Sentiment Analysis of Kenyan Healthcare Sector

1. Business Understanding

Overview

The health care sector in Kenya has been facing a lot of criticism over its performance in the last two years. This ranges from corrupt dealings at the "Afya" house, Operations by KEMSA, threats of strike by medical workers, handling of the recent coronavirus pandemic and more recently the loss of life by a doctor to coronavirus over lack of protective medical equipment.

Problem Statement

The Healthcare Sector has been a subject of criticism over the last few years calling for need to evaluate what Kenyans feel about its sub sectors

The project will therefore focus on the following objectives:

1. To classify tweets related to the kenyan healthcare sector as either Pharmaceuticals, Medical Practitioners, Health Facilities, Medical Insurance or Hospital Supplies
2. To perform sentiment analysis of the tweets on each class as either positive, negative or neutral.

Project Justification

The Ministry of Health has a mission to empower the Kenyan people to live healthy using the primary health care approach. This has seen the ministry receive the lion's share of the budget financed by taxes from the Kenyan citizens. Since a considerable amount of resources is channelled to the ministry, establishing what Kenyans feel about the HealthCare system is necessary.

Business success criteria

Above 80% accuracy in classification by the model

Assessing the situation

Requirements

1. Team work.
2. Collaboration amongst members.
3. Good internet connections for meetings.
4. Power connection.

Resource inventory: datasets and softwares used

1. We have one dataset scraped and compiled from twitter.
2. We will use Google Colab for our Analysis of the data.
3. We will use Google for Research purposes.

Constraints

1. Scraping the data was time consuming and a challenge as twitter only allows a certain number of tweets to be scrapped per day.
2. Connection Problems and Power Problems.

Risks and Contingencies

Risk	Contingency
Unavailability of a Team Member	If a team member is unavailable for any scheduled meeting, he or she will communicate to the rest of the team and the meeting will be rescheduled.
Unavailability of one or several team members.	If a team member is unavailable for a task he or she is scheduled to perform, he or she should communicate to the team leader so that task may be reassigned to another member.
Possible Crashing of Colab Notebook	If this happens, we will use Jupyter Notebook on our local machines.

Implementation Plan

To implement this analysis, we will need to follow a laid out plan in order to finish our analysis on time.

Phase	Time-Frame	Team Member
Formulation of Research Question	2 hours	
Data Scraping	1 week	
Data Understanding	2 hours	
Data Preparation	1 day	
Data Analysis	3 days	
Recommendations	2 hours	
Presentation of Findings	1 hour	T.B.D.

2. Data Understanding

Data Collection

Kenyans have particularly been noted to be good at airing their grievances over the Twitter platform notably through its "KOT" brigade (Kenyans on Twitter). This therefore, makes this platform a rich source of information which shall be scrapped for use in this project.

Data Description

The dataset has 14 columns. Some of the relevant features of this data are:

Column	Description
user	The username of the person on twitter
tweet	The tweet statement
location	Location of the user
description	Description of the user's account
friends_count	Number of friends a user has
followers_count	Number of followers a user has
statuses_count	Number of statuses a user has upto the date of the tweet
tweet_date	The date the tweet was posted
retweet_count	Number of retweets the tweet has gotten so far
likes	Number of likes the tweet has
hashtags	The hashtags featured in the tweet
labels	The specific healthcare sector the tweet is talking about

Data Preparation

In order to use the data collected the following were performed on the data:

1. Making a copy of the data for use in the analysis
2. Converting the column names to lowercase state for uniformity

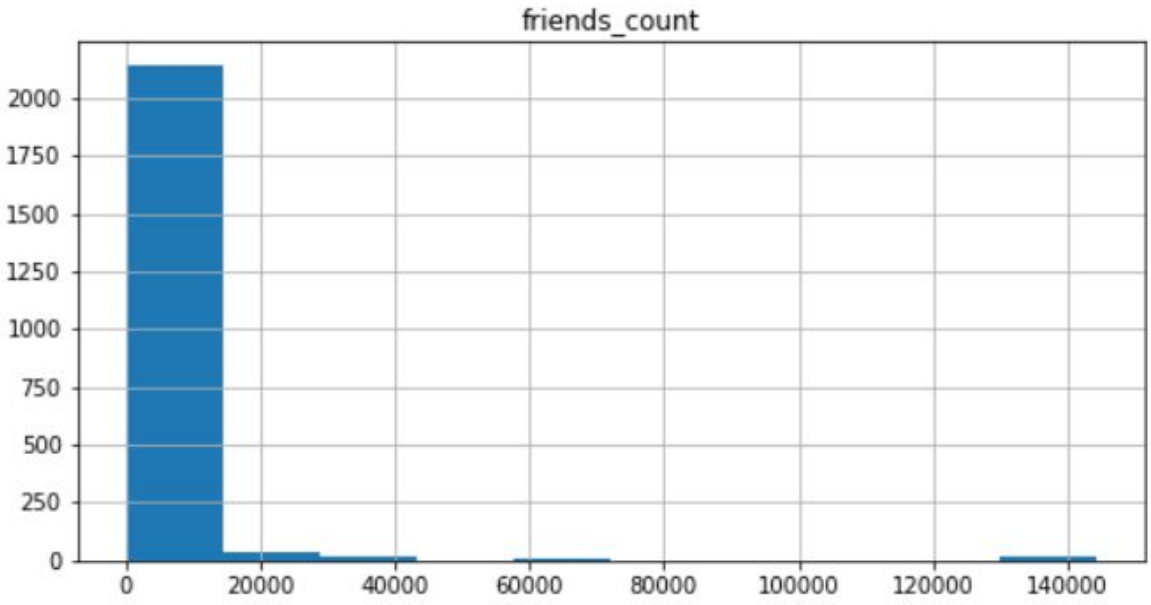
3. Checking for the presence of duplicates
4. Dropping of unnecessary columns such as `'1'`,`'user'`,`'description'`, `'hashtags'`. Dropping the 'user' column is particularly important in order to do away with personal identification data.
5. Testing for the presence and the dropping of null values
6. Correcting of the labels to be used in the data
7. Saving the cleaned data file

3. Exploratory Data Analysis Finding

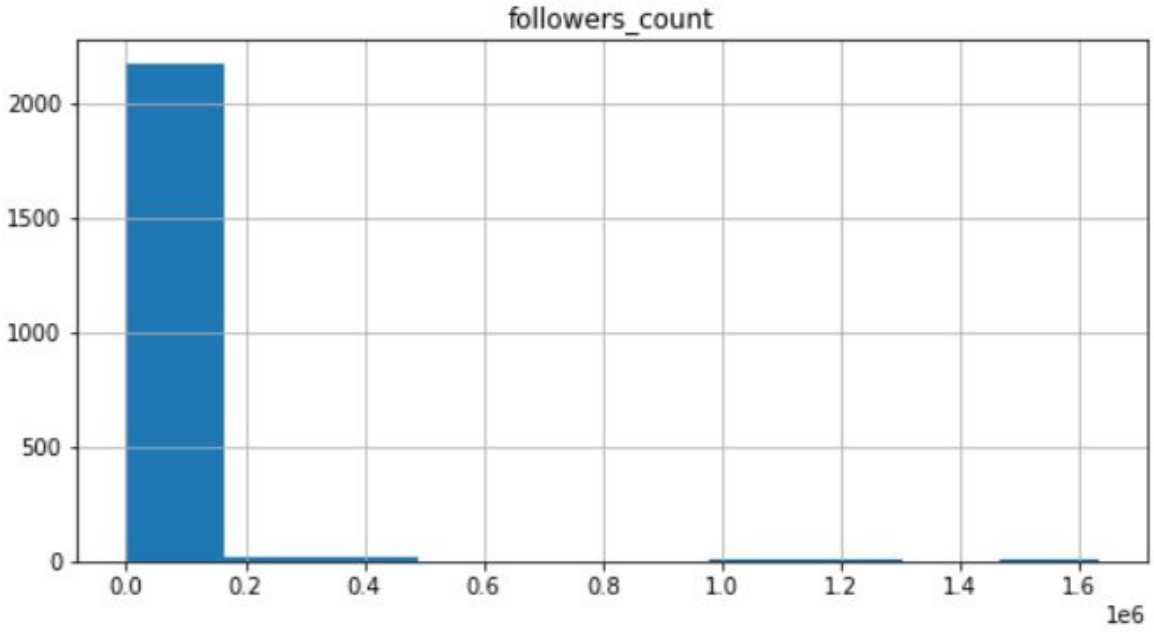
The Data Science team did an Exploratory Data Analysis just to see the trends and distribution of the data. The analysis was done on a Colab notebook Link.

Univariate Analysis Findings

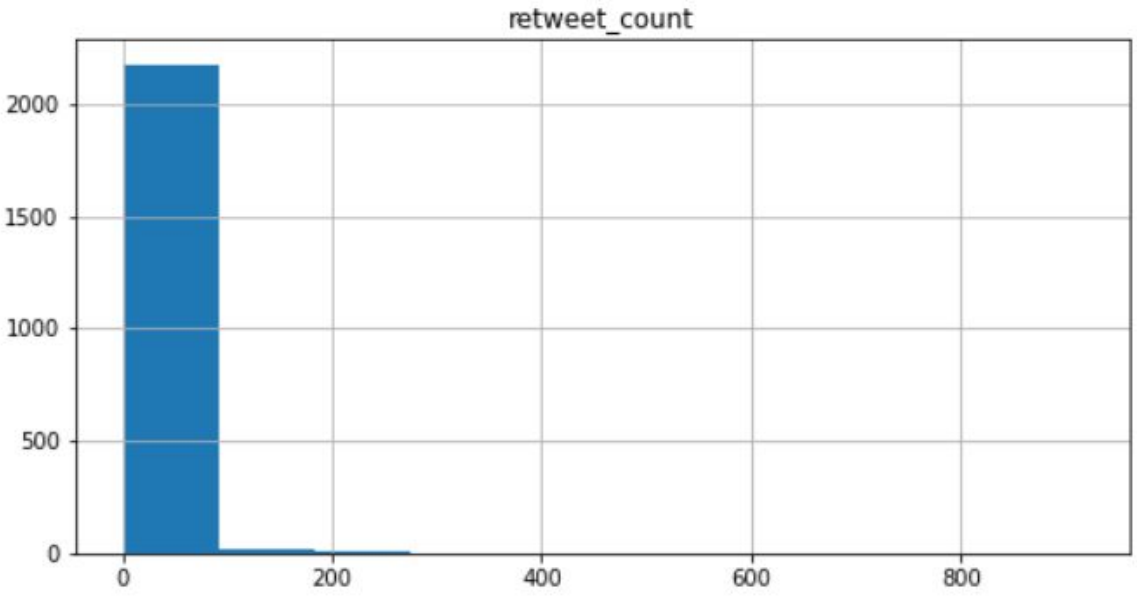
1. Over 90% of the accounts considered had between 0 - 15,000 friends making the data to be skewed to the right



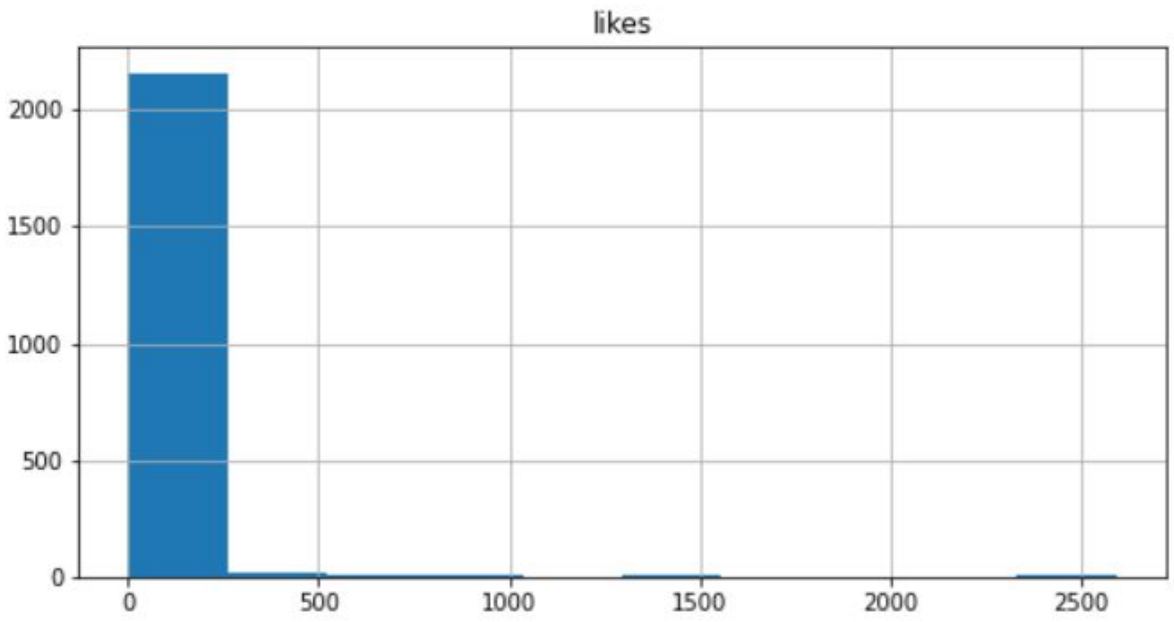
2. Less than 5% of the accounts considered had a following of over 1,000,000 persons making the data to be skewed to the right



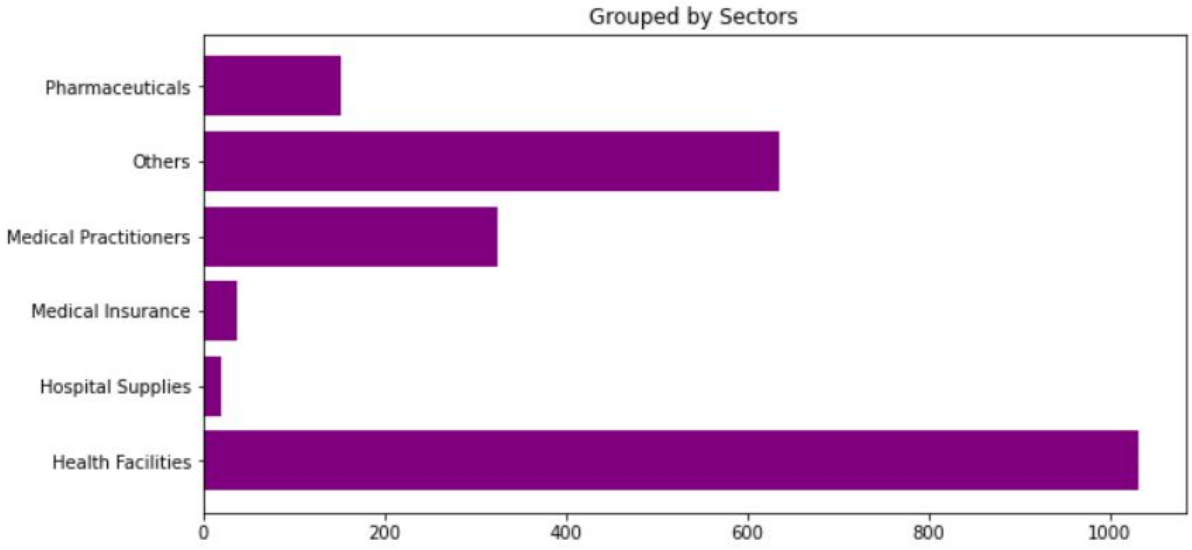
3. Less than 2% tweets were retweeted over 200 times thus the data was positively skewed



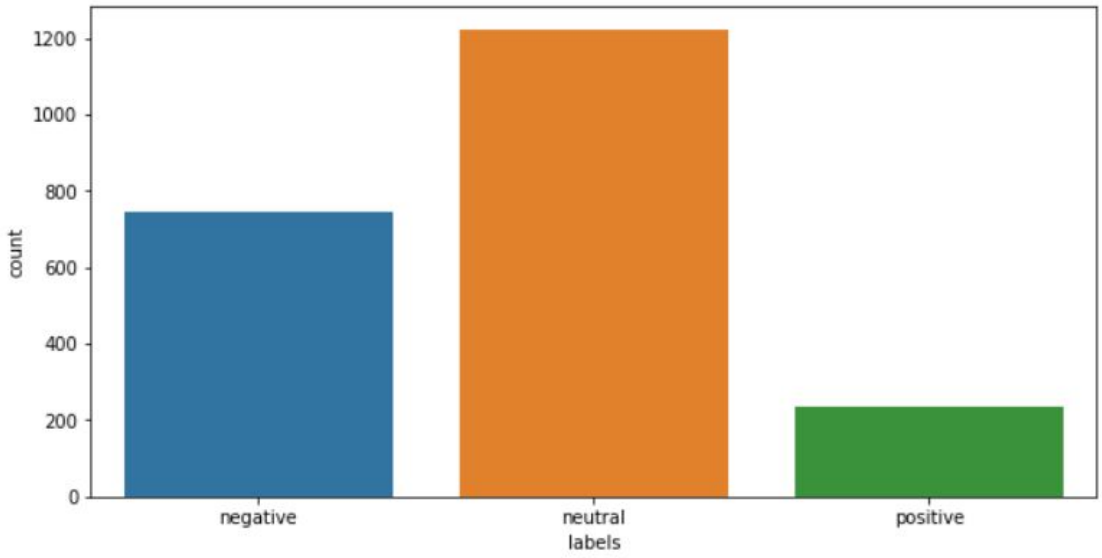
4. Less than 5% of the tweets received over 500 likes thus the data was skewed to the right.



5. Majority of the tweets were on Health Facilities

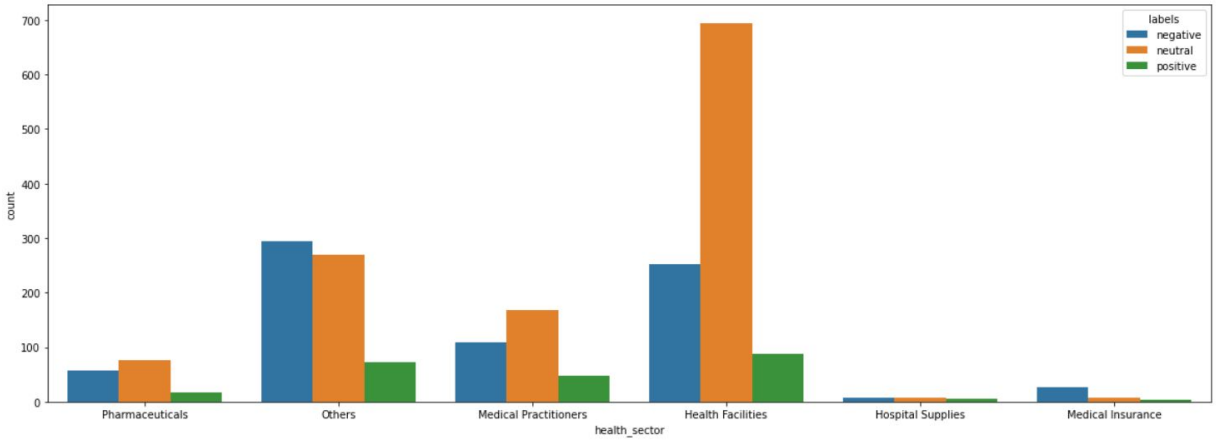


6. Over 50% of the tweets were Neutral, followed by Negative tweets and then the Positive ones.



Bivariate Analysis Findings

1. Of the five healthcare sub-sectors considered, namely Pharmaceuticals, Medical Practitioners, Health Facilities, Hospital supplies and Medical Insurance, **Medical Insurance** received the most negative reviews in comparison to its neutral and positive review it received.



2. There is a strong positive correlation (0.87) between the number of likes received and the number of retweets done.

3. There is also a moderate positive correlation between likes received on a tweet and the count of the number of followers.

4. Building the Model

Data Preprocessing

For modelling purposes the following actions were further performed on the data:

1. Selecting columns to be fed into the model. The following three columns were considered for the model `'labels'`, `'tweet'`, `'health_sector'`.
2. Further cleaning - removal of urls, retweets and cc, hashtags, user mention, emojis, html tags, extra spaces, punctuation marks, conversion to lowercase, lemmatization, removal of stopwords and converting the processed list words to string
3. Translation of words in Swahili to English
4. Transformation of the data into occurrences through the count vector.
5. Label encoding of the labels

Models

Various models were considered including:

1. Naive Bayes
2. Random Forest
3. ANN - Aspect Based Model
4. ANN - Sentiment Analysis
5. LSTM

Model	Accuracy	Precision	Recall
Naive Bayes	22.45%	55%	22%
Random Forest	55%	52%	55%
ANN - Aspect Based	72.96%	72%	73%
ANN - Sentiment Analysis	58.86%	59%	59%
LSTM	59%		

Deployment

The model was saved as a file `my_model.h5` to be deployed through StreamLit.

5. Conclusion

The Aspect Based Model with ANN had the best result and is therefore best preferred for deployment