Automatic Identification of Conspiracy Theories Using BERT

Notebook for the PAN Lab at CLEF 2024

Daniel Yacob Espinosa, Grigori Sidorov and Eusebio Ricárdez-Vázquez

Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

Abstract

In recent years, we have seen an increase in conspiracy theories on social media. Users create these ideas mainly due to ignorance and distrust of government indices. These types of problems became prominent during the Covid-19 pandemic, representing a serious public health issue. This year, PAN 2024 has organized the task "Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives", to present at CLEF 2024, in which we participated. For this task, we present a solution using a BERT configuration, with a preprocessing layer that has previously given us good results for various types of classification. We also tried combinations of n-grams of words (2-3-4) and characters (2-3-4-5).

Keywords

BERT, Conspiracy Theories, Telegram, Fake news

1. Introduction

With the evolution of the Internet comes the evolution of social media, now one of the primary means of communication and dissemination, where millions of people worldwide can exchange their opinions in real time. With the onset of the Covid-19 pandemic, millions of stories were created by social media users, primarily driven by fear or societal ignorance. Conspiracy theories on platforms such as Facebook, Twitter/X, Telegram and YouTube can be highly persuasive, exploiting users psychological and social vulnerabilities.

The lack of rigorous information verification and the ease with which these theories can be shared and amplified contribute to their persistence and spread. These narratives exist due to a combination of psychological, social, and technological factors when used on these platforms. Their impact can be significant, and addressing them requires an approach that includes education, regulation, and communication [1].

There is a study by Hendari et al. [2], which evaluates tweets from two datasets. The Cresi 2017 dataset includes all kinds of information from Twitter accounts of genuine users and spambots, and the other dataset is CoAID, which is a collection of misinformation about Covid-19, and these data are not exclusively from Twitter [3]. With these datasets, Hendari et al. demonstrate a solution perspective with an 87% F1-score using BERT. The research by Hendari et al. provides a solid foundation for future research. In this work, we will use BERT as a classification model. We believe it is possible to improve the accuracy and effectiveness in identifying misleading and conspiratorial content for social media.

A notable study using BERT is by Guo et al., which discusses the use of transformers [4]. It improves the performance of RoBERTa by using it as an independent output and testing layer efficiency. This study demonstrates the efficiency of RoBERTa for the dataset extended from the FakeNews task in 2022 in MediaEval 2021 [5]. This corpus is comprised of English tweets that contain many conspiracy theories, primarily related to Covid-19. An important point is that this dataset is only in English and

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

espinosagonzalezdaniel@gmail.com (D. Y. Espinosa); sidorov@cic.ipn.mx (G. Sidorov); ericardez@cic.ipn.mx (E. Ricárdez-Vázquez)

http://www.cic.ipn.mx/~sidorov/ (G. Sidorov); https://www.cic.ipn.mx/index.php/eusebio-ricardez-vazquez (E. Ricárdez-Vázquez)

0009-0004-9245-2350 (D. Y. Espinosa); 0000-0003-3901-3522 (G. Sidorov)

it is unbalanced, which complicates the effectiveness of categorization. Significant progress is shown with an accuracy of 91% using RoBERTa.

Due to the COVID-19 lockdown, Ana López tells us how the traffic on social networks like WhatsApp and Telegram grew [6], as certain channels were used to share messages. In this way, certain government agencies used these methods to reach more communities, keeping everyone informed about procedures while maintaining social distancing. The research primarily describes the use of these technologies to reach a wider audience, which, although not as expected, was relevant during critical public health moments. However, not everything shared through these channels is done with good intentions.

In the research of Herasimenka et al., it is shown how certain communities manage to share misinformation via Telegram, often through links to deceptive sources [7]. They also note that misinformation can spread virally even on platforms without an algorithmic timeline, such as Telegram, depending on whether communities are involved in its dissemination. Furthermore, the people who share this type of information are very specific users and not many of them are content creators. Since Telegram does not use many mechanisms to eradicate these problems, these users utilize it as a channel to disseminate such content.

This year, PAN 2024 has organized a series of tasks called "Oppositional Thinking Analysis: Conspiracy Theories vs. Critical Thinking Narratives" [8, 9], aimed at classifying conspiracy theories and critical thinking narratives using Telegram texts. These tasks seek to address the growing issue of misinformation and the proliferation of conspiracy theories on social media platforms, where information spreads rapidly and can have a significant impact on public opinion.

2. Methodology

For this work, we used a combination of BERT models to perform the binary classification. We have employed these combinations for other tasks and wanted to see how they would perform in this context, as we have primarily used them for tweets analysis as well. As part of our experiments, we utilized certain variations of BERT, such as Bertweet and DistilBERT.

2.1. Preprocessing steps

In each research project related to natural language processing that we conduct, we recommend having a data preprocessing layer. Often, this step is essential for obtaining better results that the training model cannot achieve on its own to improve accuracy. Data preprocessing includes several essential tasks such as data cleaning, normalization, tokenization, and noise removal. For this research, we performed the following steps:

- **Lowercase.** We decided to use all lowercase text to standardize the text.
- **Links.** The links in the messages were replaced with the label 'link'.
- **User Mentions.** The messages in this dataset contain the '@' sign followed by a space and then the username. Were replaced with the tag 'usermention'.
- **Hashtags.** Can identify them because they start with the '#' sign, followed by a space and then the word. Were replaced with the tag 'hashtag'.
- **Emojis.** In previous experiments, emojis have shown to improve classifier learning, so we did not make any changes to them.
- **Other symbols.** All symbols not registered in the ASCII standard were removed from the dataset.

3. Experiments

Due to our previous research experience with social media, we decided also to try a methodology based on N-grams. Previously, we obtained interesting results despite of its' simplicity [10]. While the messages in this dataset share similarities, it is crucial to note that each social media platform has

its own dynamics and peculiarities. Therefore, adapting our approach to account for these specific variations will ensure more precise and relevant results. We present the F1-Score results using this setup.

Table 1
Results of combinations N-grams characters and words accuracy

	N-grams characters / words	subtask1
English	2-3-4-5 / 2-3-4	61.22
Spanish	2-3-4-5 / 3-4	59.74

Given that the initial results could be improved, we decided to use BERT and some of its variants for the task. One of these variants is DistilBERT, which is a lighter and more efficient version of BERT. However, a significant issue with this version is that it was primarily trained on English data, which may limit its effectiveness in contexts requiring data analysis in other languages [11].

Similarly, another model we considered was RoBERTa. Although RoBERTa is known for being an extremely robust and powerful model, it also shares the limitation of being trained primarily on English data, which could affect its performance in multilingual applications [12].

To evaluate and compare the performance of these models, we decided to train and test BERT, DistilBERT, and RoBERTa under the same conditions. Specifically, we configured the experiments with a batch size of 16 and trained the models for 9 epochs. This setup allowed us to maintain a balanced use of GPU resources and training time. It is important to note that while RoBERTa is more robust and generally requires more time to fully train due to its complexity and size, this approach allowed us to efficiently manage time and available resources without overwhelming the capabilities of the GPUs used.

Based on the experiments, we decided to use BERT as the model to execute this task. The model performed very close to our expectations, and we believe that with a bit more training data, it could have achieved even better results. DistilBERT ended up in last place. While it is quick in training and execution, it may be suitable for projects that involve limited storage and technological capabilities.

Table 2
Results of BERT variations with English Dataset

BERT Model	F1-Score
DistilBERT	84.02
RoBERTa	91.74
BERT	92.18

Table 3
Results of BERT variations with Spanish Dataset

BERT Model	F1-Score
DistilBERT	77.25
RoBERTa	86.45
BERT	89.13

4. Conclusions

The implementation and analysis of models such as BERT, DistilBERT, and RoBERTa were highlighted, with a focus on evaluating their performance under comparable conditions. The results indicated that BERT was the most effective model for the task at hand, showing results close to expectations. While DistilBERT, although fast, demonstrated limitations in contexts requiring deep analysis and multilingual data. These findings underscore the importance of adapting models to the specific characteristics of each task and linguistic platform, therefore optimizing the accuracy and relevance of the results obtained.

The rise of conspiracy theories on social media poses a significant challenge to society, especially during crises like the COVID-19 pandemic. The spread of misinformation can erode trust in institutions, polarize society, and lead to dangerous behaviors. Engaging in this task not only has the potential to advance academic research in natural language processing but also contributes significantly to combating misinformation in the public sphere. We believe that with a meticulous, evidence-based approach, we can develop effective tools to distinguish between conspiracy theories and critical thinking narratives, thereby helping to mitigate the negative impact of misinformation on society.

We would like to promote the use of BERT and RoBERTa for text classification in Spanish. This could lead to discoveries and improvements that will benefit not only Spanish speakers but also other languages that share similar linguistic features. We are excited about the possibilities this initiative presents and are confident that our contributions will be valuable in the global effort to understand and counter misinformation on social media.

References

[1] M. Esayas, D. Pandey, B. Alemu, B. Pandey, S. Tareke, The negative impact of social media during covid-19 pandemic, Trends in Psychology 31 (2022). doi:10.1007/s43076-022-00192-5.

[2] M. Heidari, S. Zad, P. Hajibabaee, M. Malekzadeh, S. Hekmatiathar, O. Uzuner, J. Jones, Bert model for fake news detection based on social bot activities in the covid-19 pandemic, 2021, pp. 0103–0109. doi:10.1109/UEMCON53757.2021.9666618.

[3] L. Cui, D. Lee, Coaid: Covid-19 healthcare misinformation dataset, 2020.

[4] H. Guo, T. Huang, H. Huang, M. Fan, G. Friedland, Detecting covid-19 conspiracy theories with transformers and tf-idf (2022).

[5] L. Sweeney, M.-G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. Smeaton, M. Sultana, Overview of the mediaeval 2022 predicting video memorability task (2022).

[6] A. B. López Tárraga, Comunicación de crisis y ayuntamientos: el papel de telegram durante la crisis sanitaria de la covid -19, Revista de la Asociación Española de Investigación de la Comunicación 7 (2020) 104–126. doi:10.24137/raeic.7.14.5.

[7] A. Herasimenka, J. Bright, A. Knuutila, P. Howard, Misinformation and professional news on largely unmoderated platforms: the case of telegram, Journal of Information Technology and Politics 20 (2022) 1–15. doi:10.1080/19331681.2022.2076272.

[8] K. Damir, C. Berta, B. C. Xavier, T. Mariona, R. Paolo, R. Francisco, Overview of the oppositional thinking analysis pan task at clef 2024, Working Notes of CLEF 2024, 2024.

[9] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Condensed Lab Overview.In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, in: Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, 2024.

[10] D. Espinosa, H. Gómez-Adorno, G. Sidorov, Bots and Gender Profiling using Character Bigrams, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019). URL: https://api.semanticscholar.org/CorpusID:203626972.

[12] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, 2021. URL: https://aclanthology.org/2021.ccl-1.108.