الجامعة الاسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
يونيبرسيتي اسلام انتارابغسا مليسيا
(Company No. 101067-P)
Garden of Knowledge and Virtue

International Islamic University Malaysia (IIUM) Gombak

Natural Language Processing

Group Project Report

Topic Extractor For Twitter Threads using Latent Dirichlet Allocation (LDA)

SECTION 1

SURIANI BT. SULAIMAN

Group Members:

No.	Name	Matric No.
1	FARIS HAMIDI BIN MOHAMMAD NIZAM	1813951
2	FADHLUDDIN BIN SAHLAN	1817445
3	MUHAMMAD HAZIQ ADLI BIN ZAMZURI	1814981
4	WAN MUHAMMAD HAFIZAM BIN FAUZI	1810869

Table of content

1. Introduction

Our project is related to language modelling in Natural Language Processing. Language modelling is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. They are very useful for the purpose of document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. There are several existing algorithms that can be used to perform the topic modeling which is:

- □ Latent Semantic Analysis/Indexing (LSA/LSI)
- □ Probabilistic Latent Semantic Analysis (PLSA)
- □ Latent Dirichlet Allocation (LDA)

For this project, we are using Latent Dirichlet Allocation (LDA) as our based algorithm. We want to create an algorithm that is able to read the context of a thread in Twitter. We are using the Twitter API data streaming technique from a previous assignment as our test dataset from Twitter. This data then will be the input of our algorithm and then we will evaluate the result of the algorithm with human interpretation in terms of accuracy.

2. Problem Statements

Topic extraction lists the key phrases and concepts to give the gist of an article or document. This will help in extracting topics from a long text that takes a long time to read. Twitter is a social media platform that allows people to read the current trend, situation or news around the world. People can tweet on the platform and every tweet is limited in terms of its number of characters. However, users still can continue the tweet and produce a thread which is a set of tweets. Some Twitter threads are too long and this will take a lot of time to read them completely. Many people do not have much time to read everything but prefer the summary or want to know the topic of the threads. Other than that, there are threads that include unnecessary things which are out of the topic such as advertisements, clickbaits etc. This problem might annoy people who are wasting their time.

3. Motivation

Before doing the project, the group has no knowledge about the algorithm that we are going to use since we are not taught in the class about topic extraction but we do have knowledge in cleaning the dataset for NLP, extracting data from Twitter and also know about other available and much basic language modelling algorithms. Then, some research is performed in order to get some insight on the problem statement of our project and what algorithm is suitable for the project.

We also believe that the project is able to help people to know what topics are for the Twitter threads for them to quickly catch up about what the latest news in the world is all about.Our expectation is that the project can stream Twitter threads, at the same time, performing topics extraction on the Twitter threads that are extracted. We also expect that the Twitter threads and the topics extracted will have commonality thus achieve high accuracy on the model.

4. Related Work

Our project is using Twitter as the input for the LDA, while expecting the same result in terms of accuracy. A research done by Hoon, Jacob and Jungsuk (2021), they implement LDA to trace the trend in sustainability and Social media research. The objective of the research is to use language modelling to keep track of the trend of sustainability such as consumer behaviour, education, marketing and tourism sustainability. The result shows that LDA models can show the pattern of each sustainability topic from any journal. We can see that LDA can be used to determine a trend and create a topic for a specific purpose using journal and social media.

Krestel, Peter, and Nejdl conduct a research that aims to introduce an approach to recommend tags of resources in order to improve search that is based on Latent Dirichlet Allocation (LDA). They compare their approach of using LDA with association rules which is a state-of-the-art method for tag recommendation. From their research, they were able to conclude that the implementation of LDA produces much better accuracy when compared to association rules. LDA also recommends tags that are more specific which are useful for searching.

In the research paper written by Naomi, and Tim (2019), they had implemented LDA to trace the anti-vaccination movement of Facebook. The data were collected from six public Facebook pages using the Facebook Application Programming Interface (API) and the 'SocialMediaLab' package for the R programming language. From their perspective, understanding pockets of resistance to vaccination as a public health exercise provides important insights into how these attitudes may be effectively countered. From their research, they managed to get the results of 10 topic models that are related to the anti-vaccination movement with the related terms within each topic. From this research, we can conclude that LDA is very useful to implement topic modelling.

5. Algorithm Description

Latent Dirichlet Allocation is an unsupervised machine learning model that takes documents such as research articles, e-book, review papers as the input and finds the topics as its output. The algorithm will determine the topics based on the number of word occurrences in the document. The algorithm was trained using existing Python sklearn dataset to create possible generated topics from more than 5000 sources of documents. Below is a quick visualization on how LDA works:

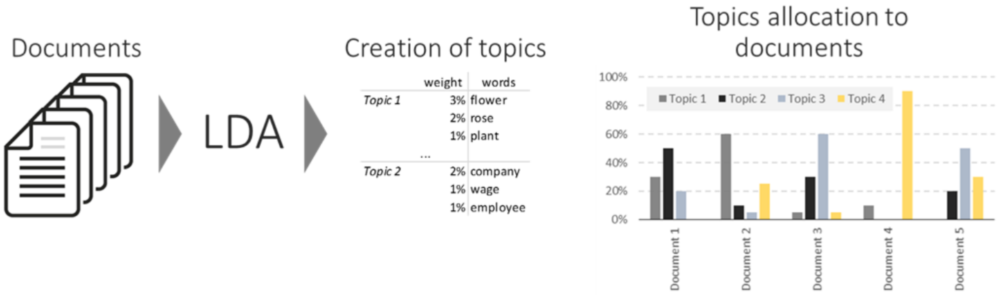


Fig. 1: Step on how LDA works

The LDA will determine the topic of the article based on the word frequency for each document. The topic creation is also based on the probability of certain words which frequently come in a sentence to create a context. As we can see from the example in the picture, the LDA determines that topic 1 has a high frequency of “flower”, “rose” and “plant”. Topic 2 has a high frequency of the words “company”, “wages” and “employee”. From the creation of the topic, the algorithm will allocate each topic on each document. For example, Document 1 is mostly discussing Topic 2 and a little bit on topic 1, the same with the other documents. Here, we can conclude that the algorithm works well on determining the topic of a document as output

6. Methodology

In our project, instead of using documents as the input. We are using Twitter threads as the input for the algorithm. We retrieved the data of the Twitter thread using Python code and then it will be stored in a csv file. Before we wanted to test the dataset using the Twitter thread, we will train the algorithm using the dataset from the sklearn package. After the algorithm finished with the training, we then proceed to test the algorithm using Twitter threads.

The algorithm then will receive the input from the csv file and will create the topic based on the words from the Twitter thread. We are expecting the algorithm to produce the same output as if we are using normal documents. Lastly, we will compare the result of the output with human interpretation as a validation method in terms of accuracy of the allocated topic on each thread.

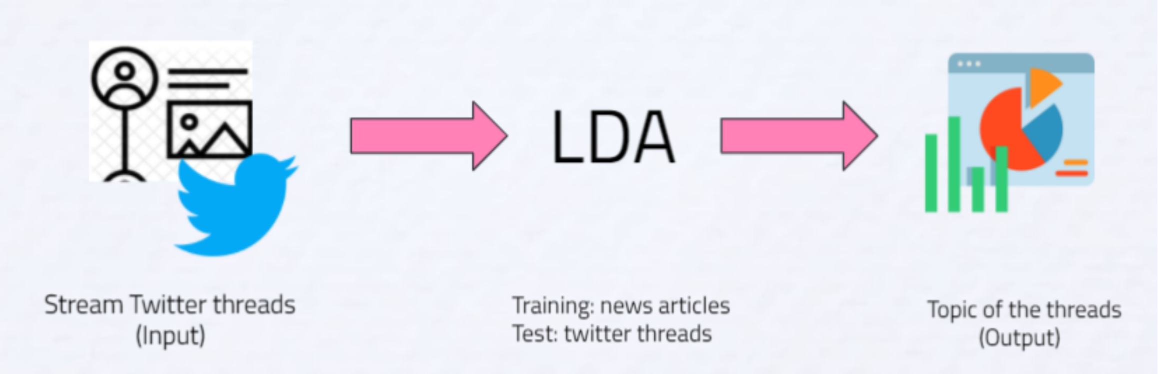
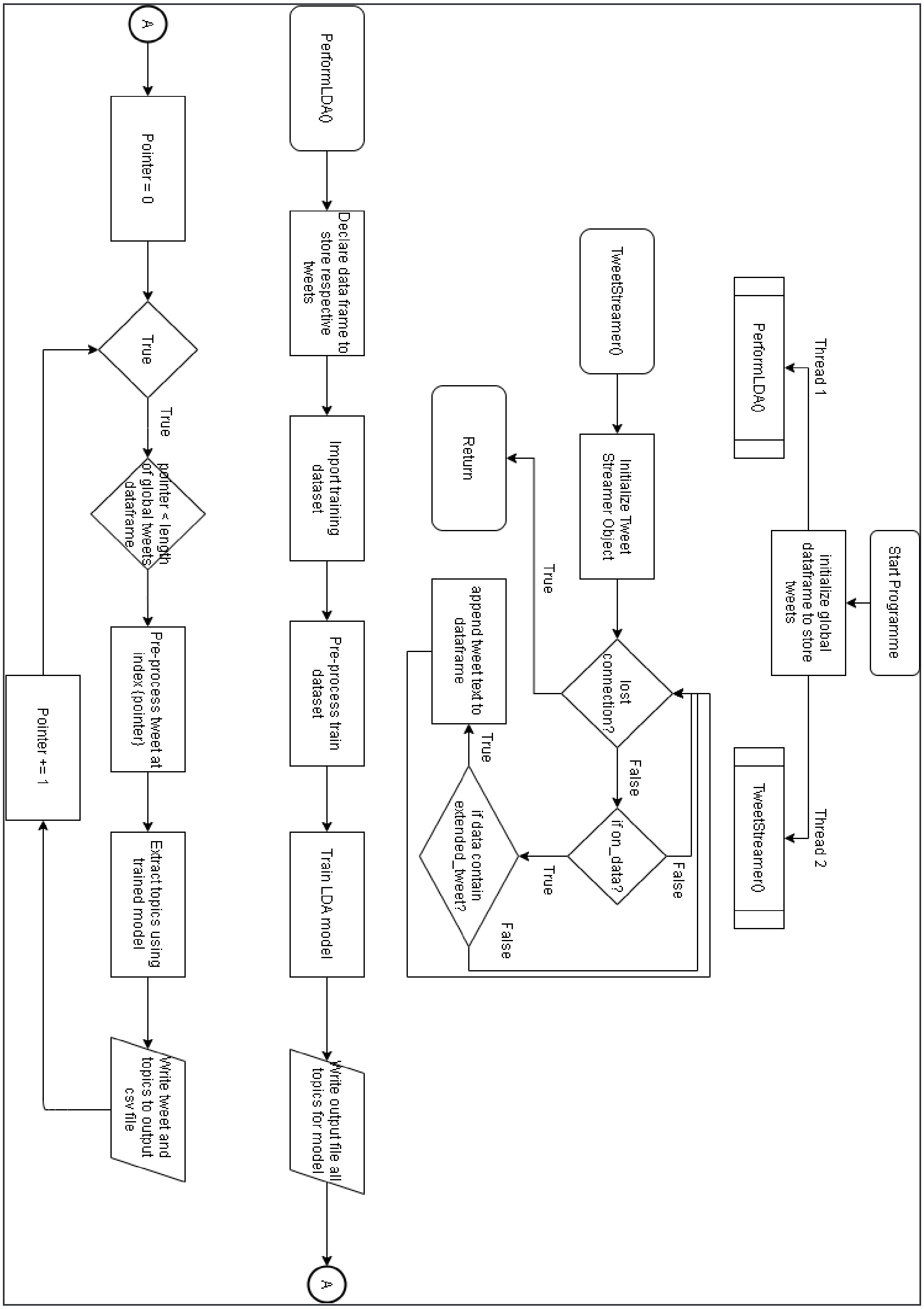


Fig. 2: Visualization of the algorithm using Twitter thread

In order to determine the performance of the model, the project will require some techniques to assess the model. Since the model is an unsupervised algorithm, we cannot determine the accuracy using machine interpretation but human interpretation is required. We will assess 100 predictions by the model and determine whether the tweet is valid with the topic assigned by the model or not. The accuracy will be calculated based on this.

7. Experimental Setup


Start Programme
initialize global dataframe to store tweets
Thread 1
PerformLDA()
Thread 2
TweetStreamer()
TweetStreamer()
Initialize Tweet Streamer Object
lost connection?
True
Return
False
if on_data?
True
False
if data contain extended_tweet?
True
append tweet text to dataframe
False
PerformLDA()
Declare data frame to store respective tweets
Import training dataset
Pre-process train dataset
Train LDA model
Write output file all topics for model
A
A
Pointer = 0
True
True
pointer < length of global tweets dataframe
Pre-process tweet at index (pointer)
Extract topics using trained model
Write tweet and topics to output csv file
Pointer += 1

8. Performance based on Human Interpretation

The performance metric of this project is accuracy. Accuracy can be calculated by evaluating from the first 100 extracted topics and tweets whether they are commonality or not between each tweet and its respective topics. Out of 100 test data involving tweets, it appears that 67 of them are considered as having commonality with their topics. This means that the model achieves 67% accuracy.

Accuracy = 67.0%

9. Result Analysis

The model seems good to determine the topics that are related to the streamed Twitter threads which the model achieved to hit 67% accuracy. This result is considered good since NLP problems often have lesser accuracy since the human language is hard to be interpreted. The topics of a Twitter thread are determined based on the 15 topics of the training model. The number of topics set for the model when training the model might be quite high since the training dataset only considers less than 15 areas of discussion. Tuning the parameter of the model might yield better performance. Training the model using a bunch of Twitter threads is also considered a way to achieve higher accuracy during testing since both training and testing data will be in the same environment. The preprocessing stage also plays an important role in this case. Enabling and disabling some preprocessing steps such as lemmatizing might cause the model yielding different accuracy.

10. Conclusion

In conclusion, the LDA algorithm is capable of extracting topics from a Twitter thread even if the model is trained by using news articles dataset. The model achieves prediction with accuracy of 67% and this is evaluated by using human interpretation on the result of the prediction. However, some questions arise on ways of making the performance better in the future. From the result, it shows that LDA is suitable to be used as the model to extract topics from Twitter threads. For future works, more varied values of parameters should be used to train the model to see the effect of the parameters to the performance of the model. Other than that, we believe that it is better to use the dataset of Twitter threads as the training data since it has the same environment as the testing data which is also Twitter thread. We hope that the future research will yield better accuracy of this research problem.

11. References

1. Lee, J. H., Wood, J., & Kim, J. (2021). Tracing the Trends in Sustainability and Social Media Research Using Topic Modeling. *Sustainability*, *13*(3), 1269. https://doi.org/10.3390/su13031269
2. Krestel R., Fankhauser P., and Nejdl W. (2009). Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems (RecSys '09). Association for Computing Machinery, New York, NY, USA, 61–68. DOI:https://doi.org/10.1145/1639714.1639726
3. Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. Information, Communication & Society, 22(9), 1310-1327.