How do disinformers write? An unsupervised approach for stylistics characterization of false information spreaders.

David Gutiérrez Marjalizo, Manuel Cano García, Alejandro Cañas Borreguero

Machine Learning, ESI Ciudad Real - Escuela Superior de Informática de UCLM, Paseo de la Universidad, Ciudad Real, 13071, España.

Contributing authors: David.Gutierrez7@alu.uclm.es; Manuel.Cano3@alu.uclm.es; Alejandro.Cañas1@alu.uclm.es;

Abstract

This project focuses on the application of machine learning techniques to classify COVID-19-related tweets as either "real" or "fake." Leveraging a dataset composed of labeled tweets, the study employs advanced natural language processing (NLP) methods, including tokenization, lemmatization, and stopword removal, to preprocess the text data. Topic modeling is performed using BERTopic, which integrates sentence embeddings, dimensionality reduction (UMAP), and clustering (HDBSCAN) to uncover underlying themes in the dataset. Visualizations, such as word clouds and distribution plots, are used to explore the frequency and structure of the data. The project also prepares the text for machine learning by converting it into numerical features using techniques like CountVectorizer. While the notebook sets the foundation for model training and evaluation, it emphasizes the importance of data exploration and preprocessing in building robust classification systems. This work highlights the potential of AI and NLP in addressing misinformation, particularly in the context of public health crises, and lays the groundwork for future advancements in text-based machine learning applications.

Keywords: machine learning, natural language processing, text classification, COVID-19 misinformation, BERTopic, topic modeling, UMAP, HDBSCAN, data preprocessing, word clouds, CountVectorizer, tweet analysis, fake news detection, public health, data visualization

Contents

1 Introduction

1.1 Background

Disinformation, defined as the intentional spread of false or misleading information, has become a global phenomenon that undermines public trust, social stability, and democratic processes. In the digital age, social media has exacerbated this issue, enabling fake news to spread rapidly and reach massive audiences. This challenge has become particularly critical in contexts such as the COVID-19 pandemic, where disinformation has influenced public health decisions, creating confusion and eroding trust in official sources.

To address this problem, Artificial Intelligence (AI) techniques, particularly Natural Language Processing (NLP) and Topic Modeling, have emerged as key tools. NLP allows for the analysis and understanding of large volumes of text, identifying patterns and anomalies that may indicate disinformation. Meanwhile, Topic Modeling helps uncover hidden themes within datasets, revealing the narratives and strategies used in disinformation campaigns. These technologies not only facilitate the detection of misleading content but also provide valuable insights to mitigate its impact.[1]

1.2 Objective of the work

Misinformation about COVID-19 impacts public behavior, vaccination rates, and trust in health authorities. Social media, especially Twitter, has become a hub for both accurate and false information. The spread of falsehoods promotes ineffective treatments, conspiracy theories, and distrust in scientific institutions [2]. Detecting misinformation in real-time ensures the dissemination of reliable information while curbing harmful narratives.

For that reason, the aim of this work is to analyze the difference between disinformation and trustworthy sources of information using NLP techniques and Topic Modeling with BERTopic in order to extract key characteristics on both sides in their way of expression, themes, or other distinguishing characteristics to put into contrast between disinformation versus credible information.[3].

2 Data Exploration

2.1 Dataset Description

The final data set used in this task was made using the following data sets: Constraint-English-Train.xlsx, English-test-with-labels.xlsx and Constraint-English-Val.xlsx to have the most different tweets possible. This dataset does have 3 columns:

- id: Corresponds to an integer number which identificates a tweet, each tweet has one.
- tweet: Corresponds to the content of the tweet itself.
- label: Corresponds whether it is a credible information (real) or disinformation (fake).

With that dataset, we have 5600 real tweets and 5100 fake tweets to do the analysis, which is quite balanced. [4].

2.2 Exploratory Data Analysis and preprocessing

First of all, we were looking to know about the length of the messages to find out if short messages could have enough semantic content to take these messages into account or not. Describing the lengths, the minimum length in the dataset is 18 characters, while the maximum is 10170 characters. Obiously the bigger is the message, the more probabilities we have to find importance in that message, for that reason, we are going to look if those short message could be important for our analysis.

Filtering into a length equal or below to 35 to see the possible impact of these short messages, we could find tweets such as "There is no pandemic", "Drinking alcohol can cure Covid-19" or "Bill Gates predicted coronavirus" which besides of having less characters, they have a strong meaning because they told about the main topic of the dataset, the COVID-19 so we are keeping also short messages to analyze the length of the tweets in order to find differences about both sides in terms of length messages.

Also, for better understanding of the task and better conclusions, we are going to split the dataset into real tweets and fake tweets which will enable us to make a prior analysis individually for later comparison between real tweets and fake ones.

3 Methodology

3.1 NLP pipeline

3.1.1 Tokenization and data cleaning

The raw text data from the tweets required extensive preprocessing to prepare it for machine learning models. Tokenization was performed to split each tweet into individual words or tokens, breaking down the text into manageable units. Following tokenization, we are going to applied data cleaning to remove those irrelevant aspects of the language in order to facilitate the task and reduce vocabulary. For data cleaning, we are going to perform the following approaches:

- Preprocess data: We are going to put each word to lowercase and remove any alphanumeric character.
- Checking for URLs: We found different source of urls used in real and fake tweets so we are keeping it because we saw more urls in real tweets than in fake tweets, so we could find some patters of behaviour related to the usage of urls.
- Checking for numbers: We can see different contexts using numbers
  - To show the number of deaths.
  - Time related
  - To refer the illness (COVID-19), in that case 19 is also a number counted
- Normalization: We are going to normalize all sort to refer the virus such as COVID19,COVID-19 into just covid

- Stepwords and puntuation marks: We are removing them because of their lack of semantic implication in a sentence.

[5].

3.1.2 Lemmatization

[6]. Lemmatization was applied to reduce words to their base or root forms, ensuring that variations of the same word (e.g., "running" and "ran") were treated as a single entity in order to reduce vocabulary, making the task more manageable and easier to analyze. Having lemmatizated both datasets we can see clearly some difference between disinformers and reliable sources using a wordcloud.

In reliable sources we saw words like 'data', 'new case', 'confirmed case', 'case', 'today' which this means that reliable sources would be related with some media or scientist context.

In disinformers, we saw words like 'lockdown', 'hospital', 'India', 'China', 'vaccine', 'death' which means that desinformer try to use contraversial words to create the more disturbance possible using the context and the emotions of the people to make damage.

3.1.3 Part of speech tagging

With this technique, we are going to remove all verbs to reduce more the vocabulary of the domain because we thought that this lexical category has less impact in the speech in comparison to adjectives or nouns, especially in this context that there are a lot of scientific vocabulary.

3.1.4 N-grams

We are use different types of n grams to see if we could extract some interesting information about the difference about real and fake tweets. We have used ngrams from 1 to 3 to see different perspectives and patterns.

3.2 Topic Modeling with BERTopic

3.2.1 Sentence embeddings with sentence-transformers

To uncover underlying themes in the dataset, topic modeling was performed using the BERTopic library. The first step involved is finding the proper embeddings in hugging face to have more accurate conclusions; We found two different approaches: a biomedical and a fake news related approach. These embeddings represent the semantic meaning of the text in a high-dimensional vector space, capturing the contextual relationships between words and phrases. By transforming the raw text into numerical embeddings, the model could better understand the content and structure of the tweets, enabling more accurate topic identification. Using different types of embeddings we could extract different perspectives about the same context and see the difference between approaches as well.

3.2.2 Dimensionality reduction using UMAP

Once the sentence embeddings were generated, dimensionality reduction was applied to simplify the data while preserving its essential structure. The Uniform Manifold Approximation and Projection (UMAP) algorithm was used to reduce the high-dimensional embeddings into a lower-dimensional space (typically 2D or 3D). This step was crucial for visualizing the data and identifying clusters of similar tweets. UMAP's ability to maintain both local and global relationships in the data made it an ideal choice for this task [6].

3.2.3 Clustering with HDBSCAN

After dimensionality reduction, the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm was employed to group the tweets into meaningful clusters. Unlike traditional clustering methods, HDBSCAN does not require specifying the number of clusters in advance and can handle noise effectively. This approach allowed for the automatic discovery of topics within the dataset, revealing patterns and themes that distinguish "real" from "fake" tweets. The resulting clusters were then visualized using tools like datamapplot to provide intuitive insights into the data [7].

3.3 Feature Engineering

3.3.1 Text vectorization with CountVectorizer

To prepare the text data for machine learning models, the tweets were converted into numerical features using the CountVectorizer from the scikit-learn library. This technique transforms the text into a matrix of token counts, where each row represents a tweet and each column corresponds to a unique word in the dataset. By creating a bag-of-words representation, the model could analyze the frequency of terms and their relevance to the classification task. This step was essential for bridging the gap between raw text and machine-readable input [3].

3.3.2 Topic representation

In Topic representation, each cluster formed represents a topic. BERTopic then identifies representative keywords for each topic using the words from the documents within that cluster. These keywords help in interpreting and labeling the topics.

3.3.3 Fine tuning

For fine tuning, we are going to use the KeyBERTInspired class to have a better topic representation for further analysis. Also, there are some other parameters in the construction of the BERTopic model which could be modified to build different topic representation:

- min-topic-size: whether we want more or less topics
- n-gram-range: we will chose de range (1,2) because previously we did not find some interesting word patterns with range 3.

- nr-topics: the objective of this parameter is to achieve the maximum number of topics without overlap between them in order to extract as much useful information as possible.

4 Results and Discussion

4.1 Real Bio vs. Fake Bio

4.1.1 Real Bio

The topics in `real_bio` are focused on factual, verifiable, and data-driven information. Below are the identified topics:

- **TOPIC -1:** Terms such as "staysafe indiawillwin", "new zealand", and "quarantine facility" suggest a focus on safety measures and statistics related to COVID-19, particularly in countries like India and New Zealand. This topic reflects an interest in disseminating official and verified information about the pandemic, with an emphasis on prevention and case tracking.
- **TOPIC 0:** Terms such as "icmrdelhi", "mohfwindia", and "indiafightscorona" are related to official organizations (Ministry of Health and Family Welfare) and social movement like indiafightscorona which encourage to fight against covid-19 in India. This topic highlights the importance of official sources and transparency in communicating COVID-19 data.
- **TOPIC 1:** Terms such as "cureddischargedmigratedactive casesdeaths" and "indiawillwin staysafe" indicate a focus on detailed statistics about COVID-19 cases, including recoveries and deaths. This topic demonstrates a commitment to accuracy and clarity in presenting numerical data.
- **TOPIC 2:** Terms such as "drharshvardhan drhvoffice" and "mohfwindia drharshvardhan" are related to public figures and health authorities. This topic reinforces trust in public figures and authorities to disseminate verified information.
- **TOPIC 3:** Terms such as "benue kaduna" and "akwa ibom" are related to geographic locations and local statistics of Nigeria. This topic reflects a focus on concrete and verifiable data at the regional level.
- **TOPIC 4:** Terms such as "quarantine number" and "arrival 2762" suggest a focus on logistics and tracking of health measures, such as quarantines and arrivals. This topic highlights the importance of organization and tracking of official measures.

4.1.2 Fake Bio

The topics in `fake_bio` are centered on misleading, exaggerated, or false information. Below are the identified topics:

- **TOPIC -1:** Terms such as "donald trump", "covid case", and "cure covid" suggest false or exaggerated claims about public figures and unverified treatments. This topic reflects a pattern of manipulating public opinion by associating prominent figures with unverified claims.

- **TOPIC 0:** Terms such as "coronavirusfacts" and "http coronavirus" are related to false claims about the pandemic, including conspiracy theories and misinformation. This topic shows how misinformation is disguised as "facts" to deceive the public.
- **TOPIC 1:** Terms such as "coronavirus bleach" and "vinegar coronavirus" suggest unverified home remedies and dangerous claims about how to "kill" the virus. This topic is particularly dangerous, as it promotes unscientific practices that could endanger health.
- **TOPIC 2:** Terms such as "covid hydroxychloroquine" and "chloroquine antimalarial" reflect false claims about unapproved medical treatments. This topic shows how medical misinformation can spread rapidly, even with potentially harmful treatments.
- **TOPIC 3:** Terms such as "test covid" and "cure covid" are related to false claims about testing and treatments. This topic reinforces the pattern of medical misinformation and promotion of unverified solutions.
- **TOPIC 4:** Terms such as "cream koolfee" and "food milkshake" suggest unverified home remedies or products. This topic reflects a focus on promoting non-medical and potentially misleading solutions.

4.2 Real News vs. Fake News

4.2.1 Real News

The topics in `real_news` are focused on factual and verifiable information. Below are the identified topics:

- **TOPIC -1:** Terms such as "bauchi kaduna" and "akwa ibom" are related to geographic locations and local statistics of Nigeria. This topic reflects a focus on concrete and verifiable data at the regional level.
- **TOPIC 0:** Terms such as "covid case", "new case", and "spread covid" are related to statistics and official reports on COVID-19. This topic highlights the importance of transparency and accuracy in disseminating data.
- **TOPIC 1:** Terms such as "coronavirusupdates covidupdates" and "indiafightscorona icmrdelhi" reflect official and verified updates about the pandemic. This topic reinforces trust in official sources for staying informed.
- **TOPIC 2:** Terms such as "discharged 1013" and "new case" are related to statistics on COVID-19 cases and recoveries. This topic demonstrates a focus on accurate and verified presentation of data.
- **TOPIC 3:** Terms such as "22788393 sample" and "2970492 case" are related to numerical data and concrete statistics. This topic reflects a commitment to transparency and accuracy in presenting data.
- **TOPIC 4:** Terms such as "arrival 2868" and "departure facility" suggest a focus on logistics and tracking of health measures. This topic highlights the importance of organization and tracking of official measures.

4.2.2 Fake News

The topics in `fake_news` are centered on misleading, exaggerated, or false information. Below are the identified topics:

- **TOPIC -1:** Terms such as "covid case", "cure covid", and "vk srinivas" suggest false or exaggerated claims about treatments and public figures. This topic reflects a pattern of manipulating public opinion by associating prominent figures with unverified claims.
- **TOPIC 0:** Terms such as "general election" and "lockdown rise" are related to misleading claims about political events and lockdown measures. This topic shows how misinformation is used to influence public perception on sensitive issues.
- **TOPIC 1:** Terms such as "delhi" and "opposition party" are related to misleading claims about politics and the economy. This topic reflects a focus on manipulating public opinion on political issues.
- **TOPIC 2:** Terms such as "state voluntary" and "covid 5185" suggest misleading claims about statistics and measures related to COVID-19. This topic shows how misinformation can distort the perception of data and official measures.
- **TOPIC 3:** Terms such as "http donaldtrump", "cure covid", and "false" are related to false claims about public figures and unverified treatments. This topic reinforces the pattern of medical misinformation and manipulation of public opinion.

4.3 Challenges and Limitations

Although Artificial Intelligence (AI) techniques, such as Natural Language Processing (NLP) and Topic Modelling, are effective for detecting disinformation, their application faces several challenges and limitations. Firstly, data quality is a key issue, as the analyzed tweets may contain noise (informal language, spelling errors) and a mix of truthful and false information, which hinders the accuracy of the model. Additionally, models may reflect biases present in the training data, limiting the generalizability of the results. The interpretation of identified topics can also be subjective and ambiguous, requiring manual validation to classify them correctly.

Another significant limitation is that models like LDA do not capture complex semantic relationships, which can lead to an incomplete representation of topics. Scalability and performance are also challenges, as processing large volumes of data in real-time is computationally expensive. Furthermore, disinformation evolves rapidly, requiring constant updates to models and datasets. Finally, the analysis of social media data raises ethical and privacy concerns that must be addressed.

In summary, while AI techniques are promising, it is necessary to address these challenges to improve the accuracy, effectiveness, and ethics of models in the fight against disinformation.[8]

5 Conclusions

5.1 Summary

The application of NLP and Topic Modeling techniques has provided valuable insights into the differentiation between truthful information and misinformation in the context of COVID-19.

Key findings include:

- **Truthful Information:** The `real` datasets emphasize factual, data-driven, and transparent information. Dominant themes include official updates, location-specific statistics, and numerical data, showcasing a commitment to accuracy and reliability.
- **Misinformation:** The `fake` datasets reveal patterns of unverified medical claims, political manipulation, and dangerous home remedies. These topics often leverage sensationalism and associations with public figures to spread false narratives.

Despite these successes, challenges persist:

- **Detection Challenges:** Ambiguity in interpretations, noisy data, and the dynamic nature of misinformation require continual advancements in modeling techniques.
- **Ethical Considerations:** The analysis of social media data raises concerns about privacy and potential misuse of AI tools, underscoring the need for responsible application of these technologies.

Future work should focus on improving data quality, enhancing model robustness, and exploring hybrid approaches that combine automated methods with human validation. Additionally, integrating advanced models capable of capturing complex semantic relationships in text will be essential for addressing the evolving landscape of misinformation.

5.2 Future work against disinformation

Nowadays, disinformation is more present than ever due to various aspects, such as the development of AI video and image generation enabling people to do whatever they like regarding of video and image; They could use it for parody or for harmless intentions, however, people could use it for harmful intentions, threatening to political personalities, spreading false rumors about health or using deepfake or fake videos.

In order to fight disinformation, new ways of detecting synthetic content must be developed to differentiate between real and false content could be a horizon to fight against a big part of disinformation.

On the other hand, not only technical advances must be made, but also social behavior and regularization must change. Governments should be encouraging about disinformation as a threat as dangerous as any other social issues such as extreme violence or traffic drugs. Indeed, it is inevitable the existence of disinformation, for that reason it is also important to the media to inform properly about the real context of a new before labeling it as "real" or "fake", otherwise, we would do disinformation due to the lack of investigation. [2].

6 Bibliography

References

[1] Kertysova, K.: Artificial intelligence and disinformation: How ai changes the way disinformation is produced, disseminated, and can be countered. Security and Human Rights **29**(1-4), 55–77 (2018) https://doi.org/10.1163/18750230-02901004

[2] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018) https://doi.org/10.1126/science.aap9559

[3] Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys **53**(5), 1–40 (2020) https://doi.org/10.1145/3395046

[4] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017) https://doi.org/10.1145/3137597.3137600

[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

[6] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint (2018) arXiv:1802.03426 [stat.ML]

[7] Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Advances in Knowledge Discovery and Data Mining, pp. 160–172 (2013). https://doi.org/10.1007/978-3-642-37456-2_14

[8] Montoro Montarroso, A., Cantón-Correa, J., Gómez Romero, J.: Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. Profesional de la Información **32**(3) (2023) https://doi.org/10.3145/epi.2023.may.22