ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗ

ΚΑΤΕΥΘΥΝΣΗ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΙΣΤΟΥ

ADVERSARIAL MACHINE LEARNING ATTACKS IN TEXT DATA

ΕΠΙΘΕΣΕΙΣ ΣΕ ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΠΟΥ ΔΙΑΧΕΙΡΙΖΟΝΤΑΙ ΚΕΙΜΕΝΙΚΑ ΔΕΔΟΜΕΝΑ

Μεταπτυχιακή εργασία

ΘΕΟΔΩΡΑ ΑΝΑΣΤΑΣΙΟΥ

Αθήνα, Φεβρουάριος 2021

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Τριμελής Εξεταστική Επιτροπή

Ριζομυλιώτης Παναγιώτης
Επίκουρος Καθηγητής,
Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Καραγιώργου Σοφία
Επισκέπτης Καθηγητής,
Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Δίου Χρήστος
Επίκουρος Καθηγητής,
Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Η Θεοδώρα Αναστασίου δηλώνω υπεύθυνα ότι:

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

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

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στην Καθηγήτρια Σοφία Καραγιώργου, επιβλέπουσα της παρούσας διπλωματικής εργασίας, για την καθοδήγηση, την υποστήριξη, την συνεργασία και την πολύτιμη συμβολή της κατά την εκπόνηση της εργασίας αλλά και για τον χρόνο που αφιέρωσε για τη συνολική εποπτεία της. Επιπλέον ευχαριστώ ιδιαιτέρως την οικογένειά μου και τον Κυριάκο.Χ για την υποστήριξη και την αγάπη τους.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

Καθώς οι τεχνολογίες της Μηχανικής Μάθησης και της Τεχνητής Νοημοσύνης αναπτύσσονται και εφαρμόζονται σε νέα συστήματα, αυξάνονται και οι ανησυχίες για την ευαισθησία και την διασφάλιση της απόδοση αυτών των εφαρμογών. Στην παρούσα διπλωματική εργασία, αναφέρονται κίνδυνοι και επιθέσεις που μπορούν να αποδυναμώσουν τα μοντέλα μηχανικής μάθησης όπως και τρόποι αντιμετώπισης ή προφύλαξης των μοντέλων αυτών. Συγκεκριμένα παρουσιάζονται επιθέσεις σε μοντέλα που διαχειρίζονται κείμενα, οι οποίες χωρίζονται σε δύο κύριες κατηγορίες. Η πρώτη κατηγορία αφορά επιθέσεις που επηρεάζουν την οπτική διαφοροποίηση των κειμένων, όπως είναι μία επίθεση σε επίπεδο χαρακτήρα των λέξεων, ενώ η δεύτερη κατηγορία αφορά επιθέσεις που επηρεάζουν την εννοιολογική τους σημασία. ‘Επειτα, παρουσιάζονται οι αντίστοιχοι μηχανισμοί άμυνας της κάθε επίθεσης, για την εξασφάλιση της προστασίας και της διατήρησης της απόδοσης των μοντέλων μηχανικής μάθησης όσο το δυνατόν περισσότερο. Ακολουθεί η εφαρμογή των τεχνολογιών μηχανικής μάθησης σε δύο διαφορετικά σενάρια, με το πρώτο να αφορά επιθέσεις σε επίπεδο χαρακτήρα, αντικρούωντας τις, με τον αντίστοιχο μηχανισμό άμυνας. Για την ανάδειξη της συγκεκριμένης επίθεσης χρησιμοποιείται το εργαλείο TextAttack, το οποίο τροποποιεί τα δεδομένα που εισέρχονται στο μοντέλο σαν δεδομένα εκπαίδευσης, ενώ ως μηχανισμός άμυνας χρησιμοποιείται το ScRNN. Στο συγκεκριμένο σενάριο γίνεται χρήση των προ εκπαιδευμένων μοντέλων BERT και DistilBert. Παράλληλα το δεύτερο σενάριο αφορά σε επίπεδο σημασιολογίας. Για την ανάδειξη της συγκεκριμένης επίθεσης χρησιμοποιείται το δίκτυο WordNet το οποίο παρέχει επίσης το εργαλείο TextAttack. Ταυτόχρονα χρησιμοποιείται και η επίθεση IGA, η οποία τροποποιεί τα δεδομένα που εισέρχονται στο μοντέλο ως δεδομένα ελέγχου, ενώ ο αντίστοιχος μηχανισμός άμυνας που επιλέχθηκε είναι ο SEM. Στο συγκεκριμένο σενάριο γίνεται χρήση των μοντέλων CNN και RNN. Ο σκοπός είναι η αξιολόγηση και η σύγκριση των αποδόσεων των μοντέλων μέσω διαφορετικών τεχνικών. Η αξιολόγηση αυτή έγινε σε τρεις διαφορετικές φάσεις : πριν από την επίθεση, μετά την επίθεση και μετά από την εφαρμογή του μηχανισμού άμυνας. Κατα την εξαγωγή των αποτελεσμάτων πραγματοποιήθηκε ο διαχωρισμός και η παρουσίαση συγκεκριμένων παραδειγμάτων από τα σενάρια αυτά. Συμπληρωματικά γίνεται αναφορά στα συστήματα Question Answering, τα οποία βασίζονται σε παρόμοιες τεχνολογίες. Παράλληλα παρουσιάζεται μια εφαρμογή τους, με την χρήση του Bert-as-a-Service σε συνδυασμό με τις επιθέσεις σε μοντέλα μηχανικής μάθησης, τόσο σε επίπεδο χαρακτήρα όσο και σε επίπεδο σημασιολογίας. Τέλος αναγράφονται τα συμπεράσματα που προέκυψαν όπως και οι μελλοντικές πιθανές επεκτάσεις.

Λέξεις κλειδιά: Αντιπαραθετική επίθεση, Αντιπαραθετική άμυνα, Μηχανική μάθηση, Εργασίες Επεξεργασίας φυσικής γλώσσας, Νευρωνικά δίκτυα.

Abstract

While Machine Learning and Artificial Intelligence technologies are growing and are being used in new systems, the concerns about their robustness and the assurance of a high level of performance of these applications will be rising. In this thesis we will investigate some dangers and attacks that can weaken machine learning models and we will see some ways of adversarial defenses for these models. More specifically, the adversarial attacks in textual data that are presented are mainly divided into two categories. The first category is about adversarial attacks that affect the visual similarity of text as an attack on character level, while the other category is about adversarial attacks that affect the semantic similarity of the words. In addition, the corresponding defense mechanism for every adversarial attack is explained and demonstrated to reassure the security and to maintain the performance of machine learning models as high as possible. Next, the application of machine learning techniques in two different scenarios is shown. The first is about the adversarial attacks in character level with the following defense mechanism. To demonstrate the adversarial attack, TextAttack framework was used for the augmentation of the model input training data, while ScRNN was used for adversarial defense. In this scenario, pre-trained models BERT and DistilBERT were used. Besides that, the second scenario, is about the semantic similarities on a word level. For the demonstration of its adversarial attacks, WordNet was used, that is also provided by the TextAttack framework. Meanwhile, IGA adversarial attack was also used to modify the input test data. As an adversarial defense SEM was chosen. In this scenario, CNN and RNN models were used. The main goal was the evaluation and the comparison of the models with the use of different techniques. The evaluation has been done in three different phases: before the attack, after the attack and after the defense mechanism. From the extraction of the results of the two scenarios, some examples were selected for demonstration. In addition a reference on Question Answering Systems was made, which are also based on similar technologies. Alongside, an application of them is demonstrated, using the Bert-as-a-Service framework, in combination with the adversarial attacks on Machine Learning models, both in terms of character level and in terms of semantic word level. In the end the conclusions that occurred are given, as well as future possible extensions.

Keywords: Adversarial Attacks, Adversarial Defenses, Machine Learning, Natural Language Processing Tasks , Neural Networks.

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΠΙΝΑΚΑΣ ΑΠΟΣΠΑΣΜΑΤΩΝ ΚΩΔΙΚΑ

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

IMDB	Internet Movie Database
BERT	Bidirectional Encoder Representations from Transformers
RNN	Recurrent Neural Network
CNN	Convolution Neural Network
IGA	Genetic based on Text Attack
SEM	Synonym Encoder Methods
QA	Question Answering
NLP	Natural Language Processing
PWWS	Probability Weighted Word Saliency
PSO	Particle Swarm Optimization
BAE	BERT-based Adversarial Examples
DNNs	Deep Neural Networks
RSE	Random Substitution Encoding
DNE	Dirichlet Neighborhood Ensemble
ML	Machine Learning
GLUE	General Language Understanding Evaluation
SQuAD	Stanford Question Answering Dataset
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
GPU	Graphics Processing unit
CLS	Class

1. Εισαγωγή

Η μηχανική μάθηση ή αλλιώς Machine Learning, βασίζεται σε εφαρμογές και μεθόδους, οι οποίες χρησιμοποιούνται πολύ συχνά στις μέρες μας, έχοντας τη δυνατότητα να μαθαίνουν και να εντοπίζουν μοτίβα και χαρακτηριστικά σε μεγάλες ποσότητες δεδομένων (Mohammed et al., 2016). Έτσι δεδομένου της εμπειρίας που αποκτούν τα μοντέλα μέσα στον χρόνο, μπορούν να κάνουν προβλέψεις για μελλοντικά δεδομένα ή να λαμβάνουν αποφάσεις, με ένα βαθμό αβεβαιότητας.

Σήμερα η επαφή με τέτοια μοντέλα μηχανικής μάθησης κάθε μέρα είναι αναπόφευκτη. Τέτοια μοντέλα υπάρχουν είτε στον χώρο εργασίας μας αλλά ακόμα και στην προσωπική μας ζωή. Μερικά παραδείγματα είναι οι "ψηφιακοί βοηθοί" ("Digital Assistants"), για αναζήτηση στο διαδίκτυο, τα προτεινόμενα προϊόντα, ταινίες, τραγούδια που βασίζονται στο ιστορικό αναζήτησης , σε προηγούμενες αγορές μας ή σε υποκατηγορίες. Όλα αυτά είναι εφαρμογές των συστημάτων συστάσεων("Recommender Systems") που βασίζονται σε μοντέλα μηχανικής μάθησης. Ακόμα και τα chatbots, που βλέπουμε όλο και πιο συχνά, χρησιμοποιούν επεξεργασία της πραγματικής γλώσσας ώστε να μπορούν να ερμηνεύσουν το κείμενο που εισάγεται ώστε να δώσουν την κατάλληλη απάντηση. Επιπλέον άλλες εφαρμογές της μηχανικής μάθησης σε συνδυασμό με την τεχνητή νοημοσύνη εμφανίζονται σε πλατφόρμες αναγνώρισης αντικειμένων, αναγνώρισης προσώπων, αναγνώρισης ήχου/φωνής, ανάλυσης συναισθημάτων (Huq and Pervin, 2020)

Παρόλα αυτά, ενώ η μηχανική μάθηση και η τεχνητή νοημοσύνη αποκτούν όλο και πιο κυριαρχικό ρόλο στον τομέα της τεχνολογικής ανάπτυξης όσο και στη ζωή μας, οι ευαισθησίες τους μπορούν να τα επηρεάζουν κατά την λήψη αποφάσεων. Ενώ παράλληλα η ασφάλειά τους μπορεί εύκολα να παραβιαστεί και έτσι η ευρωστία τους να καθίσταται αμφισβητήσιμη.

1.1. Επιθέσεις σε μοντέλα μηχανικής μάθησης

Όπως αναφέρθηκε παραπάνω υπάρχουν ποικίλα μοντέλα μηχανικής μάθησης, ανάλογα με τις ανάγκες της κάθε εφαρμογής, αλλάζουν και τα δεδομένα που χρειάζεται να επεξεργάζονται. Επομένως οι επιθέσεις στα μοντέλα, αλλάζουν επίσης ανάλογα με τον τύπο και τη μορφή των

δεδομένων που αυτά επεξεργάζονται. Από την άλλη μεριά για να μπορέσει κάποιος να χτίσει ένα μηχανισμό άμυνας για το μοντέλο του, χρειάζεται να γνωρίζει την επίθεση την οποία έχει να αντιμετωπίσει. Μέχρι σήμερα δεν έχει σχεδιαστεί καμία επίθεση που να μπορεί να χρησιμοποιηθεί σαν μηχανισμός άμυνας για όλες τις επιθέσεις. Επομένως κάθε επίθεση χρειάζεται να έχει τον δικό της μηχανισμό άμυνας.

Δεδομένου αυτού και αν λάβουμε υπόψη το πόσο ταχιά είναι η ανάπτυξη στον συγκεκριμένο τομέα, το υλικό και οι δημοσιεύσεις που υπάρχουν σε αυτό το αντικείμενο, δεν είναι πολλές. Συνεπώς δύσκολα μπορεί να βρει κάποιος κρητικές για συγκεκριμένες επιθέσεις και αντίστοιχα τις άμυνές τους ώστε να μπορέσει να της χρησιμοποιήσει με βεβαιότητα στα δικά του μοντέλα.

1.2. Αντικείμενο διπλωματικής εργασίας

Αντικείμενο της παρούσας εργασίας είναι η πειραματική μελέτη των επιθέσεων και των αντίστοιχων μηχανισμών άμυνας, σε μοντέλα μηχανικής μάθησης, τα οποία εκπαιδεύονται ώστε να επεξεργάζονται δεδομένα κειμένου. Στόχος της παρούσας εργασίας είναι σύγκριση των τριών φάσεων: πριν την επίθεση, μετά την επίθεση και μετά την άμυνα.

Μετά από έρευνα και μελέτη που προηγήθηκαν, συγκεντρώθηκαν οι απαραίτητες τεχνολογίες και τα εργαλεία ώστε να υλοποιηθούν με επιτυχία κάποια πειράματα για την μελέτη της επίπτωσης των επιθέσεων black box στην απόδοση των μοντέλων. Οι επιθέσεις αυτές εφαρμόζονται σε μοντέλα μηχανικής μάθησης που χρησιμοποιούνται για κατηγοριοποίηση δεδομένων κειμένου, ενώ παράλληλα αξιολογείται κατά πόσο ένας αντίστοιχος μηχανισμός άμυνας εξυπηρετεί ή όχι τον σκοπό του στο συγκεκριμένο σενάριο.

Το πρώτο πειραματικό σενάριο που σχεδιάστηκε και αναλύθηκε, έχει να κάνει με επίθεση σε μοντέλο μηχανικής μάθησης το οποίο κατηγοριοποιεί κείμενα (text classification). Η συγκεκριμένη επίθεση έχει σαν στόχο τους χαρακτήρες της κάθε λέξης. Δηλαδή εισαγωγή, διαγραφή, αντιστροφή, αντικατάσταση κλπ. κάποιου χαρακτήρα της λέξης ώστε να μπερδεύεται το μοντέλο κατά την κατηγοριοποίηση του συνολικού κειμένου. Στη συνέχεια, αυτή η επίθεση αντιμετωπίζεται με τον αντίστοιχο μηχανισμό άμυνας. Τα αποτελέσματα απόδοσης συγκρίνονται

με την τεχνική k-fold cross validation λαμβάνοντας ένα αριθμό accuracy, κατά τις τρεις φάσεις, πριν την επίθεση, μετά την επίθεση και μετά την άμυνα.

Κάτι αντίστοιχο ακολουθήθηκε για την σχεδίαση και την ανάλυση του δεύτερου πειραματικού σεναρίου. Το δεύτερο πειραματικό σενάριο έχει να κάνει επίσης με επίθεση σε μοντέλο μηχανικής μάθησης το οποίο κατηγοριοποιεί κείμενα, με την διαφορά ότι οι επιθέσεις έχουν να κάνουν πλέον με την σημασιολογική έννοια της κάθε λέξης. Έτσι το μοντέλο εξετάζεται στο κατά πόσο μπορεί να διαχειριστεί τέτοιου είδους αλλαγές που πλέον δεν είναι εμφανείς στο ανθρώπινο μάτι, τουλάχιστον όχι τόσο εύκολα. Στην συνέχεια προστέθηκε ένας μηχανισμός άμυνας ο οποίος υλοποιήθηκε με σκοπό την αντιμετώπιση ακριβώς αυτής της μορφής επιθέσεις. Έπειτα αξιολογείται επίσης η απόδοση του σεναρίου με cross validation στις τρεις φάσεις του πειράματος.

Γενικότερα ο σκοπός της εργασίας είναι η ανάπτυξη και η αξιολόγηση διαφόρων σεναρίων με επιθέσεις και άμυνες σε μοντέλα μηχανικής μάθησης. Ο ρόλος των οποίων μπορεί να είναι καθοριστικός για την απόδοση των εφαρμογών που χρησιμοποιούν τέτοιου είδους τεχνολογίες ενώ διαχειρίζονται σημαντικά δεδομένα.

1.3. Οργάνωση κειμένου

Στο δεύτερο κεφάλαιο αναφέρεται η έρευνα που προηγήθηκε, ώστε να υπάρχει απόλυτη κατανόηση των εννοιών και των τεχνολογιών που θα χρησιμοποιηθούν αργότερα κατά το στήσιμο των σεναρίων και των πειραμάτων. Έπειτα στο τρίτο κεφάλαιο, περιγράφεται η μεθοδολογία και η προσέγγιση του κάθε σεναρίου ξεχωριστά, ώστε εν τέλη να μπορέσει να επιτευχθεί ο σκοπός των πειραμάτων. Παράλληλα εξηγείται το κάθε βήμα και ο τρόπος που χρησιμοποιήθηκε η κάθε τεχνολογία ή το εργαλείο στα πλαίσια της εργασίας. Επιπλέον, στο τέταρτο κεφάλαιο, λαμβάνοντας υπόψη την μελέτη που προηγήθηκε, παρουσιάζονται και αναλύονται τα πειραματικά αποτελέσματα που προκύπτουν μέσα από τα διαφορετικά σενάρια. Τέλος στο πέμπτο κεφάλαιο αναφέρονται μελλοντικές κατευθύνσεις που μπορούν να εφαρμοστούν σε μελλοντικές επεκτάσεις. Ενώ παράλληλα στο πέμπτο κεφάλαιο αναγράφονται επίσης τα συμπεράσματα που προέκυψαν στην παρούσα εργασία αλλά και τα προβλήματα που αντιμετωπίστηκαν.

2. Σχετική Έρευνα και Προηγούμενες Εφαρμογές

Παρόλο που τα Νευρωνικά Δίκτυα έχουν ευρεία αναγνώριση για τις δυνατότητες που προσφέρουν κατά την εξαγωγή, τη διαχείριση πληροφοριών και γνώσης από μεγάλες ποσότητες δεδομένων, φαίνεται να είναι ευάλωτα όσον αφορά τα αντιπαραθετικά παραδείγματα. Συνεπώς για την αντιμετώπιση αυτού του προβλήματος, οι αντιπαραθετικές επιθέσεις και οι άμυνες ενάντια αυτών, έχουν πλέον γίνει το επίκεντρο των ερευνών στον συγκεκριμένο τομέα.

Τα αντιπαραθετικά παραδείγματα (Adversarial Examples), έχουν σαν στόχο την αποτυχία του μοντέλου το οποίο στοχεύουν και πιο συγκεκριμένα την διαδικασία πρόβλεψης του μοντέλου. Αυτό μπορεί να γίνει σκόπιμα αλλά και χωρίς κάποιο συγκεκριμένο σκοπό. Ουσιαστικά τα αντιπαραθετικά παραδείγματα, είναι δεδομένα τα οποία χρησιμοποιούνται συνήθως σαν δεδομένα εισόδου στο μοντέλο μηχανικής μάθησης και είναι σχεδιασμένα ώστε να το παραπλανήσουν. Ένα τέτοιο παράδειγμα είναι μία τροποποίηση της αυθεντικής εισόδου και είναι γνωστό και σαν μια αντιπαραθετική τροποποίηση (adversarial perturbation) (J.Morris, 2020). Παρόλα αυτά η δημιουργία και η χρήση των αντιπαραθετικών παραδειγμάτων σήμερα, γίνεται για να μπορούν να αξιολογούνται οι αποδόσεις των μοντέλων. Ώστε να διασφαλιστεί ότι τα συγκεκριμένα μοντέλα έχουν γερά θεμέλια, ενώ επίσης βοηθούν και στον εντοπισμό των αδυναμιών των ίδιων των μοντέλων. Από την άλλη μερία, οι επιθέσεις σε μοντέλα μηχανικής μάθησης μπορούν να επιτευχθούν εύκολα, αφού τα μοντέλα αυτά συνήθως δεν προστατεύονται από τέτοιου είδους επιθέσεις μέχρι και σήμερα. Έτσι τελικά οι αντιπαραθετικές επιθέσεις έχουν σκοπό την αλλοίωση των αποτελεσμάτων που θα εξάγει το μοντέλο.

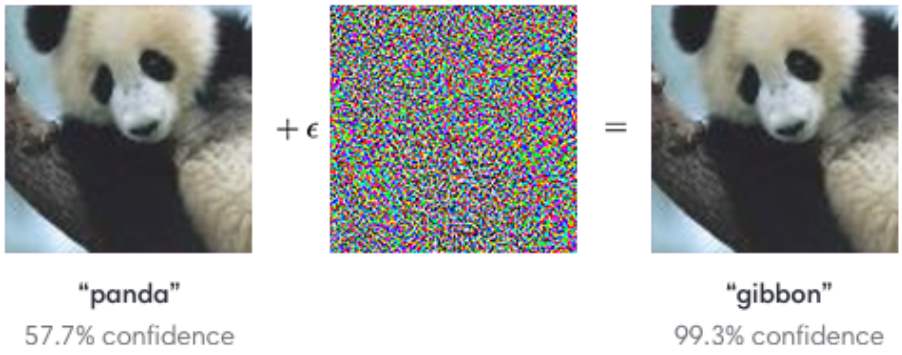
Η ασφάλεια των εφαρμογών που διαχειρίζονται σημαντικά δεδομένα και είναι βασισμένες σε μοντέλα μηχανικής μάθησης μπορεί να παραβιαστεί με την χρήση αντιπαραθετικών επιθέσεων. Η επίθεση που θα χρησιμοποιηθεί για να πετύχει η παραβίαση της ασφάλειας του ίδιου του μοντέλου, απαιτεί να είναι γνωστή η λειτουργικότητα του μοντέλου ή τουλάχιστον το τι δεδομένα εισόδου - εξόδου χρησιμοποιεί το μοντέλο. Επίσης καθοριστικό ρόλο στην όλη διαδικασία παίζει και το πώς αυτό το μοντέλο προστατεύει τα δεδομένα του και αν αυτό γίνεται.

Ένας τρόπος που μπορεί το μοντέλο να αποτύχει στις προβλέψεις του ή να παραπλανηθεί κατά την διαδικασία λήψης αποφάσεων, είναι η τροποποίηση των δειγμάτων των δεδομένων που λαμβάνει σαν είσοδο του. Είτε προσθέτοντας θόρυβο ή επιπλέον πληροφορία σε αυτά, είτε αφαιρώντας πληροφορία αλλά ακόμα και τροποποιώντας την ίδια την πληροφορία. Δεδομένου αυτού, οι επιθέσεις αυτές χωρίζονται σε δύο κύριες κατηγορίες, τις black box και τις white box επιθέσεις.

Οι black box επιθέσεις αντιπροσωπεύουν τα πιο ρεαλιστικά σενάρια, στα οποία ο επιτήδειος δεν γνωρίζει και δεν έχει άμεση πρόσβαση στην αρχιτεκτονική του μοντέλου. Έτσι το μόνο που μπορεί να τροποποιήσει είναι το input/output του μοντέλου. Οι white box επιθέσεις από την άλλη μεριά είναι πιο ισχυρές και μπορούν να επιφέρουν στο μοντέλο καταστροφικές συνέπειες, αφού ο επιτήδειος έχει γνώση ή άμεση πρόσβαση σε ολόκληρο το μοντέλο και την αρχιτεκτονική του (Liang et al., 2017).

Έτσι η ανάγκη για περισσότερη έρευνα στον τομέα της εύρεσης μηχανισμών άμυνας για τις συγκεκριμένες επιθέσεις καθίσταται αναγκαία. Ο σκοπός των αντιπαραθετικών άμυνων είναι να καταφέρουν να βοηθήσουν το μοντέλο να κρατήσει τις υψηλές αποδόσεις του τόσο στα αρχικά δεδομένα όσο και στα Adversarial examples δεδομένα.

Εφόσον οι επιτήδειοι δεν μοιράζονται τις επιθέσεις που χρησιμοποιούν και η έρευνα στην εύρεση τέτοιου είδους επιθέσεων δεν είναι ιδιαίτερα εύκολη, η διαδικασία αντιστοίχισης της κάθε επίθεσης με συγκεκριμένη άμυνα γίνεται όλο και πιο περίπλοκη. Έτσι τελικά η εκπαίδευση των μοντέλων σε επιθέσεις περιορίζεται σε συγκεκριμένες περιπτώσεις . Πέρα από αυτό εάν ένας χρήστης προσπαθήσει να εκπαιδεύσει το μοντέλο του ενάντια σε όλες τις επιθέσεις , το πιο πιθανό είναι αυτό να οδηγήσει σε παραπλάνηση και περισσότερο μπέρδεμα του μοντέλου καθώς αυτό δεν θα είναι ικανό να εκτελέσει την διαδικασία σωστά λόγω της πολύ μικρής πληροφορίας που θα έχει από τα αρχικά αυθεντικά δεδομένα.



Εικόνα 1 Παράδειγμα Adversarial Attack σε δεδομένα εικόνων (Goodfellow et al., 2014)

Ένα χαρακτηριστικό παράδειγμα για καλύτερη κατανόηση της έννοιας αυτών των επιθέσεων, είναι αυτό που φαίνεται στην εικόνα 1. Όσον αφορά τα οπτικοποιημένα δεδομένα, όπως οι εικόνες, μία επίθεση μπορεί να γίνει αλλάζοντας ένα μικρό pixel. Έτσι το μοντέλο κατηγοριοποιεί λάθος την εικόνα. Ενώσω το ανθρώπινο μάτι δεν μπορεί να αναγνωρίσει αυτή την αλλοίωση μεταξύ της αυθεντικής εικόνας και αυτής που έχει δεχτεί κάποια επίθεση, το μοντέλο μπορεί να μπερδευτεί και θεωρεί ότι το “panda” του παραδείγματος, μετά την προσθήκη θορύβου, κατηγοριοποιείται σαν “gibbon”.

Όσον αφορά τα δεδομένα κειμένου, παρουσιάζουν διαφορές στην μορφή σε σχέση με τα δεδομένα εικόνας. Αρχικά τα δεδομένα κειμένου είναι διακριτά δεδομένα ενώ οι εικόνες είναι συνεχόμενα, παράλληλα κάποιος μπορεί να αντιληφθεί τον θόρυβο με το ανθρώπινο μάτι πιο δύσκολα ενώ από την άλλη μια απλή αλλαγή ενός χαρακτήρα στο κείμενο, μπορεί να εντοπιστεί πολύ εύκολα. Επιπλέον στο κείμενο προστίθεται και η εννοιολογική σημασία, η οποία αλλάζει με την αλλαγή μιας λέξης ή ακόμα και ενός χαρακτήρα. Στην περίπτωση των εικόνων έστω και αν προσθέσουμε θόρυβο ο άνθρωπος μπορεί να αντιληφθεί τί απεικονίζεται.

Επιπλέον η σημασία των αντιπαραθετικών παραδειγμάτων ισως να διαφέρει στα δεδομένα κειμένου, σε σχέση με τα δεδομένα άλλων κατηγοριών. Καθώς οι εναλλαγές που γίνονται στα κείμενα δεν είναι διακριτικές, έτσι οι ερευνητές προτείνουν δύο διαφορετικούς εναλλακτικούς ορισμούς για τα αντιπαραθετικά παραδείγματα σε NLP, που βασίζονται στην οπτική τους ομοιότητα.

Visual Similarity (J.Morris, 2020): Μερικές NLP επιθέσεις, θεωρούν ένα αντιπαραθετικό παράδειγμα, σαν να είναι μια ακολουθία κειμένου που μοιάζει πολύ με την αυθεντική είσοδο του μοντέλου, με μερικούς χαρακτήρες διαφορετικούς, αλλά στο τέλος λαμβάνει διαφορετική πρόβλεψη κατηγορίας από το μοντέλο. Μερικές από αυτές τις επιθέσεις προσπαθούν να αλλάζουν όσο το δυνατόν λιγότερους χαρακτήρες ώστε το μοντέλο να κάνει λάθος προβλέψεις. Άλλες επιθέσεις προσπαθούν να παρουσιάσουν πιο ρεαλιστικά λάθη, όπως τα τυπογραφικά τα οποία μπορούν να γίνουν και από τον ίδιο τον άνθρωπο. Αυτές οι επιθέσεις, θεωρείται ότι μπορούν να αντιμετωπιστούν πιο εύκολα χρησιμοποιώντας rule-based ορθογραφικό έλεγχο ή με μοντέλα εκπαιδευμένα σε προτάσεις ώστε να διορθώνουν τα ορθογραφικά λάθη. Τέτοιες επιθέσεις είναι το DeepWordBug, HotFlip, Pruthi, Morpheus κτλ.

Semantic Similarity (J.Morris, 2020): Άλλες επιθέσεις NLP, θεωρούν ότι ένα αντιπαραθετικό παράδειγμα είναι χρήσιμο, όταν είναι σημασιολογικά διαφορετικό από το αυθεντικό κείμενο που θα χρησιμοποιηθεί σαν είσοδος στο μοντέλο. Με άλλα λόγια πρέπει να είναι μια παράφραση του αυθεντικού κειμένου, αλλα παράλληλα να λαμβάνει μια λάθος πρόβλεψη από το μοντέλο. Υπάρχουν μοντέλα τα οποία εκπαιδεύονται ώστε να υπολογίζουν την σημασιολογική ομοιότητα των προτάσεων και για να πετύχει μια επίθεση τέτοιου είδους, χρειάζεται να χρησιμοποιηθεί άλλο μοντέλο NLP ώστε να επιβάλλεται ότι η τροποποίηση αυτή είναι σωστή και σημασιολογικά ίδια με την αυθεντική πρόταση. Μερικές επιθέσεις βασισμένες στην σημασιολογική έννοια των προτάσεων είναι οι Bae, Bert-Attack, IGA, Kuleshov, PWWS, PSO κτλ.

Στο άρθρο του Belinkov (2018) σχολιάζονται οι διαφορετικοί τρόποι ανάλυσης NLP, ο τύπος της πληροφορίας που δέχονται τα μοντέλα, ενώ επίσης παρουσιάζονται παραδείγματα που συνάδουν ότι τα νευρωνικά μοντέλα μηχανικής μάθησης (DNNs) είναι αδύναμα και δεν προστατεύονται σε επαρκή βαθμό απο αντιπαραθετικές επιθέσεις.

Χάρη στα αντιπαραθετικά παραδείγματα που είναι η βάση των περισσότερων επιθέσεων, εντάσσεται η ανάγκη της εκπαίδευσης των μοντέλων σε τροποποιημένα κείμενα, αυτή η διαδικασία ονομάζεται αντιπαραθετική εκπαίδευση (Adversarial Training). Κατά την αντιπαραθετική εκπαίδευση ουσιαστικά το μοντέλο εκπαιδεύεται χωριστά στα αυθεντικά δεδομένα αλλά και στα δεδομένα που έχουν δεχθεί επίθεση, χρησιμοποιώντας τις αυθεντικές

κατηγορίες (labels) των δεδομένων. Κατά συνέπεια εκπαιδεύεται και με τις δύο ομάδες δεδομένων, έτσι πλέον είναι σε θέση να αναγνωρίσει πιο εύκολα τα δεδομένα και με μορφή επίθεσης αλλά και στην κανονική τους μορφή.

Παράλληλα για τα δεδομένα που αφορούν κείμενα ή επεξεργασία κειμένου (NLP), αντιμετωπίζονται τέσσερα διαφορετικά είδη επιθέσεων.

- Σε επίπεδο χαρακτήρα (character level attacks) : Εισαγωγή νέων χαρακτήρων, ειδικών χαρακτηρων, ακόμη και αριθμών, ανταλλαγή χαρακτήρων με γειτονικούς, διαγραφή χαρακτήρων, αντιστροφή χαρακτήρων.
- Σε επίπεδο λέξης (word level attacks) : Αντικατάσταση λέξεων με συνώνυμες, αντώνυμες. Εισαγωγή λέξεων με ορθογραφικά λάθη. Διαγραφή ολόκληρων λέξεων.
- Σε επίπεδο προτάσεων (sentence level attacks) : Προσθήκη επιπλέον προτάσεων για αλλοίωση του νοήματος.
- Σε όλα τα επίπεδα (Multilevel attacks) : Εφαρμογή όλων των επιπέδων.

past → pas!t	Alps → llps	talk → taln	local → loral	you → yoTu	ships → hips	actor → actr	lowered → owered
pasturing	lips	tall	moral	Tutu	dips	act	powered
pasture	laps	tale	Moral	Hutu	hops	acting	empowered
pastor	legs	tales	coral	Turku	lips	actress	owed
Task	slips	talent	morals	Futurum	hits	acts	overpowered

Εικόνα 2 Παράδειγμα Word-Level Adversarial Attack (Ebrahimi, 2018)

Στην εικόνα 2, φαίνονται παραδείγματα λέξεων που δέχονται καποιου είδους επίθεση. Έπειτα κατα την προσπάθεια της ανάκτησης της πρωταρχικής λέξης προβλέπεται άλλη λέξη. Έτσι είναι προφανή η δυσκολία που χρειάζεται να αντιμετωπιστεί ώστε να χτιστεί ένας μηχανισμός άμυνας που να επαναφέρει την λέξη που δέχθηκε επίθεση στην αρχική της μορφή. Επιπρόσθετα για τη χρήση την επεξεργασία και την ανάλυση των δεδομένων που αφορούν κείμενα, υπάρχουν μοντέλα με διαφορετικές ιδιότητες τα οποία χρησιμοποιούνται στις μέρες μας.

- Μοντέλα κατηγοριοποίησης κειμένων (Classification models) (Liang), όπως αυτό που χρησιμοποιήθηκε στα πλαίσια της εργασίας.

- Μοντέλα που δεδομένου μιας παραγράφου ή μιας ερώτησης μπορούν να εξάγουν απάντηση. (Question Answering Models) (Jia 2017)
- Μοντέλα παραγωγής συμπερασμάτων (Natural Language Inference)
- Μοντέλα που χρησιμοποιούνται για μεταφράσεις.(Machine Translation)

Παρόλη την ταχεία ανάπτυξη των επιθέσεων, φαίνεται ότι όσο αφορά τους μηχανισμούς άμυνας η έρευνα έχει μείνει σχετικά πίσω. Μέχρι τώρα υπάρχουν λίγες υλοποιήσεις σε μηχανισμούς άμυνας για επιθέσεις σε NLP διαδικασίες. Σκοπός των μηχανισμών αυτών, είναι συνήθως η αύξηση της ανθεκτικότητας των μοντέλων, με την ενσωμάτωση τροποποιήσεων σε λέξεις όπως για παράδειγμα με την τεχνική του Adversarial Training, ή με Defensive Distillation. Οι προσεγγίσεις αυτές παρόλα αυτά φαίνεται να έχουν μεγαλύτερες αποδόσεις από ότι τα βασικά μοντέλα, χωρίς να λαμβάνουν υπόψη το ενδεχόμενο της κακοήθης παρέμβασης στα μοντέλα.

Τέτοιοι μηχανισμοί άμυνας σε επίπεδο σημασιολογικής έννοιας, είναι ο SEM και ο RSE (Z.Wang and H.Wang, 2020), με την διαφορά ότι ο πρώτος μπορεί να αμυνθεί αποτελεσματικά σε επιθέσεις που είναι ήδη υλοποιημένες και εφαρμόζονται. Αφού περιορίζει το εύρος απόστασης μεταξύ των μετασχηματισμένων παραδειγμάτων δοκιμής και των περιορισμένων παραδειγμάτων εκπαίδευσης. Από την άλλη ο RSE, τροποποιεί τις ενσωματωμένες λέξεις πριν από την διαδικασία εκπαίδευσης. Ενώ επίσης προσθέτει την διαδικασία αντικατάστασης των συνωνύμων μέσα στην διαδικασία εκπαίδευσης, ώστε να δώσει στα μοντέλα περισσότερα παραδείγματα για εκπαίδευση. Έτσι οι δημιουργοί του RSE προτείνουν ανάμεσα στο επίπεδο εισόδου και στο επίπεδο Embedding, να εισαχθεί μια Dynamic Random Synonym Substitution, που εφαρμόζει το RSE.

Ένας άλλος μηχανισμός άμυνας που έχει προταθεί είναι η μέθοδος DNE (Dirichlet Neighborhood Ensemble), που χρησιμοποιείται για να εκπαιδεύεται η ανθεκτικότητα του μοντέλου, ώστε να αμύνεται σε επιθέσεις αντικατάστασης λέξεων. Κατά την εκπαίδευση των μοντέλων με την χρήση της συγκεκριμένης μεθόδου, φτιάχνετε μια εικονική ακολουθία, συγχέοντας τις συνώνυμες λέξεις μιας λέξης από την πρόταση εισόδου(Zhou et al., 2020) . Έτσι εκπαιδεύει το μοντέλο σε αυτά τα δεδομένα, πετυχαίνοντας την ενίσχυση της ανθεκτικότητας του μοντέλου σε επιθέσεις βασισμένες σε εναλλαγές συνώνυμων λέξεων, αλλά ταυτόχρονα και την διατήρηση της απόδοσής του στα αυθεντικά δεδομένα.

3. Μεθοδολογία και Προσέγγιση του Προβλήματος

3.1. Σενάριο 1 - Αλλαγή Χαρακτήρων

Για το συγκεκριμένο σενάριο χρησιμοποιήθηκαν τα προ εκπαιδευμένα μοντέλα BERT (Devlin et al., 2018), DistilBert (Sanh et al., 2019) για την εκπαίδευση και αξιολόγηση σε διαφορετικά σύνολα δεδομένων. Για την κατηγοριοποίηση και τις ανάγκες του πειράματος χρησιμοποιήθηκαν δεδομένα από το σύνολο AgNews και από το σύνολο IMBD. Ενώ στη συνέχεια για την δημιουργία αντιπαραθετικών παραδειγμάτων, που χρησιμοποιήθηκαν για αντιπαραθετική εκπαίδευση, έγινε χρήση του TextAttack (X.Morris et al., 2020) εργαλείου. Έπειτα για τον μηχανισμό άμυνας χρησιμοποιήθηκε το μοντέλο ScRNN (Pruthi et al., 2019). Τέλος για την αξιολόγηση του μοντέλου κατά τις τρεις φάσεις, χρησιμοποιήθηκε τεχνική Stratified K-Fold Cross Validation με την μετρική Accuracy.

3.1.1. Δεδομένα Κατηγοριοποίησης

Τα δεδομένα που χρησιμοποιούνται για σκοπούς κατηγοριοποίησης πρέπει να περιλαμβάνουν τουλάχιστον δύο στήλες [κείμενο][κατηγορία], ώστε να είναι εμφανές σε τι κατηγορία κατατάσσεται το συγκεκριμένο κείμενο. Είτε αφορά κάποιο αρνητικό ή θετικό σχόλιο, είτε αφορά κάποια άλλη κατηγορία. Επιπλέον αυτές οι κατηγορίες όταν πρόκειται για μοντέλα κατηγοριοποίησης, θα πρέπει να εμφανίζονται σαν αριθμητικοί χαρακτήρες ώστε να μπορούν να επεξεργαστούν ανάλογα από τα μοντέλα. Για τις ανάγκες της παρούσας εργασίας έγινε η απαραίτητη προεπεξεργασία ώστε τα δεδομένα που χρησιμοποιήθηκαν να είναι αντίστοιχης μορφής.

3.1.1.1. AG_News

Τα δεδομένα από το AG_News, περιέχουν τίτλους ειδήσεων οι οποίες χωρίζονται σε τέσσερις μεγάλες κατηγορίες World, Sports, Business, Sci/Tech οι οποίες αναφέρονται σαν 1,2,3,4 αντίστοιχα. Μερικά παραδείγματα από το σύνολο αυτό παρουσιάζονται στον πίνακα 1.

1	Britain Black Watch regiment, deployed to back up US forces outside Baghdad since the end of October, has returned to its base in Basra in southern Iraq, said a UK defense ministry spokesman.
4	It's time to get down to business for stem-cell scientists and administrators after the tremendous boost they received from the passage of California's Proposition 71. Two advisory board appointments have been made; there are 25 to go. By Kristen Philipkoski.
4	The company will pony up \$2.6 million to settle a lawsuit over allegedly flawed electronic voting systems it sold to California counties. Surprise: E-vote activists still aren't happy.
3	The WTO case was brought by the Caribbean state of Antigua and Barbuda, host to many of the online casinos whose use is illegal in the US.

Πίνακας 1 Δεδομένα από το σύνολο AG_News

3.1.1.2. IMDB

Τα δεδομένα από το IMDB περιέχουν σχόλια ταινιών τα οποία χωρίζονται σε δύο κατηγορίες positive/negative ανάλογα με το περιεχόμενό τους, τα οποία αναγράφονται σαν 0-1 αντίστοιχα.

0	I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue ….
1	Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all ….

Πίνακας 2 Δεδομένα από το σύνολο IMDB

3.1.2. Bert - DistilBert – Pretrained

Η ξαφνική άνοδος των εφαρμογών που χρησιμοποιούν μεγάλης κλίμακας προ εκπαιδευμένα μοντέλα είναι φανερή τα τελευταία χρόνια. Πλέον έχουν γίνει ένα βασικό εργαλείο σε πολλές εφαρμογές επεξεργασίας γλώσσας σε μορφή κειμένου (NLP). Τα προ εκπαιδευμένα μοντέλα παρέχουν την δυνατότητα σε κάποιο ώστε πιο εύκολα να χρησιμοποιήσει ένα ήδη υπάρχον μοντέλο για τον σκοπό που το χρειάζεται, χωρίς να πρέπει να το "κτίσει" από την αρχή. Επιπλέον τα προ εκπαιδευμένα μοντέλα αφήνουν συνήθως τα τελευταία τους επίπεδα άδεια ώστε να μπορεί να τα εκπαιδεύσει ο χρήστης χρησιμοποιώντας τα δικά του δεδομένα.

Για τις ανάγκες της εργασίας χρησιμοποιήθηκαν οι transformers ώστε να υπάρχει η πρόσβαση στα προ εκπαιδευμένα μοντέλα κατηγοριοποίησης, BERT και DistilBert.

3.1.2.1. BERT

Το 2018, η ομάδα της Google κυκλοφόρησε το State-Of-The-Art μοντέλο, ονομαζόμενο BERT. Το BERT(Bidirectional Encoder Representations from Transformers), είναι ένας μετασχηματιστής βασισμένος στην μηχανική μάθηση(ML) για συστήματα επεξεργασίας φυσικής γλώσσας(NLP). Μπορεί να χρησιμοποιηθεί και σαν ένα ήδη εκπαιδευμένο μοντέλο, το οποίο δημιουργήθηκε από την ίδια την Google. Η Google από το 2019, το χρησιμοποιεί ώστε να μπορεί η μηχανή αναζήτησής τους να κατανοεί καλύτερα τις αναζητήσεις των χρηστών.

Είναι ένα προ εκπαιδευμένο μοντέλο, το οποίο καταλαβαίνει την σημασία της λέξης, με την βοήθεια των μετασχηματιστών, σε σχέση με την υπόλοιπη πρόταση και όχι σαν ένα αυτόνομο κομμάτι. Έτσι μπορεί να θεωρηθεί ότι τα μοντέλα BERT έχουν την ικανότητα να αντιλαμβάνονται το απόλυτο νόημα του κειμένου που τους δίνεται σαν είσοδος. Το μοντέλο BERT προσπαθεί να καταλάβει το νόημα της κάθε λέξης σε σχέση με την υπόλοιπη πρόταση, ξεκινώντας να "διαβάζει" από αριστερά προς τα δεξιά την πρόταση (Naushad , 2020).

Το αυθεντικό BERT που είναι βασισμένο στην αγγλική γλώσσα, είναι διαθέσιμο σε δύο προ εκπαιδευμένους γενικευμένους τύπους Uncased/Cased τα οποία έχουν εκπαιδευτεί σε σύνολα βιβλίων, με 800 εκατομμύρια λέξεις ενώ επίσης και με 2.500 λέξεις από το Wikipedia. Ο BERT έχει χρησιμοποιηθεί για συγκεκριμένες εφαρμογές όπως GLUE (General Language

Understanding Evaluation), για την δημιουργία του συνόλου δεδομένων SQuAD, απαντήσεων σε ερωτήσεις του Stanford (Stanford Question Answering Dataset), ενώ επίσης και σαν γεννήτρια αντιπαραθετικών κειμένων για επιθέσεις (Jin et al.2019).

Μετά την κυκλοφορία του BERT, ξεκίνησαν να κυκλοφορούν και διάφορες άλλες παραλλαγές του, είτε με καλύτερη απόδοση είτε με λιγότερο υπολογιστικό κόστος ανάλογα με τις ανάγκες κάθε φορά.

```
import transformers as ppb
""' Loading the Pre-trained BERT or DistilBERT model """
if( pretrained == "distilbert"):
    #DistilBERT:
    model_class, tokenizer_class, pretrained_weights = (ppb.DistilBertModel, ppb.DistilBertTokenizer, 'distilbert-base-uncased')
if(pretrained== "bert"):
    #BERT
    model_class, tokenizer_class, pretrained_weights = (ppb.BertModel, ppb.BertTokenizer, 'bert-base-uncased')
```

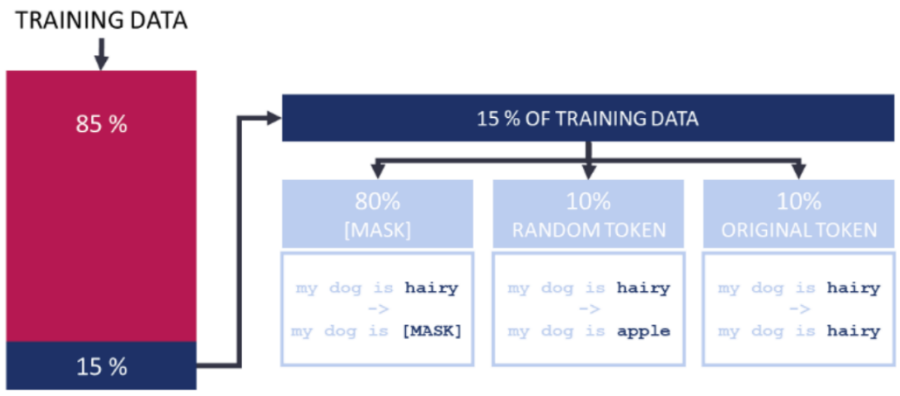
Απόσπασμα Κώδικα 1 Φόρτωση Προ Εκπαιδευμένων Μοντέλων

Χρησιμοποιώντας την βιβλιοθήκη των transformers μπορούμε να κατεβάσουμε και να χρησιμοποιήσουμε τα μοντέλα με τα απαραίτητα αρχεία τους. Ενώ επίσης και την κρυφή μνήμη που χρειάζεται το κάθε μοντέλο για να δουλέψει. Μερικές παράμετροι που χρειάζεται να αρχικοποιηθούν είναι η τιμή model_classes, που ουσιαστικά είναι το ίδιο το μοντέλο, η τιμή του configuration_classes που περιλαμβάνει όλα τα αρχεία που χρειάζεται το μοντέλο για να διαμορφώσει το κατάλληλο περιβάλλον και η τιμή tokenizer_classes στην οποία αποθηκεύεται το λεξιλόγιο του κάθε μοντέλου για να χρησιμοποιηθεί αργότερα κατά την κατηγοριοποίηση των δεδομένων (Naushad,2020).

Χρειάζεται να ακολουθηθούν δύο απαραίτητες τεχνικές, για την εκπαίδευση του BERT. Οι δύο αυτές τεχνικές εφαρμόζονται στα δεδομένα που χρησιμοποίησε το BERT για να εκπαιδευτεί και αναφέρονται πιο κάτω.

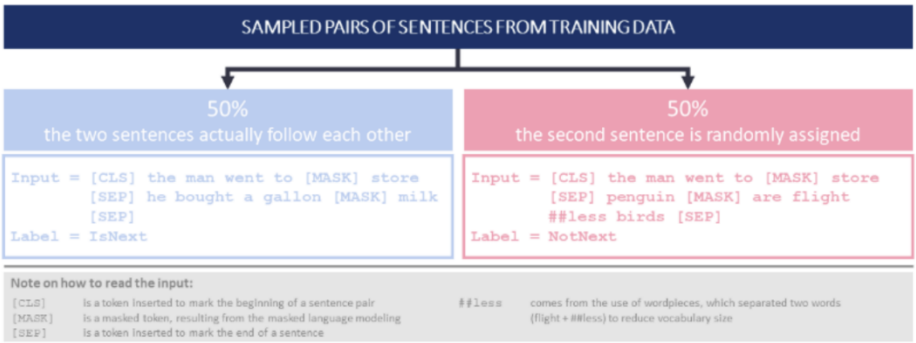
Masked Language Modeling (MLM): Για να έχει το μοντέλο μια αμφίδρομη εκπαίδευση, ο Ντελβιν το 2018, άφησε το μοντέλο να προβλέψει κωδικοποιημένες λέξεις βασισμένο μόνο στο περιεχόμενο του κειμένου. Ουσιαστικά χρειάστηκε να κρυφτεί τυχαία το 15% των λέξεων σε κάθε ακολουθία κειμένου. Από αυτό το 15% των λέξεων, μόνο το 80% αντικαταστάθηκε με

κωδικοποιημένη λέξη [MASK], το 10% αντικαταστάθηκε με τυχαίες άλλες λέξεις και στο υπόλοιπο 10% δεν έγινε καμία αλλαγή.



Σχήμα 1 Masked Language Modeling (Godbole, Grubinska & Kelnreiter, 2020)

Next Sentence Prediction (NSP): Το μοντέλο εκπαιδεύτηκε επίσης ώστε να αναγνωρίζει την σχέση μεταξύ των προτάσεων. Το μοντέλο BERT έπρεπε να αποφασίζει ανάμεσα σε ζευγάρια προτάσεων ή ακόμα και από ομάδα προτάσεων, πότε το ένα μέρος συμπληρώνει το άλλο και πότε όχι. Με ένα ποσοστό 50% για την κάθε πρόβλεψη (true είναι συνέχεια της προηγούμενης, false δεν είναι συνέχεια της προηγούμενης (Devlin et al,2018).



Σχήμα 2 Next Sentence Prediction (Godbole, Grubinska & Kelnreiter, 2020)

3.1.2.2. Transformers

Είναι μοντέλα τα οποία επεξεργάζονται λέξεις, σαν να είναι σε σχέση με όλες τις υπόλοιπες λέξεις της πρότασης αντί για μία προς μία σε σειρά. Έτσι και τα μοντέλα BERT βλέπουν τις

προηγούμενες και τις επόμενες λέξεις, ώστε να κατανοήσουν καλύτερα το νόημα μιας συγκεκριμένης λέξης.

Όπως φαίνεται στο σχήμα 3 της αναπαράστασης της αρχιτεκτονικής ενός μετασχηματιστή, στην αριστερή μεριά είναι ο κωδικοποιητής, ενώ στα δεξιά ο αποκωδικοποιητής. Πιο συγκεκριμένα ο κωδικοποιητής, είναι υπεύθυνος για την διαμόρφωση των αναπαραστάσεων της εισερχόμενης ακολουθίας. Ενώ από την άλλη ο αποκωδικοποιητής είναι αυτός που μετατρέπει το αποτέλεσμα του κωδικοποιητή πίσω σε ακολουθίες, τις οποίες θα εξάγει το μοντέλο στο τέλος. Επιπλέον και οι δύο αυτοί μηχανισμοί αποτελούνται από πολλά επίπεδα (τα Nx στο σχήμα 3) με 2 και 3 υπερεπίπεδα ο κάθε ένας. Τα επίπεδα αυτά είναι αρχικά το Multi Head Attention το οποίο είναι υπεύθυνο ώστε τα κλειδιά, οι τιμές αλλά και τα ερωτήματα να εμφανίζονται γραμμικά ώστε αργότερα να μπορέσει να εκτελεστεί παράλληλα η απαραίτητη συνάρτηση. Σαν χαρακτηριστικό η τεχνολογία αυτή, έχει το ότι καθιστάται εφικτή η χρήση διαφορετικών αναπαραστάσεων των υποχώρων σε διαφορετικούς χώρους. Έπειτα είναι και το Masked Multi Head Attention το οποίο χρησιμοποιείται μόνο από τον κωδικοποιητή. Το οποίο είναι και το πρώτο επίπεδο στον αποκωδικοποιητή και εκτελεί την multi head attention στο παραγόμενο του αποκωδικοποιητή. Το παραγόμενο αυτό είναι κωδικοποιημένο ώστε να αποτρέπονται οι προβλέψεις που είναι βασισμένες σε πληροφορία που δεν χρειάζεται να είναι γνωστή, σε συγκεκριμένη θέση. Στη συνέχεια στο δίκτυο Feed Forward εφαρμόζονται πανομοιότυποι γραμμικοί μετασχηματισμοί στην κάθε θέση ξεχωριστά (Godbole,2020) .

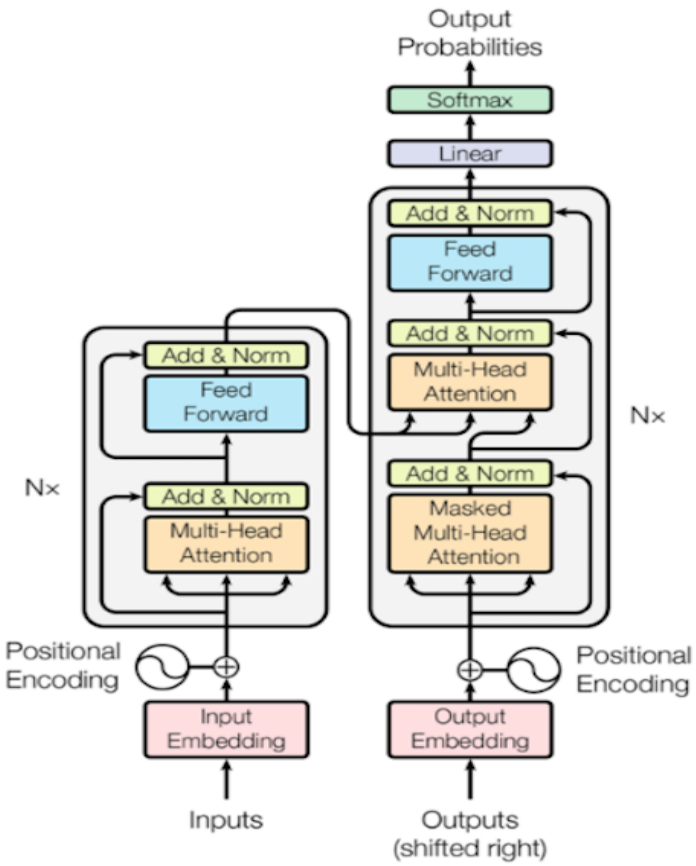
Πριν η πληροφορία εισέλθει στα επίπεδα αυτά, οι λέξεις των προτάσεων μπαίνουν σε σειρά με βάση την σχετική τους θέση σε μια ακολουθία. Αφού η συνάρτηση και οι μετασχηματισμοί εκτελεστούν απο τα επίπεδα, γίνονται άλλοι δύο μετασχηματισμοί, ένας γραμμικός (linear), και ένας softmax, ο οποίος μετατρέπει το τελικό αποτέλεσμα πίσω σε πιθανότητες.

Έστω ότι έχουμε τα παρακάτω παραδείγματα αναζήτησης από κάποιο χρήστη.

“2019 brazil traveler to usa need a visa”,

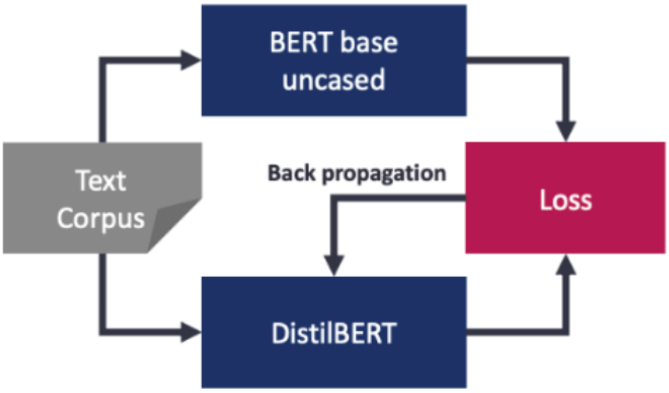
Το “to”, είναι πολύ σημαντικό σε αυτή την περίπτωση για την κατανόηση της σημασίας του κειμένου, η σημασία αυτή φαίνεται με την χρήση του BERT (Nayak , 2019).

Έτσι το BERT βοηθά να κατανοήσει η μηχανή την πρόταση, όπως το καταλαβαίνουν και οι ίδιοι οι άνθρωποι.



Σχήμα 3 Αρχιτεκτονικής ενός Μετασχηματιστή (Vaswani et al. 2017)

3.1.2.3. DistilBert



Σχήμα 4 Αναπαράστασης Εκπαίδευσης DistilBERT (Godbole, Grubinska & Kelnreiter, 2020)

Πέρα από τα ποικίλα πλεονεκτήματα που έχουν τα μεγαλύτερα μοντέλα, δημιουργούνται και πολλές ανησυχίες. Τα μεγαλύτερα μοντέλα έχουν περισσότερο κόστος και περισσότερες ανάγκες από ένα σύστημα, ανάγκες πόρων αλλά και μνήμης που μπορούν να αποτρέψουν την ευρεία υιοθεσία τους.

Έτσι δεδομένου του μεγαλύτερου μοντέλου BERT υλοποιήθηκε ένα μικρότερο μοντέλο γενικής χρήσης που βοηθά στην αναπαράσταση κειμένου, το DistilBERT. Το μοντέλο αυτό μπορεί να παρέχει καλή απόδοση σε διάφορες εφαρμογές, όπως άλλωστε και ο μεγαλύτερος ομόλογός του (Sanh, 2019). Εξάλλου αυτός είναι και ο σκοπός των μικρότερων προ-εκπαιδευμένων μοντέλων, να κρατούν την ευελιξία και την απόδοση των μεγαλύτερων μοντέλων όσο το δυνατόν περισσότερο και ταυτόχρονα να χρειάζονται μικρότερο κόστος πόρων για να χρησιμοποιηθούν.

Προσπάθησαν να μειώσουν το μέγεθος του BERT μοντέλου κατά 40%. Έτσι ο μεγαλύτερο μοντέλο BERT, χρησιμοποιείται για απόσταξη γνώσης(Knowledge distillation). Η απόσταξη γνώσης είναι μια τεχνική κατά την οποία ένα μικρότερο μοντέλο, ο μαθητής, είναι εκπαιδευμένο να αναπαράγει την συμπεριφορά ενός μεγαλύτερου μοντέλου, του δασκάλου του (Naushad , 2020).

Συγκεκριμένα ο μαθητής DistilBERT, έχει την ίδια αρχιτεκτονική με του δασκάλου του, BERT, αλλά με μικρότερο αριθμό επιπέδων ώστε να μειωθεί το μέγεθος του μοντέλου. Για να επιτευχθεί αυτό χρησιμοποιήθηκαν τεχνικές όπως η απώλεια εκπαίδευσης, η απώλεια απόσταξης, η απώλεια του MLM που χρησιμοποιούσε το BERT και η απώλεια ενσωμάτωσης συνημίτονου ώστε να ευθυγραμμιστεί η πορεία του μαθητή και του δασκάλου στα διανύσματα των κρυμμένων καταστάσεων. Αυτές οι απώλειες διασφαλίζουν ότι ο μαθητής εκπαιδεύεται όπως πρέπει και η μεταφορά γνώσης είναι αποδοτική. Επιπλέον αφαιρέθηκε η μέθοδος NSP από την εκπαίδευση του μοντέλου.

Το μοντέλο του DistilBERT εκπαιδεύτηκε στα ίδια δεδομένα όπως και το πρωταρχικό BERT μοντέλο. Εκπαιδεύτηκε σε 8 16G V100 GPU για περίπου 90 ώρες.

Το μοντέλο DistilBERT, που χρησιμοποιείται σαν μια προ-εκπαιδευμένη έκδοση του BERT, είναι 40% μικρότερο, 60% πιο γρήγορο ενώ ταυτόχρονα φαίνεται να διατηρεί το 97% των ικανοτήτων του αρχικού μοντέλου ενώ παράλληλα χρειάζεται μικρότερη υπολογιστική δύναμη για να τρέξει (Sanh et al.,2019) .

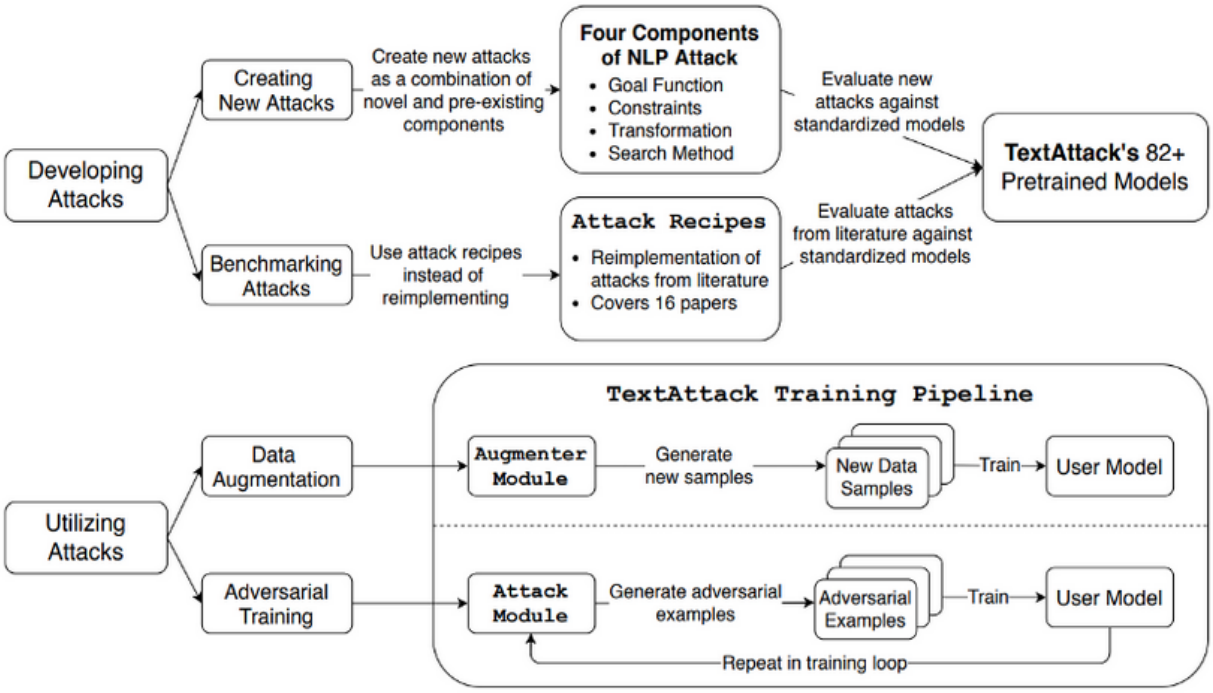
3.1.3. TextAttack

Το εργαλείο TextAttack (X.Morris, 2020) υλοποιήθηκε με σκοπό την διευκόλυνση των ερευνητών στο πεδίο τον επιθέσεων σε μοντέλα μηχανικής μάθησης. Αυτό έγινε με την παροχή ενός συνόλου από εργαλεία που μπορούν να επαναχρησιμοποιηθούν σε πολλά σενάρια, ώστε να εφαρμόζονται όσο το δυνατό περισσότερες επιθέσεις από τα θεωρητικά άρθρα. Με την χρήση του TextAttack μπορούν να εφαρμοστούν συγκεκριμένες επιθέσεις σε συγκεκριμένα μοντέλα, επιλέγοντας διαφορετικά δεδομένα και μετρικές κάθε φορά. Ενώ επίσης στο τέλος ο χρήστης μπορεί να έχει άμεση πρόσβαση στα αποτελέσματα που εξήγαγε το μοντέλο του. Η διαδικασία του Adversarial Attack μπορεί να εφαρμοστεί χρησιμοποιώντας τέσσερα βασικά συστατικά : μία συνάρτηση στόχου, κάποιους περιορισμούς, τους μετασχηματισμούς και την μέθοδο αναζήτησης.

Το εργαλείο TextAttack περιέχει κώδικα ώστε να φορτώνονται εύκολα και γρήγορα δημοφιλή NLP σύνολα δεδομένων ενώ παράλληλα υποστηρίζει ποικιλία από μοντέλα όπως π.χ BERT αλλά και άλλους μετασχηματιστές. Επίσης κάνουν εύκολη την πρόσβαση σε μεθόδους

επεξεργασίας ή να πολλαπλασιασμού των NLP δεδομένων, με χρήση augmenter. Με την χρήση αυτών τον μεθόδων μπορούν να επιλεχθούν πολλές διαφορετικές εφαρμογές (charswap, easy data augmentation, wordnet, embeddings, checklist κλπ), ενώ παράλληλα παρέχει τη δυνατότητα επιλογής σε πόσο μεγάλο βαθμό θα τροποποιηθούν τα δεδομένα.

Ενσωματώνοντας των κώδικα του TextAttack για φόρτωση δεδομένων αλλά και επεξεργασία ή πολλαπλασιασμό αυτών, με τις επιθέσεις σε μοντέλα μηχανικής μάθησης, το TextAttack παρέχει ένα περιβάλλον στους ερευνητές ώστε να κάνουν δοκιμές σε πολλά διαφορετικά σενάρια.



Σχήμα 5 Εφαρμογές του TextAttack(X. Morris et.al. 2020)

3.1.4. Μηχανισμός Επίθεσης Charswap Augmentation

Στα πλαίσια της εργασίας το εργαλείο TextAttack χρησιμοποιήθηκε αφού έγιναν οι απαραίτητες εγκαταστάσεις, ώστε να γίνουν τροποποιήσεις στα δεδομένα των παραπάνω δεδομένων, σε επίπεδο χαρακτήρα (charswap).

```
from argparse import ArgumentDefaultsHelpFormatter, ArgumentError, ArgumentParser
import csv
import os
import time

import tqdm

import textattack
from textattack.commands import TextAttackCommand

AUGMENTATION_RECIPE_NAMES = {
    "wordnet": "textattack.augmentation.WordNetAugmenter",
    "embedding": "textattack.augmentation.EmbeddingAugmenter",
    "charswap": "textattack.augmentation.CharSwapAugmenter",
    "eda": "textattack.augmentation.EasyDataAugmenter",
    "checklist": "textattack.augmentation.CheckListAugmenter",
}

recipe = "charswap"
pct_words_to_swap=0.50,
transformations_per_example=1
augmenter = eval(AUGMENTATION_RECIPE_NAMES[recipe])(
    pct_words_to_swap=0.50,
    transformations_per_example=1,
)
```

Απόσπασμα κώδικα 2 Ορισμός TextAttack και χρήση Augmenter

Η τιμή της μεταβλητής pct_words_to_swap που φαίνεται στο απόσπασμα κώδικα 2, μπορεί να πάρει τιμές από 0.1-1.0 και προσδιορίζει τον βαθμό των αλλαγών που θα έχει κάθε πρόταση. Στον πίνακα 3 παρουσιάζεται ένα παράδειγμα χρήσης του Augmenter charswap.

Κείμενο πριν την επίθεση	Economic growth in Japan slows down as the country experiences a Αναγνώριση drop in domestic and corporate spending.
Κείμενο μετά την επίθεση με συντελεστή αλλαγής λέξεων = 0.5	nconomic grpowth in apan skows down as the country experiences a dorp in domEstic and corpoxrate Cpending.

Πίνακας 3 Πρόταση πριν και μετά την επίθεση με χρήση Charswap

3.1.5. Μηχανισμός άμυνας μοντέλου κατά τις επιθέσεις σε επίπεδο χαρακτήρα

Σαν μηχανισμό άμυνας, η προσέγγιση που επιλέχθηκε για τα πλαίσια της εργασίας, αναπτύχθηκε απο τους Pruthi, Dhingra & C. Lipton (2019). Η προσέγγισή τους αυτή μπορεί να διαχειριστεί τροποποιημένα κείμενα που προορίζονται σαν επιθέσεις σε μοντέλα και περιέχουν εισαγωγές, διαγραφές, εναλλαγές χαρακτήρων, ή ακόμα και ορθογραφικά λάθη. Ουσιαστικά χρησιμοποιούν ένα μοντέλο βασισμένο στο RNN μοντέλο, με τρεις διαφορετικές στρατηγικές για αναγνώριση λέξεων από το μοντέλο.

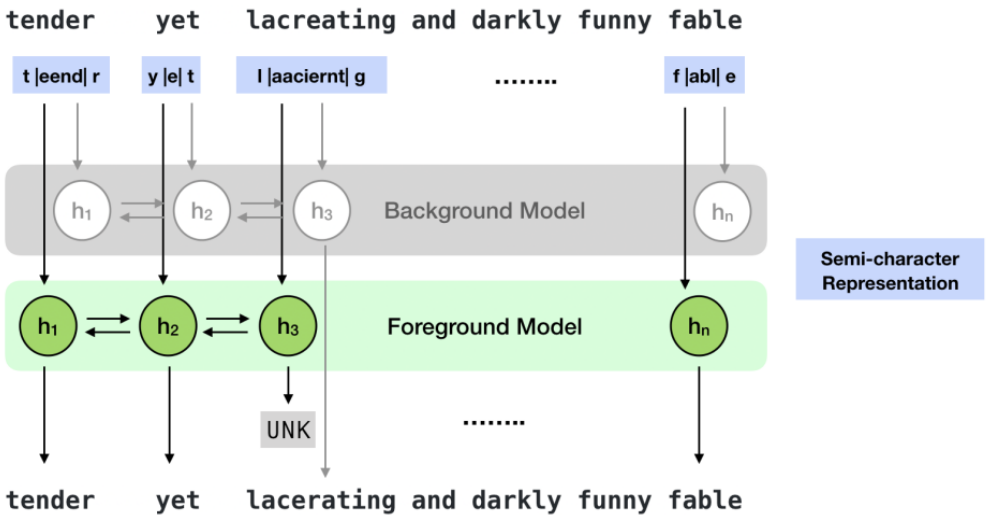
Ειδικότερα για την καταπολέμηση των ορθογραφικών λαθών που παράγονται για σκοπούς παραπλανήσεις των μοντέλων, προτείνουν να προστίθεται ένα μοντέλο αναγνώρισης λέξεων (Word Recognition Model), πριν από τον κατηγοριοποιητή. Από τα πειραματικά τους αποτελέσματα, ήταν φανερό ότι ακόμη και η αλλαγή ενός στρατηγικά επιλεγμένου χαρακτήρα μπορεί να επιφέρει μείωση στην απόδοση του μοντέλου BERT από ένα ποσοστό 90.3% στο 45%. Εφαρμόζοντας το μοντέλο αναγνώρισης λέξης σαν μηχανισμό άμυνας επανέφερε την απόδοση του BERT στο 75%.

Στις μέρες μας εξάλλου οι επιθέσεις ορθογραφικών λαθών αποτελούν ένα πραγματικό πρόβλημα καθώς μπορούν να χρησιμοποιηθούν με διάφορους τρόπους. Για παράδειγμα στα ανεπιθύμητα μηνύματα του ηλεκτρονικού ταχυδρομείου, μπορούν να χρησιμοποιηθούν αντίστοιχες επιθέσεις ώστε να επιτεθούν στους ίδιους τους διακομιστές του ηλεκτρονικού ταχυδρομείου. Προσπαθούν να προωθούν διακριτικά τροποποιημένα μηνύματα χωρίς να μπορεί να γίνει άμεση ανίχνευση λάθους, διατηρώντας πάντα το νόημα της πρότασης. Ένα άλλο παράδειγμα χρήσης των ορθογραφικά λάθος λέξεων, είναι η λογοκρισία του διαδικτύου που ωθεί τις κοινότητες να υιοθετούν παρόμοιες μεθόδους ώστε να επικοινωνούν μεταξύ τους επανειλημμένα.

Στην έρευνα τους, οι δημιουργοί του συγκεκριμένου μοντέλου αναγνώρισης λέξεων, παρουσιάζουν ότι η απόδοση ενός κατηγοριοποιητή μπορεί να μειωθεί με την τροποποίηση το πολύ δύο χαρακτήρων σε κάθε πρόταση. Ενώ παράλληλα αυτές οι τροποποιήσεις μπορούν είτε να μετατρέψουν την λέξη που θα δεχθεί την συγκεκριμένη επίθεση σε κάποια άλλη λέξη είτε θα

την μετατρέψουν σε μια λέξη η οποία δεν υπάρχει στην πραγματικότητα. Τις λέξεις αυτές στην υλοποίηση τους τις χαρακτηρίζουν σαν UNK. Συνεπώς, όπως αναφέρουν, με την χρήση του μοντέλου αναγνώρισης λέξεων υπάρχουν δύο οφέλη, από την μία το μοντέλο μπορεί να εκπαιδευτεί σε μεγαλύτερο σύνολο δεδομένων οποιαδήποτε στιγμή ανεξάρτητα, ενώ επίσης μπορεί να χρησιμοποιηθεί σε διάφορες εφαρμογές ή μοντέλα για σκοπούς κατηγοριοποίησης.

Παρόλα αυτά όμως, όσο αφορά τα στοχευμένα λάθη υπάρχουν δύο σημαντικοί παράγοντες που πρέπει να ληφθούν υπόψη για την εξάλειψή τους. Η διασφάλιση της απόδοσης του μοντέλου αναγνώρισης λέξεων και η ευαισθησία του ίδιου του κειμένου που εισάγεται στο μοντέλο.



Σχήμα 6 Αναπαράστασης αρχιτεκτονικής του μοντέλου αναγνώρισης λέξεων (Pruthi et.al, 2019)

Στο σχήμα 6, φαίνονται τα δύο υπό μοντέλα που χρησιμοποιούνται για την υλοποίηση του μοντέλου ScRNN. Αρχικά εκπαιδεύουν το foreground μοντέλο σε μικρότερο σύνολο δεδομένων και το backround μοντέλο σε μεγαλύτερα σύνολα δεδομένων όπως π.χ. το IMDB. Παράλληλα όμως και τα δύο μοντέλα εκπαιδεύονται ώστε να μπορούν να ανακατασκευάζουν την ορθή ορθογραφικά μορφή της κάθε λέξης, αλλά και του περιεχομένου της όταν αυτή αποτελείται από υπό λέξεις ή λέξης της ίδιας οικογένειας. Αυτό το πετυχαίνουν αφού κατά την εκπαίδευση

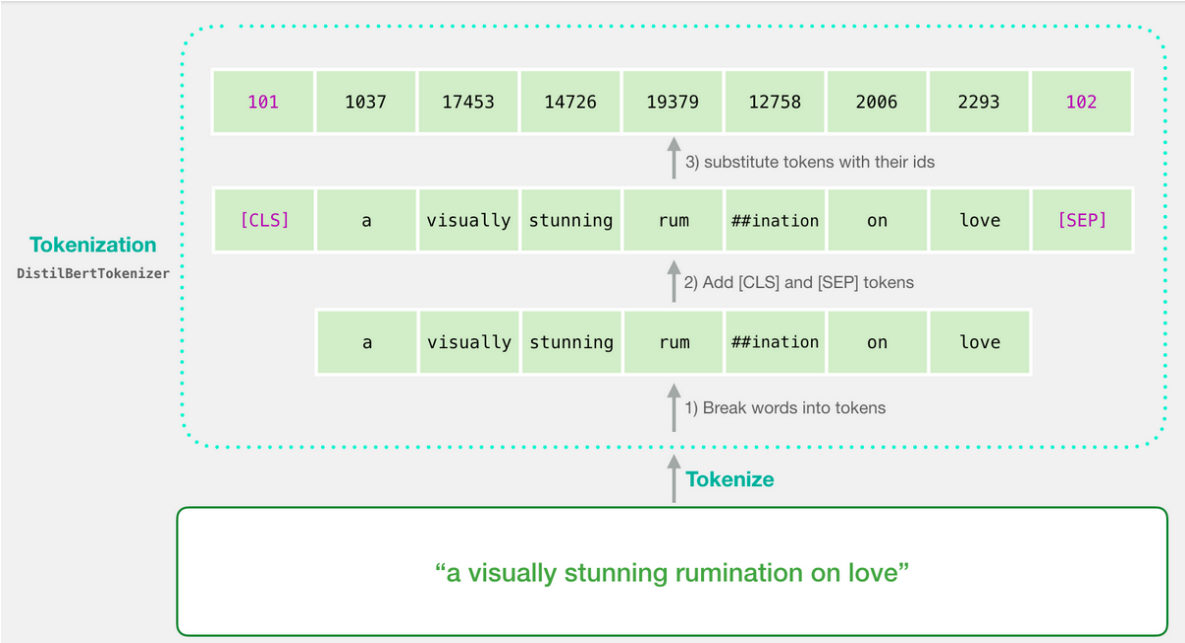
εισάγουν συνθετικά κατεστραμμένες λέξεις. Το backround μοντέλο χρησιμοποιείται όταν φτάσουν σε μια λέξη που χαρακτηρίζεται σαν UNK.

Το μοντέλο της άμυνας χρειάστηκε να εκπαιδευτεί σε δεδομένα που χρησιμοποιήθηκαν στα πλαίσια της παρούσας εργασίας ώστε να επιτευχθούν καλύτερες αποδόσεις.

3.1.6. Διαδικασίες σεναρίου επίθεσης και άμυνας

Αρχικά έχοντας αποφασίσει πιο σύνολο δεδομένων θα χρησιμοποιηθεί, το αρχείο αυτό πρέπει να σπάσει σε δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου, και σε δεδομένα που θα χρησιμοποιηθούν για τον έλεγχο του μοντέλου. Η διαδικασία αυτή γίνεται στην αρχή ώστε και στις τρεις φάσεις να χρησιμοποιείται το ίδιο ακριβώς σύνολο και τα αποτελέσματα να είναι πιο ακριβής και να είναι εύκολο να συγκριθούν.

Στη συνέχεια αφού χωριστούν τα δεδομένα, απαιτείται μια προ επεξεργασία τους ώστε να έρθουν σε μορφή στην οποία μπορεί πλέον να τα επεξεργαστεί το μοντέλο BERT/DistilBERT. Για να επιτευχθεί αυτό χρειάζεται να γίνουν τα παρακάτω βήματα.



Σχήμα 7 Προ επεξεργασίας κειμένων από DistilBertTokenizer (Alammar, 2019)

α) **Tokenizing**: Στο σημείο αυτό χρειάζεται τα δεδομένα να σπάσουν σε tokens, δηλαδή κάθε πρόταση να σπάσει σε λέξεις. Οι tokenizers που χρησιμοποιούν τα μοντέλα που επιλέχθηκαν (DistiBertTokenizer, BertTokenizer), πέρα από το απλό σπάσιμο σε λέξεις, προσπαθούν να βρίσκουν και την ρίζα τον λέξεων ώστε να γίνεται πιο εύκολα η κατηγοριοποίηση.

β) **Padding**: Στο επόμενο βήμα, οι προτάσεις πλέον είναι λίστες, σπασμένες σε tokens, άρα όλες μαζί είναι μια μεγάλη λίστα από λίστες. Καθώς κάθε πρόταση ίσως να έχει διαφορετικό μέγεθος από κάποια άλλη πρόταση, χρειάζεται να γίνει μια προ επεξεργασία ώστε να φέρουμε όλες τις προτάσεις στο ίδιο μέγεθος για να μπορεί να τις επεξεργαστεί το μοντέλο όλες με τη μία. Αυτή η διαδικασία είναι η διαδικασία του padding και φτιάχνει ένα 2-d πίνακα για τις προτάσεις.

γ) **Masking**: Το μοντέλο χρειάζεται αυτή την προ επεξεργασία γιατί αλλιώς θα μπερδευτεί με τα δεδομένα, έτσι ουσιαστικά φτιάχνει τη μάσκα για να αγνοήσει το μοντέλο το (β) όταν κάνει επεξεργασία του εισαγόμενου κειμένου.

Έπειτα στο σημείο της κατηγοριοποίησης, το μοντέλο προσθέτει ένα επιπλέον κομμάτι [CLS], το οποίο χαρακτηρίζει την κλάση-κατηγορία στην οποία ανήκει το συγκεκριμένο κείμενο, όπως φαίνεται και στο σχήμα παραπάνω (Alammar ,2019).

Η διαδικασία της προ επεξεργασίας των δεδομένων γίνεται σε κάθε μία από τις τρεις φάσεις, καθώς τα κείμενα τροποποιούνται μετά τις επιθέσεις/άμυνες και χρειάζεται να περάσουν ξανά από όλα τα βήματα ώστε να μπορέσει να τα επεξεργαστεί μετά ξανά το μοντέλο.

Τελικά χρησιμοποιώντας SGDRegressor γίνονται οι προβλέψεις σε κάθε φάση, στα ίδια δεδομένα ελέγχου. Ενώ στη συνέχεια χρησιμοποιώντας Stratified K-fold Cross Validation αξιολογείται η απόδοση του μοντέλου σε κάθε φάση.

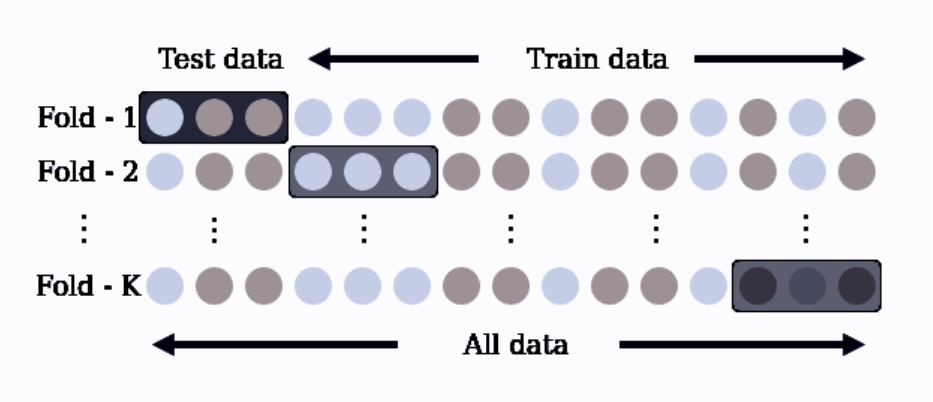
3.1.6.1. K-Fold Cross Validation

Η στατιστική μέθοδος Cross Validation, χρησιμοποιείται για να αξιολογεί τις αποδόσεις των μοντέλων μηχανικής μάθησης. Συνήθως χρησιμοποιείται σε εφαρμογές μηχανικής μάθησης καθώς είναι ένα εύκολα κατανοήσιμο και υλοποιήσιμο εργαλείο. Αποτελείται από μια δειγματοληπτική διαδικασία που αξιολογεί τα μοντέλα μηχανικής μάθησης σε περιορισμένο αριθμό από δείγματα δεδομένων (Brownlee, 2021). Η διαδικασία αυτή χρειάζεται μια

παράμετρο k, που προσδιορίζει τον αριθμό των ομάδων, στις οποίες θα χωριστούν τα δεδομένα ώστε να αξιολογηθούν. Εξαιτίας αυτού η διαδικασία καλείται συχνά και σαν k-fold cross validation. Όταν για παράδειγμα η τιμη του k=5 , τότε μπορεί να αναφερθεί σαν 5-fold cross validation. Η τιμή του k, χρειάζεται να επιλέγετε προσεκτικά σε σχέση με τα δεδομένα, διαφορετικά μπορεί να μην φέρει αντιπροσωπευτικά αποτελέσματα για το συγκεκριμένο μοντέλο.

Συμπληρωματικά το σύνολο δεδομένων που διατίθεται για αξιολόγηση σπάει σε k υποσύνολα, τα οποία κατά προσέγγιση έχουν το ίδιο μέγεθος. Στη συνέχεια το μοντέλο εκπαιδεύεται χρησιμοποιώντας k-1 υποσύνολα, που όλα μαζί θα αναπαριστούν το σύνολο δεδομένων εκπαίδευσης του μοντέλου. Το υπολειπόμενο υποσύνολο από δεδομένα, χρησιμοποιείται σαν validation σύνολο και βάση αυτού υπολογίζεται η απόδοση του μοντέλου.

Ομοίως, η διαδικασία αυτή επαναλαμβάνεται μέχρι κάθε ένα από τα k υποσύνολα να έχει χρησιμοποιηθεί και σαν validation υποσύνολο για την αξιολόγηση του μοντέλου. Έτσι λοιπόν η μέση τιμή των k αποδόσεων από τα k validation υποσύνολα που υπολογίζεται είναι και η τιμή απόδοσης του μοντέλου χρησιμοποιώντας αυτή την στατιστική μέθοδο.



Σχήμα 8 K-Fold Cross Validation(Schmid, 2020)

Όπως φαίνεται στο σχήμα 8 η διαδικασία που ακολουθείται για το 1ο fold είναι η εξής, το 1ο υποσύνολο χρησιμοποιείται σαν validation και το υπόλοιπο σαν σύνολο εκπαίδευσης, και αυτό επαναλαμβάνεται μέχρι όλα τα δεδομένα να χρησιμοποιηθούν για την αξιολόγηση του μοντέλου, δηλαδή σαν validation σύνολο.

Έπειτα μια επιπλέον δυνατότητα που προσφέρει η συγκεκριμένη τεχνική, είναι η Stratified Random Sampling, η οποία φροντίζει σε κάθε υποσύνολο να αναπαρίστανται οι ομάδες κατηγοριοποίησης, στον ίδιο βαθμό με του αρχικού συνόλου δεδομένων. Δηλαδή υπόσχεται ότι κάθε υποσύνολο είναι μια μικρότερη αναπαράσταση του μεγάλου συνόλου δεδομένων. Αυτό συνήθως χρησιμοποιείται όταν δεν υπάρχουν πολλά δεδομένα ώστε να αφιερωθούν μόνο για λόγους αξιολόγησης του μοντέλου. Για αυτό και στα πλαίσια της εργασίας χρησιμοποιήθηκε η τεχνική αυτή.

3.2. Σενάριο 2 – Αλλαγή σημασιολογίας λέξεων

Για το συγκεκριμένο σενάριο ακολουθήθηκε παρόμοια μεθοδολογία με αυτή των δημιουργών της μεθόδου SEM. Αρχικά χρειάστηκε να εκπαιδευτούν εξ αρχής τα μοντέλα. Στα πλαίσια της εργασίας τα μοντέλα εκπαιδεύτηκαν μόνο στο σύνολο δεδομένων IMDB. Εκπαιδεύτηκαν συνολικά 4 διαφορετικά μοντέλα δύο RNN, ένα χωρίς την ενσωμάτωση του κωδικοποιητή για χρήση της άμυνας και ένα με την ενσωματωμένη άμυνα και αντίστοιχα δύο CNN, με το μεγαλύτερο μέγεθος λεξιλογίου να είναι ίσο με 30000 λέξεις.

Αφού εκπαιδεύτηκαν τα μοντέλα, χρειάστηκε να τρέξει το σενάριο στο οποίο θα χρησιμοποιούνταν τα αρχικά δεδομένα ώστε να έχουμε το αρχικό Accuracy των μοντέλων. Στη συνέχεια τα μοντέλα χρησιμοποιήθηκαν ώστε να προβλέψουν τις κατηγορίες ακολουθιών κειμένου στα οποία υπήρξε επίθεση WordNet είτε επίθεση IGA, και αντίστοιχα για να προβλέψουν τιμές μετά την άμυνα ώστε τελικά να συγκριθούν μεταξύ τους. Η αξιολόγηση των αποδόσεων των μοντέλων, έγινε χρησιμοποιώντας Confusion-matrix συγκρίνοντας τις πραγματικές τιμές των κατηγοριών στις οποίες ανήκαν οι ακολουθίες με τις αντίστοιχες τιμές τις οποίες προέβλεψαν τα μοντέλα.

Για το σενάριο αυτό χρησιμοποιήθηκε το σύνολο δεδομένων IMDB. Όπως αναφέρθηκε παραπάνω (βλ 3.1.1.2) τα δεδομένα από το IMDB περιέχουν σχόλια ταινιών τα οποία χωρίζονται σε δύο κατηγορίες positive/negative ανάλογα με το περιεχόμενό τους.

3.2.1. Μοντέλα RNN – CNN για κείμενα

Ο κύριος ρόλος των Νευρωνικών Δικτύων είναι να αναπαριστούν το μέγεθος μια μεταβλητής κειμένου σαν ένα σταθερό μήκος διανύσματος. Αυτά τα μοντέλα αποτελούνται από ένα επίπεδο προβολής το οποίο χαρτογραφεί λέξεις, υπό-λέξεις, μέρη λέξεων ή κ-μέρη σε αναπαραστάσεις διανυσμάτων και μετά συνδυάζονται με τις διαφορετικές αρχιτεκτονικές των νευρωνικών δικτύων.

3.2.2.1. Μοντέλα RNN (Recurrent Neural Network)

Τα μοντέλα RNN, είναι πολύ δημοφιλές αφού χρησιμοποιούνται πολύ συχνά σε NLP προβλήματα μιας και η επαναλαμβανόμενη δομή τους είναι ιδανική για την επεξεργασία του μήκους των κειμένων. Τα μοντέλα RNN είναι ικανά να επεξεργάζονται αλληλουχίες αυθαίρετου μήκους εφαρμόζοντας αναδρομικά μια μεταβατική συνάρτηση στην εσωτερική κρυμμένη κατάσταση του διανύσματος της εισερχόμενης πρότασης (Liu , Qiu and Huang, 2016) .

Τα RNN δουλεύουν σε τρεις φάσεις (Olah,2015) α) κινούνται προς το κρυμμένο επίπεδο και κάνουν προβλέψεις, β) συγκρίνουν τις προβλέψεις με την πραγματική τους τιμή, χρησιμοποιώντας συναρτήσεις απώλειας (loss functions) και γ) χρησιμοποιώντας τις λανθασμένες τιμές σε ανάστροφη διάδοση (back propagation) υπολογίζοντας περαιτέρω την κλίση για το κάθε σημείο ή κόμβο.

Τα RNN μοντέλα χρησιμοποιούνται συνήθως για να διαχειρίζονται σειριακά δεδομένα, δηλαδή δεδομένα που διακατέχονται το ένα το άλλο με κάποια σειρά. Αυτό συμβαίνει γιατί τα μοντέλα αυτά χρησιμοποιούν επίπεδα που του δίνουν μικρής διάρκειας μνήμη. Έτσι χρησιμοποιώντας αυτή την μνήμη, μπορούν να προβλέψουν την επόμενη τιμή με περισσότερη ακρίβεια. Ο χρόνος αυτός για τον οποίο παραμένει η μνήμη εξαρτάται από τα βάρη που διαθέτει το κάθε μοντέλο και έτσι δεν είναι πάντα σταθερός.

Γενικότερα τα RNN χρησιμοποιούνται για ανάλυση συναισθημάτων, κατηγοριοποίηση ακολουθιών, για αναγνώριση φωνής κλπ.

3.2.2.2. Μοντέλα CNN (Convolutional Neural Networks)

Σκοπός τους είναι η επεξεργασία δεδομένων τα οποία υπάγονται σε ένα γνωστό δίκτυο σαν μια τοπολογία (Namatēvs, 2017). Αρχικά αναπτύχθηκαν σαν νευρωνικά δίκτυα για επεξεργασία εικόνων, όπου και πέτυχαν, αφού σήμερα είναι ευρέως χρησιμοποιημένα για να αναγνωρίζουν αντικείμενα σε εικόνες από μια προκαθορισμένη μορφή εικόνας.

Ο Goldberg(2017) ορίζει το CNN σαν ένα επίπεδο που έχει την ευθύνη να εξάγει δομές οι οποίες θα προσφέρουν χρήσιμη πληροφορία για την γενική διαδικασία πρόβλεψης. Ένα CNN είναι σχεδιασμένο να αναγνωρίζει τοπικές προβλέψεις σε μια μεγαλύτερη δομή, και συνδυάζοντας τις να παράγει μια αναπαράσταση διανύσματος σταθερού μεγέθους. Στην περίπτωση που το CNN θα χρησιμοποιηθεί για NLP, η αρχιτεκτονική του θα αναγνωρίσει τα μέρη της πρότασης τα οποία είναι προβλέψιμα, χωρίς να υπάρχει η ανάγκη να προκαθοριστεί ένα ενσωματωμένο διάνυσμα για κάθε πιθανό μέρος.

Για να χρησιμοποιηθεί ένα CNN χρειάζεται να εκπαιδευτεί πρώτα μαζί με ένα επίπεδο κατηγοριοποίησης, ώστε να αναπαράγει χρήσιμα συμπεράσματα. Ένα CNN τυπικά περιλαμβάνει δύο διαδικασίες convolutional + pooling, με το αποτέλεσμα της αλληλουχίας αυτών των διαδικασιών να είναι συνδεδεμένο με ένα εξ ολοκλήρου συνδεδεμένο επίπεδο το οποίο αρχικά είναι ίδιο με το παραδοσιακό πολύ-επίπεδο που έχουν τα νευρωνικά δίκτυα (Batista,2018).

Επιπλέον, όταν τα CNN εφαρμόζονται σε NLP, διαδικασίες που αφορούν κείμενο, παράγουν ένα πίνακα μιας διάστασης (1D) για την αναπαράσταση του κειμένου. Η πιο συχνή χρήση των CNN σε NLP δεδομένα, είναι για την κατηγοριοποίηση ακολουθιών κειμένου, κατηγοριοποιώντας τους χαρακτήρες αλλά και τις λέξεις σε προ-υπάρχουσες κατηγορίες.

3.2.2. Μηχανισμός επίθεσης IGA – Improved Genetic based on Text Attack

Οι δημιουργοί (Wang, 2019) της μεθόδου SEM σε μια προσπάθεια να αξιολογήσουν την αποδοτικότητα του συστήματος, δημιούργησαν μία επιπλέον επίθεση, την IGA. Οι τρέχουσες

επιθέσεις αντικατάστασης συνώνυμων λέξεων, είναι βασισμένες σε περιορισμούς, ώστε να υπάρχουν συνώνυμες λέξεις στην ίδια θέση τουλάχιστον μια φορά ή να αντικαθίστανται λέξεις με προκαθορισμένο αριθμό συνωνύμων στην αρχική αυθεντική πρόταση. Αυτός ο περιορισμός οδηγεί σε προβληματισμούς, όπως για παράδειγμα ποιος αριθμός συνωνύμων να επιλεχθεί ώστε να είναι κατάλληλος, αφού η κάθε λέξη μπορεί να έχει διαφορετικό αριθμό από συνώνυμα σε σχέση με κάποια άλλη λέξη. Επομένως, για να το αντιμετωπίσουν αυτό το πρόβλημα, προτείνουν την επίθεση IGA. Η IGA επιτρέπει στις συνώνυμες λέξεις να υπάρξουν στην ίδια θέση περισσότερες από μία φορές. Έτσι η επίθεση αυτή μπορεί να διασχίζει όλες τις συνώνυμες λέξεις ασχέτως από την τιμή των συνώνυμων.

3.2.3. Μηχανισμός Επίθεσης WordNet Augmentation

Σαν επιπλέον επίθεση χρησιμοποιήθηκε το WordNet, χρησιμοποιώντας το εργαλείο TextAttack το οποίο χρησιμοποιήθηκε επίσης στο πρώτο σενάριο, για την αντικατάσταση λέξεων από το εισαγόμενο κείμενο, με συνώνυμες τους λέξεις.

Το WordNet (Fellbaum, 2005) είναι μια τεράστια λεξικολογική βάση, βασισμένη στην αγγλική γλώσσα, περιλαμβάνοντας ουσιαστικά, ρήματα, επιρρήματα, επίθετα ομαδοποιημένα ανάλογα με την σημασιολογική τους έννοια. Τα σύνολα αυτά συνδέονται εσωτερικά βάση της εννοιολογικής σημασίας των λέξεων που τα απαρτίζουν, ενώ παράλληλα συνδέονται και με άλλα σύνολα με παρόμοιες σημασίες. Έτσι δημιουργούνται δίκτυα από εννοιολογικά συγγενικές λέξεις ή και φράσεις.

Η δομή του αυτή το καθιστά ένα χρήσιμο εργαλείο για υπολογισμούς που αφορούν την γλώσσα ενώ παράλληλα είναι χρήσιμο και για την επεξεργασία πραγματικής γλώσσας (NLP). Παρόλα αυτά το εργαλείο αυτό, ομαδοποιεί αυτόνομες λέξεις με κοινή έννοια χωρίς να λαμβάνει υπόψη την συνολική έννοια μιας λέξης σε ολόκληρη την πρόταση. Έτσι κάθε λέξη μπορεί να έχει πολλές διαφορετικές σημασίες και επομένως η συγκεκριμένη λέξη πολύ πιθανό να είναι μέλος σε όσα σύνολα όσα και οι έννοιες της. Άρα κάθε σύνολο με διαφορετική σημασία είναι μοναδικό και συμπεριλαμβάνει ένα σύντομο ορισμό(“gloss”) που το περιγράφει.

```
import textattack
from textattack.commands import TextAttackCommand

AUGMENTATION_RECIPE_NAMES = {
    "wordnet": "textattack.augmentation.WordNetAugmenter",
    "embedding": "textattack.augmentation.EmbeddingAugmenter",
    "charswap": "textattack.augmentation.CharSwapAugmenter",
    "eda": "textattack.augmentation.EasyDataAugmenter",
    "checklist": "textattack.augmentation.CheckListAugmenter",
}

recipe = "wordnet"

pct_words_to_swap=0.50,
transformations_per_example=1
augmenter = eval(AUGMENTATION_RECIPE_NAMES[recipe])(
    pct_words_to_swap=0.50,
    transformations_per_example=1,
)
```

Απόσπασμα κώδικα 3 Augmentation με χρήση του WordNet

Η τιμή της μεταβλητής pct_words_to_swap που φαίνεται στο απόσπασμα κώδικα 3, μπορεί να πάρει τιμές από 0.1-1.0 και προσδιορίζει τις αλλαγές που θα έχει κάθε πρόταση π.χ. για 0.5 το 50% των λέξεων της πρότασης θα αλλάξουν. Βέβαια η συγκεκριμένη επίθεση δεν εγγυάται ότι η πρόταση μετά την επίθεση θα έχει το ίδιο νόημα με πριν, αφού βλέπει την κάθε λέξη χωριστά χωρίς να βλέπει το νόημα της στην πρόταση.

Κείμενο πριν την επίθεση	this is the greatest movie ever if you have written it off with out ever seeing it you must give it a second try
Κείμενο μετά την επίθεση με συντελεστή αλλαγής λέξεων = 0.5	this is the majuscule film e'er if you have spell it off with out e'er envision it you mustiness throw it a instant test
Κείμενο πριν την επίθεση	this is the greatest movie ever if you have written it off with out ever seeing it you must give it a second try
Κείμενο μετά την επίθεση με με συντελεστή αλλαγής λέξεων = 0.2	this is the swell movie ever if you have pen it off with out ever seeing it you must open it a second taste

Πίνακας 4 Παραδείγματα χρήσης Augmenter WordNet

Στον πίνακα 4, παρουσιάζονται κάποια παραδείγματα χρήσης του WordNet, ανάλογα με το βαθμό αλλαγής που ορίζεται κάθε φορά.

3.2.4. Μηχανισμός Άμυνας SEM – Synonym Encoder Method

Για την προστασία των μοντέλων από επιθέσεις βασισμένες στην αλλαγή του νοήματος της πρότασης, χρησιμοποιώντας συνώνυμα, στην δημοσίευση "Natural Language Adversarial Attack and Defense in Word Level" (Wang et al., 2019), προτείνεται η μέθοδος SEM(Synonym Encoding Method). Η μέθοδος αυτή προσθέτει στο μοντέλο ένα κωδικοποιητή(encoder) πριν από τον κατηγοριοποιητή(classifier) και κοιτάζει για εναλλαγές ενώ αργότερα εκπαιδεύεται στα πραγματικά δεδομένα. Έτσι το πρόβλημα που χρειάζεται να αντιμετωπιστεί είναι το πως θα εντοπιστούν οι γείτονες του κάθε σημείου ή λέξης. Δεδομένου ότι τα κείμενα είναι διακριτά κομμάτια, είναι εύκολο με χρήση διαφόρων τεχνικών ή αλγορίθμων να εντοπιστούν οι γείτονες. Βασισμένοι σε αυτά, προτείνουν τη νέα αυτή μέθοδο κωδικοποίησης συνωνύμων, ώστε να εντοπίζονται οι γείτονες της εισερχόμενης ακολουθίας.

Σε αυτή την προσέγγιση, ουσιαστικά ομαδοποιώντας (clustering) και κωδικοποιώντας τα συνώνυμα, παράγεται ένα μοναδικό κωδικό κλειδί, ώστε να αναγκάζονται όλες οι συνώνυμες λέξεις να έχουν ένα κοινό κωδικό κλειδί κατά την ενσωμάτωση. Η μέθοδος αυτή αξιολογήθηκε αφού εκπαιδεύτηκε σε δεδομένα που είχαν δεχθεί επιθέσεις καθιστώντας την καλύτερη τεχνική όσον αφορά στην αποφυγή επιθέσεων σε συνώνυμα.

Συγκεκριμένα για να φτιαχτούν οι ομάδες, θεώρησαν ότι όσο πιο κοντά είναι το νόημα δύο προτάσεων τόσο πιο κοντά είναι και η απόστασή τους στο διάστημα και έτσι υποθέτουν ότι οι γείτονες της κάθε πρότασης x, είναι συνώνυμες προτάσεις. Επομένως λοιπόν καταλήγουν στις ομάδες (clusters) με τα μοναδικά κλειδιά, ενώ παράλληλα εφαρμόζουν την μέθοδο αυτή σε Glove διανύσματα, που εισάγει τους περιορισμούς για τα αντίθετα και τα συνώνυμα σε αναπαραστάσεις διανυσμάτων.

Για την αξιολόγηση της μεθόδου αυτής, επέλεξαν τυχαία 200 σωστά κατηγοριοποιημένα παραδείγματα προτάσεων, σε διαφορετικά μοντέλα από διαφορετικά σύνολα δεδομένων και τα χρησιμοποιούν για να φτιάξουν παραδείγματα με άμυνα είτε χωρίς άμυνα. Όσο πιο

αποτελεσματικός ήταν ο μηχανισμός άμυνας τόσο πιο λίγο έπεφτε η απόδοση της κατηγοριοποίησης.

3.2.5. Διαδικασίες σεναρίου επίθεσης και άμυνας

Κατά την διαδικασία υλοποίησης του σεναρίου αυτού χρειάστηκε να εκπαιδευτούν τα μοντέλα ώστε να μπορούν να χρησιμοποιηθούν για τους σκοπούς του σεναρίου σε κείμενα ελέγχου. Χρησιμοποιήθηκαν και εκπαιδεύτηκαν δύο διαφορετικοί τύποι νευρωνικών δικτύων CNN και RNN, αρχικά χρειάστηκε να εκπαιδευτούν με τα αυθεντικά δεδομένα, ενώ αργότερα όπως αναφέρθηκε παραπάνω, χρειάστηκε να εκπαιδευτούν χρησιμοποιώντας την μέθοδο άμυνας, με την επιπλέον κωδικοποίηση πριν το επίπεδο της κατηγοριοποίησης των κειμένων σε κατηγορίες. Τα μοντέλα αυτά χρησιμοποίησαν λεξιλόγιο βασισμένο στα αυθεντικά δεδομένα, με μέγεθος 30000 λέξεις το πολύ.

Οι μετρικές που χρησιμοποιήθηκαν κατά το στήσιμο και την εκπαίδευση του μοντέλου είναι οι ίδιες μετρικές που είχαν προκαθορισμένες οι δημιουργοί των μοντέλων (Wang et al., 2019) όπως φαίνεται στο απόσπασμα κώδικα 4, με την διαφορά ότι μειώθηκαν οι εποχές σε 5 λόγο του τεράστιου κόστους πόρων που θα χρειαζόταν η εκπαίδευση διαφορετικά.

```
# Model Hyperparameters
tf.flags.DEFINE_boolean("enable_word_embeddings", True, "Enable/disable the word embedding (default: True)")
tf.flags.DEFINE_integer("embedding_dim", 128, "Dimensionality of character embedding (default: 128)")
tf.flags.DEFINE_integer("hidden_size", 128, "Number of hidden layer units (default: 128)")
tf.flags.DEFINE_integer("hidden_layers", 2, "Number of hidden layers (default: 2)")
tf.flags.DEFINE_string("filter_sizes", "3,4,5", "Comma-separated filter sizes (default: '3,4,5')")
tf.flags.DEFINE_integer("num_filters", 128, "Number of filters per filter size (default: 128)")
tf.flags.DEFINE_integer("rnn_size", 128, "Number of units rnn_size (default: 128)")
tf.flags.DEFINE_integer("num_rnn_layers", 3, "Number of rnn layers (default: 3)")
tf.flags.DEFINE_float("dropout_keep_prob", 0.5, "Dropout keep probability (default: 0.5)")
tf.flags.DEFINE_float("l2_reg_lambda", 0.0, "L2 regularization lambda (default: 0.0)")

# Training parameters
tf.flags.DEFINE_integer("batch_size", 64, "Batch Size (default: 64)")
tf.flags.DEFINE_integer("num_epochs", 5, "Number of training epochs (default: 15)")
tf.flags.DEFINE_integer("evaluate_every", 200, "Evaluate model on dev set after this many steps (default: 100)")
tf.flags.DEFINE_integer("checkpoint_every", 200, "Save model after this many steps (default: 100)")
tf.flags.DEFINE_integer("num_checkpoints", 2, "Number of checkpoints to store (default: 5)")
tf.flags.DEFINE_float("grad_clip", 5, "grad clip to prevent gradient explode")
tf.flags.DEFINE_float("decay_coefficient", 2.5, "Decay coefficient (default: 2.5)")
```

Απόσπασμα κώδικα 4 Μετρικές για Εκπαίδευση Μοντέλου

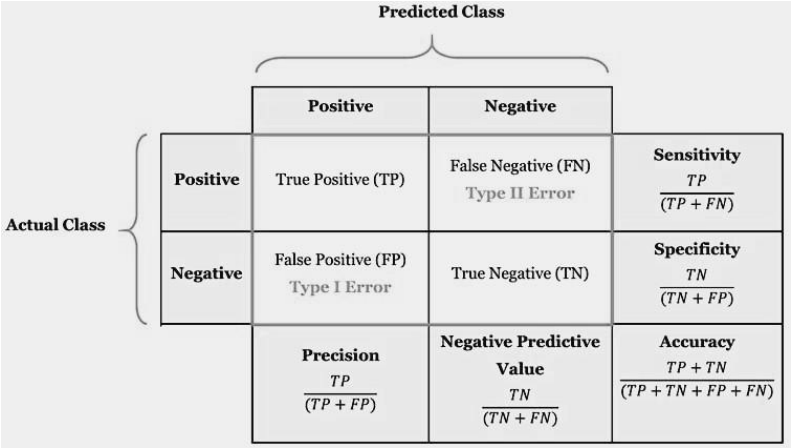
Έπειτα αφού τα μοντέλα είχαν πλέον εκπαιδευτεί δεν ήταν αναγκαίο να εκπαιδεύονται κάθε φορά που θα τρέχαμε κάποιο δοκιμαστικό σενάριο. Έτσι αργότερα χρησιμοποιήθηκε το μοντέλο που εκπαιδεύτηκε στα αυθεντικά δεδομένα, για να κάνει προβλέψεις σε δεδομένα που δεν δέχτηκαν κανένα είδος επίθεσης, σε συγκεκριμένο αριθμό δεδομένων. Στη συνέχεια το ίδιο μοντέλο χρησιμοποιήθηκε για να κάνει προβλέψεις σε δεδομένα που είχαν δεχτεί κάποιο είδος επίθεσης. Οι επιθέσεις που χρησιμοποιήθηκαν όπως προαναφέρθηκαν είναι η επίθεση IGA, και το WordNet για τροποποίηση των δεδομένων με τη βοήθεια του εργαλείου TextAttack. Έπειτα το μοντέλο με την επιπλέον κωδικοποίηση χρησιμοποιήθηκε σαν μοντέλο άμυνας, ώστε να κάνει προβλέψεις σε δεδομένα που είχαν δεχθεί επιθέσεις IGA ή WordNet.

Τα δεδομένα καταγράφονται και στις τρεις αυτές φάσεις, πριν την επίθεση, κατά την επίθεση και με τη χρήση άμυνας ώστε να συγκριθούν και να αξιολογηθούν αναλόγως.

Οι αποδόσεις των μοντέλων στα συγκεκριμένα δεδομένα στις τρεις φάσεις, αξιολογήθηκαν χρησιμοποιώντας ένα confusion matrix για τον υπολογισμό του Accuracy του μοντέλου. Αυτό έγινε συγκρίνοντας τις πραγματικές τιμές κατηγορίας στην οποία ανήκουν τα κείμενα με τις τιμές των κατηγοριών που προβλέπουν τα μοντέλα για τα συγκεκριμένα κείμενα.

3.2.6.1. Confusion Matrix

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκε confusion matrix. Είναι ένας συνοπτικός πίνακας που χρησιμοποιείται για να εκτιμά την απόδοση ενός μοντέλου κατηγοριοποίησης.



Σχήμα 9 Confusion Matrix[10]

Positive: Η τιμή είναι θετική.

Negative: Η τιμή είναι αρνητική.

True Positive: Οι τιμές που προβλέφθηκαν σωστά για στην κατηγορία “Positive”

True Negative: Οι τιμές που προβλέφθηκαν σωστά για στην κατηγορία “Negative”

False Positive: Οι τιμές που προβλέφθηκαν λάθος για στην κατηγορία “Positive”

False Negative: Οι τιμές που προβλέφθηκαν λάθος για στην κατηγορία “Negative”

Δεδομένων των αποτελεσμάτων που εμφανίζονται στον πίνακα, μπορεί να υπολογιστεί αντίστοιχα και η μετρική accuracy, με την οποία προσδιορίζεται η απόδοση των μοντέλων

4. Πειραματικά Αποτελέσματα

4.1 Σενάριο 1 - Αλλαγή Χαρακτήρων

IMDB						
Model	**BERT**			**DistilBERT**		
Training Samples	800	2400	4500	800	2400	4500
Test Samples	200	600	400	200	600	500
	Stratified K-Fold k =10 Cross Validation : Accuracy					
Original Data	0.77 (+/- 0.08)	0.74 (+/- 0.05)	0.77 (+/- 0.04)	0.77 (+/- 0.08)	0.76 (+/- 0.06)	0.78 (+/- 0.05)
Charswap Attack	0.57 (+/- 0.15)	0.61 (+/- 0.05)	0.64 (+/- 0.05)	0.57 (+/- 0.15)	0.63 (+/- 0.06)	0.63 (+/- 0.06)
ScRNN Defense	0.62 (+/- 0.08)	0.63 (+/- 0.03)	0.65 (+/- 0.04)	0.62 (+/- 0.08)	0.67 (+/- 0.04)	0.68 (+/- 0.02)

Πίνακας 5 Στατιστικά Αποτελέσματα Σεναρίου 1 στα δεδομένα IMDB

AGNews						
Model	**BERT**			**DistilBERT**		
Training Samples	800	2400	4500	800	2400	4500
Test Samples	200	600	500	200	600	500
	Stratified K-Fold k =10 Cross Validation : Accuracy					
Original Data	0.82 (+/- 0.08)	0.84 (+/- 0.04)	0.86 (+/- 0.03)	0.82 (+/- 0.08)	0.84 (+/- 0.04)	0.87 (+/- 0.02)
Charswap Attack	0.68 (+/- 0.08)	0.70 (+/- 0.06)	0.72 (+/- 0.03)	0.72 (+/- 0.06)	0.70 (+/- 0.06)	0.77 (+/- 0.02)
ScRNN Defense	0.72 (+/- 0.11)	0.73 (+/- 0.05)	0.75 (+/- 0.03)	0.75 (+/- 0.08)	0.73 (+/- 0.05)	0.81 (+/- 0.03)

Πίνακας 6 Στατιστικά Αποτελέσματα Σεναρίου 1 στα δεδομένα AG_News

Στους πίνακες 5 και 6 , αναγράφονται τα πειραματικά αποτελέσματα του πρώτου σεναρίου για δύο σύνολα δεδομένων IMDB, AG News. Διαπιστώνεται ότι η απόδοση των μοντέλων μειώνεται όταν τα δεδομένα εκπαίδευσης που δίνονται στο μοντέλο δέχονται επίθεση σε επίπεδο χαρακτήρα, ενώ αντίστοιχα η απόδοσή τους βελτιώνεται όταν τα δεδομένα εκπαίδευσης του μοντέλου δεχτούν τροποποιήσεις από τον μηχανισμό άμυνας scRNN.

Για την αξιολόγηση των αποδόσεων χρησιμοποιήθηκε 10-fold cross validation , με την μετρική accuracy, στα δεδομένα εκπαίδευσης. Στη συνέχεια ένα μικρότερο σύνολο δεδομένων το οποίο δεν χρησιμοποιήθηκε κατά την εκπαίδευση, δόθηκε στο μοντέλο ώστε να γίνουν προβλέψεις.

Το μοντέλο του μηχανισμού άμυνας ScRNN όπως προαναφέρθηκε, χρειάστηκε να εκπαιδευτεί σε δεδομένα από τα σύνολα που χρησιμοποιήθηκαν ώστε να αποδώσει καλύτερα. Παρόλα αυτά στα πλαίσια του συγκεκριμένου παραδείγματος εκπαιδεύτηκε με συνολικά 20.000 γραμμές δεδομένων και από τα δύο υποσύνολα και η διαδικασία αυτή κράτησε περίπου 6 ώρες σε φορητό υπολογιστή με χρήση CPU. Είναι φανερό ότι εάν υπήρχαν κατάλληλες εγκαταστάσεις, ώστε το μοντέλο να εκπαιδευτεί σε περισσότερα δεδομένα τα αποτελέσματα του μοντέλου της άμυνας ενδέχεται να έχουν καλύτερες αποδόσεις.

4.1.1. Σκορ-προβλεψεων

Από ένα σύνολο δεδομένων ελέγχου 200 δειγμάτων, τα οποία ήταν στην κανονική τους μορφή, δηλαδή χωρίς να τροποποιηθούν από κάποια επίθεση, έγιναν προβλέψεις χρησιμοποιώντας το μοντέλο DistilBERT. Αφού έγιναν οι προβλέψεις με SGDRegressor, στο σύνολο δεδομένων IMDB, χρησιμοποιήθηκε η συνάρτηση round() ώστε να γίνουν οι απαραίτητες στρογγυλοποιήσεις και έτσι να υπάρχει ξεκάθαρη εικόνα της κατηγορίας η οποία προβλέφθηκε. Όπως φαίνεται στον πίνακα 7 σε μία ενδεικτική εκτέλεση, μετά την συνάρτηση round(), έγινε σύγκριση των πραγματικών τιμών των κατηγοριών που ανήκουν τα κείμενα, με τις τιμές τις οποίες προέβλεψε το μοντέλο. Έτσι φαίνεται ότι το μοντέλο που εκπαιδεύτηκε στα δεδομένα που έχουν δεχθεί επίθεση, δεν μπορεί να αποδώσει σωστά τις κατηγορίες στα δεδομένα που δεν έχουν δεχθεί επίθεση. Ενώ μετά φαίνεται ότι όταν το μοντέλο εκπαιδευτεί στα ίδια δεδομένα, αλλά αφού έχουν δεχτεί ένα μηχανισμό άμυνας, έχει καλύτερη απόδοση.

	Σωστές Προβλέψεις	**Λάθος Προβλέψεις**
Μοντέλο που εκπαιδεύτηκε σε Κανονικά Δεδομένα	144	56
Μοντέλο που εκπαιδεύτηκε σε Δεδομένα με επίθεση charswap	129	71
Μοντέλο που εκπαιδεύτηκε σε Δεδομένα με άμυνα ScRNN	146	54

Πίνακας 7 Ποσοστά Προβλέψεων κατα τις τρεις φάσεις

Παρόλα αυτά μπορεί να επιβεβαιωθεί το ότι το DistilBERT συνεχίζει να έχει το 97% της συνολικής απόδοσης του δασκάλου του BERT, καθώς έχουν σχεδόν τις ίδιες αποδόσεις στα ίδια ακριβώς δεδομένα.

4.1.2. Μερικές αποκλίσεις που παρατηρήθηκαν και αξίζει να σχολιαστούν

True Class	**Sentence**	**Prediction**	**Round (1)**	**Round (2)**	**Round (3)**	**0 = Σωστή προβλεψη 1= Λάθος προβλεψη**	**Τύπος**
1	“Whatever rating I give BOOM is only because of the superb location photography of Sardinia and Rome. Otherwise, this is only for ...." “”	0.824841313262397	1	0.8	0.82	0	Επίθεση
		0.402049496867088	0	0.4	0.4	1	Άμυνα

Πίνακας 8 Παράδειγμα μεγάλης απόκλισης πρόβλεψης μοντέλου Επίθεσης-Άμυνάς

Χαρακτηριστικό παράδειγμα είναι η πρόταση του παραδείγματος που φαίνεται στον πίνακα 8, καθώς υπάρχει μεγάλη απόκλιση πρόβλεψης από το μοντέλο που έχει εκπαιδευτεί σε δεδομένα

που έχουν δεχθεί επίθεση και στο μοντέλο που έχει εκπαιδευτεί με δεδομένα που έχουν δεχθεί άμυνα. Φαίνεται ότι εν τέλη το μοντέλο που έχει εκπαιδευτεί με δεδομένα μετά από άμυνα αποκλίνει κατά 0.4 σχεδόν, προβλέπωντας λάθος την αληθινή κατηγορία της πρότασης.

4. Σενάριο 2 – Αλλαγή σημασιολογίας λέξεων

Test Samples	80			
Maximum length of sequence	150			
	IGA		**WordNet : 0.2**	
	CNN	**RNN**	**CNN**	**RNN**
No Attack	0.75	0.86	0.75	0.86
Attack	0.3572	0.204	0.625	0.68
SEM Defense	0.75	0.75	0.75	0.76

Πίνακας 9 Στατιστικά αποτελέσματα επιθέσεων 1

Test Samples	25			
Maximum length of sequence	100			
	IGA		**WordNet : 0.2**	
	CNN	**RNN**	**CNN**	**RNN**
No Attack	0.75	0.88	0.75	0.88
Attack	0.36	0.25	0.66	0.63
SEM Defense	0.75	0.76	0.74	0.78

Πίνακας 10 Στατιστικά αποτελέσματα επίθεσεων 2

Στους πίνακες φαίνονται τα πειραματικά αποτελέσματα που έγιναν σε μικρό σύνολο δεδομένων, με μεγαλύτερο μήκος πρότασης 100 λέξεις και 150 λέξεις. Διαπιστώνεται ότι και σε επίπεδο σημασιολογίας η απόδοση των μοντέλων κατά την κατηγοριοποίηση των δεδομένων χαμηλώνει όταν τα δεδομένα δέχονται επίθεση. Ενώ αντίστοιχα παρουσιάζει μια άνοδο όταν τα μοντέλα

τροποποιούνται με επιπλέον κωδικοποίηση ώστε να αναγνωρίζουν πότε μια πρόταση έχει αλλάξει σημασιολογικά ή απλά έχουν αντικατασταθεί κάποιες λέξεις της πρότασης με άλλες συνώνυμες, ώστε η απόσταση των δύο προτάσεων πριν και μετά την επίθεση να μην είναι μεγάλη.

Παρόλα αυτά παρατηρείται ότι η επίθεση με την χρήση WordNet δεν είναι τόσο αποδοτική, καθώς αλλάζει τις λέξεις με συνώνυμες χωρίς να αλλάζει το νόημα της πρότασης. Από την άλλη η επίθεση IGA φαίνεται να έχει καταστροφικές συνέπειες για το μοντέλο, ενώ επίσης είναι φανερό ότι η συγκεκριμένη άμυνα έχει φτιαχτεί ώστε να αντιμετωπίζει στον πιο μεγάλο βαθμό την συγκεκριμένη επίθεση. Έτσι επιβεβαιώνεται η θεωρία ότι η κάθε επίθεση χρειάζεται να αντικρούεται από μία αντίστοιχη άμυνα και ότι δεν μπορεί μία άμυνα να προσφέρει τα επιθυμητά αποτελέσματα για κάθε επίθεση.

Παράλληλα βλέπουμε ότι για τα μοντέλα RNN μοντέλα, ενώ φαίνεται να έχουν καλύτερη απόδοση αρχικά, μετά την επίθεση η απόδοσή τους καταστρέφεται. Ενώ επίσης στην συνέχεια φαίνεται ότι ο μηχανισμός άμυνας δεν μπορεί να επαναφέρει την απόδοση που είχαν αρχικά. Σε αντίθεση με τα CNN μοντέλα που ενώ αρχικά φαίνεται να έχουν πιο μικρή απόδοση μετά τον μηχανισμό άμυνας αποκτούν σχεδόν το 100% της αρχικής τους απόδοσης. Αυτό ίσως οφείλεται στην ίδια την φύση των μοντέλων αυτών αφού τα CNN μοντέλα θεωρούνται ισχυρότερα.

4.2.1. Μερικές αποκλίσεις που παρατηρήθηκαν

	True value	Predicted Value	Array of predict SGDRergressor	Authentic Review	Attacked by IGA Review
1	0	1	[array([0.19038194, 0.80961806], dtype=float32)]	a real hoot , unintentionally sidney portier 's character is so sweet and lovable you want to smack him nothing about this movie rings true and it 's boring to boot	a real hoot , unintentionally sidney portier 's character is so sweet and lovable you want to dope him nothing about this movie rings true and it 's boring to boot
2	1	0	[array([0.84494644, 0.1550536], dtype=float32)]	as i watched wirey spindell i couldnt but laugh at what was taking place on screen wirey sure got a lot of play from both boys and women but i was confused as to why the actor that played wirey in h s was 10 years old then the actor changed age to like 20 to play wirey when he was a senior in hs but whatever , i thought it was funny	as i watched wirey spindell i couldnt but laugh at what was taking place on screen wirey sure got a lot of play from both boys and women but i was confused as to why the actor that played wirey in h s was 10 years old thereafter the actor changed age to like 20 to play wirey when he was a senior in hs but whatever , i thought it was funny

Πίνακας 11 Κατηγοριοποίηση Δεδομένων με επιθέσεις IGA

Όπως φαίνεται στον πίνακα 11, τα μοντέλα που διαχειρίζονται κείμενα, μπορεί να αποδειχθούν πολύ ευάλωτα, καθώς η αλλαγή μόνο μιας λέξης αρκεί για να τα κάνει να πάρουν λάθος απόφαση με αρκετά μεγάλη απόκλιση. Όπως στο 1ο και το 2ο παράδειγμα του πίνακα, ώπου η απόκλιση φαίνεται να είναι σε ποσοστό 0.7%. Ποσοστό το οποίο δεν μπορεί να περάσει απαρατήρητο σε τέτοιου είδους εφαρμογές.

		True value	Predicted Value	Array of predict SGDRergressor	Authentic Review	Review Attacked by IGA
1	Απλό Μοντέλο	1	1	[array([0.49221247, 0.5077875], dtype=float32)]	as i watched wirey spindell i couldnt but laugh at what was taking place on screen wirey sure got a lot of play from both boys and women but i was confused as to why the actor that played wirey in h s was 10 years old then the actor changed age to like 20 to play wirey when he was a senior in hs but whatever , i thought it was funny	as i noticed wirey spindell i couldnt but laugh at what was taking place on screen wirey sure got a lot of play from both boys and women but i was confused as to why the actor that played wirey in h s was 10 years old then the actor changed age to like 20 to play wirey when he was a senior in hs but whatever , i thought it was funny
2	Μοντέλο με Άμυνα	1	0	[array([0.84494644, 0.1550536], dtype=float32)]		
3	Απλό Μοντέλο	0	0	[array([0.5887899 , 0.41121012], dtype=float32)]	a real hoot , unintentionally sidney portier 's character is so sweet and lovable you want to smack him nothing about this movie rings true and it 's boring to boot	a real hoot , unintentionally sydney portier 's character is so sweet and lovable you want to smack him nothing about this movie rings true and it 's boring to boot
4	Μοντέλο με Άμυνα	0	1	[array([0.19038194, 0.80961806], dtype=float32)]		

Πίνακας 12 Επιθέσεις IGA σε απλό μοντέλο και σε μοντέλο άμυνας

		True value	Predicted Value	Array of predict SGDRergressor	Authentic Review	Review Attacked by WordNet
1	Απλό Μοντέλο	0	0	[array([0.6410923 , 0.35890767], dtype=float32)]	i found this film to be a bit too depressing i do n't mind dramas , but this was a bit too much for me luckily , there does seem to be somewhat a decent outcome i suppose it was well done i 'd watch it again , but it 's nothing to rave about	i encounter this cinema to be a chip too cheerless i do n't thinker drama , but this was a turn too lots for me fortuitously , there does appear to be jolly a decently effect i reckon it was good set i 'd vigil it again , but it 's cipher to rant about
2	Μοντέλο με Άμυνα	0	1	[array([0.0211956 5, 0.97880435], dtype=float32)]		
3	Απλό Μοντέλο	0	0	[array([0.6238249 , 0.37617514], dtype=float32)]	this film features two of my favorite guilty pleasures sure , the effects are laughable , the story confused , but just watching hasselhoff in his knight rider days is always fun i especially like the old hotel they used to shoot this in , it added to what little suspense was mustered give it a 3	this take sport 2 of my front-runner hangdog pleasure sure , the force are amusing , the tarradiddle baffled , but just view hasselhoff in his dub passenger daytime is constantly sport i peculiarly care the sometime hotel they habituate to take this in , it tally to what piddling suspense was rally commit it a deuce-ace
4	Μοντέλο με Άμυνα	0	1	[array([0.2756355 4, 0.7243644], dtype=float32)]		

Πίνακας 13 Επιθέσεις WordNet σε απλό μοντέλο και σε μοντέλο άμυνας

Οι παραπάνω πίνακες 12 και 13, παρουσιάζουν χαρακτηριστικά παραδείγματα επιθέσεων με τις άμυνές τους. Στα συγκεκριμένα παραδείγματα φαίνεται η αστοχία του μοντέλου που έχει εκπαιδευτεί σαν μοντέλο άμυνας. Όπως φαίνεται στο παράδειγμα 1,2 του πίνακα με τις επιθέσεις IGA, η σημασία της κάθε λέξης χωριστά μπορεί να παίξει σημαντικό ρόλο στην πρόβλεψη του μοντέλου. Ένα άλλο χαρακτηριστικό παράδειγμα το οποίο είναι ενδιαφέρον, είναι ότι ενώ η ίδια πρόταση (παράδειγμα 3,4 του πίνακα 11) χρησιμοποιείται στο απλό μοντέλο και μετά στο μοντέλο άμυνας, φαίνεται να το μπερδεύει και έτσι να την κατατάσσει σε λάθος κατηγορία. Αυτό επίσης αποδεικνύει την ευαισθησία που έχουν τα μοντέλα μηχανικής μάθησης που διαχειρίζονται κείμενα. Αλλά και το πόσο σημαντικό είναι για τα μοντέλα να εκπαιδευτούν σε επαρκή σύνολα δεδομένων, ώστε να είναι ικανά να κάνουν σωστές προβλέψεις και να έχουν ικανοποιητική απόδοση. Παρόλα αυτά πρόκειται για μεμονωμένες περιπτώσεις και δείχνουν απλά το πόσο ευαίσθητα είναι τα μοντέλα διαχείρισης κειμένου.

Παράλληλα όσο αφορά τον πίνακα με τις επιθέσεις WordNet 13, παρατηρείται ότι η προτάσεις αλλάζουν ριζικά νόημα μετά την επίθεση, αφού το WordNet αντικαθιστά τις λέξεις σαν λέξεις και δεν λαμβάνει καθόλου το νόημα της όλης πρότασης. Εξάλλου αυτό φαίνεται και στα στατιστικά αποτελέσματα από τους πίνακες στατιστικών αποτελεσμάτων των επιθέσεων 9 και 10.

5. Επίλογος και Μελλοντικές Κατευθύνσεις

Τα μοντέλα μηχανικής μάθησης που διαχειρίζονται κείμενα, έχουν ευρεία χρήση και εφαρμογή στις μέρες μας. Για αυτό και επιβάλλεται να διασφαλίζεται η προστασία των δεδομένων που εισάγονται σε αυτά, αλλά και η προστασία των δεδομένων τα οποία εξάγονται. Μια τέτοια εφαρμογή την οποία οι περισσότεροι συναντούν καθημερινά, είναι σε συστήματα Question Answering ή σε επεκτάσεις τους όπως είναι τα chatbots. Τέτοια συστήματα έχουν άμεση επικοινωνία με τον χρήστη, αφού συνήθως τους ζητείται να απαντήσουν συγκεκριμένες ερωτήσεις ή χρειάζεται να δώσουν στον χρήστη συγκεκριμένες πληροφορίες. Έτσι έγινε μια μικρή έρευνα πάνω στα συστήματα Question Answering.

5.1 Επιθέσεις και άμυνες σε Question Answering συστήματα

Σήμερα η τεχνητή Νοημοσύνη έχει κάνει τεράστια βήματα εξέλιξης. Τα συστήματα τεχνητής νοημοσύνης είναι πλέον ικανά να αναγνωρίσουν ήχους, να λάβουν αποφάσεις ενώ παράλληλα μπορούν να χρησιμοποιηθούν από τους ανθρώπους για πιο γρήγορη, εύκολη και άμεση επίλυση ερωτημάτων σε συγκεκριμένα θέματα. Για αυτό και σήμερα όταν μιλάμε για την τεχνητή νοημοσύνη, σκεφτόμαστε εφαρμογές σαν την Alexa, Siri ή άλλους βοηθούς για εξυπηρέτηση πελατών όπως είναι τα chatbots. Αυτό που κρύβεται πίσω από την επικοινωνία τέτοιων έξυπνων εφαρμογών με τους ίδιους τους ανθρώπους, δεν είναι τίποτα παραπάνω από συστήματα Question Answering. Παρόλα αυτά τα αρχικά βήματα στον τομέα αυτών των συστημάτων ξεκίνησαν από τα τέλη της δεκαετίας του 60, με τα LUNAR και BASEBALL (Clementeena et al.,2018).

Τέτοιες τεχνολογίες μπορούν να εφαρμοστούν σε chatbots. Τα chatbots βοηθούν στην αναζήτηση σε συγκεκριμένο αντικείμενο. Έτσι η αναζήτηση πληροφορίας, η τροποποίηση κειμένων ή ακόμα και η διαχείριση κειμένων, αυτοματοποιουνται. Παράλληλα σήμερα χρησιμοποιούνται πολύ συχνά για την εξυπηρέτηση πελατών, για σκιαγράφηση του προφίλ του χρήστη, που μετέπειτα χρησιμοποιείται για προτάσεις σε προϊόντα ή πληροφορίες που θεωρεί το μοντέλο ότι θα φανούν ενδιαφέρον στον συγκεκριμένο χρήστη(Kodra, 2017). Επιπλέον

χρησιμοποιούνται για σκοπούς μάρκετινγκ, αφού προωθούν συνήθως την επιχείρηση για την οποία “εργάζονται”, διαδικτυακά.

Υπάρχουν δύο είδη από QA συστήματα, τα Close Domain και τα Open Domain . Τα πρώτα χειρίζονται συγκεκριμένο θέμα, έτσι μπορεί να είναι πιο εύκολο να διαχειρίζονται τις απαιτήσεις του χρήστη όσο αφορά την απάντηση που θα πρέπει να δοθεί. Παράλληλα μπορούν να δουλεύουν πολύ γρήγορα αφού οι ερωτήσεις βρίσκονται σε συγκεκριμένο κόμβο ώστε να μπορούν να πάρουν τις αντίστοιχες απαντήσεις πολύ εύκολα. Παρόλα αυτά η περισσότερη έρευνα γίνεται στα Open Domain QA συστήματα. Στόχος τους είναι να δίνουν την ακριβή και πιο σωστή απάντηση για την ερώτηση που λαμβάνουν από τον χρήστη. Το σύστημα βρίσκει τις απαντήσεις με τη βοήθεια της ανάκτησης πληροφοριών, την υπολογιστική γλωσσολογία, αλλά και με την αναπαράσταση γνώσης.

Έτσι γενικότερα τα QA μπορούν να περιγραφούν σαν μια τεχνολογία που δίνει την σωστή απάντηση σε μια ερώτηση αντί να για μια λίστα από πιθανές απαντήσεις όπως τις μηχανές αναζήτησης. Παρόλα αυτά μπορεί κανείς να πει ότι λειτουργούν σαν μηχανές αναζήτησης, αλλά με διαφορετική αναπαράσταση των αποτελεσμάτων τους. Οι μηχανές αναζήτησης επιστρέφουν μια λίστα με συνδέσμους από απαντήσεις ενώ τα QA δίνουν κατευθείαν απάντηση στην ερώτηση (Andrew Zola, 2020). Η διαδικασία ανάκτησης της πληροφορίας σε ένα QA σύστημα αποτελείται από τρία κύρια βήματα α) την προ επεξεργασία της ίδιας της ερώτησης, β) την ταξινόμηση των απαντήσεων ανάλογα με τον βαθμό σχετικότητας τους και γ) την εξαγωγή της ίδιας της απάντησης.

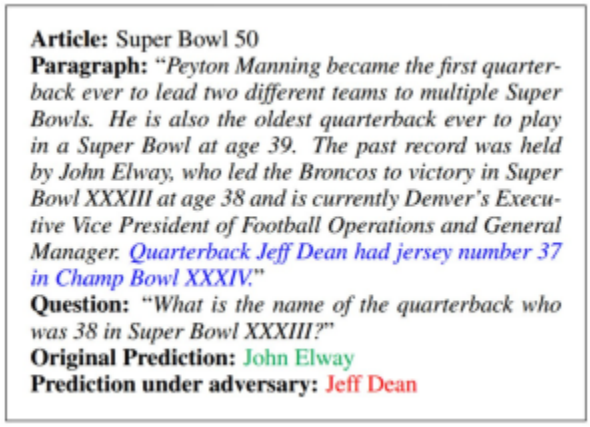
5.1.1 Επιθέσεις σε συστήματα QA

Τα QA συστήματα, χρειάζεται να είναι σχεδιασμένα με τέτοιο τρόπο ώστε να είναι προετοιμασμένα να απαντήσουν σε συγκεκριμένες ερωτήσεις , οι οποίες το πιό πιθανό θα είναι διατυπωμένες σε πραγματική ανθρώπινη γλώσσα. Έτσι είναι προφανές ότι τα πιο πολύπλοκα συστήματα χρησιμοποιούν πλέον NLP για να κατανοήσουν την ανθρώπινη γλώσσα όπως αυτή χρησιμοποιείται από τους ανθρώπους και όχι από τις μηχανές. Για ακόμα μια φορά εισάγεται η

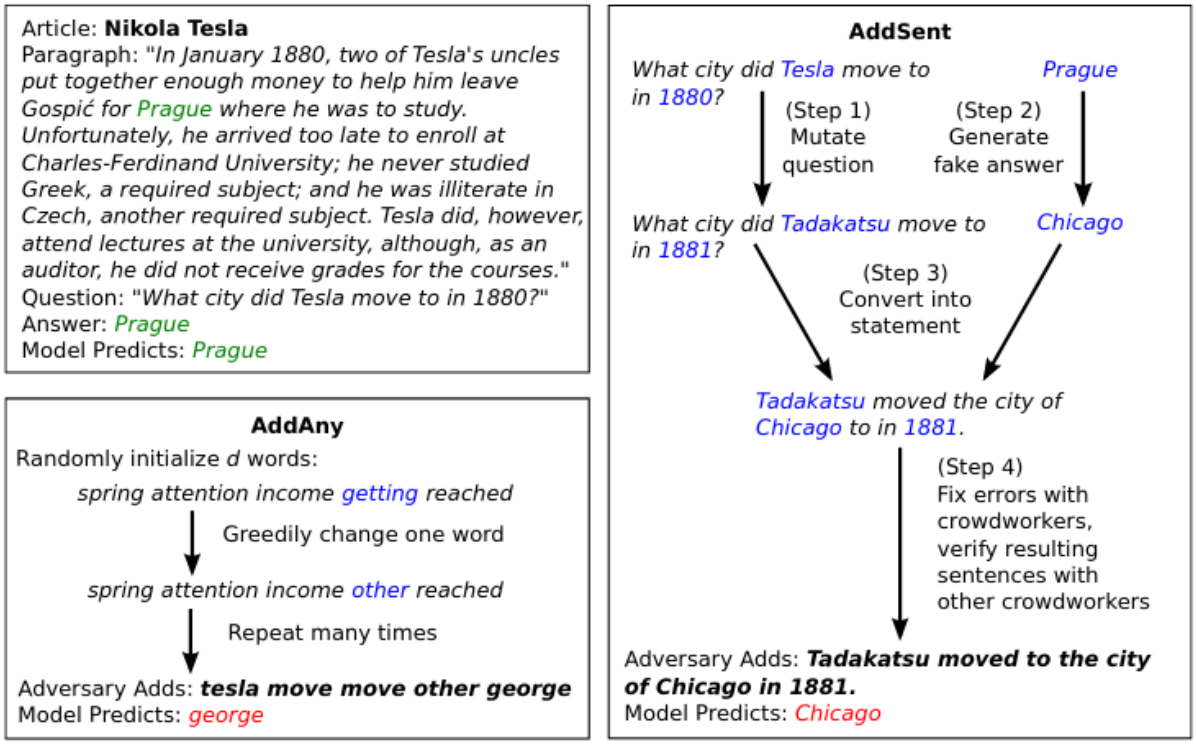
χρήση των προ εκπαιδευμένων μοντέλων που αποτελείται από δύο διαδικασίες, την προετοιμασία και την επεξεργασία των ίδιων των δεδομένων.

Το μοντέλο BERT, έχει πλέον γίνει πολύ δημοφιλής για τις εφαρμογές QA. Παράλληλα η πρωταρχική μελέτη του Devlin (Delvin et al., 2018), του δημιουργού του BERT, περιλαμβάνει μέσα αποτελέσματα της απόδοσης του ίδιου του BERT πάνω στο σύνολο δεδομένων SQuAD, όπου το σύστημα κλήθηκε να προβλέψει απαντήσεις για συγκεκριμένες ερωτήσεις, βασισμένο σε κείμενα από το Wikipedia.

Ένα παράδειγμα επίθεσης σε μοντέλα QA είναι αυτό της εικόνας 4 (Jia and Liang, 2017). Συγκεκριμένα προσπάθησαν να μπερδέψουν και να υποβιβάσουν την απόδοση του μοντέλου, προσθέτοντας στο κείμενο, το οποίο θα περνούσε σαν είσοδος στο μοντέλο, επιπλέον τροποποιημένες προτάσεις. Στην εικόνα η μπλε πρόταση μοιάζει πολύ με την ίδια την ερώτηση που δόθηκε στο μοντέλο, αλλά δεν περιλαμβάνει την σωστή απάντηση. Αυτή η επιπλέον πρόταση μπορεί να είναι κατανοητή για τον άνθρωπο, αλλά παρόλα αυτά μπορεί να αποπλανήσει το μοντέλο. Επομένως όπως φαίνεται, η απόδοση του μοντέλου μειώνεται από την αρχική 75% στο 36%, προφανώς αυτή η απλή αλλαγή και ασήμαντη αλλαγή για τον άνθρωπο, μπορεί να είναι καταστροφική για τα αποτελέσματα του μοντέλου.



Εικόνα 3 Επίθεση σε μοντέλο Κατανόησης κειμένου(Jia and Liang, 2017)



Εικόνα 4 Προσθήκη κειμένου σε μοντέλο Κατανόησης κειμένου(Jia and Liang, 2017)

Στην εικόνα 4 παρουσιάζονται 2 διαφορετικές τεχνικές ώστε να προστεθεί επιπλέον κείμενο σε παράγραφο που χρησιμοποιείται για την αναζήτηση απάντησης. Σκοπός τους είναι η επίθεση σε μοντέλα κατανόησης κειμένου, ώστε αυτά να εσκεμμένα να προβλέψουν κάτι άλλο.

5.1.2 Bert-as-a-service

Το Bert-as-a-Service (Xiao, 2018) χρησιμοποιεί το ίδιο το μοντέλο του BERT, σαν κωδικοποιητή προτάσεων (Sentence Encoding) και το παρέχει σαν υπηρεσία μέσω του ZeroMQ. Έτσι δίνει την δυνατότητα της χαρτογράφησης των προτάσεων σε αναπαραστάσεις σταθερού μήκους με την χρήση λίγου κώδικα. Η διαδικασία της κωδικοποίησης των προτάσεων, απαιτείται πολύ συχνά από NLP εφαρμογές όπως η ανάλυση συναισθημάτων ή η κατηγοριοποίηση κειμένων. Ο στόχος της είναι να μπορεί να αναπαρίσταται το μέγεθος μιας μεταβλητής συνάρτησης, ως ένα διάνυσμα σταθερού μεγέθους.

Για την υλοποίηση του Bert-as-a-Service χρησιμοποιήθηκαν τα 12/24 επίπεδα από το προ εκπαιδευμένο μοντέλο του BERT. Για να χρησιμοποιηθεί χρειάζεται να εγκατασταθεί ο Server και ο Client, ενώ επίσης να εγκατασταθεί και ένα προ εκπαιδευμένο μοντέλο του BERT. Στα πλαίσια της εργασίας χρησιμοποιήθηκε το ‘‘cased_L-12_H-768_A-12‘’.

5.1.3 Προσπάθεια επίθεσης σε QA με τη χρήση του Bert-as-a-Service

Το Σύνολο δεδομένων που χρησιμοποιήθηκε, υλοποιήθηκε την άνοιξη του 2008 από φοιτητές του πανεπιστημίου της Pittsburgh, στην Τρανσυλβανία οι οποίοι έκαναν μαθήματα σχετικά με NLP διαδικασίες(Smith, Heilman, and Hwa, 2008).

Αυτό το σύνολο δεδομένων αποτελείται από ζευγάρια ερωτήσεων και απαντήσεων με κάποια επιπλέον πεδία τα οποία δεν χρησιμοποιήθηκαν. Τα πεδία που υπήρχαν αρχικά στο σύνολο δεδομένων είναι τα: [ArticleTitle , Question, Answer, DifficultyFromQuestioner, DifficultyFromAnswerer, ArticleFile]. Τελικά χρησιμοποιήθηκαν μόνο το πεδίο Question για την ερώτηση και το πεδίο Answer για την απάντηση. Χρειάστηκε να γίνει μια μικρή προ επεξεργασία του συνόλου καθώς υπήρχαν κενά κελία έτσι διαγράφηκαν οι γραμμές με κενά κελιά. Η διαδικασία που ακολουθήθηκε είναι παρόμοια με τα προηγούμενα σενάρια. Τα δεδομένα χρησιμοποιήθηκαν για εκπαίδευση ενώ παράλληλα κάποιες επιπλέον ξεχωριστές ερωτήσεις χρησιμοποιήθηκαν για έλεγχο της απόδοσης.

Το συγκεκριμένο σενάριο αφού λαμβάνει μια ερώτηση, προσπαθεί να προβλέψει σε ποια από τις ερωτήσεις που του είχαν δοθεί για εκπαίδευση ανήκει. Ενώ στην συνέχεια αφού προβλέψει σε ποια ερώτηση αναφέρεται η ερώτηση που του δίνεται, επιστρέφει πίσω την απάντηση που της αντιστοιχεί (Zola ,2020).

Όπως διαπιστώθηκε όταν εφαρμόστηκαν οι επιθέσεις charswap/wordnet και από την μεριά της ερώτησης αλλά και στην μεριά της απάντησης, το σύστημα παρουσιάζει κάποια λάθη τα οποία οφείλονται κυρίως στην ευαισθησία που προκύπτει από τη φύση των συγκεκριμένων συστημάτων.

	Q	Prediction	A	Score
0	Why Lincoln grow a beard?	Who suggested Lincoln grow a beard?	11-year-old Grace Bedell	0.989488
1	Lincoln old in 1816?	How old was Lincoln in 1816?	seven	0.967059
2	how old was Lincoln in 1816?	How old was Lincoln in 1816?	seven	1.000000
3	Was King Victor Emmanuel III paying homage to ...	Was King Victor Emmanuel III there to pay homa...	yes	0.997284
4	Year of publishing he's collection?	When did he publish a collection?	1733	0.955545
5	Food of beetles?	What happened on plants and fungi?	They are food to beetles.	0.955738
6	Favourite food of beetles?	Are turtles pets?	yes	0.954250
7	Is it possible there more than 350	Is it possible that there are more than 350	000 species of beetles?	0.988989
8	Which year did Grover Cleveland win?	Which election did Grover Cleveland win?	1884 and 1892 presidential elections	0.988161
9	Who were Grover Cleveland's parents?	Who were Grover Cleveland's parents?	Reverend Richard Cleveland and Anne Neal.	1.000000
10	Who was mother and father of Grover Cleveland?	Who were Grover Cleveland's parents?	Reverend Richard Cleveland and Anne Neal.	0.979688
11	How many children did Grover Cleveland have?	How many children did Grover Cleveland have?	Six.	1.000000
12	did Grover Cleveland have children?	How many children did Grover Cleveland have?	Six.	0.986567
13	Was Cleveland born in Caldwell	Was Cleveland born in Caldwell	New Jersey to the Reverend Richard Cleveland ...	1.000000
14	Where did Cleveland born?He became chief engin...	He became chief engineer in the Department of ...	in 1894	0.984664
15	When did He became chief engineer in the Depar...	He became chief engineer in the Department of ...	in 1894	0.991645
16	What led Becquerel to investigate the spontane...	What led Becquerel to investigate the spontane...	phographic plates being fully exposed	1.000000
17	What made Becquerel to investigate?	What led Becquerel to investigate the spontane...	phographic plates being fully exposed	0.945026
18	Is it true that he married louise désirée lori...	Is it true that he married louise désirée lori...	yes	1.000000
19	when did he married louise désirée?	When did he marry louise désirée lorieux?	1890	0.979107

Εικόνα 5 Αποτελέσματα QA χωρίς κάποια επίθεση.

	Q	Prediction	A	Score
2	Why Lincoln grow a beard?	Who suggested Lincoln grow a beard?	11-year-old Grace Bdell	0.989487588405609
3	Lincoln old in 1816?	How old was Lincoln in 1816?	seve	0.967059314250946
4	how old was Lincoln in 1816?	How old was Lincoln in 1816?	seve	0.9999988079071
5	Was King Victor Emmanuel III paying homage to Avogadro ?	Was King Victor Emmanuel III there to pay homage to Avogadro ?	yeV	0.997283637523651
6	Year of publishing he's collection?	When did he publish a collection?	17g33	0.95554530620575
7	Food of beetles?	What happened on plants and fungi?	Thy are foJod to beetles.	0.955738127231598
8	Favourite food of beetles?	Are turtles pets?	yse	0.95425021648407
9	Is it possible there more than 350	Is it possible that there are more than 350	000 specieZ of beetleYs?	0.988988518714905
10	Which year did Grover Cleveland win?	Which election did Grover Cleveland win?	1884 and 1892 prseidential electins	0.988160967826843
11	Who were Grover Cleveland's parents?	Who were Grover Cleveland's parents?	Reverend RicharB Clevelanhd and Ane Neal.	1
12	Who was mother and father of Grover Cleveland?	Who were Grover Cleveland's parents?	Reverend RicharB Clevelanhd and Ane Neal.	0.979688405990601
13	How many children did Grover Cleveland have?	How many children did Grover Cleveland have?	Sx.	1
14	did Grover Cleveland have children?	How many children did Grover Cleveland have?	Sx.	0.986567258834839
15	Was Cleveland born in Caldwell	Was Cleveland born in Caldwell	New Jersye to the RevereNnd Rihard CWleveland and Anne Neah?	1
16	Where did Cleveland born?He became chief engineer in the Department of Bridges and Highways in what year?	He became chief engineer in the Department of Bridges and Highways in what year?	in 18C4	0.984663784503937
17	When did He became chief engineer in the Department of Bridges and Highways?	He became chief engineer in the Department of Bridges and Highways in what year?	in 18C4	0.991645336151123
18	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	phographic plateos being fully exposd	1
19	What made Becquerel to investigate?	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	phographic plateos being fully exposd	0.945025861263275
20	Is it true that he married louise désirée lorieux in 1890?	Is it true that he married louise désirée lorieux in 1890?	yse	1
21	when did he married louise désirée?	When did he marry louise désirée lorieux?	s890	0.979106843471527

Εικόνα 6 QA με απαντήσεις που έχουν δεχθεί επίθεση charswap 0.5

Όπως φαίνεται στην εικόνα 6 με τις βαθμολογίες των απαντήσεων, το μοντέλο καταφέρνει να προβλέψει σχεδόν όλες τις ερωτήσεις σωστά εκτός από τις ερωτήσεις 7,8 για τις οποίες η αυθεντική ερώτηση ήταν η “What do beetles eat?”. Παρόλα αυτά φαίνεται ότι το μοντέλο στην περίπτωση που δέχεται επίθεση στις απαντήσεις του, δεν το αντιλαμβάνεται, με αποτέλεσμα να επιστρέφει πίσω την απάντηση που έχει δεχθεί επίθεση και μάλιστα χωρίς να μειώνει την βαθμολογία της. Από την άλλη μεριά αν εφαρμόσουμε επίθεση στις ερωτήσεις του συνόλου δεδομένων ,όπως φαίνεται στην αντίστοιχη εικόνα 7, το μοντέλο όντως μπερδεύετε και προβλέπει λανθασμένες ερωτήσεις και έτσι μειώνεται και η βαθμολογία της απάντησης για την κάθε ερώτηση, αφού πλέον το μοντέλο δεν είναι σίγουρο για το τί ερώτηση πρέπει να απαντήσει.

	Q	Prediction	A	Score
2	Why Lincoln grow a beard?	What is Finland's conomy lkie?	a highly industrialised	0.964727997779846
3	Lincoln old in 1816?	Ddi LinOcoln bea John C. Breckinridge in the 1680 lection?	Yes.	0.968042075634003
4	how old was Lincoln in 1816?	Did Liccoln starot his pbolitical careerr in 1832?	Yes	0.962865650653839
5	Was King Victor Emmanuel III paying homage to Avogadro ?	Wo lots control of his arty to the agrraians and silverietes in 1E896?	Grover Cleveland	0.956818342208862
6	Year of publishing he's collection?	sI a beetle's genehal anatomy yniform?	Yes	0.941555976867676
7	Food of beetles?	Ark beetles insects?	Yes	0.962132215499878
8	Favourite food of beetles?	wAre certian species of beetles considered Tpests?	Yes	0.957130014896393
9	Is it possible there more than 350	IAs it pssible that there are more than 35c	Sorry, I didn't get you.	0.938707411289215
10	Which year did Grover Cleveland win?	Did GrDver ICeveland win the 1884 ejection?	yes	0.963038682937622
11	Who were Grover Cleveland's parents?	WLho were Grover Clevеuland's parents?	Reverend Richard Cleveland and Anne Neal.	0.971622049808502
12	Who was mother and father of Grover Cleveland?	Where was Grover Cleeland marreid?	in the Blue Room in the White House	0.96395355463028
13	How many children did Grover Cleveland have?	uow many children did GrovOr Clevealnd have?	Six.	0.977722823619842
14	did Grover Cleveland have children?	Wehre was Gyrover Cleveland married?	In the White House	0.97395920753479
15	Was Cleveland born in Caldwell	Was Clevelan born in Caldwell	New Jersey to the Reverend Richard Cleveland and Anne Neal?	0.98786723613739
16	Where did Cleveland born?He became chief engineer in the Department of Bridges and Highways in what year?	Wan Wilsno president of the Amercian Political Science AssocOation in 910 ?	Yes	0.958289265632629
17	When did He became chief engineer in the Department of Bridges and Highways?	hy was the capital of Uruguya ounded?	Uruguay's capital	0.942763864994049
18	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	What is the offical laungуage of Liehhtenstein?	German.	0.946140885353088
19	What made Becquerel to investigate?	What was Adams' politcal pary?	Sorry, I didn't get you.	0.937960922718048
20	Is it true that he married louise désirée lorieux in 1890?	s it true that he vmarried ouise désLirée lorgeux in 1890?	yes	0.983566164970398
21	when did he married louise désirée?	hat did Lleveland die from?	A heart attack	0.960853755474091

Εικόνα 7 QA με ερωτήσεις που έχουν δεχθεί επίθεση charswap 0.5

	Q	Prediction	A	Score
2	Why Lincoln grow a beard?	WHO suggested Lincoln mature a whiskers?	Grace Bedell.	0.976422786712646
3	Lincoln old in 1816?	fare Lincoln advance the election of 1860?	Yes	0.971563994884491
4	how old was Lincoln in 1816?	How previous was Lincoln in 1816?	seven	0.989794909954071
5	Was King Victor Emmanuel III paying homage to Avogadro ?	constitute mogul master Emmanuel III there to yield court to Avogadro ?	yes	0.976760506629944
6	Year of publishing he's collection?	equal a beetle's ecumenical figure uniform?	yes	0.947200894355774
7	Food of beetles?	Are turtle pets?	Yes	0.962122678756714
8	Favourite food of beetles?	Are turtle pets?	Yes	0.96336156129837
9	Is it possible there more than 350	represent it potential that there are more than 350	000 species of beetles?	0.957138478755951
10	Which year did Grover Cleveland win?	Which election did Grover Cleveland succeed?	1884 and 1892 presidential elections	0.985873878002167
11	Who were Grover Cleveland's parents?	WHO were Grover Cleveland's parent?	Cleveland was born in Caldwell	0.99338436126709
12	Who was mother and father of Grover Cleveland?	WHO were Grover Cleveland's parent?	Cleveland was born in Caldwell	0.981811821460724
13	How many children did Grover Cleveland have?	How many youngster did Grover Cleveland have?	5	0.992574989795685
14	did Grover Cleveland have children?	How many youngster did Grover Cleveland have?	5	0.983268141746521
15	Was Cleveland born in Caldwell	coif Grover Cleveland gain the 1884 election?	yes	0.960813105106354
16	Where did Cleveland born?He became chief engineer in the Department of Bridges and Highways in what year?	he get main orchestrate in the section of bridgework and highway in what year?	in 1894	0.965007245540619
17	When did He became chief engineer in the Department of Bridges and Highways?	What did Uruguay succeed in 1828?	Its independence	0.948650002479553
18	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	What contribute Becquerel to inquire the unwritten expelling of atomic radiation?	phographic plates being fully exposed	0.956766843795776
19	What made Becquerel to investigate?	What did ford aver about his biologic forefather?	his biological father was abusive and had a history of hitting his mother	0.946159243583679
20	Is it true that he married louise désirée lorieux in 1890?	personify it avowedly that he tie louise désirée lorieux in 1890?	yes	0.975547671318054
21	when did he married louise désirée?	WHO did obstruct Monroe marry?	Elizabeth Kortright	0.967647790908814

Εικόνα 8 QA με ερωτήσεις που έχουν δεχθεί επίθεση wordnet 0.5

	Q	Prediction	A	Score
2	Why Lincoln grow a beard?	Who suggested Lincoln grow a beard?	11-year-old ornament Bedell	0.989487588405609
3	Lincoln old in 1816?	How old was Lincoln in 1816?	heptad	0.967059314250946
4	how old was Lincoln in 1816?	How old was Lincoln in 1816?	heptad	0.99999988079071
5	Was King Victor Emmanuel III paying homage to Avogadro ?	Was King Victor Emmanuel III there to pay homage to Avogadro ?	yes	0.997283637523651
6	Year of publishing he's collection?	When did he publish a collection?	1733	0.95554530620575
7	Food of beetles?	What happened on plants and fungi?	They are nutrient to overhang.	0.955738127231598
8	Favourite food of beetles?	Are turtles pets?	yes	0.95425021648407
9	Is it possible there more than 350	Is it possible that there are more than 350	000 specie of overhang?	0.988988518714905
10	Which year did Grover Cleveland win?	Which election did Grover Cleveland win?	1884 and 1892 presidential election	0.988160967826843
11	Who were Grover Cleveland's parents?	Who were Grover Cleveland's parents?	clergyman Richard Cleveland and Anne Neal.	1
12	Who was mother and father of Grover Cleveland?	Who were Grover Cleveland's parents?	clergyman Richard Cleveland and Anne Neal.	0.979688405990601
13	How many children did Grover Cleveland have?	How many children did Grover Cleveland have?	sixer.	1
14	did Grover Cleveland have children?	How many children did Grover Cleveland have?	sixer.	0.986567258834839
15	Was Cleveland born in Caldwell	Was Cleveland born in Caldwell	newly jersey to the clergyman Richard Cleveland and Anne Neal?	1
16	Where did Cleveland born?He became chief engineer in the Department of Bridges and Highways in what year?	He became chief engineer in the Department of Bridges and Highways in what year?	in 1894	0.984663784503937
17	When did He became chief engineer in the Department of Bridges and Highways?	He became chief engineer in the Department of Bridges and Highways in what year?	in 1894	0.991645336151123
18	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	phographic plates being full debunk	1
19	What made Becquerel to investigate?	What led Becquerel to investigate the spontaneous emission of nuclear radiation?	phographic plates being full debunk	0.945025861263275
20	Is it true that he married louise désirée lorieux in 1890?	Is it true that he married louise désirée lorieux in 1890?	yes	1
21	when did he married louise désirée?	When did he marry louise désirée lorieux?	1890	0.979106843471527

Εικόνα 9 QA με απαντήσεις που έχουν δεχθεί επίθεση wordnet 0.5

Όσον αφορά τις εικόνες 8 και 9, στις οποίες παρουσιάζεται η εφαρμογή της επίθεσης wordnet, στο σύστημα QA. Φαίνεται ότι παρόλο που τις περισσότερες φορές το μοντέλο προβλέπει σωστή ερώτηση στην περίπτωση που δέχεται επίθεση στις ερωτήσεις που του δίνονται για εκπαίδευση μπερδεύεται προφανώς περισσότερο από ότι με την επίθεση charswap, και αυτό έχει να κάνει με την φύση του BERT. Από την άλλη μεριά όσο αφορά τις επιθέσεις στις απαντήσεις που δίνονται στο μοντέλο φαίνεται ότι συνεχίζει να επιστρέφει την αντίστοιχη απάντηση στην ερώτηση την οποία τελικά προβλέπει έστω και αν αυτή έχει δεχθεί κάποιου είδους επίθεση.

Διαπιστώνεται για ακόμα μια φορά το πόσο ευαίσθητα και ευάλωτα είναι τα συγκεκριμένα συστήματα, έστω και αν υποστηρίζονται από πανίσχυρα εργαλεία όπως το BERT. Έτσι χρειάζεται να διαμορφωθούν ώστε να προστατεύονται και να βελτιώνονται σε επίπεδο ευαισθησίας, καθώς πολύ εύκολα μπορούν να δώσουν λάθος αποτέλεσμα το οποίο για το ίδιο το σύστημα φαίνεται σωστό, ενώ είναι ξεκάθαρα λανθασμένο για τον άνθρωπο.

5.2 Μελλοντικές Επεκτάσεις

Τα QA συστήματα είναι ένα πολύ χρήσιμο εργαλείο στις μέρες μας. Τα οποία μάλιστα όπως προαναφέρθηκε είναι η βάση για την υλοποίηση chatbots (Alloatti, 2019). Επομένως χρειάζεται να γίνει περαιτέρω έρευνα στον τομέα των επιθέσεων και των μηχανισμών άμυνας, που πλέον είναι απαραίτητο να υποστηρίζουν τέτοιου είδους συστήματα. Τα συστήματα αυτά μπορούν να χρησιμοποιηθούν ακόμα και για την διαχείριση δεδομένων υγείας ή διάφορων προσωπικών δεδομένων καθιστώντας τα πάρα πολύ ευαίσθητα. Έπειτα από την άλλη οι επιτήδειοι ψάχνουν αφορμές για να επιτίθενται σε τέτοιου είδους συστήματα ώστε να υποκλέπτουν είτε να τροποποιούν αποτελέσματα και δεδομένα. Συνεπώς πρέπει να ληφθούν στο έπακρο τα απαραίτητα μέτρα από τα ίδια τα συστήματα και τους διαχειριστές τους. Ώστε να διασφαλιστεί η ασφάλεια που απαιτείται από συστήματα που διαχειρίζονται τόσο σημαντικά και ευαίσθητα δεδομένα, αλλά ακόμα και σε συστήματα που διαχειρίζονται δεδομένα επιχειρήσεων που μπορούν επίσης να γίνουν στόχος κάποιου επιτηδείου.

5.3 Σύνοψη και Συμπεράσματα

Όσον αφορά τις επιθέσεις σε διαφορετικούς τύπους δεδομένων όπως οι εικόνες και τα δεδομένα κειμένου, είναι προφανές ότι οι εικόνες είναι συνεχή δεδομένα, ενώ τα κείμενα είναι διακριτά δεδομένα. Έτσι οι επιθέσεις και οι άμυνες που προτείνονται για δεδομένα που αφορούν εικόνες δεν μπορούν να εφαρμοστούν και σε δεδομένα κειμένου. Παράλληλα μέσω της έρευνας που προηγήθηκε διαπιστώθηκε ότι, μέχρι αυτή την στιγμή, δεν υπάρχει ένας μηχανισμός άμυνας ο οποίος να προστατεύει από επιθέσεις όλων των ειδών, ενώ αντίθετα φαίνεται ότι υπάρχει ένας μηχανισμός άμυνας που προστατεύει τα δεδομένα ή τα μοντέλα μόνο από μια επίθεση κάθε φορά. Για να μπορέσει να χτιστεί ένας μηχανισμός άμυνας ώστε να αντιμετωπίζει μία επίθεση,

χρειάζεται να γίνει γνωστή σαν επίθεση και να αναλυθεί ώστε να υπάρχει πλέον η γνώση για το τί είδους ζημία κάνει στο μοντέλο ή στα δεδομένα του.

Επιπλέον, όσον αφορά τα αποτελέσματα με συγκεκριμένες επιθέσεις και αντίστοιχες άμυνες, ο κάθε ερευνητής χρησιμοποιεί διαφορετικές τεχνικές για να εφαρμόσει κάποιο πείραμα, έτσι χρησιμοποιώντας διαφορετικούς ταξινομητές, είτε ακόμα και διαφορετικά σύνολα δεδομένων καθιστά την διαδικασίας σύγκρισης αποτελεσμάτων δύσκολη. Συνεπώς, δεδομένου του ότι η έρευνα στο συγκεκριμένο τομέα είναι στα αρχικά της βήματα τα αποτελέσματα που λαμβάνει από το πείραμα της κάθε ομάδα έρευνας χρειάζεται να αναλυθούν και να εξεταστούν, εφαρμόζοντας τους γενικούς κανόνες, ώστε να επιτευχθεί η σωστή αξιολόγηση των αποδόσεων των μοντέλων. Παράλληλα μέσα από την υλοποίηση και το στήσιμο των συγκεκριμένων πειραμάτων, διαπιστώθηκε η δυσκολία και το κόστος των πόρων που απαιτείται από ένα σύστημα για την σωστή εκπαίδευση των μοντέλων, ώστε να μπορούν να αποδίδουν όπως αναμένεται.

Επιπλέον φαίνεται ότι η ερευνητική αξία των ευαισθησιών των μοντέλων που διαχειρίζονται κείμενα, λόγω της ιδιαιτερότητας των NLP διεργασιών, όλο και αυξάνεται. Από το σενάριο 1, φαίνεται ότι οι επιθέσεις σε επίπεδο χαρακτήρα, αλλά και στο σενάριο 2 με τις τροποποιήσεις σημασιολογίας, σε μικρό αλλά και σε μεγάλο βαθμό μπορούν να μειώσουν σε μεγάλο βαθμό την απόδοση των μοντέλων. Κατά συνέπεια χρειάζεται να βελτιωθεί και να ρυθμιστεί αυτή η ευαισθησία των μοντέλων από πλευράς άμυνας, ώστε να υπάρχουν υψηλότερες αποδόσεις και να είναι πιο ασφαλής και πιό έμπιστη η εξαγωγή αποτελεσμάτων.

Βιβλιογραφία

1. Allam, A.M.N., and M.H. Haggag. 2012. “The Question Answering Systems: A Survey.” International Journal of Research and Reviews in Information Sciences (IJRRIS), http://aliallam.net/upload/598575/documents/QA%20Survey%20_2012_.pdf.
2. Alloatti Francesca, Luigi Di Caro, and Gianpiero Sportelli. 2019. “Real Life Application of a Question Answering System Using Bert Language Model.” In SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference, 250–53. Association for Computational Linguistics (ACL). doi:10.18653/v1/w19-5930.
3. Aminul Huq, and Mst Tasnim Pervin. 2020. “Adversarial Attacks and Defense on Texts: A Survey.” ArXiv. https://arxiv.org/abs/2005.14108
4. Andrew Zola ,2020, “Simple Question Answering (QA) Systems That Use Text Similarity Detection in Python” https://www.kdnuggets.com/2020/04/simple-question-answering-systems-text-similarity-python.html
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaise, Illia Polosukhin. 2017 “Attention Is All You Need” https://arxiv.org/pdf/1706.03762.pdf
6. Belinkov, Yonatan, and James Glass. 2018. “Analysis Methods in Neural Language Processing: A Survey.” ArXiv. arXiv. doi:10.1162/tacl_a_00254.
7. Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian,Xirong Li, and Wenchang Shi. 2017. “Deep text classification can be fooled” . arXiv preprint arXiv:1704.08006.
8. Christopher Olah, 2015, “Understanding LSTM Networks” https://colah.github.io/posts/2015-08-Understanding-LSTMs/
9. Clementeena, A., and P. Sripriya. 2018. “A Literature Survey on Question Answering System in Natural Language Processing.” International Journal of Engineering and

Technology(UAE) 7 (2.33 Special Issue 33). Science Publishing Corporation Inc: 452–55. doi:10.14419/ijet.v7i2.33.14209.

10. Confusion Matrix, 2019, “What is Confusion Matrix and Advanced Classification Metrics?” https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html
11. Danish Pturhi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. “Combating Adversarial Misspellings with Robust Word Recognition.” In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 5582–91. Association for Computational Linguistics (ACL). doi:10.18653/v1/p19-1561. https://github.com/danishpruthi/Adversarial-Misspellings
12. Dataset_Ag_News: AG's News Topic Classification Dataset In Textdata: Download And Load Various Text Datasets". Rdrr.Io. https://rdrr.io/cran/textdata/man/dataset_ag_news.html.
13. David S.Batista, 2018, “Convolutional Neural Networks for Text Classification”, http://www.davidsbatista.net/blog/2018/03/31/SentenceClassificationConvNets/
14. Fellbaum Christiane, 2005, Wordnet,A Lexical Database For English" Wordnet.Princeton.Edu. https://wordnet.princeton.edu/
15. Goodfellow, I. J., Shlens, J., and Szegedy, C. , 2014 “Explaining and harnessing adversarial examples” arXiv preprint arXiv:1412.6572.
16. Han Hu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2019. “Adversarial Attacks and Defenses in Images, Graphs and Text: A Review.” ArXiv. https://arxiv.org/abs/1909.08072
17. Han Xiao ,2018 ,bert-as-service Documentation, Revision a6afd773.
18. IMDB Dataset Of 50K Movie Reviews". 2021. Kaggle.Com. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.
19. Jack Morris , 2020, “What are adversarial examples in NLP? Exposing blind spots in NLP models, from RoBERTa to XLNet” https://towardsdatascience.com/what-are-adversarial-examples-in-nlp-f928c574478e
20. Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” ArXiv. arXiv. https://arxiv.org/abs/1810.04805

21. Jason Brownlee, 2021. "A Gentle Introduction To K-Fold Cross-Validation". Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/.
22. Javid Ebrahimi, Anyi Rao†, Daniel Lowd, Dejing Dou,2018, “HotFlip: White-Box Adversarial Examples for Text Classification”, arXiv:https://arxiv.org/abs/1712.06751v2
23. Jay Alammar ,2019, “A Visual Guide to Using BERT for the First Time" http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
24. Jin Di, Joey Tianyi Zhou, Zhijing Jin, and Peter Szolovits. 2019. “Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment.” ArXiv. arXiv. doi:10.1609/aaai.v34i05.6311.
25. John X. Morris , Eli Lifland , Jin Yong Yoo , Jake Grigsby , Di Jin , Yanjun Qi, 2020, “TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP” arXiv: https://arxiv.org/abs/2005.05909
26. Lorena Kodra, and Elinda Kajo. 2017. “Question Answering Systems: A Review on Present Developments, Challenges and Trends.” International Journal of Advanced Computer Science and Applications 8 (9). The Science and Information Organization. doi:10.14569/ijacsa.2017.080931. ****CHALLENGES
27. Mohammed, Mohssen, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashie. 2016. Machine Learning: Algorithms and Applications. Machine Learning: Algorithms and Applications. CRC Press. doi:10.1201/9781315371658.
28. Namatēvs, Ivars. 2018. “Deep Convolutional Neural Networks: Structure, Feature Extraction and Training.” Information Technology and Management Science 20. Walter de Gruyter GmbH. doi:10.1515/itms-2017-0007.
29. Noah A. Smith, Michael Heilman, and Rebecca Hwa, 2008, “Question Generation as a Competitive Undergraduate Course Project In Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge” , Arlington, VA, Available at: http://www.cs.cmu.edu/~nasmith/papers/smith+heilman+hwa.nsf08.pdf
30. Pandu Nayak, Google Fellow and Vice President, 2019, “Understanding searches better than ever before” https://blog.google/products/search/search-language-understanding-bert/

31. Pengfei Liu, Xipeng, Qiu Xuanjing Huang ,2016 , “Recurrent Neural Network for Text Classification with Multi-Task Learning” https://arxiv.org/abs/1605.05101
32. Philipp Schmid, 2020, “K-Fold as Cross-Validation with a BERT Text-Classification Example”https://www.philschmid.de/k-fold-as-cross-validation-with-a-bert-text-classification-example
33. Raoof Naushad , 2020 , “BERT, Distilbert, Roberta, And Xlnet Simplified Explanation And Implementation”. https://medium.com/dataseries/bert-distilbert-roberta-and-xlnet-simplified-explanation-and-implementation-5a9580242c70.
34. Robin Jia and Percy Liang, 2017 “Adversarial Examples for Evaluating Reading Comprehension Systems” arXiv preprint arXiv:1707.07328.
35. Sherstinsky, Alex. 2020. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network.” doi:10.1016/j.physd.2019.132306.
36. Siddharth Godbole, Karolina Grubinska & Olivia Kelnreiter, 2020, “Economic Uncertainty Identification Using Transformers - Improving Current Methods” https://humboldt-wi.github.io/blog/research/information_systems_1920/uncertainty_identification_transformers/
37. Textattack Documentation — Textattack 0.2.15 Documentation". 2021. Textattack.Readthedocs.Io. https://textattack.readthedocs.io/en/latest/ Github: Qdata/Textattack. https://github.com/QData/TextAttack
38. Understanding LSTM Networks -- Colah's Blog". 2021. Colah.Github.Io. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
39. Victor Sanh ,2019 “Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT” https://medium.com/huggingface/distilbert-8cf3380435b5
40. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.” ArXiv. arXiv. https://arxiv.org/abs/1910.01108
41. Wang, Xiaosen, Hao Jin, and Kun He. 2019. “Natural Language Adversarial Attacks and Defenses in Word Level.” ArXiv. https://arxiv.org/abs/1909.06723 Github: https://github.com/xiaosen-wang/SEM

42. Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-wei Chang, Xuanjing Huang, 2020, "Defense against Adversarial Attacks in NLP via Dirichlet Neighborhood Ensemble" https://arxiv.org/abs/2006.11627
43. Yoav Goldberg, 2017, "Neural Network Methods for Natural Language Processing"
44. Zhang, Wei Emma, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2019. "Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey." ArXiv. https://arxiv.org/abs/1901.06796
45. Zhaoyang Wang, Hongtao Wang, 2020, "Defense of Word-level Adversarial Attacks via Random Substitution Encoding" ArXiv. https://arxiv.org/abs/2005.00446