COVID-19 South Africa (COVID19ZA) Consortium

Andani Madodonga, **Yolanda**

1. Motivation

- Covid-19 has been declared a global pandemic in 2020
- There has been an increase in misinformation and speculations about Covid-19 in social media platforms.
- The misinformation and speculations spread in social media platform mislead the society and impact the Health institution/Government
- Health Institution /Government have significant gap in risk communication strategies via social media to address the society about Covid-19.
- It is important for the government/Health institution to understand the knowledge, behaviour and beliefs of the society about Covid-19 so that they can formulate communication strategies ,to effectively communicate and understand societies perception around the pandemic.

I. Objective/AIM

The purpose of this project is to use data science and statistical techniques on a Microblog dataset, twitter, to address the following objectives:

- Identify and cluster the dataset into a local and global category.
- Identify, describe, and quantify the spread of information between users in the dataset.
- Perform sentiment analysis across various groups of identified spreads of information.

2. Methods & Results

I. Exploratory data analysis:

- The dataset was cleaned and feature importance algorithm was applied to remove insignificant columns as part of pre-processing.
- Columns with 80% missing values were also removed
- Non English tweets were translated into English
- Undetected languages were excluded
- Data was split into training, test and validation in the ratio of 6:2:2 respectively.

II. Modelling:

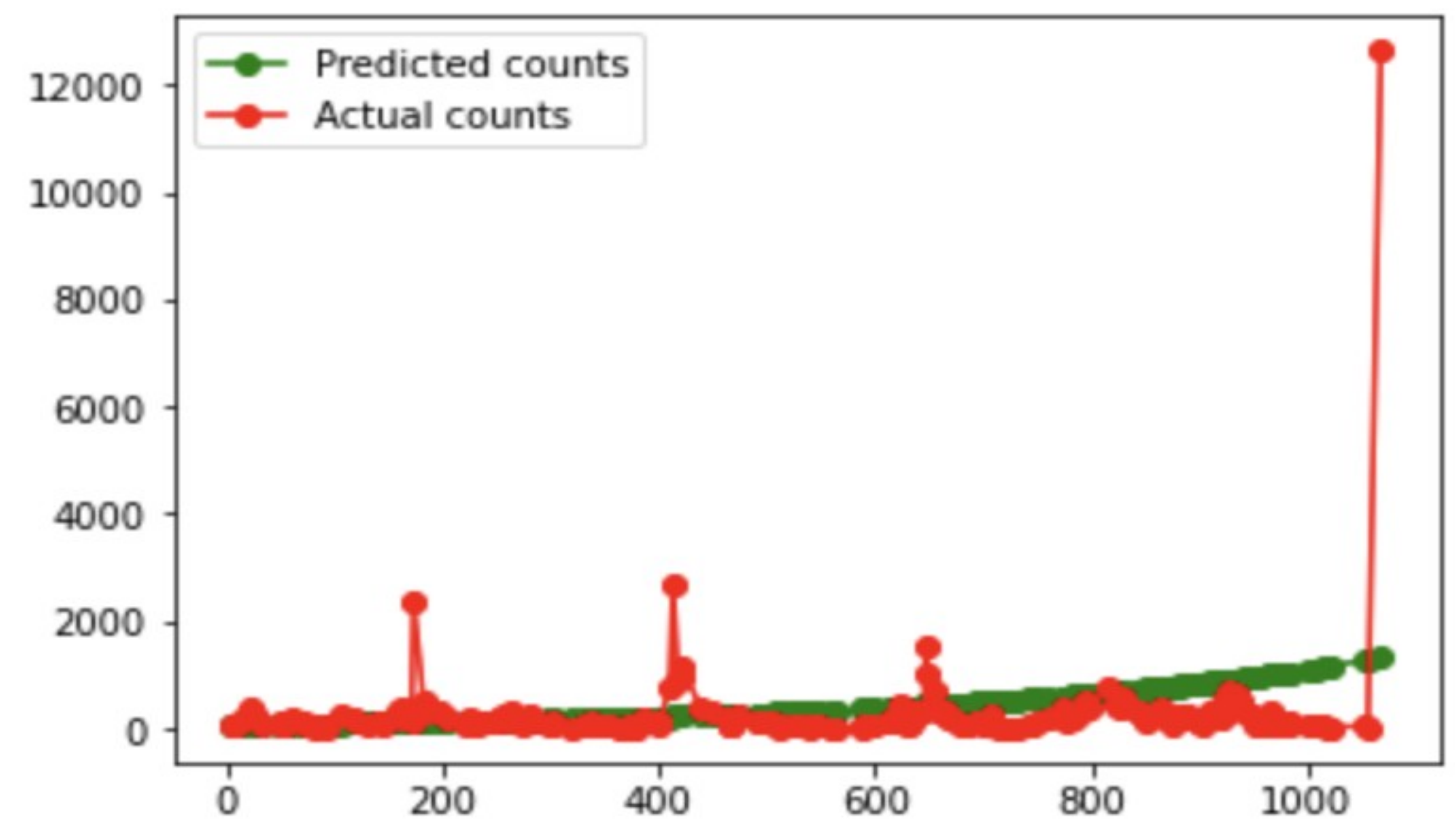
- A new dataset was created to answer and build some of the models. This dataset was created through feature engineering of the original dataset.
- Topic modelling was used to cluster similar microblogs together.
- Distribution was fit in the time series data per topic with respect to retweeted counts to determine if they follow any of the known statistical distributions.
- Several Machine learning models were trained to address the objectives, below are best performing models for specific goals:
  - **I. Random Forest**
    - Clustering local and international microblogs
    - Identifying trending microblogs.
  - **II. Xgboost**
    - Classify influential twitter users
    - Determine rate of transmission of microblogs.
- Pre-trained model VanderSentiment was utilised to

III. RESULTS

- Accuracy and F1 were used to select the best models for classification tasks and MSE was used for regression tasks Below is the Models' performances.

Cluster local vs Internation- Random Forest			
Metrics	Test	Validation	Stability
Accuracy	96%	97%	99%
F1	96%	97%	99%
Classify if Topic will Trend- Random Forest			
Metrics	Test	Validation	Stability
Accuracy	97%	97%	91%
F1	98%	98%	97%
Classify if user will be influencer-Xgboost			
Metrics	Test	Validation	Stability
Accuracy	100%	100%	100%
F1	100%	100%	100%
hourly rate of transmission-Xgboost			
Metrics	Test	Validation	
MSE	3.77	3	

- Poisson distribution was found to be the best for distribution for count/retweet rate of transmission per hour as can be seen below plot.

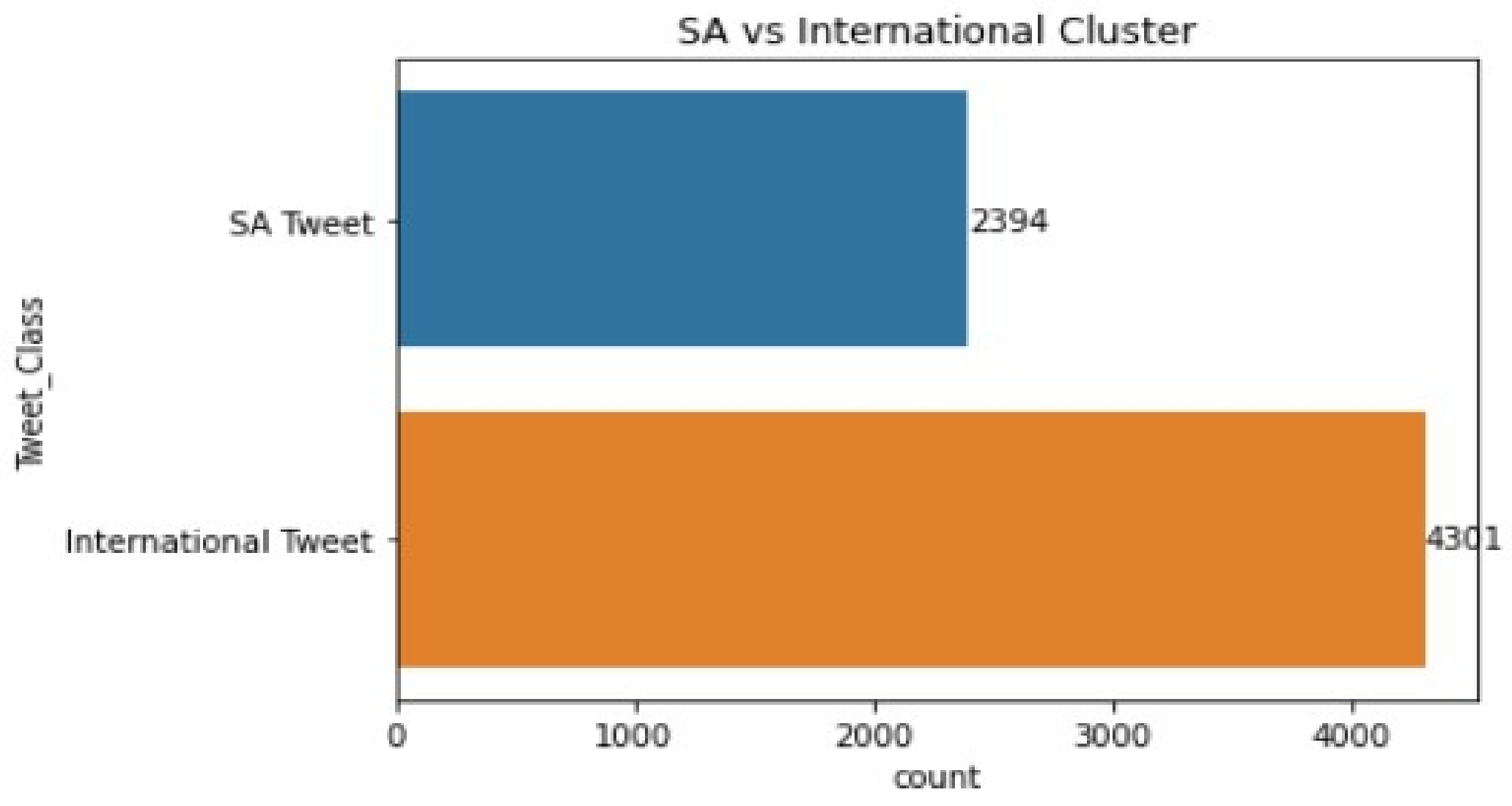


3. Model Deployment:

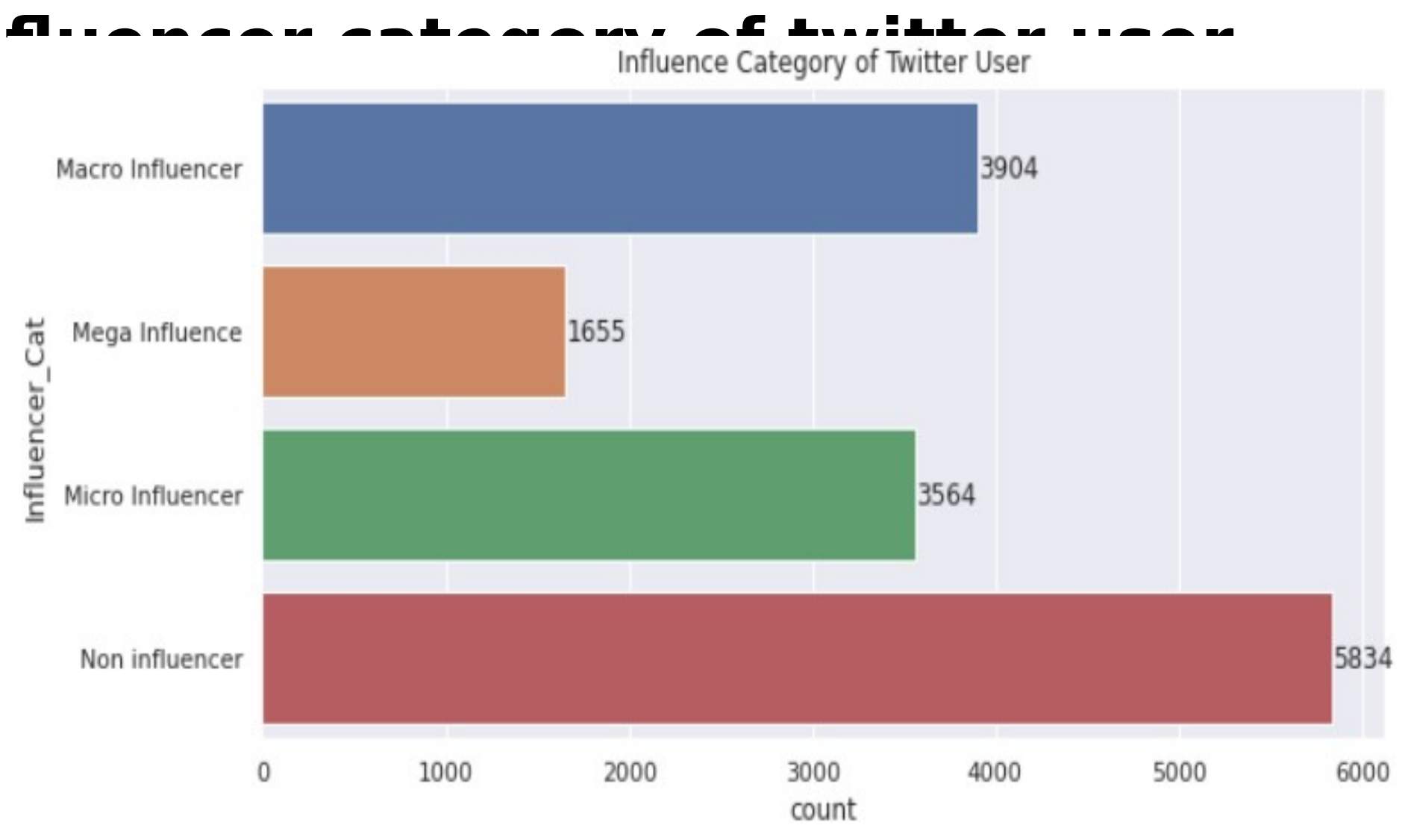
- Models were pushed into github and deployed to Streamlit
- Models predictions are working as expected
- Visuals produced from the models are incorporated into Streamlit App.
- To view and access the app please scan the QR code

4. Visualization

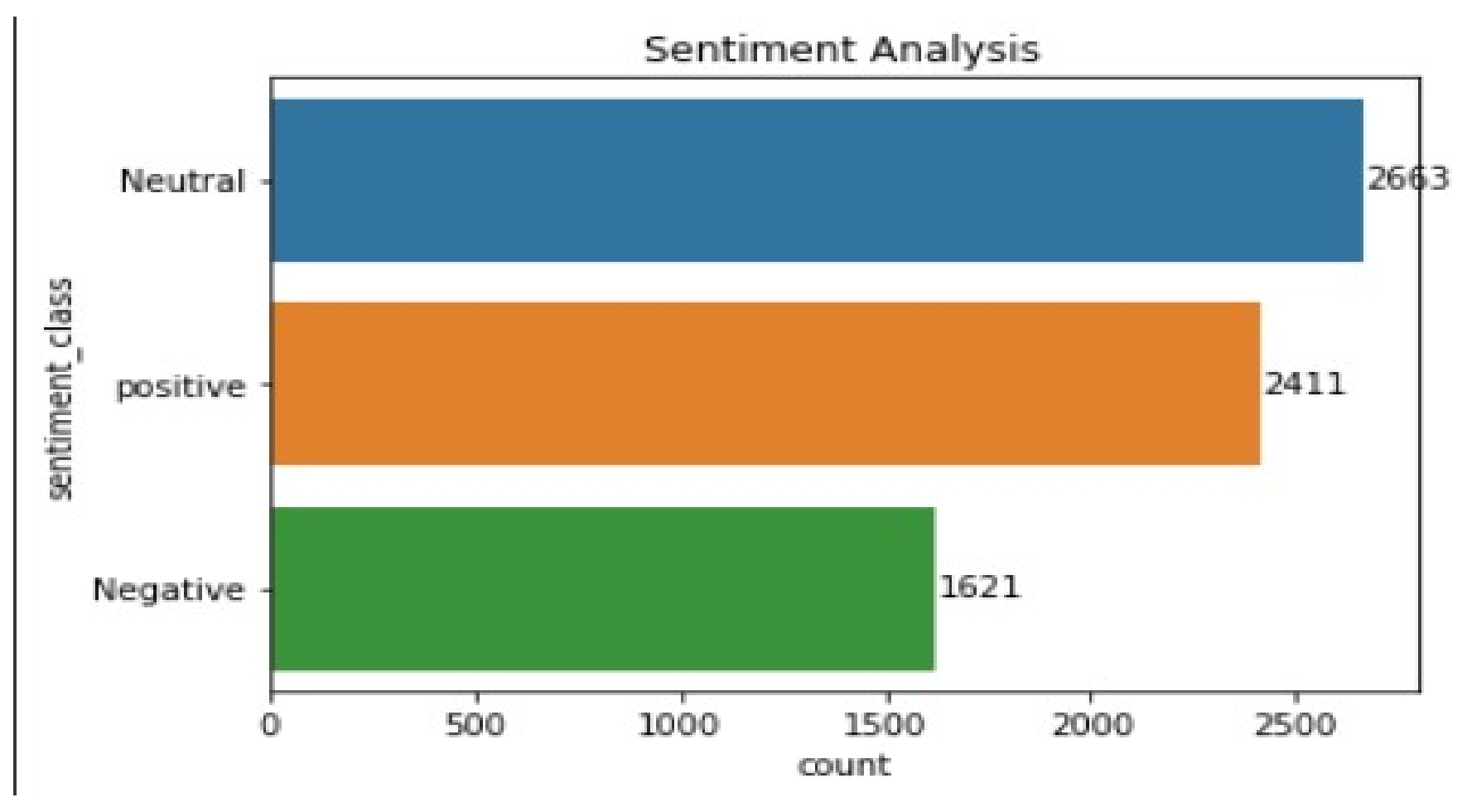
- **SA vs International cluster**



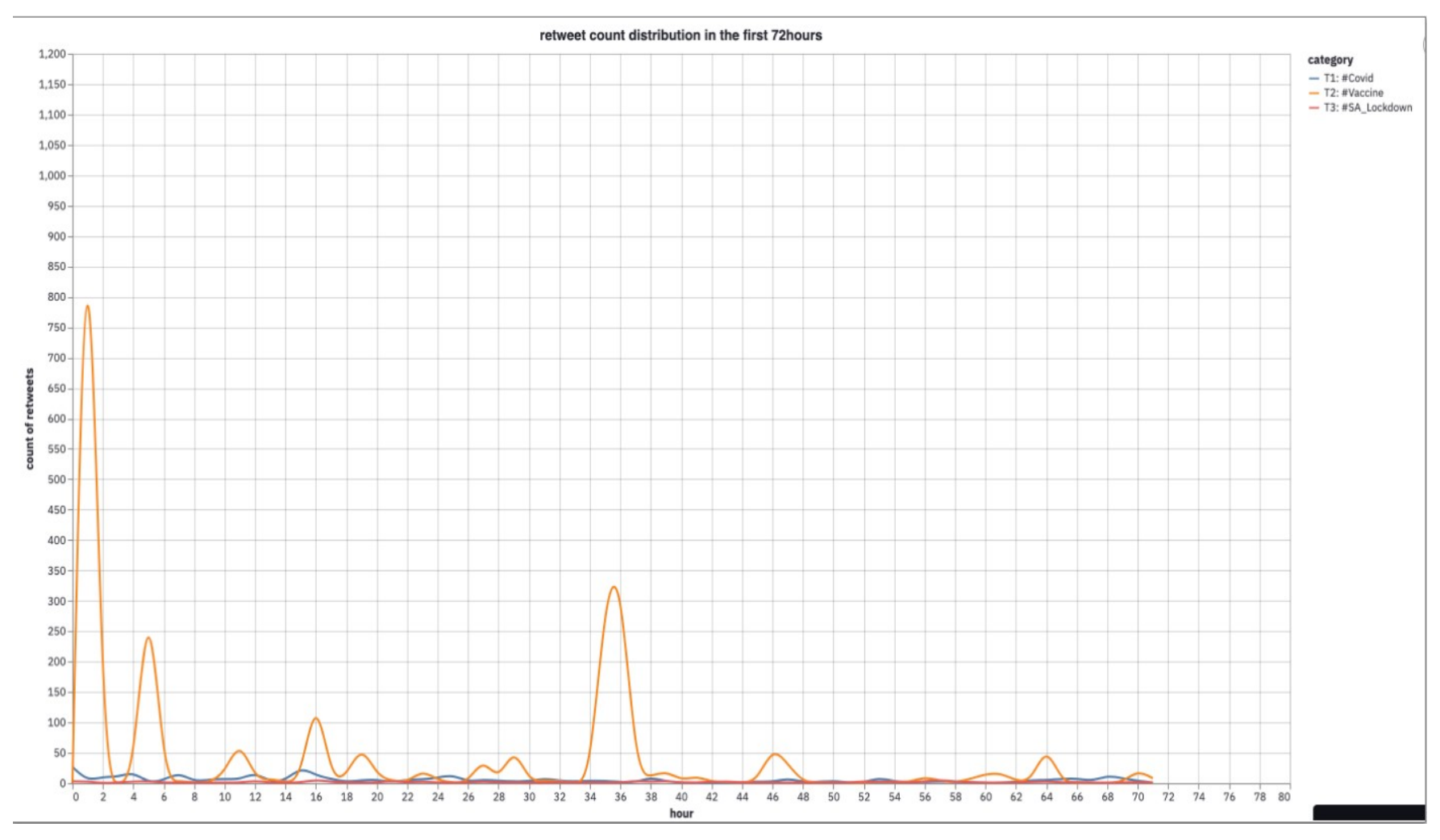
- **Influencer category of twitter user**



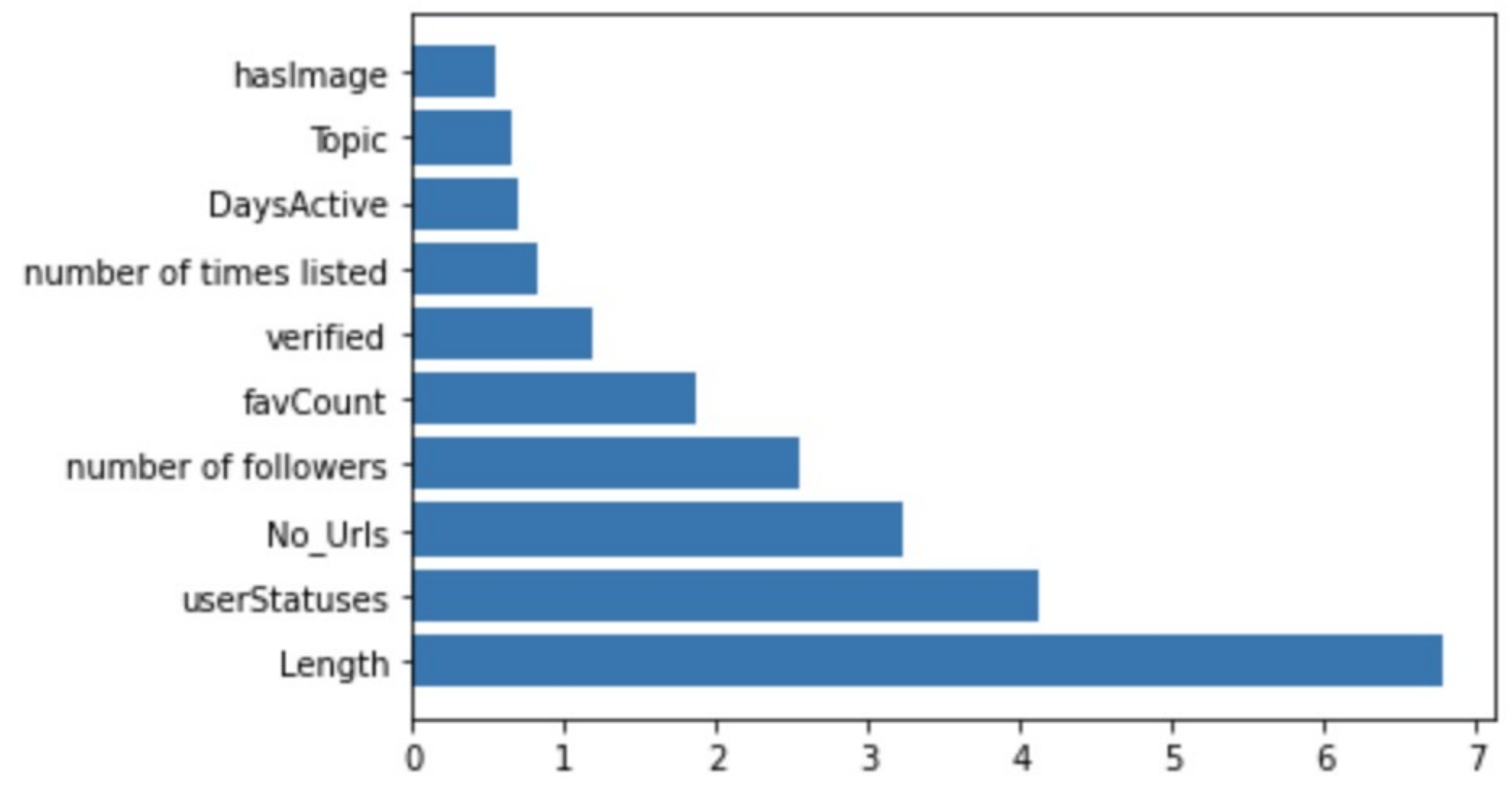
- **Sentiment Analysis**



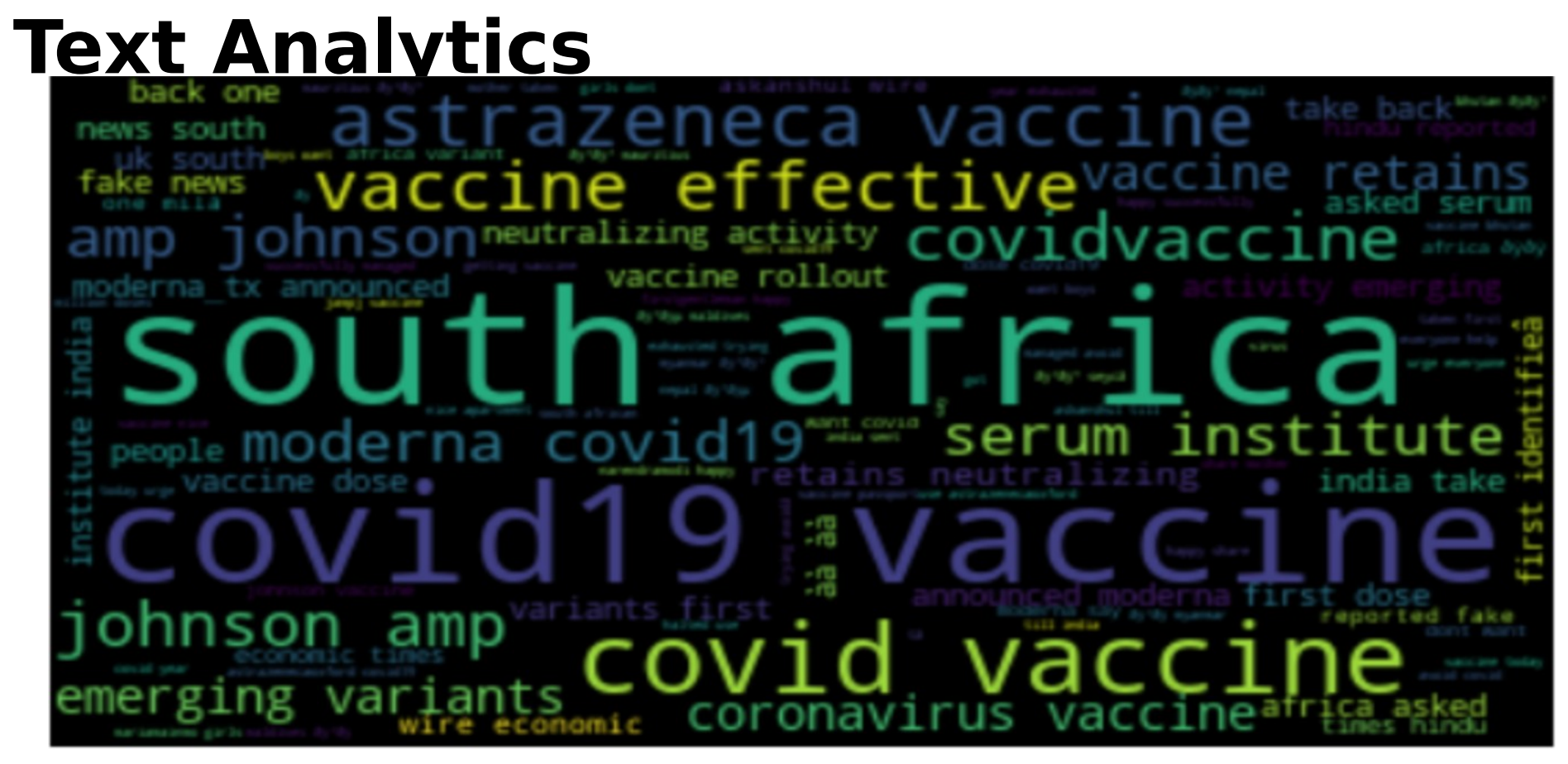
- **Distribution of the hourly Rate of transmission.**



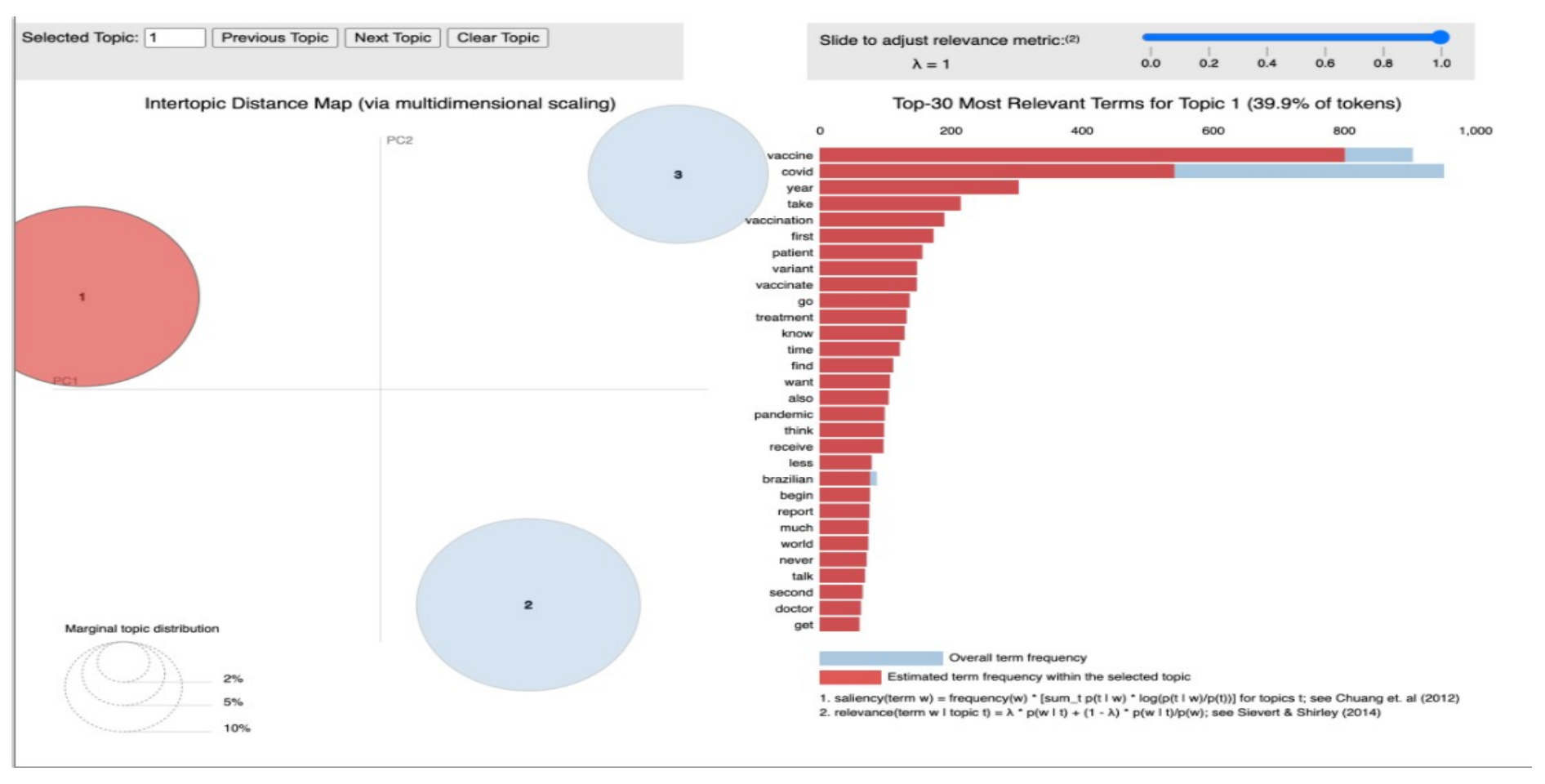
- **Trending microblogs per topic - Feature I**



- **Text Analytics**



Word cloud of top words in the SA microblog



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Most Salient words chart based on Topic modelling

4.Next steps

- VanderSentiment was not trained on Covid related tweets ,hence likely to be bias
- Accurate classification of sarcastic tweets
- Improve on model to classify local and international tweets

Scan me to go to App

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)