



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

**SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION**

BSc THESIS

**Fake News Detection with the GREEK-BERT
Model with a focus on COVID-19**

Dimosthenes P. Fioretos

Supervisors:

**Manolis Koubarakis, Professor
Despina - Athanasia Pantazi, PhD Candidate
Christos Papadopoulos, PhD Candidate**

ATHENS

NOVEMBER 2021



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Αναγνώριση Ψευδών Ειδήσεων με το μοντέλο GREEK-BERT
με εστίαση στον COVID-19**

Δημοσθένης Π. Φιορέτος

Επιβλέποντες: Μανώλης Κουμπάρακης, Καθηγητής
Δέσποινα – Αθανασία Πανταζή, Υποψήφια Διδάκτωρ
Χρήστος Παπαδόπουλος, Υποψήφιος Διδάκτωρ

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2021

BSc THESIS

Fake News Detection with the GREEK-BERT
Model with a focus on COVID-19

Dimosthenes P. Fioretos

S.N.: 1115200300248

SUPERVISORS: **Manolis Koubarakis**, Professor
Despina - Athanasia Pantazi, PhD Candidate
Christos Papadopoulos, PhD Candidate

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Ψευδών Ειδήσεων με το μοντέλο GREEK-BERT
με εστίαση στον COVID-19

Δημοσθένης Π. Φιορέτος

A.M.: 1115200300248

ΕΠΙΒΛΕΠΟΝΤΕΣ: Μανώλης Κουμπάρκης, Καθηγητής
Δέσποινα – Αθανασία Πανταζή, Υποψήφια Διδάκτωρ
Χρήστος Παπαδόπουλος, Υποψήφιος Διδάκτωρ

ABSTRACT

Fake news, while being a problem appearing since the ancient times, is one of the major political and societal issues of recent years. The issue becomes even more important by the prevalence of social media use by the general public. Especially during the COVID19 pandemic, fake news dissemination can have very serious and even fatal side effects for societies as well as individuals.

This thesis outlines our work in creating two classification models for fake news and fake social media posts, alongside a web application for studying the relationships and dissemination patterns of fake and non fake information in social media platforms. Our work is target at the Greek language and the ongoing coronavirus pandemic.

We also present an overview of the research work on which we base our models, as well as related research endeavors regarding fake news detection.

For this purpose we have reused an existing Greek fake news data set, which was part of Odysseas Trispiotis' Master in Science Thesis [\[1\]](#), and we have also created a novel data set for the purposes of this project. In the process of generating this novel data set, we have observed that finding reliable fake post sources is a hard problem, even more so to automate it. The basis of the our classification models are the state of the art BERT [\[2\]](#) and GREEK-BERT [\[3\]](#) models.

The results of the above process were very encouraging, as the final classification models reached accuracy levels greater than 90%, with similarly good scores for other traditional classification metrics, such as precision, recall, f1 score and AUROC.

SUBJECT AREA: machine learning, natural language processing, automatic data tagging

KEYWORDS: fake news detection, BERT, GREEK-BERT

ΠΕΡΙΛΗΨΗ

Οι ψευδείς ειδήσεις, αν και είναι ένα πρόβλημα που παρουσιάζεται από του αρχαίους καιρούς, είναι ένα από τα κύρια πολιτικά και κοινωνικά προβλήματα τα τελευταία χρόνια. Το πρόβλημα γίνεται ακόμα μεγαλύτερο λόγω της διείσδυσης των κοινωνικών δικτύων σε μεγάλο μέρος του πληθυσμού. Ειδικότερα κατά την διάρκεια της πανδημίας του COVID19, η διασπορά ψευδών ειδήσεων μπορεί να έχει πολύ σοβαρές και ακόμα και θανάσιμες παρενέργειες για τις κοινωνίες και τους πολίτες.

Η παρούσα εργασία περιγράφει την δουλειά γύρω από την δημιουργία δυο μοντέλων κατηγοριοποίησης ψευδών ειδήσεων και ψευδών αναρτήσεων κοινωνικών δικτύων, μαζί με μια διαδικτυακή εφαρμογή για την μελέτη των σχέσεων και των μοτίβων διάδοσης ψευδών και αληθών πληροφοριών σε πλατφόρμες κοινωνικής δικτύωσης. Η δουλειά μας χρησιμοποιεί την Ελληνική γλώσσα και στοχεύει σε πληροφορίες που έχουν σχέση με την τρέχουσα πανδημία του κορωνοϊού.

Επίσης παρουσιάζουμε μια επισκόπηση των ερευνών πάνω στις οποίες βασίζουμε τα μοντέλα μας, καθώς και άλλες έρευνες σχετικές με την αναγνώριση ψευδών ειδήσεων.

Για αυτό το σκοπό, επαναχρησιμοποιήσαμε ένα προϋπάρχον Ελληνικό σύνολο δεδομένων, το οποίο ήταν μέρος της Διπλωματικής του Οδυσσέα Τρισπιώτη [1], και επίσης δημιουργήσαμε ένα νέο σύνολο δεδομένων για τους σκοπούς αυτού του έργου. Κατά την διάρκεια της δημιουργίας αυτού του νέου συνόλου δεδομένων, παρατηρήσαμε πως η εύρεση αξιόπιστων πηγών ψευδών αναρτήσεων είναι ένα δύσκολο πρόβλημα, που γίνεται ακόμα δυσκολότερα αυτοματοποιήσιμο. Η βάση για τα μοντέλα κατηγοριοποίησης που αναπτύξαμε είναι τα μοντέλα τεχνολογίας αιχμής BERT [2] και GREEK-BERT [3].

Τα αποτελέσματα της άνωθι διαδικασίας ήταν εξόχως ενθαρυντικά, καθώς τα τελικά μοντέλα κατηγοριοποίησης έφτασαν accuracy επιπέδου μεγαλύτερου του 90%, και εξίσου καλά αποτελέσματα σε άλλες παραδοσιακές μετρικές κατηγοριοποίησης δεδομένων, όπως precision, recall, f1 score και AUROC.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: μηχανική μάθηση, επεξεργασία φυσικής γλώσσας, αυτόματη σήμανση δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: αναγνώριση ψευδών ειδήσεων, μοντέλο BERT, μοντέλο GREEK-BERT

This thesis is dedicated to the loving memory of my parents, for their love, endless support and encouragement.

ACKNOWLEDGMENTS

I would like to thank my professor and supervisor Manolis Koubarakis for offering me the opportunity to work on this worthwhile endeavor and his cooperation and valuable contribution on fulfillment of this thesis. I would also like to thank my research supervisors Christos Papadopoulos and Despina Athanasia Pantazi for their constructive suggestions during the planning and implementation of this research work.

CONTENTS

PREFACE.....	23
1. INTRODUCTION.....	25
2. RELATED WORK.....	27
2.1 BERT.....	27
2.2 GREEK-BERT.....	29
2.3 Fake News in the 21st Century.....	29
2.3.1 The Social Sciences Perspective.....	30
2.3.2 The Computer Science Perspective.....	33
3. DEVELOPMENT OF GREEK-BERT FAKE NEWS DETECTION MODEL.....	38
3.1 Technologies Used.....	39
3.1.1 Software.....	39
3.1.2 Hardware.....	42
3.2 Dataset Creation.....	43
3.2.1 Pre Existing Dataset.....	43
3.2.2 Novel Dataset.....	43
3.3 Model Development.....	48
4. EXPERIMENTAL RESULTS.....	50
4.1 Metrics Definitions.....	50
4.2 Models Performance.....	52
4.2.1 Fake News Model.....	52
4.2.2 Fake Tweets Model.....	54
5. WEB APPLICATION.....	56
5.1 Search Options.....	56
5.2 Data Retrieval Options.....	57
.....	60
6. CONCLUSION.....	61
REFERENCES.....	62

LIST OF IMAGES

Figure 1: Wardle's Fake Information Classification.....	31
Figure 2: Warlde's Types of Fake News.....	32
Figure 3: Overall Project Architecture.....	39
Figure 4: Confusion Matrix Definition and Related Performance Metrics.....	50
Figure 5: F1 Measure Definition.....	51
Figure 6: AUROC Curve Sample.....	52
Figure 7: First Model Confusion Matrix.....	53
Figure 8: First Model AUROC Curve.....	53
Figure 9: Second Model Confusion Matrix.....	54
Figure 10: Second Model AUROC Curve.....	55
Figure 11: Web Application User Interface.....	57
Figure 12: Dynamic Graph Results Example.....	59
Figure 13: Static Image Results Example.....	60

LIST OF TABLES

Table 1: Performance metrics for the fake news model.....	53
Table 2: Performance metrics for the fake tweets model.....	54

PREFACE

This thesis is part of my obligations as an undergraduate student in the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens, Greece, under the supervision and guidance of professor Manolis Koubarakis.

This research project focuses on applying state of the art natural language processing techniques, namely BERT and GREEK-BERT, in the use of fake news classification with focus on the web and the COVID19 pandemic. The principal incentive behind this venture is our intention to provide a better means of understanding of the information available on the Greek web, the dissemination of fake news, as well as the betterment of public awareness with regards to the global pandemic.

I hope you enjoy reading this thesis.

1. INTRODUCTION

Multiple definitions have been proposed for Artificial Intelligence, commonly abbreviated as AI. According to the creator of the concept, John McCarthy, artificial intelligence can be defined as "the science and engineering practices of creating intelligent machines" [4]. Another commonly accepted definition, is that "artificial intelligence is the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment" [5]. Machine learning, commonly abbreviated as ML, is the branch of artificial intelligence that is concerned with "programming computers to learn from experience should to eventually eliminate the need for much of detailed programming effort" [6]. Natural language processing, commonly abbreviated as NLP, is the branch of artificial intelligence and machine learning that studies the understanding and interpretation of human language. It allows computers to communicate with humans, read written language, process speech, analyze and understand the underlying information. Natural language processing is built upon the scientific fields of Linguistics, Computer Science and Mathematics.

Fake news can be defined as "news that conveys or incorporates false, fabricated, or deliberately misleading information, or that is characterized as or accused of doing so" [7]. However, there isn't a commonly accepted definition for the term. In general, "fake news" is not a simple concept. Different research endeavors, have proposed a plethora of alternative definitions for fake news [8,9,10]. For example, in one research [8] the concept of "deliberately misleading the reader" has been added to the definition. The more general concept of misinformation has been observed since ancient times, when Plutarch, Ctesias, Aristotle, and Thucydides have accused Herodotus as "liar" and "mythologist", for inaccuracies in the narrative of historical events that have happened before his time [11,12].

In the modern world, more than 70% of United States adult citizens are getting their news via social media networks [13], while the corresponding statistic for the European Union is 53% [14]. According to research, the very mathematical nature of social networks as graphs is favorable to the quick dissemination of information [15]. and the spread of fake news in social networks is faster and wider for every category of information, in contrast to that of real news [16].

Consequently, recognizing and neutralizing fake news is a critical problem for modern society. Artificial intelligence can offer automated and quick methods of fake news detection. The present thesis focuses on classifying news and social media posts as either real or fake, regarding the global coronavirus pandemic. As the foundation of the classification models, the Google's BERT (Bidirectional Encoder Representations from Transformers) [2] architecture has been selected and its applied translation to the Greek language, by the GREEK-BERT model [3].

BERT is an architecture built upon neural networks, for natural language processing, that was developed by the Google corporation. It is used to generate models of the human language, that take into consideration not only the words that are being used, but also the relationships between the words themselves. For example, in the phrases "the road forks in one kilometer", "the personal table fork was most likely invented in the Eastern Roman Empire" and "a GitHub fork is a copy of a repository that is stored in your account", the word "fork" has very different interpretations and this is the break through that the BERT

model brought to natural language processing. When the BERT model was first published, in 2018, it proved to have the best performance results than any other pre-existing model in a series of standardized scientific tests and metrics [17].

This document comprises of five main chapters, including this introductory chapter. This is the list of chapters and an overview of their contents:

- Chapter 2: Related Work

In this chapter we discuss the relevant research work which has been the bedrock upon which we've built our work. We explain the operating principals of the BERT model, alongside its Greek transfiguration GREEK-BERT. We also cover other research endeavors regarding fake news, mainly regarding machine learning and natural language processing, but also from a societal and journalistic perspective.

- Chapter 3: Model Development

In this chapter we cover the work of building two new models for classifying fake news articles and fake information posts in social media, with focus on the twitter platform and the coronavirus pandemic. We briefly describe the mode of operation of twitter with regards to our data set generation and go into details on how the data set was gathered and post processed before being fit for being fed to our natural language processing machine learning models for training / fine tuning.

- Chapter 4: Experimental Results

In this chapter we go over the performance metrics of classification models and present the testing results of our models.

- Chapter 5: Web Application

In this chapter we showcase the web application that was built for using our classification models, by going over its user interface and how a user can interact with it and the output types ti provides.

- Chapter 6: Conclusion

In this chapter we elaborate on our conclusions and final results, with respect to what we set out to do in the start of this project and how we achieved it.

2. RELATED WORK

In this chapter, we firstly aim to highlight the way that lead to the developmen of BERT and then analyze how BERT changed the NLP landscape. Afterward, we will present the Greek application of the BERT architecture in the GREEK-BERT model and how it can be used to solve NLP tasks in the Greek language, a research space that is currently expanding. Finally, we will also talk about the fake news phenomenon, and cover the research space around it, from both an AI and journalistic perspective.

2.1 BERT

2018 was a very important year for the field of natural language processing, due to the publication of the BERT architecture. Even though the state of the art performance that resulted by its application has since been improved even further, one bit at a time, this new way of thinking about NLP models brought new perspectives to the NLP research community.

BERT utilized two cornerstone elements of earlier researchers, which have since been used for many new advances in the field of natural language processing: the transformer architecture and unsupervised pre-training. The transformer architecture [17] is a sequence model that ignores the recurrent structure of Recurrent Neural Networks [18] in favor for a fully attention based approach [19]. BERT is also pre trained, meaning its model weights are created in advance through the training on two "fake tasks": language modeling and next sentence prediction.

Language modeling, also known as masked word prediction, is the task of predicting the correct word in a sentence, by removing a specific amount of words (specifically in the case of Bert, 15%). Next sentence prediction is the task of calculating the probability whether given two sentences, one can follow the other. The term fake is used allegorically, because the two tasks BERT is trained upon are not the actual tasks that the model will be used for, but it is a task for which a plethora of perfect quality training data exist. BERT is trained on unlabeled data extracted by the BooksCorpus [20] with 800 million words and the English version of Wikipedia with 2.5 billion words. Once the training is finished, the final layers of the neural network are removed, as they were specific to the two "fake" tasks, and then we can keep the rest of the weighted neurons to apply them on tasks we actually care about solving. Thus, BERT doesn't have to be trained each time from scratch for every new task, but merely fine tuning it to a specific task dependent data set is enough. Coming up, and applying, the notion of "fake tasks" was one of the main drivers of BERT's success. In machine learning, the idea of using knowledge gained by solving one problem, in order to solve a different problem, is called transfer learning.

The underlying transformer architecture of BERT, is a class of neural networks has shown great success in the NLP field. It adopts a stacked encoder-decoder architecture, proposed in [19]. The encoder has multiple encoding layers stacked onto each other that process the input sequence iteratively, while the decoder is comprised of multiple decoding layers, of the same number, that work in a similar way on the encoders output. Internally, the encoders and decoders do not share the same exact architecture. The encoders firstly contain a self attention layer, which looks at other tokens in the input sequence before encoding a specific word, and then the output of the self attention layer is

fed through a feed forward neural network. On the other hand, the decoder, has the same first layer, a self attention layer, but before the feed forward neural network, exists an extra attention layer that helps the decoder to focus on the important parts of the input sequence, as it was encoded by the encoding layer. One of the more important features of the transformer architecture, is that the feed forward layers can be computed in parallel, thus reducing the overall training time and allowing for bigger models to be trained, something that BERT took advantage of.

One of the most important features of the transformer architecture is self attention. First, lets start with an example. In the sentence "the rock rolled down the hill because it was round", does the "it" refer to the rock or the hill? This might be a trivial question for a human, but there is no direct way to express it algorithmically. The purpose of the self attention mechanism, is to associate the "it" token with the "rock" token. Attention is a mechanism that allows the neural network to determine which are the important parts in the input sequence, regarding the task at hand. What "self" brings to "self attention", is that only the sequence itself is examined, when determining the attention to assign to each token in the sequence. The transformer architecture, further refined the self-attention layer by adding a mechanism that is called "multi headed attention". What this does, is that it extends the model's abilities by allowing it to extract more than one attention pattern from the same sequence, thus capturing more context information from the input.

The final major idea that the BERT model was built upon, was that in order to capture the true meaning of a word, context matters. Previous word embedding representations, would map each word to a unique vector, regardless of context. As previously stated, the same word can have very different meaning, depending on the surrounding context, such as for example the word "bat" in the phrases "baseball bat" and "bitten by a bat". The solution that was introduced was to assign embeddings to words based on the context they are used in, in order to capture both the word's meaning and the context that it appears in [21].

The success of the transformer architecture, the adoption of transfer learning and the contextualized word embeddings, gave rise to a new class of models. With the usage of pre-trained language embeddings, the OpenAI transformer [22] re-used the transformer architecture, but completely remove the encoder stack, keeping only the decoders. As there are no longer encoders in the transformer, the intermediate second self attention layer has also been removed from the decoder architecture.

With all preceding research advancements, the groundwork has been laid for BERT's research team. The final missing piece, was that between ELMo and GPT-2, the bidirectionality of the language model has been lost. While ELMo's architecture has been using a bidirectional language model, GPT-2 architecture used only a feed-forward language model. BERT reinstated the bidirectionality of the language model, themselves claiming in [2] that "the bidirectional nature of our model is the single most important new contribution". BERT is using an encoder-only transformer architecture and reintroduced bidirectionality by its self-attention layer, which applies attention on both directions of the input. Future versions of GREEK-BERT plan to also include the entire corpus of the Greek legislation and the entire corpus of the EU legislation (as translated in Greek), in the training corpora.

BERT has built in support only for the English language and comes in two different configurations. The BERT-BASE model, with 12 encoders with 12 bidirectional self-

attention heads, and BERT-LARGE, with 24 encoders and 16 bidirectional self attention heads. Both models come in case sensitive and insensitive editions, generating a total of four different models.

2.2 GREEK-BERT

The most constraining factor of BERT is the language of its vocabulary, as it is based on English corpora. This has draw back has been rectified by the work of the Natural Language Processing Group of the Athens University of Economics and Business, in [3]. GREEK-BERT is a monolingual BERT based language model for the modern Greek language. GREEK-BERT uses the architecture of BERT-BASE-UNCASED, which is the lighter version of BERT that is trained on lower case words. The model is pretrained on 29 GB of text form the following corpora:

1. the Greek part of Wikipedia
2. the Greek part of the European Parliament Proceedings Parallel Corpus (Europarl)
3. the Greek part of OSCAR, a clean version of Common Crawl

Accents and other diacritics were removed and all words were converted to lower case, in order to provide the widest possible normalization. This corpora has been also used to extract a vocabulary of 35.000 BPEs -a compressed binary representation format [23]-.

Most work on transfer learning for languages other than English focuses on multilingual language modeling, in order to cover multiple languages ata once. In order to cope with multiple languages, such models rely on extended shared covabularies, which results to an underrepresentation of the Greek language, mainly because of the unique Greek alphabet. In contrast, languages that follow the Latin alphabet, such as English, French and Spanish, among others, share most sub-words, which in practice improves the overall model's performance. The representation of the Greek language in such multilingual models was less than 2% of the total vocabulary size. Since it has been observed in various research endeavors that monolingual language models perform better on most downstream tasks, GREEK-BERT is a great step in the right direction for the Greek NLP field.

With regards to performance, GREEK-BERT has been benchmarked on three common NLP downstream tasks, Part of Speech tagging, Named Entity Recognition and Natural Language Inference, against various strong baseline models. For the PoS tagging benchmark, all models performed very well, with accuracy scores over 97%. GREEK-BERT was second, with an accuracy score of 98.02 - 98.18, marginally worse than XLM-R with a score of 98.13 - 98.27. On the NER benchmark, GREEK-BERT outperforms all other models, by a large margin in most cases, with an accuracy of 84.7 - 86.7. In the third and final benchmark, that of NLI, GREEK-BERT again has the best performance, by a significant margin in comparisson to other models, with an accuracy of 77.98 - 79.2. Overall, GREEK-BERT offers state of the art performance on multiple NLP downstream tasks in the Greek language.

2.3 Fake News in the 21st Century

News media have been studied by a wide variety of scientific disciplines, such us Psychology, Political Science, Journalism/Communication and Economics, among others. Since the start of the Digital Age in the middle of the 20th century, and especially the birth

of the World Wide Web in 1989 by English scientist Tim Berners-Lee, which revolutionized the way human societies exchange information, the nature of news production and consumption has shifted en masse to the digital mediums and especially the web. This new landscape, caught the attention of Computer Science as well, as it is the main driver behind the technologies used for news media generation and delivery. The interest of Computer Scientists is targeted in both the underlying operating principles of the technologies used, such as communication networks and publication platforms, as well as the consumption, dissemination and interaction of the content itself with individuals and societies, as part of Digital Capitalism [24] and especially Surveillance Capitalism [xxx], since Information Theory and Computer Science are the driving force behind the commodification of personal data for the core purpose of profit making via online advertising companies.

2.3.1 The Social Sciences Perspective

2.3.1.1 The Digital Revolution

Traditionally, news articles were written only by professional journalists working either for news outlets, or independently. With the arrival of digital mass media, everyone with access to the World Wide Web could be a news editor. News consumption also shifted a lot in the last two decades, as the amount of people that get their news from the web has been steadily increasing and has overtaken that of those getting them from traditional, such as printed sources or television, for audiences younger than 30 years old [25], and the age limit is has an upward trend. With 78.1% of the Greek population having access to the internet by 2020 [26]. Already by 2017, 66.3% of Greek WWW citizens were using the internet as a source for news[27], and it would be safe to assume that this percentage has only risen in the last four years, as this is a global trend [28]. The final and possibly most important change in the news consumption and dissemination landscape, is the same that the internet has already revolutionized: instant worldwide communication. This way, news articles can spread with -literally- the speed of light, bringing it to more people and more quickly, than ever before.

2.3.1.2 Fake News Definition

A wide range of efforts has tried to exactly define and capture the essence of what exactly are fake news. Among them, one that tries to encapsulate not only the origin but also the target of fake news, has been proposed by Wardle, Derakhshan et. al [29]. It introduces the concept of intent, besides that of conceptual validity, and classifies fake news into three sub categories:

1. dis-information: information that is false and deliberately created to harm a person, social group, organization or country.
2. mis-information: information that is false, but not created with the intention of causing harm.
3. mal-information: information that is based on reality, used to inflict harm on a person, social group, organization or country.

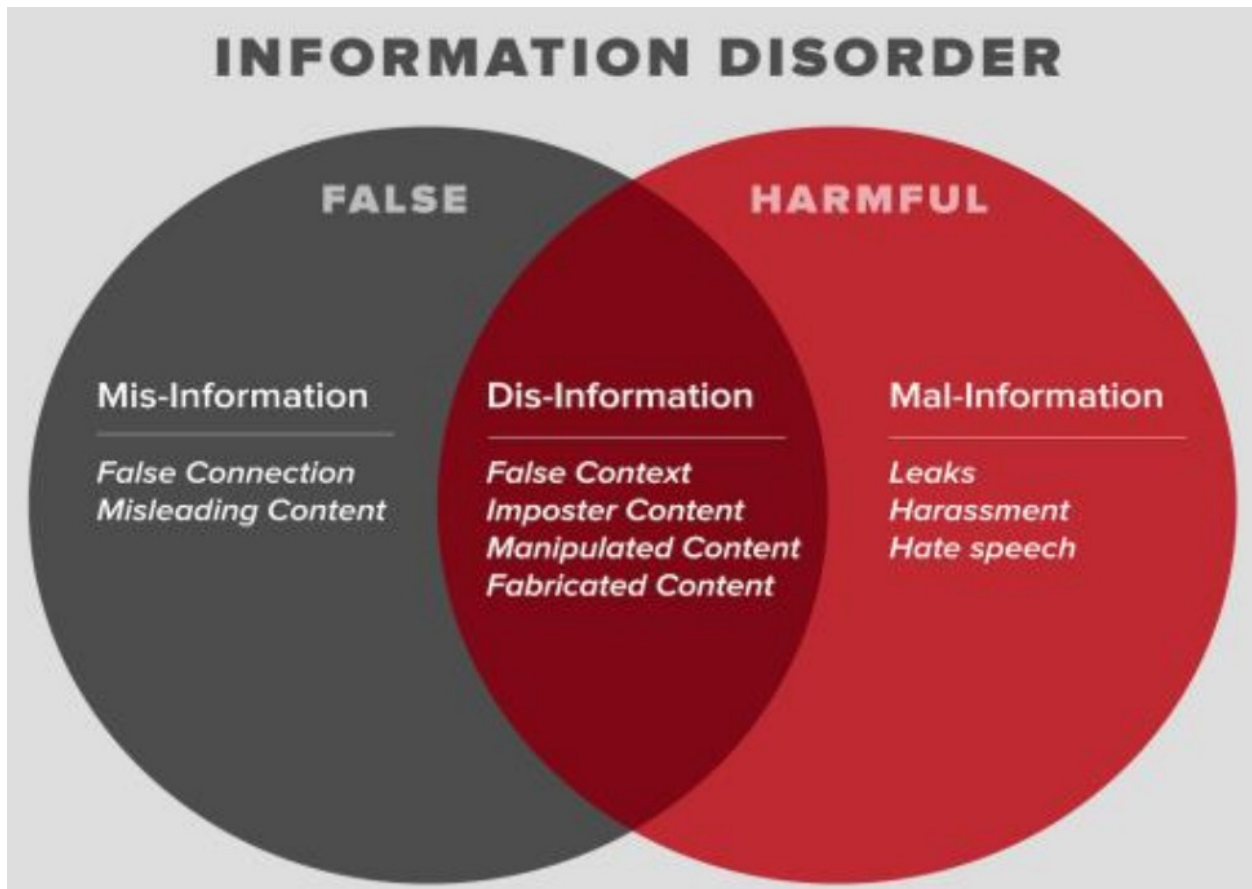


Figure 1: Wardle's Fake Information Classification

A further refined topology of types of fake news has been proposed by Wardle [30], with respect to the measure of the intent of the information producer to deceive:

1. Satire or Parody: no intention to cause harm, but has potential to fool
2. Misleading Content: misleading use of information to frame an issue or individual
3. Imposter Content: when genuine sources are impersonated
4. Fabricated Content: new contents is 100% false, designed to deceive and do harm
5. False Content: when headlines, visuals or captions don't support the content
6. False Context: when genuine content is shared with false contextual information
7. Manipulated Context: when genuine information or imagery is manipulated to deceive

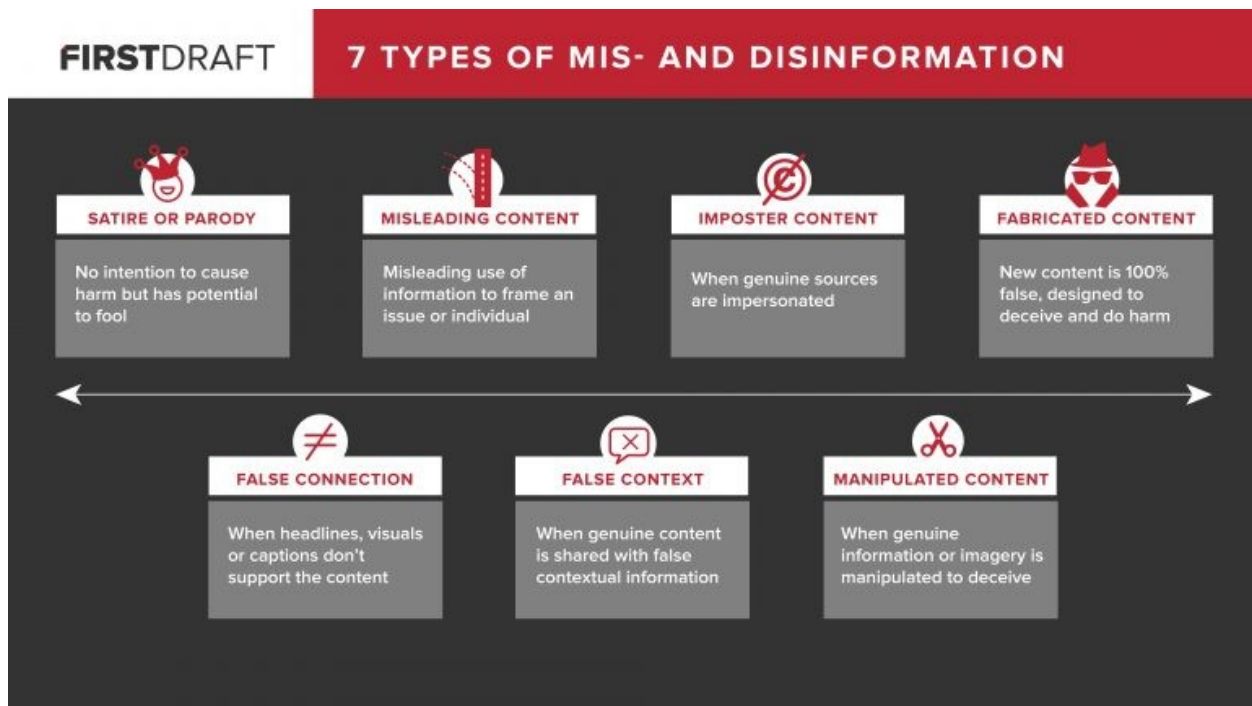


Figure 2: Warlde's Types of Fake News

2.3.1.3 Effects on Society

Permeation of fake news in our daily lives has followed the same explosive growth that the internet and the World Wide Web have brought to personal communication. While the concept of misinformation has been observed since ancient times, and has been made extremely more common in the post industrialized world, the modern fake news phenomenon is now globally viewed as one of the greatest threats to democracy and journalism [31], due to the aforementioned multiplicative effect of technology. The era of social media has been called the era of "post truth" [32]. As explained by Higgins [33], the term "post truth" describes not only an increase in the frequency of lies in public discourse, but also characterizes a society in which truth is no longer neither the norm nor an expectation. This reality has been foreshadowed by Friedrich Nietzsche in his 1873 essay "Truth and Lying in an Extra-Moral Sense":

"If someone hides an object behind a bush, then seeks and finds it there, that seeking and finding is not very laudable: but that is the way it is with the seeking and finding of "truth" within the rational sphere. If I define the mammal and then after examining a camel declare, "See, a mammal," a truth is brought to light, but it is of limited value. I mean, it is anthropomorphic through and through and contains not a single point that would be "true and universally valid, apart from man. The investigator into such truths is basically seeking just the metamorphosis of the world into man; he is struggling to understand the world as a human-like thing and acquires at best a feeling of assimilation."

In other words, Nietzsche argues that we humans create the concepts through which we define correctness and justice, thereby replacing the concept of objective truth, with the concept of subjective value, and defining reality in terms of the human will.

Another important feature of the fake news problem, is the persistence of fake information in humans [34]. It has been empirically observed that exposure to correction after the exposure to false information, does not always result in correct outcomes. An extension of this cognitive explanation for the resilience of misinformation against correction is related to the way human minds encode negative memories, by creating a positive memory with a negative tag [35]. This means that the statement "Dimosthenes is not a liar", is mentally stored as "Dimosthenes is a liar" and "not", an observation that has the side effect of possible information mislabeling in the case of simple human memory loss of the tag.

The hallmark of fake news effects on public discourse and societies in general, are the 2016 US elections. Without delving into the details of foreign country interference, as it goes beyond the scope of this dissertation, it has been observed that fake news were on purpose created and shared among the US population on social media platforms as the 2016 US presidential elections were approaching. Hot topics that divide the public have been carefully selected and polarizing articles were spread, in order to deepen the divide among the populace. Researchers found out that the top 20 fake news stories that were circulating on social media, have received more engagement than the top 20 factually correct news stories on the same platforms [35]. Based on other published research, such as Allcott and Matthew [36] and Spenkuch and Toniatti [37], it is theorized that advertisements do indeed affect voter behavior with regards to the way they vote and news in general, which includes fake news as a subset, have similar effects as political advertisements. The measure of influence of fake news on the election of President Trump is up for debate according to the researchers, who do not assess neither a favorable nor a detrimental effect.

In Greene, Nash and Murphy [38], it is proven that exposition to fake news stories can result in false memories about the events portrayed. In their results, 44% of participants reported a false memory for at least one fabricated story. The results also suggested that exposure to social identity threats may enhance the polarizing effects of fake news.

The SARS-CoV-2 novel coronavirus pandemic, has also generated a large amount of misleading and false information in its wake. It has been observed that misinformation regarding the pandemic can fundamentally distort the public's risk perception of the virus [39]. This is of major importance, as risk perception has been associated with the likelihood of following preventative health behavior [40]. Belief in conspiracies is also a major deterrent in accepting information from expert authorities as true [41]. In England, Freeman, Waite, Rosebrock and Petit [41] has linked believing in conspiracy theories regarding the coronavirus with hesitance to participate in vaccination. These effects are detrimental to the attempt of taming the pandemic via herd immunity, as only 60% of the public is fully vaccinated in Greece.

2.3.2 The Computer Science Perspective

After establishing the importance of fake news for our society, the next step is to go over the current research trends and results from the Computer Science perspective, especially with regards to automated fake news classification modeled as a natural language processing problem.

Naturally, the most effective oracle of truth is an objective human mind. Unfortunately, fact checking is a very time consuming process, and given the speed of news production and

dissemination, tackling the problem via a horde of human workers is impractical, to say the least. Consequently, the task must somehow be automated, and this is where the field of artificial intelligence and natural language processing can be of tremendous help.

The definition of fake news detection is that, given an uncategorized textual representation of a news article, a model can assign a probability score in the space $[0, 1]$, expressing the model's confidence on whether the input can be categorized as fake. The efforts based on artificial intelligence techniques are commonly based on supervised learning [42, 43, 44, 45], weakly supervised via reinforcement [46], active [47] or deep learning [48, 49, 50, 51]. Another active field of artificial intelligence research is the explainability of machine learning models, meaning why a certain classification was decided for a specific input sequence [52, 53, 54, 55]. A detailed recollection regarding practical applications of fake news detection, can be found at [56].

One of the most important parts of machine learning model training, is the existence of good, labelled datasets. In recent years, multiple datasets have been released with the purpose of fake news detection. This is a review of some of the most well known ones:

1. Yelp [57]

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the most recent dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada.

The paper with the best score in this fake review classification task is [58], achieving a 0.99 f1 score.

2. FacebookHoax [59]

This dataset comprises information related to posts from the facebook pages related to scientific news (non- hoax) and conspiracy pages (hoax) collected using Facebook Graph API. The dataset contains 15,500 posts from 32 pages (14 conspiracy and 18 scientific) with more than 2,300,000 likes.

The paper with the best score in this hoax detection classification task is [60], achieving accuracy exceeding 99%.

3. LIAR [61]

LIAR is a new, publicly available dataset for fake news detection. We collected a decade-long, 12.8K manually labeled short statements in various contexts from POLITIFACT.COM, which provides detailed analysis report and links to source documents for each case. This dataset can be used for fact-checking research as well. Notably, this new dataset is an order of magnitude larger than previously largest public fake news datasets of similar type.

The paper with the best performance in this fake news detection classification task is [62], achieving an accuracy of 0.759.

4. FakeNewsNet [63]

FakeNewsNet is collected from two fact-checking websites: GossipCop and PolitiFact containing news contents with labels annotated by professional journalists and experts, along with social context information. It also contains tweets related to each news samples.

The paper with the best performance in this fake news detection classification task is [64], achieving an F1 score of 93.71%.

5. NELA2017 [65]

This is a large political news data set, containing over 136K news articles, from 92 news sources, collected over 7 months of 2017. These news sources are carefully chosen to include well-established and mainstream sources, maliciously fake sources, satire sources, and hyper-partisan political blogs. In addition to each article we compute 130 content-based and social media engagement features drawn from a wide range of literature on political bias, persuasion, and misinformation.

The paper with the best performance in this fake news detection classification task is [66], achieving 97% accuracy.

6. Burfoot Satire News [67]

This is a novel task of determining whether a newswire article is "true" or satirical. The researchers experiment with SVMs, feature scaling, and a number of lexical and semantic feature types, and achieve promising results over the task.

The paper with the best performance in this satire detection classification task is [68], achieving 85.86% F1 score.

7. BuzzFeed News [69]

The BuzzFeed news dataset comprises a complete sample of news published in Facebook from 9 news agencies over a week close to the 2016 U.S. election from September 19 to 23 and September 26 and 27. Every post and the linked article were fact-checked claim-by-claim by 5 BuzzFeed journalists. There are two datasets of BuzzFeed news one dataset of fake news and another dataset of real news in the form of csv files, each have 91 observations and 12 features/variables.

The paper with the best performance in this fake news classification task is [70], achieving 87% accuracy.

8. Deceptive Opinion Spam Corpus [71, 72]

This corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels. This corpus contains:

- 400 truthful positive reviews from TripAdvisor (described in [71])
- 400 deceptive positive reviews from Mechanical Turk (described in [71])
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp (described in [72])
- 400 deceptive negative reviews from Mechanical Turk (described in [72])

Each of the above datasets consist of 20 reviews for each of the 20 most popular Chicago hotels.

The paper with the best performance in this fake review detection classification task is [71], achieving 90% accuracy.

9. FNC-1 [73]

FNC-1 was designed as a stance detection dataset and it contains 75,385 labeled headline and article pairs. The pairs are labeled as either agree, disagree, discuss, and unrelated. Each headline in the dataset is phrased as a statement.

The paper with the best performance in this stance detection classification task is [74], achieving 82.02% accuracy.

10. BS Detector [75]

This dataset is collected from a browser extension called BS Detector, developed for checking news veracity. It searches all links on a given web page for references to unreliable sources by checking against a manually compiled list of domains. The labels are outputs of BS Detector, rather than human annotators.

11. CREDBANK [76]

The CREDBANK corpus was collected between mid October 2014 and end of February 2015. It is a collection of streaming tweets tracked over this period, topics in this tweet stream, topics classified as events or non events, events annotated with credibility ratings.

The paper with the best performance in this credibility assessment classification task is [77], achieving accuracy 82.86%.

12. BuzzFace [78]

Scripts to create a dataset focused on news stories (which are annotated for veracity) posted to Facebook during September 2016 consisting of:

- Nearly 1.7 million Facebook comments discussing the news content
- Facebook plugin comments
- Disqus plugin comments
- Associated webpage content of the news articles

No similar data set focusing on Facebook existed as of 2016. Potential uses for the data include news veracity assessment using machine learning, social bot detection, and study of propagation of information through several various platforms.

13. COVID-19 Fake News Dataset [79]

In this paper, a dataset of social media posts and articles on COVID-19 with real and fake labels is presented. The targeted media platforms for the data collection are designated to be the ones that are actively used for social networking for peer communication and relaying information, which could be in the form of news, events, social phenomenon, etc. We collect both real news and fake claims that surfaced on social media on COVID-19 topic. Fake claims are collected from various fact-checking websites like Politifact, NewsChecker, Boomlive, etc., and from tools like Google fact-check-explorer and IFCN chatbot. Real news is collected from Twitter

using verified twitter handles. We also perform exploratory data analysis and implement four machine learning baselines.

The paper with the best performance in this fake news detection classification task is [79], achieving 93.32% F1 score.

14. COVID19-FNIR [80]

The CoVID19-FNIR dataset contains news stories related to CoVID-19 pandemic fact-checked by expert fact-checkers. CoVID19-FNIR is a CoVID-19-specific dataset consisting of fact-checked fake news scraped from Poynter and true news from the verified Twitter handles of news publishers. The data samples were collected from India, The United States of America, and European regions and consist of online posts from social media platforms between February 2020 to June 2020. The dataset went through preprocessing steps that include removing special characters and non-vital information.

15. FibVID [81]

This dataet addresses COVID-19 and non-COVID news from three key angles. First, truth and falsehood indicators are provided (T/F) for news items, as labeled and validated by several fact-checking platforms (e.g., Snopes and Politifact). Second, spurious-claim-related tweets and retweets are collected from Twitter, one of the world's largest social networks. Third, basic user information, including the terms and characteristics of "heavy fake news" user to present a better understanding of T/F claims in consideration of COVID-19 is included.

A more in depth analysis of fake news detection datasets is provided by D'Ulzian et al [82].

3. DEVELOPMENT OF GREEK-BERT FAKE NEWS DETECTION MODEL

The purpose of this dissertation is twofold: to create an NLP, GREEK-BERT based, fake news detection model and fake tweets detection model, and a web application that can be used to detect and analyze fake news dissemination and emerging patterns in social networks, with focus on twitter. The first step of the procedure is to gather the data. Since we created two models, we have used two datasets. Firstly, we reused the dataset available from Odysseas Trispiotis' Master in Science Thesis [1]. Secondly, we have generated a semi manually generated dataset of tweets in the Greek language, consisting of 1485 non fake tweets and 716 fake. The data are then carefully post processed and cleaned. Finally, the GREEK-BERT model is fine tuned on the input datasets and the two final models are generated.

The web application frontend, is comprised of a simple python web application, that offers the user an experience similar to that of the official twitter advanced search [83]. It receives input from the user for various search terms, such as tweet text, start and end date for the search window and number of specific user interactions for the searched tweets (namely likes, retweets and replies). Then, it uses twitter's api [84] for fetching the relevant results. Once the search result data are downloaded, the application's backend workers perform a number of steps to infer the status and relationships of the fetched tweets. In particular, it first classifies each tweet as either fake or non fake and then it analyzes the relationships and impact of each tweet with other tweets and users. Finally, the resulting analysis is offered to the user in a multitude of forms:

1. json data, which can be used for further analysis and post processing
2. a dynamic graph which can be interactively edited through the end user's browser
3. a static graph image that can be used as a single point of time status for the given search terms combination

The general architecture can be seen in the following image.

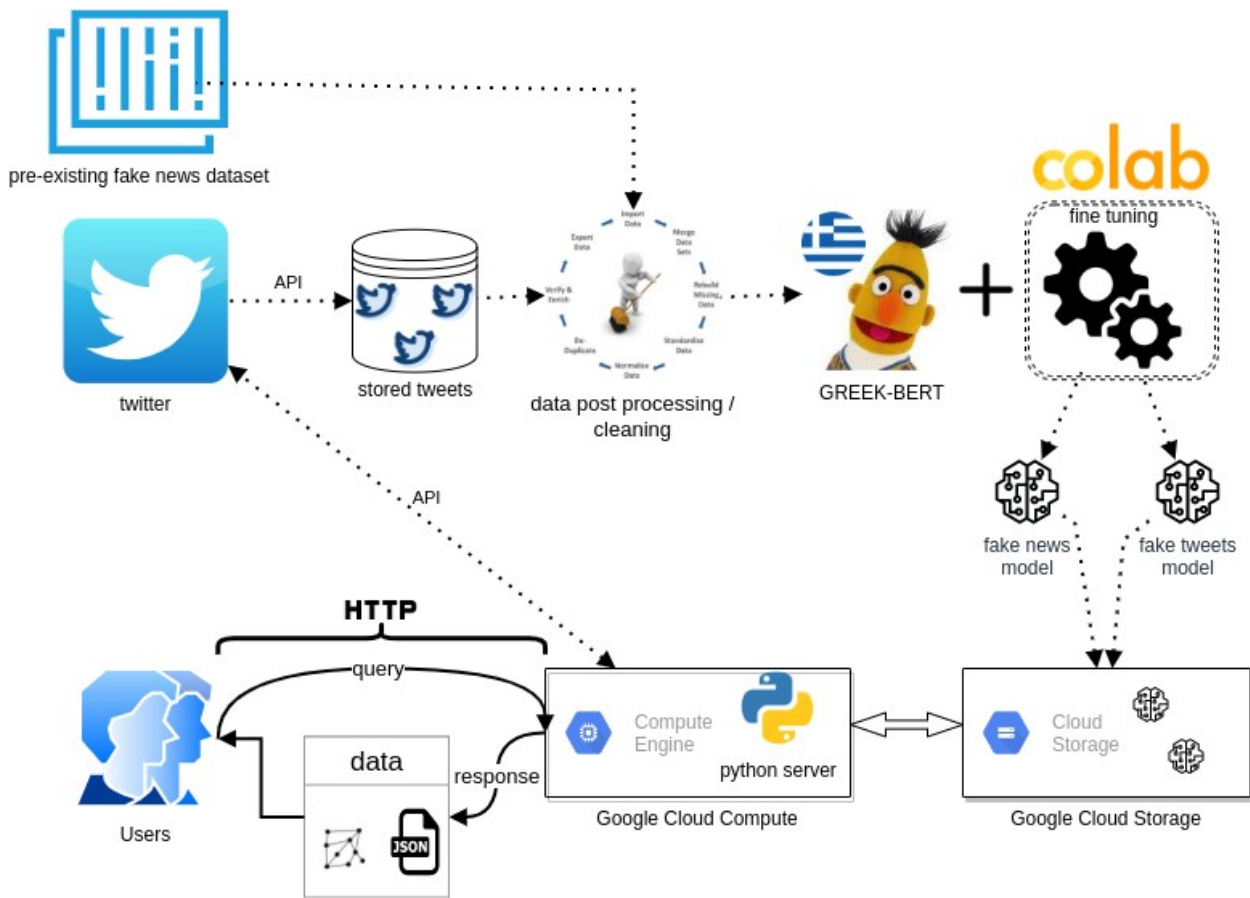


Figure 3: Overall Project Architecture

3.1 Technologies Used

For the development of the classification models and the accompanying web application, a multitude of different technology stacks and software libraries have been used. The software has been developed on the Ubuntu Linux operating system, version 20.04 LTS [85]. The base of the project has been the python programming language [86]. The model development was implemented in the Jupyter Notebook environment [87]. The model training was executed in the google colab environment [88]. For the storage of the created models, google drive [89] and google cloud storage [90] have been used. The web application has been developed with the flask python library and the web server that has been selected for serving the application is gunicorn. For hosting the application, Google Cloud Platform has been used, with the same version of the Ubuntu operating system as the one used for development. In total, this dissertation project required the combination of eighteen core libraries and tools for the python and javascript programming languages. A detailed list of the technologies used follows.

3.1.1 Software

1. Operating System: Ubuntu Linux [91]

Ubuntu is a Linux [92] distribution based on Debian [93] Linux and composed mostly of free and open source software [94]. It is officially released in three editions, Desktop, Server and Core. For the development environment the Desktop version has been used, while for the web application hosting the server edition has been used instead. Ubuntu is released every six months, with long term support releases being released every two years. The latest long term support version is 20.04 and is the one used in this project. Ubuntu is developed by Canonical [95] and a community of other developers, under a meritocratic governance model. Ubuntu is named after the Nguni [96] philosophy of the same name -ubuntu-, which is indicated to mean "show humanity unto others" with the connotation of "I am what I am because of who we all are".

2. Programming Language: Python [86]

Python is an interpreted high level general purpose programming language. Its design philosophy emphasizes code readability. Its language constructs as well as its object oriented approach aim to help programmers write clean, logical code for small and large scale projects. Python is a dynamically typed [97] and garbage collected [98] language. It supports multiple programming paradigms, included procedural, object oriented and functional programming. The creator of the language Guido van Rossum began working on it in the late 80s and it was first released in 1991. The version of the python programming language that has been used is 3.10 for the development and web application environment, while for the execution of the model training, the Google colab environment offers python 3.6. There are no major differences between the two versions and they can cooperate flawlessly.

3. Development Environment: Jupyter Notebook [87]

Project Jupyter is a project and community whose goal is to develop open source software, open standards and services for interactive computing across dozens of programming languages. It was spun off from IPython in 2014. The three core programming language supported by Jupyter are Julia, Python and R. Jupyter Notebook is an open source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization and machine learning, among others. The version of Jupyter Notebook used is 6.2.0.

4. Numerical Analysis / Matrix Operations: NumPy [99]

NumPy is a library for the python programming language, adding support for large, multi dimensional arrays and matrices, along with a large collection of high level mathematical functions to operate on these arrays. NumPy is open source software and has many contributors around the world. The version of numpy utilized by this project is 1.19.5.

5. Data Manipulation: Pandas [100]

Pandas is a software library written for the python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series data. It aims to be the fundamental high level building block for doing practical, real world data analysis in python. Pandas is free software [101]. The version of pandas used in this project is 1.1.5.

6. Machine Learning / Neural Networks: PyTorch [102]

PyTorch is an open source machine learning library based on the Torch library [103]. It is used for application such as computer vision and natural language processing. It is primarily developed by Facebook's AI research lab. It is free and open source software. The version of pytorch used by this project is 1.9.0.

7. Machine Learning / Metrics: Scikit-Learn [104]

Scikit-learn (also known as sklearn) is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms and it is designed to interoperate with python numerical and scientific libraries NumP and SciPy. The version of sklearn used in this project is 0.22.2.

8. Natural Language Processing: HuggingFace Transformers [105]

Transformers is a ste of the art natural language processing for jax, pytorch and tensorflow. It provides general purpose architectures (such as BERT, GPT-2, RoBERTa, and others) for natural language understanding and natural language generation, with over 32 pretrained models in more than 100 languages. The version of transformers used by this project is 4.12.3.

9. Data Visualization: Matplotlib [106]

Matplotlib is a plotting library for the python programming language and its numerical mathematics extension, NumPy. It provides an object oriented API for embedding plots into applications using general purpose GUI toolkits like TkInter, wxPython, Qt or GTK. The version of matplotlib utilized by this project is 3.2.2.

10. Data Visualization: Seaborn [107]

Seaborn is a python data visualization library based on matplotlib. It provides a high level interface for drawing attractive and informative statistical graphics. The version of seaborn used in this project is 0.11.2.

11. Web Application Framework: Flask [108]

Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks. The version of flask used by this project is 2.0.1.

12. HTML Template Rendering Engine: Jinja [109]

Jinja is a full-featured template engine for Python. It has full unicode support, an optional integrated sandboxed execution environment and widely used. It is the built in option for the flask framework, regarding html templating. The version of jinja used in this project is 3.0.1.

13. Static Graph Generation: NetworkX [110]

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex graphs and networks. The version of networkx utilized by this project is 2.6.3.

14. Twitter Dynamic Searching: Twint [111]

Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search

operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags and trends. The version of twint used in this project is 2.1.21.

15. Dynamic Graph Generation: Alchemy [\[112\]](#)

Alchemy.js is a graph drawing application built in d3. Alchemy.js was built so that developers could easily get up and running with Graph visualization applications, and not much over head. The version of alchemy.js used by this project is 0.3.1.

16. HTML Styling: Bootstrap [\[113\]](#)

Bootstrap is a free and open source css framework directed at responsive, mobile-first front-end web development. It contains CSS and Javascript based design templates for typography, forms, buttons, navigation and other interface components. The version of bootstrap used in this project is 4.3.1.

17. Twitter API Client Library: Tweepy [\[114\]](#)

Tweepy is An easy-to-use python library for accessing the Twitter API. The version of tweepy used by this project is 4.2.0.

18. Twitter Web-UI Client: Selenium [\[115\]](#)

Selenium is an open source umbrella project for a wide range of tools and libraries aimed at supporting web browser automation. It is primarily used for automating web application functional testing, but it is not limited to that, as it is the most widely used web automation tool, for operations such as web based administration tasks or web scrapping. The version of python-selenium used in this project is 4.0.0.

19. Selenium Backend: Geckodriver [\[116\]](#)

Geckodriver is a program that provides the HTTP API described by the WebDriver protocol to communicate with Gecko browsers, such as Firefox. It can be used to programmatically operate browsers and is one of selenium's supported backends. The version of geckodriver used by this project is 0.30.0.

3.1.2 Hardware

For the purposes of this project, namely training the GREEK-BERT based models and serving the accompanying web application, two platforms have been used. Google Colab [\[88\]](#) has been used for the training of the models and Google Cloud Platform [\[117\]](#) has been used for deploying and service the web application.

3.1.2.1 Training

Google Colaboratory, or "Colab" for short, allows writing and executing Python in a browser environment. It is backed by Google Cloud Platform and offers a Jupyter Notebook like environment, for usage in data science and AI projects. It offers free, limited time, access to CPU and, especially, GPU resources, that can tremendously help with the training time of neural network models, accelerating the training time by a factor as big as six, when compared with CPU only training. For the purposes of this project, the free tier of Google Colab has been utilized, which offers an environment with 2 Intel Xeon 2.2 GHz CPUs, 13 GB of RAM and an Nvidia Tesla K80 GPU.

3.1.2.2 Web Application

Google Cloud Platform, commonly abbreviated as GCP, is a suite of cloud computing services offered by Google. Alongside a set of management tools, it provides a series of modular cloud services, like computing, data storage and machine learning, among others. For the purposes of this project, GCP's free tier has been used. The web application deployment environment utilizes a 4 Intel Xeon 2.2 GHz CPU, 16 GB of memory and no GPU. While the usage of a dedicated GPU card would tremendously help speeding up the classification work of the backend service, unfortunately GCP's free tier services do not include a GPU offering.

3.2 Dataset Creation

For the purposes of this dissertation, two datasets were used. One, for classifying news articles and one for classifying twitter posts. For the first model, the dataset available from Odysseas Trispiotis' Master in Science Thesis [1] has been used. For the second model, a novel dataset has been created from scratch, with the usage of Twitter's API and a set of post processing, data cleaning steps. The novel dataset is targeted only in COVID-19 related true or fake twitter posts.

3.2.1 Pre Existing Dataset

The dataset used for classifying fake news articles is offered in the csv format and contains the following attributes:

1. id: unique id for a news article
2. title: the title of a news article
3. author: the author of the news article ; could be incomplete
4. text: the text of the article
5. label: a label that marks the article as potentially fake (1 for fake and 0 for non fake)

A total of 34.976 articles exist in the dataset, of which 20.515 are not fake and 14.461 are fake.

3.2.2 Novel Dataset

Automated fake news classification is a challenging problem with immense real world political and social impact. However, practical applications are limited due to the lack of proper datasets. In deep learning, a model can be as good as the data you feed it. A dataset, in order to be efficient for model training, needs to be balanced on the order of classes, and has to encapsulate enough statistical variance per class, in order to avoid the training overfitting problem. As is well known, since 1975 and the relevant paper from Charles Goodhart [118], which gave rise to Goodhart's Law, "When a measure becomes a metric, it ceases to be a good measure". There is a well know story in AI research circles about a neural network model that was trained to classify whether an image contained a wolf or a huskie [119]. Even though the model achieved high accuracy scores for images that weren't part of its training, the researchers noticed misclassification of certain clear images. Upon further investigation, it was determined that the neural network was effectively trained for classifying images based on whether there was snow in it or not, and not for the full amount of information included in the whole picture. Another well known and

frequently faced problem with the requirement of labeled, well balanced, varied datasets, is that most commonly, the English language is used for research and benchmark purposes, leading to a severe lack of proper datasets in other languages, especially for those used by a tiny fracture of the world's population, such as Greek. In particular, Greek is the 74th, out of 91, most natively spoken language in the world, which represents a total of 0.17% of the world's population [120] -important note: only languages with more than ten million native speakers are included in the list-

3.2.2.1 Twitter Background

Before we go into the specifics of how the dataset was captured, we would like to go over the details of what is twitter and how it works. This is by no means an exhaustive analysis of all of twitter's features and intricacies, but it is limited to what is relevant for the dataset creation. Twitter is an American microblogging and social networking platform, on which users post and interact with messages known as tweets.

The first building block of the platform, is the user. Each user can "follow" another user, thus subscribing to the second user's tweet feed. This operation is directed, which means that if user A follows user B, it doesn't mean that user B also follows user A. A user can also block another user, which prevents him / her from viewing the blocking user's tweets, regardless of their follow relationship or lack thereof.

The second building block of the platform, is the tweet. A tweet, is a message that can be comprised of text, urls, images and a set of special text types. The special text types are the "mention", which is the character "@" concatenated with an alphanumeric string representing a twitter user's non unique handle -also known as screen name-, and the "hashtag" which is the character "#" followed by an alphanumeric string which can be used to group posts, by different or the same users, together. The total length of a tweet must not exceed 280 characters. Any urls included in the tweet, are automatically shortened by twitter's own url shortening service.

The third and final building block of the platform, is tweet interactions. There three types of tweet interactions: like, retweet and reply. Like is a single click operation that relates a user who "likes" a tweet, with the tweet itself. Retweet can be a composite operation, as there are two types of retweet operations: you can either simply retweet an existing tweet, by which you just post it as if it was written by you -with a note to the original tweet source-, or you can retweet and include a tweet in the same time, in which case the new composite tweet will contain the new message alongside the message of the original tweet and a link to the original tweet link. Finally, there is the reply operation, in which case you link a new tweet with an existing one, generating a time ordered tree of tweets, a process that emulates a conversation with possibly multiple branching points.

3.2.2.2 Data Gathering

Initially, an attempt was made to gather tweets by manually using the twitter platform's search functionality. This proved very time consuming, as the estimated speed of tweet gathering was less than one per minute. Furthermore, contrary to what was expected by the author, it was quite difficult to find reliable sources of fake tweets, as most of the relevant search terms would result in tweets debunking, mocking or directly attacking fake news or users spreading fake news. On the other hand, reliably tracking true information, was easier than anticipated, as certain official sources or medical professionals are dedicated to providing honest scientific information on twitter.

A more robust approach was deemed necessary, thus the second attempt for gathering the data was based on the official twitter API (Application Programming Interface). Twitter offers its development platform [84], that enables real time and historical access to tweets that can be accessed by a user. This means that private tweets or blocked user's tweets cannot be viewed even via the API, as the API access has to come through a valid twitter user. Twitter's API is offered in three separate product tracks: standard, academic and business. As the business use case was out of the question with regards to this project, a request for an academic research license was placed. Unfortunately, after a long period of time, the request was denied, with twitter's representative offering the use of the standard product track instead. The differences between the standard and the academic product tracks are quite important, as the academic track most important differences are that it allows for historic searches without a time limit and with a wider range of search operators, whereas the standard level of API access only allows for the last week's worth of historic data and only a basic set of search operators. The standard API access resulted in the generation of 1403 non fake tweets and 430 fake ones.

As the amount of fake tweets was deemed to be too low and the standard API access limits enforces a monthly quota, in contrast to the academic level of access which doesn't have one, a new way of obtaining tweets, especially fake ones, was necessary. Since the official API was out of quota and the manual search was too slow to be effective, a new approach was invented, based on the two former ones. The twitter platform itself, has two separate APIs running at the same time, the official one, that drives its product tracks and the unofficial one, that drives the web and mobile application user interfaces. While mapping the required unofficial API would be a certainly doable undertaking, it would require a significant amount of research, that falls outside the main objective of the thesis. As an alternative, the browser automation tool Selenium was used. Selenium is a browser automation API, with bindings for most widely used programming languages, that can emulate human interaction with a web browser, in an automated way. By following this strategy, we could directly use twitter's web ui to automate the procedure of searching for tweets and extracting the required text from the search results by parsing the rendered html document. This approach generated another 561 fake tweets and 837 non fake. The limited number of results, especially regarding fake tweets, are only due to the searching strategy.

An important observation that can be made, is that spotting fake information dissemination in social networks user posts is a rather intractable problem. The public APIs, either free or via paid subscriptions, offered by social media companies are blatantly obvious to be intentionally limited, which is a fact that in hindsight should be expected, as the main driver of social media and internet service companies' revenue, in the new model of surveillance capitalism [122, 123], is the online advertisements market, which has an estimated market cap of 522 billion dollars worldwide, by 2025 [124].

3.2.2.3 Searching Strategy

The searching strategy for fake and non fake tweets, was the same. The classification of news articles as fake or non fake, is a well understood problem. As a design choice for our dataset, we selected the following base strategy: fake tweets discuss about fake news, whereas non fake tweets discuss about non fake news. Of course this is a fairly limited subset of all the available fake and non fake information posted on twitter, but it is the only one that can be automated in an unsupervised way, in order to generate a labeled dataset.

As already stated in the previous section, finding reliable news platforms is easy. We have selected 20 of the most known Greek online news outlets, which purposefully represent a wide range of social and political views, in order to maximize the statistical variance of our dataset. The non fake news outlets selected, in alphabetical order, are:

1. alfavita.gr
2. amna.gr
3. capital.gr
4. cnn.gr
5. dikaiologitika.gr
6. documentonews.gr
7. enikos.gr
8. ert.gr
9. huffingtonpost.gr
10. in.gr
11. koutipandoras.gr
12. lifo.gr
13. naftemporiki.gr
14. news247.gr
15. protothema.gr
16. skai.gr
17. tanea.gr
18. thepressproject.gr
19. thetoc.gr
20. tovima.gr

For fake news sources, we have used the Greek fact checking website elinikahoaxes.gr. We have gathered a set of twenty Greek news outlets that have published more than one COVID-19 related fake news story during the pandemic. The fake news outlets selected, in alphabetical order, are:

1. aetos-apokalypsis.com
2. amazonios.net
3. anastoxasmoi.gr
4. anixneuseis.gr
5. athensmagazine.gr
6. attikanea.info
7. defencenet.gr

8. dimpenews.com
9. emperorsclothes.gr
10. greeknewsondemand.com
11. hellas-now.com
12. katohika.gr
13. mainlynews.gr
14. makeleio.gr
15. oparlapipas.gr
16. press-gr.com
17. pronews.gr
18. taxidromos.gr
19. triklopodia.gr
20. voicenews.gr

In order to target COVID-19 specific posts and discussions, we have selected a set of nine COVID-19 related search terms, some of which tend to imply fake or non fake related content. The selected search terms are:

1. covid
2. κορονοϊός
3. κοροδοϊός
4. εμβόλιο
5. εμβολιασμός
6. αντιεμβολιαστές
7. κρούσματα
8. μπόλι
9. lockdown

A search has been done for a combination of each one the news outlets base url and each search term. For example, one search might be "tovima.gr covid" or "makeleio.gr lockdown". For each search up to 100 results were fetched, in order to not exhaust the API limits during development.

In the resulting dataset, which is provided in two text files, one for fake and one for non fake news, one tweet is included in each line.

3.2.2.4 Data Cleaning

As described in earlier sections, tweets include more than simple textual representations of information: it can include urls, images, user mentions, hashtags and extended character such as emoticons or special glyphs. Moreover, the tweets fetched via official

twitter API were fetched in an ordered json structure, where data was more easily extractable, while those extracted via the selenium scrapping route, were not structured at all and the html page needed to be more heavily post processed in order to extract the relevant tweet text.

In general, in both cases, the following text cleaning methods were applied:

1. remove URLs (Uniform Resource Locator)
2. remove twitter mentions
3. remove special HTML (HyperText Markup Language) escape sequences
4. convert newlines to spaces
5. remove any non alphanumeric character (except the hashtag symbol)
6. remove emoticons
7. remove duplicate or overly similar entries
8. remove tweets shorter than 50 characters

Please note that while it is common in twitter dataset post-processing to also remove hashtags from the tweets text, in this case all hashtags have been preserved, as it was assumed that they carry information relevant to the classification process.

3.3 Model Development

For the model development, initially all the datasets were uploaded to Google Drive, Google's public storage platform. Google drive can be automatically used in the Google Colab environment by using minimal code, as only two lines of code are required.

Then, a set of non pre-installed python packages are installed to the colab environment, namely transformers, pytorch_wrapper and seaborn. Also, the GREEK-BERT source code is cloned in the current working directory, as it is going to be used in later steps.

Afterwards, we load the dataset into memory, via the pandas toolchain, and a set of descriptive information is printed and plotted for the dataset. Especially regarding the fake news dataset, as the maximum sequence length supported by the BERT architecture is 512, we also calculate the amount of data lost in the process, as most articles are longer than 512 words. In particular, 83.64% of articles (28.942 out of 34.605) exceed this margin, which causes a total loss of 25.88% of our input data (3,109.972 out of 12.017.277 words). The strategy of text selection, was to use the first 512 tokens. Alternative strategies include using the last 512 token or the first and last 256. These strategies have not been tested.

Subsequently, we lowercase and remove any accents from the Greek language input and tokenize the input. Due to limitations of the Google Colab environment, with regards to maximum time of execution, only a subset of two thousand articles are used from the fake news dataset, in order to keep the training time within the limitations. The data are split into stratified training, validation and test subsets, with size 80%, 10% and 10% of the original dataset respectively.

Then, the model is fine tuned, with a learning rate of $2e-5$ for the Adam optimizer, for batch size of 8 (in order to avoid GPU memory overflowing). The training is executed for four

epochs. Then, the training summary is printed and the final model is saved to disk for later re-use.

Finally, the model is tested against the test dataset and the required performance metrics are computed and printed.

4. EXPERIMENTAL RESULTS

To evaluate the performance of our classification models, we used metrics commonly used in the machine learning research space, namely precision, recall and f1 score. These metrics were measured using the scikit-learn python library. We will first define the metrics by which we will measure our models performance and then go over the experimental results.

4.1 Metrics Definitions

In our classification problem, we have two different classes, namely fake and non fake. Fake has been numerically defined as "1" and non fake as "0", in both datasets. When a numerical representation of our textual data is fed into the model, the model outputs in which class the data belongs at. The model's decision can either be correct or incorrect and the data itself belongs to one of the two classes regardless of the model's decision. A result that is positive and correct is called True Positive or TP, whereas a positive result that is incorrect is called False Positive or FP. Correspondingly, a negative result that is correct is called True Negative or TN, whereas a negative result that is incorrect is called False Negative or FN. These four classification outcomes are the base of a model's performance and can be summarized in, what is called a confusion matrix.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 4: Confusion Matrix Definition and Related Performance Metrics

The above definitions are the base for the rest of a model's performance indicator metrics:

1. Recall: the rate of TP classifications over the total number of positive cases in our dataset, which is represented by the sum of TP and FN. Alternatively, it can be defined as the ratio of correctly predicted positive classes out of all the positive data

points in our dataset. Recall is also known as sensitivity, most commonly for binary classification tasks.

2. Precision: the rate of TP classifications over the total number of positive classifications, which is represented by the sum of TP and FP. Alternatively, it can be defined as the ratio of correctly predicted positive classes out of all the positive predictions.
3. Accuracy: the rate of our correct predictions over the total predictions made. Alternatively, it can be defines as the ratio of correct predictions.
4. F-measure: it is the harmonic mean of precision and recall and it is commonly used for comparing models. F1 score is not as intuitive to understand compared to accuracy. In essence, accuracy is a better metric if both false positives and false negatives have similar cost for our classification task, but if the costs are very different, the F1 score is a better evaluator for a model.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{tp}}{\text{tp} + \text{fp} + \text{fn}}$$

Figure 5: F1 Measure Definition

5. AUROC curve: the Area Under the Receiver Operating Characteristics curve is one of the most important evaluation metrics for checking any classification model's performance. It is a performance measurement for classification problems at various threshold settings. The ROC curve is a probability curve, with the True Positive Rate (the number of correct positive predictions over all positive datapoints) on the y-axis and the False Positive Rate (the number of incorrect false predictions over all positive datapoints) on the x-axis. AUC represents the probability that a classifier is more confident than a uniformly random classifier. Consistently bad classifiers can be inverted in order to generate proportionally good classifiers by inverting their classifications.

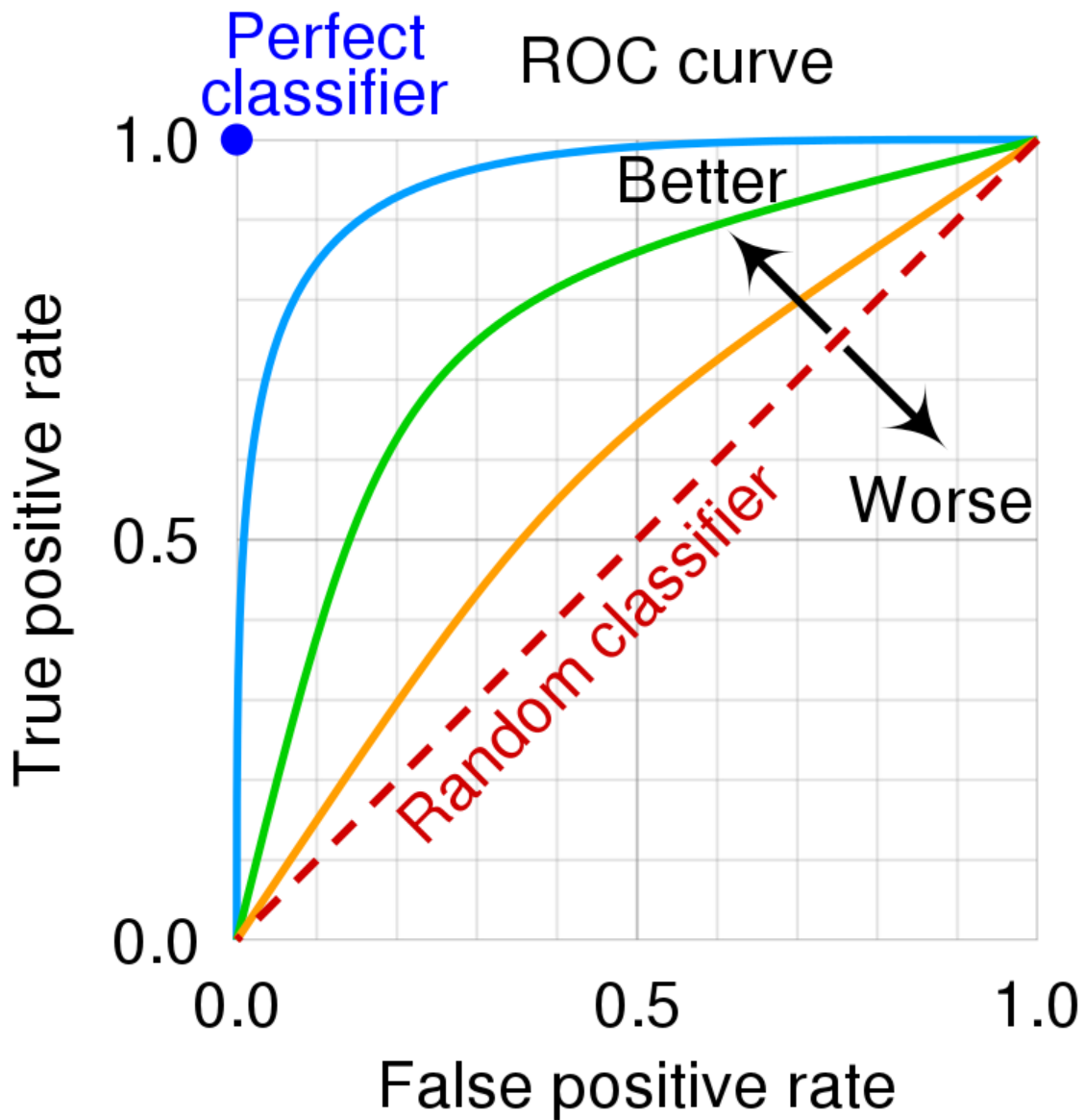


Figure 6: AUROC Curve Sample

4.2 Models Performance

4.2.1 Fake News Model

Overall, the fake news detection model performs very well on the classification task. It has a precision of 0.91 for non fake news and 0.95 for fake, which translates to 0.96 recall for non fake news and 0.89 for fake. The overall accuracy is 0.93 and the f1 score is 0.93, with almost the same performance for both fake and non fake news detection. The detailed performance can be viewed in the following table and images.

Table 1: Performance metrics for the fake news model

	Precision	Recall	F1-Score	Support
Not-fake	0.91	0.96	0.94	107
Fake	0.95	0.89	0.92	93
Accuracy			0.93	200
Macro avg	0.93	0.93	0.93	200
Weighted avg	0.93	0.93	0.93	200



Figure 7: First Model Confusion Matrix

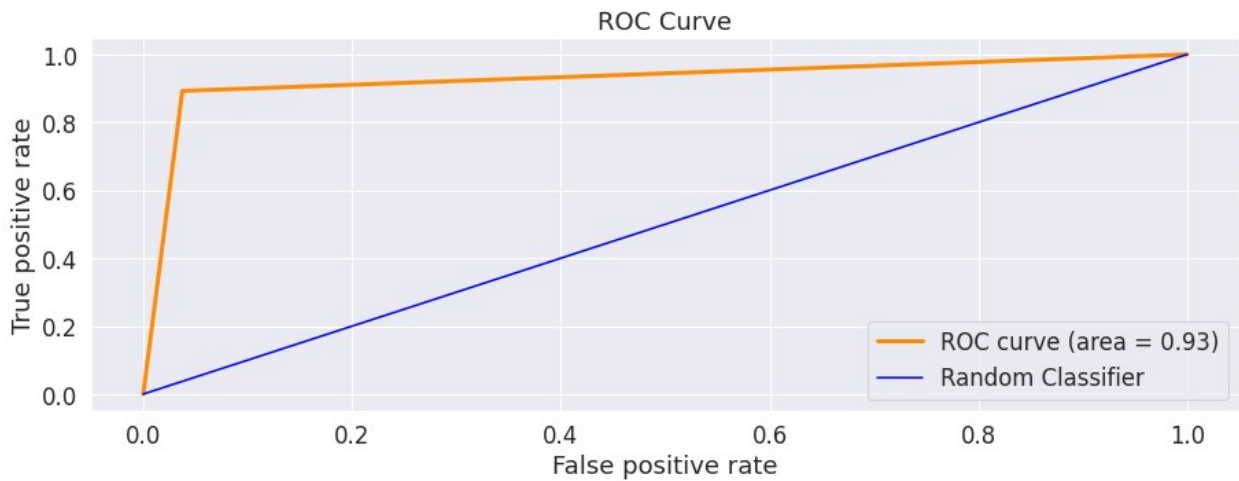


Figure 8: First Model AUROC Curve

4.2.2 Fake Tweets Model

Overall, the fake tweets detection model performs very well on its classification task as well. It has a precision of 0.93 for non fake tweets and 0.87 for fake ones, which translates to 0.94 recall for non fake tweets and 0.86 for fake. The overall accuracy is 0.91 and the f1 score is 0.90, with slightly better performance for non fake tweets detection. This observation follows the earlier remark, regarding the difficulty of obtaining fake tweets with respect to quality and quantity. The detailed performance can be viewed in the following table and images.

Table 2: Performance metrics for the fake tweets model

	Precision	Recall	F1-Score	Support
Not-fake	0.93	0.94	0.94	148
Fake	0.87	0.86	0.87	72
Accuracy			0.91	220
Macro avg	0.9	0.9	0.9	220
Weighted avg	0.91	0.91	0.91	220

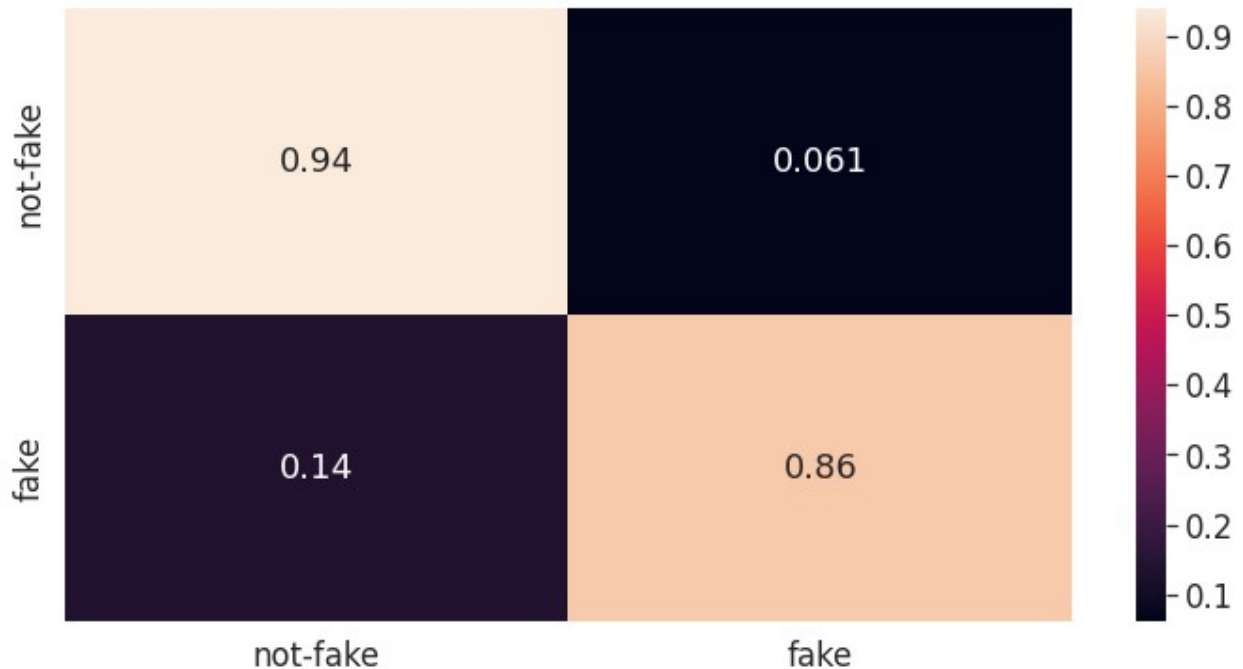


Figure 9: Second Model Confusion Matrix

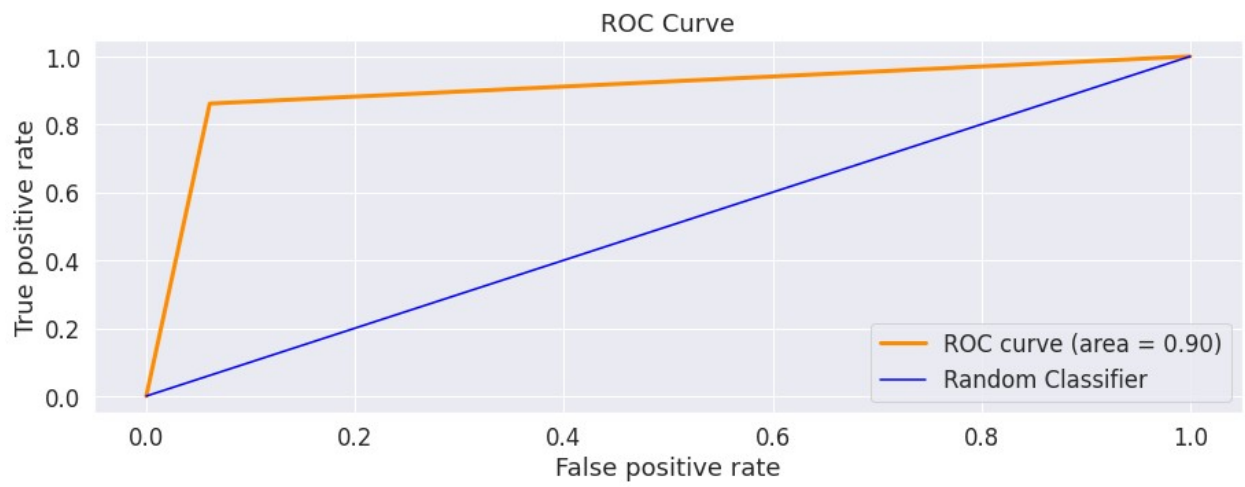


Figure 10: Second Model AUROC Curve

5. WEB APPLICATION

With the help of the NLP classification models, a web application for studying and modeling the dissemination of fake information in twitter has been developed. The application searches for tweets based on search criteria, which are described bellow. Afterwards, for each tweet matching the search criteria, it fetches the list of replies and interactions with other twitter users. Finally, it classifies each tweet as fake or non-fake based on the relevant GREEK-BERT based NLP classification model.

It needs to be pointed out the application is painfully slow. For the twitter search, the unofficial API has been used, via the twint library, as the search limitations of the official API for non research accounts produce very few results. Due to this specific design choice, the search results are unstable and very time consuming to fetch. The same search terms, may return different results, or even no results at all, on consecutive search executions. Moreover, the execution of the NLP model in a non GPU enabled platform, makes the tens or hundreds of classifications that are required for every base search, very time consuming as well. For example, the search for the word "kopovoioς" with a search window time limit of the last two days and a search limit for twenty original tweets, results on an execution time of 45 minutes.

5.1 Search Options

- Search String: the text that should be included in the tweets we're searching for. It is the only mandatory field.
- User: restrict results for a specific twitter user
- Retweets: limit results to tweets that have at least that have been retweeted at least this many times
- Replies: limit results to tweets that have at least that many replies
- Likes: limit the results to tweets that have at least that many likes
- Date (from): limit the search to tweets after this date (inclusive)
- Date (until): limit the search to tweets until this date (inclusive)

Συμβολοσειρά Αναζήτησης*

Είναι πιθανό έγκυρες αναζητήσεις να μην επιστρέψουν αποτελέσματα με την πρώτη προσπάθεια. Σε αυτήν την περίπτωση, παρακαλώ επαναλάβετε την αναζήτησή σας.

Χρήστης

Περιορισμός αναζήτησης σε συγκεκριμένο λογαριασμό/χρήστη του twitter.

Retweets

Ελάχιστος αριθμός retweet ανά tweet.

Απαντήσεις

Ελάχιστος αριθμός απαντήσεων ανά tweet.


Likes

Ελάχιστος αριθμός likes ανά tweet.

Μέγιστο Πλήθος Αποτελεσμάτων


Μέγιστος αριθμός αποτελεσμάτων.

Ημερομηνία (από)

Έναρξη παραθύρου χρονικής αναζήτησης.

Ημερομηνία (εως)

Λήξη παραθύρου χρονικής αναζήτησης.

[Αναζήτηση Διαδραστικού Γράφου](#) [Αναζήτηση Δεδομένων JSON](#) [Αναζήτηση Εικόνας](#)

Figure 11: Web Application User Interface

5.2 Data Retrieval Options

Three different types of data retrieval options are supported:

1. Interactive Graph: each tweet is represented by a node in the graph and the lines are "reply" relationships. Each node's color is either red for a "fake" classification or green for "non-fake". Node size depends on the total number of interactions it had with other tweeter users. The total number of replies, likes and retweets is counted. The graph can be traversed and modified with the mouse, while left clicking on a node shows the original tweet's text and source url.
2. PNG Image: a static image is returned, with exactly the same semantics as the interactive graph.
3. JSON Data: return the data that has been used for the graph generation. A sample is provided bellow.

Sample JSON data response:

```
{
  "edges": [
    {
      "source": "1234567890123456789",
      "target": "9876543210987654321"
    },
    ...
  ],
  "nodes": [
    {
      "caption": "username",
      "id": "1234567890123456789",
      "interactions": 80,
      "link": "https://twitter.com/username/status/1234567890123456789",
      "role": "not_fake",
      "text": "tweet text"
    },
    ...
  ]
}
```

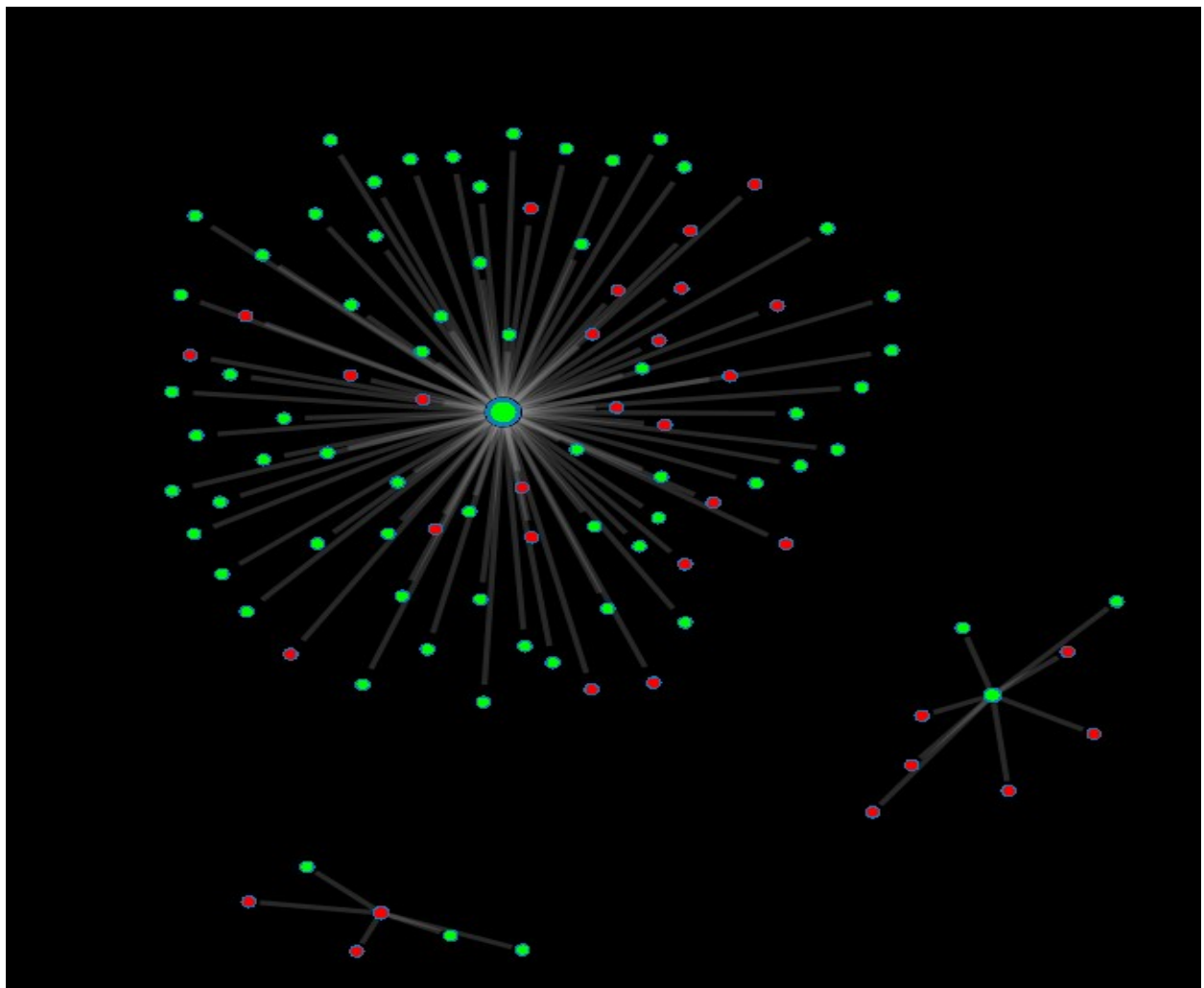


Figure 12: Dynamic Graph Results Example

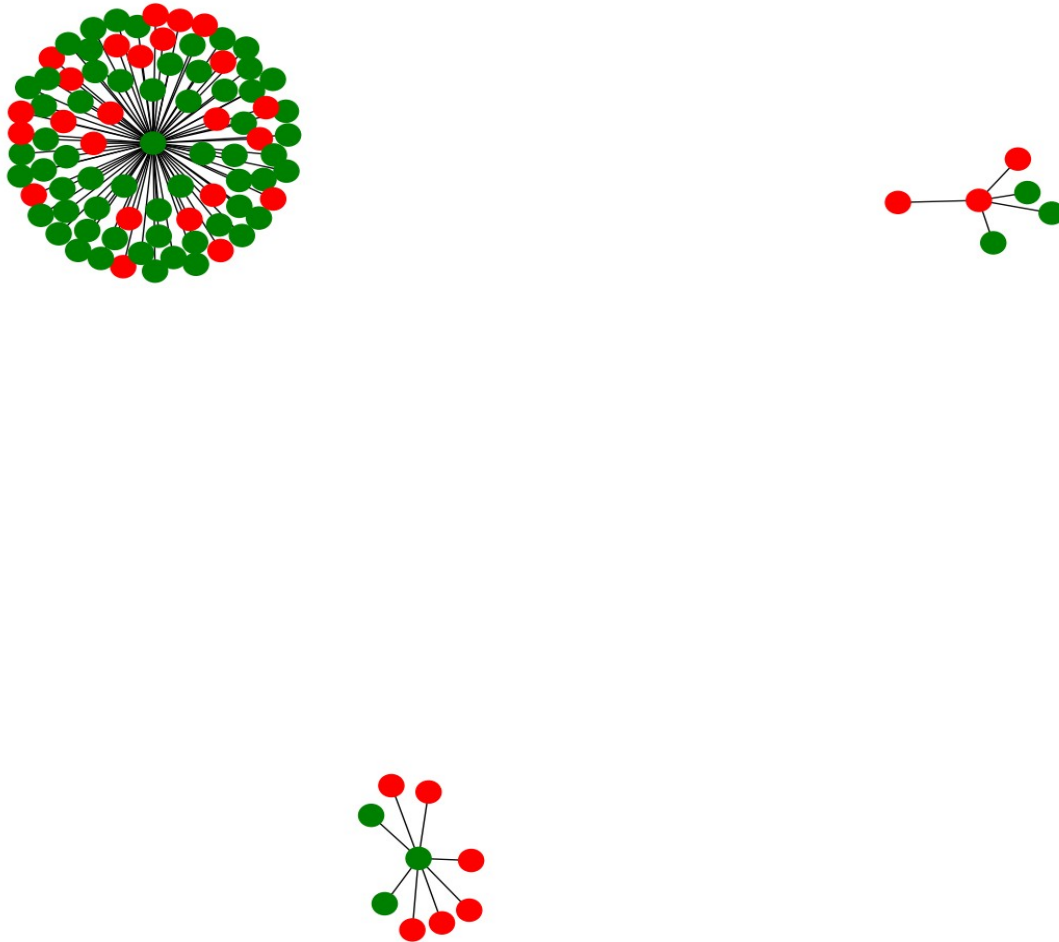


Figure 13: Static Image Results Example

6. CONCLUSION

As part of this thesis, we have introduced all the relevant technology breakthroughs that made this work possible, namely BERT and GREEK-BERT. We have also covered relevant work from other researchers, regarding fake news detection on the modern web, from news outlets and social media platforms.

As part of this thesis, a novel dataset has been created for fake information on social media with focus on the twitter platform and the COVID19 pandemic. During the new dataset creation, we found out that coming up with reliable fake information on social media is a hard problem and one that is, most probably purposefully, made harder by the social media platforms themselves.

Then, we have created two NLP classification models based on the GREEK-BERT pretrained model, for classifying fake news and fake tweets. Our models have achieved excellent performance, with 93% accuracy for the fake news model and 91% for the fake tweets one and similarly high scores for the rest of the common performance metrics, such as precision, recall, f1 score and AUROC.

Finally, we have created a web application that utilizes the fake tweets classification model for analyzing patterns of information dissemination on social media platforms, with a focus on the COVID19 pandemic.

The source code for this project is available on the GitHub platform, in the following web page: <https://github.com/dimfioretos>

REFERENCES

- [1] Odysseas Trispiotis, Real-time Fake-news Detection in Greek using a Browser Extension, master's thesis, Dept. Of Informatics and Telecommunications, NKUA, 2021
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Google AI Language, 2019
- [3] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, Ion Androutsopoulos, "GREEK-BERT: The Greeks visiting Sesame Street", Department of Informatics Athens University of Economics and Business, 2020
- [4] J. McCarthy, whatisai, 12 November 2007 ; <http://www-formal.stanford.edu/jmc/whatisai/node1.html> [Προσπελάστηκε 31/10/2021]
- [5] Stuart J. Russel, Peter Norvig, "Artificial Intelligence: A Moder Approach (2nd ed.)", 2003, p.55
- [6] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers", IBM Journal of Research and Development, 1959, pp. 210-229
- [7] Oxford English Dictionary, Oxford University Press, 2019
- [8] V. L. Rubin, Y. Chen, N. K. Conroy, Deception detection for news: three types of fakes, Proceedings of the Association for Information Science and Technology 52 (2015) 1–4.
- [9] Amjad, Maaz et al. "'Bend the Truth': Benchmark Dataset for Fake News Detection in Urdu Language and Its Evaluation". 1 Jan. 2020 : 2457 – 2469.
- [10] Edson C. Tandoc Jr., Zheng Wei Lim & Richard Ling (2018) Defining "Fake News", Digital Journalism, 6:2, 137-153, DOI: 10.1080/21670811.2017.1360143
- [11] Antonios Pepoudis, "Πλούάρχουν Περί της Ηροδότου Κακοήθειας", master's thesis, Department of History and Archeology, Aristotle University of Thessaloniki, 2013
- [12] Tasoyla Kariskaki, "Οι Ψευδολόγοι που μας Κυβερνοούν", Kathimerini Online Edition, 3 March 2019 ; <https://www.kathimerini.gr/opinion/1012832/oi-pseydologoi-poy-mas-kyvernoyn/> [Accessed on 31/10/2021]
- [13] Levy Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.", American Economic Review, 2021, 111 (3): 831-870.
- [14] European Commision, Brussels, Eurobarometer Spring 2021 Edition, 2021 ; <https://www.europarl.europa.eu/at-your-service/files/be-heard/eurobarometer/2021/spring-2021-survey/key-findings.pdf> [Προσπελάστηκε 31/10/2021]
- [15] B. Doerr, M. Fouz, T. Friedrich, Why rumors spread so quickly in social networks. Commun. ACM 55, 2012, pp. 70–75
- [16] Vosoughi S, Roy D, Aral S. , "The spread of true and false news online", Science, 9 March 2018, 9;359(6380):1146-51.
- [17] Jay Alammar, "The Illustrated Transformer", <http://jalammar.github.io/illustrated-transformer/> [Accessed on 31/10/2021]
- [18] Samuel Dupond, "A thorough review on the current advance of neural network structures". Annual Reviews in Control. 14, 2019, pp. 200–230.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need", arXiv, December 2017
- [20] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, Sanja Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books", arXiv:1506.06724, 2015, pp. 19-27
- [21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee , Luke Zettlemoyer, "Deep contextualized word representations", arXiv, March 2018.
- [22] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", OpenAI, June 2018.
- [23] Philip Gage, "A New Algorithm for Data Compression", C Users Journal, 1994.
- [24] Christian Fuchs, "Media, Communication and Society, Volume Three", Routledge, January 2022 (preprint).
- [25] Elisa Shearer, "Social media outpaces print newspapers in the U.S. as a news source", Pew Research Center, December 10, 2018 ; <https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/> [Accessed on 31/10/2021]
- [26] World Bank, "World Development Indicators", 2020 ; <http://datatopics.worldbank.org/world-development-indicators> [Accessed on 31/10/2021]

- [27] National Centre for Social Research, "The Internet in Greece", 2017 ; <https://www.ekke.gr/siemens/WIP%202017%20english-Version-EKKE.pdf> [Accessed on 31/10/2021]
- [28] Working Party on the Information Economy, "The evolution of News and the Internet", Organisation for Economic Co-operation and Development; 11 June 2010
- [29] Clair Wardle, Hossein Derakhshan, "Information Disorder: Toward an interdisciplinary framework for research and policy making", Council of Europe Report, 2017.
- [30] Clair Wardle, "Fake News: Its Complicated", <https://firstdraftnews.org/articles/fake-news-complicated/> [Accessed on 31/10/2021]
- [31] Zafarani, Reza & Zhou, Xinyi & Shu, Kai & Liu, Huan, "Fake News Research: Theories, Detection Strategies, and Open Problems", 25th ACM SIGKDD International Conference, 2019, 3207-3208. 10.1145/3292500.3332287.
- [32] A.C. Madrigal, "What Facebook did to American democracy", The Atlantic, 2017 ; <https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/> [Accessed on 31/10/2021]
- [33] Kathleen Higgins, "Post-truth: A Guide For The Perplexed", Nature, 2016
- [34] Yariv Tsfati, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, E. Lindgren, "Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis", 2020, Annals of the International Communication Association, 44:2, 157-173, DOI: 10.1080/23808985.2020.1759443
- [35] Craig Silverman, "How Viral Fake Election News Stories Outperformed Real News On Facebook", BuzzFeed, 2016 ; <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.qqE7PoA2QI> [Accessed on 31/10/2021]
- [36] Hunt Allcott, Matthew Gentzkow, "Social Media and Fake News in the 2016 Election", The Journal of Economic Perspectives Vol. 31, No. 2 (Spring 2017), pp 211-235, 2016
- [37] Jorg Spenkuch, David Toniatti, "Political Advertising and Election Outcomes", CESifo Working Paper Series 5780, 2016
- [38] Ciara Greene, Robert Nash, Gillian Murphy, "Misremembering Brexit: Partisan Bias and Individual Predictors of False Memories for Fake News Stories Among Brexit Voters", Memory Volume 29, pp. 587-604, 2021.
- [39] Nicole Krause, Isabelle Frelling, Becca Beets, Dominique Brossard, "Fact-checking as Risk Communication: the Multi-layered Risk of Misinformation in Times of COVID-19", Risk Research, pp. 1052-1059, 22 April 2020.
- [40] Dryhurst, S., Schneider, C. R., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., et al, "Risk perceptions of COVID-19 around the world", J. Risk Res. 1–13. doi: 10.1080/13669877.2020.1758193, 2020.
- [41] Uscinski, J. E., Enders, A. M., KLoftstad, C., Seelig, M., Funchion, J., Everett, C., et al, "Why do people believe COVID-19 conspiracy theories?", Harvard Kennedy School (HKS) Misinform. Rev. 1, 1–12. doi: 10.37016/mr-2020-015, 2020
- [42] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T), pages 1--4.
- [43] Pratiwi, I. Y. R., Asmara, R. A., and Rahutomo, F. (2017). Study of hoax news detection using naïve bayes classifier in indonesian language. In Proc. of the IEEE Int'l Conference on Information & Communication Technology and System (ICTS), pages 73--78.
- [44] Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 647--653.
- [45] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 422--426.
- [46] Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., and Gao, J. (2020). Weak supervision for fake news detection via reinforcement learning. In Proc. of the AAAI Conference on Artificial Intelligence (AAAI), pages 516--523.
- [47] Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. In Proc. of the IEEE Int'l Conference on Big Data (Big Data), pages 556--565.

- [48] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767.
- [49] Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 797--806.
- [50] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018b). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 849--857.
- [51] Zhang, J., Cui, L., Fu, Y., and Gouza, F. B. (2018). Fake news detection with deep diffusive network model. In *arXiv preprint arXiv:1805.08751*.
- [52] Cui, L., Shu, K., Wang, S., Lee, D., and Liu, H. (2019). defend: A system for explainable fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 2961--2964.
- [53] Lu, Y.-J. and Li, C.-T. (2020). Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. pages 505--514.
- [54] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019a). defend: Explainable fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 395--405.
- [55] Yang, F., Pentylala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. (2019). Xfake: explainable fake news detector with visualizations. In *Proc. of the ACM Web Conference (WWW)*, pages 3600--3604.
- [56] Julio Cesar Soares Dos Reis, "Towards Automatic Fake News Detection in Digital Platforms: Properties, Limitations and Applications", Universidade Federale de Minas Gerais, 2020.
- [57] <https://www.yelp.com/dataset>
- [58] Patil Gouri, Raje Swathi, "A Machine Learning Model to Predict Fake Review Using Classifier on Yelp Dataset", IJERT, September 2021.
- [59] <https://github.com/gabll/some-like-it-hoax>
- [60] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, Luca de Alfaro, "Some Like it Hoax: Automated Fake News Detection in Social Networks", School of Engineering, University of California, Santa Cruz, 25 April 2017.
- [61] William Yang Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection", *Association for Computational Linguistics*, pp. 422-426 2017
- [62] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, Huan Liu, "Unsupervised Fake News Detection on Social Media: A Generative Approach", 33rd AAAI Conference on Artificial Intelligence, February 2019.
- [63] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media", *arXiv:1809.01286*, 5 September 2018.
- [64] Shaily Bhatt, Sakshi Kalra, Naman Goenka, Yashvardhan Sharma, "Fake News Detection: Experiments and Approaches beyond Linguistic Features", *Web Intelligence and Social Computing Lab, Department of Computer Science and Information Systems, Birla Institute of Technology and Science*, 2021.
- [65] Benjamin D. Horne, William Dron, Sara Khedr, Sibel Adali, "Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape", *arXiv:1803.10124*, 2018.
- [66] Caireann Kennedy, Josephine Griffith, "Using Markup Language to Differentiate Between Reliable and Unreliable News", *IEEE*, 10.1109/DSAA49011.2020.00087, 9 October 2020
- [67] Clinton Burfoot, Timothy Baldwin, "Automatic Satire Detection: Are You Having a Laugh?", *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, August 2009.
- [68] Kumar Ravi, Ravi Vadlamani, "A Novel Automatic Satire and Irony Detection Using Ensembled Feature Selection and Data Mining", *Knowledge-Based Systems*, 10.1016/j.knosys.2016.12.018, December 2016.
- [69] <https://github.com/BuzzFeedNews/everything> [Accessed on 31/10/2021]
- [70] <https://www.kaggle.com/sohamohajeri/buzzfeed-news-analysis-and-classification> [Accessed on 31/10/2021]

- [71] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.
- [72] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- [73] <https://github.com/FakeNewsChallenge/fnc-1> [Accessed on 31/10/2021]
- [74] <https://github.com/Cisco-Talos/fnc-1> [Accessed on 31/10/2021]
- [75] <https://github.com/thiagovas/bs-detector-dataset> [Accessed on 31/10/2021]
- [76] Tanushree Mitra, Eric Gilbert, "CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations", ICWM, 2015.
- [77] Doaa Hassan Salem, "A Text Mining Approach for Evaluating Event Credibility on Twitter", 10.13140/RG.2.2.34962.71365, 2018.
- [78] Giovanni Santia, Jake Ryland Williams, "BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos", International AAAI Conference on Web and Social Media Twelfth International AAAI Conference on Web and Social Media, 2017.
- [79] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, Tanmoy Chakraborty, "Fighting an Infodemic: COVID-19 Fake News Dataset", AAAI, 2021.
- [80] Julio A.Saenz, Sindhu Reddy Kalathur Gopal, Diksha Shukla, 10.21227/b5bt-5244, July 2021.
- [81] Jisu Kim, Jihwan Aum, Sang Eun Lee, Yeonju Jang, Eunil Park, Daejin Choi, "FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period", Telematics and Informatics, Volume 64, November 2021.
- [82] Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, Patrizia Grifoni, "Fake News Detection: a Survey of Evaluation Datasets", Institute of Research on Population and Social Policies, National Research Council, Rome, Italy, June 2021.
- [83] <https://twitter.com/search-advanced> [Accessed on 31/10/2021]
- [84] <https://developer.twitter.com/> [Accessed on 31/10/2021]
- [85] <https://releases.ubuntu.com/20.04/> [Accessed on 31/10/2021]
- [86] <https://www.python.org/> [Accessed on 31/10/2021]
- [87] <https://jupyter.org/> [Accessed on 31/10/2021]
- [88] <https://colab.research.google.com/> [Accessed on 31/10/2021]
- [89] <https://www.google.com/drive/> [Accessed on 31/10/2021]
- [90] <https://cloud.google.com/storage> [Accessed on 31/10/2021]
- [91] <https://ubuntu.com/> [Accessed on 31/10/2021]
- [92] <https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git> [Accessed on 31/10/2021]
- [93] <https://www.debian.org/> [Accessed on 31/10/2021]
- [94] <https://opensource.org/licenses> [Accessed on 31/10/2021]
- [95] <https://canonical.com/> [Accessed on 31/10/2021]
- [96] https://en.wikipedia.org/wiki/Ubuntu_philosophy [Accessed on 31/10/2021]
- [97] https://developer.mozilla.org/en-US/docs/Glossary/Dynamic_typing [Accessed on 31/10/2021]
- [98] McCarthy, John (1960). "Recursive functions of symbolic expressions and their computation by machine, Part I". Communications of the ACM. 3 (4): 184–195. doi:10.1145/367177.367199. S2CID 1489409. Retrieved 2009-05-29.
- [99] <https://numpy.org/> [Accessed on 31/10/2021]
- [100] <https://pandas.pydata.org/> [Accessed on 31/10/2021]
- [101] <https://www.fsf.org/> [Accessed on 31/10/2021]
- [102] <https://pytorch.org/> [Accessed on 31/10/2021]
- [103] <http://torch.ch/> [Accessed on 31/10/2021]
- [104] <https://scikit-learn.org/stable/> [Accessed on 31/10/2021]
- [105] <https://huggingface.co/transformers/> [Accessed on 31/10/2021]
- [106] <https://matplotlib.org/> [Accessed on 31/10/2021]
- [107] <https://seaborn.pydata.org/> [Accessed on 31/10/2021]
- [108] <https://flask.palletsprojects.com/en/2.0.x/> [Accessed on 31/10/2021]
- [109] <https://jinja.palletsprojects.com/en/3.0.x/> [Accessed on 31/10/2021]
- [110] <https://networkx.org/> [Accessed on 31/10/2021]
- [111] <https://github.com/twintproject/twint> [Accessed on 31/10/2021]

- [112] <https://graphalchemist.github.io/Alchemy/#/> [Accessed on 31/10/2021]
- [113] <https://getbootstrap.com/> [Accessed on 31/10/2021]
- [114] <https://www.tweepy.org/> [Accessed on 31/10/2021]
- [115] <https://selenium-python.readthedocs.io/> [Accessed on 31/10/2021]
- [116] <https://github.com/mozilla/geckodriver/releases> [Accessed on 31/10/2021]
- [117] <https://cloud.google.com/> [Accessed on 31/10/2021]
- [118] Goodhart, Charles (1975). "Problems of Monetary Management: The U.K. Experience". Papers in Monetary Economics. 1. Sydney: Reserve Bank of Australia.
- [119] Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 1135-1144. 10.1145/2939672.2939778.
- [120] Ethnologue: Languages of the World, 2019 Edition.
- [121] <https://developer.twitter.com>
- [122] Shoshan Zuboff, "A Digital Declaration: Big Data as Surveillance Capitalism", September 15 2014, Frankfurter Allgemeine Zeitung
- [123] Zuboff, Shoshana (2015). "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization". Journal of Information Technology. Rochester, NY. 30: 75–89. doi:10.1057/jit.2015.5. SSRN 2594754.
- [124] Bradford Cornell & Aswath Damodaran (2020) The Big Market Delusion: Valuation and Investment Implications, Financial Analysts Journal, 76:2, 15-25, DOI: 10.1080/0015198X.2020.1730655