Classifying COVID-19 Misinformation on Social Media Posts Dataset

João Pedro Araujo, INSPER instituto de ensino e pesquisa, São Paulo, Brasil

I. Dataset

The "Constraint COVID-19 Fake News Dataset" was used for this study. It contains labeled social media posts classified as either real or fake, and the dataset was part of the Constraint AI 2021 shared task for misinformation detection [1]. Its primary business purpose is to assist organizations in detecting and preventing the spread of false information related to the COVID-19 pandemic. The dataset consists of two columns: the tweet text and a label indicating whether the tweet is real or fake.

It used 6,420 samples for the training set and 2,124 samples for the test set. The task was to classify the tweets into their respective categories.

II. Classification Pipeline

The classification pipeline includes the following steps:

A. Preprocessing

Text data was cleaned by removing special characters, URLs, and performing lemmatization [2]. Common stopwords were also removed to ensure that only relevant information was retained.

B. Feature Engineering

Employed Term Frequency-Inverse Document Frequency (TF-IDF) to vectorize the text data. TF-IDF assigns importance to words based on their frequency across the corpus.

C. Model Selection

The Bernoulli Naive Bayes model was used for sentiment analysis due to its simplicity and ability to handle binary features, making it effective for detecting the presence or absence of key sentiment words like "good" or "bad." Its assumption of feature independence works well in text classification, where individual words often drive sentiment [3]. However, the feature independence assumption can be problematic for certain words, as discussed in the conclusion.

III. Evaluation

The dataset was split into 80% training and 20% testing, and the model was evaluated using the balanced accuracy score to account for any class imbalance. The Bernoulli Naive Bayes model achieved a balanced accuracy of 90.09%, indicating a strong ability to distinguish between real and fake COVID-19-related tweets. To better understand the model's decision-making process, we extracted the most important words that contributed to the classification.

The following words were identified as being highly influential for each class:

- **Real tweets:** covid, coronavirus, people, pandemic, trump
- **Fake tweets:** covid, case, new, state, test

IV. Dataset Size

Given the relatively small size of the dataset for sentiment classification, no downsampling or modifications were applied to the dataset [4].

V. Conclusion

In this study, it was developed a classification pipeline to detect COVID-19-related fake news using a Bernoulli Naive Bayes model. The model achieved a balanced accuracy of 90.09%, demonstrating its effectiveness in distinguishing between real and fake tweets. Key words driving classification, such as "covid," "coronavirus," and "pandemic" in real tweets, reflected factual information, while terms like "case," "test," and "death" were more prominent in fake tweets, often used in misleading contexts.

Interestingly, "covid" appeared as an important word in both categories, highlighting the importance of context in distinguishing real from fake news. These insights into word usage can guide future improvements, such as incorporating phrase analysis or topic-specific classifiers to enhance accuracy.

Overall, the identification of these key words provides valuable insights into the linguistic features that differentiate real from fake news. This understanding can guide further improvements to the model, such as enhancing context awareness or refining topic-specific classifiers. Future work could also explore the role of phrases or word sequences in improving classification accuracy, especially for more ambiguous cases.

References

[1] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. PYKL, A. Das, A. Ekbal, S. Akhtar, and T. Chakraborty, "Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts," in *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT)*, Springer, 2021.
2

[2] M. Tomana, R. Tesara, and K. Jezeka, "Influence of Word Normalization on Text Classification," University of West Bohemia, Faculty of Applied Sciences, Plzen, Czech Republic, 2021.
3

[3] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, ”A novel classification approach based on Naïve Bayes for Twitter sentiment analysis,” College of Software, Sungkyunkwan University, Suwon, Korea, 2017.
4

[4] D. Effrosynidis, G. Sylaios, and A. Arampatzis, ”The effect of training data size on disaster classification from Twitter,” Information, vol. 11, no. 12, p. 560, 2020.