TCSA
https://dergipark.org.tr/tr/pub/tcsa
Vol.1 No.1
May 2024: 1-1
Accepted 16 May 2024
Published 6 June 2024

Transactions on Computer Science and Applications

Detecting misinformation on social networks with natural language processing

Masis Zovikoğlu [a], Uzay Çetin [*,a,b]
[a] Galatasaray University, Dept. of Computer Engineering, Istanbul, Türkiye, 34349,
[b] ORCID: https://orcid.org/0000-0002-0784-253X.

*** Corresponding author email address**: uzay00@gmail.com, **phone:** +90 5340580606

Abstract

In this study, we conduct a data-driven study to detect misinformation in social media. Our aim is to apply natural language processing (NLP) techniques to detect fake news in Turkish. To this end, we have found a publicly accessible English dataset of fake news articles, consisting of 20,800 samples and translate it into Turkish. We have applied sentence-transformer models to vectorize our textual content. Then we simply applied Logistic Regression algorithm for fake news detection with different inputs. Our observations indicate that the title of a news article holds greater significance than its content when it comes to the detection of fake news. However, enhanced detection performance can be attained through the combined utilization of both the title and content. Interestingly, our findings reveal that the removal of stopwords does not lead to improved accuracy. We also discuss that the more advanced transformer-based approaches would offer superior performance, particularly in scenarios characterized by data drift. But we leave it for future work.

Keywords: Machine Learning, NLP, Data Mining, Text Classification, Fake News Detection

1. **Introduction**

We live in the information age. Huge quantities of digital information are constantly circulating around the world (Cetin et al., 2023). All the technological infrastructures we have created and the networks to which we are connected make the dissemination of information faster and easier than ever. The crucial point is that in this information age, misinformation spreads easily alongside information (İhsan et al., 2022). Often the former prohibits the latter. As a result, access to real information becomes difficult.

There is a growing literature on fake news detection. Fake news detection is an old problem related to text categorization. There are plenty of fake news detection techniques (Kaur et Ranjan., 2022) using machine learning algorithms such as Naive Bayes, SVM. A comprehensive survey of deep learning methods for fake new detection can be found in the article (Hu et al., 2022). For more information about the taxonomy of fake new detection, please refer to the article of (İhsan et al., 2022). Another recent article (Hu et al., 2024) investigates the diffusion structure of fake news spread.

The Internet has no mechanism for verifying the content of the data that circulates. This is exactly the reason why we must build better socioscopes (Cetin and Gundogmus 2018). A socioscope is the computational device that help us to examine the digital social universe (Cetin and Gundogmus 2022). It is difficult to speak of a correlation between the prevalence of information and its accuracy. This is where the proliferation of hoaxes comes in. Social media accounts play a key role in spreading malicious information (Nasir, 2021). Bot accounts are the source of a substantial amount of information on the web. Real users can also create malicious information. We are talking here about click bait, which acts as an intermediary to attract clicks on a piece of content. Clickbait can be classified by using deep learning and information divergence (Oliva, 2021)

All these problems reduce the Internet's capacity to disseminate real knowledge. In addition, freedom of expression, which is a fundamental motto of the internet, is undermined. Because there is too much noise around, meaningful sounds are difficult to hear. As far as the social aspects are concerned, the internet is likely to be used as a tool to create public opinion and manipulate people for commercial or political ends.

The ethics of the latter are the subject of another debate. From a computational point of view, we will propose algorithms for detecting misinformation and bot accounts and measure the usefulness of our work in this area. In the light of the reasons given above, we have decided to carry out a project on the detection of disinformation from a computer science point of view.

1.1 Data Collection

Data is the principal source for any machine learning project. But it is harder than one might thing to come with an original data source. There are several methods of obtaining data. Firstly, we considered using the web, in particular Twitter, to create our dataset. Web scraping is widely acknowledged as an efficient and powerful technique for collecting big data (Zhao, 2017). Python's 'snscraper' module is an appropriate tool for this task. We have written a snippet of this module that successfully performs this function. We were able to collect recent tweets by specifying a hashtag. However, due to the winds of change at the top of the company, scraping Twitter wasn't available all the time. So, we decided to look for an existing dataset. We found an English dataset on Kaggle. Fake News Dataset is available online, https://www.kaggle.com/c/fake-news/data. The dataset was made available as part of a community prediction competition to build a system to identify unreliable news. The file contains 20800 news items in the following format:

id: unique identifier for an article

title: title of the article

author: author of the article

text: the text of the article; it may be incomplete

label : a label indicating that the article is not reliable.

1: unreliable

0: reliable

The distribution of the number of labels is as follows, almost balanced:

1 : 10413

0 : 10387

1.2 Data Translation

The shortage of datasets in low-resource languages has compelled researchers to employ innovative methods to tackle this problem. One widely adopted solution is utilizing language translation services to replicate existing datasets from resource rich languages such as English. (Ghafoor, 2021). Given that there are more sources of data in English, we assumed that the Turkish dataset could be obtained by translation. Reasonably speaking, translating a text does not change its true or false status. We were therefore able to translate the English dataset while retaining the correct labels. Once we had the English dataset, we decided to translate the data into Turkish using the Google translate API. Python has a 'googletrans' library suitable for this task. We created a code snippet to perform the translation. As we had a large dataset, the translation process created an excessive number of requests to the server, which led to timeouts. We had to adjust the code to perform the translation in batches. After completing the translation process, we obtained a labelled Turkish fake news dataset containing news worldwide.

1.3 Vectorizing Textual Data

Vectoring is the process of transforming raw data into a format that can be easily processed by algorithms. It involves representing the data in the form of vectors, which are arrays of numbers that capture the relevant characteristics or attributes of the data. In Python, we have the "SentenceTransformer" framework for calculating sentence embeddings. We imported "emrecan/bert-base-turkish-cased-mean-nli-stsb-tr" from https://huggingface.co. This is a sentence transformer template. It maps sentences and paragraphs to a dense 768-dimensional vector space and can be used for tasks such as clustering or semantic search (Reimers and Gurevych, 2019). The model was trained on Turkish-translated versions of the NLI and STS-b datasets, using example training scripts from the sentence-transformers GitHub repository.

1.4 Ethical Concerns About Data Collection

Scraping the web to collect news data may raise ethical concerns. The collected data can be considered personal information. However, such data is more newsworthy rather than sensitive information about individuals. Plus, data has already been published on public platforms by the individuals. Another vulnerability is that people who spread false news may be tagged and discredited as a result of data collection. The standard way is to check a website’s robots.txt file before scraping. Following the guideline within that file is best practice to avoid legal and ethical issues. Our scrapers also respect robots.txt files of web sites.

2. Building Model

In the context of text classification, logistic regression appears to be a prevalent choice due to its simplicity and remarkable effectiveness, especially in scenarios characterised by binary classification tasks. The basic premise involves associating vectorised representations of news content with their corresponding target labels. This methodological approach is in line with the basic principles of logistic regression, which has been widely used in various academic works, including a notable Turkish news classification project (Bozuyla, Ozcift, 2021).

2.1 Choosing Input Data

In the process of shaping our model, we've carefully considered the input data selection. Within our dataset, we're presented with "title," "content," and "author" columns. While the "author" field might typically lack meaningful context, it holds potential importance in instances where certain authors are known for propagating false information. However, our project's primary objective is to cultivate a model that not only comprehends content but also derives underlying trends from the provided information. To avoid oversimplification, wherein author names are merely associated with labels, we've opted to exclusively utilize the "title" and "content" columns as our model's input. This decision rests on the notion that headlines often encapsulate significant details, and detecting shared patterns within titles could offer insights into the identification of potentially misleading content.

It was possible to use title and content columns separately or create a third input column by combining them.

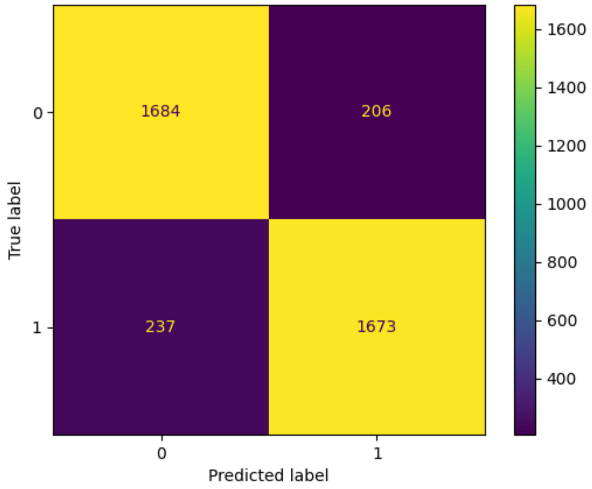
The combined input is obtained by merging title and content columns as text. The input is treated as a single text and vectorized by using sentence transformers.

3. Results

We first ran a logistic regression model with our dataset of 19,000 entries. We divided the dataset into two parts, training and testing, with an 80/20 ratio. We calculated the training and testing accuracies of the model.

3.1 Logistic Regression Trained with News Content Only

We first carried out the model training by only using news content data. In order to measure the model performance, at the end of the training, we extracted the confusion matrix over the test data. The confusion matrix can be seen in Fig.1. The purple parts show the number of inaccurate predictions. The model misses a certain proportion of fake news and assigns the fake value to some reliable news. The hit and error rates are nearly evenly distributed for both label values.



,

Fig. 1. Confusion Matrix of Logistic Regression Model - Content Only

We created and plotted a learning curve to examine in more detail how the model performance changes according to the size of the training data, in Fig.2. The curve illustrates the trade-off between training accuracy and validation accuracy. Converging curves show that after a certain level, performance can no longer be improved simply by adding more data. A large disparity between training and test accuracy indicates overfitting. While it is expected that tarin accuracy would be higher than test accuracy, the fact that the gap is closing indicates that we are refraining from overfitting.

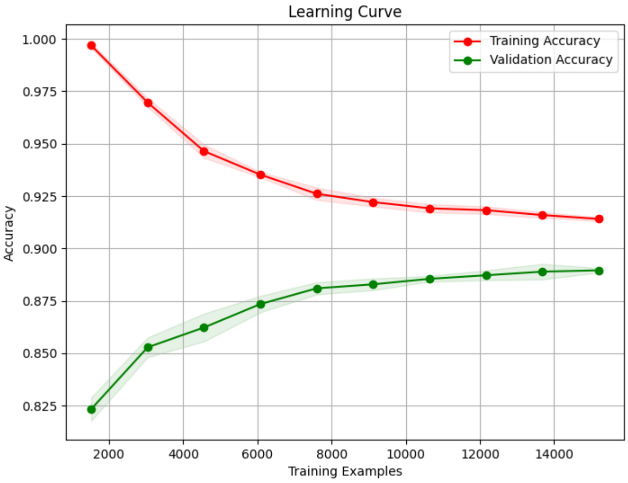


Fig. 2. Learning Curve of Logistic Regression News Content Only

3.2 Logistic Regression Trained with News Title Only

We then carried out the model training by only using news title data, in Fig.3. In order to measure the model performance, at the end of the training, we extracted the confusion matrix over the test data. The titles, representing a small portion of the textual data of a news article, were enough to obtain satisfying results at the end of the training.

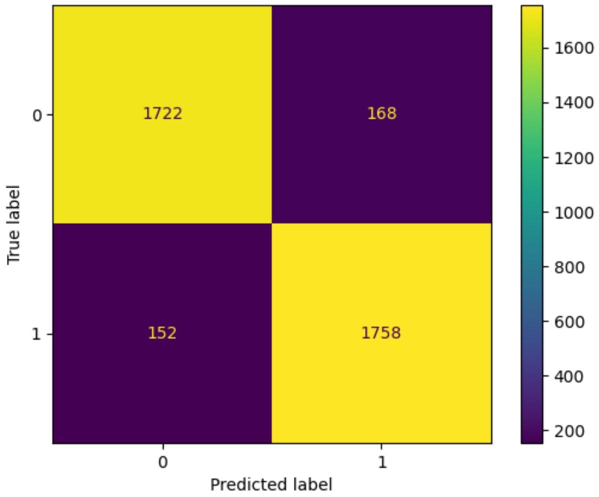


Fig. 3. Confusion Matrix of Logistic Regression Model - Title Only

The learning curves for training and validation accuracies are much better in Fig.4. compared to Fig2. These findings are consistent with the idea that the title represents the content of the news article and the alleged event. In addition, features such as sensationalised words and editorial style help to identify fake news by its headline.

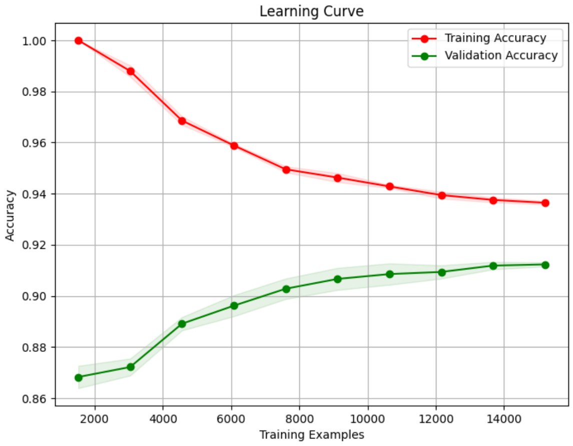


Fig. 4. Learning Curve of Logistic Regression News Title Only

3.3 Overall comparison of models with different input data of training.

We can resume the performance analysis of the different models we trained so far. Among them the logistic regression model trained by combining news title and content and keeping the stop words in the textual data, gives the best result so far. The confusion matrix can be seen in Fig.5. We have seen that title is more important than the content of the news for fake news detection. But we can achieve better results by using titles and content together. And surprisingly our results show that removing stopwords is not good for better accuracy.

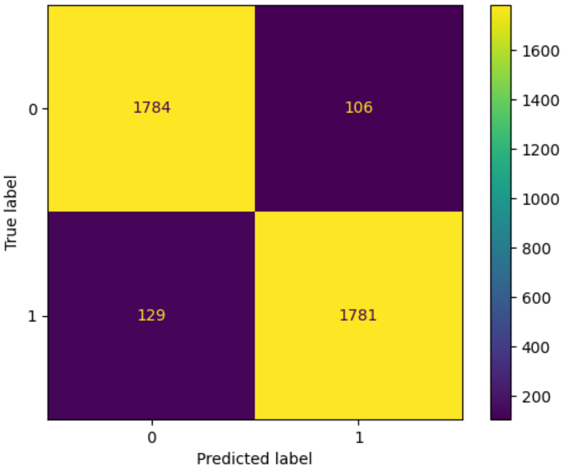


Fig. 5. Confusion Matrix of Logistic Regression Model Title+Content

The learning curves for training and validation accuracies can be seen in Fig.6. We can conclude that, the best performance for fake news deteciton, is obtained while using both title and content. When preprocessing data, stopwords should be removed only if they don't add any new information for the problem (Ghag, 2015).

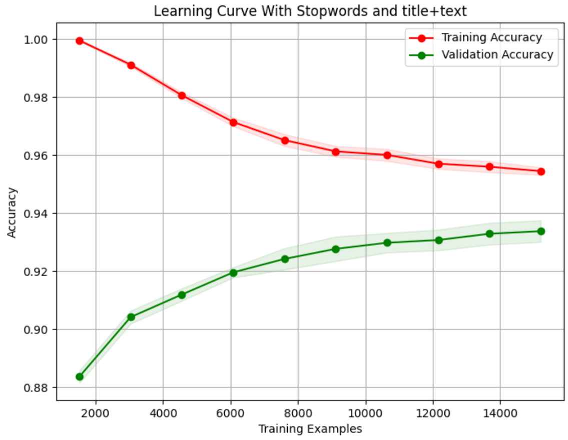


Fig. 6. Learning Curve of Logistic Regression Title+Content

For example, for a theme classification problem stopwords can be insignificant. In our case, small nuances in language can give a clue about the likelihood of a news being fake. This is because fake news can reveal itself in style as well as content. The overuse of stopwords can constitute a common pattern for fake news or reveal an unnatural narrative created to manipulate the reader.

We are now interested in a model that uses both the title and the content of a news item. This method has produced the best results so far, with an accuracy of 0.9381 as shown in Table 1. Using the title and the content all together results in better accuracy compared to the case where either one is used separately.

Table 1
Results of LR with translated news data.

Accuracy LR	With Stopwords	Without Stopwords
Title	**0.9157**	**0.9013**
Content	**0.8834**	**0.8581**
Title + Content	**0.9381**	**0.9103**

4. Future Work

So far, we have obtained a model which can classify news as fake or real and return a reliability percentage as well. In practice, a percentage score of how trustworthy a news article is may be more useful than a mere categorisation of fake and real news. We deployed our model and created a simple user interface so that the models return a prediction when users provide a news article.

We measured models' performance and obtained satisfying accuracy scores when tested with our dataset. However, when we tested the model with different data sources, such as articles from websites about different dates, countries, agendas, the predictions were not accurate all the time. The models were returning very high or very low reliability scores for the given news article. Plus, the categorisation accuracy wasn't satisfying.

This is a common issue in machine learning tasks, called data drift (Mallick et Al. 2022). Data drift occurs when a model is successful on a historical data, but its performance diminishes when used with real world data. This is because the distribution of the input data varies. The news article dataset represents a specific agenda, time, and region of the world. So, the machine learns a specific agenda depending on the dataset. So, the model parameters are insufficient to make predictions about a new agenda.

Feature selection (Cetin and Gundogmus 2019) and anomaly detection (Cetin and Tasgin 2020) methods can help to deal with data drift problem. But it needs further investigation. We can improve model performance and overcome the data drift problem by using different datasets or building a hybrid dataset which represents news from different agendas and countries. It is possible to fine tune our model with new data or to use different algorithms such as BERT (Vaswani et al., 2017). BERT has shown satisfactory results on Turkish news classification tasks. (Ozcelik 2021)

Overall, our work has shown that automating fact checking tasks with machine learning model is a viable concept.

Acknowledgments

This study was accomplished as part of Masis's intern at Summarify (https://summarify.io/) during his degree in computer engineering at Galatasaray University. We are deeply grateful to all the dedicated members of the Faculty of Engineering.

References

Çetin, U, Aslantaş, S, Gündoğmuş E, (2023). Challenges and Opportunities Related to Data Drift Problem in Sentiment, 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, pp. 86-90, doi: 10.1109/UBMK59864.2023.10286687.

Ihsan A., Nizam Bin Ayub M., Shivakumara P., Fazmidar Binti Mohd Noor N, (2022). Fake News Detection Techniques on Social Media: A Survey, Wireless Communications and Mobile Computing, vol. 2022, Article ID 6072084, 17 pages, https://doi.org/10.1155/2022/6072084

Sufanpreet K, Sandeep, R, (2024). Comparative Analysis of Supervised and Unsupervised Machine Learning Algorithms for Fake News Detection: Performance, Efficiency, and Robustness.

Hu, L., Wei, S, Zhao, Z, Wu, B, (2022). Deep learning for fake news detection: A comprehensive survey. AI Open, 3, 133-155.

Hu, B., Mao, Z., Zhang, Y. (2024). An Overview of Fake News Detection: From A New Perspective. Fundamental Research.

Cetin, U, Gundogmus, YE, (2018). A Glimpse to Turkish Political Climate with Statistical Machine Learning. In 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 537-541.

Cetin, U, Gundoğmuş, YE, (2022). A Glimpse to the Digital Social Universe in the Times of War. In 30th Signal Processing and Communications Applications Conference (SIU), pp. 1-4.

Nasir, JA, Khan, OS, Varlamis, I, (2021). Fake news detection: A hybrid CNN-RNN based Deep Learning Approach. International Journal of Information Management DataInsights,1(1),100007. https://doi.org/10.1016/j.jjimei.2020.100007

Reimers, N., & Gurevych, I, (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Bozuyla, M, Özçift A, (2022). Developing a fake news identification model with advanced deep language transformers for Turkish COVID-19 misinformation data https://journals.tubitak.gov.tr/cgi/viewcontent.cgi?article=3818&context=elektrik

Oliva, C, Palacio-Marín, I, Lago-Fernández, LF, & Arroyo, D, (2022). Rumor and clickbait detection by combining information divergence measures and deep learning techniques. Proceedings of the 17th International Conference on Availability, Reliability and Security. https://doi.org/10.1145/3538969.3543791

Ozcelik, E, Ozcelik, NS, (2021). Comparison of Traditional Classifiers and BERTurk for Fake News Identification in Turkish. Journal of Computer Science and Engineering, 9(2), 164-174. https://dergipark.org.tr/tr/download/article-file/1973019

Cetin, U, Gundogmus, YE, (2019, September). Feature selection with evolving, fast and slow using two parallel genetic algorithms. In 2019 4th International Conference on Computer Science and Engineering (UBMK), pp. 699-703.

Mallick, A, Hsieh, K, Arzani, B, Joshi, G, (2022). Matchmaker: Data drift mitigation in machine learning for large-scale systems. Proceedings of Machine Learning and Systems, 4, 77-94.

Cetin, U, Tasgin, M, (2020). Anomaly detection with multivariate K-sigma score using Monte Carlo. In 2020 5th International Conference on Computer Science and Engineering (UBMK), pp. 94-98.

Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Gomez, AN, Polosukhin, I, (2017). Attention is all you need. Advances in neural information processing systems, 30.

Ghafoor A. et al, (2021). "The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing," in IEEE Access, vol. 9, pp. 124478-124490, doi: 10.1109/ACCESS.2021.3110285.

Ghag KV, Shah, K, (2015). "Comparative analysis of effect of stopwords removal on sentiment classification. International Conference on Computer, Communication and Control (IC4), Indore, India, pp. 1-6, doi: 10.1109/IC4.2015.7375527.

Zhao, B, (2017). Web scraping. Encyclopedia of big data, 1.

Author Biographies

Masis Zovikoglu, born in Istanbul in 2000, completed bachelor's degree of computer engineering at Galatasaray University in Turkey. He spent 1 year at University of Twente, in the Netherlands, as an exchange student pursuing bachelor's degree in computer science. He started a master's degree of computer and data science in University of Lumiere Lyon2 in France in 2023.

Uzay Çetin received his master's degree in artificial intelligence from Pierre and Marie Curie University (PARIS VI) and his doctoral degree in complex systems from Bogazici University. He has joined to Galatasaray University Computer Engineering department as a faculty member in 2021. He organizes free machine learning (ML) courses to young people in Sarıyer Municipality (http://bit.ly/yapayzeka04) and he works as a ML consultant with different firms. He also organizes multi-disciplinary workshops on complex systems and data science (http://bit.ly/kahve2019). Uzay Çetin is interested in computational social science, data science, artificial intelligence, complex systems and complex networks.