Sentiment and Topic Classification Aproaches to African News Articles About COVID-19 and Vaccines

Braian Pita, Undergraduate, University of Houston | bjpitama@cougarnet.uh.edu | @braian_pita

Abstract

• This project analyzes news articles from African sources scraped from three different origins (wayback, GDELT, and our DRC Project). These articles were extracted and used to test which are the best approaches for sentiment and topic models.

• It analyzes the content of news articles through several dictionary approaches to sentiment analysis. This method's purpose is to learn if these models are effective with COVID-19 and vaccine related news, and which are the best approaches to use.

• The paper also uses Naives Bayes, Support Vector Machine (SVM), and the New York Times (NYT) trained topic model classifiers to learn which approaches are best to apply to this data.

Background

• COVID-19 has become a world-wide topic of interest as it affects the lifes of people all over the world.

• The influence of COVID-19 is evident when analysing news articles that talk about the topic., but little research exists when we explore data from African sources.

• As COVID-19 spreads worldwide, it becomes critical to produce vaccines capable of mitigating the damages to both the economy and people's life. Therefore, news articles are an important source of information that can help combat misinformation about vaccination, and create a better perspective on the eyes of the population.

Results

Figure 1. Percent distribution of labels for each approach

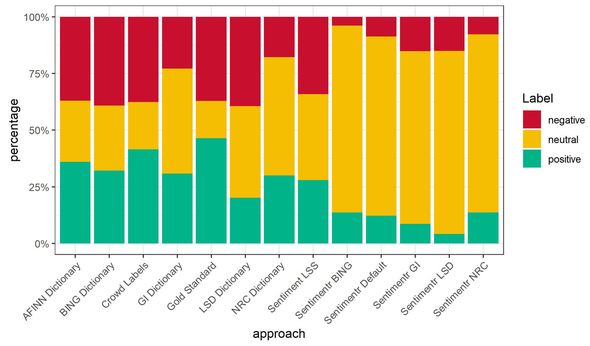


Figure 2. Overall accuracy of sentiment compared to gold standard

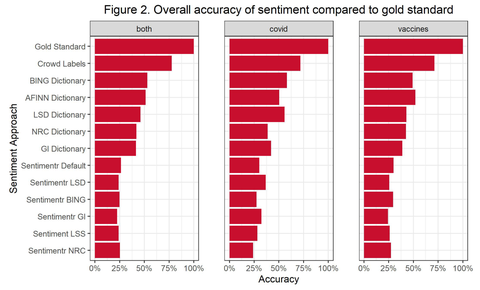


Figure 3. Confussion matrices facet per sentiment approach

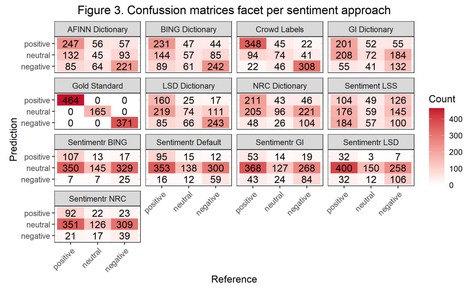


Figure 4. Heatmap comparison of approach numeric values

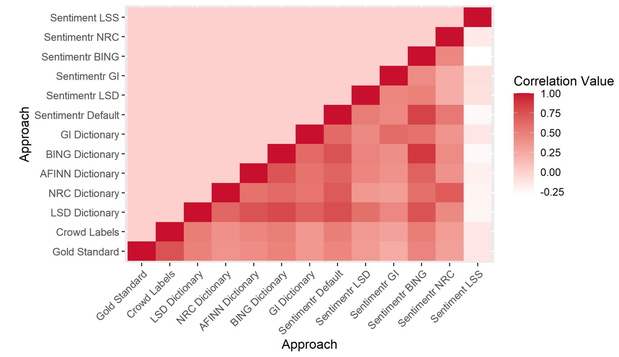


Figure 5. Confussion matrices comparing topic models

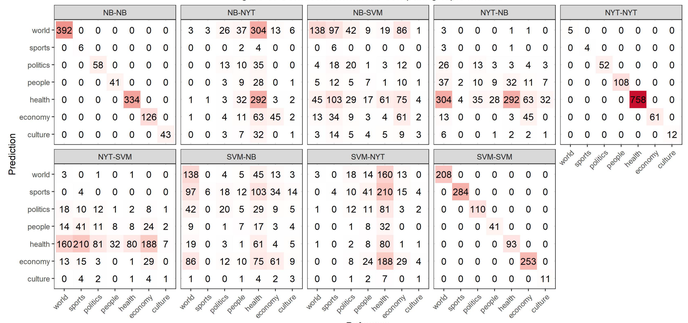
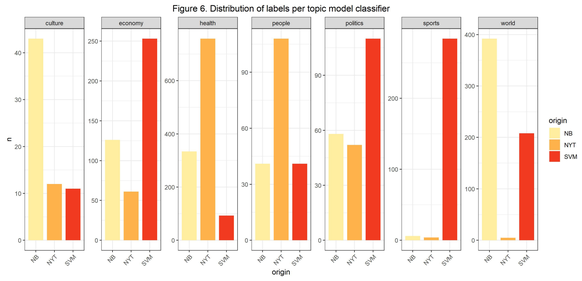


Figure 6. Distribution of labels per topic model classifier



Key findings

#1 The dictionary approaches tested are not effective for our data. The graphs show low accuracy across alal approaches. Furthermore, there is little consistency between them.

#2 Topic classification models had better results. The use of labelled data extracted from the article's link proves the SVM model to have about 0.82 accuracy, making it the most fitting model.

#3 The news articles content had a big impact in the effects of both topic classification and sentiment analysis due to context specific meaning of covid-related words.

Methods

• The news articles were extracted using our own scrapper. The links were taken from African sources used for the "DRC Project" that were found from our own crawler and both "wayback" and "GDELT".

• A gold standard was manually coded by the researchers, and a crowd labelling job was done using Amazon Mechanical Turk.

• For each news article we extracted COVID-19 and vaccine mentions. This allows us to extract a sample for the gold standard with paragraphs that are related to these topics.

Conclusion

• Sentiment analysis through dictionaries is only effective for generic and especific cases. The meaning and sentiment of words can vary based on the context. This makes it ineffective for COVID-19 and vaccine news articles.

• In a similary way, topic classification is influenced by the especificity of our data. Using our own trained models seems to work better than general purpose topic models, which are biased to classify topics as "health" because of their relation to COVID-19 or vaccination.

Email the author:

About the author:

17th Annual Undergraduate Research Day
14 April 2022, Houston, TX