Study of the Yahoo-yahoo Hash-tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms

Adebayo Abayomi-Alli
FUNAAB
Olusola Abayomi-Alli
Kaunas University of Technology: Kauno Technologijos Universitetas
Sanjay Misra (sanjay.misra@covenantuniversity.edu.ng)
Covenant University https://orcid.org/0000-0002-3556-9331
Luis Fernandez-Sanz
Universidad de Alcala de Henares

Research Article

Keywords: cloud computing, Opinion mining, Twitter, cyber-crime, content analysis, text classification

Posted Date: April 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-354801/v1

Version of Record: A version of this preprint was published at Information on March 15th, 2022. See the published version at https://doi.org/10.3390/info13030152.

Study of The Yahoo-Yahoo Hash-Tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms

Adebayo Abayomi-Alli[1], Olusola Abayomi-Alli[2], Sanjay Misra[3*](Sr. Member, IEEE), Luis Fernandez-Sanz[4]

[1]Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria.
[2]Department of Software Engineering, Kaunas University of Technology, Kaunas, Lithuania.
[3]Department of Computer Engineering, Atilim University, Ankara, Turkey and Department of Electrical and Information Engineering, Covenant University Ota, Nigeria.
[4]Department of Computer Science, University of Alcala, Madrid, Spain

abayomialla@funaab.edu.ng , olusola.abayomi-alli@ktu.edu , *sanjay.misra@covenantuniversity.edu.ng , luis.fernandez.sanz@uah.es

*Corresponding Author

Declaration:

- Availability of data and material- We are not providing any data and supplement material.
- Competing interests- We declare that we don't have any potential competing interests. Funding- No funding available for this work.
- Authors' contributions- All authors made an almost equal contribution.
- Acknowledgments- We acknowledge the support of affiliated universities of each author for providing a research environment.

Study of The Yahoo-Yahoo Hash-Tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms

ABSTRACT

Background

Social media opinion has become a medium to quickly access large, valuable, and rich details of information on any subject matter within a short period. Twitter being a social microblog site, generate over 330 million tweets monthly across different countries. Analyzing trending topics on Twitter presents opportunities to extract meaningful insight into different opinions on various issues.

Aim

This study aims to gain insights into the trending yahoo-yahoo topic on Twitter using content analysis of selected historical tweets.

Methodology

The widgets and workflow engine in the Orange Data mining toolbox were employed for all the text mining tasks. 5500 tweets were collected from Twitter using the 'yahoo yahoo' hashtag. The corpus was pre-processed using a pre-trained tweet tokenizer, Valence Aware Dictionary for Sentiment Reasoning (VADER) was used for the sentiment and opinion mining, Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) was used for topic modeling. In contrast, Multidimensional scaling (MDS) was used to visualize the modeled topics.

Results

Results showed that "yahoo" appeared in the corpus 9555 times, 175 unique tweets were returned after duplicate removal. Contrary to expectation, Spain had the highest number of participants tweeting on the 'yahoo yahoo' topic within the period. The result of Vader sentiment analysis returned 35.85%, 24.53%, 15.09%, and 24.53%, negative, neutral, no-zone, and positive sentiment tweets, respectively. The word yahoo was highly representative of the LDA topics 1, 3, 4, 6, and LSI topic 1.

Conclusion

It can be concluded that emojis are even more representative of the sentiments in tweets faster than the textual contents. Also, despite popular belief, a significant number of youths regard cybercrime as a detriment to society.

Keywords: cloud computing, Opinion mining, Twitter, cyber-crime, content analysis, text classification

1. INTRODUCTION

The continuous rise in Internet technology and various social media platforms has made it possible for effective communication and interaction among various people from diverse social and cultural backgrounds (Appel *et al.*, 2020). This growth in technology advancement has also introduced some downside once wrongly applied, known as cyber-crime (Hariani and Riadi, 2017). Social media has become a significant aspect of online activity and plays a crucial part in cybercrime and cyber terrorism-related operations (Boyer, 2014). Cyber-crime, which is one of the popular forms of deviance among youth in Nigeria, is still a serious problem affecting the country's image (Ojedokun & Eraye, 2012; Tade & Aliyu, 2011). The perpetrators are received by some people and social institutions when they make illegitimate money; hence, the increasing justification of illegality (Adeniran, 2008; Ninalowo, 2016). The phrase Yahoo-Yahoo originated from the use of Yahoo emails and Yahoo instant messenger as a dominant medium of communication between perpetrators and victims (Lazarus & Okolorie, 2019). This popular term refers to the activities that entail using computers, phones, and the Internet to defraud unsuspecting victims, especially those outside the country. The likelihood of fraudulent users integrating new approaches without necessarily applying extensive technical knowledge on the Internet could result in a fraud activity (Rossy and Ribaux, 2020).

The rising popularity of cyber-crime in Nigeria (Longe *et al.*, 2009) can be connected to the current state of economic instability, high unemployment rate among able-bodied youths, erosion of traditional values of integrity, and quick-money syndrome, etc. To curb these illegal activities, institutions such as the Economic and Financial Crimes Commission (EFCC) were established in Nigeria and have recorded several arrests and prosecution of cyber-crime suspects (Omorogbomwan, 2018). However, it is expected that with the apprehensions and prosecutions,

more understanding of the "modus operandi" of culprits will emerge. However, crime may not be static as suspects could adopt new methods when the old ones are known to the people and law enforcement agencies. Cyber-crime has gone being the notorious 419 email and SMS scams (Abayomi-Alli *et al.*, 2019) to applying more sophisticated methods making social media users vulnerable (AUC, 2016). Recently, social media platforms such as Facebook, Instagram, Twitter, Google+, and Pinterest are becoming popular for crucial data sources in research studies relating to sentiment analysis (Gupta et al., 2016; Kunwar and Sharma, 2016). It can accommodate information on different subjects, thus increasing and improving communication between them. Therefore, social media participants can form groups with a common interest and express themselves freely (Kirik and Çetinkayas, 2018).

The importance of social media opinion cannot be over-emphasized as this medium serves as the most accessible way to get large, valuable, and rich details of information (especially on the subject matter) within a short period. The Twitter platform is a social microblog site and has been reported to generate about 330 million tweets every monthly across different countries (Can and Alatas, 2019). Twitter is recently being used to mine opinion and trending topics to understand users' behaviors and attitudes by using predefined information such as user description, location, status and other attributes. Also, this platform allows the exchange of data such as text, images, videos, etc. and the potential to facilitate research over social phenomena based on sentiment analysis, using Natural Language Processing and Machine Learning techniques to interpret sentimental tendencies related to users' opinions and make predictions about real events (Hernandez-Suarez et al., 2018).

Analyzing different trending topics on Twitter may create insight into polarized opinion in varieties of issues such as politics, celebrities, national disasters, corporations, etc., for real-world event prediction. Previous studies by researchers have shown that this practice falls within the socioeconomic cyber-crime (Ibrahim, 2016), and its continued popularity can be attributed to the influence of friends (Tade & Aliyu, 2011; Ojedokun & Eraye, 2012; Arimi, 2011). The relationships between factors influencing these activities and the learning process are depicted in Figure 1. However, this study is motivated by increasing information on social media, majorly Twitter, considering the great benefit to the Government and all related stakeholders. We have

considered the effect on a developing country – Nigeria(as a case study), a fast-growing economy, the largest populated country in Africa.

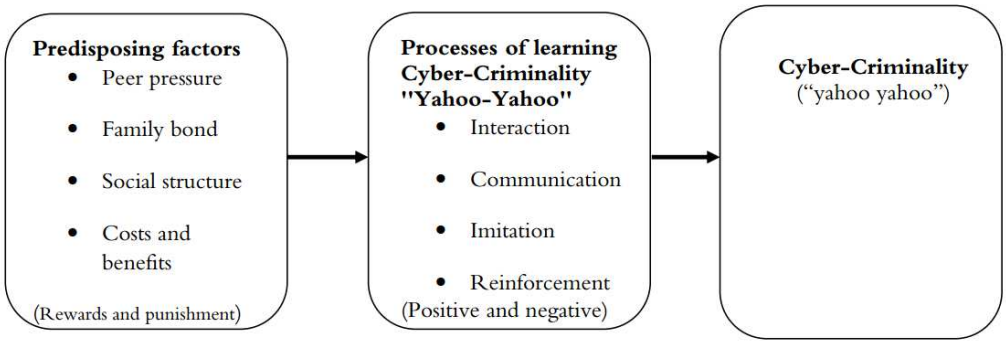


Figure 1: Relationships between predisposing factors and cyber-crime in Nigeria

This paper aims to assess the social media opinion on the trending yahoo-yahoo topic using Twitter data's content analysis. Trending tweets of the yahoo-yahoo topic were collected and analyzed for this study. The rest of the paper is organized as follows: Section 2 discusses the related work, while Section 3 provides a detailed description of the proposed method and materials used. Section 4 discussed in detail the implementation and result, and the paper concludes in Section 5.

2. RELATED WORK

This section discusses in detail the progress of previous research endeavors in identifying and analyzing cyber-crime activities using Twitter data. The specific focus is on cyber-crime and or Twitter data, and state-of-the-art methods proposed in the literature will be carefully studied for contributions and future recommendations.

Somayyeh and Masoud (2018) proposed a temporal topic detection model to infer predictive topics over time. Authors developed a dynamic vocabulary to detect topic trends rather than word dictionaries using Twitter data to predict the Chicago crime trend. The study concluded that the use of content-based features improves overall prediction performance. Hernandez-Suarez *et al.* (2018) presented a statistical analysis based on ℓ_1 regularization regression algorithm for detecting cyber-attacks using social sentiment sensors on Twitter. Kounadi *et al.* (2015) examined twitter messages for detecting homicide crime in London based on the spatial and temporal analysis. The authors adopted two pre-processing methods from link correspondence and the home estimation

model. Hariani and Riadi (2017) analyzed Twitter data for cyberbullying using naïve Bayes classifier and TF-IDF weighting. Authors claimed from their classification results was able to detect cyberbullying on social media, and the effect of this bullying is more psychological, with a prediction of 77.73%. Sharma *et al.* (2019) proposed a sentiment analysis of Twitter data using VADER method for detecting cybersecurity and cyber-crime. The authors concluded that Asian nations are majorly affected by cybersecurity challenges when compared to other EU countries. Al-garadi *et al.* (2016) proposed a supervised machine learning approach using four classifiers, namely the SVM, NB, KNN, and the random forest classifier for detecting cyber-crime on the Twitter network. The methods show that integrating SMOTE with random forest gave the best performance of 94.3% compared with the other machine learning classifier.

The application of deep learning methods has also been proposed in previous studies. Founta et al. (2019), it presented an architecture based on deep learning for detecting online multiple abusive behaviors among Twitter users. The proposed approach gave a significant performance in detection rate and increasing AUC from 92-98%. Like the previous study, Drishya et al. (2019) also applied a deep learning method based on a convolutional neural network to detect cyber-bullying using Instagram images and text data. The detected bullying words are further analyzed using the NB classifier to detect potential cyberbullying threats effectively.

Table 1: Summary of related work on cybercrime analysis using Twitter Data

Authors	Methods	Contributions	Research domain
Kounadi *et al.* (2015)	Machine learning based on logistic regression	Result shows the proposed method could be effective and reliable for investigating the crime.	Homicide detection
Hernandez-Suarez *et al.* (2018)	ℓ1 regularization regression algorithm	Proposed methods were useful to predict possible cyber-attacks.	Cyber-attack detection
Zulfikar and Suharjito (2019)	Support Vector Machine (SVM)	Significant improvement in classification accuracy	Detection Traffic Congestion
Figueira *et al.* (2019)	Ensemble method based on Linear SVM, Radial SVM, Polynomial SVM, Random Forest, and Naïve Bayes	The proposed method gave a reliable capacity to predict relevancy with an improvement in accuracy of more than 6%.	Relevance Detection
Donchenko *et al.* (2017)	Stochastic gradient descent (SGD) approach to training of SVM classifier.	Improved prediction accuracy for the detection of social tension topics in Russia	Social tension detection
Liu *et al.* (2020)	CyberEM model based on pattern clustering and an NMF-based (non-negative matrix factorization) event aggregation algorithm	The proposed model was able to discover cybersecurity events and update event aggregation online	Event detection

Van der Walt *et al.* (2018)	Random Forest algorithm	Developed a low-cost interpretative model	Identity deception
Al-garadi et al. (2016)	Synthetic minority over-sampling technique (SMOTE) approach on supervised ML (naïve Bayes (NB), support vector machine (SVM), random forest, and k-nearest neighbor (KNN))	Develop a cost-sensitive model	Cyberbullying detection
Cheng *et al.* (2019)	K-means clustering algorithm and Random Forest algorithm	The proposed methods were able to show significant prediction power in detecting cyberbullying	Cyberbullying behavior
Burnap and Williams (2015)	Ensemble machine Classification and Statistical Modelling	classification results showed very high levels of performance at reducing false positives and produced promising results with respect to false negatives	Cyber Hate Speech

Based on previous studies' progress and to the best of our knowledge, this study is to first analyze the Twitter dataset for understanding and identifying behavioral trends of the yahoo-yahoo cyber-crime trending topic.

3. RESEARCH METHOD

The research methodology employed in this study is presented in this section. The data analyzed is based on the content of the tweet and other metadata. Duplicated tweets were detected and filtered to analyze unique tweets from the tweet dataset. This study adopted Liu Hu (Hu and Liu, 2004) and Vader (Hutto and Gilbert, 2015) methods for sentiment analysis. The research approach was divided into three modules, as shown in Figure 2, which include Data Collection, Pre-processing, and Data Analysis.

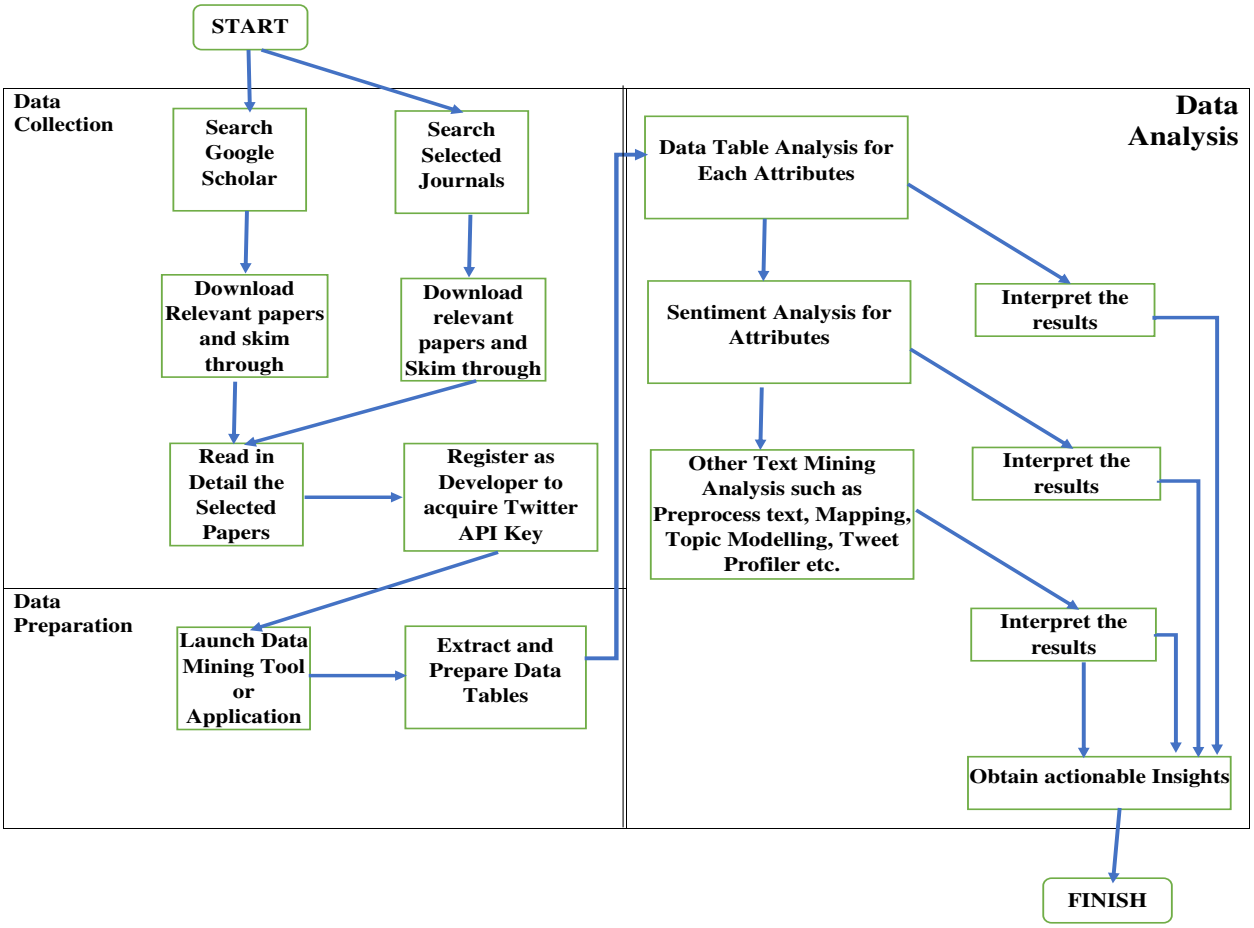
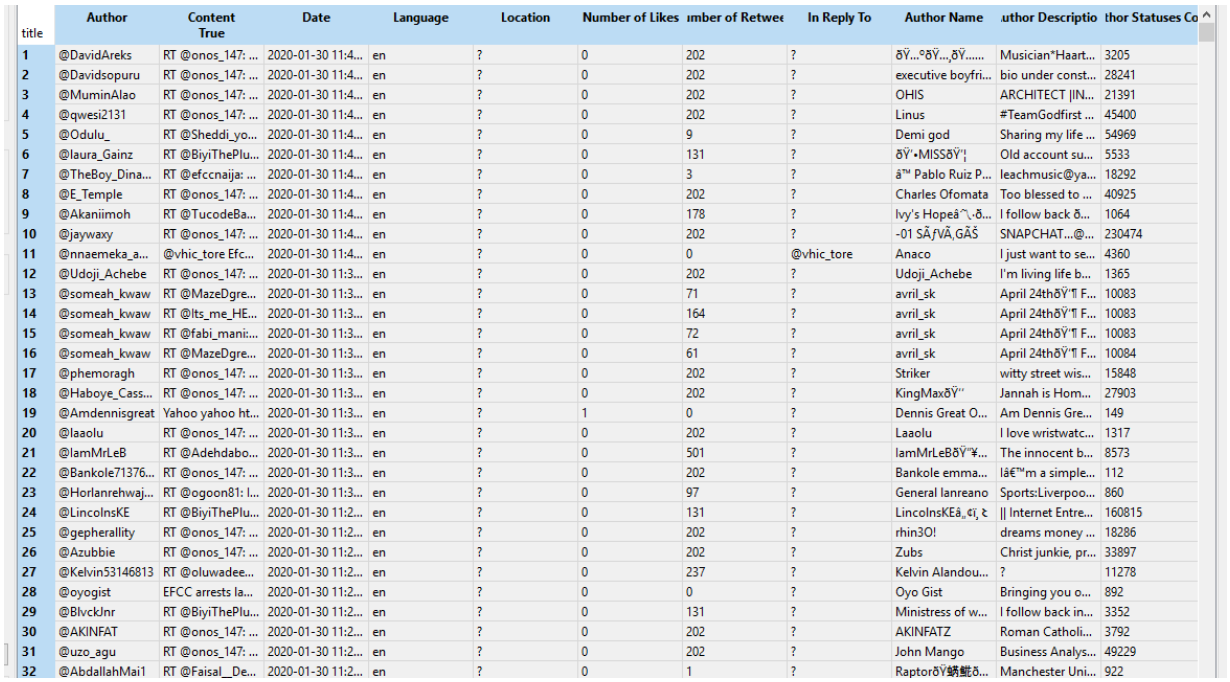


Figure 2: Proposed Content Analysis Framework

Data Collection

In this study, Twitter data was chosen based on its popularity with microblog services for sentiment and opinion analysis in detecting cyber-bullying, cyber-terrorism, etc. (Gupta *et al.*, 2016). Twitter API was employed in streaming live tweets for the past 14 days on the Orange Data mining toolbox (Demšar *et al.*, 2013). To use the Twitter API, it is required to obtain the Twitter API credentials, which contain the key and secret passwords. With the API, query parameters relating to specific keywords such as wordlist query, search by, language, allow retweets, etc., can be set, and the data obtained can be saved as Comma-Separated Value (CSV) format. For this study, our search query's keyword was "yahoo-yahoo", and a maximum tweet of 5500 tweets was returned. Figure 3 shows the tweet dataset obtained from the "yahoo yahoo" keyword on a data table containing the tweets content as well as 17 other metadata which includes Author ID, Date, Language, Location, Number

of Likes, Number of Retweets, In Reply To, Author Name, Author Description, Longitude and Latitude, etc.



title	Author	Content True	Date	Language	Location	Number of Likes	ımber of Retwee	In Reply To	Author Name	ıthor Descriptio	thor Statuses Co
1	@DavidAreks	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	ðŸ...°ðŸ...,ðŸ.....	Musician°Haart...	3205
2	@Davidsopuru	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	executive boyfri...	bio under const...	28241
3	@MuminAlao	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	OHIS	ARCHITECT |IN...	21391
4	@qwesi2131	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	Linus	#TeamGodfirst ...	45400
5	@Odulu_	RT @Sheddi_yo...	2020-01-30 11:4...	en	?	0	9	?	Demi god	Sharing my life ...	54969
6	@laura_Gainz	RT @BiyiThePlu...	2020-01-30 11:4...	en	?	0	131	?	ðŸ•MISSðŸ¦	Old account su...	5533
7	@TheBoy_Dina...	RT @efccnaija: ...	2020-01-30 11:4...	en	?	0	3	?	â™ Pablo Ruiz P...	leachmusic@ya...	18292
8	@E_Temple	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	Charles Ofomata	Too blessed to ...	40925
9	@Akaniimoh	RT @TucodeBa...	2020-01-30 11:4...	en	?	0	178	?	Ivy's Hopeâ˜·ð...	I follow back ð...	1064
10	@jaywaxy	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	-01 SÃƒVÃ‚GÃŠ	SNAPCHAT...@...	230474
11	@nnaemeka_a...	@vhic_tore Efc...	2020-01-30 11:4...	en	?	0	0	@vhic_tore	Anaco	I just want to se...	4360
12	@Udoji_Achebe	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Udoji_Achebe	I'm living life b...	1365
13	@someah_kwaw	RT @MazeDgre...	2020-01-30 11:3...	en	?	0	71	?	avril_sk	April 24thðŸ¶ F...	10083
14	@someah_kwaw	RT @Its_me_HE...	2020-01-30 11:3...	en	?	0	164	?	avril_sk	April 24thðŸ¶ F...	10083
15	@someah_kwaw	RT @fabi_mani:...	2020-01-30 11:3...	en	?	0	72	?	avril_sk	April 24thðŸ¶ F...	10083
16	@someah_kwaw	RT @MazeDgre...	2020-01-30 11:3...	en	?	0	61	?	avril_sk	April 24thðŸ¶ F...	10084
17	@phemoragh	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Striker	witty street wis...	15848
18	@Haboye_Cass...	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	KingMaxðŸ''	Jannah is Hom...	27903
19	@Amdennisgreat	Yahoo yahoo ht...	2020-01-30 11:3...	en	?	1	0	?	Dennis Great O...	Am Dennis Gre...	149
20	@laaolu	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Laaolu	I love wristwatc...	1317
21	@IamMrLeB	RT @Adehdabo...	2020-01-30 11:3...	en	?	0	501	?	IamMrLeBðŸ"¥...	The innocent b...	8573
22	@Bankole71376...	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Bankole emma...	Iâ€™m a simple...	112
23	@Horlanrehwaj...	RT @ogoon81: I...	2020-01-30 11:3...	en	?	0	97	?	General lanreano	Sports:Liverpoo...	860
24	@LincolnsKE	RT @BiyiThePlu...	2020-01-30 11:2...	en	?	0	131	?	LincolnsKEâ„¢ï¸	|| Internet Entre...	160815
25	@gepherallity	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	rhin3O!	dreams money ...	18286
26	@Azubbie	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	Zubs	Christ junkie, pr...	33897
27	@Kelvin53146813	RT @oluwadee...	2020-01-30 11:2...	en	?	0	237	?	Kelvin Alandou...	?	11278
28	@oyogist	EFCC arrests la...	2020-01-30 11:2...	en	?	0	0	?	Oyo Gist	Bringing you o...	892
29	@BIvckJnr	RT @BiyiThePlu...	2020-01-30 11:2...	en	?	0	131	?	Ministress of w...	I follow back in...	3352
30	@AKINFAT	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	AKINFATZ	Roman Catholi...	3792
31	@uzo_agu	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	John Mango	Business Analys...	49229
32	@AbdallahMai1	RT @Faisal__De...	2020-01-30 11:2...	en	?	0	1	?	RaptorðŸ¦ðŸ¦ð...	Manchester Uni...	922

Figure 3: Screenshot showing the Data Table with the Tweet contents and other metadata

Data Pre-processing

The tweets dataset was pre-processed by breaking the tweet content into smaller pieces like words, phrases, or bi-grams called tokens. Normalization was done on the tweets to generate n-grams and tags with spoken tags and partial language markings. Other pre-processing tasks carried out on the tweets include:

1. Converting all characters in the corpus to lowercase;
2. Remove all HTML tags from a string;
3. Removing all text-based diacritics and accents;
4. Removing URLs, articles, and punctuations;
5. Filtering stop words, lexicon, Regular expressions.

Sentiment Analysis

Sentiment analysis aims to extract users' emotions from texts at sentence, document or aspect/feature level. It determines the feeling of being projected from each tweet as either positive, negative, or neutral. The NLTK emotion modules in Orange contain both Liu Hu (Hu and Liu,

2004) and Vader (Hutto and Gilbert, 2015) techniques based on sentiment lexicons. The lexicon-based approach is an unsupervised machine learning method that employs a dictionary or lexicon list. Each lexicon is associated with a sentiment strength which represents a positive or negative orientation (Labille *et al.*, 2017).

The Liu Hu method (Hu and Liu, 2004) involves an examination of the lexicon. It classifies the tweets into negative, positive, and neutral sentiment while the Vader examines the lexicon and uses the thumb rule. It simply sums up the sentiment scores of all sentiment words in a tweet or sentence segment.

The Valence Aware Dictionary for Sentiment Reasoning (VADER) was proposed in Hutto and Gilbert (2015). Unlike Liu Hu, Vader has its sentiment orientation divided into four categories which are: positive, negative, neutral and final compound scores for analyzing sentiment. The compound score is calculated in Equation 1 by finding the sum of each word's valence scores in the lexicon, which are adjusted according to the rules.

$$x = \frac{x}{\sqrt{x^2 + \propto}} \qquad (1)$$

where x = sum of valence scores of constituent words, and α = Normalization constant (default value is 15)

The costs are then normalized between -1 and +1, representing the most extreme negative and most extreme positive sentiments, respectively. The Vader compound score is a single unidimensional measure of a tweet's sentiment. The VADER method was adopted for this study. Figure 4 shows the sentiment analysis and duplicate detection model.

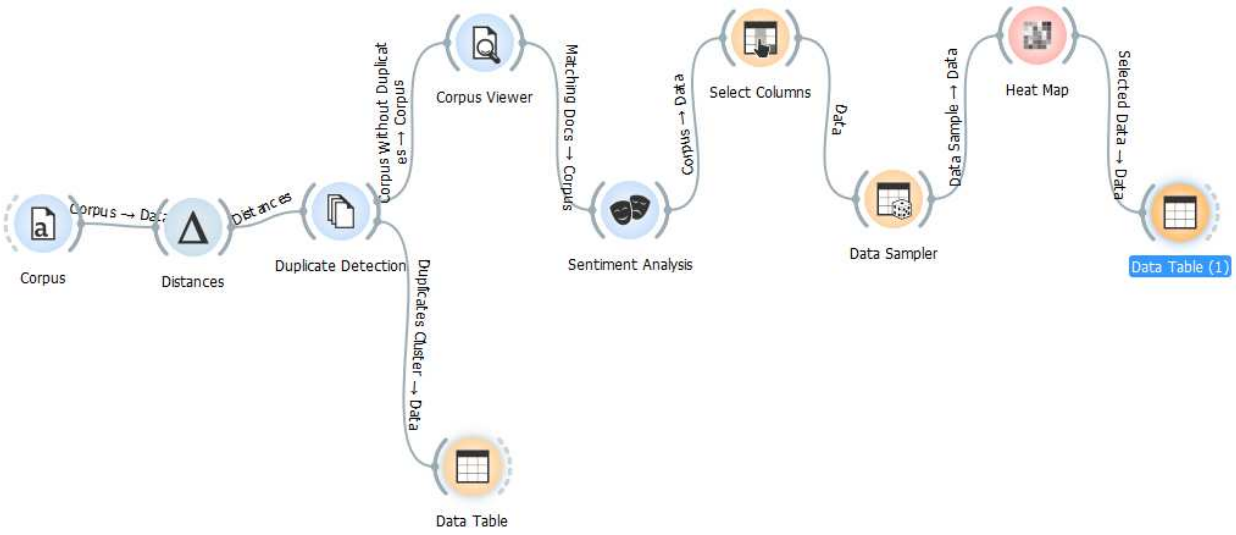


Figure 4: Sentiment Analysis and Duplicate Detection Model

Topic Modelling

Topic modeling is used to detect abstract topics in the corpus or data table based on word clusters and their respective frequency in each document or tweet as in this case study. It has been applied in natural language processing (NLP) to discover topics and extract semantic meaning from unordered documents, especially in applications such as social media, text mining, and information retrieval. In this study, we aim to use the topic to facilitate understanding the emotion and conversations between the respondent in the corpus under study. The orange topic modeling widget wraps the Gensim's topic models (Rehurek and Sojka, 2010) that contain Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Processing (HDP) algorithms, respectively. LDA is a three-level hierarchical Bayesian model in which each item of a collection is modeled as a finite mixture over an underlying set of topics. It is interpreted easily but slower than LSI. LSI model returns topics with negative and positive keywords that have negative and positive weights on the topic. The positive weights are words that are highly representative of the topic and contribute to its occurrence. For negative weights, the topic is more likely to occur if they appear less in it. The modeled topics were visualized using Multidimensional scaling (MDS), which is a low-dimensional projection of the topics as points. MDS attempt to fit distances between the points as much as possible. Figure 5 shows the workflow for topic modelling.

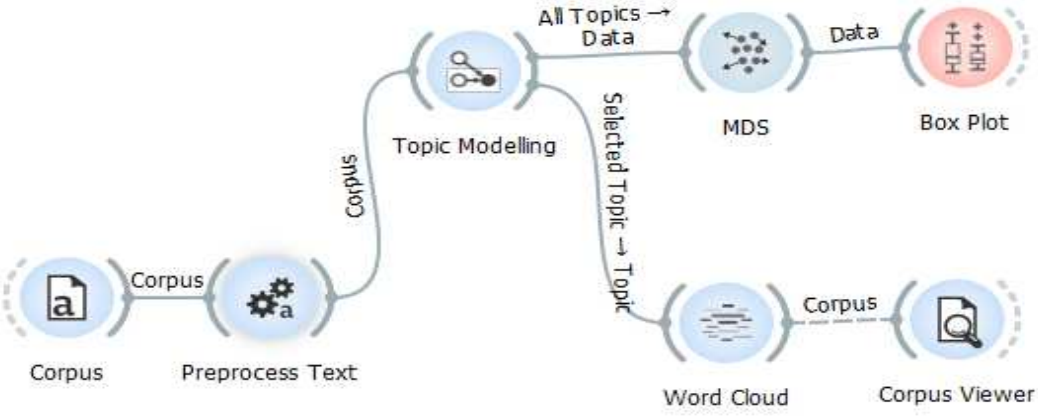


Figure 5: Workflow for the topic modeling

4. RESULTS AND DISCUSSION

As outlined in section three, the result obtained from implementing the research methodology is presented and discussed in this session. The Twitter data mining API, widgets, and workflow engine for text mining in the Orange Data mining toolbox (Demšar *et al.*, 2013), developed at the University of Ljubljana (SLO), was used primarily for the data collection and implementation of this study.

Pre-processing and Tokenization

A pre-trained tweet tokenizer was used for pre-processing of the corpus texts. By setting the document frequency range, tokens outside the range will be removed. 75,280 tokens of 3,968 types were generated using a document frequency of 0.00–1.00, while for 0.10 and 0.90, 16,620 tokens of 5 types were returned. Figure 6 shows the tokens' visualization and their frequency in the tweet dataset through a word cloud. The larger the word in the cloud, the higher its frequency. The only record of tokens with a frequency higher than 100 was stored, and 3,960 tokens appeared more than a hundred times. Table 2 shows the 12 most frequent tokens with "yahoo" on top of the chart with 9,555 frequencies.

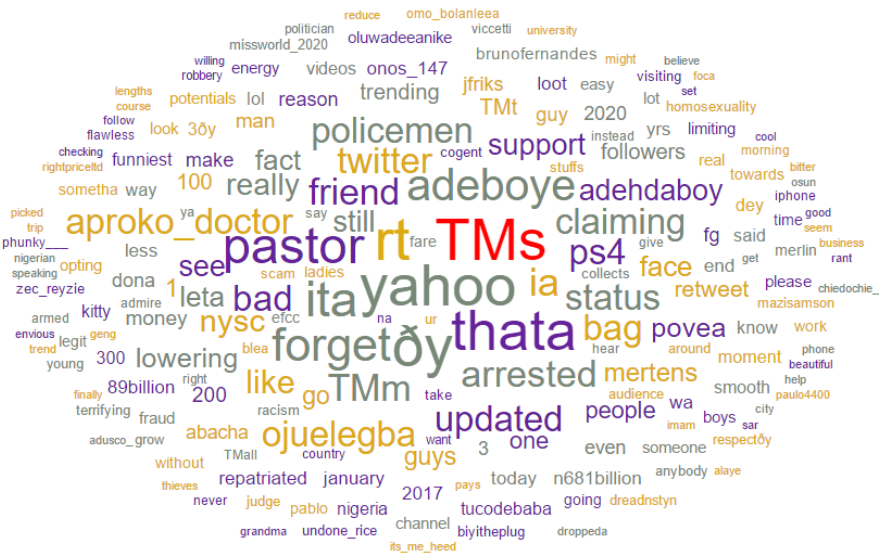


Figure 1: Word Cloud showing the tokens and their frequency/weight from the pre-processed tweet dataset

Table 2: The topmost frequent tokens with their frequency/weight

S/N	Words	Frequency/Weight
1	yahoo	9,555
2	pastor	745
3	forget	668
4	adeboye	628
5	arrested	511
6	friend	499
7	bad	498
8	status	488
9	ps4	488
10	bag	488
11	Twitter	487
12	Updated	486

One would be wondering what a token or word like a *pastor* has to do with cyber-crime. This does not mean that they are involved, but, in this context, based on the tweets mined, some authors classified the way of life of some pastors as a form of "yahoo," which means cyber-crime in Nigeria. The arrest has to do with the Landlord who harbored fraudsters in Ibadan. A man found with ps4 arrested at Ojuelegba Lagos, whom the policemen assumed to be a "yahoo-boy." Pastor Adeboye appeared among others due to a statement he made which was, *"one of my sons once told me that he was always excited to resume in the office every Monday because he would get to see his secretary again. I told him to fire (sack) her immediately. Nothing and no one is worth your*

marriage". With this, pastor Adeboye was hashtagged with "yahoo yahoo." A Twitter user, @aproko_doctor, tweeted that, *"My friend just updated on his status that policemen arrested him at Ojuelegba for having a ps4 in his bag, claiming that he was a yahoo boy*". This tweet generated many retweets that were responsible for tokens such bag, ps4, arrested, and rt. Another user, @Adehdaboy, was quite sentimental with his opinion as *"I'm not in support of Yahoo yahoo, it's really bad but let's face the fact that it's yahoo yahoo that's still lowering poverty"* while @mighty_tolu supported that *"it has saved people and promoted more business"*.

Geolocation

From Figure 7, the color bar differentiates the number of tweets originating from each country in the range 0 to 20 on a scale of 0-4, 5-9,10-14, and 15-20, respectively. The white locations had no tweets from the "yahoo yahoo" hashtag, while the colored ones had tweets of varying amounts.

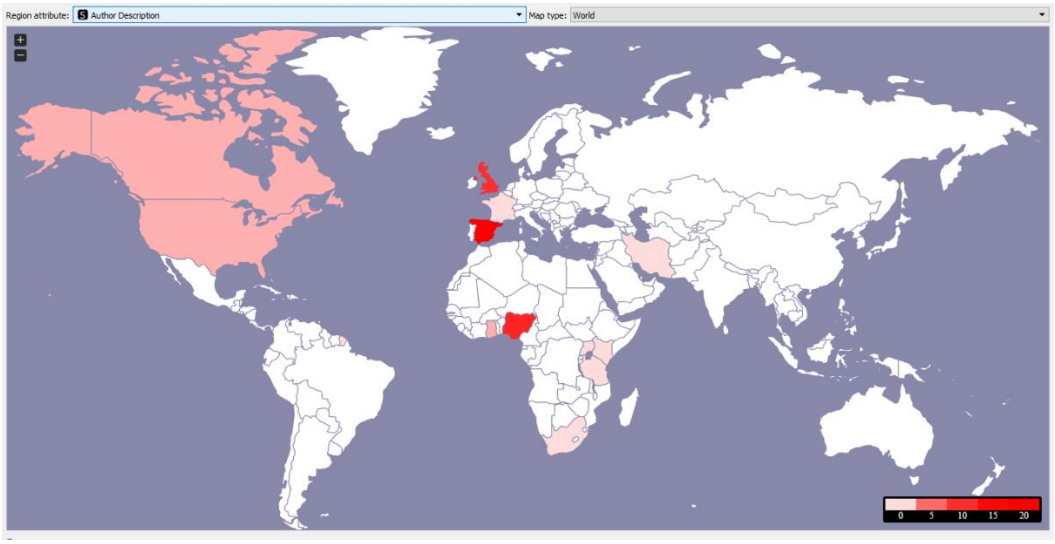


Figure 7: Map Showing Location of Author representing the frequency with colour weight

Some of the countries in the range 0-4 (color code 1) on the world map are Ghana, French Guyana, South Africa, Tanzania, Uganda, Kenya, France and Iran; Canada and the United States of America are in the range 5-9 (colour code 2) while Northern Ireland, United Kingdom, and Nigeria are in the range 10-14 (colour code 3) with 9, 10 and 11, respectively. Finally, Spain fell into the last category with colour 16. From the dataset and contrary to expectation, Spain has the highest number of tweets on the trending yahoo yahoo hashtag on Twitter, followed by Nigeria. However,

our tweet dataset also confirmed that most Twitter users prefer to have their locations as anonymous.

Results of Duplicate Detection

The dataset was filtered for unique tweets using duplicate detection to remove duplicate tweets from the 5,500 tweets.

With the linkage set to single and distance threshold =0.5, the duplicate detection workflow returned 175 unique clusters and their sizes. Where a cluster represents a unique tweet, and the size is the number of times it is retweeted or duplicated in the dataset. Thus, 175 unique tweets and 5325 duplicates were returned. Figure 8 shows the duplicate detection widget's output, while Table 3 shows the top twenty tweets, cluster, number of retweets, and content. Tweet C91 had the largest size with 484 retweets. The 175 unique tweet clusters were adopted for further analysis in the study.

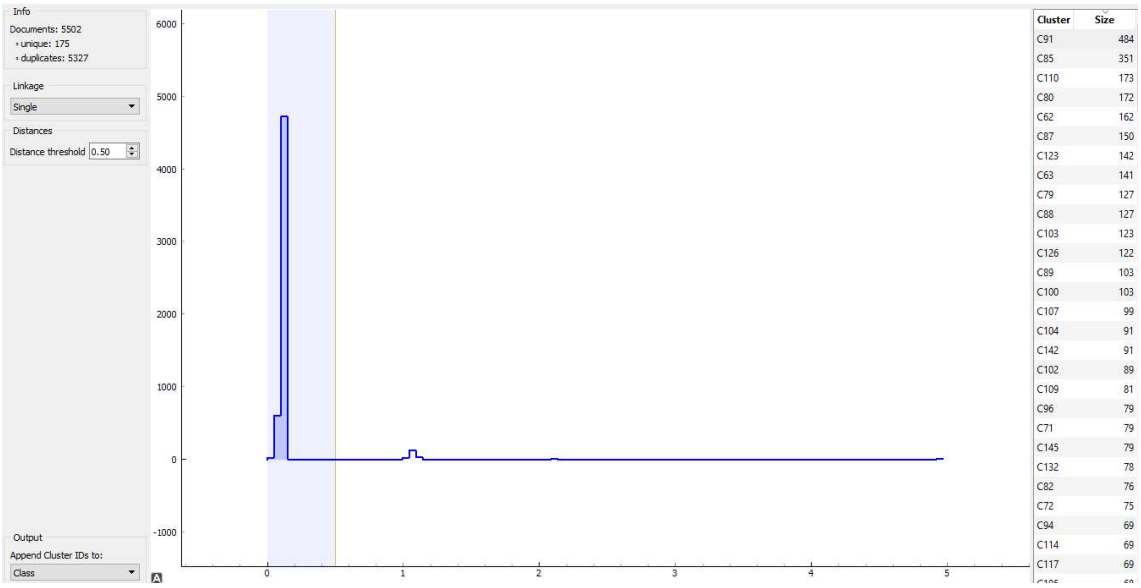


Figure 8: Showing the output of the duplicate detection widget

Table 3: Top twenty tweets showing the cluster, number of retweets, and content

S/N	Cluster	No. of retweet	Content
1	C91	484	My friend just updated on his status that policemen arrested him at Ojuelegba for having a Ps4 in his bag, claiming that he was a yahoo boy.
2	C85	351	I'm not in support of Yahoo yahoo; it's really bad but let's face the fact that it's yahoo yahoo that's still lowering poverty
3	C110	173	Forget yahoo yahoo for a moment and be as smooth as this kitty.

4	C80	172	This is one of the funniest video you will see on Twitter today.
5	C62	162	EFCC Arrests Landlord for housing Yahoo boys. This comprises of more than one form of a tweet (e.g EFCC, Bad Governance, Landlord, yahoo yahoo government etc.)
6	C87	150	Yahoo yahoo is like opting for the easy way out, limiting your potentials, why not channel that same energy towards something worthwhile and good.
7	C123	142	Grow your Twitter audience now. As we can't do fraud, we can't do Yahoo yahoo, we can't steal, and we can't be lazy
8	C63	141	This Administration is a scam. EFCC is yahoo yahoo. Every sector of this nation is in Coma. (This talks about the resignation of President Buhari, Fulani Herdsmen, Budget of $12m, EFCC and Yahoo boys)
9	C79	127	Ladies who collect T-Fare from a man and end up not visiting him without a cogent reason are the real Yahoo Yahoo.
10	C88	127	Yahoo yahoo – they will brainwash you and make you give them your money. Fraud – you will give them your money on your own free (This emphasis on difference between yahoo yahoo and fraud. Also, it contains tweets on Rochas, linkage with Government and that they are better than politicians)
11	C103	123	I'm not even going to judge anybody doing yahoo yahoo.
12	C126	122	Forget, NYSC, Yahoo yahoo, Mertens, Pablo and pastor Adeboye, Twitter people don't have respect.
13	C89	103	I don't know why Yahoo Yahoo is trending, but you all should take your time and admire this flawless make up
14	C100	103	The greatest, easiest and most legitimate form of yahoo yahoo in Nigeria is politics
15	C107	99	DO girls also do yahoo yahoo? Or is it only the boys?
16	C104	91	Problems caused by yahoo yahoo scammers government (This is on corruption, bribery, fraud, yahoo-yahoo and scammers)
17	C142	91	Legit work that pays. Say No to Yahoo Yahoo.
18	C102	89	Yahoo yahoo is bad, instead just be a pastor, imam or a politician.
19	C109	81	Between 2017 and January 2020, FG has repatriated $1.89Billion of Abacha Loot.
20	C96	79	To SARS you are doing yahoo yahoo o. the should just arrest themselves.
21	C71	79	Someone said Yahoo Yahoo is now a course in his University.

Result of Sentiment Analysis

The Vader method of sentiment analysis uses continuous polarity annotation; hence the compound score is ordered by ranking between the range of +1.0 to -1.0. Where +1 represent positive or very optimistic tweet, and -1 represent Negative tweets or very pessimistic. Figure 9 presents the graphical visualization of sentiment classification using the heat map. The result showed that 53 instances were analyzed with the Vader method; the upper bound consist of 40 cases which are classified as 19, 13, and 8 for negative, neutral, and No-Zone sentiment tweets.

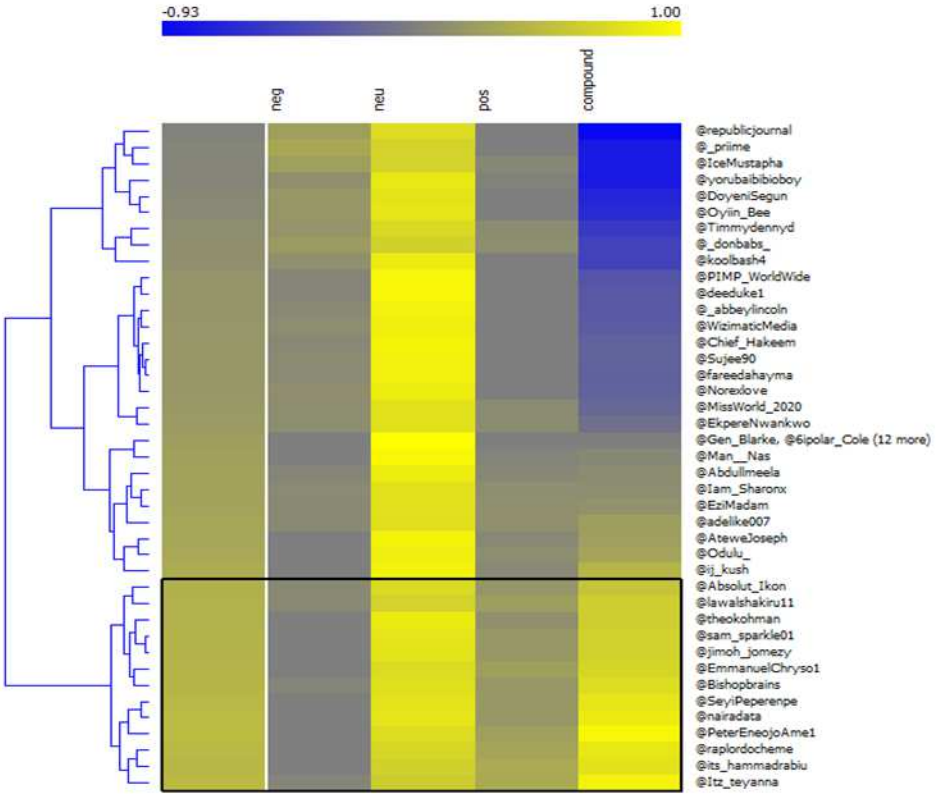


Figure 9: Heat map showing positive and negative sentiments classification by Vader method

The 19 negative sentiment tweets were negative as depicted by the blue colour on the heat map in Figure 9; their compound sentiment scores ranged from -0.093 and -0.1027, their positive sentiment scores are very low and close to zero (0), as seen in Table 4. From Table 5, the positive, negative, and compound sentiment orientation scores went down to zero (0) while the neutral scores were all one (1). This returned 13 neutral tweets, as seen in Table 5, starting with @Gen_Blarke and 12 others. The third class in the heat map's upper bound is tweeted whose compound sentiment scores are above zero (0) and below 0.5. They do not represent negative or neutral sentiments and are also below the 0.5 thresholds to be considered as positive sentiment tweets. We refer to the tweets in this class as no-zone sentiment tweets. Their compound sentiment scores ranged from 0.0772 to 0.4404, as shown in Table 6, and their color on the heat map is not distinct. The lower bound of the heat map in Figure 9 has 13 instances with yellow color. They are classified as positive tweets having compound scores in the range of 0.5526 to 0.9136, as shown in Table 7.

Subjective analysis of the tweet's sentiment classification was conducted to verify the accuracy of the Vader method. The result is shown in the actual column in Table 4–7. The negative sentiment

classification has 94.74% accuracy with only one (1) misclassification; neutral sentiment 84.62% accuracy with two (2) misclassifications; positive sentiment 100% accuracy with no misclassification. The no-zone sentiment classification had 62.5% accuracy with three (3) misclassifications.

Table 4: List of tweets classified by Vader as Negative sentiment and the actual classification

S/N	Author	Content	Neg	Neu	Pos	Compound	Actual
1	@republicjournal	The origin of online scams in Nigeria can be traced to the 1980s: beset by flailing oil prices, …	0.266	0.734	0	-0.93	Neg
2	@_priime	Dont do yahoo yahoo just do fraudulent acts, with your shirt and tie Cooperate fraud …	0.333	0.667	0	-0.7906	Neg
3	@IceMustapha	RT @Blackculture_X: Yahoo yahoo is not just a Black problem, even other race perpetrate …	0.259	0.669	0.072	-0.783	Neg
4	@yorubaibibioboy	Ed Sheeran dropped out of school and slept in the Subways. …	0.137	0.839	0.024	-0.7939	Neg
5	@DoyeniSegun	Wait wait wait, forget about Pastor Adeboye NYSC Yahoo Yahoo …	0.202	0.798	0	-0.7059	Neg
6	@Oyiin_Bee	RT @UNCLE_AJALA: What are you into, what legit business are you doing, that's stopping …	0.184	0.816	0	-0.6597	Neg
7	@Timmydennyd	Please i want to know .. Does it mean that you're into Cyber fraud (Yahoo-Yahoo) If you	0.176	0.72	0.104	-0.5423	Neg
8	@_donbabs_	Honestly Yahoo Yahoo is bad, I don't support it one bit.Instead of you to aim high,why not just …	0.232	0.653	0.115	-0.4678	Pos
9	@koolbash4	RT @laurelchinedu: Watch video of the drama that happened when an officer of @PoliceNG …	0.134	0.866	0	-0.4767	Neg
10	@PIMP_World Wide	So a lot of real hip hop fans be waiting on this to drop when you people be talking about …	0.052	0.948	0	-0.3382	Neg
11	@deeduke1	EFCC has come to stay in Ibadan, they've settled and marked Ibadan as a hub for yahoo …	0.052	0.948	0	-0.296	Neg
12	@_abbeylincoln	Iâ€™m not in support of yahoo yahoo , but u guys need to listen to NOWO by …	0.088	0.912	0	-0.2989	Neg
13	@WizimaticMedia	RT @Ayooebun: Moti drop out ðŸ¥º. Yahoo yahoo plus plus is now the way ðŸ¤□ðŸ¥º …	0.116	0.884	0	-0.2732	Neg
14	@Chief_Hakeem	Forget NYSC, Yahoo Yahoo, Pastor Adeboye. I told my mom she's invading my privacy, she …	0.076	0.924	0	-0.2263	Neg
15	@Sujee90	RT @Undone_rice: Forget #BrunoFernandes, NYSC, Yahoo yahoo, mertens, pablo and pastor …	0.095	0.905	0	-0.2263	Neg
16	@fareedahayma	RT @testyflowz: Youâ€™re asking me not to do yahoo yahoo, shey na betway go pay my …	0.102	0.898	0	-0.2359	Neg
17	@Norexlove	RT @ogoon81: Guys forget about yahoo yahoo , have you heard Emotions by #RicoSwavey …	0.128	0.872	0	-0.2263	Neg
18	@MissWorld_2020	Grow your Twitter audience no @MissWorld_2020 …	0.114	0.786	0.1	-0.1779	Neg
19	@EkpereNwankwo	RT @NwaExit: @ElvisChinedu12 @EmekaIhedioha #Biafra Brexit 2020 only …	0.123	0.773	0.105	-0.1027	Neg

Table 5: List of tweets classified by Vader as Neutral sentiment and the actual classification

S/N	Author	Content	Neg	Neu	Pos	Compound	Actual
1	@Gen_Blarke	RT @paulo4400: Someone said Yahoo Yahoo is now a course in his University! …	0	1	0	0	Neu
2	@6ipolar_Cole	RT @JFriks: Guys, guys the eagle has landed Yahoo Yahoo guys: https://t.co/AvUpWoN09U…	0	1	0	0	Neu
3	@sodiq_ololade	General sani Abacha is the biggest yahoo yahoo guy in the history of Nigeria. …	0	1	0	0	Neu
4	@REALKENI	RT @volqx__: This is the only man in Nigeria that obeys Nigerian rules and regulations …	0	1	0	0	Pos
5	@JoTechTracker	If U see this, EPP me RT my hustle plsðŸ™□ðŸ□»…	0	1	0	0	Neu
6	@TheFoodPlaceABJ	RT @obiage_li: I bake cakes, they are affordable Just A DM away #gwagwalada …	0	1	0	0	Neu
7	@v_ibiok	RT @flamagraaaa: You said yahoo yahoo twice …	0	1	0	0	Neu
8	@_kayspice	RT @onos_147: Between 2017 & January 2020, FG has repatriated $1.89Billion of Abacha…	0	1	0	0	Neg
9	@AspiringCeleb_S	Whyâ€™s Merlin and Yahoo Yahoo trending? Or is he now officially a marlian? ðŸ˜,…	0	1	0	0	Neu
10	@koolbash4	RT @KneWKeeD: Nobody is going to defraudyou in your house minding your business, they can …	0	1	0	0	Neu
11	@GistGal	Man returns wife to her parents in Edo state, says her private part is too wide.ðŸ˜,yahoo yahoo …	0	1	0	0	Neu
12	@Cheyih_Viktah	Even a nigga I know to be a yahoo boy even up till now was speaking heavily against yahoo yahoo …	0	1	0	0	Neu
13	@EziMadam	When are we gonna talk about girls Yahoo Yahoo. …	0	1	0	0	Neu

Table 6: List of tweets with no specific sentiment classification (no-zone) by Vader and the actual classification

S/N	Author	Content	Neg	Neu	Pos	Compound	Actual
1	@Man__Nas	RT @oil_shaeikh: If youâ€™re not a yahoo yahoo boy, you know them, or you have them in your …	0	0.956	0.044	0.0772	Neu
2	@Abdullmeela	RT @Al_ameen_Yabo: When online stores turn into yahoo yahoo. …	0.056	0.873	0.071	0.1027	Neu
3	@Iam_Sharonx	Your time is limited, don't waste it living someone else life…	0.099	0.78	0.121	0.1109	Neu
4	@EziMadam	If you've never Yahoo Yahoo'd your parents, uncles, aunts, siblings, spouse... Raise your hands, …	0.079	0.786	0.135	0.1511	Pos
5	@adelike007	RT @Millishield: To all Nigerian youths...there no way an enduring wealth can be achieved overni …	0.094	0.769	0.137	0.25	Pos

6	@AteweJoseph	RT @jidesanwooIu: Yahoo Yahoo might seem cool, but does this look right? Are you willing …	0	0.923	0.077	0.2523	Pos
7	@Odulu_	RT @Sheddi_younG: EFCC when they hear Yahoo yahoo boys are about to share their small …	0	0.885	0.115	0.296	Neu
8	@ij_kush	It's probably not his Pastor Adeboye's biological son. These guys call their mentees "sons" …	0	0.919	0.081	0.4404	Neu

Table 7: List of tweets classified by Vader as positive sentiment and the actual classification

S/N	Author	Content	Neg	Neu	Pos	Compound	Actual
1	@Absolut_ikon	RT @OnibonMedia: Am not saying Yahoo Yahoo is good tho but it has helped this country …	0.087	0.727	0.186	0.5526	Pos
2	@lawalshakiru11	RT @milez_wadup: Forget what Adeboye said, all man na Yahoo Yahoo, enjoy this 2 …	0.079	0.669	0.251	0.6249	Pos
3	@theokohman	If we're to be sincere with ourselves enh, you see those ladies that collect money for T-fare …	0	0.859	0.141	0.6249	Pos
4	@sam_sparkle01	RT @nairadata: While Pastor Adeboye is Trending, hope Yahoo Yahoo boys will not …	0	0.81	0.19	0.6597	Pos
5	@Jimoh_jomezy	"RT @akeula_trendy: I wish to release a thread on ""Internet scam/yahoo yahoo"" by …	0	0.797	0.203	0.6553	Pos
6	@EmmanuelChryso1	"RT @E_hmeka: Good afternoon guys …	0	0.739	0.261	0.6908	Pos
7	@Bishopbrains	"Yahoo yahoo no be a hustle. Making shoes is …	0.049	0.761	0.19	0.7425	Pos
8	@SeyiPeperenpe	I don't do Yahoo yahoo or play #BabaIjebu, I'm not Daddy G.O 's wife either I only have …	0	0.8	0.2	0.8176	Pos
9	@nairadata	"RT @nairadata: Hello, Predict the correct score of this match before 8:30pm and and …	0	0.805	0.195	0.8658	Pos
10	@PeterEneojoA	"I donâ€™t know why Yahoo Yahoo is trending, but yâ€™all should just take ur time …	0	0.747	0.253	0.9508	Pos
11	@raplordocheme	"Please come and buy what my Friend is selli …	0	0.694	0.306	0.8316	Pos
12	@its_hammadrabiu	The greatest, easiest and most legitimate form of yahoo yahoo in Nigeria is politics …	0	0.65	0.35	0.7906	Pos
13	@itz_teyanna	As you scroll through the trends yahoo yahoo and pastor Adeboye. Donâ€™t forget that we …	0.048	0.608	0.344	0.9136	Pos

Results of Topic Modelling

The LDA and LSI models were applied for topic modeling. Using a document frequency of range 0.10 and 0.90, only 16,617 tokens were returned with five (5) types. In this case, the LDA and LSI models returned only one topic with the same keywords, namely: yahoo, rt, pastor, forget, and adeboye. However, using the document frequency range of 0.00-1.00, 75,280 tokens of 3,968 types were returned. We set out for 6 topics using LDA and LSI; the topics and their keywords are shown in Figure 10.

Topic	LDA Topic Keywords
1	yahoo, rt, go, like, said, fraud, end, reason, man, real
2	yahoo, rt, bad, arrested, updated, status, ps4, bag, friend, ojuelegba
3	trending, 😂 make, merlin, nadal, know, time, someone, take, ur
4	rt, geng, 😂 set, 😂 order, @mazedgreat, everyone, sars, 10
5	money, 😭 thiem, since, #whatwentwrong, give, get, daddy, need, saying
6	yahoo, rt, pastor, retweet, forget, 😂 adeboye, 🐟 @jfriks, =

Topic	LSI Topic Keywords
1	yahoo, rt, 😂 bad, pastor, really, let, support, still, @adehdaboy
2	rt, yahoo, =, arrested, bag, friend, ps4, policemen, status, claiming
3	😂 twitter, pastor, adeboye, arrested, friend, bag, ps4, policemen, updated
4	😂 =, pastor, nysc, adeboye, rt, followers, forget, 200, 100
5	fg, 2020, january, abacha, 3, loot, 2017, repatriated, 1.89, n681billion
6	like, even, lot, going, terrifying, judge, anybody, racism, homosexuality, stuffs

Figure 10: showing LDA and LSI generated topics with keywords

The topics obtained from LDA and LSI contain words that are consistent with those that also appear more than 100 times. There are strong similarities between the keywords in the topics obtained from the LSI and LDA models. The word yahoo is highly positively representative of topics 1, 3, 4, 6 in the LDA with weights 0.170, 0.101, 0.092, 0.234, respectively, and topic 1 in LSI with a weight of 0.087. "Yahoo" is also representative of LSI topic 2 with a negative weight of -0.232 in Table 9. It can be observed from Figure 11 (subplot 5) of LDA generated topics and Table 8 that LDA topic five is made up of words with very low weights when compared to other LDA topics. Also, from Table 8, LDA topics 1, 4, and 6 have top keywords that formed the following tweets in the corpus when pulled together. The tweets created by the top words of topic 1, 4, and 6 were retweeted 72, 1418, and 501 times, respectively, in the corpus.

Topic 1: @fabi_mani, ***Someone said yahoo yahoo*** is youth empowerment, i ***weep*** for my ***country***

Topic 4: @aproko_doctor, My ***friend*** just ***updated*** on his ***status*** that policemen ***arrested*** him at ***Ojuelegba*** for having a ***ps4*** in his ***bag***, ***claiming*** that

Topic 6: @Adehdaboy, Iâm not in support of ***yahoo yahoo***, it is ***really bad*** but ***let u***s ***face*** the ***fact*** that itâ€™s ***yahoo yahoo*** thatâ€™s still lowering poverty¦

In Figure 11, the cloud of words that constitutes LDA and LSI-generated topics is presented. It was observed that words in red from the LSI topics such as racism, judge, homosexuality, terrifying, etc., have very strong negative weights towards LSI topic 6. For the LSI, only topic 1 contain all positive words, topic 2 to 5 has a mix of negative and positive words, while topic 6 has only negatively representative words as its constituents.

Table 8 and Table 9 show that the Emojis are seen as topic keywords because the corpus was pre-processed specifically with a pre-trained tweet tokenizer. The Emojis have strong weights in both LDA and LSI-generated topics with positive contributions towards their respective topics, as shown in Figure 11, Table 8, and Table 9. The Emojis express emotions and sentiments such as Laughing and rolling on the floor, Face with tears of joy, Sweat droplets, Loudly crying face, Grinning face with sweat, etc.

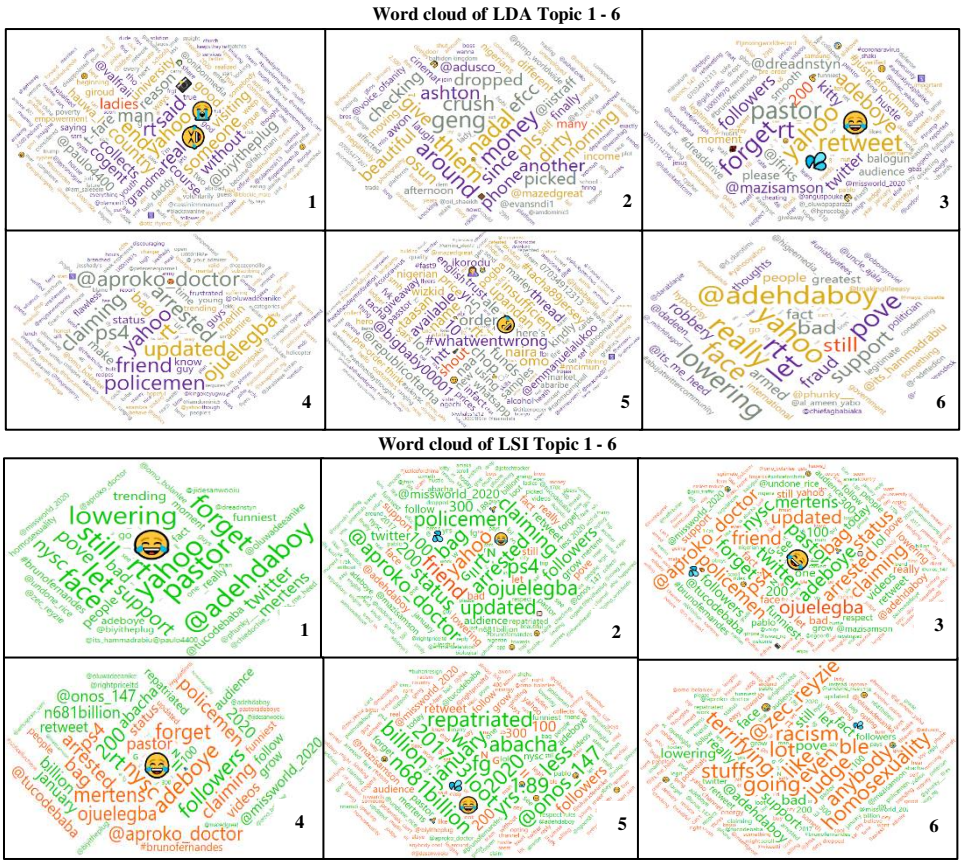


Figure 11: The cloud of words that constitutes LDA and LSI generated topics 1 to 6

Table 8: LDA selected topics with top 10 words and weights

	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6	
S/N	word	weight	word	weight	word	weight	word	weight	word	weight	word	weight
1	yahoo	0.170	nadal	0.028	yahoo	0.101	rt	0.095	rt	0.040	yahoo	0.234
2	rt	0.066	money	0.028	rt	0.062	yahoo	0.092	#whatwentwrong	0.018	rt	0.075
3	🤣	0.028	geng	0.026	pastor	0.039	arrested	0.042	order	0.013	bad	0.026
4	end	0.022	thiem	0.021	forget	0.037	updated	0.041	😅	0.011	really	0.023
5	said	0.022	around	0.020	😂	0.028	status	0.041	10	0.009	still	0.022
6	man	0.020	crush	0.018	adeboye	0.027	ps4	0.041	chop	0.009	support	0.022
7	someone	0.019	pls	0.017	🦋	0.026	bag	0.041	name	0.009	@adehdaboy	0.021
8	real	0.019	since	0.016	@jfriks	0.023	friend	0.041	available	0.009	let	0.021
9	😭	0.018	efcc	0.016	equal to	0.023	claiming	0.040	25	0.009	fact	0.021
10	country	0.018	self	0.016	followers	0.022	ojuelegba	0.040	upandan	0.008	face	0.021

Table 9: LSI selected topics with top 10 words and weights

	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6	
S/N	word	weight	word	weight	word	weight	word	weight	word	weight	word	weight
1	yahoo	0.087	rt	0.511	😂	0.403	😂	-0.382	fg	0.226	like	-0.294
2	rt	0.362	yahoo	-0.232	twitter	0.218		0.209	2020	0.224	even	-0.258
3	😂	0.092		0.200	pastor	0.195	pastor	-0.180	january	0.224	lot	-0.248
4	bad	0.067	arrested	0.194	adeboye	0.190	nysc	-0.173	abacha	0.224	going	-0.246
5	pastor	0.062	bag	0.193	arrested	-0.185	adeboye	-0.171	3	0.224	terrifying	-0.246
6	really	0.060	friend	0.193	friend	-0.185	rt	0.165	loot	0.224	judge	-0.245
7	let	0.058	ps4	0.193	bag	-0.184	followers	0.164	2017	0.224	anybody	-0.245
8	support	0.058	policeman	0.193	ps4	-0.184	forget	-0.160	repatriated	0.224	racism	-0.245
9	still	0.058	status	0.193	policeman	-0.184	200	0.156	1.89	0.224	stuffs	-0.245
10	@adehdaboy	0.057	ojuelegba	0.193	updated	-0.184	100	0.156	n681billion	0.224	homosexuality	-0.245

The topics represent the point in the Multidimensional Scaling (MDS) graph, where the size of the point is a function of the Marginal Topic Probability (MTP) for each topic extracted from the tweet corpus. The bigger the size of the point, the stronger the topic is represented by the words in the corpus. Only the LDA topics are visualized using MDS because LDA is easier to interpret than LSI even though it is more computationally intensive. The visualization of the LDA topics with MDS shows that topic 6 has the highest marginal probability of 0.244889, followed and topic three and topic 4. Figure 12 shows the LDA topics using MTP with multidimensional scaling points.

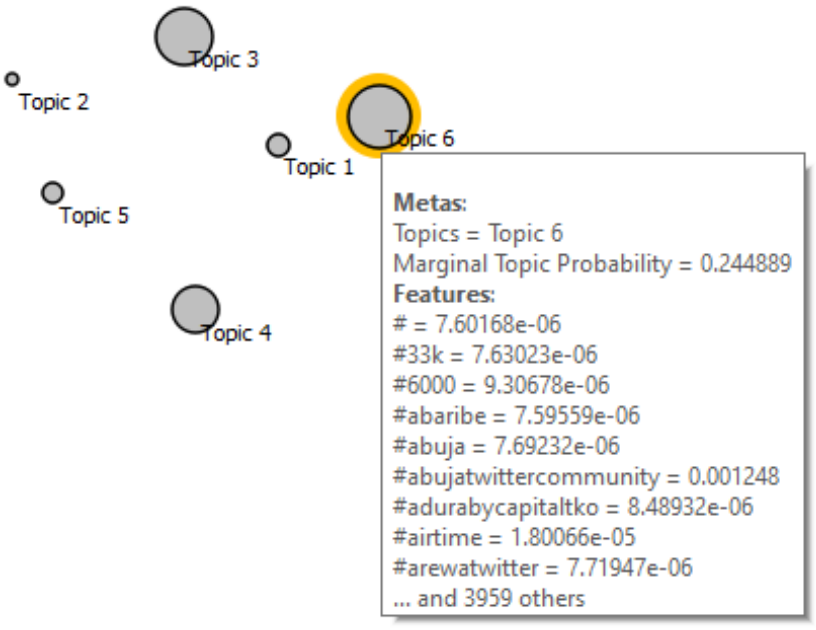


Figure 12: Visualizing LDA topics using marginal topic probability with multidimensional scaling points.

We further used the box plots to visualize the words that are most representative of each topic. The box plot sorts the variables (words) by separating the selected subgroup values. The subgroup *yes* represents the weights of the most representative words of the topic selected on the MDS graph. Table 10 shows the top ten most representative words for LDA topics 1 - 6 selected on the MDS graph and visualized on the box plot. The words are sorted by their order of relevance to the topics.

Table 10: Top ten words by order of relevance to the topic in the corpus extracted from the box plot

S/N	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	😂	money	Pastor	arrested	order	bad
2	end	geng	Retweet	updated	whatwentwrong	really
3	someone	nadal	forget	status	10	still
4	real	thiem	😂	ps4	chop	support
5	😭	around	adeboye	bag	name	@adehdaboy
6	without	crush	🫧	friend	available	let
7	ladies	since	@jfriks	claiming	25	fact
8	collects	efcc	=	ojuelegba	shout	face
9	t-fare	self	followers	policemen	upandan	lowering
10	@biyitheplug	laugh	100	@aproko_doctor	funds	pove

The second information that can be observed from the box plot is the notable separation between *yes* and *no* subgroups for topics with high MTP. The *yes* subgroup represents the words for the selected topic, while *no* subgroup is the other words in the corpus. The subplots in Figure 13 show how the selected LDA topics from the MDS graph are displayed on the box plot. The box plot changes by closing up the separation between the *yes* and *no* subgroups. Subplot 6 of Figure 13 shows the good separation between the word frequency for LDA topic 6 and all the others. The MTP of each topic is shown by the *yes* subgroup of each topic as $2.41322e^{-5}$, $2.1151e^{-5}$, 0.226381, 0.195948, 0.112567 and 0.244889 for topic 1, 2, 3, 4, 5 and 6, respectively. The gap between the subgroups is consistent with the sizes of the points and MTP values of the topics, as shown in Figure 12.

5. CONCLUSION

In this study, we conducted a content analysis of Twitter data using 5,500 tweets from the yahoo yahoo hashtag to assess social media opinion on the issue of cyber-crime issues such as "yahoo-yahoo" to the society. A convenience sample of opinions is used for the study via the social media application update on Twitter (tweets). A semi-structured Twitter data was collected from various verified and unverified authors. The result gives a detailed analysis on the sentimental view of people towards yahoo yahoo. Although the geolocation showed more users tweeted on the topic from Spain, a closer look into the corpus shows otherwise because of privacy concerns, and many users don't declare their location on Twitter. It can also be concluded that LDA and LSI modeled topics showed a more representative reflection of the tweet corpus. Although LSI is said to be more computationally demanding in literature and is often less preferable to LDA, we observed that the insight it provided by identifying negative representative words along with the positive representative ones is very significant to topic modeling and gaining insights from tweets. Emojis have strong weights in determining sentiments and contribution to topics modeling. The discussion towards yahoo-yahoo as a cybercrime was largely seen as negative to society. At the same time, equally positive and neutral sentiments were shared by 35.85%, 24.53%, 15.09%, and 24.53% for negative, neutral, No-Zone, and positive sentiment tweets, respectively.

Future Work

For future research, the Authors plan to collect; create long short-term memory (LSTM) deep learning models for sequence-to-label classification problem in tweets; test the model with new narratives to evaluate their performance; conduct internal and external validity of the study to ascertain that the result obtained are meaningful and trustworthy; collect very large historical tweet datasets on trending national issues such as #COVID-19Nigeria, #EndSARS, #LekkiMassacre, etc. for evaluating the proposed future research directions.

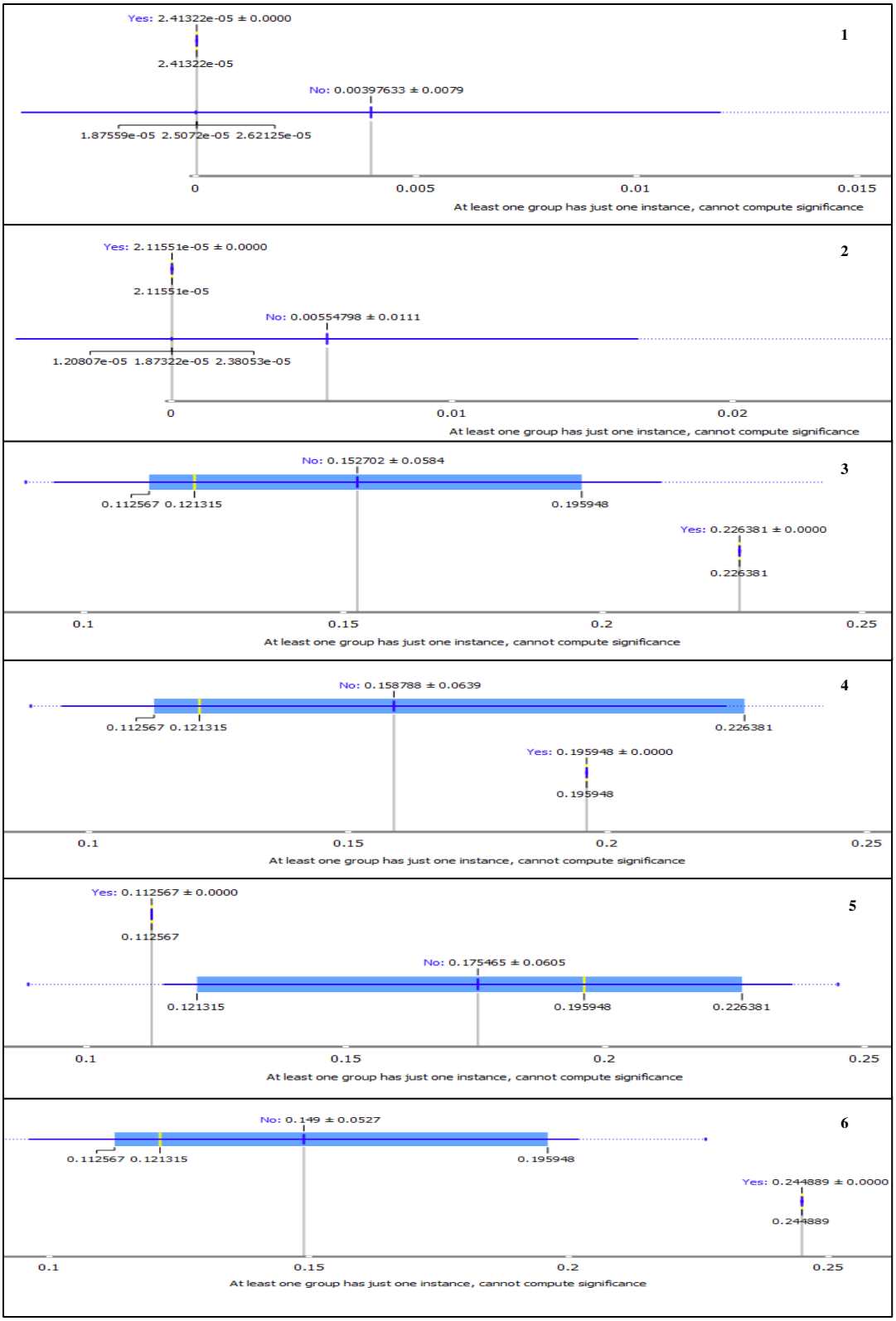


Figure 13: Box plots showing Marginal Topic Probability of the topics using Latent Dirichlet Allocation (LDA)

REFERENCES

Abayomi-Alli O., Sanjay M., Abayomi-Alli A., Odusami M. (2019). "A review of soft techniques for SMS spam classification: Methods, approaches and applications", Engineering Applications of Artificial Intelligence, 86(2019), 197-212, Elsevier B.V., Amsterdam, The Netherlands, doi:10.1016/j.engappai.2019.08.024

Adeniran, A. I. (2008). The Internet and emergence of yahoo-boys sub-culture in Nigeria. International Journal of Cyber Criminology, 2(2), 368–381.

Appel, G., Grewal, L., Hadi, R., Stephen, A. T. The future of social media in marketing. J. of the Acad. Mark. Sci. 48, 79–95 (2020). doi:10.1007/s11747-019-00695-1

Al-garadi, M. A., Varathan, K. D., Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behaviour, 63, 433-443.

Arimi, C. N. (2011). Social-economic factors influencing the crime rate in Meru Municipality Kenya. Master's (MA) Thesis, University of Nairobi, Kenya. http://erepository.uonbi.ac.ke:8080/handle/123456789/4688

AUC: African African Union Commission. (2016). Cyber Crime & Cyber Security Trends in Africa. Retrieved on the 14th of April, 2020. Available online at: https://www.thehaguesecuritydelta.com/media/com_hsd/report/135/document/Cyber-security-trends-report-Africa-en.pdf

Boyer, H. (2014). Emerging Technologies – Social Media. INALJ Virginia, URL: http://inalj.com/?p=62623

Burnap, P., Williams, M.L. (2015), Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy & Internet, 7: 223-242. doi:10.1002/poi3.85

Can, U., Alatas, B. (2019). A new direction in social network analysis: Online social network analysis problems and applications. Physica A: Statistical Mechanics and its Applications, 535 (2019), 122372. doi:10.1016/j.physa.2019.122372

Cheng, L., Guo, R., Liu, H. (2019). Robust cyberbullying detection with causal interpretation. In Companion Proceedings of the 2019 World Wide Web Conference. 169-175.

Demšar J, Curk T., Erjavec A., Gorup C., Hočevar T., Milutinovič M., *et al.* (2013). "Orange: data mining toolbox in Python". JMLR. 14 (1), 2349–2353.

Donchenko, D., Ovchar, N., Sadovnikova, N., Parygin, D., Shabalina, O., Ather, D. (2017). Analysis of comments of users of social networks to assess the level of social tension. Procedia Computer Science, 119, 359-367.

Drishya, S. V., Saranya, S., Sheeba, J. I., Devaneyan, S. P. Cyberbully Image and Text Detection using Convolutional Neural Networks. CiiT International Journal of Fuzzy Systems, 11(2), 25-30.

Figueira, Á., Guimarães, N., Pinto, J. (2019). A System to Automatically Predict Relevance in Social Media. Procedia Computer Science, 164, 105-112.

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In Proceedings of the 10th ACM Conference on Web Science, pp. 105-114.

Gupta, B., Sharma, S., Chennamaneni, A. (2016). Twitter Sentiment Analysis: An Examination of Cybersecurity Attitudes and Behaviour. Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics, Dublin. https://aisel.aisnet.org/sigdsa2016/17

Hariani K., Riadi, I. (2017). Detection of cyberbullying on social media using data mining techniques. International Journal of Computer Science and Information Security (IJCSIS), 15(3), 244-250

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J., Sanchez, V. (2018). Social sentiment sensor in Twitter for predicting cyber-attacks using ℓ1 regularization. Sensors, 18(5), 1380,1-17. doi:10.3390/s18051380

Hu, M., Liu, B. (2004). Mining opinion features in customer reviews. In Proceedings of the 19th national conference on Artificial Intelligence (AAAI'04). AAAI Press, 755–760.

Hutto C. J., Gilbert E. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, January 2015, pp. 1-10

Ibrahim, S. (2016). Social and contextual taxonomy of cybercrime: Socioeconomic theory of Nigerian cybercriminals,

International Journal of Law, Crime and Justice, Elsevier, 47(2016), 44-57. doi:10.1016/j.ijlcj.2016.07.002

Kounadi O, Lampoltshammer TJ, Groff E, Sitko I, Leitner M (2015) Exploring Twitter to Analyze the Public's Reaction Patterns to Recently Reported Homicides in London. PLoS ONE 10(3):e0121848. https://doi.org/10.1371/journal.pone.0121848

Kunwar, R. S., Sharma, P. (2016). Social media: A new vector for cyber-attack. In 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring), IEEE, pp. 1-5.

Kirik A. M., Çetinkaya A. (2018). The Use of Social Media in Online Journalism, 3rd International Eurasian Conference on Sport Education and Society, 15th-18th November 2018, Mardin, Turkey, 3, 1171-1187.

Labille K., Gauch S., Alfarhood S. (2017). "Creating Domain-Specific Sentiment Lexicons via Text Mining". In Proceedings of Workshop on Issues of Sentiment Discovery and Opinion Mining, Halifax, Canada, August 2017 (WISDOM'17), pp. 1-8. doi:10.1145/nnnnnnn.nnnnnnn

Lazarus, S., & Okolorie, G. U. (2019). The bifurcation of the Nigerian cybercriminals: Narratives of the Economic and Financial Crimes Commission (EFCC) agents. Telematics and Informatics, 40, 14–26.

Longe O. B., Abayomi-Alli A., Shaib I. O., Longe F. A. (2009). "Enhanced content analysis of fraudulent Nigeria electronic mails using e-STAT". 2009 2nd International Conference on Adaptive Science & Technology (ICAST), Accra, 2009, pp. 238-243, IEEE. doi:10.1109/ICASTECH.2009.5409717.

Liu, X., Fu, J., Chen, Y. (2020). Event Evolution Model for Cybersecurity Event Mining in Tweet Streams. Information Sciences. 524(2020), 254-276.

Ninalowo, A. (2016). Nexus of state and legitimation crisis. Prime Publications, Lagos.

Ojedokun, U. A., & Eraye, M. C. (2012). Socioeconomic lifestyles of the yahoo-boys: a study of perceptions of university students in Nigeria. International Journal of Cyber Criminology, 6(2), 1001–1013.

Omoroghomwan O. B. (2018). An Appraisal of the Activities of Economic and Financial Crime Commission (EFCC) on the Administration of Criminal Justice in Nigeria, ACTA Universitatis Danubius, 11(2), 174-193.

Rehurek R., Sojka P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, pp. 45-50.

Rossy Q. and Ribaux O. (2020). Orienting the Development of Crime Analysis Processes in Police Organisations Covering the Digital Transformations of Fraud Mechanisms, European Journal on Criminal Policy and Research, 26, 335–356, doi:10.1007/s10610-020-09438-3

Sharma, K., Bhasin, S., and Bharadwaj, P. (2019). A Worldwide Analysis of Cyber Security and Cyber Crime using Twitter. International Journal of Engineering and Advanced Technology (IJEAT), 8(6S3), 1051-1056.

Somayyeh A., Masoud M. (2018). Mining Twitter data for crime trend prediction. Intelligent Data Analysis, IOS Press, 22(1), 117-141. doi: 10.3233/IDA-163183

Tade, O., Aliyu, I. (2011). Social organization of Internet fraud among university undergraduates in Nigeria. International Journal of Cyber Criminology, 5(2), 860–875.

Van der Walt, E., Eloff, J. H., Grobler, J. (2018). Cyber-security: Identity deception detection on social media platforms. Computers & Security, 78, 76-89.

Zulfikar, M. T., Suharjito (2019). Detection Traffic Congestion Based on Twitter Data using Machine Learning, Procedia Computer Science, 157, 118-124, doi:10.1016/j.procs.2019.08.148

Figures

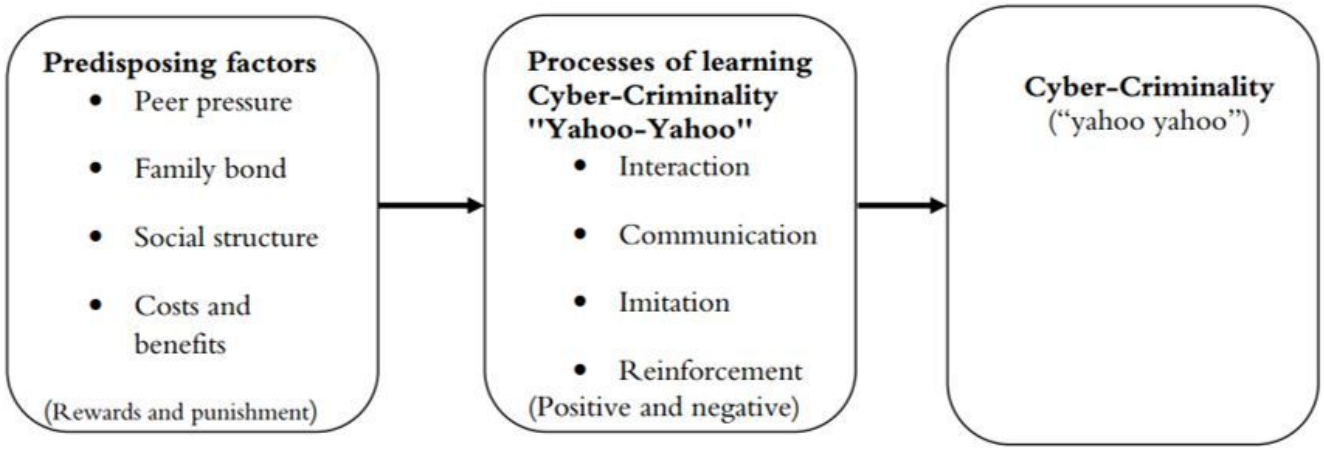


Figure 1

Relationships between predisposing factors and cyber-crime in Nigeria

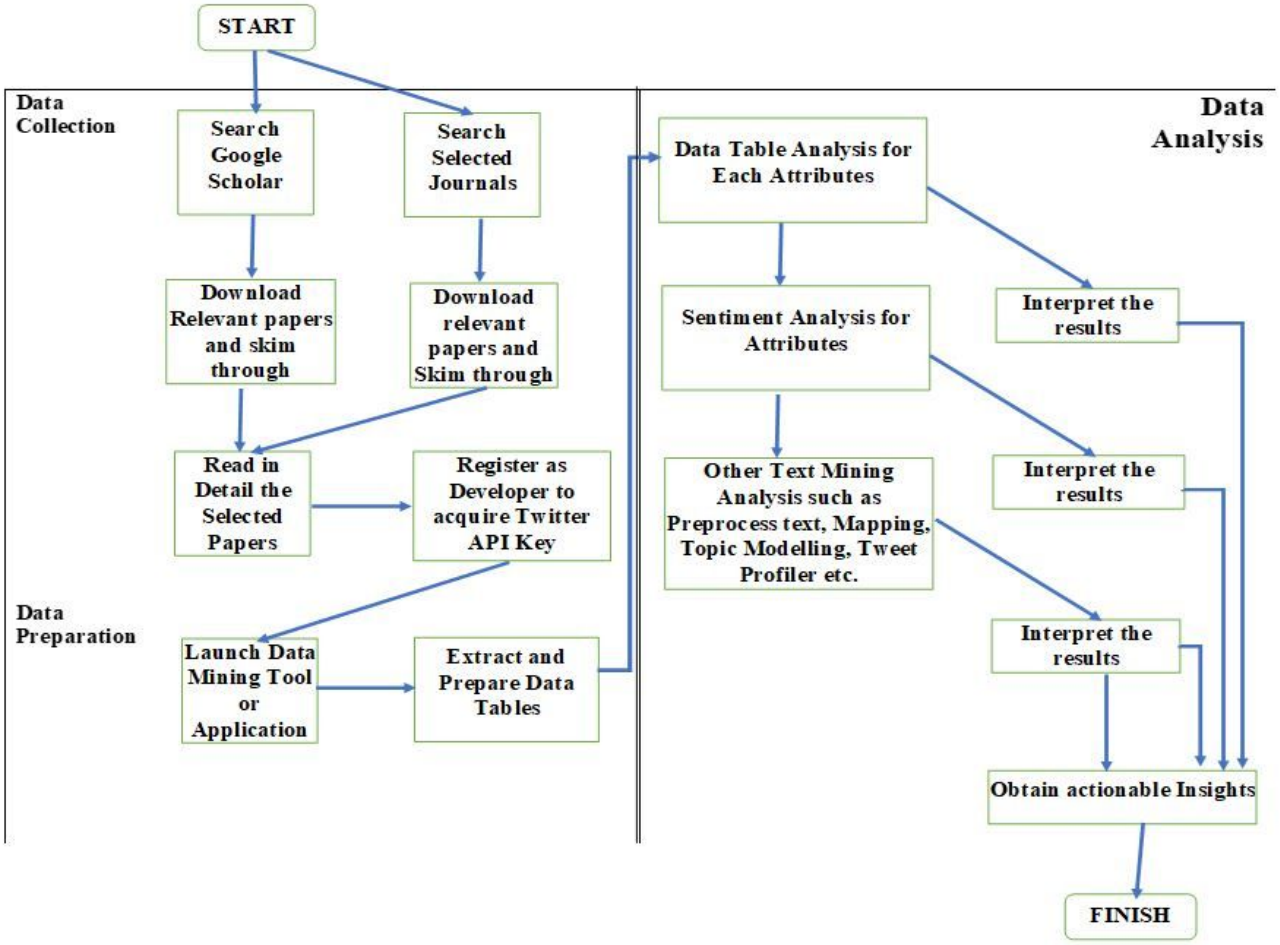


Figure 2

Proposed Content Analysis Framework

title	Author	Content True	Date	Language	Location	Number of Likes	ımber of Retwee	In Reply To	Author Name	uthor Descriptio	thor Statuses Co
1	@DavidAreks	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	ðŸ...°ðŸ...,ðŸ.....	Musician*Haart...	3205
2	@Davidsopuru	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	executive boyfri...	bio under const...	28241
3	@MuminAlao	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	OHIS	ARCHITECT |IN...	21391
4	@qwesi2131	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	Linus	#TeamGodfirst ...	45400
5	@Odulu_	RT @Sheddi_yo...	2020-01-30 11:4...	en	?	0	9	?	Demi god	Sharing my life ...	54969
6	@laura_Gainz	RT @BiyiThePlu...	2020-01-30 11:4...	en	?	0	131	?	ðŸ•MISSðŸ¦	Old account su...	5533
7	@TheBoy_Dina...	RT @efccnaija: ...	2020-01-30 11:4...	en	?	0	3	?	â™ Pablo Ruiz P...	leachmusic@ya...	18292
8	@E_Temple	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	Charles Ofomata	Too blessed to ...	40925
9	@Akaniiimoh	RT @TucodeBa...	2020-01-30 11:4...	en	?	0	178	?	Ivy's Hopeâ^\-ð...	I follow back ð...	1064
10	@jaywaxy	RT @onos_147: ...	2020-01-30 11:4...	en	?	0	202	?	-01 SÃƒVÃ‚GÃŠ	SNAPCHAT...@...	230474
11	@nnaemeka_a...	@vhic_tore Efc...	2020-01-30 11:4...	en	?	0	0	@vhic_tore	Anaco	I just want to se...	4360
12	@Udoji_Achebe	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Udoji_Achebe	I'm living life b...	1365
13	@someah_kwaw	RT @MazeDgre...	2020-01-30 11:3...	en	?	0	71	?	avril_sk	April 24thðŸ¶ F...	10083
14	@someah_kwaw	RT @Its_me_HE...	2020-01-30 11:3...	en	?	0	164	?	avril_sk	April 24thðŸ¶ F...	10083
15	@someah_kwaw	RT @fabi_mani:...	2020-01-30 11:3...	en	?	0	72	?	avril_sk	April 24thðŸ¶ F...	10083
16	@someah_kwaw	RT @MazeDgre...	2020-01-30 11:3...	en	?	0	61	?	avril_sk	April 24thðŸ¶ F...	10084
17	@phemoragh	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Striker	witty street wis...	15848
18	@Haboye_Cass...	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	KingMaxðŸ''	Jannah is Hom...	27903
19	@Amdennisgreat	Yahoo yahoo ht...	2020-01-30 11:3...	en	?	1	0	?	Dennis Great O...	Am Dennis Gre...	149
20	@laaolu	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Laaolu	I love wristwatc...	1317
21	@IamMrLeB	RT @Adehdabo...	2020-01-30 11:3...	en	?	0	501	?	IamMrLeBðŸ'¥...	The innocent b...	8573
22	@Bankole71376...	RT @onos_147: ...	2020-01-30 11:3...	en	?	0	202	?	Bankole emma...	Iâ€™m a simple...	112
23	@Horlanrehwaj...	RT @ogoon81: I...	2020-01-30 11:3...	en	?	0	97	?	General Ianreano	Sports:Liverpoo...	860
24	@LincolnsKE	RT @BiyiThePlu...	2020-01-30 11:2...	en	?	0	131	?	LincolnsKEâ„¢	|| Internet Entre...	160815
25	@gepherallity	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	rhin3O!	dreams money ...	18286
26	@Azubbie	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	Zubs	Christ junkie, pr...	33897
27	@Kelvin53146813	RT @oluwadee...	2020-01-30 11:2...	en	?	0	237	?	Kelvin Alandou...	?	11278
28	@oyogist	EFCC arrests la...	2020-01-30 11:2...	en	?	0	0	?	Oyo Gist	Bringing you o...	892
29	@BlvckJnr	RT @BiyiThePlu...	2020-01-30 11:2...	en	?	0	131	?	Ministress of w...	I follow back in...	3352
30	@AKINFAT	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	AKINFATZ	Roman Catholi...	3792
31	@uzo_agu	RT @onos_147: ...	2020-01-30 11:2...	en	?	0	202	?	John Mango	Business Analys...	49229
32	@AbdallahMai1	RT @Faisal_De...	2020-01-30 11:2...	en	?	0	1	?	RaptorðŸ...ð...	Manchester Uni...	922

Figure 3

Screenshot showing the Data Table with the Tweet contents and other metadata

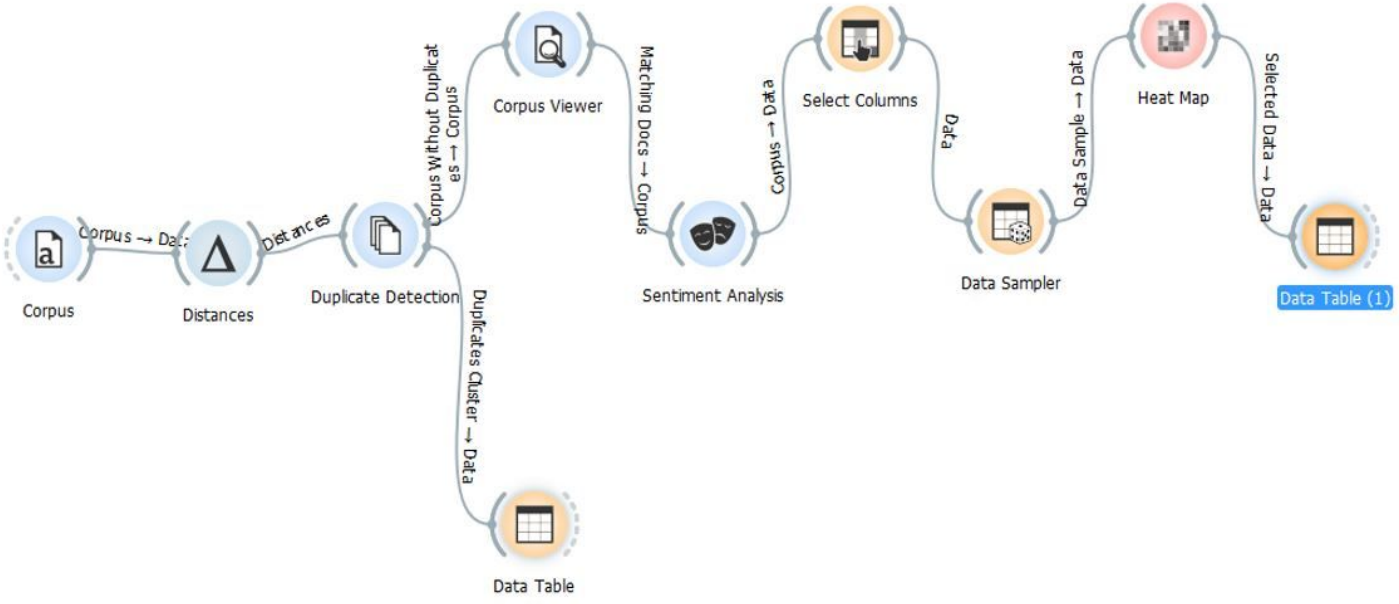


Figure 4

Sentiment Analysis and Duplicate Detection Model

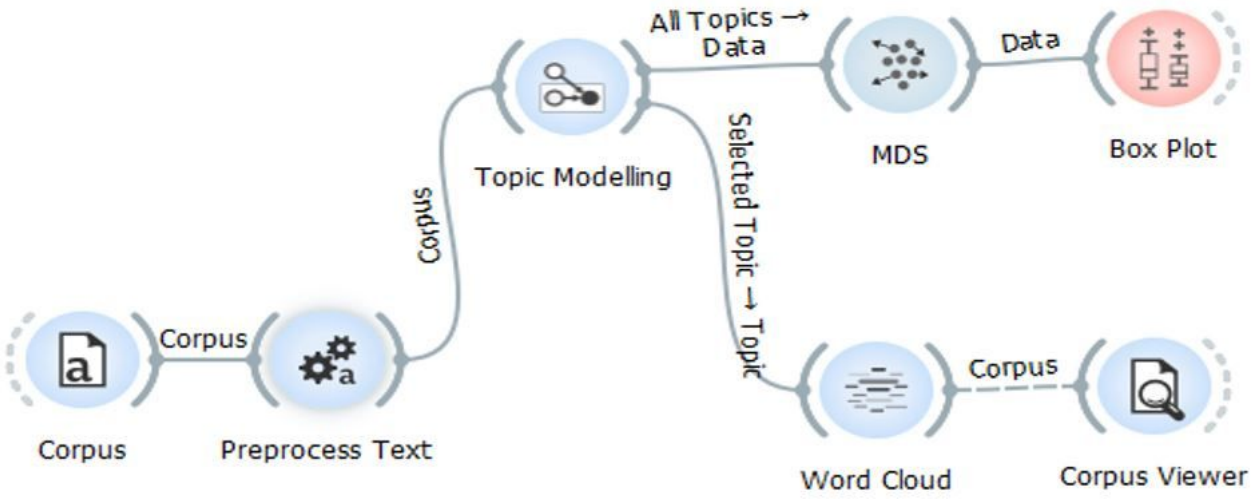


Figure 5

Workflow for the topic modeling

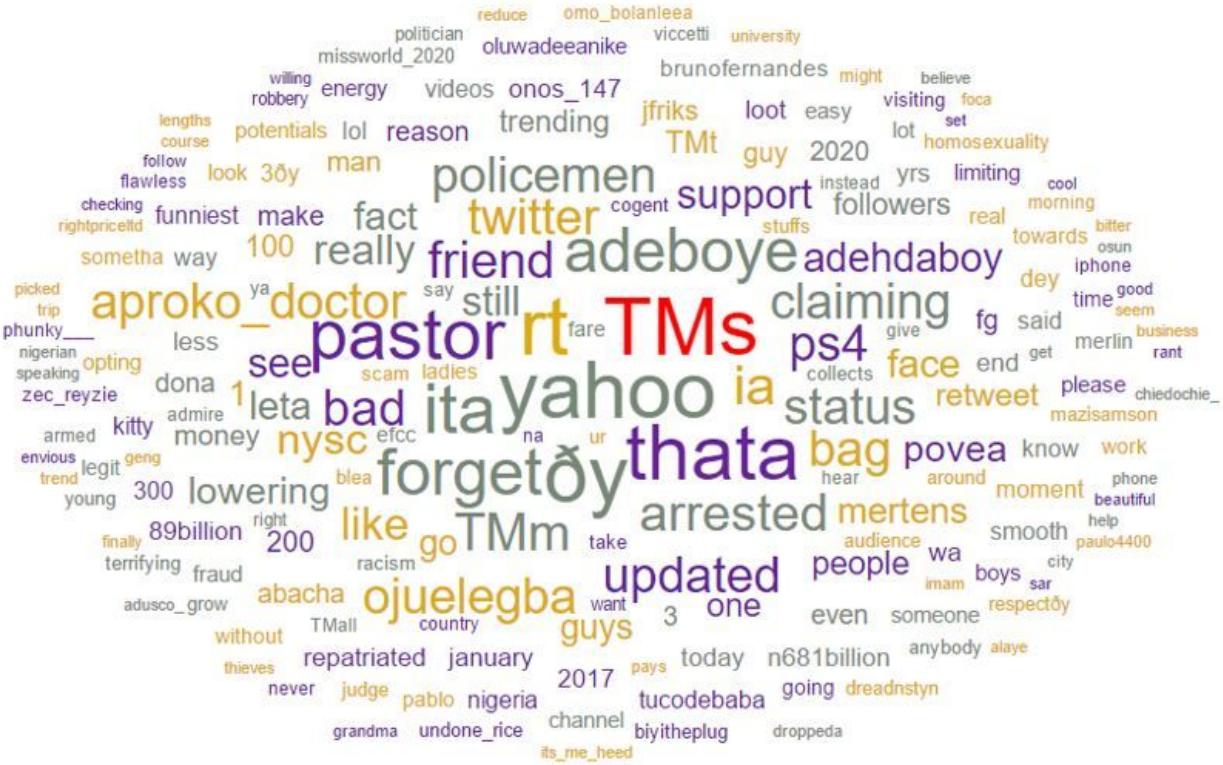


Figure 6

Word Cloud showing the tokens and their frequency/weight from the pre-processed tweet dataset

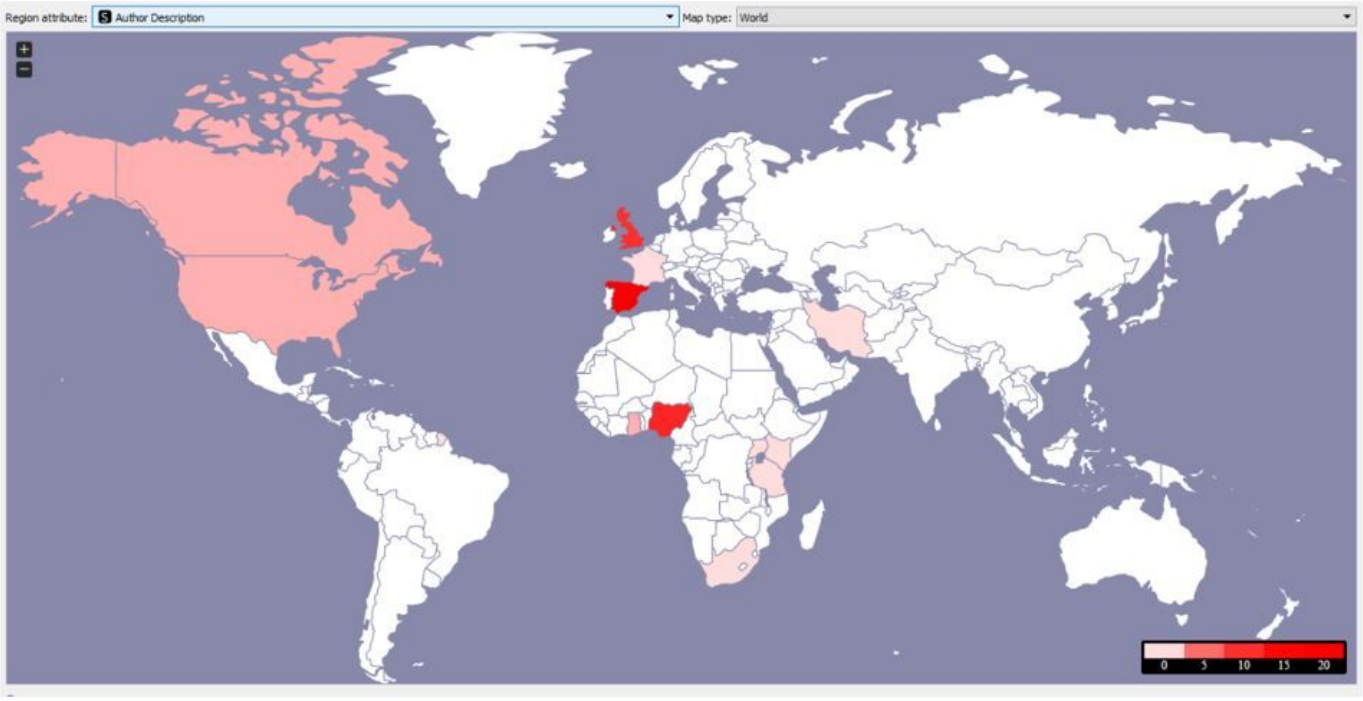


Figure 7

Map Showing Location of Author representing the frequency with colour weight Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

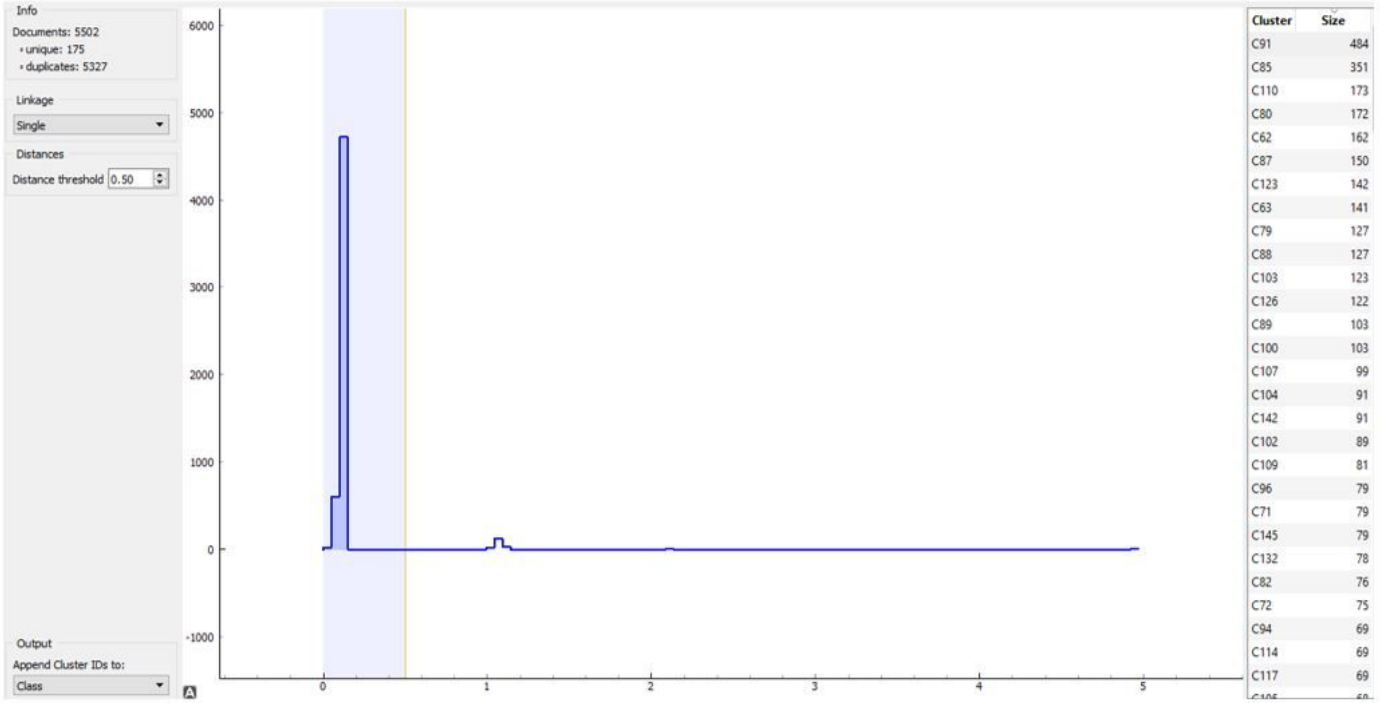


Figure 8

Showing the output of the duplicate detection widget

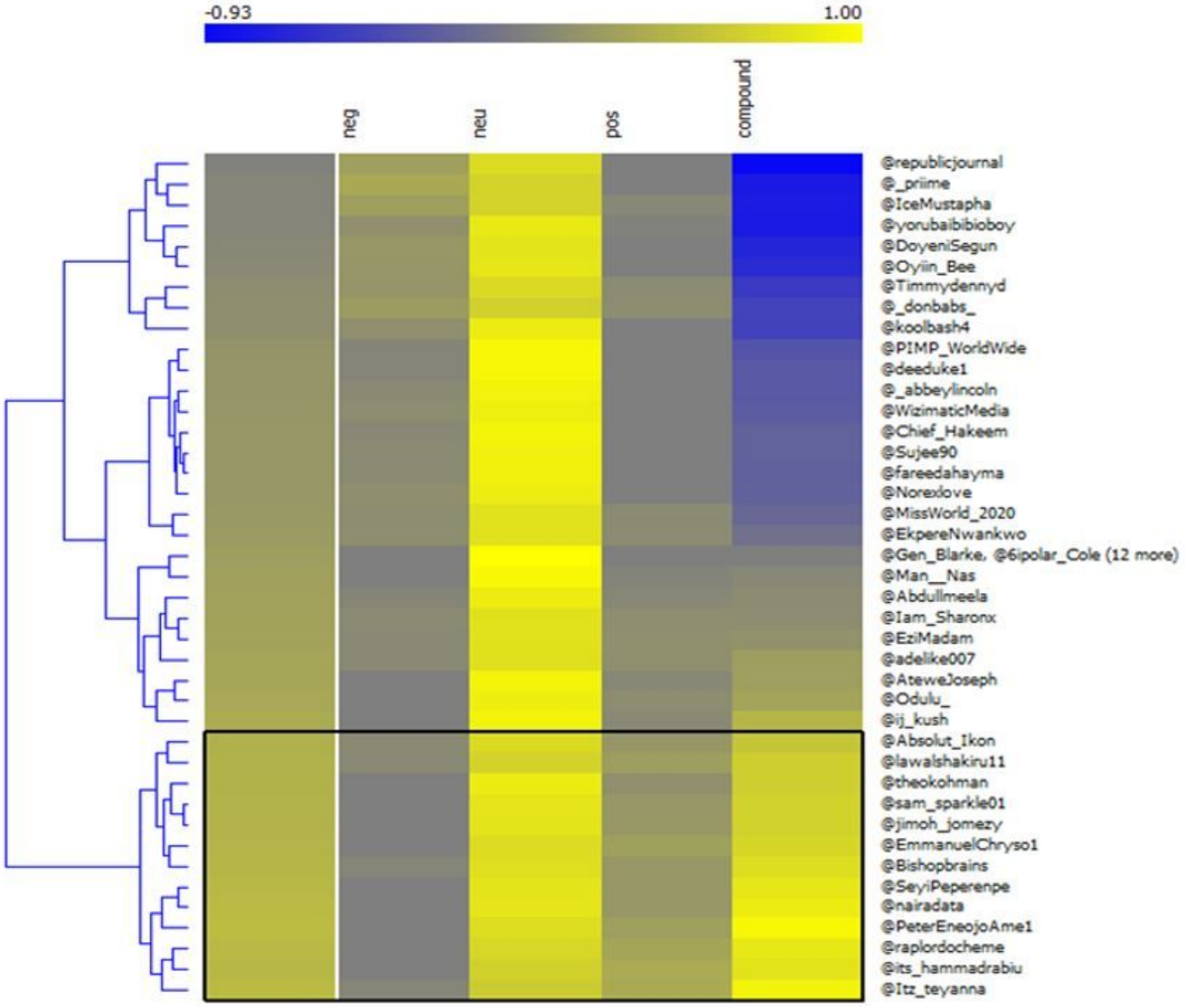


Figure 9

Heat map showing positive and negative sentiments classification by Vader method

Topic	LDA Topic Keywords
1	yahoo, rt, go, like, said, fraud, end, reason, man, real
2	yahoo, rt, bad, arrested, updated, status, ps4, bag, friend, ojuelegba
3	trending, 😂, make, merlin, nadal, know, time, someone, take, ur
4	rt, geng, 😅, set, 😔, order, @mazedgreat, everyone, sars, 10
5	money, 😭, thiem, since, #whatwentwrong, give, get, daddy, need, saying
6	yahoo, rt, pastor, retweet, forget, 😂, adeboye, 🐟, @jfriks, =

Topic	LSI Topic Keywords
1	yahoo, rt, 😂, bad, pastor, really, let, support, still, @adehdaboy
2	rt, yahoo, =, arrested, bag, friend, ps4, policemen, status, claiming
3	😂, twitter, pastor, adeboye, arrested, friend, bag, ps4, policemen, updated
4	😂, =, pastor, nysc, adeboye, rt, followers, forget, 200, 100
5	fg, 2020, january, abacha, 3, loot, 2017, repatriated, 1.89, n681billion
6	like, even, lot, going, terrifying, judge, anybody, racism, homosexuality, stuffs

Figure 10

showing LDA and LSI generated topics with keywords

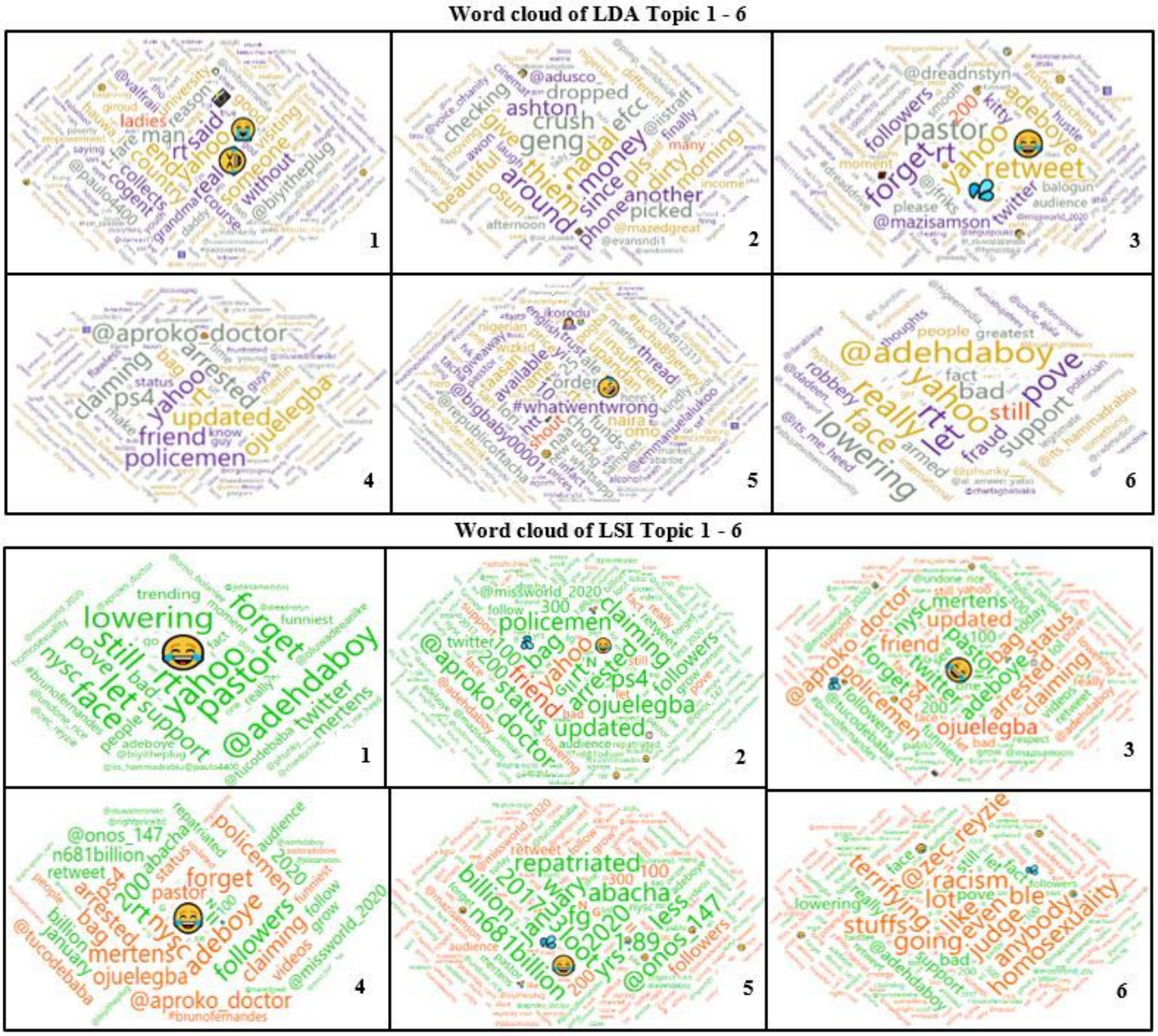


Figure 11

The cloud of words that constitutes LDA and LSI generated topics 1 to 6

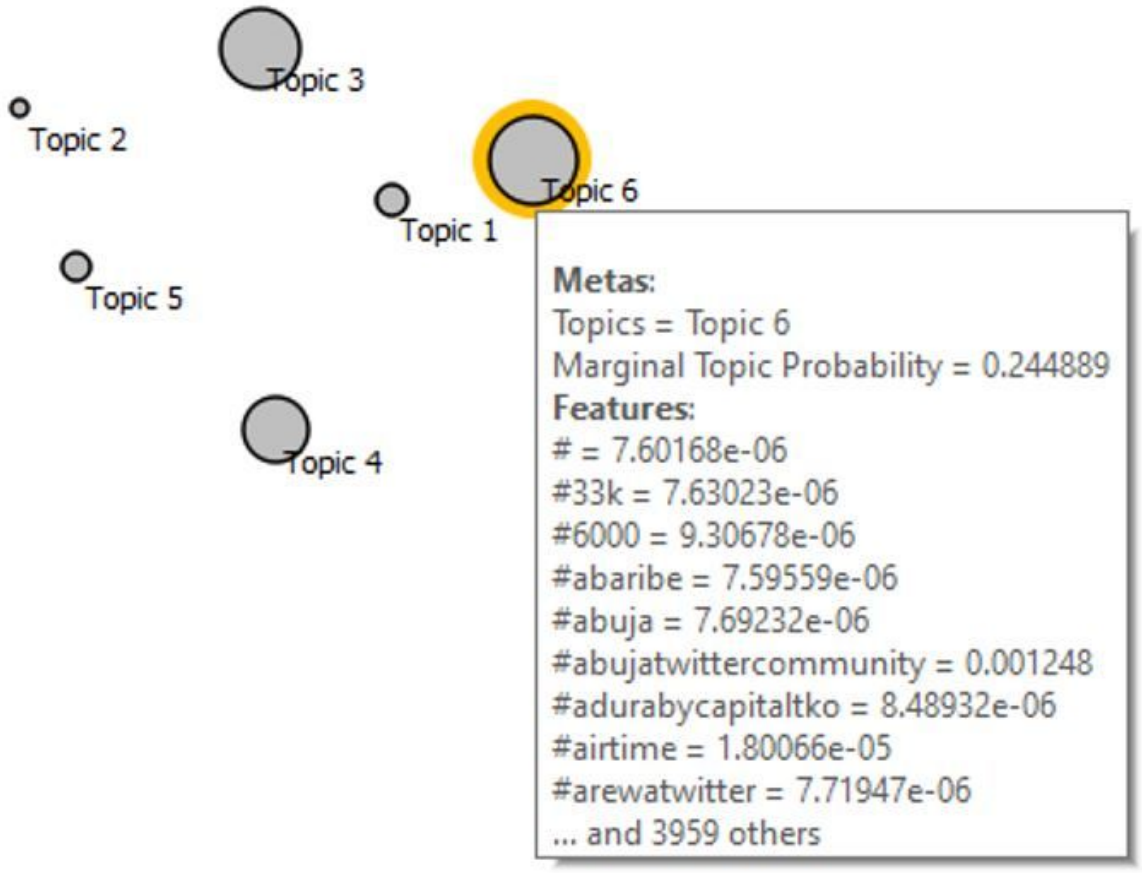


Figure 12

Visualizing LDA topics using marginal topic probability with multidimensional scaling points.

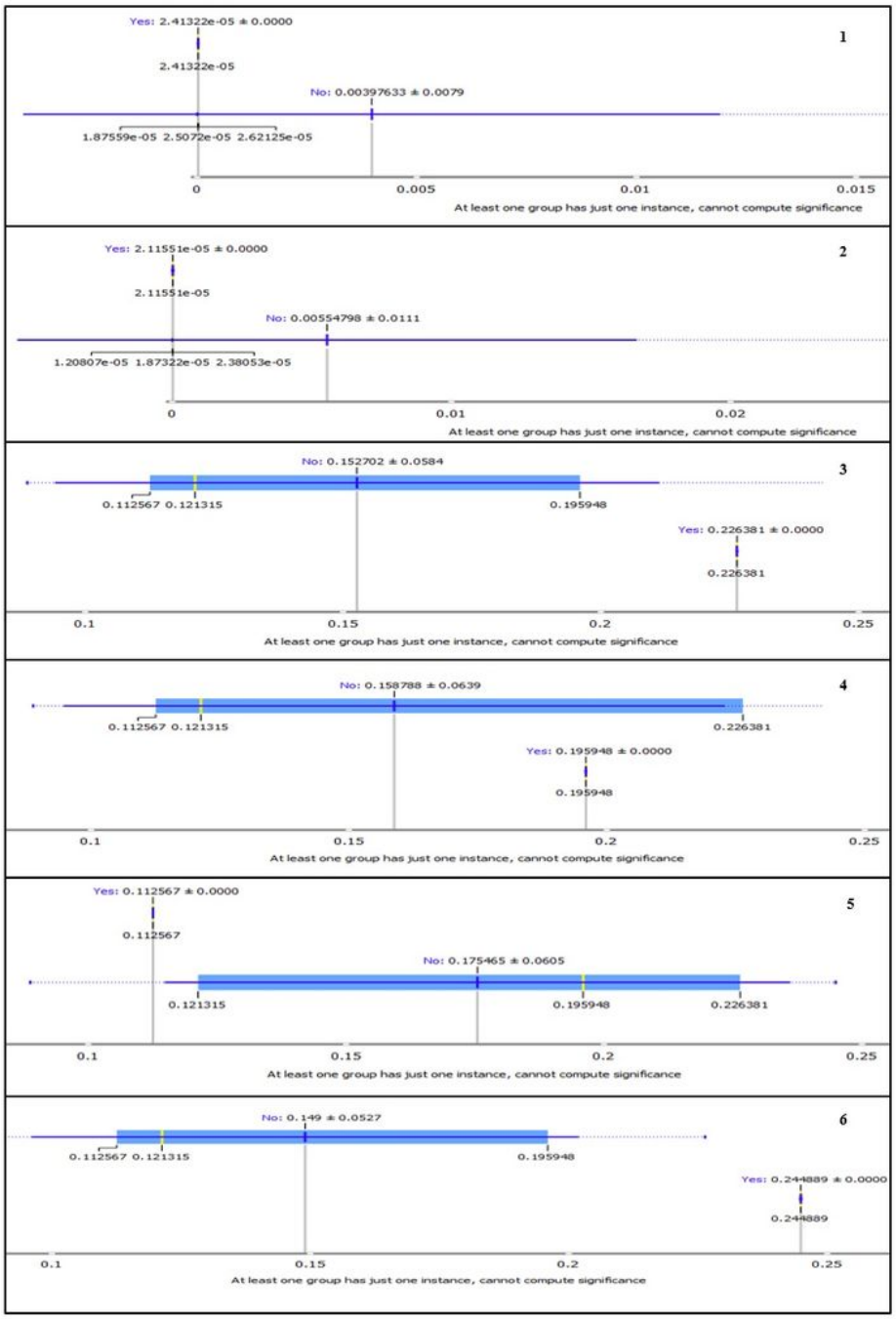


Figure 13

Box plots showing Marginal Topic Probability of the topics using Latent Dirichlet Allocation (LDA)