ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

PROFILING OPTIMAL MACHINE LEARNING MODEL TO IDENTIFY FAKE NEWS FROM SOCIAL MEDIA PLATFORMS

[1]Bahaa Eddine Elbaghazaoui, [2]Mohamed Amnai, [3]Youssef Fakhri and [4]Abdellatif Semmouri
[1,2,3]Laboratory of Computer Sciences, Faculty of Sciences Kenitra, IbnTofail University, Morocco
[4]Faculty of Sciences and Techniques, Lab. TIAD, Sultan Moulay Slimane University, Beni Mellal, Morocco

ABSTRACT:
The massive amounts of information and data released by social media have aroused more interest due to the multiple application areas they can use. In fact, misinformation and disinformation campaigns are becoming more common around the world. The explosive growth of fake news and its erosion has increased the need of detection and intervention. Therefore, identifying fake news on online social media is crucial to allow users knowing the information truth and not fall into rumours and misinformation. This paper considers the text format fake news detection methods, and at the same time elaborates the existence and reasons of fake news. In this paper, our main purpose is to identify the optimal machine learning model for profiling fake news from social media posts content. However, to achieve this goal, we need to compare many powerful machine learning models on a huge dataset, then we could identify the optimal model to profile fake news. In any case, the results that we have obtained are very important and with very high precision.

Keywords: Fake news, Machine learning, Social media, Information Diffusion.

1 INTRODUCTION

Social media users can promote ideas or spread news by sharing, liking, or reposting. Therefore, they will always be exposed to a kind of uncontrollable information, especially news from independent authors. However, the increase in fake news is becoming a worldwide problem (McGonagle, 2017). Although fake news is not new, but now it is worrying because the popularity of social media allows the interaction and dissemination of new ideas (Zhou, 2018).

With the data grow in online dissemination of misinformation, fake news detection has recently attracted a large interest from the public and researchers. Fake news is a type of propaganda where disinformation is intentionally spread through news outlets and/or social media outlets. Given that the propagation of fake news can have serious impacts such swaying elections and increasing political divide.

The widespread dissemination of fake news will have a serious negative impact on individuals and society. Firstly, fake news can break the authenticity balance of the news ecosystem (Balmas, 2014). Secondly, fake news deliberately persuades consumers to accept biased or false beliefs. Fake news is often manipulated by propagandists to convey political information or influence (Banko, 2007). In addition, fake news has changed the way people interpret and respond to real news (Journell, 2016). To help alleviate the negative impact of fake news both the public and the news ecosystem it is vital that we develop methods to automatically detect fake news on social media.

So far, the computational methods and approaches for fake news detection has relied on satirical news sources and fact-checking sites (Journell, 2017). However, fact-checking websites are usually limited to specific areas of interest, such as politics, and require human expertise. Therefore, it is difficult to obtain data sets that provide a certain degree of generalization in multiple fields.

In this paper, our main objective is to profiling the suitable machine learning model to identify fake news from online social media platforms. However, our system starts with the collection of necessary data and clean up the ambiguities and errors from our database, then and depending on the data that we have extracted, we used machine learning models to train our system to identify fake news quickly and in real-time.

By summarizing all previous thoughts, the structure of this paper is as follows. First, we outline the literature review and our main objective approach. Second, we introduce a survey on fake news spreading. Third, we illustrate the fake news with machine learning and data profiling. Then, we present our implementation. Finally, we give a global conclusion and the proposals future works.

2 LITERATURE REVIEW

In this section we describe the most important previous work related to fake news concept. This is preceded by a review of the growing number of studies that examined fake news sharing.

The emergence of the World Wide Web and the rapid spread of social media platforms (such as Facebook and Twitter) have paved the way for information dissemination that has never been seen in human history (Aldwairi, 2018). With the use of current social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading and have nothing to do with reality.

Scholars conceptualize fake news in many ways, but the meaning is almost the same. McGonagle (2017) describes fake news as deliberately fabricated information that is spread to mislead individuals into accepting lies and false facts (McGonagle, 2017). Duffy et al. (2019) Classify fake news as any information that imitates legitimate news stories but has false and misleading content (Duffy, 2019). Currently, fake

news represents as untrue information, including myths, rumors, conspiracy theories, hoaxes, and deceptive or wrong content that is intentionally or unintentionally spread on social media networks (Wang, 2019).

Perhaps the most illustrative real example is the 2016 presidential election and Donald Trump's candidacy (Journell, 2017). In fact, some media such as The Onion or The National Enquirer, TV shows, etc. usually use various aspects of the political world, either exaggerating them or creating fictional stories to make fun of today's political issues. Although these types of fake news may affect the political stance of Americans (Brownell, 2017). We can also take the example of 2020, when Facebook does not allow its fact-checking for trump to evaluate his content and instead created a label meant to refer people to credible election information as shown in the following figure 1.

The most important thing for users is that if a page or an account often produces fact-checked content. Some social media platforms such Twitter and Facebook have special policies for posts from political content. They will take punitive measures, including reducing the overall visibility of the page or block all their permissions for utilization.

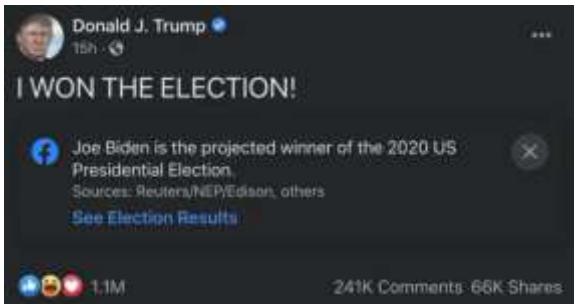

Figure 1: Facebook screenshot: direct fact-check for trump post.

In addition, in the era of the COVID-19 pandemic, research on the proliferation of fake news is emerging. Some studies have tried to recognize the connection between social media and misinformation in this pandemic era (Hou, 2020). Recent research shows that the most compelling sharing of fake news that is harmful to health in recent months is the COVID-19 pandemic (Pennycook, 2020). For example, users on social media are concerned about the impact of the coronavirus on people and countries. Specifically, a large posts and blogs number about the death toll related to the coronavirus have been published. In addition, many publications have mentioned the emotional and psychological the coronavirus impact. Users on social media may express fear and pressure for COVID-19, as well as the lack of vaccine treatments or specific antiviral treatments to prevent it. However, due to the widespread dissemination of misinformation and conspiracy theories, the sensational use of social media may pose a huge challenge to public health and the response to the epidemic. The contagious outbreak of "fake news" and "distorted evidence" in the digital world can cause large-scale panic and cause and consequences devastating in the real world, distorting evidence and hindering the response of health care workers and the public health system efforts and activities.

In this article, our main objective focus on profiling fake news from any database. Through this analyze, we will be able to predict fake news and misinformation due to a suitable machine learning model. However, and due to this study, we will be able to detect fake news and inform users to not be a victim of herd politics.

3 PROBLEMATIC

After the news is released, a large number of users may engage in its propagation over online social media networks. The user can create his thoughts on the news. Specifically, for each piece of news, many news posts can be observed and collected on social media platforms. Note that many of these unverified posts are fake. So far, attempts to exploit news propagation for fake news detection applied many features. The most of researches are released for the fake news detection with supervised manner (Julio, 2019) and few of them cared on unsupervised models (Monti, 2019). In addition, the aforementioned methods focus on identifying fake news base on single machine learning model, and use it to build a special framework. In contrast, in this paper, the key idea strives to address the problem of fake news detection in many models and identify the suitable based on time and accuracy

3 SPREADING FAKE NEWS

Figure 2: The main axis of research on fake news

Fake news is spread through social and mainstream electronic media in the form of comments, political agendas, news articles, rumors and satire (Zhao, 2020). It is widely used for misleading information, false persuasion, confusion and making it a major threat to the public's trust in online activities (social community activities, online shopping, and active media reinforcement). Due to the dynamic and heterogeneous nature of the task, natural language processing (NLP) researchers have provided multitude solutions to this problem. The diversity, speed, quantity and time delay of fake news articles are the four basic problems encountered by current academic researchers as shown in the following figure 2.

Fake news may start out as disinformation (something deliberately fabricated for a specific purpose) and end up as misinformation (false content accidentally shared by people who don't know the information is inaccurate).

The major goals behind the creation of fake news may be:

- Mishearing or misleading from real news
- Made to make money or click on the website
- Aims to promote individuals, political parties or opinions
- Misunderstandings from jokes or parody posts are regarded as facts.

If this information matches our own beliefs, we are more likely to treat it as fact because of confirmation bias. Even if this is not something we usually believe in, the shocking nature of the content may affect our emotions and make us share it.

4 DATA PROFILING AND FAKE NEWS

The core business of technology companies involves the design of identity based on data collection and data profiling (Elbaghazaoui, 2021) systems. This means that even the most resourceful news organizations can hardly deliver news through platforms such as Facebook and Twitter. However, Facebook can create audiences as well as their demands. Technology companies have most of the data needed to fully understand how information reaches its audience. This is equivalent to contemporary data blind spots, and is a factor in the erosion of trust between news organizations and the general public. If vital facts cannot be communicated to the majority of the public, then we cannot effectively have the information reality. For example, in politics we could not apply democratic protection to prevent corruption, deception and special interests.

The usefulness of data profiling on rumors and fake news is to know the information value, and also know the relevant tool which can identify the fake news from social media. However, Data profiling help to build a system to automatically detect misinformation in news. In addition, metadata features such as account name, image, followers, number of likes, tags, and release date can provide activity frequency signals as clues to suspicious activities.

The web contains data in various formats. News published online in unstructured formats such that textual news and articles, image, video, and audio. It is relatively difficult to detect and classify because it requires human expertise. However, computing techniques. However, data profiling can be used to detect anomalies that distinguish deceptive text articles from fact-based articles (Shu, 2017). It can also involve analyzing the spread of fake and real news. More specifically, data profiling could analyze how the dissemination of fake news articles on the Internet is different from that of real news (Vosoughi, 2018).

5 MACHINE LEARNING AND FAKE NEWS

IA, Machine learning and data science work for any interdisciplinary problem and usually distributed in multiple sub-disciplines that the data analysis community has organized. Fake news has incited distrust of the media, politics and established institutions around the world (Deepak, 2021). Although new technologies such as artificial intelligence (AI) may make things worse, it can also be used to combat misinformation. For example, deep fakes programs (AI systems that adapt audio, pictures and videos) make people say and do things they never did. These technologies could launch a new fake news era and spread misinformation through social media platforms.

Unfortunately, the generation and spread of fake news is driven by economic and ideological interests. Measures to combat fake news are also unlikely to prevent vested interests from finding new ways to spread fake news and other forms of false information (Tandoc, 2021). This is why research tends to focus on understanding the factors that make individuals easily misled by fake news, because these factors are often used by those behind the production and spread of fake news.

Only by knowing the complete story of the subject can one know that the news is fake. This is a daunting task, because most people don't know the complete story, they just start to believe in fake news without any verification. Many studies have focused on the detection and classification of fake news on social media platforms such as Facebook and Twitter (Vosoughi, 2018) (Allcott, 2017). At the conceptual level, fake news is divided into different types, then the knowledge is expanded to generalize machine learning (ML) models in multiple fields (Conroy, 2015) (Jwa, 2019)

By using machine learning, fake news can be detected easily and automatically (Khan, 2019) (Vedova, 2018) Once someone posts fake news, the machine learning algorithm will check the content of the post and detect it as fake news. Several researchers are trying to find the best machine learning classifier to detect fake news (Kurasinski, 2020). The accuracy of the classifier must be considered because if it fails to detect fake news, then it may cause harm to different people. The accuracy of the classifier depends on the training of the classifier. A model trained in a good way can provide higher accuracy. There are different machine learning classifiers that can be used to detect fake news.

6 EXPERIENCE

In order to apply our proposition, we chose the ISOT fake news dataset. However, the data set contains two types of articles fake news and real news. This data set is collected from real world sources. The real articles are obtained by crawling articles from "Reuters.com" (news website). For fake news articles, they were collected from unreliable websites flagged by Politifact (a fact-checking organization in the United States) and Wikipedia. The data news contains different types of articles on many topics, but the most of the articles are focused on political and world news topics (Hadeer, 2018) (Hadeer, 2017)

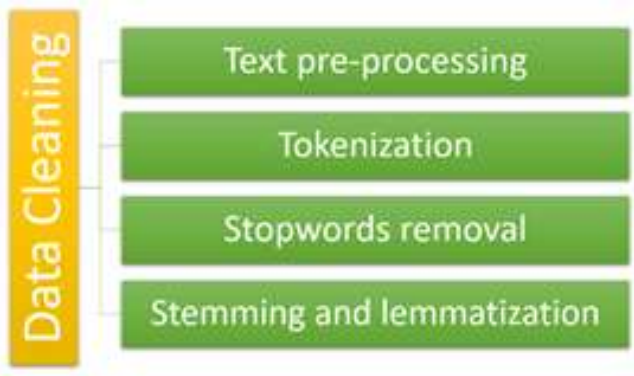

Figure 3: Cleaning process

Due to this database, we obtained a file containing the title, content, creation date and the new truth (real or fake). Indeed, according to this database, we coffle all inputs and analyze it accordance to the process shown in the figure 3 below:

In the first step, we must consider every new content. However, we start the preprocessing by removing all the noise in the text (links, numbers, emojis, special characters, etc.), and we limit the acceptable characters to 26 letters in the alphabet by removing the number form and links (Http / Https).

After that, we convert each text into Tokens "Tokenization". For example, the sentence "the earth is round" will be associated with the words one by one, resulting in the following forms ("the", "earth", "is", "round"), where each element of this list is one of the sentence Token.

This is not enough, because the terms "the" and "is" do not reflect the subject of our sentence, but the most important thing is to add free complexity to our model, and we chose a simple solution. In fact, we have to delete all the empty words "STOPWORDS", so the same sentence becomes: ("Earth", "Round").

Usually when searching for the text of a keyword, it helps if the search returns variations of that word. For example, a search for "boat" may also return "boats" and "boating". Here, "boat" will be the stem of [boat, boater, boating, boats]. This is called stemming.

Contrary to stemming, morphological restoration does not only consider word simplification, but considers the complete vocabulary of the language to perform morphological analysis on words. The lemma of "was" is "be" and the lemma of "mice" is "mouse".

After cleaning up our database, moving now to the machine learning application on our data. Indeed, we take as input the result of words extracted from news texts and the news validity as output (fake or real). According to our program that we have written with python language and with the same machine, we obtain the following table 1 as result:

Table 1: Models comparison for fake news detection

Model	Accuracy %	Duration s
LINEAR SVC	99.43	21.2639
LOGISTIC REGRESSION	98.89	38.6611
MULTINOMIAL NAIVE BAYES	93.66	22.0979
BERNOULLIS NAIVE BAYES	95.84	22.7570
GRADIENT BOOST	99.61	159.8904
XGBOOST	99.59	64.7849
DECISION TREE	99.55	41.5180
K NEAREST NEIGHBOR	90.43	101.2192
Aggressive Classifier	99.38	20.4679

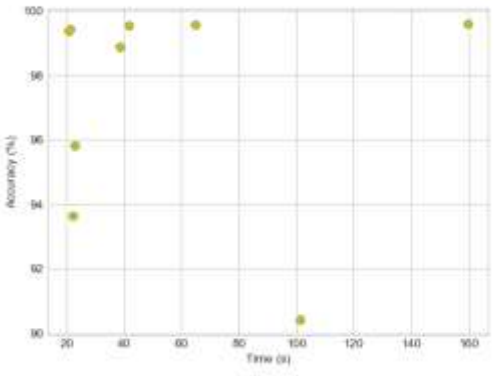

Figure 4: Models accuracy over execution duration

Depending on the results we have obtained, we find that the most reliable model is the Gradient boosting to predicts the fake news classification than other models. Indeed, each module took a duration to be executed as shown in the following figure 4 and the previous table 1.

We cannot say that Gradient boosting is the most optimal model. We find that the other models have an approximate accuracy with the best solution and with very short time. However, regarding the time execution, we think it is not a good measurement because of using different architecture. We can accelerate and enhance the performance time by using GPUs or adding cores to CPU.

7 CONCLUSION

The main background of this work is to combine the four basic tasks in big data. Data collection and storage, analysis and graphical representation, learning and prediction, and finally the processing of text data. However, throughout this project, we profiled fake news from social media based on machine learning models. Through our experience, we compare several models to determine the best model for our needs. Finally, we find that the most reliable model is Gradient boosting to predict fake news with a precision of 99.61%. In future works, we plan to add more data to our models (like images, voice messages, etc.). Also, classify news by brands, this will help us to make adaptable suggestions for each new.

REFERENCES

Tarlach McGonagle. "Fake news: False fears or real concerns?". Netherlands Quarterly of Human Rights, 35(4), 203209. (2017).

Zhou, X., & Zafarani, R. “Fake News: A Survey of Research, Detection Methods, and Opportunities”. arXiv preprint arXiv:1812.00315, academia.edu (2018).

Meital Balmas. “When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism”. Communication Research, 41(3):430-454, (2014).

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. “Open information extraction from the web”. IJCAI’07: Proceedings of the 20th international joint conference on Artifical intelligence. (2007).

Alessandro Bessi and Emilio Ferrara. “Social bots distort the 2016 us presidential election online discussion”. First Monday, 21(11), (2016).

W Journell. "Fake news, alternative facts, and Trump: Teaching social studies in a post-truth era". Social studies journal, (2017).

Duffy, A., Tandoc, E., & Ling, R. “Too good to be true, too good not to share: the social utility of fake news”. Information Communication and Society, 0(0), 115. (2019).

Wang, Y., Mckee, M., Torbica, A., & Stuckler, D. "Social Science & Medicine Systematic Literature Review on the Spread of Healthrelated Misinformation on Social Media". Social Science & Medicine, 240(September), 112552. (2019).

Judi Brownell. "Listening: Attitudes, principles, and skills". 6th Edition. Communication Studies, Humanities. (2017).

Hou, Z., Du, F., Jiang, H., Zhou, X., Lin, L., Assessment, T., & Commission, N. H. Assessment of public attention, risk perception, emotional and behavioural responses to the COVID-19 outbreak: social media surveillance in China. (2020).

Pennycook, G., McPhetres, J., Zhang, Y., & Rand, D. "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention". PsyArXiv [Working Paper], 124. (2020).

M Aldwairi, A Alwahedi. "Detecting fake news in social media networks". Procedia Computer Science, Elsevier. (2018).

Julio C. S. Reis et al. “Supervised Learning for Fake News Detection”. IEEE Intelligent Systems. Volume: 34, Issue: 2, March-April 2019.

F Monti et al. "Fake news detection on social media using geometric deep learning". arXiv:1902.06673v1 [cs.SI] 10 Feb (2019).

P Deepak, T Chakraborty, C Long. "Data Science for Fake News: Surveys and Perspectives". The Information Retrieval Series. Springer. (2021).

EC Tandoc. "Tools of Disinformation: How Fake News Gets to Deceive". Disinformation and Fake News. Springer. (2021).

S. Vosoughi, D. Roy, and S. Aral, "spread of true and false news online" Science, vol. 359, no. 6380, pp. 11461151, (2018).

H. Allcott and M. Gentzkow. "Social media and fake news in the 2016 election". Journal of Economic Perspectives, vol. 31, no. 2, pp. 211236, (2017).

N. K. Conroy, V. L. Rubin, and Y. Chen. "Automatic deception detection: methods for finding fake news". Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 14, (2015).

H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim. "exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert)". Applied Sciences, vol. 9, no. 19, (2019).

Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification". Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February (2018).

Ahmed H, Traore I, Saad S. "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques". In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138). (2017).

Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. "A benchmark study on machine learning methods for fake news detection". Computation and Language. (2019).

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. "Automatic online fake news detection combining content and social signals". FRUCT’22: Proceedings of the 22st Conference of Open Innovations Association FRUCT. Pages 272279. (2018).

Kurasinski, L. "Machine Learning explainability in text classification for Fake News detection". Malmo University Publications. (2020).

Bahaa Eddine Elbaghazaoui, Amnai Mohamed & Abdellatif Semmouri. "Data Profiling over Big Data Area: A Survey of Big Data Profiling: State-of-the-Art, Use Cases and Challenges". In book: Intelligent Systems in Big Data, Semantic Web and Machine Learning. Springer. (2021).

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu & Shlomo Havlin. “Fake news propagates differently from real news even at early stages of spreading”. EPJ Data Science volume 9, Article number: 7 (2020).