ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Διαδίκτυο των Πραγμάτων: Ευφυή Περιβάλλοντα σε Δίκτυα Νέας Γενιάς

Πρόβλεψη Δικαστικών Αποφάσεων μέσω Μοντέλων Βαθιάς Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ορφανού Λεόντιου

Επιβλέπων : Δρ. Σταματάτος Ευστάθιος

Μέλη εξεταστικής επιτροπής:

Σάμος, Οκτώβριος 2024

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Η σελίδα αυτή είναι σκόπιμα λευκή

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Πρόλογος και ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον φίλο μου Μαμμά Κωνσταντίνο ο οποίος στάθηκε σε όλο αυτό το ταξίδι κατάρτισης ως μέντορας, σύμβουλος και φίλος. Τον ευχαριστώ που μου έδειξε τον δρόμο στην κατάκτηση ενός επιτεύγματος που ήταν εφηβικό όνειρο. Χωρίς την συμπαράσταση του ίσως παρέμενε ένα ανεκπλήρωτο όνειρο.

Επιπλέον θα ήθελα να ευχαριστήσω την Ιωάννα Μαρασιώτη που με έκανε να πιστέψω σε μένα και κατάφερα να αποκτήσω αυτό τον τίτλο και να εξελιχθώ επαγγελματικά. Ήταν η έμπνευση σε όλο αυτό το ταξίδι.

Τέλος θα ήθελα να ευχαριστήσω τον καθηγητή μου Σταματατο Ευστάθιο για τις βάσεις που απέκτησα και την καθοδήγηση του στην επίτευξη ενός από τους σημαντικότερους στόχους της ζωής μου.

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Η σελίδα αυτή είναι σκόπιμα λευκή

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Περιεχόμενα

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Περίληψη

Η παρούσα εργασία αποτελεί μια έρευνα πάνω στην επίδοση των μοντέλων βαθιάς μάθησης για την πρόβλεψη δικαστικών υποθέσεων. Πιο συγκεκριμένα μελετήθηκαν ελαφριά μοντέλα νευρωνικών δικτύων στην πρόβλεψη της δικαστικής απόφασης του ευρωπαϊκού συμβουλίου της ευρώπης. Οι δικαστικές υποθέσεις οι οποίες προσφεύγουν στο συγκεκριμένο θεσμικό όργανο αφορούν την παραβίαση των δικαιωμάτων του ανθρώπου των ευρωπαίων πολιτών. Το μοντέλα αυτά μέσω από την περιγραφή της δικαστικής υπόθεσης σε ελεύθερο κείμενο κατηγοριοποιούν το συγκεκριμένο κείμενο. Στην εργασία μελετήθηκαν δύο διαφορετικές περιπτώσης κατηγοριοποίησης. Η πρώτη αφορά την πρόβλεψη της δικαστικής απόφασης με βάση το δικαστικό κείμενο σε VIOLATED και NON_VIOLATED. Οι δύο αυτές κλάσεις αναφέρονται στο είδος της τελικής ετυμηγορίας σε παραβίαση άρθρου για τα θεμελιώδη δικαιώματα του ανθρώπου ή μη παραβίαση. Η δεύτερη περίπτωση αφορά την απόδοση σπουδαιότητας στο δικαστικό κείμενο για τον σχηματισμό μελλοντικής νομολογίας. Η σπουδαιότητα του κειμένου περιλαμβάνει της ακέραιες τιμές από το 1 μέχρι το 4, με την μικρότερη τιμή να αναφέρεται σε μεγαλύτερη σπουδαιότητα και την μείωση της σπουδαιότητας όσο αυτή αυξάνεται. Στην παρούσα εργασία η σπουδαιότητα με τιμή 1 και 2 έγιναν μία κλάση με τιμή IMPORTANT και η σπουδαιότητα με τιμή 3 και 4 σε μία κλάση με τιμή NOT_IMPORTANT. Στην εργασία παρατίθενται επίσης αποτελέσματα έτερων ερευνών πάνω στην επίλυση του ίδιου προβλήματος. Από τις συγκεκριμένες εργασίες χρησιμοποιήθηκαν αποτελέσματα παραδοσιακών μοντέλων μηχανικής μάθησης - μοντέλα που δεν αποτελούν βαθιά νευρωνικά δίκτυα ως την βάση σύγκρισης (baseline). Πέρα από τα παραδοσιακά μοντέλα χρησιμοποιήθηκαν και τα πιο σύγχρονα μοντέλα των εργασιών για την συγκρισή τους με τα μοντέλα που χρησιμοποιήθηκαν στην παρούσα εργασία.

Τα μοντέλα που επιλέχθηκαν στην παρούσα εργασία είναι τα medium BERT, DistilBERT, ALBert, RoBERTa, LongFormer και BigBird. Τα πρώτα τρία μοντέλα είναι ελαφριές παραλλαγές του μεγάλου μοντέλου BERT ενώ τα τελευταία δύο είναι μοντέλα τα οποία έχουν την δυνατότητα να τροφοδοτηθούν με κείμενα με μεγάλο αριθμό λέξεων. Όλα τα μοντέλα εκπαιδεύτηκαν σε πόρους σε περιβάλλον νέφους του παρόχου Amazon (AWS) με την υποστήριξη της της ΕΔΥΤΕ. Τα δεδομενα πανω στα οποια εκπαιδεύτηκαν τα μοντελα είναι ανοιχτά και προσβάσιμα στην επίσημη σελίδα του συμβουλίου της Ευρώπης (https://hudoc.echr.coe.int.). Τα δεδομένα συλλέχθηκαν και ταξινομήθηκαν σε δεδομένα εκπαίδευσης (train data), επαλήθευσης (validation data) και δοκιμής (test data) από τους (Chalkidis et al., 2019). Τα αποτελέσματα της κατηγοριοποίησης ως προς το αποτέλεσμα της απόφασης ήταν ιδιαίτερα ενθαρρυντικά ενώ η πρόβλεψη της σπουδαιότητας του μοντέλου αποτέλεσε μία δυσκολότερη εργασία. Στην συνεχεια

γινεται μια αναλυση των αποτελεσματων των πειραματων για τα διαφορετικα προβληματα που κληθηκαν να επιλυθούν στην παρουσα εργασια. Επιπροσθέτως μελετήθηκαν οι περιορισμοί που αναδύθηκαν κατα την διενέργεια των πειραμάτων καθώς επίσης προβληματισμοί για τις προκλήσεις τέτοιου είδους προβλημάτων. Τέλος γίνεται μία παράθεση των συμπερασμάτων από τα αποτελέσματα των πειραμάτων.

Εισαγωγή

Το σύγχρονο νομικό σύστημα ενός κράτους αποτελεί τον τομέα ο οποίος αναλαμβάνει την επίλυση των διαφορών και διενέξεων τόσο μεταξύ πολιτών όσο και μεταξύ κράτους και πολίτη. Η σοβαρότητα της αποστολής και το μέγεθος της ευθύνης που περιέχει μία δικαστική απόφαση έχει οδηγήσει σε δαιδαλώδεις διαδικασίες και γραφειοκρατία μέσα από την οποία σκοπός είναι η διασφάλιση της σωστής απονομής της δικαιοσύνης. Με αυτό τον τρόπο σήμερα απασχολούνται πολλοί πόροι ανθρώπινοι και φυσικοί στην λειτουργία του νομικού συστήματος. Ενδεικτικό είναι το ηλεκτρονικό δημοσίευμα της εφημερίδας “Καθημερινής” (Γεωργιοπούλου) όπου η Ελλάδα δαπανά το 0,35 % του ΑΕΠ της για το δικαστικό σύστημα ενώ για την έκδοση μία πρωτόδικης απόφασης ο μέσος όρος αναμονής στην Ελλάδα είναι περίπου 700 μέρες. Όλοι αυτή η πολυπλοκότητα και η καθυστέρηση του συστήματος καθιστούν την απονομή δικαιοσύνης μη αποδοτική για τους πολίτες (Γεωργιοπούλου, Τ. (2023, June 10). Δικαιοσύνη: Περίπου 700 ημέρες για μια πρωτόδικη απόφαση. *kathimerini*. Retrieved October 4, 2023, from https://www.kathimerini.gr/society/562462618/dikaiosyni-peripoy-700-imeres-gia-mia-protodiki-apofasi/).

Στον αντίποδα τα τελευταία χρόνια σημειώνεται σημαντικές επιτυχίες και επιτεύγματα στον τομέας της επιστήμης των υπολογιστών και συγκεκριμένα στην Μηχανική Μάθηση. Ειδικότερα στο κομμάτι της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) έχουν γίνει άλματα την τελευταία περίοδο. Έχουν δημιουργηθεί πλατφόρμες όπου ο χρήστης μέσω υποβολής ερωτήσεων αλληλεπιδρά με μεγάλα μοντέλα επεξεργασίας φυσικής γλώσσας (Large Language Models - LLMs) όπως το BARD της Google και το CHATGPT της OpenAI μέσα από τις οποίες μπορεί πλέον πέρα από το να αποκτήσει γνώσεις, να δημιουργήσει νέο περιεχόμενο ή ακόμα και να εμπνευστεί.

Αυτές οι εξελίξεις όπως είναι φυσικό έχουν επηρεάσει πολλούς τομείς καθώς πλέον τα LLMs είναι σε θέση να παρέχουν μία συνεχή καθοδήγηση σε επιτέλεση κάποιων εργασιών. Πέρα από τον συμβουλευτικό ρόλο έχουν αυτοματοποιήσει πολλές διαδικασίες οι οποίες παλιότερα απαιτούσαν επιπλέον πόρους. Μέσα στους τομείς που έχει ξεκινήσει η χρήση τεχνικών NLP είναι και ο νομικός τομέας λόγω των όλων αυτών προβλημάτων που αναφέραμε. Εκεί το NLP μέσω των εργαλείων που διαθέτει μπορεί να επισπεύσει τον χρόνο διεκπεραίωσης μίας

υπόθεσης όπως για παράδειγμα η δημιουργία περιλήψεων από μεγάλα νομικά κείμενα, οι ερωταπαντήσεις για θέματα νομικού αντικειμένου και η εξαγωγή ονοματικών οντοτήτων (Named Entity Extraction - NER).

Μία άλλη εφαρμογή του NLP στον νομικό τομέα είναι η πρόβλεψη της δικαστικής απόφασης (Legal Judgment Decision). Στο κομμάτι αυτό έχουν γίνει πολλές εργασίες σε διάφορα νομικά συστήματα χρησιμοποιώντας ακόμη και πολυγλωσσικά μοντέλα πρόβλεψης. Αυτά τα μοντέλα μπορούν να βοηθήσουν τους πολίτες παρέχοντας οι οποίοι δεν κατέχουν νομικές γνώσεις και να βοηθήσουν την πρόσβασή τους στην δικαιοσύνη. Επιπλέον μπορούν να αποτελέσουν εργαλεία δικαστών ώστε να μελετήσουν την πιθανότητα έκβασης μίας υπόθεσης και να μπορέσουν να οδηγηθούν σε πιο συνεπείς και ενημερωμένες αποφάσεις. Τέλος τα εργαλεία αυτά μπορούν να αποτελέσουν ένα εφόδιο σε ανεξάρτητες αρχές και παρατηρητές ώστε να μπορέσουν να εντοπίσουν δικαστικές αποφάσεις οι οποίες είναι μεροληπτικές.

Στην παρούσα διπλωματική εργασία γίνεται μία δοκιμή της αποτελεσματικότητας των state-of-the-art μοντέλων NLP στην πρόβλεψη των αποφάσεων του Ευρωπαϊκού Δικαστηρίου για τα Ανθρώπινα Δικαιώματα (European Council for Human Rights). Στο συγκεκριμένο δικαστήριο έχουν την δυνατότητα να προσφύγουν ιδιώτες και μη κυβερνητικές οργανώσεις κατά κρατών μελών σχετικά με την παραβίαση των δικαιωμάτων τους όπως για παράδειγμα ελευθερία λόγου, ελευθερία της θρησκείας ή της δίκαιης δίκης. Η προσφυγή αυτή αφορά μόνο πολίτες και όχι κυβερνητικών οντοτήτων.

Το Ε.Δ.Α.Δ. Στην ιστορία του αντιμετώπισε και αυτό προβλήματα διαχείρισης του μεγάλου όγκου προσφυγών. Ειδικότερα μετά το 1989 και την πτώση του τείχους του Βερολίνου οι προσφυγές αυξήθηκαν σε μεγάλο βαθμό με αποτελεσμα να κλονιστεί η αποτελεσματικότητα του δικαστηρίου. Το 1999, 8.400 υποθέσεις τέθηκαν σε ακρόαση. Το 2009 υπήρξαν 57.200 προσφυγές, με 119.300 να βρίσκονται σε εκκρεμότητα. Λόγω της συσσώρευσης των εκκρεμών υποθέσεων, το 2010 τέθηκε σε ισχύ το Πρωτόκολλο 14 της Ευρωπαϊκής Σύμβασης για τα Δικαιώματα του Ανθρώπου, όπου ένας νέος μηχανισμός εισήχθη για να βοηθήσει στην επιτάχυνση των διαδικασιών, προκειμένου να διατηρηθεί και να βελτιωθεί η αποτελεσματικότητα του συστήματος ελέγχου μακροπρόθεσμα (*Ευρωπαϊκό Δικαστήριο Ανθρωπίνων Δικαιωμάτων*. (n.d.). Βικιπαίδια. Retrieved 10 17, 2023,).

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Ανάλυση υπάρχουσας Βιβλιογραφίας

Data Structures για NLP αλγορίθμους

Η επεξεργασία κειμένου φυσικής γλώσσας απαιτεί την μετατροπή των λέξεων και τον προτάσεων ενός κειμένου σε ένα σύνολο αριθμών τα οποία θα μπορέσουν να είναι κατανοητά από ένα υπολογιστικό σύστημα το οποίο θα μπορέσει να κατανοήσει την πληροφορία που μεταφέρουν. Έτσι με την εξέλιξη του NLP στην Μηχανική Μάθηση έχουν αναπτυχθεί διάφορες τεχνικές και αλγόριθμοι με σκοπό την μετατροπή των λέξεων σε αυτές τις συλλογές τιμών. Ο τρόπος με τον οποίο αποδίδονται οι λέξεις σε μορφή κατανοητή από τα υπολογιστικά συστήματα είναι με την κατασκευή διανυσμάτων (vectors) τα οποία στο τομέα του NLP ονομάζονται embeddings.

One-hot-vectors

Οι one-hot-vectors είναι η πιο απλή μορφή αντιπροσώπευσης της λέξης όπου ουσιαστικά η κάθε λέξη είναι ένα διάνυσμα τιμών. Το συγκεκριμένο διάνυσμα αποτελείται από μηδενικές τιμές που αντιστοιχούν στις υπόλοιπες λέξεις ενός κειμένου και από την τιμή 1 στην θέση που βρίσκεται η συγκεκριμένη λέξη μέσα στο κείμενο. (Liu et al., 2023, 8). Ετσι λοιπόν προκύπτουν αραιά διανύσματα αποτελούμενα ως επί το πλείστον με μηδενικές τιμές τα οποία δεν φέρουν κάποια πληροφορία σχετικά με την σημασιολογία της λέξης

Bag-of-Words

Με το σκεπτικό των one-hot-vectors προτάθηκε στην συνέχεια η αντιπροσώπευση ενός εγγράφου ως ένα διάνυσμα όπου κάθε θέση του διανύσματος αντιστοιχεί στην λέξη από το λεξικό του εγγράφου και η τιμή του διανύσματος αντιστοιχεί στην συχνότητα εμφάνισης της λέξης στο σύνολο του εγγράφου (Zellig, 1954, 146–162).

```
(1) John likes to watch movies. Mary likes movies too.
```

```
(2) Mary also likes to watch football games.
```

Based on these two text documents, a list is constructed as follows for each document:

```
"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"
```

Representing each bag-of-words as a JSON object, and attributing to the respective JavaScript variable:

```
BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
BoW2 = {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};
```

(*Bag-Of-Words Model*, n.d.)

Η μετατροπή των εγγράφων σε διανύσματα μέσω του μηχανισμού bag-of-words έδωσε λύσεις σε πολλά προβλήματα κατηγοριοποίησης κειμένου ενώ αποτέλεσε το βασικό εργαλείο για NLP εφαρμογές (Han Kyul et al., 2017, 336). Πέρα όμως από βασικό εργαλείο της Μηχανικής Μάθησης στο τομέα NLP το bag-of-words χρησιμοποιήθηκε εφαρμογές υπολογιστικής όρασης για την ταξινόμηση εικόνων, όπου τα διάφορα χαρακτηριστικά της κάθε εικόνας αντιπροσωπεύονται με τον ίδιο τρόπο όπως οι λέξεις σε ένα κείμενο ((Wisam et al., 2019, 201).

Οι περιορισμοί χρήσης της συγκεκριμένης μεθόδου είναι:

- η αδυναμία να αποτυπώσει την σημασία της θέσης στην πρόταση της λέξης για την σημασιολογία της,

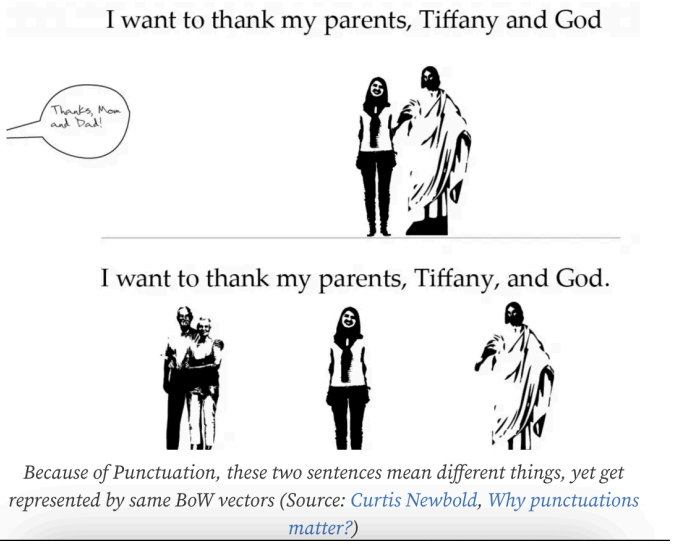
Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	**[1,1,1,1,2]**
D2	1	1	1	1	2	**[1,1,1,1,2]**

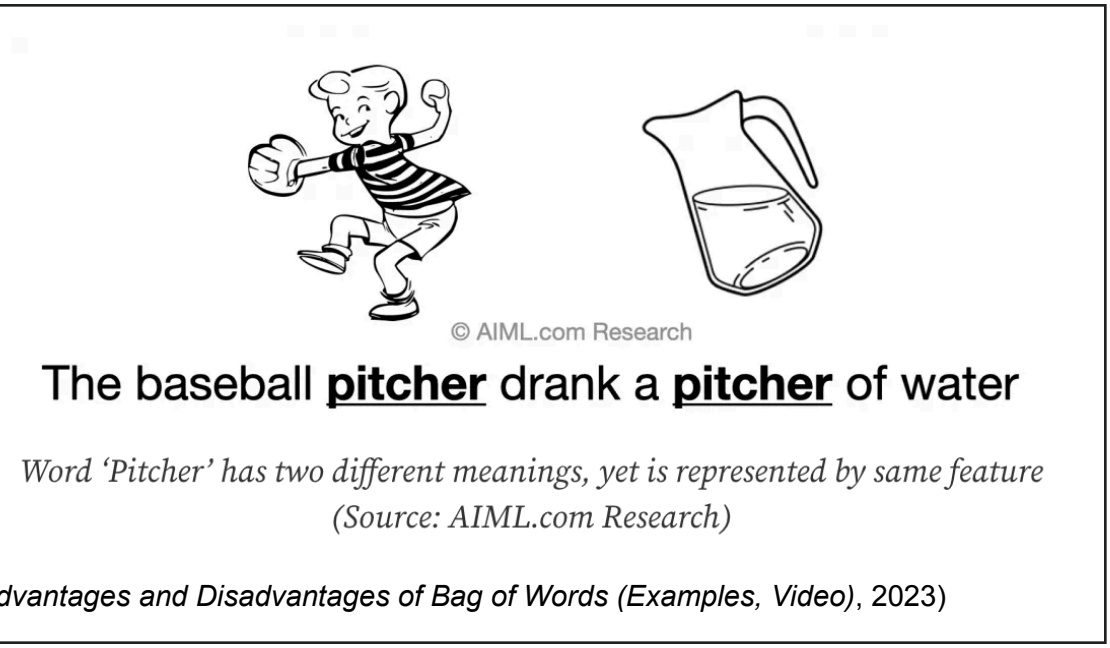
(*Advantages and Disadvantages of Bag of Words (Examples, Video)*, 2023)

- η αδυναμία να αποτυπώσει την σημασία των σημείων στίξης και της γραμματικής μέσα στο κείμενο



(*Advantages and Disadvantages of Bag of Words (Examples, Video)*, 2023)

- η αδυναμία να αποτυπώσει την πολυσημία μίας λέξης η οποία είναι κομμάτι της σημασιολογίας της



(*Advantages and Disadvantages of Bag of Words (Examples, Video)*, 2023)

- η δημιουργία μεγάλων πινάκων λόγων των πολλών διαστάσεων που αποδίδουν το σύνολο των μοναδικών λέξεων που εμπεριέχονται στο λεξικό των κειμένων
- οι πίνακες που δημιουργούνται είναι αραιοί με πολλές μηδενικές τιμές καθώς δεν περιέχεται το σύνολο του λεξικού σε ένα έγγραφο

Πέρα όμως από το από την συχνότητα εμφάνισης της λέξης χρησιμοποιούνται και άλλες τιμές όπως το TF_IDF σκορ. Η μέθοδος TF-IDF είναι μία μέθοδος με την οποία αποδίδεται σε μία λέξη η μέτρηση της σημαντικότητας της σε ένα έγγραφο ή σε μία κατηγορία. Το TF-IDF προκύπτει από την παραδοχή ότι όσο πιο συχνά εμφανίζεται μία λέξη σε μία κατηγορία ή σε ένα

έγγραφο, ενώ παράλληλα δεν εμφανίζεται στο υπόλοιπο πλήθος δεδομένων, τότε αυτή η λέξη κατέχει σημαντικό ρόλο στο έγγραφο ή στην κατηγορία στην οποία εμφανίζεται (Cai-zhi et al., 2018, 218). Ο αλγόριθμος TF-IDF προτάθηκε για πρώτη φορά το 1974 από τον Salton (Salton & Yu, 1974) και αποτελείται από δύο μέρη. Την συχνότητα της λέξης σε ένα κείμενο ή μία κλάση και την αντίστροφη συχνότητα αρχείου. Η της λέξης αναφέρεται στον αριθμό εμφανίσεων της λέξης σε ένα κείμενο ενώ η αντίστροφη συχνότητα κειμένου αντιπροσωπεύει την μέτρηση της γενικής σημασίας της λέξης (Cai-zhi et al., 2018, 218). Οι μαθηματικοί τύποι για την συχνότητα της λέξης (TF) και η αντίστροφη συχνότητα κειμένου (IDF) προκύπτουν από τους εξής μαθηματικούς τύπους:

- $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ όπου $n_{i,j}$ είναι ο αριθμός των εμφανίσεων της λέξης t_i στο έγγραφο d_j και

  $\sum_k n_{k,j}$ είναι το άθροισμα των εμφανίσεων όλων των λέξεων στο έγγραφο d_j.

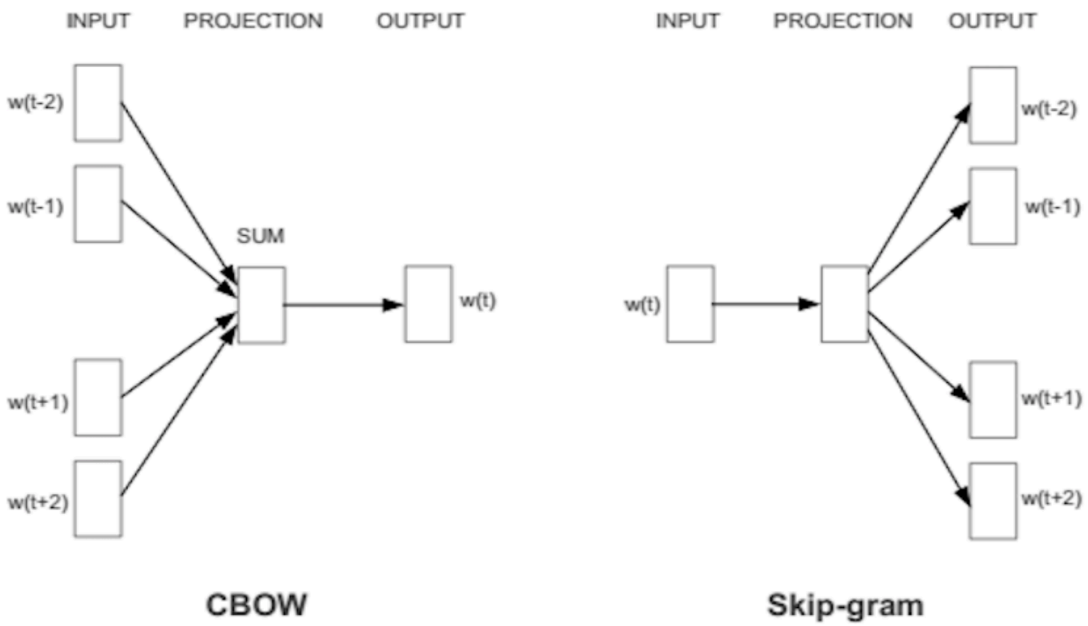
- $idf_i = log\frac{|D|}{\left|\{j: t_i \in d_j\}\right|}$ όπου $|D|$ είναι ο συνολικός αριθμός εγγράφω σε ένα σώμα κειμένων, $\left|\{j: t_i \in d_j\}\right|$ είναι ο αριθμός των εγγράφων που περιέχουν την λέξη t_i. Αν η λέξη δεν περιέχεται σε κανένα έγγραφο του σώματος κειμένων τότε ο διαιρέτης γίνεται μηδέν, οπότε σε αυτή την περίπτωση γίνεται χρήση του τύπου $1 + \left|\{j: t_i \in d_j\}\right|$

- $tfidf_{i,j} = \frac{tf_{i,j} \times idf_i}{\sqrt{\sum_{t_i \in d_j}\left[tf_{i,j} \times idf_i\right]^2}}$, όπου πλέον $tfidf_{i,j}$ είναι το βάρος της λέξης t_i. Όσο μεγαλύτερη είναι η συχνότητα της λέξη t_i σε ένα κείμενο ή κλάση και όσο μικρότερη είναι η συχνότητα της στο υπόλοιπο σώμα κειμένων ή κλάσεων τόσο μεγαλύτερο είναι το βάρος TF-IDF.

Word2Vec

Το word2vec είναι ένας αλγόριθμος για την δημιουργία των διανυσμάτων vectors τα οποία χρησιμοποιούνται στο NLP για την αντιπροσώπευση των λέξεων. Προτάθηκε από τον Mikolov το 2013 και εφαρμόστηκε αρχικά από την Google (Mikolov et al., 2013). Το word2vec είναι ουσιαστικά ένας αλγόριθμος νευρωνικού δικτύου το οποίο διακρίνεται σε δύο αρχιτεκτονικές: την CBOW και την SKIP-GRAM.

Με τις δύο αυτές αρχιτεκτονικές να είναι παρόμοιες η CBOW στοχεύει στην πρόβλεψη μίας λέξης στόχου δεδομένου του περιβάλλοντος της ενώ η SKIP-GRAM στοχεύει στην πρόβλεψη του περιβάλλοντος μίας λέξης δεδομένου ως εισόδου της λέξη στόχου.

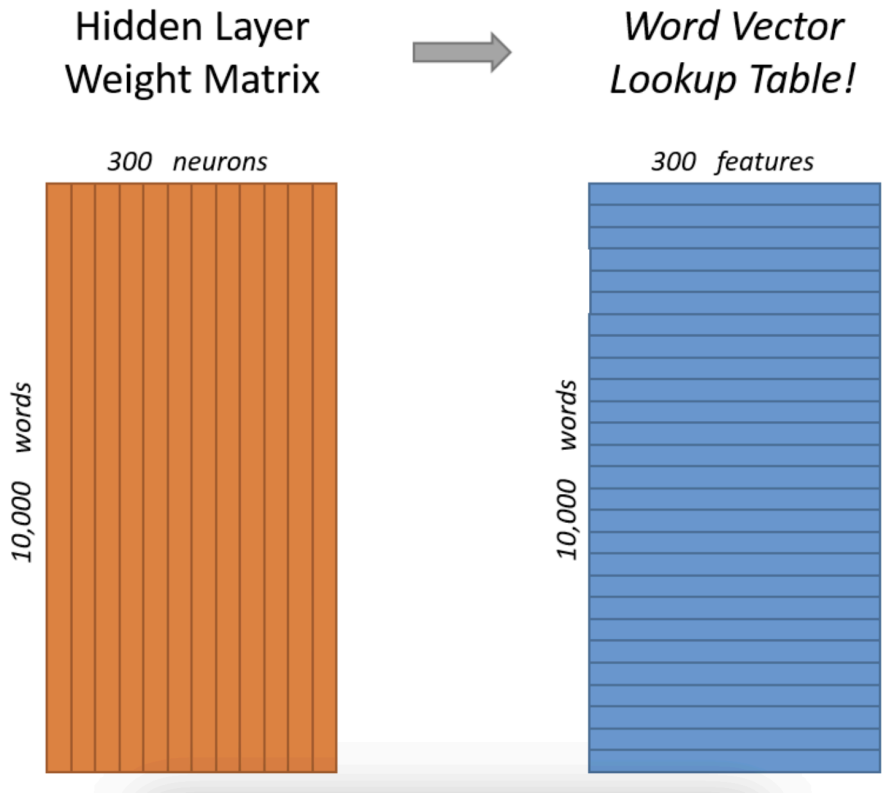


(Mikolov et al., 2013, 5)

Το word2vec είναι ένα απλό νευρωνικό δίκτυο το οποίο εκπαιδεύεται με σκοπό το output του να είναι είτε μία λέξη στόχος στην περίπτωση της αρχιτεκτονικής του CBOW είτε στην περίπτωση της αρχιτεκτονικής SKIP-GRAM από μία λέξη ως input στο σωστό περιβάλλον context της λέξης. Το word2vec όμως διαφέρει από τα κλασσικά νευρωνικά δίκτυα γιατί αυτό το οποίο χρησιμοποιεί ως vector της λέξης δεν είναι το output αλλά το διάνυσμα των βαρών του κρυφού επιπέδου του νευρωνικού δικτύου το οποίο βρίσκεται πριν το τελευταίο επίπεδο. Στο τελευταίο επίπεδο εφαρμόζεται η softmax συνάρτηση η οποία κανονικοποιεί τις τιμές του τελευταίου επιπέδου σε κλίμακα από 0 έως 1 ως σύνολο πιθανοτήτων και έτσι προκύπτει το τελικό output.

Πιο συγκεκριμένα όμως η βασική αλλαγή και καινοτομία του word2vec είναι ότι κατάφερε να μειώσει το πρόβλημα των πάρα πολλών αραιών vectors και να καταφέρει να αποδώσει τις σχέσεις μεταξύ των λέξεων. Οι μέχρι τότε vectors είχαν πάρα πολλές διαστάσεις, όσες το μέγεθος του λεξικού που χρησιμοποιείται για την εκπαίδευση του αλγορίθμου με κατά κύριο λόγο μηδενικές τιμές. Ένα απλό παράδειγμα για να την καλύτερη κατανόηση αυτής της μετάβασης στις λιγότερες διαστάσεις είναι μία υπόθεση εφαρμογής ενός ρηχού νευρωνικού δικτύου με 300 νευρώνες στο κρυμμένο επίπεδο, σε ένα λεξικό 10.000 χιλιάδων λέξεων. Η

είσοδος σε αυτό το δίκτυο είναι μία λέξη η οποία αντιπροσωπεύεται ως one-hot encoded vector με 10.000 διαστάσεις όσες δηλαδή είναι το πλήθος του λεξικού. Με την εκπαίδευση του δικτύου προσαρμόζονται κατάλληλα τα βάρη των νευρώνων στο κρυμμένο δίκτυο μέσω της τεχνικής του back propagation όπου το λάθος της πρόβλεψης από το επίπεδο της εξόδου μεταφέρεται προς τα πίσω για την αναπροσαρμογή των βαρών των νευρώνων. Με το πέρας της εκπαίδευσης τα προσαρμοσμένα βάρη του κρυμμένου επιπέδου εξάγονται ως ένας vector αντιπροσώπευσης της λέξης με 300 διαστάσεις όσο είναι δηλαδή το πλήθος των νευρώνων του κρυμμένου επίπεδου. Έτσι από ένα διάνυσμα 10.000 διαστάσεων το οποίο αποτελείται κυρίως από μηδενικά (one-hot encoded vector) καταλήγουμε σε ένα διάνυσμα αντιπροσώπευσης της λέξης 300 διαστάσεων το οποίο δεν περιέχει μηδενικές τιμές (Xin, 2016, 2).



(*Word2Vec Explained*, 2017)

Η αρχιτεκτονική CBOW όπως αναφέρθηκε παραπάνω περιλαμβάνει ως είσοδο το τις λέξεις που περιβάλλουν την λέξη στόχο και ως εκ τούτου η είσοδος σε σχέση με το προηγούμενο παράδειγμα διαφέρει. Αντίστοιχα η αρχιτεκτονική του SKIP-GRAM περιλαμβάνει ως είσοδο την λέξη στόχο με σκοπό την απόδοση ως αποτέλεσμα του σωστού περιβάλλοντος της.

Η επιτυχία του word2vec να αποτυπώσει τις σχέσεις μεταξύ των λέξεων στα παραγόμενα διανύσματα έγκειται στην παραδοχή ότι λέξεις οι οποίες μπορούν να εμφανιστούν σε παρόμοια περιβάλλοντα βρίσκονται επίσης κοντά σημασιολογικά (Goldberg & Levy, 2014, 5). Ο Mikolov

στο άρθρο του για το Word2Vec βρήκε ότι οι σχέσεις μεταξύ των καινούργιων διανυσμάτων που αντιπροσωπεύουν τις λέξεις, αποτυπώνονταν και στις σημειολογικές σχέσεις των λέξεων. Έτσι η μαθηματική πράξη *X = vector("biggest") - vector("big") + vector("small")* αποδίδει ένα διάνυσμα το οποίο βρίσκεται κοντά στο διάνυσμα της λέξης *"smallest"*. Η μέτρηση της απόστασης μεταξύ των διανυσμάτων γίνεται μέσω της εφαρμογής της μεθόδου της απόστασης της εφαπτομένης (cosine distance) (Mikolov et al., 2013, 6). Ο Mikolov αξιολογώντας την απόδοση των σχέσεων μεταξύ των λέξεων κατέληξε στον παρακάτω ενδιαφέρον πίνακα.

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

(Mikolov et al., 2013, 6)

Στα πλεονεκτήματα και μειονεκτήματα των δύο αρχιτεκτονικών συγκρίνοντας τες μεταξύ τους η αρχιτεκτονική SKIP-GRAM αποδίδει καλύτερα σε μικρό μέγεθος δεδομένων εκπαίδευσης και προσφέρει καλύτερες αναπαραστάσεις των σπάνιων λέξεων. Από την άλλη μεριά η αρχιτεκτονική CBOW είναι πιο γρήγορη στην εκπαίδευση για την αναπαράσταση των λέξεων και έχει ελαφρώς καλύτερη απόδοση στις λέξει με μεγάλη συχνότητα (Kulshrestha, 2019).

GloVe

Το GloVe είναι ακρωνύμιο προερχόμενο από τον τίτλο "Global Vectors". Παρουσιάστηκε το 2014 από το πανεπιστήμιο του Stanford (Pennington et al., 2014) και αποτελεί μία στατιστική

προσέγγιση για την δημιουργία των embeddings των λέξεων όπου οι σχέσεις μεταξύ των λέξεων στα embeddings προκύπτουν με πιο ξεκάθαρο τρόπο. Ο αλγόριθμος GloVe στηρίζεται στην σημασία που έχει η συνεμφάνιση λέξεων μέσα στο ίδιο περικείμενο - περιβάλλον.

Το πρώτο βήμα για την εφαρμογή του GloVe αλγορίθμου είναι η κατασκευή ενός δισδιάστατου πίνακα του οποίου κάθε γραμμή και στήλη αντιπροσωπεύει μία λέξη. Η τιμή του κελιού είναι ουσιαστικά ο αριθμός συνεμφανίσεων στο ίδιο περικείμενο. Έτσι έχουμε ένα πίνακα X όπου η καταχώρηση X_{ji} αντιπροσωπεύει το αριθμό εμφανίσεων της λέξης *j* στο ίδιο περιβάλλον με την λέξη *i* (Pennington et al., 2014, 2). Έτσι η γραμμή $X_i = \sum_k X_{ik}$ αντιπροσωπεύει το σύνολο των λέξεων που εμφανίζονται στο περικείμενο της λέξης *i*. Αντίστοιχα η πιθανότητα $P_{ij} = P(i|j) = \frac{X_{ij}}{X_i}$ εκφράζει την πιθανότητα η λέξη *j* να εμφανιστεί στο περικείμενο της λέξης *i*. Έτσι ο πίνακας X ο οποίος απεικονίζει τις συνεμφανίσεις μετατρέπεται σε πίνακα πιθανοτήτων ο οποίος μπορεί μέσα από μαθηματικές πράξεις να δώσει πληροφορίες για τις ίδιες τις λέξεις, όπως το ακόλουθο παράδειγμα που παρουσιάζεται στην εργασία του (Pennington et al., 2014, 1534)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Very small or large:
solid is related to ice but not steam, or
gas is related to steam but not ice

close to 1:
water is highly related to ice and steam, or
fashion is not related to ice or steam.

(Hui, 2019)

Από τον παραπάνω πίνακα του παραδείγματος μπορούμε να βγάλουμε συμπεράσματα από την διαίρεση των πιθανοτήτων. Αν το αποτέλεσμα της διαίρεσης των πιθανοτήτων είναι πολύ μεγάλο μπορούμε να συμπεράνουμε ότι η λέξη της πιθανότητας που βρίσκεται στον αριθμητή έχει μεγαλύτερη σχέση με την εξεταζόμενη λέξη (η λέξη *ice* με την εξεταζόμενη λέξη *solid*) ενώ αν το αποτέλεσμα είναι μικρό τότε αντίστροφα η λέξη που βρίσκεται στον παρονομαστή είναι αυτή που σχετίζεται περισσότερο (η λέξη *steam* με την εξεταζόμενη λέξη *gas*). Αν το αποτέλεσμα της διαίρεσης είναι είναι κοντά στο 1 τότε οι δύο λέξεις είτε σχετίζονται πολύ με την

εξεταζόμενη λέξη (οι λέξεις *ice* και *steam* με την εξεταζόμενη λέξη *water*) είτε είναι άσχετες με την λέξη αυτή (οι λέξεις *ice* και *steam* με την εξεταζόμενη λέξη *fashion*) .

Όπως μπορεί να γίνει κατανοητό από τον παραπάνω τύπο, η συσχέτιση λέξεων στο GloVe εξάγεται από την αναλογία των πιθανοτήτων αυτών των λέξεων να εμφανίζονται στατιστικά μαζί στο υπό εξέταση σώμα κειμένων. Στην συνέχεια μέσω παραγοντοποίησης του πίνακα επιτυγχάνεται η μείωση των διαστάσεων του πίνακα με σκοπό κάθε γραμμή του πίνακα να αποτελεί το διάνυσμα μίας λέξης με τις τιμές των χαρακτηριστικών του. Ο GloVe αλγόριθμος χρησιμοποιεί την συνάρτηση των ελάχιστων τετραγώνων με σκοπό την ελαχιστοποίηση της διαφοράς του εσωτερικού γινομένου των διανυσμάτων των δύο λέξεων και του λογάριθμου του αριθμού συνεμφάνισης τους (*GloVe Explained*, n.d.).

Όπως και στην περίπτωση του Word2Vec έτσι και το GloVe μπορεί να εξετάσει τις σχέσεις μεταξύ των λέξεων μέσω της εξετασης της ομοιότητας της εφαπτομένης των διανυσμάτων των λέξεων. Μάλιστα στο άρθρο (Pennington et al., 2014, 1537 - 1539) παρουσιάζονται βελτιωμένες επιδόσεις του GloVe διανυσμάτων σε σχέση με το Word2Vec.

Το βασικό μειονέκτημα του GloVe αλγορίθμου οπως και με τον αλγόριθμο Word2Vec είναι η αδυναμία του να διακρίνει την πολυμορφία των λέξεων όπου λέξεις με την ίδια μορφή έχουν διαφορετική σημασιολογία (Uzila, 2022).

Fasttext

Το fastext είναι μία μέθοδος η οποία βασίστηκε στο Word2Vec για την κατασκευή embeddings και παρουσιάστηκε το 2016 από (Bojanowski et al., 2016). Η διαφοροποίηση του FastText συγκριτικά με το Word2Vec έγκειται στο ότι το FastText αναζητάει πέρα από τις δομές σε επίπεδο λέξης, τις δομές της γλώσσας εντός των ορίων της λέξης. Έτσι εντοπίζονται προθήματα, επιθήματα τα οποία έχουν σημασιολογικό ρόλο σε μία γλώσσα. Αυτή η τεχνική αποφέρει πολύ καλά αποτελέσματα σε διαχυτικές γλώσσες, όπου εμφανίζεται πλούσια μορφολογία. Επιπλέον σε περίπτωση όπου συναντάται λέξη η οποία δεν περιλαμβάνεται στο λεξικό που χρησιμοποιήθηκε για την δημιουργία των embeddings, υπάρχει η δυνατότητα η λέξη αυτή να διαιρεθεί σε μορφήματα τα οποία είναι γνωστά και να μπορέσει να εξαχθεί συμπέρασμα με αυτόν τον τρόπο και για άγνωστες λέξεις (Bojanowski et al., 2016, 1).

Ο τρόπος με τον οποίο το επιτυγχάνει αυτό ο αλγόριθμος FASTTEXT είναι μέσω της διαίρεσης της λέξης σε n-grams όπου n είναι αριθμός των χαρακτήρων που επιλέγουμε για να πραγματοποιήσουμε τον αλγόριθμο. Έτσι δημιουργούνται διανύσματα για κάθε υποσύνολο n-gram το οποίο περιέχει η λέξη. Εν συνεχεία το διάνυσμα της λέξης αποτελείται από το άθροισμα των διανυσμάτων των n-gram. Επιπλέον λέξεις και υποσύνολα λέξεων τα οποία έχουν την ίδια μορφή έχουν διαφορετικά διανύσματα. Αυτός ο τρόπος επιτρέπει σε διαφορετικές λέξεις

να μοιράζονται παρόμοιες αντιπροσωπεύσεις καθώς επίσης και μία άγνωστη λέξη να αναλύεται σε γνωστά n-grams και να επιτυγχάνεται η αντιπροσώπευση της σε διάνυσμα (Bojanowski et al., 2016, 3).

Διανύσματα με βάση το περικείμενο (Contextuals Embeddings)

Τα διανύσματα με βάση το περικείμενο (Contextuals Embeddings) αντιπροσωπεύουν μια σημαντική πρόοδο στην επεξεργασία της φυσικής γλώσσας αποτυπώνοντας το νόημα των λέξεων ή των φράσεων δυναμικά με βάση το περικείμενο - πλαίσιο τους σε μια πρόταση ή μεγαλύτερο κείμενο. Σε αντίθεση με τις παραδοσιακές στατικές ενσωματώσεις λέξεων, όπως το Word2Vec ή το GloVe, που εκχωρούν ένα μόνο σταθερό διάνυσμα σε κάθε λέξη ανεξάρτητα από τη χρήση της, τα embeddings με βάση τα συμφραζόμενα είναι ευαίσθητες στο πλαίσιο, επιτρέποντας στην ίδια λέξη να έχει διαφορετικές αναπαραστάσεις ανάλογα με τις γύρω λέξεις. Για παράδειγμα, η λέξη "τόνος" στις φράσεις "ψάρεψε ένα τόνο" και "ένα τόνο βαμβάκι" θα έχει διαφορετικά embeddings που αντανακλούν τη μοναδική της σημασία σε κάθε πλαίσιο. Αυτές οι ενσωματώσεις δημιουργούνται από προεκπαιδευμένα γλωσσικά μοντέλα όπως το BERT, το RoBERTa και το GPT, τα οποία αξιοποιούν βαθιές νευρικές αρχιτεκτονικές και μηχανισμούς προσοχής. Αυτά τα μοντέλα επεξεργάζονται ολόκληρες προτάσεις αμφίδρομα, λαμβάνοντας υπόψη τόσο τις προηγούμενες όσο και τις επόμενες λέξεις για να οικοδομήσουν μια λεπτή κατανόηση των σημασιολογικών σχέσεων. Αυτή η αμφίδρομη ανάλυση διασφαλίζει ότι καταγράφονται λεπτές παραλλαγές στη χρήση λέξεων, τον τόνο και τη γραμματική δομή. Οι ενσωματώσεις με βάση τα συμφραζόμενα έχουν αποδειχθεί ιδιαίτερα ωφέλιμες για εργασίες που απαιτούν βαθιά κατανόηση του κειμένου, συμπεριλαμβανομένης της ανάλυσης συναισθημάτων, της αναγνώρισης ονομαστικών οντοτήτων και της ταξινόμησης εγγράφων. Επιπλέον, η ικανότητά τους να προσαρμόζονται σε κείμενα που αφορούν συγκεκριμένους τομείς, όπως νομικά, ιατρικά ή τεχνικά σώματα, τα καθιστά ιδιαίτερα ευέλικτα. Κωδικοποιώντας μια πιο ολοκληρωμένη και με βάση τα συμφραζόμενα αναπαράσταση της γλώσσας, οι ενσωματώσεις με βάση τα συμφραζόμενα βελτιώνουν σημαντικά την απόδοση των μοντέλων μηχανικής εκμάθησης σε πολύπλοκες εργασίες επεξεργασίας φυσικής γλώσσας (Peters et al., 2018).

Αρχιτεκτονικές NLP

Παραδοσιακές Μέθοδοι NLP

Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται στο NLP εδώ και πολλά χρόνια και συνεχίζουν να διαδραματίζουν σημαντικό ρόλο στο πεδίο. Αυτοί οι αλγόριθμοι είναι

συνήθως απλούστεροι και λιγότερο υπολογιστικά δαπανηροί από τους αλγόριθμους βαθιάς μάθησης και μπορούν να είναι αποτελεσματικοί για μια ποικιλία εργασιών NLP.

Οι κοινοί παραδοσιακοί αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται στο NLP περιλαμβάνουν:

Support Vector Machine (SVM): Τα SVM είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης και παλινδρόμησης. Τα SVM λειτουργούν βρίσκοντας ένα υπερεπίπεδο που διαχωρίζει τα σημεία δεδομένων σε δύο κατηγορίες με τέτοιο τρόπο ώστε το περιθώριο μεταξύ του υπερεπίπεδου και των σημείων δεδομένων να μεγιστοποιείται.

Ταξινομητές Naïve Bayes: Οι ταξινομητές Naïve Bayes είναι ένας τύπος αλγορίθμου μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης. Οι ταξινομητές Naïve Bayes λειτουργούν υποθέτοντας ότι τα χαρακτηριστικά των δεδομένων είναι ανεξάρτητα μεταξύ τους και χρησιμοποιώντας το θεώρημα του Bayes για τον υπολογισμό της πιθανότητας ενός σημείου δεδομένων να ανήκει σε μια συγκεκριμένη κλάση.

Δέντρα αποφάσεων (Decision Trees): Τα δέντρα αποφάσεων είναι ένας τύπος αλγορίθμου μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης και παλινδρόμησης. Τα δέντρα αποφάσεων λειτουργούν κατασκευάζοντας μια δομή που μοιάζει με δέντρο που αντιπροσωπεύει τις διαφορετικές διαδρομές που μπορεί να ακολουθήσει ο αλγόριθμος για να κάνει μια πρόβλεψη.

Λογιστική παλινδρόμηση (Logistic Regression): Η λογιστική παλινδρόμηση είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης. Η λογιστική παλινδρόμηση λειτουργεί προσαρμόζοντας μια λογιστική συνάρτηση στα δεδομένα και στη συνέχεια χρησιμοποιώντας τη συνάρτηση για την πρόβλεψη της πιθανότητας ενός σημείου δεδομένων που ανήκει σε μια συγκεκριμένη κλάση.

Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για μια ποικιλία εργασιών NLP, όπως:

Ταξινόμηση κειμένου: Οι εργασίες ταξινόμησης κειμένου περιλαμβάνουν την πρόβλεψη της κατηγορίας ενός εγγράφου κειμένου, όπως εάν είναι ανεπιθύμητο ή όχι ανεπιθύμητο. Οι παραδοσιακοί αλγόριθμοι μηχανικής εκμάθησης μπορούν να χρησιμοποιηθούν για εργασίες ταξινόμησης κειμένου εκπαιδεύοντας ένα μοντέλο σε ένα επισημασμένο σύνολο δεδομένων εγγράφων κειμένου.

Ανάλυση συναισθήματος: Οι εργασίες ανάλυσης συναισθήματος περιλαμβάνουν τον προσδιορισμό του συναισθήματος ενός κειμένου, όπως εάν είναι θετικό, αρνητικό ή ουδέτερο. Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για εργασίες

ανάλυσης συναισθήματος εκπαιδεύοντας ένα μοντέλο σε ένα επισημασμένο σύνολο δεδομένων εγγράφων κειμένου με τις αντίστοιχες ετικέτες συναισθήματος.

Αναγνώριση ονοματικής οντότητας (NER): Οι εργασίες NER περιλαμβάνουν τον προσδιορισμό οντοτήτων σε ένα κείμενο, όπως άτομα, οργανισμούς και τοποθεσίες. Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για εργασίες NER εκπαιδεύοντας ένα μοντέλο σε ένα επισημασμένο σύνολο δεδομένων εγγράφων κειμένου με τις αντίστοιχες επώνυμες ετικέτες οντοτήτων τους.

Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης είναι ένα ισχυρό εργαλείο για εργασίες NLP. Είναι απλοί, αποδοτικοί και αποτελεσματικοί για μια ποικιλία εργασιών. Ωστόσο, οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης έχουν ορισμένους περιορισμούς όπως η περιορισμένη εκφραστικότητα διότι δεν είναι σε θέση να μάθουν πολύπλοκα μοτίβα σε δεδομένα τόσο αποτελεσματικά όσο οι αλγόριθμοι βαθιάς μάθησης και η απαίτηση χαρακτηριστικών τα οποία κατασκευάζονται χειροκίνητα, το οποίο μια χρονοβόρα και εντατική διαδικασία.

Απλά Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για την επίλυση ποικίλων προβλημάτων, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας (NLP). Τα απλά νευρωνικά δίκτυα είναι μια καλή επιλογή για εργασίες NLP όπου η απλότητα, η αποτελεσματικότητα και η ευρωστία είναι σημαντικά.

Το νευρωνικό δίκτυο είναι ένα μαθηματικό μοντέλο που εμπνέεται από τον ανθρώπινο εγκέφαλο. Ένα απλό νευρωνικό δίκτυο στη μηχανική μάθηση λειτουργεί λαμβάνοντας εισόδους, περνώντας τις μέσω μιας σειράς διασυνδεδεμένων κόμβων και παράγοντας εξόδους. Κάθε κόμβος στο δίκτυο εκτελεί μια απλή μαθηματική πράξη, όπως τον πολλαπλασιασμό των εισόδων με βάρη και την πρόσθεσή τους. Εκεί προστίθεται και μία επιπλέον τιμή η οποία αναφέρεται ως bias. Στη συνέχεια, οι έξοδοι ενός στρώματος κόμβων περνούν στο επόμενο στρώμα κόμβων και ούτω καθεξής περνώντας μέσα από συναρτήσεις ενεργοποίησης. Οι συναρτήσεις ενεργοποίησης χρησιμοποιούνται για την εισαγωγή μη γραμμικότητας στο δίκτυο. Αυτό είναι σημαντικό γιατί επιτρέπει στο δίκτυο να μάθει πιο σύνθετες σχέσεις μεταξύ των εισόδων και των εξόδων. Ορισμένες κοινές λειτουργίες ενεργοποίησης περιλαμβάνουν το sigmoid, το tanh και το ReLU.

Τα βάρη των συνδέσεων μεταξύ των κόμβων προσαρμόζονται κατά τη διάρκεια της εκπαιδευτικής διαδικασίας. Αυτό γίνεται συγκρίνοντας τις εξόδους του δικτύου με τις επιθυμητές εξόδους για ένα δεδομένο σύνολο εισόδων. Τα βάρη ρυθμίζονται με τέτοιο τρόπο ώστε οι έξοδοι του δικτύου να γίνονται όλο και πιο ακριβείς με την πάροδο του χρόνου. Αυτό γίνεται συγκρίνοντας τις εξόδους του δικτύου με τις επιθυμητές εξόδους για ένα δεδομένο σύνολο

εισόδων μέσω της διαδικασία backpropagation. Το backpropagation λειτουργεί με τη διάδοση του σφάλματος από το επίπεδο εξόδου του δικτύου πίσω στο επίπεδο εισόδου. Αυτό επιτρέπει στο δίκτυο να προσαρμόζει τα βάρη του με τέτοιο τρόπο ώστε το σφάλμα να μειώνεται.

Βαθιά Νευρωνικά Δίκτυα

Τα βαθιά νευρωνικά δίκτυα (DNN) έχουν φέρει επανάσταση στον τομέα του NLP τα τελευταία χρόνια. Τα DNN έχουν επιτύχει αποτελέσματα αιχμής σε ένα ευρύ φάσμα εργασιών NLP, συμπεριλαμβανομένης της αυτόματης μετάφρασης, της σύνοψης κειμένου, της απάντησης σε ερωτήσεις και της ανάλυσης συναισθημάτων. Τα DNN είναι σε θέση να μάθουν περίπλοκες σχέσεις μεταξύ των λέξεων μιας πρότασης και της σημασίας της πρότασης στο σύνολό της. Αυτό οφείλεται στο γεγονός ότι τα DNN έχουν μεγάλο αριθμό παραμέτρων που μπορούν να εκπαιδευτούν για την καταγραφή αυτών των σχέσεων.

Τα DNN συνήθως εκπαιδεύονται χρησιμοποιώντας μια προσέγγιση εποπτευόμενης μάθησης. Αυτό σημαίνει ότι δίνεται στο δίκτυο ένα σύνολο δεδομένων εκπαίδευσης που αποτελείται από εισόδους και επιθυμητές εξόδους. Στη συνέχεια, το δίκτυο εκπαιδεύεται ώστε να ελαχιστοποιεί το σφάλμα μεταξύ των προβλέψεών του και των επιθυμητών εξόδων. Μόλις εκπαιδευτεί ένα DNN, μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις για νέα δεδομένα.

Ένα από τα κύρια πλεονεκτήματα των DNN είναι ότι μπορούν να μάθουν από δεδομένα χωρίς να πρέπει να χαρακτηριστούν με ανθρώπινη παρέμβαση. Αυτό σημαίνει ότι τα DNN μπορούν να χρησιμοποιηθούν για την επίλυση μιας μεγάλης ποικιλίας εργασιών NLP, ακόμα κι αν δεν υπάρχουν διαθέσιμες λειτουργίες οι οποίες να έχουν προκαθοριστεί ρητά. Ένα άλλο πλεονέκτημα των DNN είναι ότι μπορούν να είναι πολύ ακριβή, ειδικά όταν εκπαιδεύονται σε μεγάλα σύνολα δεδομένων.

Ωστόσο ένα μειονέκτημα των DNN είναι η μεγάλη υπολογιστική ισχύ που απαιτούν για την εκπαίδευση τους. Επιπλέον τα DNN μπορεί να είναι δύσκολο να ερμηνευτούν, πράγμα το οποίο σημαίνει ότι είναι δύσκολο να μελετηθούν οι μηχανισμοί και οι παράγοντες που οδήγησαν σε μία πρόβλεψη.

Ένας τύπος DNN που είναι ιδιαίτερα κατάλληλος για εργασίες NLP είναι το επαναλαμβανόμενο νευρωνικό δίκτυο (RNN). Τα RNN είναι σε θέση να επεξεργάζονται διαδοχικά δεδομένα, όπως το ελευθερο κείμενο, διατηρώντας μια κατάσταση που ενημερώνεται καθώς το δίκτυο επεξεργάζεται τα δεδομένα. Αυτό επιτρέπει στα RNN να μαθαίνουν δεδομένα σε σειρά στα οποία υπάρχουν αλληλεξαρτήσεις μεταξύ τους. Δύο δημοφιλείς τύποι RNN είναι το δίκτυο Long Short-Term Memory (LSTM) και το δίκτυο Gated Recurrent Unit (GRU). Τα δίκτυα LSTM και GRU είναι σε θέση να μάθουν μακροπρόθεσμες εξαρτήσεις σε δεδομένα πιο αποτελεσματικά από τα παραδοσιακά RNN καθώς μπορούν και να αναπαριστούν την εξάρτηση

των δεδομένων στο κοντινό πλαίσιο που τα περιβάλλει ενώ η εξάρτηση αυτή αποδυναμώνεται με την απομάκρυνση από την θέση που βρίσκεται το δεδομένο αυτό.

Αρχιτεκτονική Μετασχηματιστή - Transformer

Τα παραδοσιακά μοντέλα NLP, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα συνελικτικά νευρωνικά δίκτυα (CNN), είχαν περιορισμούς στην επεξεργασία διαδοχικών αλληλοεξαρτώμενων δεδομένων όπως η γλώσσα. Τα RNN υστερούν κάθως με την πάροδο της ακολουθίας των δεδομένων οι διαβαθμίσεις που παράγονται είτε εξαφανίζονται είτε εκρήγνυνται, γεγονός που τα καθιστά αναποτελεσματικά για εξαρτήσεις μεγάλης εμβέλειας. Τα CNN, ενώ ήταν καλύτερα στην παράλληλη επεξεργασία, αλλά δεν είχαν την ικανότητα να συλλαμβάνουν αποτελεσματικά διαδοχικές πληροφορίες.

Η αρχιτεκτονική του μετασχηματιστή είναι ένας τύπος βαθιού νευρωνικού δικτύου που έχει φέρει επανάσταση στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP). Οι μετασχηματιστές είναι σε θέση να επιτύχουν αποτελέσματα αιχμής σε ένα ευρύ φάσμα εργασιών NLP. Η transformer αρχιτεκτονική εισήχθει για πρώτη φορά στην εργασία "Attention Is All You Need" των Ashish Vaswani et al. το 2017.

Η βασική καινοτομία της transformer αρχιτεκτονικής του μετασχηματιστή είναι ο μηχανισμός αυτοπροσοχής (self attention), ο οποίος επιτρέπει στο μοντέλο να σταθμίζει τη σημασία διαφορετικών λέξεων σε μια πρόταση κατά την επεξεργασία κάθε λέξης. Σε αντίθεση με τα RNN και τα CNN, τα οποία επεξεργάζονται λέξεις με διαδοχικό ή σταθερό τρόπο, η αυτοπροσοχή λειτουργεί παράλληλα, καθιστώντας την εξαιρετικά αποτελεσματική. Για κάθε λέξη σε μια πρόταση, ο μηχανισμός αυτοπροσοχής δημιουργεί τρία διανύσματα: key, query και value. Αυτά τα διανύσματα μαθαίνονται κατά τη διάρκεια της εκπαιδευτικής διαδικασίας όπου key αντιπροσωπεύει τη σημασία της λέξης σε σχέση με άλλες λέξεις της σειράς, το query αντιπροσωπεύει τη λέξη που βαθμολογείται ως προς τη σημασία και το value διατηρεί τις πραγματικές πληροφορίες που σχετίζονται με τη λέξη.

Για τον υπολογισμό της σημασία μιας λέξης σε μια πρόταση, υπολογίζεται μια βαθμολογία παίρνοντας το εσωτερικό γινόμενο του διανύσματος query της τρέχουσας λέξης και το διάνυσμα key κάθε άλλης λέξης της πρότασης. Στη συνέχεια, αυτές οι βαθμολογίες ταξινομούνται συνάρτησης softmax για να ληφθεί μια κατανομή βάρους σε όλες τις λέξεις της πρότασης. Το σταθμισμένο άθροισμα των διανυσμάτων value δίνει την έξοδο για την τρέχουσα λέξη. Αυτή η διαδικασία επαναλαμβάνεται για κάθε λέξη της σειράς, επιτρέποντας στο μοντέλο να μάθει και να εφαρμόσει διαφορετικά μοτίβα προσοχής ανάλογα με το πλαίσιο.

Ένα από τα δυνατά σημεία της αρχιτεκτονικής του Transformer είναι η ικανότητά του να εστιάζει σε διαφορετικές πτυχές της ακολουθίας εισόδου ταυτόχρονα. Αυτό επιτυγχάνεται μέσω

του Multi-Head Attention. Στο Multi-Head Attention, το μοντέλο δημιουργεί πολλαπλά σύνολα διανυσμάτων query, key και value και υπολογίζει τις βαθμολογίες παράλληλα. Κάθε κεφαλή μπορεί να μάθει διαφορετικές πτυχές της εισόδου, επιτρέποντας στο μοντέλο να συλλάβει ένα ευρύ φάσμα πληροφοριών.

Εφόσον το αποτέλεσμα των transformer δεν επηρεάζεται από την μετάθεση των δεδομένων στην περίπτωση του NLP, δεν κωδικοποιούν από την φύση τους τις πληροφορίες θέσης των λέξεων στην ακολουθία εισόδου. Για να αντιμετωπιστεί αυτό, οι μετασχηματιστές χρησιμοποιούν κωδικοποίηση θέσης με τα embeddings των λέξεων για να προσθέσουν πληροφορίες θέσης στις αναπαραστάσεις εισόδου. Αυτές οι κωδικοποιήσεις θέσης παρέχουν στο μοντέλο πληροφορίες σχετικά με τη θέση της λέξης στην ακολουθία, επιτρέποντάς του να κατανοήσει την διαδοχικότητα των δεδομένων.

Μετά τα επίπεδα self attention, ο transformer χρησιμοποιεί feed-forward νευρωνικά δίκτυα για περαιτέρω επεξεργασία. Αυτά σύμφωνα με την αρχιτεκτονική που περιγράφηκε παραπάνω βοηθούν το μοντέλο να συλλάβει πολύπλοκες σχέσεις μέσα στα δεδομένα. Έτσι ολοκληρώνεται η δομή ενός μοντέλου NLP στο οποίο έχει εφαρμοστεί η αρχιτεκτονική transformer. Αυτή η δομή μπορεί να επαναληφθεί μέσα σε ένα μοντέλο πράγμα με το οποίο δημιουργεί στοιβάδες transformer (stacking layers). Πολλαπλά επίπεδα δικτύων self-attention και feed-forward νευρωνικών δικτύων που στοιβάζονται το ένα πάνω στο άλλο. Αυτό επιτρέπει στο μοντέλο να μάθει ιεραρχικές αναπαραστάσεις των δεδομένων εισόδου, καταγράφοντας χαρακτηριστικά χαμηλού και υψηλού επιπέδου.

Η αρχιτεκτονική του μετασχηματιστή είναι ένα ισχυρό εργαλείο για το NLP. Έχει επιτύχει αποτελέσματα αιχμής σε ένα ευρύ φάσμα εργασιών NLP και χρησιμοποιείται για την ανάπτυξη νέων και καινοτόμων εφαρμογών NLP.

Transformer Μοντέλα

BERT

Ένα μοντέλο που έχει δείξει πολλά υποσχόμενα αποτελέσματα στην επεξεργασία φυσικής γλώσσας είναι το μοντέλο BERT (Bidirectional Encoder Representations from Transformers). Είναι μια τεχνική επεξεργασίας φυσικής γλώσσας τελευταίας τεχνολογίας που έχει φέρει επανάσταση σε διάφορες εργασίες που σχετίζονται με τη γλώσσα. Αναπτύχθηκε από την Google και έχει επιτύχει αξιοσημείωτη επιτυχία σε εργασίες όπως η ταξινόμηση κειμένου, η αναγνώριση οντοτήτων (named entity recognition - NER), η ανάλυση συναισθημάτων (sentiment analysis) και η απάντηση ερωτήσεων.

Η αρχιτεκτονική που κρύβεται κάτω από το μοντέλο BERT βασίζεται στην transformer αρχιτεκτονική, η οποία αφορά μοντέλα σχεδιασμένα για να καταγράφουν τις συμφραζόμενες σχέσεις μεταξύ των λέξεων σε μια πρόταση. Σε αντίθεση με τα παραδοσιακά μοντέλα που επεξεργάζονται κείμενο με διαδοχικό τρόπο, οι μετασχηματιστές επιτρέπουν την παράλληλη επεξεργασία λέξεων, επιτρέποντας μια πιο ολοκληρωμένη κατανόηση του περιβάλλοντος.

Ένα από τα βασικά χαρακτηριστικά του BERT είναι η αμφίδρομη φύση του. Σε αντίθεση με τα κλασικά μοντέλα που λαμβάνουν υπόψη μόνο το αριστερό ή το δεξί πλαίσιο μιας λέξης, το BERT εξετάζει και τις δύο κατευθύνσεις ταυτόχρονα. Αυτή η αμφίδρομη εκπαίδευση επιτρέπει στο μοντέλο να καταγράφει τις εξαρτήσεις μεταξύ των λέξεων πιο αποτελεσματικά και να δημιουργεί πιο ακριβείς αναπαραστάσεις.

Το BERT εκπαιδεύεται χρησιμοποιώντας μία μοντελοποίηση γλώσσας με μάσκα (masked language model - MLM), όπου τυχαίες λέξεις σε μια πρόταση καλύπτονται και το μοντέλο έχει την αποστολή να προβλέψει τις καλυμμένες λέξεις με βάση το πλαίσιο. Επιπλέον, το BERT χρησιμοποιεί επίσης έναν στόχο πρόβλεψης επόμενης πρότασης (next sentence prediction - NSP), όπου παρέχονται ζεύγη προτάσεων στο μοντέλο και πρέπει να προβλέψει εάν η δεύτερη πρόταση είναι η πραγματική επόμενη πρόταση της ακολουθίας.

Για την εκπαίδευση του BERT, απαιτείται τεράστιος όγκος δεδομένων κειμένου. Συνήθως, χρησιμοποιούνται σώματα μεγάλης κλίμακας όπως η Wikipedia και το BooksCorpus. Ο τεράστιος όγκος δεδομένων επιτρέπει στο BERT να μάθει ένα ευρύ φάσμα γλωσσικών προτύπων και σημασιολογικών σχέσεων. Κατά τη φάση της βελτιστοποίησης του, το BERT εκπαιδεύεται σε συγκεκριμένες εργασίες. Αυτή η διαδικασία περιλαμβάνει την προσθήκη επιπέδων συγκεκριμένων εργασιών πάνω από το προεκπαιδευμένο μοντέλο BERT και τη λεπτομερή ρύθμιση ολόκληρης της αρχιτεκτονικής σε δεδομένα για συγκεκριμένες εργασίες. Με την βελτιστοποίηση του, το BERT προσαρμόζει τις μαθημένες αναπαραστάσεις του στη συγκεκριμένη εργασία, με αποτέλεσμα βελτιωμένη απόδοση.

Ένα από τα κύρια πλεονεκτήματα του BERT είναι η ικανότητά του να χειρίζεται έννοιες και αποχρώσεις που εξαρτώνται από το πλαίσιο στη γλώσσα. Εξετάζοντας ολόκληρο το πλαίσιο μιας πρότασης, ο BERT μπορεί να συλλάβει λεπτές σημασιολογικές σχέσεις και να κατανοήσει καλύτερα τη σημασία των λέξεων με βάση το περιβάλλον τους.

Το μοντέλο BERT είναι μια πρωτοποριακή προσέγγιση στην επεξεργασία φυσικής γλώσσας που αξιοποιεί transformer αρχιτεκτονική και αμφίδρομη εκπαίδευση για να επιτύχει επιδόσεις αιχμής σε διάφορες εργασίες που σχετίζονται με τη γλώσσα. Η ικανότητά του να συλλαμβάνει πληροφορίες με βάση τα συμφραζόμενα και να χειρίζεται σύνθετα γλωσσικά μοτίβα το καθιστά ένα ισχυρό εργαλείο για την κατανόηση και την επεξεργασία της ανθρώπινης γλώσσας.

Medium Bert

Το MediumBERT είναι μια μικρότερη, πιο αποτελεσματική παραλλαγή του BERT που εξισορροπεί τις απαιτήσεις απόδοσης και πόρων, μεταξύ BERT-Base και συμπαγών μοντέλων όπως το DistilBERT. Ενώ το BERT-Base έχει 12 επίπεδα μετασχηματιστή, ένα κρυφό μέγεθος 768 νευρώνων και περίπου 110 εκατομμύρια παραμέτρους, το MediumBERT συνήθως μειώνει τον αριθμό των επιπέδων μετασχηματιστή σε 6–8, μειώνει το κρυφό μέγεθος των νευρώνων σε 512–640 και μπορεί να έχει 8–12 κεφαλές προσοχής, με αποτέλεσμα μέτρηση παραμέτρων περίπου 50–85 εκατομμύρια. Αυτή η βελτιωμένη αρχιτεκτονική επιτρέπει στο MediumBERT να προσφέρει ταχύτερη εξαγωγή συμπερασμάτων, χαμηλότερη χρήση μνήμης και καταλληλότητα για περιβάλλοντα με περιορισμένους πόρους, διατηρώντας παράλληλα ισχυρή απόδοση, συχνά με πτώση ακρίβειας μόνο 2–5% σε σύγκριση με το BERT-Base. Το MediumBERT είναι ιδανικό για εργασίες ή εφαρμογές μέτριας κλίμακας που απαιτούν επεξεργασία σε πραγματικό χρόνο, όπου το BERT-Base μπορεί να είναι πολύ μεγάλο, αλλά μικρότερα μοντέλα όπως το DistilBERT ενδέχεται να μην έχουν επαρκή χωρητικότητα.

DistilBERT

Το DistilBERT είναι ένα μοντέλο γλώσσας που έχει συγκεντρώσει σημαντική προσοχή για την αποδοτικότητα και την αποτελεσματικότητά του σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας (NLP). Είναι μια αποσταγμένη έκδοση του αρχικού μοντέλου BERT, που έχει σχεδιαστεί για να επιτυγχάνει παρόμοια απόδοση ενώ είναι υπολογιστικά πιο αποδοτική.

Το μοντέλο DistilBERT χρησιμοποιεί και αυτό την transformer αρχιτεκτονική, η οποία έχει φέρει επανάσταση στις εργασίες NLP καταγράφοντας σχέσεις με βάση τα συμφραζόμενα μεταξύ των λέξεων σε μια πρόταση. Αυτό το μοντέλο χρησιμοποιεί μια αμφίδρομη προσέγγιση, επιτρέποντάς του να λαμβάνει υπόψη τόσο τις προηγούμενες όσο και τις επόμενες λέξεις κατά την πρόβλεψη της σημασίας μιας λέξης. Αξιοποιώντας τον μηχανισμό αυτοπροσοχής (self-attention) της transformer αρχιτεκτονικής, το DistilBERT μπορεί να καταγράψει περίπλοκες εξαρτήσεις μεταξύ των λέξεων, οδηγώντας σε ανώτερη κατανόηση της γλωσσικής σημασιολογίας.

Ένα σημαντικό πλεονέκτημα του μοντέλου DistilBERT είναι το μικρότερο μέγεθος σε σύγκριση με το αρχικό μοντέλο BERT. Επιτυγχάνει αυτή τη μείωση με απόσταξη της γνώσης από το μεγαλύτερο μοντέλο σε ένα μικρότερο, διατηρώντας παράλληλα τη συνολική του απόδοση. Αυτή η διαδικασία συμπίεσης περιλαμβάνει την εκπαίδευση ενός μικρότερου μοντέλου ώστε να μιμείται τη συμπεριφορά του μεγαλύτερου μοντέλου, μεταφέροντας έτσι τις

γνώσεις του. Κατά συνέπεια, το DistilBERT κληρονομεί την ικανότητα κατανόησης πολύπλοκων γλωσσικών προτύπων και σημασιολογικών σχέσεων από το αρχικό μοντέλο BERT.

Εκτός από το συμπαγές του μέγεθος, ένα άλλο χαρακτηριστικό γνώρισμα του μοντέλου DistilBERT είναι ο ταχύτερος χρόνος συμπερασμάτων του. Αυτή η βελτιωμένη απόδοση το καθιστά πιο προσιτό και πρακτικό για εφαρμογές σε πραγματικό χρόνο. Η μείωση του μεγέθους του μοντέλου έχει επίσης ως αποτέλεσμα χαμηλότερες απαιτήσεις μνήμης, καθιστώντας ευκολότερη την ανάπτυξη σε συσκευές με περιορισμένους υπολογιστικούς πόρους.

Για την εκπαίδευση του μοντέλου DistilBERT, χρησιμοποιείται ένα μεγάλο σύνολο δεδομένων κειμένου, επιτρέποντας στο μοντέλο να μάθει από ένα ευρύ φάσμα γλωσσικών προτύπων και σημασιολογικών αποχρώσεων. Η εκπαίδευση περιλαμβάνει δύο βασικά βήματα: προεκπαίδευση και την βελτιστοποίηση. Κατά τη διάρκεια της προεκπαίδευσης, το μοντέλο μαθαίνει να προβλέπει καλυμμένες λέξεις σε μια πρόταση, επιτρέποντάς του να συλλαμβάνει τις συμφραζόμενες σχέσεις μεταξύ των λέξεων. Στο επόμενο στάδιο τελειοποίησης, το μοντέλο εκπαιδεύεται σε συγκεκριμένες μεταγενέστερες εργασίες, όπως ανάλυση συναισθήματος ή απάντηση ερωτήσεων, για να το προσαρμόσει σε πιο συγκεκριμένες εφαρμογές NLP.

Συνολικά, το μοντέλο DistilBERT προσφέρει μια αποτελεσματική και αποτελεσματική λύση για διάφορες εργασίες NLP. Η ικανότητά του να αποτυπώνει περίπλοκες σχέσεις με τα συμφραζόμενα, σε συνδυασμό με το μικρότερο μέγεθος και τον ταχύτερο χρόνο συμπερασμάτων, το καθιστά πολύτιμο εργαλείο στον τομέα της επεξεργασίας φυσικής γλώσσας.

ALBert

Το μοντέλο ALBERT (A Lite BERT) είναι μια παραλλαγή της αρχιτεκτονικής BERT (Αμφίδρομες αναπαραστάσεις κωδικοποιητή από μετασχηματιστές), που έχει σχεδιαστεί για να βελτιώνει τόσο την επεκτασιμότητα όσο και την απόδοση προεκπαιδευμένων μοντέλων γλώσσας μεγάλης κλίμακας. Το συγκεκριμένο μοντέλο παρουσιάστηκε στην εργασία «ALBERT: A Lite BERT for Self-supervised Learning of Language Representations» (Lan et al., 2020). Ο πρωταρχικός στόχος του ALBERT είναι να μειώσει τις ανάγκες μνήμης και την υπολογιστική πολυπλοκότητα του BERT, διατηρώντας ή βελτιώνοντας την απόδοσή του σε εργασίες κατανόησης φυσικής γλώσσας. Το ALBERT αντιμετωπίζει αυτές τις προκλήσεις εισάγοντας δύο βασικές αρχιτεκτονικές καινοτομίες: παραμετροποίηση παραγοντοποιημένης ενσωμάτωσης και κοινή χρήση παραμέτρων μεταξύ επιπέδων. Αυτές οι τροποποιήσεις μειώνουν δραστικά τον αριθμό των παραμέτρων χωρίς να θυσιάζεται η απόδοση του μοντέλου, καθιστώντας το ALBERT μια πιο αποδοτική υπολογιστικά εναλλακτική του BERT.

Στο BERT, το μέγεθος του στρώματος ενσωμάτωσης (το οποίο αντιστοιχίζει τα διακριτικά εισόδου σε πυκνά διανύσματα) συνδέεται με το μέγεθος του κρυφού στρώματος. Αυτό μπορεί να οδηγήσει σε μεγάλο αριθμό παραμέτρων, ειδικά όταν το μέγεθος του λεξιλογίου είναι μεγάλο. Ο ALBERT αποσυνδέει το μέγεθος του λεξιλογίου που ενσωματώνεται από το μέγεθος του κρυφού στρώματος εισάγοντας μια διαδικασία παραγοντοποίησης δύο βημάτων. Αρχικά τα διακριτικά εισόδου ενσωματώνονται σε ένα χώρο χαμηλότερης διάστασης (η διάσταση ενσωμάτωσης είναι πολύ μικρότερη). Στη συνέχεια, ένα επίπεδο προβολής αντιστοιχίζει αυτά τα embeddings σε έναν κρυφό χώρο υψηλότερης διάστασης που χρησιμοποιείται από τα στρώματα του μετασχηματιστή. Αυτή η παραγοντοποίηση μειώνει σημαντικά τον αριθμό των παραμέτρων στο επίπεδο των embeddings, καθιστώντας το μοντέλο πιο αποτελεσματικό χωρίς να επηρεάζει την ικανότητά του να μαθαίνει ουσιαστικές αναπαραστάσεις.

Ένας από τους μεγαλύτερους παράγοντες στον αριθμό παραμέτρων του BERT είναι η χρήση ανεξάρτητων παραμέτρων σε κάθε στρώμα του μετασχηματιστή. Ο ALBERT μετριάζει αυτό με την κοινή χρήση παραμέτρων σε όλα τα επίπεδα του μοντέλου. Συγκεκριμένα, τα βάρη των επιπέδων τροφοδοσίας και των επιπέδων προσοχής μοιράζονται μεταξύ των επιπέδων, που σημαίνει ότι οι ίδιες παράμετροι επαναχρησιμοποιούνται σε όλο το δίκτυο. Αυτό μειώνει δραστικά τον συνολικό αριθμό των παραμέτρων, επιτρέποντας στον ALBERT να κλιμακωθεί σε βαθύτερες αρχιτεκτονικές χωρίς έκρηξη στη χρήση μνήμης.

Τα βασικά πλεονεκτήματα του ALBERT σε σχέση με το μεγαλύτερο μοντέλο προκάτοχο του BERT είναι το μικρότερο μέγεθος του μοντέλου, η καλύτερη επεκτασιμότητα του, και η δυνατότητα του για καλύτερη γενίκευση.

Μειωμένο μέγεθος μοντέλου. Μέσω παραγοντοποιημένων ενσωματώσεων και κοινής χρήσης παραμέτρων σε πολλαπλά επίπεδα, το ALBERT μειώνει τον αριθμό των παραμέτρων σε σύγκριση με το BERT, διατηρώντας παράλληλα ανταγωνιστική ή ανώτερη απόδοση. Για παράδειγμα, η μεγαλύτερη παραλλαγή του ALBERT, το ALBERT-xxlarge, έχει πολύ λιγότερες παραμέτρους από το BERT-large, αλλά επιτυγχάνει αποτελέσματα αιχμής σε πολλά σημεία αναφοράς NLP.

Βελτιωμένη επεκτασιμότητα. Η μείωση παραμέτρων του ALBERT επιτρέπει την αποτελεσματική εκπαίδευση των βαθύτερων μοντέλων. Με λιγότερες παραμέτρους, το μοντέλο καταναλώνει λιγότερη μνήμη και εκπαιδεύεται πιο γρήγορα, καθιστώντας δυνατή την ανάπτυξη του ALBERT σε μεγαλύτερη κλίμακα.

Καλύτερη γενίκευση. Εκτός από τη συμπίεση μοντέλων, ο ALBERT εισάγει την πρόβλεψη σειράς προτάσεων (SOP) ως μια νέα εργασία προεκπαίδευσης, η οποία βελτιώνει την κατανόησή του για τη συνοχή των προτάσεων και τις σχέσεις μεταξύ των προτάσεων. Αυτή η

βελτίωση σε σχέση με την εργασία πρόβλεψης επόμενης πρότασης (NSP) του BERT βοηθά τον ALBERT να αποδίδει καλύτερα σε εργασίες που απαιτούν συλλογισμό πολλών προτάσεων.

RoBERTa

Το μοντέλο RoBERTa είναι ένα προηγμένο μοντέλο επεξεργασίας φυσικής γλώσσας (NLP) που έχει κερδίσει σημαντική προσοχή για την αποτελεσματικότητά του στην πρόβλεψη αποφάσεων κρίσης. Αναπτύχθηκε από το Facebook AI Research και βασίζεται στην transformer αρχιτεκτονική, για την κατανόηση της φυσικής γλώσσας.

Στον πυρήνα του, το μοντέλο RoBERTa χρησιμοποιεί μια μεγάλης κλίμακας προεκπαιδευτική διαδικασία σε διάφορες μορφές σωμάτων κειμένου από το Διαδίκτυο. Αυτή η φάση προεκπαίδευσης περιλαμβάνει την πρόβλεψη λέξεων που λείπουν μέσα στις προτάσεις, γνωστή και ως μοντελοποίηση γλώσσας με μάσκα (masked language modeling). Με την εκπαίδευση σε έναν τεράστιο όγκο δεδομένων κειμένου, η Roberta μαθαίνει τις στατιστικές ιδιότητες της γλώσσας, συμπεριλαμβανομένων της γραμματικής, της σειράς λέξεων και των εξαρτήσεων από τα συμφραζόμενα.

Μετά την προεκπαίδευση, το μοντέλο της Roberta υποβάλλεται σε μια φάση βελτιστοποίησης, όπου εκπαιδεύεται σε συγκεκριμένες εργασίες, όπως για παράδειγμα η πρόβλεψη απόφασης δικαστηρίου. Η βελτιστοποίηση περιλαμβάνει την εκπαίδευση του μοντέλου σε ένα μικρότερο, ειδικό για κάθε εργασία σύνολο δεδομένων που φέρει ετικέτα με τα επιθυμητά αποτελέσματα. Αυτή η διαδικασία επιτρέπει στο μοντέλο να προσαρμόσει τις γνώσεις του και να βελτιστοποιήσει τις παραμέτρους του για τη συγκεκριμένη εργασία.

Ένα από τα βασικά πλεονεκτήματα του μοντέλου RoBERTa έγκειται στην ικανότητά του να συλλαμβάνει τις συμφραζόμενες πληροφορίες λέξεων και προτάσεων. Λαμβάνοντας υπόψη τις γύρω λέξεις και τις σχέσεις τους, το RoBERTa μπορεί να κατανοήσει το νόημα και την πρόθεση πίσω από το κείμενο. Αυτή η κατανόηση των συμφραζομένων επιτρέπει στο μοντέλο να κάνει ακριβείς προβλέψεις σχετικά με τις δικαστικές αποφάσεις, αξιοποιώντας τη γνώση που έχει αποκτήσει κατά τη διάρκεια της προεκπαίδευσης και της βελτιστοποίησης. Επιπλέον, η RoBERTa ενσωματώνει μια τεχνική που ονομάζεται εκμάθηση πολλαπλών εργασιών, όπου εκπαιδεύεται σε πολλαπλές σχετικές εργασίες ταυτόχρονα. Αυτή η προσέγγιση βοηθά το μοντέλο να γενικεύει καλύτερα και βελτιώνει τη συνολική του απόδοση σε διάφορες εργασίες NLP.

Συμπερασματικά, το μοντέλο της RoBERTa λειτουργεί αξιοποιώντας την προεκπαίδευση σε ένα μεγάλο σύνολο δεδομένων κειμένου, ακολουθούμενη από βελτιστοποίηση σε σύνολα δεδομένων για συγκεκριμένες εργασίες. Μέσω της ικανότητάς της να συλλαμβάνει πληροφορίες

με βάση τα συμφραζόμενα και να εκμεταλλεύεται τη μάθηση πολλαπλών εργασιών, το RoBERTa διαπρέπει στον τομέα του NLP.

LongFormer

Το μοντέλο Longformer είναι μια παραλλαγή της αρχιτεκτονικής Transformer ειδικά σχεδιασμένη για να χειρίζεται μεγάλα έγγραφα πιο αποτελεσματικά ξεπερνώντας την τετραγωνική πολυπλοκότητα του παραδοσιακού μηχανισμού αυτοπροσοχής του Transformer. Το Longformer εισήχθη από (Beltagy et al., 2020). Το μοντέλο είναι ικανό να επεξεργάζεται πολύ μεγαλύτερες ακολουθίες σε σύγκριση με τα αρχικά μοντέλα Transformer και BERT, τα οποία συνήθως περιορίζονται από τις υπολογιστικές τους απαιτήσεις σε μικρότερες ακολουθίες (π.χ. 512 tokens στο BERT).

Το Longformer αντιμετωπίζει αυτόν τον περιορισμό εισάγοντας έναν νέο, πιο αποτελεσματικό μηχανισμό προσοχής που κλιμακώνεται γραμμικά με το μήκος της ακολουθίας, επιτρέποντάς του να χειρίζεται έγγραφα με χιλιάδες διακριτικά διατηρώντας παράλληλα ισχυρή απόδοση σε εργασίες κατανόησης φυσικής γλώσσας.

Οι βασικές καινοτομίες στο Longformer είναι το Sliding Window Attention, τα μοτίβα διευρυμένης προσοχής και η προσοχή σε συγκεκριμένα tokens

Η βασική καινοτομία του Longformer έγκειται στον μηχανισμό Sliding Window Attention. Οι παραδοσιακοί μετασχηματιστές υπολογίζουν την αυτοπροσοχή σε όλα τα διακριτικά στην ακολουθία εισόδου, με αποτέλεσμα πολυπλοκότητα των διαδικασίας να ανεβαίνει εκθετικά ως προς το μήκος της ακολουθίας (δηλ.$O(n^2)$, όπου n είναι το μήκος της ακολουθίας. Αυτό είναι υπολογιστικά ακριβό και ανέφικτο για μεγάλες ακολουθίες. Το Longformer το αντικαθιστά με έναν μηχανισμό τοπικής προσοχής, όπου κάθε διακριτικό παρακολουθεί μόνο ένα παράθυρο σταθερού μεγέθους γειτονικών κουπονιών (π.χ. 128 μάρκες). Αυτό μειώνει την υπολογιστική πολυπλοκότητα σε γραμμική ως προς το μήκος της ακολουθίας (δηλ. $O(n)$, επιτρέποντάς του να επεξεργάζεται πολύ μεγαλύτερες ακολουθίες, ενώ εξακολουθεί να καταγράφει τοπικές εξαρτήσεις μέσα στο έγγραφο.

Εκτός από το Sliding Window Attention, το Longformer υποστηρίζει μοτίβα διευρυμένης προσοχής, τα οποία επιτρέπουν στα token να παρακολουθούν μη γειτονικά διακριτικά σε προκαθορισμένα διαστήματα. Αυτό επεκτείνει το πεδίο υποδοχής κάθε διακριτικού πέρα από τους άμεσους γείτονές του, βοηθώντας στην καταγραφή εξαρτήσεων μεγαλύτερης εμβέλειας χωρίς να αυξάνει σημαντικά το μέγεθος του παραθύρου προσοχής.

Για εργασίες που απαιτούν προσοχή σε ολόκληρο το έγγραφο (όπως η απάντηση ερωτήσεων ή η σύνοψη), το Longformer εισάγει την καθολική προσοχή για ένα επιλεγμένο υποσύνολο

διακριτικών. Αυτά τα διακριτικά (π.χ. διακριτικά ερωτήσεων σε μια εργασία απάντησης ερωτήσεων) μπορούν να παρακολουθήσουν όλα τα άλλα διακριτικά της σειράς, ενώ άλλα διακριτικά συνεχίζουν να χρησιμοποιούν την προσοχή του συρόμενου παραθύρου. Αυτή η υβριδική προσέγγιση διασφαλίζει ότι τα βασικά διακριτικά μπορούν να συλλέγουν πληροφορίες από ολόκληρο το έγγραφο, διατηρώντας παράλληλα αποτελεσματικό τον συνολικό υπολογισμό.

Τα πλεονεκτήματα του LongFormer είναι η μείωση της πολυπλοκότητας των υπολογισμών σε γραμμική, ο αποτελεσματικός χειρισμός μεγάλων εγγράφων και τα ευέλικτα μοτίβα προσοχής. Το Sliding Attention Window μειώνει την πολυπλοκότητα που απαιτείται για την προσοχή του μοντέλου από την εκθετική (τετραγωνική) $O(n^2)$ του BERT σε γραμμική $O(n)$ στο Longformer, επιτρέποντάς του να χειρίζεται πολύ μεγαλύτερες ακολουθίες (έως πολλές χιλιάδες tokens).Το Longformer είναι ιδιαίτερα κατάλληλο για εργασίες που απαιτούν την επεξεργασία μεγάλων εγγράφων, όπως η ταξινόμηση εγγράφων, η σύνοψη και η απάντηση σε ερωτήσεις μεγάλου πλαισίου. Επιτρέποντας την αποτελεσματική προσοχή σε μεγάλες ακολουθίες, μειώνει την κατανάλωση μνήμης και επιταχύνει την εξαγωγή συμπερασμάτων. Το Sliding Window Attention του Longformer είναι ευέλικτο, υποστηρίζοντας συνδυασμούς του, διευρυμένης και συνολικής προσοχής. Αυτό επιτρέπει στο μοντέλο να καταγράφει τόσο τοπικές εξαρτήσεις όσο και σχέσεις μακράς εμβέλειας, καθιστώντας το εξαιρετικά ευέλικτο σε διάφορες εργασίες NLP.

BigBird

Το BigBird είναι ένα εξαιρετικά αποδοτικό μοντέλο που βασίζεται σε μετασχηματιστές και έχει σχεδιαστεί για να χειρίζεται μεγάλες ακολουθίες πιο αποτελεσματικά από τους παραδοσιακούς μετασχηματιστές όπως το BERT. Το BigBird παρουσιάστηκε από (Zaheer et al., 2020). Ξεπερνά το ζήτημα της τετραγωνική πολυπλοκότητα αυτοπροσοχής χρησιμοποιώντας έναν μηχανισμό αραιής προσοχής που μειώνει το κόστος υπολογισμού και μνήμης. Αυτό επιτρέπει στο BigBird να επεξεργάζεται αλληλουχίες μήκους έως 8.192 tokens και πέρα, κάτι που είναι κρίσιμο για εφαρμογές όπως η σύνοψη εγγράφων, η απάντηση σε ερωτήσεις μεγάλου πλαισίου και η μοντελοποίηση ακολουθιών DNA.

Η καινοτομία του BigBird έγκειται στην ικανότητά του να προσεγγίζει τον μηχανισμό πλήρους προσοχής των παραδοσιακών μετασχηματιστών, διατηρώντας παράλληλα θεωρητικές εγγυήσεις από τη θεωρία γραφημάτων, καθιστώντας το ταυτόχρονα επεκτάσιμο και ισχυρό για επεξεργασία μεγάλης ακολουθίας.

Το BigBird βασίζεται στην αρχιτεκτονική του Transformer τροποποιώντας τον μηχανισμό αυτοπροσοχής ώστε να είναι πιο αποδοτικός υπολογιστικά, ιδιαίτερα για μεγάλες ακολουθίες. Ο παραδοσιακός μετασχηματιστής κλιμακώνεται τετραγωνικά με μήκος ακολουθίας $O(n^2)$, καθιστώντας το μη πρακτικό για μεγάλα έγγραφα. Το BigBird το αντιμετωπίζει αυτό εισάγοντας

την αραιή προσοχή με έναν συνδυασμό καθολικών, τυχαίων και συρόμενων μηχανισμών προσοχής παραθύρων. Το BigBird μειώνει την εκθετική πολυπλοκότητα της αυτοπροσοχής χρησιμοποιώντας έναν μηχανισμό αραιής προσοχής. Σε αυτόν τον μηχανισμό, αντί να παρακολουθεί κάθε token της ακολουθίας, παρακολουθεί μόνο ένα υποσύνολο άλλων token. Αυτό το υποσύνολο επιλέγεται με βάση έναν συνδυασμό τοπικής προσοχής (συρόμενο παράθυρο), τυχαίας προσοχής και συνολικής προσοχής. Στο Sliding Window Attention Κάθε token παρακολουθεί ένα παράθυρο σταθερού μεγέθους γειτονικών token, καταγράφοντας τοπικές εξαρτήσεις. Στην τυχαία προσοχή (Random Attention) κάθε token παρακολουθεί επίσης ένα τυχαίο σύνολο token σε ολόκληρη την ακολουθία, το οποίο βοηθά στην αποτύπωση παγκόσμιων εξαρτήσεων χωρίς να χρειάζεται να παρακολουθείτε κάθε token. Στην καθολική προσοχή (Global Attention) ορισμένα token, που ορίζονται ως «σημαντικά» (π.χ. διακριτικό [CLS] για εργασίες ταξινόμησης), παρακολουθούν κάθε άλλο token της σειράς. Αυτό διασφαλίζει ότι τα βασικά token εξακολουθούν να έχουν πλήρη πρόσβαση σε ολόκληρη την ακολουθία, καθιστώντας το BigBird κατάλληλο για εργασίες όπως η σύνοψη και η απάντηση ερωτήσεων. Αυτή η αραιή προσοχή μειώνει την πολυπλοκότητα της προσοχής από την τετραγωνική $O(n^2)$ σε γραμμική $O(n)$ όπου n είναι το μήκος της ακολουθίας.

Ο μηχανισμός προσοχής του BigBird βασίζεται στη θεωρία γραφημάτων, ειδικά στα γραφήματα επέκτασης. Παρά την αραιή προσοχή του, το BigBird προσεγγίζει την πλήρη προσοχή διασφαλίζοντας ότι κάθε token συνδέεται έμμεσα με κάθε άλλο token της ακολουθίας. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να αποτυπώνει αποτελεσματικά τόσο τις τοπικές όσο και τις παγκόσμιες εξαρτήσεις, διατηρώντας την ισχυρή απόδοση μοντέλων πλήρους προσοχής όπως το BERT. Χάρη στην αραιή προσοχή, το BigBird μπορεί να χειριστεί ακολουθίες έως και 8.192 tokens ή περισσότερα χωρίς σημαντική αύξηση στο υπολογιστικό κόστος και στη χρήση μνήμης. Αυτή η επεκτασιμότητα το καθιστά ιδανικό για εργασίες που περιλαμβάνουν μεγάλες εισροές, όπως η επεξεργασία ερευνητικών εγγράφων, νομικών εγγράφων ή βιολογικών αλληλουχιών (π.χ. DNA).

Έτσι τα πλεονεκτήματα του BigBird είναι η μείωση του υπολογιστικού κόστους της της προσοχής από τετραγωνικό σε γραμμικό, καθιστώντας εφικτή την επεξεργασία πολύ μεγαλύτερες ακολουθίες από τους τυπικούς μετασχηματιστές όπως το BERT ή το GPT. Επιπλέον παρά τη χρήση αραιής προσοχής, το BigBird διατηρεί τις θεωρητικές δυνατότητες των μοντέλων πλήρους προσοχής. Αυτό επιτυγχάνεται με την εξασφάλιση παγκόσμιας συνδεσιμότητας μεταξύ των διακριτικών, επιτρέποντάς του να συλλαμβάνει εξαρτήσεις τόσο μικρής όσο και μεγάλης εμβέλειας στα δεδομένα. Τέλος το BigBird έχει επιδείξει ισχυρή απόδοση σε μια μεγάλη ποικιλία εργασιών, όπως επεξεργασία φυσικής γλώσσας (NLP), εργασίες όπως σύνοψη εγγράφων, απάντηση ερωτήσεων και ταξινόμηση κειμένου, βιοϊατρικές

εφαρμογές, εργασίες όπως η επεξεργασία γονιδιωματικών αλληλουχιών, όπου η ανάλυση μακρών αλληλουχιών DNA είναι κρίσιμη και μπορεί να εφαρμοστεί σε δεδομένα εικόνας και βίντεο που περιλαμβάνουν ακολουθίες οπτικών πλαισίων ή ακολουθίες εικονοστοιχείων, επεκτείνοντας τη χρήση του πέρα από το κείμενο.

Προηγούμενες Εργασίες

Υπάρχει πληθώρα υλοποιημένων εργασιών πάνω στον τομέα πρόβλεψης δικαστικών αποφάσεων. Οι εργασίες αυτές έχουν διεξαχθεί σε σε νομικά κείμενα διάφορων γλωσσών όπως αγγλικά (Strickson & Iglesia, 2020), τουρκικά (Turan et al., 2023), ινδικά (Shirsat et al., 2021), γαλλικά, γερμανικά και ιταλικά (Niklaus et al., 2021). Το κίνητρο πίσω από την διεξαγωγή αυτού του είδους εργασιών είναι η επιτάχυνση απόδοσης μίας απόφασης καθώς ένα κοινό χαρακτηριστικό των δικαστικών συστημάτων είναι η μεγάλες καθυστερήσεις λόγω της δαιδαλώδης γραφειοκρατίας.

Επιπλέον έχουν γίνει εργασίες πάνω στα δεδομένα τα οποία χρησιμοποιήθηκαν στην παρούσα εργασία. Στην εργασία του (Aletras et al., 2016) έγινε προσπάθεια πρόβλεψης της κλάσης για παραβίασης η μη των δικαιωμάτων που προστατεύονται από το συμβούλιο της Ευρώπης όπου χρησιμοποιήθηκε το SVM μοντέλο με linear kernel πετυχαίνοντας ακρίβεια 73 %. Επιπλέον στην εργασία του (Chalkidis et al., 2019) έγινε προσπάθεια επίλυσης του ίδιου προβλήματος καθώς επίσης και της πρόβλεψης της σπουδαιότητας της κάθε κλάσης χρησιμοποιώντας νευρωνικά δίκτυα και ένα SVM μοντέλο ως baseline. Τις καλύτερες επιδόσεις πέτυχε το HIER - BERT το οποίο έχει την δυνατότητα να λαμβάνει απεριόριστο μέγεθος κειμένου ως είσοδο.

	P	R	F1
MAJORITY	32.9 ± 0.0	50.0 ± 0.0	39.7 ± 0.0
COIN-TOSS	50.4 ± 0.7	50.5 ± 0.8	49.1 ± 0.7
Non-Anonymized			
BOW-SVM	71.5 ± 0.0	72.0 ± 0.0	71.8 ± 0.0
BIGRU-ATT	87.1 ± 1.0	77.2 ± 3.4	79.5 ± 2.7
HAN	88.2 ± 0.4	78.0 ± 0.2	80.5 ± 0.2
BERT	24.0 ± 0.2	50.0 ± 0.0	17.0 ± 0.5
HIER-BERT	**90.4** ± 0.3	**79.3** ± 0.9	**82.0** ± 0.9

`Επίσης στην ίδια εργασία έγινε προσπάθεια προβλεψης του επιπέδου σπουδαιότητας ενός κειμένου όπου τα αποτελέσματα δεν ήταν πολύ καλά. Πιο συγκεκριμένα το άθροισμα των τετραγώνων του λάθους δεν έπεσε κάτω από το 40 %. Το άρθρο καταλήγει στο ότι η νομική σπουδαιότητα ενός κειμένου οφείλεται όχι μόνο στην περιγραφή των γεγονότων σε ελεύθερο κείμενο αλλά και στην προηγούμενη γνώση των νομικών και δικαστικών.

	MAE	SPEARMAN's ρ
MAJORITY	**.369** ± .000	N/A*
BOW-SVR	.585 ± .000	.370 ± .000
BIGRU-ATT	.539 ± .073	.459 ± .034
HAN	.524 ± .049	.437 ± .018
HIER-BERT	.437 ± .018	**.527** ± .024

Μεθοδολογία

Στην παρούσα εργασία αναπτύχθηκαν μοντέλα μηχανικής μάθησης τα οποία εκπαιδεύτηκαν και αξιολογήθηκαν στο σύνολο δεδομένων του Ευρωπαϊκού Δικαστηρίου Ανθρωπίνων Δικαιωμάτων (European Court of Human Rights - ECtHR). Τα δεδομένα περιλαμβάνουν την περιγραφή μίας υπόθεσης σε ελεύθερο κείμενο καθώς και το αποτέλεσμα της δίκης το οποίο μπορεί να είναι η παραβίαση άρθρου ή άρθρων της Οικουμενικής Διακήρυξης των Δικαιωμάτων των Ανθρώπων. Έτσι κάθε δικαστική υπόθεση συνοδεύεται και από το αντίστοιχο άρθρο ή άρθρα τα οποία παραβιάστηκαν. Στην εργασία το πρόβλημα αυτό αντιμετωπίστηκε ως ένα δυαδικό πρόβλημα κατηγοριοποίησης σε “ΠΑΡΑΒΙΑΣΗ” (VIOLATED) και “ΜΗ ΠΑΡΑΒΙΑΣΗ’ (NON_VIOLATED) άρθρου της Διακήρυξης Ανθρωπίνων Δικαιωμάτων όπως έχει γίνει και στις εργασίες των (Aletras et al., 2016), (Medvedeva et al., 2020) και (Medvedeva et al., 2021). Στην περίπτωση (Aletras et al., 2016) και (Medvedeva et al., 2020) η δυαδική κατηγοριοποίηση έγινε για κάθε άρθρο ξεχωριστά.

Dataset

Το δεδομένα τα οποία χρησιμοποιήθηκαν αποτελούν την συλλογή ECHR. Τα δεδομένα αποτελούν περιγραφές σε ελεύθερο κείμενο δικαστικών υποθέσεων στο Ευρωπαϊκό Δικαστήριο. Τα δεδομένα αυτά εξήχθησαν από το την ιστοσελίδα του Ευρωπαϊκού Δικαστηρίου Ανθρωπίνων Δικαιωμάτων (https://hudoc.echr.coe.int/) με δημιουργούς τους Αλετράς Νικόλας και Χαλκίδης Ηλίας (Aletras & Chalkidis, 2019).

Τα δεδομένα περιέχουν περίπου 11,5 χιλιάδες δικαστικές υποθέσεις. Κάθε δικαστική υπόθεση περιέχει μία λίστα με γεγονότα σε ελεύθερο κείμενο, μία λίστα με άρθρα που παραβιάζονται αν τυχόν υπάρχουν και μία τιμή σημαντικότητας της συγκεκριμένης υπόθεσης. Η βαθμολογία σπουδαιότητας σε κάθε υπόθεση αντανακλά την συμβολή της υπόθεσης στην δημιουργία νομολογίας. Η κλίμακα της σημαντικότητας είναι μία βαθμολογία από το 1 (πολύ σημαντική) έως το 4 (ασήμαντη).Τα δεδομένα είναι χωρισμένα σε δεδομένα εκπαίδευσης (training),

επαλήθευσης (dev) και δοκιμής (test). Τα training δεδομένα περιέχουν υποθέσεις από το 1959 έως το 2013 και τα test δεδομένα περιέχουν δικαστικές υποθέσεις από το 2014 έως το 2018. Τα training και dev δεδομένα είναι ισορροπημένα περιέχοντας ίσο αριθμό υποθέσεων ανάμεσα σε δικαστικές υποθέσεις με αποτέλεσμα παραβίασης άρθρου της ευρωπαϊκής σύμβασης για την προστασία των ανθρωπίνων δικαιωμάτων και υποθέσεις χωρίς παραβίαση. Η ισορροπία ανάμεσα στα δύο είδη δικαστικών υποθέσεων συμβάλει στην αποφυγή μεροληψίας του μοντέλου μηχανικής μάθησης κατά την εκπαίδευση προς την μία από τις δύο κλάσεις. Τα test δεδομένα περιέχουν 66% περισσότερες υποθέσεις με παραβίαση κάποιου άρθρου όπως συμβαίνει στην βάση δεδομένων του Ευρωπαϊκού Δικαστηρίου.

Για την δυαδική διάκριση των δικαστικών υποθέσεων ανάμεσα σε υποθέσεις με παραβίαση κάποιου άρθρου και σε υποθέσεις χωρίς παραβίαση εξετάστηκε η στήλη των άρθρων που σχετίζονται με την κάθε υπόθεση. Με την κλάση 0 χαρακτηρίστηκαν οι υποθέσεις που δεν περιλαμβάνουν κάποιο άρθρο που παραβιάστηκε (NON_VIOLATED) και με την κλάση 1 οι υποθέσεις που περιλαμβάνουν κάποιο άρθρο που παραβιάστηκε (VIOLATED).

Στατιστικά του ECHR Data

Το train σύνολο των δεδομένων περιέχει 7100 υποθέσεις, το validation σύνολο 1380 υποθέσεις και το test σύνολο 2998 υποθέσεις. Στον παρακάτω πίνακα εμφανίζονται ο μέσος όρος facts, λέξεων και χαρακτήρων ανά υπόθεση για κάθε υποσύνολο των δεδομένων (train, validation, test).

Subset	Cases(C)	Facts/C	Words/C	Characters/C
Train	7100	42.74	2310.67	13979.18
Validation	1380	45.07	2466.95	14901.10
Test	2998	31.40	1912.21	11583.34

Violation - Binary Classification

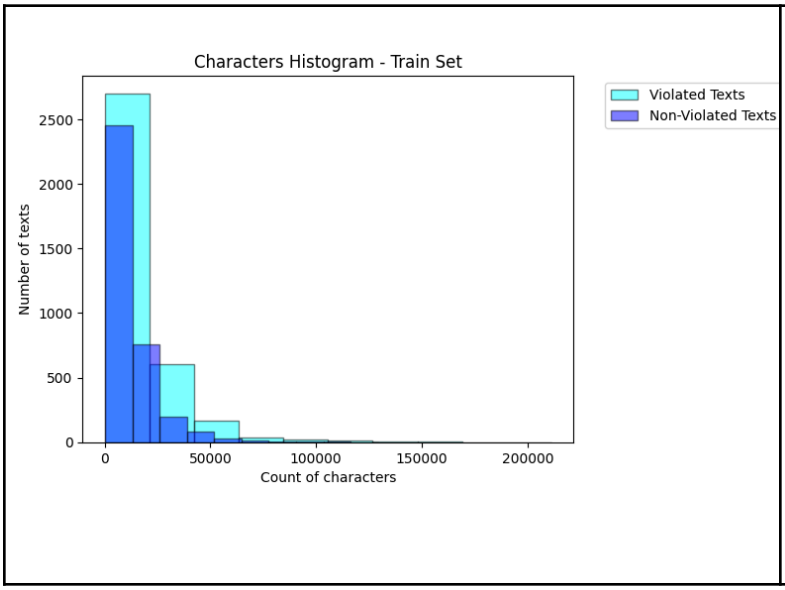
Η κλάση 0 (NON_VIOLATED) περιέχει 3549 παραδείγματα εκπαίδευσης στο train set και 690 παραδείγματα εκπαίδευσης στο validation set με τα ακόλουθα στατιστικά στοιχεία

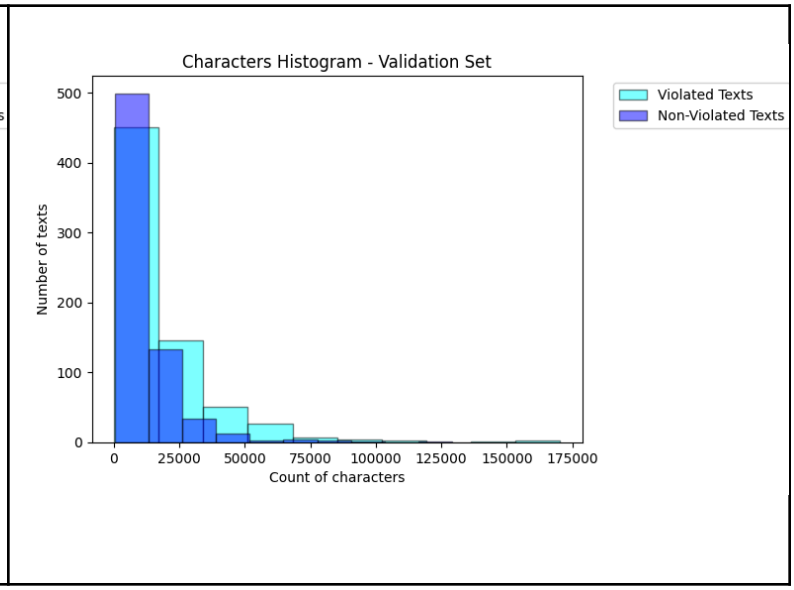
Subset	Cases(C)	Facts/C	Words/C	Characters/C
Train	3549	35.69	1958.18	11855.18
Validation	690	35.65	1964.00	11875.5

Η κλάση 1 (VIOLATED) περιέχει 3551 παραδείγματα εκπαίδευσης στο train set και 690 παραδείγματα εκπαίδευσης στο validation set με τα ακόλουθα στατιστικά στοιχεία

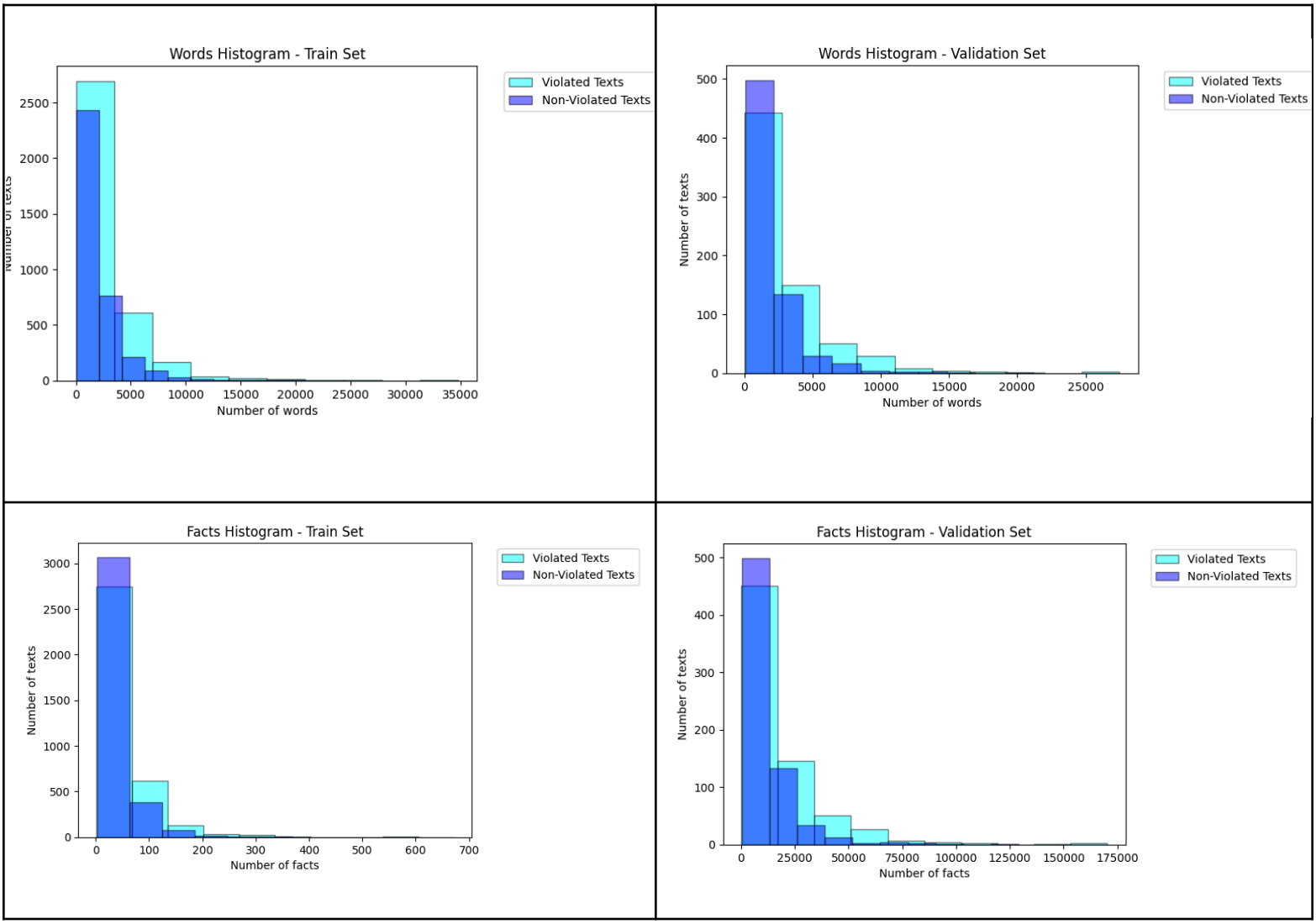
Subset	Cases(C)	Facts/C	Words/C	Characters/C
Train	3551	49.79	2662.96	16102.46
Validation	690	54.50	2969.91	17927.30

Τα παρακάτω ιστογράμματα παρουσιάζουν τον αριθμό των χαρακτήρων, τον αριθμό των λέξεων και τον αριθμό των γεγονότων σε κάθε κλάση για κάθε υποσύνολο δεδομένων



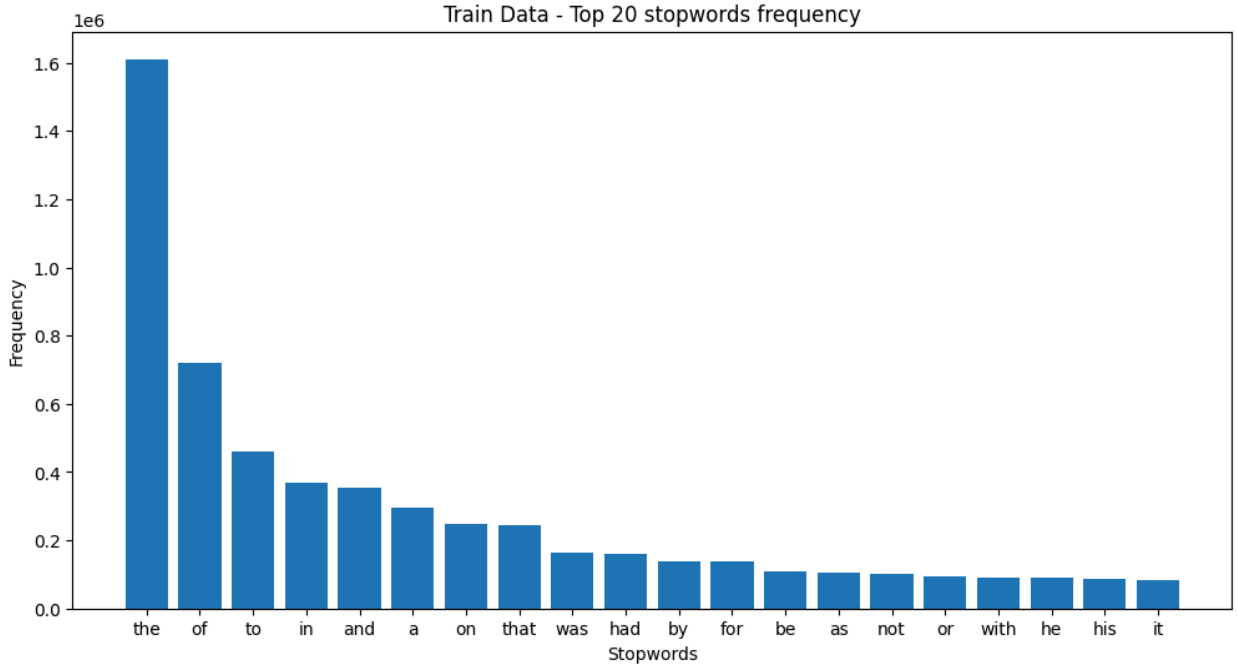


Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

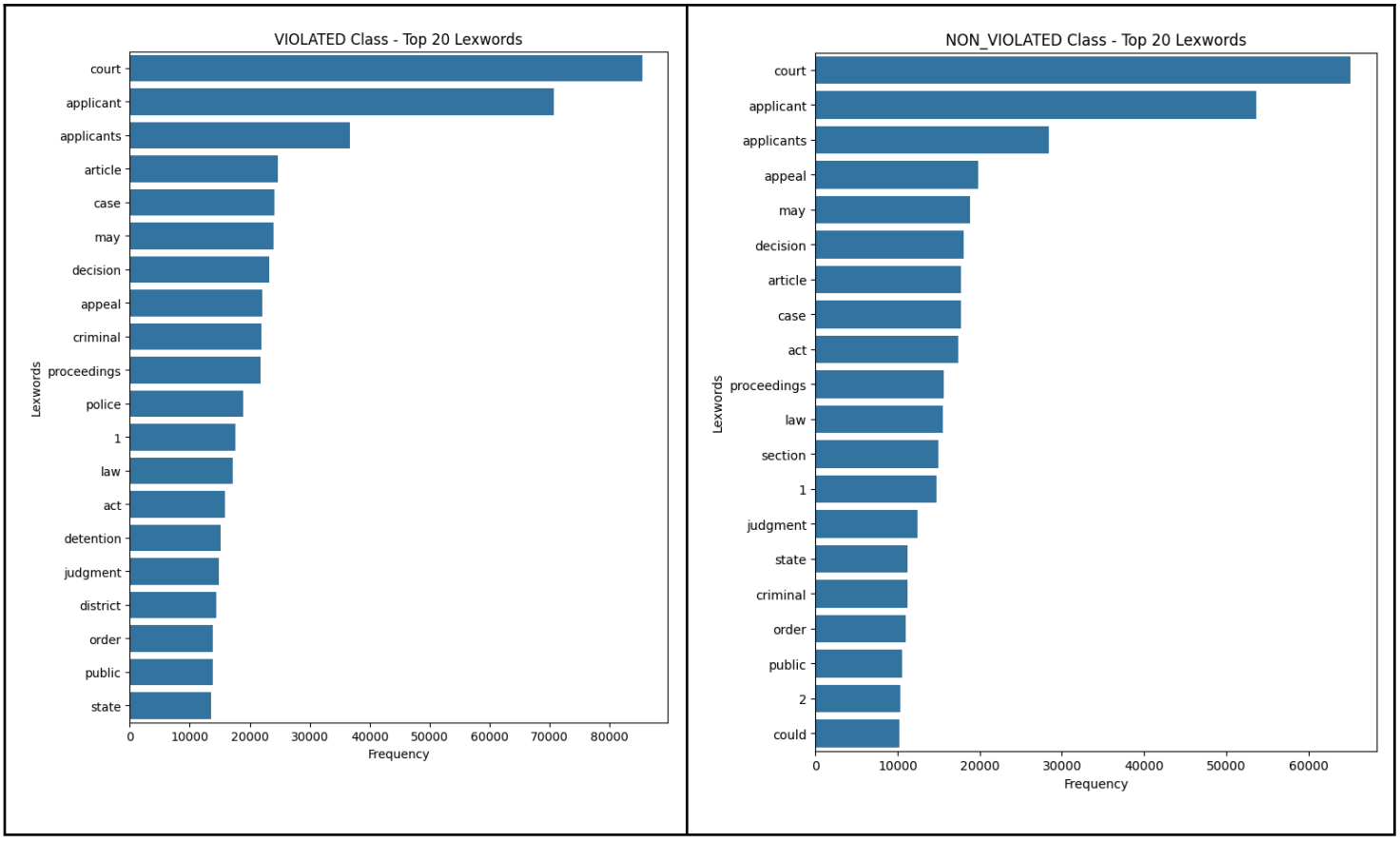


Βασικό εργαλείο της επεξεργασίας φυσικής γλώσσας ο εντοπισμός των λέξεων εκείνων που έχουν βαρύνουσα σημασία στον καθορισμό μίας κλάσης. Η σημασία αυτή μπορεί να εντοπιστεί στην συχνότητα των λέξεων αυτών σε μία μόνο από τις δύο κλάσεις. Αυτό σημαίνει ότι λέξεις που εμφανίζονται και στις δύο κλάσεις έχουν μικρή σημασία. Τέτοιες λέξεις είναι συνήθως λέξεις με λειτουργική σημασία, επιτελούν δηλαδή γραμματικές λειτουργίες η λέξεις γενικής χρήσης. Αυτές ονομάζονται ενδιάμεσες λέξεις (stopwords) στην ορολογία της επεξεργασίας φυσικής γλώσσας. Στην συνέχεια παρουσιάζεται η συχνότητα των ενδιάμεσων λειτουργικών λέξεων (stopwords) οι οποίες δεν έχουν σημασιολογικό περιεχόμενο στο σύνολο εκπαίδευσης

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων



Μετά τις ενδιάμεσες λέξεις παρουσιάζονται οι λέξεις που δεν ανήκουν στην κατηγορία stopwords στο σύνολο εκπαίδευσης



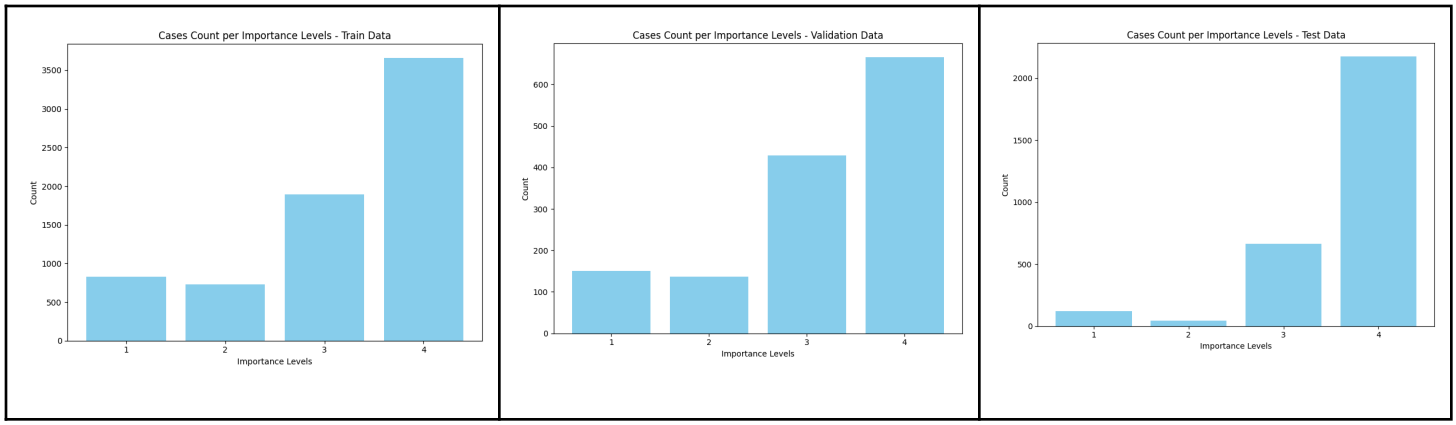
Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Importance - Multiclass Classification

Η σημαντικότητα μία υπόθεσης αφορά ένα επιπλέον χαρακτηριστικό των ποινικών υποθέσεων. Συγκεκριμένα η τιμή αυτή παίρνει διακριτές τιμές σε μία κλίμακα από το 1 μέχρι το 4. Το χαρακτηριστικό αυτό αφορά το πόσο σημαντική είναι η συγκεκριμένη ποινική υπόθεση στην διαμόρφωση νομολογίας. Όσο πιο μικρή είναι η τιμή του importance τόσο πιο σημαντική είναι η υπόθεση. Για την εξαγωγή στατιστικών διαγραμμάτων και την σύγκριση χαρακτηριστικών μεταξύ των 4 κλάσεων έγινε χρήση του διάμεσου κατά την ομαδοποίηση κατά κλάση. Από την στατιστική ανάλυση των δεδομένων προέκυψαν οι παρακάτω πίνακες και διαγράμματα ανά κλάση.

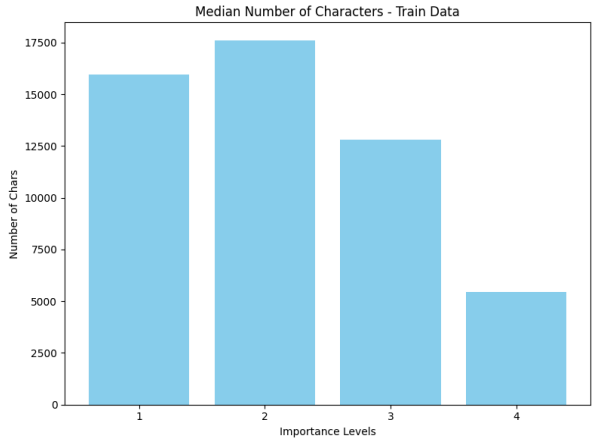
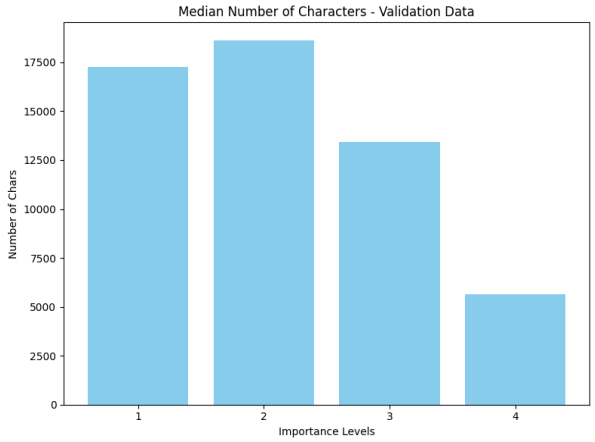
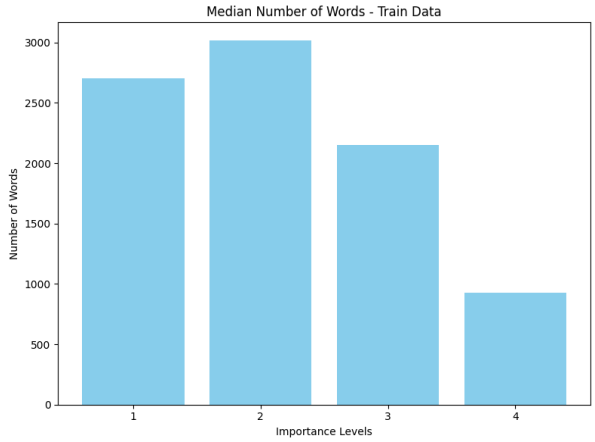
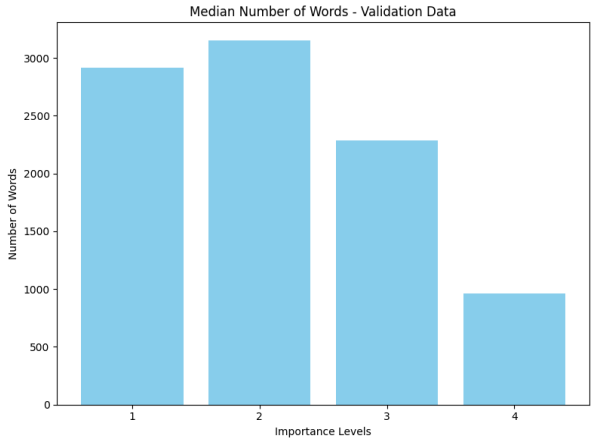
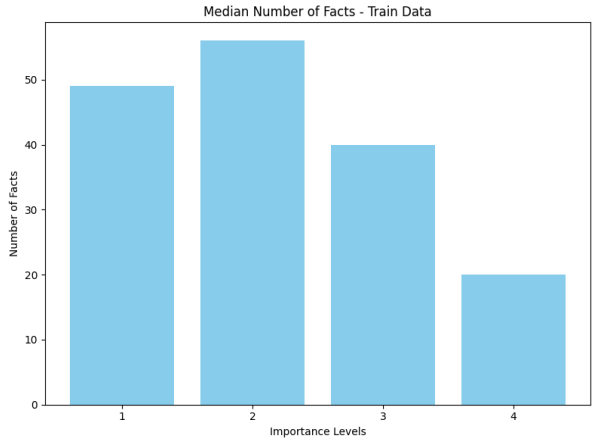
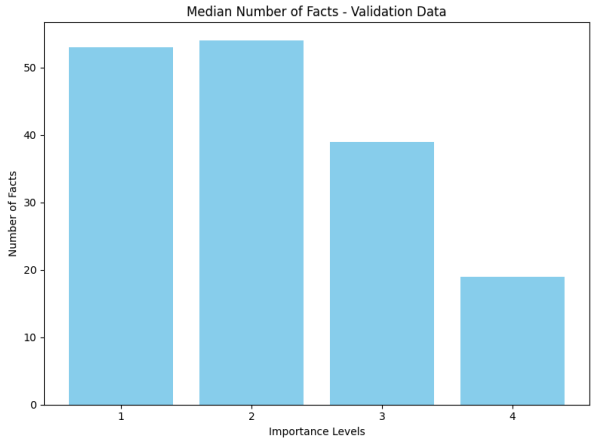
Ο παρακάτω πίνακας και διάγραμμα αφορά το πλήθος των ποινικών υποθέσεων ανά set δεδομένων και ανά τιμή importance

	Train Data	Validation Data	Test Data
1	826	150	120
2	726	136	42
3	1891	429	662
4	3657	665	2174



Στην συνέχεια τα παρακάτω διαγράμματα παρουσιάζουν το διάμεσο για κάθε κλάση όσον αφορά το πλήθος των χαρακτήρων, τον λέξεων και των γεγονότων που έχει η κάθε υπόθεση για το train και validation set.

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Train Set	Validation Set
	
	
	

Πανεπιστήμιο Αιγαίου
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

Προεπεξεργασία Δεδομένων

Τα ECHR δεδομένα πριν χρησιμοποιηθούν ως είσοδος των νευρωνικών δικτύων για την εκπαίδευση τους για την απόδοση μίας πρόβλεψης σχετικά με την κατηγορία του κειμένου υπόκεινται σε μία προεπεξεργασία.

Το πρώτο στάδιο είναι η παραγωγή ενός ενιαίου κειμένου από τα γεγονότα (facts) του παραδείγματος εκπαίδευσης. Τα γεγονότα είναι στο σύνολο των δεδομένων είναι οργανωμένα σε λίστες και όχι σε ένα ενιαίο κείμενο. Τα γεγονότα ενοποιούνται σε ένα ενιαίο κείμενο το οποίο αποτελεί την περιγραφή της δικαστικής υπόθεσης με ελεύθερο κείμενο φυσικής γλώσσας.

Στην συνέχεια ακολουθεί η μεταμόρφωση των χαρακτήρων σε πεζούς έτσι ώστε οι λέξεις να μπορούν να ομαδοποιηθούν. Τα κεφαλαία γράμματα σε δικαστικά έγγραφα εξυπηρετούν γραμματικές λειτουργίες όπως ότι η λέξη βρίσκεται στην αρχή της πρότασης ή ότι είναι κύριο όνομα. Παρόλα αυτά δεν διαφοροποιούν την σημασία της λέξης οπότε σημασιολογικά παραμένει ίδια. Στην συνέχεια αφαιρούνται οι ενδιάμεσες λέξεις (stopwords) οι οποίες δεν έχουν βαρύνουσα σημασιολογική σημασία ως προς το περιεχόμενο του κειμένου έτσι ώστε να μειωθεί το μέγεθος το κειμένου διατηρώντας τις λέξεις με την μεγαλύτερη επιρροή ως προς την απόδοση μίας κλάσης. Σημειώνεται ότι τα πειράματα έγιναν τόσο στα κείμενα που αφαιρέθηκαν οι λειτουργικές λέξεις καθώς επίσης και στα ίδια κείμενα χωρίς την αφαίρεση των ενδιάμεσων λέξεων.

Εφόσον λοιπόν καταλήξουμε στο επεξεργασμένο κείμενο το οποίο θέλουμε να τροφοδοτήσουμε το μοντέλο μηχανικής μάθησης είτε για την εκπαίδευση του είτε ως προς την απόδοση πρόβλεψης ακολουθεί η μετατροπή του σε διάνυσμα μέσω της διαδικασίας του tokenization. Κάθε αλγόριθμος διαθέτει τον δικό του τρόπο με τον οποίο μετατρέπει το κείμενο σε διανύσματα. Έτσι τα διανύσματα αυτά είναι διαφορετικά για κάθε μοντέλο μηχανικής μάθησης αναφερόμενα στο ίδιο κείμενο. Αυτό εξαρτάται από την την μέθοδο διανυσματοποίησης καθώς επίσης και από το σύνολο δεδομένων που έχει εκπαιδευτεί το μοντέλο μηχανικής μάθησης. Τα διανύσματα βοηθούν στην απόδοση και σύγκρισης σημασίας σε επίπεδο μηχανής. Έτσι για το κάθε λέξη αποκτά διαφορετικό νόημα ανάλογα με το είδος του αντικειμένου των εγγράφων στο οποίο έχει εκπαιδευτεί το μοντέλο μηχανικής μάθησης.

Στην συνέχεια αφού όλο το training και validation set έχει μεταμορφωθεί σε διανύσματα εισάγονται στο μοντέλο μηχανικής μάθησης για την εκπαίδευση του για την κατηγοριοποίηση των κειμένων ανάλογα με την κλάση τους. Η εκπαίδευση του μοντέλου γίνεται υπολογίζοντας την ακρίβεια των προβλέψεων, δηλαδή το σύνολο των σωστά προβλεπόμενων κλάσεων ωπς προς το σύνολο των δεδομένων. Με αυτό τον τρόπο το μοντέλο εξειδικεύεται ως προ την συγκεκριμένη εργασία της πρόβλεψης υποθέσεων οι οποίες αποτελούν παράβαση των θεμελιωδών ευρωπαϊκών δικαιωμάτων μέσα από την παράθεση των γεγονότων σε ελεύθερο

κείμενο. Έτσι το μοντέλο μηχανικής μάθησης εκπαιδευτεί ως προς την συγκεκριμένη εργασία του δίνονται με την μορφή μεταμορφωμένων διανυσμάτων τα δεδομένα από το test set.

Έτσι το μοντέλο αποδίδει κατηγοριοποιήσεις σε αυτά τα δεδομένα, οι οποίες συγκρίνονται με τις κατηγορίες των κειμένων. Με αυτό τον τρόπο δημιουργείται μία ανάλυση με τις διάφορες μεθόδους αξιολόγησης. Στην παρούσα εργασία περιλαμβάνονται πρότυπα αξιολόγησης ως προς το σύνολο του test set αλλά και ως προς την κάθε κλάση ξεχωριστά.

Βελτιστοποίηση των μοντέλων μηχανικής μάθησης

Το περιβάλλον στο οποίο εκπαιδεύτηκαν τα νευρωνικά δίκτυα είναι το περιβάλλον νέφους της Amazon (AWS). Οι πόροι για την χρησιμοποίηση των εργαλείων και των πόρων του AWS χορηγήθηκαν με την υποστήριξη της ΕΔΥΤΕ (https://grnet.gr/services/cloud-services/aws/). Πιο συγκεκριμένα χρησιμοποιήθηκαν οι εικονικοί πόροι ml.g5.8xlarge οι οποίοι έχουν τα κάτωθι χαρακτηριστικά

Single GPU VMs	Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)
	g5.8xlarge	1	24	32	128	1x900	25	16

Τα μοντέλα βελτιστοποιήθηκαν σύμφωνα με της κάτωθι παραμέτρους χρησιμοποιώντας ως metric για την μέτρηση της επίδοσης το accuracy.

learning_rate=2e-5,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
num_train_epochs=2,
weight_decay=0.01,
evaluation_strategy="epoch"

Οι ανωτέρω παράμετροι προτείνονται από το αποθετήριο μοντέλων μηχανικής μάθησης “transformers” το οποίο χρησιμοποιείται στην γλώσσα προγραμματισμού Python (https://huggingface.co/docs/transformers/en/tasks/sequence_classification). Από το

συγκεκριμένο αποθετήριο ελήφθησαν τα προεκπαιδευμένα μοντέλα μηχανικής μάθησης για την εκπαίδευση τους και την διενέργεια προβλέψεων.

Μετρικές Αξιολόγησης

Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν στην παρούσα εργασία είναι το macro average precision, το recall και το F1-Score. Οι συγκεκριμένες μετρικές αξιολόγησης χρησιμοποιήθηκαν στο (Chalkidis et al., 2019, 3) και θα μπορέσουν να συγκριθούν τα αποτελέσματα των μοντέλων της προαναφερόμενης εργασίας με αυτά της παρούσας εργασίας.

Το macro average precision δίνει ίση βαρύτητα σε κάθε κατηγορία, γεγονός που το καθιστά ανθεκτικό στην περίπτωση που έχουμε ανισορροπία μεταξύ των κατηγοριών. Διασφαλίζει ότι κάθε αποτέλεσμα αντιμετωπίζεται με την ίδια σημασία, ανεξάρτητα από τη συχνότητά του στο σύνολο δεδομένων. Στο συγκεκριμένο σύνολο δεδομένων στο test σετ υπάρχει ισορροπία μεταξύ των κλάσεων τόσο ως προς την κλάση του importance όσο και ως προς την κλάση του violated ή non_violated. Πέρα όμως από την macro διάσταση των συγκεκριμένων μετρικών υπολογίστηκε επίσης και η micro διάσταση η οποία δίνει αντίστοιχο βάρος στην μετρική κάθε κλάσης όση και η συχνότητα εμφάνισης της στο σύνολο δεδομένων.

Αποτελέσματα Έρευνας Ανάλυση και Συζήτηση

Αποτελέσματα Πρόβλεψης Δικαστικής Απόφασης

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα αξιολόγησης των γλωσσικών μοντέλων (LLM) που δοκιμάστηκαν στην πρόβλεψης της δικαστικής απόφασης για την αξιολόγηση της απόδοσής τους σε μια σειρά μετρήσεων. Τα LLM που αξιολογήθηκαν σε αυτήν τη μελέτη περιλαμβάνουν medium-BERT, ALBERT, RoBERTa, LongFormer, BigBird και αξιολογούνται χρησιμοποιώντας τυπικές μετρήσεις απόδοσης, όπως accuracy, precision, recall, και F1 score και άλλες τόσο σε macro όσο και σε micro επίπεδο.

Από την παρακάτω παράθεση των αποτελεσμάτων προκύπτει το στο accuracy σε micro επίπεδο το medium bert πέτυχε τα καλύτερα αποτελέσματα με 87 %. Παρόλα αυτά δεν είχε μεγάλη διαφορά από τα αμέσως επόμενα μοντέλα τα οποία στην πλειοψηφία τους πέτυχαν accuracy 86% εκτός από το distilBERT το οποίο πέτυχε 83%. Από την επίδοση των μοντέλουν μπορούμε να παρατηρήσουμε ότι τα μοντέλα μπόρεσαν να αντιληφθούν δομές της γλώσσας μέσα στα κείμενα τα οποία σχετίζονται με το αποτέλεσμα της δίκης σε VIOLATED ή NON_VIOLATED. Επίσης το mediumBERT πέτυχε εξαιρετικά αποτελέσματα στο precision και στην κλάση NON_VIOLATED πράγμα το οποίο υποδεικνύει την βεβαιότητα την πρόβλεψης όταν

πρόκειται για την κλάση NON_VIOLATED. Στην πρόβλεψη της κλάσης VIOLATED όλα τα μοντέλα είχαν τον ίδιο βαθμό επιτυχίας 83% στο precision. Επίσης το macro average του f1 score όλων το μοντέλων ήταν στο 83% πράγμα το οποίο δείχνει ότι τα μοντέλα έχουν την δυνατότητα να αναγνωρίζουν και τις δύο κλάσεις, λαμβάνοντας υπόψη ότι στο test set των δεδομένων υπήρχε ανισορροπία των κλάσεων.

Τα παρακάτω αποτελέσματα αναφέρονται στην πρόβλεψη των νομικών κειμένων από τα οποία έχουν αφαιρεθεί οι ενδιάμεσες λέξεις (stopwords).

Model		Precision	Recall	f1-score	support
albert	NON_VIOLATED	0.95	0.62	0.75	1024
	VIOLATED	**0.83**	0.98	0.90	1974
	accuracy			0.86	2998
	macro avg	0.89	0.80	**0.83**	2998
	weighted avg	0.87	0.86	**0.85**	2998
distil bert	NON_VIOLATED	0.81	0.63	0.71	1024
	VIOLATED	**0.83**	0.92	0.87	1974
	accuracy			0.82	2998
	macro avg	0.82	0.78	0.79	2998
	weighted avg	0.82	0.82	0.82	2998
roberta	NON_VIOLATED	0.98	0.61	0.75	1024
	VIOLATED	**0.83**	0.99	**0.91**	1974
	accuracy			0.86	2998
	macro avg	**0.91**	**0.80**	**0.83**	2998
	weighted avg	**0.88**	0.86	**0.85**	2998
longformer	NON_VIOLATED	0.94	0.62	0.75	1024
	VIOLATED	**0.83**	0.98	0.90	1974
	accuracy			0.86	2998
	macro avg	0.89	**0.80**	**0.83**	2998

	weighted avg	0.87	**0.86**	0.85	2998
bigbird	NON_VIOLATED	0.96	0.62	0.75	1024
	VIOLATED	**0.83**	0.99	0.90	1974
	accuracy			0.86	2998
	macro avg	0.90	**0.80**	**0.83**	2998
	weighted avg	0.88	0.86	**0.85**	2998
Medium bert	NON_VIOLATED	**0.99**	0.62	0.76	1024
	VIOLATED	**0.83**	**1.00**	0.91	1974
	accuracy			**0.87**	2998
	macro avg	0.91	0.81	**0.83**	2998
	weighted avg	0.89	**0.87**	**0.86**	2998

Εν συνεχεία τα παρακάτω αποτελέσματα αναφέρονται στα κείμενα από τα οποία δεν έχουν αφαιρεθεί οι ενδιάμεσες λέξεις (stopwords)

Model		Precision	Recall	f1-score	support
albert	NON_VIOLATED	0.96	**0.62**	0.75	1024
	VIOLATED	0.83	0.99	0.90	1974
	accuracy			0.86	2998
	macro avg	0.89	0.80	**0.83**	2998
	weighted avg	0.88	0.86	0.85	2998
distil bert	NON_VIOLATED	0.84	**0.62**	0.71	1024
	VIOLATED	0.83	0.94	0.88	1974
	accuracy			0.83	2998
	macro avg	0.83	0.78	0.79	2998
	weighted avg	0.83	0.83	0.82	2998
roberta	NON_VIOLATED	0.87	**0.67**	0.75	1024

	VIOLATED	**0.85**	0.95	0.89	1974
	accuracy			0.85	2998
	macro avg	0.86	0.81	0.82	2998
	weighted avg	0.85	0.85	0.85	2998
longformer	NON_VIOLATED	0.98	0.61	**0.76**	1024
	VIOLATED	0.83	0.99	**0.91**	1974
	accuracy			0.86	2998
	macro avg	**0.91**	0.80	**0.83**	2998
	weighted avg	0.88	0.86	0.85	2998
bigbird	NON_VIOLATED	**0.99**	0.62	**0.76**	1024
	VIOLATED	0.83	**1.00**	**0.91**	1974
	accuracy			**0.87**	2998
	macro avg	**0.91**	**0.81**	**0.83**	2998
	weighted avg	**0.89**	**0.87**	**0.86**	2998
Medium bert	NON_VIOLATED	0.94	0.62	0.75	1024
	VIOLATED	0.83	0.98	0.90	1974
	accuracy			0.86	2998
	macro avg	0.89	0.80	**0.83**	2998
	weighted avg	0.87	0.86	0.85	2998

Αποτελέσματα Πρόβλεψης Σπουδαιότητας Υπόθεσης

Στην προσπάθεια επίλυσης του προβλήματος πρόβλεψης της σπουδαιότητας της δικαστικής υπόθεσης τα αποτελέσματα δεν ήταν τόσο ενθαρρυντικά όσο στο προηγούμενο πρόβλημα. Παρόλο που βλέπουμε πολύ υψηλά νούμερα στο accuracy (πάνω από 90 %), υπάρχει μία αδυναμία των μοντέλων να ξεχωρίσουν τις δύο κλάσεις. Στα δεδομένα υπάρχει μεγάλη ανισορροπία πράγμα το οποίο σημαίνει ότι το macro accuracy δεν μπορεί να αποτελέσει ασφαλής τρόπος για την εξαγωγή συμπερασμάτων.Παρατηρώντας το precision στην κλάση IMPORTANT βλέπουμε ότι τα μοντέλα παρουσιάζουν ιδιαίτερη αδυναμία στην επιτυχή

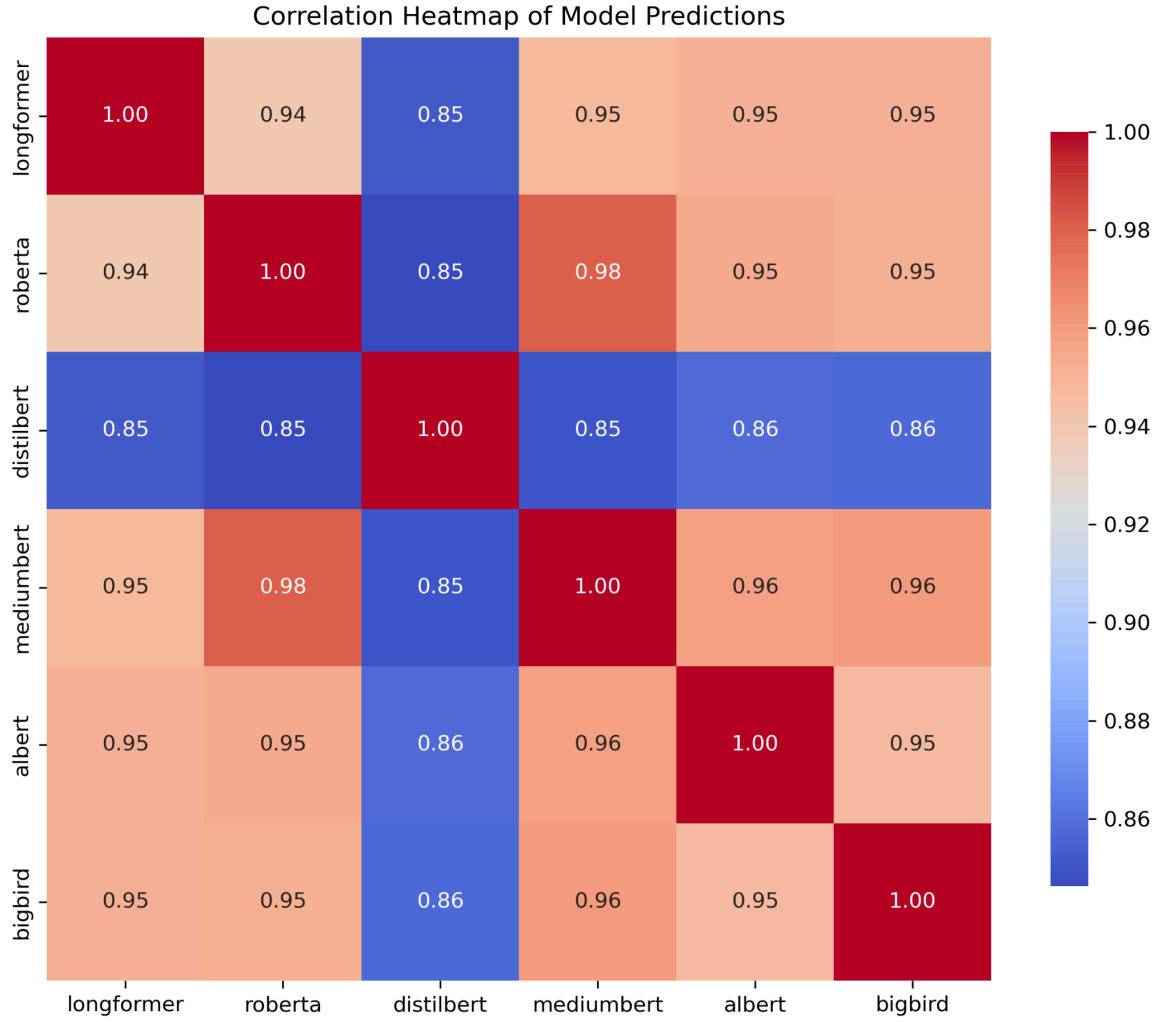
πρόβλεψη της συγκεκριμένης κλάσης. Εξαίρεση σε αυτόν τον κανόνα αποτελεί το longformer το οποίο επιτυγχάνει precision 67% σε αυτή την κλάση. Παρόλα αυτά εξετάζοντας το macro f1 score παρατηρούμε ότι καλύτερη επίδοση είχε το distil BERT με 60% πράγμα το οποίο υποδεικνύει την καλύτερη ικανότητα του στην διάκριση των κλάσεων. Όπως αναφέρθηκε και στην προηγούμενη ενότητα η πρόβλεψη της σπουδαιότητας ενός κειμένου είναι ιδιαίτερα σύνθετο αντικείμενο και δεν επαφύεται αποκλειστικά στην περιγραφή του κειμένου σε ελεύθερο κείμενο. Η κατηγοριοποίηση των σπουδαιότητας απαιτεί εμπειρία πάνω στον τομέα της νομικής καθώς επίσης και εξειδικευμένες γνώσεις, πράγμα το οποίο κάνει το πρόβλημα πιο πολύπλοκο. Αυτή η πολυπλοκότητα δεν μπορεί να εξαχθεί από τα παρόντα μοντέλα στις δομές του περιγραφικού κειμένου της υπόθεσης.

Model		Precision	Recall	f1-score	support
albert	IMPORTANT	0.09	0.88	0.16	162
	NOTIMPORTANT	0.98	0.46	0.63	2836
	accuracy			0.49	2998
	macro avg	0.54	0.67	0.39	2998
	weighted avg	0.94	0.49	0.60	2998
distil bert	IMPORTANT	0.36	0.17	0.23	162
	NOTIMPORTANT	0.95	0.98	0.97	2836
	accuracy			0.94	2998
	macro avg	0.65	0.57	0.60	2998
	weighted avg	0.92	0.94	0.93	2998
roberta	IMPORTANT	0.22	0.12	0.16	162
	NOTIMPORTANT	0.95	0.97	0.96	2836
	accuracy			0.93	2998
	macro avg	0.58	0.55	0.56	2998
	weighted avg	0.91	0.93	0.92	2998
longformer	IMPORTANT	0.67	0.09	0.15	162
	NOTIMPORTANT	0.95	1.00	0.97	2836
	accuracy			0.95	2998

	macro avg	0.81	0.54	0.56	2998
	weighted avg	0.93	0.95	0.93	2998
bigbird	IMPORTANT	0.38	0.09	0.14	162
	NOTIMPORTANT	0.95	0.99	0.97	2836
	accuracy			0.94	2998
	macro avg	0.66	0.54	0.56	2998
	weighted avg	0.92	0.94	0.93	2998
Medium bert	IMPORTANT	0.75	0.02	0.04	162
	NOTIMPORTANT	0.95	1.00	0.97	2836
	accuracy			0.95	2998
	macro avg	0.85	0.51	0.50	2998
	weighted avg	0.94	0.95	0.92	2998

Σύγκριση Αποτελεσμάτων Μοντέλων

Από την εξέταση των αποτελεσμάτων των πειραμάτων των μοντέλων μηχανικής μάθησης προέκυψε ότι τα μοντέλα πραγματοποιούν σε μεγάλο ποσοστό τα ίδια λάθη στις προβλέψεις όπως φαίνεται στην παρακάτω εικόνα.



Με εξαίρεση το distilbert τα λάθη ανάμεσα στα υπόλοιπα μοντέλα τα αποτελέσματα στα υπόλοιπα μοντέλα ταιριάζουν πάνω από 90 τοις εκατό. Τα αποτελέσματα του distilbert ταιριάζουν πάνω από 80 τοις εκατό με τα αποτελέσματα των υπόλοιπων μοντέλων.

Όπως προέκυψε από την ανάλυση των λαθών των μοντέλων η συντριπτική πλειοψηφία των

	bigbird_errors	distilbert_errors	albert_errors	roberta_errors	medium_errors	longformer_errors
NON_VIOLATED	93.01 %	70.68 %	92.36 %	97.56 %	98.24 %	90.8 %
VIOLATED	6.99 %	29.32 %	7.64 %	2.44 %	1.76 %	9.2 %

λαθών έγινε στην κλάση NON_VIOLATED

Το μεγαλύτερο ποσοστό των λαθών ανήκουν στην χρονολογία 2016 και 2017, ενώ το μικρότερο μέρος των λαθών στο 2018

Επιπλέον η συντριπτική πλειοψηφία των λαθών ανήκει στην κατηγορία σημαντικότητας 3 και τέσσερα

	bigbird_errors	distilbert_errors	albert_errors	roberta_errors	medium_errors	longformer_errors
1	9.4 %	8.08 %	9.55 %	10.24 %	10.3 %	10.61 %
2	2.41 %	2.26 %	2.15 %	1.95 %	2.01 %	2.36 %
3	37.83 %	29.51 %	37.95 %	37.8 %	37.69 %	35.85 %
4	50.36 %	60.15 %	50.36 %	50.0 %	50.0 %	51.18 %

Το μεγαλύτερο μέρος των λαθών έγινε σε κείμενα με μεγάλο αριθμό γεγονότων (facts). Ο μεγάλος αριθμός γεγονότων ορίζεται ως αριθμός των γεγονότων μεγαλύτερου του διαμέσου του συνόλου.

	bigbird_errors	distilbert_errors	albert_errors	roberta_errors	medium_errors	longformer_errors
HIGH	59.52 %	47.18 %	58.71 %	60.0 %	59.05 %	58.96 %
LOW	40.48 %	52.82 %	41.29 %	40.0 %	40.95 %	41.04 %

Το μεγαλύτερο μέρος των λαθών έγινε σε κείμενα με μεγάλο αριθμό λέξεων. Ο μεγάλος αριθμός λέξεων ορίζεται ως αριθμός των λέξεων μεγαλύτερου του διαμέσου του συνόλου.

	bigbird_errors	distilbert_errors	albert_errors	roberta_errors	medium_errors	longformer_errors
HIGH	66.75 %	52.82 %	67.06 %	67.8 %	67.59 %	66.27 %
LOW	33.25 %	47.18 %	32.94 %	32.2 %	32.41 %	33.73 %

Το μεγαλύτερο μέρος των λαθών έγινε σε κείμενα με μεγάλο αριθμό χαρακτήρων. Ο μεγάλος αριθμός χαρακτήρων ορίζεται ως αριθμός των χαρακτήρων μεγαλύτερου του διαμέσου του συνόλου.

	bigbird_errors	distilbert_errors	albert_errors	roberta_errors	medium_errors	longformer_errors
HIGH	66.75 %	52.63 %	66.83 %	67.8 %	67.59 %	66.04 %
LOW	33.25 %	47.37 %	33.17 %	32.2 %	32.41 %	33.96 %

Συμπεράσματα

Τα αποτελέσματα από τα πειράματα δείχνουν ότι τα μοντέλα mediumBERT, DistilBERT, RoBERTa, ALBERT, Longformer και BigBird, ήταν εξαιρετικά αποτελεσματικά στην πρόβλεψη του αποτελέσματος της δικαστικής απόφασης, επιτυγχάνοντας ακρίβειες και βαθμολογίες F1 πάνω από 85%. Αυτό υποδηλώνει ότι αυτά τα μοντέλα είναι σε θέση να συλλάβουν τα σχετικά πρότυπα και τα χαρακτηριστικά που είναι απαραίτητα για την πρόβλεψη των δικαστικών αποτελεσμάτων με υψηλό βαθμό ακρίβειας, ιδιαίτερα σε δομημένα ή καλά καθορισμένα νομικά κείμενα. Επιπλέον η πλειονότητα των μοντέλων, με εξαίρεση το medium-BERT και το RoBERTa απέδωσε καλύτερα σε κείμενα τα οποία δεν είχαν προεπεξεργαστεί με την αφαίρεση των stopwords. Γίνεται μνεία ότι την καλύτερη επίδοση πέτυχε το μοντέλο BigBird το οποίο βελτιστοποιήθηκε σε κείμενα όπου δεν αφαιρέθηκαν οι ενδιάμεσες λέξεις.

Ωστόσο, τα μοντέλα δυσκολεύτηκαν όταν επιφορτίστηκαν με την πρόβλεψη της σημασίας μιας υπόθεσης που βασίζεται σε περιγραφές ελεύθερου κειμένου. Αυτό δείχνει ότι ενώ οι LLM υπερέχουν στην πρόβλεψη του αποτελέσματος όταν τους παρέχονται σαφείς και δομημένες πληροφορίες. Παρόλα αυτά αντιμετωπίζουν προκλήσεις στην αξιολόγηση αφηρημένων εννοιών, όπως η σημασία μιας υπόθεσης, η οποία πιθανότατα απαιτεί βαθύτερη κατανόηση των νομικών αρχών, των προηγούμενων υποθέσεων και των συμφραζόμενων αποχρώσεων. Η αδυναμία ακριβούς κατάταξης ή πρόβλεψης της σημασίας περίπτωσης υποδηλώνει ότι πρόσθετες τεχνικές, όπως η εκπαίδευση σε συγκεκριμένο τομέα ή η ενσωμάτωση εξωτερικών πηγών γνώσης, μπορεί να είναι απαραίτητες για τη βελτίωση της απόδοσης σε αυτόν τον τομέα.

Οι προβλέψεις τους ήταν επίσης σε μεγάλο βαθμό πανομοιότυπες ενώ τα μοντέλα υιοθετούν διαφορετικές τεχνικές προσοχής στο κείμενο και διαφέρουν στο μέγεθος κειμένου που μπορούν να καταναλώσουν. Αυτή η συνέπεια υποδηλώνει ότι τα μοντέλα καταγράφουν παρόμοια μοτίβα και χαρακτηριστικά στο νομικό κείμενο και ότι το έργο της πρόβλεψης των δικαστικών αποτελεσμάτων μπορεί να είναι αρκετά καλά καθορισμένο ώστε αυτά τα μοντέλα που βασίζονται σε transformer αρχιτεκτονικές να συγκλίνουν στα ίδια αποτελέσματα. Επιπλέον το μοντέλο RoBERTa απέδωσε πολύ καλύτερα στην κλάση VIOLATED όταν εκπαιδεύτηκε σε κείμενα χωρίς την αφαίρεση ενδιάμεσων λέξεων παρουσιάζοντας καλύτερη ισορροπία ανάμεσα στις δύο κλάσεις παρόλο που η συνολική του απόδοση ήταν χαμηλότερη συγκριτικά με την έκδοση του εκπαιδευμένο σε κείμενα με την αφαίρεση των ενδιάμεσων λέξεων.

Περιορισμοί

Η παρούσα εργασία πραγματοποιήθηκε με πόρους σε περιβάλλον νέφους της Amazon (AWS). Οι πόροι για την χρησιμοποίηση των εργαλείων και των πόρων του AWS χορηγήθηκαν

με την υποστήριξη της ΕΔΥΤΕ (https://grnet.gr/services/cloud-services/aws/). Παρόλη την προαναφερόμενη χρημοτοδότηση δεν ήταν δυνατή η χρήση μεγαλύτερων LLM (π.χ. mistral) λόγω του ότι απαιτούσαν επιπλέον κόστος για την χρήση τους. Επιπλέον η ανάπτυξη μοντέλων πολύ μεγαλύτερων δεν μπορούσε να γίνει στους παρόντες πόρους που χρησιμοποιήθηκαν έτσι ώστε η σύγκριση των LLM μοντέλων να περιλαμβάνει περισσότερα μοντέλα. Αυτό κόστισε στον μεγαλύτερο πλουραλισμό των μοντέλων που θα μπορούσαν να χρησιμοποιηθούν.

Επιπλέον όπως αναφέρει στην εργασία της η (Medvedeva & Mcbride, 2023) τα συστήματα πρόβλεψης βασίζονται κυρίως σε γεγονότα υποθέσεων που εξάγονται από αποφάσεις. Συγκεκριμένα, αυτά τα μοντέλα μηχανικής μάθησης τροφοδοτούνται από το κείμενο της ενότητας «γεγονότων» των υποθέσεων (και μερικές φορές και το σκεπτικό του δικαστηρίου) ως στοιχεία εισόδου μαζί με τις αντίστοιχες αποφάσεις (π.χ. παραβίαση/καμία παραβίαση άρθρου της Ευρωπαϊκής Σύμβασης για Ανθρώπινα Δικαιώματα) ως ετικέτες. Τα συστήματα είναι εκπαιδευμένα να «προβλέπουν» αποφάσεις, μαθαίνουν μοτίβα μέσα στην είσοδο που αντιστοιχούν σε συγκεκριμένες ετικέτες. Στη συνέχεια αξιολογούνται σε ένα σετ δοκιμών (δηλαδή «αόρατα» περιστατικά περιστατικών που δεν συμπεριλήφθηκαν στο σετ εκπαίδευσης). Τα «γεγονότα» που αποτελούν το σύνολο δοκιμών, όπως αυτά του σετ εκπαίδευσης, εξάγονται από δημοσιευμένες τελικές κρίσεις. Οι ετικέτες που χρησιμοποιούνται για την αξιολόγηση της απόδοσης των συστημάτων στο δοκιμαστικό σύνολο μπορούν επίσης να βρεθούν σε αυτές τις κρίσεις. Παρόλο που τέτοια συστήματα μπορεί να επιτύχουν υψηλή απόδοση, δεν δοκιμάζονται ποτέ σε «πραγματικά» δεδομένα, δηλαδή δεδομένα που είναι πραγματικά διαθέσιμα σε δυνητικούς τελικούς χρήστες πριν ληφθεί η κρίση ή εκδοθεί. Ο πραγματικός χρήστης αυτών των συστημάτων (π.χ. ο δικηγόρος ενός διαδίκου) δεν έχει πρόσβαση στην ίδια διατύπωση των γεγονότων που είναι διαθέσιμη και αναφέρεται στην απόφαση. Αυτή η διατύπωση δεν μπορεί να υπάρξει μέχρι να εκδοθεί η απόφαση. Έτσι τα μοντέλα εκπαιδεύονται σε κείμενα τα οποία περιγράφουν το σκεπτικό των δικαστών και κατ' επέκταση οδηγούν στην απόδοση της απόφασης.

Αυτό το κείμενο είναι το τελικό συμπέρασμα από την μελέτη ολόκληρου του φακέλου της υπόθεσης. Αυτό έχει σαν αποτέλεσμα να δημιουργείται μία ψευδαίσθηση για καλή απόδοση των μοντέλων μηχανικής μάθησης καθώς μαθαίνουν τις δικαστικές υποθέσεις από κείμενα τα οποία είναι κατασκευασμένα να δικαιολογήσουν την τελική απόφαση. Αυτό σημαίνει ότι η απόφαση έχει ήδη παρθεί. Η πραγματική πρόκληση είναι η εκπαίδευση του μοντέλου πάνω στην ανεπηρέαστη παράθεση γεγονότων που γίνεται στον φάκελο της υπόθεσης. Αυτό σημαίνει ότι τα μοντέλα μηχανικής μάθησης πρέπει να τροφοδοτηθούν με πολύ μεγάλο όγκο

πληροφορίας δεδομένου ότι ένας φάκελος μπορεί να περιέχει χιλιάδες σελίδες κειμένου. Σκοπός της προσπάθειας πρόβλεψης της δικαστικής απόφασης μέσω μηχανικής μάθησης πρέπει να είναι η παρόμοια με διαδικασία του δικαστή. Η μελέτη δηλαδή του φακέλου με σκοπό την εκπόνηση των γεγονότων αυτών που οδηγούν στην λήψη μίας συγκεκριμένης απόφασης.

Πηγές

Advantages and Disadvantages of Bag of Words (examples, video). (2023, March 29). AIML.com. Retrieved October 9, 2023, from https://aiml.com/what-are-the-advantages-and-disadvantages-of-bag-of-words-model/

Aletras, N., & Chalkidis, I. (2019, 06 02). *ECHR Dataset - ACL 2019*. Internet Archive. Retrieved 06 15, 2024, from https://archive.org/details/ECHR-ACL2019

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Publishing*, *2*(e93), 19. https://doi.org/10.7717/peerj-cs.93

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016, 10 24). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Comput. Sci*. https://doi.org/10.7717/peerj-cs.93

Bag-of-words model. (n.d.). Wikipedia. Retrieved October 9, 2023, from https://en.wikipedia.org/wiki/Bag-of-words_model

Beltagy, I., Peters, M. E., & Cohan, A. (2020, 12 2). Longformer: The Long-Document Transformer. *arxiV*, *2*(-), 1-17. https://doi.org/10.48550/arXiv.2004.05150

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arxiv*, *1*, 1-7. https://doi.org/10.48550/arXiv.1607.04606

Cai-zhi, L., Yan-xiu, S., Zhi-qiang, W., & Yong-Quan, Y. (2018). "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent

Robotic and Control Engineering (IRCE). *IEEE*, 218-222. 978-1-5386-7417-8. 10.1109/IRCE.2018.8492945

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural Legal Judgment Prediction in English. *arxiv*, 7. https://doi.org/10.48550/arXiv.1906.02059

GloVe Explained. (n.d.). Papers With Code. Retrieved October 13, 2023, from https://paperswithcode.com/method/glove

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arxiv*, 5. https://doi.org/10.48550/arXiv.1402.3722

Han Kyul, K., Hyunjoong, K., & Sungzoon, C. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *sciencedirect*, *266*, 336-352. https://doi.org/10.1016/j.neucom.2017.05.046.

Hui, J. (2019, October 21). *NLP — Word Embedding & GloVe. BERT is a major milestone in creating… | by Jonathan Hui | Medium*. Jonathan Hui. Retrieved October 13, 2023, from https://jonathan-hui.medium.com/nlp-word-embedding-glove-5e7f523999f6

Kulshrestha, R. (2019, November 24). *NLP 101: Word2Vec — Skip-gram and CBOW | by Ria Kulshrestha*. Towards Data Science. Retrieved October 13, 2023, from https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arxiV*, *6*(-), 1-16. https://doi.org/10.48550/arXiv.1909.11942

Liu, Z., Lin, Y., & Sun, M. (Eds.). (2023). *Representation Learning for Natural Language Processing*. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1600-9

Medvedeva, M., & Mcbride, P. (2023, 12). Legal Judgment Prediction: If You Are Going to Do It, Do It Right. *Proceedings of the Natural Legal Language Processing Workshop 2023*, *1*(-). 10.18653/v1/2023.nllp-1.9

Medvedeva, M., Üstün, A., Xu, X., Vols, M., & Wieling, M. (2021). Automatic judgement forecasting for pending applications of the European Court of Human Rights.

Proceedings of the Fifth Workshop on Automatec Semantic Analysis of Information in Legal Text (ASAIL 2021), 12-23. https://www.researchgate.net/publication/352742175_Automatic_Judgement_Forecasting_for_Pending_Applications_of_the_European_Court_of_Human_Rights

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*, 237-266. https://doi.org/10.1007/s10506-019-09255-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arxiv*, 12. https://doi.org/10.48550/arXiv.1301.3781

Niklaus, J., Chalkidis, I., & Stürmer, M. (2021). Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. *In Proceedings of the Natural Legal Language Processing Workshop 2021*, -(-), 19-35. 10.18653/v1/2021.nllp-1.3

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *ACL Anthology*, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018). Dissecting Contextual Word Embeddings: Architecture and Representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1499–1509. https://doi.org/10.48550/arXiv.1808.08949

Salton, G. M., & Yu, C. T. (1974). On the construction of effective vocabularies for information retrieval. *ACM SIGIR Forum*, *9*(3), 48-60. https://doi.org/10.1145/951761.951766

Shirsat, K., Keni, A., Chavan, P., & Gosavi, M. (2021, 05). Legal Judgement Prediction System. *International Research Journal of Engineering and Technology (IRJET)*, *8*(5), 734-738. https://d1wqtxts1xzle7.cloudfront.net/70098361/IRJET_V8I5149-libre.pdf?1632295283=&response-content-disposition=inline%3B+filename%3DIRJET_Legal_Judgement_Prediction_System.pdf&Expires=1727003290&Signature=AWFZF5iBUX1HoiB-JABsn7qDN5JyL~vUFdJIOiF428cunGMck-

Strickson, B., & Iglesia, B. (2020, 04 20). Legal Judgement Prediction for UK Courts. *ICISS '20: Proceedings of the 3rd International Conference on Information Science and Systems*, 204-209. https://doi.org/10.1145/3388176.3388183

Turan, T., Küçüksille, E., & Kemaloğlu Alagöz, N. (2023, 9 30). Prediction of Turkish Constitutional Court Decisions with Explainable Artificial Intelligence. *Dergi Park Akademik*, *7*(2), 128-141. https://doi.org/10.30516/bilgesci.1317525

Uzila, A. (2022, September 11). *GloVe and fastText Clearly Explained: Extracting Features from Text Data*. Level Up Coding. Retrieved October 13, 2023, from https://levelup.gitconnected.com/glove-and-fasttext-clearly-explained-extracting-features-from-text-data-1d227ab017b2

Wisam, Q. A., Musa, A. M., & Bilal, A. I. (2019). An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. *International Engineering Conference (IEC)*, 200-204. 10.1109/IEC47844.2019.8950616

Word2Vec Explained. (2017, March 23). Hacker's Blog. Retrieved October 10, 2023, from https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

Xin, R. (2016). word2vec Parameter Learning Explained. *arxiv*, *4*, 1-21. https://doi.org/10.48550/arXiv.1411.2738 Focus to learn more

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020, 12 6). Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, *1*(-), 17283 - 17297. https://doi.org/10.48550/arXiv.2007.14062

Zellig, H. S. (1954). *Distributional Structure* (10th ed., Vol. 2-3). WORD. https://doi.org/10.1080/00437956.1954.11659520