aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 6, Sayı 2 | Kış 2022
Volume 6, No 2 | Winter 2022, 239-251

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

PRE-PROCESSING TECHNIQUES FOR MOVIE REVIEW SENTIMENT ANALYSIS: A COMPARATIVE STUDY FOR BEST FEATURE SET DETERMINATION

Sohayla E. ALY
Department Information Technology, Altınbaş University, İstanbul, Turkey,
Sohayla.elsayed234@gmail.com, ORCID: 0000-0003-1033-4461

Oğuz BAYAT
Department of Software Engineering, Altınbaş University, İstanbul, Turkey
oguz.bayat@altinbas.edu.tr RCID: 0000-0001-5988-8882

Adil D. DURU
Department of Coaching Education, Marmara University, İstanbul, Turkey
deniz.duru@marmara.edu.tr ORCID: 0000-0003-3014-9626

GELİŞ TARİHİ/RECEIVED DATE: 12.10.2022 KABUL TARİHİ/ACCEPTED DATE: 03.11.2022

Abstract

Sentiment analysis is considered as the process to extract the overall expression, opinions, or feelings from reviews about something such as products, services, or movies. A pre-processing is considered as a crucial phase in sentiment analysis for text mining because it allows us to analyse the reviews according to its intended meaning by removing all of the appendages which are the words that do not affect the semantic from sentences. And therefore, the number of features will decrease and thus accuracy will increase. Accordingly, we have decided to evaluate our experiment in identifying the best influencing technique of pre-processing for several features through making a comparison between the features and by combining them together to reach the best result based on features number for each pre-processing technique and classification accuracy. this comparison was done by using three algorithms for classification SVM, NB and DT after applying tools for feature selection and feature extraction with three techniques for tokenization. We concluded that there are some of these techniques that have a negative effect like lemmatization and the part of them is not due to any difference, other, which is a little part, have an effect

Keyword: Sentiment analysis, Pre-processing Techniques, Machine Learning Approach. Feature Selection.

FİLM DUYGU ANALİZİ İÇİN ÖN İŞLEME TEKNİKLERİ: EN İYİ ÖZELLİK SETİNİ BELİRLEMEK İÇİN KARŞILAŞTIRMALI BİR ÇALIŞMA

Özet

Duygu analizi, ürünler, hizmetler veya filmler gibi bir şey hakkındaki incelemelerden genel ifade, görüş veya duyguları çıkarma süreci olarak kabul edilir. Bir ön işleme, metin madenciliği için duygu analizinde çok önemli bir aşama olarak kabul edilir, çünkü semantiği etkilemeyen kelimelerin tüm eklerini cümlelerden çıkararak yorumları amaçlanan anlamına göre analiz etmemize izin verir. Bu nedenle öznitelik sayısı azalacak ve dolayısıyla doğruluk artacaktır. Buna göre, her bir ön işleme tekniği için özellik numarasına dayalı olarak en iyi sonuca ulaşmak, özellikler arasında bir karşılaştırma yaparak ve bunları bir araya getirerek birkaç özellik için en iyi ön işleme tekniğini belirlenilmiştir. Bu karşılaştırma, üç tokenleştirme tekniği ile özellik seçimi ve özellik çıkarımı için araçlar uygulandıktan sonra sınıflandırma için üç algoritma (SVM, NB ve DT) kullanılarak yapıldı. Bu tekniklerden lemmatizasyon gibi olumsuz etkisi olan bazılarının olduğu ve bunların bir kısmının herhangi bir farklılıktan kaynaklanmadığı, küçük bir kısmının etkisinin olduğu sonucuna varılmıştır.

Anahtar Kelime: Duyarlılık analizi, Ön İşleme Teknikleri, Makine Öğrenimi Yaklaşımı. Öznitelik Seçimi.

1. INTRODUCTION

Over the last period of time, sentiment analysis became the most important process in blogs and microblogging for finding information that indicates user's opinions about a product or a service. But comments are not truthful all the time. According to [Boschetti vd.], the noise can reach 40 %as a total percentage in a dataset, meaning that data does not contain any useful data for the analysis. Actually, it is considered as an unstructured data on the Internet —especially in our case on Movies Review— which have the most amounts of noise. User probably make mistakeson spelling and use some abbreviations and slag. But usually it is not essential for all terms tocontain benefit or a meaning. so sometimes can be ignored or replaced it to get more accuracy.

There are some studies denoting the effective role of pre-processing technique that contains two phases. The first phase is data cleansing phase which includes some tasks such as removing URL and removing punctuation and usernames, the second phase is considered as processing phase and sometimes is considered as transformation phase for text and this phase includes tasks like lemmatizing, normalization, stop words removal, and stemming in order to prepare the texts which was to be classified. pre-processing step is possible lead to weaknessin the total meaning of the sentence e.g. more of exclamation mark(!!!!) refer to surprise or suddenly according to [1] in this case the meaning indicate to strong emotion and during remove it is possible that it affected in the general meaning of the sentence so the accuracy is to be reduction.

The aims in this study benchmark to collect frequent pre-processing techniques from studies have been recover before accordingly (Symeonidis, Efffrosynidis and Arampatzis) and compare some steps of pre-processing with others and concatenate them together to examine them to test and find their influence on the number of features and test the accuracy of sentiment classification to reach the best way to get the least number of feature and the highest accuracy. We performed our experiment by different steps

involving feature extraction, feature selection and classification. Feature selection depends on a bag of word that were tokenized with three different types of split technique such as bigram, unigram and trigram in which term frequency and inversed document frequency (tf*idf) methods are being used, as a way to convert each document to avector that contains the weight for each feature in the document. The selected features are chosen according to the most used features in domain for Information Retrieval.

Feature extraction contain three category which are: filter, wrapper, and embedded methods based on (Parlar and Ozel) we apply Filter methods depending on statistical measures such as Chi square .

The final step of sentiment analysis text classification to determine the polarities of the reviewed document. Machine learning is one of the most commonly approaches which used in sentiment classification. Beside lexicon based and linguistic method (Pang and Lee), The most commonalgorithms in machine learning are Naïve Bayes (NB), Support Vector Machine (SVM) and decision Tree (DT). Sentiment classification depends on the classification model that was learned from the training.

In conclusion, our work aims to obtain the most suitable technique for sentiment analysis to reach to the best way to reduce the number of feature and that give high accuracy

This paper is planned as follows. Section II is a previous work on pre-processing techniques, Section III denotes the common pre-processing technique, Section IV explains methodologythat was used in the study of sentiment analysis, in this section we discussed some of thetechnique such as feature extraction and feature selection methods in sentiment analysis .Section V offers and discusses our experiment results. Finally, it provides our conclusions withthe suggestions for future work.

RELATED WORK

There are some various studies that have covered the pre-processing methods and had proposed in literature the extraction of unstructured data from text.

Some of them worked on pre-processing techniques separately and examined their impacts onaccuracy, however there were no focus on examining the impacts of the techniques when concatenating them together.

In (Krouska, Trousas, Viryou), some of the pre-processing techniques were applied on three different dataset in twitterand determined the performance for each dataset by using four different classification algorithms which are (NB, SVM, KNN, C4.5) to evaluate their impacts on classification accuracy, Afterword they determined the effect of pre-processing on classification for twitterdataset .In their experiment they tried to improve the accuracy so they applied feature selectiontechnique and display their work as bag of word(Bow) by using NGram Tokenization to compare word (unigram, bigram, trigram) and applied some of selection attribute. Finally, they reached that the sentiment analysis accuracy was improved and the performance of unigram and trigram became better compared to the other representations results and feature extractionimproved the classification accuracy in comparison with using all created

attributes. The weakness in this study was that their work did not go deeper in studying the best choice of algorithm to feature extraction such as Infogain and Chi square.

The orientation of this paper (Haddi, Liu, Shi) is to be about how to reduce the noise from the online textso we can apply some role of pre-processing techniques to prepare the text for classification On two datasets which are about movie reviews. pre-process steps have been split into threeprocesses, the first one denotes to data cleansing and processing which was contain severalsteps such as space removal, replace abbreviation, stemming, stop words removal and negationhandling. The second one was transformation such as feature selection which wasimplemented by using tf-idf technique. The third and the last process is called filtering in whichsome functions were applied to select the effective features in text such as chi-squared thatwas used to remove irrelevant features and finally we reach that chi-squared improved theclassification accuracy. They also used an SVM classifier to denotative to improve theclassifier's performance.

According to (Boschetti vd.) They display 15 pre-processing techniques and make a compression between them to examine each in them how can effect on number of features and accuracy. they apply it in Twitter datasets. And employed with three machine learning algorithms such as Linear SVC, Naïve Bayes and Logistic Regression. In conclusion they reach to that there were some of them affect as positivity such as stemming and remove number and replace repetition of punctuation, and some of them affect as negativity such as spelling correction and negation handling. So, we take this paper as a big reference for us but we addition some of technique to improvement the sentime classification such as feature selection like tf-idf and feature extraction to reduction the dimensionality of matrix

They apply some of Pre-processing techniques that are stopword removal, lowercaseconversion, and stemming to know their impact on various of aspect such as classification accuracy , text domain, text language ,and dimension reduction were also explored by (Uysal and Gunal) . theyapplied their experiment on two domain such as e-mails and news with two language Turkish and English. And evaluated it with micro-F1 score by using SVM classifier. They show that the researcher must tend to try all possible combination to reach to the best result of accuracy

2. FREQUENT TECHNIQUES OF PRE-PROCESSING

The 8 pre-processing steps which were applied will be defined in detail. Their arrangement in the implementation is very important so we are going to present them by the recommended order. We will display it briefly to explain some information about each technique such as thebenefit of applying the technique and we will give examples for each one and we will mentionrelated situations that used it before and the correct arrangement that it must be in.

2.1. Remove Url and Usermention

On social online text, most sentences contain a URL but specially in movie's review it is rare,but we remove it because it doesn't indicate any sentiment if there are any in the review.

When a user mention an actor name, film name and/or a hashtag symbol it does not add any addition or change in the meaning of the sentence because it does not denote any emotions or sentiments, Although it does not express any sentiment, it is sometimes used to know what thesubject of a text is about, is it a person , a place or an event, for example, @earthquake, @Obama. Some research prefer to replace user-mention with tags "AT_USER" as, e.g [15] ,but this case is far from our study. In our work, we remove the URL tag if found and we also removed user mention and hashtag symbols. This technique is almost use in all of the texts. So,it should be done before any other technique.

2.2. Replace Abbreviation

Recently, the use of blogs or microblogs have become more common that it has led to a lot ofunstructured data and their writing contain a great deal of abbreviation. As this abbreviation are frequently used to refer to sentence or phrases. Therefore, it must be replaced to understandits meaning correctly. We manually used a list that contains 1539 words and phrases, and all of them have their replacements. We present Some examples like "BTW", "CYA" and "LYLC", some of them indicate to sentiment like "LYLC" refer to "love you like crazy"

And the other are neutral so that "BTW" refer to "by the way" and" CYA" meaning "see you".So this step must be applied first before removing any noise data .This study [10]used this technique for Sentiment classification and analysis

2.3. Remove Stopword

Stopwords are works on words with high frequencies in sentences like (of, is are). In our case,we used movie review as dataset so it was possible that we would see "movie" word very frequently, therefor, in our case we considered "movie" word as a Stopword.

we considered it unnecessary to examine them, because they don't contain a lot of valuable data for Sentiment classification. therefore, it is better to remove them to reduce the number offeatures. the usual stopwords provided by [11]

2.4. Stemming

it is the used method to remove the ending of a word with the purpose of spoting their root form or stem. Because of this the dimensionality will be reduced by merging many similar words with each other. In our work use porter stemmer that was provided by [12] as a method that usually produces good results.

For example: stemming for root word "like" includes:(likes, liked, liking)

2.5. Lemmatizing

Another method to reduce the number of features was to analyze a word and remove its ending. Accordingly, producing its base form or lemma as it is found in the dictionary. Lemmatizationis also a method of integrating many words into one. In our work the wordNetLemmatizer is used

For example:'Caring' -> lemmatization ->care

2.6. Replacing repetition of punctuation

We have identified three punctuation signs, which are the exclamation marks, question marks and stop marks. The presence of these sign indicates extreme emotions and has an effect on the total meaning of a sentence. So, in our work we replaced those marks with its the suitable expressions. for example, exclamation marks "!!!" were replace with the word "shocked". Whereas, possible it may denote negative or positive emotion depending on the sentence meaning, also question marks "???" sometimes expresses negative emotions. This technique must be applied before removing punctuation technique. In study of [14], this technique is used to organize the language and simplify the vocabulary to represent feeling.

2.7. Remove punctuation

One of the most used symbols as it is used to split the document into sentence or phrases. So, in NLP, the punctuation-removal considered as the most important step in pre-processing step. Therefore, we preferred to remove it because it does not refer to any sentiment, for example: How we will arrive to the garden? In this example the punctuation doesn't carry any expression or emotion so in our work we remove it as a way of data cleansing.

3. MOTHODOLGY:

We implemented our experiment on movie review which have been published in1. it consists of the labels which divided to positive, and negative. We chose to examine the comparison between some pre-processing steps to study their impact on the number of feature and classification accuracy by implementing tf-idf and chi-squared to extract and select the features and we used three different classification algorithms. with the predefined classes of positive, negative.

3.1. Pre-processing technique

Table 1: Applied technique for pre-processing

Technique 1	Without pre-processing
Technique 2	Data cleansing (remove URL, remove punctuation, and remove user-mention)
Technique 3	Data cleansing, Stemming, Stopword Removal
Technique 4	Data cleansing, Stemming, Stopword Removal, Replace abbreviation and repeat punctuation
Technique 5	Data cleansing, Lemmatizing, Stopword Removal, Replace abbreviation and repeat punctuation
Technique 6	Data cleansing, Stemming, Replace abbreviation and repeat punctuation

3.2. Movie review data set

This corpus benchmark to movie review. which contains 2000 reviews: divided into 1000positive and 1000 negative review. it used in[16]

3.3. Tokenization

As an preliminary stage in pre-processing, most of the researches consider applying tokenization , e.g. (Krouska, Troussas, Viryou)

Tokenization by (Basile vd) is a task which split the sentence to words, based on the technique that was used if it was to be unigram ,bigram or trigram. (Atkinson, Salas and Figueroa) defined tokenization as a type of lexical test which divided content into words, expressions, or other important components called tokens. Basically, the tokenization process is a usual technique for Natural Language Processing (NLP)

3.4. Feature selection and extraction

There are some ways to evaluate the features in a bag-of-words representation. We chose to use Term Frequency-Inverse Document Frequency (TF.IDF) as a technique for feature selection that was improved in () , it presented by TF.IDF = $f \log(N/df)$, where f is the numberof occurrences in the document, N is the number of documents, and df is the number of documents that contain this feature (Na vd.), we used Term Co-occurrence (unigrams, bigrams or tri-grams) to recognize the data

Feature extraction: the selection of suitable features to reduce the problem of dimensionality.This can be done by using the following methods:

Based on [19] chi (χ^2) method that achieves a high performance and accuracy,

chi squared (χ^2) statistic is a **test** that measures how expectations compare to actualobservational data (or model results).

3.5. Algorithms of machine learning approach

Three classification algorithms were applied in our work, were to be as Decision Tree (DT),the Support Vector Machine (SVM) and Naïve Bayes (NB).

3.5.1. Naïve Bayes

The algorithm which consider as simplest algorithm to probability classification. It uses BayesTheorem to predict the probability that a certain feature set belongs to a particular label.

$$p(label|\text{features}) = \frac{\text{p(label)} * \text{p(features |label)}}{\text{p(features)}} \quad (6)$$

P(label) is the Previous probability of a label. P (features| label) is the Previous probability to classify a special feature set as a label. P(features) is the Previous probability that a given feature set happening.

3.5.2. Support vector Machine

SVM classification algorithm consider as most popular machine learning algorithms for linearproblem and consider as the simplest and faster algorithm

3.5.3. Decision Tree Classification Algorithm

Decision tree may be a graph with branches representing each result that can be reached through a decision. The details created by the call tree model that are human clear, and the call is simply explained.

4. FRAMEWORK

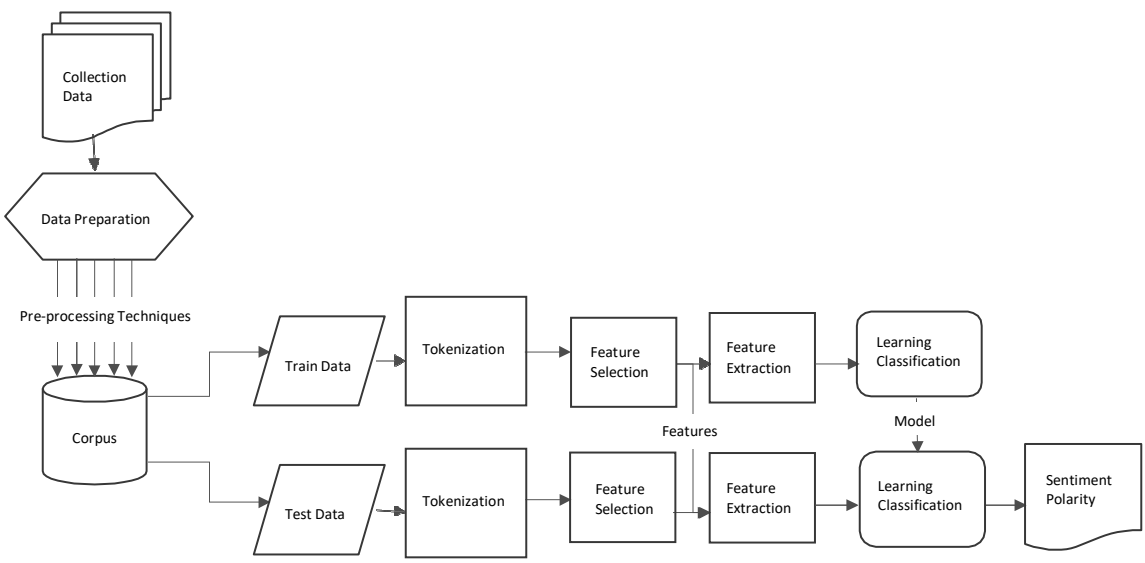

Figure 1: Sentiment analysis framework

5. DISCUSSION AND RESULT

After describing the specific pre-processing techniques that was used, machine learningalgorithms and evaluation measures, we present and discuss the results of the experiment on the dataset to see the real impact of text pre-processing steps on number of features and accuracy. Consider Table 1 that shows all techniques of pre-processing which we used. Next, we discuss the best way and worst way. For each algorithm in order to reach the smallest number of features by three type of tokenization is shown in Table 2.

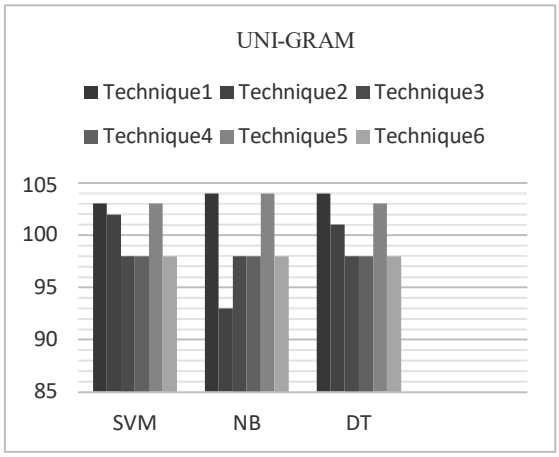

Figure2: Comparison the number of features for apply technique by three algorithms with unigram

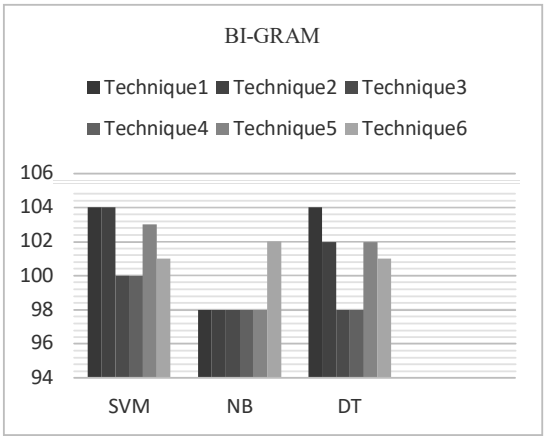

Figure3: Comparison the number of features for apply technique by three algorithms with bigram

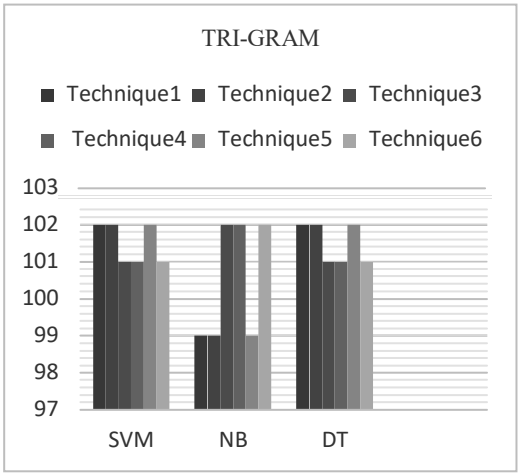

Figure4: Comparison the number of features for technique which applied by three algorithms with trigram

We have concluded that SVM and DT algorithm are closer to each other in their result, using stemming is the better than lemmatizing with this algorithm in order to reach the smallest number of features, as shown by using technique 4 and 5, and stopword-removal it was affectednegatively when removed as shown by using technique 4,6.

Table 2: worst and best technique for each algorithm

	Worst	Best
SVM	Technique 1,2	Technique 3,4
NB	Technique 6	Technique 2
DT	Technique 1,2	Technique 3,4

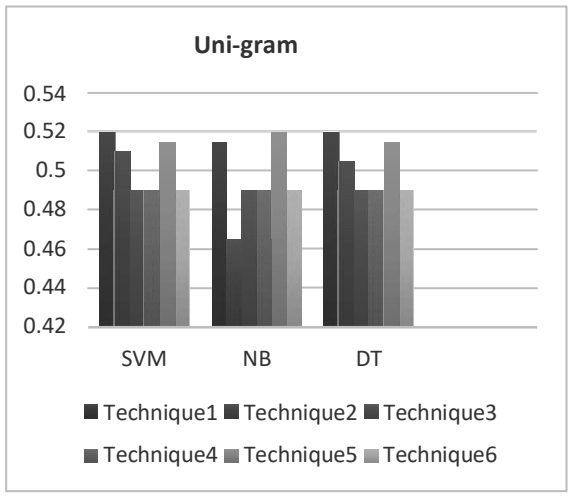

Figure5: Accuracy percentage for pre-processing techniques which applied in all three machinelearning algorithms by using unigram on movie datasets

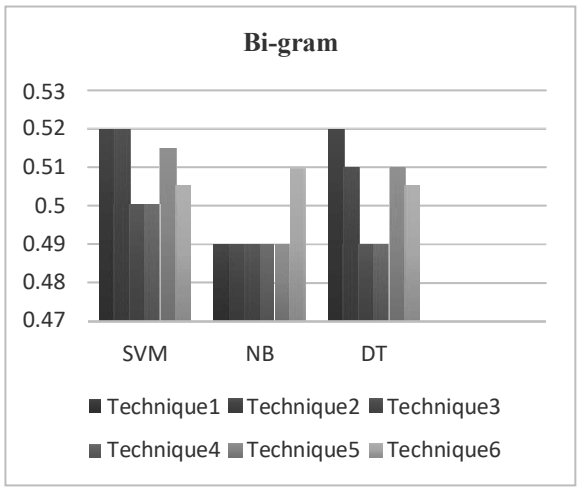

Figure6: Accuracy percentage for pre-processing techniques which applied in all three machinelearning algorithms by using bigram on movie datasets

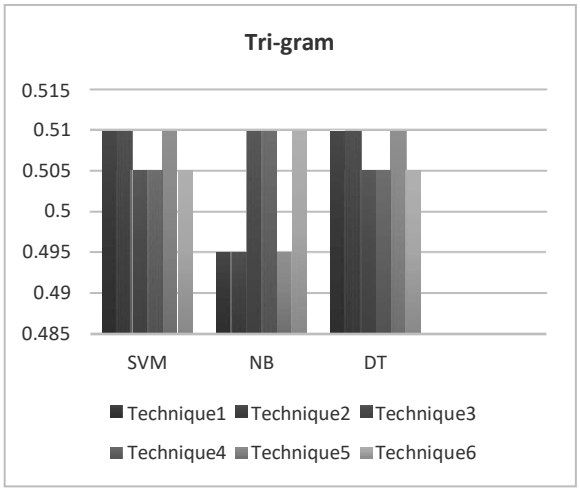

Figure7: Accuracy percentage for pre-processing techniques which applied in all three machinelearning algorithms by using trigram on movie datasets

5 pre-processing techniques for the three-algorithm that was obtainable from Table 1. From our experiment we noticed that there is no significant correlation between accuracy and numberof features. technique1 and technique 2 led to increase the number of features with SVM andDT and technique 6 with DT in addition to there are high performance achieved. The techniquesthat greatly reduce feature such as 3 and 4, provided better results of accuracy.

Next, we discuss the average for f-measure of each of the techniques. These numbers are presented in Table 3.

Table 3: F-measure for each technique in three different type of tokenization

	Unigram						Bigram						Trigram					
SVM	Teq1	Teq2	Teq3	Teq4	Teq5	Teq6	Teq1	Teq2	Teq3	Teq4	Teq5	Teq6	Teq1	Teq2	Teq3	Teq4	Teq5	Teq6
	0.40	0.40	0.33	0.33	0.40	0.33	0.38	0.38	0.35	0.35	0.37	0.36	0.36	0.36	0.35	0.35	0.36	0.35
NB	0.37	0.32	0.33	0.33	0.38	0.33	0.33	0.33	0.33	0.33	0.33	0.36	0.33	0.33	0.36	0.36	0.33	0.36
DT	0.40	0.39	0.33	0.33	0.40	0.33	0.38	0.36	0.33	0.33	0.36	0.36	0.36	0.36	0.35	0.35	0.36	0.35

6. CONCLUTION

Our study covered what is the effect of different pre-processing techniques on the number of feature and accuracy. The framework contains of two stages, The first stage is to be pre- processing data which divided to two phases, the first to be a data cleansing including remove punctuation, remove user name and mention and remove URL. The second phase was appliedseveral text processing techniques including: replacing the abbreviation words and repeat punctuation, removing stopwords, lemmatizing, and stemming.

The second stage text classification using Support Vector Machine (SVM), Decision Tree (DT) classifiers and Naïve Bayes (NB) by applying three type of n -gram (unigram, bigram, trigram)after feature selection approaches, (TD-IDF) and feature extraction (chi squared). We have been estimated the classifier's

performance by calculate the average of precision, recall and F-score. The text processing techniques were all applied on movie review data set.

In Our works show that, In the sentiment analysis there was some of techniques that was provided better results of classification for movie review dataset that was used. The recommended techniques are technique 3 which contain Data cleansing, Stemming, StopwordRemoval, and technique 4 which include Data cleansing, Stemming, Stopword Removal, Replace abbreviation and repeat punctuation. The non- recommended techniques 1 which is no pre-processing and technique 5 which indicates to Data cleansing, lemmatizing, stopword removal, replace abbreviation and repeat punctuation. in conclusion there was some of pre- processing techniques are an effective for sentiment classification. And also according to special experiment we noticed that there is no significant correlation between accuracy and number of features

7. FUTURE WORK

In the future we plan to test the more machine learning algorithms with different feature selection methods. We will make effort to incorporate more of these techniques to achieved better results. Moreover, another future aim will to be a test those techniques on different datasets dataset such as twitter and make more of comparison between reminding techniques

REFERENCE

Abbasi, A., France, S. Zhang, Z. and Chen, H. "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," vol. 23, no. 3, pp. 447–462, 2011.

Uysal A.K. and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.

Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," *IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl.*, 2016.

Pang B. and L. Lee. "Sentiment analysis using subjectivity summarization," 2004.

by combining features-based coreferencing and memory-based learning q," *Inf. Sci. (Ny).*, vol. 299, no. 1130035, pp. 20–31, 2015.

Guzman E. and W. Maalej. "How do users like this feature? A fine grained sentiment analysis of App reviews," *2014 IEEE 22nd Int. Requir. Eng. Conf. RE 2014 - Proc.*, pp. 153–162, 2014.

Haddi, E. X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *First Int. Conf. Inf. Technol. Quant. Manag.*, vol. 17, pp. 26–32, 2013.

Loper E. and S. Bird, "NLTK: The Natural Language Toolkit," 2002.

Boschetti, F., M. Romanello, A. Babeu, D. Bamman, and G. Crane, "Research and Advanced Technology for Digital Libraries," *Res. Adv. Technol. Digit. Libr.*, vol. 5714, no. September, pp. 156–167, 2009.

Paltoglou, G. "Sentiment Analysis in Social Media," no. June, pp. 3–17, 2014.

Atkinson, J., G. Salas, and A. Figueroa, "Improving opinion retrieval in social media, Technology, "An improved TF-IDF approach for text classification," vol. 4, no. 60082003, pp. 49–55, 2005.

Na, J., H. Sui, C. Khoo, S. Chan, and Y. Zhou, "ISKO 04- Effectiveness of Simple Linguistic Processing in Automatic.pdf," pp. 49–54, 2004.

Porter, M.F. "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130– 137, 1980.

Basile, P., V. Basile, M. Nissim, N. Novielli, and V. Patti, "Encyclopedia of Social Network Analysis and Mining," *Encycl. Soc. Netw. Anal. Min.*, no. January, 2017.

Nakov, P. "Semantic Sentiment Analysis of Twitter Data," no. June, pp. 30–38, 2017.

Symeonidis, S., D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre- processing techniques and their interactions for twitter sentiment analysis," *Expert Syst. Appl.*, vol. 110, pp. 298–310, 2018.

Parlar T. and S. A. Ozel, "A new feature selection method for sentiment analysis of Turkish reviews," *Proc. 2016 Int. Symp. Innov. Intell. Syst. Appl. INISTA 2016*, no. December 2017, 2016.

T. Parlar, "A New Feature Selection Method for Sentiment Analysis of Turkish Reviews," no. December 2017, 2016.

Wu, W., B. Zhang, and M. Ostendorf, "Automatic generation of personalized annotation tags for Twitter users," *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, pp. 689–692, 2010.