EasyChair Preprint

№ 1711

Sentiment Analysis Application and Natural Language Processing for Mobile Network Operators' Support on Social Media

Kingsley Ogudo and Dahj Muwawa Jean Nestor

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 18, 2019

Sentiment Analysis Application and Natural Language Processing for Mobile Network Operators' Support on Social Media

***Abstract*—Social Media have become a mixed platform of emotional expressions on services and products reviews. While Network Operators focus on Quality of Service and customer experience, mostly built upon complaints and performance indicators, subscribers and/or followers are mostly expressing their emotions on twitter and other social media. On one side, understanding followers' sentiment and perception on offered and applied services can help the South African Mobile Network Operators to anticipate network and customer problems, embracing proactive measures rather than reactive ones on improving service and network quality. On the other side, the rise of text mining, sentiment analysis and the global Natural Language Processing expands the needs of Data Analysis across textual statistics. In this paper, we leverage on Natural Language Processing (NLP), using sentiment analysis and text mining to analyze Mobile Network Operators, in this case CellC followers' using the R platform. We use the polarity model of sentiment Analysis to determine the level of potential detraction and promotion across South African Mobile Network Operator (MNO), based on public tweets. The systematic analysis, as approached in this paper, creates a bi-directional link between customers or followers and the MNO, extracting from sets of unstructured information, relevant signification.**

Keywords—Social Media, Sentiment Analysis, Text Mining, Mobile Network Operators, Twitter, Proactive Measures, Data Analysis, Network Quality, Natural Language Processing (NLP) .

I. INTRODUCTION

Mobile Network Operators have the duty to ensure optimal network performance, high rate of service availability and uninterrupted network function, in order to attract more customers and protect revenue. As a deduction, the justification of high investment on Quality of Services (QoS), Customer Experience (CEM) and Service Quality (SQM) tools [1]. The mentioned tools rely more on customer complaints and identified network outages. However, with the rise of Social media, customers and other followers have the ability to express their happiness or discontent in front of the entire world, which raises spontaneous interactions [2] of followers -customers- between them and with the Mobile Network Operators. In this paper, the topic of quality goes behind the business model of South African MNOs, by exploring the public opinions on every topic on the operators. Twitter is an interactive micro-blogging social media platform which contains considerable and rich users' information. With the availability of twitter REST API (Access Programming Interface), it makes it not easier but possible to extract several tweets on specific people or businesses. South African Mobile Network Operators are very active on twitter and constantly interact with followers on different topics, including customer support queries. In this paper, tweets are collected from the most popular South African MNOs, analyzed and visualized to extract from different posts, the sentiment of the users. With the objective to determine attitudes of the posters or followers, sentiment analysis is an integral part of text mining and Natural Language Processing (NLP). The determination of polarity in the texts or posts categorizes the feeling as either positive, negative or even neutral [3]; nevertheless, the veracity of analyzing opinions and attitudes hidden inside chunk of texts depends on the admission of human perception and judgment, underlined by model performance indicators such as recall and precision [4]. Mobile Network Operators provides mobile cellular services across the country, both voice and data services. Their revenue is mostly determined by the subscribers' usage of network resources (calls, internet, SMS etc.); which means the best the quality of the network and services offered, the likely for the MNO to have more subscribers. A country hosting multiple MNOs is driven by services and quality competition, which sees subscribers churning or moving from one operator to another, impacting MNO revenue [5]. MNOs have long relied on in-house platforms (CEM, trouble-tickets) to apprehend customer problems. This can be defined by the number of customer tickets opened or the number frustrated customer care calls. Based on MNO business and operation processes, solving a customer complaint is not as spontaneous as acknowledging it. Hence, reactive rather than pro-active network improvement methodologies are adopted. With the rise of the internet era and the affordability of internet data bundles, the delay in processing and fixing network and service problems could result in social media reaction on the MNO timelines. Be that as it may, grasping the "what people think about "us" or about our service", the "what are people talking about on out timeline?" and or "what is the sentiment expressed by our followers?" provides a competitive edge on analyzing followers' behaviors. The analysis supports MNOs to identify and classify followers as promoters or detractors. How likely it is for our followers to promote or detract our services to others, locate the areas of the countries in which most of the negative tweets come from, allowing the MNOs to adopt a more pro-active methodology, then reactive. Data is also visualized in a simplistic form to understand and have an effective data insight. Different visualization methods including wordcloud, barplots, line trends, area plots are all used to facilitate the understanding of information. In this research we follow both a lexicon approach to study text polarity and a machine learning approach using the Latent Dirichlet Allocation (LDA) to evaluate different topics on the timelines.

II. BACKGROUND AND RELATED WORK

Mining textual data to extract sentiment has been a challenging aspect of Natural Language Processing (NLP) due to the unstructured form of text and the discovery of information patterns in the text; most machine learning algorithms have been applied to numerical and categorical variable on structured data frames. Thus, the full dependence on NLP and complex techniques and algorithms to manipulate textual data, discover useful pattern of information. These techniques deviate marginally from legacy languages perceived by computers and machines [6]. Many researches have been conduction and practically applied to real world problems in the past 2 decades. Rincy Jose and Varghese S Chooralil [7] used NLP and sentiment analysis to predict election result on twitter data using classification approach. The study implements sentiment analysis on election prediction using four algorithms, SentiWordnet, Naïve Bayes, HMM and ensemble approach. The implementation of the study uses ensemble approach on two candidates using three weeks tweets data and predicted political sentiment of followers. The proposition of a framework for sentiment analysis has been proposed also by K. Zvarevashe and O.O. Olugbara by analyzing unlabeled Hotel review data from public datasets [8]. Two models are used for their research, an intuitive model in which data is labelled manually by the user agent upon perception, transformed to a standard processed format and a classification algorithm is applied on the processed data and the sentiment polarity model in which label is assigned by using sentiment polarity positive or negative score; data is then transformed and prediction algorithms are applied on the processed data. K Zvarevashe and O.O. Olugbara have relied on survey reviews to build the sentiment model; our research focuses on the application of NLP in the Telecommunications arena. In the objective to also analyze sentiments on microblog texts, Jie Li and Lirong Qu introduces a dependency parsing model which considers the relationship migration of sentiment and adjusted distance of sentiment in microblog text; the study structures the sentiment and accentuate on the correlation between emotion and modifier [9]. Equivalently, a dependency parsing model was introduced by Feng Shi et al. to analyze text in microblog, deriving emotion word and relationship dependency in text to compute the sentiment in each sentence [10]. Eissa M. Alshari et al. [11] uses the SentiWordNet [12], an important lexical appliance for sentiment analysis in order to enrich the Word2vec lexical dictionary. The study expands the intersection of the SentiWordNet and Corpus library, assigning polarities to non-opinion words on one by learning the other. The work of M.F.A. Bashri and K. Retno [13] develops the concept of sentiment analysis by using the Latent Dirichlet Allocation (LDA) and Wordcloud eye of the mind, to analyze and assign topic polarity on text. The study is based on student reviews and comments on educational institution.

LDA is a probabilistic model in which any text is overseen as an arbitrary combination of latent topics, each topic is then modeled as a set of probabilities and distributed on vocabularies [14]. The model is generative with the following process to generate a document, as defined by D.M. Blei [14].

- Choose N as distributed words in a text or document.
- Choose a random LDA variable θ
- For N in document w,
    - Select a topic t_n randomly
    - Select a word w_n randomly from the corresponding topic t_n.

The model illustrates the concept that a text or document displays multiple topics. The generative process of the LDA is then mathematically expressed by the equation:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d)\, p(w_{d,n}|\beta_{1:K}, z_{d,n})$$

In which $\beta_{1:K}$ are the topics, θ_d is the proportion of topic for the d^{th} document or text which means $\theta_{d,k}$ is the proportion of topic k in document d; z_d is the topic assignment for the d^{th} document; and w_d is the observed word in the d^{th} document. The LDA equations for computing the probability and proportion of topic is provided in this research using the Gibbs sampling algorithm [15], computing the posterior of the model:

$$\beta_{ij} = \frac{C_{ij}^{WT} + \alpha}{\sum_{k=1}^{W} C_{kj}^{WT} + W_\alpha}, \quad \theta_{dj} = \frac{C_{dj}^{DT} + \varphi}{\sum_{k=1}^{T} C_{dk}^{DT} + W_\varphi}$$

Where C^{WT} and C^{DT} are counts matrices which respectively relates to the count of Word-Topic and Document-Topic. The Gibbs sampling is a typical illustration of a Markov Chain Monte Carlo approach in which the sampling of a variable is done at a time, while the current value of the variable remains unchanged.

Similar concept and approach are applied to the twitter data on MNOs timelines, where tweets exhibit multiple topics and the distribution of topic structure is computed given all the collected tweets. The evaluation of our model is based on the standard machine learning performance parameters such as the precision and recall [5] and the area under the curve based on sampled data.

An important aspect of NLP and sentiment analysis is the efficient understanding of lexical words, their prefixes and roots words, which is referred to as stemming. The sentiment analysis in this research is based on the Porter's stemming algorithm, which for all the words in the text, suffixes are stripped [16]. Effective, simple and implemented in several NLP applications, the algorithm follows a set of rules and conditions to reduce and deduct input words to its root. As described by A. Singh et al. [17], and in the original article by M.F. Porter [16], the application of rules relies on analyzing the suffix of a word and alternating that specific suffix should some conditions be fulfilled. A summary of the algorithm is shown in Fig. 1.

Vowel are A, E, I, O, U and consonants are the rest of the alphabet letters out of the mentioned vowel. The condition is given by:

$$\text{(condition) } S1 \rightarrow S2$$

$S2$ replaces $S1$ if the stem of the later (suffix of a word) fulfils a specific condition. Variable m is the number of times a vowel-consonant combination repeats itself.

Steps 1 to 3 apply when m>0.

Step 1a.

SSES > SS	caresses > caress
IES > I	pies > pi
S >	dogs > dog
…	

Step 1b.

EED > EE	agreed > agree
ING >	motoring > motor
…	

Step 1c.

Y >	Happy > happi
	Sky > sky

Step 2.

ATIONAL > ATE	relational > relate
TIONAL > TION	conditional > condition
ALISM > AL	feudalism > feudal
…	

Step 3.

ICATE > IC	triplicate > triplic
ATIVE >	formative > form
ICAL > IC	electrical > electric
…	

Step 4 and step 5 apply when m > 1:

Step 4.

AL >	revival > reviv
ANCE >	allowance > allow
ENCE >	inference > infer
…	

Step 5a.

E >	probate > probat
…	

Step 5b.

(m >1 and *d and *L) > single letter	
	Controll > control
	Roll > roll

Fig. 1 Porter's Algorithm stemming rules summary

Hierarchical Clustering Algorithm Concept:

The hierarchical clustering method groups predictors of parameters in to groups in which the structure is a tree hierarchy. Clusters are represented by nodes, referred to as dendrograms [18]. The objective in hierarchical clustering is to assess the similarity between two classified parameters. Parameters with higher similarity are placed in the same cluster or group. The similarity or dissimilarity is measured by the distance between elements. The distance can be calculated using the Minkowski distance:

$$L_p(X_a, X_b) = \left(\sum_{i=1}^{n} \left| X_{i,a} - X_{i,b} \right|^p \right)^{1/p} \quad ; \forall p \geq 1, p \in \mathbb{Z}^+$$

If $p = 2$, which is finding the nearest neighbor between 2 elements, the distance is given by the Euclidian formula:

$$L_2(X_a, X_b) = \sqrt{\left(X_{1,a} - X_{1,b}\right)^2 + \left(X_{2,a} - X_{2,b}\right)^2}$$

The elements with the shortest distance are merged to make a cluster. As proposed by Z. Nazari et al. [19], the mean μ is calculated for each built cluster, distance from μ to the cluster elements is computed, and using the maximum distance $L(max)$, distance between different clusters is computed to merge or associate clusters.

Natural Language Processing using the word2vector Model and glmnet algorithm:

Introduced by Mikolov et al. [20] , word2vector has become popular in Natural Language Processing and is built on Neural Network [21]. The model uses continuous bag-of-words and skip-gram methodologies to train the textual datasets. In the first methodology, given neighboring words W_n, the model predicts the current word W_c. And in the second methodology, W_n is predicted using W_c. The workflow of the model is detailed in .

Word2vec is therefore a supervised Machine Learning technique.

Glmnet is an R package which fits linear model through maximum probability using penalty method. The algorithm answers the following equation [22]:

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha) || \frac{\beta| |_2^2}{2} + \alpha ||\beta||_1 \right]$$

Where α controls the elastic-net penalty, λ is the tuning variable, which monitor the penalty strength. [22], [23] Provides more details on the package and the mathematical background of the algorithm origin.

In this research paper, we use the word2vec model and the glmnet algorithm to predict from the continuous bag of words methodology, the binary outcome based on tweet sentiment (positive or negative) computed using the Porter algorithm.

III. Methodology

Applying text mining, sentiment analysis or any sort of NLP model relies on a well-defined process to facilitate the operations required to effectively mine the textual data. Fig. 1 shows the methodology used in this research paper. The main steps include Data Collection, Data preprocessing, Data Analysis, Topic Modeling and Prediction. However, a set of subprocesses are also involved as illustrated in Fig. 1.

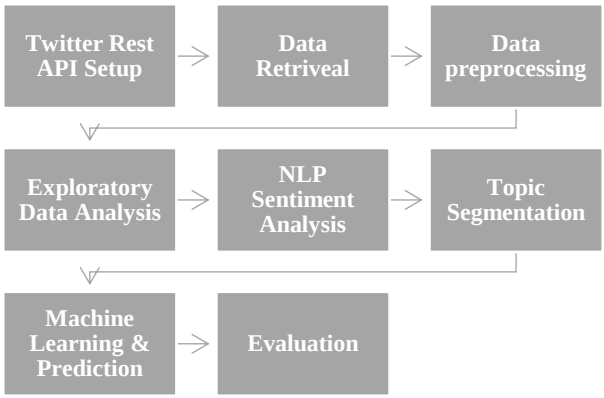

Fig. 2 Research Methodology

Twitter allows public access to its data, subject to registration and permission to retrieve, in real time or historical manner, public tweets. In order to access the twitter public data, one needs to have a twitter account and ensure the authentication process for security access is maintained [18], by using the different security keys generated by twitter, as shown on Fig. 2.

Fig. 3 Twitter Rest API Setup Process

After successful authentication, a direct connection is set to the twitter public data where we can retrieve the tweets. We collect the tweets coming from three of South African popular Mobile Network Operators (@CellC_Support, @Vodacom111 and @MTNzaService). The MNO with the highest number of tweets will be used for further Analysis.

In order to format the collected textual data, a preprocessing method is used to clean and remove unnecessary information, stop words (such as "and", "the", "or" etc.) and URLs. Tokenization is used to unnest the words on the tweets. Preprocessing converts the data to a format suitable for Machine learning and deep analysis. An EDA (Exploratory Data Analysis) is then executed to find important patterns of information and correlation between different parameters in the tweets; for example, identifying the number of followers on each timeline and accessing a possible correlation between tweets and twitters (here also referred to as followers). We also attach in this step geographical information to the tweets to segment tweets based on their geo-location, a great indication to the MNOs of regions twitting a lot and about what. Sentiment Analysis is then applied to the processed tweets to classify sentiments using different algorithms. Topic segmentation step allows the discovery most discussed topics for each MNO using machine learning algorithm, in our case the LDA, identify the mostly used words in tweets and predict or classify the polarity in users' tweets. Both predominant attitude and probabilistic topic modeling are used.

IV. Implementation and Experiment's result

After successful authentication and collecting information tweets from the MNO supports or services timelines, the below implementation and analysis is executed in parallel for all operators. Tweets extraction period for all Operators: 14/12 – 24/12/2018. The most important variables used for analysis are shown below, after converting the retrieved tweets to dataframes. Structured Twitter data contains 16 variables, but not all of them are used in our research.

'data.frame': **var** obs. of 16 variables:
$ text : the tweets and retweets of users
$ favorited : Favored tweets or liked
$ favoriteCount: Number of times a tweet has been liked
$ created : Date of tweet creation
$ id : Tweets ID
$ statusSource : http/http links to tweets
$ screenName : twitters or followers who have posted
$ retweetCount : count of Retweeted posts
$ isRetweet : Binary value to indicate if a post is a retweet
$ retweeted : Binary value to indicate if a post has been retweeted or not
$ longitude : Geo-code Longitude
$ latitude : Geo-code Latitude

Apart from the tweets collected, we also collect user information with the below capital variables:

'data.frame': # obs. of # variables:
$ description : user description
$ statusesCount : status count
$ followersCount : number of the followers' followers
$ friendsCount : number of the followers' friends
$ name : followers name
$ created : Date of creation
$ screenName : followers names
$ location : location
$ lang : language
$ id : user id

A. Exploratory Data Analysis of Tweets

In this EDA section, more insight on the tweeter data is given. Collected data is stored as dataframes and saved as R Objects for future analysis. Fig. 4 displays the posted and collected tweets during the selected time frame for each operator, which indicates the level of interaction with the support team. 640 tweets were collected from CellC, which will be used for further processing. Fig. 5 provides the normalized most persuasive, influential or popular followers for the MNO. The level of influence is computed by the number of followers each follower has on twitter, indicated by the field or variable "followersCount" of the followers' datasets.

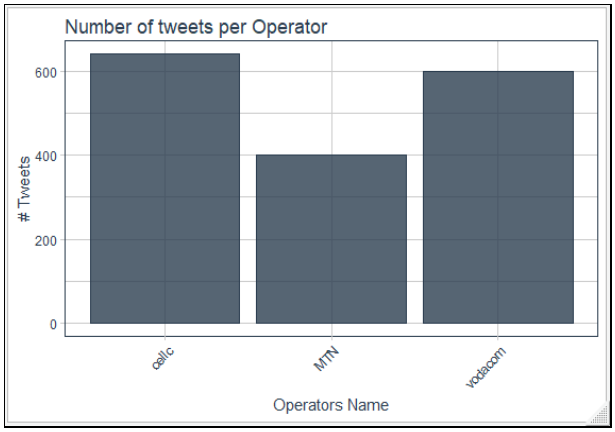

Fig. 4 Number of tweets per MNO

Followers graphs show that most of the operators have followers who have high number of second level followers, with $\log_2 F(o) > \sim 10$, which means followers with $\sim 10^3$ followers. $F(o)$ is the count of followers each follower has got.

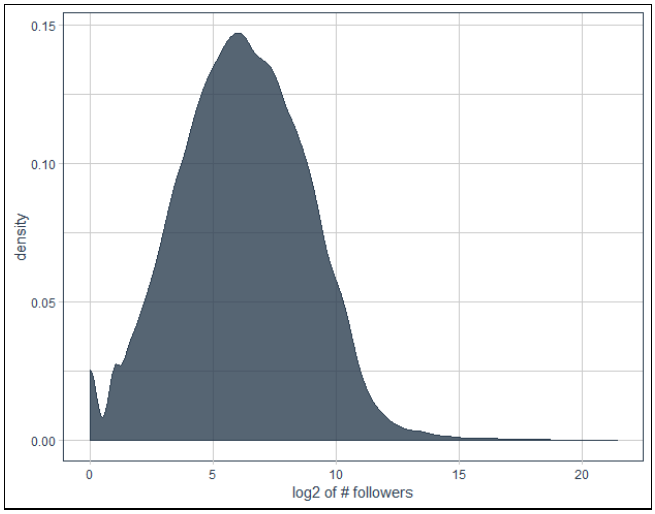

Fig. 5 Most persuasive Followers by Popularity

Fig. 6 analyzes the relationship and correlation between tweets and followers for the MNO with the highest number of tweets. Correlation between a normalized number of followers and how often they tweet is shown. The graphs show the followers vs. tweets for the 10 days during which the tweets have been collected; most of the people with larger number of followers (x-axis) tend to tweet more (y-axis). An increasing graph is observed.

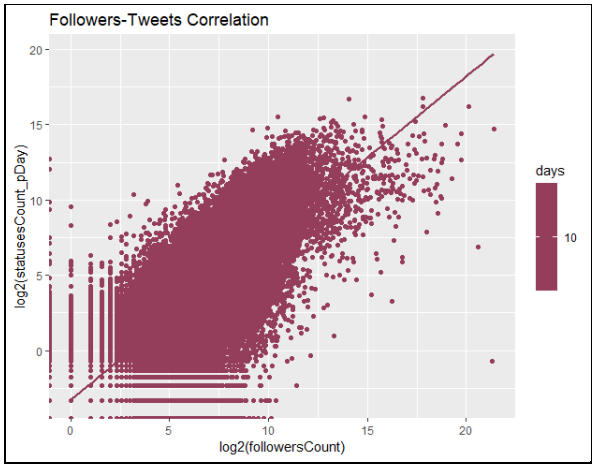

Fig. 6 Followers - tweets correlation

B. Sentiment Analysis

We use the Porter algorithm introduced in section II, summarized in Fig. 1 to apply words categorization and splitting. We highlight words which will be associated to positive and negative connotation during the analysis. The algorithm. Wordcloud is used in Fig. 7 to show the most used words in tweets. The words are stemmed using the algorithm. Words such as "monei", "pai", "updat" "servic" … are linked to "money", "pay", "update", "service" respectively which aligns with rules step 1c and 5a of the Porter algorithm.

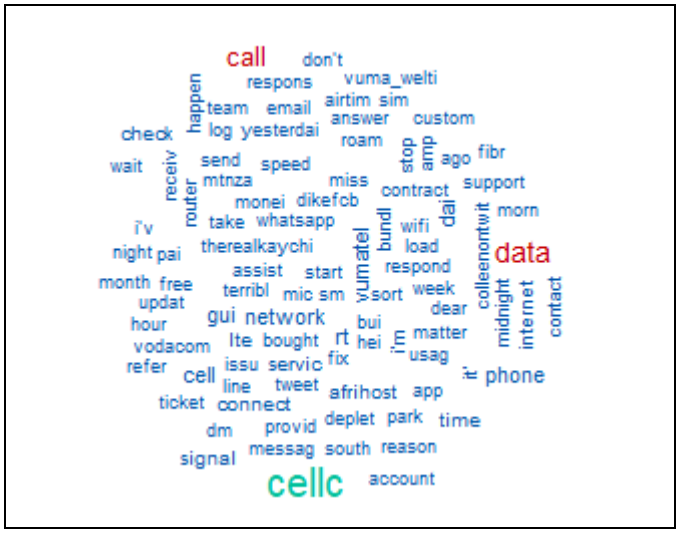

Fig. 7 Most used words in the Operator's tweets

From network perspectives, we highlight distinguished words highly used: CellC, call, data. In order to analyze the sentiment, negated words are identified, words that are preceded by "not" or "no", which are used to analyze sentiment. Words such as "not solved", "not apologized", "no support" and "not agree" can be deducted from the graph. Fig. 9 displays the computed prevailing sentiment and feeling in the followers' tweets. The dominant feeling lies towards negativity than positivity.

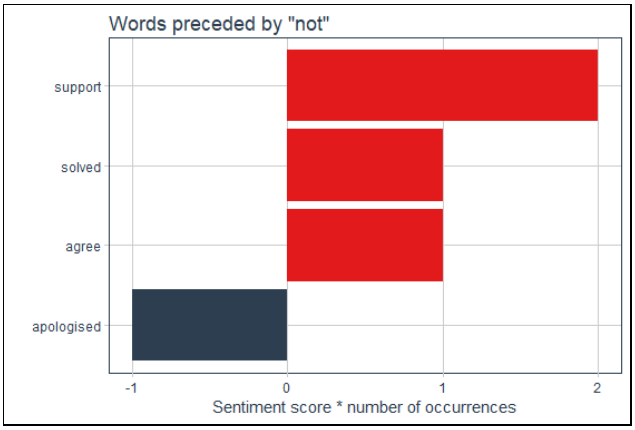

Fig. 8 Negated Words

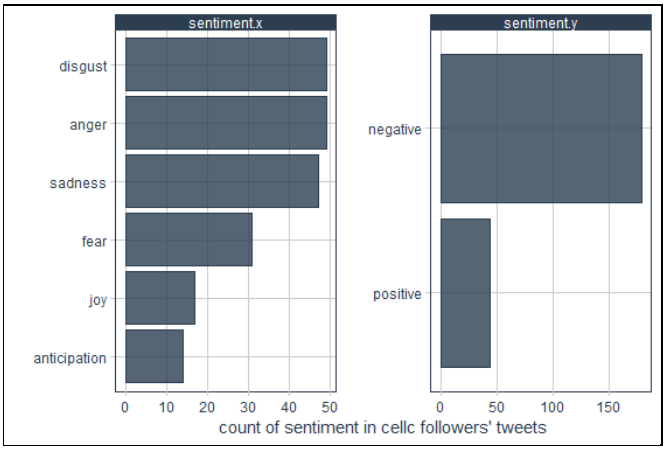

Fig. 9 Followers sentiment Analysis

Fig. 10 shows the words categorization, the positive and negative words classified by the algorithm.

Fig. 11 shows topic categorization of followers' interest. The topic has been subset to 5 types. In both categories cellc_support takes the toll.

Examination of text polarity is executed using the "quanteda" package. The whole sentiment polarity is negative looking at the average polarity and the mean polarity.

all	total.sentences	total.words	ave.polarity	sd.polarity	stan.mean.polarity
1 all	3	26535	-5.587	9.677	-0.577

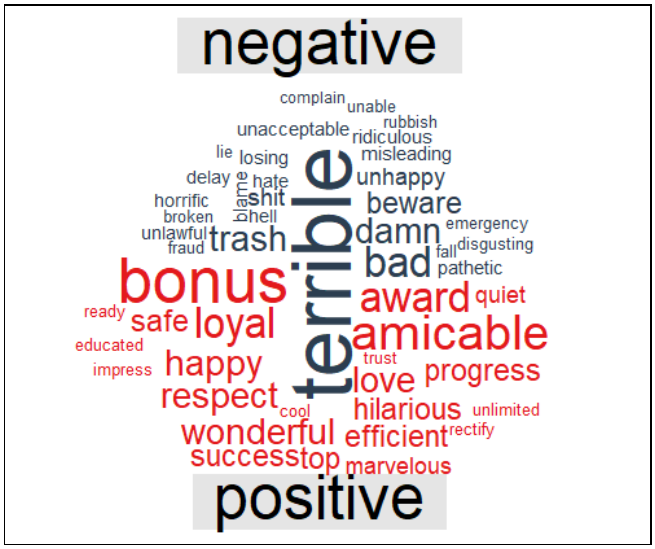

Fig. 10 Common negative and positive words

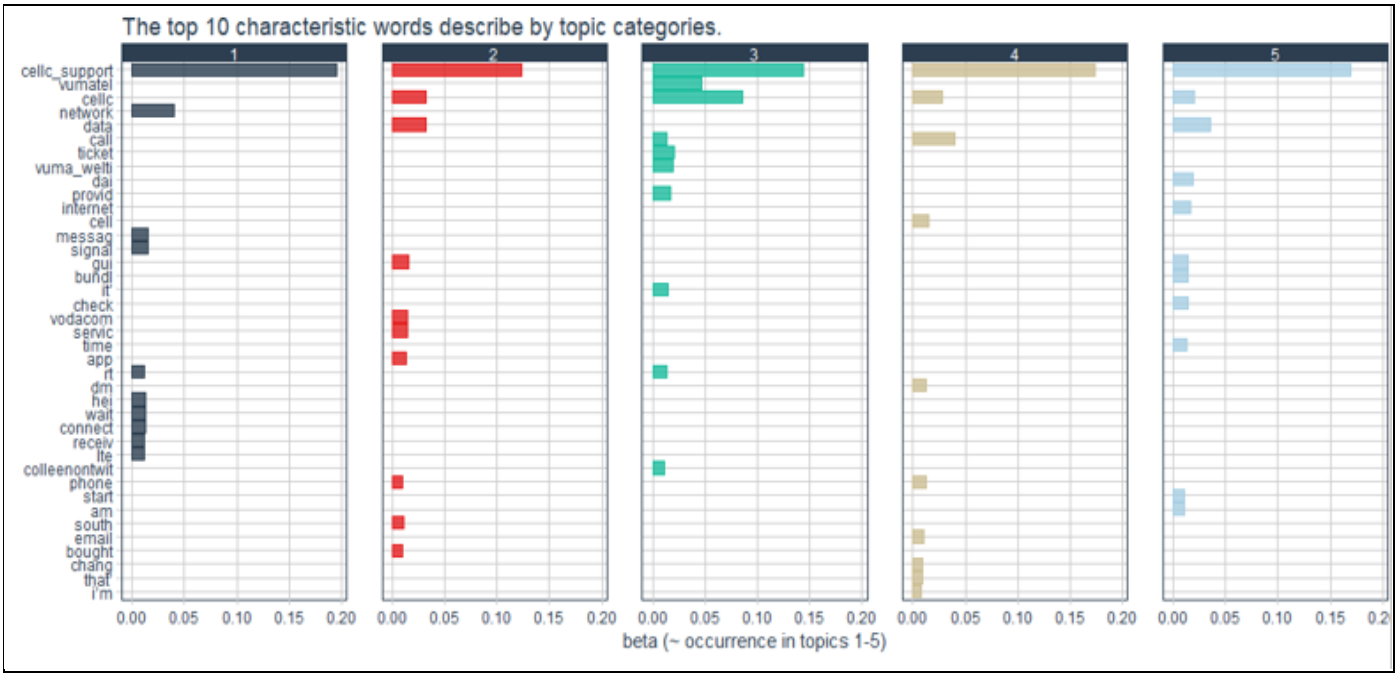

Fig. 11 Topic Modeling graph

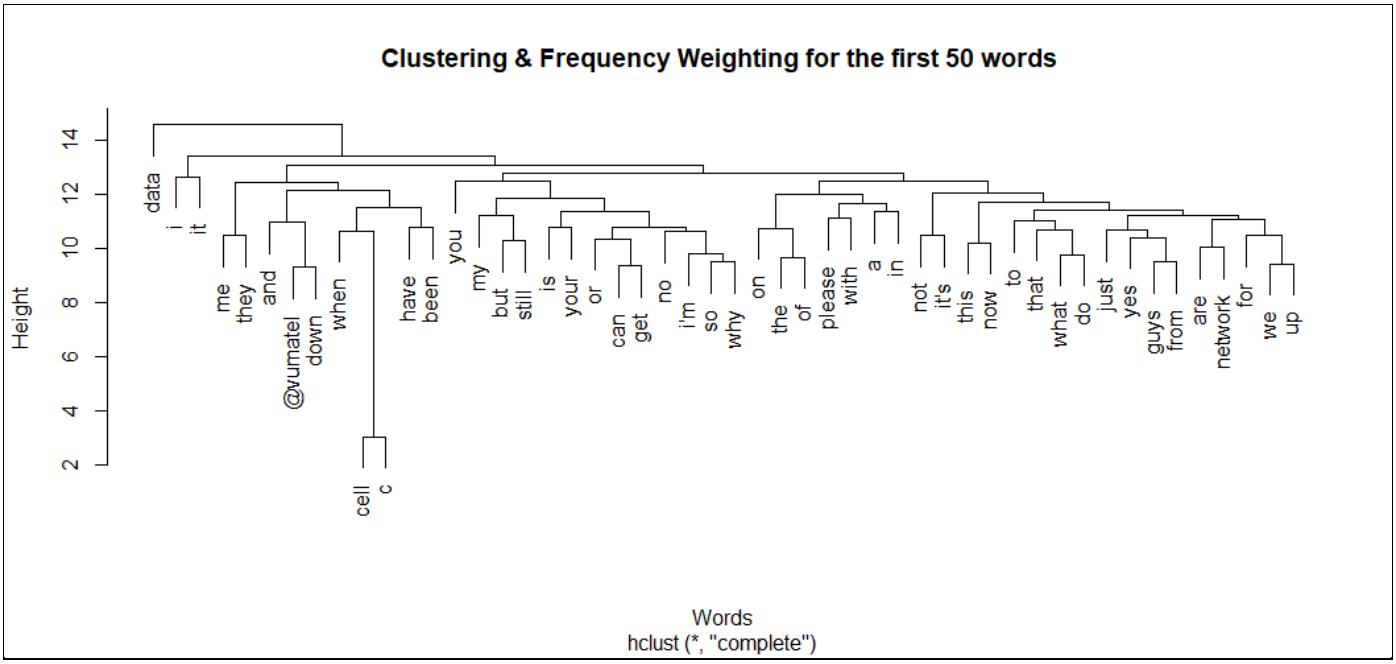

Figure 12 Hierarchical Clustering of words

The top 5 tweets with the highest emotional indexes are shown in Table 1, which describes the most positive tweets during the selected time period. The 5 most negative tweets (with the lowest emotional indexes) are shown in Table 2.

Table 1 The 5 most positive tweets

text	created	emotionalindex
@CellC_Support Just spoke to Tulani at your call centre and he was extremely efficient. Great service from him. Butâ€¦ https://t.co/BEwZ2tNhow	2018-12-21T16:10:25Z	**0.98386991**
@CellC_Support Thank you very much, I have and I'm quite happy with you guysðŸ¤œðŸ¤›ðŸ˜ŽðŸ‘Œ	2018-12-22T23:23:47Z	**0.929516**
@CellC_Support Fantastic, its indeed a good buy.	2018-12-21T20:57:40Z	**0.70710678**
@CellC_Support @CellC please can you explain how the cell C to cell C free calls work as I thought it was unlimitedâ€¦ https://t.co/GUFfn0sTXL	2018-12-22T04:11:31Z	**0.62554324**
@CellC_Support Thank you for the prompt response, I have sent a DM.	2018-12-20T14:33:31Z	**0.5547002**

Table 2 The 5 most negative tweets

text	created	emotionalvalence
@CellC_Support Signal defaults to 3G and is very very slow	2018-12-15T14:43:09Z	**-0.7839295**
@CellC_Support You guys don't seem serious about fixing this problem coz it's bothering me even now	2018-12-20T04:53:11Z	**-0.6790998**
@CellC_Support Itâ€™s the exact same problem with weak coverage. Someone called me and said they would revert back toâ€¦ https://t.co/s7tpwRtw76	2018-12-17T17:35:55Z	**-0.6546537**
@CellC_Support I'm getting tired of reversing your UNLAWFUL debits to my account.	2018-12-15T08:15:55Z	**-0.5547002**
@CellC_Support Still??? Slow	2018-12-20T17:21:22Z	**-0.5**
@CellC_Support It's unacceptable	2018-12-16T17:54:41Z	**-0.5**

C. Machine Learning Application

We apply machine learning techniques on the text data collected in order to predict and cluster tweets. Two machine learning techniques are used: Unsupervised Machine learning using hierarchical clustering and supervised machine learning

using the word2vector algorithm, based on deep learning. Fig. 12 shows the clustering of words according to their weight and distance factor, modelled using the hierarchical clustering algorithm.

Applying the supervised word2vec algorithm:
We predict using deep learning method, the sentiment in binary outcome 1 for positive and 0 for negative. Only two fields are used for the supervised prediction, the "text" and the "sentiment.y" field which classifies if a tweet's sentiment is 1 or 0. In order to implement the below analysis in R platform, the *text2vec* library needs to be loaded.

The following Steps are executed:

1. Data is partitioned to training and testing datasets: 70 % for training and 30% for testing. D_T is the partitioned data from the main dataset D, v is the predictor variable, d_{tr} is the training dataset and d_{ts} is the testing dataset.

$$D_T = createDataPartition(D\$v, p = 0.7, list = FALSE) \quad (1)$$
$$d_{tr} = D[D_T, \ ] \ \& \ d_{ts} = D[-D_T, \ ] \quad (2)$$

2. Tokenization of the dataset: let t_i be the tokenized training and testing datasets t_{tr} and t_{ts},

$$t_i = itoken(d_i\$v, preprocessor = tolower, tokenizer = word\_tokenizer) \quad (3)$$

3. Vocabulary is created using the text2vec library: let Tv be the vocabulary matrix and Tv' the vectorized Tv.

$$Tv = create\_vocabulary(t_i) \quad (4)$$
$$Tv' = voac\_vectorizer(Tv) \quad (5)$$

4. Model definition based on the term matrix: let dtm_i be the term matrix for the tokenized data set t_i and I, the defined model. We calculate the term matrix for the training and testing tokenized datasets.

$$dtm_i = create\_dtm(t_i, Tv') \quad (7)$$
$$I = TfIdf\$new() \quad (8)$$

5. Model fitting: let M_i be the fitted model and its application transformation on the training and testing document matrix dtm_i

$$M_i = fit\_transform(dtm_i, I) \quad (9)$$

6. Performance Evaluation on training dataset: we fit a generalized linear model (the GLMNET) to train and evaluate the model performance. Let C be the training fit classifier, k the chosen constant for K-fold cross validation, e & e' are constant used to determine how fast and accurate the model can train.

$$C = cv.glmnet(x = M_i, y = d_{tr}[['v']], family = 'binomial', alpha = 1, type.measure = 'auc', nfolds = k, thresh = e, maxit = e') \quad (10)$$

The Area Under the Curve (AUC) is show in Fig. 13, with the performance index, the optimal coordinates of the graph of 0.91077 (91.077%) on the training set.

7. Prediction and evaluation on testing dataset: let P be the predicted value and $M_i = M_{ts}$, the transformation on the testing set.

$$P = predict(C, M_{ts}, type = 'response')[,1] \quad (11)$$

The accuracy on the new dataset (the testing dataset) is given by the below expression:

$$glmnet :::: auc(as.numeric(d_{ts}\$v), P) \quad (12)$$

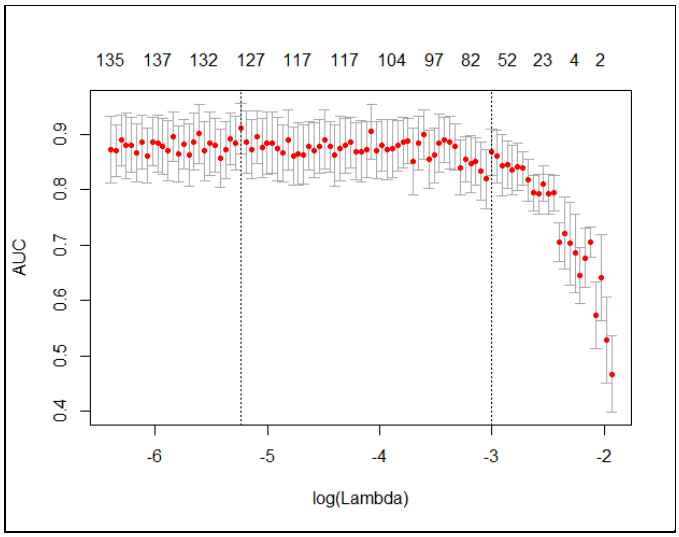

Fig. 13 AUC evaluation of the Textual Data (NLP)

```
[1] "max AUC =  0.91077"
```

```
[1] "Performance on Test Data =  0.96516"
```

The model performs well on the new dataset.

D. Word Pairs' Link Analysis

The word pairs' link analysis allows us to study the networking of word groups. Link analysis forms an integral part of graph modeling and internet analytics, identifying relationship between different vertices and edges of the dataset [19]. Fig. 14 illustrates the words pair relationship, with the edges density or color indicating the bond strength between the words. The figure shows how key word pairs fit in the entire network. The graph is obtained using igraph and ggraph libraries in R

platform. Strong relation is observed between “CellC” and “CellC_Support”, and weak relations are observed elsewhere. For example, between “sole” and “reason” which makes the pair word: sole reason …

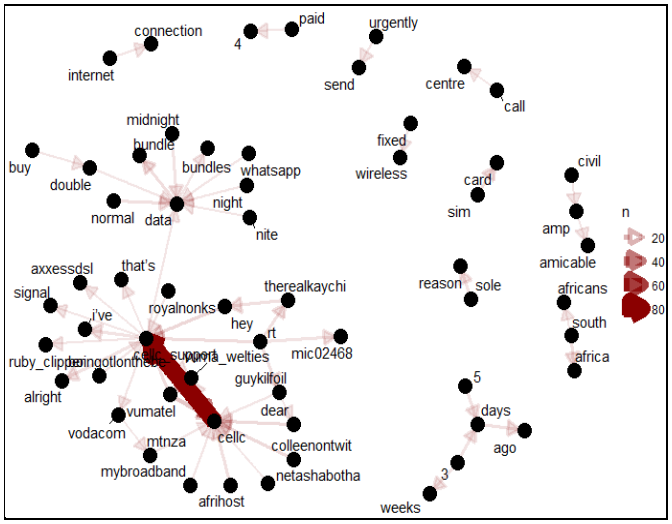

Fig. 14 Link Analysis of the word pairs on tweets

V. CONTRIBUTION OF THE STUDY

Natural Language Processing is transforming several domains in life, from complex language translation and robotics to applications such as people review analysis. Telecommunications is one of the areas which is not spared, due to the live feeds and online implementation of support services. This study benefits MNOs in adopting pro-active quality maintenance built on followers’ sentiments. When the network becomes slow for example, before congestion detection on inside tools, customers have already start complaining about slow internet and or high number of drops etc. keeping trac of social media activities provides an advantage to the MNO to stay closer to the customers.

VI. CONCLUSION

In this paper, an in-depth social media analysis has been conducted on Mobile Network Operator to determine the level of satisfaction on the services, from the customers’ standpoint. Most often, MNOs relies on Network monitoring tools to address Quality of Services. Using Natural Language Processing, we can beyond traditional QoS and QoE, by focusing on customer feelings. The study has shown how sentiment analysis and machine learning techniques can be applied to Telecommunications Operators, which impacts the Net Promoter Score. Based on the tweet’s analysis and attitude deduction, how likely it is for customers to promote or demote MNO services. The study combines various models and algorithms to derive an optimal analysis of textual tweets on MNO supports. In this analysis, we can observe that the tweet polarity is turned more towards negativity and based on the emotional index, we have determined the most friendly and unfriendly or negative tweets. Using Machine Learning techniques, we cluster and predict the sentiment of followers.

VII. REFERENCES

[1] K. A. Ogudo and D. M. J. Nestor, "Modeling of an Efficient Low Cost, Tree Based Data Service Quality Management for Mobile Operators Using in-Memory Big Data Processing and Business Intelligence use Cases," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1-8, 2018.

[2] S. K. Ravindran and V. Garg, "Fundamentas of Mining," in *Mastering Social Media Mining with R*, Birmingham - Mumbai, PACKT Publishing, 2015, pp. 1-20.

[3] N. D. Valakunde and M. S. Patwardhan, "Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process," *IEEE, International Conference on Cloud and Ubiquitous Computing and Emerging Technologies*, pp. 188-192, 2013.

[4] B. Liu, "Sentiment Analysis and Opinion Mining," *Morgan & Claypool Publishers*, pp. 5-155, 2012.

[5] D. M. J. Nestor and K. A. Ogudo, "Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1-10, 2018.

[6] S. Pattanayak, "Natural Language Processing using Recurrent Neural Network," in *Pro Deep Learning with Tensorflow, A Mathematical Approach to Advanced Artificial Intelligence in Python*, Bangalore, Karnataka, Springer Science and Business Media New York, Apress Media, 2017, pp. 223-278.

[7] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp. 64-67, 2016.

[8] K. Zvarevashe and O. O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," *2018 Conference on Information Communications Technology and Society (ICTAS)*, pp. 1-4, 2018.

[9] J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pp. 776-779, 2017.

[10] F. Shi, Y. Fu, Y. Feng and al., "Blog Sentiment Orientation Analysis Based on Dependency Parsing [J]," *Journal of Computer Research & Development*, vol. 39, no. 11A, pp. 146-148, 2012.

[11] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha and M. Alkeshr, "Effective Method for Sentiment Lexical Dictionary Enrichment Based on Word2Vec for Sentiment Analysis," *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 1-5, 2018.

[12] D. S. Tomar and P. Sharma, "A text polarity analysis using sentiwordnet based an algorithm," *(IJCSIT) International Journal of Computer Science and Information Technologies*, 2016.

[13] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization," *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, pp. 1-5, 2017.

[14] D. M. Blei, "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, pp. 77-84, 2012.

[15] R. Kusumaningrum, H. Wei, R. Manurung and A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image," *Journal of Applied Remote Sensing*, vol. 8, 2014.

[16] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-7, 1980.

[17] A. Singh, N. Kumar, S. Gera and A. Mittal, "Achieving magnitude order improvement in Porter stemmer algorithm over multi-core architecture," *2010 The 7th International Conference on Informatics and Systems (INFOS)*, pp. 1-8, 2010.

[18] Nisha and P. J. Kaur, "Cluster quality based performance evaluation of hierarchical clustering method," *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, pp. 649-653, 2015.

[19] Z. Nazari, D. Kang, M. R. Asharif, Y. Sung and S. Ogawa, "A new hierarchical clustering algorithm," *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 148-152, 2015.

[20] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv*, 2013.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *In Proceedings of International Conference on Neural Information Processing Systems*, pp. 3111-3119, 2013.

[22] J. Friedman, T. Hastie and R. Tibshirani, "Regularization Paths for Generalized Linear," *via Co ordinate Descent Journal of Statistical Software*, vol. 33, no. 1, pp. 1-22, 2010.

[23] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, "Regularization Paths for Cox's Proportional Hazards Model," *via Coordinate Descent Journal of Statistical Software*, vol. 39, no. 5, pp. 1-13, 2011.

[24] A. Bifet, G. Holmes and B. Pfahringer, "MOA-TweetReader: real-time analysis in twitter streaming data," *LNCS 6926 Springer-Verlag*, p. 4660, 2011.

[25] A. Silberstein, A. Machanavajjhala and R. Ramakrishnan, "Feed following: The big data challenge in social applications," in *in Databases and Social Networks, ser. DBSocial '11*, New York, NY: ACM, 2011, pp. 1-6.