TURKISH CLICKBAIT DETECTION IN SOCIAL MEDIA VIA MACHINE LEARNING ALGORITHMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ŞURA GENÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

AUGUST 2021

Approval of the thesis:

TURKISH CLICKBAIT DETECTION IN SOCIAL MEDIA VIA MACHINE LEARNING ALGORITHMS

Submitted by ŞURA GENÇ in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Sciences Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Dr. Ceyhan Temürcü
Head of Department, **Cognitive Science** _____

Asst. Prof. Dr.Elif Sürer
Supervisor, **Modeling and Simulation, METU** _____

Asst. Prof. Dr. Murat Perit Çakır
Co-Supervisor, **Cognitive Science, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science, METU _____

Asst. Prof. Dr. Elif Sürer
Modeling and Simulation, METU _____

Asst. Prof. Dr. Murat Perit Çakır
Cognitive Science, METU _____

Asst. Prof. Dr. Murat Ulubay
Business Dept., Yıldırım Beyazıt University _____

Asst. Prof. Dr. Barbaros Yet
Cognitive Science, METU _____

Date: *26.08.2021*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Şura Genç

Signature : _____

ABSTRACT

TURKISH CLICKBAIT DETECTION IN SOCIAL MEDIA VIA MACHINE LEARNING ALGORITHMS

Genç, Şura

MSc., Department of Cognitive Sciences

Supervisor: Asst. Prof. Dr. Elif Sürer

Co-Supervisor: Asst. Prof. Dr. Murat Perit Çakır

August 2021, 83 pages

Clickbait strategy, mostly used in headlines and teaser messages, aims to attract people's attention, and make them click on the link by using intriguing expressions with various text-related features. Clickbait, which has become very common especially in social media in recent years, is a major problem for the flow of information. Since the information promised in the clickbait headline is generally not included in the main text, clickbait headlines disappoint readers and is problematic for ethics of journalism. In this thesis, we constructed a Turkish dataset –ClickbaitTR– with 48,060 samples, including headlines of Turkish news sources extracted from Twitter, and made it publicly available. Various machine learning algorithms such as Artificial Neural Network (ANN), Logistic Regression (LR), Random Forest (RF), Long Short-Term Memory Network (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Ensemble Classifier (EC) were applied on the dataset for detecting the clickbait headlines. The results show that the BiLSTM has the best performance in detecting clickbait headlines with 97% accuracy followed by the LSTM, the ANN, and the Ensemble Classifier with 93% accuracy. In addition to a successful clickbait detection performance, in this thesis, linguistic and psychological analysis of clickbait sentences were presented with a focus on psychological mechanisms such as curiosity and interest. This thesis contributes to clickbait detection studies with the largest clickbait dataset and best clickbait detection performance in Turkish.

Keywords: Clickbait detection, dataset formation, news headlines, machine learning, artificial neural networks.

ÖZ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE SOSYAL MEDYADA TÜRKÇE CLİCKBAİT TESPİTİ

Genç, Şura

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi Elif Sürer

Yardımcı Tez Yöneticisi: Dr. Öğr. Üyesi Murat Perit Çakır

Ağustos 2021, 83 sayfa

Çoğunlukla başlıklarda ve tanıtım mesajlarında kullanılan clickbait (tık tuzağı) stratejisi, metinle ilgili özelliklere sahip bazı merak uyandırıcı ifadeleri kullanarak insanların dikkatini çekmeyi ve onların bağlantıya tıklamalarını sağlamayı amaçlamaktadır. Son yıllarda özellikle sosyal medyada oldukça yaygın hale gelen tık tuzağı, bilgi akışı açısından büyük bir sorun teşkil etmektedir. Clickbait başlıkta vaat edilen bilgiler genellikle ana metinde yer almadığından, clickbait başlıklar okuyucuları hayal kırıklığına uğratmaktadır ve gazetecilik etiği açısından sorunludur. Bu tezde, Twitter'dan alınan Türk haber kaynaklarının manşetlerini de içeren 48.060 örnek ile bir Türkçe veri seti –ClickbaitTR– oluşturulmuş ve veri seti açık kaynak olarak paylaşılmıştır. Clickbait haber başlıklarının tespit edilmesi için bu veri seti üzerinde Yapay Sinir Ağları, Lojistik Regresyon, Rastgele Orman, Uzun-Kısa Süreli Bellek Ağı, Çift Yönlü Uzun-Kısa Süreli Bellek ve Topluluk Öğrenmesi gibi çeşitli makine öğrenmesi algoritmaları uygulanmıştır. Sonuçlar %97 doğruluk oranına sahip olan Çift Yönlü Uzun-Kısa Süreli Bellek algoritmasının clickbait başlıkları tespit etmede en iyi performansa sahip olduğunu ve ardından %93 doğrulukla Uzun-Kısa Süreli Bellek, Yapay Sinir Ağları ve Topluluk Öğrenmesi algoritmalarının geldiğini göstermektedir. Bu tezde başarılı bir clickbait tespiti performansının yanı sıra, clickbait cümlelerin merak ve ilgi gibi psikolojik mekanizmalarına odaklanılarak dilbilimsel ve psikolojik analizi sunulmuştur. Bu tez, en büyük clickbait veri seti ve Türkçe'deki en iyi clickbait tespiti performansı ile clickbait tespiti çalışmalarına katkıda bulunmaktadır.

Anahtar Sözcükler: Tık tuzağı tespiti, veri seti oluşturma, haber başlıkları, makine öğrenmesi, yapay sinir ağları.

To Oğuzhan, who never stopped resisting.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my supervisor Asst. Prof. Dr. Elif Sürer for encouraging me by seeing the potential in the earlier version of this study and for supporting me in every way at every stage of the study. I would like to thank my co-supervisor, Asst. Prof. Dr. Murat Perit Çakır for his contributions and support throughout my thesis. I am extremely grateful to my supervisors for their guidance and endless patience.

Besides my supervisor, I would like to thank Cem Bozşahin, who expands my knowledge by sharing his ideas and experiences about science and cognitive science during my graduate.

I would also like to thank my mother Bahar, my brother Ahmet Emin and my sister Kübra, who were always ready to help. I owe special thanks to my sister Beyza for being there for me whenever I need her.

I also owe special thanks to my second family, Filiz, Mustafa, Oğuzhan, Mervenur, and Metehan Türkmen, who always encourage me, appreciate my efforts, and share my burden.

Many thanks to my friends Deniz Korkmaz, Görkem Göven, Deniz Pala, Celil Sevincik, Gizem Özdemir and Şeyma Nur Ertekin for their friendship and always being a source of help.

This study owes to my best friend and soul-mate, Hamit Başgöl. It would not be possible to complete this thesis without his intellectual richness, unofficial supervision, and sensibility for my troubles; we did it together.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xiii
CHAPTERS
INTRODUCTION 1
1.1. Motivation of the Study 1
1.2. Purpose of the Thesis 2
1.3. Contributions and Novelties 3
1.4. Outline of the Thesis 3
LITERATURE REVIEW 5
2.1. Clickbait 5
2.1.1. Structure of Clickbait 5
2.1.2. Varieties of Clickbait 7
2.2. Cognitive Mechanisms of Clickbait 7
2.2.1. The Role of Curiosity in Clickbait Strategy 7
2.2.2. The Role of Interest in Clickbait Strategy 10
2.3. Clickbait Detection 11
2.3.1. Clickbait Detection in English 11
2.3.2. Clickbait Detection in Other Languages 13
2.3.3. Proposed Study 14
METHODOLOGY 17

3.1. Data Collection....17
3.1.1. Background Information on Twitter Usage in Turkey....17
3.1.2. Platform Selection....19
3.1.3. Publisher Selection....21
3.1.4. Resources with Clickbait News Headlines....21
3.1.5. Resources with Non-Clickbait News Headlines....24
3.1.6. Data Extraction....24
3.2. Dataset Construction....28
3.2.1. Scanning the Collected Data....28
3.2.2. Data Pre-processing....35
3.2.3. Dataset Validation....37
3.3. Data Analysis....38
3.3.1. Feature Selection....38
3.3.2. Machine Learning Models....38
3.3.3. Confidence Analysis....41
RESULTS....43
4.1. Artificial Neural Network....43
4.2. Logistic Regression....45
4.3. Random Forest....46
4.4. Long Short-Term Memory....47
4.5. Bidirectional Long Short-Term Memory....47
4.6. Ensemble Classifier....49
4.7. Model Validation....51
4.8. Statistical Tests....52
4.9. Model Confidence....52
DISCUSSION AND CONCLUSION....59
5.1. Further Studies....65
REFERENCES....67
APPENDICES....73
APPENDIX A....73
APPENDIX B....76

APPENDIX C 78
APPENDIX D 81
APPENDIX E.....82

LIST OF TABLES

Table 1: News sources retweeted by Limon Haber and Spoiler Haber. 23
Table 2: The sources and quantities of clickbait and non-clickbait tweets. 26
Table 3: The features presented in the dataset 27
Table 4: The overall summary of the data acquisition procedure. 27
Table 5: The favorite and retweet frequencies of three tweets posted by Limon Haber. 36
Table 6: Hyperparameters of MLP 1 and MLP 2. 45
Table 7: Results of the algorithms used in this study 50
Table 8: Examples of news headlines from the dataset with the confidence values of the models and raters.....56

LIST OF FIGURES

Figure 1: Word cloud showing the most frequent words in clickbait data. 29
Figure 2: Word cloud showing the most frequent words in non-clickbait data 31
Figure 3: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from the dataset 33
Figure 4: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 2 (magazine) from the clickbait data 33
Figure 5: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from the non-clickbait data.....34
Figure 6: Comparison between dataset and human scores.....37
Figure 7: Model accuracy of the first MLP model.......43
Figure 8: Model accuracy of the second MLP model 44
Figure 9: Feature importances obtained from the second MLP 2 model. 46
Figure 10: Receiver operating characteristic (ROC) curve of the Logistic Regression model.....47
Figure 11: Feature importances obtained from the Random Forest model.....48
Figure 12: Model accuracy of the LSTM model.....48
Figure 13: Model accuracy of the BiLSTM model.....49
Figure 14: Comparison of performances of Logistic Regression, Random Forest, ANN1, ANN2, LSTM, BiLSTM and Ensemble Classifier 50
Figure 15: The accuracy scores of the models and three raters (calculated by taking the mean value of raters as the ground truth.).....51
Figure 16: Comparison between the confidence scores of the models and three annotators. 53
Figure 17: Confidence frequencies of the models and annotator mean. 54
Figure 18: Confidence-based features of the BiLSTM model. 55
Figure 19: Confidence-based features of the three raters.....56

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
MLP	Multilayer Perceptron
LR	Logistic Regression
RF	Random Forest
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
EC	Ensemble Classifier
CNN	Convolutional Neural Networks
LDA	Latent Dirichlet Allocation
NBC	Naive Bayes Classifier
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TANH	Hyperbolic Tangent Function
NLP	Natural Language Processing
ADAM	Adaptive Momentum

CHAPTER 1

INTRODUCTION

Headlines and teaser messages of online content consist of texts, images, or links about the content and provide people with a short introduction message about the content of the main material (news, video, and article), highlighting the critical points. However, they are also used to attract people to see the actual content by giving limited, incomplete, or incorrect information about it, aiming to draw people's interest rather than summarizing the main points. Reading such headlines or teaser messages contradicts people's demand for information about the content and drives them to read the main content. This strategy, which aims to deceive the reader and creates a discrepancy between the aims of the reader and the publisher, is called clickbait (Potthast et al., 2018). In other words, clickbaits are headlines and teaser messages that arouse people's curiosity and direct them to click on a link that will lead them to specific content. These short text messages present information that is not included in the primary material or reflects weak information. The number of clicks an online content receives is closely related to the income it receives from advertisements. For this reason, the clickbait strategy is used very often in online content. Besides, using this strategy, the content title becomes a self-feeding advertisement by advertising itself, which is highly preferred by publishers.

1.1. Motivation of the Study

With the widespread use of social media worldwide, sharing news by news outlets on social media has also become widespread. This means that more and more people start accessing the news on various social media platforms. A 1-year survey conducted in 2019 shows that 57% of internet users worldwide indicated that social media platforms had been increasingly used for accessing information. Social media is used by 53.6% of the world's population, with an average of about two and a half hours daily (Research by Global Web Index, 2019). 28% of people across all countries get the daily news from a website or app, and young people (18–24) are more than twice as likely to access news via social media (Reuters Institute Digital News Report, 2020). This situation brings some discussions about the reliability and accuracy of the news shared on social media. The accuracy of the content and

presentation of the news is still a major problem to be addressed. Although people worldwide view social media as the most unreliable source for news and global concerns about misinformation are rising, many people continue to use social media to access news. More than 50 percent of Poland, Bulgaria, Hungary, Romania, Slovakia, Portugal, and Croatia access daily news from social media platforms, even though less than 35% of adults in Europe find social media reliable (Statistica, 2019).

The prevalence of the use of social media to access news in Turkey shows similarities with the situation in the world. According to recent statistics, over 71 million internet users in Turkey will access social networks in 2025, up from 54.34 million social media users in 2020. Furthermore, Reuters Institute Digital News Report 2020 shows that 85% of people access news online in Turkey while 58% of them used social media as a source of news.

These projections show that social media platforms are used extensively throughout the world and Turkey as a source of news. This indicates that news headlines designed for clickbait posted on social media will continue to be a major problem for users who follow the news on these networks. On the one hand, many users become a part of this process simply by reading these false, exaggerated, provocative, or incomplete headings. On the other hand, many users support, intentionally or not, these news sources that violate journalism ethics by clicking on the content directed by clickbait headlines. The primary motivation for this study is to identify Turkish clickbait news headlines that lead to these fundamental problems related to the reliability of information and ethics. Besides, insufficient data to understand the variations and mechanisms of the clickbait strategy in Turkish and compare these mechanisms with those in other languages is one of the motivations of this thesis.

1.2. Purpose of the Thesis

As will be explained in detail in the platform selection section, Twitter has been chosen to create a dataset and detect clickbait in this thesis. Twitter is ranked sixteenth with 353 million users worldwide among other social media platforms in 2021. It has 23% of the global active usage among Facebook with 63%, Youtube with 61%, WhatsApp with 48%, Facebook Messenger with 38%, and Instagram with 36% global usage. Clickbait strategy is also a major problem for Twitter users who use it to access online news because news channels' Twitter accounts frequently use clickbait headlines. Statistics show that 10 percent of users who use social media to access news were Twitter users in February 2019 (Statista, 2019).

Turkey is ranked seventh based on the number of Twitter users with 13.6 million users in 2021, indicating that clickbait usage in Twitter accounts of news channels may constitute a major problem for users in Turkey (Statista, 2019). These data reveal that clickbait news headlines are also a growing problem in Turkey, and therefore there is a need for more research on this issue. The datasets that contain Turkish news headlines and especially the data of a platform like Twitter where news

sources frequently share news every day are essential in this respect. Although clickbait studies have been done in various languages such as English (Uddin Rony et al., 2017). Thai (Wongsap et al., 2018) and Chinese (Zheng et al., 2018), Turkish studies on this issue are still needed. Geçkil et al.'s (2018) dataset is the only Turkish dataset including Turkish news headlines for clickbait detection.

One of the main objectives of this thesis is to create a dataset containing sufficient data to identify Turkish clickbait sentences and understand the basic mechanisms in these sentences. Another aim of this thesis is to develop models that can detect Turkish clickbait news headlines and compare their detection performance. Finally, one of the aims of this thesis is to determine the basic features of Turkish clickbait sentences and examine the importance of these features.

1.3. Contributions and Novelties

In this thesis, a dataset —ClickbaitTR— containing 48,060 Turkish news headlines was formed and made publicly available to be used in other studies. ClickbaitTR is the largest and most comprehensive dataset available in Turkish, which provides a rich resource for Turkish clickbait studies and cross-language clickbait studies (Genç & Surer, 2021).

Besides, several machine learning algorithms were applied to this dataset for comparative analysis on Turkish clickbait detection. The results show that Logistic Regression performs with an accuracy of 0.85 with an F1-score of 0.85; Random Forest performs with an accuracy of 0.86 an F1-score of 0.86; Long Short-Term Memory (LSTM) performs with an accuracy of 0.93 and F1-score of 0.93; Artificial Neural Network (ANN) performs with an accuracy of 0.93 with and F1-score of 0.94; Ensemble Classifier performs with an accuracy of 0.93 and F1-score of 0.94; and Bidirectional Long Short-Term Memory (BiLSTM) performs with an accuracy of 0.97 and F1-score of 0.96. Among these models, BiLSTM, LSTM, ANN and Ensemble Classifier present the best results in Turkish clickbait detection.

Finally, important features obtained from ANN, Random Forest, BiLSTM algorithms and annotators provide critical information about the structure and mechanism of Turkish clickbait sentences. These critical features are thought to be a starting point in other Turkish clickbait studies.

1.4. Outline of the Thesis

The organization of the thesis is as follows: Chapter 2 presents a detailed literature review of the various datasets, including clickbait datasets formed with Twitter data and the relevant clickbait studies; Chapter 3 gives the details of tweets posted by Turkish news sources and explains the data collection process including the selection processes of social media platform and news sources. This chapter clarifies the dataset construction process from the collected data giving details about the

properties of the data and the data pre-processing. This chapter also introduces the data analysis process, including six different machine learning algorithms and confidence analysis. Chapter 4 introduces the results of six different analyses conducted for comparison using Artificial Neural Network, Logistic Regression, Random Forest, Long Short-Term Memory, Bidirectional Long Short-Term Memory, and Ensemble Classifier algorithms and discusses them in detail. In Chapter 5, a detailed discussion on the analyses, the conclusion, and future work can be found.

CHAPTER 2

LITERATURE REVIEW

2.1. Clickbait

Clickbait refers to headlines or teaser messages that use intriguing or complex phrases or patterns being designed to attract readers to particular content. These headlines or teaser messages, which are mostly shared with the link of the main content, can also be considered as ads that aim to increase the click rates of readers or social media users. In recent years, the widespread use of social media and access to news from social media platforms have led to the more frequent preference of such news headlines by news sources. This strategy, which has many different types and tools, has a specific structure, and it is essential to understand it in detail and to strengthen the studies for identifying clickbait headlines. In the next section, certain studies to understand the structure of clickbait and non-clickbait news headlines can be found.

2.1.1. Structure of Clickbait

In studies where clickbait sentences are analyzed, the basic properties of clickbait and non-clickbait sentences have been investigated. When these properties are examined, it is seen that text-related strategies are commonly used in clickbait sentences, and they are essential for detecting clickbaits. Studying fake news, Hardalov et al. (2016) detected certain credibility features essential for distinguishing credible news from fake news. They chose twenty crucial features based on the literature and conducted several experiments using these features' combinations to select the best features. When these features are examined, it is seen that some of these features reflect the clickbait news headlines. For example, clickbait headlines include more negative words such as not, no, and nobody, more uppercase letters, and more exclamation marks.

In clickbait detection studies, several studies aimed to determine the fundamental properties that make a sentence clickbait. Chakraborty et al. (2016) carried out a

linguistic analysis comparing the clickbait and non-clickbait headlines using the Stanford CoreNLP tool, and they detected specific semantic and syntactic nuances typical in clickbait sentences. They discussed the properties of clickbait and non-clickbait headings separately at the sentence level, word level, and word group level.

In their study, the sentence structure of clickbait and non-clickbait sentences were compared based on three features: the length of the headlines, the length of the words, and the length of the syntactic dependencies. According to this study, clickbait headlines (an average of 10 words) are longer than non-clickbait headlines (an average of 7 words) in English. This is because non-clickbait sentences mostly include content words referring to specific locations or persons omitting the function words. In contrast, clickbait sentences mostly contain function words such as determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers, and question words, as well as content words. On the other hand, the average word length is shorter in clickbait headlines because of shorter function words. A syntactic dependency reflects the relation between two words (a governing and a dependent word) of a sentence and plays a critical role in semantic interpretation (Gamallo Otero, 2008). According to Chakraborty et al. (2016)'s study, the distance between governing words and dependent words is longer in clickbait sentences, meaning they have longer syntactic dependencies. They explained this by the fact that clickbait sentences contain more complex phrasal sentences, which is similar to the fact that clickbaits generally include both content words and function words.

In Chakraborty et al. (2016)'s study, word patterns of clickbait and non-clickbait headlines were examined based on the part of speech (POS) tags and word n-grams which refer to the consistent word sequences in the text. They extracted 1, 2, 3, and 4-grams from their dataset and found that while 19% of non-clickbait data contains n-grams, 65% of clickbait data contains n-grams such as "see what happened" or "guess who." On the other hand, when they investigated speech tags, they discovered that while clickbait headlines include more determiners and adverbs, non-clickbait headlines mostly include proper nouns, making sense because non-clickbaits generally refer to specific persons and locations. Besides, they found that personal and possessive pronouns such as she and her are frequently used in clickbait headlines.

Another study investigating the features of clickbait teaser messages is Potthast et al. (2016)'s clickbait detection study. They analyzed both the news headline in the tweet and text features of the web pages linked from tweets. Character n-grams, word n-grams, hashtags (#), mentions (@), image tags for images, number of dots, and mean word length were among the most important features. These results are consistent with Chakraborty et al.'s (2016) and Hardalov et al. (2016)'s studies regarding n-grams and mean word length. On the other hand, the number of hashtags (#) and the number of mentions (@) are Twitter-specific features distinctive for clickbait and non-clickbait headlines extracted from Twitter.

2.1.2. Varieties of Clickbait

There are other types of clickbait besides those in which the title does not reflect the main content or reflects the information in the main content in an exaggerated manner. Biyani, Tsioutsiouliklis, and Blackmer (2016) pointed out eight clickbait varieties with different strategies: exaggeration, formatting, teasing, inflammatory, graphic clickbait, bait-and-switch, ambiguous, and wrong. “Whoever does this in trouble! Punishment after punishment...” is an example of an exaggeration since it presented the content, which mentioned that the driver’s license was confiscated for a while and a fine was imposed due to drifting with his car, exaggeratedly. “FLASH! Sale banned! 18 tons caught at the border...” is an example of formatting, which refers to the overuse of capital letters and punctuation marks in the headline. “She buried an egg and a banana in the ground, look what happened?” is an example of teasing, which means deliberately removing basic information about the content from the title. “This will make you say, “No way. He has a BMW but looks what he’s done!” is an example of inflammatory clickbaits due to using rude or provocative expressions in the headline. “Did you know this celebrity? She would look like this if she didn’t get plastic surgery!” graphic clickbait, which refers to headlines reflecting the content in a disturbing or salacious way. “The date when Clubhouse will be activated on Android devices has been announced.” is an example of bait-and-switch clickbait because it promises information, the date when Clubhouse will be activated on Android devices, which is not included in the main text. “The world’s most dangerous countries were announced: Turkey’s ranking is surprising!” is an ambiguous clickbait headline clickbaits because it can be interpreted in two different ways: Turkey’s ranking is pretty good, or it is too bad. “Bad news from the famous actor Birkan Sokullu!” is an example of wrong clickbait headlines because it reflects the news in the wrong way, although Birkan Sokullu said that he hurt his foot while walking and that he is fine now.

In addition to these eight different strategies, Pujahari and Sisodia stated three variants of clickbait headlines as incomplete, headline cloning, and URL redirection. In incomplete clickbait headlines, a part of the title is deliberately left incomplete in order to increase people's curiosity, like in the following one: “The police found him. After he did not speak for a long time...”. Headline cloning refers to using a headline for various contents, although it does not reflect the content. For example, “Have the driver's license exam results been announced?” is used not only for content related to driver's license exam results, but also for content that is not directly related to the exam. The clickbait headlines in which URL redirection redirects people to an unrelated website increase the number of clicks on that website.

2.2. Cognitive Mechanisms of Clickbait

2.2.1. The Role of Curiosity in Clickbait Strategy

The sense of curiosity is an essential part of gaining new knowledge, producing scientific knowledge, lifelong development, literature, and art by motivating people

to examine the new stimulus and learn more about it. It is also frequently used for commercial purposes (Loewenstein, 1994) like it is in the clickbait headlines due to its ability to push people to acquire new knowledge. In scientific developments, in addition to the ability of people to establish causal relationships, the curiosity that initiates the information-seeking behavior is also essential (Gottlieb et al., 2013). Exploratory behaviors triggered by curiosity differ from other motor behaviors in that it changes the observer's epistemic state rather than her external world. These behaviors aim to reduce or eliminate uncertainty by making changes in the epistemic state of the person. Actions driven by randomness, novelty, uncertainty, or surprise can be classified as intrinsically motivated since these features of the environment reward them. For example, sensory novelty enhances neural responses in certain brain areas such as temporal, frontal, visual lobe, and dopaminergic areas. Like novelty, the surprising stimulus is intrinsically motivated because it is salient and causes attentional attraction. Although novel or surprising stimuli lead agents to curiosity-relieving behaviors, they do not guarantee to learn and may create an information gap in the agent (Gottlieb et al., 2013).

Loewenstein (1994) defined the knowledge gap as the discrepancy between one's current state of knowledge and the new information encountered. In line with Berlyne (1960)'s proposed characteristics, he argues that when people meet with new, complex, surprising, and ambiguous stimuli, they will behave in a way that try to reduce this informational gap between what they know and what they encounter by resolving the uncertainty. It is thought that one of the main factors why clickbait works may be that the attention of the people reading the clickbait text is focused on this information gap in their mind (Potthast et al., 2016). This also explains why people who read main content after clickbait headlines are frustrated. People want to reduce the information gap after reading these headlines, but they are not satisfied because the content often does not match the information that the clickbait headline promises.

According to Berlyne (1966), there are four types of curiosity: perceptual curiosity, epistemic curiosity, specific exploration, and diversive exploration. Perceptual curiosity refers to conditions triggered by novel, complex, surprising, or ambiguous stimuli. Epistemic curiosity is described as the desire to acquire knowledge. Specific exploration refers to the condition stemming from lack of information and uncertainty, leading people to seek information. It can also be said that this is a general definition of curiosity. Diversive exploration arises when a subject has no stimulus to be curious about, namely in a monotonous environment. Considering these varieties of curiosity, clicking the links of clickbait news headlines can be classified as both perceptual curiosity and epistemic curiosity since they generally present novel, complex, surprising, or ambiguous stimulus patterns. Curiosity for clickbait headlines can be triggered by the odd punctuation patterns or intriguing expressions used in these headlines and the information gap created by the content of the title.

Berlyne (1960) proposed that the new stimuli evoking curiosity in people may be investigated by some of their characteristics such as novelty, surprisingness, complexity, and ambiguity, which play a critical role in the relationship between people's internal knowledge and the external information surrounding them. Novelty refers to the characteristic of stimulus or situations that contain new information for the person who encounters them. "Scientists did research! Does listening to music affect the climate?" is an exemplary news headline for novelty because it contains a question that most people have probably never considered or encountered before. Surprisingness indicates that the event or stimulus is unexpected. For example, "Attention: Fish are crossing over the street!", a headline about flood news can be given as an example of this characteristic because it gives unexpected information about the behavior of the fish. Complexity refers to situations involving multiple components that must be evaluated at the same time. For example, "They saw it in the pool! They anesthetized it when no one could pull it out..." is a complex news headline containing many curious questions like who saw something in the pool, what they saw, why they wanted to take it out of the pool and why they could not, and why they knocked it out. The stimuli or the situations that can be interpreted in two different ways because of a lack of information constitute an ambiguity (Loewenstein, 1994). This characteristic of the stimulus or the situation creates a difference between what one knows and what one is interested in knowing about the stimulus, which is defined as a knowledge gap by Loewenstein (1994). For example, "Gunshot sounds in London, the capital of England!" is an ambiguous news headline because it may refer to an armed attack or a military drill.

Different types of clickbait detected by Biyani et al. (2016) and Pujahari and Sisodia (2020) can be evaluated in terms of the characteristics of curiosity such as novelty, complexity, surprisingness, and ambiguity. For instance, graphic clickbaits can be associated with surprisingness and novelty as they try to raise readers' curiosity by presenting unbelievable, disturbing, or salacious titles. Similarly, wrong clickbaits can be related to novelty and surprisingness since they arouse curiosity with false information. On the other hand, the exaggeration strategy in clickbaits may be associated with the surprisingness aspect of stimuli. Complexity can be encountered in formatting clickbaits in which textual features are changed inappropriately or inflammatory clickbaits in which rude or provocative expressions are used. These headings, which include incorrect use of punctuation rules or unfamiliar rude expressions, may seem difficult and complicated to understand. Some types of clickbait such as teasing, bait-and-switch, ambiguous, and incomplete can be seen as ambiguous since they work by trimming the details or including unclear and confusing expressions. When clickbait strategies are examined in this way from different characteristics of curiosity, it can be seen that clickbait does not have a single mechanism and encapsulates the sense of curiosity in all aspects.

There are two approaches for explaining curiosity-driven behavior as novelty-based theories and complexity theories. Novelty-based theories hypothesize that curiosity is an organism's motivation to seek new stimuli, and learning about it is rewarding. Some neuroscientific studies show that the reward-responsive areas in the brain are

sensitive to novel stimuli (Düzel et al., 2010). In addition, some computational studies show that searching for novelty can be efficient (Tang et al., 2017). Although all these studies support novelty-based theories, these theories have some weaknesses. For example, although novelty-based theories suggest that the agent's discovery of new stimuli is optimal for the agent, new stimuli may not always be informative and useful and may direct the agent to ambiguous results (Dubey & Griffiths, 2020).

On the other hand, complexity theories suggest that stimuli or events of medium complexity trigger curiosity. Loewenstein's (1994) knowledge gap hypothesis mentioned above can also be classified in this group. Since the complexity of the stimulus is related to the agent's prior knowledge about that stimulus, curiosity is usually represented by an inverted U-shaped function. The individuals feel more curious about the stimuli that they are moderately confident (stimuli with medium complexity). The shortage of these theories is that they cannot explain how the agent is curious about the stimuli for which they had no prior knowledge (novel stimuli).

Dubey and Griffiths reconciled these two accounts showing an information-seeking rational agent. Results show that curiosity is determined by the context presented by the environment, together with the frequency of encountering a stimulus/task before (past exposure) and the likelihood of encountering a stimulus/task later (future occurrences). If past exposure and future occurrences to a task or stimulus are independent, low confidence increases the value of information, which refers to novelty-based curiosity (Brändle et al., 2020). For example, if these results are considered in terms of clickbait news headlines, people who are not interested in the agenda (i.e., those who have not been exposed to the agenda too much) may be more curious about the clickbait headlines when there is an event on the agenda that they are likely to encounter in the future. On the other hand, if future occurrences and past exposure are highly related, then the moderate confidence about the information increases the value of information, which refers to complexity-based curiosity (Brändle et al., 2020). For example, if a person who is highly concerned with the agenda sees clickbait news headlines about events she has read frequently in the past, she will be more curious about clickbait news headlines on these events. When all these discussions and results are examined, it is seen that curiosity is not determined by a single factor but depends on many variables, such as the agent's internal state and the characteristics of the stimulus. For this reason, these factors should be taken into consideration when examining clickbait news headlines and studies in which different theories can be reconciled should be designed.

2.2.2. The Role of Interest in Clickbait Strategy

Interest is an important part of curiosity. Based on the idea that evaluations (appraisals) about an event generate emotions, Silvia states that certain features of an event or a stimulus causes appraisals that evoke interest. Those features are novelty, complexity, and comprehensibility of an event, consistent with Berlyne's (1960) framework (Silvia, 2005). Silvia's research (2005) shows that encountering more

complicated stimuli with the cues indicating that the stimuli can be understood arose people's interest. These findings indicate that when people encounter a new and complex stimulus or event, they also evaluate whether their knowledge is enough to comprehend it. These features can address mechanisms underlying clickbait headlines since these headlines promise new and sophisticated information.

2.3. Clickbait Detection

2.3.1. Clickbait Detection in English

With the widespread use of clickbait in all online platforms, the number of studies on clickbait detection has been increased. Researchers are trying to identify clickbaits in video or text data obtained from different social media platforms such as Twitter, Youtube, and Facebook.

Qu et al. (2018) investigated the clickbaits in the teaser information of videos creating a YouTube clickbait dataset. Teaser information of videos includes title, thumbnail, the cover images of videos, and the first 123 characters of the video description. They realized that this teaser information is not sufficient for clickbait detection, and therefore, the further details of the video should also be evaluated. In this dataset, 109 YouTube videos were annotated as clickbait and non-clickbait by two reviewers based on the title, description, thumbnail, and video comments after watching the video.

Shang et al. (2019) developed Online Video Clickbait Protector (OVCP) to identify clickbait videos by reviewing the comments of people who watched the video. They collected a dataset consisting of Youtube videos to evaluate the performance of OVCP. Three annotators labeled the videos in the dataset in terms of their clickbaitness. This dataset consists of the label of videos in terms of clickbaitness, title, description, thumbnail, comments, and comment threads of videos.

Apart from Youtube datasets, there are also various clickbait datasets created with data from Facebook. Rony et al. (2017) created a dataset consisting of 1.67 million Facebook posts posted by 153 media organizations. They applied their word-embedding-based clickbait detection model to this dataset, and the model performed 98.3% accuracy.

Chakraborty et al. (2016) constructed another dataset containing 32,000 headlines of news articles gathered from The Guardian, New York Times, BuzzFeed, Upworthy, ViralNova, Thatscoop, WikiNews, Scoopwhoop, and ViralStories. The articles labeled by three annotators are represented as an almost equal number of clickbaits and non-clickbaits in the dataset. On this dataset, they applied three machine learning algorithms as Support Vector Machines (SVM), Decision Trees, and Random Forests, and SVM performed best with an accuracy of 93%. Anand, Chakraborty, and Park (2017) also applied BiLSTM on the same dataset, and it performed with an accuracy of 98%.

Apart from platforms such as Youtube or Facebook, Twitter has also been used to obtain data in clickbait studies, as in this thesis. Because Twitter contains many news-related tweets or teaser messages with clickbait, it is a very convenient platform for collecting data. One of the early studies presenting a machine learning approach to clickbait detection in a social media stream is Potthast et al. (2016)'s study in which they created a dataset with Twitter data and applied three different machine learning algorithms on it. Firstly, they determined the accounts to be used for data extraction based on their retweet numbers. Then, the data extracted from Twitter accounts of official news outlets such as BBC News, Business Insider, Huffington Post, and BuzzFeed was annotated by three reviewers as clickbait or not for creating Twitter Clickbait Corpus. On this dataset, they applied three learning algorithms, Logistic Regression, Naive Bayes, and Random Forest. The Random Forest classifier achieved the best performance with 0.76 precision and 0.76 recall.

By expanding the previous dataset and detailing the annotation process, Potthast et al. (2018) created a new Webis Clickbait Corpus 2017. As in the previous study, the accounts from which data will be retrieved were selected based on retweet numbers and 38,517 teaser messages from accounts of 27 news publishers. Tweets and their media attachments were recorded for six months from the Twitter accounts of selected news sources such as Independent, Guardian, Businessinsider, Foxnews, CNN, and Washingtonpost. In the evaluation process, crowd workers first evaluated the clickbaitness of tweets for a fee. Annotators have been warned not to mistakenly consider tweets about gossip as clickbait and pay attention to the images presented with the headlines. Secondly, the referees were asked to note the words that caused them to label the tweets as clickbait. Finally, they were asked to evaluate the tweets on a four-option Likert scale instead of a binary classification task. The use of a non-binary task to detect the clickbaitness of the headlines is one of the most notable features of this study.

Chakraborty et al. (2017) created another dataset with the headlines taken from Twitter. The clickbait headlines in this dataset were collected from the top three newspapers (New York Times, Washington Post, and India Times) and one online news media outlet (Huffington Post) which also have 38 secondary accounts. The headlines in these 38 secondary accounts were also collected. On the other hand, non-clickbait data were extracted from 27 primary and secondary accounts of the five outlets (BuzzFeed, Upworthy, ViralNova, ScoopWhoop, and ViralStories). With the collection of the tweets of these accounts as well as their retweets, the largest dataset in English with 288K tweets and 11.4M retweets was formed over a period of 8 months.

Chakraborty et al. (2016)'s dataset was used in Lopez-Sanchez et al. (2018)'s study to provide a method adapting itself to users. They developed a Case-Based Reasoning (CBR) method combined with wor2vec, word-embeddings, and deep metric learning techniques for adaptable clickbait detection.

Except for clickbait identification and prevention work in English, there are also studies in which clickbait is used in a different context, such as the study of Bhowmik et al. (2019). They proposed that clickbait messages can engage readers with reliable health-related information rather than misleading items. They designed an experimental setup in which participants would be asked if they would like to read the article after they were presented with articles with clickbait and non-clickbait titles. They planned to ask participants if they found the article reliable and would like to share it. They offered a different perspective for clickbait titles by proposing such research on a dataset to be created for this particular purpose.

2.3.2. Clickbait Detection in Other Languages

Although clickbait detection studies are mostly done on English datasets, there are also studies with datasets created in other languages such as Chinese (Zheng et al., 2018) and Thai (Wongsap et al., 2018).

Zheng et al. (2018) constructed a dataset including 14,922 headlines taken from four famous Chinese news websites (Tencent, 163, Sohu, Sina) and well-known blogs. They proposed a Clickbait Convolutional Neural Network (CBCNN) consisting of Word2Vec models and a CNN model. They applied this model to Chinese news headlines preprocessed with stop-word filtering, part-of-speech filtering, and segmentation. CBCNN model performed with an 80.50% accuracy.

Wongsap et al. (2018) applied Decision Tree, Support Vector Machine, and Naïve Bayes on a dataset with 5,000 Thai news headlines which were labeled as clickbait and non-clickbait by two users, and another dataset consisting of special characters such as '!', '?', and '#.' The results showed that the Decision Tree classifier gave the 99.90% accuracy on the special characters dataset, which is the best performance in their study. On the other hand, the Decision Tree classifier performed 84.79% accuracy on the news headlines.

William and Sari (2020) constructed a dataset, CLICK-ID, consisting of 15,000 annotated Indonesian news headlines gathered from 12 Indonesian online news publishers. They applied BiLSTM and CNN on this dataset. Results show that BiLSTM performs with an 81% accuracy while the CNN performs with a 79% accuracy on the stemmed words of the dataset.

Geçkil et al. (2018) formed a Turkish dataset with 2000 clickbait and non-clickbait news headlines extracted from Twitter, and 2000 clickbait and non-clickbait news gathered from the websites of the selected news outlet. The data of BBC Turkish and Anadolu Agency were labeled as non-clickbait since they were considered as being unlikely to be clickbait, while the data of other media institutions such as Hurriyet, Vatan, and Sabah Daily Newspaper were labeled as clickbait. They applied the TF-IDF method on this dataset, which gave an 87% accuracy.

2.3.3. Proposed Study

It can be seen from the literature that various datasets containing clickbait and non-clickbait news headlines or teaser messages in different languages have been created, and clickbait sentences have been studied with different methods in many studies. These studies have purposes such as identifying clickbait sentences, understanding the mechanism of this strategy, and developing warning methods against these sentences. Although many datasets have English clickbait and non-clickbait samples, other languages, including Turkish, still need new studies and datasets. One of the main motivations of this study is that there is only one dataset for Turkish clickbait detection (Geckil et al., 2018).

This thesis consists of three main points which contribute to clickbait research from different aspects. Firstly, the Turkish clickbait dataset, ClickbaitTR, consisting of 48,060 samples, was created for clickbait detection (Genç & Surer, 2021). The social media platform Twitter, which is widely used globally and in Turkey, was preferred to collect data. As it will be explained later, many news sources share daily news from their Twitter accounts, and these accounts provide a suitable environment for creating datasets. For the clickbait news headlines to be included in the dataset, Limon Haber (i.e., Lemon News in English) and Spoiler Haber (i.e., Spoiler News in English) detect and share the clickbait news daily on Twitter were preferred. For non-clickbait headlines, Evrensel Newspaper and Diken Newspaper, which rarely make clickbait, were preferred. ClickbaitTR dataset was created by preprocessing the data extracted from these accounts. This dataset, divided into two categories as clickbait and non-clickbait, provides a rich resource for other Turkish clickbait detection studies and cross-language comparison studies.

Secondly, in this study, clickbaits were detected using six different machine learning algorithms: Artificial Neural Network, Random Forest, Logistic Regression, Long Short-Term Memory, Bidirectional Long Short-Term Memory, and Ensemble Classifier. Among the analyzes, Artificial Neural Network, Long Short-Term Memory, Bidirectional Long Short-Term Memory, and Ensemble Classifier gave the best performance in Turkish clickbait detection. The results show that Logistic Regression performs with an accuracy of 0.85 with an F1-score of 0.85; Random Forest performs with an accuracy of 0.86 an F1-score of 0.86; Long short-term memory (LSTM) performs with an accuracy of 0.93 and F1-score of 0.93; Artificial Neural Network (ANN) performs with an accuracy of 0.93 with and F1-score of 0.94; Ensemble Classifier performs with an accuracy of 0.93 and F1-score of 0.94; and Bidirectional long short-term memory (BiLSTM) performs with an accuracy of 0.97 and F1-score of 0.96.

Finally, the twenty most important properties, which are important in distinguishing clickbait sentences from the non-clickbait sentences, were identified via the Random Forest and Artificial Neural Network algorithms. The important features obtained from these algorithms show that in Turkish clickbait sentences, features such as the number of dots (.), the number of at signs (@), the number of hashtags (#), the

number of exclamation marks (!), the number of question marks (?) are also distinctive for clickbaits, as well as the length of the sentence, the word length and the number of capital letters. Besides, certain words such as explain (açıkla), look (bakın), flash (flaş), development (gelişme), interesting (ilginç), horror (dehşet), stated (belli), celebrity (ünlü), last (son), and she or he (o) occupy a critical place in the mechanism of Turkish clickbait sentences.

This thesis presents one of the first efforts for clickbait detection in Turkish. Clickbaits in Turkish news headlines have initially been studied in Geçkil et al. (2018)'s study, but there are some limitations. The limitations of that study have been overcome in this thesis—the insufficient number of samples, improper labeling of clickbait and non-clickbait data, and the lack of variety in the news sources. Firstly, in Geçkil et al. (2018)'s study, the news headlines of Anadolu News Agency and BBC Turkish were labeled as non-clickbait, but the information coming from the Twitter data of Limon Haber and Spoiler Haber showed that Anadolu News Agency and BBC Turkish frequently use clickbait headlines. In this thesis, the headlines of these news organizations are in the clickbait category. Secondly, the study by Geçkil et al. (2018) collected news headlines from five news sources. However, using the data of a wider range of news sources is necessary to understand the mechanism of clickbait. In this thesis, the dataset includes 60 different news organizations which cover most of the Turkish news sources, including the top-five newspapers (Sözcü, Hürriyet, Sabah, Posta, and Milliyet Newspaper) (Gazete Tirajları, 2019) and a lot of online news media outlets (Duvar, Diken and T24 Newspaper). Finally, Geçkil et al. (2018)'s dataset includes 4000 samples consisting of 2000 news headlines and 2000 news articles, which can be considered a relatively small amount of data for clickbait detection. On the other hand, 48,060 samples were used in this thesis, which presents the largest clickbait dataset in Turkish.

CHAPTER 3

METHODOLOGY

3.1. Data Collection

ClickbaitTR consists of 48,060 news headlines (24,031 clickbaits, 24,029 non-clickbaits) gathered from Twitter. Clickbait data were taken from the Twitter accounts of Limon Haber and Spoiler Haber, while non-clickbait data were taken from the Twitter accounts of Evrensel Newspaper and Diken Newspaper.

In this section, information is given about the use of Twitter in Turkey and the content of the agenda on Twitter by giving examples of Turkish tweets in order to give information about the content of the dataset created in this thesis. Then, it is explained how the platform and publisher selection is made for the dataset, and the necessary details about these publishers are presented. Finally, it is explained how data is extracted from this platform and preprocessed for the analysis.

3.1.1. Background Information on Twitter Usage in Turkey

Like the rest of the world, almost all news sources in Turkey have a Twitter account, and news is regularly shared from these accounts. In these accounts, the news is tweeted continuously throughout the day, not at certain times, for keeping up with the rapidly changing agenda. Trend topics reflect which topics are popular and are being discussed on Twitter at a given moment. Since these trends show the current agenda of the country, not the topics of the last few days, it is possible to follow the events developing at a specific location and the most up-to-date version of the shared news following the trend topics. They can be an important source of information about both world events and local events. When the trending topics in recent years are examined, it is seen that the news about politics and the economy in Turkey attracted the attention of Twitter users. Political and economic trend topics in Turkey cover women's rights, violence against women, LGBTQ rights, current developments in the Turkish economy, the value of the Turkish currency, events in the parliament, and statements by politicians.

In this thesis, the headlines mentioned as non-clickbait are the headlines that contain essential details about the news article's content, which shows that these headlines have content that matches the content of the main news article. It is not common to misuse punctuation marks and pronouns instead of nouns in these titles. After reading a non-clickbait title, people understand what the news is about and what content it has, and they decide whether to read the whole news article or not based on the information in the title. It can be said that non-clickbait Turkish news headlines generally have content related to politics and the economy. Since this news always has a large target audience, the clickbait strategy may not be used very often in these matters. The fact that the following Turkish news headlines (with their English translations) about education and economy contain detailed figures and dates indicates that these headings are non-clickbait:

a) Üniversitelerde 60 bin 674 kontenjan boş kaldı / 60 thousand 674 quotas remained vacant at universities.

b) MEB, 30 Ekim Cuma ve 2 Kasım Pazartesi günü okulların tatil edildiğini açıkladı. / Ministry of Education announced that schools were closed between Friday, October 30 and Monday, November 2.

In the following non-clickbait Turkish news headline, the statement made by a famous actor's doctor about his health condition is given:

c) Doktoru ve oğlu açıkladı: Yoğun bakımdaki Cüneyt Arkın'ın durumu iyi / His doctor and his son explained: Cüneyt Arkın, who is in intensive care, is in good condition.

Another piece of news below contains the full name of the famous theater actress instead of trying to arouse curiosity in people using pronouns. Therefore, this title can be considered non-clickbait:

d) Tiyatronun önemli ismi Nurhan Karadağ hayatını kaybetti. / Nurhan Karadağ, an important figure in theater, passed away.

The headlines referred to as clickbait in this thesis mostly contain incomplete, exaggerated, unrealistic information or information that is not included in the news article. In addition, the inappropriate use of punctuation marks (i.e. ?!!), being too short, and containing slang words are some of the characteristics of such titles. These titles are usually related to daily events, magazines, and sports topics. For example, in the Turkish news headlines below, no information has been given about the subjects of the news, and the result of the events that these subjects have experienced has been left incomplete to arouse curiosity. These titles can therefore be considered in the clickbait category:

e) Aracından gelen sesleri fark etti, kaputu açınca şaşkına döndü! / He noticed the sounds coming from his car; he was astonished when he opened the hood!

f) Yolcunun hostese verdiği not paniğe neden oldu. / Passenger's note to the hostess caused panic.

In the following clickbait Turkish news headlines, sentences were intentionally left incomplete, and a gap was created between what the readers knew and what they wanted to know:

g) Ayağınıza siyah çay ve sirke sürerseniz... / If you rub black tea and vinegar on your feet...

h) Deney kötü bitti! Termometre kırıldı, öğretmen ve öğrenciler... / The experiment turned out badly! The thermometer broke; and then teacher and students...

Some headlines try to grab readers' attention by containing intriguing questions:

i) Okullar ne zaman, hangi tarihte açılacak ve 3 aylık yaz tatili uzayacak mı? / When will the schools open, and will the 3-month summer vacation be extended?

j) Kim bunu ekmeğin içine koyar? Marketten aldığı ekmekten çıktı. / Who puts this in bread? It came out of the bread he bought from the market.

3.1.2. Platform Selection

Social media platforms include a wide range of accounts, from individual user accounts to corporate accounts of organizations and companies. Especially organizations (i.e., news media organizations) actively use social media to increase their visibility online. For this reason, posts of such accounts are very suitable materials for many different types of datasets. Many news companies also care about online journalism and share the news on various social media accounts daily. This posted news usually includes the news headline and the news article's link and/or image. As online journalism is increasingly widespread and is a competitive environment, the shared news headlines are intended to be exciting and catchy, and therefore the clickbait strategy is frequently used in these news headlines. Clickbait news headlines are shared by other users as well, and their circulation is accelerated. Since it is aimed to include Turkish news headlines in the dataset to be created for this thesis, it was thought that it would be appropriate to use one of the social media platforms to get data. Twitter, which has a more convenient API and whose user interactions are much more accessible, has been chosen among the widely used social media platforms.

As stated before, Twitter is ranked sixteenth with 353 million users worldwide among other social media platforms in 2021, which is 23% of the global active users of all social media platforms. These numbers are indicative of how popular Twitter has been lately. Users can share tweets (posts shared on Twitter) where they express an opinion, or share videos, photos, or information on any subject, with a limit of

280 characters on Twitter. The users can view the posts of other users and share (retweet) the posts (tweets) shared (tweeted) by other accounts if those accounts are not protected. Tweets are public by default on Twitter, and anyone can view and interact with the tweets of a public account. If an account protects its Tweets, only accounts allowed by this protected account can view and interact with its tweets. Data obtained from Twitter to create a dataset is collected from public accounts in this study.

In recent years, public Twitter accounts of various organizations or news outlets have become important data sources for various research domains such as Psychology, Linguistics, Computer Science, and Sociology. Twitter is used for gathering data for clickbait detection like it is in this thesis, as well. Reviewing the datasets obtained for different purposes using Twitter data and the research made with these datasets will show how valuable and essential the data collected from this platform can be. These studies are on specific topics ranging from mining Twitter for adverse drug reaction mentions (Ginn et al., 2014) to identifying fake news (Atodiresei et al., 2018).

Sentiment analysis can be given as the first example of this wide range of research using Twitter data. In sentiment analysis, the feelings people reflect in the text are detected with computational processes. Studies in extracting subjective judgments from the text have recently increased with the growing methods in natural language processing (Lin & He, 2009). The tweets being gathered for sentiment analysis datasets are usually annotated as positive, negative, neutral, or mixed, allowing researchers to analyze the sentiment of the users for many purposes (Saif et al., 2013), such as analyzing the judgments of the clients (Di. S. Sisodia & Reddy, 2018).

Natural language processing is another domain for which tweets are used in analyzing various subjects such as fake news identification using Twitter accounts of news channels (Atodiresei et al., 2018) or clickbait detection on news headlines on Twitter (Potthast et al., 2016; Zhou, 2017) as done in this thesis.

In other studies, Twitter data have also been used in pharmacology and medicine to investigate the frequency of drug side effects and catch unknown side effects (Hsu et al., 2017). Researchers gathered Twitter comments on specific drug names to form a dataset (Ginn et al., 2014). Besides, Achrekar et al. (2011) conducted a study critical for public health to predict influenza-like illness (ILI) patterns in the population and provide real-time assessment of the disease using tweets and retweets about it.

There are also various studies in which the data obtained from Twitter are used to investigate social dynamics. For example, Hernandez-Suarez et al. (2019) developed a “social sensor” methodology for detecting natural disasters, which examines the tweets that include information about a natural disaster and its location —i.e., toponym. Toponyms obtained from Twitter data proved to be informative and practical for detecting disaster areas and quickly responding to the needs. In another study that examined social dynamics using Twitter data, refugee migration patterns

were analyzed in geotagged tweets (Hübl et al., 2017). It is expected that enriched datasets gathered from Twitter will help to study refugees or other social movements more efficiently.

Considering the diversity of the studies made with the data taken from Twitter, it is seen that Twitter is a very convenient source for collecting data and creating a dataset. In addition, as explained in detail in Chapter 2, Twitter has become a frequently used resource for generating datasets for clickbait detection studies (Chakraborty et al., 2017; Potthast et al., 2018, 2016).

There are also some specific reasons for choosing Twitter for data collection in this study. Firstly, almost all Turkish news organizations actively use Twitter and share daily news by tweeting. Twitter is a rich source in terms of news headlines and titles, making this platform crucial in creating a dataset for clickbait detection. Secondly, these news sources can only include news headlines and links in tweets due to character limitations (280 characters). This means that the data to be collected has a suitable and ready-made structure for clickbait detection. Finally, since many news accounts on Twitter have regularly shared daily news for a very long time, this platform contains data that covers a wide range of time, which means that the content of the data offers a considerable variety for analysis.

3.1.3. Publisher Selection

In order to create a ready-to-use dataset for the analyses, the headlines in the dataset were collected so that they can be categorized as clickbait and non-clickbait. For this reason, accounts of news organizations on Twitter and accounts that retweet the news even though they were not official news organizations were investigated. At the end of the investigation, Limon Haber (Limon News in English) and Spoiler Haber (Spoiler News in English) were selected as the accounts to retrieve clickbait news headlines. Although they are not official news organizations, these accounts identify and share the clickbait headlines examining official news outlets' news headlines and news articles. For this reason, these accounts, which share the news they label as clickbait every day, provide a rich and reliable source for the analysis to be made in this thesis.

On the other hand, Evrensel Newspaper and Diken Newspaper were selected as sources for non-clickbait news headlines. After the data of Limon Haber and Spoiler Haber was examined, it was determined that these two newspapers rarely use clickbait. Since the clickbait strategy has been widely used on all platforms lately, finding news sources that do not share clickbait news headlines is challenging. For this reason, it was determined that only these two newspapers among Turkish news sources on Twitter would be eligible to collect non-clickbait headlines.

3.1.4. Resources with Clickbait News Headlines

As explained in the Publisher Selection section, Limon Haber and Spoiler Haber were chosen to collect clickbait Turkish news headlines. Limon Haber, which is not

an official news source, regularly detects and shares the clickbait news headlines that aim to make the users click on the news article link by arousing their curiosity. There are two ways to share other accounts' tweets on Twitter: retweet and quote tweets. If Twitter users want to share another user's tweet as it is, they retweet that tweet. If users want to share another user's tweet by quoting it and commenting on it, they use the quote tweet option. Limon Haber shares the news tweeted by official news sources by quote tweets and warns the users by providing the necessary information about that news content. This account informs users that the content and title of the news do not match, the information promised in the news headline is missing in the news article, or the news headline is exaggerated according to the content of the news. This account, which voluntarily detects clickbait for journalism ethics, offers a rich source of clickbait news headlines.

The Twitter account of Limon Haber contains 43,300 tweets (May 07, 2021). This account contains headlines from almost all official Turkish news sources and shares other news sources that do online journalism. The sources of news whose headlines were evaluated and shared as clickbaits by Limon Haber are listed in Table 1. The fact that this account shares headlines from 41 different news sources shows that the data of this account provides a wide variety and the excessive number of news headlines for the dataset. The news headlines quoted by this account are considered clickbaits.

Many Twitter users share the news headlines they read during the day, which they consider as clickbait, by mentioning/tagging Limon Haber. A mention is a tweet containing another accounts's username with the "@" symbol to address that account on the tweet. These users can be thought of as evaluators/annotators who read the news headlines with articles and decide whether the headline is clickbait or not. They also prove how accurate their labeling is, by providing information about the content of the news, what is missing in the news, what is used in the headline to arouse curiosity. These mentions tweeted by other users are retweeted by Limon Haber and therefore are included in the Twitter data of Limon Haber. This sharing and interactions strengthen the judgments that the titles taken from Limon Haber's account are clickbait.

Spoiler Haber, which is not an official news source, detects and shares the clickbait news headlines to warn the users against them. It uses the quote tweet option to share news headlines of news media outlets. When this account's tweets are examined, it can be seen that it posts more sports-related clickbait headlines. Spoiler Haber's account includes 15,400 tweets (May 07, 2021). This account posts clickbait headlines less frequently than Limon Haber.

The account of Spoiler Haber also covers the news of 44 Turkish news sources, which makes it a rich resource for the ClickbaitTR dataset. The Turkish news sources whose news headlines are retweeted by this account can be seen in Table 1.

Table 1: News sources retweeted by Limon Haber and Spoiler Haber.

News Sources Retweeted by Limon Haber		News Sources Retweeted by Spoiler Haber	
Bloomberg HT	BirGün Newspaper	Bloomberg HT	Eurosport TR
Star Newspaper	T24	Star News	TRT Sports
Hürriyet Newspaper	A Sports	Al Jazeera Turkish	TRT News
Hürriyet Gündem	Demirören News	Hürriyet Newspaper	Sabah News
Hürriyet Economy	BBC Turkish	Hürriyet Kelebek	BirGün Newspaper
Hürriyet Kelebek	OdaTV	Vatan Newspaper	Business HT
Vatan Newspaper	MedyaTava	NTV	T24
NTV	ABC	NTV Art and Culture	A Sports
NTV Art and Culture	Cumhuriyet News	NTV Money	Şampiy10
NTV Science	Gerçek Gündem	NTV Sports	AA Sports
NTV Sports	Milliyet Newspaper	NTV Life	Demirören News
CNN Turkish	Aydınlık Newspaper	NTV Health	TGRT Haber
Sözcü Newspaper	Halk TV	NTV Science	BBC Turkish
Spor Arena	Tele 1 TV	CNN Turkish	OdaTV
Habertürk Technology	Yeni Akit Newspaper	Sözcü Newspaper	MedyaTava
Posta Newspaper	Mynet	Spor Arena	ABC
Ihlas News Agency	Yeniçağ Newspaper	Habertürk Spor	Kelebek Magazine
Sputnik Turkey	DHA Art and Culture	Posta Newspaper	Fotomaç
Sol Haber	DHA Sports	Ihlas News Agency	Ajansspor Meydan
Sabah Newspaper	Futbol Arena	Sputnik Turkey	Cadde Milliyet
AA Sports		Anadolu Agency	Sporx
		Sol Haber	SKOR

3.1.5. Resources with Non-Clickbait News Headlines

As explained in the Publisher Selection section, Evrensel Newspaper and Diken Newspaper were chosen to collect non-clickbait Turkish news headlines. Evrensel Newspaper is an official news source whose Twitter account consists of 420,900 tweets (May 07, 2021). Diken Newspaper is an official news source whose Twitter account consists of 280,100 tweets (May 07, 2021).

There are two reasons why these news sources' Twitter accounts are considered as non-clickbait. The first reason is that these accounts are recommended by Limon news, which regularly shares clickbait news headlines. Annotators of the Limon News account stated that Evrensel Newspaper and Diken Newspaper rarely use clickbait and deliberately avoid this strategy. The second reason why these newspapers' tweets are included in the non-clickbait category is that Limon Haber and Spoiler Haber rarely refer to the tweets of Evrensel and Diken Newspapers in their Twitter account. Of the news headlines that Limon Haber detects and shares as clickbait, only 6 are the headlines of Evrensel Newspaper and 22 of Diken Newspaper. On the other hand, none of the tweets taken from the account of Spoiler Haber include the headline of Evrensel Newspaper or Diken Newspaper. These data prove that these two newspapers do not prefer the clickbait strategy, and therefore, their Twitter data can contribute to the non-clickbait news headlines category.

3.1.6. Data Extraction

For the dataset in this thesis, clickbait headlines are extracted from the Twitter accounts of Limon and Spoiler Haber, while non-clickbait headlines are extracted from the Twitter accounts of Evrensel and Diken Newspapers.

Being developed in 2012, the Twitter API (developer) provides an environment for managing the information on Twitter, such as tweets, users, and direct messages, and the interaction of developers with Twitter, such as posting and retweeting. Twitter API is used to form a part of the dataset in this thesis. Due to the limitations and restrictions imposed by Twitter, only a limited amount of data can be reached on Twitter using the Twitter API. These limitations cover the number of tweets that can be retrieved, the number of tweets that can be gathered from a Twitter account, and the number of posts that can be sent.

Since Twitter API allows retrieving only a limited number of tweets (approximately 3,200 tweets for each account), the Twitter accounts whose data are planned to be used are contacted via email and asked to share their Twitter data. Among these accounts, Limon Haber and Evrensel Newspaper contributed to this study by sharing their tweet data included in the Twitter data. On the other hand, since no response was received from Spoiler News and Diken Newspaper during the dataset construction process, the data of these accounts were manually gathered from Twitter. In the data publishing process, after the dataset was created and the analyses were made, Spoiler News and Diken Newspaper responded to the request and declared that they accept the sharing of their data.

Tweepy (tweepy) was used to access the quote tweets in the tweet data sent by Limon Haber. It is a Python library for accessing the Twitter API. It is usually used to make a new research on Twitter data in Python. Many basic operations of Twitter can be done through this library.

Extracted tweets of Limon Haber cover the period between April 30, 2016, and December 20, 2018, while extracting the tweets of Evrensel Newspaper are between November 18, 2013, and December 25, 2018. As stated earlier, Diken Newspaper did not respond to the request for tweet data in the dataset construction phase, and the data of this account were gathered manually to cover tweets from May 2, 2018, to December 24, 2018. Similarly, as there was no response from Spoiler Haber, its data were gathered manually to cover tweets from December 31, 2015, to December 2, 2018 (see Table 4).

The dataset consists of 48,060 samples, and clickbaits (24,031) and non-clickbaits (24,029) are represented equally in the dataset. The sources and quantities of clickbait and non-clickbait tweets with the source information can be seen in Table 2.

Since Limon Haber and Evrensel Newspaper sent their tweet data to be used in this study, the data of these accounts contain more detailed information than the data of Spoiler Haber and Diken Newspaper. The data from these two Twitter accounts contain other essential attributes besides the headlines, the names of the news sources that share these headlines, and the date on which they were posted. Firstly, retweet and tweet information about each headline is included in the data of these two sources. This information indicates whether the headlines have been tweeted or retweeted by the account. Secondly, the data covers whether the post (tweet or retweet of the account) includes any symbols, hashtags (#), or user mentions denoted by at sign (@). If there are any hashtags, symbols, or user mentions in the post, their quantity is specified. The name and ID of the user mentioned in the post are also designated. Thirdly, the URL of the posts is stated in the data. Fourthly, the text range of the post, indicating how many characters are used in that post, is included in the data. Finally, the information about how many times the post was liked and retweeted by the other users is found in the data.

In addition to tweeting the headlines of other news sources in the form of a quote tweet, Limon Haber also retweets the tweets of other users who share the clickbait headlines by tagging Limon Haber. However, Evrensel Newspaper shares its news because it is an official news source. Therefore, although the data of Limon Haber consists of tweets and retweets, Evrensel Newspaper's data only includes tweets.

Among all this information included in the raw data obtained from Limon Haber and Evrensel Newspaper, the following attributes are presented in the ClickbaitTR dataset: The name of the news source, Tweet ID, the date of tweet, the text of tweet, how many likes, and retweet a tweet has.

As it is explained earlier, tweets of Spoiler Haber and Diken Newspaper are gathered manually in the data construction process since these accounts could not be made contact with. Spoiler Haber, like Limon Haber, tweets the clickbait headlines of other news sources as quote tweets. The data gathered from Spoiler Haber and Diken Newspaper include tweets; there is no retweet in this data. These manually acquired data contain information about how many replies, retweets, and likes a tweet has, as well as the name of the source and the date of the headline was tweeted. The tweets of these accounts do not contain Tweet IDs as this data is taken manually.

Among all this information included in the raw data obtained from Spoiler Haber and Diken Newspaper, the following attributes are presented in the ClickbaitTR dataset: The name of the news source, the date of tweet, the text of tweet, how many likes, and retweet a tweet has. The number of likes and retweets that Diken Newspaper's headlines get are partially included in the dataset since there is missing information in the data taken from Diken Newspaper. Among all these attributes, those included in the dataset are shown in Table 3. The organization of the dataset can be found in Appendix D.1. Besides, the information about the data acquisition process can be seen in Table 4.

Table 2: The sources and quantities of clickbait and non-clickbait tweets.

Source	**Clickbait**	**Non-clickbait**	**Total**
Limon Haber	22,133	0	22,133
Spoiler Haber	1898	0	1898
Evrensel Newspaper	0	13,093	13,093
Diken Newspaper	0	10,936	10,936
Total	24,031	24,029	48,060

Table 3: The features presented in the dataset. The checkmark (✓) indicates that the feature is included in the dataset for the corresponding news source, while the cross mark (X) indicates the opposite. The dash (–) indicates that the feature is partially included in the dataset's corresponding news source.

Source	**Tweet ID**	**Date**	**Tweet**	**Number of Likes**	**Number of Retweets**
Limon Haber	✓	✓	✓	✓	✓
Spoiler Haber	X	✓	✓	✓	✓
Evrensel Newspaper	✓	✓	✓	✓	✓
Diken Newspaper	X	✓	✓	–	–

Table 4: The overall summary of the data acquisition procedure.

Properties	
Platform	Twitter
Overall period of acquired data	18 November 2013 – 30 July 2019
Period of data from Limon Haber	30 April 2016 – 20 December 2018
Period of data from Spoiler Haber	31 December 2015 – 2 December 2018
Period of data from Evrensel Newspaper	18 November 2013 – 25 December 2018
Period of data from Diken Newspaper	2 May 2018 – 24 December 2018
Number of extracted tweets	315,135
News sources	Limon Haber, Spoiler Haber, Evrensel Newspaper, Diken Newspaper
Content of tweets	Text
Sampling strategies	Taking tweet data of publishers from them by request and getting data from the Twitter account of publisher manually
Number of samples	48,062

3.2. Dataset Construction

3.2.1. Scanning the Collected Data

Data visualization may provide crucial information about the data immediately and reveal a general picture. Especially in text-based data, it can enable to see the main points of the data that cannot be discovered by other types of analyzes. As a data visualization method, word clouds are often used to get general information about text data. A word cloud is a cluster of words represented in different sizes. The size of the words is directly proportional to their frequency in the data. The larger a word appears in the word cloud, the more often it appears in the text data. This information is helpful in getting an idea of the importance of that word in the data. In addition, some comparisons between words in the data can be made using this information.

After the data extraction process, word clouds were created to obtain information about the dataset's content, examine the frequency of words in the dataset, and determine how appropriate the clickbait and non-clickbait data were to their respective categories. For this purpose, two separate analyzes were performed for these two different categories of the dataset, and the words of each data category were visualized to see the frequencies of the words in these categories.

When two word clouds obtained as a result of the two analyzes are examined, it is seen that the clickbait data contains non-specific words such as question words or pronouns (i.e., what or she). These words reflect the basic clickbait mechanism, as the clickbait strategy aims to arouse curiosity by not providing the necessary details, giving incomplete information, or asking questions.

The most prominent words in the clickbait data are explain (açıkla), she/he (o), and flash (flaş), attention (dikkat), stated (belli), first (ilk), what (ne), and judgment (karar) which are consistent with the main words being used in Turkish news headlines. Immediately after these words, in terms of size, the following words have the same frequency (size) in the word cloud: see (gör), celebrity (ünlü), here (işte), why (neden), about (ilgili), shock (şok), and more (daha) (see Figure 1).

Figure 1: Word cloud showing the most frequent words in clickbait data such as explain (açıkla), flash (flaş), what (ne), stated (belli), she or he (o) and attention (dikkat). (Adapted from (Genç and Surer, 2021)).

The information obtained from word clouds about Turkish clikbait news headlines reflects the quality of these headlines very well. For example, the following sentence is an example of a powerful clickbait news headline that uses a combination of word roots flash (flaş), celebrity (ünlü), and explain (açıkla).

a) 5.8'lik İstanbul depremiyle ilgili ünlü uzmandan #FLAŞ açıklama! / #FLASH statement from the famous expert about 5.8 Istanbul earthquake!

In another clickbait title below, it is seen that him as a pronoun is used and no information was given about the identity of the person:

b) Bu köyde onun dışındaki herkes elektrik parası ödüyor... / In this village, everyone except him pays for electricity...

The clickbait examples below include the other essential words for being clickbait such as attention (dikkat), judgement (karar), here (işte), first (ilk), stated (belli), why (neden), see (bakın), what (ne), more (daha), about (ilgili):

c) Araba alacaklar dikkat: Karar Resmi Gazete'de... / Car buyers attention: The judgement is in the Official Newspaper...

d) Yüzüklerin Efendisi dizi oluyor; işte yayınlanacağı kanal / Lord of the Rings is becoming a series; here is the channel where it will be broadcast

e) Atv'nin iddialı dizisinin final tarihi belli oldu! / The final date of Atv's ambitious series has been announced!

f) Şüpheli paketten bakın ne çıktı / See what the suspicious package turned out

g) Dünyaca ünlü kahinden tüyler ürperten bir kehanet daha! / Another gruesome prophecy from the world's famous oracle!

h) Uzmanlardan çok kritik uyarı! Sakın ama sakın önümüzdeki 2-3 ay tavuk döner yemeyin! Peki neden?… / Highly critical warning from experts! Do not eat chicken doner for the next 2-3 months! So why?…

i) Dağın zirvesine yapılan görenleri şaşkına çevirdi! / It stunned those who saw the top of the mountain!

j) Dünyayı tehdit eden sızıntıyla ilgili açıklama geldi! / The statement came about the leak threatening the world!

On the other hand, the word cloud of non-clickbait data shows that this data category contains informative words such as woman or public. These words are consistent with the fact that non-clickbait sentences contain the necessary details and are not intended to arouse curiosity in readers. The outstanding words in the non-clickbait data are police (polis), newspaper (gazete), woman (kadın), judgment (karar), child (çocuk), and ABD. Immediately after these words, the following words are having the same frequency: made (yapıl), explain (açıkla), and give (ver). The following words that appear slightly smaller than these words have the same size in the word cloud: government (hükümet), case (dava), and public (halk) (see Figure 2).

Figure 2: Word cloud showing the most frequent words in non-clickbait data such as police (polis), newspaper (gazete), woman (kadın), judgement (karar) and child (çocuk). (Adapted from (Genç and Surer, 2021)).

The non-clickbait headlines below contain keywords detected by word clouds such as police, newspaper, woman, made, explain, judgment, and child. Interestingly, two of these words (judgment and explain) are also used frequently in clickbait titles. This may be because explanations about innovations, details, or a judgment on a topic reflect the primary purpose of journalism.

a) ABD polisinden 7 yaşındaki çocuğa ters kelepçe / Back handcuffs on 7-year-old boy from US police.

b) İtalya'da bir grup yaşlı kadın, internette para toplayıp hayatlarında ilk kez denizi gördü / A group of old women in Italy collected money on the internet and saw the sea for the first time in their lives.

c) Bayram tatili verisi: Kredi kartıyla 6.1 milyar lira ödeme yapıldı / Holiday data: 6.1 billion lira was paid by credit card

d) THY Atatürk Havalimanı'ndan son uçuş saatini açıkladı: 6 Nisan 02.00 / THY announced the last flight time from Atatürk Airport: 6 April 02.00

To gain more detailed information about the content of the data, the main topics of the dataset were extracted and analyzed using the LDA (Latent Dirichlet Allocation) method, a generative probabilistic model for topic modeling. Topic modeling is usually used for discovering the main themes in a dataset to understand and organize the data. In this study, topic modeling is done to see how accurate it is to group the dataset as clickbait data and non-clickbait data according to the Twitter accounts from which it was collected. LDA (Blei et al., 2003) is a technique that finds specific themes from data by grouping various word distributions. The basic idea is that several words represent the latent topics in the data and these topics can be extracted by grouping these words based on the bag-of-words paradigm and word-document counts (Rehurek & Sojka, 2010).

The LDA is applied to the dataset three times for the whole dataset, for the clickbait data, and for the non-clickbait data separately (passes = 5, alpha = 0.01, eta=0.01) to discover the general topics of the dataset and to see whether the dataset is correctly labeled as clickbait and non-clickbait by learning the content of clickbait and non-clickbait data.

Firstly, LDA is used on the entire dataset, and the results show that its general topics are the following: 1) Politics, 2) Woman, child and law, 3) Crime and security, 4) Economy, 5) Education. Out of these five topics, policy reflects the most common theme in the dataset, which can be seen from the most relevant terms (words) of the first topic (Figure 3) such as president (başkan), candidate (aday), president of the republic (cumhurbaşkanı), parliament (meclis), government (hükümet), municipality (belediye). As shown in Figure 3, these words have the highest estimated term frequency within the selected topic (Topic 1), which means they represent the first topic best. Other topics that best represent the dataset can be seen in Appendix A.

Secondly, LDA is used on the clickbait data, and the general topics represented with the estimated word frequencies are the following: 1) Sports and football, 2) Magazine, 3) Politics, and 4) Economy. It can be seen from Figure 4 that the sports and football topic is represented by the words like match (maç), football (futbol), and league (lig). On the other hand, this topic also includes the clickbait-related words detected in word clouds as such as flash (flaş), here it is (işte), latest (son), stated (belli), explain (açıkla), and judgment (karar). Topics in the clickbait data are not clearly distinguishable because most of the words representing topics are non-specific words related to clickbait (see Figure 4). Other topics that are extracted from the clickbait data can be seen in Appendix B.

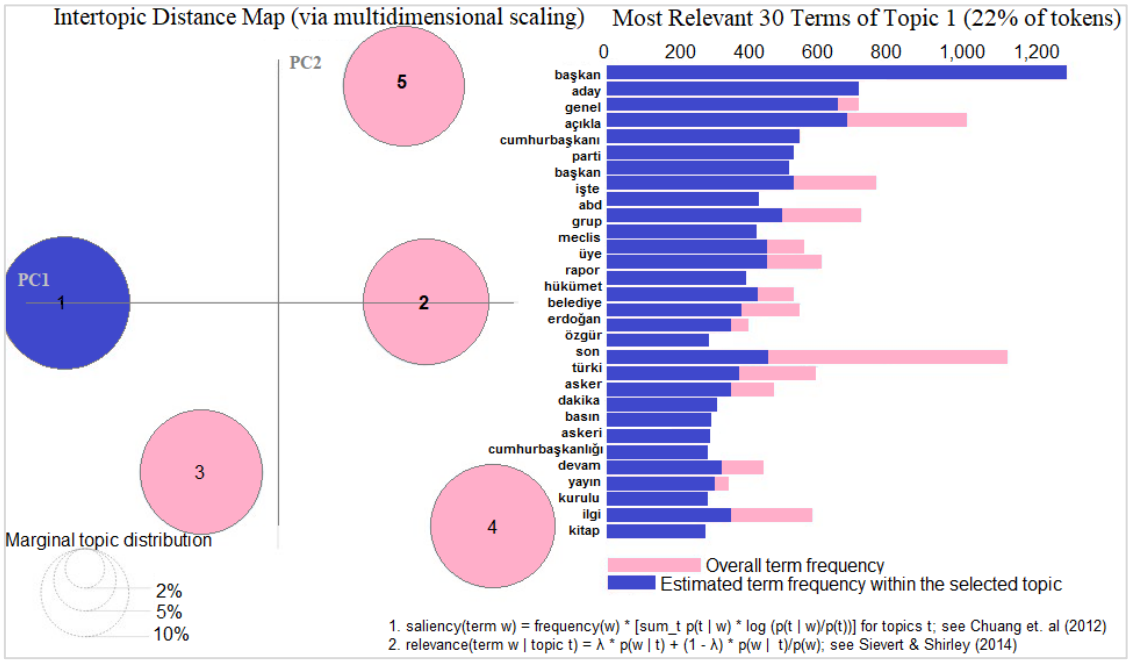


Figure 3: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from the dataset. (Adapted from (Genç and Surer, 2021)).

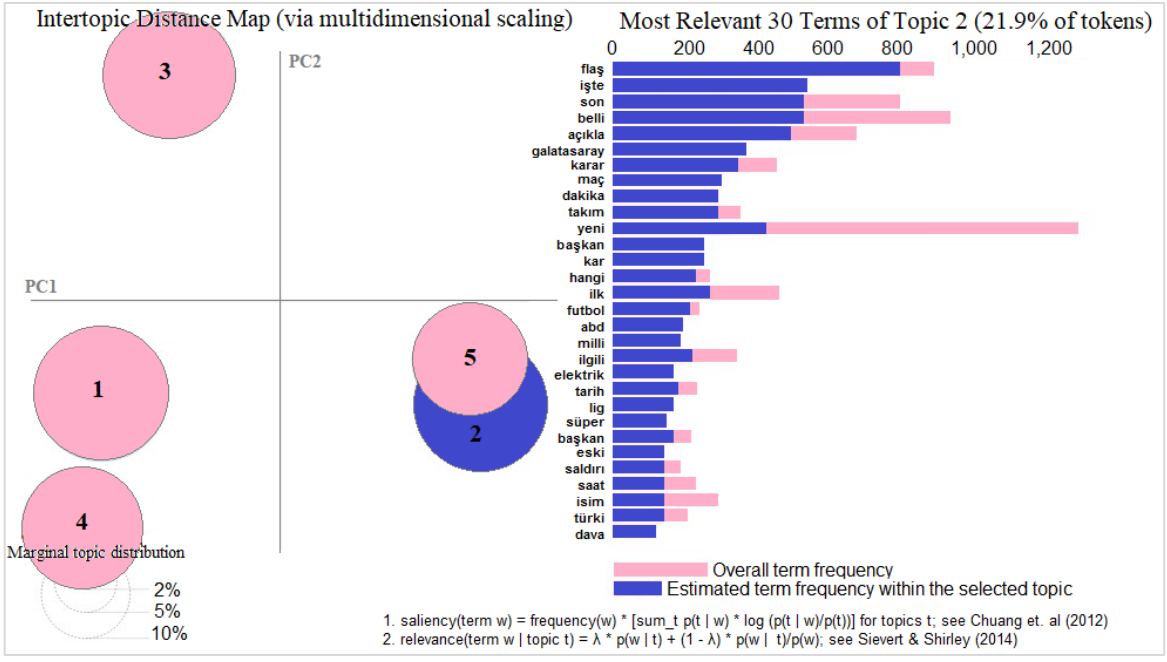


Figure 4: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 2 (magazine) from the clickbait data. (Adapted from (Genç and Surer, 2021)).

Finally, according to the LDA analysis, the following topics are the best representative of the non-clickbait data: 1) Politics, 2) Education and economy, 3) Law, 4) Security and politics, and 5) Law and politics. The frequent words of the first topic are president (başkan), parliament (meclis), candidate (aday), president of the republic (cumhurbaşkanı), law (yasa), and public (halk) (see Figure 5). Other topics that best represent the non-clickbait data can be found in Appendix C.

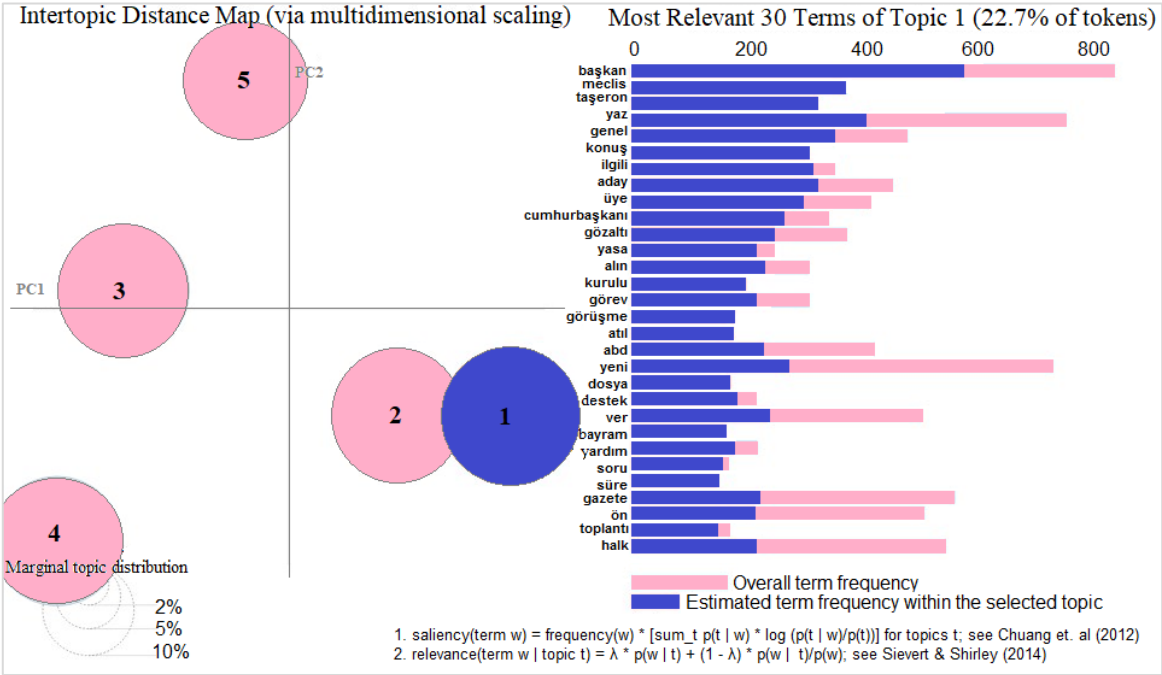


Figure 5: Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from the non-clickbait data. (Adapted from (Genç and Surer, 2021)).

3.2.2. Data Pre-processing

A stem (base) morpheme is a constituent that cannot be dived into other morphemes and is the uninflected part of a word to which affixes are attached to form new words. Affixes are the morphemes that are attached to the beginning (prefixes) or end (suffixes) of words. Affixes can be categorized as inflectional and derivational affixes based on several factors.

Firstly, inflectional affixes do not make any changes in the parts of the speech (category) of the stem morpheme. For example, if a plural marker is attached to a noun such as kitap (book in English), the new word form kitaplar (books in English) will still be a noun. On the other hand, derivational affixes often change the parts of the speech of the stem morpheme. For example, yaz is a verb in Turkish (write in English) and it becomes a noun yazar (writer in English) by taking the derivational suffix -ar. Not all derivational suffixes change the category of the word like it is in the kitap and kitaplık (book and bookshelf in English). Secondly, derivational suffixes are always attached to a base morpheme before inflectional suffixes. Namely, derivational suffixes form new bases to which other derivational and inflectional suffixes can be attached. For instance, the plural form kitaplık (bookshelf) is the kitaplıklar (bookshelves). Finally, inflectional affixes do not make any radical changes in the meaning of the stem, and there is a similarity between the meanings of stem and the new form (stem + inflectional affix). On the other hand, derivational affixes may produce a new form (stem + derivational affix) whose meaning is unpredictable. Inflectional affixes produce word sets which are called paradigms such as know - knows - knowing (bil - bilir - biliyor in Turkish) (Akmajian, 2017). As it can be seen from this example, all of these three words have parallel meanings.

Since the words in the paradigms have similar meanings, it is convenient to represent them as a single stem for the clickbait analysis. Using the words with different inflectional suffixes in any algorithm would decrease the accuracy of the training process and make it impossible to find the frequency of unique words in the news headlines. Dividing the words as stems and suffixes gives us a chance to reduce the dimensionality of the data without removing information because it provides a compressed version of the word paradigms. For example, in Turkish clickbait headlines, happened (oldu), will happen (olacak), and happening (oluyor) are used frequently and they have the same root, happen (ol). It was thought that identifying those words used in different forms in clickbait headlines would increase computational efficiency in the analyses. It is important to note that all inflectional affixes in Turkish are suffixes. The inflectional suffixes of the words in the dataset are removed for the clickbait detection process, but the derivational suffixes are left as they are.

Kalbur Project (Aksoy, 2017) has been developed to separate Turkish words into their roots and suffixes. The project's main focus is on roots rather than suffixes because word roots are needed so much in artificial learning studies and big data

processing (Aksoy). Kelime-bol library of the Kalbur Project compares the words in the dataset from their first character to the last character with the root list and collects the possible roots in a list. Then the remaining characters from the possible root of each word are searched in the list of suffixes. If these character groups are included in the list of suffixes and are suitable for the possible root, they are parsed into the word stem and suffixes. This library has done quite well in separating words in the ClickbaitTR dataset from their suffixes, which can be seen from the word clouds and results of the LDA analysis.

The dataset consists of the following information: the name of the Twitter accounts the headlines were gathered (source), identity number (id) of tweets, the release date of tweets, the text of news headline (tweet), how many times a tweet was liked by other (number of likes) users and how many times a tweet was shared by other users (number of retweets). There is no tweet id information in the data of Spoiler Haber and Diken Newspaper since the tweets of these two accounts were taken manually. For the same reason, some tweets of these two accounts do not cover the number of retweets and likes. At the end of the data pre-processing, tweets are represented as "source, tweet-id, created-at, full-text, favorite-count, retweet-count," which can be seen in Table 5. ClickbaitTR dataset is available at https://github.com/clickbaittr/turkish-clickbait-dataset (Clickbaittr, 2021).

Table 5: The favorite and retweet frequencies of three tweets posted by Limon Haber.

Source	**Tweet id**	**Created at**	**Full text**	**Favorite count**	**Retweet count**
Limon	867794265937760256	5/25/2017 17:27	Interpol'ün aradığı İngiliz kadın, Marmaris'te bulundu	6	0
Limon	867679183731916800	5/25/2017 09:50	Bakan Elvan açıkladı! Bir hafta içinde yasalaşıyor	6	2
Limon	867801404701671424	5/25/2017 17:55	Bir otomobil devine daha dava şoku	7	1

3.2.3. Dataset Validation

In order to evaluate the accuracy of labeling the headlines in the dataset as clickbait and non-clickbait, three independent annotators were asked to rate the clickbaitness of the news headlines in the dataset. Evaluators scored 10% of the test set on a scale of 1-10 (1: non-clickbait, 10: clickbait). This information coming from human is essential because is was used for validating the labels of dataset and validating the results of the machinel learning models.

The reliability of scores from raters was calculated using Cohen's Kappa (Cohen, 1960). Cohen's Kappa is a quantitative measure of inter-rater reliability which shows the agreement between raters on a classification task (Di Eugenio & Glass, 2004). The raters' scores for 900 of the news headlines in the test set and the average of the scores from these three raters were compared with Cohen's Kappa. Besides, the scores coming from raters were compared with the labels of headlines in the dataset for validating the suitability of labeling process. The results of the inter-rater reliability testing can be seen in Figure 6. The Kappa value for the dataset and human rater mean (κ = .75) show that there is a substantial agreement between the labels of the dataset and the decisions of the three raters.

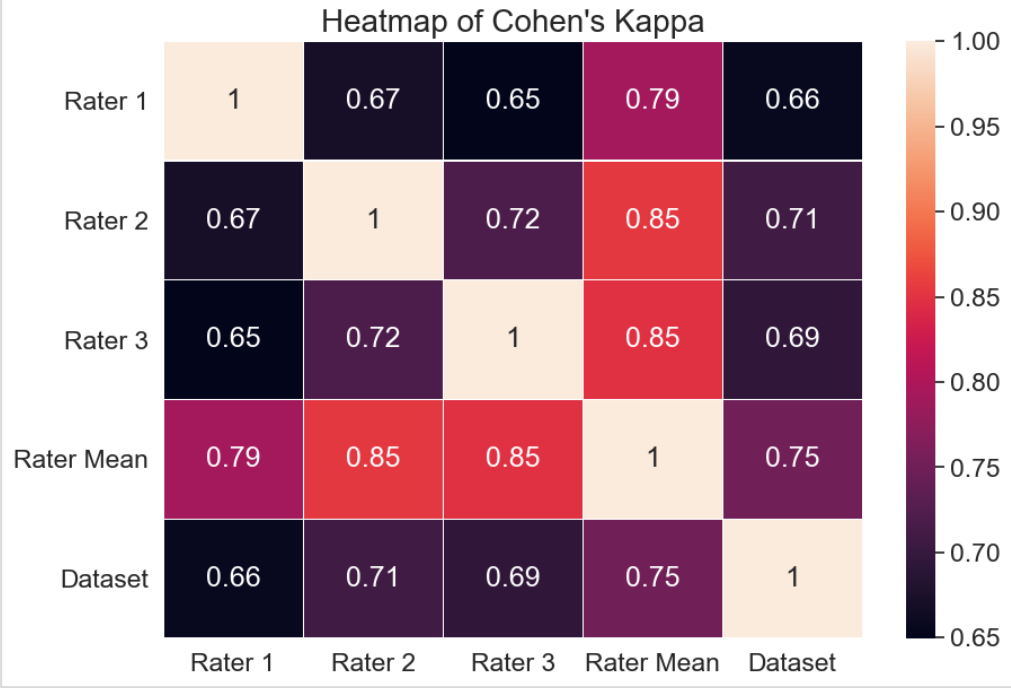


Figure 6: Comparison between dataset and human scores.

3.3. Data Analysis

3.3.1. Feature Selection

Features to be used in addition to words in machine learning algorithms are selected by investigating the literature. Chakraborty et al. (2016) indicated that odd punctuation patterns are typically used in clickbait headlines, which seems parallel with the structure of Turkish news headlines in the dataset. Based on this study, special characters were chosen as a part of features that would be added to the analyses.

Potthast et al. (2016)'s study is one of the studies in which the features of clickbait and non-clickbait headlines were found and sorted by importance. In their study, as the dataset consists of tweets obtained from Twitter, it is thought that some of the features they detected can be used in the analysis in this thesis. Based on the detected features in their study, the number of hashtags (#), question marks (?), exclamation marks (!), dots (.), at signs (@), and the other special characters were determined as features to be included in the feature vector of this study. In their study, as the dataset consists of tweets obtained from Twitter, it is thought that some of the features they detected can be used in the analysis in this thesis. A separate feature called Special Characters was also created to include all kinds of special characters except these characters.

Hardalov et al. (2016) studied fake news, and certain features were detected to distinguish credible news from fake news. The number of upper cases, the length of words, and the length of tweets (i.e., length of the headlines) were selected as other features to be used in this study based on Hardalov et al. (2016)'s study. It should be noted that the length of words and the length of tweets were also mentioned as essential features in Potthast et al. (2016)'s study.

At the end of the feature selection process, these nine features (number of hashtags, number of question marks, number of exclamation marks, number of dots, number of at signs, number of other special characters, number of upper cases, length of words, and length of tweets) were added to the feature vector which represents the frequencies of words in the news headlines. The feature vector will be explained in detail in the next section.

3.3.2. Machine Learning Models

In order to understand whether the selected nine features have a positive effect on analyses, two models were trained on a smaller version of the dataset in the preliminary study (Genc & Surer, 2019). In this preliminary, a Multi-layer Perceptron classifier (MLPClassifier) from Scikit-learn was trained for both models (Pedregosa et al., 2011) on the dataset consisting of 39,201 samples. In order to decide the best parameters for the models, GridSearchCV's suggestions were used. GridSearchCV is a hyperparameter tuning tool of the Scikit-learn library, which searches through predefined hyperparameters and helps to find the best parameters for the model.

For the first MLP model, 80% of the data were separated for training, and 20% were separated for the test set. It was trained on a feature vector with 10,338 dimensions, including all word roots and nine selected features. On the other hand, the second MLP model was trained on nine features, and 80% of the data were separated for training, and 20% were separated for the test set in this second model. Namely, the first model was trained on the feature vector where news headlines were represented by words and nine features, while the second model was trained on the feature vector where only nine features represented news headlines. The results showed that the first model performs with an accuracy of 0.91 with an F1-score of 0.91, while the second model performs with an accuracy of 0.83 with an F1-score of 0.83 (Genc & Surer, 2019).

This preliminary study was one of the first efforts to clickbait detection in Turkish news headlines. On the one hand, those results showed that the selected nine features were essential in distinguishing Turkish clickbait news headlines from non-clickbait headlines, and clickbait detection was made with 83% accuracy even by using only these features. On the other hand, when nine features were used together with word roots obtained from news headlines, it gave the best result in Turkish clickbait detection with 91% accuracy. This inference constitutes a basis for using these nine features in this thesis study as well.

In this thesis, a dataset consisting of 48,060 Turkish news headlines was constructed to analyze for clickbait detection. The sources and quantities of clickbait and non-clickbait data used in the analyses can be found in Table 2. After the data pre-processing step, a feature vector with 10,890 dimensions was created for machine learning algorithms. This vector consists of frequencies of different words in each news headline and selected nine features in the dataset uniquely represented on it. Namely, each news headline (tweet) in the dataset was represented uniquely in this 10,890-dimensional feature vector, including the nine selected features.

In order to detect clickbait news headlines in the data and compare models, six different machine learning algorithms were trained on the feature vector. These algorithms are Artificial Neural Network, Logistic Regression (Yu et al., 2011), Random Forest (Breiman, 2001), Long Short-Term Memory Network (Hochreiter & Schmidhuber, 1997), Bidirectional Long Short-Term Memory Network (Schuster & Paliwal, 1997) and Ensemble Classifier (Sisodia, 2019).

Artificial neural network, a nonlinear algorithm, is good at learning the representations of the data, predicting the output, and doing classifications. In this study, a multilayer perceptron (MLP), which is an artificial neural network with three or more layers, was implemented because it is widely used in natural language processing (NLP) studies (Schwenk & Gauvain, 2005).

Logistic regression is a linear method for binary classification and is applicable in different areas such as document classification and natural language processing (NLP). It was considered appropriate to include this algorithm in the analysis of this thesis since

it was also used in clickbait studies (Potthast et al., 2016). On the other hand, the random forest was used in this well-known clickbait detection study and performed the best out of the three algorithms used. It is an ensemble of several decision trees for various tasks such as classification and regression. The random forest algorithm is also included in the analyses of this study because it is good at classifying clickbait and non-clickbait titles (Chakraborty et al., 2016; Potthast et al., 2016).

On the other hand, for a dataset containing sequential data, neural networks that can learn from sequential data such as LSTM are used. For complex tasks with sequence prediction problems such as speech recognition and machine translation, LSTM is preferred and applied successfully. Because there are news headlines of different lengths that are sequential In the ClickbaitTR, it was decided to train in LSTM and MLP.

BiLSTM is another type of neural networks that can learn from sequential data using LSTM in two directions, backward and forward. BiLSTM can exploit information from both the past and future due to its bidirectional mechanism. BiLSTM can be very useful on a dataset containing sentences like ClickbaitTR because information can be lost as LSTM runs from the beginning to the end of the sentences. However, BiLSTM can overcome this problem by running backward as well as running forward. Anand et al. (2017) applied BiLSTM on a large dataset consisting of clickbait and non-clickbait tweets, and the BiLSTM performed with high accuracy (98%). This result shows training BiLSTM on the dataset would show a good performance.

Ensemble learning is another method in machine learning that might give a better performance making predictions based on decisions of multiple algorithms. Ensemble learning techniques applied for clickbait detection previously and gave an accuracy of 91.16% (Sisodia, 2019). Due to its high performance, it was decided to use ensemble learning algorithms on ClickbaitTR, as well.

The libraries included in Scikit-learn (Pedregosa et al., 2011) were used for using Logistic Regression and Random Forest algorithms, while Keras (Chollet, 2015) was used for Artificial Neural Network (ANN), Long Short-Term Memory Network (LSTM), Bidirectional Long Short-Term Memory Network (BiLSTM) and Ensemble Classifier algorithm. A 10-fold cross-validation procedure was used in the implementation of all of the algorithms. On the other hand, the neural network determined the important features by calculating the importance of each feature. It was expected that, as the importance of a given feature increases, the absolute strength of connections between the input neuron that corresponds to that feature and the first hidden layer also increases in correlation. The absolute value of connections of each input neuron to neurons in the first hidden layer was summed to quantify the importance of each feature.

3.3.3. Confidence Analysis

In addition to developing machine learning models that can successfully distinguish clickbait sentences from non-clickbait sentences, it is also critical to understand how these models decide and their confidence in their decisions. The explainability of models is essential for understanding the mechanism of the clickbait strategy and its linguistic structure. Examining the results of the models by discretization does not provide an accurate assessment of the performance of the models. In this way, it is not possible to observe which sentences the models call clickbait or non-clickbait with what degree of certainty. In this thesis, along with training machine learning algorithms for clickbait detection, several analyses were also applied based on confidence values and parameters of machine learning models to reveal how those complex models work and the linguistic structure of clickbait. The same analyses were applied to the data of human raters, and the models were compared with human data.

Confidence scores show the distance between the threshold of the activation function and the model's estimation probability, which reflects the probability of a sample to fall in a certain class. The probability of a sample being in a category far from the threshold means high confidence, while the probability of this probability being close to the threshold means low certainty. After all the models were trained, a confidence analysis was performed to understand how confident the models were in their decision to classify news headlines. For this, the prediction probabilities of each model were obtained; these probabilities were examined and compared with each other. The confidence values of Logistic Regression and Random Forest were obtained with the Scikit-learn library, while the confidence values of ANN, LSTM, and BiLSTM were obtained using the Keras library.

In order to determine the features that models give importance in distinguishing the news headlines, the features that the models decided on with high confidence were detected for the final process of the confidence analysis. These features were extracted using the first 900 samples of the test set to keep the number of headlines scored by the raters equal to the number of headlines used by the algorithms. The confidence scores of the all models were scaled between -1 and 1 in order to represent the non-clickbait decisions with negative values and the clickbait decisions with positive values. Since the model confidence for clickbait headlines are represented by positive values and the model confidence for non-clickbait sentences are represented by negative values, the important features were detected separately for clickbait and non-clickbait categories. Thus, the important features that affect the output confidence of each model while deciding for these two categories were examined separately for the clickbait and non-clickbait data.

It was thought that the frequency of words and the frequency of special characters used in sentences were not equivalent, and therefore, a word and a special character should not have the same weight. For example, the features such as the use of multiple special characters (i.e., !!! or …) or the average word length, especially used in clickbait

sentences, have a lower weight than multiple usages of a word. For this reason, the weights of the words and the weights of special characters are balanced by applying the log2 function to the frequencies of the special characters. Thus, features such as average word length or average tweet length are prevented from becoming essential among important features.

CHAPTER 4

RESULTS

4.1. Artificial Neural Network

A Multi-Layer Perceptron (MLP) as an Artificial Neural Network algorithm was trained the 10,890-dimensional feature vector. Firstly, the same parameters as in the preliminary study (mentioned in the Data analysis) were used. This first trained model performed with a mean accuracy of 0.78 with an accuracy of 0.78 across 10-folds. The performance of the first MLP model on a 10,890-dimensional feature vector can be found in Figure 7.

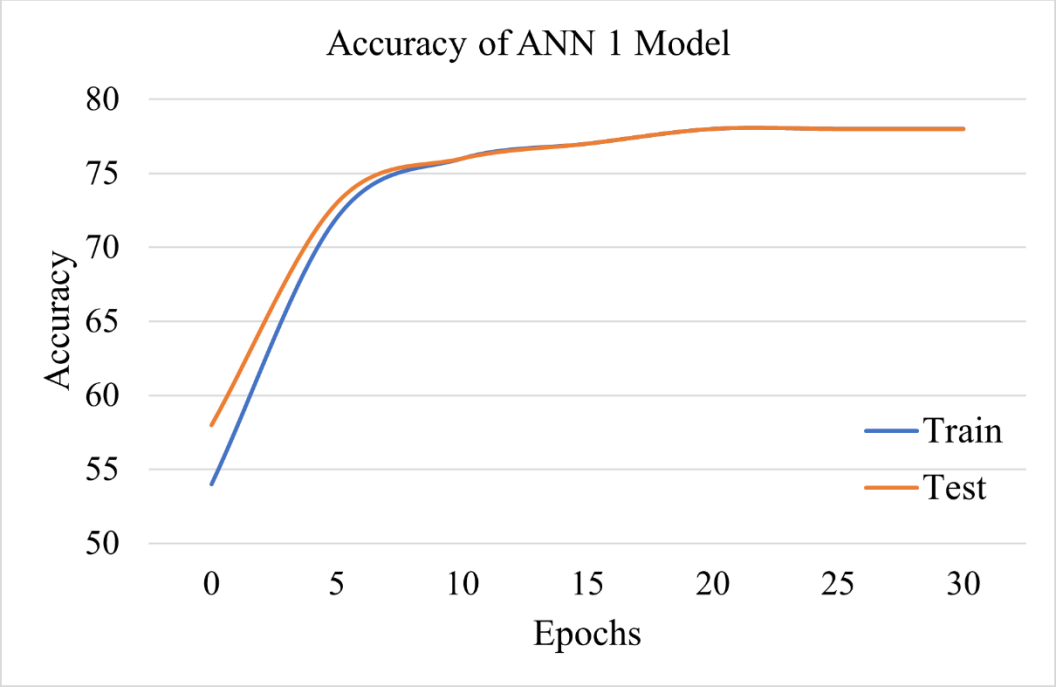


Figure 7: Model accuracy of the first MLP model. The hyperparameters of this model are presented in Table 6. (Adapted from (Genc and Surer, 2021)).

In the preliminary study, 91% accuracy was obtained with the MLP model, which was trained on a 10329-dimensional feature vector with the same hyperparameters. On the other hand, in this study, the larger dataset and, therefore, the larger feature vector size changed the results of the trained MLP model. Considering that the parameters of this model are not suitable for the feature vector of this dimension, a new MLP model has been trained by changing the parameters. While the previous model consists of 1 hidden layer, this new model has two hidden layers. Although the optimization method and activation function are left the same (SGD and tanh, respectively), the learning rate has been increased from 0.0001 to 0.01. While the batch is left the same (120), the number of epochs has been increased from 30 to 50—this new MLP model with a mean accuracy of 0.93 across 10-folds. The hyperparameters used in the two MLP models can be seen in Table 6. The performances of the second MLP model on the 10,890-dimensional feature vector can be found in Figure 8.

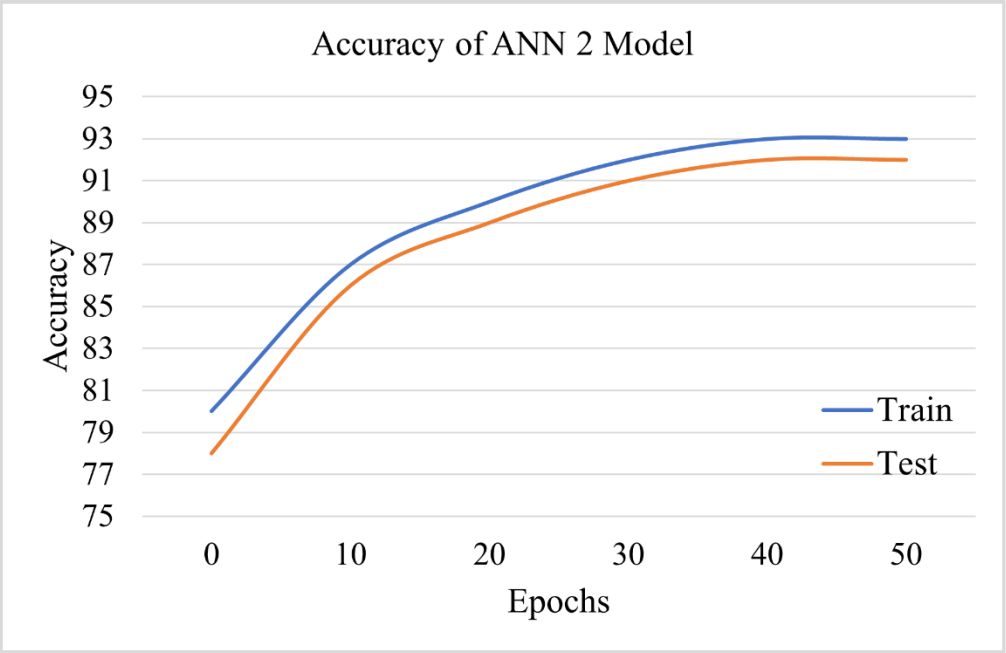


Figure 8: Model accuracy of the second MLP model. The hyperparameters of this model are presented in Table 6. (Adapted from (Genç and Surer, 2021)).

The feature importance derived from the second MLP model can be found in Figure 9. The other special characters, the length of a tweet, flash (flaş), the number of dots, the number of upper cases, the length of words, look (bakın), the number of at signs, development (gelişme), the number of hashtags, interesting (ilginç), horror (dehşet), number of exclamation marks, shock (şok), famous (ünlü), last (son) were among the 20 most essential features in distinguishing clickbait news headlines.

Table 6: Hyperparameters of MLP 1 and MLP 2.

Hyperparameters	**MLP1**	**MLP2**
Number of epochs	30	50
Batch size	120	120
Hidden layer size	10	10, 5
Alpha	0.0001	0.01
Optimization	SGD	SGD
Activation function	tanh	tanh

4.2. Logistic Regression

The Logistic Regression was another algorithm used on the 10,890-dimension feature vector in this study. This model performs with a mean accuracy of 0.88 across 10-folds. The receiver operating characteristic (ROC) curve of the model is very good at distinguishing between the true positives and negatives, which can be seen in Figure 10, showing that the curve is far from the diagonal line.

4.3. Random Forest

The Random Forest model was trained using the same feature vector with 10,890-dimension. It performs with a mean accuracy of 0.90 across 10-folds. According to feature importance of this model (Figure 11), the number of upper cases, the number of dots, the length of words, the number of at signs, explain (açıkla), the number of exclamation marks, stated (belli), she or he (o) and flash (flaş), famous (ünlü), look (bakın), the number of question marks (?), last (son), the other special characters, development (gelişme), judgment (karar) are distinctive for clickbait detection.

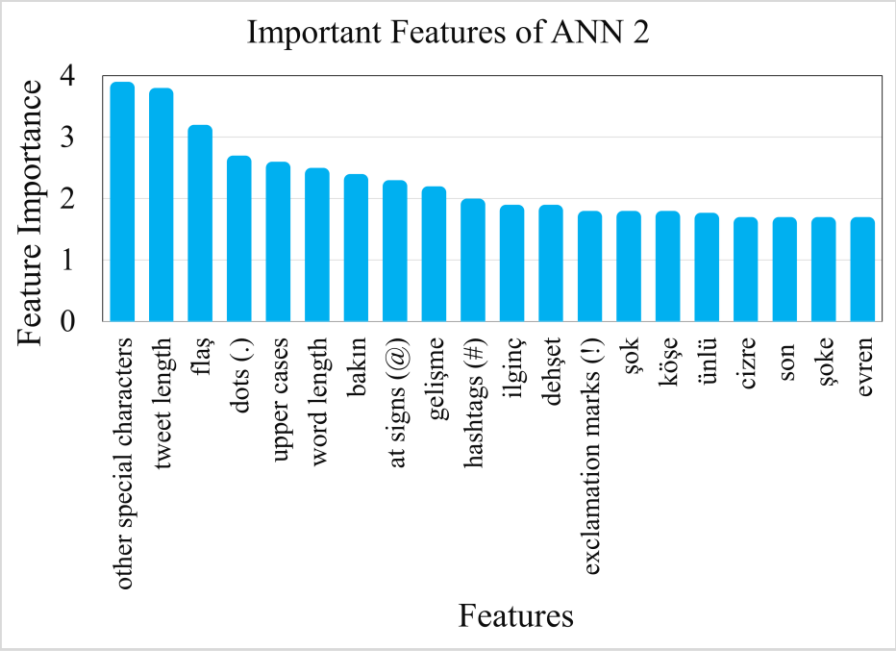


Figure 9: Feature importances obtained from the second MLP 2 model. The hyperparameters of this model are presented in Table 6. (Adapted from (Genç and Surer, 2021)).

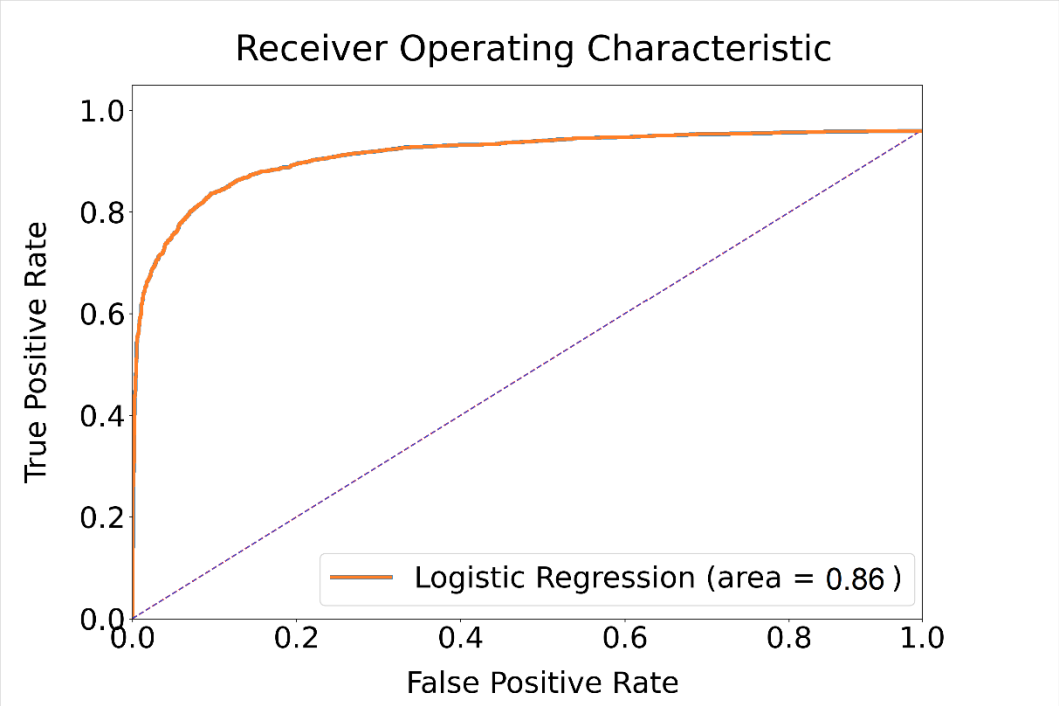


Figure 10: Receiver operating characteristic (ROC) curve of the Logistic Regression model. (Adapted from (Genç and Surer, 2021)).

4.4. Long Short-Term Memory

The LSTM was developed using a Deep Learning library, Keras. This model was trained with the 10,890-dimension feature vector as in the previous algorithms, and it performs with a mean accuracy of 0.93 across 10-folds. The performance of the LSTM can be found in Figure 12.

4.5. Bidirectional Long Short-Term Memory

The BiLSTM was trained as another analysis on the same feature vector using Keras. This model performs with a mean accuracy of 0.97 across 10-folds, as shown in Figure 13.

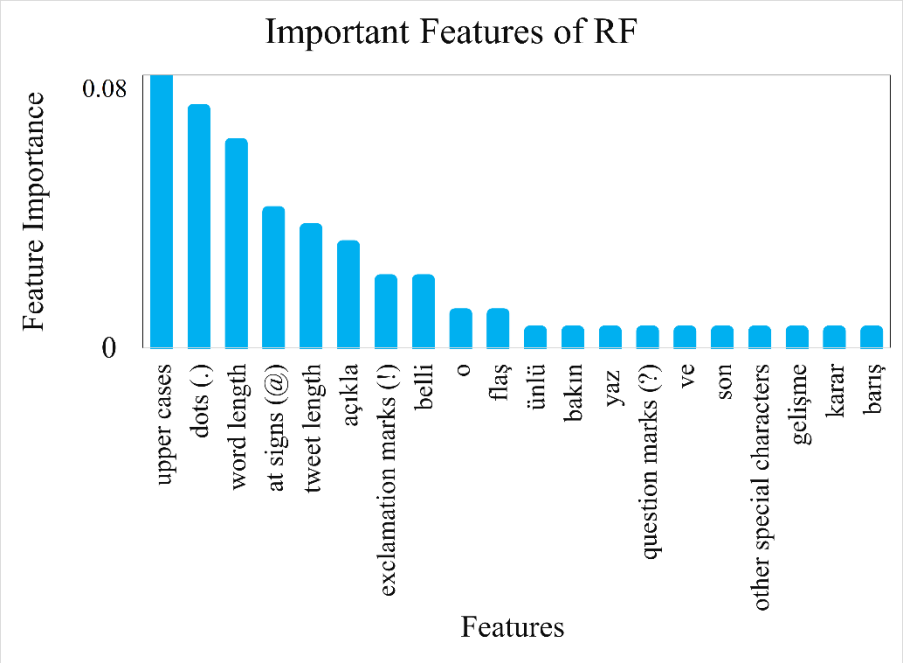


Figure 11: Feature importances obtained from the Random Forest model. (Adapted from (Genç and Surer, 2021)).

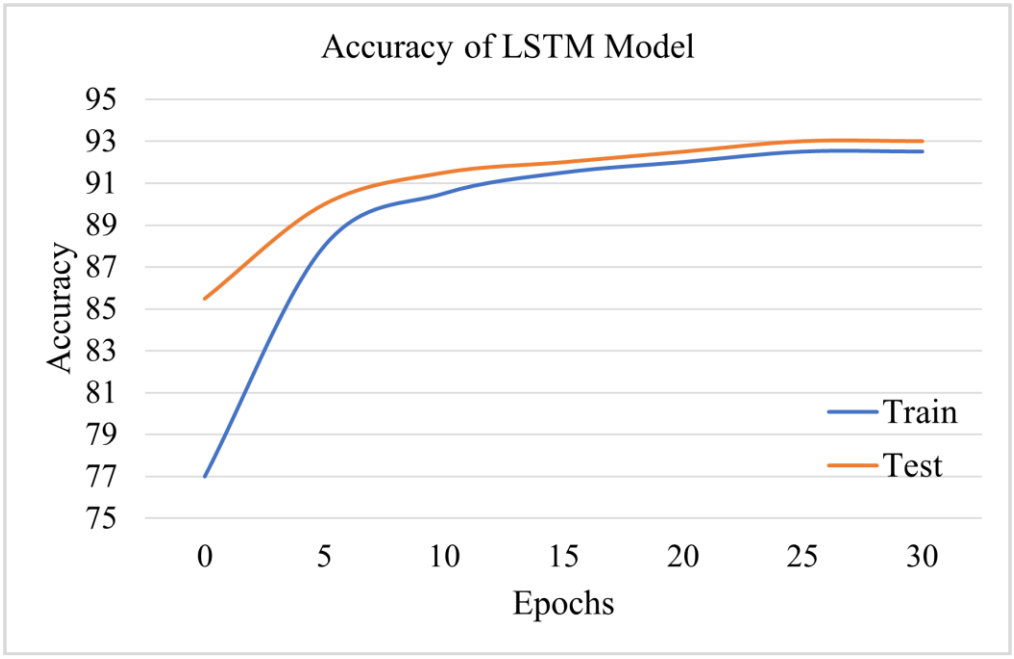


Figure 12: Model accuracy of the LSTM model. (Adapted from (Genç and Surer, 2021)).

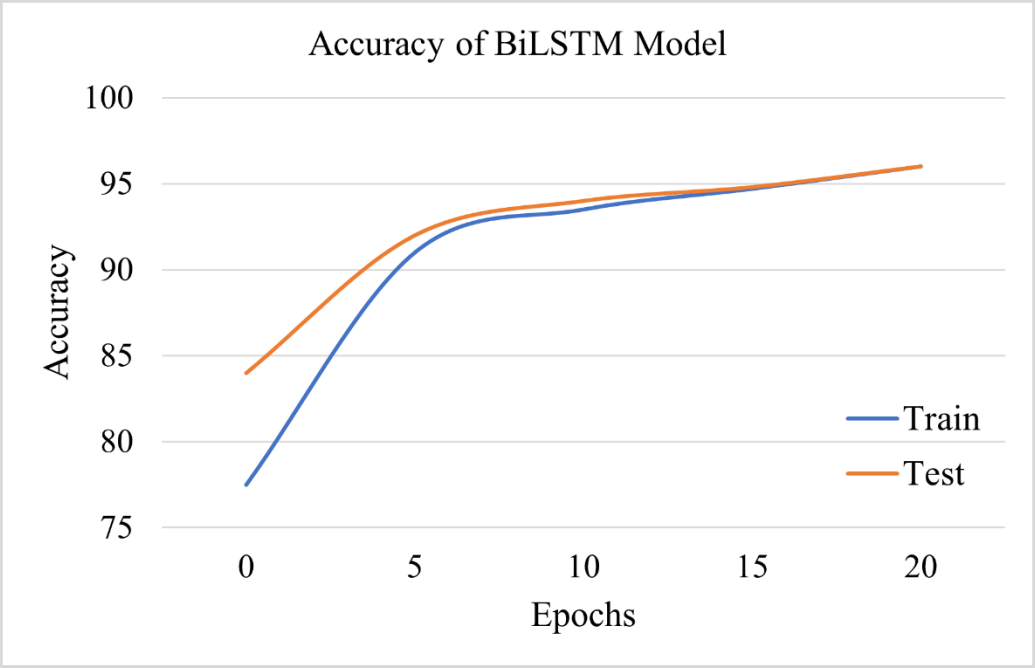


Figure 13: Model accuracy of the BiLSTM model. (Adapted from (Genç and Surer, 2021)).

4.6. Ensemble Classifier

Finally, the Bagging Classifier (Sisodia, 2019) algorithm was developed as the Ensemble Classifier. In the Ensemble Classifier, ANN, Random Forest, Logistic Regression, LSTM, and BiLSTM algorithms are trained together. Ensemble Learning combined the predictions of all machine learning models developed in this study (LR, RF, ANN1, ANN2, LSTM, BiLSTM) and exploited their predictive power. This model performed with a mean accuracy of 0.93 across 10-folds.

Figure 14 shows the 10-fold cross-validation performance of all models developed in our study (Logistic Regression, Random Forest, ANN1, ANN2, LSTM, BiLSTM, and Ensemble Classifier). The comparison of the results of these algorithms can be found in Table 7. The results obtained from the analyses show that the BiLSTM has the best performance on our dataset. After the BiLSTM, the Ensemble Classifier, the second ANN, and LSTM perform well on the ClickbaitTR.

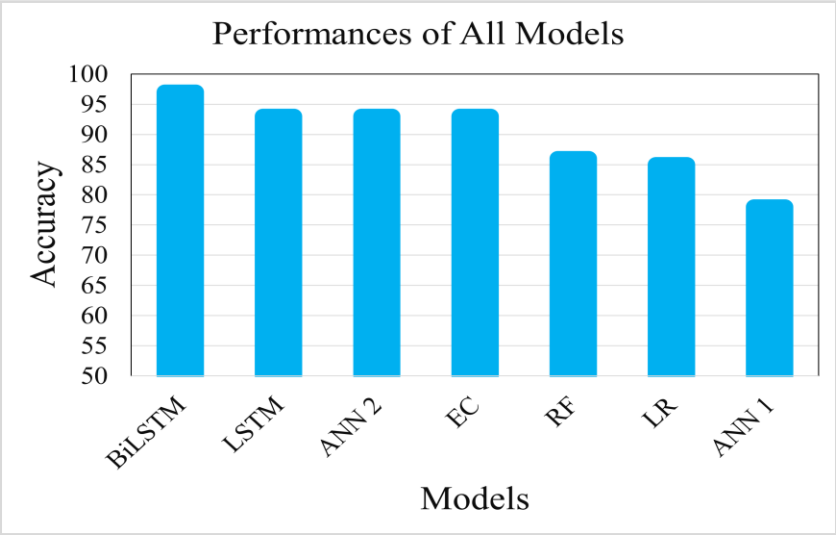


Figure 14: Comparison of performances of Logistic Regression, Random Forest, ANN1, ANN2, LSTM, BiLSTM and Ensemble Classifier. (Adapted from (Genç and Surer, 2021)).

Table 7: Results of the algorithms used in this study (Artificial Neural Networks (ANN), Logistic Regression (LR), Random Forest (RF), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Ensemble Classifier (EC)). All of the algorithms have 10-fold cross-validation, a feature vector size of 10,890, and a sample size of 48,060.

Metrics	**ANN1**	**ANN2**	**LR**	**RF**	**LSTM**	**BiLSTM**	**EC**
Test/validation accuracy	0.78	0.93	0.85	0.86	0.93	0.97	0.93
Precision	0.78	0.94	0.85	0.86	0.93	0.96	0.94
Recall	0.78	0.94	0.85	0.86	0.93	0.96	0.94
F1 score	0.78	0.94	0.85	0.86	0.93	0.96	0.94

4.7. Model Validation

The scores given by the three raters to the 900 news headlines in the test set were used as labels for the headlines, and the outputs of all models were compared with these labels. It was thought that this accuracy analysis, in which the scores of the raters were used as ground truth, will provide further information about the performance of the models. The accuracy scores of the models and three raters can be seen in Figure 15. According to the results coming from 900 samples, rater 2 has an accuracy of 0.93, rater 3 has an accuracy of 0.92, rater 1 has an accuracy of 0.89, the BiLSTM has an accuracy of 0.83, the LSTM has an accuracy of 0.81, the LR has the accuracy of 0.80, the RF has the accuracy of 0.80, and the ANN has the accuracy of 0.79.

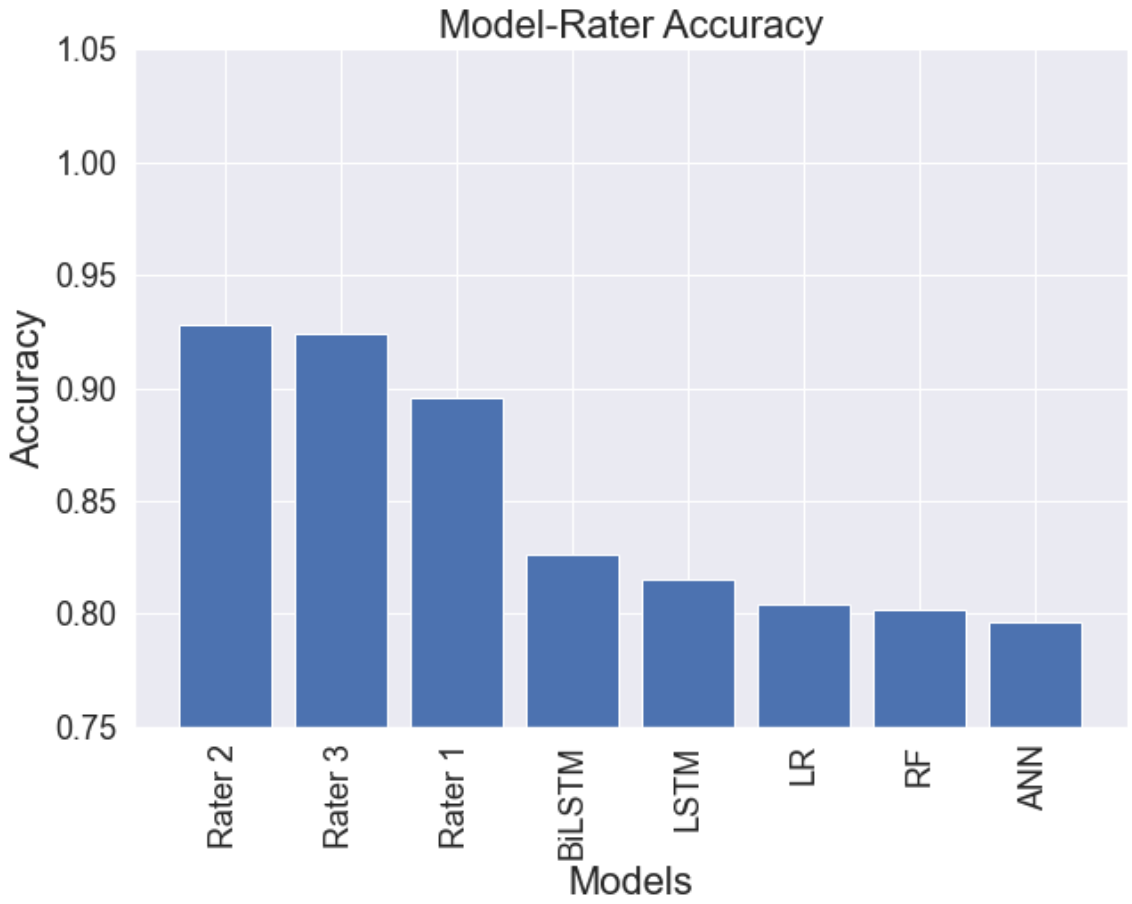


Figure 15: The accuracy scores of the models and three raters (calculated by taking the mean value of raters as the ground truth.)

4.8. Statistical Tests

The results of the different algorithms in the data analysis process were tested using the Mann–Whitney U test in order to see whether there are significant differences between their performances. According to the results of the tests, the performance of first ANN (MLP1) (mean = 0.783, SD = 0.002) is significantly different than the performances of second ANN (MLP2) (mean = 0.935, SD = 0.012), Random Forest (mean = 0.857, SD = 0.003), Logistic Regression (mean = 0.855, SD = 0.005), LSTM (mean = 0.929 , SD = 0.076), BiLSTM (mean = 0.967, SD = 0.033) and Ensemble (mean = 0.935, SD = 0.012) algorithms ($p < .001$). On the other hand, the second ANN performs significantly different than the first ANN, Random Forest, Logistic Regression, LSTM, and BiLSTM algorithms ($p < .001$). But no significant difference found between the performances of the second ANN and Ensemble Classifier.

The Random Forest model's performance is different from the first ANN, the second ANN, LSTM, BiLSTM, and Ensemble Classifier ($p < .001$) while it has the same performance as the Logistic Regression model on the ClickbaitTR dataset.

The performance of Logistic Regression is significantly different than LSTM ($p < .05$) and ANN1, ANN2, BiLSTM, and Ensemble Classifier ($p < .001$). Finally, the results show that LSTM, ANN2, and Ensemble Classifier having the best performance after BiLSTM on the dataset have the same performance.

4.9. Model Confidence

The results of the correlation analysis between confidence values of algorithms show that LSTM and BiLSTM have similar performance in terms of confidence when classifying news headlines as clickbait or non-clickbait. Besides, the confidence values of Random Forest and ANN have a strong correlation. On the other hand, the mean confidence value representing the confidence of three raters on news headlines is more similar to the performances of the LSTM and BiLSTM models. The comparison between the confidence scores of the models and three annotators can be seen in Figure 16.

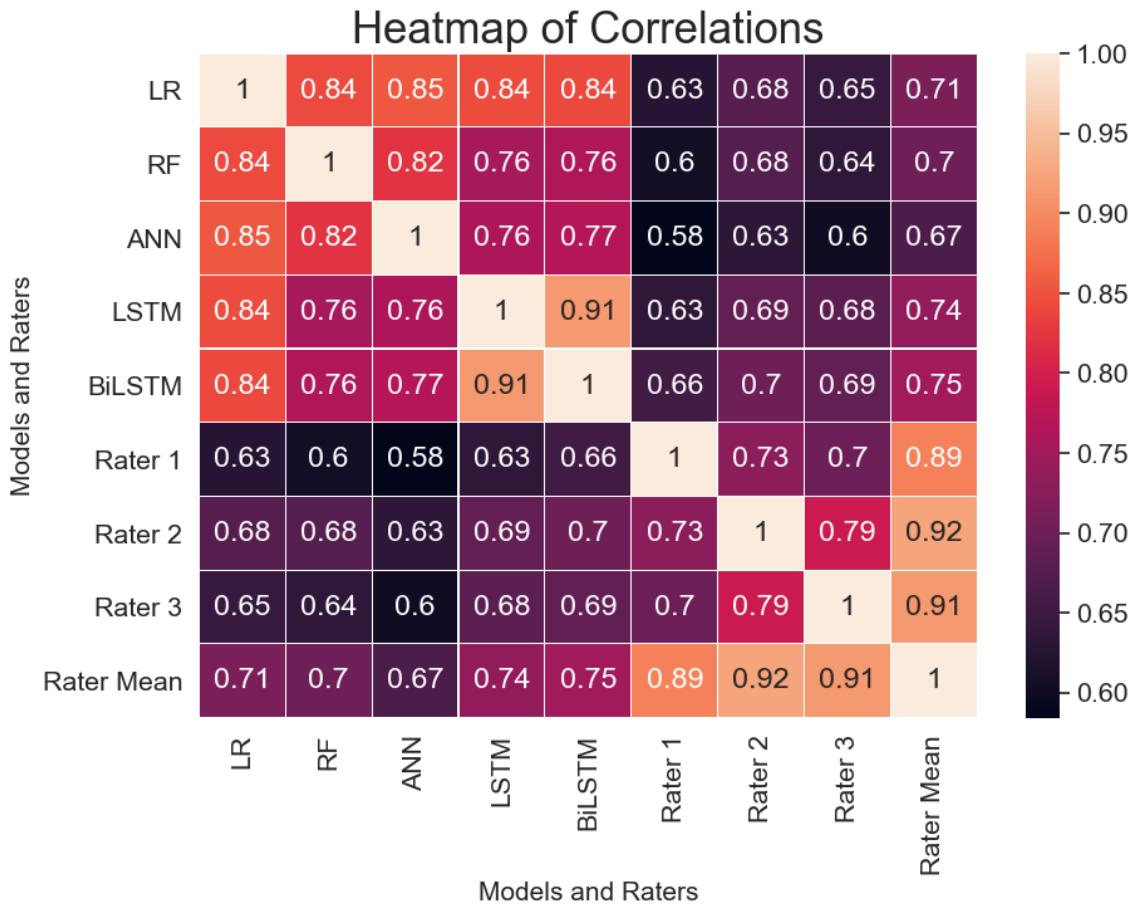


Figure 16: Comparison between the confidence scores of the models and three annotators.

To see the confidence values of the models and raters in more detail, figures showing the confidence frequencies were created (see Figure 17). Looking at Figure 17, it can be seen that ANN made the decision with the highest confidence among the models. However, the average confidence of the raters also seems to be higher than the confidence scores of the models.

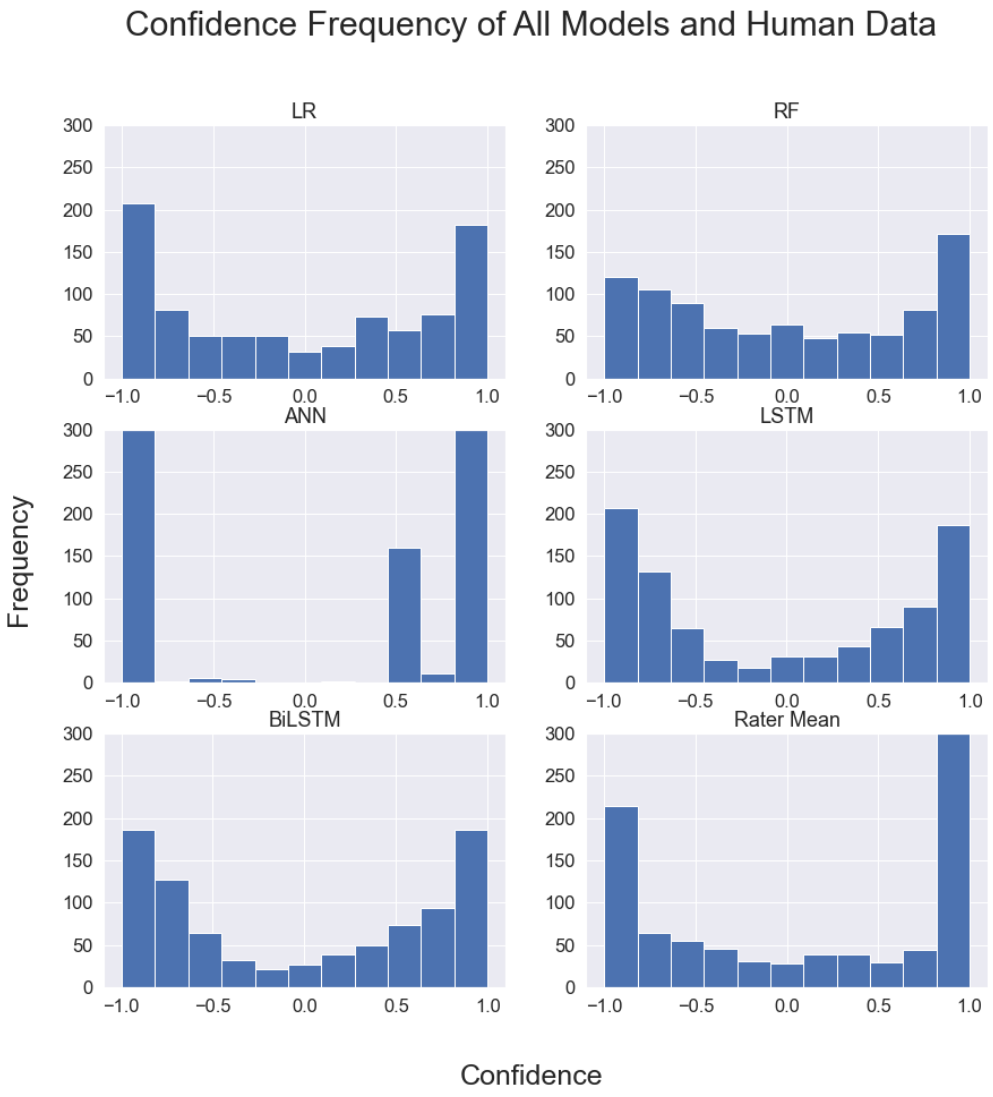


Figure 17: Confidence frequencies of the models and annotator mean.

The confidence-based important features extracted from BiLSTM are the number of dots, the number of exclamation marks, the number of question marks, explain (açıkla), the number of hashtags, she/he (o), flash (flaş), stated (belli), last (son) and attention (dikkat) for clickbait headlines, and conflict (çatışma), more (daha), newspaper (gazete), government (hükümet), say (de), so (diye), percent (yüzde), much or until (kadar), thorn (diken) and parliament (meclis) for non-clickbait headlines (see Figure 18). Similarly, the confidence-based important features extracted from the data coming from annotators are the number of dots, the number of exclamation marks, explain (açıkla), the number of question marks, the number of hashtags, happen (ol), stated (belli), she/he (o), flash

(flaş) and attention (dikkat) for clickbait headlines, and get (al), there is (var), war (savaş), lose (kaybet), get (ede), vote (oy), (hükümet), newspaper (gazete), corner (köşe) and thorn (diken) for non-clickbait headlines (see Figure 19). The confidence-based important features of the LR model can be seen in Appendix E.1., the confidence-based important features of the RF model can be seen in Appendix E.2., the confidence-based important features of the ANN model can be seen in Appendix E.3., and the confidence-based important features of the LSTM model can be seen in Appendix E.4.

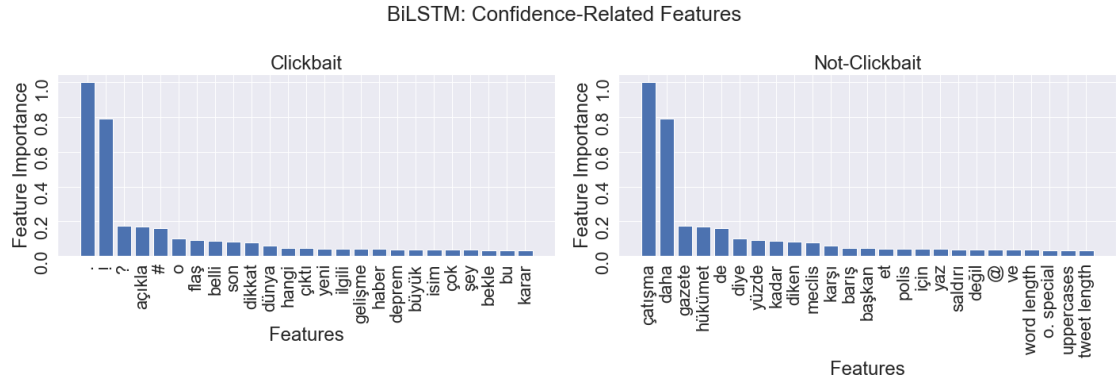


Figure 18: Confidence-based features of the BiLSTM model.

Finally, some examples that were easy and hard to classify (i.e., borderline cases) were reported in Table 8. The following news headline (labeled as clickbait) has a confidence of over 90% of being a clickbait as a result of all models and raters: Good news for students! A second right… (Öğrencilere müjdeli haber! İkinci bir hak…). On the other hand, the following news headline (labeled as non-clickbait) has a confidence of 0.27 in the LSTM model, 0.27 in the BiLSTM model, -1.0 in the ANN model, 0.12 in the LR model, 0.0 in the RF model, and 0.71 in the raters' scores, which makes it a borderline case although it was classified with high confidence by the raters: Metrobus was stopped after a suicide bomb threat, the passenger with a tattoo on his wrist was searched. (Canlı bomba ihbarı üzerine metrobüs durduruldu, bileği dövmeli yolcu arandı).

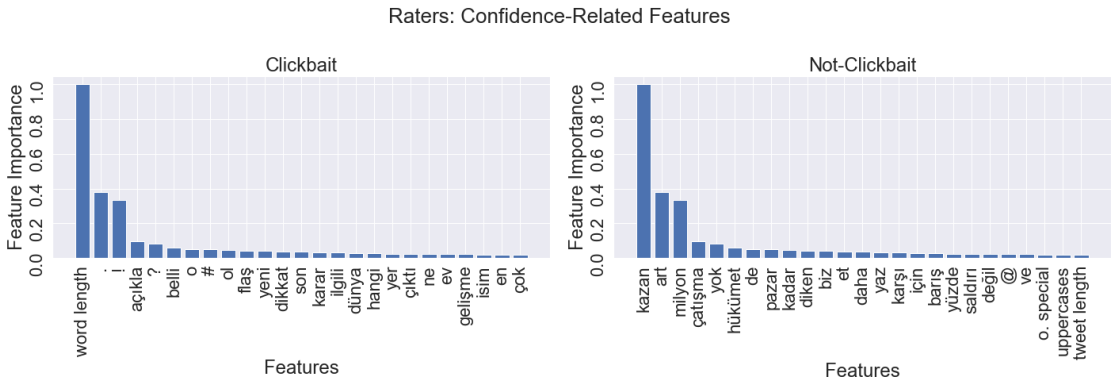


Figure 19: Confidence-based features of the three raters.

Table 8: Examples of news headlines from the dataset with the confidence values of the models and raters.

News Headlines	Label	LSTM	BiLSTM	ANN	LR	RF	Rater
Öğrencilere müjdeli haber! İkinci bir hak…	1	0.99	0.99	1.0	0.95	0.96	1.0
Uyumlu bir çift misiniz?	1	0.96	0.96	0.64	0.85	0.64	1.0
Aman dikkat! Sakın merak edip internette bu siteye girmeyin	1	0.91	0.92	1.0	1.0	0.92	1.0
Galatasaray'dan bomba transfer atağı! İlk sırada o var	1	0.93	0.92	1.0	0.97	0.86	1.0
Eğitime kar engeli! 5 ilde tatil…	1	0.98	0.98	1.0	0.85	0.98	0.85
Gece uyumadan önce bir çay kaşığı bal yerseniz…	1	0.94	0.95	1.0	0.66	0.58	1.0
SON DAKİKA: Yunanistan konusunda Avro Bölgesi zirvesinde anlaşma çıktığı açıklandı.	0	-0.89	-0.89	-1.0	-0.66	-0.86	-0.33
BİM'den "Ürünler sahte" diye ihtarname çeken Apple'a: Bilirkişi "Orijinal dedi"	0	-0.92	-0.91	-1.0	-0.75	-0.76	-0.78

Kemal Dinç konser kayıtlarından oluşan 25 parçalık bir albümle dinleyicilerinin karşısına çıkmaya hazırlanıyor	0	-0.93	-0.92	-1.0	-1.0	-0.78	-1.0
Foça'da denize petrol sızmasıyla ilgili şüpheli gemi inceleniyor.	0	-0.72	-0.72	-1.0	-0.68	-0.26	-0.48
'Çav Bella' yorumu alay konusu olan Hilal Cebeci'den 'analiz'li yanıt: Ben de solcuyum; ülkede solculuk, devrimcilik bitmiş!	0	-0.88	-0.87	-1.0	-0.90	-0.72	-1.0
Yeni Zelanda parlamentosu silah yasası değişikliğini onayladı	0	-0.74	-0.74	-0.99	-0.40	-0.80	-0.63
Polisin "Lastik yakacaktı" izlenimiyle 3 yıl hapis cezası!	0	0.10	0.08	0.48	-0.38	-0.24	0.92
Şiddetli rüzgar ağaçları devirdi	1	0.39	-0.46	-0.62	-0.21	-0.04	0.70
Mersin Üniversitesi'nde öğrenciler 500 davaya bakacak avukat arıyor!	0	0.05	-0.06	-1.0	-0.84	-0.44	-0.04
Kahve zinciri Starbucks ABD'de ilk işaret dilli mağazasını açtı	0	0.34	-0.78	0.89	0.41	0.54	-0.93
Büyük operasyon! Piyasaya süreceklerdi	1	0.20	0.20	-1.0	-0.68	0.09	1.0
Canlı bomba ihbarı üzerine metrobüs durduruldu, bileği dövmeli yolcu arandı	0	0.27	0.27	-1.0	0.12	0.0	0.71

CHAPTER 5

DISCUSSION AND CONCLUSION

In this thesis, a clickbait dataset with 48,060 Turkish samples, ClickbaitTR, was created and publicly released, and six different models were developed in order to make comparisons between different algorithms on this dataset. Artificial Neural Network (ANN), Long Short-Term Memory Network (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Logistic Regression (LR), Random Forest (RF), and Ensemble Classifier (EC), machine-learning algorithms applied on the 10,890-dimension feature vector, were proven to give successful results in detecting Turkish clickbait and non-clickbait news headlines on the ClickbaitTR.

On the other hand, in the preliminary study of this thesis, two models were developed, one of which was trained on the feature vector containing all the features (10,338-dimension), and the other was trained with only nine selected features to see whether the selected features are distinctive for clickbait detection. The results of these two Artificial Neural Network models showed that using nine features as well as the word roots extracted from news headlines in the dataset increases the accuracy of the models. Having the information coming from these results, nine features were added to the analyses of all algorithms in this thesis.

In this study, the first MLP (ANN1) model was trained using the hyperparameters of the MLP model developed in the preliminary study. Results showed that those hyperparameters were not convenient for the current MLP model due to the smaller size of the dataset and the feature vector in the previous study. While the MLP model of the preliminary study performed with 91% accuracy, the current MLP model (ANN1) performed with 78% accuracy. After getting these results, another MLP model (ANN2) was trained with different hyperparameters, and it reached a 93% accuracy. Among the distinctive features of ANN2, the most notable ones are the other special characters, the length of a tweet, flash (flaş), the number of dots, the number of upper cases, the length of words, look (bakın), the number of at signs, development (gelişme), the number of hashtags, interesting (ilginç), horror (dehşet), number of exclamation marks, shock (şok), famous (ünlü), and last (son). Eight of the nine preselected features, and shock,

flash and famous from the words that stand out in word clouds, were distinctive. On the one hand, this result emphasizes the importance of these features; on the other hand, it shows that the ANN2 model is successful in utilizing them.

LSTM (93% accuracy), BiLSTM (97% accuracy), and Ensemble Classifier algorithms (93% accuracy) have the best performance on our dataset as well as the second MLP model. Although the Ensemble Learning model combines a diverse set of learners and makes use of their predictions, it seems that the performance of this model is not as high as BiLSTM, and it appears to be the same as LSTM and ANN (see Figure 14). This is because the performances of LR and RF models lower the average of model predictions, including the performances of powerful models such as BiLSTM, LSTM and ANN.

On the other hand, the Logistic Regression and Random Forest algorithms, applied on a 10,890-dimension feature vector, have the worst performance on the ClickbaitTR dataset, with 85% and 86% accuracy. The following features can be seen among the most distinctive features of the Random Forest model: the number of upper cases, the number of dots, the length of words, the number of at signs, explain (açıkla), the number of exclamation marks, stated (belli), she or he (o) and flash (flaş), famous (ünlü), look (bakın), the number of question marks (?), last (son), the other special characters, development (gelişme), judgment (karar). These features are consistent with the words such as flash, famous, explain, and judgment which were detected in the word clouds created to discover the details of the dataset. Besides, it is seen that twelve of the most important features that ANN2 and Random Forest models learned in training are common. These features include the number of other special characters, tweet length, word length, the number of upper cases, the number of exclamation marks, the number of at signs, the number of dots, flash (flaş), look (bakın), development (gelişme), famous (ünlü), and last (son). Many of these features show that ANN and RF learned similar set of features and patterns. Given that the RF is less computationally expensive, further development of this model for clickbait detection may provide good performance.

After the dataset construction process, the data was scanned to find out the content of the Turkish news and understand the main themes in them. For this purpose, two different methods were used as creating word clouds and applying the LDA method on the dataset. It is important that the dataset created covers the general content of Turkish news, and the most frequent words detected by the word clouds are compatible with the content of the Turkish news. It is quite informative to see such pattern on the data before making any categorization.

The most frequent words of clickbait headlines such as explain (açıkla), flash (flaş), what (ne), stated (belli), she/he(o), and attention (dikkat) can be seen in the examples of the Turkish news such as "The most harmful social media platform for the mental health of young people was explained...", “FLASH! Very important warning about your health!”, “Mysterious 15-meter creature... Nobody could figure out what it is”, and “Attention! Meteorology reported its time. In Istanbul...”. On the other hand, the most frequent words of non-clickbait headlines such as police (polis), newspaper (gazete),

woman (kadın), and judgment (karar) can be seen in the examples of the Turkish news such as “Arda Turan fired a gun in the hospital: He said to the police, ‘It fired while putting it in the holster on my waist’ ”. “Women defend their lives in solidarity”, and “Supreme Cour’s judgment: If the bank is notified within 24 hours of the card being stolen, the withdrawn money will be refunded.”. Providing consistent information about the content of Turkish news headlines, these results show that the ClickbaitTR is an important source for further studies.

Besides reflecting the content of Turkish news, the information coming from the word clouds, and important features of ANN2 and RF models are in line with the studies of Biyani et al. (2016), and Pujahari and Sisodia (2020). Each of these features and their various combinations are used to create varieties of clickbait headlines classified by Biyani et al. (2016), and Pujahari and Sisodia (2020). “She buried an egg and a banana in the ground, look what happened?” is an example of teasing, which means deliberately removing basic information about the content from the title, and its meaning is provided by the word “what.” In the same way, “FLASH! Sale banned! 18 tons caught at the border...” is an example of formatting, which refers to the overuse of capital letters and punctuation marks. In this headline, flash is written in capital letters, and it is tried to arouse curiosity with both capital letters and the word flash.

After developing six different models for clickbait detection, it was thought that investigating the confidence of models in their decisions is also critical for understanding the mechanisms of clickbait strategy. According to the results, the LSTM and BiLSTM detect clickbait headlines with similar confidence values as well as the Random Forest and ANN algorithms. Besides, the LSTM and BiLSTM algorithms’ decisions have similar confidence values to three annotators. On the other hand, according to confidence frequency histograms of the models, ANN and annotators present the most confident decisions on clickbait detection. It can also be seen that LSTM and BiLSTM have high confidence scores, which makes them powerful models for clickbait detection, considering that their accuracy values are also high. Looking at the correlations between models and people's confidence values in Figure 16 (heatmap of correlations), it seems that the models have a higher correlation with the Rater Mean. That is, the correlations of the models with Rater 1, Rater 2 and Rater 3 seem less than their correlation with Rater Mean. This is because the models behave more like human mean. When the scores of the people are examined, it is seen that while Rater 2 and Rater 3 give very high scores to the news headlines, Rater 1 gives relatively low scores, and reduces the Rater Mean. When the confidence values of the models are examined, it is seen that they act like Rater 1 and do not output very high confidence values. So instead of making decisions with extreme confidence scores like Rater 2 and Rater 3, they give more intermediate scores like Rater 1.

When the features that most affect the confidence scores of the BiLSTM model, which has the best performance on the ClickbaitTR dataset, are examined, it can be seen that the features such as the number of dots, the number of exclamation marks, the number of question marks, explain (açıkla), the number of hashtags, she/he (o), flash (flaş), stated

(belli), last (son) and attention (dikkat) for clickbait headlines are distinctive in terms of confidence scores. These features are evident in the following examples of headlines where the model decides with high confidence to be clickbait: Good news for students! A second right… / Öğrencilere müjdeli haber! İkinci bir hak… (classified as clickbait with 0.99 confidence); Snowfall barrier to education! Snow break in 5 cities… / Eğitime kar engeli! 5 ilde tatil… (classified as clickbait with 0.98 confidence); Attention! Do not go to this site on the internet out of curiosity / Aman dikkat! Sakın merak edip internette bu siteye girmeyin (classified as clickbait with 0.92 confidence); Surprising transfer from Galatasaray! He's in the first place / Galatasaray'dan bomba transfer atağı! İlk sırada o var (classified as clickbait with 0.92 confidence) (see Table 8). The use of features such as the number of dots, the number of exclamation marks, attention (dikkat), he (o) has been identified by the model to be related to the clickbaitness of these sentences, which is consistent with the information from the three raters' confidence scores. According to these scores, the number of dots, the number of exclamation marks, explain (açıkla), the number of question marks, stated (belli), she/he (o), the number of hashtags, flash (flaş), and attention (dikkat) are also important features for people to classify headlines as clickbait with high confidence (see Figure 19).

On the other hand, when looking at the confidence-based features of BiLSTM and people for non-clickbait sentences, content words such as conflict, newspaper, summer, win, million or market are seen. It is clear that these words are random and different from words used to arouse curiosity such as flash or attention. However, these features also include some of the selected features such as other special characters, the number of at signs, the number of capital letters, and the length of the tweet. Of these features, other special characters are critical for the evaluation of non-clickbait sentences, because these sentences are written correctly according to spelling rules and contain many other special characters such as quotation marks or apostrophes. The fact that the number of capital letters in these sentences is also among the confidence-based features of non-clickbait sentences is that these sentences frequently contain certain words starting with a capital letter and referring to specific information such as the proper names or location names. In these sentences, these proper names are tagged in tweets using the at signs (@) frequently. This explains why the at sign is among the confidence-based features for non-clickbait sentences.

It is highly informative to examine the confidence values of models on some specific news headlines. The following example can be investigated: 3 years imprisonment with the impression of the police "would burn a tire"! / Polisin "Lastik yakacaktı" izlenimiyle 3 yıl hapis cezası!. This news headline was labeled as non-clickbait in the dataset but was labeled as clickbait by raters with 92% confidence. This sentence was classified as clickbait by ANN (0.48), was classified as non-clickbait by RF (-0.24) and LR (-0.38). It can be said that LSTM (0.10) and BiLSTM (0.08) were confused about this sentence. The reason why the ANN can classify this example correctly may be because it learns specific features such as exclamation marks. Here is another example: Students at Mersin University are looking for lawyers to handle 500 cases! / Mersin

Üniversitesi'nde öğrenciler 500 davaya bakacak avukat arıyor!. Although this headline was labeled as non-clickbait in the dataset, raters (-0.04) were undecided about it. Although this headline cannot be classified with high confidence by LSTM (0.05) and BiLSTM (-0.06), the correct classification by ANN (-0.84) with high confidence may also be due to its specific features such as an exclamation mark. Finally, the following example can be examined: Strong winds knocked down trees / Şiddetli rüzgar ağaçları devirdi. In this news headline, which is labeled as clickbait in the dataset and seen as clickbait by the raters, there is no feature that can enable it to be evaluated as clickbait by the models. This headline, which is classified as non-clickbait by all models except LSTM (0.39), is misclassified with the highest confidence value by the ANN (-0.62), which may indicate that the ANN model works by learning some specific features rather than capturing a wider range of features. This also explains why the accuracy score comparing the ANN with the rater mean very low compared to other models. On the other hand, the BiLSTM, which showed the best performance on the dataset, was the model that showed the best accuracy performance when compared to the rater mean (Figure 15). Accuracy scores calculated by assuming the rater mean as the ground truth are consistent with the performances of the models, although they provide insufficient information as they are obtained from only 900 samples.

When the borderline examples are examined, some features of clickbait headlines that the human cognition pays attention most and some features of clickbait headlines that the models do not learn can be seen. It is observed that the headlines that have no special characters, punctuation marks or the words such as flash or attention were not classified with high confidence values by models. For example, the headline "Metrobus was stopped after a suicide bomb threat, the passenger with a tattoo on his wrist was searched" has none of the nine features selected from the literature or confidence-based features detected in confidence analysis. All models, except from ANN, are confused about this sentence. On the other hand, this headline, which was classified as non-clickbait in the dataset, was classified as nonclickbait with 0.71 confidence value by the annotators, as well. This result shows that people pay attention to certain patterns that models cannot learn when evaluating the clickbaitness of the sentences. In the example above, the readers are informed that there could be a bomb at the metrobus, but the result of the event is not stated. This headline may be classified as clickbait by people because of the knowledge gap created by this critical security problem. Although there is a high correlation between the confidence values of people and the confidence values of the BiLSTM model, there are many examples that are labeled as with high confidence by people and low confidence by the BiLSTM. That is, the experimental studies for understanding which pattents people pay attention in clickbait headlines can provide essential information on human cognition and mechanisms of curiosity.

One of the strengths of this thesis is addressing the limitations of the first clickbait study in Turkish, such as the number of samples, the suitability and variety of samples, and solutions. Firstly, Geçkil et al. (2018) collected the non-clickbait data from news sources such as BBC Turkish and Anadolu News Agency, which were found to be doing clickbait and evaluated in the clickbait category in this thesis. Secondly, the number of

samples (48,060) gathered from the Twitter accounts of various news sources is higher than the number of samples (4000) in the study of Geçkil et al. (2018). Finally, they used the headlines from five news sources while this dataset contains samples from 60 different news sources.

Another strength of this study is that it provides a suitable dataset for Turkish clickbait studies as well as cross-language studies. The spread of accessing news via social media makes it necessary to examine the clickbait strategies used by online news channels on social media networks. Clickbait is one of the most problematic strategies used by news channels because it creates a gap between the objectives of the reader and the news source and contradicts the basic motivation of journalism. Therefore, it is important to examine the patterns and structure of clickbait and non-clickbait sentences in different languages, and this makes the existence of datasets on clickbait in various languages very critical.

Besides all these strengths, the ClickbaitTR, created by collecting data from many different sources, provides a Turkish dataset for a variety of areas such as Cognitive Science, Computer Science, Psychology, and Linguistics. For example, this dataset includes many kinds of clickbait sentences written to arouse curiosity in people, which makes it suitable for psychology studies on curiosity and interest and natural language processing. So, it may provide stimuli to experimental studies designed for different purposes ranging from training artificial learning algorithms to conducting human experiments.

Besides the strengths of this study, there are also some limitations. The most crucial of these is about labeling the headlines in the dataset. The news headlines taken from Diken Newspaper in the dataset were operationally labeled as non-clickbait before applying machine learning algorithms. However, when the data of this source is reviewed, it is seen that it contains some clickbait headlines like the other news sources. The assumption that all titles of this resource are non-clickbait is an important limitation. In order to solve this limitation, the whole training set and test set should be evaluated by human raters before developing clickbait detection models.

A comprehensive clickbait dataset including 48,060 clickbait and non-clickbait headlines of Turkish news sources is created and presented publicly in this thesis. Twitter, which includes active accounts of many news sources, was preferred as the platform to extract data. Firstly, because News media outlets frequently use Twitter, a wide range of news headlines that have been posted over a wide range of time was obtained. Secondly, the largest dataset consisting of Turkish clickbait and non-clickbait sentences was created, and six different models were developed to achieve the best performance in Turkish clickbait detection. Finally, the analysis results and the concept of clickbait were discussed in terms of curiosity and interest, which may provide a new perspective in terms of the mechanisms of the clickbait strategy and curiosity. In all these aspects, this thesis contributes to the linguistic and psychological analysis of clickbait sentences as well as the clickbait detection research.

5.1. Further Studies

In future studies for Turkish clickbait detection, other textual features of tweets such as user comments and hashtags can be used in the analysis. First, comments to a news headline in a tweet made by other users can provide valuable information about whether that headline is clickbait or whether it seems reliable to people. Especially when the comments made to the clickbait news headlines are examined, it is seen that some words such as click, title, trick, news are frequently used in these comments. The following headline (a) shared by a Turkish news source on Twitter and the comments under it can be given as an example:

a) "And he announced his decision to resign / Ve istifa kararını açıkladı"

   "Follow @LimonHaber. Do not click in vain."

   "So, is this news now? You do not need a game like this to get attention."

   "You better not make those cheap tricks to get clicks as Hürriyet or Sabah newspapers do."

   "Is this the important news?"

   "This is a shame. Are you trying to prove that you are not different from the other newspapers?"

   "A very bad headline, officially HAMMING. The OdaTV does not need it. Be yourself."

   "It is really not nice that you are doing this. Look at the news site, it is not more than the magazine..."

   "Look at the news you made to get two clicks more!"

   "What kind of title is this?"

Second, the hashtags used in the tweets and retweets of news headlines can be another source of information about the content of that tweet and its general effect on users. For example, the hashtags #BREAKINGNEWS and #FLASH are frequently used in clickbait news headlines, and the hashtags #fakenews and #clickbait written by other users can be a part of future analyses in clickbait detection.

Thirdly, besides the textual features, certain non-textual features such as images and videos presented with the news headline can distinguish clickbait and non-clickbait headlines. For example, in general, it can be said that non-clickbait news headlines are shared with an image of the person or place that is the subject of the news, while

clickbait news headlines generally share images that are unrelated to the content of the news.

On the other hand, when the Turkish news headlines in Twitter are examined, it can be seen that the number retweets of the clickbait headlines are more than the number of their likes. This observation may indicate that the clickbaits are widely shared by users or commercial accounts although the users who read the main text of those news do not like the content. Therefore, in future studies, the data obtained from Twitter should be analyzed with retweet and like numbers other non-textual features.

In future studies, analyzing the psychological mechanisms of clickbaits conducting experimental research with human participants may also be useful in understanding the nature of clickbait. Although the concept of clickbait was investigated in terms of curiosity and interest in this thesis, other cognitive mechanisms related to how this strategy works so well should also be explored. The varieties of clickbait categorized in two studies (Biyani et al., 2016; Pujahari & Sisodia, 2020) can be subjected to different analyses and experiments to investigate the psychological mechanisms of clickbait. For example, while inflammatory clickbaits may be arousing angry or violent feelings as well as curiosity in people, teasing clickbaits may be working just by curiosity. Working on clickbait with different emotions can contribute to the understanding of the emotional functions of humans as well as its contribution to the understanding of the nature of clickbait.

Finally, in this thesis, the different types of clickbait detected by Biyani et al. (2016) can be evaluated in terms of some of the different characteristics of stimuli, stated by Berlyne (1960), such as novelty, complexity, surprisingness, and ambiguity. The fact that there are different types of clickbaits triggered by different attributes such as novelty or complexity might lead researchers to future studies that can provide a basis for reconciling the novelty-based and complexity theories.

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). Predicting flu trends using twitter data. *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, 702–707. https://doi.org/10.1109/INFCOMW.2011.5928903

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (2001). *An introduction to language and communication.* Cambridge, MA: MIT Press.

Aksoy A. (2017). *Kalbur Project.* https://github.com/ahmetax/kalbur

Alexa. (n.d.). *The top 500 sites on the web.* https://www.alexa.com/topsites

Anand, A., Chakraborty, T., & Park, N. (2017). We used neural networks to detect clickbaits: You won't believe what happened next! *Lecture Notes in Computer Science Advances in Information Retrieval,* 541–547.

Atodiresei, C. S., Tănăselea, A., & Iftene, A. (2018). Identifying fake news and fake users on Twitter. *Procedia Computer Science, 126*, 451–461. https://doi.org/10.1016/j.procS.2018.07.279

Berlyne, D. (1966). Curiosity and exploration. *Science, 153*(3731), 25–33.

Bhowmik, S., Rony, M. M. U., Haque, M. M., Swain, K. A., & Hassan, N. (2019). Examining the role of clickbait headlines to engage readers with reliable health-related information. *ArXiv, Raphael*.

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016). "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. *Thirtieth AAAI Conference on Artificial Intelligence*, 94–100.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Brändle, F., Wu, C. M., & Schulz, E. (2020). Spotlight what are we curious about ? *Trends in Cognitive Sciences*, *24*(9), 685–687. https://doi.org/10.1016/j.tics.2020.05.010.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1201/9780429469275-8

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 9–16. https://doi.org/10.1109/ASONAM.2016.7752207

Chakraborty, A., Sarkar, R., Mrigen, A., & Ganguly, N. (2017). Tabloids in the era of social media? Understanding the production and consumption of clickbaits in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW). https://doi.org/10.1145/3134665

Chollet F. (2015). Keras: Deep Learning library for Theano and TensorFlow. *Data Sci Cent.* 7(8), T1.

Clickbaittr. (2021). *Turkish Clickbait Dataset.* GitHub. https://github.com/clickbaittr/turkish-clickbait-dataset.

Cohen 1960 DO - 10.1177/001316446002000104, J. D. A.-A. 1. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46 ST-A coefficient of agreement for nominal. http://epm.sagepub.com

Di Eugenio, B., & Glass, M. (2004). Squibs and discussions - The Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. https://doi.org/10.1162/089120104773633402

Diken [@DikenComTr]. (2013, June). *Diken Gazetesi* [Twitter account]. Twitter. https://twitter.com/DikenComTr

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476. https://doi.org/10.31234/osf.io/wg5m6

Düzel, E., Bunzeck, N., Guitart-Masip, M., & Düzel, S. (2010). Novelty-related motivation of anticipation and exploration by dopamine (NOMAD): Implications for healthy aging. *Neuroscience and Biobehavioral Reviews*, *34*(5), 660–669. https://doi.org/10.1016/j.neubiorev.2009.08.006

Evrensel Gazetesi. [@evrenselgzt]. (2009, December). *Evrensel Gazetesi* [Twitter account]. Twitter. https://twitter.com/evrenselgzt

Gamallo Otero, P. (2008). The meaning of syntactic dependencies. *Linguistik Online*, *35*(3), 33–53. https://doi.org/10.13092/lo.35.522

Gazete Tirajları. (n.d.). *Weekly Circulations.* http://gazetetirajlari.com

Geckil, A., Mungen, A. A., Gundogan, E., & Kaya, M. (2018). A clickbait detection method on news sites. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 932–937. https://doi.org/10.1109/ASONAM.2018.8508452

Genc, S., & Surer, E. (2019). Detecting "clickbait" news on social media using machine learning algorithms. *27th Signal Processing and Communications Applications Conference, SIU 2019*, 1–4. https://doi.org/10.1109/SIU.2019.8806257

Genç, Ş., & Surer, E. (2021). ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms. *Journal of Information Science*, 016555152110077. https://doi.org/10.1177/01655515211007746

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., Oconnor, K., Sarker, A., Smith, K., & Gonzalez, G. (2014). Mining Twitter for adverse drug reaction mentions : A corpus and classification benchmark. *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM), 1*.

Gottlieb, J., Oudeyer, P., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention : Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

GWI. (2019). *Research by Global Web Index.* https://www.gwi.com/reports/social-2019

Hardalov, M., Koychev, I., & Nakov, P. (2016). In search of credible news. *17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), September*, 172–180. https://doi.org/10.1007/978-3-319-44748-3_17

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & Villalba, L. J. G. (2019). Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors (Switzerland)*, *19*(7). https://doi.org/10.3390/s19071746

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural*

Computation, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hsu, D., Moh, M., & Moh, T. S. (2017). Mining frequency of drug side effects over a large twitter dataset using apache spark. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 915–924. https://doi.org/10.1145/3110025.3110110

Hübl, F., Cvetojevic, S., Hochmair, H., & Paulus, G. (2017). Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *6*(10), 1–23. https://doi.org/10.3390/ijgi6100302

Limon Haber. [@LimonHaber]. (2016, April). *Limon Haber* [Twitter account]. Twitter. https://twitter.com/LimonHaber

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *International Conference on Information and Knowledge Management, Proceedings*, 375–384. https://doi.org/10.1145/1645953.1646003

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.

López-Sánchez, D., Herrero, J. R., Arrieta, A. G., & Corchado, J. M. (2018). Hybridizing metric learning and case-based reasoning for adaptable clickbait detection. *Applied Intelligence*, *48*(9), 2967–2982. https://doi.org/10.1007/s10489-017-1109-7

Newman, N., Fletcher, R., Schulz, A., Andı, S., and Nielsen, R. K. (2020). Reuters Institute digital news report 2020. *Reuters Institute for the Study of Journalism (RISJ)*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Bertrand, T. (2011). Scikit-learn: Machine learning in Python. *Journal Of Machine Learning Research*, *12*(9), 2825–2830. https://doi.org/10.1289/EHP4713

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., Hagen, M., & Stein, B. (2018). Crowdsourcing a large corpus of clickbait on Twitter. *Proceedings of the 27th International Conference on Computational Linguistics*, 1498–1507. https://www.aclweb.org/anthology/C18-1127

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. *European Conference on Information Retrieval*, 810--817. https://doi.org/10.1007/978-3-319-30671-1_72

Pujahari, A., & Sisodia, D. S. (2020). Clickbait detection using multiple categorization techniques. *ArXiv*.

Qu, J., Hißbach, A. M., Gollub, T., & Potthast, M. (2018). Towards crowdsourcing clickbait labels for YouTube videos. *CEUR Workshop Proceedings*, *2173*.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, May 2010*, 45–50. https://doi.org/10.13140/2.1.2393.1847

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, the STS-Gold. *CEUR Workshop Proceedings*, *1096*, 9–21.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Schwenk, H., & Gauvain, J. (2005). Training neural network language models. *Proceedings OfHuman Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), October*, 201–208.

Shang, L., Zhang, D. (Yue), Wang, M., Lai, S., & Wang, D. (2019). Towards reliable online clickbait video detection: A content-agnostic approach. *Knowledge-Based Systems*, *182*, 104851. https://doi.org/10.1016/j.knosys.2019.07.022

Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion*, *5*(1), 89–102. https://doi.org/10.1037/1528-3542.5.1.89

Sisodia, D. S. (2019). Ensemble learning approach for clickbait detection using article headline features. *Informing Science*, *22*(2019), 31–44. https://doi.org/10.28945/4279

Sisodia, Di. S., & Reddy, N. R. (2018). Sentiment analysis of prospective buyers of mega online sale using tweets. *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017*, 2734–2739. https://doi.org/10.1109/ICPCSI.2017.8392217

Spoiler Haber. [@spoilerhaber]. (2013, March). *Spoiler Haber* [Twitter account]. Twitter. https://twitter.com/spoilerhaber

Statista. (n.d.). *Platform GNBD.* http://www.statista.com

Tang, H., Houthooft, R., Foote, D., Stooke, A., & Chen, X. (2017). # Exploration : A study of count-based exploration for deep reinforcement learning. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, arXiv:1611.04717

Tweepy. (n.d.). *An easy-to-use Python library for accessing the Twitter API.*

https://www.tweepy.org

Twitter API. (n.d.). *Use Cases, Tutorials, Documentation Twitter Developer*. https://developer.twitter.com

Uddin Rony, M. M., Hassan, N., & Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 232–239. https://doi.org/10.1145/3110025.3110054

William, A., & Sari, Y. (2020). CLICK-ID : A novel dataset for Indonesian clickbait headlines. *Data in Brief*, *32*, 106231. https://doi.org/10.1016/j.dib.2020.106231

Wongsap, N., Lou, L., Jumun, S., Prapphan, T., Kongyoung, S., & Kaothanthong, N. (2018). Thai clickbait headline news classification and its characteristic. *2018 International Conference on Embedded Systems and Intelligent Technology and International Conference on Information and Communication Technology for Embedded Systems, ICESIT-ICICTES 2018*, 1–6. https://doi.org/10.1109/ICESIT-ICICTES.2018.8442064

Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*(1–2), 41–75. https://doi.org/10.1007/s10994-010-5221-8

Zheng, H. T., Chen, J. Y., Yao, X., Sangaiah, A. K., Jiang, Y., & Zhao, C. Z. (2018). Clickbait convolutional neural network. *Symmetry*, *10*(5), 1–12. https://doi.org/10.3390/sym10050138

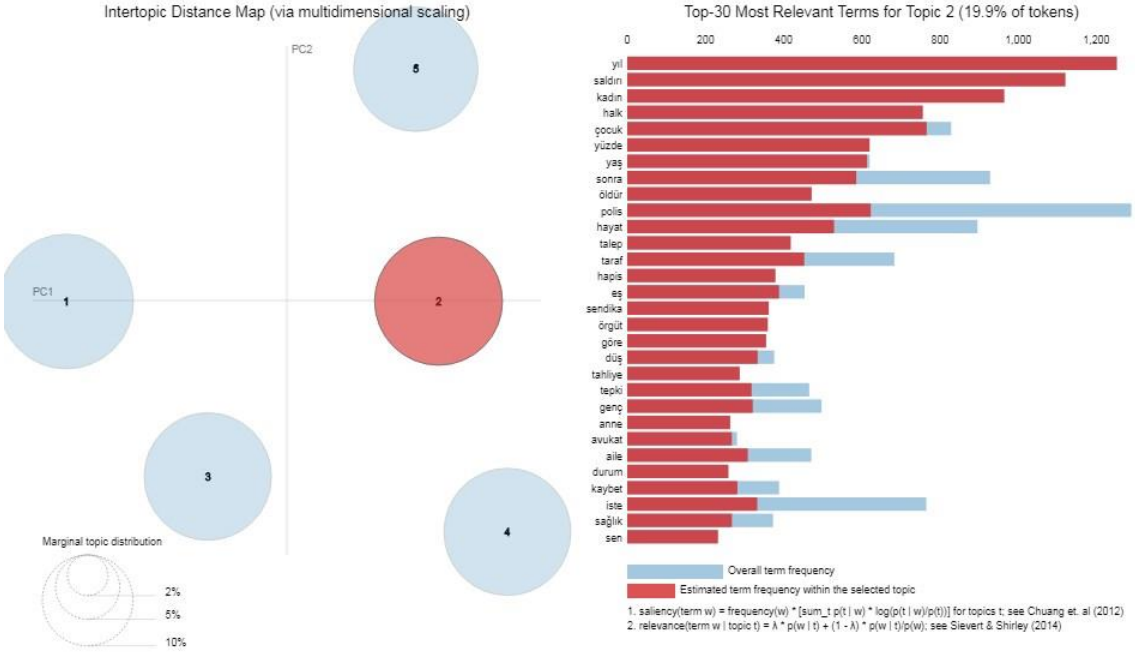
Zhou, Y. (2017). Clickbait detection in tweets using self-attentive network. *ArXiv*.

APPENDICES

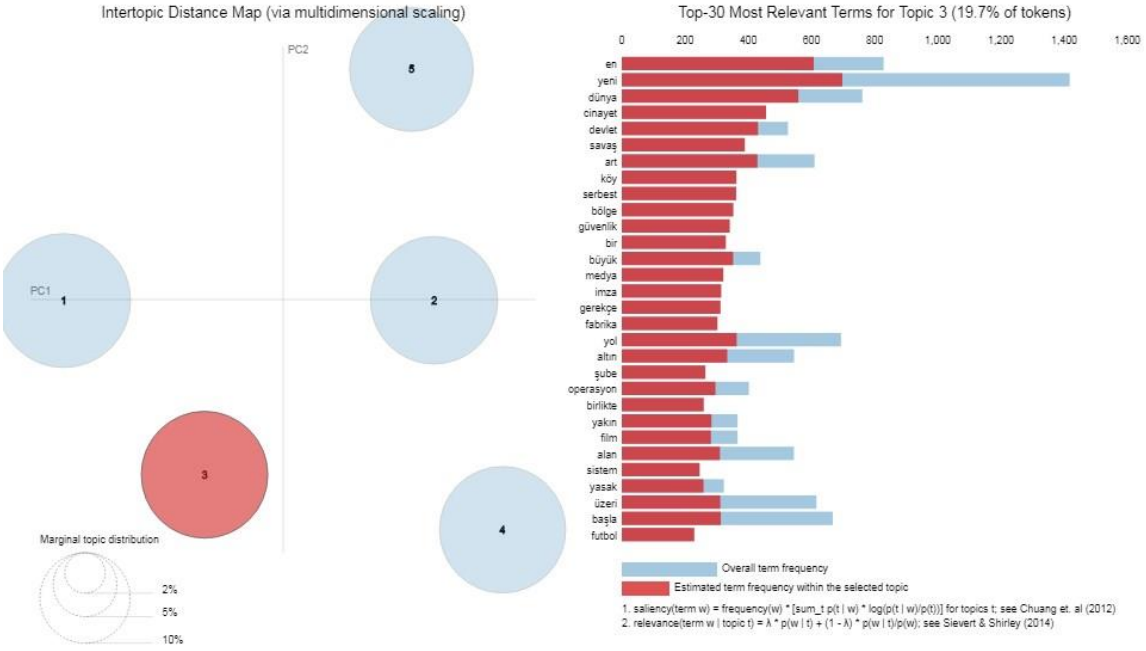
APPENDIX A

INTERTOPIC DISTANCE MAPS FROM THE DATASET

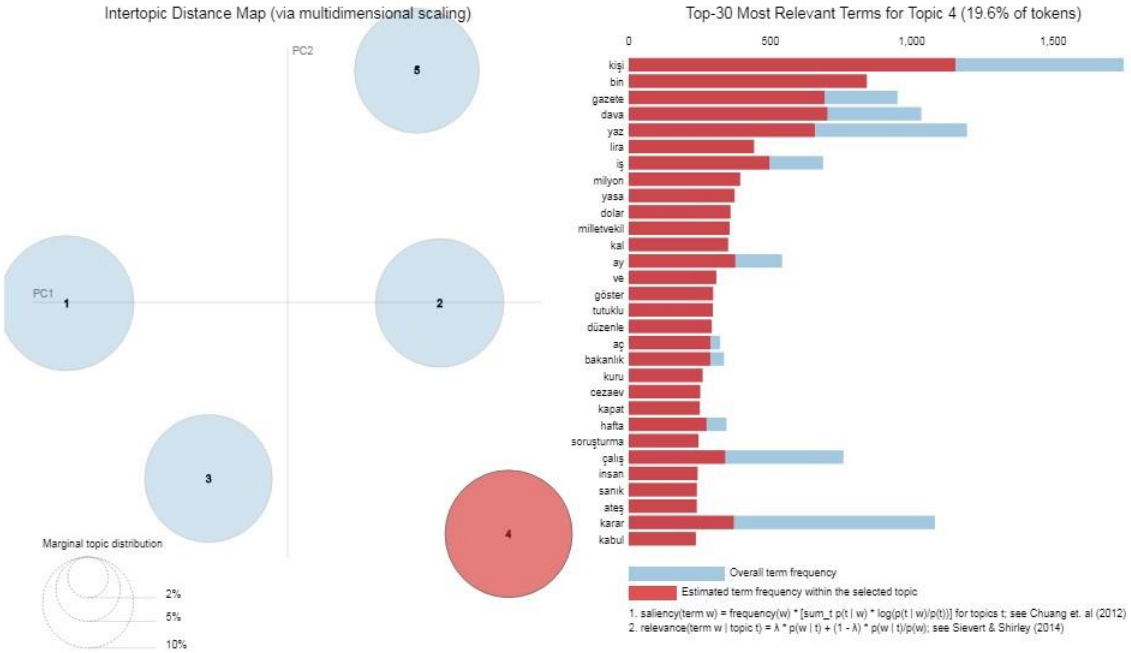
Appendix A.1. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 2 (woman, child, and law) from the dataset.



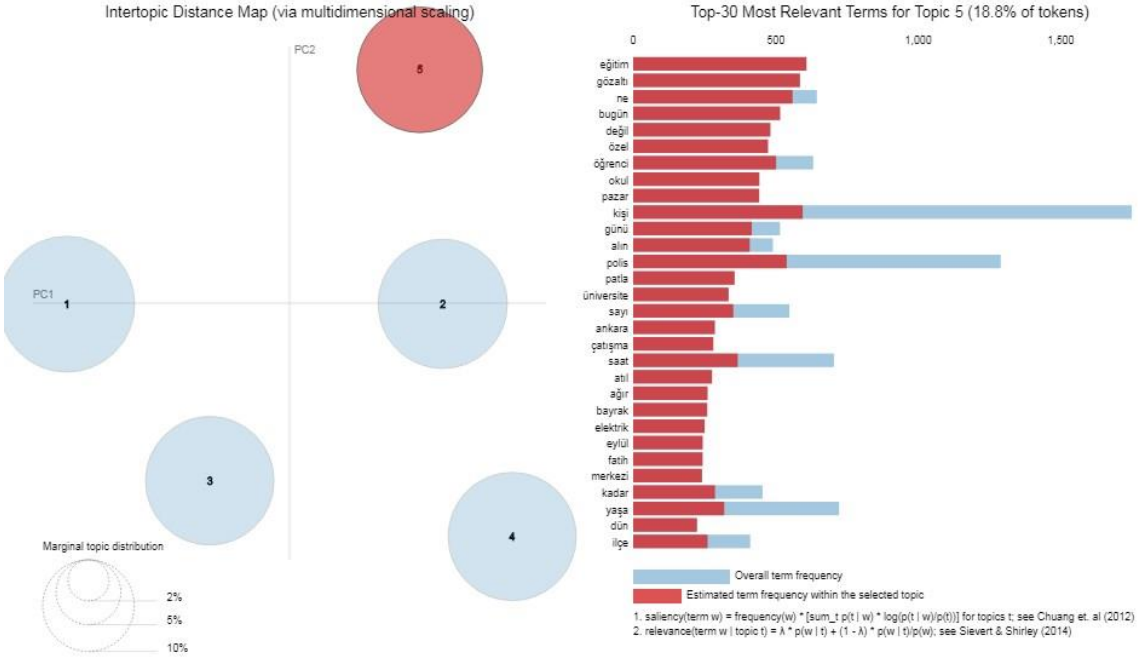
Appendix A.2. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 3 (crime and security) from the dataset.



Appendix A.3. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 4 (economy) from the dataset.



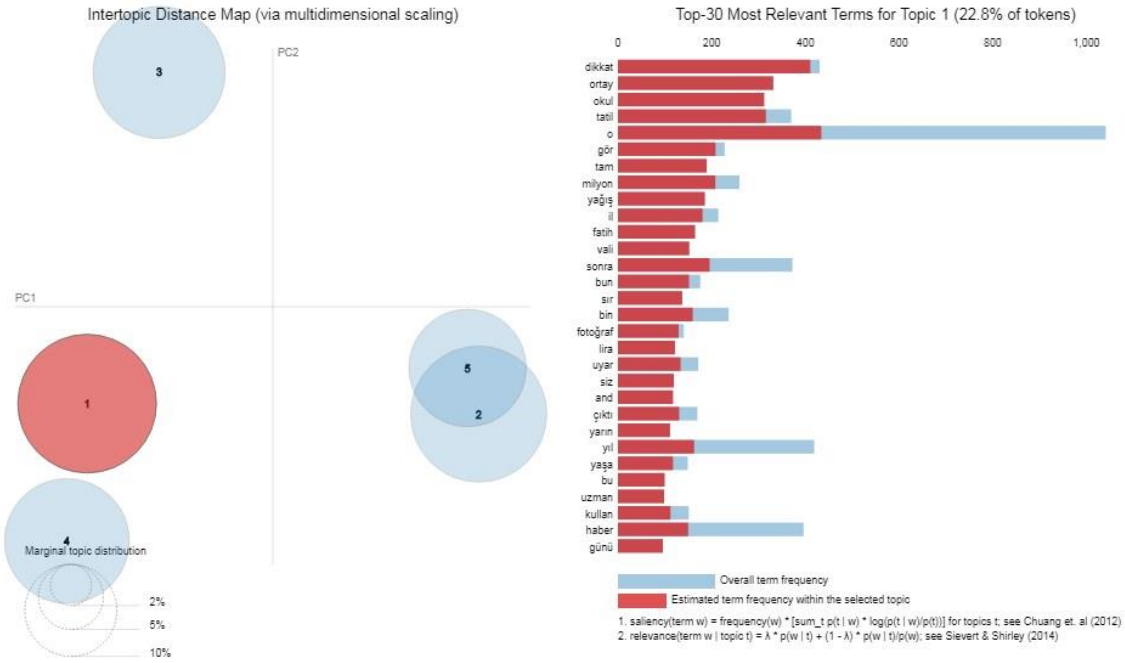
Appendix A.4. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 5 (education) from the dataset.



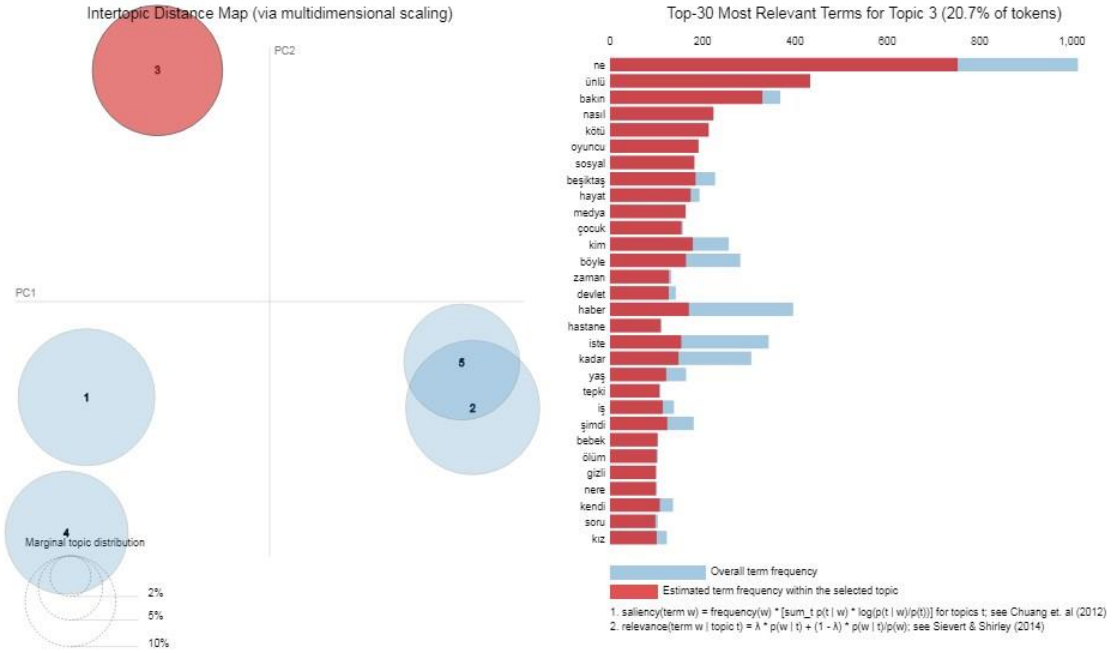
APPENDIX B

INTERTOPIC DISTANCE MAPS FROM THE CLICKBAIT DATA

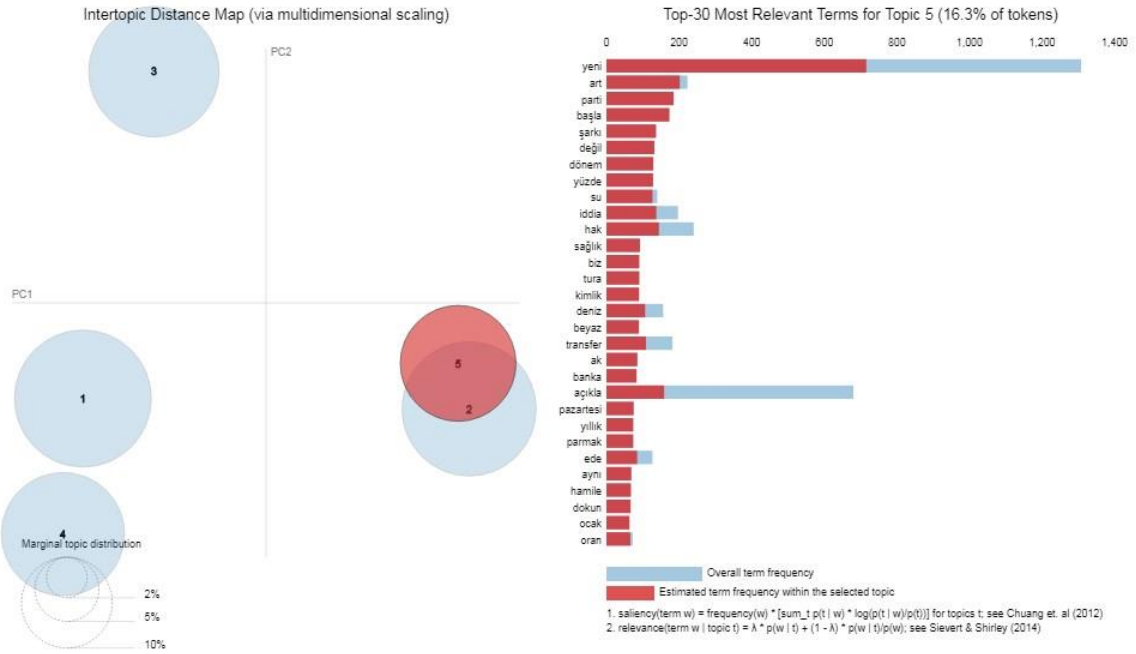
Appendix B.1. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (sports and football) from the clickbait data.



Appendix B.2. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 3 (politics) from the clickbait data.



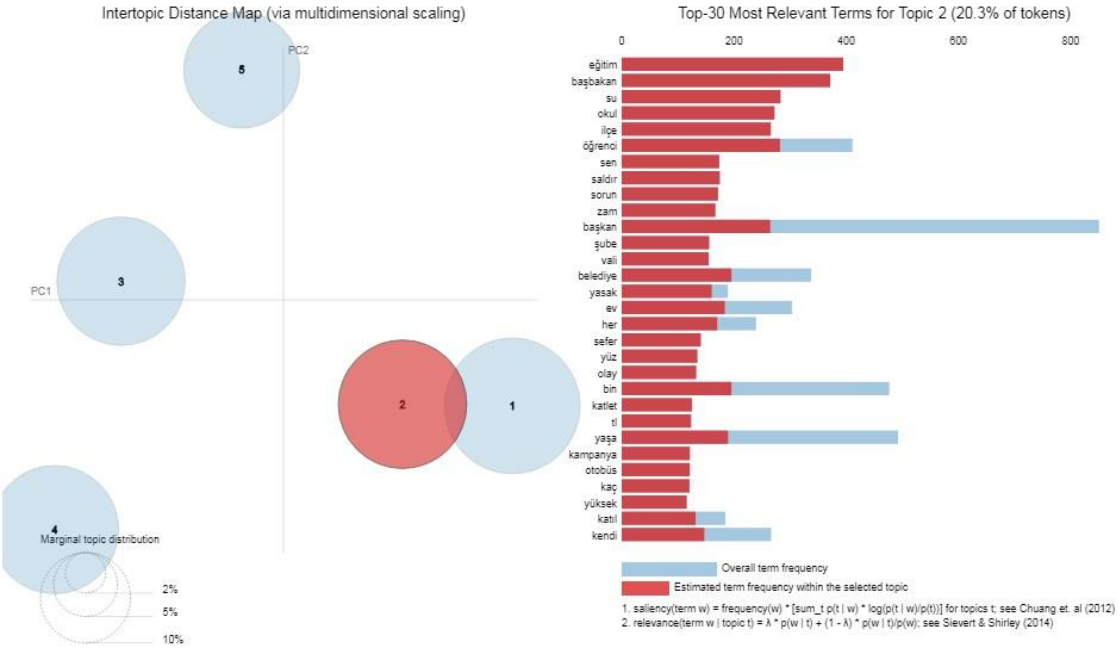
Appendix B.3. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 5 (economy) from the clickbait data.



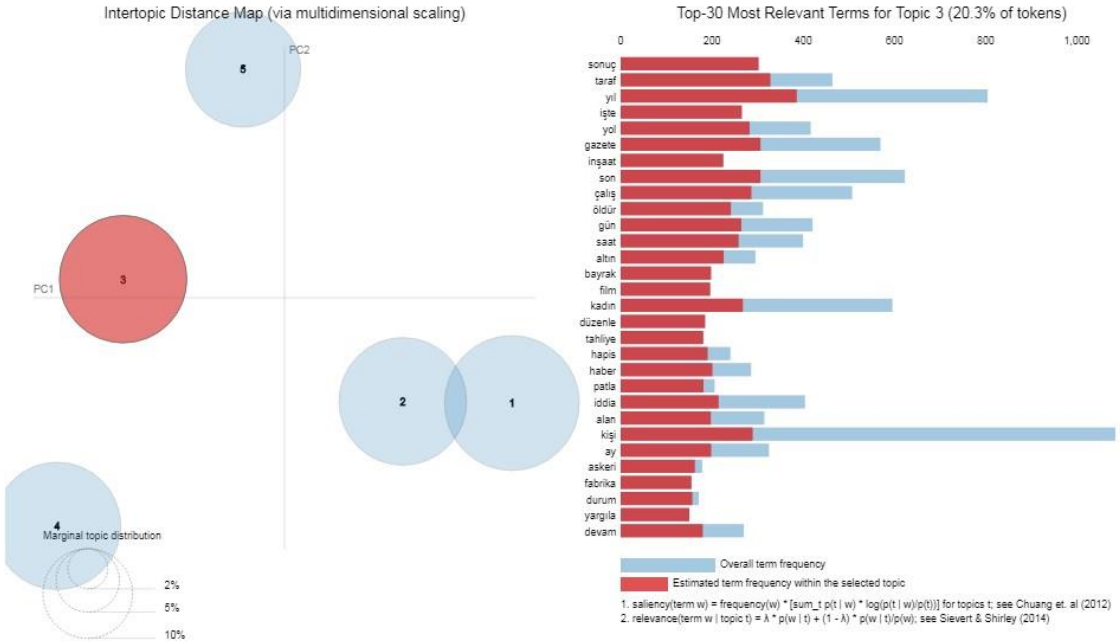
APPENDIX C

INTERTOPIC DISTANCE MAPS FROM THE NON-CLICKBAIT DATA

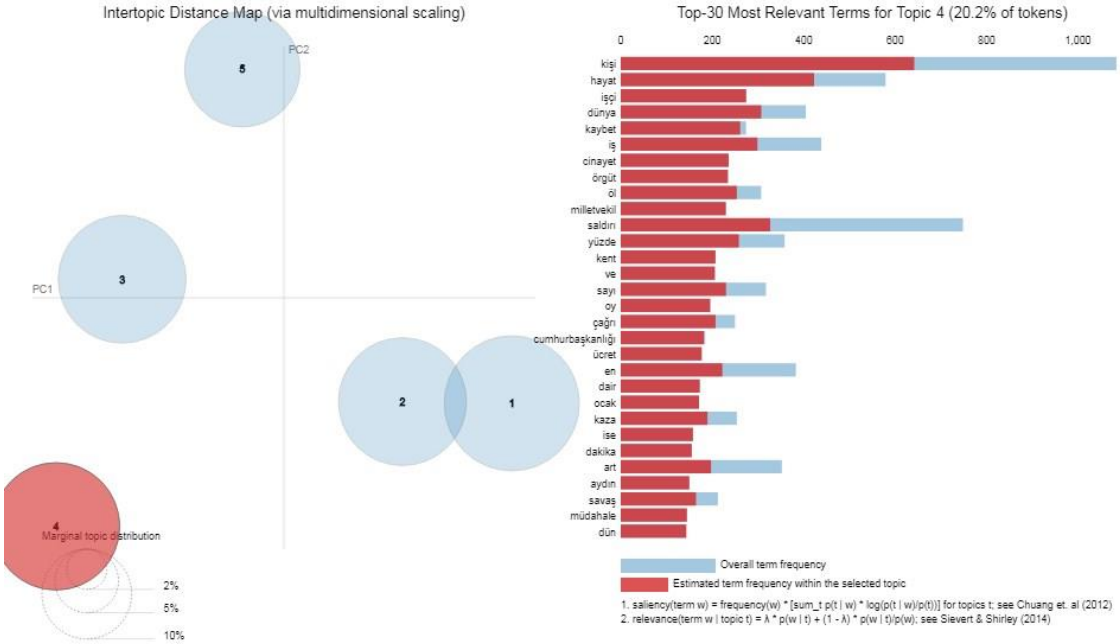
Appendix C.1. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 2 (education and economy) from the non-clickbait data.



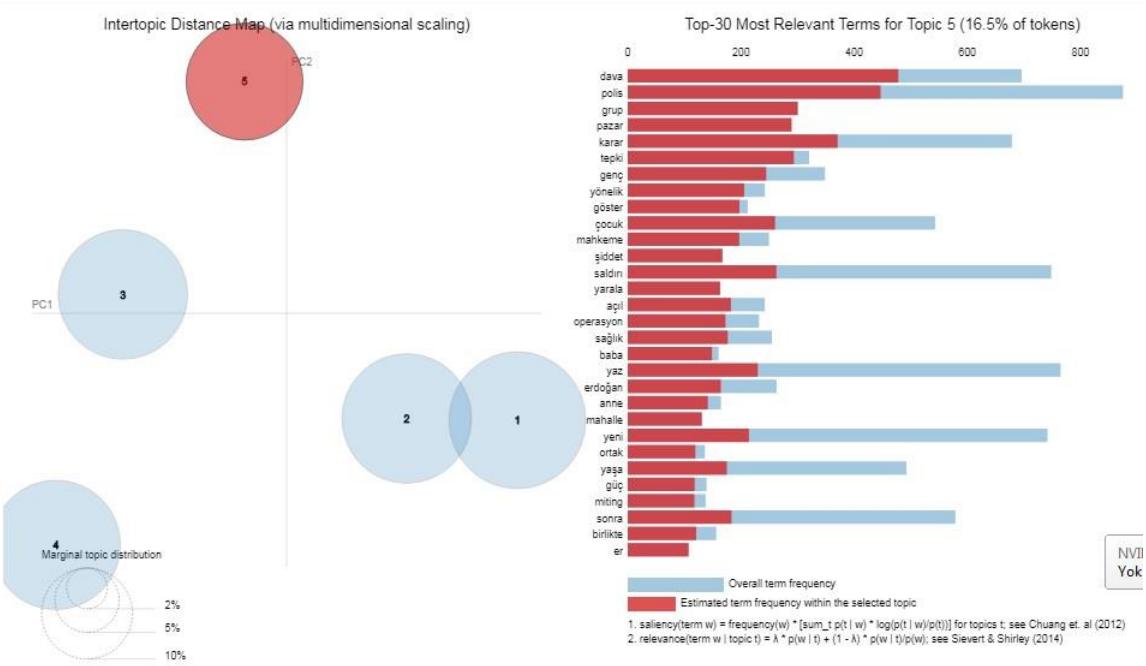
Appendix C.2. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 3 (law) from the non-clickbait data.



Appendix C.3. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 4 (security and politics) from the non-clickbait data.



Appendix C.4. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 5 (law and politics) from the non-clickbait data.



APPENDIX D

THE ORGANIZATION OF THE DATASET

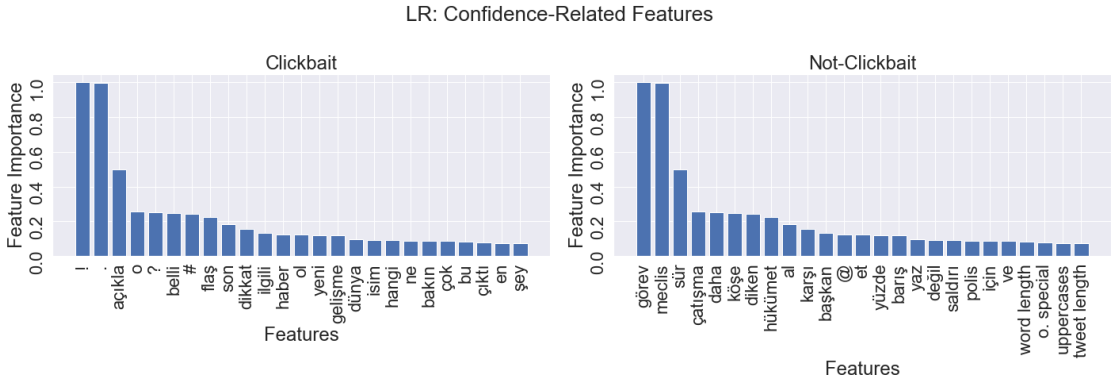
Appendix D.1. The organization of the dataset. A view of the clickbait data from the Limon Haber data

	source	tweet_id	created_at	full_text	favorite	retweet_
2	Limon	867679000(	5/25/2017 17:27	Interpol'ün aradığı İngiliz kadın, Marmaris'te bulundu	6	0
3	Limon	867801000(	5/25/2017 9:50	Bakan Elvan açıkladı! Bir hafta içinde yasalaşıyor https://t.co/9Fr4VnFuBs https://t.co/LbNp54TUDe	6	2
4	Limon	867801000(	5/25/2017 17:55	Bir otomobil devine daha dava şoku.. https://t.co/tMeJj3iF7q https://t.co/zsaz9Rt5Hn	7	1
5	Limon	121428000(	1/6/2020 20:09	#SONDAKİKA Ankara'da okullar yarın tatil mi? Vali Vasip Şahin'den açıklama geldi. https://t.co/ThrCoTcWh0	6	0
6	Limon	121172000(	12/30/2019 18:43	Kriz büyüyor! 3 diplomata sınır dışı kararı https://t.co/tVGfQ88FXc https://t.co/X6iYdEq9m4	3	0
7	Limon	868157000(	5/26/2017 17:30	Instagram'da yeni dönem başlıyor! Özel mesaj atarken artık bakın neyle karşılaşacaksınız?	5	1
8	Limon	121426000(	1/6/2020 18:40	Fotoğrafın NASA uydusu tarafından çekildiği ve Avustralya'daki yangınların anlık durumunu gösterdiği iddiası doğru..	3750	1251
9	Limon	118093000(	10/6/2019 19:39	Tehlikeli yolculuk! Hepsi yola saçıldı https://t.co/p6hDXdEf6F https://t.co/aCgiYrDaLI	3	0
10	Limon	121449000(	1/7/2020 10:12	4 yıllık bekleyiş... Anlatırken gözyaşlarına boğuldular https://t.co/OU4lWQH2fL	1	0
11	Limon	118086000(	10/6/2019 15:00	Denizli'de dehşet anları! Bıçaklanan şahıs herkesi şaşkına çevirdi… https://t.co/6HsmqoEMRH https://t.co/ULZW5shV	5	1
12	Limon	867447000(	5/24/2017 18:29	Yolcunun hostese verdiği not paniğe neden oldu https://t.co/NH7HB26gVc https://t.co/tyWSeuOWKK	6	0
13	Limon	867691000(	5/25/2017 10:37	Sabah görenler önce şaşırdı, sonra nedenini öğrendi https://t.co/Evsz4nLuni https://t.co/aB2oJochj9	5	2
14	Limon	118088000(	10/6/2019 16:14	5 yaşındaki çocuğa 'tokat' atmıştı! Cezası belli oldu	5	1
15	Limon	121426000(	1/6/2020 19:08	Ankara'da yarın okullar tatil mi? https://t.co/RbQIVCGzNI https://t.co/IVg7jLejsj	22	6
16	Limon	867825000(	5/25/2017 19:30	Ücretsiz otel, ücretsiz servis, ücretsiz internet... Bu ilçede her şey bedava! https://t.co/Ii5RZcFZkN https://t.co/Qf5Eazl	31	10

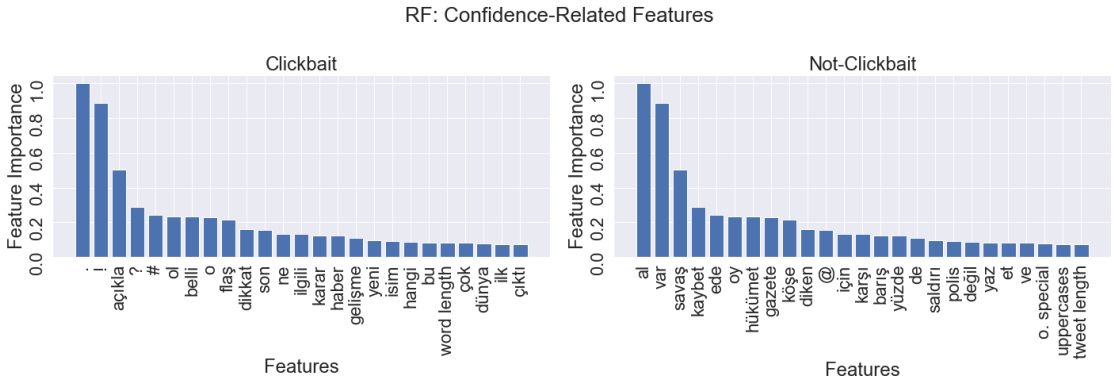
APPENDIX E

THE CONFIDENCE-BASED FEATURES OF THE MODELS

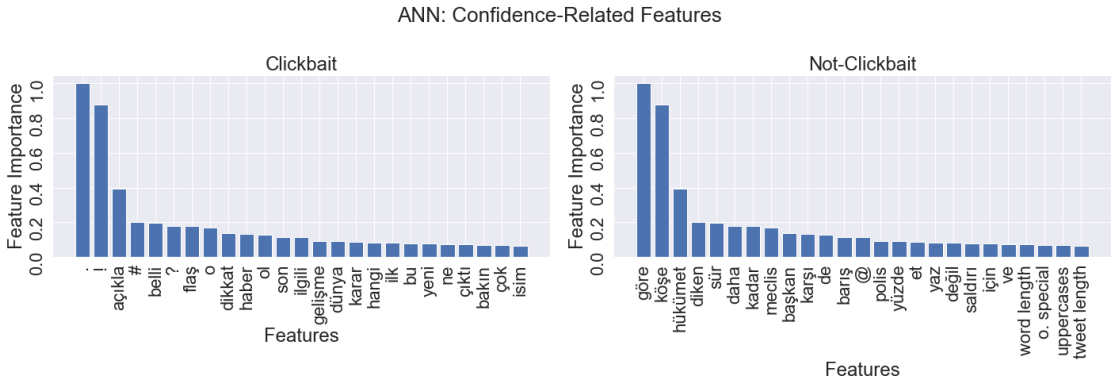
Appendix E.1. Confidence-based important features of the Logistic Regression model.



Appendix E.2. Confidence-based important features of the Random Forest model.



Appendix E.3. Confidence-based important features of the ANN model.



Appendix E.4. Confidence-based important features of the LSTM model.