Machine Learning Progress Task 3: A deep insight on the writing techniques of disinformers

Elena Ballesteros Morallón, Mario Jiménez Redondo, Miguel Ángel Ruiz Arreaza

Contributing authors: elena.ballesteros2@alu.uclm.es; Mario.Jimenez9@alu.uclm.es; MiguelAngel.Ruiz9@alu.uclm.es;

Abstract

This study applies Natural Language Processing (NLP) techniques to analyze linguistic differences between real and disinformative COVID-19-related tweets. Using tokenization, N-Gram extraction, Named Entity Recognition, Topic Modeling, and Sentiment Analysis, we identify key patterns in a dataset of verified news and fact-checked fake claims.

Results reveal that real tweets are longer, more detailed, and focus on factual reporting, while fake tweets are concise, emotionally charged, and often target political figures. Sentiment analysis highlights higher negative sentiment in fake tweets, and topic modeling shows diverse clusters for real tweets versus narrow, sensational themes in fake ones. These findings emphasize the potential of NLP tools in combating disinformation by identifying patterns in fake news dissemination.

Keywords: COVID-19, disinformation, Topic Modeling, Natural Language Processing

1 Introduction

1.1 Data Science and Natural Language Processing

The huge volume of information in the present era has allowed human beings to generate techniques to study, detect and process patterns in data, giving birth to the concept of Artificial Intelligence and Machine Learning. Shortly after, the concept of Natural Language Processing (NLP), a set of techniques for processing human language extracts of any kind (audio, text, image...).

In this paper, results from multiple NLP techniques such as tokenization, N-Gram extraction (extracting the most frequent sequences of words), Named Entity Recognition, Topic Modeling (with topics extracted using BERTopic and summarized using an LLM) and Sentiment Analysis (Analysis on patterns found on text to find some hidden sentiment) are presented in order to accomplish the goals described below.

1.2 Aim of this report

Disinformation, misinformation and other sources of informative chaos are increasingly present on social media. With entire teams working to verify the veracity of many claims a day, it has now become a real issue that has established the beginning of a trust and authority crisis (either for users whenever they come across any news through these information channels, or for scientific authorities). Disinformation is an informative disorder or manipulation that has the intention of generating chaos in the material world, and/or generate a specific reaction towards an individual or collective of individuals, like some propaganda applied to mainstream news.

The existence of disinformation has moved many individuals to have the objective to tackle it somehow. This paper has the aim of using a variety of data science and natural language processing techniques to search for significant writing pattern differences between real information producers and disinformers.

2 Method

Data is extracted from a dataset about COVID-19 Fake News[1]. It combines both real news and fake claims that surfaced on social media on COVID-19 topics during 2020. Fake claims are collected from various fact-checking websites whereas real news are collected from Twitter using verified Twitter handles. It's crucial to take into account the source of our data for the analysis.

2.1 Exploratory Data Analysis (EDA)

Understanding the dataset and extracting key insights is essential before applying text transformations. Here, we summarize the main observations.

1. **Message length**: after measuring the number of characters, words and average word length, we can conclude that real messages are longer, with more content and detail. This suggests they intend to communicate a complete message while fake writers want to express an idea without much elaboration.
2. **Message content**: the number of stopwords is higher in real tweets, which supports the above statement that it is a complete text. As we are in a social media context, it's interesting to check the number of links and hashtags attached to the messages. The analysis shows that fake tweets usually contain one or no links and hashtags whereas real ones are more likely to have some. Also, we can examine the most repeated words with the following chart.

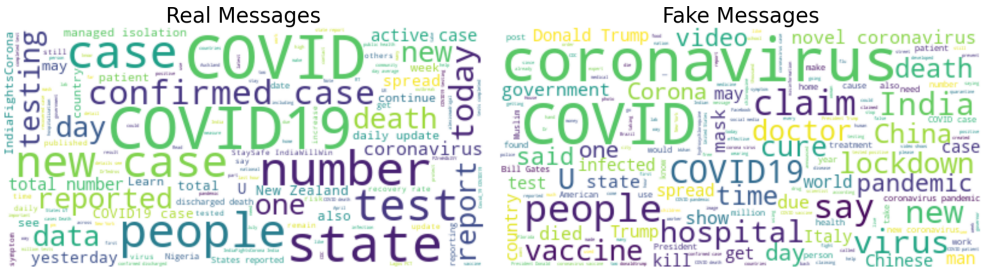


Fig. 1: Word Cloud

3. **Text complexity**: We have used textstat, which is a Python library, to score the readability index of a text indicating how readable it is and what type of audience it is intended for. Result shows that fake messages are easier to read while real ones are more complex. This supports our previous findings: Fake messages tend to be more concise and lack elaboration, aiming for quick dissemination of information. On the other hand, real tweets have a richer and more elaborate style, as they are supposed to be written by people with higher levels of education or journalists.

2.2 Data Preprocessing

Following the EDA step, the data was cleaned by removing unnecessary characters. First, we applied tokenization, a process that consists in splitting the text into individual words. Additionally, given the nature of the text, an extra step was taken to remove links as well as symbols commonly used on social media, such as hashtags (#) and mentions (@). Stopwords were also eliminated to ensure a cleaner and more meaningful analysis. After this, we obtained a clearer ranking compared to the word cloud. We have found something interesting:

- Real messages: The top word for naming the virus is "covid-19".
- Fake messages: The top word for naming the virus is "coronavirus".

This difference means that while real messages are more formal and use the official name of the virus, fake messages use the most popular name and even others like "Corona" or "COVID". It might be interesting to check if names like "China virus" or "Chinese virus" appear together, to do this we'll study N-grams.

N-grams are contiguous sequences of 'n' items, that can be characters, words, or even syllables. They play an important role in capturing patterns and relationships within a sequence of words. We'll only work with bigrams and trigrams, as they are the most common n-grams used in text analysis. Results can be found in section 3.1.

2.3 NLP Techniques

2.3.1 Named Entity Recognition (NER)

This is a technique that analizes a list of tokens (in our case, previously taken from all of the tweets in the dataset) and processes them into a tree, that classifies different parts of the token list into a series of entities, categorized by the token list.

We performed the NER process by using the nltk library, and then identifying all named entities found in the trees, and extracting them into a list of strings, with each string representing an entity that was mentioned by the corresponding tweet. This is done for all different categories of entities, and in aggregate for all of them.

By formatting the named entity data like this, we can use a bar plot to identify which entities are most frequently mentioned.

This is an improvement over the use of things like word frequency, or identifying n-grams, because compared with those methods, it allows entities to have a variable number of words, and avoids the presence of useless words in the statistic that are not used to mention any entity.

The main disadvantage of this approach would be the case sensitive approach to recognizing named entities, since it can cause issues if the word is uppercased, but not referencing an entity, or if an entity is in lowercase (e.g. trump vs Trump).

2.3.2 Sentiment Analysis

Sentiment analysis uses machine learning models capable of extracting patterns from a piece of text that can be used to derive its emotional tone. Since some sources claim that headings and messages with disinformation purposes tend to contain expressions that appeal to emotion, it would be interesting to explore the emotional load within the tweets of the dataset to gain insight into whether this applies to social media. Results

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a pre-trained model from the Python Natural Language Tool Kit (NLTK for short) specialized in extracting emotion scores from text data in a user-friendly format and workflow. Its convenience as a tool for Sentiment Analysis was the main reason why it was used to extract emotion scores from the COVID tweets.

2.4 Topic Modelling

Topic Modelling is described as a set of Natural Language Processing techniques that allow the classification of pieces of text from a certain dataset into different categories (called "topics"), usually through a model that uses pre-trained models that map the semantic distance of each token in a piece of text. The process of division of the data into topics can be seen as an instance of unsupervised learning, as the final model aggregates data into clusters without employing labelled data as a guide.

The most used and user-friendly Python library that enables support for topic modelling is BERTopic. To train a BERTopic model, a very specific workflow must be followed:

1. The model processing semantics (the Embeddings Model) shall be loaded and trained on the sample data. In this case, two Embedding Models were trained: one for real tweets and another one for fake tweets. The base model that was used for training the models was "COVID Twitter BERT v2"[2], a model developed by digitalepidemiologylab which contains several tweets on COVID news, so it is of remarkable convenience as it perfectly aligns with the kind of data to be processed.

2. The problem involves various tokens that make processing unfeasible in its current dimensionality. This means that dimensions shall be reduced by use of an unsupervised learning model. UMAP is the most used algorithm that's the reason why it has been used in the research. The model's configuration takes 15 neighbours and reduces the problem to 10 dimensions using the Cosine Similarity Method - tokens with high similarity, which can be interpreted as words or expressions with similar semantic value, are merged into a single category.
3. Once dimensionality is reduced, the next step is to make clusters out of the embedding model. For this case, the HDBSCAN algorithm has been employed, with a minimum cluster size of 15 elements and using Euclidean distance within clusters as criteria.
4. It is important that topics are created and separated into tokens. This is why a CountVectorizer element has been used for this purpose, using unigrams as a minimum analysable unit and bigrams as the maximum. The Stopwords database contains words in English.
5. Topics must be represented with a significant name to be clearly detectable in graphs. In order to provide support to such representation, a Class TF-IDF transformer has been used in conjunction with a Large Language Model that gathers the topic's keywords and returns a label summarising its meaning. The LLM chosen was "FLAN T5 BASE"[3], a model made by Google available at HuggingFace.
6. At the last step, all the models are aggregated into a pipeline which represents a BERTopic model, that is trained to obtain the final topic selection.

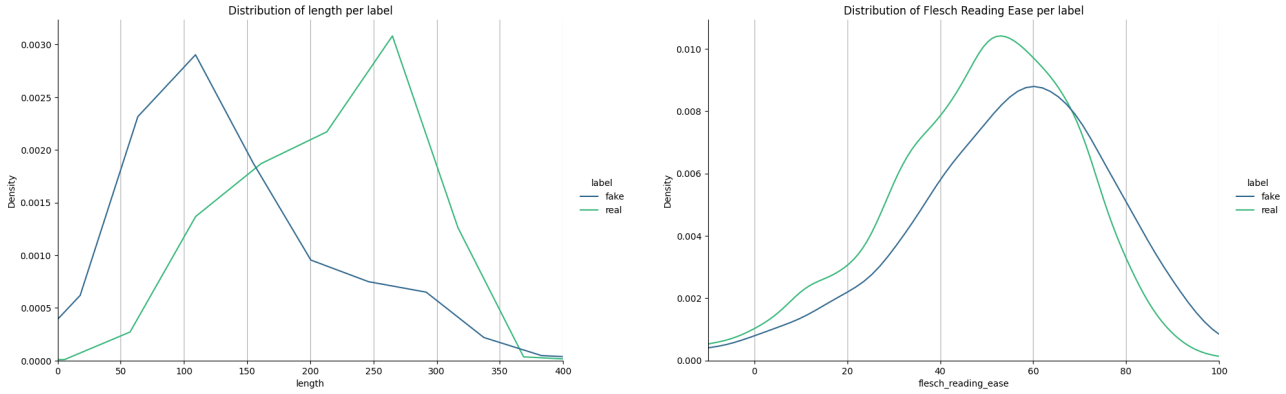
3 Results and Discussion

3.1 NLP Processing Results

This section contains the graphs obtained during the analysis and how these findings highlight the linguistic differences between real and fake messages.

Regarding message length, the number of characters in fake tweets is less than 200 while real messages have its peak at around 300 characters. This fact also affects the number of words that are higher in real texts with around 30 to 40 words per message, while most fake texts have between 10 to 20 words. These results suggest that real messages are usually longer and more detailed than fake messages. Therefore, we can extract that real messages contain more information than fake messages with more structured text. We have to take into account that the tweets hadn't been processed at that point, so it's also counting links.

Another measure to distinguish between the writing style of both groups is the text complexity. The blue line, that represents the fake messages, is tailed to the right, meaning that the text is easier to read. On the other hand, the green line, representing the real messages, is tailed to the left, meaning that the text is more difficult to read.



(a) Tweets length

(b) Reading Ease

The study of N-grams has helped us understand which words appear most frequently in the dataset and provides context before performing topic modeling. The top 3 most frequent bigrams for real messages are "new cases", "confirmed cases" and "total number". These bigrams express the idea of a report or a daily update to show the current situation of the pandemic. On the other hand, the top 3 most frequent bigram for fake messages are "novel coronavirus", "Donald Trump" and "Bill Gates". Trigram analysis confirms some of the previous insights. Real messages refer to "daily update published" and "number confirmed cases". Again, these are related to a report and with an informative intent. While in fake messages the most frequent trigram is "President Donald Trump".

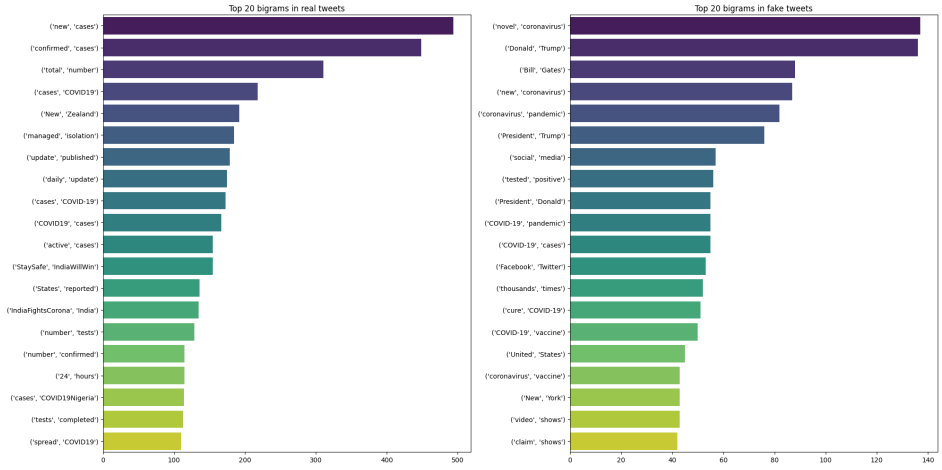


Fig. 3: Bigrams per label

3.2 Named Entity Recognition Results

It is notable that fake tweets tend to mention China and India, in particular Wuhan, and tend to mention Facebook, and NEWS (in all caps) . On the other hand, true tweets tend to mention the CDC, New Zealand, Nigeria, and also India (although, not as frequently as fake tweets)

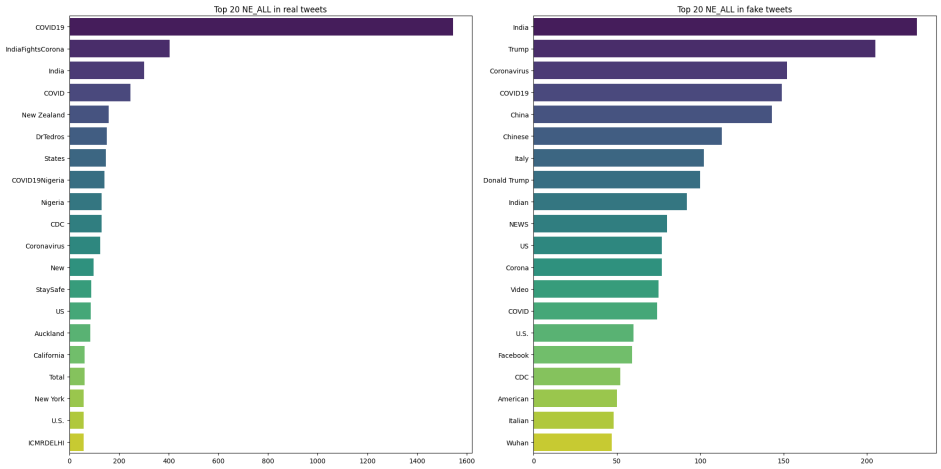


Fig. 4: Frequency of different Named Entities per label

3.3 Sentiment Analysis Results

After processing the sentiment load within the tweets, a pie chart was extracted with the proportions of each kind of sentiment (positive, negative and neutral)

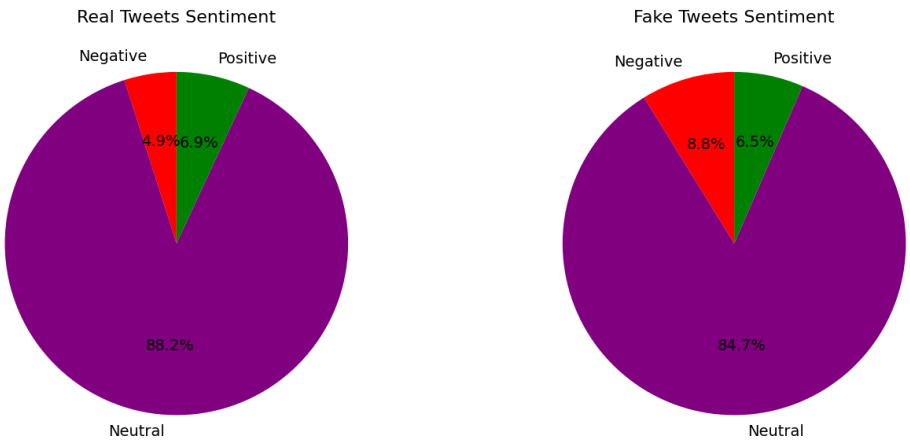


Fig. 5: Pie chart expressing the sentiment proportions for both real and fake tweets.

Both charts show a high rate of neutral tweets. The most significant difference on the charts is the proportion of negative tweets in the fake instances, with a rate of 8.8% in comparison to the 4.9% present in real instances.

According to Andrés Montoro et al[4], posts containing disinformation tend to have a greater emotional load than those which do not, often containing linguistic patterns that appeal to negative emotions like rage, disgust or fear in order to be catchier. This may be slightly proven here, where the proportion of negative emotions in disinformative tweets exceeds the rate of the real tweets. Nevertheless, the proportion of neutral emotional load remains high on both sides, which may be caused either by the nature of the data or by a low precision score of the model used.

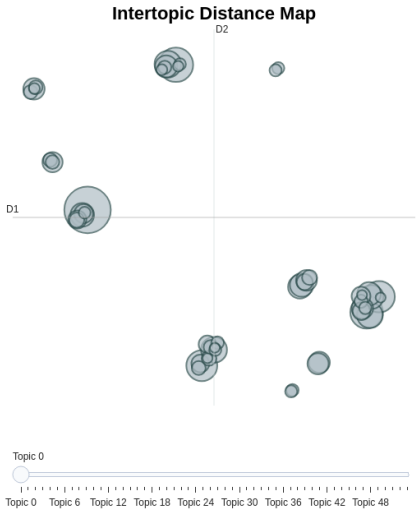
3.4 Topic Modeling Results

The representations of the results from Topic Modelling are diverse, however, there are two main representations where significant differences can be spotted between the representation for real tweets and the one for fake tweets: Intertopic distance maps and the top word scores representation.

3.4.1 Intertopic Distance Maps

Intertopic distance maps show a representation that describes the distribution of the topics in the semantic space, representing topic sizes (the larger the circle, the more tweets that fall under the category of that topic) and their semantic distance from other topics. For instance, topics that have a similar semantic load will be really close to each other.

(a) Intertopic Distance Map for real tweets

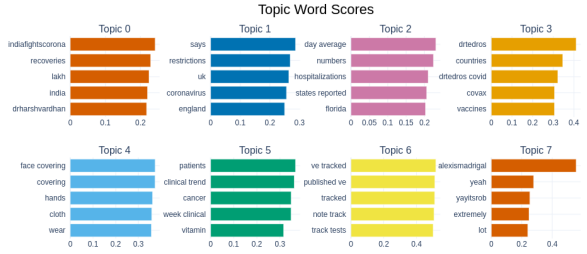


(b) Intertopic Distance Map for fake tweets

In the distance maps, particularly in the case of real tweets, is visible that the vast majority of the topic distribution consists of small clusters with reasonable separation from other clusters, which have a huge topic concentration each. This gives out the idea that there are well-separated niches, with reasonably similar content. Fake tweets, however, describe a map with two great clusters, greatly separated by the map's diagonal space. This means that the focus of fake tweets falls under two very distinct niches, with a considerably large topic load in each niche. The size of the topics is also to be mentioned, with some clusters reaching sizes even greater than any of the instances in the real tweets. This means that tweets containing disinformation would be categorized within one of a very narrow selection of niches, and that has an objective to harm in particular.

3.4.2 Top Word Scores representation

This representation ranks the density of occurrences of a word in a topic and represents a subplot of the N largest topics and the most significant words, that is, the ones that have a larger density of occurrences.



(a) Top Word Scores for real tweet topics.

(b) Top Word Scores for fake tweet topics.

This representation clearly shows the most significant differences in the writing style. Whilst the representation for real tweets is dominated by words related to the medicine industry, as well as very sounded news like country restrictions or case reports, the one for fake tweets shows a dominance of specific words like "chloroquine" (a compound that supposedly could be consumed in order to eradicate COVID) and other words that specifically target countries and world leaders. The clearest case is the existence of a whole topic focused on one particular individual: the President of the United States, Donald Trump. This gives out the idea that disinformers often tend to write about these people rather than topics focused on medicine or restrictions so that the news could reach further and have a great emotional impact on a target audience (that is, in this case, the niche of users believing in conspiracies involving politicians).

4 Conclusions

This study highlights key linguistic differences between real and fake COVID-19-related tweets, offering insights into disinformation patterns. Real tweets are longer, more detailed, and use formal language, while fake tweets are shorter, simpler, and more emotionally charged. Sentiment analysis revealed that fake tweets often evoke negativity, while topic modeling showed distinct thematic focuses—real tweets address public health, while fake ones emphasize conspiracies and political narratives. These findings can be usuful to create automated disinformation detection tools or campaigns to combat misinformation.

References

[1] Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Fighting an infodemic: COVID-19 fake news dataset. vol. 1402, pp. 21–29. https://doi.org/10.1007/978-3-030-73696-5_3 . http://arxiv.org/abs/2011.03327 Accessed 2025-01-22

[2] Müller, M., Salathé, M., Kummervold, P.E.: Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503 (2020)

[3] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling Instruction-Finetuned Language Models. arXiv (2022). https://doi.org/10.48550/ARXIV.2210.11416 . https://arxiv.org/abs/2210.11416

[4] Montoro-Montarroso, A., Cantón-Correa, J., Rosso, P., Chulvi, B., Panizo-Lledot, A., Huertas-Tato, J., Calvo-Figueras, B., Rementeria, M.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