International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-11, September 2020

Text Preprocessing Method on Twitter Sentiment Analysis using Machine Learning

Jenifer Mahilraj,Getahun Tigistu,Sisay Tumsa

Abstract*: In real world, twitter sentimental analysis (TSA) acting a major role in observing the public opinion about customer side. TSA is complex compared to general sentiment analysis due to pre-processing of text on Twitter. The maximum limit on the number of characters allowed on Twitter is 280. In this article we discuss the influence of the text pre-processing technique on the classification efficiency of emotions in two kinds of classification problems and summarize the classification efficiency of the four pre-processing methods. This paper contributes to the consumer satisfaction classification sentiment analysis and is useful in evaluating the details in the context of the amount of tweets where views are somewhat unstructured and are either positive or negative, or somewhere in between. We first pre-processed the dataset, then extracted the adjective from the dataset with some meaning called the feature vector, then selected the feature vector list and subsequently applied machine learning based classification algorithms namely: Naive Bayes, Random Forest and SVM along with WordNet based Semantic Orientation which extracts synonyms and similarity for the features of content. Experiments display that the accuracy (Acc) and average F1-measure (F1-M) of the classification classifier on Twitter are enhanced by using methods of pre-processing the extension of acronyms and swapping negation, but barely deleting numbers or stop words.

Keywords : Classification Efficiency,Data mining, Deep learning,Sentimental analysis.

I. INTRODUCTION

Today, the era of the Internet has altered the way people precise their views and opinions. Nowadays this is mainly done via blog posts, product review websites, online forums, social networks etc. Today millions of people use social networks like Facebook, Twitter, and Google Plus etc. to share their feelings, opinions and views about your daily life. A life. We receive interactive media through online communities, where customers inform and inspiration others through forums. Social networks generate a large amount of atmospheric data in the form of tweets, blog posts, comments, reviews, etc.

Revised Manuscript Received on September 30, 2020.
***** Correspondence Author
Jenifer Mahilraj*, Faculty of Computing and Software engineering, Amit, Arbaminch university, Arbaminch, Ethiopia Email:jenifer.mahilraj@amu.edu.et,jenimetil@gmail.com
Getahun Tigistu,, Faculty of Computing and Software engineering, Amit, Arbaminch university, Arbaminch, Ethiopia Email:getahun.tigistu@amu.edu.et
SisayTumsa,,Faculty of Computing and Software engineering, Amit, Arbaminch university, Arbaminch, Ethiopia Email: sisay.tumsa@amu.edu.et

In addition, social network companies offer the opportunity to contact their customers for advertising. People mostly rely on user-generated content over the Internet to make decisions. For example, if someone needs to buy a product or use a service, they first look at their reviews on the Internet and discuss them on social networks before making a decision.

The amount of content made by users is too large for the average user to analyse. It is therefore necessary to automate this. Various mood study systems are widely used. A sentimental analysis (SA) informs the user before the purchase whether the product information is satisfactory. Vendors and companies use their analytics data to recognize their products and services so that they can be offered according to the needs of the user. Methods of extracting text information primarily focus on processing, finding, or analysing existing facts. The facts have an objective element, but there is another textual content that expresses subjective features. These materials are primarily opinions, feelings, perceptions, attitudes and feelings that are a core part of SA. This opens up many complex opportunities for developing new applications, mainly due to the huge growth in information available on online sources such as blogs and social networks. For example, the commendations of the components projected by the system can be considered using SA considering positive or negative opinions about these factors. SA can be defined as a process that systematizes the extraction of attitudes, opinions, and feelings from text, language, tweets and database sources through the Natural Language Process (NLP). In sensory analysis, opinions in the text are divided into categories such as "positive" or "negative" or "neutral". It is also mentioned to as the essence of subjectivity analysis, dismantling and evaluation. Opinions, feelings, attitudes and beliefs are interchangeable, but there are differences.

- Opinion: A assumption open to dispute
- View: subjective attitude
- Belief: considered acceptance and intellectual assent
- Sentiment: opinion signifying one's feelings

II. RELATED WORK

OrenEtzioni et al. [1] had recommended that the customers are often enforced to wade through many online reviews to make an informed product choice. This document introduces OPINE, an automated information retrieval system that removes reviews to model the features of the main products, its reviewer ratings and the corresponding quality in the products.

Compared to past work, OPINE achieves 22% higher accuracy and only 3% less feedback for function extraction work. New use of open punctuation marks to find opinions and their polarity.

Jingjing Liu [6] had displayed that this paper presents a parse-and-Para parse paradigm to assess the degrees of sentiment for product reviews. Cognitive recognition is well understood; However, the previous work only provides positive and negative values for most binary poles, and the polarity of emotions changes slightly when a deviation is found. Pain due to cellular characteristics such as Unigram / Bigram also complicate the classification of emotions, since language structures such as innate distance dependence are often ignored. In this article, we propose an adverb-adjective-noun sentence extraction approach that is based on the sentence structure by dividing a hierarchical representation by dividing a phrase. We also offer a general solution to simulate the contribution of adverbs and reject the assessment of sensation. Applying the pros and cons based on aspects of restaurant reviews, we were able to achieve a relative 45% improvement in recall with analytics and improved clarity. ChenliangL [3] had recommended that many public or private organisations have been described to generate and monitor targeted twitter streams to gather and understand user's opinion about the organisation. A Twitter target stream is usually created by filtering tweets with user choice criteria, for example, tweets posted by users in a nominated region or tweets matching one or more predefined keywords. It then tracks the targeted Twitter flow to gather and understand user opinions about the organization. There is an urgent need to identify and respond to the crisis with such a targeted flow. Such an application requires a designated object detection system for Twitter that can automatically detect emerging nomenclature that may be related to the crisis. In this article, we are introducing a new, automated two-step system for finding objects named for the Twitter stream called TwiNER. In the first phase, global contexts from Wikipedia and Web N-Gram Corpus are used to split tweets into valid section phrases using dynamic programming algorithms. Each such tweet area is a nominated candidate. It turns out that the assigned objects in the target stream typically display green assets and how the target stream is created. In the second step, TwiNER creates a random working model to check the quality of communication in the local context received from the Twitter stream. High-level sections are likely to be actual designated objects. We rated TwiNER on two sets of real tweets that mimic the target flow.

TwiNER is evaluated on the basis of basic truths and comparative efficiency is achieved in both threads as in the traditional method. Various TwiNER methods were also tested to test our idea of combining global contexts and local contexts. Minyi Guo [7] had exhibited that the topic in recent years. The aim of this work is to clarify the views or opinions on tweets that are seriously understood as a problem of text classification based on machine learning. Some methods use manually marked data to train fully serviced models, while some models use noisy tags such as emoticons and hash tags for training. In general, we can retrieve a limited amount of training data for fully serviced models because manually marking tweets is very time consuming and time consuming. For models with noisy markings, it is difficult to work satisfactorily due to noise on the markings, but it is easy to obtain a large amount of training data. Therefore, the best strategy for learning is to use both manually marked data and noise marked data. How easy it is to integrate these two different types of data into the same learning environment is still a problem. In this article, we introduce a new model, the Emoticons Smooth Language Model, to solve this problem. The basic idea is to train language models based on manually marked data and then use noisy emotional data to smooth them out. Experiments with real data sets show that soft emoticon language models can efficiently integrate data from both by using only one of these methods to overtake these methods. Avirupsil [2] had recommended that recognizing and associating names with structured data is a fundamental part of test analysis. Existing approaches typically take these two steps using a pipelined architecture: they use a named entity recognition system to determine the boundaries of mentions in the text, and entity binding systems for linking mentions to records are semi-structured or structured repositories such as Wikipedia. However, the tasks are interrelated and each scheme can greatly benefit from the information provided by the others. We offer a general ideal for a nominal entity identification system and an existing communication system that uses a large number of reference elements from a candidate in a particular nominal existence identification system and a large number of candidate component relationships from a system of reference candidate candidates. Organizes together to make general predictions. A system for identifying named entities and a system for linking existing elements, experiments on three data sets, a system for identifying named entities significantly exceeds or approaches the features of a modern system for identifying named entities, surpassing references to six competing entities. This test name object identification system and object binding system offer 60% error reduction compared to the next named object search system 68% error lessening compared to the next best object communication system..

III. PROPOSED WORK

This is an important task, since it will clean the dataset by reducing its complexity. This is the important task in the sentiment analysis because it will neglect the unnecessary phrases in the tweets, for this reason, this pre-processing is also called as a CLEANING PROCESS.

- Reduce the token
- Remove stop words
- Remove meaningless words
- Remove hash tag
- Remove punctuation
- Remove URL Corpus
- Replacing negative mention
- Reverting words
- Removing emoji's

A. Model Building:N-Grams (N-G)

N-G is a sequence of N elements in a given text or speech template, N-G is usually composed of a text or speech corpus. N-G Language Model: Use the previous N-1 words in a sequence to guess the next word. N-G Types: Unigram, Bigram, Trigram, etc.…

IV. CLASSIFICATION

The evaluation by analysing the target class in a classification is a training data set. This can be achieved by finding the right limits for each target class. In general, training data sets are used to obtain optimal boundary conditions with which each target class can be determined.

A. Support Vector Machine (SVM)

SVM is a types of machine learning model based on the idea of classifying data with large fields.

This tool has a solid theoretical foundation and the classification procedures based on it provide good generalization presentation.

Due to good classification accuracy, standard implementations are slow and not easily scalable.

Therefore, they cannot be applied to large data mining requests.

Usually you need a huge sum of support vectors. Thus, the training and classification time is longer.

Input: Input data matrix, class info

Output:

- Set of Basis vectors

- Start
- Repeat
- For every candidate sample – samples not in current set of BVs
- Include it in the model efficiently. Observe the generalization presentation on the residual points.
- End for candidate samples
- Add that point to the BVs list that gave better test error.
- Till the ending criterion end.

B. Naive Bayes

Rev. Thomas Bayes after named as Bayes theorem. Works on conditional probability. Conditional probability is the probability that something is occurring because something else has already occurred. Using conditional probability, we can compute the probability using prior knowledge of the event.

The conditional probability is calculated by using below formula,

$$P(H/E) = \frac{P\left(\frac{E}{H}\right) * P(H)}{P(E)} \quad (1)$$

Where,

- P represent as probability.
- H represent as hypothesis.
- E represent as evidence.
- P (H) is the P of hypothesis H being true. This is known as the prior P.
- P (E) is the P of the evidence regardless of the H.
- P (E/H) is the P of the evidence given that H is true.
- P (H/E) is the P of the hypothesis given that the E is there.

C. Random Forest (RF)

This is an integrated classifier consisting of several decision-making trees, and each tree produces a class with a mode of class output. This is one of the most accurate learning algorithms available.

Random selection of features to create a collection of decision-making plants with controlled variation.

Pseudo code of the proposed Methodology

Input: Labeled Dataset

Output: positive and negative polarity with synonym of words and similarity between words

Step-1 Pre-Processing the tweets:

Pre-processing ()
Remove URL:
Remove special symbols
Convert to lower:

Step-2 Get the Feature Vector List:

For w in words:
Replace two or more words
Strip:
If (w in stopwords) Continue
Else:
Append the file
Return feature vector

Step-3 Extract Features from Feature Vector List:

For word in feature list
Features=word in tweets_words
Return features

Step-4 Combine Pre-Processing Dataset and Feature Vector List

Pre-processed file=path name of the file
Stopwords=file path name
Feature Vector List=file path of feature vector list

Step-5 Training the step 4 Apply classifiers classes

Step-6 Find Synonym and Similarity of the Feature Vector

For every sentences in feature list
Extract feature vector in the tweets ()
For each Feature Vector: x
For each Feature Vector: y
Find the similarity(x, y)
If (similarity>threshold)
Match found
Feature Vector: x= Feature Vector: y
Classify (x, y)

Print: sentiment polarity with similar feature words

D. Generates Decision trees.

The construction of each tree is based on the following steps. Let the number of training cases in the classification be n, and the number of variables be m. The number of input variables used to determine the solution by the node of the tree m; I should be less than m. Choose the training time for this plant and select the boot sample to choose, replacing n available in all n cases. Rate them using the remaining event class to evaluate the tree error. For each node in the tree, randomly select the counter node that will decide on this node.

Calculate the best separation based on these meter variables in the training set. Each plant is fully grown and not cut, so you can create a normal classification of plants. A new template is projected under the tree. This training is conducted at the terminal node. This comes in the final stages. The average voice of all trees is recorded in all processes. Like random forest forecasts.

V. PERFORMANCE EVALUATION

Ultimately, a system that can recognize emotions and predict their authenticity, and possibly the end result, is also truly valuable and useful. Classification accuracy factors were calculated for the data set. For example, in the problem of classifying two classes, positive and negative, in the same forecast there are four possibilities. A truly positive indicator and a truly negative indicator are the correct classification. When a false positive result occurs, when it is truly negative, a positive prognosis arises as a positive one. When the result is positive, it is taken as negative, and when it is erroneous, it is negative.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

A. Datasets

Pre-processing can have different effects in different contexts. Words and URLs that are not discriminatory in one context may contain some meaningful information in another context. This article describes the effects of pre-processing on five diverse Twitter records used in other emotion analysis materials.

Stanford Twitter Sentiment Test (STS) record. This was commented on manually and contains182 positive, 177 negative, and 139 neutral tweets. Although the test suite at Stanford is comparatively small, it has often been used for various assessment tasks.

The SemEval2014 record was provided in Task 9 by SEMEval2014. The record contains a tweet ID, which is identified with positive, negative and neutral tags. Some tweets were not available for download. This leaves us 11042 tweets for testing. Stanford Twitter Cement Gold Record (STS Gold). The doctoral record is automatically commented on by three doctoral students, both at the level of the tweets and at the level of existence.

The Sentiment Strength Dataset (SS-Twitter) contains 422,422 tweets, which are characterized by their positive and negative emotional strength. The Mood Assessment Dataset (SE-Twitter) contains tweets that have been humanized by three mechanical Turkish workers with emotional labels. Table 1 shows the distribution of the tweets across the five records selected from these price tags.

VI. EXPERIMENTAL VERIFICATION

This division reports the result obtained after several kinds of pre-processing methods. The accuracy development of one pre-processing method can be considered as:

$$Acc_{improvement} = Acc_{baseline}\ Acc_{compared} \quad (3)$$

The average F1- measure improvement of one pre-processing method was calculated as

$$F1_{improvement} = Average\ F1_{baseline} - Average\ F1_{compared} \quad (4)$$

Gain/Loss percentage in accuracy and F1 measure is calculated to compare the classifier's accuracy and F1 measures individually. By doing this way in which pre-processing techniques which classifiers are high and which classifiers are low is calculated and compared. We are using prior Polarity feature model for the calculation of accuracy and F1 measure. Then we are using the datasets they are STS-Gold and Film Review. The sentiment classification used here is binary. For example, here we have used removing URL and not removing URL likewise we have calculated for all the five Pre-processing techniques. The feature model used in prior probability. Then the classifiers used in this section are Naive Bayes. Support Vector machine, Random Forest.

Table 1 Gain/Loss in Acc and Average F1-M for not removing URLs using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.93	0.25
		SVM	1.43	1.2
		RF	1.88	1.35
F1-Measure	Prior Polarity	NB	1.59	1.65
		SVM	1.95	0.65
		RF	0.29	1.45

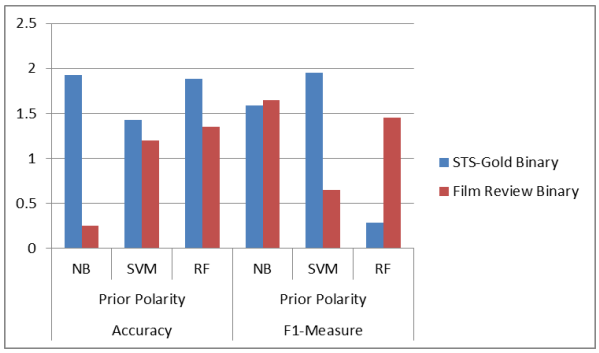

Figure 1 Gain/Loss in Acc and Average F1-M for not removing URLs using three classifiers for binary

Table 2 Gain/Loss in Acc and Average F1-M for removing URLs using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.14	0.21
		SVM	1.33	1.25
		RF	1.20	1.46
F1-Measure	Prior Polarity	NB	1.65	1.55
		SVM	1.67	0.56
		RF	0.99	1.46

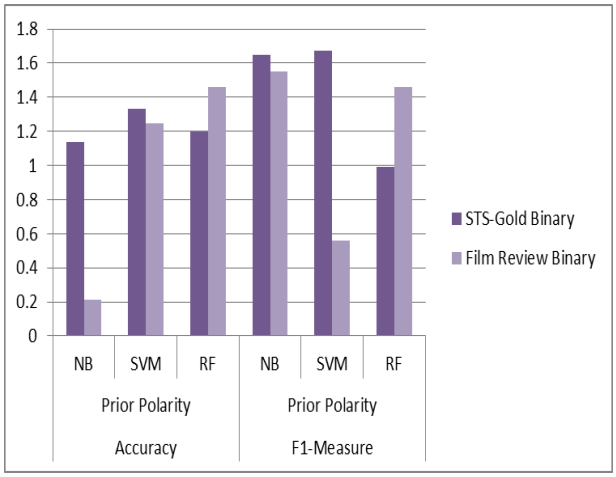

Figure 2 Gain/Loss in Acc and Average F1-M for removing URLs using three classifiers for binary

In Table 1 and Figure 1 have done removing URL using three classifiers but in Table 2 and Figure 2 we have done without removing URL. In Table 1 for STS-Gold the accuracy is high in NB than SVM and RF.

The F1 measure is high in SVM than NB and RF. In film review the accuracy is high in RF than NB and SVM. The F1 measure is high in NB than RF and SVM. Table 2 for STS-Gold accuracy is high in SVM when compared to RF and NB. Then for F1 measure SVM is high when compared to NB and RF. In film review the accuracy is high in RF than NB and SVM.

The F1 measure is high in NB than RF and SVM.

Table 3 Gain/Loss in Acc and Average F1-M for not removing stop words using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.14	1.36
		SVM	1.26	0.25
		RF	1.66	1.52
F1-Measure	Prior Polarity	NB	0.55	1.25
		SVM	1.26	0.78
		RF	1.47	1.85

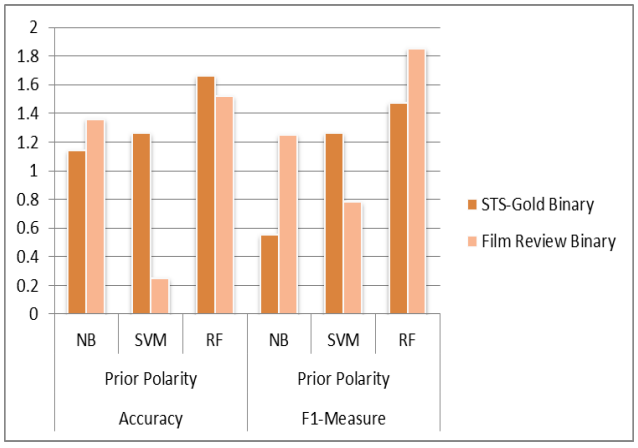

Figure 3 Gain/Loss in Acc and Average F1-M for not removing stop words using three classifiers for binary

Table 4 Gain/Loss in Acc and Average F1-M for removing stop words using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.78	1.45
		SVM	0.99	1.96
		RF	1.73	1.55
F1-Measure	Prior Polarity	NB	1.28	1.56
		SVM	1.33	1.28
		RF	0.55	0.24

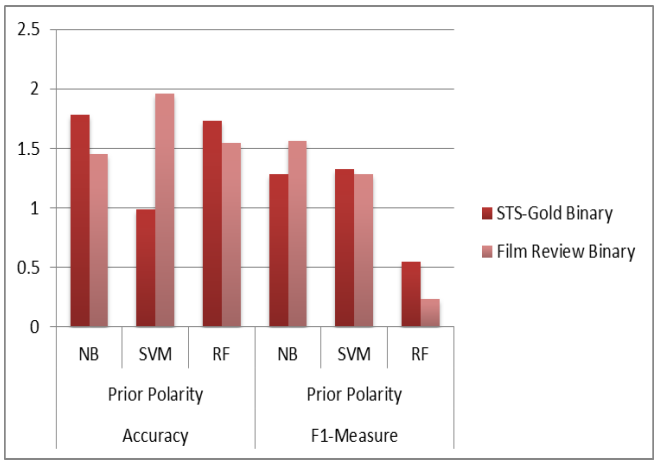

Figure 4 Gain/Loss in Acc and Average F1-M for removing stop words using three classifiers for binary

In Table 3 and Figure 3 we have done not removing stop words using three classifiers but in Table 4 and Figure 4 we have done with removing stop words.

In Table 3 for STS-Gold the accuracy is high in RF than SVM and NB. Then for F1 measure is high when compared to SVM and NB. In film review the accuracy is high in RF than NB and SVM.

The F1 measure is high in NB than RF and SVM. Table 4 for STS-Gold accuracy is high in SVM when compared to RF and NB.

Then for F1 measure SVM is high in NB when compared to SVM and RF. In film review the accuracy is high in SVM than NB and RF.

The F1 measure is high in NB than RF and SVM.

Table 5 Gain/Loss in Acc and Average F1-M for not removing numbers using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.23	1.52
		SVM	1.63	1.96
		RF	1.58	0.45
F1-Measure	Prior Polarity	NB	1.55	1.25
		SVM	1.69	1.85
		RF	0.25	0.66

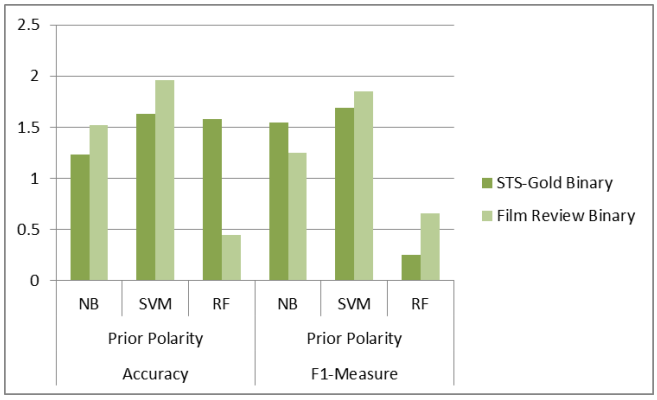

Figure 5 Gain/Loss in Acc and Average F1-M for not removing numbers using three classifiers for binary

Table 6 Gain/Loss in Acc and Average F1-M for removing numbers using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.55	1.58
		SVM	1.53	1.25
		RF	1.22	0.78
F1-Measure	Prior Polarity	NB	1.25	1.35
		SVM	1.32	1.45
		RF	0.89	1.36

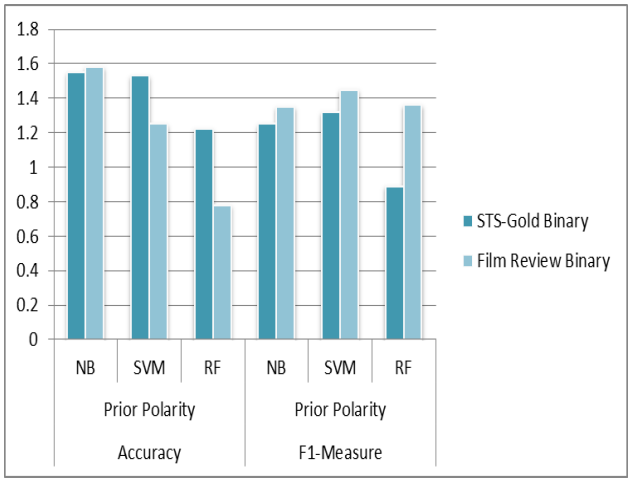

Figure 6 Gain/Loss in Acc and Average F1-M for removing numbers using three classifiers for binary

In Table 5 and Figure 5 we have done without removing numbers using three classifiers but in Table 6 and Figure 6 we have done removing numbers. In Table 5 for STS-Gold the accuracy is high in SVM than NB and RF.

The F1 measure is high in SVM than NB and RF. In film review the accuracy is high in SVM than RF and NB. The F1 measure is high in SVM than NB and RF.

In Table 6 for STS-Gold the accuracy is high in NB than SVM and RF.

The F1 measure is high in SVM than NB and RF. In film review the accuracy is high in NB than RF and SVM. The F1 measure is high in RF than NB and SVM.

Table 7 Gain/Loss in Acc and Average F1-M for not reverting repetition using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	0.22	1.66
		SVM	1.85	1.52
		RF	0.23	0.79
F1-Measure	Prior Polarity	NB	0.56	1.45
		SVM	1.32	0.25
		RF	1.85	1.46

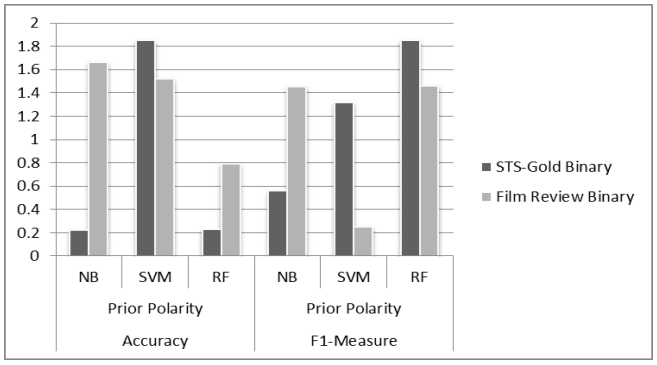

Figure 7 Gain/Loss in Acc and Average F1-M for not reverting repetition using three classifiers for binary

Table 8 Gain/Loss in Acc and F1-M for reverting repetition using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.89	1.22
		SVM	1.25	0.85
		RF	0.58	1.66
F1-Measure	Prior Polarity	NB	0.69	1.58
		SVM	1.56	1.74
		RF	0.68	0.25

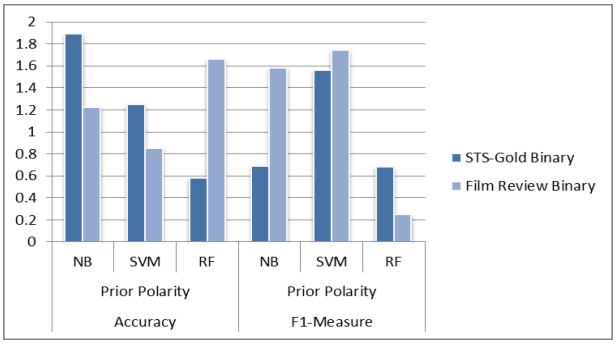

Figure 8 Gain/Loss in Acc and F1-M for reverting repetition using three classifiers for binary

In Table 7 and Figure 7 we have done without removing repetition using three classifiers but in Table 8 and Figure 8 we have done removing repetition.
Table 7 for STS-Gold the accuracy is high in SVM when compared to RF and NB.
Then for F1 measure RF is high when compared to NB and SVM. In film review the accuracy is high in NB than RF and SVM.
The F1 measure is high in RF than NB and SVM. In Table 8 for STS-Gold the accuracy is high in NB than SVM and RF. The F1 measure is high in SVM than NB and RF.
In film review the accuracy is high in RF than NB and SVM. The F1 measure is high in SVM than NB and RF.

Table 9 Gain/Loss in Acc and Average F1-M for not expanding acronyms using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	0.55	0.25
		SVM	0.22	1.96
		RF	0.58	0.27
F1-Measure	Prior Polarity	NB	1.88	1.66
		SVM	1.25	0.85
		RF	0.55	1.36

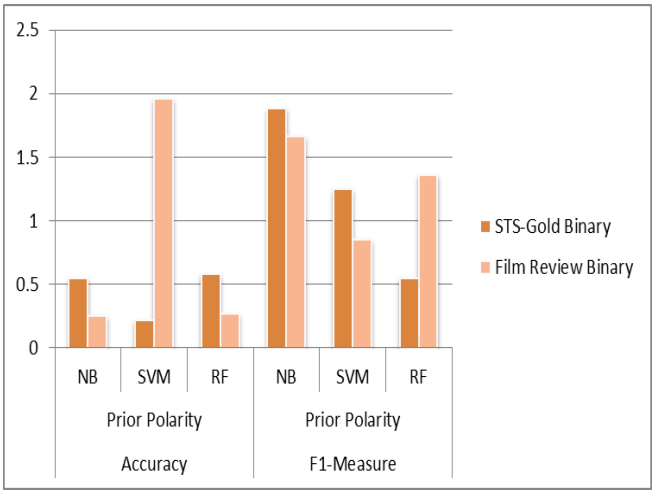

Figure 9 Gain/Loss in Acc and Average F1-M for not expanding acronyms using three classifiers for binary

Table 10 Gain/Loss in Acc and Average F1-M for expanding acronyms using three classifiers for binary

Gain /Loss%	Feature model	Classifiers	STS-Gold Binary	Film Review Binary
Accuracy	Prior Polarity	NB	1.36	1.25
		SVM	1.69	0.89
		RF	0.52	1.58
F1-Measure	Prior Polarity	NB	0.25	1.69
		SVM	1.56	1.58
		RF	1.68	0.28

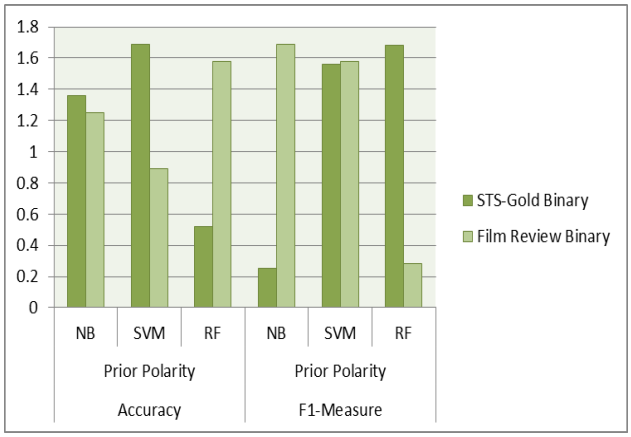

Figure 10 Gain/Loss in Acc and Average F1-M for expanding acronyms using three classifiers for binary

In Table 9and Figure 9, we have done for not expanding acronyms using three classifiers but in Table 10 and Figure 10 we have done for expanding acronyms. In table 9 for STS-Gold the accuracy is high in RF when compared to SVM and NB. Then for F1 measure NB is high when compared to SVM and RF. In film review the accuracy is high in SVM than NB and RF. The F1 measure is high in NB than RF and SVM.
In Table 10 for STS-Gold the accuracy is high in SVM than RF and NB. The F1 measure is high in RF than NB and SVM.
In film review the accuracy is high in RF than NB and SVM. The F1 measure is high in NB than RF and SVM.

V.CONCLUSION

In this article, we discuss how these six different preprocessing methods affect Twitter polar taxonomy. We are conducting a series of experiments with four classifiers to test the effectiveness of several pre-processing methods for five Twitter posts. Experimental results show that removing URLs, removing stop words, and removing numbers have minimal effects on classification performance. In addition, changing the position of the rejection and widening the abbreviation can improve the classification accuracy. So, deleting words, numbers, and URLs is a great way to stop noise, but it doesn't affect performance. Rejecting rejection is effective for mood analysis. We select the appropriate preprocessing methods and functional models for different classifiers to categorize Twitter's mood. In this paper we suggested a set of machine techniques playing through Semantic Sentence Recognition and Twitter related product ratings. Important goal is Using twitter api to evaluate a vast number of reviews And are branded already. The naïvebias and RF by technique Gives us a better outcome than maximum entropy and is SVM Is subjected to unigram model that results better than the other techniques.

REFERENCES

1. Ana Maria Popescu and Oren Etzioni ”*Extending Product Features and Opinion from Reviews*” October 2005. Association for Computational Linguistics.
2. Avirup Sil and Alexander Yates ”*Re-ranking for Joint Named -Entity Recognition and Linking* “ Copyright 2012 ACM 978-1-4503-1472-5/12/08.

3. Chenliang Li and Jianshu Weng. "*TwiNER: Named Entity Recognition in Targeted Twitter Stream* "vol.27,pp:558-570,2012.
4. GeorgiousPaltoglou and Mike Thelwall "*A Study of Information Retrieval Weighing schemes for sentiment analysis*" 11-16 July 2010 . c 2010 Association for Computational Linguistics.
5. Gui Xiaolin and Zhao Jianqiang (2017)"*Comparison Research on Test Pre-processing Methods on Twitter Sentiment Analysis*" IEEE ,2169 -3536, 2017 .
6. Jingjing Liu and Stephanie Seneff *"Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm*" August 2009 . ACL and AFNL.
7. Kun-Lin Liu , Wu-Jun Li and Minyi Guo .(2012)"*Emotion Smoothed Language Models for Twitter Sentiment Analysis" Copyright c 2012,Association for the Advancement of Artificial* Intelligence(www.aaai.org).
8. S.M.N.Arifin ,S. Dasgupta and V. Ng "*Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews,"* in Proceeding of the COLING/ACL on main conference Poster Sessions, ser. COLING-ACL '06.Stroudsburg ,PA, USA: Association for ComputationalLinguistics,2006,pp.611-618.

AUTHOR PROFILE

Jenifer Mahilraj received her B.Tech degree in Information Technology from Karpagam College of Engineering and M.E degree in Computer Science and Engineering from Karpagam University ,Coimbatore, india. She is currently pursuing her doctorate research and working as a senior faculty member in Faculty of computing & Software engineering, AMIT, Arbaminch University, Ethiopia. Her research interests include Data science, semantic analysis, Machine learning, Deep science, Cloud computing.

Getahun Tigistu received his Master degree from Addis ababa university and and working as a senior faculty member in Faculty of computing & Software engineering, AMIT, Arbaminch University, Ethiopia. His research interests include image processing, Machine learning, Artifical intelligence.

Sisay Tumsa received his Master degree from Arbaminch university and and working as a senior faculty member in Faculty of computing & Software engineering, AMIT, Arbaminch University, Ethiopia. His research interests includeData mining, Machine learning, Artifical intelligence.