Multidisciplinary Applied Research and Innovation Vol. 3 No. 2 (2022) 84-91

MARI

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/mari
e-ISSN :2773-4773

Opinion Analysis Based on TNF (Textual Noise Fixing) Algorithm

Alen Lam Ming Feng, Lee Jin Hong, Soh Ying Ying, Abdul Halim Omar*

Department of Information Technology, Centre for Diploma Studies,
Universiti Tun Hussein Onn Malaysia, Pagoh Higher Education Hub,
84600 Pagoh, Johor, MALAYSIA

DOI: https://doi.org/10.30880/mari.2022.03.02.012
Received 31 March 2022; Accepted 31 May 2022; Available online 28 July 2022

Abstract: This project objective is to develop TNF (Textual Noise Fixing) algorithm. The main goal of this project is to successfully create a pre-processing algorithm that can clean the noise of data set to solve the problem which is computer does not understand human languages, particularly truncated words, spelling errors, and the usage of symbols in sentences. In this project, we use an algorithm track and manage people's comments by mining sentimental tendencies from online people's comments. We collect people's comments from Facebook post and develop pre-processing algorithm which is cleaning, fix short forms, tokenization and lowercase. We will fix all the short forms, remove unwanted symbol, separate term and make a lower-case transformation for the sentences. After pre-processing process, the next step is sentiment clustering. We use K-means clustering to divide an object into k clusters. Through the above steps, the results will be show by histogram. Based on the result, we can see all the transformation in every steps. We hope that in the future, we can use this project to easily analyse user's comments.

Keywords: natural language processing, data mining, k-means, cluster

1. Introduction

This study called opinion analysis based on the TNF (text noise repair) algorithm. The current outbreak of Covid-19 in Malaysia has brought much trouble to Malaysians and impacted businesses. This indirectly resulted in our wanting to collect people's comments from the Internet. It uses an algorithm to track and manage people's comments by digging out sentimental tendencies from people's comments on the Internet. Usually began to study natural language processing in the 1950s. Alan Turing published an article in 1950 called "Computer Machinery and Intelligence", in which Alan Turing proposed the intelligence standard now known as the Turing Test. However, in 1950, Alan Turing developed the test-the Turing test, which is the ability of a machine to show intelligent behaviour equivalent to or indistinguishable from humans. Natural Language Processing (NLP) is the automatic manipulation of text or speech through software. Speech recognition, understand natural language, interpretation and generate natural language are all included in NLP. In addition, until the 1980s, most NLP systems used complex handwritten rule sets. But NLP has undergone a revolution because the

*Corresponding author: halimomar@uthm.edu.my

language processing machine learning (ML) algorithm was introduced at that time, which occurred in the late 1980s [1].

2. Literature Review

2.1 Introduction to Natural Language Processing

Natural language processing is a separate part of artificial intelligence. Natural language processing (NLP) can help computers interpret, understand, and exploit human languages. NLP extracts computer science and computational linguistics content, filling the gap between human communication and computer understanding. Machine code or machine language is also known as the native language of the computer. At the lowest level of the device, communication is through logic rather than through words, operations generated by millions of zeros and ones [2].

2.2 Types of Natural Languages Processing

Natural language processing is a separate part of artificial intelligence, and it is planned to understand the nature of intelligence. Has successfully produced a new type of intelligent machine that can react similarly to humans. Robots, language recognition, image recognition, natural language processing, expert systems and so on are the scope of research in this field. The theory and technology of artificial intelligence are becoming more and more mature, and the field of application is constantly expanding. The information of human consciousness and thinking can be simulated by artificial intelligence. Artificial intelligence thinks like humans. Artificial intelligence is not human intelligence, but it is also possible that artificial intelligence can surpass human intelligence. Artificial intelligence is a challenging science, including a vast range of sciences, consisting of different fields. The main goal of studying artificial intelligence is to enable machines to complete complex tasks that usually require human intelligence [3].

2.3 Cleaning

Table 1: Example of data cleaning

Description	Input Text	Output Text
Remove unwanted symbol	Jom makan!!!	Jom makan
Remove unwanted symbol	Aduh...	Aduh
Remove unwanted symbol	Betul tak???	Betul tak

Data cleaning is the process of correcting or deleting data from a dataset that is inaccurate, corrupted, improperly formatted, duplicated, or incomplete. But for our project in this part, we use data cleaning to clean the punctuation in the sentences. We must constantly clean the data before we can fit a machine learning or statistical model. With confused data, no model can produce useful outcomes. Here is the example for cleaning in python [4]. Some examples of data cleaning are shown in **Table 1**.

2.4 Data Mining

Data mining is a process for extracting and detecting patterns in huge data sets that combines machine learning, statistics, and database systems. Data mining is the process of examining and analyses large amounts of data in order to uncover important patterns and trends. Database marketing, credit risk management, fraud detection, spam email cleaning, and even detecting user sentiment are all possible applications. We need a huge data to complete our project. So, in this project, we choose to collect the comments from Facebook post because the Facebook platform is one of the largest social media across the globe.

3. Methodology

This section will explain the methodology used in this project. This section will introduce and mention the research of the various components involved and the theoretical analysis of the procedures and concepts related to the topic. This section proposes a framework as a guide for researchers to achieve the research goals and ensure that this pre-processing work well.

3.1 Methodology Process

This project the main idea is to describe the proposed algorithms. The methodology process of this project are Data Selection, Text Document Pre-Processing, Proposed Algorithm, Automatic Sentiment Clustering and Evaluation. **Figure 1** shows framework that using as method process.

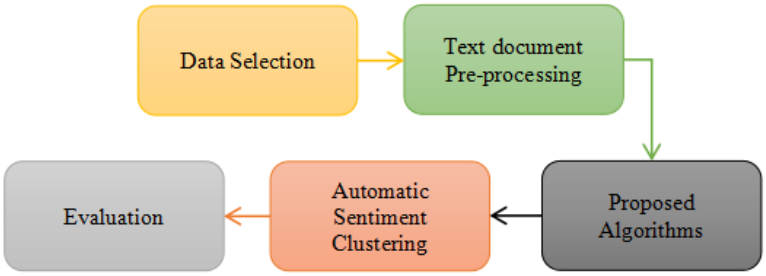


Figure 1: Framework of method process

3.2 Data Selection

Regarding data selection, the comment dataset is from the comments on the social media Facebook platform because it is the largest social media platform with 2.85 billion users. In order to let everyone, understand the people's views and thoughts on the impact of Covid-19 on business. We found many positive and negative comments, so we collected 500 real comments from netizens. The focus of the comments was on the impact of Covid-19 on business. The topic is the response of netizens after former Prime Minister Dato Seri Najib Razak uploaded a statement about people requesting temporary closure of the factory and Lockdown. For more organized data, all comments are collected and isolated. All these comments are manually fetched and placed in Microsoft Excel [5].

3.3 Text Document Pre-Processing

The text document is very important because any unnecessary or less meaningful punctuation is deleted to support sentiment analysis to produce satisfactory results. We use three methods. **Table 2** shows some examples of standard text pre-processing the task for a text document.

Table 2: Text document pre-processing

Pre-processing	Description	Input Text	Output Text
Data Cleaning	Removed unwanted symbol	Hello, how are you?	Hello how are you
Tokenization	Separate term	I like food	“I” “like” “food”
Lower Case	Lower case transformation	HELLO WORLD	Hello world

The first method is data cleaning. An important step in the text cleaning process is to remove unnecessary characters or punctuation marks, such as periods, commas, question marks, and so on.

Inject unnecessary characters or punctuation marks in the code, so that output we can get a sentence without characters or punctuation marks. After cleaning up 500 comments, we put the data into an excel file.

Next, the second method is text tokenization, which is to decompose a complex sentence into words. The tokenization process is to mark a large amount of text as sentences; sentences are marked as words. First, we open RapidMiner Studio, use "process document from files" to put data files into file directories, and then use tokenization to decompose complex sentences into words.

In addition, the third method is Transform Case. Transform Case converts all uppercase characters to lowercase. If there are no uppercase characters in the content, the original string is returned. After continuing to use tokenization, then use transform case to convert all characters to lowercase.

3.4 Proposed Algorithm

Changing the misspelt words can enable the proposed algorithm to produce good results for sentiment analysis. To fix the short format, we need to restore some shortened text to the original text. People like to use shortened text to express what want to say because this can reduce the number of words in the text that need to enter, and people can quickly enter what people want to express. So, in most comments, you can see that netizens use many short forms. These shortened texts were basically invented and disseminated on the Internet. The shortened text can also be referred to as web terminology. To have a good result algorithm, we used python to restore the shortened text to the original text. **Table 3** shows some examples of fix short form.

Table 3: Fix short form

Before	After
blm	Belum
aq	Aku
xde	Tidak ada

Filtering stop words is also a way to achieve good results in sentiment analysis. Filtering stop words means filtering out the meaningless and useless words that are ignored by automatic programming. It can improve search efficiency. This method can be safely ignored and does not change the meaning of the sentence. We use tools in RapidMiner Studio to filter stop words. **Table 4** has some examples of different sentences with stop word and sentences without stop word.

Table 4: Different sentence

With Stop Word	Without Stop Word
I like to eat french fries	Like, eat, french fries
Listening to the song is a lot of fun	Listening, song, fun
He is kind	Kind

To calculate the statistics of keywords, we use Term Frequency-inverse Document Frequency (TF-IDF). It is used to measure the importance of a word in a document.

TF: The frequency of a word in the article, the higher the frequency, the more likely it is a keyword.

IDF: A word covering many documents means that the word is less important. The closer the DF index is to 0, it means that the word is very common [6].

Actually, TF−IDF is equal to TF∗IDF. 500 comments, average each comment has 30 words. The first step is to find the word covid. The word covid appears twice. The second step calculates the TF, the number of times the term appears in the document divided by the document's total number of terms. The third step is to find out how many times covid appears in all 500 comments. The fourth step is to calculate the IDF index and use the log() to divide the total number of documents divided by the number of documents containing entries to obtain the IDF index. The last step is TF∗IDF. Multiply TF by IDF to know the importance of the word in a file set. **Figure 2** is the example for TF-IDF algorithm.

Step 1: Average each comment has 30 words, where in the word covid appear 2 times.
Step 2: TF (2 / 30) = 0.066
The algorithm of TF is (the number of times the term appears in the document) divided by (the total number of terms in the document)
Step 3: A document has 500 comments, where in the word covid appear 44 times.
Step 4: IDF (500 / 2) = 2.70
The algorithm of IDF is log (total number of documents) divided by (number of documents containing entries).
Step 5: TF-IDF is the product of these quantities
0.066 * 2.70 = 0.1782

Figure 2: TF-IDF algorithm

3.5 Automatic Sentiment Clustering

There are hundreds of sentiment clustering used in this field. What we are going to use today is called cluster analysis, it is the task of grouping a group of objects in such a way and that the objects in the same group we called a cluster. If are more similar to each other than objects in other groups (clusters). In this section we plan to use K-means clustering. What is K-means clustering? The K-means algorithm was originally Inspired by Macqueen (1967). K-means is a technique in cluster analysis work Assign the term or object to the nearest K point or centroid. **Figure 3** is the example for K-means algorithm.

Step 1: Partition the items into K initial clusters
Step 2: Scan the list of items and assign each item to its centroid(mean) is the closest. Each time an item is reassigned, the cluster will be recalculated. The mean or centroid of the clusters that received the item and the clusters that lost the item.
Step 3: Repeat step 2 over and over again until no more reallocations are made.

Figure 3: K-means algorithm

K-Means clustering aims to divide n objects into k clusters, where each object belongs to the cluster with the nearest mean. This method just produces k different clusters with the largest possible difference. The optimal number of clusters k leading to the largest separation (distance) is not known a priori and must be calculated from the data. The goal of K-Means clustering is to minimize the total intra-cluster variance, or squared error function [7].

$$J = k \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \qquad \textbf{Eq. 1}$$

Eq.1 is the K-means basic equation, “J” is the objective function, “K” is the number of clusters, “n” is the number of cases, “x” is the case of I, “c” is the centroid for cluster of j. By calculating the nearest mean to every centroid, it will partition the data according to the space or distance of each cluster to the specified centroid. One of the advantages of this method is There is no need to calculate the distance metric between all pairs of objects. Therefore, when dealing with very large data sets, this process seems more effective or practical. The user initializes the number of k clusters to be considered. The area unit

2 approaches in K-means procedure and it's varied on however the procedure begins the partitioning. the primary approach is to try to random partitioning of subjects into teams repetitively. the choice is to start out with a further set of beginning points till the centres of clusters area unit shaped.

3.6 Accuracy Table

The accuracy table stores the raw number of true positives, false positives, true negatives, and false negatives at many probabilities thresholds. The accuracy table is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. The following are actual class and predicted class tables. Respectively, **Table 5** is the actual class, and **Figure 4** is the accuracy predicted class.

True positives (TP): These are cases in which we predicted yes, and they do.

True negatives (TN): We predicted no, and they do not have.

False positives (FP): We predicted yes, but they do not actually have.

False negatives (FN): We predicted no, but they actually do.

Table 5: Actual Class

	Class = Yes	Class = No
Class = Yes	True Positive	False Negative
Class = No	False Positive	True Negative

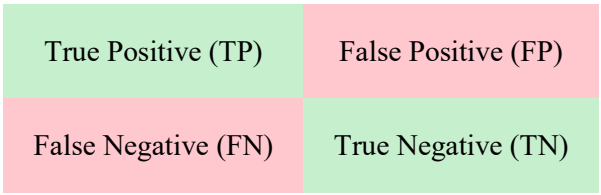


Figure 4: Predicted Class [8]

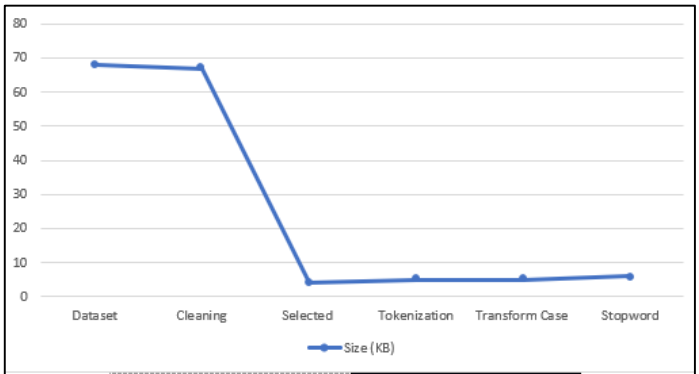
3.7 Evaluation

Evaluation is the last stage, it plays a vital role, any errors in it. If repair or correction is needed, it should be based on Functions created for sentiment analysis to obtain satisfactory results. This Part discusses the measurement or verification of the results obtained using the proposed algorithm. The measurement will be performed by using F measurement. The following table shows the analysis of forecast accuracy levels.

The True Positives (TP), they are positive values for correct predictions This means that the value of actual class is true, and the predicted value is the same is true for classes. Then, for example, if actual class value indicates the comments were positive, and the predicted class told the same thing. Next, really A negative number (TN) is a negative value predicted correctly, which means the value of the actual class is false, and the value of the predicted class is false. in order to. Explain if the actual class says the reviews are negative and predictive the class told the same thing. False positive (FP) an actual class is false, and the predicted class is true. If the actual class value indicates that the review is negative, but the predictive class will tell us that the review is positive. Although false negative (FN) is the actual class is true, but the predicted class is false. If the actual class value Forecast accuracy level Actual class, class = yes, class = no, Category = yes true positive false negative, Class = true negative without false positives Indicates that the comment is positive, and the predictive class tells that the comment is Negative.

4. Results and Discussion

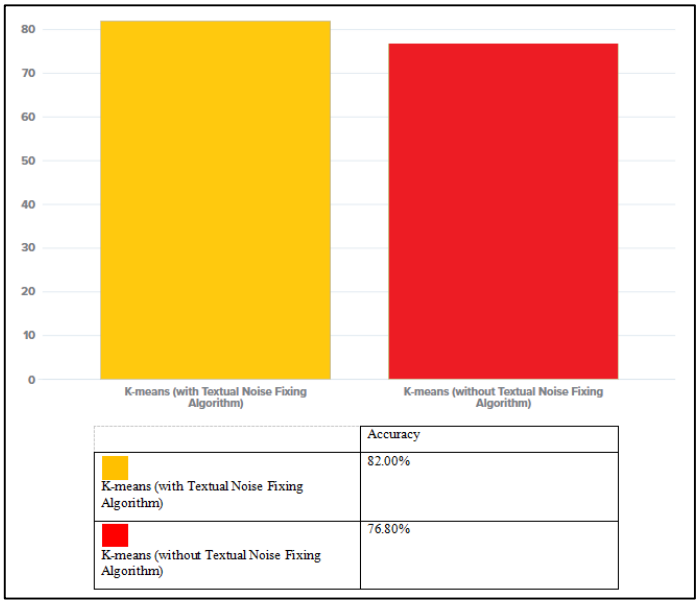
Figure 5 shows the size after every process. All the comments are collected from Facebook and the size is 68 kb. 68 kb is the size of all comments we choose which is 500 comments from Facebook page Najib Razak posted on 17th of June at 9.29p.m. After cleaning process, the size changed from 68kb to 67 kb because we clean all the unwanted symbol in the sentences. After selected the positive and negative data, it reduces to 4kb because only this 4 kb data has been chosen. When tokenize, the size increase from 4 kb to 5 kb because the word was separated, and it will take some space. Next, the lowercase transformation will make sure all the words will be lowercase and none of the words will be uppercase. Lastly, the last process is stopword filtering.



	Size (KB)
Dataset	68
Cleaning	67
Selected	4
Tokenization	5
Transform Case	5
Stopword	6

Figure 5: Reduction size of dataset

Figure 6 show the accuracy dataset of k-means with Textual Noise Fixing Algorithm and k-means without Textual Noise Fixing Algorithm. The accuracy dataset of k-means without Textual Noise Fixing Algorithm is 82%. The accuracy dataset of k-means with Textual Noise Fixing Algorithm is 76.80%.



	Accuracy
K-means (with Textual Noise Fixing Algorithm)	82.00%
K-means (without Textual Noise Fixing Algorithm)	76.80%

Figure 6: Dataset accuracy

5. Conclusion

In conclusion, the purpose of this project is to collect the comments of users and classify them. After classification, we performed methodology to make the data can be analyse by the computer. In this project, the steps of methodology presented such as cleaning, fix short form, tokenization, data mining, and also translate. After the methodology, our result was successfully produced and displayed. We hope that in the future, we can use this project to easily analyse user comments and display them through graphs.

References

[1] P. Canuma, 'The brief history of NLP', *DataDrivenInvestor*, 2019. https://medium.datadriveninvestor.com/the-brief-history-of-nlp-c90f331b6ad7 (accessed Aug. 02, 2021).

[2] SAS, 'Natural language processing (NLP): What is it and why is it important'. https://www.sas.com/zh_cn/insights/analytics/what-is-natural-language-processing-nlp.html (accessed Aug. 02, 2021).

[3] J. Spacey, '9 Examples of Natural Launguage Processing', *Simplicable*, 2016. https://simplicable.com/new/natural-language-processing (accessed Aug. 02, 2021).

[4] Lianne & Justin, 'Data Cleaning in Python: the Ultimate Guide', *Towards Data Science*, 2020. https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d (accessed Aug. 02, 2021).

[5] N. Razak, 'Tutup Kilang', *Facebook*, 2021. https://www.facebook.com/najibrazak/posts/10157957349045952 (accessed Aug. 02, 2021).

[6] B. Stecanella, 'What Is TF-IDF?', *MonkeyLearn*, 2019. https://monkeylearn.com/blog/what-is-tf-idf/ (accessed Aug. 02, 2021).

[7] S. Sayad, 'K-Means Clustering'. https://www.saedsayad.com/clustering_kmeans.htm (accessed Aug. 02, 2021).

[8] 'Classification: Accuracy', *Google Developers*. https://developers.google.com/machine-learning/crash-course/classification/accuracy (accessed Aug. 02, 2021).