SENTIMENT ANALYSIS USING TOPIC MODELLING BASED ON THE PERFORMANCE OF SENTIMENT CLASSIFIER

[1]Fazidah @ Idah Binti Wahit, [2]Nor Samsiah Binti Sani
[1,2]CAOT, FAKULTI TEKNOLOGI SAINS MAKLUMAT, UKM
P97716@ukm.edu.my

ABSTRACT: *Sentiment analysis is a very important field of social media that deals with the identification and analysis of sentimental contents that is generally available in social media. Twitter is one of such social medias used by many users about some topics in the form of tweets. The first problem is a wide number of comments increasingly conducted almost every day have resulted in social media being more difficult if not becoming more complex in analyzing the sentiments. The second problem is The Bag Of Word (BOW) representation is unable to recognize synonyms from a given term set and unable to recognize semantic relationships between terms. Third problem is the weaknesses of text/sentiment classifiers is concerned with the process of extraction, where some of the different words have the same weight. Latent Dirichlet Allocation (LDA) is topic modelling and sentiment analysis by examining Twitter plain text data in English. Due to a large number of short texts available, effective topic models for extracting hidden thematic structure from short texts have become essential for many tasks requiring a semantic understanding of textual content, like short text clustering short text classification and sentiment analysis The basic idea is to treat the documents as mixtures of topics in the topic model, and each topic is viewed as a probability distribution of the words. The framework to develop the proposed improvement using Topic Modelling that includes numerous stages: (1) acquiring the standard datasets (Twitter Sanders dataset)and preprocessing of the dataset using text processing methods. (2) employing the Topic Modelling methods, and (3) analyze and evaluate the selected features from the previous process sentiment classification in classifiers. This study also does light analysis of features through the Shap Visualization and Statistically experiment. In the Shap Visualization this reseach extracted significant topics from the clusters, split them into positive, negative sentiments and identified the most frequent topics using the proposed model. Based on the result using different classifier which are KNN, DT and NB after the implementation of Topic Modelling, it shows that LDA is an appropriate model to improve traditional feature extraction method. Thereafter, different classifications were employed on these datasets which enabled the exploration of the performance of traditional classification. Result show improvement on the improvement of feature extraction using Topic Modelling as it showed higher performance compared to traditional methodologies which are BOW.*

Keywords: Sentiment analysis, Topic Modelling, sentiment classifier, Social media data

1. INTRODUCTION

Sentiment analysis and opinion mining is the field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language [1]. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining[2]. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter and other social networks. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis[3].

However, A wide number of comments increasingly conducted almost every day have resulted in social media more difficult if not being complexed in analyzing sentiments. The weaknesses of text/sentiment classifiers is concerned with the process of extraction, where some of the different words have the same weight [4]. However, topic modelling as extraction have not been used for extraction in text/sentiment classifiers, which avoids in giving the same weight for different words. Other problem related to text/sentiment classifiers is the problem of dimensionality of features[5]. One of the most popular unsupervised machine learning approaches for feature based opinion mining is topic modeling. In machine learning and natural language processing, topic modeling is type of statistical modeling for discovering the abstract topics that occurs in a collection of documents [6]. Topic modeling is unsupervised learning method and it assumes each document is consists of a mixture of topics and each word is the probability of distribution over words [7].

The first aspect of the problem for feature extraction that need improvement is the weakness of extracting important information “words or terms” from comments to distinguish between different topics or types of sentiments. To extract terms from text Syntactic terms (Bag-of-Words) and Semantics sense are the best levels to be adopted (Bag-of-Concepts ‘Bag of Word’ with their semantic) [8].

Vast amount of text classification studies make use of the bag-of-words model [9] to represent text documents where the exact ordering of words or terms in the documents is ignored but the number of term occurrences is considered. Each distinct term in a document collection consequently constitutes an individual feature. Thus a document is represented by multi-dimensional feature vector where each dimension of the vector corresponds to the weighted value for a distinct word within the document collection, which is also known as the vector space model [10]

Different approaches are proposed to solve problems of feature based opinion mining. Liu [11] classifiying these approaches into four categories: 1.Finding frequent nouns and noun phases, 2. Using opinion and target relations, 3. Using supervised learning, 4. Using topic models and mapping implicit aspects. Schouten and Frasincar [12] discussed taxonomy for aspect-level sentiment analysis approaches.

Brody and Elhadad [13] proposed LDA based model for feature based opinion mining and summarization. Model

extracts topics as features at sentence level. Polarity identification is done for each feature using seed adjectives with known polarity.

The main objective of this study is to assess the use of Topic Modelling to extract features and provide an upper-bound to the performance of sentiment classifier. Feature are tunned using an incremental approach that have five different k values 50, 100, 300 and 500 on four corpora. The main questions investigated here are: (i) how topic modelling can improve the result among four different k values (ii) what are the best feature among different k values (iii) how is the sentiment in the topic modelling.

This paper focuses on topic-level sentiment analysis in which the topic is extracted from a sentence and then sentiment analysis with references to the extracted topic and corresponding sentence is performed.

2. RELATED WORKS

2. 1 Topic Modelling based methods

Sentiment mining is a part of data mining which process the Electronic text and tag the words into three categories that are positive, negative and neutral. Different techniques are used for sentiment analysis, classification and summarization. Different techniques are used for sentiment summarization such as Data mining, classification of Text, Information Retrieval and Summarization of Text shows general structure of sentiment analysis. Sentiment analysis can be achieved at various levels, the levels are: Phrase Level, Aspect Level, Sentence Level, Document Level, Natural Language Processing. Depending upon nature of use the level of Sentiment analysis is selected [14].

Sentence level classification deals with the considering polarity of each sentences. Document level classification can also be applied to sentence level classification to classify the sentences in polarity. Here also we have to consider the subjectivity and objectivity of the sentence. Subjective sentences contain words related to particular domain. Single sentence contains single opinion about single domain. Complex sentence are also commented in reviews. In such case sentence level classification cannot be useful. Sentence level classification deals with the positive, negative and neutral sentiments. Sentence level classification deals with the subjectivity classification. For Example, "I brought a Canon Camera last week. At initial stage, everything was good. The pictures were high quality and clearer, although it was a bit bulky. Then it stopped working today". The first sentence contains no opinion as it simply states a fact. All other sentences expresses implicit and explicit opinions. The last sentence "Then it stopped working today" is objective sentence but the current used methodology cannot express opinion for the above sentences even though it carries negative sentiment or undesirable sentiment [15].

Machine learning methodology consist of supervised, unsupervised and semi supervised categories .Each category is again sub divided into as shown in Figure 1. Supervised Learning methodology predicts attribute classes on the basis of given set of training values. It contains training and testing dataset. Training dataset is smaller which contains same attributes as testing datasets. It is more efficient and accurate. A training dataset creates model testing on test corpora contains the same attributes but no predicted attributes. Accuracy of model checks how accurate it is to make prediction. Classification is a supervised learning used to find the relationship among attributes. Prediction hit rate is used to measure the accuracy of extracted rules that shows how true they are to make prediction by applying on test data [16].

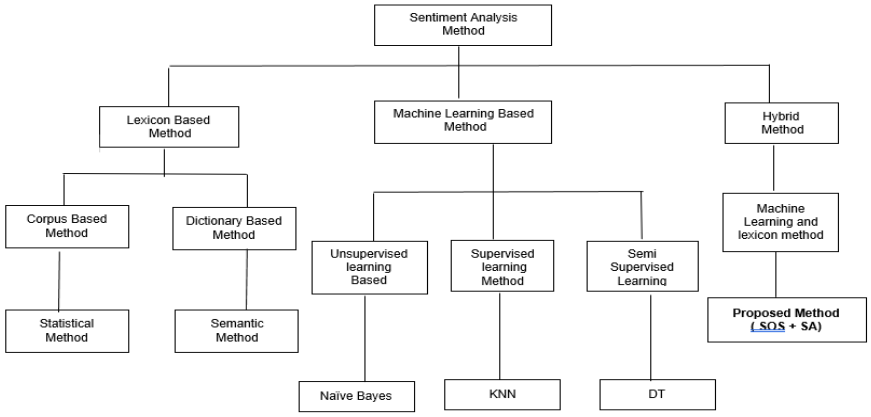


Figure 1: Category of sentiment analysis

2.2.1 Fundamental Topic Models

Due to its wide range of applicability in multitude of text mining and natural language processing tasks, significant research work has been done in topic modeling. Probabilistic topic models such as LDA [17], LSI [18], PLSA [19] etc are used for analyzing and extracting latent topics from large collection of text dataset which work on the assumption that each document is posit as mixture of topics. Every topic is represented using probability distribution over words. Latent topic parameters in these graphical models have direct connections with observed parameters, which represent words in a document.

LDA is a widely preferred probabilistic generative model. Due to the intractable nature of exact inference method, LDA uses a convexity-based variational method based on approximate technique. LDA based topic models can be easily deployed in complex architectures due to their modular nature. LSI, on the other hand, does not support this property.

LSI is based on SVD method and extracts implicit semantic structure in the document collection.

PLSA is a modified version of latent semantic analysis model. It is based on statistical latent class model. However, these models suffer from the following major drawbacks. (1) Exact inference in such models is intractable and therefore slow or inaccurate approximation techniques are required for computing the posterior distributions over topics. (2) These models cannot capture the essence of distributed representations.

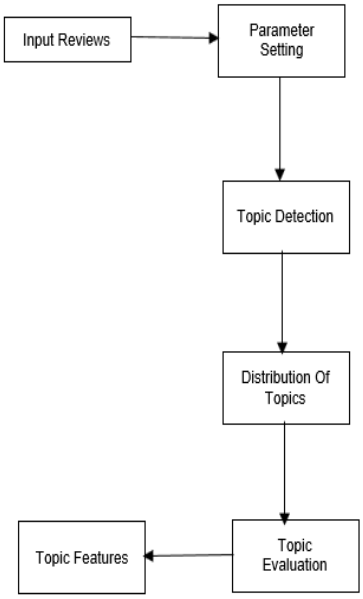


Figure 2 : Process Of Topic Modelling

2.2.2 Scalable Topic Models

Online LDA proposed in [20] follows non-Markovian Gibbs sampling and maintains weight-matrix history in the generative process based on the homogeneity of domain. However, intertopic differences are not handled by this model. Topic model puts forth in [21] performing dynamic prediction of future trends for temporally sequenced data and supports scalable topic modeling. The conventional inference process requires multiple passes over the document for topic modeling. This inference process proves to be inefficient when topic modeling is applied over large scale data. Additionally, documents may contain large number of short text documents and it raises ambiguity in merely using words of documents for analyzing the data. To address these issues, Hennig et al. [22] patented a method for topic modeling that trains a model by sequential processing and incorporates to use features of documents (author of the document, document metadata, author location, etc.) for topic modeling. To support scalable topic modeling, the method of regularized latent semantic indexing has been patented in [23]. This method allows an equation involving approximation of the term-document matrix to be executed in parallel. For memory-limited environments, Pathak et al. [24] proposed disk-based topic modeling approach, namely, BlockLDA, which applies space reduction technique and local scheduling technique for minimizing the disk I/O and supports scalable topic modeling even if data and model do not fit into the memory. Deep probabilistic autoencoder for topic modeling proposed in [25] supports scalable Bayesian inference on big data. This model is based on stochastic gradient Markov chain Monte Carlo (MCMC).

2.2.3 Sentiment Classifier: Data Tree, Naive Bayes, KNN

2.2.4 LDA Topic Model

Topic models are generative probabilistic models and unsupervised like clustering methods. Represent topics as multinomial distribution of words. Few of the topic models are latent dirichlet allocation (LDA), latent sematic analysis (LSA) and Non-negative matrix factorization (NMF). LDA works on probabilistic graphical modelling and NMF works on linear algebra. LSA uses singular value decomposition (SVD) method to reduce the document term matrix which is input to topic models.

Topic model framework is shown in Figure 3. Topic model take text reviews as input and latent parameters including number of topics are set. Topics are detected automatically using probability distribution between words and topics. Distribution of topics are produced as output and topics are evaluated by using parameters like perplexity, log likelihood and topic coherence. Topic vectors are derived from topic distributions which are the features to sentiment classification and prediction. Perplexity is the evaluation criteria of estimating the LDA model. Smaller the perplexity value, model will generate topics with high performance.

3. PROPOSED METHODOLOGY

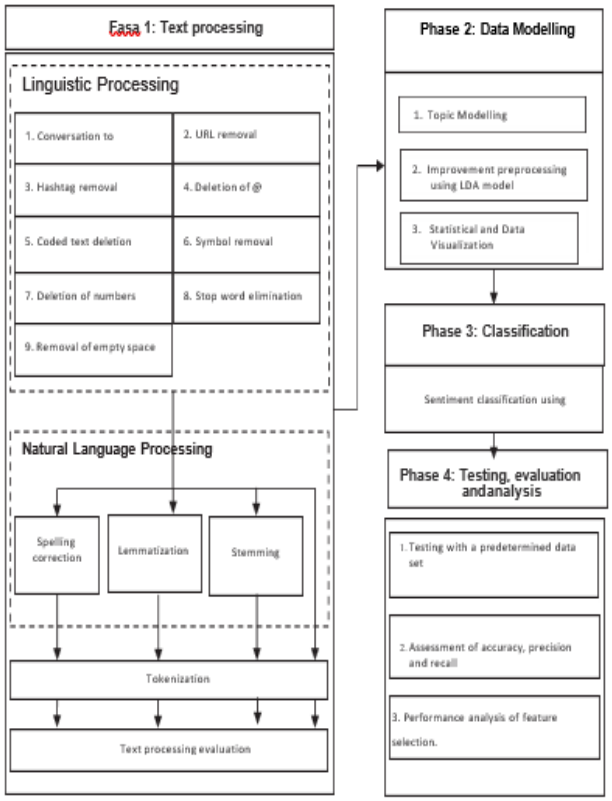


Figure 3: Research methodology

3.1. Phase 1: Text Processing

3.1.1 *Text Processing Methods*

In this phase all the text data is cleansed off. All the unnecessary white spaces, tabs, newline character is removed from the text. The URLs from the tweets are removed. The RT tag mentioned before every retweeted tweet is removed. All punctuations, numbers are also removed from the tweets. Unnecessary sparse terms are removed. The stop words are removed from the tweets. All text is converted to lowercase to have consistent messages. Stemming is performed on each word of tweet

3.1.2 *Tokenization*

After filtering the noise from that dataset, all that was left were raw words in the sentences. These words individually have some meaning and may consist of emotion or sentiment expressed by the user in the tweet. In natural language processing the process or steps for breaking down sentences into words and punctuation marks is known as tokenization [26]. The goal for generating the list of words by separating the string is called tokenizing sentences into words [27]. Here, to tokenize the words, the Natural Language Toolkit (NLTK) tokenize package Tweet Tokenizer module is used. The choice for selecting tokenizer depends on the characteristic of data and the language. The algorithm for word tokenization using Tweet Tokenizer is shown in Table 1.

Table 1. Algorithm for word tokenization

Input: Filtered tweets Output: Tokenize words
For all words in Processed Tweets Tokenize the word passing to Tweet Tokenizer Method and append Tokenize Sentence Return Tokenize Sentence

3.1.3 *Text Processing Evaluation*

This evaluation process is performed using a set of features generated from the tokenization process, which are evaluated through the accuracy of sentiment classification. The classification algorithm used in this process is KNN. This evaluation is used to identify the combination of text processing methods that can produce the highest sentiment classification accuracy which wouldthen be selected and used in this study.

Phase 2: Data Modelling

Topic modelling is an essential algorithm used in sentiment analysis. A topic model is a probabilistic model that discovers the main themes in a collection of documents. The basic idea is to treat the documents as mixtures of topics in the topic model, and each topic is viewed as a probability distribution of the words. When using a topic model as a sentiment analysis tool, each topic is viewed as a collection of words and each document can be viewed as a set of topics with different proportions depending on the frequency that terms appear. The topics are what the documents talk about. Intuitively a document is about a topic and the topic words tend to appear more often than other words [28]. For example, considering the themes in a collection of recipes in a cookbook, words like "sugar", "teaspoon", "oil" will appear more frequently, while in technical IT documents, words like "computer", "research", "algorithms" will appear more often. In addition, words like "is", "are", "a" will appear in all the documents, regardless of the themes, and the method of removing these kinds of words are called “stopwords”. “Stopword” is discussed in the section on data preprocessing.

3.1.4 Phase 3: Classification of Sentiment

Phase 3 entails the classification of sentiments. In this phase, the sentiment classification method will be implemented through the KNN algorithm, using a set of features selected in the feature selection phase. KNN, DT and NB.

3.1.5 Phase 4: Testing, Evaluation, and Analysis

Phase 4 consisted of testing, evaluating, and analyzing, which was carried out based on the results of sentiment classification. The classification results were tested based on the confusion matrix obtained from those results. The confusion matrix presents information pertaining to the actual number of a class and the number of predictions generated by the classification algorithm.

Three performance testing criteria were used, namely accuracy, precision, and retrieval. For comparison and evaluation of the algorithms used in this study, the tests that were carried out by several studies have been selected.

4.0 Data Collection (Sentiment Classification)

Data cleaning is one of the most important processes for obtaining accurate experimental results. To test our model, this work used the Twitter Sanders dataset [29]. In this work, we have used pre-labeled (with positive, negative and neutral opinion) tweets on particular topics for sentiment classification. Opinion score of each tweet is calculated using feature vectors. These opinion score is used to classify the tweets into positive, negative and neutral classes. Then using various machine learning classifiers the accuracy of predicted classification with respect to actual classification is being calculated and compared using supervised learning model.

The dataset was divided into a 80% (4000) training set and a 20% (1000) testing set. First, unnecessary columns were removed. The second process included some spelling corrections, removing weird spaces in the text, html tags, square brackets, and special characters represented in text and contraction. They were then handled with emoji by converting them to the appropriate meaning of their occurrence in the document.

4.1 Dataset

These are 5500 hand-classified tweets on 4 topics. These tweets are labeled as positive, negative, neutral and irrelevant. Among which 1786 irrelevant tweets are not considered in this work because they are irrelevant to the topic and they are not in English language. In this corpus, there are 570, 654, 2503, positive, negative and neutral tweets respectively.

4.2 Experimental Design

This section presents the topic modeling and experimental setup of sentiment polarity classification based on the Sandler dataset. We implemented and tested the proposed method in Mathlab on a PC with a 3.20-GHz CPU, 32 GB of RAM, and two Nvidia GeForce.

5.0 PREPROCESSING

Thereafter, we put all the text to lowercase, and removed text in square brackets, links, punctuation, and words containing numbers. Next, we removed stop words because having these makes our analysis less effective and confuses our algorithm. Subsequently, to reduce the vocabulary size and overcome the issue of data sparseness, stemming, lemmatization, and tokenization processes were applied. We normalized the text in the dataset to transform the text into a single canonical form. With the aim of achieving better document classification, we also performed count vectorization for the bag-of-words (BOW) model. The BOW model can be used to calculate various measures to characterize the text. For this calculation process, the term frequency-inverse document frequency (TF-IDF) is the best method. Basically, TF-IDF reflects the importance of a word [30].

5.1 FEATURE SELECTION USING DATA MODELLING

After preprocessing, the LDA model was applied. LDA is a three-level hierarchical model that creates probabilities at the word level, on the document level, and on the corpus level. Corpus level means that all documents exist in the dataset. Then, a model was developed for identifying unique words in the initial documents and the number of unique words after removing rare and common words. For the analysis of visualizing document relationships, the LDA developed model was applied at the corpus level, which was used for whole document visualization. The LDA application greatly reduced the dimensionality of the data

Hence, this score considers both false positives and false negatives. Intuitively, it is not as easy to understand as the accuracy, but F1 is more commonly used than accuracy. It can be seen that the deep learning models performed better in sentiment classification than SVM, which is a traditional method. The F1 score calculated using following Equation (12).

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Precision is defined as the number of true positives over the number of true positives plus the number of false positives. Recall is defined as the number of true positives over the number of true positives plus the number of false negative.

5.2 TOPIC MODELLING BASED ON SENTIMENT ANALYSIS CLASSIFIER MODEL

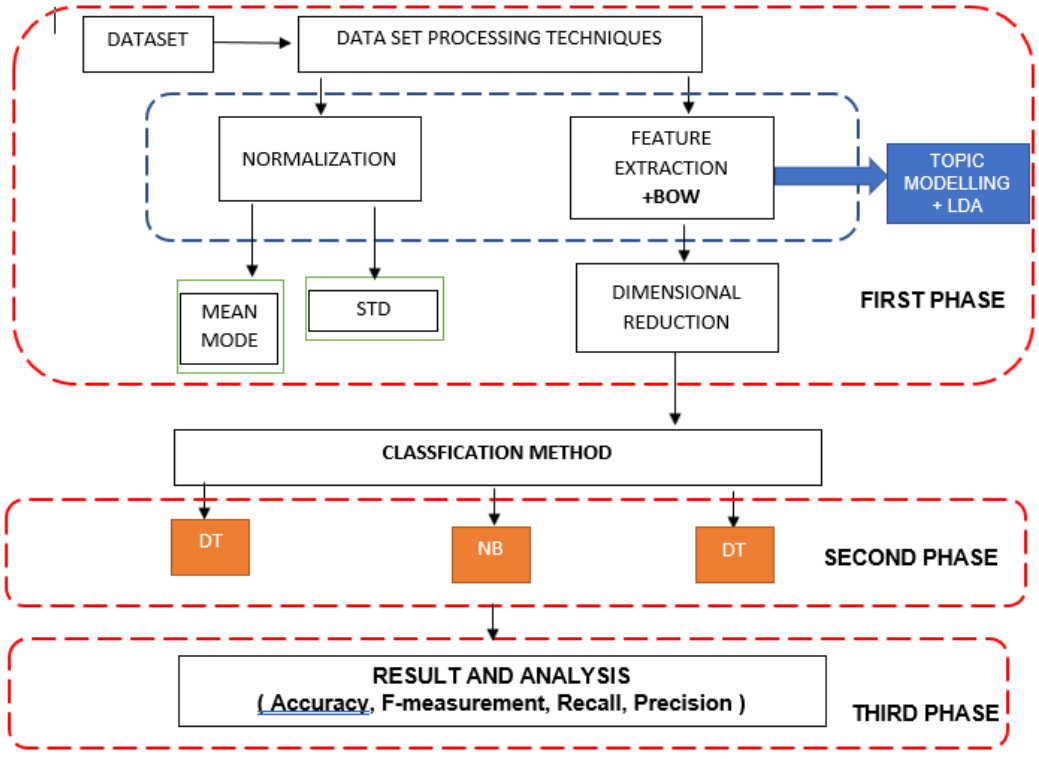


Figure 4: Experimental Design

5.2.1 FIRST PHASE

This research splits the data around 20%-80% between testing and training stages. Under supervised learning, the experiment splits a dataset into a training data and test data in assuming, however, the main objectives can conclude that research will conclude to use testing and validation sets (and you should conclude this), crafting them using train test split is easy; we split the entire dataset once, separating the training from the remaining data and then again to split the remaining data into testing and validation sets.

In a BoW, the dimensionality of word space may be enormous and the BoW reflects only the words of the original texts. In contrast, the most important thing people expect to know about a document is the themes rather than words. The aim of topic modelling is to improve the feature extraction and discover the themes that run through a corpus by analysing the words of the original texts. Latent Dirichlet Allocation(LDA) is an approach ,based on definite theorem

to capture significant inter as well as intra document statistical structure via mixing distributional assumes that document arise from multiple topics, a topic is defined as distribution over a vocabulary .Corpus is associated with predefined number of topics k, and each document in corpus contain these topics with different proportion.

Topic Modeling aims to learn these topics from data or corpus. In this phase , the research will use as variable model, prevalent in machine learning from decades. Generative model does not consider order of words in producing documents, so purely based on bag of words (BOW) approach. Topic modeling is to automatically discover the topics from a collection of documents, so in a data set, the K topic distribution must be learned through statistical inference.

Normalization process will use the evaluation metrics of statistics measures mean, mix and standard deviation for multi-class problems. Quantitative evaluation with standard metric on English data set are also done on both probabilistic and non-probabilistic (Latent Dirichlet allocation,Latent Semantic analysis) topic modelling method

5.2.2 SECOND PHASE

This research used supervised machine learning approach. Different machine learning classifiers have been used by us on our model to classify the tweets into their respective classes of sentiments. Following are the machine learning classifiers used in this study

5.2.2.1 Naive Bayes

The Naive Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is the probabilistic model and it permits us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems. One of main reason that this research used Bayesian classification because the alghorithm provides useful learning algorithms and past knowledge and observed data can be combined. It helps to provide a useful perspective for understanding and also evaluating many learning algorithms. This helps to determine exact probabilities for hypothesis and also it is robust to noise in input data.

To perform the classifier, it uses the concepts of mixture models. A mixture model is capable of establishing the probability of the component that it consists of Bayes theorem to perform as a probabilistic classifier. Another name that a naïve Bayes is known as simple Bayes or independence Bayes. The probability P is defined as follows:

$$P(m|n) = \frac{\mathrm{P(n|m)\,P(m)}}{\mathrm{P(n)}}$$

P(C | X) is posterior probability,
P(X | C) is likelihood,
P(C) is class prior probability,
P(X) is predictor prior probability.

5.2.2.2 Decision Tree

A DT essentially classifies the training data into sets. These sets are formed by branching on the attribute values of the examples in the training data. A perfect DT is constructed when all training examples at a node in the tree are under the same classification.

The decision tree construction process usually works in a top-down manner, by choosing an attribute test condition at each step that best splits the records Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. Each leaf represents class labels associated with the instance. Instances in the training set are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Starting from the root node of the tree, each node splits the instance space into two or more sub-spaces according to an attribute test condition. Then moving down the tree branch corresponding to the value of the attribute, a new node is created. This process is then repeated for the subtree rooted at the new node, until all records in the training set have been classified.

The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, Ck is as follows :

Step 1. If all the objects in **S** belong to the same class, for example $\mathbf{C}_i$, the decision tree for **S** consists of a leaf labelled with this class.

Step 2. Otherwise, let T be some test with possible outcomes O_1, O_2,…, O_n. Each object in **S** has one outcome for T so the test partitions **S** into subsets $\mathbf{S}_1$, $\mathbf{S}_2$,… $\mathbf{S}_n$ where each object in $\mathbf{S}_i$ has outcome O_i for T. **T** becomes the root of the decision tree and for each outcome O_i we build a subsidiary decision tree by invoking the same procedure recursively on the set $\mathbf{S}_i$.

5.2.2.3 KNN

The classification in this study uses the K nearest Neighbours method with feature selection. Feature selection in classification is expected to be more efficient by reducing the amount of data analyzed by identifying features which will then be processed based on the classifier model that has been gene rated from the training process. In this study, the data is divided into two parts, namely training data and testing data

The use of K- Nearest Neighbours aims to class if new objects based on learning data that is closest to the new object. The K - Nearest Neighbour algorithm technique is easy to implement. In this case, the amount of data or commonly referred to as the closest neighbour is determined by the user which is stated by k.

As for the steps in the K Nearest Neighbor method: 1) Calculates Euclidean distances Euclidean distance formula

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

Where d () is euclidean distance, is record ti i, record to j and data to r.2) Sort by Euclidean distance value, 3) Determine the k nearest classification record, and 4) The target output is the majority class.

5.2.4 THIRD PHASE

The training data will load to classifier model in this study run the classifier and after training the model already trained and send the testing data then make evaluation of the model using four comparison parameters namely – accuracy, recall, precision, and f-score and for both feature extraction methods namely for the analysis. then from the evaluation this

research can prove the implementation of model has improvement. Through this classification evaluation, it helps us understand the strengths and limitations of proposed models when making predictions in new situations, model performance is essential for machine learning.

The proposed model can be conceptually divided into three components: (1) a topic prevalence model, which controls how words are allocated to topics as a function of covariates and statistical analysis (2) a topical content model, which different machine learning classifiers has been used and (3) result and analysis evaluation .

6.0 RESULT ANALYSIS OF COMPARISON SENTIMENT CLASSIFIERS

This study has three types of results a) Statistical analysis b) Shape Visualization c) Comparison between three classifiers model. The (a) and (b) analysis description will be focus on the importance of the features that can be describe in graphical visualization while the (c) analysis result will prove the improvement on Feature Extraction using LDA Topic Modelling has significant impact on the classifier compare to the traditional Feature Extraction using BOW model.

Based on the Figure 5(a) shows the result of mean statistical result that show the first and the last has a higher peak where the middle feature has the lower peak values except column 28 and 27 .That means the highest value indicate. This feature analysis are calculated based on the mean of terms or features occurrences in the classes.

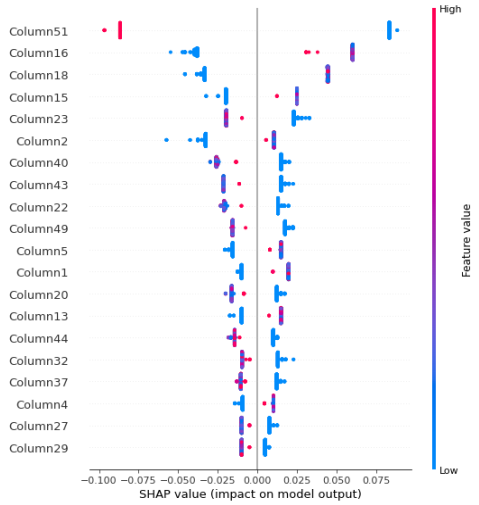
Figure 5(c) is a perfectly symmetrical distribution, the mean and the median are the same because result that show the first and the last has a higher peak where the middle feature has the lower peak values. This means that the distribution is bimodal and from this graph it shows that filtering out features that do not contain peaking values can help significantly prunes the amount of data to process and improves the quality of the results by removing noisy input.

Standard deviation (SD) is the most common way to present the variation of the features. In this graph distributions with relatively sharp peaks and long tails will have a high value of features then distributions with relatively flat peaks and short tails will have a low value of features. The highest peak was located in the middle of the graph distribution then low at the beginning and end of the features then it can be found that the graph distribution was opposite with the mean and mode. This is because the calculation for Standard Deviation indicates how accurately the mean represents the features of data and the features that has been used in this research has a higher standard deviation value to indicates greater spread in the feature of data.

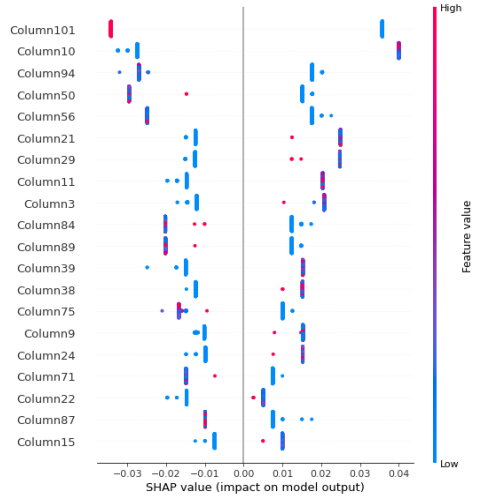
6.1 Statistical Analysis Of Features

. 6.2 Shape Visualization

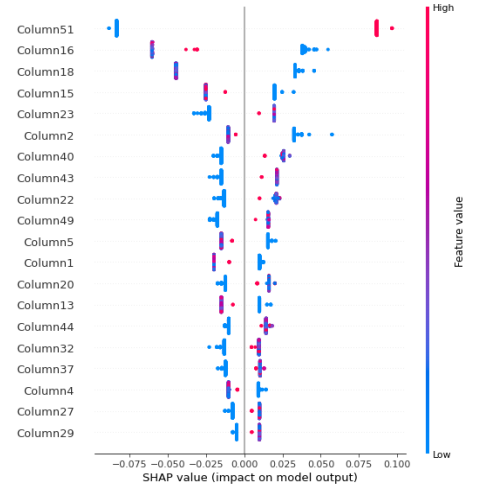
The summary plot below shows the top 2000 features based on their feature importance for the predictions. The SHAP value on the x-axis shows whether the feature effected a higher or lower prediction probability. Each dot represents a different test observation and the color of the dot is how important that feature was for that particular prediction



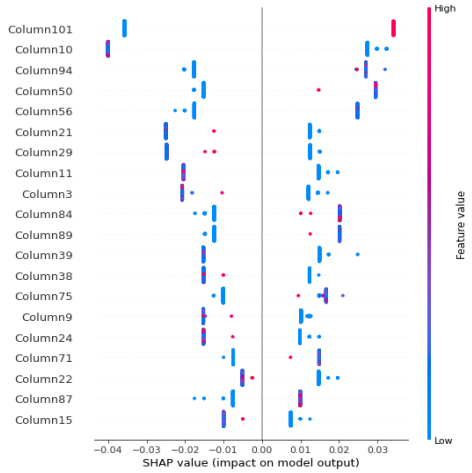
(a) SHAP Visualization English Dataset E50 Positive



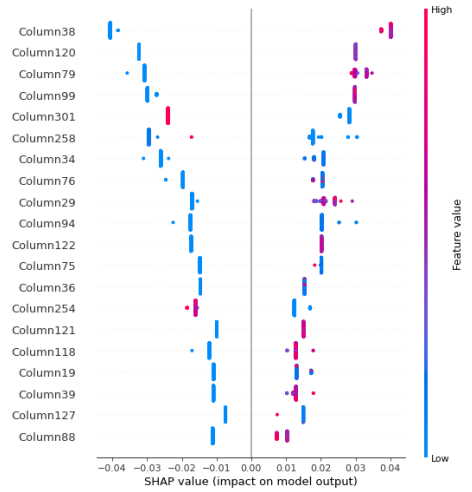
(b)SHAP Visualization English Dataset E50 Negative



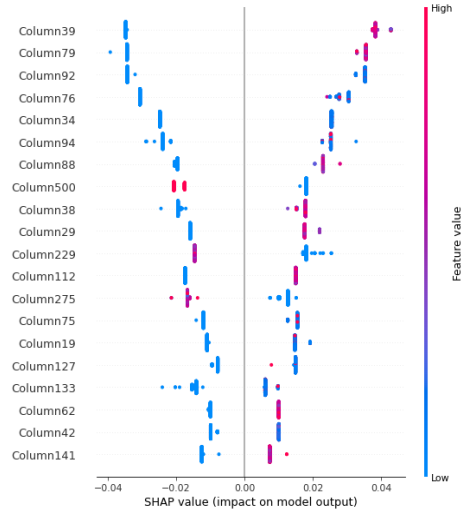
(c)SHAP Visualization English Dataset E100 Positive



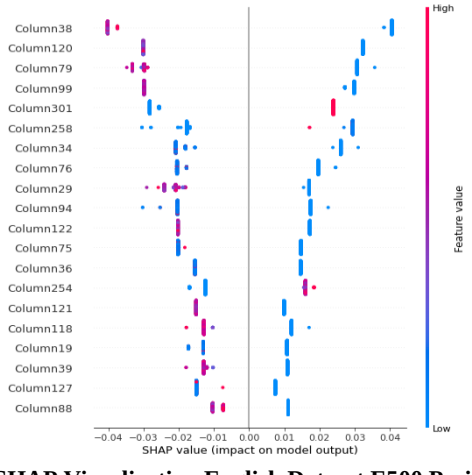
(d)SHAP Visualization English Dataset E100



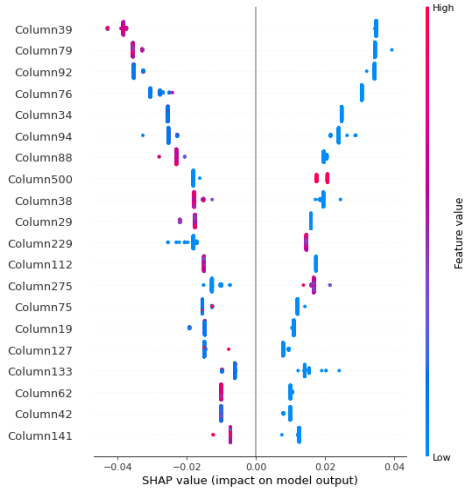
(e)SHAP Visualization English Dataset E300 Positive



(f)SHAPVisualization English Dataset E300 Negative



(g)SHAP Visualization English Dataset E500 Positive



(h)SHAP Visualization English Dataset E500 Negative

Figure 6 : SHAP Visualization on features of Microsoft in Sanders Data

However, this research measured shapley additive explanations (SHAP) values of various tokens to determine positive and negative sentiments more effectively. SHAP is a game theoretic technique to interpret the findings of any machine learning model. Therefore, the result of E50, E100, E300, E500 using Twitter Sanders dataset has been evaluated in each features and explored which tokens are responsible to classify positive and negative sentiments. Figure 6 shows the probability of SHAP values for different tokens in different eight clusters.

Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue. From the graph it shows that positive and negative features are well balanced or significant for each features in E50, 100, 300 and 500 respectively and this result prove the Sanders dataset can be the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on this research problem.

6.3 Comparison between three classifiers model.

The result of Table 1 shows accuracy , F Score , recall and precision of Decision Tree classifier based on 50 ,100 , 300 feature selection respectively . In addition, there is BOW as a traditional feature selection which present the lowest accuracy lower than the LDA feature selection. From this study, it was found that Feature E300, had the best accuracy, precision, and recall rates of 0.99 respectively, as shown in Tab. 1

Table 1 : Result of Decision Tree Classifier (Topic Modeling)

DT				
Feature	**Accuracy (100%)**	**F-measurement**	**Recall**	**Precision**
E50	99.4	0.994186	0.99	0.99
E100	97.9	0.978852	0.98	0.97
E300	0.999	0.999022	0.99	0.99
E500	82.5	0.839006	0.89	0.79
EBOW	57.1	0.447876	0.33	0.69

On the other hand , for the KNN classifiers E50 gives the highest result better than another feature 95.4 , 0.98 and 0.94 respectively. Once again, the result EBOW feature give the lowest accuracy compared to any features which improve by the Topic Modelling technique.

Table 2 : Result of KNN Classifier (Topic Modeling)

KNN				
Model	**Accuracy**	**F-measurement**	**Recall**	**Precision**
E50	95.4	0.957009	0.98	0.94
E100	84.5	0.847591	0.87	0.83
E300	47	0.539931	0.65	0.46
E500	50.6	0.288184	0.2	0.53
EBOW	56.1	0.699109	0.97	0.54

The Tab. 3. From this table, it is apparent that E500 and E300 obtained the highest classification accuracy of 0.99 respectively followed by E50 and E100 with the lowest value is EBOW. This small gap in accuracy difference indicates that the dataset used in this study balanced for classification of F- Measurement , it can be seen that only small gaps between the result value which are 0.001

Table 3 : Result of NBClassifier (Topic Modeling)

NB				
Model	**Accuracy**	**F-measurement**	**Recall**	**Precision**
E50	0.999	0.998981	0.99	0.99
E100	0.996	0.996016	0.99	0.99
E300	0.999	0.999022	0.99	0.99
E500	0.99	0.99	0.99	0.99
EBOW	0.636	0.630081	0.6	0.66

8.0 CONCLUSION

The importance of this research is that it is motivated by the need to increase the performance of text classifiers such as sentiment analysis and evaluate the effect of languages on sentiment analysis by tracking and detecting the weakness on text classifiers. Thus, the sentiment analysis that proposed in this thesis is anticipated to be able to identify the sentiment in English language using Text Classifiers in order to deal with them as soon as possible with high performance. The proposed of sentiment Classifiers aims to detect and identify the sentiments comments available in our social media with improved performance and this is important since currently, the methods of text classifiers being conducted on the events and topics of sentiments comments are still not satisfying. This work can be used to define the patterns of comments or detect the sentiments more effectively by using topic modelling as extraction. Finally, the significance of this study is the analysis in terms of sentiments comments and features was extracted from positive and negative tweets and identified high frequency information features transmitted and commented as the response to the result of improvement in classifiers.

REFERENCES

[1] Cambria E. Affective computing and sentiment analysis. IEEE Intell. Syst. 2016;31:102–107. doi: 10.1109/MIS.2016.31.

[2] Zhang H., Wheldon C., Dunn A.G., Tao C., Huo J., Zhang R., Prosperi M., Guo Y., Bian J. Mining twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J. Am. Med. Inform. Assoc. 2020;27:225–235

[3] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[4] M. Zhiwei, M.M. Singh, Z.F. Zaaba : Email spam detection: a method of metaclassifiers stacking : The 6th International Conference on Computing and Informatics (2017), pp. 750-757

[5] Kumar, H. K., & Harish, B. S. (2020). A new feature selection method for sentiment analysis in short text. Journal of Intelligent Systems, 29(1), 1122-1134.

[6] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.

[7] Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 353-362.

[8] Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. Journal of Hospitality Marketing & Management, 26(7), 675-693.

[9] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. Elektronika ir Elektrotechnika, 19(5), 67-72.

[10] Fidan, H. 2020. Grey Relational Classification of Consumers' Textual Evaluations in E-Commerce. Journal of theoretical and applied electronic commerce research 15(1): 48-65.

[11] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635.

[12] Schouten, K. (2018). Semantics-Driven Aspect-Based Sentiment Analysis (No. EPS-2018-453-LIS).

[13] Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 804-812).

[14] Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. Complex Adaptive Systems Modeling, 4(1), 1-19.

[15] Kamale, M. A., Ghode, M. S., Dhainje, P. B., & Moholkar, M. A. (2014). A Survey on Feature-Sentiment Classification Techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 2(12), 3972-3978.

[16] Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., & Hampapur, A. (2014). Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies, 45, 17-26.

[17] Fraj, M., Hajkacem, M. a. B. & Essoussi, N. 2019. Ensemble method for multi-view text clustering. *International Conference on Computational Collective Intelligence*, pp.219-231.

[18] Friedman, J. H. 2017. *The elements of statistical learning: Data mining, inference, and prediction.* springer open

[19] Froud, H., Benslimane, R., Lachkar, A. & Ouatik, S. A. 2010. Stemming and similarity measures for Arabic Documents Clustering. *2010 5th International Symposium On I/V Communications and Mobile Network*, pp.1-4

[20] Pathak, A. R., Pandey, M., & Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. Applied Soft Computing, 108, 107440.

[21] Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q. & Tian, G. 2019. Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems 61(2): 1123-1145.

[22] P. Hennig., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. International Journal of Hospitality Management, 55(2016), 16-24.

[23] Pathak, A. R., Pandey, M., & Rautaray, S. (2020). Adaptive framework for deep learning based dynamic and temporal topic modeling from big data. Recent Patents on Engineering, 14(3), 394-402.

[24] Pathak, A. R., Pandey, M., & Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. Applied Soft Computing, 108, 107440.

[25] Zhang, H., Chen, B., Cong, Y., Guo, D., Liu, H., & Zhou, M. (2020). Deep autoencoding topic model with scalable hybrid Bayesian inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12), 4306-4322.

[26] Alhawarat, M. & Hegazi, M. 2018. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. *IEEE Access* 6: 42740-42749.

[27] Alian, M. & Awajan, A. 2020. Factors affecting sentence similarity and paraphrasing identification. International Journal of Speech Technology 23(4): 851-859.

[28] Poria S., Majumder N., Hazarika D., Cambria E., Gelbukh A., Hussain A. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. IEEE Intell. Syst. 2018;33:17–25. doi: 10.1109/MIS.2018.2882362. conference Name: IEEE Intelligent Systems

[29] Boon-Itt S., Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis

and topic modeling study. JMIR Public Health Surveill. 2020

[30] Yaacob, N.M., Basari, A.S.H., Salahuddin, L., Ghani, M.K.A., Doheir, M., Elzamly, A. Electronic personalized health records [E-Phr] issues towards acceptance and adoption (2019) International Journal of Advanced Science and Technology, 28 (8), pp. 1-9.

[31] Doheir, M., Basari, A. H., Elzamly, A., Hussin, B., Yaacob, N., & Al-Shami, S. S. A. (2019). The new conceptual cloud computing modelling for improving healthcare management in health organizations. International Journal of Advanced Science and Technology, 28(1), 351-362.

[32] Doheir, M., Kadhim, A., Samah, K. A. F. A., Hussin, B., & Basari, A. S. H. (2014). Extension of NS2 framework for wireless sensor network. Advanced Science Letters, 20(10-12), 2097-2101. doi:10.1166/asl.2014.5638

[33] Figueiredo, E., Macedo, M., Siqueira, H. V., Santana Jr, C. J., Gokhale, A. & Bastos-Filho, C. J. 2019. Swarm intelligence for clustering—A systematic review with new perspectives on data mining. Engineering Applications of Artificial Intelligence 82: 313-329