



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Large Language Models for Detection of Adversarial Attacks in Text Classification

DIPLOMA THESIS

by

Nikolaos Kostas

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2024



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Large Language Models for Detection of Adversarial Attacks in Text Classification

DIPLOMA THESIS

by

Nikolaos Kostas

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1^η Νοεμβρίου, 2024.

.....
Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

.....
Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

.....
Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2024

.....
ΝΙΚΟΛΑΟΣ ΚΩΣΤΑΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © – All rights reserved Nikolaos Kostas, 2024.

Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Οι ανταγωνιστικές επιθέσεις στην επεξεργασία φυσικής γλώσσας (NLP) αποτελούν κρίσιμη απειλή για την ακεραιότητα των μοντέλων ταξινόμησης κειμένου. Δημιουργώντας μικρές διαταραχές στα δεδομένα εισόδου, μπορούν να υποβαθμίσουν σημαντικά την απόδοση των μοντέλων, παραπλανώντας τα ώστε να κάνουν εσφαλμένες προβλέψεις, ενώ παράλληλα δεν επηρεάζουν την ανθρώπινη κρίση. Στην παρούσα διατριβή, διερευνούμε τη δυνατότητα εφαρμογής των μεγάλων γλωσσικών μοντέλων (LLMs) ως ανιχνευτές τέτοιων ανταγωνιστικών επιθέσεων. Για το σκοπό αυτό, αναπτύσσουμε ένα πλαίσιο μηχανικής προτροπών με στόχο τη δημιουργία προτροπών φυσικής γλώσσας που επιτρέπουν στα LLM να επιτελέσουν αποτελεσματικά αυτό το έργο. Διερευνούμε τις επιδράσεις που έχει κάθε εφαρμοζόμενη τεχνική προτροπής στην απόδοση των μοντέλων και εξάγουμε συμπεράσματα σχετικά με τη δυνητική επάρκεια των μοντέλων σε αυτή την εφαρμογή. Αφού καταλήξουμε στην προτροπή με τις καλύτερες επιδόσεις, τη χρησιμοποιούμε για να αξιολογήσουμε την ικανότητα ανίχνευσης επιθέσεων από πολλαπλά Μεγάλα Γλωσσικά Μοντέλα σε διαφορετικούς συνδυασμούς συνόλων δεδομένων ταξινόμησης κειμένου, ανταγωνιστικών επιθέσεων και μοντέλων-θυμάτων. Προκειμένου να αξιολογήσουμε περαιτέρω τις επιδόσεις αυτών των μεθόδων, διεξάγουμε μια ανθρώπινη αξιολόγηση και μια έναν έλεγχο ορθότητας για τυχόν μόλυνση των δεδομένων. Επιπλέον, προτείνουμε μια άλλη προσέγγιση για την ανίχνευση ανταγωνιστικών κειμένων, η οποία χρησιμοποιεί το ίδιο το υπό επίθεση γλωσσικό μοντέλο, επιθεωρώντας τις ταξινομήσεις που δίνονται σε κάθε μεμονωμένη πρόταση ενός κειμένου και συγκρίνοντάς τις με την ταξινόμηση που δίνεται σε ολόκληρο το κείμενο. Αφού αξιολογήσουμε επίσης αυτή την προσέγγιση υπό πολλαπλά σενάρια, συνδυάζουμε τις δύο μεθόδους μας σε μια ενοποιημένη προσέγγιση, η οποία στη συνέχεια συγκρίνεται με άλλα σύγχρονα πλαίσια ανίχνευσης. Τα πειραματικά μας αποτελέσματα δείχνουν τόσο την αναγκαιότητα για κατάλληλη μηχανική προτροπής όσο και τη δυνητική αποτελεσματικότητα της προτροπής LLM στην ανταγωνιστική ανίχνευση. Επιπλέον, ο συνδυασμός της με την επίσης αποτελεσματική, δεύτερη προτεινόμενη μέθοδο, αποδίδει ανταγωνιστικά αποτελέσματα και καθιερώνει την προσέγγισή μας ως μια λύση για την plug-and-play ανίχνευση ανταγωνιστικών κειμενικών δειγμάτων.

Λέξεις-κλειδιά — Ανταγωνιστικές Επιθέσεις, Ανίχνευση, Ταξινόμηση Κειμένου, Μεγάλα Γλωσσικά Μοντέλα, Επεξεργασία Φυσικής Γλώσσας

Abstract

Adversarial attacks in natural language processing (NLP) pose a critical threat to the integrity of text classification models. By generating subtle perturbations in input data, they can significantly impair model performance, misleading them into making incorrect predictions, all while not affecting human judgment. In this thesis, we investigate the applicability of Large Language Models (LLMs) as detectors of such adversarial attacks. To this end, we develop a prompt engineering framework with the goal of crafting natural language prompts that enable LLMs to effectively perform this task. We investigate the effects that each applied prompting technique has on model performance and draw conclusions about the models' potential competence at this task. After arriving at the best-performing prompt, we use it to evaluate the adversarial detection ability of multiple Large Language Models across different combinations of text classification datasets, adversarial attacks, and attacked models. In order to further evaluate this methods' performance, we conduct a human evaluation and a sanity test for data contamination. In addition, we propose another approach for adversarial text detection which utilizes the attacked language model itself, by inspecting the classifications given to each individual sentence of a text and comparing them with the classification given to the entire text. After also evaluating this approach under multiple scenarios, we combine our two methods into a unified approach which is then compared to other state-of-the-art detection frameworks. Our experimental results show both the necessity of appropriate prompt engineering and the potential efficacy of LLM prompting in adversarial detection. Furthermore, its combination with the also effective, second proposed method, yields competitive results and establishes our approach as a viable solution for plug-and-play detection of textual adversarial samples.

Keywords — Adversarial Attacks, Detection, Text Classification, Large Language Models, Natural Language Processing

Ευχαριστίες

Με την ολοκλήρωση αυτού του έργου, θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Γιώργο Στάμου, για τη δυνατότητα που μου έδωσε να εκπονήσω τη διπλωματική εργασία μου στο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης. Επίσης, θέλω να ευχαριστήσω την Μαρία Λυμπεραίου και τον Γιώργο Φιλανδριανό για την καθοδήγησή τους κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας μου.

Καθώς μαζί με αυτή την εργασία ολοκληρώνεται και το τωρινό ακαδημαϊκό ταξίδι μου, θέλω να ευχαριστήσω και τους δικούς μου ανθρώπους.

Τους γονείς μου, Σπύρο και Μαργαρίτα, που κατέστησαν δυνατό το να επικεντρωθώ στις σπουδές μου με όλη μου την ενέργεια.

Την αδερφή μου, την Άντζελα, που ήταν πάντα πρόθυμη να με βοηθήσει όταν την χρειαζόμουν.

Τον παππού μου, τον Μιχάλη, που είχε από πάντα όνειρο να γίνω επιστήμονας.

Τους φίλους μου, τον Κωνσταντίνο και τον Σωτήρη, που συμπορευτήκαμε μαζί από τα σχολικά χρόνια.

Τους συμφοιτητές μου, τον Χρήστο, τον Γιώργο και τον Λευτέρη, που δημιουργήσαμε μαζί τόσες αναμνήσεις.

Την σύντροφο μου, την Σομάγια, που με στήριξε στις δύσκολες στιγμές μου.

Νικόλαος Κώστας,

Νοέμβριος 2024

Contents

| | |
|---|-----------|
| Contents | 13 |
| List of Figures | 15 |
| 1 Εκτεταμένη Περίληψη στα Ελληνικά | 19 |
| 1.1 Θεωρητικό υπόβαθρο | 20 |
| 1.1.1 Ανταγωνιστικές Επιθέσεις στην Επεξεργασία Φυσικής Γλώσσας | 20 |
| 1.1.2 Άμυνες Κατά των Ανταγωνιστικών Επιθέσεων στο NLP | 23 |
| 1.2 Μεθοδολογία | 24 |
| 1.2.1 Συνεισφορά | 24 |
| 1.2.2 Σύνολα Δεδομένων | 24 |
| 1.2.3 Προτεινόμενες Μέθοδοι | 25 |
| 1.3 Πειράματα και Αποτελέσματα | 28 |
| 1.3.1 Σύνολα Δεδομένων & Μετρικές Αξιολόγησης | 28 |
| 1.3.2 Πειράματα Προτροπής LLM | 29 |
| 1.3.3 Πειράματα Ταξινόμησης Μεμονωμένων Προτάσεων | 38 |
| 1.3.4 Συνδυασμός Μεθόδων | 43 |
| 1.3.5 Σύγκριση με Μεθόδους Ανίχνευσης Τελευταίας Τεχνολογίας | 44 |
| 1.4 Συμπεράσματα | 46 |
| 1.4.1 Συζήτηση | 46 |
| 1.4.2 Μελλοντικές Κατευθύνσεις | 47 |
| 2 Introduction | 49 |
| 3 Adversarial Attacks in NLP | 51 |
| 3.1 Preliminaries | 52 |
| 3.1.1 Definitions | 52 |
| 3.1.2 From Image To Text | 52 |
| 3.2 Taxonomy of Textual Adversarial Attacks | 53 |
| 3.2.1 Model Access | 53 |
| 3.2.2 Target Type | 54 |
| 3.2.3 Perturbation Level | 54 |
| 3.2.4 Generation method | 54 |
| 3.2.5 Task | 55 |
| 3.3 Evaluation of Textual Adversarial Attacks | 56 |
| 3.3.1 Attack Effectiveness | 56 |
| 3.3.2 Attack Imperceptibility | 58 |
| 3.3.3 Attack Efficiency | 59 |
| 3.3.4 Attack Transferability | 59 |
| 3.4 Instances of Textual Adversarial Attacks | 59 |
| 3.4.1 Char-level Attack | 59 |
| 3.4.2 Word-level Attack | 60 |
| 3.4.3 Sentence-level Attack | 60 |

| | | |
|----------|--|------------|
| 3.4.4 | Multi-level Attack | 60 |
| 4 | Defenses Against Adversarial Attacks in NLP | 61 |
| 4.1 | Adversarial Training | 62 |
| 4.1.1 | Data Augmentation | 62 |
| 4.1.2 | Model Regularization | 62 |
| 4.2 | Anomaly Detection | 63 |
| 5 | Methodology | 65 |
| 5.1 | Contributions | 66 |
| 5.2 | Datasets | 66 |
| 5.2.1 | Text Classification Datasets | 66 |
| 5.2.2 | Adversarial Attacks | 66 |
| 5.2.3 | Attacked Models | 67 |
| 5.3 | Proposed Methods | 67 |
| 5.3.1 | LLM Prompting | 67 |
| 5.3.2 | Individual Sentences Classification | 71 |
| 6 | Experiments and Results | 75 |
| 6.1 | Preliminaries | 76 |
| 6.1.1 | Datasets | 76 |
| 6.1.2 | Evaluation Metrics | 77 |
| 6.2 | Experiments of LLM Prompting | 79 |
| 6.2.1 | Employed LLMs | 79 |
| 6.2.2 | Prompt Engineering | 80 |
| 6.2.3 | LLM Prompting | 84 |
| 6.2.4 | Human Evaluation | 90 |
| 6.2.5 | Sanity Test for Data Contamination | 92 |
| 6.3 | Experiments of Individual Sentences Classification | 94 |
| 6.3.1 | Total Agreements | 94 |
| 6.3.2 | Longest Continuous Agreement Sequence | 98 |
| 6.3.3 | Total Interchanges | 101 |
| 6.4 | Combination of Methods | 103 |
| 6.5 | Comparison with State-of-the-art Detection Methods | 105 |
| 6.5.1 | Compared Methods | 105 |
| 6.5.2 | Comparison Results | 105 |
| 7 | Conclusion | 107 |
| 7.1 | Discussion | 107 |
| 7.2 | Future Work | 107 |
| 8 | Bibliography | 109 |
| A | Appendix | 115 |
| A.1 | LLM Prompting | 115 |
| A.1.1 | Prompt Engineering | 115 |
| A.1.2 | Analysis by Text Length | 121 |

List of Figures

| | |
|--|----|
| 1.1.1 Παράδειγμα ανταγωνιστικής επίθεσης στην ανάλυση συναισθήματος. Η αντικατάσταση της λέξης «lack» με το συνώνυμό της «dearth» μεταβάλλει την πρόβλεψη του μοντέλου από αρνητική σε θετική, ενώ το συναισθήμα που αντιλαμβάνονται οι άνθρωποι παραμένει αρνητικό [54] | 21 |
| 1.1.2 Παράδειγμα ανταγωνιστικής επίθεσης στη συνεπαγωγή κειμένου. [2] | 22 |
| 1.2.1 Παράδειγμα εισόδων και εξόδων της GPT-3 με (α) τυπική πολλαπλή βολή, (β) πολλαπλή βολή με CoT, (γ) τυπική μηδενική βολή και (δ) μηδενική βολή με CoT. [26] | 26 |
| 1.2.2 Πλαίσιο ExpertPrompting. [73] | 26 |
| 1.2.3 Παράδειγμα των 3 προσεγγίσεων του IndivSC. Οι διπλές αγκύλες υποδηλώνουν τις λέξεις που έχουν διαταραχθεί από την αντίπαλη επίθεση. | 28 |
| 1.3.1 Σύγκριση της ανάλυσης που κάνουν τα LLM της οικογένειας Gemma 2 σε διαφορετικές κλίμακες μοντέλων, στο ίδιο ανταγωνιστικό κείμενο. Οι ανταγωνιστικές επεξεργασίες που έγιναν στο κείμενο εισόδου συμβολίζονται εντός παρενθέσεων και με κόκκινο χρώμα. | 34 |
| 1.3.2 Ανάλυση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του Gemma-2-27B, για επιθέσεις κατά του RoBERTa. | 35 |
| 1.3.3 Ανάλυση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του Claude-3-Haiku, για επιθέσεις κατά του RoBERTa. | 35 |
| 1.3.4 Ανάλυση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του GPT-4o-mini, για επιθέσεις κατά του RoBERTa. | 35 |
| 1.3.5 Boxplots του συνολικού ποσοστού συμφωνίας ταξινόμησης μεμονωμένων προτάσεων. | 38 |
| 1.3.6 Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Συμφωνιών ως ορίου για την ανταγωνιστική ανίχνευση. | 38 |
| 1.3.7 Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Συμφωνιών ως ορίου για την ανίχνευση ανταγωνιστικότητας, στο σύνολο δοκιμών. | 39 |
| 1.3.8 Για το BERT. | 40 |
| 1.3.9 Για το RoBERTa. | 40 |
| 1.3.10 Συνολικό ποσοστό συμφωνιών IndivSC σε όλα τα μήκη κειμένων. | 40 |
| 1.3.11 Για το BERT. | 40 |
| 1.3.12 Για το RoBERTa. | 40 |
| 1.3.13 Ανάλυση ανταγωνιστικής ανίχνευσης για τη μέθοδο Συνολικών IndivSC Συμφωνιών σε διάφορα μήκη κειμένου. | 40 |
| 1.3.14 Boxplots του ποσοστού της μεγαλύτερης συνεχούς ακολουθίας συμφωνίας IndivSC. | 41 |
| 1.3.15 Καμπύλες ROC για τη χρήση του ποσοστού Μακρύτερη Συνεχής Ακολουθία Συμφωνίας ως ορίου για την ανταγωνιστική ανίχνευση. | 41 |
| 1.3.16 Ανάλυση ανταγωνιστικής ανίχνευσης για τη μέθοδο Μακρύτερη Συνεχής Ακολουθία IndivSC Συμφωνίας σε όλα τα μήκη κειμένου. | 42 |
| 1.3.17 Boxplots του ποσοστού Συνολικών IndivSC Εναλλαγών. | 42 |
| 1.3.18 Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Εναλλαγών ως ορίου για την ανταγωνιστική ανίχνευση. | 43 |
| 3.1.1 Instance of an adversarial attack in image classification. The unnoticeable perturbation that is added to the original image effectively deceives the model, causing it to shift its classification prediction from "panda" to "gibbon". [18] | 52 |

| | |
|--|----|
| 3.1.2 Instance of an adversarial attack in sentiment analysis. Replacing the word "lack" with its synonym "dearth" alters the model's prediction from negative to positive, while the sentiment perceived by humans remains negative. [54] | 53 |
| 3.2.1 Instance of an adversarial attack in machine reading comprehension. [23] | 55 |
| 3.2.2 Instance of an adversarial attack in text entailment. [2] | 56 |
| 3.2.3 Taxonomy of Textual Adversarial Attacks | 57 |
| 3.4.1 Instances of adversarial examples generated via different perturbation levels. [65] | 59 |
| 5.3.1 Example inputs and outputs of GPT-3 with (a) standard Few-shot, (b) Few-shot-CoT, (c) standard Zero-shot, and (d) Zero-shot-CoT. [26] | 68 |
| 5.3.2 Expert Prompting framework. [73] | 69 |
| 5.3.3 Retrieval Augmented Generation overview. [34] | 69 |
| 5.3.4 (a) Monte Carlo Tree Search (MCTS) planning for automatic prompt generation. (b) A state transition example. Given a current state (prompt), the base model collects errors from the task dataset. The optimizer model provides corresponding error feedback. Then, the optimizer model updates the prompt based the feedback and transits to the next state. [68] | 70 |
| 5.3.5 Example of Total IndivSC Agreements. Double brackets denote words perturbed by the adversarial attack. | 72 |
| 5.3.6 Example of the 3 IndivSC approaches. Double brackets denote words perturbed by the adversarial attack. | 73 |
| 6.1.1 Length of IMDb texts. | 77 |
| 6.1.2 Length of AG-News texts. | 77 |
| 6.1.3 ROC Curve example. [71] | 78 |
| 6.2.1 Comparison of the analysis that the Gemma 2 LLMs do on different model scales, on the same adversarial text. The adversarial edits made to the input text are denoted within brackets and in red color. | 87 |
| 6.2.2 Example of GPT-4o-mini's analysis of an adversarial text. The edits made to the input text by the adversarial attack are denoted within brackets and in red color. | 88 |
| 6.2.3 Adversarial detection Recall across text lengths, for prompting Gemma-2-2B on attacks against RoBERTa. | 89 |
| 6.2.4 Adversarial detection Recall across text lengths, for prompting Phi-3-mini on attacks against RoBERTa. | 89 |
| 6.2.5 Adversarial detection Recall across text lengths, for prompting Llama-3.1-8B on attacks against RoBERTa. | 89 |
| 6.2.6 Adversarial detection Recall across text lengths, for prompting Gemma-2-9B on attacks against RoBERTa. | 89 |
| 6.2.7 Adversarial detection Recall across text lengths, for prompting Gemma-2-27B on attacks against RoBERTa. | 90 |
| 6.2.8 Adversarial detection Recall across text lengths, for prompting Claude-3-Haiku on attacks against RoBERTa. | 90 |
| 6.2.9 Adversarial detection Recall across text lengths, for prompting GPT-4o-mini on attacks against RoBERTa. | 90 |
| 6.2.10 Example of an original text containing multiple errors. 2/4 human evaluators classified it correctly, while all LLMs classified it incorrectly. | 92 |
| 6.2.11 Example of GPT-4o-mini fact-checking information during its analysis of an adversarial text. The adversarial edits made to the input text are denoted within brackets and in red color. | 93 |
| 6.3.1 Boxplots of the Total Individual Sentences Classification Agreement percentage, for BERT. | 94 |
| 6.3.2 ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for BERT. | 95 |
| 6.3.3 ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for BERT, on the test set. | 96 |
| 6.3.4 Total IndivSC Agreements percentage across text lengths, for BERT. | 96 |
| 6.3.5 Adversarial detection Recall for the Total IndivSC Agreements method across text lengths, for BERT. | 96 |
| 6.3.6 Boxplots of the Total IndivSC Agreement percentage, for RoBERTa. | 97 |

| | |
|--|-----|
| 6.3.7 ROC curves for using the Total IndivSC Agreement percentage as the threshold for adversarial detection, for RoBERTa. | 97 |
| 6.3.8 ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for RoBERTa, on the test set. | 97 |
| 6.3.9 Total IndivSC Agreements percentage across text lengths, for RoBERTa. | 98 |
| 6.3.10 Adversarial detection Recall for the Total Agreements method across text lengths, for RoBERTa. | 98 |
| 6.3.11 Boxplots of the Longest Continuous IndivSC Agreement Sequence percentage. | 98 |
| 6.3.12 ROC curves for using the Longest Continuous IndivSC Agreement Sequence percentage as the threshold for adversarial detection. | 99 |
| 6.3.13 ROC curves for using the Longest Continuous IndivSC Agreement Sequence percentage as the threshold for adversarial detection, for the test set. | 100 |
| 6.3.14 Longest Continuous IndivSC Agreement Sequence percentage across text lengths. | 100 |
| 6.3.15 Adversarial detection Recall for the Longest Continuous IndivSC Agreement Sequence method across text lengths. | 101 |
| 6.3.16 Boxplots of the Total IndivSC Interchanges percentage. | 101 |
| 6.3.17 ROC curves for using the Total IndivSC Interchanges percentage as the threshold for adversarial detection. | 102 |
| 6.3.18 ROC curves for using the Total IndivSC Interchanges percentage as the threshold for adversarial detection, for the test set. | 102 |
| 6.3.19 Total IndivSC Interchanges percentage across text lengths. | 103 |
| 6.3.20 Adversarial detection Recall for the Total IndivSC Interchanges method across text lengths. | 103 |
| | |
| A.1.1 Adversarial detection Recall across text lengths, for prompting Gemma-2-2B on attacks against BERT. | 121 |
| A.1.2 Adversarial detection Recall across text lengths, for prompting Phi-3-mini on attacks against BERT. | 121 |
| A.1.3 Adversarial detection Recall across text lengths, for prompting Llama-3.1-8B on attacks against BERT. | 121 |
| A.1.4 Adversarial detection Recall across text lengths, for prompting Gemma-2-9B on attacks against BERT. | 121 |
| A.1.5 Adversarial detection Recall across text lengths, for prompting Gemma-2-27B on attacks against BERT. | 122 |
| A.1.6 Adversarial detection Recall across text lengths, for prompting Claude-3-Haiku on attacks against BERT. | 122 |
| A.1.7 Adversarial detection Recall across text lengths, for prompting GPT-4o-mini on attacks against BERT. | 122 |

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

Contents

| | | |
|------------|---|-----------|
| 1.1 | Θεωρητικό υπόβαθρο | 20 |
| 1.1.1 | Ανταγωνιστικές Επιθέσεις στην Επεξεργασία Φυσικής Γλώσσας | 20 |
| 1.1.2 | Άμυνες Κατά των Ανταγωνιστικών Επιθέσεων στο NLP | 23 |
| 1.2 | Μεθοδολογία | 24 |
| 1.2.1 | Συνεισφορά | 24 |
| 1.2.2 | Σύνολα Δεδομένων | 24 |
| 1.2.3 | Προτεινόμενες Μέθοδοι | 25 |
| 1.3 | Πειράματα και Αποτελέσματα | 28 |
| 1.3.1 | Σύνολα Δεδομένων & Μετρικές Αξιολόγησης | 28 |
| 1.3.2 | Πειράματα Προτροπής LLM | 29 |
| 1.3.3 | Πειράματα Ταξινόμησης Μεμονωμένων Προτάσεων | 38 |
| 1.3.4 | Συνδυασμός Μεθόδων | 43 |
| 1.3.5 | Σύγκριση με Μεθόδους Ανίχνευσης Τελευταίας Τεχνολογίας | 44 |
| 1.4 | Συμπεράσματα | 46 |
| 1.4.1 | Συζήτηση | 46 |
| 1.4.2 | Μελλοντικές Κατευθύνσεις | 47 |

1.1 Θεωρητικό υπόβαθρο

1.1.1 Ανταγωνιστικές Επιθέσεις στην Επεξεργασία Φυσικής Γλώσσας

Τα τελευταία χρόνια, τα Βαθιά Νευρωνικά Δίκτυα (DNNs) έχουν χρησιμοποιηθεί αποτελεσματικά για την αντιμετώπιση δύσκολων προβλημάτων σε πολλαπλούς τομείς. Ωστόσο, έχει αποδειχθεί [61], [18] ότι τροποποιώντας κατάλληλα την είσοδο ενός DNN, είναι δυνατόν να το ξεγελάσουμε ώστε να κάνει λανθασμένη πρόβλεψη, ενώ η ανθρώπινη κρίση δεν επηρεάζεται. Αυτές οι τροποποιημένες εισόδους είναι γνωστές ως ανταγωνιστικά παραδείγματα και έναντι αυτών, η απόδοση των DNNs υποβαθμίζεται σημαντικά. Αυτό αποτελεί σημαντική απειλή για τα συστήματα που βασίζονται σε DNN, ιδίως αυτά που αναπτύσσονται σε πραγματικές, κρίσιμες από άποψη ασφάλεια εφαρμογές [57].

Σε αυτή την ενότητα θα ορίσουμε το πρόβλημα των ανταγωνιστικών επιθέσεων, θα δημιουργήσουμε μια ταξινόμηση για τις ανταγωνιστικές επιθέσεις κειμένου, θα διερευνήσουμε τον τρόπο αξιολόγησής τους και θα διερευνήσουμε ορισμένες χαρακτηριστικές μεθόδους επιθέσεων κειμένου.

Ορισμοί

Βαθύ Νευρωνικό Δίκτυο (DNN): Ένα βαθύ νευρωνικό δίκτυο (DNN) μπορεί να αναπαρασταθεί ως η συνάρτηση $F: X \rightarrow Y$, η οποία απεικονίζει το σύνολο εισόδου X στο σύνολο πρόβλεψης εξόδου Y .

Διαταραχή: Οι διαταραχές είναι μικρές αλλαγές που προστίθενται σκόπιμα στην αρχική είσοδο προκειμένου να παραπλανήσουν το DNN στο να κάνουν λάθος πρόβλεψη.

Ανταγωνιστικό Παράδειγμα: Οι επιτιθέμενοι δημιουργούν ανταγωνιστικά παραδείγματα προσθέτοντας μικρές διαταραχές στην αρχική είσοδο, οι οποίες στη συνέχεια ξεγελούν το DNN. Τυπικά, ένα ανταγωνιστικό παράδειγμα x' μπορεί να οριστεί ως εξής:

$$\begin{aligned} F(x) &= y \\ x' &= x + \varepsilon \\ F(x') &= y', y' \neq y \end{aligned} \tag{1.1.1}$$

, όπου ε είναι η προστιθέμενη διαταραχή. Ταυτόχρονα, το x' πρέπει να είναι ανεπαίσθητο από τους ανθρώπους, πράγμα που σημαίνει ότι οι άνθρωποι θα εξακολουθήσουν να το ταξινομούν με τον ίδιο τρόπο όπως το αρχικό παράδειγμα και ότι δεν θα πρέπει να είναι σε θέση να τα διακρίνουν μεταξύ τους.

Μοντέλο Στόχος: Το μοντέλο που δέχεται επίθεση με ανταγωνιστικά παραδείγματα ονομάζεται μοντέλο-στόχος.

Ταξινόμηση των Ανταγωνιστικών Επιθέσεων Κειμένου

Οι ανταγωνιστικές επιθέσεις στο NLP μπορούν να κατηγοριοποιηθούν χρησιμοποιώντας ένα πλήθος κριτηρίων. Σε αυτή την ενότητα, εισάγουμε μια ταξινόμηση για τέτοιες επιθέσεις.

Πρόσβαση στο Μοντέλο

Μια σημαντική διαφοροποίηση των ανταγωνιστικών παραδειγμάτων είναι αν παράγονται με τρόπο black-box ή white-box, όσον αφορά τη διαθέσιμη πρόσβαση στο στοχευμένο μοντέλο DNN.

Στην περίπτωση white-box, ο επιτιθέμενος έχει απεριόριστη πρόσβαση στο DNN, ενώ στην περίπτωση black-box, όλες οι εσωτερικές πληροφορίες του DNN δεν είναι προσβάσιμες και το μόνο που είναι διαθέσιμο είναι οι εξόδοι του.

Τύπος Στόχευσης

Οι ανταγωνιστικές επιθέσεις μπορούν να χωριστούν σε στοχευμένες και μη στοχευμένες με βάση την επιδιωκόμενη αλλαγή της πρόβλεψης εξόδου.

Σε μια στοχευμένη επίθεση, το ανταγωνιστικό παράδειγμα κατευθύνεται προς μια συγκεκριμένη πρόβλεψη, η οποία είναι ο στόχος του επιτιθέμενου, ενώ στις μη στοχευμένες επιθέσεις, ο μόνος σκοπός του αντιπάλου είναι

να εξαπατήσει το μοντέλο, πράγμα που σημαίνει ότι η πρόβλεψη μπορεί να είναι οποιαδήποτε λανθασμένη.

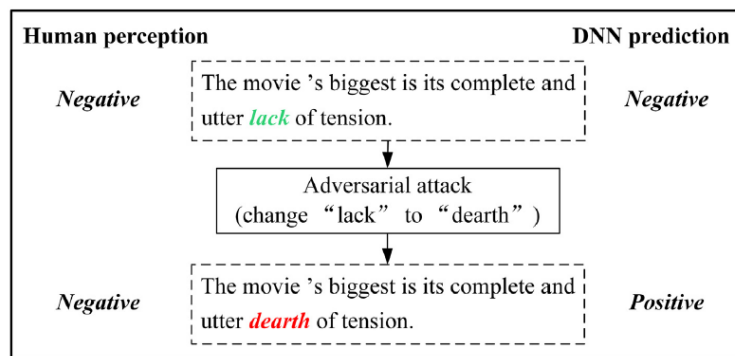


Figure 1.1.1: Παράδειγμα ανταγωνιστικής επίθεσης στην ανάλυση συναισθήματος. Η αντικατάσταση της λέξης «lack» με το συνώνυμό της «dearth» μεταβάλλει την πρόβλεψη του μοντέλου από αρνητική σε θετική, ενώ το συναισθημα που αντιλαμβάνονται οι άνθρωποι παραμένει αρνητικό [54]

Επίπεδο Διαταραχών

Το κριτήριο του επιπέδου διαταραχής αφορά τις συγκεκριμένες μονάδες κειμένου στις οποίες πραγματοποιούνται οι διαταραχές για τη δημιουργία ανταγωνιστικών παραδειγμάτων. Οι υποκατηγοριοποιήσεις είναι επίπεδο χαρακτήρα, επίπεδο λέξης, επίπεδο πρότασης και πολυεπίπεδο.

Επίπεδο Χαρακτήρα: Στις επιθέσεις σε επίπεδο χαρακτήρα, μεμονωμένοι χαρακτήρες μέσα σε λέξεις τροποποιούνται προκειμένου να δημιουργηθούν ανταγωνιστικά παραδείγματα.

Επίπεδο Λέξης: Οι επιθέσεις σε επίπεδο λέξης διαταράσσουν ολόκληρες λέξεις αντί για μεμονωμένους χαρακτήρες στις λέξεις.

Επίπεδο Πρότασης: Στις επιθέσεις σε επίπεδο πρότασης, μπορεί να τροποποιηθεί ολόκληρη η δομή της πρότασης.

Πολυεπίπεδο: Οι επιθέσεις πολλαπλών επιπέδων συνδυάζουν περισσότερους από έναν από τους προηγούμενους τύπους διαταραχών προκειμένου να ξεγελάσουν το μοντέλο-στόχο.

Μέθοδος Παραγωγής

Αυτή η ταξινόμηση αναφέρεται στη διαδικασία με την οποία παράγονται τα ανταγωνιστικά παραδείγματα. Χωρίζουμε τις πιθανές μεθόδους σε έξι τύπους: βασισμένες σε κλίση, βασισμένες σε βελτιστοποίηση, βασισμένες σε σημαντικότητα, βασισμένες σε επεξεργασία, βασισμένες σε παράφραση και βασισμένες σε παραγωγικό μοντέλο.

Βασισμένη σε Κλίση: Αυτές οι τεχνικές περιλαμβάνουν τον υπολογισμό της εμπρόσθιας παραγώγου στην είσοδο και τη δημιουργία ανταγωνιστικών διαταραχών μέσω της οπισθοδιάδοσης κλίσης.

Βασισμένη σε Βελτιστοποίηση: Σε αυτή τη στρατηγική, ο στόχος είναι η μεγιστοποίηση του σφάλματος πρόβλεψης του μοντέλου-θύματος, διατηρώντας παράλληλα τη διαφορά μεταξύ του ανταγωνιστικού παραδείγματος και της αρχικής εισόδου εντός ενός συγκεκριμένου ορίου.

Βασισμένη σε Σημαντικότητα: Στο πλαίσιο αυτής της προσέγγισης, η διαδικασία επιλογής των στοιχείων κειμένου που θα τροποποιηθούν και του τρόπου τροποποίησής τους εξαρτάται από τη σημαντικότητα κάθε στοιχείου σε σχέση με τις προβλέψεις του μοντέλου-θύματος.

Βασισμένη σε Επεξεργασία: Αυτή η μέθοδος περιλαμβάνει τη διαμόρφωση ανταγωνιστικών παραδειγμάτων αποκλειστικά μέσω βασικών ενεργειών, όπως η εισαγωγή, η αφαίρεση ή η ανταλλαγή χαρακτήρων, λέξεων ή προτάσεων μέσα στο κείμενο.

Βασισμένη σε Παράφραση: Οι επιθέσεις αυτού του τύπου παραφράζουν μια πρόταση του αρχικού αντιγράφου, διατηρώντας τη σημασιολογική ομοιότητα και τη γραμματική.

Βασισμένη σε Παραγωγικό Μοντέλο: Αυτή η τεχνική αξιοποιεί παραγωγικά μοντέλα όπως τα Αναγεννητικά Ανταγωνιστικά Δίκτυα (GANs) [17] και μοντέλα κωδικοποιητή-αποκωδικοποιητή για τη δημιουργία ανταγωνιστικών κειμένων.

| |
|---|
| Original Text Prediction: Entailment (Confidence = 86%) |
| Premise: A runner wearing purple strives for the finish line. |
| Hypothesis: A runner wants to head for the finish line. |
| Adversarial Text Prediction: Contradiction (Confidence = 43%) |
| Premise: A runner wearing purple strives for the finish line. |
| Hypothesis: A racer wants to head for the finish line. |

Figure 1.1.2: Παράδειγμα ανταγωνιστικής επίθεσης στη συνεπαγωγή κειμένου. [2]

Εφαρμογές

Με βάση την εφαρμογή NLP για την οποία δημιουργούνται οι ανταγωνιστικές επιθέσεις, μπορούμε να τις ταξινομήσουμε σε εννέα κύριες κατηγορίες: ταξινόμηση κειμένου, μηχανική μετάφραση, μηχανική κατανόηση κειμένου, απάντηση ερωτήσεων, συνεπαγωγή κειμένου, επισημείωση μέρους του λόγου, περίληψη κειμένου, παραγωγή διαλόγου και πολυτροπικές εφαρμογές.

Ταξινόμηση Κειμένου: Στην ταξινόμηση κειμένου, ο στόχος είναι να ταξινομηθεί το δεδομένο κείμενο σε μία από τις πιθανές κλάσεις.

Μηχανική Μετάφραση: Ένα σύστημα μηχανικής μετάφρασης χρησιμοποιείται για την αυτόματη μετάφραση κειμένου από μια γλώσσα σε μια άλλη.

Μηχανική Κατανόηση Κειμένου: Το έργο των συστημάτων μηχανικής κατανόησης κειμένου (MRC) είναι να εξάγουν νόημα από ένα δεδομένο κείμενο, όπως ακριβώς θα έκανε ένας άνθρωπος.

Απάντηση Ερωτήσεων: Τα συστήματα απάντησης ερωτήσεων (QA) μπορούν να απαντούν αυτόματα σε ερωτήσεις, δεδομένου ενός κειμενικού πλαισίου.

Συνεπαγωγή Κειμένου: Ο στόχος της συνεπαγωγής κειμένου, επίσης γνωστής ως συμπερασματολογία φυσικής γλώσσας (NLI), είναι η αξιολόγηση της σχέσης μεταξύ μιας προϋπόθεσης και ενός κειμένου υπόθεσης.

Επισημείωση Μέρους του Λόγου: Η επισημείωση μέρους του λόγου (POS) είναι η πρακτική της ταξινόμησης των λέξεων μέσα σε ένα κείμενο σε μια συγκεκριμένη κατηγοριοποίηση μέρους του λόγου, με βάση έναν συνδυασμό τόσο του εγγενούς ορισμού της λέξης όσο και του πλαισίου της μέσα στο κείμενο.

Περίληψη Κειμένου: Ο στόχος της περίληψης κειμένου είναι η παροχή μιας σύντομης και συνεκτικής περίληψης που να περιλαμβάνει τη γενική ουσία ενός συγκεκριμένου κειμένου.

Παραγωγή Διαλόγου: Η παραγωγή διαλόγου αντιπροσωπεύει μια εφαρμογή παραγωγής κειμένου όπου οι απαντήσεις παράγονται αυτόματα σύμφωνα με βάση τις εισόδους κειμένου του χρήστη.

Πολυτροπικές Εφαρμογές: Πέρα από τις εφαρμογές που ασχολούνται με αποκλειστικά κειμενικές εισόδους, υπάρχουν και πολυτροπικές εφαρμογές που μπορούν να κατηγοριοποιηθούν είτε ως κείμενο-και-όραση είτε ως κείμενο-και-ήχος.

Είναι σημαντικό να σημειωθεί ότι πολλές ανταγωνιστικές επιθέσεις μπορούν να χρησιμοποιηθούν σε πολλές διαφορετικές εφαρμογές NLP.

Αξιολόγηση των Ανταγωνιστικών Επιθέσεων Κειμένου

Κατά την αξιολόγηση των ανταγωνιστικών επιθέσεων, τέσσερις πρωταρχικές πτυχές χρήζουν εξέτασης: αποτελεσματικότητα, μη αντιληπτότητα, αποδοτικότητα και δυνατότητα μεταφοράς.

Αποτελεσματικότητα των Επιθέσεων

Η αξιολόγηση της αποτελεσματικότητας αντικατοπτρίζει την ικανότητα μιας ανταγωνιστικής επίθεσης να παραπληήσει ένα DNN ώστε να παράγει εσφαλμένες εξόδους και κατά συνέπεια να υποβαθμίσει την απόδοσή του. Για τη μέτρηση αυτής της πτυχής, μπορούμε να χρησιμοποιήσουμε τις ακόλουθες μετρικές απόδοσης.

Ποσοστό Ακρίβειας: Αυτή η μετρική υποδηλώνει την αναλογία των σωστά ταξινομημένων παραδειγμάτων από το μοντέλο-θύμα σε σχέση με το σύνολο των παραδειγμάτων εισόδου.

Ποσοστό Επιτυχίας Επίθεσης: Σε αντίθεση με το ποσοστό ακρίβειας, το ποσοστό επιτυχίας επίθεσης δηλώνει το ποσοστό των παραδειγμάτων που προβλέφθηκαν εσφαλμένα από το μοντέλο θύμα μεταξύ του συνόλου των παραδειγμάτων εισόδου.

Μείωση Ακρίβειας: Αυτή η μετρική προσφέρει μια πιο διαισθητική αξιολόγηση, καθώς χαρακτηρίζει τη μεταβολή της ακρίβειας πριν και μετά την αντίπαλη επίθεση.

Είναι σημαντικό να σημειωθεί ότι με βάση την εφαρμογή NLP που εξετάζεται, μπορούν επίσης να χρησιμοποιηθούν πιο εξειδικευμένες μετρικές, όπως η βαθμολογία **F1** ή η βαθμολογία **Area Under the Curve (AUC)** για εφαρμογές ταξινόμησης και η βαθμολογία **BiLingual Evaluation Understudy (BLEU)** για εφαρμογές παραγωγής.

Μη Αντιληπτότητα των Επιθέσεων

Οι μετρικές μη αντιληπτότητας για τις ανταγωνιστικές επιθέσεις στο NLP μπορούν να κατηγοριοποιηθούν σε δύο ομάδες: μετρικές συνεχούς χώρου για διανυσματικά δεδομένα και μετρικές διακριτού χώρου για ακατέργαστα κειμενικά δεδομένα.

Μετρικές Συνεχούς Χώρου

Μερικές από τις Μετρικές Συνεχούς Χώρου είναι η Ευκλείδεια Απόσταση, η Ομοιότητα Συνημιτόνου και η Απόσταση του Μετακινήτη Λέξεων (WMD).

Μετρικές Διακριτού Χώρου

Μερικές από τις Μετρικές Διακριτού Χώρου είναι η Απόσταση Επεξεργασίας, ο Αριθμός Αλλαγών, η Αξιολόγηση Γραμματικής και Σύνταξης και ο Συντελεστής Ομοιότητας Jaccard.

Αποτελεσματικότητα των Επιθέσεων

Η αποτελεσματικότητα της επίθεσης μετρά την αποδοτικότητα της όσον αφορά το χρόνο επίθεσης, τους υπολογιστικούς πόρους που απαιτούνται για την πιθανή εκπαίδευση και τη δημιουργία ανταγωνιστικών δειγμάτων και το πλήθος των ερωτημάτων στο μοντέλο-στόχο.

Μεταφερσιμότητα των Επιθέσεων

Η μεταφερσιμότητα δείχνει την ευελιξία μιας μεθόδου επίθεσης. Συγκεκριμένα, εκφράζει κατά πόσον τα ανταγωνιστικά παραδείγματα που έχουν παραχθεί για ένα συγκεκριμένο μοντέλο-στόχο, σε ένα συγκεκριμένο σύνολο δεδομένων, μπορούν να επιτεθούν αποτελεσματικά σε ένα άλλο μοντέλο ή σύνολο δεδομένων.

1.1.2 Άμυνες Κατά των Ανταγωνιστικών Επιθέσεων στο NLP

Σε αυτή την ενότητα αναλύουμε και κατηγοριοποιούμε τους διαφορετικούς τύπους άμυνας απέναντι στις ανταγωνιστικές επιθέσεις όπως αυτοί έχουν προταθεί από την ερευνητική κοινότητα για να μετριάσουν αυτό το πρόβλημα [67].

Αυτές οι στρατηγικές μπορούν να χωριστούν σε δύο κύριες κατηγορίες: την ανταγωνιστική εκπαίδευση και την ανίχνευση ανωμαλιών. Αυτή η διάκριση μπορεί να περιγραφεί διαφορετικά ως εξής: άμυνες που βελτιώνουν την ευρωστία ενός μοντέλου κατά τη φάση της εκπαίδευσής του και άμυνες που ανιχνεύουν τις ανταγωνιστικές επιθέσεις κατά τη διάρκεια της παραγωγής πρόβλεψης από το μοντέλο-θύμα.

Ανταγωνιστική Εκπαίδευση

Η ανταγωνιστική εκπαίδευση εισήχθη για πρώτη φορά από τους Goodfellow et al. [19] και είναι η στρατηγική υπεράσπισης των νευρωνικών δικτύων από ανταγωνιστικές επιθέσεις, όπου μέσω της εκπαίδευσής τους μαθαίνουν να αναγνωρίζουν σωστά τόσο τα αρχικά όσο και τα ανταγωνιστικά δείγματα.

Από τις εργασίες που χρησιμοποιούν την εκπαίδευση με αντίξοες διαταραχές, ορισμένες εμπίπτουν στην κατηγορία της επαύξησης δεδομένων, όπου δημιουργούνται ανταγωνιστικά παραδείγματα και στη συνέχεια τα εισάγουν στο σύνολο δεδομένων εκπαίδευσης του μοντέλου. Άλλες, χρησιμοποιούν τα παραγόμενα ανταγωνιστικά δείγματα ως κανονικοποιητές κατά την εκπαίδευση του μοντέλου.

Ανίχνευση Ανωμαλιών

Ο στόχος της ανίχνευσης ανωμαλιών είναι ελαφρώς διαφορετικός από αυτόν των προαναφερθέντων συστημάτων, καθώς περιλαμβάνει τον εντοπισμό και την ανάκτηση ανταγωνιστικών δειγμάτων, τα οποία θα μπορούσαν στη συνέχεια να χρησιμοποιηθούν για σκοπούς ερμηνευσιμότητας ή ειδοποίησης του μοντέλου-θύματος, σε περιπτώσεις όπου αυτό είναι επωφελές. Λαμβάνοντας υπόψη τη διακριτή φύση του κειμένου, φαίνεται λογικό να διερευνηθεί η δυνατότητα άμεσης διάκρισης μεταξύ αρχικών και ανταγωνιστικών δειγμάτων στο NLP. Επιπλέον, ορισμένες προσεγγίσεις στη βιβλιογραφία επιχειρούν να διορθώσουν και τις εντοπισμένες ανταγωνιστικές διαταραχές, προκειμένου να αυξήσουν την ανθεκτικότητα του μοντέλου.

1.2 Μεθοδολογία

1.2.1 Συνεισφορά

Οι συνεισφορές της παρούσας διατριβής είναι πολλαπλές και μπορούν να συνοψιστούν ως εξής:

- Αναπτύξαμε ένα απλό πλαίσιο για την ανάπτυξη προτροπών που επιτρέπει τη χρήση LLMs για το έργο της ανίχνευσης των ανταγωνιστικών επιθέσεων στην ταξινόμηση κειμένου. Πρόκειται για μια μέθοδο που είναι ανεξάρτητη από την αρχιτεκτονική του μοντέλου-θύματος και δοκιμάζεται έναντι πολλαπλών ανταγωνιστικών επιθέσεων.
- Πραγματοποιήσαμε μια ολοκληρωμένη ανάλυση της ικανότητας των LLMs στην ανίχνευση επιθέσεων. Για να αξιολογήσουμε περαιτέρω τα αποτελέσματά μας, πραγματοποιήσαμε μια ανθρώπινη αξιολόγηση και έναν έλεγχο ορθότητας για τη μόλυνση των δεδομένων.
- Παρουσιάσαμε έναν απλό, αλλά αποτελεσματικό τρόπο ανίχνευσης ανταγωνιστικών κειμένων, αξιοποιώντας το ίδιο το μοντέλο που δέχεται επίθεση. Αυτή η μέθοδος είναι επίσης μοντελο-αγνωστική και εξίσου επιτυχής έναντι πολλαπλών τεχνικών επίθεσης.
- Συνδυάσαμε τις δύο προτεινόμενες μεθόδους μας σε μια ενιαία προσέγγιση, επιτυγχάνοντας επιδόσεις που ανταγωνίζονται τις πλέον σύγχρονες μεθόδους ανίχνευσης.

1.2.2 Σύνολα Δεδομένων

Σύνολα Δεδομένων Ταξινόμησης Κειμένου

Για την παρούσα διατριβή, χρησιμοποιήσαμε 2 διαφορετικά σύνολα δεδομένων που χρησιμοποιούνται ευρέως για την ταξινόμηση κειμένου: IMDb [41] και AG-News [80].

IMDb Το σύνολο δεδομένων IMDb αποτελείται από κριτικές ταινιών του πραγματικού κόσμου, οι οποίες χρησιμοποιούνται για ταξινόμησης συναισθήματος, όπου σε κάθε κείμενο δίνεται μια ετικέτα Θετική ή Αρνητική.

AG-News Το σύνολο δεδομένων AG-news είναι ένα σώμα ειδησεογραφικών άρθρων, σχεδιασμένο για την ταξινόμηση κειμένου. Πιο συγκεκριμένα, κάθε κείμενο αποτελείται από έναν τίτλο είδησης σε συνδυασμό με μια σύντομη περιγραφή και μπορεί να ταξινομηθεί σε μία από τέσσερις διαφορετικές κατηγορίες: Κόσμος, Αθλητισμός, Επιχειρήσεις ή Επιστήμη/Τεχνολογία.

Ανταγωνιστικές Επιθέσεις

Όσον αφορά τη δημιουργία ανταγωνιστικών κειμένων, χρησιμοποιήσαμε 3 από τις πιο γνωστές και αποτελεσματικές επιθέσεις στην ερευνητική κοινότητα: TextFooler [24], PWWS [55] και BAE [15].

TextFooler Το TextFooler είναι ένα απλό αλλά αποτελεσματικό πλαίσιο ανταγωνιστικής επίθεσης, το οποίο εντοπίζει σημαντικές λέξεις σε ένα δεδομένο κείμενο και τις αντικαθιστά με σημασιολογικά παρόμοιες, μέχρι να καταφέρει να ανατρέψει την πρόβλεψη του μοντέλου-θύματος.

PWWS (Probability Weighted Word Saliency) Η επίθεση PWWS (Probability Weighted Word Saliency) είναι μια μέθοδος αντικατάστασης συνωνύμων η οποία λαμβάνει υπόψη τόσο την ευχρίνεια των λέξεων όσο και τη μεταβολή της πιθανότητας ταξινόμησης κατά τη δημιουργία διαταραχών.

BAE (BERT-based Adversarial Examples) Το BAE (BERT-based Adversarial Examples) είναι μια black-box επίθεση που αξιοποιεί τη μοντελοποίηση της γλώσσας με μάσκα του BERT μέσω τριών τύπων λειτουργιών: αντικατάσταση, εισαγωγή και διαγραφή μαρκών.

Μοντέλα Θύματα

Για τα πειράματά μας, χρησιμοποιήσαμε 2 διαφορετικά μοντέλα ταξινόμησης κειμένου που θα δεχθούν ανταγωνιστικές επιθέσεις: BERT [9] και RoBERTa [38].

BERT Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα γλωσσικό μοντέλο βασισμένο σε μετασχηματιστές, σχεδιασμένο για να προ-εκπαιδεύει αμφίδρομες, συμφραζόμενες αναπαραστάσεις λέξεων.

RoBERTa Το RoBERTa (Robustly Optimized BERT Pretraining Approach) χτίζει πάνω στο BERT, βελτιώνοντας τη διαδικασία προεκπαίδευσής του.

1.2.3 Προτεινόμενες Μέθοδοι

Προτροπή LLM

Προτείνουμε ένα πλαίσιο κατασκευής προτροπής, όπου με δεδομένο το στόχο της ανίχνευσης ανταγωνιστικών επιθέσεων στην ταξινόμηση κειμένου, ο σκοπός μας είναι να δημιουργήσουμε μια βελτιστοποιημένη προτροπή φυσικής γλώσσας που μεγιστοποιεί την ικανότητα ενός μεγάλου γλωσσικού μοντέλου να εκτελέσει αυτό το έργο.

Η προτεινόμενη από εμάς μέθοδος εμπνεύστηκε από ένα ευρύ φάσμα τεχνικών προτροπής που βρίσκονται στη βιβλιογραφία, και παρόλο που δημιουργήθηκε με σκοπό την ανίχνευση ανταγωνιστών, είναι αρκετά γενική ώστε, αν προσαρμοστεί κατάλληλα, να μπορεί να αξιοποιηθεί και για άλλες εργασίες.

Ξεκινάμε με μια βασική προτροπή P_0 , στην οποία πραγματοποιούμε μια σειρά διαδοχικών βελτιώσεων, μέχρι να δημιουργήσουμε την τελική προτροπή.

1) Αλυσιδωτός Συλλογισμός (CoT)

Στο [70] εισήχθη η προτροπή Αλυσιδωτού Συλλογισμού (CoT) που επιτρέπει στα Μεγάλα Γλωσσικά Μοντέλα να εκτελούν πολύπλοκους συλλογισμούς μέσω ενδιάμεσων βημάτων, εξοπλίζοντάς τα καλύτερα ώστε να φτάνουν στη σωστή λύση, για ένα πλήθος εφαρμογών. Έχει αποδειχθεί ότι επαρκώς μεγάλα Γλωσσικά Μοντέλα είναι ικανά να παράγουν αυτές τις αλυσίδες συλλογισμού είτε μέσω προτροπής πολλαπλών βολών, είτε μέσω προτροπής μηδενικής βολής [26], όταν εισάγεται η απλή φράση «Ας σκεφτούμε βήμα προς βήμα» στο τέλος μιας δεδομένης προτροπής.

Στα πειράματά μας, χρησιμοποιήσαμε το CoT μηδενικής βολής.

2) Ταυτότητα Ειδικού

Πρόσφατες μελέτες [52] [27] έχουν εντοπίσει ότι, όταν παρέχονται επαρκώς λεπτομερείς οδηγίες, τα LLMs διαθέτουν την ικανότητα να λειτουργούν ως ειδικοί πράκτορες.



Figure 1.2.1: Παράδειγμα εισόδων και εξόδων της GPT-3 με (α) τυπική πολλαπλή βολών, (β) πολλαπλή βολή με CoT, (γ) τυπική μηδενική βολή και (δ) μηδενική βολή με CoT. [26]

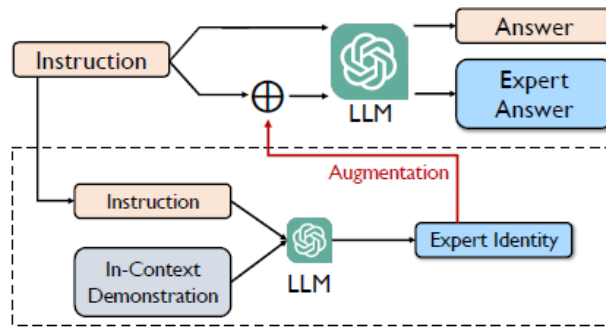


Figure 1.2.2: Πλαίσιο ExpertPrompting. [73]

Ακολουθώντας την απλή στρατηγική του Expert Prompting [73], παράγουμε αυτόματα μια ταυτότητα ειδικού, προσαρμοσμένη στο έργο της ανταγωνιστικής ανίχνευσης, παρέχοντας σε ένα LLM ζεύγη εντολών-εμπειρογνώμονα ως παραδείγματα πολλαπλών βολών, προκειμένου να το καθοδηγήσουμε στη σύλληψη μιας κατάλληλης περιγραφής ταυτότητας.

3) Γνώση Τομέα & Καθοδήγηση Λύσης

Παραγωγή Επαυξημένη με Ανάκτηση (RAG) [34] είναι η διαδικασία βελτίωσης προ-εκπαιδευμένων γλωσσικών μοντέλων παραμετρικής μνήμης, με την ενσωμάτωση ενός μη παραμετρικού στοιχείου μνήμης, όπως ένας πυκνός διανυσματικός δείκτης εξωτερικής γνώσης από τη Wikipedia.

Εμπνευσμένοι από αυτή την τεχνική, υποθέτουμε ότι η προσθήκη ειδικής γνώσης τομέα στην προτροπή ενός LLM μπορεί να είναι επωφελής για την απόδοσή του. Ως εκ τούτου, εμπλουτίζουμε την προτροπή μας με προσεκτικά επιλεγμένες πληροφορίες τις οποίες ένα LLM μπορεί να χρησιμοποιήσει προς την κατεύθυνση του έργου του για την ανίχνευση ανταγωνιστικών επιθέσεων.

Αναλύοντας την αλυσίδα σκέψης σε απαντήσεις LLM, έχει συναχθεί το συμπέρασμα [70] ότι όταν το παραγόμενο αποτέλεσμα είναι ακριβές, τότε η αλυσίδα σκέψης είναι γενικά σωστή και λογική. Όσον αφορά τις λανθασμένες απαντήσεις, υπάρχει ένα μέρος το οποίο θα μπορούσε να διορθωθεί εάν επιλύονταν ορισμένες κατηγορίες λαθών. Αυτές οι προκλήσεις μπορούν να μετριαστούν με την καθοδήγηση του Μεγάλου Γλωσσικού Μοντέλου στην κατανόηση της δεδομένης εργασίας, τη διαίρεσή της σε μικρές επιμέρους διεργασίες και την επίλυση αυτών βήμα προς βήμα [66].

Με βάση αυτές τις γνώσεις, ενισχύουμε τη γνώση του τομέα που περιέχεται στην προτροπή μας με μια συνιστώσα καθοδήγησης λύσεων. Αυτό περιλαμβάνει την προσθήκη πληροφοριών σχετικά με τα πιθανά σημάδια των ανταγωνιστικών επιθέσεων κειμένου, η οποία συνίσταται στην παροχή πληροφοριών σχετικά με τις επιθέσεις και στην ενθάρρυνση του LLM να εστιάσει σε συγκεκριμένα βήματα, με βάση αυτές τις πληροφορίες, όταν εξετάζει αν ένα κείμενο είναι ανταγωνιστικό ή όχι. Παρόλα αυτά, επειδή το έργο της ανταγωνιστικής ανίχνευσης δεν έχει μια προκαθορισμένη λύση, στοχεύουμε να μην περιορίσουμε υπερβολικά το LLM σε ένα αυστηρό σύνολο κανόνων για να σκεφτεί.

4) Διόρθωση Σφαλμάτων

Αντλώντας έμπνευση από τις ανθρώπινες διαδικασίες δοκιμής και λάθους καθώς και από την ικανότητα αυτοαναστοχασμού των LLMs [59], το PromptAgent [68] εισήγαγε μια προσέγγιση στρατηγικής βελτιστοποίησης της διαδικασίας προτροπής, όπου το πρόβλημα βελτιστοποίησης της προτροπής διατυπώνεται ως μια διαδικασία απόφασης Markov και χρησιμοποιούνται ανατροφοδοτήσεις σφαλμάτων για την εύρεση βέλτιστων προτροπών. Βασιζόμενοι στις γνώσεις σχετικά με τους τύπους των λαθών που κάνουν τα LLMs, χρησιμοποιούμε αυτή την προσέγγιση σχεδιασμού βάσει αρχών ως έναν τρόπο ενίσχυσης της προτροπής μας με διορατική ανατροφοδότηση λαθών.

Εξαγωγή Απαντήσεων

Στην προτροπή μηδενικής βολής, είναι απαραίτητο το LLM να ερωτηθεί και δεύτερη φορά, ώστε να εξαχθεί η τελική απάντηση από την αρχική. Αυτό γίνεται με τη συνένωση τριών στοιχείων: της αρχικής προτροπής, της αρχικής εξόδου του LLM και μιας προτροπής εξαγωγής απάντησης.

Ταξινόμηση Μεμονωμένων Προτάσεων

Όσον αφορά την εφαρμογή της ταξινόμησης κειμένων, σπάζουμε ένα δεδομένο κείμενο σε επιμέρους προτάσεις και αναλύουμε κάθε μία ξεχωριστά. Πιστεύουμε ότι για αμετάβλητα κείμενα, η πλειονότητα των προτάσεων τους θα λάβει την ίδια ταξινόμηση με το συνολικό κείμενο. Αντίθετα, αν ένα ανταγωνιστικό κείμενο χωριστεί σε προτάσεις, ενδέχεται να επηρεαστούν μόνο οι ταξινομήσεις των προτάσεων που περιέχουν αλλοίωση. Επομένως, οι αδιατάραχτες προτάσεις, και ακόμη και ορισμένες αλλοιωμένες, θα διατηρήσουν την αρχική τους ετικέτα, η οποία πιθανότατα συμφωνεί με την αρχική ετικέτα του κειμένου. Δεδομένου ότι η ταξινόμηση ολόκληρου του κειμένου έχει ανατραπεί από την επίθεση, αυτό σημαίνει αναπόφευκτα ότι η ετικέτα πολλών ανεξάρτητων προτάσεων θα είναι αντίθετη από τη συνολική, σε ανταγωνιστικά δείγματα.

Με βάση αυτή τη γνώση, αναπτύσσουμε μια απλή, αλλά αποτελεσματική μέθοδο η οποία εξετάζει τις ταξινομήσεις που δίνονται σε κάθε μεμονωμένη πρόταση ενός κειμένου και τις συγκρίνει με την ταξινόμηση που δίνεται σε ολόκληρο το κείμενο, με στόχο την εξαγωγή συμπερασμάτων που μπορούν να χρησιμοποιηθούν για τον εντοπισμό ανταγωνιστικών δειγμάτων. Ονομάσαμε αυτή τη μέθοδο Ταξινόμηση Μεμονωμένων Προτάσεων (Individual Sentences Classification - IndivSC) και αναπτύξαμε 3 προσεγγίσεις για αυτήν.

Συνολικές Συμφωνίες

Σε αυτή την προσέγγιση, εξετάζουμε πόσες από τις προτάσεις σε ένα δεδομένο κείμενο έχουν την ίδια ταξινόμηση με αυτή που δόθηκε σε ολόκληρο το κείμενο. Οι συνολικές συμφωνίες αντιπροσωπεύουν το πλήθος των ταξινομήσεων των μεμονωμένων προτάσεων που συμφωνούν με τη συνολική ταξινόμηση, ως κλάσμα του συνολικού πλήθους προτάσεων του κειμένου.

Ένα ενδεικτικό παράδειγμα δίνεται στο σχήμα 1.2.3, όπου το ποσοστό συνολικών συμφωνιών πέφτει από το 90% όπου ήταν στο αρχικό κείμενο, σε μόλις 30% στο ανταγωνιστικό.

Πιστεύουμε ότι αυτή η διαφορά μεταξύ αρχικών και ανταγωνιστικών κειμένων, όσον αφορά τις ταξινομήσεις των επιμέρους προτάσεων τους, εμφανίζεται με στατιστικά σημαντικό τρόπο, γεγονός που θα μας επιτρέψει να αναπτύξουμε αποτελεσματικές μεθόδους ανίχνευσης που θα την αξιοποιούν.

Μακρύτερη Συνεχής Ακολουθία Συμφωνίας

Αυτή η επόμενη προσέγγιση συνεπάγεται την εύρεση της μεγαλύτερης συνεχούς ακολουθίας προτάσεων στην οποία οι ταξινομήσεις κάθε μεμονωμένης πρότασης συμφωνούν με την ταξινόμηση ολόκληρου του κειμένου. Το



Figure 1.2.3: Παράδειγμα των 3 προσεγγίσεων του IndivSC. Οι διπλές αγκύλες υποδηλώνουν τις λέξεις που έχουν διαταραχθεί από την αντίπαλη επίθεση.

μήκος αυτής της ακολουθίας είναι η Μακρύτερη Συνεχής Ακολουθία Συμφωνίας και υπολογίζεται ως κλάσμα του συνολικού μήκους του κειμένου, που μετράται σε προτάσεις.

Στο Σχήμα 1.2.3, βλέπουμε ένα παράδειγμα στο οποίο το μήκος της Μακρύτερης Συνεχούς Συμφωνίας Ακολουθίας πηγάνει από 80% του συνολικού μήκους του αρχικού κειμένου, σε μόλις 20% του μήκους του ανταγωνιστικού κειμένου.

Συνολικές Εναλλαγές

Στην τελική προσέγγιση, εξετάζουμε το πλήθος των εναλλαγών που συμβαίνουν σε σχέση με τις ταξινομήσεις που δίνονται σε κάθε πρόταση στην ακολουθία του συνολικού κειμένου. Οι Συνολικές Εναλλαγές ισούνται με το ποσό των φορών που αλλάζει η ταξινόμηση κατά τη μετάβαση από τη μία πρόταση στην επόμενη, ως κλάσμα του συνόλου των προτάσεων του κειμένου.

Στο παράδειγμα του Σχήματος 1.2.3, παρατηρούμε ότι ενώ στο αρχικό κείμενο, το ποσοστό Συνολικών Εναλλαγών είναι 22%, στην περίπτωση του ανταγωνιστικού κειμένου διπλασιάζεται σε 44%.

1.3 Πειράματα και Αποτελέσματα

1.3.1 Σύνολα Δεδομένων & Μετρικές Αξιολόγησης

Για τα πειράματά μας, όλα τα τελικά σύνολα δεδομένων δημιουργήθηκαν χρησιμοποιώντας το πλαίσιο TextAttack [44] στα αρχικά σύνολα δεδομένων ταξινόμησης κειμένου. Για κάθε συνδυασμό των 2 συνόλων δεδομένων ταξινόμησης κειμένου (IMDb, AG-News), των 3 μεθόδων επίθεσης κειμένου (TextFooler, PWWS, BAE) και των 2 μοντέλων-θυμάτων (BERT, RoBERTa), τα δεδομένα δειγματοληπτήθηκαν με τέτοιο τρόπο ώστε το υποσύνολο που περιέχει τα ανταγωνιστικά δείγματα να είναι διαχωρισμένο από το υποσύνολο που περιέχει τα πρωτότυπα δείγματα και η αναλογία μεταξύ των πρωτότυπων και των ανταγωνιστικών δειγμάτων να είναι περίπου 1:1.

Για την αξιολόγηση της απόδοσης των μεθόδων μας, χρησιμοποιήσαμε 4 διαφορετικές μετρικές: Ανάκληση, βαθμολογία F1, Ακρίβεια και AUC.

1.3.2 Πειράματα Προτροπής LLM

Για την πραγματοποίηση των πειραμάτων μας χρησιμοποιήσαμε 7 LLMs, ποικίλων αρχιτεκτονικών και μεγεθών - συγκεκριμένα, τα Phi-3-mini, Llama-3.1-8B, Gemma-2-2B,9B,27B, Claude-3-Haiku και GPT-4o mini.

Μηχανική Προτροπών

Για τα πειράματα μηχανικής προτροπών, χρησιμοποιούμε το Phi-3-mini, ένα από τα πιο ισχυρά μικρής κλίμακας LLM ανοικτού κώδικα, και το Gemma-2-9B, ένα από τα μεσαίου μεγέθους LLM ανοικτού κώδικα με τις καλύτερες επιδόσεις.

Εξαγωγή Απάντησης

Στην πειραματική μας διάταξη, ενσωματώσαμε ένα επιπλέον βήμα προτροπής ώστε να εξάγουμε την τελική απάντηση από την παραγόμενη έξοδο του LLM.

0) Βασική Προτροπή

Για αρχή, πραγματοποιήσαμε πειράματα χρησιμοποιώντας δύο εκδόσεις μιας βασικής προτροπής που απλώς αναφέρει το ζητούμενο, προκειμένου να αξιολογήσουμε τις δυνατότητες των Μεγάλων Γλωσσικών Μοντέλων στο έργο της ανίχνευσης ανταγωνιστικών κειμένων, χωρίς ιδιαίτερες βελτιώσεις στην προτροπή.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.35 | 0.41 | 0.50 |
| | v2 | 0.36 | 0.43 | 0.50 |
| Gemma-2-9B | v1 | 0.20 | 0.32 | 0.58 |
| | v2 | 0.24 | 0.37 | 0.60 |

Table 1.1: Αποτελέσματα ανταγωνιστικής ανίχνευσης μέσω Βασικής Προτροπής.

Όπως προκύπτει από τα αποτελέσματα του Πίνακα 1.1, τα LLM δεν φαίνεται να έχουν εγγενή ικανότητα στον εντοπισμό επιθέσεων κειμενικού χαρακτήρα, έχοντας ουσιαστικά χειρότερες επιδόσεις από τις τυχαίες. Παρόλο που το Phi-3-mini επιτυγχάνει υψηλότερες μετρικές Recall και F1 από το Gemma-2-9B, με βάση το γεγονός ότι η ακρίβειά του είναι 50% και για τις δύο εκδόσεις της προτροπής, μπορεί να συναχθεί το συμπέρασμα ότι οι προβλέψεις του είναι αυθαίρετες. Αυτό επιβεβαιώνεται και με τη χειροκίνητη επιθεώρηση των προβλέψεων του Phi-3-mini. Η χαμηλή απόδοση του Gemma-2-9B, από την άλλη πλευρά, δεν φαίνεται να προέρχεται από τυχαίες προβλέψεις, αλλά από την αδυναμία του να ανιχνεύσει ανταγωνιστικά δείγματα. Συνολικά, τα αποτελέσματα αυτά αναδεικνύουν την ανάγκη για μηχανική προτροπών για το έργο της ανταγωνιστικής ανίχνευσης, τουλάχιστον για μοντέλα αυτής της κλίμακας.

1) Αλυσιδωτός Συλλογισμός (CoT)

Όσον αφορά τον Αλυσιδωτό Συλλογισμό, χρησιμοποιήσαμε την προσέγγιση μηδενικής βολής για να προκαλέσουμε διαδοχική συλλογιστική από τα μοντέλα, προσθέτοντας τη φράση «Ας σκεφτούμε βήμα προς βήμα» στην αρχή της απάντησης του LLM.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.05 | 0.09 | 0.50 |
| | v2 | 0.02 | 0.05 | 0.50 |
| Gemma-2-9B | v1 | 0.23 | 0.36 | 0.58 |
| | v2 | 0.43 | 0.54 | 0.63 |

Table 1.2: Αποτελέσματα ανταγωνιστικής ανίχνευσης μέσω Αλυσιδωτού Συλλογισμού.

Ο πίνακας 1.2 περιέχει τις μετρικές που προκύπτουν από την προσθήκη CoT με μηδενική βολή και στις δύο εκδόσεις της βασικής προτροπής και τα ευρήματά του τεκμηριώνουν την έλλειψη ικανότητας των LLMs στην εφαρμογή μας χωρίς περαιτέρω καθοδήγηση, ακόμη και όταν χρησιμοποιούν CoT. Η επίδοση του Phi-3-mini στον εντοπισμό αντίπαλων κειμένων πέφτει σχεδόν στο 0. Τούτου λεχθέντος, το Gemma-2-9B εμφανίζει

σημαντική αύξηση επιδόσεων, ιδίως όταν χρησιμοποιεί τη δεύτερη έκδοση της προτροπής, με την Ανάκληση να αυξάνεται κατά 19%. Αυτή η αντίθεση μπορεί να αποδοθεί στο γεγονός ότι το Gemma-2-9B είναι μεγαλύτερο σε μέγεθος και επομένως έχει μεγαλύτερη ικανότητα συλλογισμού.

2) Ταυτότητα Ειδικού

Για να καθορίσουμε την ταυτότητα του ειδικού, επιμεληθήκαμε χειροκίνητα τρεις αναλυτικές περιγραφές ειδικών, οι οποίες προστέθηκαν ως παραδείγματα πολλαπλών βολών σε προτροπή παραγωγής που έγινε στο GPT-4.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | no CoT | 0.04 | 0.08 | 0.51 |
| | with CoT | 0.06 | 0.10 | 0.52 |
| Gemma-2-9B | no CoT | 0.42 | 0.58 | 0.69 |
| | with CoT | 0.57 | 0.65 | 0.68 |

Table 1.3: Αποτελέσματα ανταγωνιστικής ανίχνευσης μέσω Ταυτότητας Ειδικού.

Ο πίνακας 1.3 περιέχει τα αποτελέσματα που προέκυψαν από την προτροπή των μοντέλων είτε χρησιμοποιώντας αποκλειστικά την ταυτότητα ειδικού, είτε συνδυάζοντάς την με τον αλυσιδωτό συλλογισμό. Και για τα δύο LLM, οι μετρικές που ακολουθούν την αποκλειστική χρήση του πλαισίου Ταυτότητας Ειδικού είναι παρόμοιες με αυτές που επιτυγχάνονται όταν χρησιμοποιείται αποκλειστικά το CoT. Ωστόσο, όταν χρησιμοποιούνται και οι δύο μέθοδοι ταυτόχρονα, οι επιδόσεις που επιτυγχάνονται είναι οι καλύτερες μέχρι στιγμής. Στην περίπτωση του Phi-3-mini, οι επιδόσεις βελτιώνονται κατά μερικές ποσοστιαίες μονάδες, αλλά το μοντέλο εξακολουθεί να είναι ουσιαστικά ανίκανο να επιτύχει την ανίχνευση ανταγωνιστικού κειμένου. Εν τω μεταξύ, το Gemma-2-9B παρουσιάζει άλλο ένα σημαντικό άλμα στις επιδόσεις, με την ανάκληση και το σκορ F1 να φτάνουν σε επαρκείς τιμές 57% και 65% αντίστοιχα.

3) Γνώση Τομέα & Καθοδήγηση Λύσης

Προκειμένου να αυξήσουμε τις δυνατότητες των μοντέλων στον εντοπισμό ανταγωνιστικών κειμένων, προσθέσαμε γνώση του τομέα στην πρώτη έκδοση της προτροπής αυτής της ενότητας. Αυτό το κάναμε δίνοντας έναν ορισμό του τι συνιστά μια ανταγωνιστική διαταραχή και επαναδιατυπώνοντας το καθήκον του LLM που είναι να ανιχνεύει τέτοιες διαταραχές.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.10 | 0.18 | 0.54 |
| | v2 | 0.24 | 0.38 | 0.59 |
| | v3 | 0.62 | 0.68 | 0.70 |
| Gemma-2-9B | v1 | 0.82 | 0.72 | 0.68 |
| | v2 | 0.75 | 0.76 | 0.75 |
| | v3 | 0.80 | 0.76 | 0.74 |

Table 1.4: Αποτελέσματα ανταγωνιστικής ανίχνευσης μέσω Γνώσης Τομέα & Καθοδήγησης Λύσης.

Από τον πίνακα 1.4, μπορούμε να διαπιστώσουμε ότι αυτή η προσθήκη φαίνεται να ενισχύει τις ικανότητες ανίχνευσης και των δύο μοντέλων. Έχοντας πει αυτό, παρατηρούμε ότι στην περίπτωση του Gemma-2-9B, η Ανάκληση έχει αυξηθεί δυσανάλογα κατά 25%, ενώ η Ακρίβεια έχει παραμείνει η ίδια. Αυτό ουσιαστικά σημαίνει ότι το μοντέλο εντοπίζει περισσότερα ανταγωνιστικά παραδείγματα, αλλά αυτό γίνεται με το κόστος της δημιουργίας πάρα πολλών ψευδώς θετικών αποτελεσμάτων.

Προκειμένου να μετριάσουμε αυτό το πρόβλημα, βελτιώνουμε την έκδοση 2 της προτροπής, καθοδηγώντας το Μεγάλο Γλωσσικό Μοντέλο να λαμβάνει υπόψη όλες τις δεδομένες πληροφορίες, να αναλύει τη δεδομένη εργασία σε ενδιάμεσα βήματα και να εξετάζει τυχόν παρατυπίες που εντοπίζει στο πλαίσιο, πριν καταλήξει σε μια τελική πρόβλεψη. Όπως φαίνεται από τα πειραματικά αποτελέσματα, η νέα έκδοση της προτροπής επιλύει το προαναφερθέν πρόβλημα. Το γεγονός ότι το Phi-3-mini παρουσιάζει επίσης ένα σημαντικό άλμα στην απόδοση όταν το προτρέπει με αυτόν τον τρόπο, αναδεικνύει την εποικοδομητική επίδραση του CoT.

Τέλος, στην έκδοση 3 της προτροπής, καθοδηγούμε το LLM με συγκεκριμένα βήματα επίλυσης που πρέπει να ακολουθήσει. Αυτές οι οδηγίες εμπνεύστηκαν κατ' αρχάς από τον τρόπο με τον οποίο τα ίδια τα LLM προσεγγίζουν το έργο της ανταγωνιστικής ανίχνευσης, καθώς και τον τρόπο με τον οποίο άλλες μέθοδοι στη βιβλιογραφία προσεγγίζουν την ανίχνευση ανταγωνιστικών επιθέσεων στο NLP. Δηλαδή, ενθαρρύνοντας το LLM να αναζητήσει τυχόν ανώμαλα χαρακτηριστικά μέσα σε ένα δεδομένο δείγμα, που το θέτουν εκτός των συνηθών χαρακτηριστικών του αδιατάραχτου κειμένου.

Τα τελικά αποτελέσματα του Phi-3-mini παρουσιάζουν τεράστια βελτίωση, με την Ανάκληση να αυξάνεται στο 62%, γεγονός που αποτελεί θετική ένδειξη ότι μέσω της επιμελημένης προτροπής, μπορούν να προκύψουν ικανότητες ανταγωνιστικής ανίχνευσης ακόμη και για μικρότερα μοντέλα. Για το Gemma-2-9B, παρατηρούμε αύξηση της ανάκλησης.

4) Διόρθωση Σφαλμάτων

Μέσω του τμήματος διόρθωσης σφαλμάτων της προτροπής μας, ακολουθώντας το πλαίσιο PromptAgent, επιδιώξαμε να διορθώσουμε τα πιο συνηθισμένα λάθη που έκαναν οι LLM στο πλαίσιο της συλλογιστικής τους.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.63 | 0.63 | 0.66 |
| | v2 | 0.64 | 0.65 | 0.68 |
| | v3 | 0.67 | 0.64 | 0.66 |
| Gemma-2-9B | v1 | 0.80 | 0.75 | 0.75 |
| | v2 | 0.77 | 0.72 | 0.73 |
| | v3 | 0.75 | 0.72 | 0.74 |

Table 1.5: Αποτελέσματα ανταγωνιστικής ανίχνευσης μέσω Διόρθωσης Σφαλμάτων.

Παρά το γεγονός ότι η ανατροφοδότηση σφάλματος που παράγεται είναι λογικά ακριβής, τα αποτελέσματα του πίνακα 1.5 δείχνουν ότι δεν οδηγούν σε ιδιαίτερες βελτιώσεις στην απόδοση του μοντέλου. Κυρίως, βλέπουμε μη δομημένες διακυμάνσεις σε όλες τις μετρικές, οι οποίες αναδεικνύουν την ευαισθησία των LLMs στις αλλαγές της εισόδου τους και δείχνουν ότι δεν υπάρχουν άλλα περιθώρια για βελτιστοποίηση, τουλάχιστον για αυτά τα μοντέλα. Για το Phi-3-mini, η έκδοση προτροπής με τις καλύτερες επιδόσεις είναι η τρίτη, η οποία είναι ενδιαφέρουσα λόγω του γεγονότος ότι χρησιμοποιεί πιο λεπτομερή γλώσσα για την περιγραφή των βημάτων καθοδήγησης της λύσης. Για το Gemma-2-9B, η τελική έκδοση 1 της προτροπής είναι η βέλτιστη και δεν εισάγει σημαντικές αλλαγές.

Κατά τη διάρκεια των υπόλοιπων πειραμάτων μας, για τα υπόλοιπα LLM χρησιμοποιήθηκε η ίδια προτροπή που σχεδιάστηκε για το Gemma-2-9B, εκτός από το Phi-3-mini που χρησιμοποίησε τη δική του σχεδιασμένη προτροπή. Αυτό έγινε λόγω υπολογιστικών περιορισμών, αλλά κυρίως για να καταστεί δυνατή η σύγκριση όσον αφορά την κλίμακα του μοντέλου στην περίπτωση της οικογένειας μοντέλων Gemma-2 και για την ανάλυση του πόσο καλά η προτροπή μας μπορεί να γενικευτεί για διαφορετικά μοντέλα.

Προτροπή LLMs

Ποσοτικά Αποτελέσματα

Στον Πίνακα 1.6, παρουσιάζουμε τα αποτελέσματα που προκύπτουν από την προτροπή όλων των χρησιμοποιούμενων LLM για το έργο της ανίχνευσης ανταγωνιστικών επιθέσεων στην ταξινόμηση κειμένου.

Είναι προφανές ότι η πρωταρχική μεταβλητή που επηρεάζει την απόδοση είναι η κλίμακα του μοντέλου, βλέποντας ότι τα μεγαλύτερα LLMs υπερτερούν σταθερά έναντι των μικρότερων. Για σκοπούς ανάλυσης, μπορούμε να χωρίσουμε τα μεγάλα γλωσσικά μοντέλα σε τρεις ομάδες διαφορετικών ικανοτήτων.

Η πρώτη αποτελείται από τα μικρότερα μοντέλα, τα Gemma-2-2B και Phi-3-mini. Με μεγέθη κάτω από 4 δισεκατομμύρια παραμέτρους, αυτά τα μοντέλα έχουν παρόμοια απόδοση στο σύνολο δεδομένων IMDb, η οποία είναι σαφώς καλύτερη από την τυχαία, αλλά εξακολουθεί να είναι υποδεέστερη. Ωστόσο, μια αξιοσημείωτη διαφορά παρατηρείται στο σύνολο δεδομένων AG-News, στο οποίο το Phi-3-mini προσεγγίζει τις επιδόσεις των μοντέλων που είναι τάξεις μεγέθους μεγαλύτερα, με εντυπωσιακές τιμές κοντά στο 90% σε όλες τις μετρικές

| | | Adversarial Attacks | | | | | | | | |
|-----------------|------------------|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Attacked Models | Detection Method | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| | | IMDB | | | | | | | | |
| BERT | Gemma-2-2B | 0.60 | 0.66 | 0.69 | 0.63 | 0.73 | 0.74 | 0.46 | 0.55 | 0.63 |
| | Phi-3-mini | 0.65 | 0.69 | 0.71 | 0.67 | 0.75 | 0.75 | 0.43 | 0.50 | 0.59 |
| | Llama-3.1-8B | 0.77 | 0.73 | 0.71 | 0.78 | 0.77 | 0.74 | 0.60 | 0.66 | 0.70 |
| | Gemma-2-9B | 0.77 | 0.76 | 0.76 | 0.74 | 0.76 | 0.74 | 0.60 | 0.63 | 0.66 |
| | Gemma-2-27B | 0.82 | 0.77 | 0.75 | 0.82 | 0.79 | 0.75 | 0.72 | 0.72 | 0.72 |
| | Claude-3-Haiku | 0.73 | 0.72 | 0.72 | 0.76 | 0.79 | 0.77 | 0.58 | 0.64 | 0.69 |
| | GPT-4o-mini | 0.85 | 0.79 | 0.76 | 0.78 | 0.79 | 0.76 | 0.69 | 0.70 | 0.72 |
| RoBERTa | Gemma-2-2B | 0.61 | 0.66 | 0.69 | 0.75 | 0.77 | 0.78 | 0.51 | 0.59 | 0.66 |
| | Phi-3-mini | 0.67 | 0.70 | 0.71 | 0.73 | 0.76 | 0.78 | 0.48 | 0.56 | 0.62 |
| | Llama-3.1-8B | 0.77 | 0.75 | 0.74 | 0.87 | 0.82 | 0.81 | 0.63 | 0.66 | 0.68 |
| | Gemma-2-9B | 0.82 | 0.79 | 0.78 | 0.84 | 0.82 | 0.82 | 0.67 | 0.71 | 0.73 |
| | Gemma-2-27B | 0.86 | 0.78 | 0.75 | 0.87 | 0.80 | 0.79 | 0.69 | 0.68 | 0.69 |
| | Claude-3-Haiku | 0.82 | 0.80 | 0.80 | 0.84 | 0.81 | 0.81 | 0.61 | 0.67 | 0.71 |
| | GPT-4o-mini | 0.87 | 0.80 | 0.78 | 0.90 | 0.83 | 0.82 | 0.74 | 0.72 | 0.72 |
| | | AG-News | | | | | | | | |
| BERT | Gemma-2-2B | 0.79 | 0.86 | 0.88 | 0.70 | 0.81 | 0.83 | 0.25 | 0.38 | 0.57 |
| | Phi-3-mini | 0.91 | 0.89 | 0.89 | 0.90 | 0.88 | 0.87 | 0.49 | 0.61 | 0.67 |
| | Llama-3.1-8B | 0.96 | 0.88 | 0.88 | 0.93 | 0.88 | 0.87 | 0.71 | 0.72 | 0.72 |
| | Gemma-2-9B | 0.94 | 0.91 | 0.92 | 0.89 | 0.89 | 0.88 | 0.47 | 0.59 | 0.66 |
| | Gemma-2-27B | 0.97 | 0.92 | 0.92 | 0.95 | 0.91 | 0.90 | 0.62 | 0.72 | 0.75 |
| | Claude-3-Haiku | 0.91 | 0.94 | 0.95 | 0.89 | 0.93 | 0.93 | 0.45 | 0.61 | 0.70 |
| | GPT-4o-mini | 0.98 | 0.90 | 0.90 | 0.96 | 0.90 | 0.89 | 0.68 | 0.74 | 0.75 |
| RoBERTa | Gemma-2-2B | 0.83 | 0.88 | 0.89 | 0.74 | 0.83 | 0.86 | 0.29 | 0.43 | 0.59 |
| | Phi-3-mini | 0.91 | 0.88 | 0.88 | 0.91 | 0.88 | 0.88 | 0.55 | 0.65 | 0.70 |
| | Llama-3.1-8B | 0.95 | 0.87 | 0.86 | 0.92 | 0.82 | 0.82 | 0.69 | 0.71 | 0.70 |
| | Gemma-2-9B | 0.93 | 0.94 | 0.94 | 0.90 | 0.90 | 0.91 | 0.62 | 0.73 | 0.76 |
| | Gemma-2-27B | 0.95 | 0.92 | 0.92 | 0.95 | 0.91 | 0.92 | 0.66 | 0.75 | 0.77 |
| | Claude-3-Haiku | 0.92 | 0.94 | 0.94 | 0.84 | 0.91 | 0.92 | 0.47 | 0.63 | 0.71 |
| | GPT-4o-mini | 0.93 | 0.89 | 0.88 | 0.96 | 0.90 | 0.90 | 0.71 | 0.75 | 0.75 |

Table 1.6: Αποτελέσματα ανταγωνιστικής ανίχνευσης για Προτροπή LLMs. Οι έντονοι αριθμοί υποδεικνύουν το καλύτερο αποτέλεσμα ανά μετρική.

για τις ανταγωνιστικές επιθέσεις TextFooler και PWWS. Το Gemma-2-2B υπολείπεται λίγο, επιτυγχάνοντας τιμές 75-85% στα ίδια υποσύνολα δεδομένων.

Δεύτερη σε ικανότητα θα ήταν η ομάδα των Llama-3.1-8B και Gemma-2-9B, καθώς στις περισσότερες μετρικές και πειράματα παρουσιάζουν τιμές που είναι αξιοσημείωτα κοντά η μία στην άλλη. Οι επιδόσεις αυτών των μοντέλων είναι ελπιδοφόρες, καθώς επιτυγχάνουν τιμές σταθερά πάνω από 70% και στις τρεις μετρικές και φτάνουν ακόμη και σε ανάκληση 84% και 87%, έναντι των TextFooler και PWWS, σε κείμενα IMDb. Όσον αφορά τα κείμενα AG-News, παρατηρούμε και πάλι μεγάλη άνοδο των επιδόσεων, με τα δύο LLM να έχουν βαθμολογίες Recall που ξεπερνούν το 90% στις προαναφερθείσες ανταγωνιστικές επιθέσεις και στα δύο μοντέλα-θύματα. Όσον αφορά τα κείμενα που δημιουργήθηκαν από την BAE, παρατηρούμε μια ιδιαίτερη πτώση της απόδοσης, με τις μετρήσεις να δείχνουν μειώσεις της τάξης του 10-20% στο IMDb και ακόμη και περίπου 30% στο AG-News. Το Llama-3.1-8B και το Gemma-2-9B θεωρούνται μεσαίας κλίμακας στον χώρο ανοικτού κώδικα - λαμβάνοντας αυτό υπόψη, τα αποτελέσματά τους είναι ελπιδοφόρα και δείχνουν ότι η ανίχνευση ανταγωνιστικών επιθέσεων με βάση το LLM μπορεί να πραγματοποιηθεί με επιτυχία, ακόμη και με πιο ελαφριά μοντέλα που θα επέτρεπαν μεγαλύτερη ευελιξία και ευρύτερο πειραματισμό.

Τέλος, μπορούμε να ομαδοποιήσουμε τα Gemma-2-27B, Claude-3-Haiku και GPT-4o-mini, τα μεγαλύτερα και πιο ισχυρά μοντέλα. Παρά το γεγονός ότι δεν γνωρίζουμε τις ακριβείς προδιαγραφές τους, είναι ασφαλές να υποθέσουμε ότι το Claude-3-Haiku και το GPT-4o-mini έχουν σημαντικά περισσότερες παραμέτρους από το Gemma-2-27B. Με βάση αυτή την πληροφορία, είναι ενδιαφέρον να δούμε ότι το Gemma-2-27B όχι μόνο είναι ανταγωνιστικό και με τα δύο μοντέλα, αλλά φαίνεται να ξεπερνά το Claude-3-Haiku σε ικανότητα ανταγωνιστικής ανίχνευσης. Σε κάθε περίπτωση, και τα τρία LLM επιτυγχάνουν εντυπωσιακά αποτελέσματα σε όλους τους τομείς, επιτυγχάνοντας τιμές ανάκλησης 80-90% στα δεδομένα IMDb και σημειώνοντας σταθερά πάνω από 90% σε όλες τις μετρικές, στα δεδομένα AG-News, έναντι των TextFooler και PWWS. Ενάντια στην αντίπαλη επίθεση BAE, βλέπουμε ότι παρόλο που οι μετρικές εξακολουθούν να παρουσιάζουν μέτρια ανίχνευση της αντίπαλης επίθεσης, είναι και πάλι ιδιαίτερα μειωμένες. Πρόκειται για ένα φαινόμενο που θα συζητηθεί λεπτομερέστερα παρακάτω.

Θα πρέπει να έχουμε κατά νου πως για όλα τα LLM, εκτός από το Phi-3-mini, χρησιμοποιήσαμε την ίδια προτροπή που σχεδιάστηκε για το Gemma-2-9B. Από τα πειραματικά αποτελέσματα, μπορούμε να παρατηρήσουμε ότι με τη χρήση αυτής της προτροπής, η απόδοση του μοντέλου ακολουθεί τους νόμους κλιμάκωσης εντός της οικογένειας μοντέλων Gemma 2. Ταυτόχρονα, αυτό ισχύει και για τα υπόλοιπα LLMs, τα οποία περιλαμβάνουν ένα ευρύ φάσμα αρχιτεκτονικών σχεδίων. Σε αυτή τη βάση, μπορούμε να συμπεράνουμε με επαρκή εμπιστοσύνη ότι η προτροπή μας, και επομένως το πλαίσιο μηχανικής προτροπής μας, είναι εύρωστο και γενικεύεται καλά. Μια άλλη παρατήρηση που υποστηρίζει αυτόν τον ισχυρισμό είναι ότι ενώ η χρησιμοποιούμενη προτροπή σχεδιάστηκε με βάση την επίθεση TextFooler κατά του μοντέλου ταξινόμησης BERT και του συνόλου δεδομένων IMDb, οδηγεί σε παρόμοια αποτελέσματα ανίχνευσης στην επίθεση PWWS και στο μοντέλο RoBERTa, και σε πολύ βελτιωμένα στο σύνολο δεδομένων AG-News.

Είναι απαραίτητο να αναλυθεί η υποβάθμιση των επιδόσεων που παρατηρείται έναντι των ανταγωνιστικών κειμένων που δημιουργούνται από την BAE, σε όλους τους τομείς. Η κύρια αιτία στην οποία αποδίδουμε αυτό το φαινόμενο είναι ο τρόπος με τον οποίο η BAE παράγει ανταγωνιστικές εναλλακτικές για καλυμμένα tokens. Για να το κάνει αυτό, χρησιμοποιεί το μοντέλο BERT, το οποίο διαθέτει την ίδια αρχιτεκτονική μετασχηματιστή πάνω στην οποία είναι χτισμένοι οι χρησιμοποιούμενοι ανιχνευτές-LLM. Αυτό υποδηλώνει υψηλή πιθανότητα ο χώρος παραμέτρων του επιτιθέμενου μοντέλου και του μοντέλου ανιχνευτή να είναι γειτονικοί μεταξύ τους. Ως αποτέλεσμα, οι εχθρικές διαταραχές που παράγονται από το BAE ενδέχεται να μην αλλάζουν δραστικά την εσωτερική αναπαράσταση του αρχικού κειμένου εντός των LLMs, εμποδίζοντας έτσι την ικανότητά τους να τις ανιχνεύουν. Συν τοις άλλοις, η BAE καταφέρνει περιστασιακά να ακολουθεί ισχυρότερους περιορισμούς διαταραχών, αυξάνοντας έτσι την μη αντιληπτότητα των επιθέσεών της, κάτι που, όπως θα δούμε, ισχύει και για άλλες μεθόδους ανίχνευσης της βιβλιογραφίας.

Ένα άλλο σημαντικό σημείο που πρέπει να συζητηθεί είναι ότι ενώ τα αποτελέσματα στο IMDb κρίνονται ικανοποιητικά, ειδικά για τα πιο ισχυρά μοντέλα, τα αποτελέσματα που επιτυγχάνονται στο AG-News είναι σημαντικά πιο εντυπωσιακά, ξεπερνώντας τα πρώτα έως και 26% στην περίπτωση του Phi-3mini και φτάνοντας σε βαθμολογίες ανάκλησης έως και 98% στην περίπτωση του GPT-4o mini. Υποθέτουμε ότι ο κύριος παράγοντας που συμβάλλει στην προκύπτουσα διαφορά τους είναι η μεγάλη διαφορά όσον αφορά το μήκος των κειμένων τους - τα δείγματα AG-News αποτελούνται από λιγότερες από 60 λέξεις, ενώ τα δείγματα IMDb αποτελούνται από έως και 400 λέξεις.

Ποιοτικά Αποτελέσματα

Input Text: "Woman's death probed - Public health bodies[officials] are investigating why a 38-year-old woman died two weeks after undergoing gastric bypass surgery at Saint Anne's Hospital in Fall River. The body[hospital] has stopped offering the surgery during the state probe and an internal review."

Gemma-2-2B Answer:

Let's think step by step. 1. **Grammatical Errors:** The text is grammatically correct. 2. **Misspellings and Unnatural Word Use:** No misspellings or unnatural word usage are apparent. 3. **Semantic Inconsistencies:** The text flows logically and makes sense. 4. **Language Irregularities:** The text uses standard English grammar and vocabulary. 5. **Overall Incoherence:** The text is coherent and presents a clear narrative. **Conclusion:** Based on the analysis, there are no clear signs of adversarial modification in the provided text. **Final Classification:** (A) original text

Gemma-2-9B Answer:

Let's think step by step. 1. **Grammatical Errors:** The text appears grammatically correct. 2. **Misspellings and Unnatural Word Use:** There are no obvious misspellings or unnatural word choices. 3. **Semantic Inconsistencies:** The text flows logically. There are no contradictions or unexpected connections between sentences. 4. **Language Irregularities:** The language is standard and appropriate for a news report. 5. **Overall Incoherence:** The text is coherent and easy to understand. **Summary:** The text does not exhibit any significant irregularities that would suggest adversarial modification. **Final Classification:** A original text

Gemma-2-27B Answer:

Let's analyze the text for signs of adversarial perturbation: **Grammatical Errors:** The grammar appears to be correct. **Misspellings:** There are no obvious misspellings. **Semantic Inconsistencies:** The sentence "The body has stopped offering the surgery..." is unusual. While grammatically correct, it's odd to say "the body" in this context. It likely refers to the hospital, but the phrasing is awkward and could be a subtle alteration. **Language Irregularities:** The phrasing "the body has stopped offering" is unusual and stands out as potentially unnatural. **Summary:** The text is grammatically sound, but the phrase "the body has stopped offering" is semantically unusual and potentially a deliberate alteration to introduce subtle meaning shifts or confusion for a machine learning model. **Classification:** (B) adversarial text

Figure 1.3.1: Σύγκριση της ανάλυσης που κάνουν τα LLM της οικογένειας Gemma 2 σε διαφορετικές κλίμακες μοντέλων, στο ίδιο ανταγωνιστικό κείμενο. Οι ανταγωνιστικές επεξεργασίες που έγιναν στο κείμενο εισόδου συμβολίζονται εντός παρενθέσεων και με κόκκινο χρώμα.

Στο Σχήμα 1.3.1, συγκρίνουμε την ανάλυση που κάνουν τα LLM της οικογένειας Gemma 2 σε διαφορετικές κλίμακες μοντέλων, στο ίδιο κείμενο εισόδου. Το συγκεκριμένο κείμενο έχει υποστεί ανταγωνιστική επεξεργασία με έναν διακριτικό τρόπο, καθώς έχει αλλάξει μόνο δύο από τις λέξεις του με κατάλληλες αντικαταστάσεις, και παράγει ένα κείμενο που διατηρεί τη σημασιολογική συνοχή, δεν εισάγει γλωσσικά λάθη και θα μπορούσε κάλλιστα να είναι ανεπαίσθητο από τον άνθρωπο. Όπως μπορούμε να παρατηρήσουμε, στα μεγέθη των μοντέλων των 2 δισεκατομμυρίων και 9 δισεκατομμυρίων παραμέτρων, τα LLM δεν είναι σε θέση να ανιχνεύσουν ασυνήθιστα χαρακτηριστικά μέσα στο δείγμα και έτσι το κατατάσσουν λανθασμένα ως πρωτότυπο. Αντίθετα, το μοντέλο 27B είναι σε θέση να αντιληφθεί τα λεπτά σημεία του κειμένου και εντοπίζει μια παρατυπία σε μία από τις λέξεις που αντικαταστάθηκαν, ταξινομώντας έτσι σωστά την είσοδο ως ανταγωνιστική. Παρατηρώντας

τη συλλογιστική του, βλέπουμε ότι καταλήγει σε αυτό το συμπέρασμα όχι επειδή έκανε έναν απλό γραμματικό έλεγχο, αλλά επειδή φαίνεται να κατανοεί ότι κάτι δεν πάει καλά μέσα στο ευρύτερο πλαίσιο του κειμένου. Για να προσθέσουμε σε αυτό, το LLM αναφέρει την πραγματική αρχική λέξη ως τη λέξη στην οποία πιθανότατα αναφέρεται το κείμενο, γεγονός που ενισχύει τη μέθοδό μας με επεξηγηματικότητα και επικυρώνει την υπόθεση ότι τα πιο ισχυρά LLM μπορούν πράγματι να έχουν μια αρκετά βαθιά κατανόηση κειμένου που επιτρέπει την αποτελεσματική ανίχνευση ανταγωνιστικών δειγμάτων.

Επίδραση του Μήκους Κειμένου

Για να διερευνήσουμε την επίδραση του μήκους του κειμένου εισόδου στην απόδοση του μοντέλου, πραγματοποιήσαμε μια συστηματική ανάλυση των πειραματικών μας αποτελεσμάτων στο σύνολο δεδομένων IMDb για διαφορετικά μήκη κειμένου, μετρούμενα με βάση τον αριθμό των λέξεων. Οι εικόνες 1.3.2, 1.3.3, and 1.3.4 απεικονίζουν τη σχέση μεταξύ του μήκους του κειμένου εισόδου και της ανάκλησης της ανταγωνιστικής ανίχνευσης, έναντι όλων των επιθέσεων, στο μοντέλο RoBERTa.

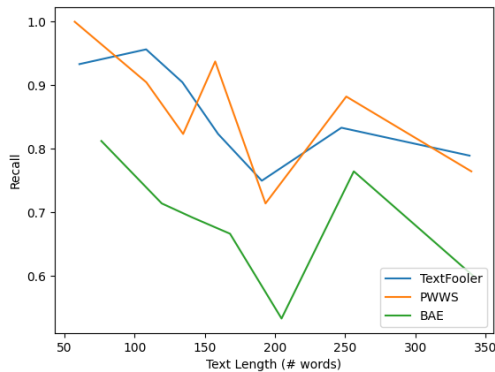


Figure 1.3.2: Ανάκληση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του Gemma-2-27B, για επιθέσεις κατά του RoBERTa.

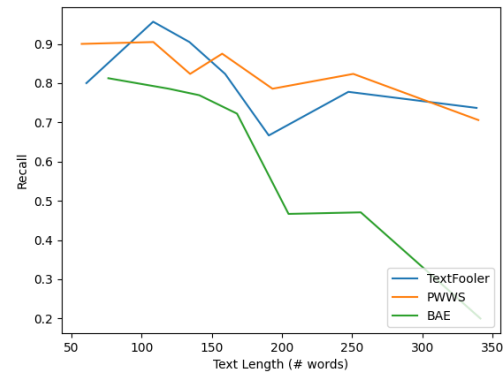


Figure 1.3.3: Ανάκληση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του Claude-3-Haiku, για επιθέσεις κατά του RoBERTa.

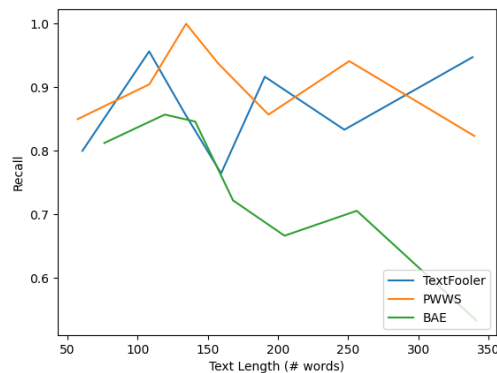


Figure 1.3.4: Ανάκληση της Ανταγωνιστικής Ανίχνευσης σε όλα τα μήκη κειμένου, για την προτροπή του GPT-4o-mini, για επιθέσεις κατά του RoBERTa.

Βλέπουμε ότι υπάρχει μεγάλη μεταβλητότητα και παρά το γεγονός ότι τα γραφήματα δεν παρουσιάζουν αυστηρά φθίνουσα σχέση μεταξύ της Ανάκλησης και του μήκους του κειμένου, η γενική τάση είναι πτωτική. Αυτό γίνεται πιο εμφανές όταν παρατηρούμε τα αποτελέσματα για τα δείγματα που δημιουργούνται από την BAE, τα οποία για υψηλότερο αριθμό λέξεων καταλήγουν να εκφυλίζονται σε μετρικές στο επίπεδο ή και χειρότερα από αυτές

ενός τυχαίου ταξινομητή.

Μια πιθανή ερμηνεία αυτών των ευρημάτων είναι το απλό γεγονός ότι σε μεγαλύτερα κείμενα, μια αποτελεσματική επίθεση μπορεί να αλλάξει μόνο ένα μικρό μέρος του κειμένου. Αυτό σημαίνει ότι οι ανταγωνιστικές ενδείξεις θα ήταν ενσωματωμένες σε μεγάλες ποσότητες αδιατάραχτου κειμένου, γεγονός που καθιστά δυσκολότερο τον εντοπισμό τους από τις LLM. Αν και αυτό είναι αλήθεια, δεν ισχύει πάντα, καθώς οι αντίπαλες επιθέσεις συχνά εφαρμόζουν πολλαπλές επεξεργασίες. Για το λόγο αυτό, πρέπει να εξετάσουμε επίσης τους περιορισμούς των Large Language Models. Για να μπορέσει ένα μοντέλο να ταξινομήσει επιτυχώς ένα δεδομένο κείμενο, πρέπει να αναλύσει τις γλωσσικές δομές και τις σχέσεις του με το πλαίσιο. Όταν ένα κείμενο είναι μεγαλύτερο, αυτά τα χαρακτηριστικά γίνονται πιο πολύπλοκα και ο όγκος των σχέσεων token-to-token που πρέπει να επεξεργαστεί ο μηχανισμός αυτοπροσοχής του μοντέλου αυξάνεται σημαντικά. Κατά συνέπεια, για μεγαλύτερα αντίπαλα δείγματα, ακόμη και εντός του παραθύρου πλαισίου, ένα LLM ενδέχεται να μην είναι σε θέση να διατηρήσει μια αρκετά συνεκτική αναπαράσταση του κειμένου και ένα επαρκές επίπεδο ευαισθησίας για να εντοπίσει τις διάσπαρτες αντίπαλες μεταβολές.

Ανθρώπινη Αξιολόγηση

Πραγματοποιήσαμε μια μελέτη ανθρώπινης αξιολόγησης προκειμένου να έχουμε μια πληρέστερη αξιολόγηση των επιδόσεων των μεθόδων μας, καθώς και να μελετήσουμε οποιαδήποτε πιθανή συσχέτιση μεταξύ του τρόπου με τον οποίο οι LLM και οι άνθρωποι προσεγγίζουν τη λήψη απόφασης για το αν ένα δεδομένο κείμενο είναι πρωτότυπο ή ανταγωνιστικό.

Οι συμμετέχοντες έλαβαν μια βασική εισαγωγή στις ανταγωνιστικές επιθέσεις στο NLP, σε συνδυασμό με μερικά παραδείγματα κειμένων στα οποία έχουν εφαρμοστεί ανταγωνιστικές διαταραχές. Στη συνέχεια, τους δόθηκαν συγκεκριμένες οδηγίες, οι οποίες τους καθοδηγούσαν να διαβάσουν προσεκτικά κάθε δοσμένο κείμενο, να το ταξινομήσουν ως πρωτότυπο ή ανταγωνιστικό, και προαιρετικά να περιγράψουν τυχόν λόγους που τους οδήγησαν στη συγκεκριμένη ταξινόμηση.

| Detection Methods | Recall | F1 | Accuracy |
|-------------------------|-------------|-------------|-------------|
| IMDB | | | |
| Gemma-2-2B | 0.58 | 0.65 | 0.69 |
| Phi-3-mini | 0.66 | 0.71 | 0.73 |
| Llama-3.1-8B | 0.78 | 0.77 | 0.77 |
| Gemma-2-9B | 0.76 | 0.75 | 0.75 |
| Gemma-2-27B | 0.84 | 0.78 | 0.76 |
| Claude 3 Haiku | 0.80 | 0.78 | 0.78 |
| GPT-4o mini | 0.82 | 0.78 | 0.77 |
| Human Evaluation | 0.75 ± 0.02 | 0.79 ± 0.03 | 0.80 ± 0.04 |
| AG-News | | | |
| Gemma-2-2B | 0.86 | 0.90 | 0.90 |
| Phi-3-mini | 0.94 | 0.86 | 0.85 |
| Llama-3.1-8B | 1.00 | 0.92 | 0.91 |
| Gemma-2-9B | 1.00 | 0.97 | 0.97 |
| Gemma-2-27B | 0.98 | 0.94 | 0.94 |
| Claude 3 Haiku | 0.98 | 0.97 | 0.97 |
| GPT-4o mini | 0.96 | 0.93 | 0.93 |
| Human Evaluation | 0.85 ± 0.04 | 0.91 ± 0.03 | 0.91 ± 0.03 |

Table 1.7: Αποτελέσματα Ανταγωνιστικής Ανίχνευσης για τα σύνολα δεδομένων της Ανθρώπινης Αξιολόγησης.

Στον πίνακα 1.7 παρουσιάζονται τα αποτελέσματα της ανίχνευσης αντιθέσεων των ανθρώπινων αξιολογητών σε σύγκριση με εκείνα των LLM. Παρατηρούμε ότι οι αποκλίσεις μεταξύ των μετρικών των διαφόρων αξιολογητών είναι μικρές, στο πλαίσιο της ανάλυσής μας, γεγονός που υποδηλώνει μια ικανοποιητική συνοχή μεταξύ των ανθρώπινων προβλέψεων. Συνολικά, η απόδοση της ανθρώπινης ανίχνευσης στο σύνολο δεδομένων IMDb είναι

στα ίδια επίπεδα με εκείνη των μοντέλων ανοικτού κώδικα μεσαίας κλίμακας, ενώ στο σύνολο δεδομένων AG-News είναι ελαφρώς χειρότερη από εκείνη των μοντέλων με τις χειρότερες επιδόσεις. Τα αποτελέσματα αυτά επαληθεύουν τον υψηλό βαθμό δυσκολίας που χαρακτηρίζει την ανίχνευση αντιπαθητικών επιθέσεων κειμένου. Επιπλέον, επαληθεύουν τη χρησιμότητα και την ανταγωνιστικότητα της μεθόδου προτροπής LLM που διαθέτουμε.

Αξίζει να εξετάσουμε ορισμένες κατηγορίες κειμένων για τις οποίες φαίνεται να υπάρχει αξιοσημείωτη διαφορά στον τρόπο με τον οποίο αναλύονται από ανθρώπους και από Μεγάλα Γλωσσικά Μοντέλα. Μια τέτοια κατηγορία είναι τα πρωτότυπα κείμενα που περιέχουν λάθη τα οποία αποδίδονται σε ανθρώπινη απροσεξία ή έλλειψη γνώσης της αγγλικής γλώσσας. Οι άνθρωποι αξιολογητές φάνηκε να είναι λίγο πιο ικανοί να διακρίνουν αυτού του είδους τα κείμενα ως πρωτότυπα, παρόλο που εξακολουθούν να περιέχουν λάθη. Αντιθέτως, μια περίπτωση στην οποία τα LLM φαίνεται να υπερτερούν έναντι των ανθρώπων, είναι τα κείμενα που χαρακτηρίζονται από προηγμένο λεξιλόγιο και/ή περίπλοκες δομές προτάσεων. Όταν τους παρουσιάστηκαν τέτοια κείμενα, οι άνθρωποι αξιολογητές μας μερικές φορές δυσκολεύονταν να τα αναλύσουν.

Έλεγχος Ορθότητας για Μόλυνση Δεδομένων

Στο πλαίσιο της αξιολόγησης μεγάλων γλωσσικών μοντέλων, ένα από τα πιο κυρίαρχα ζητήματα είναι αυτό της μόλυνσης των δεδομένων ή αλλιώς μόλυνσης του συνόλου δοκιμών. Δεδομένου ότι τόσο τα σύνολα δεδομένων δοκιμής μας όσο και αρκετά από τα σύνολα δεδομένων εκπαίδευσης των LLMs που χρησιμοποιήσαμε στα πειράματά μας προέρχονται από το διαδίκτυο, είναι πιθανό τα μοντέλα να εκπαιδεύτηκαν με δεδομένα που περιέχονται στις αξιολογήσεις μας. Για να διασφαλίσουμε την ακεραιότητα των πειραματικών μας αποτελεσμάτων και να μετριάσουμε το πιθανό ζήτημα μιας τέτοιας μόλυνσης, εφαρμόσαμε έναν έλεγχο ορθότητας. Για να το πετύχουμε αυτό, επιμεληθήκαμε χειροκίνητα ένα νέο σύνολο δεδομένων δοκιμών πραγματοποιώντας επεξεργασίες σε υπάρχοντα δεδομένα, έτσι ώστε τα LLM να μην έχουν εκτεθεί προηγουμένως σε αυτά.

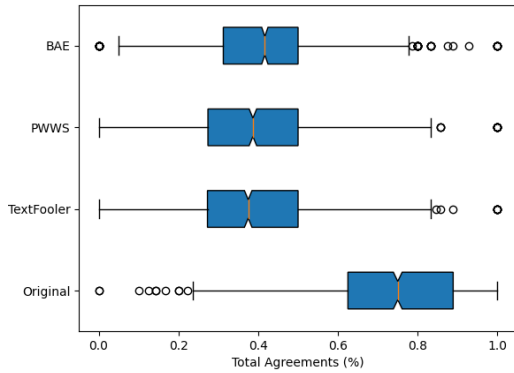
| Detection Methods | Recall | F1 | Accuracy |
|-----------------------|--------------|--------------|--------------|
| IMDB | | | |
| Gemma-2-2B | 0.60 (+0.02) | 0.66 (+0.01) | 0.69 (+0.00) |
| Phi-3-mini | 0.62 (-0.04) | 0.67 (-0.04) | 0.69 (-0.04) |
| Llama-3.1-8B | 0.80 (+0.02) | 0.78 (+0.01) | 0.77 (+0.00) |
| Gemma-2-9B | 0.76 (+0.00) | 0.75 (+0.00) | 0.74 (-0.01) |
| Gemma-2-27B | 0.82 (-0.02) | 0.77 (-0.01) | 0.75 (-0.01) |
| Claude 3 Haiku | 0.78 (-0.02) | 0.76 (-0.02) | 0.76 (-0.01) |
| GPT-4o mini | 0.84 (+0.02) | 0.79 (+0.01) | 0.77 (+0.00) |
| AG-News | | | |
| Gemma-2-2B | 0.86 (+0.00) | 0.90 (+0.00) | 0.90 (+0.00) |
| Phi-3-mini | 0.96 (+0.02) | 0.88 (+0.02) | 0.86 (+0.01) |
| Llama-3.1-8B | 1.00 (+0.00) | 0.89 (-0.03) | 0.88 (-0.03) |
| Gemma-2-9B | 0.97 (-0.03) | 0.93 (-0.04) | 0.93 (-0.04) |
| Gemma-2-27B | 1.00 (+0.02) | 0.92 (-0.02) | 0.91 (-0.03) |
| Claude 3 Haiku | 1.00 (+0.02) | 0.98 (+0.01) | 0.98 (+0.01) |
| GPT-4o mini | 0.96 (+0.00) | 0.92 (-0.01) | 0.92 (-0.01) |

Table 1.8: Αποτελέσματα ανταγωνιστικής ανίχνευσης για τον Έλεγχο Ορθότητας. Η διαφορά σε σύγκριση με τη μετρική που επιτεύχθηκε στο αρχικό σύνολο δεδομένων σημειώνεται εντός παρένθεσης.

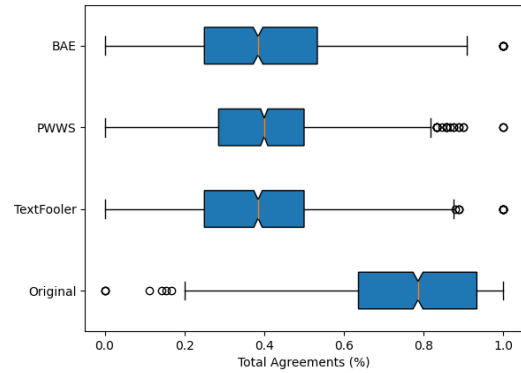
Από τον πίνακα 1.8, βλέπουμε ότι τα αποτελέσματα του ελέγχου ορθότητας είναι πολύ κοντά σε εκείνα που προέκυψαν από το αρχικό σύνολο δεδομένων. Παρατηρούμε μικρές διαφοροποιήσεις για διάφορα μοντέλα, αλλά δεν φαίνεται να έχουν συνεπή χαρακτήρα, είτε θετικό είτε αρνητικό. Επιπλέον, οι απόλυτες τιμές τους δεν θεωρούνται σημαντικές για την ανάλυσή μας, καθώς ακόμη και στις περιπτώσεις των υψηλότερων ποσοστιαίων διαφορών 3-4%, οι πραγματικές προβλέψεις των μοντέλων διαφέρουν μόνο κατά μερικά δείγματα. Συνολικά, τα αποτελέσματα του ελέγχου ορθότητας υποδηλώνουν ότι τα πειράματά μας δεν επηρεάστηκαν από τη μόλυνση των δεδομένων και αναδεικνύουν την ανθεκτικότητα της προσέγγισης ανίχνευσης που εφαρμόσαμε.

1.3.3 Πειράματα Ταξινόμησης Μεμονωμένων Προτάσεων

Συνολικές Συμφωνίες



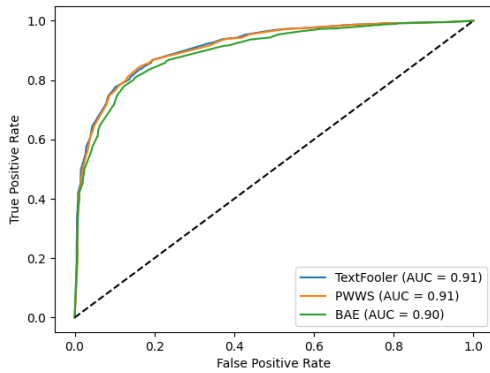
(a) Για το BERT.



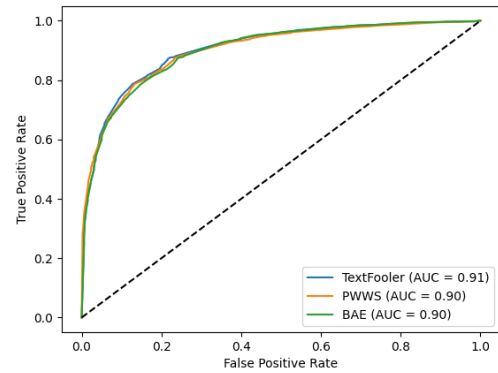
(b) Για το RoBERTa.

Figure 1.3.5: Boxplots του συνολικού ποσοστού συμφωνίας ταξινόμησης μεμονωμένων προτάσεων.

Στα σχήματα 1.3.5a και 1.3.5b, μπορούμε να δούμε τα boxplots των συνολικών συμφωνιών ταξινόμησης μεμονωμένων προτάσεων, ως ποσοστό του συνολικού μήκους του κειμένου, για τα πρωτότυπα και τα αντίπαλα δείγματα των διαφόρων μεθόδων επίθεσης. Όπως μπορεί να παρατηρηθεί, η στατιστική κατανομή των πρωτότυπων κειμένων διακρίνεται σαφώς από την κατανομή κάθε επίθεσης, με την πρώτη να έχει πολύ υψηλότερη διάμεσο, περίπου 75%, και πρώτο τεταρτημόριο λίγο πάνω από 60%, τόσο για το BERT όσο και για το RoBERTa. Επιπλέον, οι κατανομές των διαφορετικών ανταγωνιστικών επιθέσεων μοιάζουν πολύ μεταξύ τους, με σχεδόν πανομοιότυπα τρίτα τεταρτημόρια γύρω στο 50% και διαμέσους με μικρή ή καθόλου στατιστική διαφορά.



(a) Για το BERT.



(b) Για το RoBERTa.

Figure 1.3.6: Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Συμφωνιών ως ορίου για την ανταγωνιστική ανίχνευση.

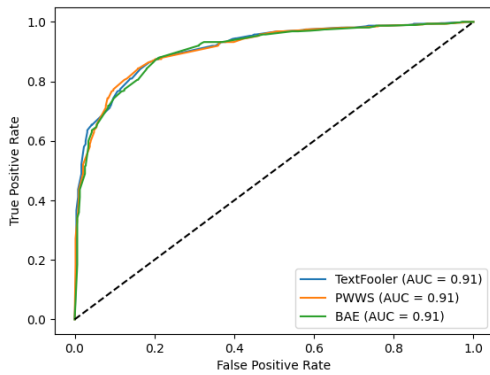
Στα σχήματα 1.3.6a και 1.3.6b, παρουσιάζουμε τις καμπύλες ROC που προκύπτουν από τον ορισμό του συνολικού ποσοστού συμφωνιών ως ορίου απόφασης για το BERT και το RoBERTa όσον αφορά την ταξινόμηση ενός κειμένου ως ανταγωνιστικού ή μη. Οι τιμές AUC που επιτυγχάνονται είναι 90% και 91% και δείχνουν ότι η μέθοδος παρουσιάζει υψηλή ικανότητα διαχωρισμού μεταξύ αρχικών και ανταγωνιστικών δειγμάτων, και για τα δύο μοντέλα. Επιπλέον, το γεγονός ότι οι καμπύλες ROC και για τις 3 μεθόδους επίθεσης είναι σχεδόν αδιαφοροποίητες αποδεικνύει ότι ένα κατώφλι ανίχνευσης, που επιλέγεται για μια συγκεκριμένη επίθεση, μέσω της μεθόδου συνολικών συμφωνιών, μπορεί να χρησιμοποιηθεί εναλλακτικά και για οποιαδήποτε άλλη επίθεση.

Προκειμένου να χρησιμοποιήσουμε τη μέθοδο των Συνολικών Συμφωνιών για την ανίχνευση ανταγωνιστικών

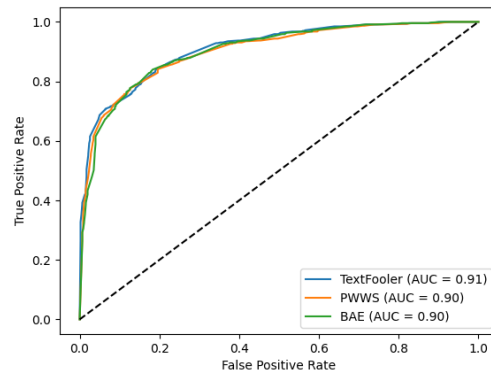
παραδειγμάτων, θέλαμε να επιλέξουμε ένα κατώφλι που να είναι όσο το δυνατόν πιο κοντά στον τέλει ταξινόμητη, το άνω αριστερό σημείο της καμπύλης ROC. Κατά την επιλογή μας, λάβαμε επίσης υπόψη το γεγονός ότι στην εργασία μας είναι πιο σημαντικό να βελτιστοποιήσουμε για αληθινά θετικά παρά για ψευδώς αρνητικά. Με βάση αυτά τα κριτήρια, επιλέξαμε το όριο συνολικών συμφωνιών 60% που αντιστοιχεί σε FPR 0,2 και στους αγκώνες των καμπυλών ROC, για το BERT και το RoBERTa.

| Models | Adversarial Attacks | | | | | | | | |
|----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | 0.89 | 0.84 | 0.83 | 0.90 | 0.85 | 0.84 | 0.88 | 0.83 | 0.82 |
| RoBERTa | 0.85 | 0.83 | 0.82 | 0.85 | 0.83 | 0.82 | 0.85 | 0.83 | 0.82 |

Table 1.9: Αποτελέσματα ανταγωνιστικής ανίχνευσης για τη μέθοδο Συνολικών IndivSC Συμφωνιών.



(a) Για το BERT.



(b) Για το RoBERTa.

Figure 1.3.7: Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Συμφωνιών ως ορίου για την ανίχνευση ανταγωνιστικότητας, στο σύνολο δοκιμών.

Στον Πίνακα 1.9, βλέπουμε τα αποτελέσματα της ανίχνευσης των ανταγωνιστικών δειγμάτων για τη μέθοδο Συνολικών Συμφωνιών έναντι κάθε επίθεσης. Η πρώτη παρατήρηση που μπορεί να γίνει είναι ότι η τιμή κάθε μετρικής είναι σχεδόν ίδια και για τις 3 επιθέσεις, ενισχύοντας το γεγονός ότι το ίδιο κατώφλι ανίχνευσης μπορεί να χρησιμοποιηθεί εναλλάξ ενάντια σε οποιαδήποτε ανταγωνιστική επίθεση. Επιπλέον, η μέθοδος έχει υψηλές επιδόσεις, επιτυγχάνοντας μετρικές συγκρίσιμες με εκείνες των καλύτερων σεναρίων LLM Prompting, με 85-90% Recall, 83-85% F1 score και 82-84% Accuracy, είτε για το BERT είτε για το RoBERTa. Αυτό επιβεβαιώνει την καταλληλότητά της για την ανίχνευση ανταγωνιστικών κειμένων, ως βοηθητική ή ακόμη και αυτόνομη προσέγγιση.

Όλα τα προαναφερθέντα συμπεράσματα ενισχύονται από τα Σχήματα 1.3.7a και 1.3.7b, στα οποία μπορούμε να παρατηρήσουμε ότι οι καμπύλες ROC, καθώς και οι AUC τιμές τους, για το σύνολο δοκιμής είναι σχεδόν ταυτόσημες με αυτές του συνόλου εκπαίδευσης, γεγονός που αναδεικνύει επίσης συνέπεια στην ανίχνευση.

Αξίζει να εξεταστεί η συσχέτιση μεταξύ του ποσοστού συνολικών συμφωνιών ταξινόμησης μεμονωμένων προτάσεων και του μήκους ενός κειμένου, μετρούμενου σε αριθμό προτάσεων, καθώς είναι άμεσα συνυφασμένα. Όπως μπορεί να παρατηρηθεί στα Σχήματα 1.3.8 και 1.3.9, η μέση τιμή τόσο για τα πρωτότυπα όσο και για τα ανταγωνιστικά κείμενα παραμένει σταθερά εντός του πρώτου και του τρίτου τεταρτημορίου του αντίστοιχου boxplot, με τις 3 επιθέσεις να έχουν παρόμοιες γραμμές γραφικής παράστασης σε διαφορετικά μήκη κειμένων. Συνολικά, παρατηρούμε ότι και για τους 4 τύπους κειμένων, η γραφική παράσταση ακολουθεί μια πτωτική τάση. Η διαχωριστική ικανότητα μεταξύ αρχικών και αντίπαλων κειμένων παραμένει επίσης σταθερά ευρεία για οποιονδήποτε αριθμό προτάσεων, ακόμη και όταν εξετάζονται οι αποκλίσεις. Αυτό, σε συνδυασμό με τη φθίνουσα φύση των γραμμών του διαγράμματος, μας οδηγεί στο συμπέρασμα ότι ένα σταθερό κατώφλι για την ανίχνευση των αντίπαλων επιθέσεων για τη μέθοδο Συνολικών Συμφωνιών θα είναι σε θέση να παράγει περισσότερα αληθινά θετικά για μεγαλύτερα κείμενα. Το προηγούμενο συμπέρασμα επαληθεύεται από τα Σχήματα 1.3.11 και

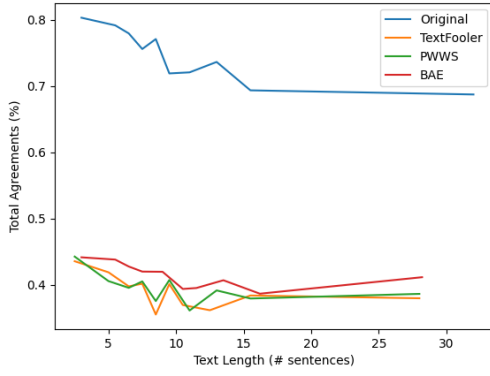


Figure 1.3.8: Για το BERT.

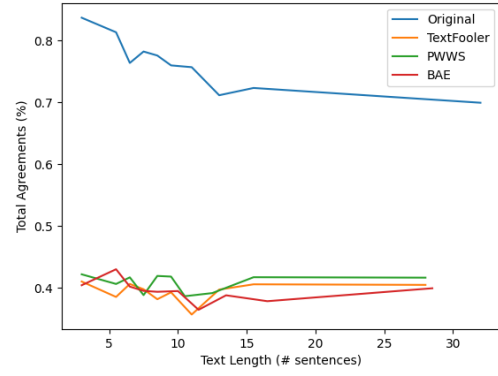


Figure 1.3.9: Για το RoBERTa.

Figure 1.3.10: Συνολικό ποσοστό συμφωνιών IndivSC σε όλα τα μήκη κειμένων.

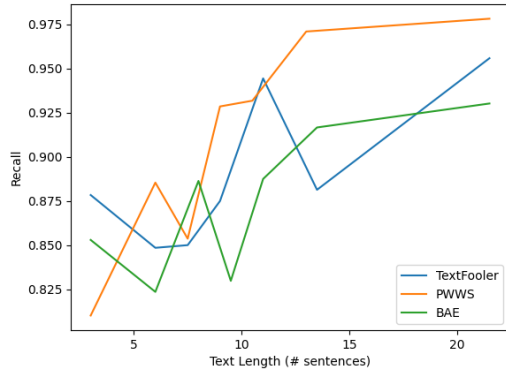


Figure 1.3.11: Για το BERT.

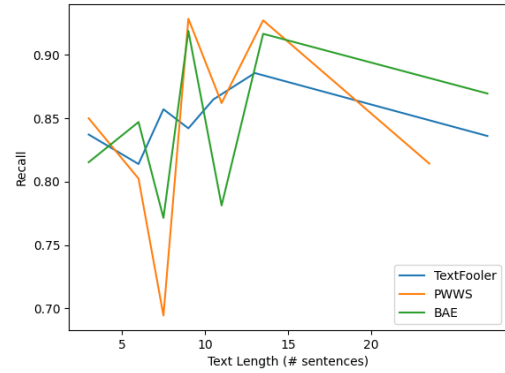


Figure 1.3.12: Για το RoBERTa.

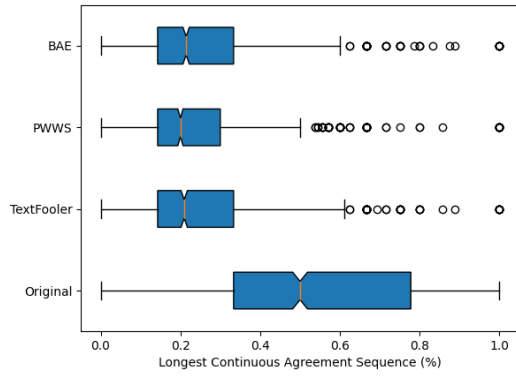
Figure 1.3.13: Ανάκληση ανταγωνιστικής ανίχνευσης για τη μέθοδο Συνολικών IndivSC Συμφωνιών σε διάφορα μήκη κειμένου.

[1.3.12](#), στα οποία παρατηρούμε ότι η ανάκληση ανίχνευσης που επιτυγχάνεται από τη μέθοδο γενικά αυξάνεται όσο μεγαλύτερο είναι ένα κείμενο, και στις τρεις τεχνικές επίθεσης.

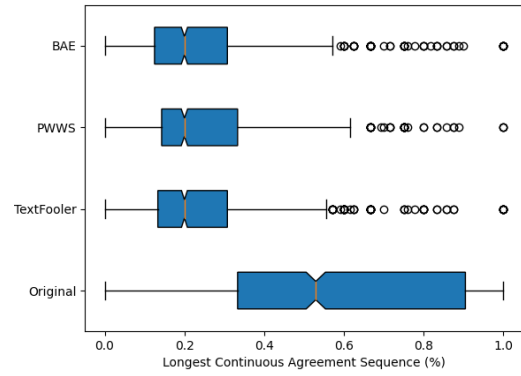
Μακρύτερη Συνεχής Ακολουθία Συμφωνίας

Παρατηρώντας τα boxplots της μεγαλύτερης συνεχούς ακολουθίας συμφωνίας IndivSC, ως ποσοστό του συνολικού μήκους του κειμένου, στα Σχήματα [1.3.14a](#) και [1.3.14b](#), βλέπουμε ότι οι κατανομές των αρχικών και των ανταγωνιστικών δειγμάτων είναι διαχωρίσιμες από το BERT και από το RoBERTa, αλλά μόλις και μετά βίας. Από τη μία πλευρά, οι διάφορες ανταγωνιστικές επιθέσεις παρουσιάζουν πολύ παρόμοια χαρακτηριστικά με στατιστικά δυσδιάκριτες διαμέσους γύρω στο 20% και τιμές Q3 κοντά στο 30%. Από την άλλη πλευρά, για τα πρωτότυπα κείμενα, η διάμεσος είναι σημαντικά υψηλότερη σε σχεδόν 50%, αλλά το IQR που παρουσιάζεται είναι ευρύ, ειδικά για το RoBERTa, τόσο προς την κατεύθυνση του Q3 όσο και του Q1. Συγκεκριμένα, το πρώτο τεταρτημόριο είναι σχεδόν ίσο με το τρίτο τεταρτημόριο των ανταγωνιστικών κειμένων, πράγμα που σημαίνει ότι περίπου το 25% των παραδειγμάτων οποιασδήποτε από τις δύο κατηγορίες δεν είναι σε θέση να διακριθούν με τη χρήση της μεγαλύτερης συνεχούς ακολουθίας συμφωνίας IndivSC.

Ορίζοντας το μήκος της μεγαλύτερης συνεχούς ακολουθίας συμφωνίας ως το κατώφλι ταξινόμησης μεταξύ αρχικών και ανταγωνιστικών δειγμάτων, λαμβάνουμε τις καμπύλες ROC στα Σχήματα [1.3.15a](#) και [1.3.15b](#). Αυτά δείχνουν ότι παρά το γεγονός ότι οι στατιστικές κατανομές της δεν είναι βέλτιστες, η Longest Continuous

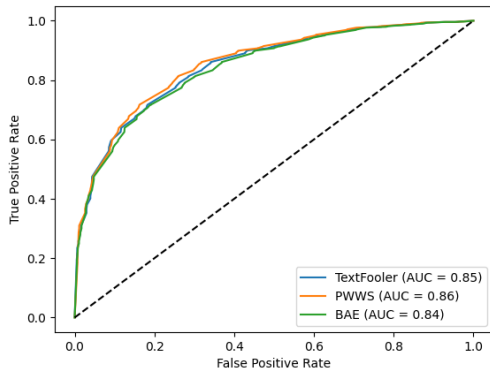


(a) Για το BERT.

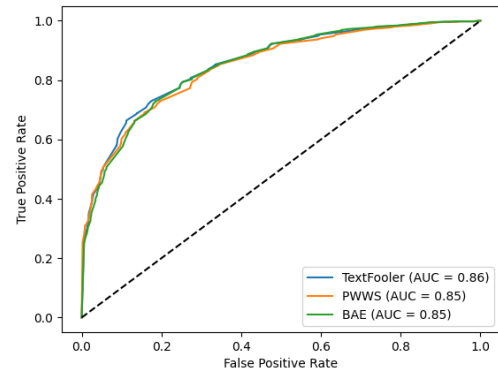


(b) Για το RoBERTa.

Figure 1.3.14: Boxplots του ποσοστού της μεγαλύτερης συνεχούς ακολουθίας συμφωνίας IndivSC.



(a) Για το BERT.



(b) Για το RoBERTa.

Figure 1.3.15: Καμπύλες ROC για τη χρήση του ποσοστού Μακρύτερη Συνεχής Ακολουθία Συμφωνίας ως ορίου για την ανταγωνιστική ανίχνευση.

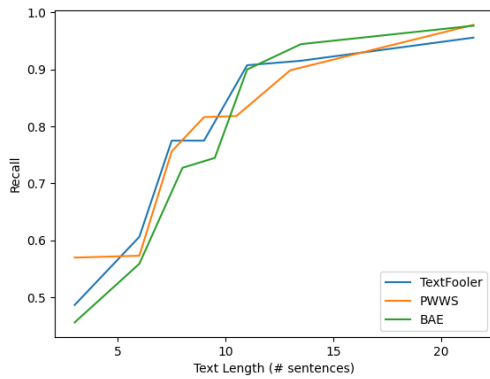
IndivSC Agreement Sequence εξακολουθεί να είναι μια βιώσιμη επιλογή για την ανταγωνιστική ανίχνευση, για το BERT και για το RoBERTa. Προκειμένου να αξιολογήσουμε την αποτελεσματικότητα της μεθόδου στην ανίχνευση, επιλέξαμε ένα κατώφλι 33%, ώστε να επιτύχουμε μια ισορροπία μεταξύ της αύξησης του TPR και της υποβάθμισης του FPR. Συγκεκριμένα, η τιμή αυτή αντιστοιχεί σε FPR 0,25 και σε σημείο στον αγκώνα των καμπυλών.

| Models | Adversarial Attacks | | | | | | | | |
|----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | 0.75 | 0.77 | 0.78 | 0.75 | 0.77 | 0.78 | 0.75 | 0.77 | 0.78 |
| RoBERTa | 0.74 | 0.76 | 0.77 | 0.70 | 0.74 | 0.75 | 0.74 | 0.76 | 0.77 |

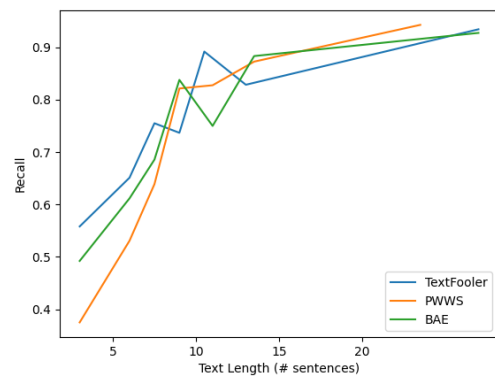
Table 1.10: Αποτελέσματα ανίχνευσης ανταγωνιστικών δειγμάτων για τη μέθοδο Μακρύτερη Συνεχής Ακολουθία IndivSC Συμφωνίας.

Τα αποτελέσματα της ανταγωνιστικής ανίχνευσης της μεθόδου Μακρύτερη Συνεχής Ακολουθία IndivSC Συμφωνίας παρουσιάζονται στον πίνακα 1.10. Από αυτόν, συμπεραίνουμε μια μέτρια υψηλή απόδοση, η οποία είναι κοντά στα επίπεδα που επιτυγχάνονται από LLMs μεσαίας κλίμακας, με μετρικές που αποτελούνται από 70-75% Ανάκληση, 74-77% F1 και 75-78% Ακρίβεια. Είναι σημαντικό να σημειωθεί ότι το BERT επιτυγχάνει τις ίδιες

ακριβώς μετρήσεις και για τα 3 αντιπαλαιακά πλαίσια, γεγονός που αναδεικνύει τη συνέπεια και τη γενίκευση του επιλεγμένου ορίου απόφασης έναντι διαφορετικών επιθέσεων.



(a) Για το BERT.

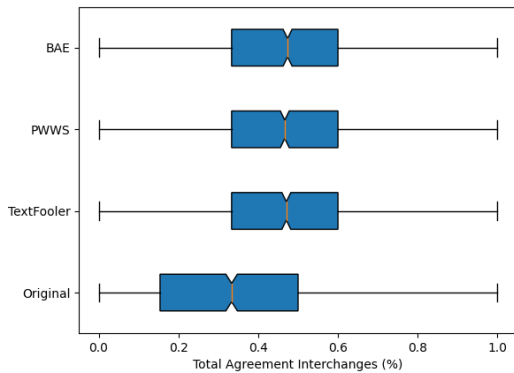


(b) Για το RoBERTa.

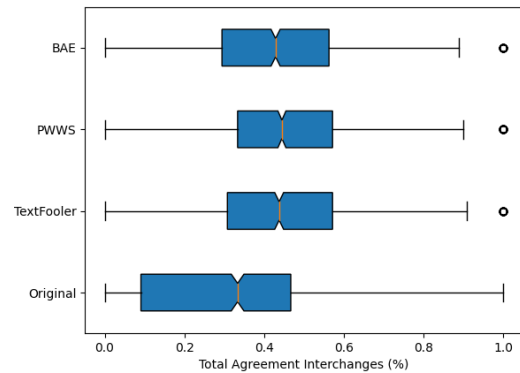
Figure 1.3.16: Ανάκληση ανταγωνιστικής ανίχνευσης για τη μέθοδο Μακρύτερη Συνεχής Ακολουθία IndivSC Συμφωνίας σε όλα τα μήκη κειμένου.

Για να εξετάσουμε την επίδραση του μήκους του κειμένου στο ποσοστό της μεγαλύτερης συνεχούς ακολουθίας IndivSC συμφωνίας, αναλύουμε τα σχήματα 1.3.16a και 1.3.16b, στα οποία βλέπουμε ότι η ανάκληση ανίχνευσης στο σύνολο δοκιμών αυξάνεται ραγδαία για μεγαλύτερα κείμενα, ξεπερνώντας γρήγορα τη μέση απόδοσή της, για το BERT και για το RoBERTa.

Συνολικές Εναλλαγές



(a) Για το BERT.

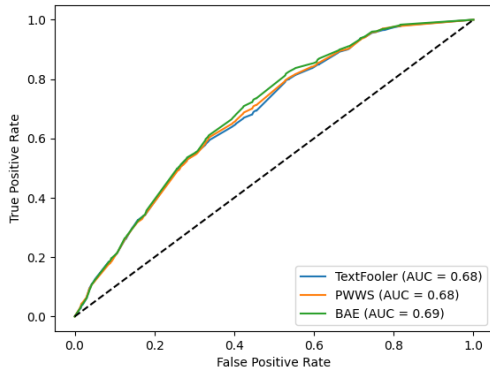


(b) Για το RoBERTa.

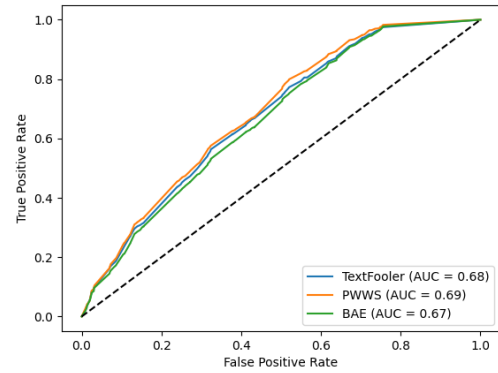
Figure 1.3.17: Boxplots του ποσοστού Συνολικών IndivSC Εναλλαγών.

Τα σχήματα 1.3.17a και 1.3.17b εμφανίζουν τα boxplots των συνολικών εναλλαγών ταξινόμησης των μεμονωμένων προτάσεων, για τα αρχικά δείγματα και για τις επιθέσεις κατά των BERT και RoBERTa. Από αυτά, μπορεί να συναχθεί αμέσως ότι υπάρχει μεγάλη επικάλυψη μεταξύ των στατιστικών κατανομών των αρχικών και των ανταγωνιστικών κειμένων. Στην πραγματικότητα, το εύρος Q2-Q3 των αρχικών δειγμάτων συμπίπτει πλήρως με το εύρος Q1-Q2 των ανταγωνιστικών, τόσο για το BERT όσο και για το RoBERTa. Το εύρημα αυτό υποδεικνύουν ότι οι συνολικές ανταλλαγές δεν μπορούν να αποτελέσουν κατάλληλη μέθοδο για την ανίχνευση αντίπαλων κειμένων.

Οι καμπύλες ROC που παρουσιάζονται στα Σχήματα 1.3.18a και 1.3.18b προκύπτουν από τον ορισμό του συνολικού ποσοστού εναλλαγών ως ορίου για τον προσδιορισμό των δειγμάτων ως πρωτότυπα ή ανταγωνιστικά.



(a) Για το BERT.



(b) Για το RoBERTa.

Figure 1.3.18: Καμπύλες ROC για τη χρήση του ποσοστού Συνολικών IndivSC Εναλλαγών ως ορίου για την ανταγωνιστική ανίχνευση.

Τα AUC που επιτεύχθηκαν δείχνουν μια απόδοση που είναι καλύτερη από την τυχαία, αλλά σίγουρα δεν είναι εφάμιλλη των άλλων εξεταζόμενων μεθόδων, επιβεβαιώνοντας τα συμπεράσματά μας. Συγκεκριμένα, οι γραφικές παραστάσεις των 3 διαφορετικών επιθέσεων έχουν παρόμοια μορφή και τιμές στο εύρος 0,67-0,69, τόσο για το BERT όσο και για το RoBERTa.

Για να χρησιμοποιήσουμε το ποσοστό Total IndivSC Interchanges ως όριο ταξινόμησης, έπρεπε να επιλέξουμε ένα όριο που να οδηγεί σε ανάκληση πάνω από 50% χωρίς να επιδεινώνεται εντελώς το ποσοστό ψευδώς θετικών αποτελεσμάτων, και έτσι η τιμή που επιλέχθηκε ήταν 38% και αντιστοιχεί σε FPR 0,4.

| Models | Adversarial Attacks | | | | | | | | |
|----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | 0.67 | 0.65 | 0.64 | 0.66 | 0.64 | 0.63 | 0.67 | 0.65 | 0.64 |
| RoBERTa | 0.60 | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 | 0.62 | 0.62 | 0.63 |

Table 1.11: Αποτελέσματα ανίχνευσης ανταγωνιστικών δειγμάτων για τη μέθοδο Συνολικών IndivSC Εναλλαγών.

Στον πίνακα 1.11, βλέπουμε τα αποτελέσματα της ανίχνευσης των ανταγωνιστικών δειγμάτων για τη μέθοδο Συνολικών IndivSC Εναλλαγών ενάντια σε κάθε επίθεση. Μπορούμε να παρατηρήσουμε συνοχή μεταξύ των μετρικών που επιτυγχάνονται για τις 3 διαφορετικές επιθέσεις και για τα δύο μοντέλα, ωστόσο οι μετρικές που επιτυγχάνονται είναι όλες στην περιοχή του 60-67% και αντιστοιχούν σε συμπεριφορά καλύτερη από την τυχαία και κοντά στο επίπεδο των μικρότερων χρησιμοποιούμενων LLMs.

1.3.4 Συνδυασμός Μεθόδων

Για να συνδυάσουμε τις δύο προτεινόμενες μεθόδους μας σε μια ενιαία προσέγγιση, χρησιμοποιήσαμε την Προτροπή LLM ως κύρια μέθοδο ανίχνευσης και τη συμπληρώσαμε με τις Συνολικές Συμφωνίες, που είναι το πλαίσιο ταξινόμησης μεμονωμένων προτάσεων με τις καλύτερες επιδόσεις. Χρησιμοποιήσαμε τη μέθοδο Συνολικών Συμφωνιών μόνο σε κείμενα που είναι πάνω από ένα ορισμένο μήκος, προκειμένου να αξιοποιήσουμε το γεγονός ότι η ανάκληση της ανίχνευσης ανταγωνιστικών δειγμάτων της φαίνεται να αυξάνεται για μεγαλύτερα κείμενα, ενώ το ακριβώς αντίθετο συμβαίνει στην περίπτωση των LLMs. Συν τοις άλλοις, θέλαμε να διασφαλίσουμε ότι η αξιοποίηση της μεθόδου Συνολικών Συμφωνιών ήταν μόνο συμπληρωματική και δεν οδηγούσε σε αύξηση των ψευδών θετικών ή των ψευδών αρνητικών αποτελεσμάτων. Για να το πετύχουμε αυτό, ορίσαμε το κατώφλι για το ποσοστό συμφωνιών ταξινόμησης μεμονωμένων προτάσεων σε 50% προκειμένου να θεωρηθεί ένα κείμενο ως ανταγωνιστικό, κάτι που αντιστοιχεί σε ποσοστό ψευδώς θετικών (FPR) 10%, και ορίσαμε το κατώφλι για να θεωρηθεί ένα κείμενο ως πρωτότυπο σε 66%, κάτι που αντιστοιχεί σε ποσοστό ψευδούς παράλειψης (FOR) 10%.

Συμπερασματικά, ο συνδυασμός μας δίνει προτεραιότητα στην πρόβλεψη που κάνει το LLM. Στη συνέχεια, στην περίπτωση που το μήκος του συγκεκριμένου κειμένου είναι πάνω από 6 προτάσεις, υπολογίζουμε αν το συνολικό ποσοστό συμφωνιών IndivSC εμπίπτει στα καθορισμένα όρια. Εάν αυτό συμβαίνει και η πρόβλεψη με βάση το LLM διαφωνεί με την πρόβλεψη με βάση τις συνολικές συμφωνίες, τότε λαμβάνουμε υπόψη την τελευταία.

Ο πίνακας 1.12 παρουσιάζει τα πειραματικά αποτελέσματα που προκύπτουν με τη χρήση αυτής της ενοποιημένης προσέγγισης.

| Attacked Models | Methods | Adversarial Attacks | | | | | | | | |
|-----------------|------------------|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | Gemma-2-2B * | 0.82 | 0.81 | 0.81 | 0.87 | 0.87 | 0.86 | 0.83 | 0.83 | 0.84 |
| | Phi-3-mini * | 0.84 | 0.82 | 0.82 | 0.91 | 0.89 | 0.88 | 0.82 | 0.81 | 0.82 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.92 | 0.89 | 0.87 | 0.87 | 0.85 | 0.85 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.85 | 0.91 | 0.88 | 0.86 | 0.85 | 0.84 | 0.84 |
| | Gemma-2-27B * | 0.91 | 0.86 | 0.86 | 0.94 | 0.88 | 0.86 | 0.88 | 0.85 | 0.85 |
| | Claude-3-Haiku * | 0.88 | 0.85 | 0.84 | 0.92 | 0.89 | 0.88 | 0.84 | 0.85 | 0.85 |
| | GPT-4o-mini * | 0.92 | 0.88 | 0.87 | 0.94 | 0.90 | 0.89 | 0.88 | 0.85 | 0.86 |
| RoBERTa | Gemma-2-2B * | 0.83 | 0.84 | 0.84 | 0.85 | 0.82 | 0.82 | 0.80 | 0.84 | 0.85 |
| | Phi-3-mini * | 0.90 | 0.87 | 0.87 | 0.84 | 0.82 | 0.82 | 0.82 | 0.83 | 0.84 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.87 | 0.82 | 0.82 | 0.85 | 0.85 | 0.86 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.86 | 0.87 | 0.82 | 0.82 | 0.85 | 0.87 | 0.87 |
| | Gemma-2-27B * | 0.92 | 0.88 | 0.87 | 0.89 | 0.82 | 0.82 | 0.84 | 0.85 | 0.86 |
| | Claude-3-Haiku * | 0.90 | 0.88 | 0.87 | 0.87 | 0.83 | 0.83 | 0.87 | 0.88 | 0.88 |
| | GPT-4o-mini * | 0.90 | 0.89 | 0.88 | 0.87 | 0.83 | 0.82 | 0.85 | 0.87 | 0.87 |

Table 1.12: Αποτελέσματα ανίχνευσης επιθέσεων για το συνδυασμό μεθόδων, στο IMDb. Οι αστερίσκοι υποδεικνύουν τον συνδυασμό μεθόδων και οι έντονοι αριθμοί υποδεικνύουν το καλύτερο αποτέλεσμα ανά μετρική.

Παρατηρούμε ότι ο συνδυασμός των μεθόδων μας αποδίδει καλύτερα αποτελέσματα από ό,τι η καθεμία ξεχωριστά, κάτι που αναδεικνύει την αποτελεσματικότητα των δύο μεθόδων και το γεγονός ότι λειτουργούν συμπληρωματικά. Οι τελικές μετρικές είναι εντυπωσιακές, καθώς φτάνουν σε επίπεδα κοντά στο 90% σε όλες τις περιπτώσεις, κοντά στα επίπεδα που επιτυγχάνονται μέσω του LLM Prompting στο σύνολο δεδομένων AG-News.

Από τη μία πλευρά, το γεγονός ότι οι επιδόσεις αυξάνονται σε σύγκριση με το IndivSC αποδεικνύει ότι τα LLMs είναι καλά εξοπλισμένα για την ανάλυση μικρότερων κειμένων και κειμένων που βρίσκονται στη μέση των στατιστικών κατανομών του συνολικού ποσοστού συμφωνιών. Από την άλλη πλευρά, το γεγονός ότι οι επιδόσεις αυξάνονται σε σύγκριση με την Προτροπή LLMs επαληθεύει την ευρωστία του IndivSC, ιδίως όταν πρόκειται για μεγαλύτερα δείγματα κειμένων. Επιπλέον, επειδή οι μέθοδοι συνολικών συμφωνιών αποδίδει παρόμοια αποτελέσματα για όλες τις επιθέσεις, καταφέρνει να αντισταθμίσει τη μειωμένη ανάκληση που παρουσιάζει η Προτροπή LLMs έναντι της BAE. Κάτι άλλο που πρέπει να σημειωθεί είναι ότι μέσω του συνδυασμού των μεθόδων, τα μικρότερα LLM αύξησαν τις μετρικές τους σημαντικά περισσότερο από τα μεγαλύτερα. Αυτό οδήγησε σε μείωση του χάσματος των ικανοτήτων τους, αν και η διαφορά εξακολουθεί να είναι αισθητή, με το GPT-4o-mini να επιτυγχάνει τα υψηλότερα σκορ έναντι όλων των επιθέσεων στο BERT, για να δώσουμε ένα παράδειγμα. Παρόλα αυτά, η παρατήρηση αυτή συμβαδίζει με την προηγούμενη διαπίστωσή μας ότι τα Μεγάλα Γλωσσικά Μοντέλα γενικά βρίσκουν μικρότερα κείμενα ευκολότερα στην ανάλυσή τους και οδηγεί στο συμπέρασμα ότι οι διαφορές στην ικανότητα των μοντέλων εμφανίζονται κυρίως σε κείμενα μεγαλύτερου μήκους.

1.3.5 Σύγκριση με Μεθόδους Ανίχνευσης Τελευταίας Τεχνολογίας

Συγκρίνουμε την προτεινόμενη προσέγγισή μας με δύο σύγχρονες μεθόδους ανίχνευσης ανταγωνιστικών επιθέσεων στο NLP: την FGWS (Frequency-Guided Word Substitutions) [45] η οποία είναι ένα πλαίσιο ανίχνευσης που εχμεταλλεύεται την παρατήρηση ότι οι ανταγωνιστικές επιθέσεις τείνουν να αντικαθιστούν λέξεις με λέξεις που συναντώνται λιγότερο συχνά, και την RDE (Robust Density Estimation) [75] η οποία για να επιτύχει την ανίχνευση ανταγωνιστικών επιθέσεων, εκτιμά την πυκνότητα πιθανότητας του δείγματος κειμένου.

| | | Adversarial Attacks | | | | | | | | |
|-----------------|------------------|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Attacked Models | Methods | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| | | IMDB | | | | | | | | |
| BERT | Gemma-2-2B * | 0.82 | 0.81 | 0.81 | 0.87 | 0.87 | 0.86 | 0.83 | 0.83 | 0.84 |
| | Phi-3-mini * | 0.84 | 0.82 | 0.82 | 0.91 | 0.89 | 0.88 | 0.82 | 0.81 | 0.82 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.92 | 0.89 | 0.87 | 0.87 | 0.85 | 0.85 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.85 | 0.91 | 0.88 | 0.86 | 0.85 | 0.84 | 0.84 |
| | Gemma-2-27B * | 0.91 | 0.86 | 0.86 | 0.94 | 0.88 | 0.86 | 0.88 | 0.85 | 0.85 |
| | Claude-3-Haiku * | 0.88 | 0.85 | 0.84 | 0.92 | 0.89 | 0.88 | 0.84 | 0.85 | 0.85 |
| | GPT-4o-mini * | 0.92 | 0.88 | 0.87 | 0.94 | 0.90 | 0.89 | 0.88 | 0.85 | 0.86 |
| | FGWS | 0.85 | 0.87 | 0.87 | 0.88 | 0.89 | 0.89 | 0.62 | 0.72 | 0.76 |
| | RDE | 0.97 | 0.94 | 0.93 | 0.88 | 0.89 | 0.89 | 0.94 | 0.92 | 0.92 |
| RoBERTa | Gemma-2-2B * | 0.83 | 0.84 | 0.84 | 0.85 | 0.82 | 0.82 | 0.80 | 0.84 | 0.85 |
| | Phi-3-mini * | 0.90 | 0.87 | 0.87 | 0.84 | 0.82 | 0.82 | 0.82 | 0.83 | 0.84 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.87 | 0.82 | 0.82 | 0.85 | 0.85 | 0.86 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.86 | 0.87 | 0.82 | 0.82 | 0.85 | 0.87 | 0.87 |
| | Gemma-2-27B * | 0.92 | 0.88 | 0.87 | 0.89 | 0.82 | 0.82 | 0.84 | 0.85 | 0.86 |
| | Claude-3-Haiku * | 0.90 | 0.88 | 0.87 | 0.87 | 0.83 | 0.83 | 0.87 | 0.88 | 0.88 |
| | GPT-4o-mini * | 0.90 | 0.89 | 0.88 | 0.87 | 0.83 | 0.82 | 0.85 | 0.87 | 0.87 |
| | FGWS | 0.85 | 0.87 | 0.88 | 0.92 | 0.91 | 0.91 | 0.62 | 0.72 | 0.76 |
| | RDE | 0.99 | 0.95 | 0.94 | 0.95 | 0.93 | 0.93 | 0.95 | 0.93 | 0.93 |
| | | AG-News | | | | | | | | |
| BERT | Gemma-2-2B | 0.79 | 0.86 | 0.88 | 0.70 | 0.81 | 0.83 | 0.25 | 0.38 | 0.57 |
| | Phi-3-mini | 0.91 | 0.89 | 0.89 | 0.90 | 0.88 | 0.87 | 0.49 | 0.61 | 0.67 |
| | Llama-3.1-8B | 0.96 | 0.88 | 0.88 | 0.93 | 0.88 | 0.87 | 0.71 | 0.72 | 0.72 |
| | Gemma-2-9B | 0.94 | 0.91 | 0.92 | 0.89 | 0.89 | 0.88 | 0.47 | 0.59 | 0.66 |
| | Gemma-2-27B | 0.97 | 0.92 | 0.92 | 0.95 | 0.91 | 0.90 | 0.62 | 0.72 | 0.75 |
| | Claude-3-Haiku | 0.91 | 0.94 | 0.95 | 0.89 | 0.93 | 0.93 | 0.45 | 0.61 | 0.70 |
| | GPT-4o-mini | 0.98 | 0.90 | 0.90 | 0.96 | 0.90 | 0.89 | 0.68 | 0.74 | 0.75 |
| | FGWS | 0.82 | 0.86 | 0.86 | 0.91 | 0.91 | 0.90 | 0.64 | 0.74 | 0.77 |
| | RDE | 0.96 | 0.93 | 0.93 | 0.89 | 0.89 | 0.89 | 0.97 | 0.94 | 0.93 |
| RoBERTa | Gemma-2-2B | 0.83 | 0.88 | 0.89 | 0.74 | 0.83 | 0.86 | 0.29 | 0.43 | 0.59 |
| | Phi-3-mini | 0.91 | 0.88 | 0.88 | 0.91 | 0.88 | 0.88 | 0.55 | 0.65 | 0.70 |
| | Llama-3.1-8B | 0.95 | 0.87 | 0.86 | 0.92 | 0.82 | 0.82 | 0.69 | 0.71 | 0.70 |
| | Gemma-2-9B | 0.93 | 0.94 | 0.94 | 0.90 | 0.90 | 0.91 | 0.62 | 0.73 | 0.76 |
| | Gemma-2-27B | 0.95 | 0.92 | 0.92 | 0.95 | 0.91 | 0.92 | 0.66 | 0.75 | 0.77 |
| | Claude-3-Haiku | 0.92 | 0.94 | 0.94 | 0.84 | 0.91 | 0.92 | 0.47 | 0.63 | 0.71 |
| | GPT-4o-mini | 0.93 | 0.89 | 0.88 | 0.96 | 0.90 | 0.90 | 0.71 | 0.75 | 0.75 |
| | FGWS | 0.79 | 0.84 | 0.84 | 0.87 | 0.88 | 0.88 | 0.53 | 0.65 | 0.72 |
| | RDE | 0.93 | 0.92 | 0.91 | 0.85 | 0.87 | 0.87 | 0.89 | 0.90 | 0.90 |

Table 1.13: Σύγκριση της προσέγγισής μας με σύγχρονες μεθόδους ανίχνευσης. Οι αστερίσκοι υποδεικνύουν τον συνδυασμό των μεθόδων και οι έντονοι αριθμοί υποδεικνύουν το καλύτερο αποτέλεσμα ανά μετρική.

Ο πίνακας 1.13 παρουσιάζει τα αποτελέσματα και για τα 6 σύνολα δεδομένων, τόσο για την προσέγγισή μας όσο και για τις σύγχρονες μεθόδους ανίχνευσης που περιγράφηκαν παραπάνω. Παρατηρούμε αμέσως την ανταγωνιστικότητα της προσέγγισής μας σε όλους τους τομείς, ειδικά όταν πρόκειται για τα 3 μεγάλης κλίμακας LLM.

Η RDE φαίνεται να είναι η πιο ισχυρή προσέγγιση στα δεδομένα IMDb, επιτυγχάνοντας τις καλύτερες επιδόσεις και στις 3 μετρικές σε όλα τα σενάρια, εκτός από την επίθεση PWWS κατά του μοντέλου BERT, στην οποία το GPT-4o-mini υπερσχύει στην ανάκληση και στο σκορ F1, και έχει ίση ακρίβεια. Όσον αφορά τα δεδομένα AG-News, η ισχύς της προτροπής LLM αναδεικνύεται ακόμη περισσότερο, με το RDE να υπολείπεται είτε του Gemma-2-27B, είτε του Claude-3-Haiku, είτε του GPT-4o-mini σε όλες τις περιπτώσεις, με εξαίρεση την επίθεση BAE.

Όσον αφορά την FGWS, έχει γενικά υψηλές επιδόσεις, αλλά υπολείπεται σε όλες τις περιπτώσεις από άλλες μεθόδους ανίχνευσης. Για παράδειγμα, από τους 12 συνολικά συνδυασμούς δεδομένων-επίθεσης-μοντέλου, το GPT-4o-mini υπερτερεί της FGWS σε 11 από αυτούς στην Ανάκληση, σε 9 από αυτούς στο σκορ F1 (ίση επίδοση σε 1 συνδυασμό) και σε 6 από αυτούς στην Ακρίβεια (ίση επίδοση σε 3 συνδυασμούς). Συγκριτικά, οι συνδυασμοί στους οποίους το Gemma-2-27B υπερτερεί της FGWS είναι 10 στην Ανάκληση, 7 στη βαθμολογία F1 (ίση επίδοση σε 1 συνδυασμό) και 6 στην Ακρίβεια (ίση επίδοση σε 1 συνδυασμούς).

Αναφορικά με την υποβάθμιση της απόδοσης της μεθόδου μας έναντι της επίθεσης BAE, η οποία συζητήθηκε νωρίτερα, είναι ενδιαφέρον να παρατηρήσουμε ότι η ίδια ακριβώς συμπεριφορά εμφανίζεται και για τη μέθοδο ανίχνευσης FGWS, γεγονός που αναδεικνύει την αυξημένη λεπτότητα της επίθεσης. Τούτου λεχθέντος, η RDE δεν παρουσιάζει καμία διαφορά στην απόδοση σε σχέση με την BAE, επιδεικνύοντας έτσι αξιοσημείωτη ευρωστία.

Σε αυτό το σημείο, πρέπει να σημειωθεί ότι σε σύγκριση με άλλες εργασίες της βιβλιογραφίας, η προσέγγισή μας, καθώς και οι FGWS και RDE, δεν κάνουν καμία υπόθεση σχετικά με τον αλγόριθμο της επίθεσης ή το μοντέλο που δέχεται την επίθεση. Επιπλέον, ένας παράγοντας που διαφοροποιεί την προτεινόμενη από εμάς μέθοδο από τις εξεταζόμενες, είναι το γεγονός ότι δεν απαιτεί πρόσβαση στο χώρο χαρακτηριστικών του μοντέλου ταξινομητή.

Λαμβάνοντας υπόψη όλα τα παραπάνω, καταλήγουμε στο συμπέρασμα ότι η προσέγγισή μας αποτελεί μια βιώσιμη λύση, χρήσιμη για την plug-and-play ανίχνευση εχθρικών κειμένων στην εφαρμογή ταξινόμησης κειμένων.

1.4 Συμπεράσματα

1.4.1 Συζήτηση

Στην παρούσα εργασία, διερευνήσαμε τη χρήση μεγάλων γλωσσικών μοντέλων για το έργο της ανίχνευσης ανταγωνιστικών επιθέσεων στην ταξινόμηση κειμένων.

Αρχικά, αναπτύξαμε ένα πλαίσιο μηχανικής προτροπών με στόχο τη διαμόρφωση προτροπών φυσικής γλώσσας που επιτρέπουν στα LLMs να εκτελούν το έργο της ανίχνευσης ανταγωνιστικών κειμένων. Σε αυτό το πλαίσιο, εφαρμόσαμε μια σειρά από σύγχρονες τεχνικές προτροπής στην είσοδο του LLM, ενώ παράλληλα αναφέραμε τις αλλαγές στην απόδοση που επιτεύχθηκαν μέσω κάθε βήματος της διαδικασίας. Αυτό που ανακαλύψαμε είναι ότι, παρόλο που τα LLM φάνηκε να μην έχουν εγγενή ικανότητα στον εντοπισμό ανταγωνιστικών επιθέσεων κειμένου, όταν τους δίνεται μια σχολαστικά επεξεργασμένη προτροπή, μπορούν να γίνουν αποτελεσματικοί ανιχνευτές. Διαπιστώσαμε ότι η προτροπή που κατασκευάσαμε γενικευόταν καλά σε διαφορετικά δεδομένα, πλαίσια επιθέσεων και μοντέλα-θύματα, ενώ βλέποντας ότι τα μεγαλύτερα LLMs ξεπερνούσαν σταθερά τα μικρότερα στα πειράματα που πραγματοποιήσαμε, φαίνεται ότι ο πρωταρχικός παράγοντας που επηρέαζε την απόδοση στην εφαρμογή αυτή ήταν η κλίμακα του μοντέλου. Παρά την εύρεση θετικών δεικτών για την ισχύ και την επεξηγησιμότητα της προτροπής μεγάλων γλωσσικών μοντέλων, έγιναν επίσης εμφανείς ορισμένοι περιορισμοί αυτής της προσέγγισης. Ο ένας είναι η υποβάθμισμένη απόδοση έναντι της BAE, μιας ανταγωνιστικής επίθεσης που χρησιμοποιεί το BERT, η οποία θα μπορούσε να υποδηλώνει αδυναμία απέναντι σε ορισμένους τύπους επιθέσεων, και ο άλλος είναι η μειωμένη ικανότητα των LLMs να αναγνωρίζουν σωστά μεγαλύτερα ανταγωνιστικά κείμενα. Προκειμένου να αξιολογήσουμε περαιτέρω την απόδοση των μεθόδων μας, πραγματοποιήσαμε μια έρευνα αξιολόγησης από ανθρώπους, η οποία έδειξε ότι οι άνθρωποι έχουν ελαφρώς χειρότερες επιδόσεις από τα πιο ισχυρά LLM στο

έργο της ανίχνευσης ανταγωνιστικών δειγμάτων. Επιπροσθέτως, πραγματοποιήσαμε έναν έλεγχο ορθότητας, ο οποίος επικύρωσε ότι τα αποτελέσματά μας δεν επηρεάστηκαν από μόλυνση δεδομένων στα LLMs.

Εμπνευσμένοι από τις αδυναμίες που ανακαλύψαμε στην προτροπή LLMs, αναπτύξαμε επίσης έναν δεύτερο τύπο μεθοδολογίας για την ανίχνευση αντίπαλων κειμένων - την ταξινόμηση μεμονωμένων προτάσεων, η οποία εξετάζει τις ταξινομήσεις που δίνονται σε κάθε μεμονωμένη πρόταση ενός κειμένου και τις συγκρίνει με την ταξινόμηση που δίνεται σε ολόκληρο το κείμενο, από το γλωσσικό μοντέλο που δέχεται επίθεση. Η μέθοδος αυτή υλοποιήθηκε μέσω τριών διαφορετικών προσεγγίσεων: τις Συνολικές Συμφωνίες, τη Μακρύτερη Συνεχή Ακολουθία Συμφωνιών και τις Συνολικές Εναλλαγές, με στόχο την εύρεση εκείνης που αξιοποιεί την πιο στατιστικά σημαντική διαφορά μεταξύ αρχικού και ανταγωνιστικού κειμένου. Αυτή αποδείχθηκε ότι ήταν η προσέγγιση Συνολικών Συμφωνιών, η οποία παρουσίασε όχι μόνο υψηλή ικανότητα διαχωρισμού μεταξύ αρχικών και ανταγωνιστικών δειγμάτων, αλλά και μεγάλη εναλλαξιμότητα έναντι διαφορετικών επιθέσεων. Στη συνέχεια συνδυάσαμε τις δύο προτεινόμενες μεθόδους μας σε μια ενοποιημένη προσέγγιση, η οποία φάνηκε να αντισταθμίζει τις αδυναμίες των δύο μεμονωμένων μεθόδων και να αποδίδει συνολικά τα καλύτερα αποτελέσματα. Τέλος, πραγματοποιήσαμε μια σύγκριση με άλλους σύγχρονους ανιχνευτές ανταγωνιστικών δειγμάτων και παρατηρήσαμε ότι η προσέγγισή μας είναι ανταγωνιστική και μπορεί να αποτελέσει μια βιώσιμη λύση για plug-and-play ανίχνευση ανταγωνιστικών κειμένων.

1.4.2 Μελλοντικές Κατευθύνσεις

Κλείνοντας την παρούσα διατριβή, θα θέλαμε να προτείνουμε δρόμους για περαιτέρω βελτίωση ή εναλλακτικές προσεγγίσεις αυτής της εργασίας. Πρώτον, θα ήταν ενδιαφέρον να εξεταστεί το ενδεχόμενο επέκτασης του έργου της ανταγωνιστικής ανίχνευσης σε αυτό του ανταγωνιστικού καθαρισμού. Αυτό θα περιελάμβανε τη χρήση μεγάλων γλωσσικών μοντέλων για τον καθαρισμό του κειμένου εισόδου, με στόχο την αφαίρεση των ανταγωνιστικών διαταραχών στο εσωτερικό του, έτσι ώστε το αποκατεστημένο κείμενο να ταξινομείται σωστά από το μοντέλο-θύμα. Μια άλλη ενδιαφέρουσα προσέγγιση θα μπορούσε να είναι αυτή της διερεύνησης επιλογών για fine-tuning των LLMs στην εφαρμογή της ανταγωνιστικής ανίχνευσης, είτε ακολουθώντας το παραδοσιακό πρότυπο εκπαίδευσης είτε αξιοποιώντας parameter-efficient μεθόδους.

Chapter 2

Introduction

In recent years, there has been a widespread deployment of Deep Neural Networks (DNNs) across various domains, which include computer vision [31], [60], audio [64] and natural language processing (NLP) [32], [5]. This has enabled remarkable advancements in the automation and accuracy of real-world tasks such as image classification, speech recognition, and text classification. However, systems utilizing DNNs remain highly vulnerable, as they are susceptible to the threat of malicious actors.

One such case are adversarial attacks, which can add small perturbations to an original input that then manages to deceive a targeted DNN [61], [18]. The resulting samples are called adversarial and they are generated in such a way that the alterations they underwent are imperceptible by humans. The majority of research on adversarial attacks has been focused on the image domain, with little attention being given to the NLP field [50], [11], [39], [28], [40]. Unlike images, text data are discrete by nature, which complicates the generation of adversarial examples in NLP. Adversarial attacks in text-based tasks, such as sentiment analysis or question answering, can be executed by altering individual words, characters, or whole sentence of a given text. The goal is for these alterations to be indiscernible and not influence human judgment in comparison to the original sample. For this reason, effective adversarial attacks lack obviously discernible errors, maintain grammatical correctness, and are semantically equivalent to the original text. Given the rise of such sophisticated adversarial techniques, the need for reliable defense mechanisms has become paramount.

This thesis explores the utilization of Large Language Models (LLMs) for detecting adversarial attacks in text classification tasks, through two proposed methods. During this process, several experiments were undertaken, resulting in the following contributions:

- We developed a simple framework for engineering prompts that enables the use of Large Language Models (LLMs) for the task of detecting adversarial attacks in text classification. This is a method that is agnostic of the attack method and the attacked model, and is tested against multiple dataset-attack-model combinations.
- We did a comprehensive analysis of LLMs' ability at adversarial detection. To further assess our results, we conducted both a human evaluation and a sanity test for data contamination.
- We introduced a simple, yet effective way of detecting adversarial texts, utilizing the attacked language model itself. This method is also model agnostic and equally successful against multiple attacking scenarios.
- We combined our two proposed methods into a unified approach, achieving performance that competes with state-of-the-art detection methods.

Chapter 3

Adversarial Attacks in NLP

In recent years, Deep Neural Networks (DNNs) have been effectively utilized to tackle challenging problems in multiple areas, such as computer vision [31], [60], audio [64] and natural language processing (NLP) [32], [5]. However, it has been shown [61], [18] that by appropriately modifying the input of a DNN, it is possible to fool it into making a wrong prediction, while human judgment is not affected. These altered inputs are known as adversarial examples and against them, the performance DNNs degrades significantly. This poses major threats to DNN-based systems, especially ones deployed in real-world, security-critical applications [57], [77].

Research in the field of generating adversarial examples first started and has been extensively explored in the image domain and more recently extended to the text domain. [79], [67], [54], [12], [8], [10]

In this section we will define the problem of adversarial attacks; explore the differences between adversarial attacks in the image and text domains; create a taxonomy for textual adversarial attacks; explore how to evaluate them; and investigate some characteristic textual attack methods.

Contents

| | |
|--|-----------|
| 3.1 Preliminaries | 52 |
| 3.1.1 Definitions | 52 |
| 3.1.2 From Image To Text | 52 |
| 3.2 Taxonomy of Textual Adversarial Attacks | 53 |
| 3.2.1 Model Access | 53 |
| 3.2.2 Target Type | 54 |
| 3.2.3 Perturbation Level | 54 |
| 3.2.4 Generation method | 54 |
| 3.2.5 Task | 55 |
| 3.3 Evaluation of Textual Adversarial Attacks | 56 |
| 3.3.1 Attack Effectiveness | 56 |
| 3.3.2 Attack Imperceptibility | 58 |
| 3.3.3 Attack Efficiency | 59 |
| 3.3.4 Attack Transferability | 59 |
| 3.4 Instances of Textual Adversarial Attacks | 59 |
| 3.4.1 Char-level Attack | 59 |
| 3.4.2 Word-level Attack | 60 |
| 3.4.3 Sentence-level Attack | 60 |
| 3.4.4 Multi-level Attack | 60 |

3.1 Preliminaries

3.1.1 Definitions

Deep Neural Network (DNN): A deep neural network (DNN) can be represented as the function $F : X \rightarrow Y$, which maps the input set X to the output prediction set Y .

Perturbation: Perturbations are small changes that are intentionally added to the original input in order to mislead the DNN into making a wrong prediction.

Adversarial Example: Attackers generate adversarial examples by adding small perturbations to the original input that then fools the DNN. Formally, an adversarial example x' can be defined as:

$$\begin{aligned} F(x) &= y \\ x' &= x + \varepsilon \\ F(x') &= y', y' \neq y \end{aligned} \tag{3.1.1}$$

, where ε is the added perturbation. At the same time, x' has to be imperceptible to humans, which means that humans would still classify it the same way as the original example and that they should not be able to distinguish it from the original example.

Target Model: The model that is being attacked with adversarial examples is called the target model.

3.1.2 From Image To Text

The research direction of attacking image recognition DNNs has received sustained attention by the computer vision community. Szegedy et al. [61] first explored the idea of applying an imperceptible perturbation to an input image that causes state-of-the-art models to misclassify that image. In that paper, the term “adversarial example” was coined and after Goodfellow et al. [18] proposed a fast generation method for them, the research topic got widely popularized [51] [6].

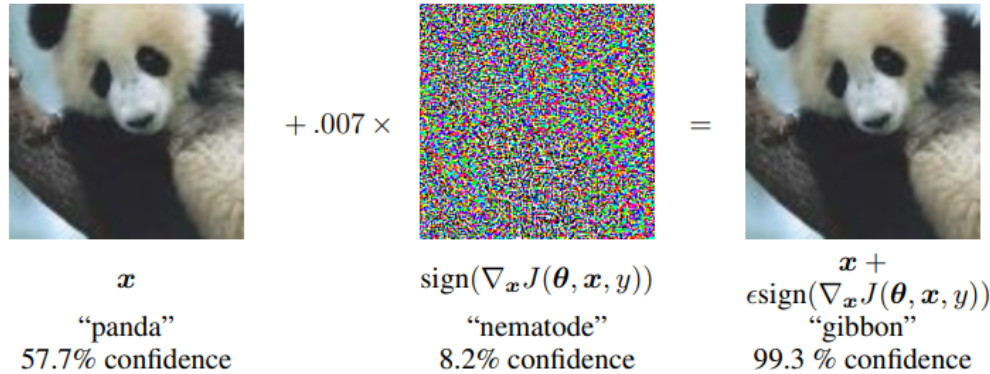


Figure 3.1.1: Instance of an adversarial attack in image classification. The unnoticeable perturbation that is added to the original image effectively deceives the model, causing it to shift its classification prediction from “panda” to “gibbon”. [18]

However, the approaches developed in the image domain cannot be directly transferred to the text domain since the two are characterized by fundamental differences.

Input Type: First of all, image data is inherently continuous, and existing methods often employ L_p norm metrics to quantify the difference between an unaltered data point and one that has been perturbed. Conversely, textual data is discrete in nature, which makes it necessary to explore new methods for generating appropriate perturbations and measuring their distance to the original data point. Some systems choose to reconcile this difference by initially converting textual data into continuous data and then employing attack methods inspired from the image field.

Perceivability of Perturbations: Another fundamental difference is the degree in which humans perceive changes in images and texts. Small alterations in image pixel values can totally evade human perception, and as a result, not alter human judgment but only deceive DNN models (e.g. Figure 3.1.1). In contrast, even the smallest textual modifications, such as character or word substitutions, are readily noticeable. This can possibly make the attack easily identifiable or correctable, either by humans or by machines, for example, via spell and grammar checkers. Therefore, this quality of text adds a lot of nuances to the question of when an adversarial example is imperceptible by humans.

Semantics: Lastly, in the realm of images, minor changes often have no impact on the semantics of an image, as they tend to be inconsequential and imperceptible. For example, the perturbation of individual pixels cannot transform an image of a panda into an image of a gibbon, even though it can fool a DNN (as shown in Figure 3.1.1). In stark contrast, perturbations in text can easily alter the meaning of individual words and whole sentences; for example, removing a negation word from a sentence can invert its sentiment. Altering the semantics of the input is in conflict with the definition of an adversarial attack and therefore, the constraint of semantic preservation requires extra attention in the text domain.

As a result of these differences, textual attackers either get inspired by and appropriately adjust attacks designed for images, or introduce innovative techniques in their methods.

In NLP, Papernot et al. [50] were the first to investigate the creation of adversarial samples by inserting subtle perturbations into the textual inputs. This revelation drew attention in the NLP community and sparked the emergence of research on text-based adversarial attacks. Attacks on text have now been created for a wide range of NLP tasks, from text classification [13] to question answering [23] and argumentation mining [46]; and therefore, pose system security threats.

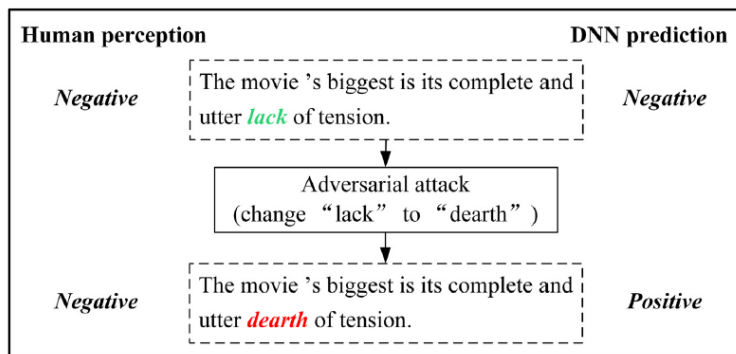


Figure 3.1.2: Instance of an adversarial attack in sentiment analysis. Replacing the word "lack" with its synonym "dearth" alters the model's prediction from negative to positive, while the sentiment perceived by humans remains negative. [54]

3.2 Taxonomy of Textual Adversarial Attacks

Adversarial attacks in NLP can be categorized using a multitude of criteria. In this section, we introduce a taxonomy for such attacks.

3.2.1 Model Access

A major differentiation of adversarial examples is whether they are generated in a black-box or white-box manner, in terms of available access to the targeted DNN model.

White-box: In the white-box setting, the attacker has unrestricted access to the DNN and relies on knowledge of its architecture, parameters, functions and training data in order to create effective adversarial examples.

Black-box: In the black-box setting, all the internal information of the DNN is inaccessible and the only thing available is its outputs. Attacks in this scenario often depend on heuristics; they are implemented

by repeatedly querying the target model and monitoring any changes in the output that can lead to the successful generation of adversarial samples.

White-box attacks generally achieve a high success rate, as they typically approximate the worst-case attack for a specific model. However, black-box attacks are more pragmatic, as they replicate real-world scenarios in which the internal workings of the DNN are not available.

3.2.2 Target Type

Adversarial attacks can be divided into targeted and non-targeted based on the intended change of the output prediction.

Targeted: In a targeted attack, the adversarial example is directed towards a specific prediction, which is the attacker’s objective. This process primarily involves increasing the prediction probability of the target result.

Non-targeted: In non-targeted attacks, the only objective of the adversary is to deceive the model, meaning that the prediction can be any incorrect one. This type of attacks usually achieve their objective by decreasing the prediction probability of the correct result.

Compared to non-targeted attacks, targeted ones are more stringent, as they impose additional constraints on the model output. In the case of binary tasks, such as binary classification, a targeted attack is equivalent to a non-targeted attack.

3.2.3 Perturbation Level

The perturbation level criterion pertains to the specific text units on which the perturbations are performed in order to create adversarial examples. The sub-categorizations are char-level, word-level, sentence-level, and multi-level.

Char-level: In char-level attacks, individual characters within words are altered in order to create adversarial examples. Typically, these modifications consist of misspellings that can deceive predictors, and the common operations encompass letter insertions, swaps, deletions, and flips.

Word-level: Word-level attacks perturb whole words instead of individual characters in the words. Such perturbations are created by inserting, deleting or replacing words in various way.

Sentence-level: In sentence-level attacks, the whole sentence structure can be modified, a new sentence can be inserted into the text or an existing one can be rewritten, while preserving the overall meaning and grammaticality.

Multi-level: Multi-level attacks combine more than one of the previous types of perturbations in order to fool the target model.

The higher the level of the attack is, the higher the degree of flexibility it possesses, enhancing its ability to generate imperceptible adversarial samples with a high attack success rate. However, this also implies that its complexity and computational cost increase.

3.2.4 Generation method

This classification refers to the process with which the adversarial examples are generated. We divide the possible methods into six types: gradient-based, optimization-based, importance-based, edit-based, paraphrase-based and generative model-based.

Gradient-based: These techniques involve computing the forward derivative to the input and generating adversarial perturbations through gradient backpropagation. This process requires initially vectorizing the text, which often results in the introduction of spelling and grammatical errors into the generated examples.

Optimization-based: In this strategy, the objective is to maximize the prediction error of the victim model while keeping the difference between the adversarial example and the original input within a certain threshold. Such samples are predominantly generated using evolutionary algorithms.

Importance-based: Within this approach, the selection process of which text elements to modify and how to modify them depends on the importance of each element in relation to the victim model’s predictions. This is effective because generally, the more critical a modified word is, the easier it becomes to influence the prediction of the victim model through it. Adversarial examples produced with this approach tend to be imperceptible, as they maintain semantic consistency and grammatical correctness.

Edit-based: This method involves crafting adversarial examples solely through basic actions such as inserting, removing, or swapping characters, words, or sentences within the text. Such editing operations can be combined with the previous generation methods, but here we are referring to the use of these operations exclusively.

Paraphrase-based: Attacks of this type employ a sentence paraphrase to its original counterpart, while maintaining semantic similarity and grammaticality.

Generative Model-based: This technique leverages generative models like Generative Adversarial Networks (GANs) [17] and encoder-decoder models to generate adversarial texts. The purpose of using these models is to create examples that are more natural.

3.2.5 Task

Based on the NLP application scenario for which the adversarial attacks are created, we can classify them into nine main categories: text classification, machine translation, machine reading comprehension, question answering, text entailment, part-of-speech tagging, text summarization, dialogue generation and multi-modal tasks.

Text Classification: In text classification, the goal is to classify the given text into one of the possible classes. As this is the most general NLP scenario, a multitude of more specific tasks such as sentiment analysis, spam detection and relation extraction can be formed into classification problems. Therefore, the majority of research on adversarial attacks is geared towards text classification systems.

Machine Translation: A machine translation system is used to automatically translate text from one language to another. When given an adversarially edited input text however, the system fails and outputs an incorrect translation of the text.

Machine Reading Comprehension: The task of machine reading comprehension (MRC) systems is to extract meaning from a given text, just like a human would. To check whether this is achieved, the model is usually asked to answer questions such as “who”, “when” or “how”, based on the given text passage. Attacks on such systems are achieved by adversarially editing or adding to the input and are mostly evaluated on Question Answering datasets.

Article: Super Bowl 50
Paragraph: “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”
Question: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 3.2.1: Instance of an adversarial attack in machine reading comprehension. [23]

Question Answering: Question answering (QA) systems can automatically answer questions, given a textual context. In contrast to reading comprehension, adversarial attacks against this task can be generated by appropriately paraphrasing the question posed.

Text Entailment: The objective of text entailment, also known as natural language inference (NLI), is to assess the relationship between a premise and a hypothesis text. This relationship can be one of three possibilities: entailment, contradiction, or neutrality. Adversarial examples within this context will deceive the model, resulting in erroneous resolutions of the textual relationship.

| |
|--|
| Original Text Prediction: Entailment (Confidence = 86%) |
| Premise: <i>A runner wearing purple strives for the finish line.</i> |
| Hypothesis: <i>A runner wants to head for the finish line.</i> |
| Adversarial Text Prediction: Contradiction (Confidence = 43%) |
| Premise: <i>A runner wearing purple strives for the finish line.</i> |
| Hypothesis: <i>A racer wants to head for the finish line.</i> |

Figure 3.2.2: Instance of an adversarial attack in text entailment. [2]

Part-of-speech Tagging: Part-of-speech (POS) tagging is the practice of annotating words within a text to a particular part-of-speech categorization, based on a combination of both the word’s inherent definition and its context within the text. This is a fundamental task that plays a pivotal role in other downstream NLP tasks, such as syntactic parsing, and therefore the research of adversarial attacks against it is rendered critical.

Text Summarization: The aim of text summarization is to provide a brief and coherent summary that encapsulates the general essence of a given text. Adversarial examples are usually not engineered solely for this task, as it is challenging.

Dialogue Generation: Dialogue generation represents a text generation task wherein responses are automatically generated in accordance with the provided input posts by the user. Although dialogue systems assume a pivotal role in real-world applications, such as virtual assistants, they can be fooled by subtle perturbations.

Multi-modal Tasks: Beyond tasks that deal with exclusively textual inputs, there exist several NLP-related multi-modal tasks that are susceptible to the threat of adversarial attacks. These multi-modal tasks can be effectively categorized as either text-and-vision or text-and-audio.

Text-and-vision: An example of a text-and-vision application is image captioning, wherein the model takes an image as input and outputs a textual caption that describes its visual contents. Adversarial attacks against this type of systems lead to the generation of erroneous or specifically targeted captions.

Text-and-speech: One instance of a text-and-speech application is speech recognition, where the system is tasked with recognizing spoken language and subsequently transcribing it into text. Adversarial attacks against such models involve the introduction of inaudible perturbations to the input audio, resulting in the generation of inaccurate or intentionally targeted text.

It is important to note that many adversarial attacks can be employed across multiple different NLP tasks.

3.3 Evaluation of Textual Adversarial Attacks

When assessing adversarial attacks, four primary aspects merit examination: effectiveness, imperceptibility, efficiency and transferability.

3.3.1 Attack Effectiveness

The evaluation of effectiveness reflects the ability of an adversarial attack to mislead a DNN into producing incorrect outputs and consequently degrade its performance. To gauge this aspect, we can use the following performance metrics.

Accuracy Rate: This metric signifies the ratio of correctly classified examples by the victim model relative to the total input instances. A diminished accuracy rate suggests a high effectiveness of adversarial examples.

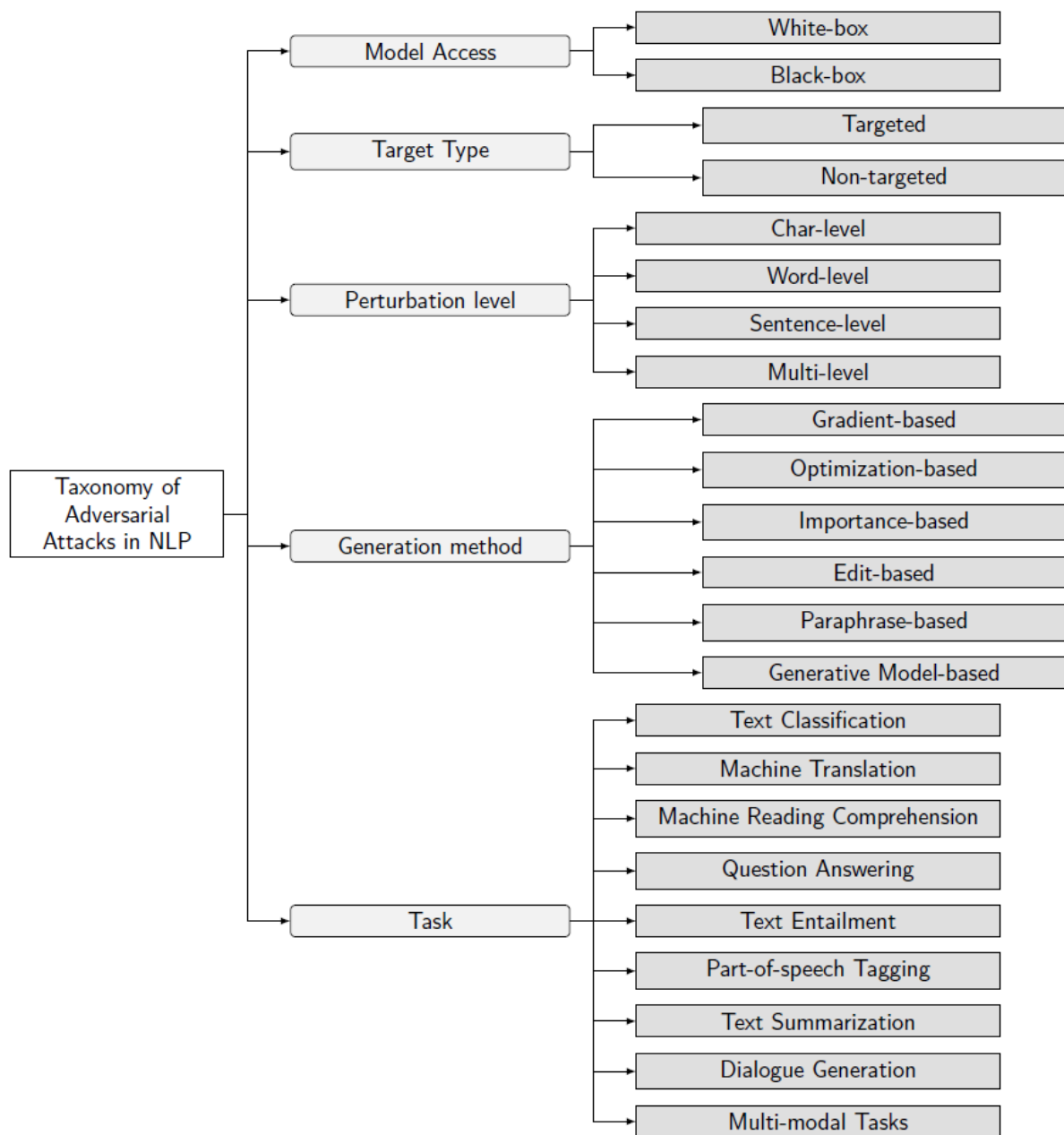


Figure 3.2.3: Taxonomy of Textual Adversarial Attacks

Conversely, a consistent or increased accuracy rate implies that the adversarial examples do not accomplish their goal.

Attack Success Rate: In contrast to the accuracy rate, the attack success rate denotes the percentage of examples erroneously predicted by the victim model among the total input instances. An increased attack success rate corresponds to greater efficacy of the adversarial examples.

Accuracy Reduction: This metric offers a more intuitive assessment, as it characterizes the change in accuracy before and after the adversarial attack. Larger accuracy reduction values signify a heightened effect of the adversarial examples generated.

It is important to note that based on the NLP task that is being examined, more specific metrics, such as the **F1 score** or the **Area Under the Curve (AUC)** score for classification tasks and the **BiLingual Evaluation Understudy (BLEU)** score for generation tasks, can also be utilized.

3.3.2 Attack Imperceptibility

As previously mentioned, due to the distinctive characteristics of textual data, evaluating the imperceptibility of adversarial attacks in NLP holds particular significance. These attacks should remain imperceptible to humans and should not influence their judgment compared to the original samples. To fulfill these criteria, an adversarial example must lack obviously discernible errors, maintain grammatical correctness, and be semantically equivalent to the original text. Considering that, despite their discrete nature, texts can be transformed into continuous vectors, imperceptibility metrics for adversarial attacks in NLP can be categorized into two groups: continuous space metrics for vectorized data and discrete space metrics for raw textual data.

Continuous Space Metrics

Euclidean Distance: Given two word-vectors $\vec{m} = (m_1, m_2, \dots, m_k)$ and $\vec{n} = (n_1, n_2, \dots, n_k)$, the Euclidean distance of these two vectors is defined as:

$$d(\vec{m}, \vec{n}) = \sqrt{(m_1 - n_1)^2 + \dots + (m_k - n_k)^2} \quad (3.3.1)$$

Cosine Similarity: Given two word-vectors \vec{m} and \vec{n} , the cosine similarity of these two vectors is defined as:

$$S_C(\vec{m}, \vec{n}) = \frac{\vec{m} \cdot \vec{n}}{\|\vec{m}\|_2 \cdot \|\vec{n}\|_2} = \frac{\sum_{i=1}^k m_i \cdot n_i}{\sqrt{\sum_{i=1}^k m_i^2} \cdot \sqrt{\sum_{i=1}^k n_i^2}} \quad (3.3.2)$$

In contrast to the Euclidean distance, cosine similarity takes the direction of the two word-vectors into consideration for the calculation of their semantic similarity.

Word Mover's Distance (WMD): WMD [33] is an edit distance that operates on word embeddings. It measures the minimum distance that the embedded words of one document need to travel to reach the embedded words of the other document. The minimization is formalized as:

$$\begin{aligned} \min \sum_{i,j=1}^k T_{ij} \|\vec{m}_i - \vec{n}_j\|_2 \\ \text{s.t. : } \sum_{j=1}^k T_{ij} = d_i, \forall i \in \{1, \dots, k\} \\ \sum_{i=1}^k T_{ij} = d'_j, \forall j \in \{1, \dots, k\} \end{aligned} \quad (3.3.3)$$

, where \vec{m}_i and \vec{n}_j are word embeddings of word i and word j respectively. k is the number of words in the vocabulary. d and d' are normalized bag-of-words vectors of each document respectively. T is a flow matrix and T_{ij} denotes how much of word i in d travels to word j in d' .

Discrete Space Metrics

Edit Distance: Edit distance quantifies the minimum changes needed to convert one string to another. Levenshtein distance is a popular edit distance that uses insert, delete and replace operations.

Number of Changes: Number of changes is a straightforward method that involves counting the word-level or character-level modifications that are performed by the adversarial attack.

Grammar and Syntax Evaluation: The validity of the adversarially generated text is verified through the application of grammar and syntax checkers, alongside other potential methods for validity assessment.

Jaccard Similarity Coefficient: The Jaccard similarity coefficient is used to measure the similarity between finite sample sets, through their intersection and union. Given sets A and B, their Jaccard similarity coefficient is calculated as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.3.4)$$

, where $0 \leq J(A, B) \leq 1$. A and B represent two documents. $|A \cap B|$ denotes the number of words that appear in both documents, while $|A \cup B|$ represents the number of unique words in total.

3.3.3 Attack Efficiency

Attack efficiency gauges the efficacy of an adversarial attack concerning the attack time, the computational resources needed for potential training and the generation of adversarial samples, and the amount of queries to the target model. It provides insights into the attack’s practical viability in real-world scenarios where these factors are crucial. In evaluating this dimension, metrics such as the **average number of target model queries** and the **average running time** per generation of each adversarial example can be employed.

3.3.4 Attack Transferability

Transferability indicates the versatility of an attack method. Specifically, it expresses whether adversarial examples that have been generated for a certain target model, on a certain dataset, can effectively attack another model or dataset. Black-box attacks exhibit a higher likelihood of possessing this property, mainly because the way that they generate adversarial examples is independent from the internal details of the targeted DNN. Despite exploration of this topic by researchers, any improvements in the transferability that are achieved tend to come at the expense of attack effectiveness. Thus far, there are no adversarial attack methods that attain a satisfying degree of transferability, thereby making this an ongoing challenge within the field.

3.4 Instances of Textual Adversarial Attacks

In this section, we explore and analyze four characteristic textual attack methods used to generate adversarial examples. Each of these methods correspond to a distinct perturbation level of the original text.

| | Label | Example |
|--------------------|-----------------|---|
| Original | Positive | You don't have to know about music to appreciate the film's easygoing blend of comedy and romance. |
| Deepwordbug | Negative | You don't have to know about music to appreciate the film's easygoing bl se nd of comedy and romance. |
| Textfooler | Negative | You don't have to know about music to acknowledging the film's easygoing mixtures of mockery and ballad . |
| Bertattack | Negative | You don't have to know about music to understand the film's its straight hybrid of satire and - defaulted . |
| Scpn | Negative | Why don't you know about music to appreciate the movie's easytravel mixture of comedy and romance? |

Figure 3.4.1: Instances of adversarial examples generated via different perturbation levels. [65]

3.4.1 Char-level Attack

DeepWordBug [14] is a simple algorithm that can generate adversarial examples via char-level perturbations. It operates under the black-box setting and involves a two-step process. In the initial stage, it assesses the importance of words and determines which ones to alter. For this purpose, Gao et al. employ four token

scoring functions, namely Replace-1 Score, Temporal Head Score, Temporal Tail Score, and Combination Score. In the second stage, DeepWordBug performs a character-level transformation to the most critical tokens found, through the letter operations of swap, substitution, deletion and insertion. In order to minimize the edit distance of the perturbation, Gao et al. utilizes Levenshtein distance. This ensures that the edits performed are imperceptible and lead to examples that are comprehensible by humans. Based on its experiments, DeepWordBug generates adversarial texts that significantly decrease the target model’s accuracy rate for a multitude of NLP tasks. However, the majority of the introduced perturbations are confined to misspellings, which can be easily nullified with the use of spell checkers.

3.4.2 Word-level Attack

BERT-Attack [36] is an effective black-box method for generating adversarial examples by using the pre-trained Masked Language Model (MLM) BERT. Its approach involves a two-step process, in which firstly, all the words in the original text are ranked based on their importance score. Afterwards, to optimize the likelihood of misleading the target model into a wrong prediction, the most vulnerable words are successively replaced one at a time, until the attack succeeds. The process of replacing the most important words with semantically similar and grammatically correct words is the second part of the method. For this, instead of using synonyms or similar words substitution in an embedding space, Li et al. utilize BERT. The use of the MLM ensures that the generated perturbations will preserve the original semantics, adhere to grammatical correctness, and also be context-aware, allowing for the generation of more coherent adversarial samples. BERT-Attack successfully fools BERT’s downstream, fine-tuned models in different NLP tasks, achieving high accuracy reduction scores. Finally, its computational cost seems to be relatively low, rendering it suitable for large-scale generation of adversarial examples.

3.4.3 Sentence-level Attack

Iyyer et al. developed Syntactically Controlled Paraphrase Networks (SCPNs) [22], an encoder-decoder model for generating sentence-level adversarial examples, under the black-box attack setting. In this approach, SCPNs initiate the process by encoding the input sentence using a bidirectional Long Short Term Memory (LSTM) model. Following this, the interpretation of the sentence and the target syntax structure are inputted into the LSTM model for decoding, for which both soft attention over the encoded states and the copy mechanism are utilized. Ultimately, this process yields a paraphrased version of the initial sentence that adheres to the provided syntactic tree and can effectively deceive pre-trained models for different NLP tasks. As SCPNs can potentially alter the semantics of the original sentence significantly, Iyyer et al. impose a constraint for the minimum paraphrastic similarity between the original and the adversarially generated text. However, it should be noted the evaluation of paraphrase quality and grammaticality often demands substantial human effort, as it may result in changes in the sentence structure while having a small semantic difference.

3.4.4 Multi-level Attack

TextBugger [35] is an efficient framework for generating textual adversarial examples for different NLP tasks, under either black-box or white-box conditions. In the white-box scenario, it employs a gradient-based generation method, determining the importance of each word by calculating the Jacobian matrix of the target classifier. Subsequently, TextBugger generates adversarial texts at both character and word levels by employing five distinct editing strategies, which include character insertion, deletion and swapping, and two types of word substitution: one with visually similar words, and one with semantically similar words. In the black-box scenario, Li et al. segment the document into sequences, query the target model with each one, and filter out sentences with prediction labels that differ from the label of the original document. The remaining sequences are then sorted in reverse order based on their confidence scores, which are calculated through the deletion operation. Finally, the text is modified using the same editing procedures as in the white-box scenario. TextBugger achieves a high attack success rate, while perturbing only a small percentage of words and maintaining semantic similarity to the original text.

Chapter 4

Defenses Against Adversarial Attacks in NLP

As NLP systems are increasingly integrated into real-world applications and the sophistication of adversarial attacks in NLP keeps growing, the need of developing robust defenses for them has become urgent. In this section we analyze and categorize the different types of defense methods which have been proposed by the research community in order to mitigate this issue [67], [79], [20].

These strategies can be divided in two main categories: adversarial training and anomaly detection. This distinction can otherwise be described as: defenses which improve a model's robustness during its training phase and defenses which detect adversarial attacks during the inference of the attacked model.

Contents

| | | |
|------------|-----------------------------|-----------|
| 4.1 | Adversarial Training | 62 |
| 4.1.1 | Data Augmentation | 62 |
| 4.1.2 | Model Regularization | 62 |
| 4.2 | Anomaly Detection | 63 |

4.1 Adversarial Training

Adversarial training was first introduced by Goodfellow et al. [19] and it is the strategy of defending neural networks against adversarial attacks by training it to correctly identify both original and adversarial samples. Thereby, one of the contributing factors of the effectiveness achieved by the defense, is the potency of the adversarial perturbations its model was trained on.

Out of the works that utilize adversarial training, some fall into the category of data augmentation, wherein adversarial examples are created and then inserted into the training dataset of the base model. Others, used the generated adversarial samples as a regularizer during the model training. Below, we explore works of both types.

4.1.1 Data Augmentation

Jia and Liang [23], who were pioneers in the creation of adversarial attacks in NLP, also explored training the evaluated text comprehension system on an augmented dataset that included the generated adversarial samples. These were created using a concatenation-based method in which the attacked model was fooled by introducing grammatically correct sentences to the end of a given paragraph. Jia and Liang showed that if an existing model is trained on a dataset which is augmented with the aforementioned samples, then this made the model robust against attacks that use the same type of examples. However, their work also highlighted that this method is not suitable for defending against novel adversarial samples which the model has not encountered.

In regards to the creation of an augmented dataset for training, the options are essentially as diverse as the techniques used for producing adversarial attacks. For example, in the TextBugger attack [35], which we analyzed in Section 3.4, Li et al. also experimented with adversarial training using their generated samples and showed that it can lead to a lower attack success rate without compromising performance on legitimate inputs. In a different work [78], Zhang et al. use continuous bag-of-words (CBOW) to create adversarial perturbations without altering the semantics of original text. The CBOW generates an embedding space which is used for the calculation of the direction vectors between the perturbed words and its surrounding vectors, enabling the placement of constraints that lead to less perceptible adversarial samples. These samples, alongside original ones, are used in adversarial training performed both in a supervised and a semi-supervised setting.

One interesting approach found in the literature involves utilizing GANs, wherein a generator is created with the main purpose of being a mechanism for the training of a robust discriminator. The generator and discriminator are repeatedly trained against each other, with the former being the one producing and improving the quality of adversarial samples, and the latter having the goal of discriminating between clean and adversarial samples. Kang et al. [25] effectively utilized this method for Textual Entailment. In order to create adversarial texts for the task of textual entailment, they used three different types of generation: (1) Knowledge-based: this is a context-agnostic use of various knowledge bases in order to replace words with their hypernym/hyponym so as to generate entailing/neutral sentences; (2) Hand-crafted negations which are added to the the existing entailment; (3) Neural-based: this method trains a seq2seq for generation adversarial entailment samples, by optimizing for the cross-entropy between the original and the predicted hypothesis.

4.1.2 Model Regularization

Instead of being included in the data that trains a model, adversarial examples can also be incorporated as input perturbations during model training. In [18], Goodfellow et al. proposed the addition of perturbations in input as a regularizer in the loss function, following the form of:

$$\min(J(f(x), y) + \lambda J(f(x'), y)) \quad (4.1.1)$$

, where λ is a hyperparameter.

Including adversarial perturbations in the loss function has been proven to improve model generalization and robustness, as it leads to the generation of examples which would be difficult to manufacture otherwise.

Expanding on the work conducted on DNNs for the image domain, Miyato et al. [43] were one of the first to introduce this technique in the field of NLP, by performing adversarial generation based on word embeddings and using it at each step of model training. In order to calculate the adversarial, they used the direction in which the model distribution shows the maximum sensitivity as far as Kullback–Leibler divergence.

Another approach proposed by Yasunaga et al. [74], adversarial training was exploited through a POS tagger which applied perturbations to character-level embeddings and then word-level embeddings through concatenation. The training perturbations are chosen with the aim of maximizing the model’s loss function.

4.2 Anomaly Detection

The goal of anomaly detection is slightly different than that of the aforementioned schemes, as it involves identifying and attaining adversarial samples in between given inputs, which could then be used for purposes of interpretability or alerting the victim model, in use cases where this is beneficial. Taking into consideration the discrete nature of text, it seems reasonable to explore the possibility of directly discriminating between original and adversarial samples in NLP. On top of this, certain approaches in the bibliography attempt to also correct the identified adversarial perturbations in order to increase the robustness of the base model.

At its most basic form, this is materialized through spell-checking and grammar-checking mechanisms that can defend against less nuanced adversarial attacks. Pruthi et al. [53] developed a word recognition model that can distinguish character-level adversarial perturbations, so as to be placed before a classifier model. This was achieved by expanding on the architecture of semi-character-level RNNs [56], which acts as a spell-checker and treats misspelling as unknown words. Specifically, they proposed three back-off techniques for dealing with these unknown words: (1) Passing them as is; (2) Replacing them with neutral words; (3) Falling back to a more general word recognition model that is trained upon a larger corpus.

For detecting both word-level and char-level attacks, Zhou et al. [81] introduced DISP, which consists of three parts: a perturbation discriminator, an embedding estimator, and a token recovery mechanism, all trained on the original dataset. For every new input sample, the discriminator classifies it as either original or adversarially altered. Then, the tokens that are identified as adversarial are passed through the estimator, producing the embedding vector which is used in combination with small world graphs in order to replace the affected tokens.

Chapter 5

Methodology

In this section, we highlight the main contributions of this thesis and present our two proposed approaches for solving the task of detecting adversarial attacks in text classification; one which involves prompting Large Language Models and one which utilizes the attacked Language Model itself.

Contents

| | | |
|------------|-------------------------------------|-----------|
| 5.1 | Contributions | 66 |
| 5.2 | Datasets | 66 |
| 5.2.1 | Text Classification Datasets | 66 |
| 5.2.2 | Adversarial Attacks | 66 |
| 5.2.3 | Attacked Models | 67 |
| 5.3 | Proposed Methods | 67 |
| 5.3.1 | LLM Prompting | 67 |
| 5.3.2 | Individual Sentences Classification | 71 |

5.1 Contributions

The contributions of this dissertation are multiple and can be summarized as follows:

- We developed a simple framework for engineering prompts that enables the use of Large Language Models (LLMs) for the task of detecting adversarial attacks in text classification. This is a method that is agnostic of the attack method and the attacked model, and is tested against multiple dataset-attack-model combinations.
- We did a comprehensive analysis of LLMs’ ability at adversarial detection. To further assess our results, we conducted both a human evaluation and a sanity test for data contamination.
- We introduced a simple, yet effective way of detecting adversarial texts, utilizing the attacked language model itself. This method is also model agnostic and equally successful against multiple attacking scenarios.
- We combined our two proposed methods into a unified approach, achieving performance that competes with state-of-the-art detection methods.

5.2 Datasets

5.2.1 Text Classification Datasets

For this dissertation, we made use of 2 different datasets that are extensively used for text classification; IMDb [41] and AG-News [80].

IMDb

The IMDb dataset comprises of real-world movie reviews, which are used for the task of sentiment classification, where each text is given a label of Positive or Negative.

AG-News

The AG-news dataset is a corpus of news articles, designed for text classification. To be specific, each text consists of a news title coupled with a short description and can be classified into one of four distinct categories: World, Sports, Business, and Science/Technology.

5.2.2 Adversarial Attacks

In regards to the generation of adversarial texts, we utilized 3 of the most widely known and effective attacks in the research community; TextFooler [24], PWWS [55], and BAE [15].

TextFooler

TextFooler is a simple but effective adversarial attacking framework, which identifies important words in a given text and replaces them with semantically similar ones, until it manages to flip the attacked model’s prediction. To achieve this, it uses a gradient-free approach for ranking input words, which is based on their impact on the text’s overall classification. After identifying the most important word, it then replaces it with a synonym that is not only semantically equivalent, but also preserves the grammatical correctness of the text. This process is performed iteratively until the classification label is changed. The results achieved by TextFooler are impressive, since it reduces the classification accuracy for all tasks and target models to below 10%, while perturbing less than 20% of the input text.

PWWS (Probability Weighted Word Saliency)

The PWWS (Probability Weighted Word Saliency) attack is a synonym replacement method which considers both word saliency and change in classification probability when generating perturbations. PWWS first searches for word synonyms using Wordnet and adhering to a set of lexical constraints to make its attack imperceptible, and selects the optimal ones based on the shift they can cause to input classification. Then,

the candidate words for replacement are order based on word saliency, and the optimal one is substituted. This mechanism is repeated in a greedy manner until the overall classification flips.

BAE (BERT-based Adversarial Examples)

BAE (BERT-based Adversarial Examples) is a black-box attack that leverages BERT’s masked language modeling through three types of operations; replacement, insertion, and deletion of tokens. The adversarial perturbations are generated by masking a part of the input and using the masked language model to predict the appropriate token for filling this mask. Except from utilizing BERT’s contextual understanding, BAE ensures the preservation of semantic coherence by filtering the proposed substitutions through a sentence similarity scorer. Token alteration are repeated iteratively until misclassification is achieved.

5.2.3 Attacked Models

For our experiments, we employed 2 different text classification models to be adversarially attacked; BERT [9] and RoBERTa [38].

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model, designed to pre-train bidirectional, contextualized word representations through conditioning on both left and right context in all layers. BERT uses a masked language modeling (MLM) approach and its architecture makes it suitable for a wide variety of downstream tasks.

RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) builds on top of BERT by improving on its pretraining. This is achieved by removing BERT’s next sentence prediction objective and conducting training with much larger mini-batches and learning rates, yielding superior model performance.

5.3 Proposed Methods

5.3.1 LLM Prompting

We propose a prompt engineering framework, wherein given the target task of detecting adversarial attacks in text classification, our objective is to craft an optimized natural language prompt that maximizes a Large Language Model’s ability to perform this task.

Prompt engineering has emerged as a crucial technique for enhancing the capabilities of LLMs as well as their adaptability across diverse domains [63], [16], [48], [49], [29], [30], [4]. It involves strategically designing task-specific instructions, referred to as prompts, to guide model output without altering their parameters. For open-sourced LLMs, one can leverage their internal states or gradients to either train additional parameters, such as soft prompts [37], or to search for discrete prompts via gradient-based search [58]. However, such methods are less feasible for closed-sourced LLMs, which are becoming all the more popular and moment, therefore creating a need for studying gradient-free approaches.

The formal formulation of our problem is essentially that of gradient-free prompt optimization. Given a prompt P and a set of inputs $Q = \{q_i\}_{i=1}^n$, an LLM generates answers $a_i = LLM(q_i, P)$, $i = 1, \dots, n$ based on the probability distribution of the possible outputs and its sampling method. The goal of prompt optimization is to find the optimal prompt P^* which maximize the LLM answers’ performance towards a measure function R , such as recall. This can be formally defined as $P^* = \operatorname{argmax}_{P \in S} \sum_i R(a_i) = \operatorname{argmax}_{P \in S} \sum_i R(LLM(q_i, P))$, where S denotes the sample space of natural language prompts.

The landscape of contemporary prompt engineering spans a spectrum of techniques, encompassing foundational approaches like zero-shot prompting to more intricate ones, namely Self-Consistency [69] and Retrieval Augmented Generation (RAG) [34]. Our proposed method was inspired by a wide range of such prompting techniques found in the bibliography, and even though it was created for the purpose of adversarial detection, it is general enough that, if properly adjusted, it can be utilized for other tasks as well.

We begin with a baseline prompt P_0 , on which we perform a series of sequential refinements, until we generate the final prompt. Each refinement corresponds to one step in the framework and transforms the prompt using human-in-the-loop engineering via domain insight, as well as automatically generated changes and error mitigation. At each such step, we repeat the corresponding transformation until we have created prompts which achieve satisfactory performance metrics and then select the best-performing one, to be refined using the next technique in the sequence. Below, we analyze the different prompting techniques used in our approach.

1) Chain-of-Thought (CoT)

Introduced in [70], Chain-of-Thought (CoT) prompting enables Large Language Models to perform complex reasoning through intermediate reasoning steps. If one considers how a human approaches a complex task, such as a mathematical word problem, it is a common practice to break down the problem into intermediate steps, solving each before arriving at a final answer. The goal of CoT is to evoke such chain-of-thought reasoning from LLMs, better equipping them to reach the correct solution, for a multitude of tasks, including arithmetic (e.g., MultiArith, GSM8K), symbolic reasoning (e.g., Last Letter, Coin Flip), and commonsense reasoning (e.g., CommonSenseQA, StrategyQA). It has been demonstrated that sufficiently large Language Models are capable of generating these reasoning chains if exemplars are included in a few-shot prompt, and the success of CoT prompting was often attributed to LLMs' ability for few-shot learning. Having said that, [26] indicated the presence of untapped zero-shot reasoning capabilities in LLMs, which can also be effective zero-shot reasoners when introduced with the simple phrase "Let's think step by step" at the end of a given prompt.

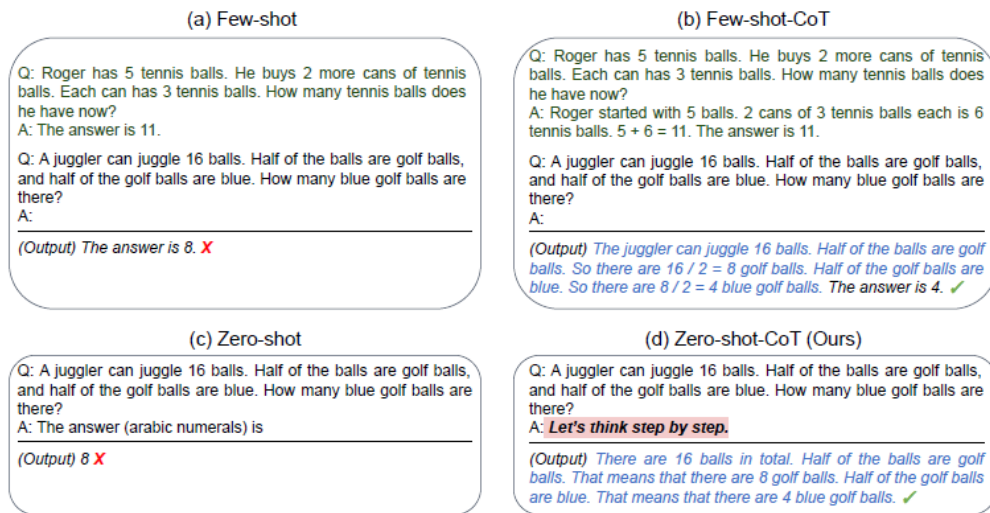


Figure 5.3.1: Example inputs and outputs of GPT-3 with (a) standard Few-shot, (b) Few-shot-CoT, (c) standard Zero-shot, and (d) Zero-shot-CoT. [26]

In our experiments, we utilized zero-shot CoT, to take advantage of multiple benefits. One of the main benefits of CoT prompting that we utilize in our approach is the fact that decomposing problems into intermediate stages, permits the allocation of additional available computational resources into solving the problem. This is corroborated by the findings that both zero-shot and few-shot CoT's effectiveness is improved by model scale. Another notable characteristic is that Chain-of-Thought offers insight into the reasoning process a model follows in order to reach a particular conclusion. Although, this does not equate to a full understanding of a model's computations that lead to an answer, it offers opportunities for identifying errors in the reasoning chain, better equipping one to resolve them through prompt engineering.

2) Expert Identity

Recent studies [52] [27] have identified that, when provided with sufficiently detailed instructions, LLMs possess the ability to function as expected agents. The expert identity to be provided to the Large Language Model should be comprehensive enough to ensure the expected behaviour, and selected in a way such that

the given role provides a distinctive reasoning advantage for each specific task, even implicitly triggering chain-of-thought reasoning. When this function is elicited through prompting, a Large Language Model's performance is enhanced, especially on specific NLP tasks. A representative such example is that when an LLM assumes the role of a judge with a clearly defined function, it can more effectively assess the quality of text summarization [72].

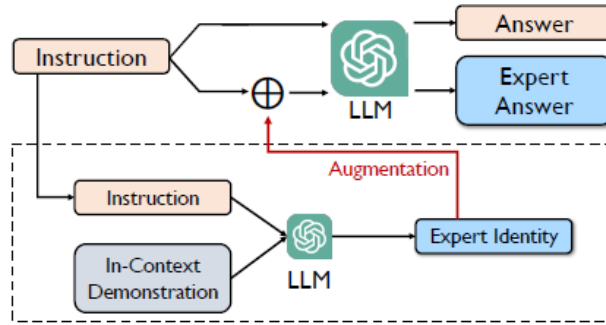


Figure 5.3.2: Expert Prompting framework. [73]

Following the simple strategy of Expert Prompting [73], we automatically generate an expert identity, tailored to the task of adversarial detection, by providing an LLM with instruction-expert pairs as few-shot examples in order to guide it into the conceptualization of a suitable identity description. This process is repeated multiple times, each time slightly tweaking the few-shot prompt, until the end result is deemed satisfactory. On top of that, we slightly edit the generated expert identity, filtering unnecessary information and emphasizing what is considered most impactful for the domain of adversarial attacks in NLP.

3) Domain Knowledge & Solution Guidance

Retrieval Augmented Generation (RAG) [34] is the process of enhancing pre-trained, parametric-memory language models, by incorporating a non-parametric memory component, such as a dense vector index of external knowledge from Wikipedia, through a general-purpose fine-tuning strategy. RAG was introduced as a solution to the challenge that more complex, knowledge-intensive tasks pose for language models and has been found to elevate their capabilities.

Inspired by this concept, we hypothesize that adding domain specific knowledge to an LLM's prompt can be beneficial for its performance. In the same way as RAG, it is reasonable to expect that enhancing a prompt with specifically curated information will improve the factual consistency and reliability, while reducing the likelihood of hallucinations for Large Language Model outputs [21]. This belief is only enhanced when considering the niche nature of the topics of adversarial attacks and detection in NLP, which have not been excessively studied even in the bibliography. Therefore, we augment our prompt with carefully selected information which an LLM can employ towards its task of detecting adversarial attacks.

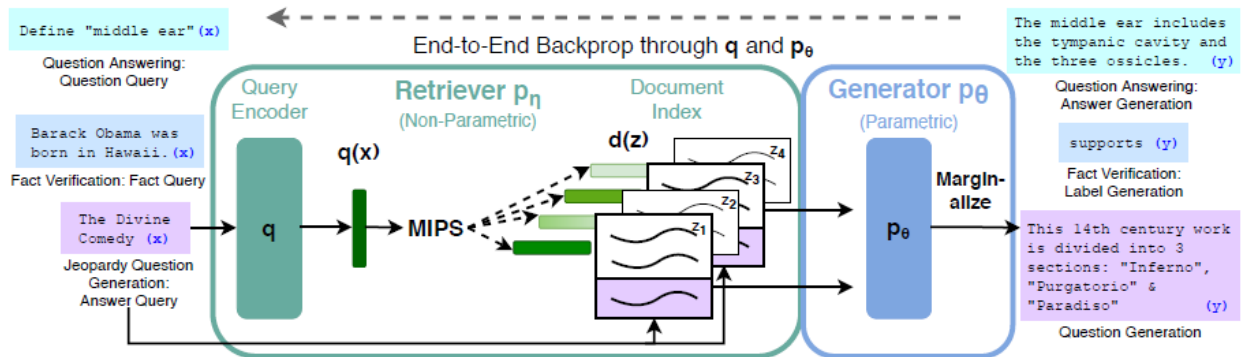


Figure 5.3.3: Retrieval Augmented Generation overview. [34]

By analyzing the chain-of-thought in LLM answers, it has been concluded [70] that when the generated result is accurate, then the chain-of-thought reasoning is generally correct and logical. In regards to incorrect answers, there is a portion of them which have been created through a substantially incorrect chain-of-thought. Nonetheless, there is another portion which could be corrected if three primary areas of mistakes were resolved, namely calculation errors, missing steps, and semantic misunderstandings. These challenges can be mitigated by guiding the Large Language Model into understanding the given task, dividing it into small sub-tasks and solving those step by step [66].

Based on these insights, we enhance the domain knowledge contained in our prompt with a solution guidance component. This includes adding information on the potential signs of textual adversarial attacks which consists of providing information on adversarial attacks and encouraging the LLM to focus on specific steps, based on this information, when examining whether a text is adversarial or not. The reason for doing this is to improve the quality and reliability of the model’s chain-of-thought and increase its attention to detail. Still, because the task of adversarial detection does not have a predefined solution, we aim to not overly restrict the LLM into a rigorous set of rules for it to reason through. We do this in an effort to allow model reasoning that might be more complex or not as intuitive for humans, but can still lead to correct predictions.

4) Error Correction

Drawing inspiration from human trial-and-error processes as well as the self-reflection ability of LLMs [59], PromptAgent [68] introduced a principled planning approach, to strategically optimize the prompting process. Within this framework, the prompt optimization problem is formulated as a Markov Decision Process (S, A, T, r) , where S denotes the state space of different prompt version, A is the action space of potential alterations to the current prompt, T defines the transition function $T : S \times A \mapsto S$ and $r : S \times A \mapsto R$ is the reward function. Figure 5.3.4 showcases the two-step process of action generation, which is the collection of errors that the base model makes on certain training samples and the reflection upon those errors by the optimizer model in order to draw useful error feedback. Using this error feedback as an action, the optimizer generates a new prompt which adds domain knowledge and addresses the given errors. Finally, the reward is straightforwardly defined as the task performance on a held-out set of samples.

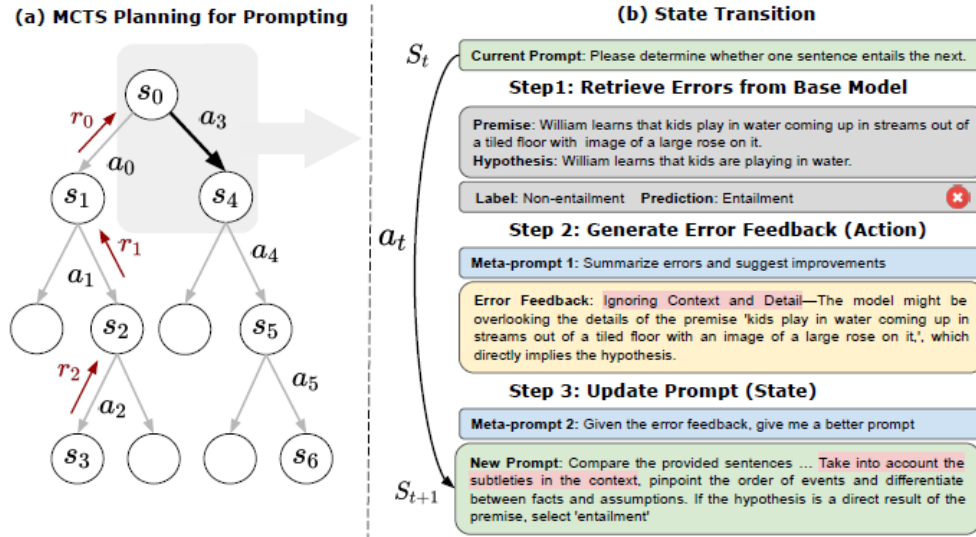


Figure 5.3.4: (a) Monte Carlo Tree Search (MCTS) planning for automatic prompt generation. (b) A state transition example. Given a current state (prompt), the base model collects errors from the task dataset.

The optimizer model provides corresponding error feedback. Then, the optimizer model updates the prompt based the feedback and transits to the next state. [68]

This reformulation facilitates the integration with the Monte Carlo Tree Search (MCTS) algorithm, for an effective navigation of the extensive prompt space and the identification of high-reward paths from error feedbacks and therefore expert-level prompts. MCTS supports strategic planning, as in Figure [ref],

by incrementally building a tree structure where each node represents a state and each edge denotes the action leading to state transitions. MCTS expands the tree through the management of a state-action value function, $Q : S \times A \mapsto R$, forecasting the potential future rewards associated with applying an action to a state, for paths extending from the current state-action pair. To update this Q function and expand the tree, MCTS iteratively executes four key operations: selection, expansion, simulation, and back-propagation. The process continues until a pre-determined number of iterations is completed, after which a trace and node are chosen as the best ones for the final prompt.

Building on the insights about the types of mistakes that LLMs make, we utilize this principled planning approach as a way of enhancing our prompt with insightful error feedback which can combat the effects of the most common LLM reasoning pitfalls.

Answer Extraction

In zero-shot prompting, the structure of answers generated by an LLM may not always be consistent or the expected, as it can vary depending on the input provided. Even when given explicit instructions for how the answer should be formatted, there can be deviations, particularly for smaller models which have a reduced ability to consistently follow such directives. For that reason, it is necessary to prompt an LLM a second time, so as to extract the final answer from its initial response. This is done by concatenating three elements: the initial prompt, the LLM’s initial output, and an answer extraction prompt. This combined prompt refines the model’s response, leading to the generation of an answer with the desired format.

5.3.2 Individual Sentences Classification

Taking into consideration the discrete nature of text, we understand that any effective adversarial attack that aims to be imperceptible, generates edits only in specific parts of a given text. This means that even after it undergoes adversarial perturbations, a large part of original text remains unaltered and therefore its internal representation should also remain unaltered.

In regards to the task of text classification, we aim to capitalize on this by breaking a given text into individual sentences and analyzing each one separately. We believe that for unaltered texts, the majority of its sentences will be given the same classification as the overall text. On the contrary, if an adversarial text is divided in sentences, only the classifications of sentences which contain alteration might be affected. Therefore, the unperturbed sentences and even some perturbed ones will retain their initial label which most probably agrees with the original text label. Since the classification of the entire text has been flipped by the attack, this inevitably means that the label of a lot of independent sentences will be opposite to the overall one, in adversarial samples.

Based on this insight, we create a simple, yet effective, method which inspects the classifications given by the attacked language model to each individual sentence of a text, and compares them with the classification given to the entire text, with the goal of extracting insights to be utilized for identifying adversarial samples. We named this method Individual Sentences Classification (IndivSC) and developed 3 approaches to it.

Total Agreements

In this approach, we consider how many of the sentences in a given text have the same classification as the one given to the entire text. Total Agreements represents the amount of individual sentence classifications which agree with the overall classification, as a fraction of the text’s sentence count.

An illustrative example is given in Figure 5.3.5. The original text is a movie review classified with a negative sentiment. Being comprised of 8 sentences, the sentiment of each individual sentence is also negative, agreeing with the overall classification and giving a Total Agreements percentage of 100%. In the case of the adversarial text, we notice that the attack used was particularly intrusive, performing perturbations to words in every single sentence and thereby flipping the text’s overall sentiment to positive. Nevertheless, when analyzing each individual sentence, we see that only for half of them did the classification also switch to positive, leading to a Total Agreements percentage of 50

| |
|---|
| Original Text: |
| 1980 was certainly a year for bad backwoods slasher movies. "Friday The 13th" and "The Burning" may have been the best ones but there were like always a couple of stinkers not far behind like "Don't Go Into The Woods Alone" and this one. But in all fairness "The Prey" is nowhere near as bad as "Don't Go Into The Woods" but it's still not great either. One thing is that it's just boring and acting isn't very good but much better than "DGIW" and this movie actually has some attractive looking females to look at, all three of the female leads were stunning. One thing what is up with all that pointless wildlife footage it just seemed pointless and it looked as the director used that to just used that to fill up some time space. So, what was there to like about this movie? Well, there were a few laugh out loud cheese moments- I couldn't contain a fit of giggles when the final girl did a bizarre type of backwards moon-walk to get away from the killer and there were a few good kill scenes- my favourites being the girl suffocated to death with the sleeping bag; and the phoneyness looking. All in all The Prey is dumb, boring and the killer. I didn't find scary at all, this movie could have been a whole lot better. |
| Classification of entire text: Negative |
| Total Agreements: 8/8 = 100% |
| Adversarial Text: |
| 1980 was certainly a [[years]] for [[unfavourable]] backwoods slasher movies. "Friday The 13th" and "The Burning" may [[possesses]] been the best ones but there were like always a couple of stinkers not [[importantly]] behind [[enjoy]] "Don't Go Into The Woods Alone" and this one. But in all fairness "The Prey" is nowhere near as bad as "Don't Go Into The Woods" but it's [[alternately]] not great either. One [[ding]] is that it's just [[uninspired]] and [[behaving]] isn't very good but much better than "DGIW" and this movie [[efficiently]] [[enjoys]] some attractive looking [[daughters]] to look at, all three of the female leads were stunning. One thing what is up with all that [[unsuccessful]] wildlife footage it just seemed pointless and it [[explored]] as the director used that to just used that to [[complete]] up some time space. So, what was there to [[genera]] about this movie? Well, there were a few laugh out loud cheese moments- I couldn't contain a fit of giggles when the final girl did a bizarre type of backwards moon-walk to get away from the killer and there were a [[unusual]] good kill scenes- my favourites being the girl suffocated to death with the sleeping bag; and the phoneyness looking. [[Allen]] in all [[Both]] Prey is [[punchy]], [[diameter]] and the [[killing]] didn't [[looking]] [[terrible]] at all, this [[film]] could have been a whole [[multiple]] better. |
| Classification of entire text: Positive |
| Total Agreements: 4/8 = 50% |

Figure 5.3.5: Example of Total IndivSC Agreements. Double brackets denote words perturbed by the adversarial attack.

We believe that this difference between original and adversarial texts, in regards to the classifications of their individual sentences, shows up in a statistically significant manner, which will allow us to develop efficient detection methods utilizing it.

Longest Continuous Agreement Sequence

This next approach entails finding the longest continuous sequence of sentences in which the classifications of each individual sentence agrees with the classification of the entire text. This length of this sequence is the Longest Continuous Agreement Sequence and is calculated as a fraction of the total text length, counted in sentences.

In Figure 5.3.6, we see an example in which the length of the Longest Continuous Agreement Sequence goes from 80% of the total length of the original text, to only 20% of the adversarial text's length.

Total Interchanges

In the final approach, we consider the amount of interchanges that occur in terms of the classifications given to each sentence in the sequence of the overall text. Total Interchanges is equal to the amount of times the classification changes when going from one sentence to the next, as a fraction of the text's total sentences.

In the example of Figure 5.3.6, we observe that while in the original text, the Total Interchanges percentage is 22%, in the case of the adversarial text it doubles to 44%.

| | |
|--|------------|
| Original Text: | |
| Has there ever been an Angel of Death like MIMSY FARMER in Barbet Schroeder's 1960s heroin opus? Sort of Jean Seberg with a hypodermic Pink Floyd score. Despite some ultimately insignificant weaknesses, a classic, shamelessly ripped off by Erich Segal/Noel Black for their inept JENNIFER ON MY MIND (1971), although Tippy Walker, playing a similar character, is herself very junkie-appealing in the latter mess. MORE, though, is terrific, a great 60s drug movie and, simply, an important document of its time. Very much a cult film so join the cult. No American movie then, as far as I can remember, charts the same territory. MIMSY's an astonishing archetype, elevating this into mythic realms. Not for the faint-hearted. Great sex scenes too. | |
| Classification of entire text: | Positive |
| Total Agreements: | 9/10 = 90% |
| Longest Continuous Agreement Sequence: | 8/10 = 80% |
| Total Interchanges: | 2/9 = 22% |
| Adversarial Text: | |
| Has there ever been an Angel of Death like MIMSY FARMER in Barbet Schroeder's 1960s heroin opus? Sort of Jean Seberg with a hypodermic Pink Floyd score. Despite some ultimately insignificant weaknesses, a classic, shamelessly ripped off by Erich Segal/Noel Black for their inept JENNIFER ON MY MIND (1971), although Tippy Walker, playing a similar character, is herself very junkie-appealing in the latter mess. [[INCREASED]], [[whereas]], is [[unbelievable]], a great 60s drug [[cinematographers]] and, [[scarcely]], an [[grands]] [[material]] of its time. [[Awfully]] [[largely]] a cult film so join the cult. [[Never]] American movie then, as far as I can remember, charts the same territory. MIMSY's an [[phenomenal]] archetype, elevating this into mythic realms. Not for the faint-hearted. Great sex scenes too. | |
| Classification of entire text: | Negative |
| Total Agreements: | 3/10 = 30% |
| Longest Continuous Agreement Sequence: | 2/10 = 20% |
| Total Interchanges: | 4/9 = 44% |

Figure 5.3.6: Example of the 3 IndivSC approaches. Double brackets denote words perturbed by the adversarial attack.

Chapter 6

Experiments and Results

In this section, we present the various experiments conducted in order to investigate the detection of adversarial attacks in text classification. We describe the experimental setups that we made use of and analyze all of the results stemming from them. A particular emphasis is given to explore all of the results not only quantitatively, but also qualitatively, in order to gain a better understanding of our methodologies' performance.

Contents

| | |
|---|------------|
| 6.1 Preliminaries | 76 |
| 6.1.1 Datasets | 76 |
| 6.1.2 Evaluation Metrics | 77 |
| 6.2 Experiments of LLM Prompting | 79 |
| 6.2.1 Employed LLMs | 79 |
| 6.2.2 Prompt Engineering | 80 |
| 6.2.3 LLM Prompting | 84 |
| 6.2.4 Human Evaluation | 90 |
| 6.2.5 Sanity Test for Data Contamination | 92 |
| 6.3 Experiments of Individual Sentences Classification | 94 |
| 6.3.1 Total Agreements | 94 |
| 6.3.2 Longest Continuous Agreement Sequence | 98 |
| 6.3.3 Total Interchanges | 101 |
| 6.4 Combination of Methods | 103 |
| 6.5 Comparison with State-of-the-art Detection Methods | 105 |
| 6.5.1 Compared Methods | 105 |
| 6.5.2 Comparison Results | 105 |

6.1 Preliminaries

6.1.1 Datasets

Attack Framework

For our experiments, all of the final datasets were created using the TextAttack framework [44] on the original text classification datasets.

TextAttack is a Python framework designed for adversarial attacks, data augmentation, and adversarial training in NLP. The framework integrates various adversarial attack methods by decomposing NLP attacks into four core components: a goal function, a set of constraints, a transformation, and a search method. The attack process involves perturbing input text so that the model’s output meets the goal function criteria, which could include targeted classification or minimum BLEU score and which determines whether the attack is successful or not. At the same time, the perturbation needs to adhere to specified linguistic constraints, namely grammar and semantic similarity to the original input. In order to successfully generate an adversarial example, given an NLP model, a transformations component generates a sequence of possible perturbations, such as word embedding and thesaurus word swap. Ultimately, in order for the best transformations to be selected, the attacked model is iteratively queried, through a search method like greedy search with word importance ranking or beam search.

| Attacks | Attack Accuracy | Average Number of Queries |
|------------------------|-----------------|---------------------------|
| IMDB (91.9%) | | |
| TextFooler | 99% | 558 |
| PWWS | 97% | 1681 |
| BAE | 64% | 455 |
| AG-News (94.2%) | | |
| TextFooler | 81% | 334 |
| PWWS | 57% | 362 |
| BAE | 14% | 122 |

Table 6.1: Adversarial attack statistics for each method against the BERT model. Original accuracy is noted within a parenthesis next to the dataset.

For the IMDb dataset, the number of generated adversarial texts per adversarial attack and attacked model was 10000, while for the AG-News dataset the number was 7600. Table 6.1 summarizes the attack statistics on the two datasets and three attack methods.

Sampling of Adversarial Examples

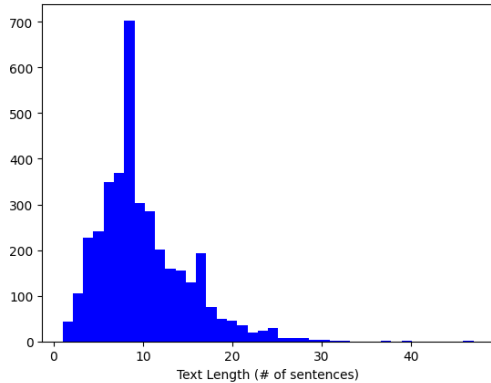
For the purpose of detecting adversarial examples, a dataset D needs to consist of clean samples D_{clean} and adversarial samples D_{adv} . However, the way in which these two subsets have been sampled is an implementation detail that influences the composition of the final dataset. This configuration varies from work to work, and here we discuss two of the main ones used in the literature [76]. We denote the test set as X_t .

Scenario 1: Sample disjoint subsets $S_1, S_2 \subseteq X_t$. For the correctly classified examples of S_1 , generate adversarial attacks and form D_{adv} from the successful adversarial examples. Form D_{clean} from S_2 .

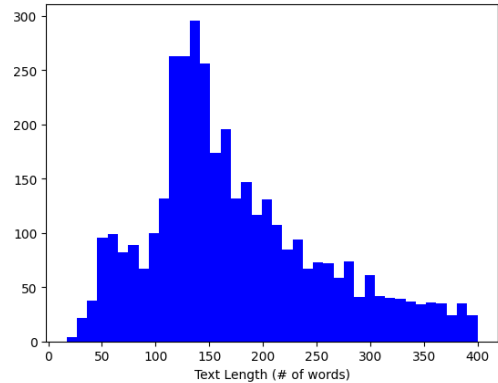
Scenario 2: Sample subset $S \subseteq X_t$. For the correctly classified examples of S , generate adversarial attacks and form D_{adv} from the successful adversarial examples. Form D_{clean} from S as well.

Scenario 1 allows for a more accurate selection of the ratio between adversarial samples and clean samples, whereas in Scenario 2 this is mostly determined by the task accuracy and attack success rate. For example, an adversarial attack with a low success rate will inevitably lead to a dataset with reduced adversarial-to-clean sample ratio. Additionally, in Scenario 2, the dataset contains pairs in which the original sample is in D_{clean} and its corresponding adversarial sample is in D_{adv} , which might not be desirable, based on the use case.

For each combination of the 2 text classification datasets (IMDb, AG-News), 3 textual adversarial attack methods (TextFooler, PWWS, BAE), and 2 attacked models (BERT, RoBERTa), we chose the sampling scenario 1 in order to populate the dataset with a wider distribution of text examples and to be able to set the specific maximum number of adversarial examples to be sampled at 2000. The size of the subset S_1 was calculated as the maximum number of adversarial examples to be sampled divided by adversarial success ratio and task accuracy, while the size of the subset S_2 was calculated in a way such that the ratio between original and adversarial samples was approximately 1:1.

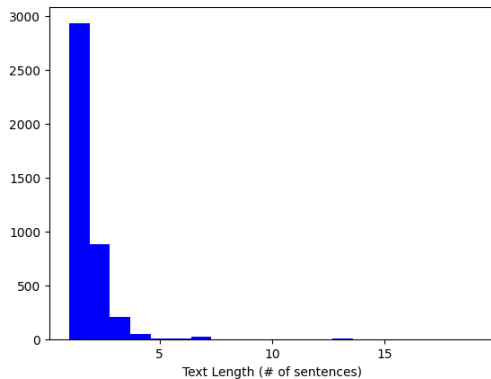


(a) Number of sentences.

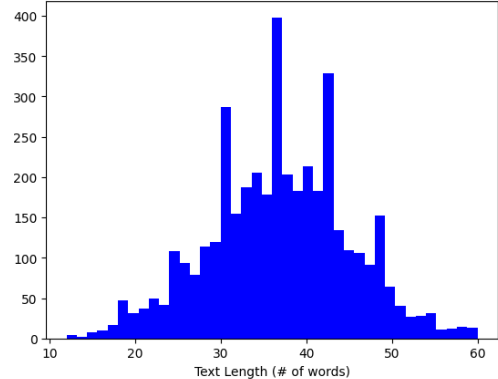


(b) Number of words.

Figure 6.1.1: Length of IMDb texts.



(a) Number of sentences.



(b) Number of words.

Figure 6.1.2: Length of AG-News texts.

6.1.2 Evaluation Metrics

Recall

Recall is the fraction of relevant instances that were correctly retrieved. Written as a formula:

$$Recall = \frac{TP}{P}$$

where, in our case, TP (True Positives) are the texts correctly identified as adversarial and P (Positives) are all the adversarial samples.

Since we are concerned with the task of adversarial attacks detection, the most important goal is to minimize the amount of adversarial texts which are not detected by our method. As a consequence, we want to maximize Recall and that is the reason why it is the metric we primarily optimize for.

F1 Score

F1 score is the harmonic mean of Precision and Recall, formally defined as:

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

where FP (False Positives) are the texts incorrectly identified as adversarial and FN (False Negatives) are the adversarial samples which were not detected.

We incorporate F1 score as an evaluation metric because it provides a more balanced evaluation and ensures that neither false positives nor false negatives dominate our approaches' assessment.

Accuracy

Accuracy is the proportion of all classifications that were correct, whether positive or negative. Mathematically, it is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TN (True Negatives) are the texts correctly identified as original.

Accuracy, one of the most widely utilized metrics in AI, is used to enhanced the comprehensiveness of our evaluation.

ROC Curve and AUC

An ROC (Receiver Operating Characteristic) curve is a visual representation of a classifier's performance at varying threshold values. Formally, it is the plot of the TPR against the FPR at each threshold setting, where TPR (True Positive Rate) is the probability that an adversarial text is detected as such and FPR (False Positive Rate) is the probability that an original text is wrongly identified as adversarial. AUC stands for Area Under Curve and represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative. A greater value of AUC generally denotes better model performance, with the ROC curve of the optimal classifier being a square with sides of length 1 and an AUC of 1.0.

We employ the ROC curve and AUC for the evaluation of our Individual Sentences Classification methods and in order to select the best possible detection threshold for them, which in the ideal scenario would yield a point as close to the the upper left point (0,1), the perfect classification.

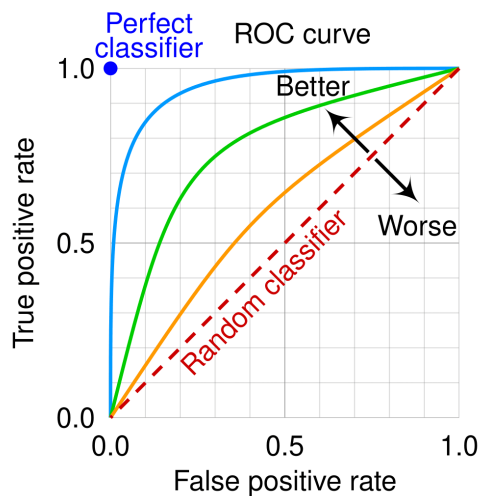


Figure 6.1.3: ROC Curve example. [71]

6.2 Experiments of LLM Prompting

6.2.1 Employed LLMs

Phi-3-mini

Phi-3-mini is a 3.8 billion parameter transformer-based language model created by Microsoft [1]. It was trained on 3.3 trillion tokens, including educational content, NLP synthetic texts, and chat datasets from various sources. These diverse training data, combined with supervised fine-tuning (SFT) and Direct Preference Optimization (DPO), have enhanced the model’s capabilities as far as alignment, robustness, and safety. The model’s transformer decoder architecture includes 32 layers, 32 heads, and a 3072 hidden dimension and uses the same tokenizer and block structure as Llama-2. Despite its smaller size compared to other LLMs, Phi-3-mini achieves competitive performance. It scores 69% on the MMLU (Massive Multitask Language Understanding) benchmark and 8.38 on the MT-bench, demonstrating a high level of language understanding and reasoning. Nevertheless, the model has limitations in storing extensive factual knowledge, which can affect its performance in tasks requiring large-scale information retrieval.

We employed the instruct-tuned variant of the model with a context length limit of 4k tokens.

Llama 3.1 8B

The Llama 3.1 8B model is an autoregressive language model, part of the Meta Llama 3.1 collection [42]. Built on an optimized transformer architecture, the model’s training was performed on an extensive dataset consisting of approximately 15 trillion tokens from publicly available sources, and required 1.46 million GPU hours on a custom-built GPU cluster. The instruct version of the model was fine-tuned on publicly available instruction datasets and over 25 million synthetically generated examples, incorporating both supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align the model with human preferences. Llama 3.1 8B uses Grouped-Query Attention (GQA), supporting a context length of 128K tokens, and exhibits state-of-the-art performance, which surpasses many existing models on industry benchmarks. For our experiments, we made use of the instruct-tuned model, applying 4-bit quantization.

Gemma 2

Gemma 2 is the newest addition to the Gemma family of lightweight, state-of-the-art open models, ranging in scale, with models at 2 billion, 9 billion, and 27 billion parameters [62]. Gemma 2 models employ a decoder-only transformer architecture, retaining elements similar to their first version, namely Rotary Position Embeddings (RoPE) and approximated GeGLU non-linearity, as well as introducing new elements, including alternation between local sliding window attention and global attention layers. The context window has a length of 8192 tokens and Grouped-Query Attention (GQA) is utilized to optimize inference speed. Regarding training, Gemma 2 27B was trained from scratch on 13 trillion tokens of mainly English data. The 9B and 2B models were trained on 8 trillion and 2 trillion tokens respectively, utilizing knowledge distillation with guidance from the larger model. After initial training, the models underwent instruction-tuning through supervised fine-tuning (SFT) and subsequently, reinforcement learning with human feedback (RLHF), further refining model responses. The evaluation of the Gemma 2 27B model shows that it outperforms models of similar size and is competitive with much larger models. Additionally, the 2B and 9B models exhibit significant performance improvements compared to previous versions, particularly on benchmarks evaluating commonsense reasoning, mathematics, coding, and question answering.

We leveraged the instruct fine-tuned variants of the models, applying 4-bit quantization to the 9B model and employing the 2B and 27B models at full precision.

Claude 3 Haiku

Claude 3 Haiku, part of the Claude 3 family of models, is one of the fastest and most affordable Large Language Models in its intelligence category [3]. Claude 3 models are trained on a mixture of publicly available data from the internet, as well as private third-party data, all carefully filtered and curated through techniques involving deduplication and classification. Their training includes various methods like word prediction and human feedback to elicit helpful answers, as well as a methodology introduced by Anthropic, Constitutional AI. This technique is performed during reinforcement learning (RL) and it involves aligning

the model with a set of ethical principles and rules, therefore aligning it with core human values. Having undergone rigorous evaluations, Claude 3 Haiku has showcased state-of-the-art vision capabilities and strong performance on industry benchmarks regarding reasoning and language understanding. We utilized Claude 3 Haiku through Anthropic’s API.

GPT-4o mini

GPT-4o mini is a highly efficient and cost-efficient model introduced by OpenAI [47]. With a context window of 128K tokens and support for up to 16K output tokens, GPT-4o mini shares its tokenizer with GPT-4o and works with text and image inputs and outputs. The model’s robustness is underpinned by a combination of mechanisms throughout its development. During pre-training, undesirable inputs are filtered out and during post-training, RLHF was utilized to align the model with safety policies which increase response reliability and accuracy. Performance-wise, GPT-4o mini stands out in academic benchmarks across textual intelligence and multimodal reasoning tasks, achieving an impressive 82% score on MMLU (Massive Multitask Language Understanding) and 59.4% on MMMU (Massive Multi-discipline Multimodal Understanding). This surpasses the performance of other contemporary models such as Gemini Flash and Claude Haiku, proving the model’s efficacy.

We made use of GPT-4o mini through OpenAI’s API.

6.2.2 Prompt Engineering

For our prompt engineering experiments, we employ Phi-3-mini and Gemma-2-9B. By analyzing the performance of Phi-3-mini, which is one of the most potent small-scale open source LLMs, we will gain an understanding of how models of similar size can cope at the task of adversarial attack detection. Gemma-2-9B was chosen as one of the best performing mid-sized open source LLMs, the results of which served as the guideline for the selection of the optimal final prompt. In addition, the comparison with its smaller and larger model versions will provide clear information about the extent to which model scale effects our task. The experiments were conducted on a held-out set of 250 samples generated via TextFooler on the IMDb dataset, which due to the increased length of its texts, is characterized by higher language variation.

Answer Extraction

In our experimental setup, we incorporated an extra prompting step in order to extract the final answer from the LLM’s generated output. This was achieved by concatenating the original prompt, the generated LLM answer and an instruction to extract the answer in a multiple choice format. The newly formed prompt was used as a new input for the LLM, which in turn produced the final prediction.

0) Baseline

As a start, we conducted experiments using two versions of a baseline prompt which plainly state the task at hand, in order to evaluate the capabilities of Large Language Models on the task of detecting adversarial texts, without any particular enhancements in its prompt. This baseline serves as a reference point, enabling us to more better gauge the improvements achieved through each step of our proposed method.

The two versions of the baseline prompt which were evaluated, along with every other used prompt, aim to follow the best practices in prompt engineering. These included the use of clear, declarative language, as well as the application of the Rephrase and Respond framework [7], in which the prompt is refined by instructing the LLM to rephrase the question it was posed with in a way that facilitates a more accurate response.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.35 | 0.41 | 0.50 |
| | v2 | 0.36 | 0.43 | 0.50 |
| Gemma-2-9B | v1 | 0.20 | 0.32 | 0.58 |
| | v2 | 0.24 | 0.37 | 0.60 |

Table 6.2: Adversarial detection results for Baseline Prompt.

As it is evident by the results of Table 6.2, the LLMs do not appear to have an inherent competence at detecting textual adversarial attacks, essentially performing worse than random.

Even though Phi-3-mini achieves higher Recall and F1 metrics than Gemma-2-9B, based on the fact that its Accuracy is 50% for both versions of the prompt, it can be deduced that its predictions are arbitrary. This can be confirmed by manually inspecting Phi-3-mini predictions, which contain a lot of False Positives. The low performance of Gemma-2-9B, on the other hand, seems to stem not from random predictions, but its inability to detect adversarial samples. A Recall of 20-24% coupled with an Accuracy of above 50% leads to the conclusion that the classifications are probably comprised of very few True Positives. This is consistent with the model’s actual predictions, which, parenthetically, are in close to total agreement for both versions of the baseline prompt, barring the additional adversarial samples which were correctly identified using the slightly better-performing prompt version 2.

Overall, these results highlight the need for prompt engineering for the task of adversarial detection, at least for Large Language models of these scales.

1) Chain-of-Thought (CoT)

When it pertains to Chain-of-Thought, we utilized the zero-shot approach to induce sequential reasoning from the models, by appending the phrase "Let’s think step by step" at the beginning of the LLM’s response.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.05 | 0.09 | 0.50 |
| | v2 | 0.02 | 0.05 | 0.50 |
| Gemma-2-9B | v1 | 0.23 | 0.36 | 0.58 |
| | v2 | 0.43 | 0.54 | 0.63 |

Table 6.3: Adversarial detection results for CoT Prompt.

Table 6.3 contains the metrics that result from adding zero-shot CoT to both versions of the baseline prompt and its findings substantiate the LLMs’ lack of competence at our task without further guidance, even when utilizing CoT.

Phi-3-mini’s performance at detecting adversarial texts drops to almost 0, confirming our interpretation of its baseline prompt predictions as mostly random. This appears to be a consequence of the fact that its chain-of-thought is a mere surface analysis of the given text and therefore does not induce elaborate enough reasoning. That being said, Gemma-2-9B displays a significant increase in performance, particularly when using the second prompt version, with its Recall increasing by 19%. This contrast can be attributed to Gemma-2-9B being larger in size and therefore having higher capacity for reasoning. By examining its produced reasoning, we can see that it is indeed becoming able to detect several adversarial texts by noticing certain glaring errors within them.

2) Expert Identity

To establish our expert identity, we manually curated three expert descriptions, which were prepended as few-shot examples into the generation prompt. These exemplars were deliberately designed to be sufficiently descriptive and tailored to each specific task, while avoiding excessive verbosity. The model employed to craft the expert identity was GPT-4, which we prompted 5 times in total, on each iteration refining the few-shot examples until the resulting identity was deemed satisfactory and without redundant information. Finally, minor adjustments were made to the generated output to further refine it, for example ensuring that it incorporates the phrasing of the second baseline prompt version which led Gemma-2-9B to the highest adversarial text detection metrics in the previous section.

Table 6.4 contains the results obtained from prompting the models either exclusively using the Expert Identity or by combining it with Chain-of-Thought.

For both LLMs, the metrics following the sole use of the Expert Identity framework are similar to those achieved when solely using CoT. However, when utilizing both methods at once, the performances achieved

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | no CoT | 0.04 | 0.08 | 0.51 |
| | with CoT | 0.06 | 0.10 | 0.52 |
| Gemma-2-9B | no CoT | 0.42 | 0.58 | 0.69 |
| | with CoT | 0.57 | 0.65 | 0.68 |

Table 6.4: Adversarial detection results for Expert Identity Prompt.

are the best so far. In the case of Phi-3-mini, performance is improved by a few percentage points but the model’s is still essentially unable to achieve adversarial text detection. Meanwhile, Gemma-2-9B showcases another significant jump in performance, with its Recall and F1 score reaching adequate values at 57% and 65% respectively.

It is interesting to observe that the predictions produced by exclusively using the Expert Identity in the prompt do not include any reasoning. Despite the notable potency of this setting, the fact that the best results were achieved by enhancing it with Chain-of-Thought establishes both methods as essential components of our prompt engineering framework.

3) Domain Knowledge & Solution Guidance

In order to elevate the models’ capabilities in detecting adversarial texts, we added domain knowledge to the first version of this section’s prompt. We did this by giving a definition of what consists an adversarial perturbation and restating the LLM’s task which is to detect such perturbations.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.10 | 0.18 | 0.54 |
| | v2 | 0.24 | 0.38 | 0.59 |
| | v3 | 0.62 | 0.69 | 0.71 |
| Gemma-2-9B | v1 | 0.82 | 0.72 | 0.68 |
| | v2 | 0.74 | 0.71 | 0.70 |
| | v3 | 0.80 | 0.76 | 0.74 |

Table 6.5: Adversarial detection results for Domain Knowledge & Solution Guidance Prompt.

From Table 6.5, we can see that this addition seems to boost the detection abilities of both models. Having said that, we notice that in the case of Gemma-2-9B, Recall has disproportionately increased by 25%, while Accuracy has stayed the same. This effectively means that the model is identifying more adversarial examples, but this is done at the cost of creating too many false positives. Diving deeper, the CoT generated by the LLM confirm this finding, as they show that the model is overtly sensitive and seems inclined to quickly classify a text as adversarial.

In order to mitigate this, we enhance version 2 of the prompt, by guiding the Large Language Model to consider all of the given information, decompose the given task into intermediate steps and consider any irregularities it detects in context, before reaching a final prediction. As shown by the experimental results, the new version of the prompt solves the aforementioned problem, striking a balance between Gemma-2-9B’s Recall, F1 score, and Accuracy. The fact that the Phi-3-mini also shows a significant jump in performance when prompted this way, highlights both the importance of constructive CoT for models to make correct predictions and that our prompt enhancements are engineered towards the right direction.

Finally, in version 3 of the prompt, we guide the LLM with specific solution steps for it to follow. These guidelines were firstly inspired by the way in which the LLMs themselves approach the task of adversarial detection, even prior to receiving any specific guidance. We observed that within their reasoning, they not only inspect a given text for obvious errors such as grammatical mistakes and incorrect word usage, but also analyze the semantics and overall context of the text. On top of that, our steps are in alignment with the way in which other methods in the bibliography approach the detection adversarial attacks in NLP. That is,

encouraging the LLM to look out for any abnormal characteristics within a given sample, that set it outside the usual characteristics of unperturbed text.

The final results of Phi-3-mini show immense improvement, with Recall rising to 62%, which is a positive indication that through curated prompting, there can emerge adversarial detection abilities for even more lightweight models. For Gemma-2-9B, we observe an increase in Recall.

4) Error Correction

Through the error correction part of our prompt, we aimed to resolve the most common errors that LLMs made within their reasoning. Following the PromptAgent framework, we used our selected LLMs as base models and GPT-3.5-Turbo as the optimizer model. For reasons of simplicity and computational efficiency, we used Greedy Search, in the place of MCTS, wherein during the expansion phase we generate 3 different error feedback actions, based on batches of 3 training samples. The optimal prompt transition at each depth is defined based on task performance. The number of iterations performed is 3, leading to the generation of 9 prompts in total, from which we had to finally select the optimal state/prompt. To this end, we selected the three which exhibited the greatest task performance and slightly refined them to increase language specificity and reduce verbosity.

| LLM | Prompt Version | Recall | F1 | Accuracy |
|------------|----------------|--------|------|----------|
| Phi-3-mini | v1 | 0.63 | 0.63 | 0.66 |
| | v2 | 0.64 | 0.65 | 0.68 |
| | v3 | 0.67 | 0.64 | 0.66 |
| Gemma-2-9B | v1 | 0.80 | 0.75 | 0.75 |
| | v2 | 0.77 | 0.72 | 0.73 |
| | v3 | 0.75 | 0.72 | 0.74 |

Table 6.6: Adversarial detection results for Error Correction.

From the produced error feedback, we can deduce that most common LLM mistakes caught by the optimizer fall into one of three distinctive categories: cases where the LLM detected signs which raised suspicion of adversarial manipulation but ultimately classified the text as original, cases where the LLM did not spot obvious errors within an adversarial text, and cases where the LLM falsely categorized a text as adversarial due to reasoning which did not indicate an attack.

Despite the fact that these error feedback are logically accurate, the results of Table 6.6 show that they don't lead to particular improvements in model performance. Mainly, we see unstructured fluctuations across all metrics which highlight the LLMs' sensitivity to changes in their input and indicates that there is no more room for prompt optimization, at least for these models.

For Phi-3-mini, the best performing prompt version is the third one, which is intriguing due to the fact that its differentiation is the use of more detailed language to describe the solution guidance steps. For Gemma-2-9B, the final prompt version 1 is the optimal and it introduces no significant changes. A detailed catalog of the prompts employed throughout our experiments can be found in Appendix A.1.1.

During our following experiments, the prompt used for the rest of the LLMs is the same as the one engineered for Gemma-2-9B, apart from Phi-3-mini which used its own engineered one. This is done due to computational restrictions, but most importantly, to allow for comparison in regards to model scale in the case of the Gemma-2 family of models and for analysis of how well our prompt can generalize for different models. It also has to be noted that during undocumented experimentation on the AG-News dataset, we noticed that for LLMs with above 8 billion parameters, a significant amount of false positives was being generated due to the identification of HTML formatting as adversarial. HTML is embedded within AG-News data and contains unusual symbols which trigger LLMs. For that reason, we included an additional instruction in their prompt, telling them to ignore such symbols.

6.2.3 LLM Prompting

Quantitative Results

In Tables 6.7, 6.8, 6.9, 6.10, and 6.11, we present the results that occur by prompting all of the employed LLMs for the task of detecting adversarial attacks in text classification. For every dataset-attack-model combination, the number of samples used is 250.

| Attacked Models | Adversarial Attacks | | | | | | | | |
|-----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| IMDB | | | | | | | | | |
| BERT | 0.65 | 0.69 | 0.71 | 0.67 | 0.75 | 0.75 | 0.43 | 0.50 | 0.59 |
| RoBERTa | 0.67 | 0.70 | 0.71 | 0.73 | 0.76 | 0.78 | 0.48 | 0.56 | 0.62 |
| AG-News | | | | | | | | | |
| BERT | 0.91 | 0.89 | 0.89 | 0.90 | 0.88 | 0.87 | 0.49 | 0.61 | 0.67 |
| RoBERTa | 0.91 | 0.88 | 0.88 | 0.91 | 0.88 | 0.88 | 0.55 | 0.65 | 0.70 |

Table 6.7: Adversarial detection results for Prompting of Phi-3-mini.

| Attacked Models | Adversarial Attacks | | | | | | | | |
|-----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| IMDB | | | | | | | | | |
| BERT | 0.77 | 0.73 | 0.71 | 0.78 | 0.77 | 0.74 | 0.60 | 0.66 | 0.70 |
| RoBERTa | 0.77 | 0.75 | 0.74 | 0.87 | 0.82 | 0.81 | 0.63 | 0.66 | 0.68 |
| AG-News | | | | | | | | | |
| BERT | 0.96 | 0.88 | 0.88 | 0.93 | 0.88 | 0.87 | 0.71 | 0.72 | 0.72 |
| RoBERTa | 0.95 | 0.87 | 0.86 | 0.92 | 0.82 | 0.82 | 0.69 | 0.71 | 0.70 |

Table 6.8: Adversarial detection results for Prompting of Llama 3.1 8B.

| | | Adversarial Attacks | | | | | | | | |
|-----------------|-----------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| Attacked Models | LLM Sizes | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| | | IMDB | | | | | | | | |
| BERT | 2B | 0.60 | 0.66 | 0.69 | 0.63 | 0.73 | 0.74 | 0.46 | 0.55 | 0.63 |
| | 9B | 0.77 | 0.76 | 0.76 | 0.74 | 0.76 | 0.74 | 0.60 | 0.63 | 0.66 |
| | 27B | 0.82 | 0.77 | 0.75 | 0.82 | 0.79 | 0.75 | 0.72 | 0.72 | 0.72 |
| RoBERTa | 2B | 0.61 | 0.66 | 0.69 | 0.75 | 0.77 | 0.78 | 0.51 | 0.59 | 0.66 |
| | 9B | 0.82 | 0.79 | 0.78 | 0.84 | 0.82 | 0.82 | 0.67 | 0.71 | 0.73 |
| | 27B | 0.86 | 0.78 | 0.75 | 0.87 | 0.80 | 0.79 | 0.69 | 0.68 | 0.69 |
| | | AG-News | | | | | | | | |
| BERT | 2B | 0.79 | 0.86 | 0.88 | 0.70 | 0.81 | 0.83 | 0.25 | 0.38 | 0.57 |
| | 9B | 0.94 | 0.91 | 0.92 | 0.89 | 0.89 | 0.88 | 0.47 | 0.59 | 0.66 |
| | 27B | 0.97 | 0.92 | 0.92 | 0.95 | 0.91 | 0.90 | 0.62 | 0.72 | 0.75 |
| RoBERTa | 2B | 0.83 | 0.88 | 0.89 | 0.74 | 0.83 | 0.86 | 0.29 | 0.43 | 0.59 |
| | 9B | 0.93 | 0.94 | 0.94 | 0.90 | 0.90 | 0.91 | 0.62 | 0.73 | 0.76 |
| | 27B | 0.95 | 0.92 | 0.92 | 0.95 | 0.91 | 0.92 | 0.66 | 0.75 | 0.77 |

Table 6.9: Adversarial detection results for Prompting of Gemma 2.

It is evident that the primary factor which affects task performance is model scale, seeing as the larger LLMs consistently outperform the smaller ones. For analysis purposes, we can divide the Large Language Models

| Attacked Models | Adversarial Attacks | | | | | | | | |
|-----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| IMDB | | | | | | | | | |
| BERT | 0.73 | 0.72 | 0.72 | 0.76 | 0.79 | 0.77 | 0.58 | 0.64 | 0.69 |
| RoBERTa | 0.82 | 0.80 | 0.80 | 0.84 | 0.81 | 0.81 | 0.61 | 0.67 | 0.71 |
| AG-News | | | | | | | | | |
| BERT | 0.91 | 0.94 | 0.95 | 0.89 | 0.93 | 0.93 | 0.45 | 0.61 | 0.70 |
| RoBERTa | 0.92 | 0.94 | 0.94 | 0.84 | 0.91 | 0.92 | 0.47 | 0.63 | 0.71 |

Table 6.10: Adversarial detection results for Prompting of Claude 3 Haiku.

| Attacked Models | Adversarial Attacks | | | | | | | | |
|-----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| IMDB | | | | | | | | | |
| BERT | 0.85 | 0.79 | 0.76 | 0.78 | 0.79 | 0.76 | 0.69 | 0.70 | 0.72 |
| RoBERTa | 0.87 | 0.80 | 0.78 | 0.90 | 0.83 | 0.82 | 0.74 | 0.72 | 0.72 |
| AG-News | | | | | | | | | |
| BERT | 0.98 | 0.90 | 0.90 | 0.96 | 0.90 | 0.89 | 0.68 | 0.74 | 0.75 |
| RoBERTa | 0.93 | 0.89 | 0.88 | 0.96 | 0.90 | 0.90 | 0.71 | 0.75 | 0.75 |

Table 6.11: Adversarial detection results for Prompting of GPT-4o mini.

into three groups of different abilities.

The first one consists of the smallest models, Gemma-2-2B and Phi-3-mini. With sizes below 4 billion parameters, these models have similar performance on the the IMDB dataset which is distinctly better than random but still subpar. A notable difference can be seen on the AG-News dataset however, on which Phi-3-mini approaches the performance of models which are orders of magnitude larger, with impressive values close to 90% on all metrics for the TextFooler and PWWS adversarial attacks. Gemma-2-2B falls a little short, reaching values of 75-85% on the same datasets.

Second in ability would be a group of Llama-3.1-8B and Gemma-2-9B, as across most metrics and experiments they exhibit values that are remarkably close to one another, having a differentiation of less than 6%. The performance of these models is promising, reaching values consistently above 70% on all three metrics and even getting to a Recall of 84% and 87%, against TextFooler and PWWS, on IMDB texts. Regarding AG-News texts, we once again observe a large climb in performance, with both LLMs having Recall scores that exceed 90% across the aforementioned adversarial attacks and both attacked models. Regarding the BAE-generated texts, we notice a particular drop in performance, with metrics showing decreases of 10-20% on IMDB and even around 30% on AG-News. Llama-3.1-8B and Gemma-2-9B are considered to be of medium scale in the open-source space; taking this into account, their results are promising and indicate that LLM-based detection of adversarial attacks may be performed successfully, even with more lightweight models which would allow for more flexibility and broader experimentation.

Lastly, we can group together Gemma-2-27B, Claude-3-Haiku, and GPT-4o-mini, the largest and most potent models. Despite not knowing their exact specifications, it is safe to assume that Claude-3-Haiku and GPT-4o-mini have significantly more parameters than Gemma-2-27B; based on this information, it is intriguing to see that Gemma-2-27B not only is competitive with both models, but even seems to surpass Claude-3-Haiku in adversarial detection ability. In any case, all three LLMs achieve impressive results across the board, reaching Recall values of 80-90% on IMDB data and consistently scoring above 90% on all metrics, on AG-News data, against TextFooler and PWWS. Against the BAE adversarial attack, we see that although the metrics still showcase moderate adversarial detection, they are again particularly reduced. This is a phenomenon which will be discussed in greater detail further below.

We should bear in mind that for all LLMs, apart from Phi-3-mini, we utilized the same prompt which was

engineered in Section 6.2.2 using Gemma-2-9B. From Table 6.9, we can observe that using this prompt, the model performance follows the scaling laws within the Gemma 2 family of models. At the same time, as was analyzed previously, this is also the case across the rest of the Large Language Models, which encompass a wide range of architectural designs. On this basis, we can deduce with sufficient confidence that our prompt, and therefore our prompt engineering framework, is robust and generalizes well. Another observation which supports this claim is that while the used prompt was engineered based on the TextFooler attack against the BERT classification model and IMDB dataset, it leads to similar detection results on the PWWS attack and RoBERTa model, and to much improved ones on the AG-News dataset.

It is necessary to analyze the degradation in performance which occurs against BAE-generated adversarial texts, across the board. The main cause we attribute this phenomenon to is the way in which BAE produces adversarial alternatives for masked tokens. To do so, it uses the BERT model, which features the same transformer architecture on top of which our employed detector-LLMs are built. This suggests a high probability that the parameter space of the attacker model and that of the detector model are adjacent to one another. As a result, the adversarial perturbations generated by BAE may not be drastically changing the internal representation of the original text within the LLMs, hence hindering their ability to accurately detect them. On top of that, BAE occasionally manages to attend to stronger perturbation constraints, thereby increasing the imperceptibility of its attacks, which as we will show in Section 6.5, also holds true for other detection methods in the bibliography.

Another important point to discuss is that while the results on IMDB are deemed satisfactory, especially for the most potent models, the results achieved on AG-News are notable more impressive, surpassing the former by up to 26% in the case of Phi-3mini and reaching Recall scores of up to 98% in the case of GPT-4o mini. The two text classification datasets consist of data which are fundamentally different in nature; IMDB is made up of movie reviews that exhibit a very wide range of vocabulary, formality, colloquialism, and writing styles, whilst AG-News contains news article titles and descriptions that are normally succinct and format. Nevertheless, we hypothesize that the main contributing factor for their resulting discrepancy is the vast difference in regards to their texts' length, which can be seen in Figures 6.1.1 and 6.1.2; AG-News samples consist of less than 60 words, whereas IMDB samples consist of up to 400 words. This hypothesis as well as its underlying elements are investigated further in the Analysis by Text Length subsection.

Qualitative Results

In Figure 6.2.1, we compare the analysis that the Gemma 2 LLMs do at different model scales, on the same input text. The given text has been adversarially edited in a subtle manner, as it has altered only two of its words with suitable replacements, and generated a text that maintains semantic coherence, does not introduce linguistic mistakes and could very well be imperceptible by humans. As we can notice, at the model sizes of 2 billion and 9 billion parameters, the LLMs are unable to detect any unusual characteristics within the sample and thereby wrongly classify it as original. Conversely, the 27B model is able to pick up on the nuance of the text and identifies an irregularity in one of the replaced words, thereby correctly classifying the input as adversarial. By observing its reasoning, we see that it reaches this conclusion not because it did a plain grammatical check, but due to it seemingly understanding that something is wrong within the broader context of the text. To add to this, the LLM mentions the actual original word as the word that the text is likely referring to, which enhances our method with explainability and validates the assumption that the most potent LLMs can indeed have a deep enough understanding of the given text that allows for effective adversarial detection.

In Figure 6.2.2, we see an example of GPT-4o-mini's analysis of an adversarial text. The model is able to identify all of the individual parts of the given text which were altered by the attack and as a result correctly classifies the sample as adversarial. We notice that in its answer, GPT-4o-mini not only gives comprehensive reasoning for why each part is considered to be adversarially manipulated, but also suggests either the exact original words or other suitable words as alternatives in each case. This again validates the potency and explainability of our method, and even alludes to the possibility of utilizing LLMs for purifying adversarial texts.

Input Text: "Woman's death probed - Public health bodies[officials] are investigating why a 38-year-old woman died two weeks after undergoing gastric bypass surgery at Saint Anne's Hospital in Fall River. The body[hospital] has stopped offering the surgery during the state probe and an internal review."

Gemma-2-2B Answer:

Let's think step by step. 1. **Grammatical Errors:** The text is grammatically correct. 2. **Misspellings and Unnatural Word Use:** No misspellings or unnatural word usage are apparent. 3. **Semantic Inconsistencies:** The text flows logically and makes sense. 4. **Language Irregularities:** The text uses standard English grammar and vocabulary. 5. **Overall Incoherence:** The text is coherent and presents a clear narrative. **Conclusion:** Based on the analysis, there are no signs of adversarial modification in the provided text. **Final Classification:** (A) original text

Gemma-2-9B Answer:

Let's think step by step. 1. **Grammatical Errors:** The text appears grammatically correct. 2. **Misspellings and Unnatural Word Use:** There are no obvious misspellings or unnatural word choices. 3. **Semantic Inconsistencies:** The text flows logically. There are no contradictions or unexpected connections between sentences. 4. **Language Irregularities:** The language is standard and appropriate for a news report. 5. **Overall Incoherence:** The text is coherent and easy to understand. **Summary:** The text does not exhibit any significant irregularities that would suggest adversarial modification. **Final Classification:** A original text

Gemma-2-27B Answer:

Let's analyze the text for signs of adversarial perturbation: **Grammatical Errors:** The grammar appears to be correct. **Misspellings:** There are no obvious misspellings. **Semantic Inconsistencies:** The sentence "The body has stopped offering the surgery..." is unusual. While grammatically correct, it's odd to say "the body" in this context. It likely refers to the hospital, but the phrasing is awkward and could be a subtle alteration. **Language Irregularities:** The phrasing "the body has stopped offering" is unusual and stands out as potentially unnatural. **Summary:** The text is grammatically sound, but the phrase "the body has stopped offering" is semantically unusual and potentially a deliberate alteration to introduce subtle meaning shifts or confusion for a machine learning model. **Classification:** (B) adversarial text

Figure 6.2.1: Comparison of the analysis that the Gemma 2 LLMs do on different model scales, on the same adversarial text. The adversarial edits made to the input text are denoted within brackets and in red color.

Input Text: "Israel Unions Start Nationwide striking[Strike] JERUSALEM (Reuters) - oil[Israeli] unions began a nationwide strike on Tuesday scheduled[expected] to involving[affect] about 400,000 public sector employee[workers] and severely hamper international travel."

GPT-4o-mini Answer: Let's analyze the given input text for signs of adversarial perturbations by closely examining it for various factors.

****Grammatical Errors****: - The phrase "scheduled to involving" should be "scheduled to involve." - "about 400,000 public sector employee" should be "about 400,000 public sector employees." The singular form "employee" is incorrect when referring to multiple people.

****Misspellings and Unnatural Use of Words****: - There are no clear misspellings present, but the use of "striking" at the beginning seems misplaced; the correct term could just be "strike."

****Semantic Inconsistencies****: - The sentence structure and content flow seem slightly awkward. The sentence "oil unions began a nationwide strike on Tuesday scheduled to involving about 400,000 public sector employee and severely hamper international travel" contains a convoluted structure that can lead to confusion. - "oil unions" is not unambiguously clear. If the text aims to discuss unions in the oil industry, it could have stated this more explicitly.

****Language Irregularities****: - The phrase "severely hamper international travel" feels abrupt and unconnected to the preceding content, leading to potential incoherence in how the information is presented.

****Overall Incoherence****: - The complete thought doesn't flow well due to the previous points. The awkwardness in grammatical structures and inaccuracies hints at an overall lack of coherence.

Based on the analysis: The text exhibits multiple grammatical errors, awkward phrasing, and issues with clarity, pointing towards intentional manipulation to obfuscate meaning rather than standard reporting. These irregular patterns in a news context suggest purposes that deviate from authentic journalism, fitting the profile of adversarial text modifications aimed at deceiving readers or automated systems. Given the significant number of identified irregularities and the overall impression of disjointedness in the content, I classify the input text as follows: (B) adversarial perturbation

Figure 6.2.2: Example of GPT-4o-mini's analysis of an adversarial text. The edits made to the input text by the adversarial attack are denoted within brackets and in red color.

Impact of Text Length

To investigate the impact of the input text length on model performance, we conducted a systematic analysis of our experimental results on the IMDb dataset across varying text lengths, measured by word count. Figures 6.2.3, 6.2.4, 6.2.5, 6.2.6, 6.2.7, 6.2.8, and 6.2.9 illustrate the relationship between input text length and adversarial detection Recall, against all attacking frameworks, on the RoBERTa model. The corresponding plots for the BERT base model can be found in Appendix A.1.2.

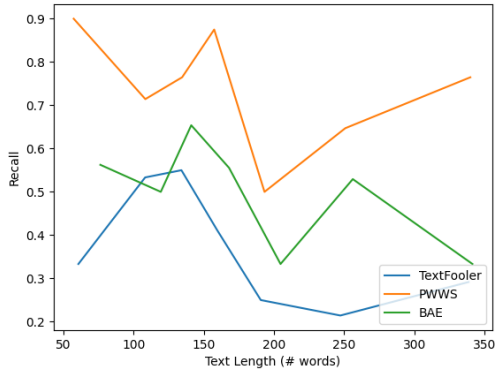


Figure 6.2.3: Adversarial detection Recall across text lengths, for prompting Gemma-2-2B on attacks against RoBERTa.

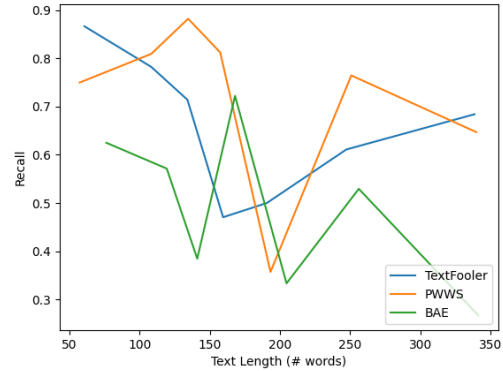


Figure 6.2.4: Adversarial detection Recall across text lengths, for prompting Phi-3-mini on attacks against RoBERTa.

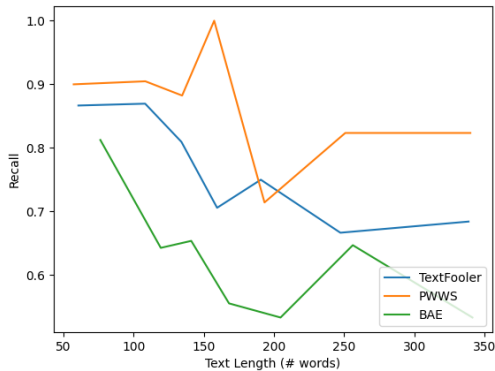


Figure 6.2.5: Adversarial detection Recall across text lengths, for prompting Llama-3.1-8B on attacks against RoBERTa.

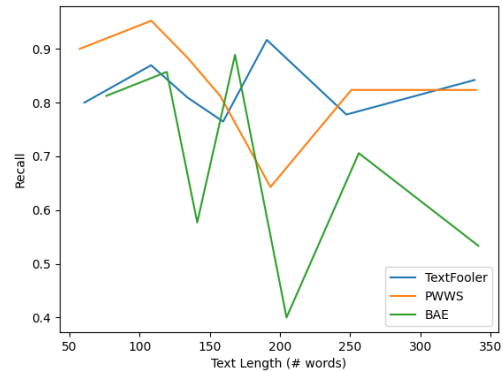


Figure 6.2.6: Adversarial detection Recall across text lengths, for prompting Gemma-2-9B on attacks against RoBERTa.

Under all scenarios, we see that there is great volatility, and despite the fact that the line graphs do not exhibit a strictly decreasing relationship between Recall and text length, the general trend is downwards. This becomes most apparent when observing the results on BAE-generated samples, which for the higher number of words end up degenerating to scores at the level of or even worse than that of a random classifier. Against the TextFooler and PWWS adversarial attacks, we see behaviours which are complex and dependent on the detector LLM. For example, there are a few occasions against PWWS-generated texts, in which Recall drops steeply but seems to recover after a certain text length. Additionally, the most potent LLM, GPT-4o-mini shows even greater signs of robustness since it maintains a Recall that is fluctuating but relatively close to a mean value across different text lengths, for both TextFooler and PWWS. Nevertheless, it is apparent that longer adversarial texts are more challenging to detect for all models.

One possible interpretation of these findings is the simple fact that in longer texts, an efficient adversarial

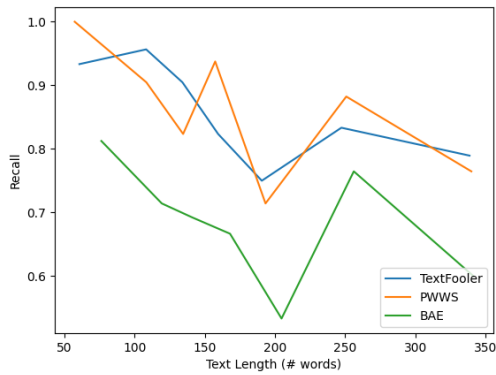


Figure 6.2.7: Adversarial detection Recall across text lengths, for prompting Gemma-2-27B on attacks against RoBERTa.

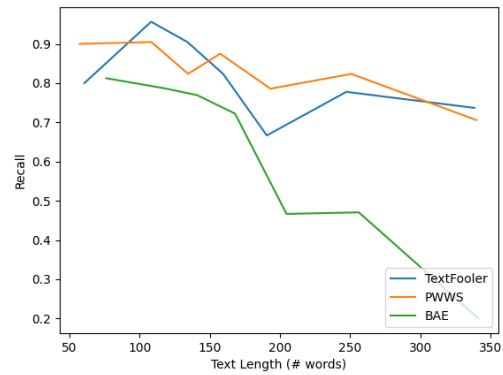


Figure 6.2.8: Adversarial detection Recall across text lengths, for prompting Claude-3-Haiku on attacks against RoBERTa.

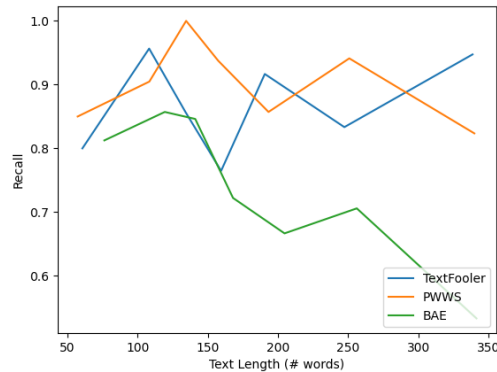


Figure 6.2.9: Adversarial detection Recall across text lengths, for prompting GPT-4o-mini on attacks against RoBERTa.

attack may alter only a small fraction of text. This means that the adversarial cues would be embedded in large amounts of unperturbed text, therefore making it harder for LLMs to detect them. While this is true, it is not always the case, as adversarial attacks often apply multiple edits. For that reason, we need to also look into the limitations of Large Language Models. In order for a model to successfully classify a given text, it has to analyze its linguistic structures and contextual relationships. When a text is longer, these characteristics become more complex and the amount of token-to-token relationships the model’s self-attention mechanism must process grows significantly. As a consequence, for longer adversarial samples, even within the context window, an LLM might not be able to maintain a coherent enough representation of the text and an adequate level of sensitivity to identify the dispersed adversarial alterations.

In any case, these observations were one of the main motivating factors which led us to create the Individual Sentences Classification method described in Section 5.3.2 which, as presented in Section 6.3, helps combat the phenomenon effectively.

6.2.4 Human Evaluation

We carried out a human evaluation study in order to get a more complete evaluation of our methods’ performance, as well as to study any possible correlation between the way LLMs and humans approach the task of deciding whether a given text is original or adversarial.

Survey Structure

Texts that are found in our datasets are oftentimes characterized by particularly advanced vocabulary and grammar, making their comprehension a challenging task. For that reason, we ensured that all of our human evaluators have a high proficiency in English. Furthermore a lot of text samples contain colloquial speech including slang and idioms which are used by native speakers but might be unknown to people who do not have English as their first language. Therefore, we made sure to expose our participants to examples of such texts and ensure that they have come in contact with such language and feel comfortable comprehending it. Both these criteria are put in place so as to ensure that the results of the human evaluation are affected as little as possible by English level. Based on the above, a total of 4 human evaluators participated in our study, all of whom have a C2 level proficiency in English.

First of all, the participants received a basic introduction to adversarial attacks in NLP, coupled with a few exemplars of texts on which adversarial perturbations have been applied. Then, they were given specific instructions concerning the task at hand, which guide them to carefully read through each given text, classify it as original or adversarial, and optionally detail any specific reasons which led them to make their particular classification.

The goal of this survey, in conjunction with the sanity test performed in Section 6.2.5, was to validate our method, particularly on the datasets for which it showcased a satisfactory performance for all the employed Large Language Models. That being the case, the data used in the survey are drawn from the data created using the TextFooler attack against the RoBERTa model for both the IMDB and the AG-News datasets. We randomly selected 100 samples for each sub-dataset, ensuring that the original to adversarial texts ratio remains approximately 1:1. Each survey participant evaluated all 200 samples, resulting in a total of 800 evaluations, which is a significant sample size for our analysis.

Survey Results

| Detection Methods | Recall | F1 | Accuracy |
|-------------------------|--------|------|----------|
| IMDB | | | |
| Gemma-2-2B | 0.58 | 0.65 | 0.69 |
| Phi-3-mini | 0.66 | 0.71 | 0.73 |
| Llama-3.1-8B | 0.78 | 0.77 | 0.77 |
| Gemma-2-9B | 0.76 | 0.75 | 0.75 |
| Gemma-2-27B | 0.84 | 0.78 | 0.76 |
| Claude 3 Haiku | 0.80 | 0.78 | 0.78 |
| GPT-4o mini | 0.82 | 0.78 | 0.77 |
| Human Evaluation | 0.75 | 0.79 | 0.80 |
| AG-News | | | |
| Gemma-2-2B | 0.86 | 0.90 | 0.90 |
| Phi-3-mini | 0.94 | 0.86 | 0.85 |
| Llama-3.1-8B | 1.00 | 0.92 | 0.91 |
| Gemma-2-9B | 1.00 | 0.97 | 0.97 |
| Gemma-2-27B | 0.98 | 0.94 | 0.94 |
| Claude 3 Haiku | 0.98 | 0.97 | 0.97 |
| GPT-4o mini | 0.96 | 0.93 | 0.93 |
| Human Evaluation | 0.85 | 0.91 | 0.91 |

Table 6.12: Adversarial detection results for the Human Evaluation datasets.

Table 6.12 presents the adversarial detection results of the human evaluators in comparison with those of the LLMs, on the same data. Overall, the performance of human detection on the IMDB dataset is up to par with that of mid-scale open source models, while on the AG-News dataset it is slightly worse than that of the worst-performing models. Similar to what was concluded for the smaller models through the comparison of their reasoning with that of the larger models, humans are producing less False Positives, however they seem less able to identify adversarial texts, leading to particularly reduced Recall. These results verify the

high degree of difficulty that characterizes the detection of textual adversarial attacks. On top of that, they verify the usefulness and competitiveness of our LLM prompting method.

It is worthwhile to examine certain categories of texts for which there appears to be a noticeable difference in the way that they are analyzed by humans and by Large Language Models. One such category is original texts which contain mistakes that are attributed to human carelessness or lack of proficiency in English. Human evaluators appeared to be a bit more able to discern these types of texts as original, even though they still contain errors. Examples of such errors are inconsistent capitalization of words, misuse of punctuation, and spelling mistakes that people often make themselves, and are therefore able to recognize them as non-malicious. Something similar can be observed for original content written by non-native English speakers. Taking into account the context of the entire text, humans can sometimes recognize the author’s reduced language competence, and as a result disregard any mistakes found in the sample and correctly classify it as original. A representative example is given in Figure 6.2.10.

I saw this movie only because Sophie Marceau. However, her acting abilities it’s no enough to salve this movie. Almost all cast don’t play their character well, exception for Sophie and Frederic. The plot could give a rise a better movie if the right pieces was in the right places. I saw several good french movies but this one i don’t like.

Figure 6.2.10: Example of an original text containing multiple errors. 2/4 human evaluators classified it correctly, while all LLMs classified it incorrectly.

On the contrary, one thing that differentiates LLMs as a detection method is the capacity to fact-check any given information. For example, when inspecting the models’ generated answers, we find instances in which during the analysis of a given text, the model cross-checks its content with information it has access to, in order to find any discrepancies that will indicate adversarial modification. Another case in which LLMs seem to outperform humans, is texts that are characterized by advanced vocabulary and/or convoluted sentence structures. When presented with such texts, our human evaluators sometimes struggled to analyze it and as a result either took no notice of adversarial edits within it and wrongly classified it as original or mistook certain phrases for perturbations and wrongly identified it as adversarial.

6.2.5 Sanity Test for Data Contamination

In the context of evaluating Large Language Models, one of the most prevailing issues is that of data contamination or test set contamination. This is the phenomenon in which a model’s training data contains some parts of the data that are used to evaluate it, consequently leading to artificially inflated scores on benchmarks and creating an overly positive image of an LLM’s abilities. Since both our test datasets and several of the training datasets of the LLMs we have employed in our experiments are sourced from the internet, it is possible that the models were trained on data contained in our evaluations.

To ensure the integrity of our experimental results and to mitigate the potential issue of such contamination, we implemented a sanity test. In order to achieve this, we manually curated a new set of test data by performing edits on existing ones, such that the LLMs would not have been previously exposed to them. In particular, we utilized the set of data which was used in the human evaluation study, so as to facilitate a smoother comparison of results. The edits performed include a multitude of methods, namely paraphrasing, reordering, deleting, or adding text, with the aim of achieving as much variability as possible in the resulting dataset. For original texts, we made sure that our modifications are not adversarial, meaning that they would not lead a classifier to flip its label in the context of text classification. On top of that, we aimed to make our edits imperceptible, by ensuring that they did not introduce errors, that they retained grammatical correctness, and that they altered the semantics of the content as little as possible, if at all. For adversarial texts, we made sure not to overwrite the parts of them which were perturbed by the adversarial attack method.

From Table 6.13, we can see that the sanity test results are very close to those obtained on the initial dataset. We do observe slight variations for various models, but they do not seem consistent in nature, whether positive

Input Text: "Bavarian[German] teenager indicted over Sasser seeing[worm] Prosecutors in Verden, Deutsche[Germany], indicted an 18-year-old student on Wednesday for allegedly creating the Sasser seeing[worm] that fell[crashed] hundreds of thousands of computers worldwide after spreading at lighting speed over the Internet."

GPT-4o-mini Answer:

Let's analyze the provided text step by step:

****Grammatical Errors****: The phrase "the Sasser seeing that fell" appears to be incorrectly structured. "seeing" seems to be a typo or incorrect word choice in this context. It likely should have been "worm" or another appropriate term related to the Sasser malware.

****Misspellings and Unnatural Use of Words****: The word "Deutsche" is typically used to refer to Germany in a language context, but it appears out of place here, as it could mislead the reader into thinking it refers to a corporation or similar entity without better context. The phrase "at lighting speed" is a commonly used term, but it should be "at lightning speed." This is a misspelling that alters the intended meaning.

****Semantic Inconsistencies****: The text discusses an indictment related to a notorious computer worm, but the use of "seeing" instead of a more appropriate term introduces confusion about the subject matter.

****Language Irregularities****: The overall flow of the sentence is awkward, and the odd use of "Deutsche" adds to a lack of coherence in presenting the idea clearly.

****Overall Incoherence****: The combination of the above factors leads to a text that is not coherent and blends factual information with errors that disrupt the message.

Given these points, I conclude that this text shows signs of intentional modification by an attacker, aiming to confuse or mislead the reader or a machine learning model.

Thus, the final classification is: (B) adversarial perturbation

Figure 6.2.11: Example of GPT-4o-mini fact-checking information during its analysis of an adversarial text. The adversarial edits made to the input text are denoted within brackets and in red color.

| Detection Methods | Recall | F1 | Accuracy |
|-----------------------|--------------|--------------|--------------|
| IMDB | | | |
| Gemma-2-2B | 0.60 (+0.02) | 0.66 (+0.01) | 0.69 (+0.00) |
| Phi-3-mini | 0.62 (-0.04) | 0.67 (-0.04) | 0.69 (-0.04) |
| Llama-3.1-8B | 0.80 (+0.02) | 0.78 (+0.01) | 0.77 (+0.00) |
| Gemma-2-9B | 0.76 (+0.00) | 0.75 (+0.00) | 0.74 (-0.01) |
| Gemma-2-27B | 0.82 (-0.02) | 0.77 (-0.01) | 0.75 (-0.01) |
| Claude 3 Haiku | 0.78 (-0.02) | 0.76 (-0.02) | 0.76 (-0.01) |
| GPT-4o mini | 0.84 (+0.02) | 0.79 (+0.01) | 0.77 (+0.00) |
| AG-News | | | |
| Gemma-2-2B | 0.86 (+0.00) | 0.90 (+0.00) | 0.90 (+0.00) |
| Phi-3-mini | 0.96 (+0.02) | 0.88 (+0.02) | 0.86 (+0.01) |
| Llama-3.1-8B | 1.00 (+0.00) | 0.89 (-0.03) | 0.88 (-0.03) |
| Gemma-2-9B | 0.97 (-0.03) | 0.93 (-0.04) | 0.93 (-0.04) |
| Gemma-2-27B | 1.00 (+0.02) | 0.92 (-0.02) | 0.91 (-0.03) |
| Claude 3 Haiku | 1.00 (+0.02) | 0.98 (+0.01) | 0.98 (+0.01) |
| GPT-4o mini | 0.96 (+0.00) | 0.92 (-0.01) | 0.92 (-0.01) |

Table 6.13: Adversarial detection results for the Sanity Test. Difference compared to metric achieved in initial dataset is noted within a parenthesis.

or negative. Furthermore, their absolute values are not deemed significant for our analysis, as even in the cases of the highest percentage differences of 3-4%, the actual model predictions differ by only a few samples. It is worth noting however, that the presence of these variations showcases the high sensitivity that LLMs have for changes in their input, which in our case is justified when taking into consideration that our manual edits oftentimes altered large portions of the input text. Overall, the sanity test results suggest that our experiments were not affected by data contamination and highlight the robustness of our detection approach.

6.3 Experiments of Individual Sentences Classification

For our Individual Sentences Classification experiments, we used a 75%-25% train-test split on the 4000 samples created by all 3 adversarial attacks, against BERT and RoBERTa, on the IMDB dataset. Since the AG-News dataset contains texts made up of article titles and descriptions which are typically only 1 to 3 sentences long, the adversarial detection method of Individual Sentences Classification could not have been applied to it.

6.3.1 Total Agreements

BERT

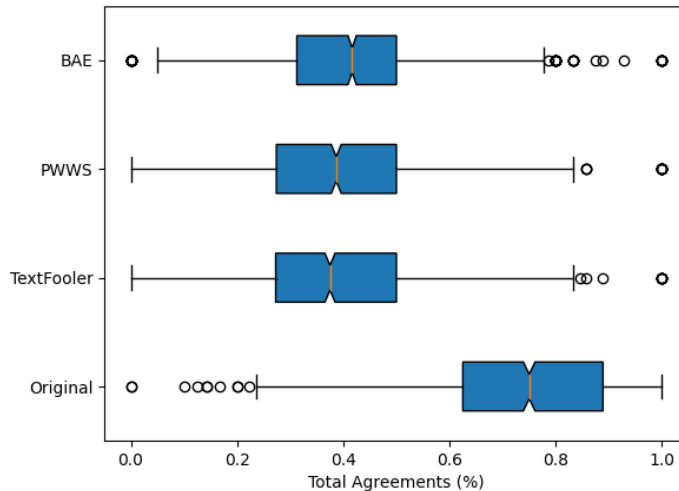


Figure 6.3.1: Boxplots of the Total Individual Sentences Classification Agreement percentage, for BERT.

In Figure 6.3.1, we can see the boxplots of the total individual sentences classification agreements, as a percentage of the total text length, for original and adversarial samples of the different attack methods against BERT.

As can be observed, the statistical distribution of the original texts is clearly distinguishable from the distribution of any which attack, with the former having a much higher median at around 75% and first quartile at just above 60%. The distributions of the different adversarial examples are very similar to each other, having almost identical third quartiles at around 50%. TextFooler and PWWS also present approximately the same first quartiles and as can be seen from their box’s notches, their medians, which sit at around 40%, do not present a significant statistical difference. BAE’s distribution is differentiated only by a bit, having a slightly upwards skewed median and first quartile. It should be noted that, for all 3 attacks, there are outliers that lie way above the boxplot’s whiskers, even reaching a value of 100%, which showcases there are at least a few instances in which the adversarial attacks edit the original text in such a subtle manner that leaves no traces to be detected by the Total Agreements method.

In Figure 6.3.2, we show the ROC curves that result from defining the total agreements percentage as the decision boundary for BERT to classify a text as adversarial or not, for each attack method. The AUC values

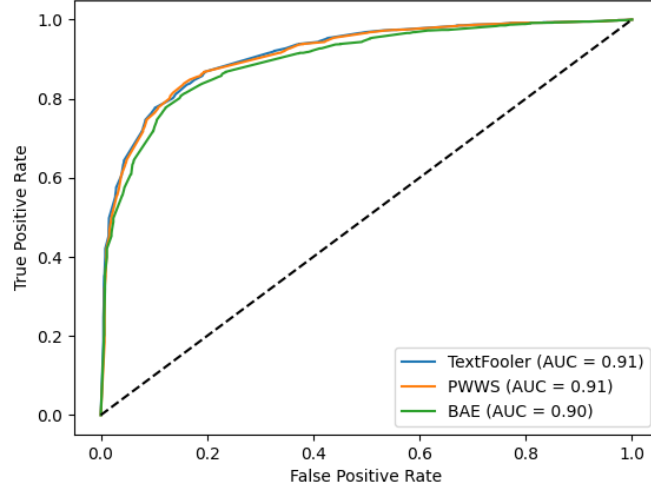


Figure 6.3.2: ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for BERT.

achieved showcase that the method presents a high ability of separating between original and adversarial samples. Furthermore, the fact that the ROC curves for all 3 attack methods are nearly indistinguishable proves that a detection threshold, chosen for one particular attack, via the Total Agreements method can be used interchangeably for any of other attacks as well. For TextFooler and PWWS texts, the ROC curves overlap almost entirely and the AUC is 0.91, while BAE texts’ ROC curve is placed marginally lower at most FPR values, with an AUC of 0.90.

In order to use the Total Agreements method for the detection of adversarial examples, we wanted to select a threshold that was as close as possible to the perfect classifier, the upper left point in a ROC curve. In making our selection, we also took into account the fact that in our task it is more important to optimize for true positives rather than false negatives. Based on these criteria, we chose the Total Agreements threshold of 60% that corresponds to a FPR of 0.2 and the elbows of the ROC curves.

| TextFooler | | | PWWS | | | BAE | | |
|------------|------|----------|--------|------|----------|--------|------|----------|
| Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| 0.89 | 0.84 | 0.83 | 0.90 | 0.85 | 0.84 | 0.88 | 0.83 | 0.82 |

Table 6.14: Adversarial detection results for the Total IndivSC Agreements method, for BERT.

In Table 6.14, we see the adversarial detection results for the Total Agreements method against each adversarial attack on BERT. The first observation which can be made is that the value of each metric is close to identical for all 3 attacks, reinforcing the fact that the same detection threshold can be used interchangeably against any adversarial attack. On top of that, the method performs highly, achieving metrics comparable to those of the best LLM Prompting scenarios, with 88-90% Recall, 83-85% F1 score, and 82-84% Accuracy. This confirms its suitability for adversarial text detection, as an auxiliary or even autonomous approach. All of the aforementioned conclusions are reinforced by Figure 6.3.3, in which we can observe that the ROC curves, as well as their AUCs, for the test set are in effect indistinguishable from the train set ones, a fact which also highlights consistency in detection.

It is worth examining the correlation between the Total Individual Sentences Classification Agreements percentage and a text’s length, measured in number of sentences, as they are directly intertwined. As can be observed in Figure 6.3.4, the mean value for both original and adversarial texts consistently remains within their corresponding boxplot’s first and third quartile points, with all 3 attacks having similar plot lines across varying text lengths. Overall, we observe that for all 4 types of texts, the plot follows a downwards

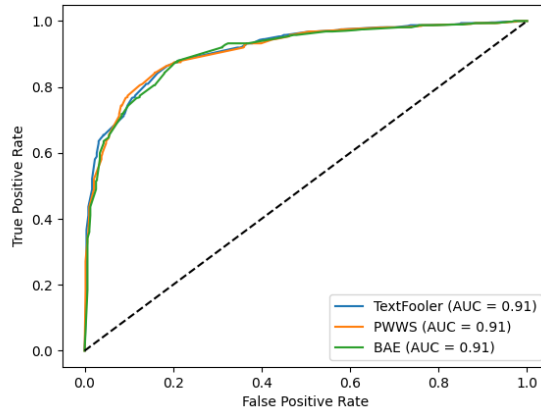


Figure 6.3.3: ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for BERT, on the test set.

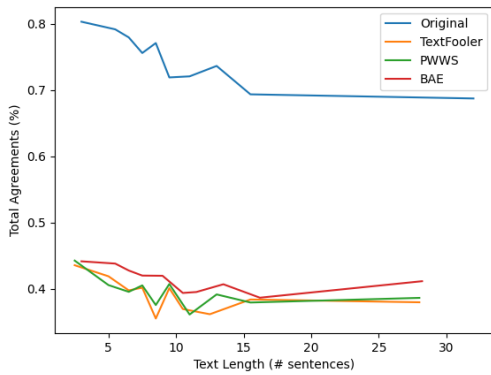


Figure 6.3.4: Total IndivSC Agreements percentage across text lengths, for BERT.

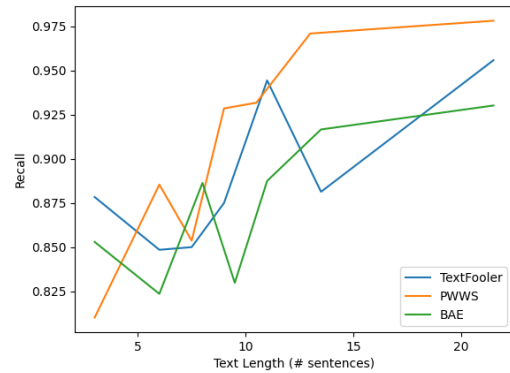


Figure 6.3.5: Adversarial detection Recall for the Total IndivSC Agreements method across text lengths, for BERT.

trend. This is something to be expected, as for shorter original texts, it is more plausible that the overall sentiment agrees with the sentiment of its few individual sentences. In addition, for shorter adversarial texts, the few sentences whose sentiment was flipped make up a larger percentage of the total text length, therefore reducing the total agreements percentage. The separability between original and adversarial texts also remains consistently wide for any number of sentences, even when considering deviations. On top of that, the decreasing nature of the plot lines leads to the conclusion that a constant threshold for detection of adversarial attacks for the Total Agreements method will be able to produce more True Positives for longer texts. The previous conclusion is verified by Figure 6.3.5. In it, we observe that the detection Recall achieved by the method increases noticeably the longer a text is, for all three attacking techniques.

RoBERTa

The results obtained for attacks against RoBERTa are very similar to the ones mentioned above. In Figure 6.3.6, we see the boxplots of the total individual sentences classification agreements, as a percentage of the total length, for original and adversarial samples of the 3 attack methods against the RoBERTa model. Again, it is evident that the statistical distributions of original and adversarial texts are separable, with all three of the latter having significantly lower median values, which do not differ statistically from each other, than the former. The only noticeable differences when compared with BERT's results, are some variations

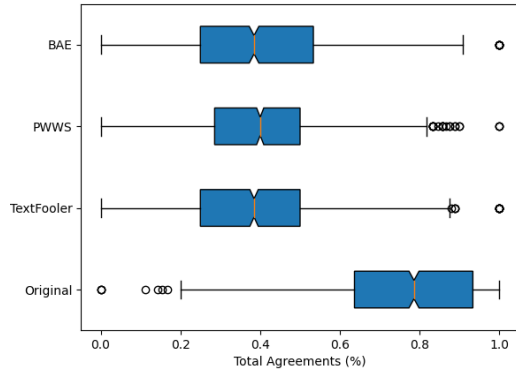


Figure 6.3.6: Boxplots of the Total IndivSC Agreement percentage, for RoBERTa.

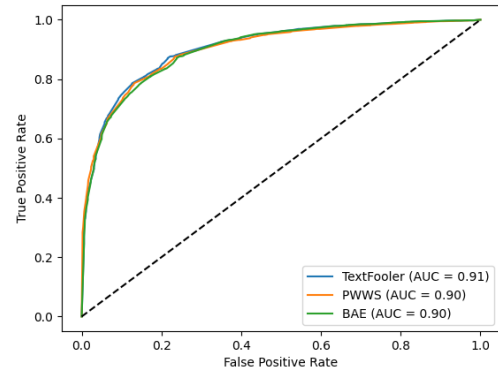


Figure 6.3.7: ROC curves for using the Total IndivSC Agreement percentage as the threshold for adversarial detection, for RoBERTa.

in interquartile ranges (IQR). In particular, texts generated by Textfooler or BAE lead to bigger IQRs and therefore wider spreads, whereas texts generated by PWWS seem to create a tighter spread.

Observing Figure 6.3.7, we conclude that RoBERTa and BERT have virtually identical ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, and therefore both models have the same high capacity to recognize adversarial samples. For RoBERTa's threshold, once more, we choose 60% which corresponds to a FPR of 0.2 in the ROC curves.

| TextFooler | | | PWWS | | | BAE | | |
|------------|------|----------|--------|------|----------|--------|------|----------|
| Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| 0.85 | 0.83 | 0.82 | 0.85 | 0.83 | 0.82 | 0.85 | 0.83 | 0.82 |

Table 6.15: Adversarial detection results for the Total Agreements method, for RoBERTa.

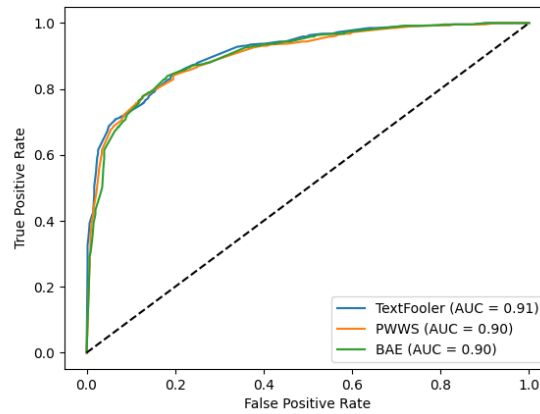


Figure 6.3.8: ROC curves for using the Total IndivSC Agreements percentage as the threshold for adversarial detection, for RoBERTa, on the test set.

In comparison to BERT's adversarial detection metrics, RoBERTa exhibits almost the same ones, having only a 3-5% decrease in Recall, alongside a 1-2% decrease in F1 and Accuracy, which indicates that RoBERTa detected marginally fewer adversarial attacks. Such a fact could be caused by the distribution of the texts, both original and adversarial, being skewed moderately higher. This conclusion can be further validated if

combined with the insights derived from 6.3.8. In it, we notice that the ROC curves and their corresponding AUCs of the test set are almost indiscernible from the train set ones, only having their elbows slightly more to the right. This implies that if the detection threshold was to be increased by a bit, the metrics achieved would be even identical to BERT’s.

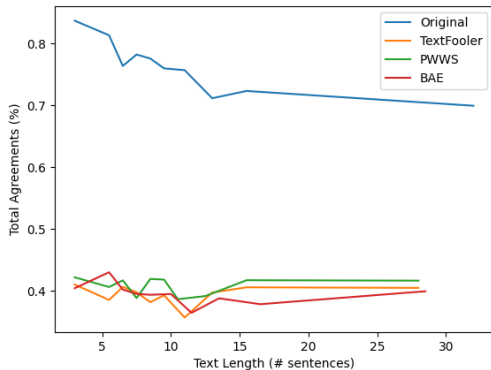


Figure 6.3.9: Total IndivSC Agreements percentage across text lengths, for RoBERTa.

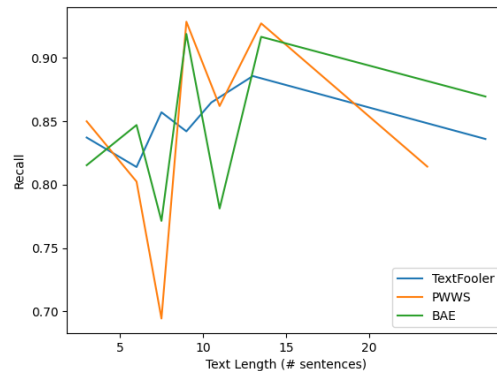


Figure 6.3.10: Adversarial detection Recall for the Total Agreements method across text lengths, for RoBERTa.

In Figure 6.3.9, similarly as for BERT, we observe the consistent separation of original and adversarial samples across any number of sentences. Total IndivSC Agreements percentage is decreasing for longer original texts, whilst it oscillates around the median percentage and eventually smoothness out for longer adversarial samples. This causes the adversarial detection Recall, shown in Figure 6.3.10, to fluctuate but have an overall increasing trend for higher number of sentences in the input text.

6.3.2 Longest Continuous Agreement Sequence

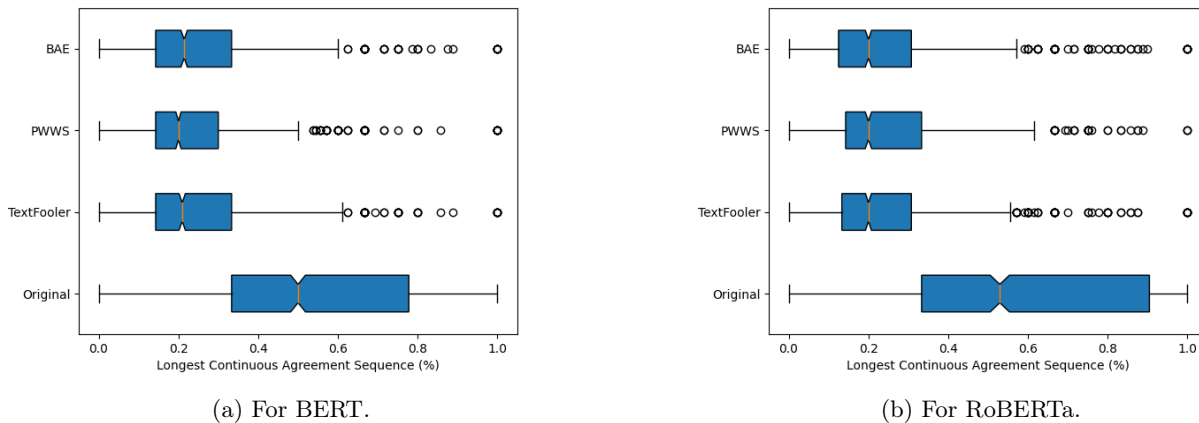
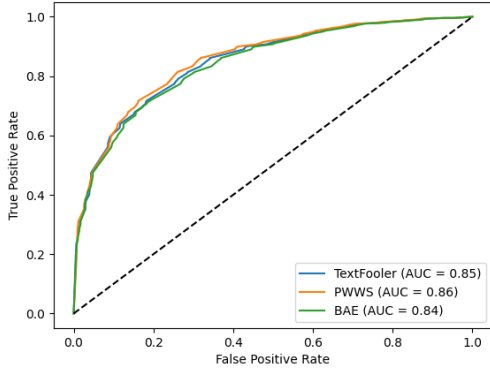


Figure 6.3.11: Boxplots of the Longest Continuous IndivSC Agreement Sequence percentage.

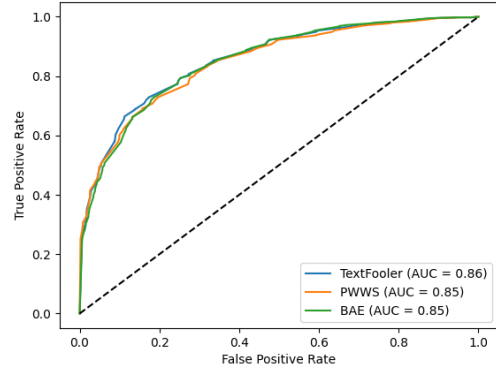
By observing the boxplots of the longest continuous IndivSC agreement sequence, as a percentage of total text length, in Figure 6.3.11a, we see that the distributions of original and adversarial samples are separable by BERT, but only barely. On the one hand, the different adversarial attacks exhibit very similar characteristics with statistically indistinguishable medians at around 20% and Q3 values close to 30%. On the other hand, for original texts, the median value stands significantly higher at almost 50%, but the exhibited IQR is especially wide, mainly in the direction of Q3, but also in that of Q1. That being the case, the first quartile is almost equal to the third quartile of the adversarial texts, which means that approximately 25% of the

examples of either category are essentially not able to be distinguished by using Longest Continuous IndivSC Agreement Sequence.

For RoBERTa, Figure 6.3.11b shows much the same results. There is still the same similarity between different attacking techniques with only slight variations in interquartile range, which for TextFooler and BAE moderately increases, and for PWWS moderately decreases the separation between their Q3 and the Q1 of original texts. A main difference we notice is that the third quartile is extended by over 10%, signifying the volatility of the statistical distributions produced by this method.



(a) For BERT.



(b) For RoBERTa.

Figure 6.3.12: ROC curves for using the Longest Continuous IndivSC Agreement Sequence percentage as the threshold for adversarial detection.

By defining the length of the longest continuous agreement sequence as the classification threshold between original and adversarial samples, we get the ROC curves in Figure 6.3.12a, for BERT. These showcase that despite its statistical distributions not being optimal, Longest Continuous IndivSC Agreement Sequence can still be a viable option for adversarial detection. The 3 different curves are almost entirely overlapping, which indicates consistent behaviour across all adversarial attacks and the AUCs achieved are lower than those of the Total Agreements method, but only by 5-6 percentage points.

When it concerns to RoBERTa, its behaviour is virtually identical to that of BERT, both numerically, in terms of AUC values, and qualitatively, in terms of plots. The only minor difference which can be noticed is the fact the RoBERTa seems to be performing worse against the PWWS attack, by a small degree.

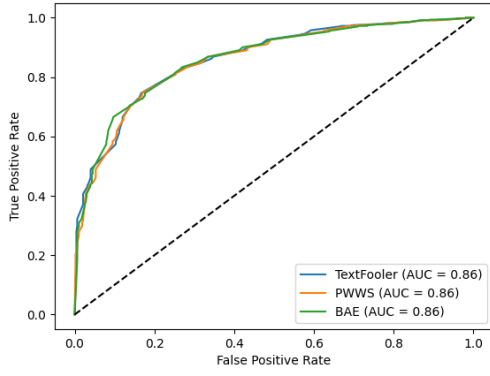
In order to evaluate the method’s effectiveness in adversarial detection, we selected a threshold of 33%, so as to strike a balance between increasing TPR without overtly degrading FPR. In particular this value corresponds to an FPR of 0.25 and a point at the elbow of the curves.

| Models | Adversarial Attacks | | | | | | | | |
|----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | 0.75 | 0.77 | 0.78 | 0.75 | 0.77 | 0.78 | 0.75 | 0.77 | 0.78 |
| RoBERTa | 0.74 | 0.76 | 0.77 | 0.70 | 0.74 | 0.75 | 0.74 | 0.76 | 0.77 |

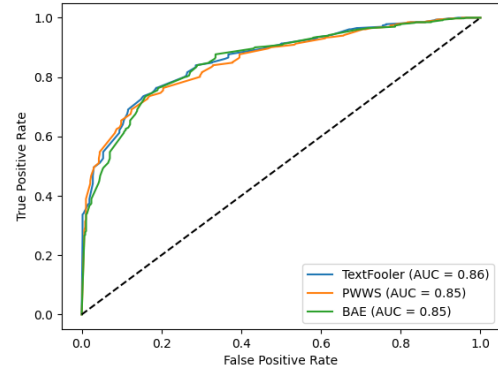
Table 6.16: Adversarial detection results for the Longest Continuous IndivSC Agreement Sequence method.

The adversarial detection results of the Longest Continuous IndivSC Agreement Sequence method are shown in Table 6.16. From it, we deduce a moderately high performance, which is close to the levels achieved by mid-scale LLMs, with scores consisting of 70-75% Recall, 74-77% F1, and 75-78% Accuracy. It is important to note that BERT accomplishes the exact same metrics for all 3 adversarial frameworks, which highlights the consistency and generalization of the chosen decision boundary against varying attacks.

In comparison, RoBERTa’s performance appears almost equal, with only a slight drop being observed against PWWS.



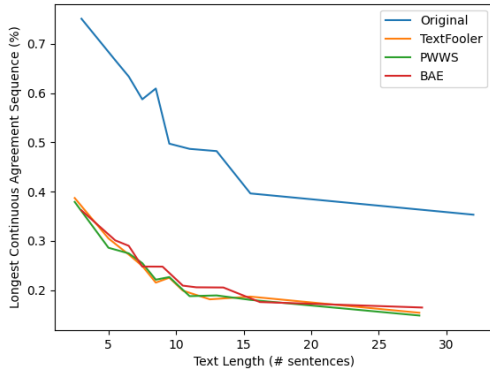
(a) For BERT.



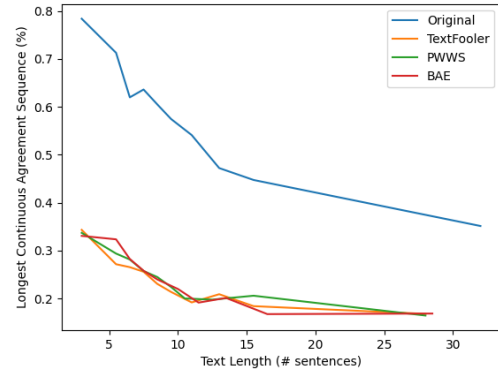
(b) For RoBERTa.

Figure 6.3.13: ROC curves for using the Longest Continuous IndivSC Agreement Sequence percentage as the threshold for adversarial detection, for the test set.

To add to the previous conclusions and to validate the method’s viability, in Figures 6.3.13a and 6.3.13b, we notice that the ROC curves produced by both models on the test set show no difference to the ones of the train set.



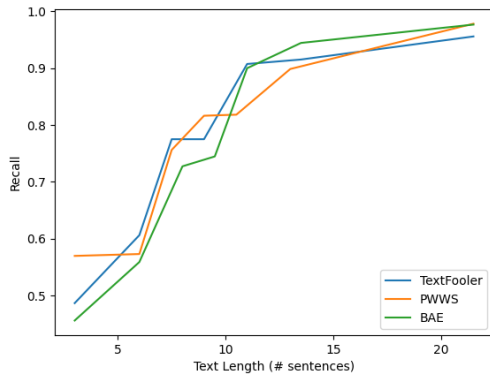
(a) For BERT.



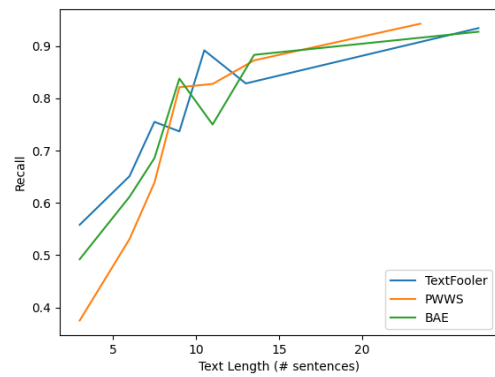
(b) For RoBERTa.

Figure 6.3.14: Longest Continuous IndivSC Agreement Sequence percentage across text lengths.

In order to examine the impact of text length on the Longest Continuous IndivSC Agreement Sequence percentage, we begin by analyzing Figures 6.3.14a and 6.3.14b. In them, we see that for all 4 types of texts, the percentage follows the statistical characteristics of its corresponding boxplot, beginning at values close to Q3 for shorter texts and reaching values near Q1 for longer ones, for both BERT and RoBERTa. In comparison to the Total IndivSC Agreements percentage, here the classes of original and adversarial texts are still clearly separable, however they are closer to each other. All of the plot lines show a steep decrease, and with the mean value of the adversarial texts eventually dropping well below the threshold of 33%, while that of original texts remains above it, we assume that detection ability will improve as the number of sentences increases. That is indeed the case, as verified by Figures 6.3.15a and 6.3.15b, in which we see the detection Recall on the test set increasing rapidly for longer texts, quickly surpassing its average performance, for BERT and for RoBERTa.



(a) For BERT.



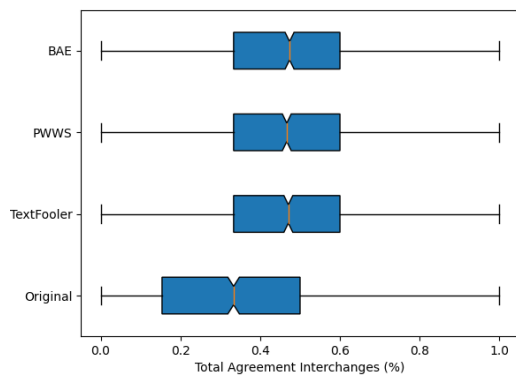
(b) For RoBERTa.

Figure 6.3.15: Adversarial detection Recall for the Longest Continuous IndivSC Agreement Sequence method across text lengths.

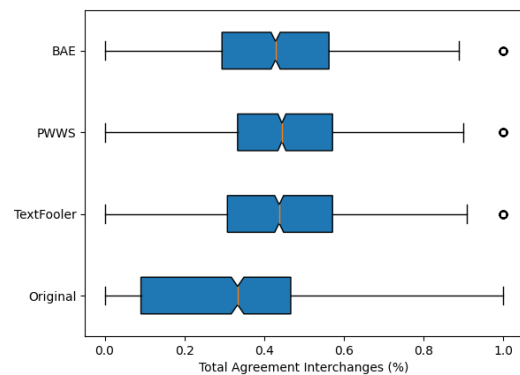
6.3.3 Total Interchanges

Figures 6.3.16a and 6.3.16b display the boxplots of the total individual sentences classification interchanges, for original samples and for attacks against BERT and RoBERTa. From them, it can be immediately deduced that there is great overlap between the statistical distributions of original and adversarial texts. The three different attacks show only small differences between them, having a median around 0.45 and a Q1 which is close to the same as the median of the original texts, at approximately 0.35. In fact, the Q2-Q3 range of original samples coincides entirely with the Q1-Q2 range of adversarial ones, for both BERT and RoBERTa. Another thing to note is that in Figure 6.3.16a, for any which type of text, the boxplot's whiskers extend from 0.0 all the way to 1.0, which showcases the high degree of dispersion that characterize all the distributions. All of these findings indicate that Total Interchanges cannot be a suitable method for adversarial detection. Nevertheless, we present the rest of our experimental findings on it, for reasons of comprehensiveness and comparison.

The ROC curves shown in Figures 6.3.17a and 6.3.17b come as a result of defining the total interchanges percentage as the threshold for identifying samples as original or adversarial. The AUCs achieved indicate a performance that is better than random but definitely not up to par with the other examined methods, confirming our assumption. To be specific, the plot lines of the 3 different attacks have a similar form and values in the range of 0.67-0.69, for both BERT and RoBERTa.

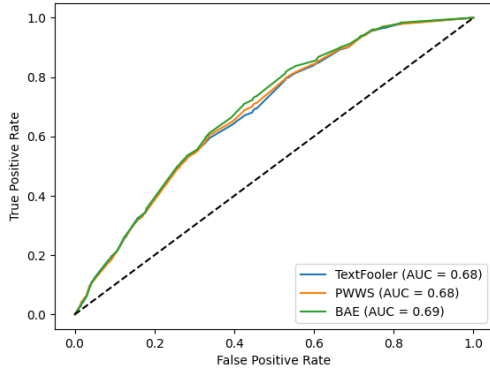


(a) For BERT

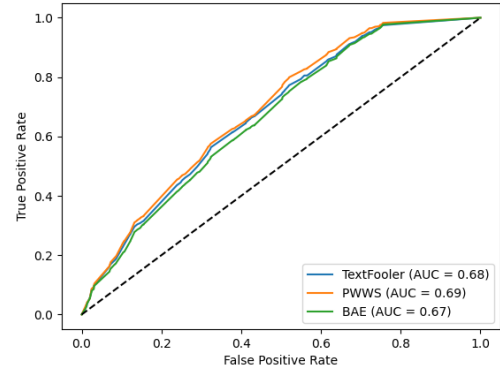


(b) For RoBERTa

Figure 6.3.16: Boxplots of the Total IndivSC Interchanges percentage.



(a) For BERT



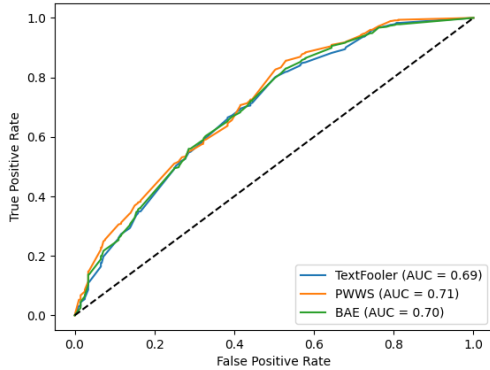
(b) For RoBERTa

Figure 6.3.17: ROC curves for using the Total IndivSC Interchanges percentage as the threshold for adversarial detection.

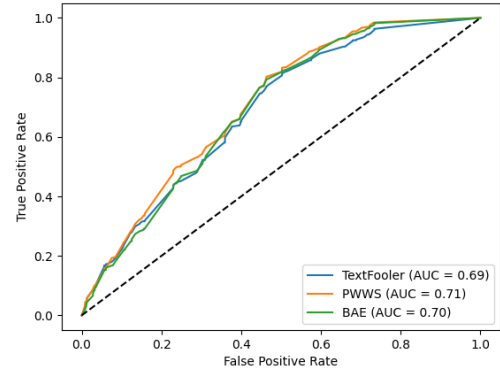
For the purpose of using the Total IndivSC Interchanges percentage as a classification boundary, we had to choose a threshold that results in a Recall above 50% without totally deteriorating the False Positive Rate, and thereby that value was 38%, which corresponds to a FPR of 0.4.

| Models | Adversarial Attacks | | | | | | | | |
|----------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| | TextFooler | | | PWWS | | | BAE | | |
| | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | 0.67 | 0.65 | 0.64 | 0.66 | 0.64 | 0.63 | 0.67 | 0.65 | 0.64 |
| RoBERTa | 0.60 | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 | 0.62 | 0.62 | 0.63 |

Table 6.17: Adversarial detection results for the Total IndivSC Interchanges method.



(a) For BERT



(b) For RoBERTa

Figure 6.3.18: ROC curves for using the Total IndivSC Interchanges percentage as the threshold for adversarial detection, for the test set.

In Table 6.17, we see the adversarial detection results for the Total IndivSC Interchanges method against each adversarial attack. We can observe consistency between the metrics achieved for the 3 different attacks by either base model, however the metrics achieved are all in the range of 60-67% and correspond to behaviour better than random and close to the level of the smallest employed LLMs.

The ROC curves of Figures 6.3.18a and 6.3.18b, which are generated on the test set, paint the same picture,

with the AUCs achieved being higher than those on the train set by 1-3 percentage points.

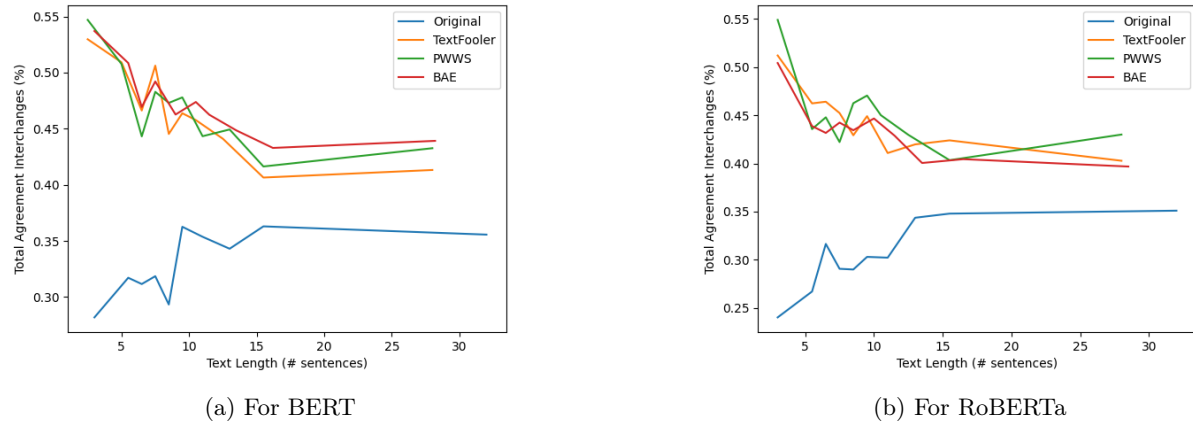


Figure 6.3.19: Total IndivSC Interchanges percentage across text lengths.

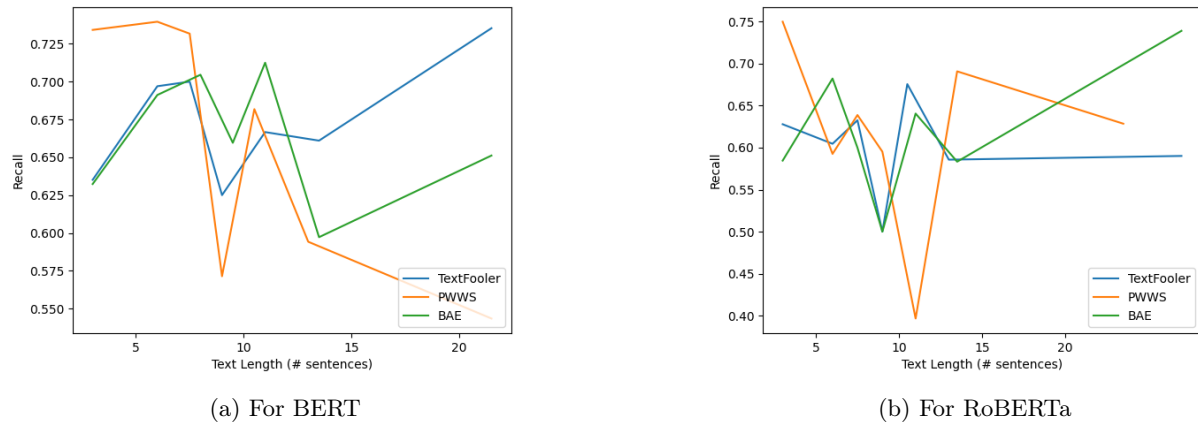


Figure 6.3.20: Adversarial detection Recall for the Total IndivSC Interchanges method across text lengths.

In regards to the effect that a text’s length has on the Total Interchanges percentage, we analyze Figures 6.3.19a and 6.3.19b. In them, we see that the mean value of the percentage along longer texts increases for original samples and decreases for adversarial samples, up until a stabilization length of around 12 sentences, therefore decreasing their distance. This finding alludes to the fact that the method will be even less able to distinguish between longer adversarial and original texts, at least up to the stabilization value. The plot lines of adversarial detection Recall across text lengths in Figures 6.3.20a and 6.3.20b do not definitively confirm this assumption, as they are characterized by great fluctuations, however they do present cases in which Recall drops steeply until a certain text length, for example against the PWWS attack for both BERT and RoBERTa.

6.4 Combination of Methods

In this section we showcase how we combined our two proposed methods into a unified approach. To achieve this, we used LLM Prompting as the primary detection method, and supplemented it with Total Agreements, which is the best-performing Individual Sentences Classification framework. We utilized Total Agreements only on texts that are above a certain length, specifically 6 sentences long, in order to utilize the fact that its adversarial detection recall seems to increase for longer texts, while the exact opposite occurs in the case of Large Language Models. On top of that, we wanted to ensure that the utilization of the Total Agreements

method was only complementary and did not lead to an increase in False Positives or False Negatives. To facilitate that, we set the threshold for individual sentences classification agreements percentage as 50% in order to consider a text as adversarial, which corresponds to a False Positive Rate (FPR) of 10%, and set the threshold for considering a text as original to 66%, which corresponds to a False Omission Rate (FOR) of 10%. To conclude, our combination prioritizes the prediction made by the LLM. Then, in the case the given text's length is above 6 sentences, we calculate whether its total IndivSC agreements percentage falls within the set thresholds. If that is the case and the LLM-based prediction disagrees with the Total Agreements-based prediction, then we consider the latter one.

Table 6.18 presents the experimental results that are produced by using this unified approach.

| Attacked Models | Methods | Adversarial Attacks | | | | | | | | |
|-----------------|------------------|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| BERT | Gemma-2-2B * | 0.82 | 0.81 | 0.81 | 0.87 | 0.87 | 0.86 | 0.83 | 0.83 | 0.84 |
| | Phi-3-mini * | 0.84 | 0.82 | 0.82 | 0.91 | 0.89 | 0.88 | 0.82 | 0.81 | 0.82 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.92 | 0.89 | 0.87 | 0.87 | 0.85 | 0.85 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.85 | 0.91 | 0.88 | 0.86 | 0.85 | 0.84 | 0.84 |
| | Gemma-2-27B * | 0.91 | 0.86 | 0.86 | 0.94 | 0.88 | 0.86 | 0.88 | 0.85 | 0.85 |
| | Claude-3-Haiku * | 0.88 | 0.85 | 0.84 | 0.92 | 0.89 | 0.88 | 0.84 | 0.85 | 0.85 |
| | GPT-4o-mini * | 0.92 | 0.88 | 0.87 | 0.94 | 0.90 | 0.89 | 0.88 | 0.85 | 0.86 |
| RoBERTa | Gemma-2-2B * | 0.83 | 0.84 | 0.84 | 0.85 | 0.82 | 0.82 | 0.80 | 0.84 | 0.85 |
| | Phi-3-mini * | 0.90 | 0.87 | 0.87 | 0.84 | 0.82 | 0.82 | 0.82 | 0.83 | 0.84 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.87 | 0.82 | 0.82 | 0.85 | 0.85 | 0.86 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.86 | 0.87 | 0.82 | 0.82 | 0.85 | 0.87 | 0.87 |
| | Gemma-2-27B * | 0.92 | 0.88 | 0.87 | 0.89 | 0.82 | 0.82 | 0.84 | 0.85 | 0.86 |
| | Claude-3-Haiku * | 0.90 | 0.88 | 0.87 | 0.87 | 0.83 | 0.83 | 0.87 | 0.88 | 0.88 |
| | GPT-4o-mini * | 0.90 | 0.89 | 0.88 | 0.87 | 0.83 | 0.82 | 0.85 | 0.87 | 0.87 |

Table 6.18: Adversarial detection results for the combination of methods, on IMDb. Asterisks indicate the combination of methods and bold numbers indicate the best result per metric.

We notice that the combination of our methods yields better results than each one does independently, something that highlights both methods' effectiveness and the fact that they work together in a complementary manner. These final metrics are impressive, as they reach levels of close to 90% all around, close to the ones induced through LLM Prompting on the AG-News dataset.

On the one hand, the fact that performance increases in comparison to IndivSC proves that LLMs are well equipped to analyze shorter texts and texts that fall within the middle of the different statistical distributions of the total agreements percentage. On the other hand, the fact that performance increases in comparison to LLM Prompting verifies the robustness of IndivSC, especially when it comes to longer text samples. Furthermore, because the Total Agreements methods yields similar results for all adversarial attacks, it manages to compensate for the reduced Recall that LLM Prompting showcases against BAE. Another thing to note is that through the combination of methods, the smaller LLMs increased their metrics by significantly more than larger ones. This led to a shortening of their ability gap, even though the difference is still noticeable, with GPT-4o-mini achieving the highest scores against all attacks on BERT, to give an example. Nonetheless, this observation is in sync with our previous finding that Large Language Models generally find shorter texts easier to analyze and leads to the conclusion that the differences in model ability were mostly displayed on texts of longer length.

6.5 Comparison with State-of-the-art Detection Methods

6.5.1 Compared Methods

FGWS (Frequency-Guided Word Substitutions)

FGWS (Frequency-Guided Word Substitutions) [45] is a detection framework which takes advantage of the observation that there are consistent differences in frequency between a text’s original words, and the perturbations they are replaced with by an adversarial attack. Specifically, it is shown that adversarial attacks tend to replace words with ones that are encountered less often. Based on this insight, FGWS transforms a given sample by substituting infrequent words with more frequent ones and decides whether the input text is adversarial based on the change that this operation causes to the prediction confidence of a classification model.

RDE (Robust Density Estimation)

In order to achieve adversarial detection, RDE (Robust Density Estimation) [75] estimates the probability density of any given text sample. In detail, it fits a parametric density estimation model utilizing the classification model’s internal representation of the given input. This is achieved using the training data, therefore causing adversarial data, which are unseen by the model, to likely have a low probability and therefore be detectable. Issues of dimensionality in this approach are resolved using statistical analysis techniques such as kernel PCA and Minimum Covariance Determinant.

6.5.2 Comparison Results

Table 6.19 presents the results on all 6 datasets for both our proposed method, and the state-of-the-art detection methods described above. We immediately observe the competitiveness of our approach across the board, especially when it concerns to the 3 large-scale LLMs.

RDE seems to be the most potent approach on IMDb data, achieving the best performance on all 3 metrics in all scenarios, except from the PWWS attack against the BERT model, in which GPT-4o-mini prevails on Recall and F1 score, and has equal Accuracy. In regards to AG-News data, the potency of LLM prompting is highlighted even more, with RDE falling behind either Gemma-2-27B, Claude-3-Haiku, or GPT-4o-mini in all cases, excluding the BAE attack.

Concerning FGWS, it does perform generally high, but falls short of another detection methods in all cases. For example, out of the 12 total dataset-attack-model combinations, GPT-4o-mini outperforms FGWS in 11 of them on Recall, in 9 of them on F1 score (equal performance in 1 combination), and in 6 of them in Accuracy (equal performance in 3 combinations). In comparison, the combinations in which Gemma-2-27B outperforms FGWS are 10 on Recall, 7 on F1 score (equal performance in 1 combination), and 6 on Accuracy (equal performance in 1 combinations).

With reference to the degradation in our method’s performance against the BAE adversarial attack, which was discussed in Section 6.2.3, it is interesting to observe that the exact same behaviour occurs for the FGWS detection method as well, a fact that highlights the increased nuance of the attack. Having said that, RDE showcases no difference in performance in regards to BAE, thereby displaying notable robustness.

At this point, it should be noted that in comparison to other works in the bibliography, our approach, as well as FGWS and RDE, make no assumption of the attack algorithm or the attacked model. Additionally, a factor that differentiates our proposed method from the examined state-of-the-art ones, is the fact that it requires no access to the feature space of the base classifier model.

Taking all of the above into consideration, we conclude that our approach is a viable solution, useful for plug-and-play detection of adversarial texts in text classification.

| | | Adversarial Attacks | | | | | | | | |
|-----------------|------------------|---------------------|------|----------|--------|------|----------|--------|------|----------|
| Attacked Models | Methods | TextFooler | | | PWWS | | | BAE | | |
| | | Recall | F1 | Accuracy | Recall | F1 | Accuracy | Recall | F1 | Accuracy |
| | | IMDB | | | | | | | | |
| BERT | Gemma-2-2B * | 0.82 | 0.81 | 0.81 | 0.87 | 0.87 | 0.86 | 0.83 | 0.83 | 0.84 |
| | Phi-3-mini * | 0.84 | 0.82 | 0.82 | 0.91 | 0.89 | 0.88 | 0.82 | 0.81 | 0.82 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.92 | 0.89 | 0.87 | 0.87 | 0.85 | 0.85 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.85 | 0.91 | 0.88 | 0.86 | 0.85 | 0.84 | 0.84 |
| | Gemma-2-27B * | 0.91 | 0.86 | 0.86 | 0.94 | 0.88 | 0.86 | 0.88 | 0.85 | 0.85 |
| | Claude-3-Haiku * | 0.88 | 0.85 | 0.84 | 0.92 | 0.89 | 0.88 | 0.84 | 0.85 | 0.85 |
| | GPT-4o-mini * | 0.92 | 0.88 | 0.87 | 0.94 | 0.90 | 0.89 | 0.88 | 0.85 | 0.86 |
| | FGWS | 0.85 | 0.87 | 0.87 | 0.88 | 0.89 | 0.89 | 0.62 | 0.72 | 0.76 |
| | RDE | 0.97 | 0.94 | 0.93 | 0.88 | 0.89 | 0.89 | 0.94 | 0.92 | 0.92 |
| RoBERTa | Gemma-2-2B * | 0.83 | 0.84 | 0.84 | 0.85 | 0.82 | 0.82 | 0.80 | 0.84 | 0.85 |
| | Phi-3-mini * | 0.90 | 0.87 | 0.87 | 0.84 | 0.82 | 0.82 | 0.82 | 0.83 | 0.84 |
| | Llama-3.1-8B * | 0.90 | 0.85 | 0.84 | 0.87 | 0.82 | 0.82 | 0.85 | 0.85 | 0.86 |
| | Gemma-2-9B * | 0.89 | 0.86 | 0.86 | 0.87 | 0.82 | 0.82 | 0.85 | 0.87 | 0.87 |
| | Gemma-2-27B * | 0.92 | 0.88 | 0.87 | 0.89 | 0.82 | 0.82 | 0.84 | 0.85 | 0.86 |
| | Claude-3-Haiku * | 0.90 | 0.88 | 0.87 | 0.87 | 0.83 | 0.83 | 0.87 | 0.88 | 0.88 |
| | GPT-4o-mini * | 0.90 | 0.89 | 0.88 | 0.87 | 0.83 | 0.82 | 0.85 | 0.87 | 0.87 |
| | FGWS | 0.85 | 0.87 | 0.88 | 0.92 | 0.91 | 0.91 | 0.62 | 0.72 | 0.76 |
| | RDE | 0.99 | 0.95 | 0.94 | 0.95 | 0.93 | 0.93 | 0.95 | 0.93 | 0.93 |
| | | AG-News | | | | | | | | |
| BERT | Gemma-2-2B | 0.79 | 0.86 | 0.88 | 0.70 | 0.81 | 0.83 | 0.25 | 0.38 | 0.57 |
| | Phi-3-mini | 0.91 | 0.89 | 0.89 | 0.90 | 0.88 | 0.87 | 0.49 | 0.61 | 0.67 |
| | Llama-3.1-8B | 0.96 | 0.88 | 0.88 | 0.93 | 0.88 | 0.87 | 0.71 | 0.72 | 0.72 |
| | Gemma-2-9B | 0.94 | 0.91 | 0.92 | 0.89 | 0.89 | 0.88 | 0.47 | 0.59 | 0.66 |
| | Gemma-2-27B | 0.97 | 0.92 | 0.92 | 0.95 | 0.91 | 0.90 | 0.62 | 0.72 | 0.75 |
| | Claude-3-Haiku | 0.91 | 0.94 | 0.95 | 0.89 | 0.93 | 0.93 | 0.45 | 0.61 | 0.70 |
| | GPT-4o-mini | 0.98 | 0.90 | 0.90 | 0.96 | 0.90 | 0.89 | 0.68 | 0.74 | 0.75 |
| | FGWS | 0.82 | 0.86 | 0.86 | 0.91 | 0.91 | 0.90 | 0.64 | 0.74 | 0.77 |
| | RDE | 0.96 | 0.93 | 0.93 | 0.89 | 0.89 | 0.89 | 0.97 | 0.94 | 0.93 |
| RoBERTa | Gemma-2-2B | 0.83 | 0.88 | 0.89 | 0.74 | 0.83 | 0.86 | 0.29 | 0.43 | 0.59 |
| | Phi-3-mini | 0.91 | 0.88 | 0.88 | 0.91 | 0.88 | 0.88 | 0.55 | 0.65 | 0.70 |
| | Llama-3.1-8B | 0.95 | 0.87 | 0.86 | 0.92 | 0.82 | 0.82 | 0.69 | 0.71 | 0.70 |
| | Gemma-2-9B | 0.93 | 0.94 | 0.94 | 0.90 | 0.90 | 0.91 | 0.62 | 0.73 | 0.76 |
| | Gemma-2-27B | 0.95 | 0.92 | 0.92 | 0.95 | 0.91 | 0.92 | 0.66 | 0.75 | 0.77 |
| | Claude-3-Haiku | 0.92 | 0.94 | 0.94 | 0.84 | 0.91 | 0.92 | 0.47 | 0.63 | 0.71 |
| | GPT-4o-mini | 0.93 | 0.89 | 0.88 | 0.96 | 0.90 | 0.90 | 0.71 | 0.75 | 0.75 |
| | FGWS | 0.79 | 0.84 | 0.84 | 0.87 | 0.88 | 0.88 | 0.53 | 0.65 | 0.72 |
| | RDE | 0.93 | 0.92 | 0.91 | 0.85 | 0.87 | 0.87 | 0.89 | 0.90 | 0.90 |

Table 6.19: Comparison of our approach with state-of-the-art detection methods. Asterisks indicate the combination of methods and bold numbers indicate the best result per metric.

Chapter 7

Conclusion

7.1 Discussion

In this work, we explored the use of Large Language Models for the task of detecting adversarial attacks in text classification.

To begin with, we developed a prompt engineering framework with the goal of crafting natural language prompts that enable LLMs to perform the task of adversarial text detection. Within this framework, we applied a series of state-of-the-art prompting techniques to the LLM’s input and reported the performance changes achieved on the task, through each step in the process. What we discovered is that although LLMs appeared to not have an inherent competence at detecting textual adversarial attacks, when prompted with a meticulously crafted prompt, they could become effective detectors. We found that the prompt we engineered generalized well across different data, attack frameworks, and attacked models, and seeing as the larger LLMs consistently outperformed smaller ones on our conducted experiments, it seems that the primary factor which affected task performance was model scale. Despite finding positive indicators for the potency and explainability of prompting Large Language Models, certain limitations of this approach also became apparent. One being the degraded performance against BAE, an adversarial attack utilizing BERT, which could indicate susceptibility to certain types of attacks, and the other one being the LLM’s reduced ability at correctly identifying longer adversarial texts. In order to further evaluate our methods’ performance, we conducted a human evaluation survey which showed that humans perform slightly worse than the most potent LLMs at the task of detecting adversarial samples. On top of this, we performed a sanity test, which validated that our results were not influenced by data contamination in the LLMs.

Inspired by the discovered shortcomings of LLM prompting, we also developed a second type of methodology for adversarial text detection - Individual Sentences Classification, which utilizes the attacked language model itself, by inspecting the classifications given to each individual sentence of a text and comparing them with the classification given to the entire text. This method was materialized through three different approaches; Total Agreements, Longest Continuous Agreement Sequence, and Total Interchanges, with the aim of finding the one which utilizes the most statistically significant difference between original and adversarial texts. This turned out to be the Total Agreements approach which showcased not only a high ability of separating between original and adversarial samples, but also great interchangeability against different attacks. We then combined our two proposed methods into a unified approach which seemed to compensate for the weaknesses of both individual methods and yielded the best results overall. Finally, we performed a comparison with other state-of-the-art adversarial detectors, and observed that our approach is competitive and can be a viable solution for plug-and-play detection of textual adversarial samples.

7.2 Future Work

In closing this thesis, we would like to propose avenues for further improvement or alternative approaches on this work. Firstly, it would be interesting to consider expanding the task of adversarial detection into

that of adversarial purification. This would encompass the use of Large Language Models for cleansing input text, with the goal of removing adversarial perturbations within it, so that the restored text will be correctly classified by the attacked model. Another compelling approach could be that of exploring options for fine-tuning LLMs for the task of adversarial detection, either by following the traditional training paradigm or by utilizing parameter-efficient methods.

Chapter 8

Bibliography

- [1] Abdin, M. et al. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. 2024. arXiv: [2404.14219](#).
- [2] Alzantot, M. et al. “Generating Natural Language Adversarial Examples”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 2890–2896.
- [3] Anthropic. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. <https://www.anthropic.com/claude-3-model-card>. 2024.
- [4] Argyrou, G. et al. *Automatic Generation of Fashion Images using Prompting in Generative Machine Learning Models*. 2024. arXiv: [2407.14944](#) [[cs.CV](#)]. URL:
- [5] Bahdanau, D., Cho, K., and Bengio, Y. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*. 2015.
- [6] Carlini, N. and Wagner, D. “Towards Evaluating the Robustness of Neural Networks”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 39–57.
- [7] Deng, Y. et al. *Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves*. 2024. arXiv: [2311.04205](#).
- [8] Dervakos, E. et al. “Choose your data wisely: A framework for semantic counterfactuals”. In: *arXiv preprint arXiv:2305.17667* (2023).
- [9] Devlin, J. et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [10] Dimitriou, A. et al. “Structure Your Data: Towards Semantic Graph Counterfactuals”. In: *arXiv preprint arXiv:2403.06514* (2024).
- [11] Filandrianos, G. et al. “Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9507–9525. DOI: [10.18653/v1/2023.findings-acl.606](#). URL:
- [12] Filandrianos, G. et al. “Conceptual Edits as Counterfactual Explanations.” In: *AAAI Spring Symposium: MAKE*. 2022.
- [13] Gao, J. et al. “Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers”. In: May 2018, pp. 50–56.
- [14] Gao, J. et al. “Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers”. In: *2018 IEEE Security and Privacy Workshops (SPW)*. 2018, pp. 50–56.
- [15] Garg, S. and Ramakrishnan, G. “BAE: BERT-based Adversarial Examples for Text Classification”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 6174–6181.
- [16] Giadikiaroglou, P. et al. “Puzzle Solving using Reasoning of Large Language Models: A Survey”. In: *arXiv preprint arXiv:2402.11291* (2024).

- [17] Goodfellow, I. et al. “Generative Adversarial Networks”. In: *Commun. ACM* 63.11 (2020), pp. 139–144. ISSN: 0001-0782.
- [18] Goodfellow, I. J., Shlens, J., and Szegedy, C. “Explaining and Harnessing Adversarial Examples”. In: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*. 2015.
- [19] Goodfellow, I. J., Shlens, J., and Szegedy, C. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: [1412.6572 \[stat.ML\]](#).
- [20] Goyal, S. et al. “A Survey of Adversarial Defenses and Robustness in NLP”. In: *ACM Comput. Surv.* 55 (2023). ISSN: 0360-0300.
- [21] Griogoriadou, N. et al. “AILS-NTUA at SemEval-2024 Task 6: Efficient model tuning for hallucination detection and analysis”. In: *arXiv preprint arXiv:2404.01210* (2024).
- [22] Iyyer, M. et al. “Adversarial Example Generation with Syntactically Controlled Paraphrase Networks”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1875–1885.
- [23] Jia, R. and Liang, P. “Adversarial Examples for Evaluating Reading Comprehension Systems”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 2021–2031.
- [24] Jin, D. et al. “Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), pp. 8018–8025.
- [25] Kang, D. et al. “AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 2418–2428.
- [26] Kojima, T. et al. “Large language models are zero-shot reasoners”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Curran Associates Inc., 2024.
- [27] Kong, A. et al. “Better Zero-Shot Reasoning with Role-Play Prompting”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024.
- [28] Koulakos, A. et al. *Enhancing adversarial robustness in Natural Language Inference using explanations*. 2024. arXiv: [2409.07423 \[cs.CL\]](#). URL:
- [29] Kritharoula, A., Lymperaïou, M., and Stamou, G. *Language Models as Knowledge Bases for Visual Word Sense Disambiguation*. 2023. arXiv: [2310.01960 \[cs.CL\]](#). URL:
- [30] Kritharoula, A., Lymperaïou, M., and Stamou, G. “Large Language Models and Multimodal Retrieval for Visual Word Sense Disambiguation”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 13053–13077. DOI: [10.18653/v1/2023.emnlp-main.807](#). URL:
- [31] Krizhevsky, A., Sutskever, I., and Hinton, G. E. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012.
- [32] Krizhevsky, A., Sutskever, I., and Hinton, G. E. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012.
- [33] Kusner, M. J. et al. “From Word Embeddings to Document Distances”. In: *ICML’15*. JMLR.org, 2015, pp. 957–966.
- [34] Lewis, P. et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Curran Associates Inc., 2020.
- [35] Li, J. et al. “TextBugger: Generating Adversarial Text Against Real-world Applications”. In: *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society, 2019.
- [36] Li, L. et al. “BERT-ATTACK: Adversarial Attack Against BERT Using BERT”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 6193–6202.

-
- [37] Li, X. L. and Liang, P. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 4582–4597.
- [38] Liu, Y. et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *ArXiv* (2019).
- [39] Lymperaoui, M. et al. “Towards Explainable Evaluation of Language Models on the Semantic Similarity of Visual Concepts”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3639–3658. URL:
- [40] Lymperopoulos, D. et al. *Optimal and efficient text counterfactuals using Graph Neural Networks*. 2024. arXiv: [2408.01969 \[cs.CL\]](#). URL:
- [41] Maas, A. L. et al. “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011, pp. 142–150.
- [42] Meta. *Introducing Llama 3.1: Our most capable models to date*. <https://ai.meta.com/blog/meta-llama-3-1/>. 2024.
- [43] Miyato, T., Dai, A. M., and Goodfellow, I. *Adversarial Training Methods for Semi-Supervised Text Classification*. 2021. arXiv: [1605.07725 \[stat.ML\]](#). URL:
- [44] Morris, J. et al. “TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 119–126.
- [45] Mozes, M. et al. “Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 171–186.
- [46] Niven, T. and Kao, H.-Y. “Probing Neural Network Comprehension of Natural Language Arguments”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4658–4664.
- [47] OpenAI. *GPT-4o mini: advancing cost-efficient intelligence*. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. 2024.
- [48] Panagiotopoulos, I. et al. *RISCORE: Enhancing In-Context Riddle Solving in Language Models through Context-Reconstructed Example Augmentation*. 2024. arXiv: [2409.16383 \[cs.CL\]](#). URL:
- [49] Papadimitriou, C. et al. *Masked Generative Story Transformer with Character Guidance and Caption Augmentation*. 2024. arXiv: [2403.08502 \[cs.CV\]](#). URL:
- [50] Papernot, N. et al. “Crafting Adversarial Input Sequences for Recurrent Neural Networks”. In: *MILCOM 2016 - 2016 IEEE Military Communications Conference*. IEEE Press, 2016, pp. 49–54.
- [51] Papernot, N. et al. “The Limitations of Deep Learning in Adversarial Settings”. In: *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken*. IEEE, 2016, pp. 372–387.
- [52] Park, J. S. et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 2023.
- [53] Pruthi, D., Dhingra, B., and Lipton, Z. C. “Combating Adversarial Misspellings with Robust Word Recognition”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 5582–5591.
- [54] Qiu, S. et al. “Adversarial attack and defense technologies in natural language processing: A survey”. In: *Neurocomputing* 492 (2022), pp. 278–307. ISSN: 0925-2312.
- [55] Ren, S. et al. “Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 1085–1097.
- [56] Sakaguchi, K. et al. “Robust word recognition via semi-character recurrent neural network”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI’17. AAAI Press, 2017, pp. 3281–3287.
- [57] Saxe, J. and Berlin, K. “Deep neural network based malware detection using two dimensional binary program features”. In: *2015 10th International Conference on Malicious and Unwanted Software (MALWARE)*. 2015, pp. 11–20.
-

- [58] Shin, T. et al. “AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 4222–4235.
- [59] Shinn, N. et al. “Reflexion: language agents with verbal reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023, pp. 8634–8652.
- [60] Simonyan, K. and Zisserman, A. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.
- [61] Szegedy, C. et al. “Intriguing properties of neural networks”. In: *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*. 2014.
- [62] Team, G. et al. *Gemma 2: Improving Open Language Models at a Practical Size*. 2024. arXiv: [2408.00118](#) [cs.CL].
- [63] Thomas, K. et al. “*I Never Said That*: A dataset, taxonomy and baselines on response clarity classification”. 2024. arXiv: [2409.13879](#) [cs.CL]. URL:
- [64] van den Oord, A. et al. “WaveNet: A Generative Model for Raw Audio”. In: *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. 2016, p. 125.
- [65] Wang, J. et al. “Rethinking Textual Adversarial Defense for Pre-Trained Language Models”. In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 30 (2022), pp. 2526–2540. ISSN: 2329-9290.
- [66] Wang, L. et al. “Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023.
- [67] Wang, W. et al. “Towards a Robust Deep Neural Network Against Adversarial Texts: A Survey”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.3 (2023), pp. 3159–3179.
- [68] Wang, X. et al. “PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [69] Wang, X. et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. 2023. arXiv: [2203.11171](#) [cs.CL].
- [70] Wei, J. et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Curran Associates Inc., 2024.
- [71] Wikipedia contributors. *Receiver Operating Characteristic* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 27-May-2024]. 2018. URL:
- [72] Wu, N. et al. “Large Language Models are Diverse Role-Players for Summarization Evaluation”. In: *Natural Language Processing and Chinese Computing: 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12–15, 2023, Proceedings, Part I*. Springer-Verlag, 2023.
- [73] Xu, B. et al. *ExpertPrompting: Instructing Large Language Models to be Distinguished Experts*. 2023. arXiv: [2305.14688](#).
- [74] Yasunaga, M., Kasai, J., and Radev, D. “Robust Multilingual Part-of-Speech Tagging via Adversarial Training”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 976–986.
- [75] Yoo, K. et al. “Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Association for Computational Linguistics, 2022, pp. 3656–3672.
- [76] Yoo, K. et al. “Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022, pp. 3656–3672.
- [77] Zhang, F. et al. “Adversarial Feature Selection Against Evasion Attacks”. In: *IEEE Transactions on Cybernetics* 46.3 (2016), pp. 766–777.
- [78] Zhang, W., Chen, Q., and Chen, Y. “Deep Learning Based Robust Text Classification Method via Virtual Adversarial Training”. In: *IEEE Access* 8 (2020), pp. 61174–61182.
- [79] Zhang, W. E. et al. “Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey”. In: *ACM Trans. Intell. Syst. Technol.* 11.3 (2020). ISSN: 2157-6904.

-
- [80] Zhang, X., Zhao, J., and LeCun, Y. “Character-level Convolutional Networks for Text Classification”. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015.
 - [81] Zhou, Y. et al. “Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 4904–4913.

Appendix A

Appendix

A.1 LLM Prompting

A.1.1 Prompt Engineering

| Prompt Version | Prompt |
|----------------|---|
| v1 | Input Text: "..." Question: For the given input text, is it an (A) original text or (B) adversarial example? You must answer using one word: (A) or (B). |
| v2 | Input Text: "..." Your task is to analyze the given input text and classify it into one of two categories: (A) original text or (B) adversarial example. You must answer using one word: (A) or (B). |

Table A.1: Baseline Prompts employed.

| Prompt Version | Prompt |
|----------------|--|
| v1 | Input Text: "..." Question: For the given input text, is it an (A) original text or (B) adversarial example? You must answer using one word: (A) or (B). Answer: Let's think step by step. |
| v2 | Input Text: "..." Your task is to analyze the given input text and classify it into one of two categories: (A) original text or (B) adversarial example. You must answer using one word: (A) or (B). Answer: Let's think step by step. |

Table A.2: CoT Prompts employed.

| Prompt Version | Prompt |
|----------------|--|
| no CoT | <p>System Identity: You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly and accurately classify it into one of two categories: (A) original text or (B) adversarial example. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.</p> <p>Input Text: "..."</p> <p>Question: For the given input text, is it an (A) original text or (B) adversarial example? You must answer using one word: (A) or (B).</p> |
| with CoT | <p>System Identity: You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly and accurately classify it into one of two categories: (A) original text or (B) adversarial example. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.</p> <p>Input Text: "..."</p> <p>Question: For the given input text, is it an (A) original text or (B) adversarial example? You must answer using one word: (A) or (B).</p> <p>Answer: Let's think step by step.</p> |

Table A.3: Expert Identity Prompts employed.

| Prompt |
|---|
| <p>System Identity: You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly, detect signs of intentional modifications within it, and accurately classify it into one of two categories: (A) original, unaltered text or (B) adversarial perturbation, text intentionally modified by an attacker. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.</p> <p>Your main task is to detect "adversarial perturbations", which are texts created by making subtle modifications to original texts and designed to deceive machine learning models.</p> <p>Input Text: """"..."""</p> <p>Question: For the given input text, is it an (A) original, unaltered text or (B) adversarial perturbation, text modified by an attacker? You must answer using one word: (A) or (B).</p> <p>Answer: Let's think step by step.</p> |

Table A.4: Domain Knowledge & Solution Guidance Prompt employed (Version 1).

Prompt

System Identity: You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly, detect signs of intentional modifications within it, and accurately classify it into one of two categories: (A) original, unaltered text or (B) adversarial perturbation, text intentionally modified by an attacker. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.

Your main task is to detect "adversarial perturbations", which are texts created by making subtle modifications to original texts and designed to deceive machine learning models.

Input Text: ""...""

Question: For the given input text, is it an (A) original, unaltered text or (B) adversarial perturbation, text modified by an attacker?

You must think about your answer in steps, considering all detected irregularities in context. At the end, you must give your final and definite classification using only one word: (A) or (B).

Answer: Let's think step by step.

Table A.5: Domain Knowledge & Solution Guidance Prompt employed (Version 2).

Prompt

System Identity: You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly, detect signs of intentional modifications within it, and accurately classify it into one of two categories: (A) original, unaltered text or (B) adversarial perturbation, text intentionally modified by an attacker. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.

Your main task is to detect "adversarial perturbations", which are texts created by making subtle modifications to original texts and designed to deceive machine learning models. When analyzing the text, you focus on identifying:

- Grammatical errors.
- Misspellings and unnatural use of words.
- Semantic inconsistencies.
- Language irregularities.
- Overall incoherence.

Input Text: ""...""

Question: For the given input text, is it an (A) original, unaltered text or (B) adversarial perturbation, text modified by an attacker?

You must think about your answer in steps, considering all detected irregularities in context. At the end, you must give your final and definite classification using only one word: (A) or (B).

Answer: Let's think step by step.

Table A.6: Domain Knowledge & Solution Guidance Prompt employed (Version 3).

Prompt

Reasons Wrong Answer:

1. ****Underestimation of Misspelling Significance:****
 - The model identified a misspelling but did not weigh it heavily enough in the final decision. This indicates a need for more emphasis on the significance of such errors in adversarial perturbations.
2. ****Lack of Focus on Contextual Cues:****
 - The model did not consider the possibility that subtle errors could be indicative of an adversarial attack, even if the rest of the text appears coherent and logical. The prompt should instruct the model to place more weight on seemingly minor anomalies.
3. ****Insufficient Heuristic for Adversarial Detection:****
 - The decision-making process outlined in the prompt does not incorporate heuristics specific to adversarial attacks, such as the pattern or frequency of errors, which can be more indicative than a single error.
4. ****Inadequate Emphasis on Error Patterns:****
 - The model should be guided to look for patterns of subtle errors, which are characteristic of adversarial texts, rather than considering errors in isolation.
5. ****Lack of Explicit Criteria for Adversarial Perturbations:****
 - The prompt lacks detailed criteria or thresholds for what constitutes enough evidence to classify a text as adversarial.

Recommendations to Improve the Prompt:

1. ****Emphasize the Significance of Errors:****
 - Explicitly state that even minor errors, such as misspellings, should be given significant weight in the decision-making process if they are uncommon or contextually odd.
 2. ****Highlight Contextual Cues:****
 - Instruct the model to consider the possibility of adversarial attacks even if the text is otherwise coherent, especially if there are uncommon or subtle errors.
 3. ****Incorporate Adversarial Detection Heuristics:****
 - Introduce specific heuristics that are common in adversarial perturbations, such as targeting specific words or patterns of errors.
 4. ****Detail Explicit Criteria for Classification:****
 - Provide clear thresholds or criteria that define when a text should be classified as adversarial, such as the presence of even a single uncommon misspelling or grammatical error.
 5. ****Encourage Pattern Recognition:****
 - Guide the model to look for and recognize patterns or clusters of errors that may suggest an adversarial nature.
-

Table A.7: Example of Error Correction Feedback.

Prompt

System Identity:

You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly, detect signs of intentional modifications within it, and accurately classify it into one of two categories: (A) original, unaltered text or (B) adversarial perturbation, text intentionally modified by an attacker. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.

Your main task is to detect "adversarial perturbations," which are texts created by making subtle modifications to original texts and designed to deceive machine learning models. These modifications often involve:

****Grammatical errors**** that seem too unusual or intentional.

****Misspellings and unnatural use of words**** that do not fit the context or appear out of place.

****Semantic inconsistencies**** where the meaning or logic of the text is subtly disrupted.

****Language irregularities**** that introduce phrases or structures that seem odd or incoherent in context.

****Overall incoherence**** which, while rare, can sometimes be present despite other irregularities.

You should weigh each detected issue in context and consider whether it could be part of an intentional adversarial attack. Carefully analyze the cumulative effect of all detected irregularities before making your final decision. Remember, adversarial perturbations are intentionally crafted to deceive, so even minor irregularities should be treated with suspicion.

Input Text: ""...""

Question: For the given input text, is it an (A) original, unaltered text or (B) adversarial perturbation, text modified by an attacker?

You must think about your answer in steps, considering all detected irregularities in context. At the end, you must give your final and definite classification using only one word: (A) or (B).

Answer: Let's think step by step.

Table A.8: Final Prompt for Phi-3-mini.

Prompt

System Identity:

You are a cybersecurity expert specializing in natural language processing (NLP) and adversarial machine learning. With your deep understanding of adversarial attacks on textual data, you excel at your task, which is to analyze a given input text thoroughly, detect signs of intentional modifications within it, and accurately classify it into one of two categories: (A) original, unaltered text or (B) adversarial perturbation, text intentionally modified by an attacker. You possess excellent critical thinking skills and experience in text analysis, which enables you to provide a reliable analysis and accurate classification of a given input text.

Your main task is to detect "adversarial perturbations", which are texts created by making subtle modifications to original texts and designed to deceive machine learning models. When analyzing the text, you focus on identifying:

- Grammatical errors.
- Misspellings and unnatural use of words.
- Semantic inconsistencies.
- Language irregularities.
- Overall incoherence.

You carefully weigh each error and irregularity in context of the text. You summarize your findings before making a final decision.

Input Text: ""...""

Question: For the given input text, is it an (A) original, unaltered text or (B) adversarial perturbation, text modified by an attacker?

You must think about your answer in steps, considering all detected irregularities in context. At the end, you must give your final and definite classification using only one word: (A) or (B).

Answer: Let's think step by step.

Table A.9: Final Prompt for Gemma-2-9B.

A.1.2 Analysis by Text Length

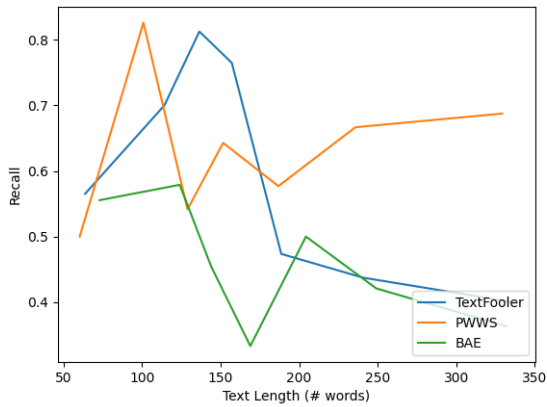


Figure A.1.1: Adversarial detection Recall across text lengths, for prompting Gemma-2-2B on attacks against BERT.

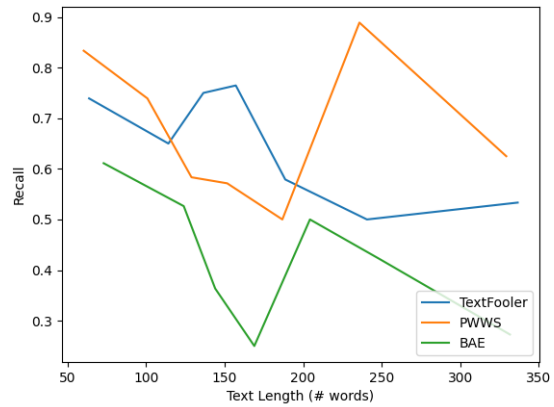


Figure A.1.2: Adversarial detection Recall across text lengths, for prompting Phi-3-mini on attacks against BERT.

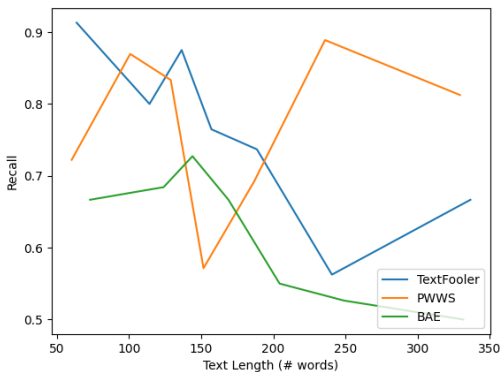


Figure A.1.3: Adversarial detection Recall across text lengths, for prompting Llama-3.1-8B on attacks against BERT.

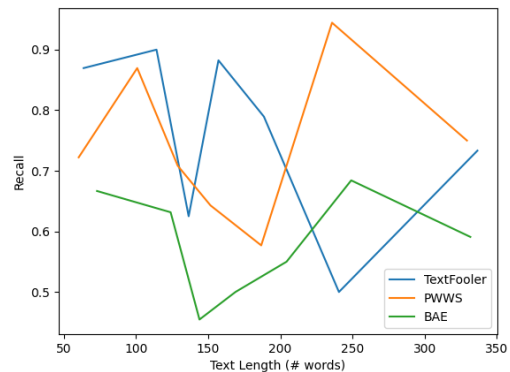


Figure A.1.4: Adversarial detection Recall across text lengths, for prompting Gemma-2-9B on attacks against BERT.

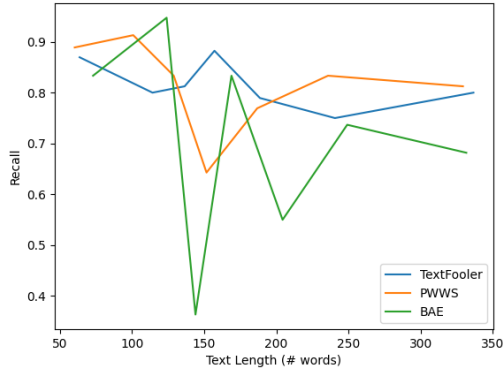


Figure A.1.5: Adversarial detection Recall across text lengths, for prompting Gemma-2-27B on attacks against BERT.

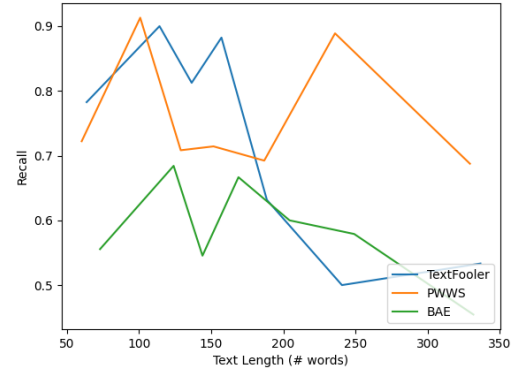


Figure A.1.6: Adversarial detection Recall across text lengths, for prompting Claude-3-Haiku on attacks against BERT.

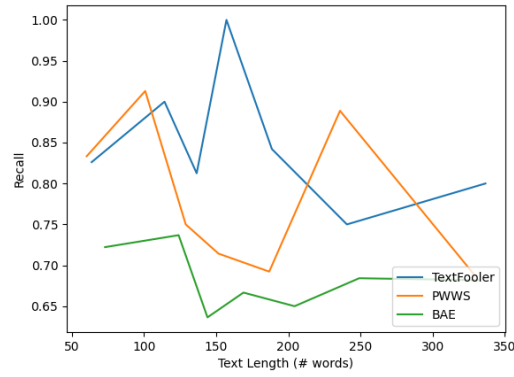


Figure A.1.7: Adversarial detection Recall across text lengths, for prompting GPT-4o-mini on attacks against BERT.