Detection of Climate Change Fake News in Mainstream Media Via NLP and Naïve Bayes

INTO GEORGE MASON UNIVERSITY

Hamza Habib | hhabib2@gmu.edu
Sriya Bhavaraju | sbhavara@gmu.edu
Juan Pablo Mendieta | jmendiet@gmu.edu

Introduction

Nowadays, global warming is a topic that is gaining more importance and relevance not only among environmentalists but also within a large variety of engineering fields. However, new large-scale misinformation diffusion threats are appearing to be a novel problem that needs to be addressed. [1] states that this problem is taking place specially across social media, mainstream media and even government announcements. [1] hints about the possible outcomes Natural Language Processing (NLP) could bring. NLP is used alongside other Artificial Intelligence (AI) algorithms in order to obtain more detailed, accurate, and trustful results when analyzing and finding language patterns or word occurrences in written texts. Within this context, one of the most popular algorithms is the Naïve Bayes (NB) which is specially used in projects that involves text classification analysis. According to [2], this preference is due to its simple structure, easy implementation and its overall effectiveness. **Thus, the main purpose of this project is to explain how NLP and the NB algorithm can be used to identify misinformation about climate change in mainstream media sources.**

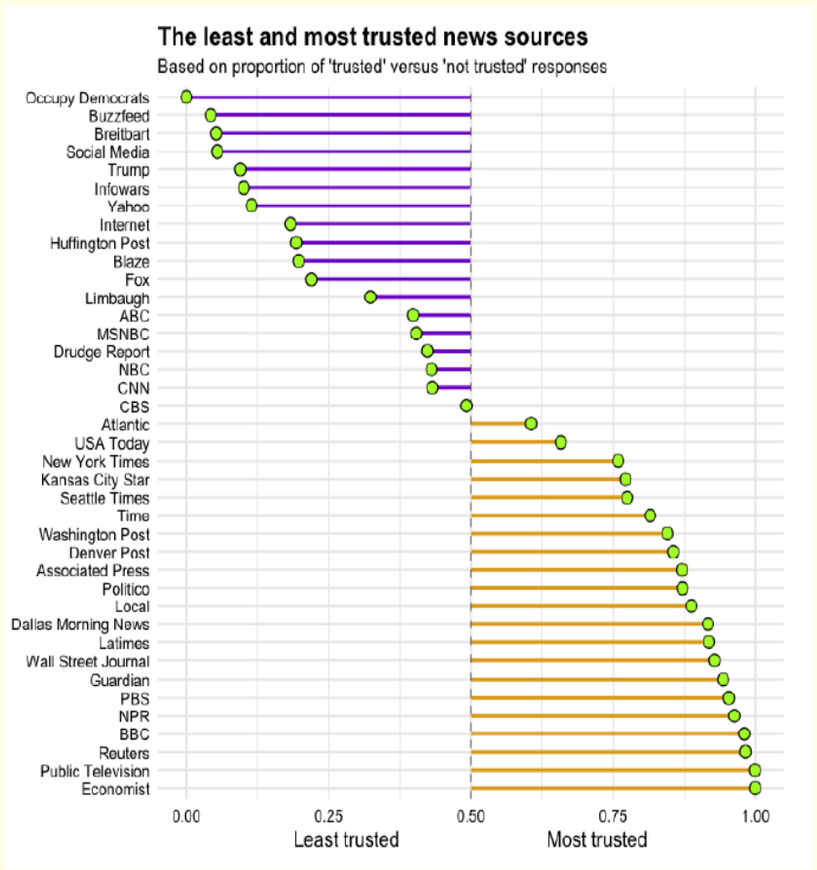

Figure 1. Least and most trusted news sources

Data Preparation

In order to analyze the data obtained from the different sources we need to perform an appropriate cleaning process. Due to the nature of the news available in the websites, the articles can contain useless terms or characters that we will be removing to get a homogeneous style within the data set. In NLP, these useless terms are referred as stop words [3]. For example, in the sentence "Can pollution be measured" the words "can" and "be" are irrelevant to NLP search engines and they can be taken out. As a result, the following actions need to be done prior the analysis:

1. Remove Stop Words
2. Convert to lowercase
3. Remove punctuation
4. Remove numbers

Data Classification

In order to use the Naïve Bayes (NB) algorithm, we will need to separate the data into two parts: test data and train data that is going to be classified. This separation is important since in machine learning problems getting more training data significantly improves the performance of a learning algorithm and therefore increase the accuracy of the results [5]. For this proposal we are going to use 80% of the collected articles for training and 20% for testing. The test data is then matched, and it gets assigned to the group that it belongs.

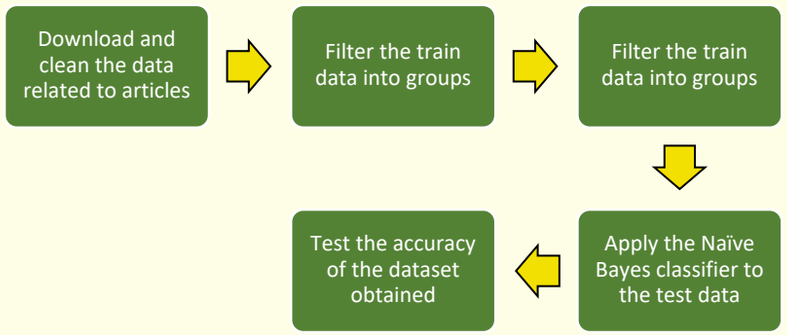

Figure 2. Classification flow

Visualization of Ham and Spam with Word Cloud

Word Clouds are aimed to provide a concise way to summarize the content of any text document [4]. We will use words contained in fake and real news articles previously identified. These words will be drawn with a font size that indicates its frequency throughout the data analyzed. Therefore, the larger a word is visually represented the more common that word was in the resulting datasets.

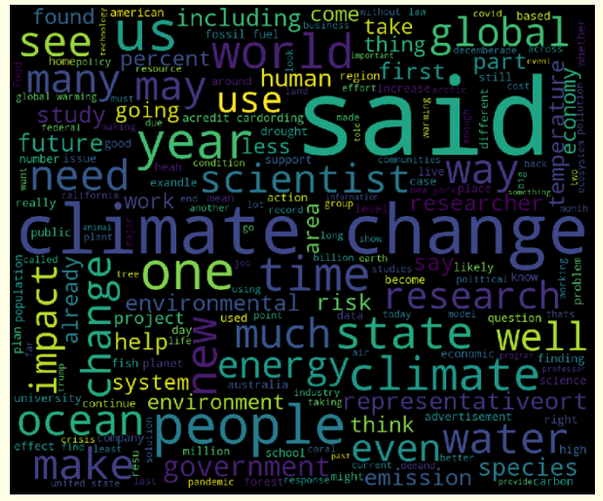

Figure 3. Real News Word Cloud example

References

[1] Farrell, Justin. "The Growth of Climate Change Misinformation in US Philanthropy: Evidence from Natural Language Processing." Environmental Research Letters 14.3 (2019): 1–10. Web.

[2] P. Kaviani and S. Dhotre, "Short Survey on Naive Bayes Algorithm". International Journal of Advance Engineering and Research Development (IJAERD), 30-Nov.-2017.

[3] B. Klatt, K. Krogmann, V. Kuttruff, "Developing Stop Word Lists for Natural Language Program Analysis", Proceedings of the 16th Workshop Software Reengineering and Evolution, pp. 38-39, 2014

[4] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," 2010 IEEE Pacific Visualization Symposium (PacificVis), 2010.

[5] Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 900–903. https://doi.org/10.1109/UKRCON.2017.8100379