


ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms

Journal of Information Science
1–20
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01655515211007746
journals.sagepub.com/home/jis


Şura Genç

Department of Cognitive Science, Graduate School of Informatics, Middle East Technical University, Turkey

Elif Surer 

Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Turkey

Abstract

Clickbait is a strategy that aims to attract people's attention and direct them to specific content. Clickbait titles, created by the information that is not included in the main content or using intriguing expressions with various text-related features, have become very popular, especially in social media. This study expands the Turkish clickbait dataset that we had constructed for clickbait detection in our proof-of-concept study, written in Turkish. We achieve a 48,060 sample size by adding 8859 tweets and release a publicly available dataset – ClickbaitTR – with its open-source data analysis library. We apply machine learning algorithms such as Artificial Neural Network (ANN), Logistic Regression, Random Forest, Long Short-Term Memory Network (LSTM), Bidirectional Long Short-Term Memory (BiLSTM) and Ensemble Classifier on 48,060 news headlines extracted from Twitter. The results show that the Logistic Regression algorithm has 85% accuracy; the Random Forest algorithm has a performance of 86% accuracy; the LSTM has 93% accuracy; the ANN has 93% accuracy; the Ensemble Classifier has 93% accuracy; and finally, the BiLSTM has 97% accuracy. A thorough discussion is provided for the psychological aspects of clickbait strategy focusing on curiosity and interest arousal. In addition to a successful clickbait detection performance and the detailed analysis of clickbait sentences in terms of language and psychological aspects, this study also contributes to clickbait detection studies with the largest clickbait dataset in Turkish.

Keywords

Data analysis; dataset formation; clickbait detection; neural networks

1. Introduction

Clickbait refers to headlines or teaser messages that arouse people's curiosity and make them click on the link. These headlines, consisting of a short introduction message about the content of the main material (news, video, article, etc.), present information that is not included in the actual content or reflects weak information about the content. Advertising is an essential income for online content, which is closely related to the number of clicks that online content receives. Using teaser messages as ads for online content is a highly preferred strategy by publishers since they allow content to become its advertisement. Teaser messages consisting of texts or images about online content, as well as a link, may attract people to see the actual content. Besides, the text in the teaser messages usually has limited information about the actual content aiming to draw people's interest. Incomplete information in these kinds of teaser messages conflicts with the reader's demand to reach the information about the content and direct the reader to the actual content. This strategy, which intends to deceive the reader and creates a contradiction between the aims of the reader and the publisher, is called clickbait [1].

Corresponding author:

Elif Surer, Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey.
Email: elifs@metu.edu.tr

Curiosity plays a critical role in the process of acquiring new knowledge, motivating people to investigate the new stimulus. Besides its importance in children's development, acquiring scientific knowledge, literature and art, the sense of curiosity has recently been manipulated for the purpose of commercial interests [2] like it is in the clickbait strategy. Berlyne [3] argued that some of the characteristics of stimuli such as novelty, complexity, surprisingness and ambiguity could explain which stimuli will evoke curiosity in people. These characteristics determine the relationship between people's internal knowledge and the external information they encounter. For example, ambiguity refers to the situations or stimuli that can have more than one interpretation due to a lack of information [2]. This introduces a gap in one's knowledge, which can be defined as the difference between what one knows about the stimulus and what one is interested in knowing. At the same time, surprisingness reflects the unexpectedness of the new stimulus. The headline 'Gunshot sounds in London, the capital of England!' is an example of ambiguity because those sounds may be originating from a drill or an attack. However, the sentence 'Attention: Fish are crossing over the street!', which is the headline of flooded news, is an example of the surprisingness of a stimulus because this situation is unexpected for us. Designating this discrepancy between one's current state of knowledge and the new information as a knowledge gap, Loewenstein [2] argues that if people are exposed to new, complex, surprising and ambiguous information, people will tend to behave in a way that they can fill this gap and resolve the uncertainty and complexity in new information. This motivation may be one of the main reasons why people click on clickbait news headlines and why they feel disappointed after reading the full content of the news. Clickbait headlines direct people to click on the link in order to reduce their curiosity; however, the content that the title directs people often does not contain information that will satisfy the curiosity created by the title.

Biyani et al. [4] found that there are eight clickbait varieties in which different strategies are used: (1) 'Exaggeration' refers to the exaggerated presentation of the content in the title; (2) 'Formatting' is the overuse of capital letters and punctuation marks in the title; (3) 'Teasing' means trimming the details of the content; (4) 'Inflammatory' indicates using rude or provocative expressions; (5) 'Graphic clickbait' refers to talking about salacious or hard-to-believe content; (6) 'Bait-and-switch' is presenting information that is not included in the original content; (7) 'Ambiguous' clickbaits give vague information about the content and (8) 'Wrong' clickbaits redirect to content that provides false information. Besides these clickbait strategies, Pujahari and Sisodia [5] detected three more categories that represent the clickbait headlines. These strategies are the following: (1) 'Incomplete' refers to that there is missing information in the title; (2) 'Headline cloning' means using the same headlines for other contents and (3) 'URL redirection' leads people to an unrelated web page.

Clickbait strategy has been widely used in news headlines posted on the Twitter accounts of news channels recently – a major problem for people who often use social media as a news source. Statistics show that 10% of users who use social media to access news were Twitter users in February 2019 [6]. The clickbait strategy can become a problem for society as well as for individuals in terms of the rapid dissemination of misinformation. Bakshy et al. [7] indicated that the word-of-mouth information shared by ordinary users is spread quickly. Also, the users, who have too many followers and have had an impact on people in the past, can be the initiators of the immense cascades of missing or false information. These results show that missing or exaggerated news with clickbait headlines can create information pollution in society. Turkey is ranked seventh based on the number of Twitter users with 8.56 million users in 2019, which indicates that clickbait usage in Twitter accounts of news channels may constitute a major problem for users in Turkey as well. It is clear that there is a need for studies on Turkish datasets about the use of clickbait in news headlines. To investigate this problem, many datasets have been created recently, and many studies have been performed on these datasets using machine learning algorithms.

Although clickbait studies have been done on datasets in various languages, most of them are in English, which is a limitation for understanding the nature of the problem. The fact that there is only one Turkish dataset [8] on clickbait is one of the main motivations of this study. In this study, we form and release a publicly available dataset – ClickbaitTR – including 48,060 samples and apply several machine learning algorithms on this dataset while analysing the results thoroughly. 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 with an F1-score of 0.86; 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 BiLSTM performs with an accuracy of 0.97 and F1-score of 0.96.

The organisation of this article is as follows: section 2 presents a detailed literature survey and provides the outline of our previous study; section 3 shows the usage examples from Turkish tweets and explains the data collection process elaborating the criteria on which the platform and news sources are selected, the properties of selected sources for clickbait and non-clickbait data and how the data are collected from these sources; section 4 clarifies how the dataset is constructed from the collected data, what properties the obtained data have and how these data are processed in order to have a structured dataset; section 5 introduces the results of four different analyses conducted for comparison using

ANN, Logistic Regression, Random Forest and LSTM algorithms as well as our previous neural network models and discusses them in detail; and finally, in section 6, the conclusion and future work can be found.

2. Background

The dataset in this study is constructed by gathering data, in Turkish, from Twitter. Twitter is one of the most popular social networks on which users can share their posts – tweets on Twitter – with a limit of 280 characters. Tweets can also include videos and photos, and the users can view the posts of other users. Statistics show that the number of monthly active Twitter users worldwide is over 275 million in 2019, and this number was 330 million in the first quarter of 2019 [6]. Twitter users spend an average of 6 min on the site [9]. These numbers are indicative of how widely Twitter is used.

In recent years, tweets posted by the public, organisations or news channels have become important data sources for a wide range of research domains such as Computer Science, Psychology, Linguistics, Sociology and commercial competition. Researchers from different fields carried out studies on specific topics using various techniques such as machine learning algorithms on Twitter data [10–12]. For example, one of the areas where researchers analysed these kinds of datasets created through Twitter is sentiment analysis. Sentiment analysis is a computational process for detecting the feelings people reflect in the text. Recently, interest in extracting subjective judgements from the text has increased, thanks to the development of natural language processing methods [13]. The datasets for sentiment analysis, being annotated as positive, negative, neutral or mixed, allow researchers to analyse the sentiment of the tweets for many purposes [14], such as analysing the sentiment of the clients [15]. In other studies, Twitter data have also been used in pharmacology and medicine to investigate drug side effects [10,16], to predict influenza-like illness [ILI] patterns in the population [17], to detect natural disasters [18] and to observe social dynamics such as refugee migration patterns [19]. Fake news identification [11] and clickbait detection [12,20] on the news are other areas of study for which tweets can be used in the analysis.

2.1. Related work

As clickbait became an important strategy encountered on almost all online platforms, various datasets have been created to detect clickbait in different kinds of data such as texts or videos by gathering data from other social media channels as well as Twitter in recent years. For example, creating a YouTube clickbait dataset, Qu et al. [21] concluded that the teaser information (title, thumbnail and the first 123 characters of the description) about videos gives inadequate information in terms of clickbaitness, which requires evaluating further details on the video. Two reviewers annotated 109 YouTube videos as clickbait and non-clickbait based on the information coming from the title, description, thumbnail and comments of the video, as well as the comments acquired by watching the video. For another dataset consisting of YouTube videos, three annotators labelled the videos by watching them, and the dataset included information about clickbaitness, title, description, thumbnail, comments and comment threads of videos [22]. Lopez-Sanchez et al. [23] developed a Case-Based Reasoning (CBR) method for automatic clickbait detection. The system they proposed can categorise the headlines as clickbait or not by adapting itself to users, which provides a strong clickbait detection on different datasets.

Clickbait strategies and features that Pujahari and Sisodia [5] detected mainly focused on the clickbait headlines. These features can be investigated in terms of novelty, complexity, surprisingness and ambiguity, which are the features of stimuli that evoke curiosity, as it is stated earlier. Graphic and wrong clickbaits can be seen as examples of novel stimuli since they arouse curiosity by presenting unbelievable, new or false information. Formatting and inflammatory clickbait strategies can be examples of complexity aspect since they provide sentences that are difficult to understand by changing the structure of expressions or by changing their textual features. However, the surprisingness aspect of stimuli that evokes interest may be provided by exaggeration. Teasing, bait-and-switch, ambiguous and incomplete clickbaits can be seen as the ambiguity aspects of attractive stimuli. Apart from these nine categories, headline cloning and URL redirection are different strategies in which nine other forms of clickbait are repeatedly produced. Investigating clickbait within this framework shows that it does not have a single mechanism, and it includes different strategies that encompass the sense of curiosity in all aspects.

As Twitter is abundant with news-related tweets, some of which are considered clickbait, it is appropriate to use this platform to obtain clickbait detection data. Potthast et al.'s [12] study on clickbait detection is one of the earliest studies tried on a dataset consisting of Twitter data of news sources using three different machine learning algorithms. They constructed a Twitter Clickbait Corpus extracting 2992 tweets (150 tweets per account) from the 20 of the most productive official news outlet accounts selected based on the number of their retweets such as BBC News, Business Insider, Huffington Post and BuzzFeed, and three people annotated these tweets as clickbait or not. Besides Twitter Clickbait

Corpus, Potthast et al. [1] also constructed a large-scale dataset, Webis Clickbait Corpus 2017, making the annotation process more detailed and meticulous. First, crowd workers, who annotate tweets for a fee, cautioned not to misclassify gossip tweets as clickbait and pay special attention to the images in the tweets. Second, for each tweet, annotators wrote down the words that most often caused them to consider that tweet as clickbait. Finally, tweets were not classified as clickbait or non-clickbait; instead, each tweet is rated on a Likert-type scale with four choices that indicate the degree of its clickbaitness. They collected 38,517 teaser messages from Twitter accounts of 27 news publishers whose importance was determined in terms of retweets. Tweets of these accounts, including text and media attachments, and except videos, were recorded on a daily basis in a 6-month period. Designing a non-binary annotation task for assessing clickbaitness of teaser messages, they provided suitable headlines for analysis.

Another dataset consisting of teaser messages (i.e. tabloid) of Twitter accounts of media organisations was constructed by Chakraborty et al. [24]. They chose the top three newspapers (New York Times, Washington Post and India Times) and one online news media outlet (Huffington Post) to collect data. They found that these news sources do not have only one Twitter account; they also have multiple secondary accounts. They collected tweets from the accounts of these four media organisations and their 38 secondary Twitter accounts for clickbait data, while they gathered non-clickbait data from 27 primary and secondary Twitter accounts of the five outlets (BuzzFeed, Upworthy, ViralNova, ScoopWhoop and ViralStories). They also collected retweets of these gathered tweets for further analysis. Consisting of 288K tweets and 11.4M retweets collected over a period of 8 months, this corpus provides a large dataset for clickbait detection in English.

While many of the studies on clickbait are about detecting and preventing it, there is also a study that plans to use the clickbait strategy to increase the reads of useful information. Bhowmik et al. [25] created a study plan for investigating whether clickbait headlines can be used for engaging readers with reliable health-related information. They plan to conduct an experiment in which the participants will be presented with both clickbait and non-clickbait headlines with health-related articles and then asked whether they want to click on the article. After the participants would read the articles, they would be asked whether they believe in the information in the article and whether they want to share it. Constructing a dataset for this specific type of clickbait, this proposed study might be an important attempt for exploring a new aspect of clickbaits.

There is only one dataset, including clickbait and non-clickbait news headlines in Turkish [8]. They collected clickbait headlines and news of those headlines from Twitter accounts and websites of BBC Turkish and Anadolu Agency, while they gathered non-clickbait data from Twitter accounts and websites of media organisations such as Hürriyet Newspaper and Vatan Newspaper. This dataset consists of 2000 news content and 2000 news headlines, which were manually labelled. The tweets of BBC Turkish and Anadolu Agency were labelled as non-clickbait since they were considered as being unlikely to be clickbait. However, the tweets of other media organisations were labelled as clickbait since they were thought of as being probably to be clickbait.

It can be seen from the literature that many studies have been conducted recently in order to detect clickbait on data from news channels and take precautions against clickbait, such as developing browser extensions that can detect clickbait and warn users. Although there are many datasets containing news headlines from English news sources, detailed and comprehensive Turkish datasets are still needed on this subject.

The classification of the Turkish news headlines in terms of being clickbait or non-clickbait has initially been studied in Geçkil et al.'s [8] study. This study presents one of the first efforts on clickbait detection in Turkish news headlines using the Twitter platform for dataset construction. There are some limitations in terms of an insufficient number of samples and proper classification of samples, which have been overcome by this study. First, in Geçkil et al.'s [8] study, the data obtained from clickbait content sources such as Anadolu News Agency and BBC Turkish were in the non-clickbait category; but in this study, tweets of these sources are in the clickbait category based on the information coming from the data obtained from Limon Haber (i.e. Lemon News in English) and Spoiler Haber (i.e. Spoiler News in English). Second, in the study of Geçkil et al. [8], clickbait tweets were collected from three different news sources, while in this study, the clickbait data were gathered from 50 different news sources which cover most of the Turkish news sources, including the top-five newspapers (Sözcü, Hürriyet, Sabah, Posta and Milliyet Newspapers) [26] and a lot of online news media outlets (Duvar, Diken and T24 Newspapers). Finally, the number of samples used in the analyses in this study (48,060 samples) is considerably higher than the number of samples in the study of Geçkil et al. [8] (4000).

2.2. Our previous study: SIU Study

The preliminary version of this study is presented at 27th Signal Processing and Communications Applications Conference (SIU) [27], and for the sake of clarity, it will be named as SIU Study from this point. In the SIU Study [27],

as a proof of concept, we have shown on a dataset of 39,201 tweets that the Multi-layer Perceptron (MLP) classifier (Hinton, 1989) performs well for clickbait detection. The clickbait data were extracted from Twitter accounts of Limon Haber [28] and Spoiler Haber [29], which identify and share misleading and intriguing news spots for users on Twitter, while the non-clickbait data were taken from Twitter accounts of Evrensel [30] and Diken [31] Newspapers which were recommended by Limon Haber. Evrensel Newspaper is the 26th newspaper of Turkey in terms of weekly circulation [32], and Diken Newspaper is an online-only newspaper.

Two different models were developed using an ANN with 39,201 samples and nine features that were found important in the SIU Study. These features were the number of hashtags (#), question marks (?), exclamation marks (!), dots (.), at signs (@) which mean mentions, the other special characters, the number of upper cases, the length of words and the length of tweets. After the feature selection process, the frequencies of these nine features were added to our vector containing the words, and each tweet in the dataset is represented uniquely in this feature vector, which has 10,329 dimensions. Two ANN algorithms were trained within the SIU Study. The first model was trained using this feature vector, and the second model was trained using only the selected nine features. The first ANN model for clickbait detection performed with an accuracy of 0.91 with an F1-score of 0.91, while the second ANN model performed with an accuracy of 0.83 with an F1-score of 0.83. The preliminary results of the first model provided the highest scores available on the Turkish dataset. In the previous study, the dataset was smaller and was not made public. Besides this, there was no available framework for public users to reuse our code.

3. Data collection

The approach we propose will be referred to as JIS throughout the remainder of this article. In this study, we expanded our dataset, constructed in SIU Study. In terms of sample size, we added 8859 new tweets (5827 clickbait and 3032 non-clickbait) and added five new analyses to compare the performances of different methods on the dataset. In SIU study, there were 39,201 tweets (18,204 clickbaits and 20,997 non-clickbaits). With 5827 new tweets from Limon Haber, 2156 tweets from Evrensel Newspaper and 876 tweets from Diken Newspaper, the number of tweets in the dataset reached 48,060 (24,031 clickbaits and 24,029 non-clickbaits). We also share the dataset and the framework code on how to use those methods in an open-source project repository. In this section, to better explain the context and the Twitter usage in Turkey, we start by briefly giving examples from the Turkish tweets, then explain our platform and publisher selection, followed by clickbait and non-clickbait resources, and finally, data extraction.

3.1. Background information on Twitter usage in Turkey and some examples

Most Turkish news sources have Twitter accounts, and daily news from these accounts are shared regularly at any time of the day. Especially the news about politics and the economy attract the attention of users, and they are frequently shared and discussed. Parallel to the news posted on Twitter, trend topics on Twitter are generally related to political and economic events such as current developments in the Turkish economy, the value of the Turkish currency, violence against women, events in the parliament and statements by politicians.

Turkish news headlines, which are non-clickbait on Twitter, are generally related to politics and the economy, and these headings contain important details of the news. For example, the following Turkish news headlines (with their English translations) about the economy and education contain detailed figures:

‘Uluslararası silah ticareti son 15 yılda 28 milyar dolardan 85 milyar dolara tırmandı’./International arms trade has risen from \$28 billion to \$85 billion in the last 15 years. ‘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 news headline, detailed information on a politician’s quote was presented:

‘Ankara belediye başkanından afet yorumu: 500 yılda bir görülebilecek bir afet, bunun önlemi olamaz’./The Ankara mayor comments about the disaster: ‘A disaster that can be seen every 500 years cannot be a measure of this’.

Another piece of news is about a famous theatre actress, and the headline gives the full name of the actress instead of using pronouns to intrigue people:

‘Tiyatronun önemli ismi Nurhan Karadağ hayatını kaybetti’./Nurhan Karadağ, an important figure in theatre, passed away.

However, clickbait news headlines are mostly related to daily events and magazines, and these headlines are either vague or contain incomplete information. For example, in the following, there are some examples from the news headlines that aim to arouse curiosity in people:

‘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! ‘Yolcunun hostese verdiği not paniğe neden oldu’./Passenger’s note to the hostess caused panic.

In the following news headlines, evoking curiosity about the rest of the news by leaving the sentences uncompleted is the main goal:

‘19 yaşındaki genç kadın bu araçta sıkıştı, yardım bahanesiyle yanına gelenler ise ...’/A 19-year-old young woman was stuck in this car, and those who came to her for help ... ‘Deney kötü bitti! Termometre kırıldı, öğretmen ve öğrenciler ...’/The experiment turned out badly! The thermometer broke; and then teacher and students ...

The other examples of the Turkish clickbait headlines call people’s attention by asking questions:

‘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? ‘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.

3.2. Platform and publisher selection

The formation of datasets, including the Turkish news headlines and the content referred to by those headlines in the literature, is important in terms of detecting language-specific patterns in the studies in which the clickbait strategy is examined. Given that this dataset would contain headlines of Turkish news sources, platforms, where news headlines could be acquired, were investigated. As a result of this examination, it was considered that social media platforms where clickbait content is frequently produced and shared are suitable for data collection. Among the social media platforms, Twitter, which has a more convenient application programming interface (API) and whose Follow relations are much more accessible, has been determined to be a convenient platform for collecting the headlines for the dataset.

Being launched in 2012, the Twitter API [33] was developed for the management of the information on Twitter such as tweets, users and direct messages, and the interaction of developers with Twitter such as posting and retweeting. In this study, a part of the dataset was collected with the help of Twitter API. Twitter imposes restrictions and limitations on retrieving data via and the use of Twitter API, which determines 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. Thus, Twitter is an appropriate platform for data collection since there are tweets from a wide variety of news sources as well as individual accounts posting daily on Twitter, and it aggregates headlines from multiple sources. This suitable environment makes Twitter one of the main sources for the creation of datasets to be created, especially for clickbait detection [1,12,24].

There are three main factors for choosing Twitter for data collection. First, the majority of Turkish news sources have an account on Twitter, and these accounts share news on Twitter regularly. Thus, news sources have excessive data in their Twitter accounts, and the amount of these data is suitable for creating a large dataset. Second, news sources can only share headlines and links of news on Twitter because there is a character limitation (280 characters) for Twitter posts. The fact that the accounts of news sources contain only news headlines and links indicates that the structure of the data to be collected is also suitable for the dataset to be created. Finally, there are long-term data in the Twitter accounts of news sources, which means that the data to be collected cover a wide range of time.

Since it is thought that the dataset with two different categories as clickbait and non-clickbait headlines will provide a ready-to-use dataset for the analyses to be performed, the sources to collect the data are selected to provide these two categories. To do so, the news headlines from four different Twitter accounts are taken for constructing a dataset. In this dataset, data taken from two of these four Twitter accounts (Limon Haber [28] and Spoiler Haber [29]) represent clickbait news headlines, while data from the other two (Evrensel Newspaper [30] and Diken Newspaper [31]) represent non-clickbait headlines.

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 Newspaper	TRT Sports
Hürriyet Newspaper	A Sports	Al Jazeera Turkish	TRT News
Hürriyet Gündem	Demirören News	Hürriyet Newspaper	Sabah Newspaper
Hürriyet Economy	BBC Turkish	Hürriyet Kelebek	BirGün Newspaper
Hürriyet Kelebek	OdaTV	Vatan Newspaper	Business HT
Vatan Newspaper	MedyaTava	NTV	T24
NTV	ABC	TV Art and Culture	A Sports
NTV Art Culture	Cumhuriyet	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	Tele1 TV	CNN Turkish	OdaTV
Habertürk Tech.	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 Culture	Posta Newspaper	Fotomaç
Sol Haber	DHA Sports	Ihlas News Agency	Ajansspor Meydan
Sabah Newspaper	Futbol Arena	Sputnik Türkiye	Cadde Milliyet
AA Sports		Anadolu Agency	Sporx
		Sol Haber	SKOR

3.3. Resources with clickbait news headlines

Although Limon Haber is not an official news source, it cites news headlines that contain clickbait and draws attention to the lack of information that causes these headlines to be evaluated as clickbait or shares the information that is left missing in the news headline in order to arouse curiosity.

The Twitter account of Limon Haber contains 26,200 tweets (24 June 2019). Clickbait news headlines shared by Limon Haber are obtained from various Turkish news sources. This account has a large amount of data that contain the headlines of many Turkish news sources, which can be seen in Table 1. All of the news headlines quoted by this account are considered as clickbaits.

However, many Twitter users who want to support the account of Limon Haber share the title of the news with this account when they find out that the content of the news does not match the title of the news they read. These users can be considered as evaluators/reviewers for determining and labelling whether the news headlines are clickbait or not. Being retweeted by Limon Haber, these headlines identified by various users strengthen the judgement that the news headlines in this account are clickbaits.

Spoiler Haber is not an official news source and warns the users against the clickbait news headlines by sharing/retweeted these headlines as the Limon Haber account does. Spoiler Haber account shares clickbait sports news headlines more often than the other clickbait headlines. Spoiler Haber's account includes 14,800 tweets (24 June 2019).

The account of Spoiler Haber also covers most of the Turkish news sources. The Turkish news sources whose news headlines are retweeted by this account can be seen in Table 1.

3.4. Resources with non-clickbait news headlines

Evrensel and Diken Newspapers are official news sources whose Twitter accounts consist of 317,000 tweets and 212,000 tweets, respectively (30 June 2019). There are two reasons for the headlines of those two news sources to be considered as non-clickbaits. The first one is that they are highly recommended by the writer of Limon Haber, who deals with detecting clickbait headlines in Turkish news, on the grounds that the news headlines of these two newspapers are rarely clickbait. The second reason is that when the Twitter data of Limon Haber and Spoiler Haber are examined, it is found that there are only a few headlines of Evrensel and Diken Newspapers in these data. Among the news headlines that Limon Haber has labelled as clickbait, only 6 of them belong to Evrensel Newspaper and 22 belong to Diken Newspaper. However, the headlines that Spoiler Haber has labelled as clickbait does not include any headlines of Evrensel and Diken

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.

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
Amount of acquired 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
Sampled tweets	48,062

Newspapers. These numbers indicate that these two newspapers rarely do clickbait and their Twitter data are suitable to be used as non-clickbait samples.

3.5. Data extraction

Clickbait headlines are gathered from the Twitter accounts of Limon Haber and Spoiler Haber, while non-clickbait headlines are obtained from the Twitter accounts of Evrensel and Diken Newspapers. The tweet data of the selected accounts were requested from the owners of the accounts since only a limited number of tweets can be reached with Twitter API [33] (approximately 3200 tweets for each account). Limon Haber and Evrensel Newspaper shared their tweet data for this study. Tweepy [34], which is a Python library for accessing the Twitter API, is used in order to access the quoted tweets in the data received from Limon Haber.

After the dataset is created, these two news sources are contacted via email, and their permissions are obtained for sharing their data. The data have an almost equal representation of clickbait and non-clickbait headlines. The sources and amounts of clickbait and non-clickbait data can be seen in Table 2, the features presented in the dataset are in Table 3, and the information about the data acquisition process can be seen in Table 4.

Twitter data of an account are prepared and sent by Twitter with the permission of that account. As it is stated before, Twitter data of Limon Haber and Evrensel Newspaper are shared by these accounts for this study. While Limon Haber's data consist of tweets and retweets, Evrensel Newspaper's data consist of only tweets since it shares its own news headlines. The data obtained from these accounts include important attributes of tweets apart from the name of the source and

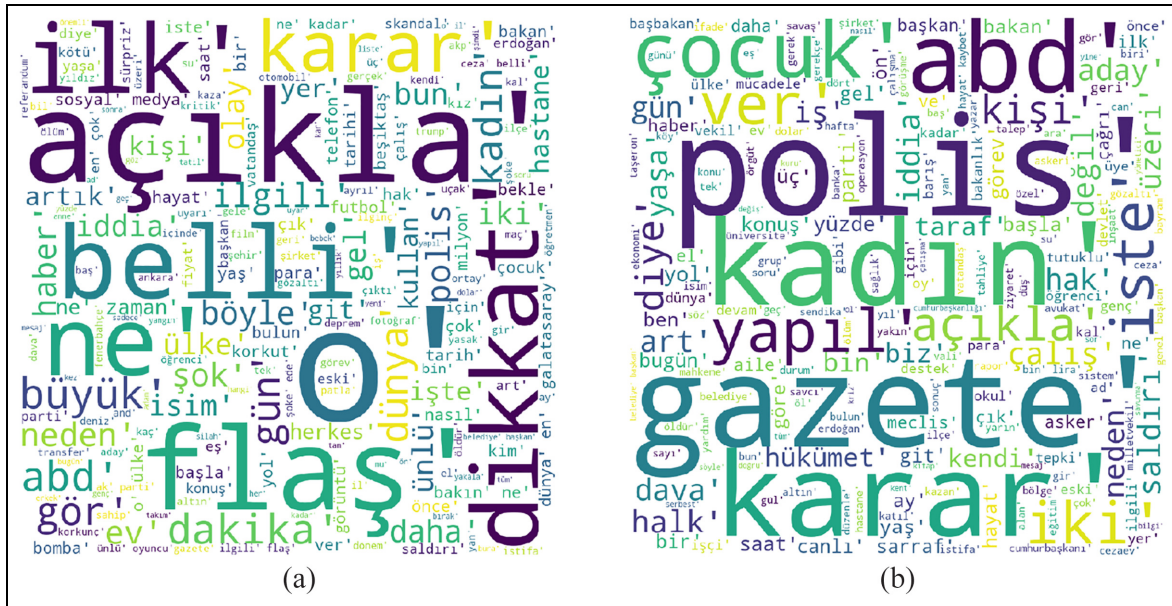


Figure 1. (a) Most frequent words in clickbait data such as explain (açıklıkla), flash (flaş), what (ne), stated (belli), she or he (o) and attention (dikkat). (b) Most frequent words in non-clickbait data such as police (polis), newspaper (gazete), woman (kadın), judgement (karar) and child (çocuk).

the date on which the post has been shared. In these data, first, it is indicated whether a headline is retweeted or tweeted by the account. Second, it is specified whether the tweet/retweet includes any hashtags, symbols or user mentions. If there are any hashtags, symbols or user mentions in the tweet/retweet, the amount of them is also specified, as well as the name and ID of the user mentioned. Third, the URL of the tweet/retweet is stated. Fourth, the text range of the tweet/retweet is included in the data, which indicates how many characters are used in that post. Finally, the information about how many likes and retweets the post has is included in the data.

As it is also mentioned before, Twitter data of Spoiler Haber and Diken Newspaper are gathered manually since the owners of the accounts could not be reached. Spoiler Haber's data consist of tweets and retweets, and Diken Newspaper's data consist of tweets, including the news headlines of this news source. Manually acquired data contain information about how many replies, retweets and likes a post has, apart from the name of the source and the date of the post. Among all these features, those included in our data are shown in Table 3.

4. Dataset construction and analysis

4.1. Scanning the collected data

Word clouds are created to examine the frequency of words contained in the dataset and to see how accurate the separation of the data as clickbait or non-clickbait is done. Two different analyses are conducted for the data containing only clickbait headlines and for the data containing only non-clickbaits in order to see the most common words in each one. The non-clickbait data include informative words such as police (polis), newspaper (gazete), woman (kadın), judgement (karar) and child (çocuk), while the clickbait data include non-specific words such as question words or pronouns such as explain (açıklıkla), flash (flaş), what (ne), stated (belli), she or he (o) and attention (dikkat). The word clouds for clickbait and non-clickbait data can be seen in Figure 1.

To better understand the data, the first five topics out of other topics are extracted from the dataset using LDA (Latent Dirichlet allocation). The LDA method [35] is used three times for the entire dataset, for the data containing only clickbait headlines and for the data containing only non-clickbaits (passes = 5, alpha = 0.01 and eta = 0.01) in order to see, separately, what topics the clickbait and non-clickbait headlines cover, as well as the general topics in all the data. First, when the results of the analysis are examined, it is seen that the dataset generally includes the following topics: (1) politics, (2) woman, child and law, (3) crime and security, (4) economy and (5) education. Politics is the best representative of the dataset among these five topics, which can be understood from the words related to politics such as president

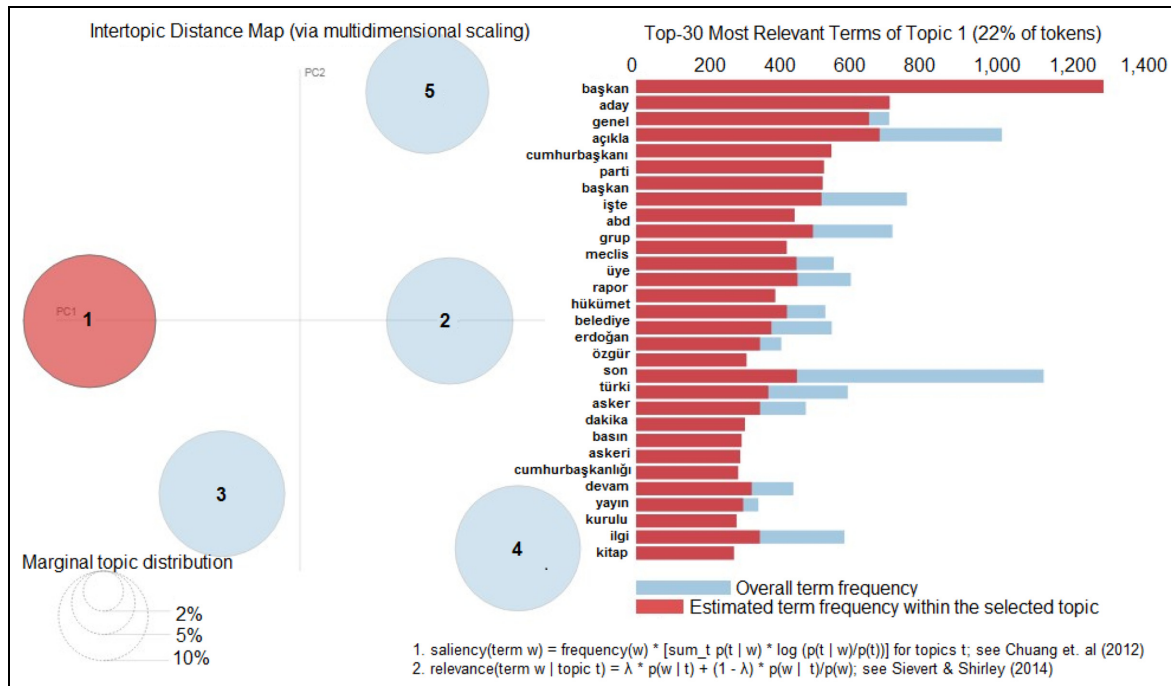


Figure 2. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from the dataset.

(başkan), candidate (aday), president of the republic (cumhurbaşkanı), parliament (meclis), government (hükümet) and municipality (belediye). The first topic and the words it includes can be seen in Figure 2.

Second, when the clickbait data are analysed, it is seen that the data include the following topics: (1) sports and football, (2) magazine, (3) politics and (4) economy. However, when the words that are relevant for the extracted topics are examined, it is seen that most of the words are used to arouse people's curiosity instead of being relevant to the topic, which makes the topics not clearly distinguishable. The words like match (maç), football (futbol) and league (lig) represent the 'sports and football' topic while the words like flash (flaş), here it is (işte), latest (son), stated (belli), explain (açıkla) and judgement (karar) reflect the profile of the clickbaits. The second topic and the word it includes can be seen in Figure 3.

Finally, as expected, certain topics can be clearly distinguished in the non-clickbait data, and these are the following: (1) politics, (2) education and economy, (3) law, (4) security and politics and (5) law and politics. The first topic and the word it includes can be seen in Figure 4.

4.2. Pre-processing

Stem/base morpheme is a morphological constituent that forms new words being attached by affixes [36]. It is the unflected part of a word that can be used with various inflectional morpheme attachments. Inflectional morphemes are affixes that do not make any changes in the parts of the speech of the stem; namely, if they are attached to a noun, the new word form will be a noun. However, they do not make any radical changes in terms of semantic relations. They do not form a new word whose meaning is totally different than the stem, and there is a regular relation between the meanings of stem and stem + inflectional affix. Inflectional affixes produce paradigms, which are word sets such as know – knows – knowing (bil – bilir – biliyor in Turkish) [36]. Since the stem of word paradigms has the same meaning, it is convenient to represent stems of words existing in the news headlines. It is thought that the words in the dataset should be separated from the inflectional suffixes and represented only by their stems since the presence of the words having the same meaning but different inflectional suffixes in the dataset would decrease the accuracy of the training process in any analysis. The stepping step combines the information gathered for similar words that may have similar meanings. In addition, it gives us a chance to reduce the dimensionality of the data without removing information. For example, in Turkish, happened (oldu), will happen (olacak), happening (oluyor) have the same root and happen (ol). All these words

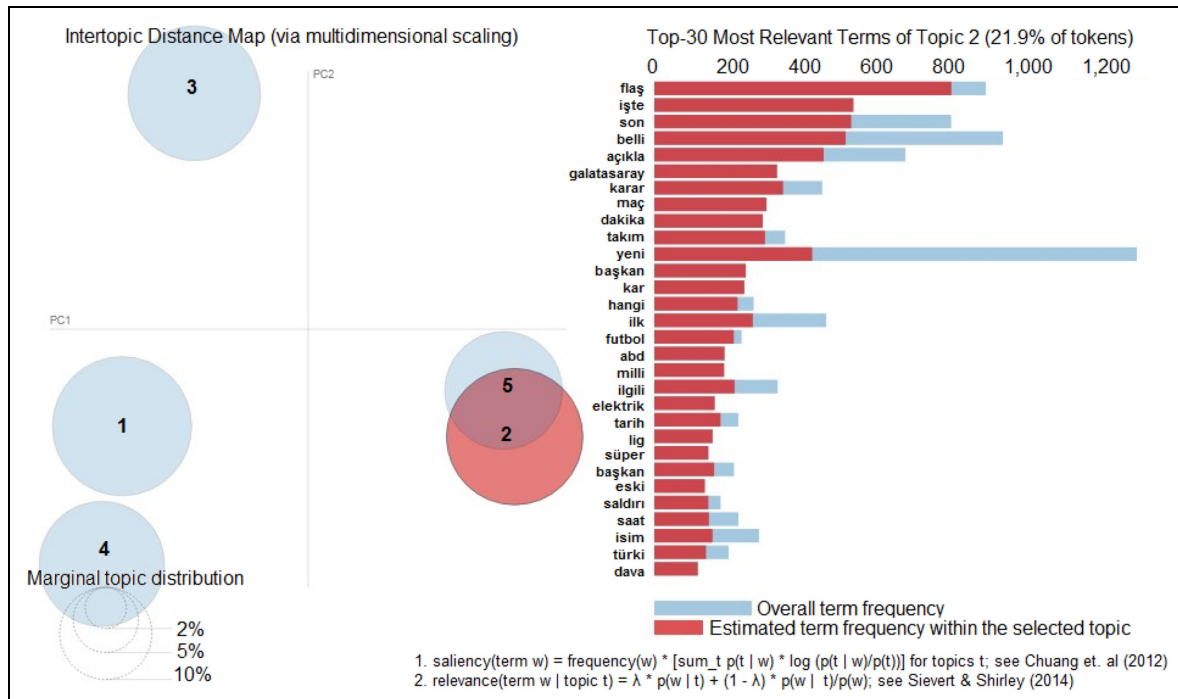


Figure 3. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 2 (magazine) from the clickbait data.

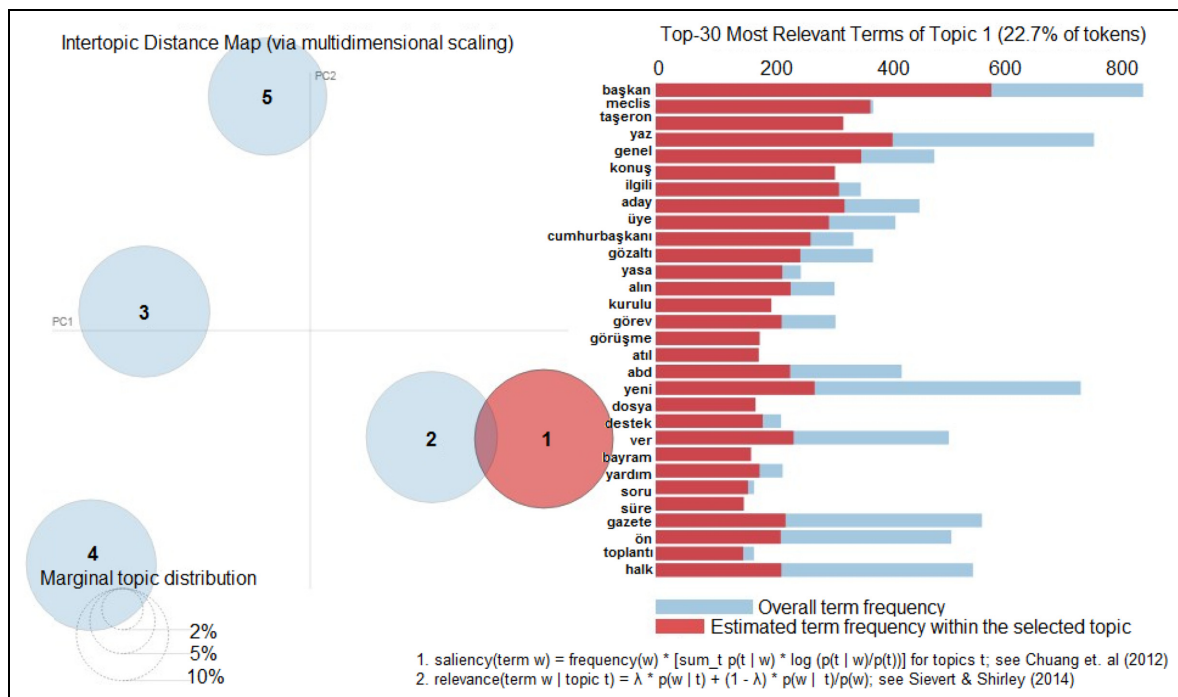


Figure 4. Intertopic distance map (via multidimensional scaling) and the Top-30 most relevant terms of Topic 1 (politics) from non-clickbait data.

tend to be used in clickbait tweets. The stemming step enables us to identify word patterns used in different forms in clickbait headlines like this happen (ol), while also increasing computational efficiency.

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

Source	Tweet id	Created at	Full text	Favourite 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 yasalaşüyor	6	2
Limon	867801404701671424	5/25/2017 17:55	Bir otomobil devine daha dava şoku	7	1

Table 6. Hyperparameters of MLP 1 and MLP 2. The values of the hyperparameters of MLP models for SIU and JIS were the same.

Hyperparameters	MLP 1	MLP 2
Number of epochs	30	50
Batch size	120	120
Hidden layer size	10	10, 5
Alpha	0.0001	0.01
Optimisation	SGD	SGD
Activation function	tanh	tanh

MLP: Multi-layer Perceptron; SGD: Stochastic gradient descent.

The ‘kelime-bol’ library of the Kalbur Project [37] developed by Ahmet Aksoy was used for this process. Kalbur Project was developed with the aim of separating Turkish words into their stems and suffixes. It is important to note that all inflectional affixes in Turkish are suffixes. The inflectional suffixes of the words in the dataset are removed, but the derivational suffixes that produce words with different meanings are left as they are.

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

4.3. Data analysis

In order to see the results of using a Machine Learning algorithm on the dataset we created, we attempted to implement six algorithms that detect clickbaits in Turkish news. Preparing the dataset for the analysis, a vector with 10,890 dimensions containing frequencies of different words of headlines in the dataset was created. Feature selection was made manually by investigating the literature. It was indicated that clickbait news uses odd punctuation patterns (i.e. !?, !!!), which increases the number of special characters they include [38]. Based on this information, it was thought that special characters might have a critical effect on the results of the analyses, and they should be treated as features. In Potthast et al. [12] study, the features for clickbait and non-clickbait tweets were determined and sorted by importance. According to this sorting, the number of hashtags (#), question marks (?), exclamation marks (!), dots (.), at signs (@) which mean mentions and the other special characters were selected as important features for this study. The feature Other Special Characters covered all of the special characters except the selected ones. Since fake news is common in social media, Hardalov et al. [39] developed a model in order to distinguish credible news from fake news. Examining the features in this study, the number of upper cases was added to the features to be used in this study. Other important features – included in Hardalov et al.’s [39] and Potthast et al.’s [12] studies such as the length of words and the length of tweets (i.e. length of the headlines) – were also selected for the study.

After selecting the features to be used in this study from Hardalov et al.'s [39] and Potthast et al.'s [12] studies, the frequencies of these nine features were added to our vector containing the words, and each tweet in the dataset is represented uniquely in this feature vector which has 10,890 dimensions. ANN, Logistic Regression [40], Random Forest [41], Long Short-Term Memory Network (LSTM) [42], Bidirectional Long Short-Term Memory Network (BiLSTM) [43] and Ensemble Classifier [44] were trained using the feature vector which has 10,890 features including nine features of clickbaits.

5. Results and discussion

Two models were trained on the subset of this dataset in the previous study [27]. The first one was trained by tweets that were represented by 10,338 features in the feature vector. The second model was trained only by nine features in order to see whether there is a positive effect of representing our tweets with word frequencies and of comparing the relative importance of the selected features. MLP classifier from Scikit-learn was used for both models [45]. The best models suggested by GridSearchCV were used in training and for each model. Before training each model, the data were divided randomly into datasets for training and testing. For the first model, 80% of data were separated for training, and 20% of data were separated for the test set. The models were trained on a dataset of 39,201 tweets. Before training, the importance of the abovementioned nine features was calculated by summing the absolute values of the connections of each input neuron to the neurons in the first latent layer. This study presents one of the first efforts to clickbait detection in Turkish news headlines using machine learning algorithms. The previous study's preliminary results show that using an ANN, and our model performs with an accuracy of 0.91 with an F1-score of 0.91, which is the highest score in Turkish datasets [27].

In this study, we expanded the dataset in terms of sample size, added 8859 new tweets (5827 clickbait and 3032 non-clickbait). The sources and new amounts of clickbait and non-clickbait data can be found in Table 2. In this study, six different analyses are carried out in order to make a comparison between models trained on the dataset we created. The libraries included in Scikit-learn [45] are used for using Logistic Regression [40] and Random Forest [41] algorithms while Keras [46] is used for ANN, LSTM [42], BiLSTM [43] and Ensemble Classifier [44] algorithm. *K*-fold cross-validation is used in the implementation of all of the algorithms. In order to apply certain statistical tests for further analysis, a 10-fold cross-validation procedure is selected. However, in order to get the important features from the neural network, we calculated the importance of each feature. We 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. To quantify the importance of each feature, we basically sum the absolute value of connections of each input neuron to neurons in the first hidden layer.

First, an MLP, which is an ANN algorithm, is implemented with the 10,890-dimension feature vector. The same hyperparameters are used as in the SIU Study defined in section 2.2. This first MLP model performs with a mean accuracy of 0.78 across 10-folds. Second, another MLP trained using the same feature vector with hyperparameters other than that in the SIU Study, and this second MLP model performs with a mean accuracy of 0.93 across 10-folds. The hyperparameters used in two MLP models can be seen in Table 6. The performances of the models can be found in Figures 5 and 6. However, feature importances obtained from the second MLP model can be seen in Figure 7.

Third, the Logistic Regression algorithm is used on the 10,890-dimension feature vector. This model performs with a mean accuracy of 0.88 across 10-folds. The receiver operating characteristic (ROC) curve of the model, which can be seen in Figure 8, shows that the model is very good at distinguishing between the true positives and negatives since the curve is far from the diagonal line. Fourth, the Random Forest model is trained using the same feature vector. It performs with a mean accuracy of 0.90 across 10-folds. Feature importance obtained from the Random Forest model can be seen in Figure 9.

Fourth, LSTM is developed using Keras, which is a Deep Learning library. As in the previous analyses, this model was trained with the 10,890-dimension feature vector, and it performs with a mean accuracy of 0.93 across 10-folds, which can be seen in Figure 10.

As the fifth analysis, BiLSTM is applied to the same feature vector using Keras, and it performs with a mean accuracy of 0.97 across 10-folds, which can be seen in Figure 11.

Finally, Bagging Classifier [44] algorithm is used to develop the Ensemble Classifier. In the Ensemble Classifier, ANN, Random Forest, Logistic Regression, LSTM and BiLSTM algorithms are used, and the Ensemble Classifier performed with a mean accuracy of 0.93 across 10-folds. A comparison of the performances of all models developed in this study (Logistic Regression, Random Forest, ANN1, ANN2, LSTM, BiLSTM and Ensemble Classifier) can be seen in Figure 12. The results of these algorithms can be seen in Table 7. The results show that the second ANN algorithm with two layers [5,10] performs better than the first ANN algorithm with one layer [10] on our dataset. Besides, this second

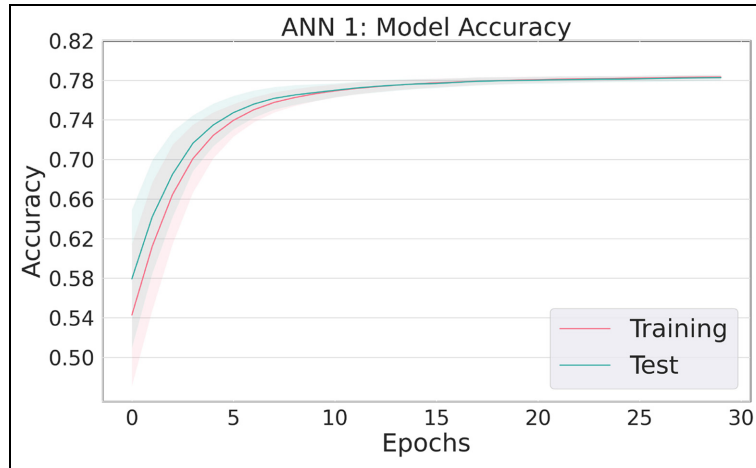


Figure 5. Model accuracy of the first MLP model of the JIS (MLP 1). The hyperparameters of this model are presented in Table 6.

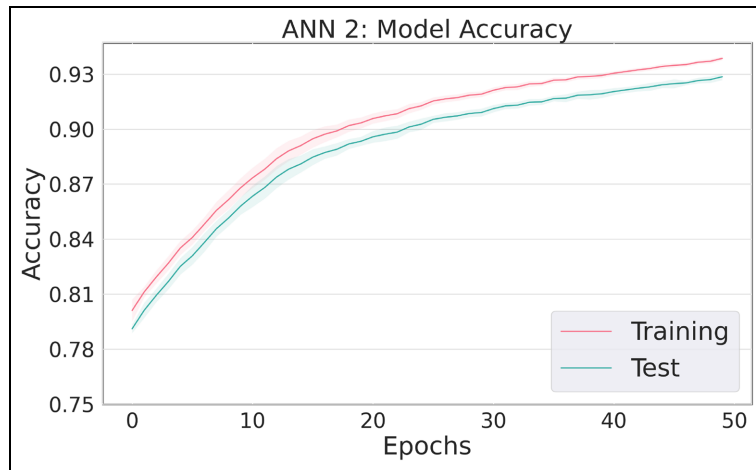


Figure 6. Model accuracy of the second MLP model of the JIS (MLP 2). The hyperparameters of this model are presented in Table 6.

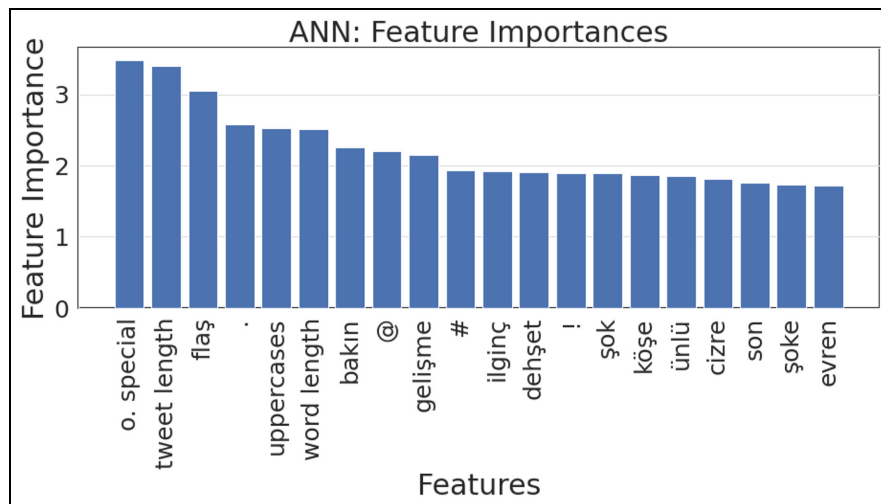


Figure 7. Feature importances obtained from the second MLP model of the JIS (MLP 2). The hyperparameters of this model are presented in Table 6.

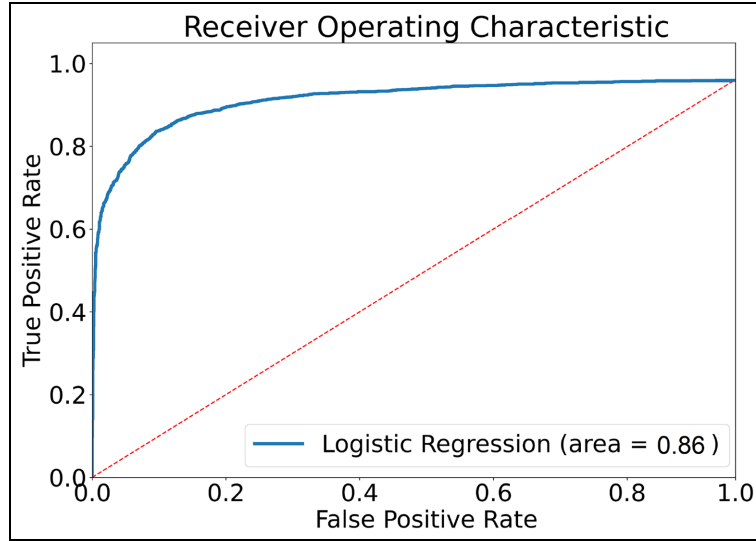


Figure 8. Receiver operating characteristic (ROC) curve of the Logistic Regression model.

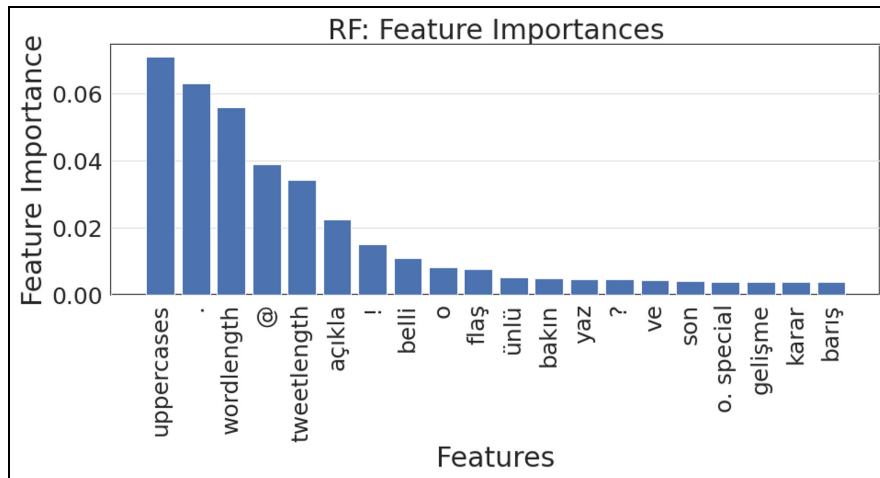


Figure 9. Feature importances obtained from the Random Forest model.

ANN has higher accuracy than the Logistic Regression and Random Forest algorithms. Finally, the BiLSTM has the best performance on our dataset, followed by the Ensemble Classifier and the second ANN.

In order to see whether there are significant differences between the performances of algorithms used in this study, the Mann–Whitney U test is applied to the results obtained for different algorithms. According to the results of the Mann–Whitney U test, there is a significant difference between ANN1 (mean = 0.783, standard deviation (SD) = 0.002), and ANN2 (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 ($p < 0.001$). However, ANN2 is significantly different than ANN1, Random Forest, Logistic Regression, LSTM and BiLSTM algorithms ($p < 0.001$). However, there is no significant difference between the performances of ANN2 and Ensemble Classifier. The Random Forest model's performance is different from the performances of ANN1, ANN2, LSTM, BiLSTM and Ensemble Classifier ($p < 0.001$) while it has the same performance as the Logistic Regression model. However, the performance of Logistic Regression is significantly different than LSTM ($p < 0.05$) and other algorithms ($p < 0.001$), except the Random Forest algorithm. The results show that LSTM, ANN2 and Ensemble Classifier have the same performance algorithms on our dataset.

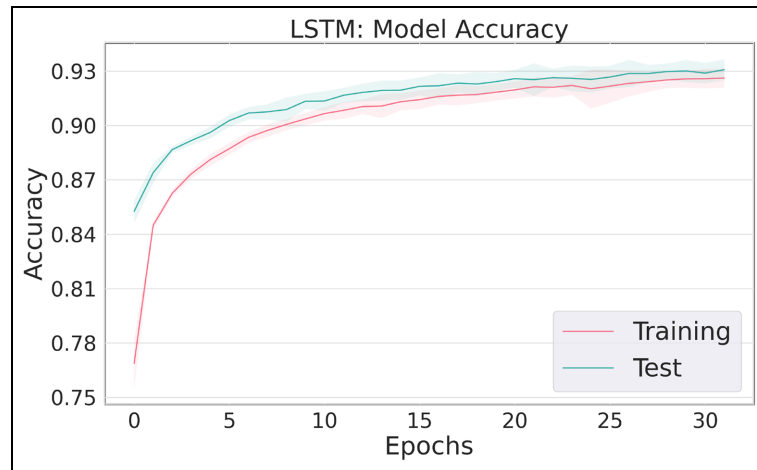


Figure 10. Model accuracy of the LSTM model.

Table 7. Results of the algorithms used in this study. All of the algorithms have a 10-fold cross-validation, a feature vector size of 10,890 and a sample size of 48,060.

Metrics	ANN1	ANN2	Logistic Regression	Random Forest	LSTM	BiLSTM	Ensemble
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

ANN: Artificial Neural Network; LSTM: Long Short-Term Memory Network; BiLSTM: Bidirectional Long Short-Term Memory.

The first MLP model of the JIS Study (78% accuracy), which is implemented with the hyperparameters in the SIU Study, performs worse than the MLP model in the SIU Study (91% accuracy). However, the second MLP model (93% accuracy) trained in this study (JIS Study) with different hyperparameters performs better than the MLP model in the SIU Study (91% accuracy). The hyperparameters of these two models can be seen in Table 6. The most important features of the MLP model of the JIS Study are the other special characters, the tweet length, flash (flaş), the number of dots (.), the number of upper cases, the length of words, look (bakın), the number of at signs (@), development (gelişme) and the number of hashtags (#), interesting (ilginç), horror (dehşet) and the number of exclamation marks (!) (see Figure 7), which shows that six of the nine features selected for this study are among the most important features of this model.

When the Logistic Regression and Random Forest algorithms are applied on a 10,890-dimension feature vector, it is seen that Random Forest has an accuracy of 86% while the Logistic Regression has an accuracy of 85% on our dataset. Looking at the most important features of the Random Forest model (see Figure 9), it can be seen that features such as 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) are distinctive for the model. These and other important features are consistent with the word clouds created to discover the details of the dataset (see Figure 1). However, 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 (93% accuracy) of the JIS Study (see Table 7).

In this study, we publicly released the ClickbaitTR – a new Clickbait dataset in Turkish – and conducted new analyses in order to make comparisons between different algorithms on our dataset. As a result of the analysis, MLP, Logistic Regression, Random Forest, LSTM, BiLSTM and Ensemble Classifier showed successful performances in detecting Turkish clickbait and non-clickbait sentences as well as the MLP classifier of our SIU Study. Besides, the updated MLP model performed better than the initial MLP trained in the SIU Study. Also, the SIU Study's MLP model performed worse on the extended dataset, which might be indicating that the hyperparameters in the SIU Study are not appropriate for the new extended dataset.

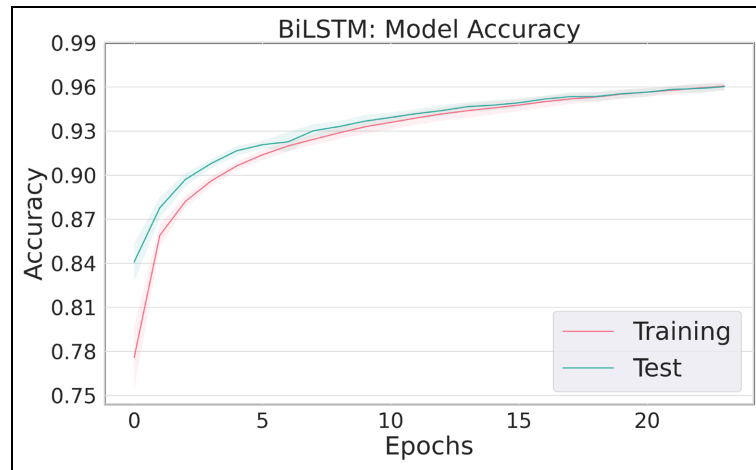


Figure 11. Model accuracy of the BiLSTM model.

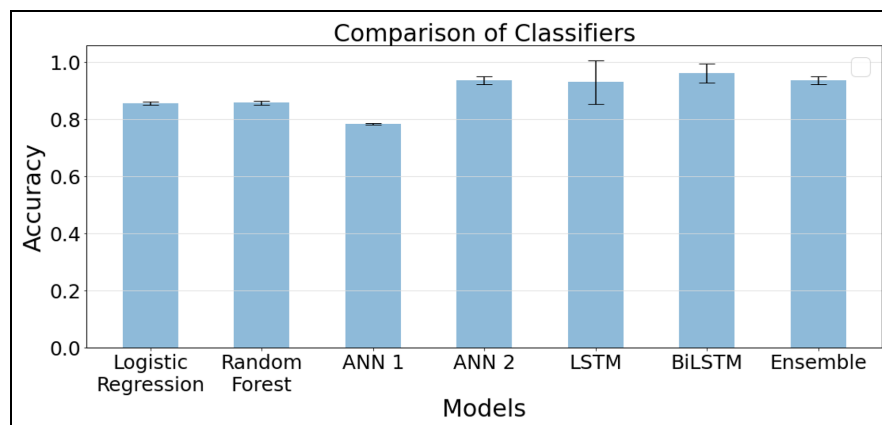


Figure 12. Comparison of performances of Logistic Regression, Random Forest, ANNI, ANN2, LSTM, BiLSTM and Ensemble Classifier.

Word clouds were created to get more detailed information about the content of the dataset used in this study, and the LDA method applied to the dataset gives important information about the general content and nature of Turkish news sources. The word clouds that show the most frequent words in clickbait and non-clickbait headlines in our dataset are compatible with the content of the Turkish news. The most frequent words of clickbait headlines such as explain (açıkla), flash (flaş), what (ne), stated (belli), she or he (o) and attention (dikkat) can be seen in the examples of the Turkish news such as ‘Is there life on the Red Planet Mars? Scientists explained ...’, ‘FLASH! Warning after warning from Meteorology!’, ‘See what it means if the skin of a lemon is thick!’ and ‘Attention to those who made a lease agreement in December: The rate of increase on the rents has been stated!’. However, the most frequent words of non-clickbait headlines such as police (polis), newspaper (gazete), woman (kadın), judgement (karar) and child (çocuk) can be seen in the examples of the Turkish news such as ‘Minister’s promise to doctor to prevent violence: Police departments are established in emergency services.’, ‘Women defend their lives in solidarity’ and ‘Three-year-old locks his father’s iPad by 2067’. These examples, which provide consistent information about the content of Turkish news headlines, are important sources for other studies on Turkish news sources.

The information coming from the word clouds and LDA analysis is in line with the studies of Biyani et al. [4], and Pujahari and Sisodia [5]. The examples for clickbait headlines above and other examples such as ‘See what really came out? The truth is different.’ or ‘These images shocked Turkey! There has just been a flash development ...’ present the different varieties of clickbait strategy which are put forward by these studies.

One of the strengths of this study is providing solutions for the limitations of the first dataset in Turkish on the number of samples and the suitability of non-clickbait samples. In this study, the data obtained from news sources such as BBC Turkish and Anadolu News Agency were evaluated in the clickbait category, contrary to previous work. Another strength of the study is that the number of samples (48,060) obtained from various news sources is higher than the number of samples (4000) in the study of Geçkil et al. [8]. Besides, while the previous dataset used examples from three news sources, this dataset contains news headlines from 52 different news sources.

Another strength of this study is that it provides the opportunity for investigating the mechanisms of the clickbait strategy in the Turkish language and an important contribution to the cross-linguistic analysis. The spread of getting news via social media makes it necessary to examine meticulously the strategies used by news channels broadcasting on social media platforms. Clickbait is perhaps the most problematic strategy 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 in different languages, and datasets constructed in different languages are critical for these kinds of studies.

However, our broad dataset, gathered from many different news sources, provides the opportunity to work on the Turkish dataset for fields such as Psychology, Linguistics, Cognitive Science and Computer Science, and it is suitable for analysing not only clickbait but also on other different topics. Since the clickbait strategy works by evoking curiosity in people, this dataset containing clickbait and non-clickbait sentences is suitable for psychology studies as well as for natural language processing. The dataset can be used for experimental studies using many machine learning algorithms to investigate concepts such as curiosity, interest and motivation. Besides, it is also suitable for examining the relationship between some experimental data to be collected from humans and computational models obtained from machine learning studies.

6. Conclusion and future work

In this study, first, we expand and publicly release a clickbait dataset in Turkish, including clickbait and non-clickbait headlines of Turkish news sources and present a very comprehensive resource – ClickbaitTR – consisting of 48,060 news headlines for further studies. As the source for data extraction, Twitter, which includes active accounts of many news sources, is preferred. This enables us to obtain a wide range of news headlines that have been posted over a wide range of time. Second, we achieve the best performance in clickbait detection by performing six different types of analyses on the expanded dataset. Finally, we contribute to the psychology of curiosity by investigating the characteristics of clickbait sentences and the topics that clickbait sentences are most relevant. In all these aspects, this study contributes to the linguistic and psychological analysis of clickbait sentences and the detection of the sentences in which clickbait strategy is used.

In future studies, different textual features such as user comments for tweets can be used as a feature in the analysis. Also, images as non-textual features in the tweets can be important indicators in distinguishing clickbait and non-clickbait headlines. In this study, we examine and discuss the clickbait in terms of curiosity and interest, but analysing other emotions may also be useful in understanding the nature of clickbait. Eight clickbait varieties categorised by Biyani et al. [4] and three clickbait varieties Pujahari and Sisodia [5] can be subjected to different analyses, and different psychology experiments can be done in order to investigate the emotional 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.

Acknowledgements

The authors thank Assistant Professor Murat Perit Çakır (METU Cognitive Science) for the final proofreading.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Elif Surer  <https://orcid.org/0000-0002-0738-6669>

References

- [1] Potthast M, Gollub T, Komlossy K et al. Crowdsourcing a large corpus of clickbait on twitter. In: *Proceedings of the 27th international conference on computational linguistics*, Santa Fe, NM, 20–26 August 2018, pp. 1498–1507. New York: ACL.
- [2] Loewenstein G. The psychology of curiosity: A review and reinterpretation. *Psychol Bull* 1994; 116(1): 75.
- [3] Berlyne DE. *Mcgraw-hill Series in Psychology. Conflict, arousal, and curiosity*. New York: McGraw-Hill, 1960.
- [4] Biyani P, Tsioutsoulis K and Blackmer J. ‘8 amazing secrets for getting more clicks’: Detecting clickbaits in news streams using article informality. In: *Proceedings of the thirtieth AAAI conference on artificial intelligence*, Phoenix, AZ, 21 February 2016, pp. 94–100. Reston, VA: AAAI.
- [5] Pujahari A and Sisodia DS. Clickbait detection using multiple categorisation techniques. *J Inform Sci* 2019; 47: 118–128.
- [6] Platform GNBD. *Statista*, n.d., <http://www.statista.com> (accessed July, 2019).
- [7] Bakshy E, Hofman JM, Mason WA et al. Everyone’s an influencer: quantifying influence on twitter. In: *Proceedings of the fourth ACM international conference on web search and data mining*, Hong Kong, China, 9–12 February 2011, pp. 65–74. New York: ACM.
- [8] Geçkil A, Müngen AA, Gündogan E et al. A clickbait detection method on news sites. In: *Proceedings of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining*, Barcelona, 28–31 August 2018, pp. 932–937. New York: IEEE.
- [9] Alexa. *The top 500 sites on the web*, n.d., <https://www.alexa.com/topsites> (accessed July, 2019).
- [10] Ginn R, Pimpalkhute P, Nikfarjam A et al. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, Reykjavik, 31 May 2014, pp. 1–8. CiteSeerx.
- [11] Atodiresei CS, Tănăsescu A and Iftene A. Identifying fake news and fake users on twitter. *Proced Comput Sci* 2018; 126: 451–461.
- [12] Potthast M, Köpsel S, Stein B et al. Clickbait detection. In: *European Conference on Information Retrieval*, Padua, 20–23 March 2016, pp. 810–817. Cham: Springer.
- [13] Lin C and He Y. Joint sentiment/topic model for sentiment analysis. In: *Proceedings of the 18th ACM conference on information and knowledge management*, Hong Kong, China, 2–6 November 2009, pp. 375–384. New York: ACM.
- [14] Saif H, Fernandez M, He Y et al. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: *Proceedings of the 1st international workshop on emotion and sentiment in social and expressive media: approaches and perspectives from AI*, Turin, 3 December 2013, pp. 1–13. CEUR-WS.
- [15] Sisodia DS and Reddy NR. Sentiment analysis of prospective buyers of mega online sale using tweets. In: *Proceedings of the 2017 IEEE international conference on power, control, signals and instrumentation engineering*, Chennai, India, 21–22 September 2017, pp. 2734–2739. New York: IEEE.
- [16] Hsu D, Moh M and Moh TS. Mining frequency of drug side effects over a large twitter dataset using apache spark. In: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining*, Sydney, NSW, Australia, 31 July–3 August 2017, pp. 915–924. New York: IEEE.
- [17] Achrekar H, Gandhe A, Lazarus R et al. Predicting flu trends using twitter data. In: *Proceedings of the 2011 IEEE conference on computer communications workshops*, Shanghai, China, 10–15 April 2011, pp. 702–707. New York: IEEE.
- [18] Hernandez-Suarez A, Sanchez-Perez G, Toscano-Medina K et al. Using twitter data to monitor natural disaster social dynamics: a recurrent neural network approach with word embeddings and kernel density estimation. *Sensors* 2019; 19(7): 1746.
- [19] Hubl F, Cvetojevic S, Hochmair H et al. Analyzing refugee migration patterns using geotagged tweets. *ISPRS Int J Geo-Inform* 2017; 6(10): 302.
- [20] Zhou Y. Clickbait detection in tweets using self-attentive network, 2017, <https://arxiv.org/abs/1710.05364>.
- [21] Qu J, Hißbach AM, Gollub T et al. Towards crowdsourcing clickbait labels for YouTube videos. In: *Proceedings of the HCOMP 2018 works in progress and demonstration papers*, Zürich, 5–8 July 2018, pp. 1–4. CEUR-WS.
- [22] Shang L, Zhang DY, Wang M et al. Towards reliable online clickbait video detection: a content-agnostic approach. *Knowl Based Syst* 2019; 182: 104851.
- [23] Lopez-Sanchez D, Herrero JR, Arrieta AG et al. Hybridizing metric learning and case-based reasoning for adaptable clickbait detection. *Appl Intel* 2018; 48(9): 2967–2982.
- [24] Chakraborty A, Sarkar R, Mrigen A et al. Tabloids in the era of social media? Understanding the production and consumption of clickbaits in twitter. In: *Proceedings of the ACM on human-computer interaction*, Portland, OR, 6 December 2017, pp. 1–21. New York: ACM.
- [25] Bhowmik S, Rony MM, Haque MM et al. Examining the Role of Clickbait Headlines to Engage Readers with Reliable Health-related Information, 2019, <https://arxiv.org/abs/1911.11214>
- [26] Tirajları G. *Weekly Circulations*, n.d., <http://gazetetirajlari.com> (accessed August, 2020).
- [27] Genç Ş and Surer E. Detecting ‘clickbait’ news on social media using machine learning algorithms. In: *Proceedings of the 2019 27th signal processing and communications applications conference (SIU)*, Sivas, 24–26 April 2019, pp. 1–4. IEEE.

- [28] @LimonHaber. *Limon Haber*, n.d., <https://twitter.com/LimonHaber> (accessed August, 2020).
- [29] @spoilerhaber. *Spoiler Haber*, n.d., <https://twitter.com/spoilerhaber> (accessed August, 2020).
- [30] @evrenselgzt. *Evrensel Newspaper*, n.d., <https://twitter.com/evrenselgzt> (accessed August, 2020).
- [31] @DikenComTr. *Diken Newspaper*, n.d., <https://twitter.com/DikenComTr> (accessed August, 2020).
- [32] Radar M. *Circulation Report for the week of 17 August–23 August*, n.d., <https://www.medyaradar.com/tirajlar> (accessed August, 2020).
- [33] API T. *Use Cases, Tutorials, Documentation Twitter Developer*, n.d., <https://developer.twitter.com> (accessed July, 2019).
- [34] Tweepy. *An easy-to-use Python library for accessing the Twitter API*, n.d., <https://www.tweepy.org> (accessed July, 2019).
- [35] Rehurek R and Sojka P. Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 workshop on new challenges for NLP Frameworks*, Valletta, 22 May 2010, pp. 45–50. University of Malta.
- [36] Akmajian A, Farmer AK, Bickmore L et al. *Linguistics: An introduction to language and communication*. Cambridge, MA: MIT Press, 2017.
- [37] Aksoy A. *Kalbur Project*, 2017, <https://github.com/ahmetax/kalbur> (accessed February, 2020).
- [38] Chakraborty A, Paranjape B, Kakarla S et al. Stop clickbait: detecting and preventing clickbaits in online news media. In: *Proceedings of the 2016 IEEE/ACM international conference on advances in social networks analysis and mining*, Davis, CA, 18–21 August 2016, pp. 9–16. New York: IEEE.
- [39] Hardalov M, Koychev I and Nakov P. In search of credible news. In: *Proceedings of the international conference on artificial intelligence: methodology, systems, and applications*, Varna, Bulgaria, 7–10 September 2016, pp. 172–180. New York: Springer.
- [40] Yu HF, Huang FL and Lin CJ. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach Learn* 2011; 85(1–2): 41–75.
- [41] Breiman L. Random forests. *Mach Learn* 2001; 45(1): 5–32.
- [42] Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9(8): 1735–1780.
- [43] Schuster M and Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Pr* 1997; 45(11): 2673–2681.
- [44] Sisodia DS. Ensemble learning approach for clickbait detection using article headline features. *Inform Sci Int J Emerg Transdiscipl* 2019; 22: 31–44.
- [45] Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–2830.
- [46] Chollet F. Keras: Deep Learning library for Theano and TensorFlow. *Data Sci Cent* 2015; 7(8): T1.