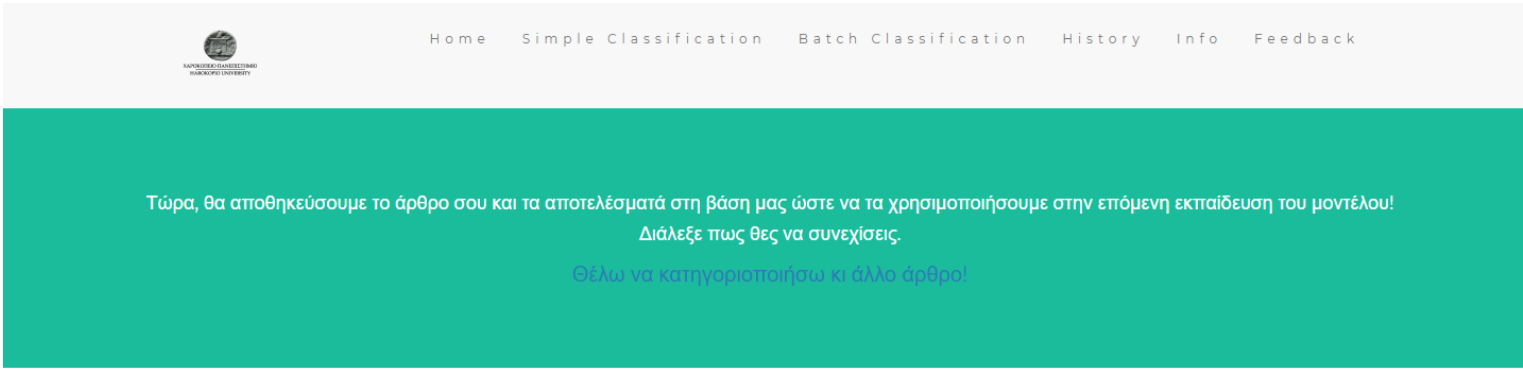
Εικόνα 27: Σελίδα αποθήκευσης αποτελεσμάτων Simple Classification

Στην περίπτωση που ο χρήστης δεν συμφωνεί με την κατηγοριοποίηση, εμφανίζεται η παρακάτω οθόνη δίνοντάς του τη δυνατότητα να επιλέξει σε ποιες κατηγορίες θα κατηγοριοποιούσε το συγκεκριμένο κείμενο.



Εικόνα 28: Σελίδα αλλαγής αποτελεσμάτων Simple Classification

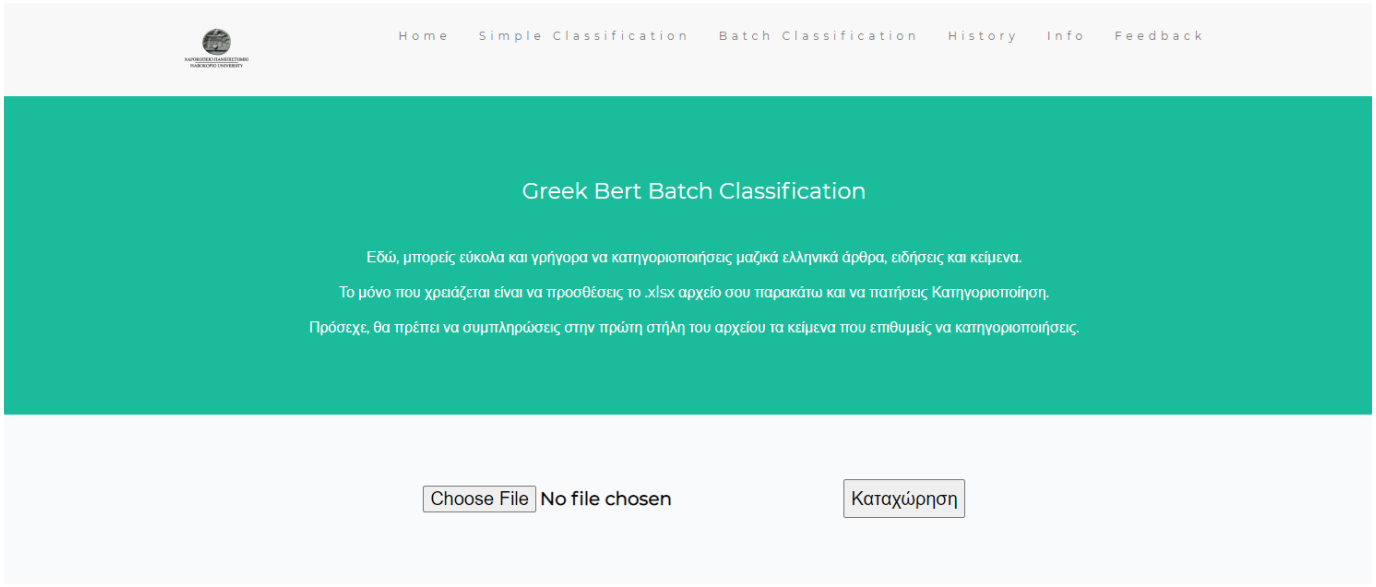
Στην συνέχεια και αφού γίνει η επιλογή των κατηγοριών και πατηθεί το κουμπί Υποβολή γίνεται αποθήκευση του κειμένου και των κατηγοριών στο αρχείο excel και δίνεται στον χρήστη η δυνατότητα να κατηγοριοποιήσει και άλλο κείμενο. Σε αυτό το σημείο θα πρέπει να σημειωθεί όπως και νωρίτερα πως σε περίπτωση που στο αρχείο excel αποθήκευσης των κειμένων υπάρχουν 10 εγγραφές, ενεργοποιείται ένα background process μέσω του οποίου γίνεται εκπαίδευση του μοντέλου από το τελευταίο checkpoint με τα νέα δεδομένα χωρίς αυτό να εμποδίζει τους χρήστες να προχωρήσουν σε εκ νέου κατηγοριοποίηση κειμένων.



Εικόνα 29: Σελίδα αποθήκευσης αλλαγών Simple Classification

4.2.3) Επεξήγηση Batch Classification

Στην περίπτωση που ο χρήστης επιλέξει να κατηγοριοποιήσει περισσότερα από ένα κείμενα ταυτόχρονα, τότε θα εμφανιστεί η παρακάτω οθόνη που περιέχει λεπτομέρειες χρήσης αυτής της επιλογής και θα δώσει τη δυνατότητα επιλογής του αρχείου που περιέχει τα κείμενα.



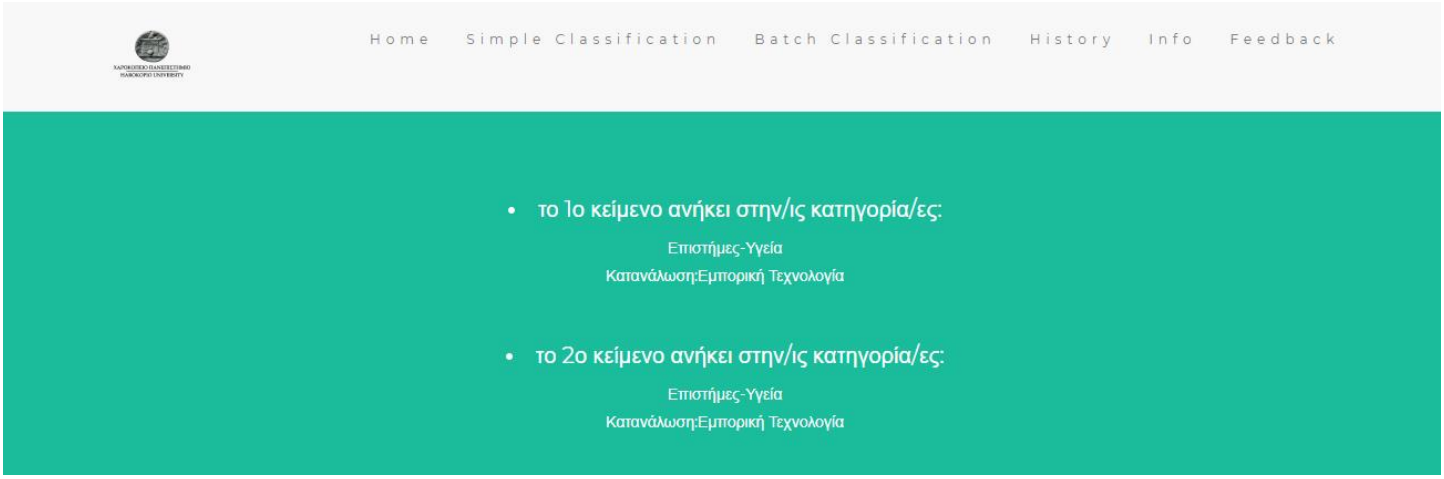
Εικόνα 30: Σελίδα Batch Classification

Σε αυτό το σημείο θα πρέπει να τονιστεί πως το αρχείο θα πρέπει να είναι excel (.xlsx) και να περιέχει τα κείμενα στην πρώτη στήλη του αρχείου όπως φαίνεται στην παρακάτω εικόνα.

	A	B	C	D	E	F
1	χρηματοδοτηση των τραπεζων ο γενικος δεικτης υποχωρησε και εκλεισε στις μοναδες στο επικεντρο των ρευστοποιησεων βρεθηκαν οι τραπεζικες μετοχες ειδικοτερα η εθνικη τραπεζα υπεστη					
2	των εκατ λιρων η εκατ ευρω η adele ειναι η πλουσιοτερη βρετανιδα τραγουδιστρια με την περιουσια της να αγγιζει τα εκατ λιρες η εκατ ευρω η χρονη τραγουδιστρια συμφωνα με την sun μονο περσι					
3	μελλοντικους επιβατες αρκετα μεγαλους και πρακτικους χωρους απριλιου η νεα mercedes benz glb που ειδαμε στην εκθεση αυτοκινητου στη σαγκαη μπορει να εχει το προσωνυμιο concept οπως η					
4	greek pm alexis tsipras requests prior agenda discussion in parliament over the state of the greek judicial system pic of request letter the greek judicial system is a major issue of contentionproto thema					
5	παπουτσια ψηλοτακουνα η φλατ glam chic η sneakers ποια ζευγαρια δεν θα αποχωριστουμε ουτε στιγμη αυτη την ανοιξη παπουτσια στις αποχρωσεις του ουρανιου τοξου αν σας αρεσουν τα εντονα					
6	κηδεια ενος κοινου τους φιλου ειναι απο τις ελαχιστες φορες που ο φακος τους συλλαμβανει μαζι και χωρις τα τρια τους παιδια μετα το πολυκροτο διαζυγιο τους τον ιουνιο αρκετα λυπημενοι					
7	βγαζει το ταμειο αξιοποιησης ιδιωτικης περιουσιας του δημοσιου με αγγελια του στον economist τρια ακινητα μεγαλης αξιας που κατεχει το ελληνικο δημοσιο στο εξωτερικο τα κτιρια που βγαινουν προς					
8	διπλασιαστηκαν σε εκταση χθες ανακοινωσαν οι αρχες οι οποιες εκτιμουν πως η κατασταση στην αλμπερτα παραμενει απροβλεπτη και επικινδυνη ... τα μεσανυχτα οι φωτιες που συνεχιζουν να					
9	harley οσο και αν μοιαζει ιεροσυλια ποιες πασιγνωστες μοτοσικλετες θα μπουν στην πριζα νοεμβριου ducati κτμ και αλλοι ευρωπαιοι κατασκευαστες μοτοσικλετων εχουν στα σκαρια τις ηλεκτρικες					

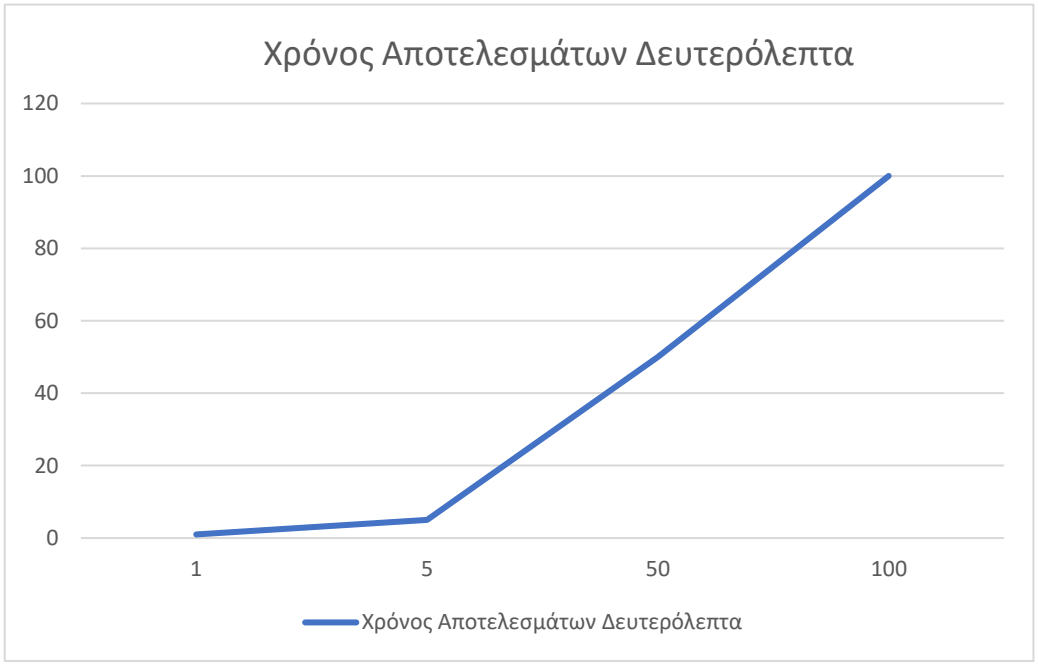
Εικόνα 31: Παράδειγμα ορθού excel για batch classification

Αφού γίνει upload το αρχείο και πατηθεί το Submit, θα εμφανιστούν στην οθόνη οι κατηγορίες που ανήκουν τα κείμενα που περιείχε το αρχείο.



Εικόνα 32: Σελίδα αποτελεσμάτων batch classification

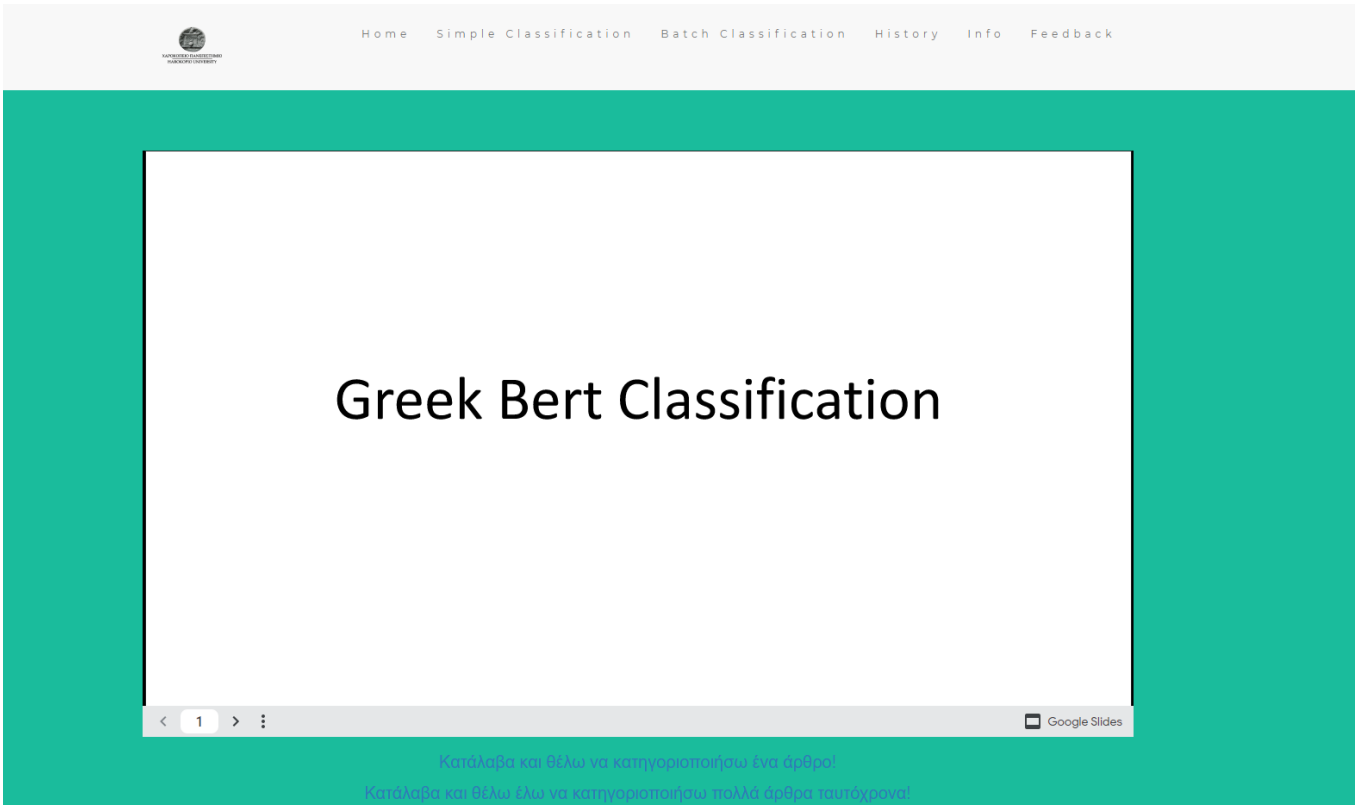
Ο χρόνος που απαιτείται είναι περίπου ένα δευτερόλεπτο για κάθε κείμενο όπως φαίνεται και παρακάτω:



Εικόνα 33: Χρόνος αποτελεσμάτων σε δευτερόλεπτα

4.2.4) Επεξήγηση Info

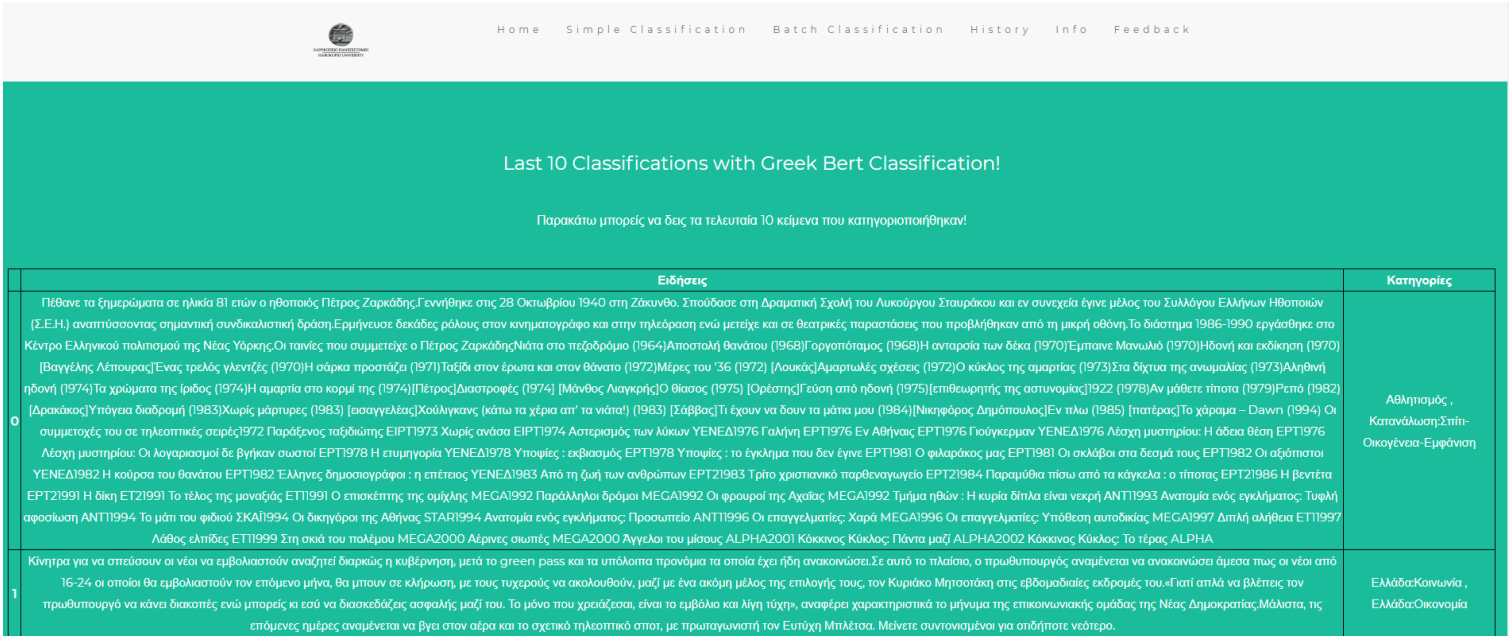
Στην περίπτωση που ο χρήστης θα επιλέξει την επιλογή Info, θα εμφανιστεί μια παρουσίαση με πληροφορίες και οδηγίες χρήσης της σελίδας, οι 17 κατηγορίες που μπορούν να κατηγοριοποιηθούν τα κείμενα, πληροφορίες για το μοντέλο BERT καθώς και θα δοθεί η δυνατότητα μεταφοράς ενός σελίδες κατηγοριοποίησης ενός ή πολλών κειμένων.



Εικόνα 34: Σελίδα Info

4.2.5) Επεξήγηση History

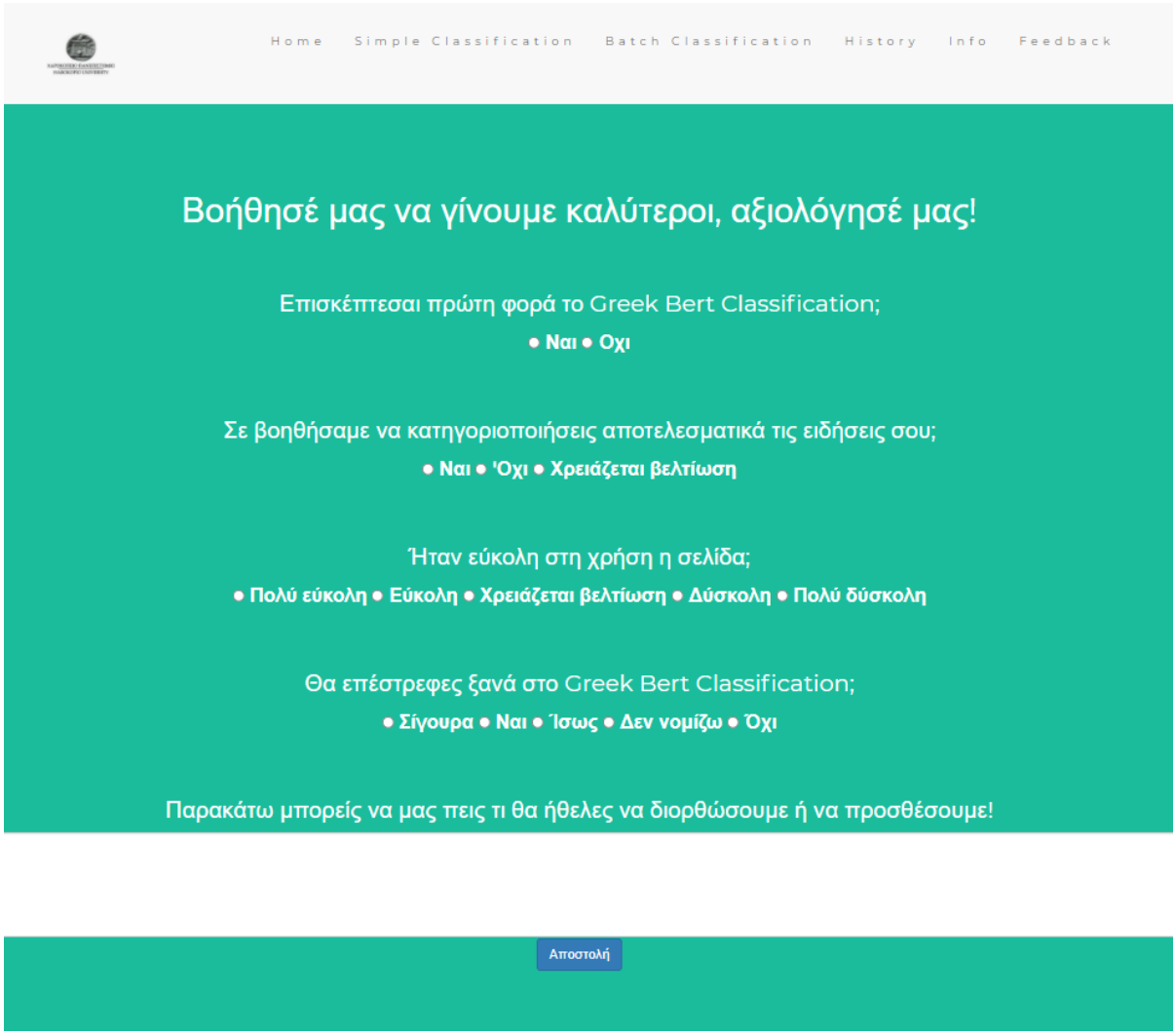
Στην περίπτωση που ο χρήστης επιλέξει την επιλογή History θα έχει τη δυνατότητα να δει τις 10 τελευταίες ειδήσεις που κατηγοριοποιήθηκαν.



Εικόνα 35: Σελίδα History

4.2.6) Επεξήγηση Feedback

Τέλος, στην περίπτωση που ο χρήστης επιλέξει την επιλογή Feedback, θα μπορεί να αξιολογήσει την εφαρμογή. Οι απαντήσεις αποθηκεύονται αυτόματα σε ένα excel αρχείο για μεταγενέστερη χρήση.



Εικόνα 36: Σελίδα Feedback

5) Συμπεράσματα και επόμενα βήματα

Σκοπός αυτής της διπλωματικής ήταν καταρχήν να γίνει μια εισαγωγή στις έννοιες τις μηχανικής μάθησης και της τεχνητής νοημοσύνης και στο μεγάλο βαθμό που τις συναντάμε στην καθημερινή ζωή των ανθρώπων. Μέσα από την ανάλυση της εξόρυξης δεδομένων έγινε κατανοητή η τεράστια σημασία του συγκεκριμένου κλάδου, αφού όσο περνάει ο καιρός το μέγεθος των δεδομένων αυξάνεται και η μελέτη τους δυσκολεύει. Επίσης έγινε αναφορά στην μεγάλη σημασία της αυτόματης ανάλυσης και κατηγοριοποίησης κειμένων και γενικότερα των δεδομένων αφού ο τεράστιος όγκος απαιτεί όλο και περισσότερο χρόνο. Στη συνέχεια έγινε ανάλυση του προ-εκπαιδευμένου μοντέλου BERT και στην αποτελεσματικότητα του σε προβλήματα ανάλυσης και κατηγοριοποίησης κειμένων καθώς επίσης και στην πολύ γρήγορη δυνατότητα του να εκπαιδεύεται σε καινούρια δεδομένα. Τέλος με την ανάπτυξη του μοντέλου κατηγοριοποιητή έγινε αντιληπτή η μεγάλη δυνατότητα που μας δίνουν τέτοιου είδους εκπαιδευμένα μοντέλα σε δύσκολα και μεγάλα προβλήματα γλιτώνοντας μας σημαντικό χρόνο.

Το μοντέλο που δημιουργήσαμε φάνηκε να ανταποκρίνεται στις ανάγκες του προβλήματος που θελήσαμε να επιλύσουμε αρχικά, φτάνοντας με την αρχική του εκπαίδευση ένα αρκετά ικανοποιητικό αποτέλεσμα μετρικών. Με την δυνατότητα που του προσθέσαμε να επανεκπαιδεύεται με τα νέα δεδομένα που προσθέτουν οι χρήστες, καταφέραμε να βάλουμε θεμέλια για ένα μοντέλο που θα φτάσει τα μέγιστα σημεία επιτυχίας στην κατηγοριοποίηση των κειμένων ανάλογα με τις ανάγκες των χρηστών που το χρησιμοποιούν. Όσον αφορά την εφαρμογή που υλοποιήσαμε φαίνεται να ανταποκρίνεται θετικά στον όγκο κατηγοριοποίησης και χρήσης.

Στα επόμενα βήματα, χρήσιμες λειτουργίες θα ήταν η προσθήκη της δυνατότητας στους χρήστες να μπορούν να αποφασίσουν για τα αποτελέσματα της batch διαδικασίας καθώς επίσης και η κατηγοριοποίησης της batch διαδικασίας να γίνεται πιο γρήγορα, αφού το 1 δευτερόλεπτο για κάθε κείμενο μπορεί να οδηγήσει σε μεγάλο χρονικό διάστημα στην περίπτωση των πολλών κειμένων. Επίσης δίνοντας την δυνατότητα στους χρήστες να αξιολογούν και να προτείνουν νέες δυνατότητες για την εφαρμογή, ευελπιστούμε να λάβουμε ιδέες που θα βοηθήσουν στην βελτίωσή της.

6)Βιβλιογραφία

[1] Natural Language Processing (NLP) for Machine Learning

https://towardsdatascience.com/natural-language-processing-nlp-for-machinelearning-d44498845d5b

[2] BERT(Language Model) https://en.wikipedia.org/wiki/BERT_(language_model)

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://arxiv.org/abs/1810.04805

[4] https://huggingface.co/

[5] https://en.wikipedia.org/wiki/Word_embedding

[6]https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lowerdimensional-space

[7] https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

[8] https://iq.opengenus.org/bert-base-vs-bert-large/

[9] https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets

[10] https://en.wikipedia.org/wiki/PyTorch

[11] https://www.tensorflow.org/tensorboard

[12] Aptè, C., Damerau, F.J. and Weiss, S.M. 1994. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12, 3, 233-251.

[13] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” in Proceedings of the Association for Computational Linguistics Conference, 2018.

[14] http://blog.datumbox.com/

[15] I. Kehayov, «Εξόρυξη δεδομένων για κατηγοριοποίηση κειμένων,» 2011

[16] M. H. Dunham, Β. Βρύκιος και Γ. Θεοδωρίσης, Data Minining, Εκδόσεις νεων τεχνολογιών, 2004.

[17] Clifton, C. (n.d.). Data mining. https://www.britannica.com/technology/data-mining

[18]https://www.europarl.europa.eu/news/el/headlines/society/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai

[19] https://searchenterpriseai.techtarget.com/definition/machine-learning-ML

[20] https://developers.google.com/machine-learning/guides/text-classification

[21] Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

[22] Uszkoreit, J. (2017). Transformer: A novel neural network architecture for language understanding. Google AI Blog, 31.

[23] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[24] https://www.tandfonline.com/doi/full/10.1080/09296174.2021.1885872

[25] https://www.sciencedirect.com/science/article/pii/S0952197616000117

[26] http://lpis.csd.auth.gr/publications/katakis_ecmlpkdd08_challenge.pdf

[27] https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence