Tracking the Outlook of Strengths in Tweet Opinions about the COVID-19 Vaccine in Tanzania's Twittersphere

Augustine Malija

March 2022

1 Motivation

As the COVID-19 vaccine rolls out in Tanzania, how should the vaccination commission set out its strategy to implement the vaccine policy? Above intention to treat statistical estimates, having a clear outlook of opinion-strengths citizenry hold helps the commission in its strategy. This research seeks to showcase the timely variation in such opinions and uncover drivers for such positions. Data from Tanzania's Twittersphere will be analyzed through a semi-supervised machine learning approach to Natural Language Processing as drawn from a Support Vector Machines flavor.

2 Background

The rise of COVID-19 has led to divided opinions towards it. Provided that it originated from richer countries, it's acceptance in African countries remains an interesting one. Following the discovery of its vaccine, heated discussions have happened in these countries. Most of these discussions occur in social media such as Tweeter, Facebook and Instagram. As a prominent one, Tweeter has been a good platform for such debates. People have been tweeting to register their opinions regarding the vaccine. This communicates stances they take towards the vaccine. It is thus of paramount importance to understand the composition of these stances at a macro level in order to inform policy implementation. Stance detection makes for a technique to aid this. Advances in research methodology have conceived stance detection techniques that identify stances from text. Novel texts subjected to stance detection are in form of tweets. These techniques apply natural language processing and machine learning to classify text to their respective stance classes. Supervised, semi-supervised and unsupervised classifiers can be built to predict what stance each tweet belongs to. Patterns can then be extracted from such predictions to generate visualizations that communicate relationships in the data. By far, Support Vector Machines

(SVMs) have proven to be one of the efficient methods yielding more precise results. Like relatable methods, they require a small sample of tweets to be manually annotated before pre-processing punctuations and building a classifier model to test annotation accuracy. For the sake of this research, a semi supervised classifier model will be built to predict the rest of unlabeled tweets following the aforementioned supervised classifier. To complete this research, the following tasks will need to be done;

- Developing guidelines to annotate sample tweets to their stance classes.
- Constructing both supervised and semi-supervised model architectures.
- Finalize the study by writing a report.

3 Related Work

Contributing to the baseline of this study, Augustine Malija detected stances from climate science tweets in continental Europe and the Americas. 10 million English tweets containing the word "climate change" were accessed from Tweepy from which 50,000 were subjected for analysis. 2000 tweets were annotated to "in favor, against and neutral classes" before being subjected to a supervised learning model for accuracy. Support Vector Machines(SVM)'s supervised and semi supervised baseline models were built. Moreover, Bidirectional Encoder Representations from Transformers (BERT) deep learning model was further developed. Results suggest that major events in the climate science calendar play a big role in changing stances from negative to positive stances. Informational tweets about climate science classified as neutral recorded a weak but positive effect in nudging negative stances.

Kyle Glandt et al detects stances from COVID-19 tweets that discussed topics pertaining to the pandemic. 30,331,993 tweets from February 27th until August 20th 2020 were collected. These tweets contained "coronavirus", "covid-19" and "corona virus" keywords lockdown, socialdistancing and washhands hashtags. 6,133 tweets were annotated to "in-favor, against and neutral" classes before being subjected to classification models to assess baseline performance. Baseline models employed include BiLSTM, Kim-CNN, TAN, ATGRU, GCAE and BERT. Self-training and domain adaptation approaches are further used to classify tweets that have not been assigned to classes. Results show that the pre-trained COVID-19-Twitter-BERT model constitute strong results.

4 Data

Twitter offers access to data for research purposes directly through their research platform. This data is accessible for free subject to the submission of a research proposal. The focus of this research purely aligns to data access requirements. These requirements are; being a researcher at an academic institution, having clearly defined research objectives and using the acquired twitter data for non-commercial purposes. Tweets including words "chanjo ya uviko", "chanjo",

"COVID vaccine", "COVID-19" or "chanjo ya korona" will be requested. To catch clear patterns in the discussion, tweets from April 2020 to date will be requested. This is because Tanzania started to roll out the vaccine during this time.

5 Methods

A novel Support Vector Machines methodology will be employed to detect vaccine stances. Both supervised and semi-supervised models will be developed to classify these stances. From millions of tweets during this two-year time, a random sample of 50,000 tweets will be drawn to form a dataset subject to analysis. Before model development, a sample of 2000 tweets will be annotated to three stances, "in favor, against and neutral". To determine which stance class a tweet falls to, an annotation guideline will be developed. A specialized annotation platform, IO Annotator, will be contracted to allow the use of their platform for labelling tweets. The annotated tweets will then be pre-processed to a clean format ready to be subjected to classifier models. A supervised classifier will then be built to determine the accuracy level of the annotation process. With a satisfactory annotation accuracy, a semi supervised model will be built to predict stances for un-annotated tweets. From here operators to extract descriptive statistics necessary to offer insights on trends will be applied. These will also offer a guide towards developing visualizations that communicate a story the data tells.

6 Research Translation

To reach necessary decision makers, dissemination of this research will use the following platforms.

• The ministry for heath and vaccination commission. Presenting to these bodies will reach key decision and policy makers who are involved in strategizing the distribution of the vaccine. This will provide them with a clue on stage wise investments in order to ensure an equitable distribution doing away with redundant vaccines and the level of effort needed to increase awareness.

• Research Talks. To enhance knowledge, the results of this research will be disseminated to research talks such as Utafiti Wetu. Researchers and students with interest get to learn and offer feedback on the design of methods of this research. It is also from here that researchers get to come up with how they can advance the work further.

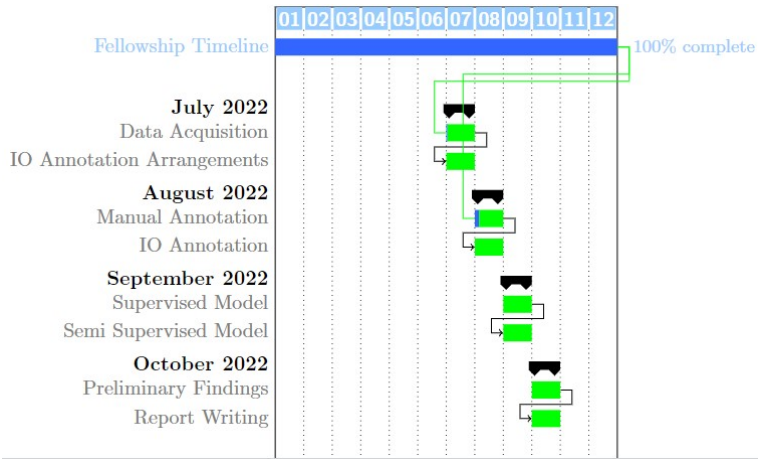

Figure 1: Timeline