Namugera *et al. Comput Soc Netw* (2019) 6:3
https://doi.org/10.1186/s40649-019-0063-4

Computational Social Networks

RESEARCH Open Access

Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda

Frank Namugera[1*], Ronald Wesonga[1,2] and Peter Jehopio[1]

*Correspondence:
frank.namugera@aims-senegal.org
[1] School of Statistics and Planning, Makerere University, P.O. Box 7062, Kampala, Uganda
Full list of author information is available at the end of the article

Abstract

Unstructured data generated from sources such as the social media and traditional text documents are increasing and form a larger proportion of unanalysed data especially in the developing countries. In this study, we analysed data received from the major print and non-print media houses in Uganda through the Twitter platform to generate non-trivial knowledge by using text mining analytics. We also explored the determinants of derived sentiments in Twitter messaging. The results show that sentiments generated from tweets derived from the main print media houses (Daily Monitor and New Vision) were positively correlated, so were the sentiments from the non-print media (NBS TV and NTV) for the study period. Most of the sentiments on topics of security, politics and economics were found to be negative, while those on sports were positive. Furthermore, the tweet sentiment statistical logistic model revealed that negative sentiments were determined by the retweet status, retweet count and source of the tweets. Moreover, the positive sentiments were determined by the topic of discussion, type of media house and other sources of tweets ($p < 0.05$). Therefore, we recommend further extensions on the predictive statistical models to classify sentiments from social media based on the concept of big data analytics.

Keywords: Text mining, Twitter social media, Traditional media, Sentiments, Classification, Statistical models

Introduction

Text or unstructured data comprise approximately 80% of the data generated from vast fields including business, research and life science [1]. The nature of such data poses management and methodological challenges during analysis. However, if well handled it could be a vital source of knowledge for planning and decision making in many aspects [2]. Improvement in technology enhances increase in text databases and hence making the study of text mining a core field in data analysis because it deals with technologies of extracting new non-trivial knowledge from the huge textual datasets. Seemingly, the rate at which the data are being generated is increasing more than the rate at which the same data are being analysed. All documents that are generated in the form of unstructured format contain useful information in its raw form. However, due to its magnitude, this type of data has become a complex process for individuals to conduct summaries and statistical analysis for such information [3].

Social media is a group of internet-based applications that improved on the concept and technology of Web 2.0 which enables the formation and exchange of user-generated content [4]. With growing use of social media from applications such as Facebook, Twitter, Instagram, Myspace and LinkedIn, a lot of information is being exchanged in these platforms in most developing countries like Uganda. These platforms contain a lot of untapped potential of generating non-trivial knowledge and are beginning to be used to effect changes in social, political and economic arenas. Information retrieval, information extraction, trends analysis, classification, associations are among the machine learning and statistical methodologies that could be developed to generate relevant and timely information from the social media platforms. In Uganda, little efforts have been devoted into the use of information on social media platforms. This implies that there exist large amounts of information which have not been analysed and utilised. In this study, we sought to identify new knowledge from social media using text mined from the traditional media houses.

With social media, data are generated and disseminated in the public domain. This type of data is of interest to many stakeholders and beneficiaries within different economies and sectors. This is mainly because of the unstructured and uncensored modes of delivery which may trigger different sentiments from the users into the public. Data from social media could be the reason for rise or downfall of many companies, governments and departments within organisations. Social media has changed the livelihoods of people with communication ranging from healthcare, religion, sports, economics and politics. Communication on social media is in real time and the aspect of timely communication and accuracy in reporting is very important in areas like health [5], tourism [6], security and education. New surveys and research show that the use of social media is on the rise and is being promoted by the increase in the access to smart phones, which are very portable (and hence usable everywhere) [7].

Sentiment analysis is also at the forefront of informing public interest and can be used as a monitoring tool for the evaluation of projects, productions and decisions. Ignoring sentiments generated from the social media content may lead to an outburst of wrong decisions that result from not listening to people's feelings about an issue. Sentiment analysis can be approached as one or a combination of supervised, semi-supervised and unsupervised classification tasks. Lexicon and machine learning are popular approaches for sentimental classification. The lexicon-based approach [8–10] uses dictionaries of words annotated with their semantic orientations. The learning-based approach [8, 11] requires creating a model by training a classifier with labelled examples and finally the utilisation of the combination of both approaches [12]. A recent review on these techniques was presented in [11].

Two performance issues discovered with regard to lexicon approaches are (1) how to deal with context-dependent words and (2) how to address multiple entities with varying orientations within a single sentence. One of the suggested approaches for tackling these issues is the use of holistic lexicon [9] which involves exploiting external evidence and linguistic conventions of natural language expressions. The machine learning approach is reported to outperform the lexicon approach, yet suffers a general drawback of labelling large training data.

Table 1 Summary statistics of retrieved tweets

Source	Twitter handle	Number of tweets	Percentage of tweets	Number of words
NTV	@ntvuganda	7354	26.0	131,379
NBS TV	@nbs	10,194	36.0	189,728
NewVision paper	@newvisionwire	5794	20.0	96,145
Daily monitor paper	@monitor	4999	18.0	83,545

Text mining is an automated technique that uses computational algorithms to extract meaning and patterns from already existing text [13–15]. Primary research has traditionally been conducted by communicational studies such as surveys and interviews designed to collect data directly from consumers [16]. Text mining, however, allows for similar analysis by exploiting existing information online. Thus, the method discovers new knowledge by analysing and identifying the relevant information from large amounts of currently existing unstructured data. In addition, text mining aims at identifying relationships between words in sentences rather than just finding words in the way of a search engine.

The main objective of this study was to develop text mining models for knowledge discovery and sentimental surveillance on Twitter messages in Uganda. From related studies in relation to Twitter and text mining, different technologies have been used to achieve different objectives [17]. Asur and Huberman [18] used the twitter sentiments to predict the future income from the box office revenue for movies markets. This same study also used the rate at which tweets were received to predict the same revenues. Zhao [19] used the retweet count to study the most retweeted texts over the network. This retweet information is very important since it can be used as a predictor in models as [20] used it to build a model to predict whether a tweet message would be retweeted in the future. Yang and Wang [21] combined the sentiment with time series data to study the behaviour of football fans during a game. O'Connor et al. [22] used both sentiment and twitter time series for opinion mining.

Methods and data sources

Data sources

For this study, web crawling was used to extract tweets from the Twitter platform. Using the dominant print and non-print media houses in Uganda as the keywords, tweets were extracted from *@nbstv* for NBS TV, *@ntv* for NTV , *@newvisionwire* for New Vision and *@monitor* for Daily Monitor tweets. These covered a period of 9 days from 11/7/2017 upto 20/7/2017 as shown in Table 1.

Data processing

Generally, the text mining process is divided into four stages which included defining the concepts and context for mining, collecting data, dictionary construction, analysis and visualisation [15].

Since the twitter text data were unstructured, we performed some preprocessing to make the results more accurate and useful. These steps included changing text into

Table 2 LDA algorithm

Algorithm: Latent Dirichlet Allocation.
1: Initialise K the number of topics
2: Input: Document Corpus
3: Choose topic $\theta_i \sim \text{Dir}(\alpha) \forall i \in \{1 \dots M\}$ with α a sparse parameter.
4. Draw $\varphi_k \sim \text{Dir}(\beta)\ \forall k \in \{1 \dots K\}$
5. For each word position (i, j) where $\forall j \in \{1 \dots N_i\}$ and $\forall j \in \{1 \dots M\}$
Draw a topic index $Z_{i,j} \sim \text{Multinomial}(\theta_i)$
Draw a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$
6: Exit

lower-case, deleting words that were meaningless in the context and that occurred too frequently. We also deleted suffixes to maintain only the main body of a word if it occurred in a conjugated form [23].

After performing these steps to bring structure to the content, there were various time series and descriptive functions that we applied to the data before performing sentiment analysis and classification [24].

Statistical modelling

Topic modelling

An unsupervised classification algorithm with words randomly assigned under different topics was employed. Different approaches, namely Latent Dirichlet Allocation (LDA), probabilistic latent semantic analysis and correlation for topic modelling, are mainly used in topic modelling. Alghamdi and Alfalqi [25] give a detailed review; however, for this work LDA was used. The assumption was that these topics had been drawn from a Dirichlet distribution. For each tweet in our corpus,[1] we assumed the following algorithm to derive topics from the text as used in recent studies [25, 26]. Let M denote the number of tweets and N the number of words in a tweet. Below, we define the model parameters for the algorithm in Table 2 and Eq. 1:

Definition of model parameters:

- α is the parameter of Dirichlet prior on the per-document topic distribution.
- β is the parameter of Dirichlet prior on the per-topic word distribution.
- φ_k is the word distribution for word k.
- θ_m is the topic distribution for document m.
- z_{mn} is the topic for nth word in document m.
- w_{mn} is the specific word.

The LDA generative process is given in Table 2. The basic model structure was to assume that tweets are represented as random mixtures over latent topics, where each topic is characterised by a distribution over all the words [26]. The multinomial distribution deals with the categorical unordered variables.

Figure 1 presents the graphical model representation of the LDA model in plate notation where the outer plate represents documents/tweets, while the inner plate represents the repeated word positions in a given tweet.

[1] Collection of documents/tweets.

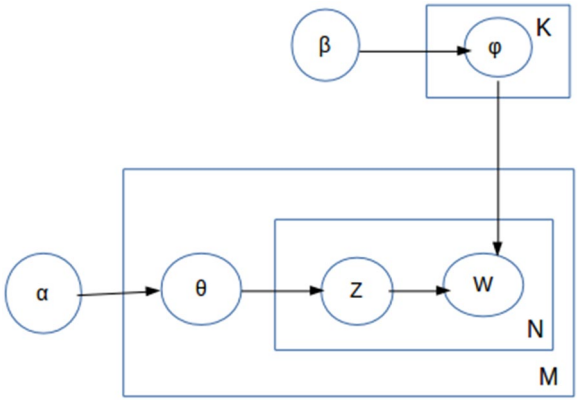

Fig. 1 Graphical representation of the LDA process and algorithm [25, 26]. The boxes are "plates" representing replicates

The probability of drawing a word from the tweets/corpus is given below [27],

$$p(w|\alpha,\beta) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta) p(z_n|\theta) \right) p(\theta;\alpha)\, \mathrm{d}\theta, \tag{1}$$

where the probabilities are explained in the algorithm above. The assumption is that the number of topics k is known implying that θ lies in $(k-1)$ dimension $\forall\theta_i \geq 0$, $\sum_i \theta_i = 1$.

Logistic regression model

A logistic regression model was developed to verify the determinants of the Twitter sentiments from the topic modelling, retweet_count, retweet_status and source of tweet with the sentiment of the final text.

In related work, different studies have used regression analysis models for twitter analysis and we discuss a few of them. Asur and Huberman [18] used a linear regression model to predict the box office revenue from movies given the sentiment polarity, rate of tweeting and distribution parameter. Hong and Davison [20] built a logistic regression model to predict whether a tweet would be retweeted and hence the retweet count plays a greater role. O'Connor et al. [22] also used lagged linear least squares model to predict poll outcomes by observing how fast the sentiments changed towards news events using the daily sentiment ratio as the predictor variable. Yang and Counts [28] used the Cox proportional hazards regression model to predict how information diffusion on Twitter through users' ongoing social interactions denoted by "@username" mentions by using the properties of the network that predict the speed, scale and range of information propagating through Twitter.

For this study, the outcome variable, $Y_i \sim B(n_i, \pi_i)$, is a binary outcome which is either positive or negative sentiment.

$$y_i = \begin{cases} 0 & \text{negative sentiment} \\ 1 & \text{positive sentiment} \end{cases}.$$

Table 3 Number of tweets by media houses per day

	Day	Weekday	NTV	NBS	Daily monitor	New vision
1	2017-07-11	Tues	608	998	200	329
2	2017-07-12	Wed	832	1046	464	641
3	2017-07-13	Thurs	968	1434	720	545
4	2017-07-14	Fri	639	1234	353	549
5	2017-07-15	Sat	546	699	556	451
6	2017-07-16	Sun	635	705	598	375
7	2017-07-17	Mon	902	942	547	608
8	2017-07-18	Tues	596	1097	524	504
9	2017-07-19	Wed	1171	1334	649	589
10	2017-07-20	Thurs	457	705	388	1203

Because of the nature of outcome, we proposed fitting a logistic model for the statistical modelling as given in Eq. 2,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i'\beta \tag{2}$$

$$\pi_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \quad , \tag{3}$$

where $i's$ are identical and independent observations, n is the number of predictor variables, X are the predictors and β are the coefficients.

Results and findings

Twitter time series analysis

Considering the number of tweets in the study period, we obtained the most anticipated topics or the most active media houses and also ascertained the gravity of the issues that were being discussed during that period. Table 3 shows a daily time series of the tweets per traditional media house over the period under study.

From Table 3, the number of tweets by the print and non-print media houses is presented. It was observed that Thursday 2017/07/13 registered the highest number of tweets across most media houses including *NTV, NBS and Daily Monitor*. A similar trend was noticed for the non-print media, i.e *NTV and NBS* on 2017/07/19 as seen from Fig. 2. However, on 2017/07/20 the *New Vision* print media registered a relatively very high number of tweets, thus requiring more analysis on the word usage.

We noticed that two peaks/spikes exist around 13 and 19 July as shown in Fig. 2 from the different media houses. These peaks in the time series show increased user activity and engagement on the different media platforms. Of interest was to discover the topics under discussion which caused the hikes in general activity.

Tweet rate per hour

We investigated the tweet behaviour for these media houses to see when the most tweets were received and ascertain the number of users on the platforms. The plots in

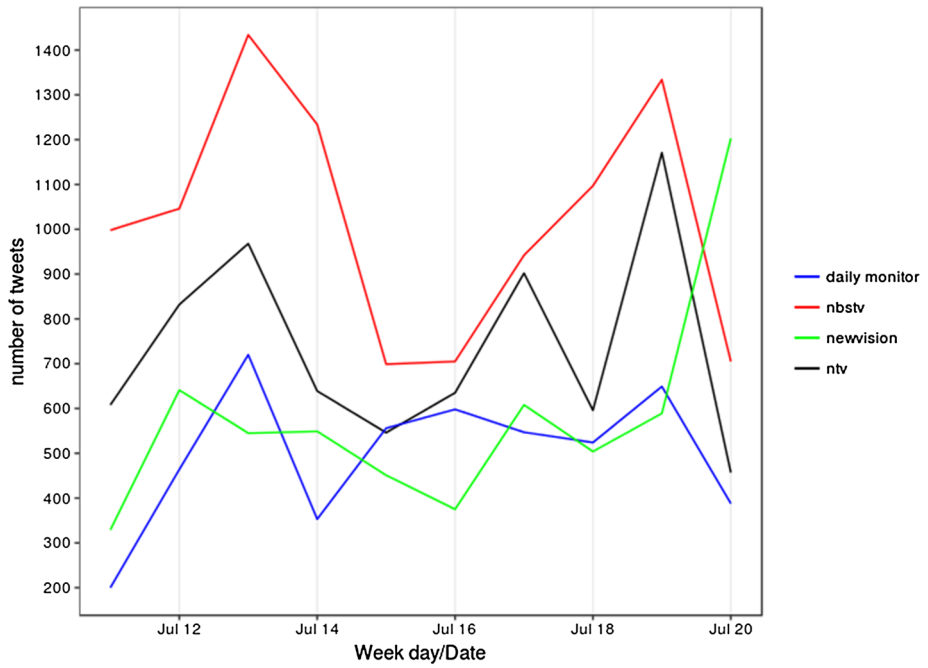

Fig. 2 Twitter time series for the 9 days from all the media houses

Fig. 3 display the tweeting behaviour of the televisions (*NTV* and *NBS TV*) and newspapers (*New Vision* and *Daily Monitor*) media houses. In Fig. 3, the box plots present the spread/distribution of the frequencies/count of tweets for the different hours during the day (0:00–23:59 h). This plot also shows some outliers that were observed on some days during the study period.

Considering the newspaper media houses, the most active hours were the morning hours starting from 03:00 h. This is the time when most newspaper media houses get to release the day's stories or new series of newspapers for the day. This implies that discussions commence once the newspapers are released. The trends of number of tweets gradually increase from the morning hours until a maximum is attained. It is after this point that we noticed a decline in the number of tweets being received through both media houses. The difference is that the tweet density for *Daily Monitor* is higher than *New Vision* paper. Table 4 represents the median count for each media house as represented in Fig. 3.

Similarly, for the television media houses there was a similar trend in the activity hours of the day. These activity hours were determined by the nature of programs posted. This means that similar programs were broadcast at similar times since the wave of activity hours was almost the same. However, the tweet density for *NBS* was heavier than *NTV*.

Followers/retweet count

Retweets are generated in the form of passing on information to another user. This is similar to making reference to an existing tweet while making a new tweet. Retweets may create new information about the matter being discussed in relation to the first/initial tweets. However, the number of retweets indicates the gravity or rate of public

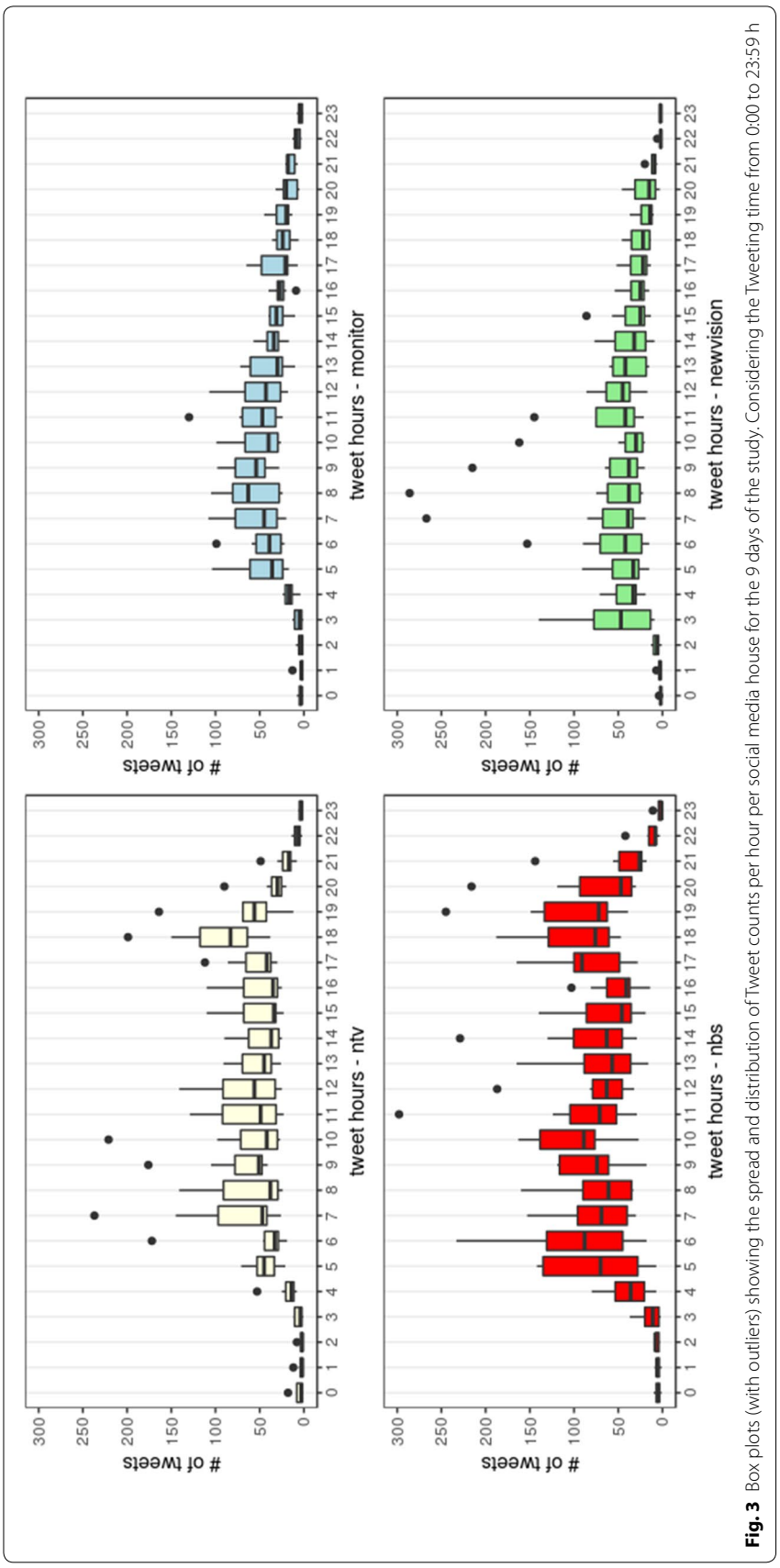

Fig. 3 Box plots (with outliers) showing the spread and distribution of Tweet counts per hour per social media house for the 9 days of the study. Considering the Tweeting time from 0:00 to 23:59 h

Table 4 The median number of tweets received per hour for the media houses over the study period

Hour	New vision paper	Monitor paper	NBS TV	NTV
0.0	2.0	3.0	5.0	3.0
1.0	2.5	3.0	6.0	2.0
2.0	6.0	4.0	8.0	2.0
3.0	47.0	5.0	11.0	4.0
4.0	33.0	17.0	36.0	14.0
5.0	33.0	36.0	70.0	45.0
6.0	42.0	39.0	88.0	33.0
7.0	39.0	45.0	69.0	47.0
8.0	38.0	63.0	61.0	38.0
9.0	38.0	54.0	74.0	51.0
10.0	30.0	40.0	89.0	42.0
11.0	42.0	47.0	71.0	49.0
12.0	45.0	43.0	63.0	56.0
13.0	42.0	30.0	57.0	45.0
14.0	32.0	34.0	63.0	37.0
15.0	25.0	31.0	46.0	34.0
16.0	25.0	27.0	41.0	35.0
17.0	22.0	21.0	91.0	42.0
18.0	22.0	24.0	76.0	83.0
19.0	14.0	20.0	72.0	56.0
20.0	15.0	20.0	47.0	30.0
21.0	9.0	18.0	26.0	17.0
22.0	2.0	8.0	8.0	7.5
23.0	2.0	4.0	1.0	3.0

interest and participation in the discussions that came in through the initial tweet. Figure 4 shows line plots of how retweets were comparing with the original tweets with the retweet status (is_retweet) showing which were retweets (is_retweet = True) and also the new tweets (is_retweet = False).

For the different days monitored, the plot shows that on average the retweets were higher than original tweets for the different social media houses. This implies that many discussions were derived from what was being tweeted causing high public participation. Of interest would be to discover the emotional attachments being generated together with the retweets.

Sentiment analysis

The study of emotions is part of Natural Language Processing (NLP) and text mining. In this study, we examined how sentiments varied over time in the conversations during twitter usage in the print and non-print media houses. We achieved this by tokenising and attaching a sentimental score to every word/term within the lexicon and using a calibration[2] of 50 consecutive terms/words in the order in which the tweets were being received.

[2] Calibration—after tokenising the tweets into single terms, we grouped the consecutive words into indices. All indices have a constant number of term/words.

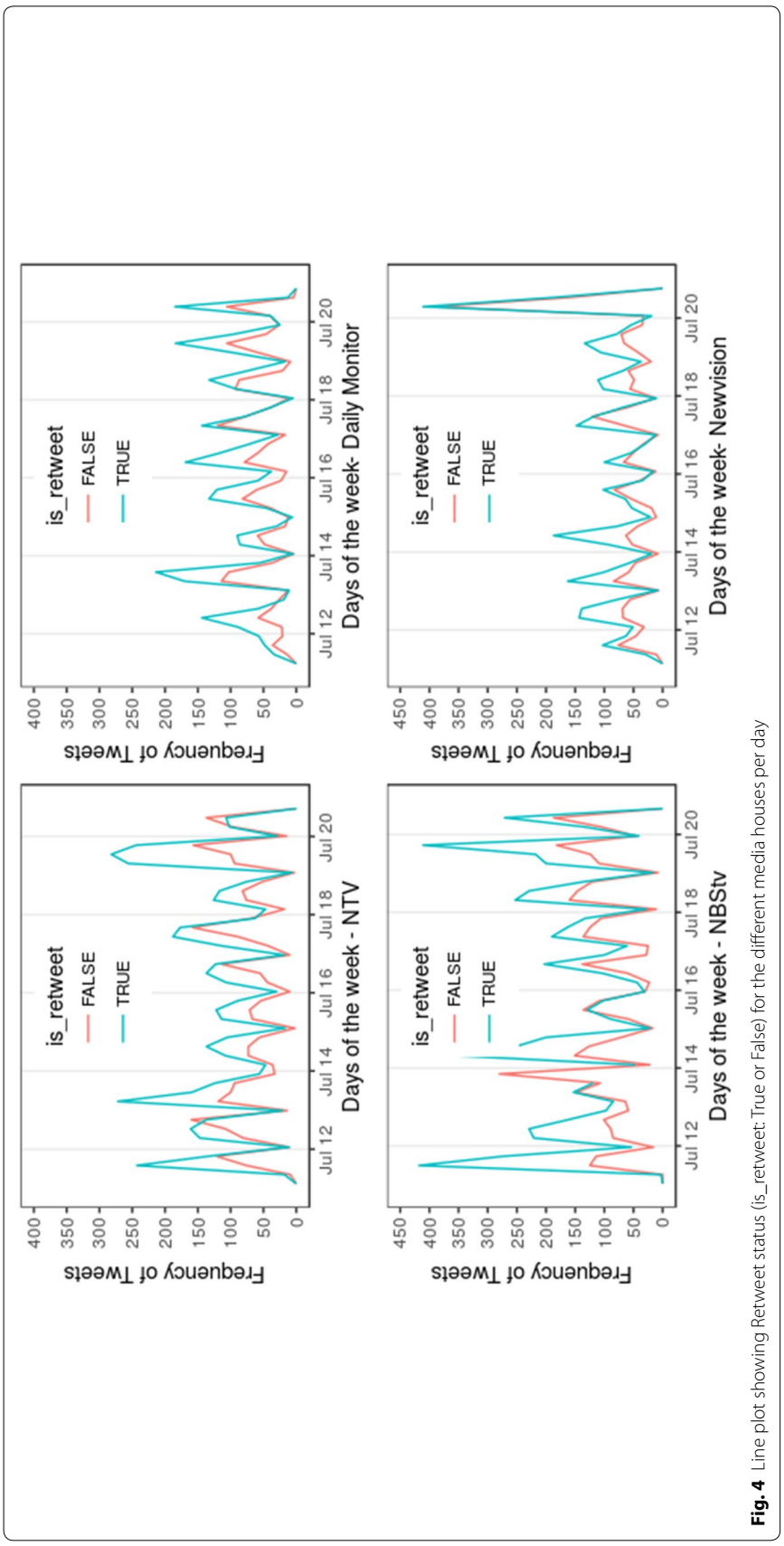

Fig. 4 Line plot showing Retweet status (is_retweet: True or False) for the different media houses per day

In this study, the sentiment score used was adapted from *"bing"* lexicon in R language since there was no predefined newspaper lexicon. The lexicon approach was used because words were separated logically into *positive:* $(+1)$, *neutral:* (0) and *negative:* (-1) as the polarity and the corresponding extent levels compared to other lexicons. Lexicon generation was based on different approaches and Liu [29] explains the dictionary approach and corpus-based approach. Using classification algorithms in the dictionary/lexicon approach, we present the results of the sentiment analysis in Fig. 5. The plot represents the cumulative sentiment score per index on the y-axis and the different indices on the x-axis. From this plot where the bars are below 0 on the y-axis, it implies that the sentiment in that index is cumulatively negative and where the bars are above the 0, it implies that the general sentiment in the index is positive where an index contains 50 words.

Of interest were the trends in sentiments over time in the different media houses. They all portrayed a similar trend of sentiments over the same time frame alternating from negative polarity to positive as seen in Fig. 5. It was noticed that in all media houses, there were general negative sentiments up to index 50. After this point, a change in trend of sentiments to a positive inclination in all media houses which was short lived for about 10 indices was observed. The trend then changed direction to the negative sentiments. *New Vision paper*, however, shows a different trend towards the end as compared to the rest. As the rest end at a negative note, it ends with a positive trend direction.

There is a positive relationship between *Daily Monitor* and *New Vision* tweets; similarly, a positive relationship between the *NBS TV* and *NTV* tweets is observed in Fig. 5. However, the study considered the *Pearson correlation* coefficients to assess the extent of the relationship. To allow for the computation, the first 100 indices were considered. Table 5 shows significant positive relationships exist between the print media/newspaper Twitter pages and also in the non-print/television Twitter pages, respectively.

This relationship was further justified to be significant by the p values computed in Table 5 at the $\alpha = 0.05$ level of significance. From this analysis, it can be concluded that newspapers have a similar trend of stories and report about similar opinions so do the non-print media houses in Uganda. Conversely, print media and non-print media do not tweet on the same topics.

Sentiment and time series analysis

The sentiment analysis and time series plots displayed features of interest in the analysis. In the time series analysis as discussed in "Twitter time series analysis" section, we discovered two peaks/spikes: the first in 12–14 July and the other in 18–20 July. This implies that there was active participation in the discussions that were handled in all the media houses generally though the *New Vision* portrayed a different path in the end.

Sentiment analysis on the other hand has also a unique trend that cuts across all media houses. In "Sentiment analysis" section, it was noticed that there were 3 sections in the

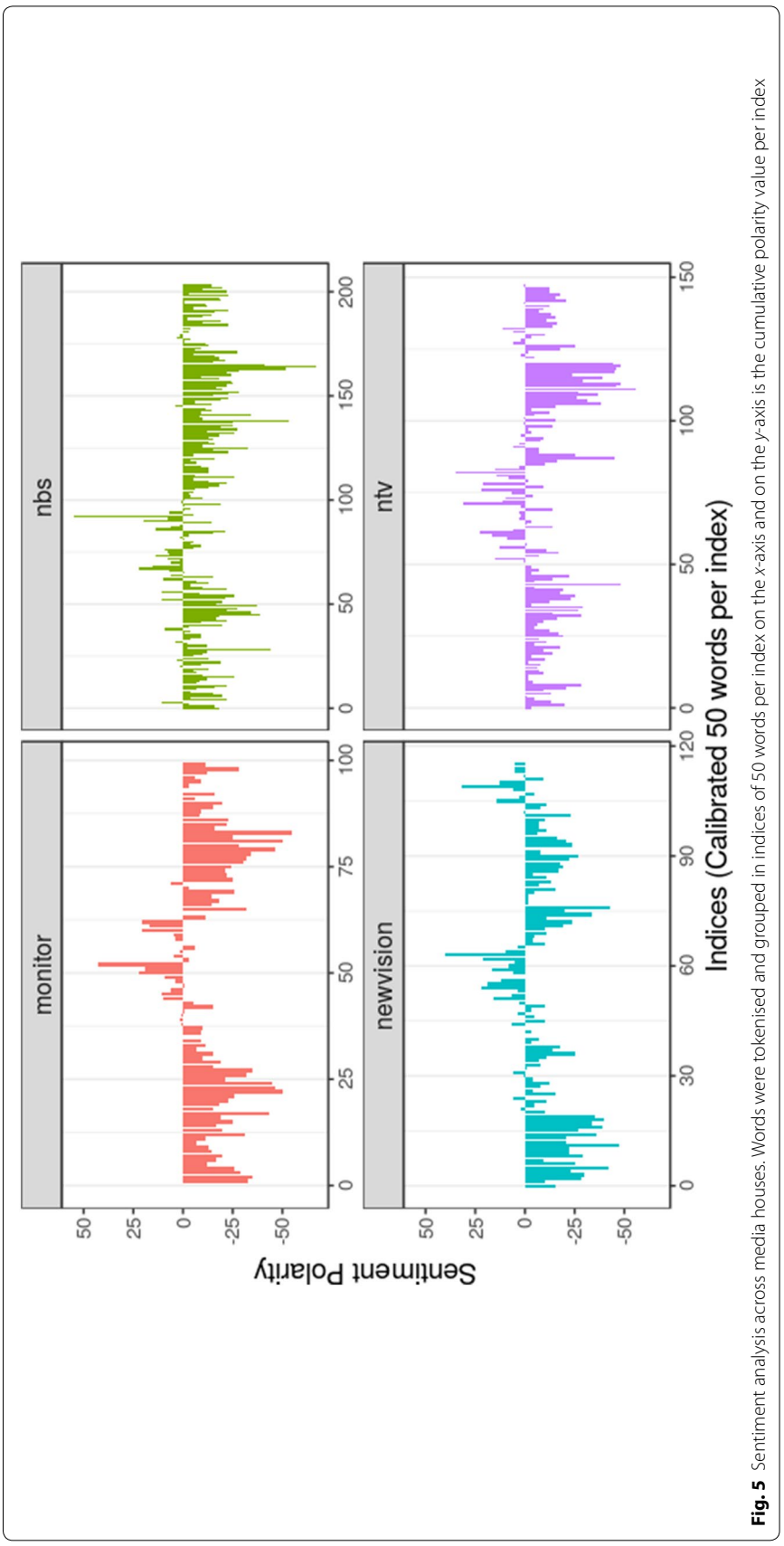

Fig. 5 Sentiment analysis across media houses. Words were tokenised and grouped in indices of 50 words per index on the *x*-axis and on the *y*-axis is the cumulative polarity value per index

Table 5 Pearson correlation coefficients table comparing sentiments of the major media houses

	Monitor		Newvision		NBS		NTV	
	ρ	*p* value	*ρ*	*p* value	*ρ*	*p* value	*ρ*	*p* value
Monitor			0.44	0.00	− 0.03	0.79	0.11	0.28
Newvision	0.44	0.00			− 0.00	0.98	0.01	0.89
NBS	− 0.03	0.79	− 0.00	0.98			0.23	0.02
NTV	0.11	0.28	0.01	0.89	0.23	0.02		

plots: two sections of negative sentiments and one section of positive sentiments in between the two types of media. In relation to time series analysis, we can hypothetically conclude that when the sentiments are generally negative, there is a high participation in discussion with many tweets and retweets as seen from Fig. 4 that shows a high number of retweets. Figure 6 is a combined plot for both time series analysis and sentiment analysis. On the upper side of the plot, the line plots represent the time series of tweets within the different social media in relation to the sentiment which are presented by bar plots on the lower side. The sentiments have been calibrated in indices of per day and the sentiment score provided on the *y*-axis is the cumulative daily sentiment score.

We noted that negative sentiments draw more attention from the public compared to the positive sentiment. Hence, sensitivity to the topics that are being discussed is key in trying to understand the cause of the Twitter time series direction.

Topic modelling

We employed the Latent Dirichlet Allocation (LDA) to discover the hidden topics discussed in the different media houses. For this study, four topics were pre-assigned to the algorithm and hence words were allocated into these 4 topics. Results of this algorithm are summarised in Table 6. The study then considered the words that were allocated under each topic to assign meaning to the bag of words under each topic.

Since these are national media houses, the discussions within them seem to be similar as would be expected/assumed. The difference in *New Vision* could be an issue of news bias; that is why there was a difference in the reporting and topics. This also can have an issue to do with the ownership of the media houses and the editing skills of these media houses or reporting interest.

Sentiments in the modelled topics

On discovering the probable topics of discussion, we then focused on understanding the underlying general sentiments of the public behind the topics. Therefore, the study considered the daily sentiment scores for the different topics. The topics that appeared to be outstanding in terms of text content included *politics, economics/banking, security/police, sports*.

Figure 7 is a plot showing the sentiment break down per topic per day. On the *y*-axis is the cumulative polarity value and on the *x*-axis are the days considered. The plot shows that *politics* and *economics/banking* had a general negative sentiment for

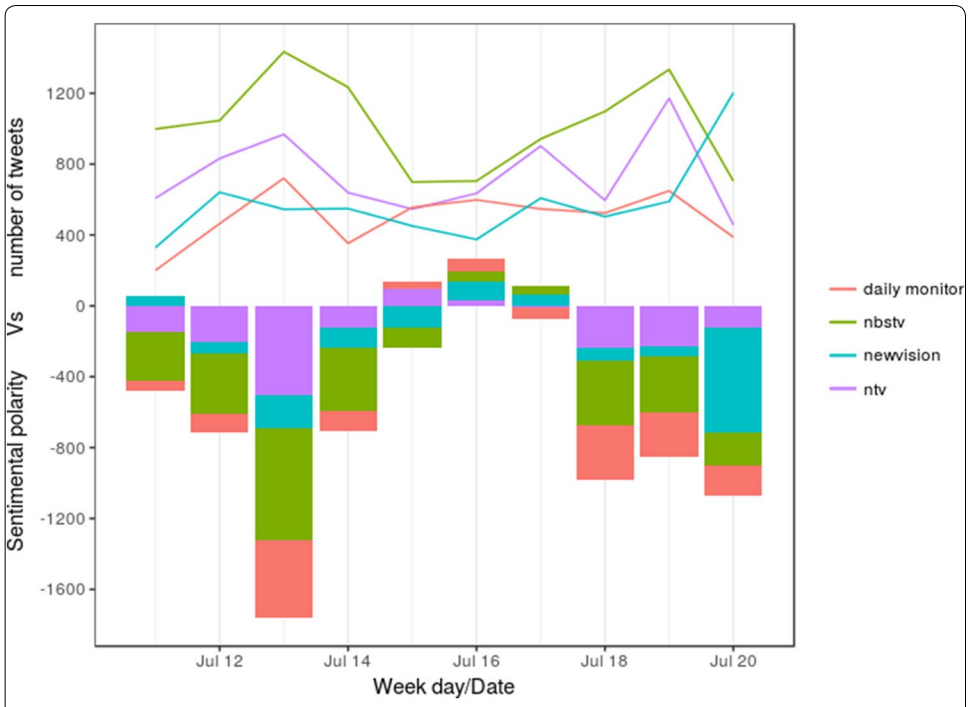

Fig. 6 A joint plot showing the relationship between sentiment (lower plot) and time series (upper plot) analysis for the Tweets obtained from the different media houses with respect to time. The *y*-axis represents both the number of tweets against the sentiment score for the respective days

the period of time considered for the data collection, whereas the other topics vary between negative and positive with the time variation.

Sports discussions in Uganda are well anticipated and this is explained by a generally positive sentiment trend as shown in Fig. 7. Depending on the purpose of the intelligence, one may choose to focus on any of the topics to further calibrate the words in a more spread way instead of days.

Predicting sentiments

Using a logistic model on the variables used in the study, we evaluated the determinants of the nature of sentiments from the tweets. The data presented in Table 1 were considered, with only positive and negative sentiment tweets by polarity and neutral tweets being excluded. The R language package *syuzhet* with the *bing* lexicon was used to derive sentiments from tweets and attaching the polarity of each tweet. The average retweet count was 20. Table 7 shows summaries of attributes use in the model.

Hypotheses testing

We sought to test the effect of parameters on sentiment polarity from the contingency tables. The Pearson's chi-square test and analysis of variance were, respectively, used to determine whether there was a significant difference between the outcome variable (sentiment polarity) and the independent variable in one or more categories.

Table 6 Topics modelled from the tweets with LDA for the different media houses

Media	Topic 1	Topic 2	Topic 3	Topic 4
NTV	Police and security	Politics	Banking	Others issues
Daily monitor	Police and people	Politics	Banking	Others issues
NBS TV	People	Politics	Banking	Others issues
New vision	Apologetics	Other issues	Politics	Banking

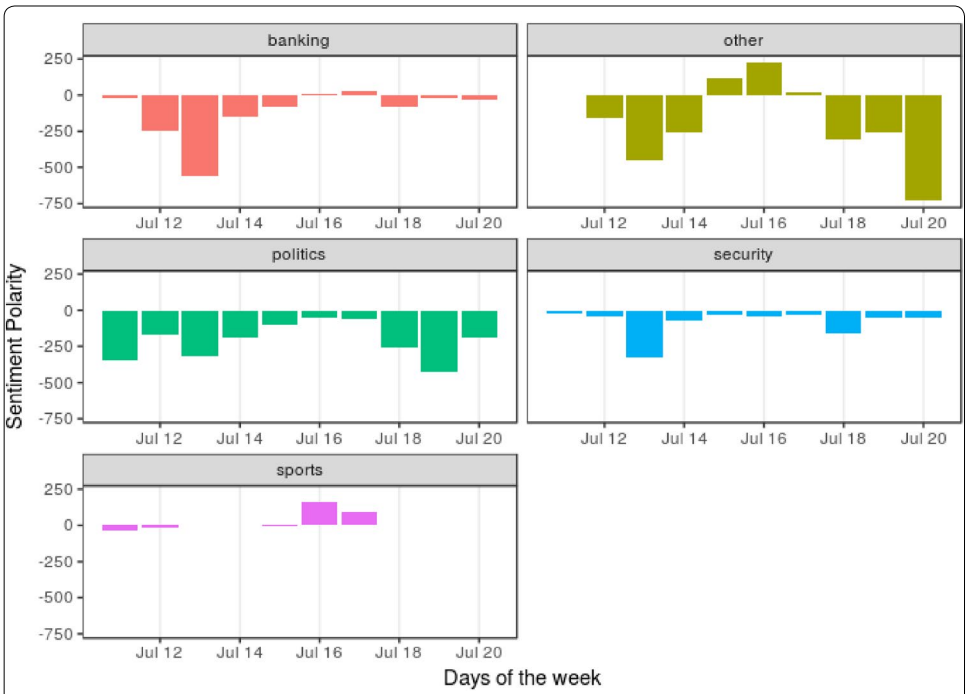

Fig. 7 Topic modelling and sentiment comparison over time. The bars represent cumulative sentiment polarity per day

The results in Table 8 show that there are significant empirical relationships between the outcome variable (sentiment polarity) and the independent variables since the p values are < 0.05 at 95% level of significance. The table presents the degrees of freedom, chi-square (χ^2) value and the p value. Basing on these results, all the variables were included in the model for predicting sentiments since they all independently contribute to determining the sentiment polarity.

Tweet sentiment statistical model

To assess the determinants of sentiment polarity derived from a tweet, a logistic regression model was fitted. Results of the model show that most of the variables according to Table 9 were significant ($p < 0.05$) within the model. In the same table, we presented the predictors, model coefficients, p values and odds ratio.

Table 7 Summary statistics of attributes to be used for logistic modelling

is_retweet		Media		Topic		Sentiment		Source	
FALSE	4610	Monitor	2429	Banking	1366	Negative	7435	Android	8335
TRUE	8434	Nbs	4505	Other	7970	Positive	5609	Iphone	1285
		Ntv	3298	Politics	2744			Mobileweb	532
		Vision	2813	Security	731			Other	1072
				Sports	233			Web	1820

Table 8 Bivariate analysis: comparing the relationship significance between each independent variable with the outcome variable—sentiment polarity

Variable	Test	Degrees of freedom	χ^2 value	*p* value
is_retweet	χ^2test	1	310.38	< 0.05
retweet_count	Anova	1	–	< 0.05
Topic	χ^2 test	4	944.14	< 0.05
Source	χ^2 test	4	19.749	< 0.05
Media	χ^2 test	3	23.347	< 0.05

Interpretation of the model

The results show that retweets compared to new tweets are more likely to have a negative sentiment with the odds ratio of 0.889 as seen from Table 9, implying that negative sentiments attract a lot of public attention resulting into prolonged discussions. Hence for any retweet message, chances are higher for it to be negative than positive. The fitted model shows that with retweet_count is a weak determinant of sentiment polarity given that the OR ~ 1; however, p value < 0.05. With the significant *p value*, it implies that tweets that have a higher retweet count are more likely to be negative.

For the identified topics, compared with the topic on economics/business the odds ratio varies significantly. Tweets on security are generally predictors of negative sentiments, whereas other topics are general predictors of positive sentiments, implying that the odds of belonging to a positive sentiment are higher than for the topic on economics and very significant for the topic on *sports*. It is expected that with higher certainty, the tweets about sports will be positive compared to all the other topics identified.

Regarding the source of the tweets, the model shows that the three sources, i.e iphone, mobileweb and web, were not significant determinants of sentiment polarity in new tweets since $p > 0.05$. However, for *"other"* tweet sources there is a higher likelihood for tweeting topics with positive sentiments (OR = 1.246) as compared to Android tweet sources which have a higher likelihood for tweeting negative tweets.

Lastly for the media categories, compared to the *Daily Monitor newspaper* media house where tweets through this media house are more likely to have negative sentiments, the OR > 1 for *NBS, NTV, New vision paper*. This implies that tweets arising through other media houses have higher chances of tweeting messages that have positive sentiments as compared to *Daily Monitor newspaper* media house.

Table 9 Coefficient of determination values and odds ratios using the different independent variables from the Twitter data

Variables	Estimate	Pr (>| *z* |)	OR (odds ratio)
(Intercept)	− 0.669	0.000	0.512
is_retweet:TRUE	− 0.118	0.016	0.889
retweet_count	− 0.009	0.000	0.991
topic:other	0.839	0.000	2.313
topic:politics	0.001	0.989	1.001
topic:security	− 0.628	0.000	0.534
topic:sports	2.189	0.000	8.928
media:nbs	0.141	0.024	1.152
media:ntv	0.179	0.008	1.196
media:vision	0.148	0.031	1.159
source_tweets:iphone	0.054	0.464	1.056
source_tweets:mobileweb	− 0.010	0.927	0.990
source_tweet:sother	0.220	0.006	1.246
source_tweets:web	− 0.019	0.767	0.981

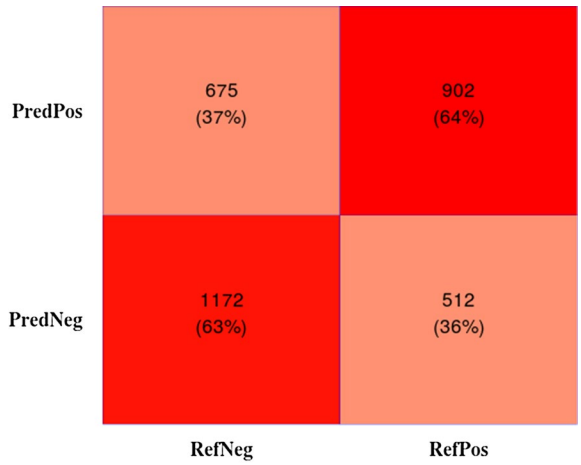

Fig. 8 Confusion matrix for the predictive results showing the predicted (Pred) values against the reference tweets (Ref)

Model validation and classification

The logistic regression model from the previous section was trained on a sample of 9783 observations and validated it on a test set of 3261 observations. Results from the model validation process show that 64% of the positive tweets and 63% of the negative tweets were correctly classified by the logistic model.

Overall, accuracy level of the model was 64% as computed from the confusion matrix given by plot Fig. 8 also known as the contingency table. The confusion matrix was formed from the four outcomes produced as a result of binary classification from the model prediction results. The classification indicates the true positives, false positives, true negatives as well as the false negatives. Results imply that the model can be used to predict the sentiment polarity of new tweets with an accuracy level of 64%.

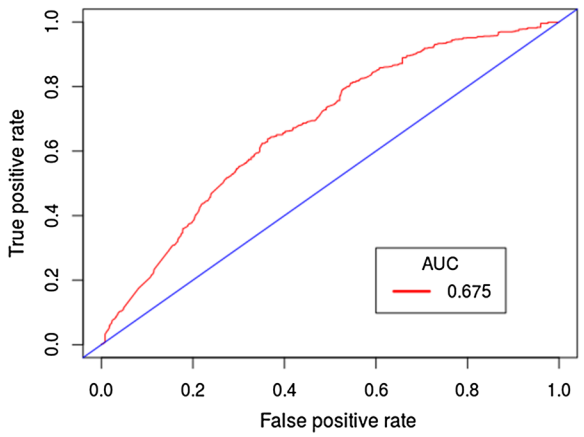

Fig. 9 The receiver operating characteristic (ROC) curve. The Area Under Curve is 0.675

To conclude the model validation process, we constructed the receiver operating characteristic (ROC) curve. This helps to see how well the logistic model classifies the positive and negative sentiments in tweets. On the y-axis is the true positive rate and on the x-axis is the false positive rate. Figure 9 shows the ROC curve together with a 45° diagonal line which is the threshold. The closer the curve is towards to the left, the higher the accuracy of the model and the closer the model is towards the diagonal, which implies that the model is weak. According to the plot in Fig. 9, the area under the curve (AUC) is a representation of the model accuracy. The best model has AUC $= 1$ and if the model is close to the diagonal, it implies that the AUC $\simeq 0.5$ which renders the model useless. From this ROC analysis, the AUC > 0.5 which renders our model useful.

For comparison purposes, we used other algorithms for classifying the sentiments. The algorithms are Random Forest, Naïve Bayes and decision tree method. The Naïve Bayes is a probabilistic method based on Baye's theorem with the assumption of independence among the attributes. Decision tree classification takes on the tree structure comprising the decision nodes (split point) and leaf nodes representing the classification. On the other hand, Random Forest a more robust method works by constructing many decision trees while taking several samples of the data and constructing a model for each data sample. The best model is chosen by averaging through the constructed models from the multiple samples. The results obtained by other methods compare competitively with the logistic regression classification. The Random Forest method shows supremacy among other algorithms, whereas Naïve Bayes method is out performed by all the other methods (Table 10).

Despite the strength of all these classification models, the balanced accuracy is still low. The models are characterised by high specificity and low sensitivity. This implies that the Twitter data possess challenges while performing classification algorithms. Ahmad e al. [30] did a comparative analysis on two different datasets using the support vector machine (SVM). They, however, discovered that their results clearly showed the dependency of SVM performance upon input datasets. Similar to this study, all methods have shown similar results on the same dataset.

Table 10 Comparative analysis of the logistic regression model performance with Decision trees, Naïve Bayes and Random Forest classification methods

Method	Reference			
	Correct rate (%)	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)
Random Forest	71	66	77	72
Decision Tree	65	62	68	65
Logistic	64	63	64	64
Naïve Bayes	60	51	70	61

Conclusion

Results from this study showed that Tweets are more prevalent under non-print media, like TVs than print media. The main print (newspaper) media, *Daily Monitor* and *New vision,* were positively correlated in generating topics with similar sentiments over time from the tweets, so were the non-print (television) media, *NBS TV* and *NTV,* for the same period. The sentiments for topics on security, politics and economics were generally negative, while sentiments on sports were positive. Hence, there is limited addition in knowledge by one following up on both the newspaper or Television media types. It is therefore advisable that for one to get a rich context of the social media content, they should follow one media house per Newspaper (print media) and Television media (non-print media).

The negative sentiments in Twitter messaging were found to be determined by mainly the retweet status, retweet count and source of the tweet (iphone, mobile and web), whereas the positive sentiments were determined by topic of discussion, type of media house and other sources of tweets. Hence, these can be used to predict the sentiment polarity of a new Tweet.

The study has also shown that tweets having a negative polarity are highly anticipated within the public. This implies that information spreads faster if it carries negative sentiments compared to positive sentiments. Therefore, it is recommended that such tweets should be monitored with precautionary measures by the different sectors including health, business, education and security. Generally, monitoring the social media topics to ensure that negative sentiments are quickly addressed before they get to diffuse within the public may be paramount to deter any potential risks.

This study focused on modelling the sentiments and topics under twitter platform for a given period from the traditional media houses in Uganda. However, this work can be extended to include more data on the prevailing social and economic characteristics. Further study could be done to improve the model classification results. Adding more attributes in the model could improve its performance making it a vital source of new and novel ideas for timely decision making and help in collecting appropriate feedback on new policies, products, politics, business and many other fields within the social media platforms.

Authors' contributions

FN developed the concept, and FN and RW drafted the manuscript. JP, FN and RW reviewed the final manuscript. All authors read and approved the final manuscript.

Author details
[1] School of Statistics and Planning, Makerere University, P.O. Box 7062, Kampala, Uganda. [2] Department of Statistics, Sultan Qaboos University, P.O. Box 36, Al-Khod 123, Muscat, Sultanate of Oman.

Acknowledgements
We are thankful to the School of Statistics and Planning, Makerere University for the cordial working environment that enabled the study to be carried out successfully. Frank is thankful to Asabea Lawrencia Osei and Charlotte Olivia Namagembe for research assistance rendered. The authors also thank the anonymous editors who reviewed and advised on the changes to be made which made this article better.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The dataset supporting the conclusions of this article can be availed on request.

Consent for publication
All the authors have consented to the publication of this manuscript.

Ethics approval and consent to participate
The analysed tweets were collected through the public Twitter API and are subject to the Twitter terms and conditions.

Funding
No funding was received to conduct this study.

Publisher's Note

Received: 18 February 2018 Accepted: 9 March 2019
Published online: 10 April 2019

References

1. Kanimozhi KV, Venkatesan M. Unstructured data analysis-a survey. Int J Adv Res Comput Commun Eng. 2015;4:223–5.
2. Grar M, Cherepnalkoski D, Mozeti I, Kralj Novak P. Stance and influence of twitter users regarding the brexit referendum. Comput Social Netw. 2017;4:6.
3. Fan W, Wallace L, Rich S, Zhang Z. Tapping the power of text mining. Commun ACM. 2006;49(9):76–82.
4. Kaplan Andreas M, Haenlein M. Users of the world, unite! the challenges and opportunities of social media. Business Horiz. 2010;53(1):59–68.
5. McNab C. What social media offers to health professionals and citizens. Bull World Health Org. 2009;87(8):566.
6. Dwivedi Mridula, Yadav Anil, Venkatesh Umashankar. Use of social media by national tourism organizations: a preliminary analysis. Inf Technol Tour. 2011;13(2):93–103.
7. Greenwood S, Perrin A, Duggan M. Social media update 2016. Pew Res Center. 2016;11:2.
8. Pang B, Lee L, et al. Opinion mining and sentiment analysis. Found Trend Inf Retriev. 2008;2(1–2):1–135.
9. Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining. ACM; 2008, p. 231–40.
10. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Linguist. 2011;37(2):267–307.
11. Bhuta S, Doshi A, Doshi U, Narvekar M. A review of techniques for sentiment analysis of twitter data. In: 2014 International Conference on issues and challenges in intelligent computing techniques (ICICT), IEEE; 2014. p. 583–91.
12. Balage F, Pardo T. Nilc_usp: a hybrid system for sentiment analysis in twitter messages. In SemEval@ NAACL-HLT. 2013, p. 568–72.
13. Lau KN, Lee KH, Ho Y. Text mining for the hotel industry. Cornell Hotel Restaur Adm Q. 2005;46(3):344–62.
14. Clark J. Text mining and scholarly publishing. In: Publishing Research Consortium, 2013.
15. Gémar G, Jiménez-Quintero JA. Text mining social media for competitive analysis. Tour Manag Stud. 2015;11(1):84–90.
16. Xiang Z, Qianzhou D, Ma Y, Fan W. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tour Manag. 2017;58:51–65.
17. Demetris A, Constantine D. Co-evolutionary dynamics in social networks: a case study of twitter. Comput Soc Netw. 2015;2:14.
18. Asur S, Huberman BA. Predicting the future with social media. In: 2010 IEEE/WIC/ACM International conference on web intelligence and intelligent agent technology (WI-IAT), IEEE. 2010, p. 492–99.
19. Zhao Y. Analysing twitter data with text mining and social network analysis. In Proceedings of the 11th Australasian data mining and analytics conference (AusDM 2013), 2013.
20. Hong L, Davison BD. Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. ACM; 2010, p. 80–8.
21. Yang Y, Wang X. World cup 2014 in the twitter world: a big data analysis of sentiments in us sports fans' tweets. Comput Hum Behav. 2015;48:392–400.
22. O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. From tweets to polls: linking text sentiment to public opinion time series. ICWSM. 2010;11(122–129):1–2.

23. Enz CA, Verma R. Introduction to the Cornell hospitality research summit special issue: the new science of service innovation in a multipartner world, 2016.
24. Kotu V, Deshpande B. Predictive analytics and data mining: concepts and practice with rapidminer. New York: Morgan Kaufmann; 2014.
25. Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. I J ACSA. 2015;6(1):147–53.
26. Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
27. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. In: Advances in neural information processing systems. 2002, p. 601–8.
28. Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in twitter. 2010.
29. Liu B. Sentiment analysis and human language technologies. Synth Lect Hum Lang Techn. 2012;51(1):1–167.
30. Ahmad M, Aftab S, Ali I. Sentiment analysis of tweets using svm. Int J Comput Appl. 2017;177(5):25–9.