World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:8, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:8, 2022 publications.waset.org/10012641.pdf

Topic Modeling Using Latent Dirichlet Allocation and Latent Semantic Indexing on South African Telco Twitter Data

Phumelele P. Kubheka, Pius A. Owolawi, Gbolahan Aiyetoro

Abstract**—**Twitter is one of the most popular social media platforms where users share their opinions on different subjects. Twitter can be considered a great source for mining text due to the high volumes of data generated through the platform daily. Many industries such as telecommunication companies can leverage the availability of Twitter data to better understand their markets and make an appropriate business decision. This study performs topic modeling on Twitter data using Latent Dirichlet Allocation (LDA). The obtained results are benchmarked with another topic modeling technique, Latent Semantic Indexing (LSI). The study aims to retrieve topics on a Twitter dataset containing user tweets on South African Telcos. Results from this study show that LSI is much faster than LDA. However, LDA yields better results with higher topic coherence by 8% for the best-performing model in this experiment. A higher topic coherence score indicates better performance of the model.

Keywords**—**Big data, latent Dirichlet allocation, latent semantic indexing, Telco, topic modeling, Twitter.

I. Introduction

ESTABLISHED in 2006, Twitter is one of the most popular social media platforms around the globe. Different kinds of information can be obtained from the different user interactions on social media [1]. Some of this information can be obtained from messages that are posted on the Twitter platform, namely tweets [2]. Users share their opinions on different topics in the form of tweets, the volumes of data generated through this platform are categorized as Big Data. Big Data is described as a set of voluminous and/or complex data that is not always possible to handle using traditional computing technologies. The South African Twitter population in 2021 is around 9.3 million users, this is 15.48% of the country's population [3].

The tweet documents hold valuable information that can be used by different entities. South African Telecommunication companies can use the tweet texts to retrieve the topics and user opinions on their service delivery. From the processed data, they can be able to identify areas of improvement in their services, product offerings and identify areas where they are doing well in the network. This paper explores this opportunity that is inherent in the availed data to develop its topic modeling.

II. Literature Review

A. Topic Modeling

Topic modeling is an unsupervised method for the classification of topics in documents. The two types of available topic models use a probabilistic and linear approach to obtain the main theme, topic or subject in a cluster of documents [4]. This study is focused on two algorithms, LDA which is a probabilistic topic modeling algorithm and LSI which is a linear approach. LDA uses a generative approach while LSI uses singular value decomposition. The similarities in the models are that they both result in vector delineations of terms and documents [5]. There is plenty of research on topic modeling utilizing Twitter data and other data non-related to Twitter. Nailing and Sheela [6] used LDA to detect cyberbullying in Twitter, in their research they were able to retrieve bullying key terms with the highest LDA score in tweets. Leah et al. [7] use non-Twitter related data to perform topic modeling using LDA, they apply LDA on Indonesian song lyrics to determine the top 10 themes with a dataset containing 193 different songs. The conclusion from the above research indicates that LDA can successfully retrieve topics from given datasets. Qomariyaha et al. [4] used both LSI and LDA for topic modeling on government twitter data and they find LDA to provide better performance scores.

B. Latent Dirichlet Allocation

LDA is considered to be the most accurate topic modeling algorithm. The basic concept with LDA is that documents represent multiple topics. The topics are a distribution over a set vocabulary using the bag-of-words model [4], [7], [8]. Fig. 1 represents the LDA Model used in this paper. In the model, there are two corpus level parameters which are α and β. Alpha (α) is a representation of document-topic density. The greater the value of α indicates a greater number of topics in the documents. Beta (β) represents the topic-word density, the greater the value of β indicates topics are formulated from most of the words contained in the corpus. Θ is a document level variable, z and w are variables at word level. The complete formula for the combined distribution *Dirichlet (α)* and *Dirichlet (β)* is represented in (1) [11]:

$$p(w,z,\theta,\phi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi) \quad (1)$$

The model generative steps can be summarized as:

1) For each topic, choose words that are most likely $\phi \sim$ *Dirichlet (β).* K represents the number of topics.

Phumelele P. Kubheka, Pius A. Owolawi, and Gbolahan Aiyetoro are with the Department of Computer Systems Engineering, Tshwane University of Technology, South Africa (e-mail: phmllkbhk@gmail.com, owolawipa@tut.ac.za, g.aiyetoro@ieee.org).

Open Science Index, Computer and Information Engineering Vol:16, No:8, 2022 publications.waset.org/10012641.pdf

2) For each document, choose the distribution of topics that should be in the document $\theta \sim Dirichlet\ (\alpha)$. M represents the documents used in the LDA model.

3) For each word, *w*:

(i) The likelihood of it belonging to a given topic *z* is determined.

(ii) A word/token *w* is chosen randomly. *N* represents the collection of words in a document.

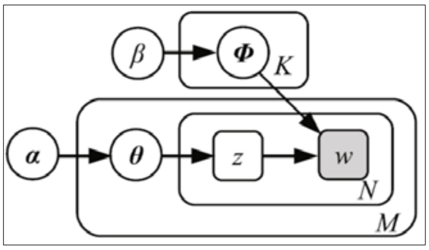


Fig. 1 LDA Probabilistic Model

C. Latent Semantic Indexing

The basic concept behind LSI is to assess the relationship between words in documents. The occurrence of terms in numerous documents indicates some degree of relatability which can mean the terms have the same meaning or indicate a common latent concept. The logic behind LSI is to first generate a key-term document matrix. The key-term document matrix is a 2-dimensional grid that shows the per word frequency in the documents within the given data set [9]. Fig. 2 represents the computation steps of the LSI Model [11]. LSI computes the topics by doing a matrix decomposition on the document-term matrix using SVD [11].

Feature Weighting

This paper uses Term Frequency-Inverse Document Frequency (TF-IDF) as a weighting algorithm. Each token acquired from the data pre-processing step is treated as a candidate term. Each term in a document has a different degree of importance. Terms with a higher TF (Term Frequency) value in TF-IDF have higher importance in the documents (Tweets).

The TF value is calculated using (2), where $n_{i,j}$ indicates the occurrence of a word in a document and $\sum_k n^{k,j}$represents the total number of words in a document.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n^{k,j}} \tag{2}$$

DF (Document Frequency) calculates the occurrence of each word in multiple documents, words appearing more often in the document set will have a higher DF value. Words with a higher DF do not carry a lot of value because they appear too often or are too common in the document set. The DF value is calculated using (3), where $|d_j\ D: t_j \in\ d_j|$ indicates the number of documents |D| that key term t_j occurs.

$$DF_{i,j} = \frac{|d_j\ D: t_j \in d_j|}{|D|} \tag{3}$$

IDF is the inverse of the DF, it measures the relevance of a word in all documents. A higher IDF value indicates the word rarely appears thus more important to the document. The IDF value is calculated using (4):

$$IDF_{i,j} = \log \frac{|D|}{|d_j \in D: t_j \in d_j|} \tag{4}$$

Finally, using equation for TF (2) & IDF (4), TF-IDF is defined as:

$$\mathrm{TFIDF} = \mathrm{TF} \times \mathrm{IDF} \tag{5}$$

Singular Value Decomposition

SVD is a matrix factorization technique, it represents a matrix in the product of two matrices. The formula for SVD can be seen in (6) [12]:

$$X = U\Sigma V^* \tag{6}$$

X is the matrix to be decomposed. It is a mxn matrix. U is the mxm left singular vector. Σ is a mxn matrix with non-negative real number on the diagonal. V is the nxn right singular vector. * indicates the complex conjugate transpose.

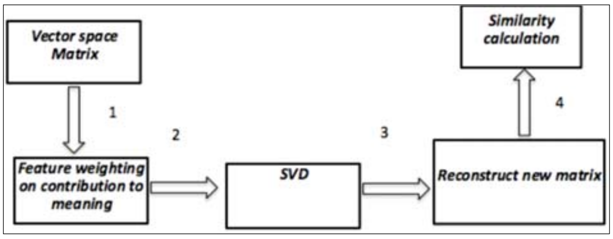


Fig. 2 LSI Algorithm

III. Methodology

This section of the paper will follow the structure as detailed below:

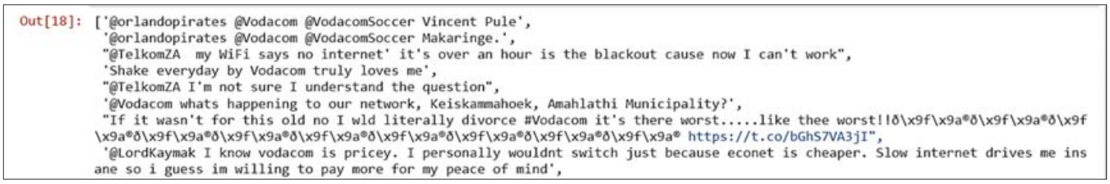
a. Data gathering,
b. Data pre-processing,
c. Parse the processed Data into LDA and LSI and Select model with a better evaluation score.

A. Data Gathering

The data used in this paper are collected from Twitter using the Twitter developer API. The widely used API in the Python framework is called Tweepy open-source, it provides access to the Twitter API. The dataset consists of tweets belonging to South African Telcos. A total of 5379 English tweets were collected. Fig. 3 shows a sample of the raw tweets collected using the Tweepy API.

B. Data Pre-processing

The pre-processing step is the most challenging. This step is critical to ensure data quality and cleanliness. The unstructured text (raw tweets) is passed through different stages for making the data more meaningful [2], [7].



```
Out[18]: ['@orlandopirates @Vodacom @VodacomSoccer Vincent Pule',
          '@orlandopirates @Vodacom @VodacomSoccer Makaringe.',
          "@TelkomZA  my WiFi says no internet' it's over an hour is the blackout cause now I can't work",
          'Shake everyday by Vodacom truly loves me',
          "@TelkomZA I'm not sure I understand the question",
          '@Vodacom whats happening to our network, Keiskammahoek, Amahlathi Municipality?',
          "If it wasn't for this old no I wld literally divorce #Vodacom it's there worst.....like thee worst!!ð\x9f\x9a®ð\x9f\x9a®ð\x9f
         \x9a®ð\x9f\x9a®ð\x9f\x9a®ð\x9f\x9a®ð\x9f\x9a®ð\x9f\x9a®ð\x9f\x9a®ð\x9f\x9a® https://t.co/bGhS7VA3jI",
          '@LordKaymak I know vodacom is pricey. I personally wouldnt switch just because econet is cheaper. Slow internet drives me ins
         ane so i guess im willing to pay more for my peace of mind',
```

Fig. 3 Raw Twitter Data Sample

The below steps are followed to generate the cleaned data:

- Case folding: this is where all words in each document are converted to lower case. The LDA model considers the frequency of each letter in each document. If a word appears multiple times in a document using a different case combination, it is treated as a different word.
- Remove mentions and links to make the data more meaningful for later analysis.
- Remove punctuations and digits: Topic modeling does not consider digits; as a result, punctuation and digits were all removed for better model training.
- Remove Stopwords: Stopwords are common terms that add no value in modeling the different topics, and thus, they were all removed from the vocabulary. These include words such as "a", “and” "this" etc.

C. Topic Coherence Score

In order to establish a well-performed model, topic coherence score becomes a metric. This metric is adopted to evaluate the two given models in this paper. The calculated coherence score indicates the best number of topics for the model training, additionally, it will give an indication on which model performs best or gives the best accuracy for a given number of topics [4]. Equation (7) shows the formula used to calculate the coherence score. The higher the coherence score the better the model. In this paper the C_v measure is used represented in (7):

$$Coherence\ (V) = \sum_{(v_i, v_j) \in V} score(v_{i,} v_j, \varepsilon), \qquad (7)$$

V represents the terms that describe each topic, ε is the smoothing constant to ensuring the calculated score is a real number.

IV. Results

A. Most Frequent Terms

The 20 most common terms in the cleaned dataset are shown in Fig. 3. From the term frequency graph, we see meaningful information. It is known that the dataset contains Tweet data belonging to South African Telecommunications companies, the below frequent terms also indicate the same, words such as network, data are common telecommunications terms, also observed is the word Vodacom appearing which is one of the South African Telecommunications companies.

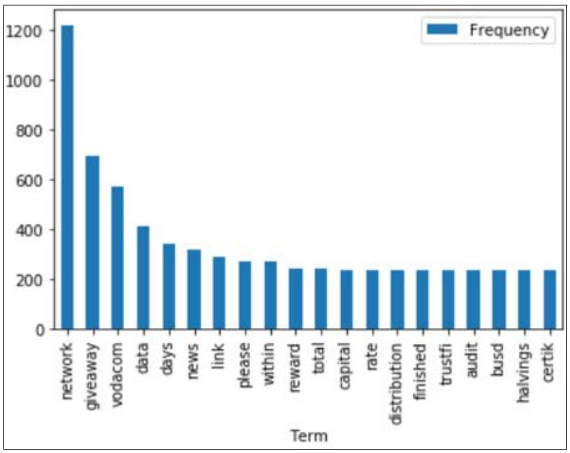
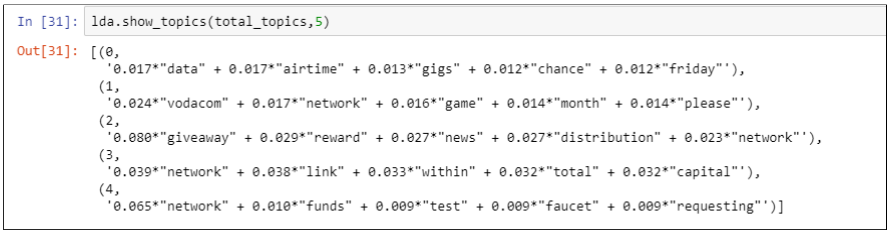


Fig. 4 20 Most Occurring Terms after Data Pre-processing

B. Model Evaluation

Table I shows the computed coherence score for LDA and LSI. Both models were trained with 5, 10 and 15 number of topics (k) to determine the best number of topics for each model. From the computed results, it was observed that the LDA model performs consistently for each given number of topics. The highest coherence score for LDA is 0.753; this is also the highest score from all the trained LDA and LSI models. From this observation, a conclusion can be made that LDA performs better than LSI for any number of given topics in this dataset. In addition, from this observation, one can also hypothesize that for larger datasets, LDA requires a longer computation time. The computing time is also dependent on the computing environment.



```
In [31]: lda.show_topics(total_topics,5)

Out[31]: [(0,
           '0.017*"data" + 0.017*"airtime" + 0.013*"gigs" + 0.012*"chance" + 0.012*"friday"'),
          (1,
           '0.024*"vodacom" + 0.017*"network" + 0.016*"game" + 0.014*"month" + 0.014*"please"'),
          (2,
           '0.080*"giveaway" + 0.029*"reward" + 0.027*"news" + 0.027*"distribution" + 0.023*"network"'),
          (3,
           '0.039*"network" + 0.038*"link" + 0.033*"within" + 0.032*"total" + 0.032*"capital"'),
          (4,
           '0.065*"network" + 0.010*"funds" + 0.009*"test" + 0.009*"faucet" + 0.009*"requesting"')]
```

Fig. 5 Top 5 words in each Topic

C. Best Model Analysis

From the results in Table I, LDA shows better evaluation scores, with the best performing model (k = 5) having a Coherence score of 0.753.

TABLE I
LDA AND LSI ALGORITHM RESULTS

Algorithm	Topics(k)	Compute Time(s)	Coherence(v)
LDA	5	4.08	0.753
LSI	5	0.52	0.689
LDA	10	2.77	0.731
LSI	10	0.52	0.683
LDA	15	4.13	0.717
LSI	15	0.45	0.704

Fig. 5 represents the word index generated from the LDA model using the show_topics method. For each topic the top 5 words are shown. The output represents the most probable words in each topic. The words that appear on the left are more indicative of the topic with words appearing towards the right contributing less to the topic. From words shown in each topic on Fig. 5, the top 3 topics can be labelled as in Table II.

TABLE II
TOP 3 TOPICS

Topic	Name
Topic 0	Data & Airtime
Topic 1	Vodacom Network
Topic 2	Competition

The inter-topic distance map in Fig. 6 shows the different topic clusters and the distance between each of the topics. The presented topics do not overlap. The size of the cluster is determined by the number of documents in it, the more documents the bigger the cluster [10]. From Fig. 6, it is seen that topic 1 has the most documents with the word vodacom having the greatest number of occurrences.

For each topic, a word cloud is generated to visualize the words appearing in each topic. The larger the size of the word the more it appears in a particular topic. Fig. 7 shows the word cloud for each topic.

V. CONCLUSION

In this paper, LDA and LSI techniques are used to perform topic modelling on SA Telco Twitter Data. From the results obtained, LDA gives better results thus making it a preferred topic modeling technique for Twitter data. The results show that tweets collected contain words most associated with Telco. For future, a much larger dataset can be used for better topic interpretability. The emotions in each topic are not clear by the words in the topic alone. The LDA topic modeling technique can be used as a feed to sentiment analysis techniques; this way the sentiment in each topic can be extracted. The resulting analysis can indicate whether the tweeter users are more positive, negative or neutral about the specific topics.

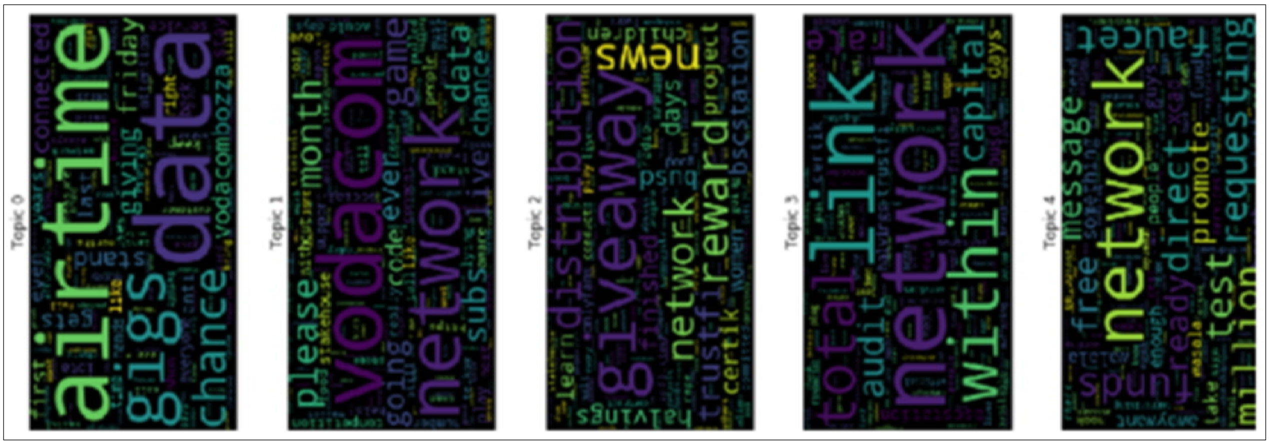


Fig. 7 Word cloud representation for each topic

REFERENCES

[1] A. Madan and U. Ghose, "Sentiment Analysis for Twitter Data in the Hindi," in *11th International Conference on Cloud Computing, Data Science & Engineering*, 201, p. 1.

[2] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet," in *International Conference on Electrical Engineering and Computer Science*, 2019.

[3] S. Writer. "The biggest and most popular social media platforms in South Africa, including TikTok." Business Tech. https://businesstech.co.za/news/internet/502583/the-biggest-and-most-popular-social-media-platforms-in-south-africa-including-tiktok/ (accessed 09/16/2021, 2021).

[4] S. Qomariyah, N. Iriawan, and K. Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis," in *The 2nd International Conference on Science, Mathematics, Environment, and Education*, 2019AIP

[5] B. Chris. "Can latent Semantic Indexation be regarded as way to do topic modeling." https://www.researchgate.net/post/Can-latent-Semantic-Indexation-be-regarded-as-as-way-to-do-topic-modeling (accessed 22/09/2021, 2021).

[6] K. Nalini and L. J. Sheela, "Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to," *Indian Journal of Science and Technology,* vol. 9(28), 2016.

[7] E. Laoh, I. Surjandari, and L. R. Febirautami, "Indonesian's Song Lyrics Topic Modelling using Latent Dirichlet Allocation," in *5th International Conference on Information Science and Control Engineering*, 2018.

[8] A. F. Hidayatullah, S. K. Aditya, Karimah, and S. t. Gardini, "Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA)," in *IOP Conference Series Materials Science and Engineering*, 2019.

[9] I. Antonellis and E. Gallopoulos, "Exploring term-document matrices

from matrix," in *SIAM Conference of Data Mining*, 2006.

[10] M. Asghari, A. S. Elmaghraby, and D. Sierra-Sosa, "Trends on Health in Social Media: Analysis using Twitter Topic Modeling," presented at the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2018.

[11] A. Rozeva and S. Zerkova, "Assessing semantic similarity of texts – Methods and algorithms " in AIP Conference Proceedings 1910, 060012 (2017), 2017, doi: https://doi.org/10.1063/1.5014006.

[12] S. L. Brunton and J. N. Kutz, "Singular Value Decomposition," in Data-Driven Science and Engineering, 2019, pp. 3-46.