JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES
VOL. 2, ISSUE 1, PP. 1-12, JUNE 2021.
Available at: www.journal.research.fanap.com
DOI: https://doi.org/10.22034/jaisis.2021.279225.1025

NB VS. SVM: A CONTRASTIVE STUDY FOR SENTIMENT CLASSIFICATION ON TWO TEXT DOMAINS

Razieh Asgarnezhad[1,*], S. Amirhassan Monadjemi[2]
[1]*Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran*
[2] *Senior Lecturer, School of continuing and lifelong education, National University of Singapore, 119077, Singapore*

ABSTRACT

Thanks to the availability of websites like Twitter, user-generated content is being published on the Internet every second. Sentiment Classification is one of the most attractive fields in text mining, which classifies reviews into positive and negative classes. Pre-processing is an important goal when these textual contexts are employed through machine learning techniques. Without effective pre-processing methods, inaccurate results will be achieved. This article aims to investigate the role of pre-processing in the Sentiment Classification problem. The main idea in this paper comes from using sampling techniques. This paper suggests classifying the tweets and reviews using supervised classifiers. We applied a set of pre-processing stages consisting of n-grams and samplings on two well-known datasets. Our results are worthwhile for companies to monitor the people's sentiment about their brands and for many other applications. We have provided further evidence to confirm the superiority of our model. Experimental results reveal that the proposed model outperforms the existing methods and can improve the performance of Sentiment Classification in terms of accuracy, precision, recall, and F1 criteria.

***KEYWORDS*:** Text mining, Sentiment classification, Supervised methods, Movie review, Twitter

1. INTRODUCTION

With the growth of various methods in text mining, this problem was the focuse of attention. One of the most widespread fields in Sentiment Analysis is Sentiment Classification (Xia et al., 2016). Sentiment Classification problem exploits and analyses the unstructured data. It automatically extracts the reviews in proper context from websites and classifies the polarity of text in terms of positive and negative classes. It causes a decision-making problem to be settled automatically. Therefore, a significant model is necessary to facilitate the information capture. In the current study, a novel model for Sentiment Classification is proposed.

To capture the beneficial information and determine the polarity of the reviews, most of the existing methods employed Sentiment Classification approaches. These approaches consist of Machine Learning (ML), Lexicon-Based, and Hybrid approaches. The ML approaches used supervised, unsupervised, and semi-supervised methods. In 2015, we compared the validity of these approaches (Asgarnezhad & Mohebbi, 2015). It inherits high accuracy from supervised ML algorithms and achieving stability for the lexicon-based approach.

A supervised approach automatically learns from the training set and can evaluate the accuracy of classification. However, the absence of a labeled dataset can be a big drawback. Hence, the first stage is the collection of the training set. The key stages are classifier and feature selections, which determine the

* *Corresponding Author, email: razyehan@gmail.com*

RECEIVED: 2 APRIL, 2021; **ACCEPTED:** 14 MAY, 2021; **PUBLISHED ONLINE:** 17 MAY, 2021

classification performance. Authors showed that Support Vector Machine (SVM) (Tu et al., 2012), Naive Bayes (NB) (Dinu & Iuga, 2012), and maximum entropy (ME) (Appel et al., 2016; Tripathy et al., 2016) successfully used in most of the researches and performed well in the SC. Striking studies were applied to weigh frequencies like Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TFIDF) (Martineau & Finin, 2009; Montejo-Ráez et al., 2014) for pre-processing. Part of speech (POS) is a feature, which separates tokens. POS is typically used in Lexicon-Based approaches. However, POS introduced a few new pieces of information and added additional complexity (Khan et al., 2016). Other studies applied features based on co-occurrence terms like unigram and bigram (Tu et al., 2012; Tripathy et al., 2016). Besides, the increase of the value n in n-gram can decrease the accuracy (Tripathy et al., 2016). It revealed that words were useful indicators for polarity classification (Tu et al., 2012). However, the usual bag of words (BOW) did not present multiple relationships. Hence, we added n-grams to obtain features inside words. We used sampling techniques to decrease the dimension of data in pre-processing stages.

This study proposes a novel approach called NS using base classifiers for Sentiment Classification on two datasets. The NS model suggests different combinations of weighting schemas and sampling techniques to improve the classification performance. The difference between our study and other studies is the diversity of pre-processing techniques like sampling, n-grams, and weighing schemas. In contrast to the others, no data reduction technique can be compared to our model in this context. The current study is simple and does not append additional costs. The Polarity Movie Dataset (PMD) and Twitter-Sanders-Apple (TSA) were considered to evaluate our model. However, the Twitter datasets used in the literature are not openly available and labeled, except for TSA.

The motivation of this study is to investigate the impact of sampling in conjunction with n-grams, and weighting schemas in this context. Two supervised ML approaches like SVM and NB were employed. TF and TFIDF schemas were used as weighting schemas. The obtained results show the difference between the two datasets. The highest accuracy was obtained through TF and SVM on the PMD. The highest accuracy was obtained through TFIDF and NB on the TSA. Therefore, the NS model captures the meaningful results; because the proposed model is an independent-domain approach.

The remaining of this article is organized as follows: Sec. 2 and Sec. 3 provide related works and our approach, respectively. Experimental results are presented in Sec. 4. Finally, the paper ends with a conclusion in Sec. 5.

2. Related Works

Sentiment Classification has attracted a great deal of attention in recent years. A large number of methods are proposed for improving classification performance. These methods differ from each other in the architecture of the classifier, algorithm parameters, or pre-processing methods. Here, the summarization of some of the existing articles are of concern:

In 2002, authors (Pang et al., 2002) used NB, SVM, and ME on the movie dataset. In 2005, authors (Whitelaw et al., 2005) proposed a semi-automatically lexicon-based approach. In 2006, authors (Kennedy & Inkpen, 2006) applied unigram and bigram features to reveal the higher accuracy of their method. Hence, we added n-grams to the BOW to extract features based on word relationships. Specifically, we employ sampling techniques for pre-processing stages. In 2009, authors (Martineau & Finin, 2009) investigated BOW and SVM by applying a delta TFIDF to classify the reviews. Also, Ohana and Tierney applied unsupervised and supervised techniques for receiving semantic term–document information (Onan et al., 2016).

In 2011, authors (Kouloumpis et al., 2011) probed the effect of linguistic features in Twitter reviews. They applied a supervised method to conduct the training phase. In 2012, authors (Dinu & Iuga, 2012) investigated the operation of the NB for the best way of features selection from texts in Sentiment Classification. In (Tu et al., 2012), authors applied lexical words and POS tags to perform experiments on the Movie Dataset and achieved redeeming results in document-level SC. However, POS tags cannot propose new information and

adds unnecessary complexity. Also, authors (Mudinas et al., 2012) developed a concept-level Sentiment Classification using both lexicon and learning approaches.

In 2014, authors (Nguyen et al., 2014) presented a new feature, rating-based feature (RBF), with n-gram to evaluate a supervised ML approach in document-level Sentiment Classification on two popular Standard Polarity Movie Datasets. In 2016, the authors (Tripathy et al., 2016) observed that the increment of the value n in n-gram decreased the accuracy. In 2017, a new feature extraction method was proposed (Seyyedi & Minaei-Bidgoli, 2017). They investigated Information Gain (IG) and Chi-square Statistic as well-known feature selection methods in the text classification task in the spam detection field.

The present authors (Asgarnezhad et al., 2018) proposed a model for Twitter Sentiment Classification in 2018. They investigated the role of weighting feature techniques in Sentiment Classification using supervised methods on the Twitter dataset. In 2019, a heuristic-based model for feature optimization was used by Rosenthal, Farra, & Nakov, 2019 to improve the Sentiment Classification performance. Five classifiers consisting of NB, SVM, K-nearest neighbor multilayer perceptron, and decision tree were applied to extract features. The highest accuracy of 76.5% was achieved for SVM on the SemEval 2016 benchmark dataset. Also, the current authors published works in this scope on the different datasets (Monadjemi et al., 2020; Asgarnezhad et al., 2020a; Asgarnezhad et al., 2020b; Asgarnezhad et al., 2019).

3. THE NS MODEL

Here, a novel model for document-level Sentiment Classification is proposed. This method studies binary classification. The proposed model investigates the effects of sampling, n-grams, and weighing schemas using base classifiers for Sentiment Classification. TF and TFIDF schemas are applied in the current study as well. Besides, it merges with samplings and n-grams. The proposed model applies two supervised methods to manage document-level Sentiment Classification. After pre-processing and choosing the best features, we employ NB and SVM as the classifier because these classifiers are used successfully in literature. Optimization selection, genetic algorithm, and sampling Techniques were added to improve the performance of individual classifiers. Fig. 1 presents the stages of the NS model.

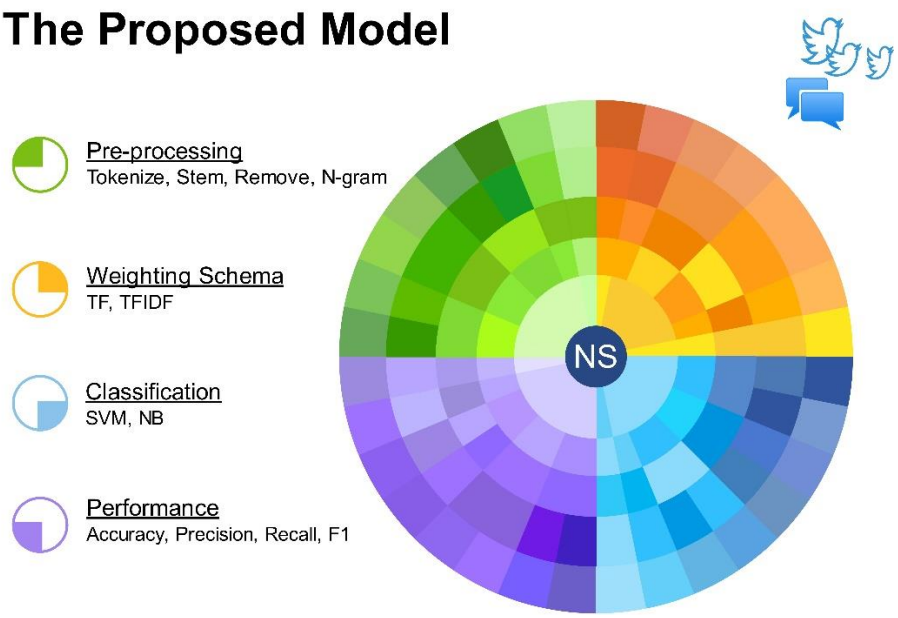

Fig. 1. The stages of NS model

3.1. *Pre-processing*

First, characters, words, and useless tokens are tokenized. The term length threshold value helps to omit the words based on their lengths. With a threshold, some words with improper length were removed in the text and

increased the overall performance. When we analyzed the datasets, we observed that many words do not appear. Therefore, it is better to remove them in our analysis. Next, words' roots were traced. It includes separating "ed" in the past tenses, ing in a present participle, etc. According to the stop word list, stop words are eliminated from the text. Finally, we used n-grams during our experiments. Two weighing schemas are applied to produce word vectors.

TF describes the relative frequency of a word in the document. TFIDF was probed by splitting the number of records. These schemas are of concern (Manning et al., 2008):

$$TF = \frac{F_{t,d}}{F_d} \tag{1}$$

$$TFIDF = TF \,.\, Log\,(\frac{N}{F_t}) \tag{2}$$

where t is the word in document d, TF is the frequency of word t in document d, N is the number of documents, and F_t is the number of documents in a collection consisting of word t.

3.2. *Sampling*

Numerous sophisticated methods existed which aimed at choosing the best informative samples in the document (Habernal et al., 2015; Forman, 2003; Zheng et al., 2004; Uchyigit, 2012). Here, a brief description of these methods is of concern.

Mutual Information (MI): The MI method regarding two identifiers can be a mechanism for the mutual dependence associated with these identifiers. The MI is used to estimate the occurrence probability of a feature in the objective class in relationship to the overall occurrence probability (Schütze et al., 2008). The schema is of concern:

$$MI_w = (c_w \times N) / (c_w + c_{\bar{w}})(c_w + \bar{c}_w) \tag{3}$$

where c_w , $\bar{c}_w$, $c_{\bar{w}}$, $\bar{c}_{\bar{w}}$ are the document of class c and contain word w, the document of class c and contains word w, the document of class c and does not contain word w, the document without class c and does not contain word w, respectively. Also, N is the total number of the documents.

Information Gain (IG): Here, the presence or absence of a feature in an original document is necessary. This metric defines the number of bits for this required information to divide the appropriate class for the document (Sebastiani, 2002). These schemas are of concern:

$$\begin{aligned} IG_w = &-P(c)\log_2 P(c) + P(\bar{c})\log_2 P(\bar{c}) - (P(w)(-P(c_w)\log_2 P(c_w) - P(\bar{c}_w)\log_2 P(\bar{c}_w))) \\ &+(p(\bar{w})(-P(c_{\bar{w}})\log_2 P(c_{\bar{w}}) - P(\bar{c}_{\bar{w}})\log_2 P(\bar{c}_{\bar{w}}))) \end{aligned} \tag{4}$$

$$\begin{aligned} &P(c_w) = c_w / (c_w + \bar{c}_w), \;\; P(\bar{c}_w) = \bar{c}_w / (c_w + \bar{c}_w) \\ &P(c_{\bar{w}}) = c_{\bar{w}} / (c_{\bar{w}} + \bar{c}_{\bar{w}}) \;, \;\; P(\bar{c}_{\bar{w}}) = \bar{c}_{\bar{w}} / (c_{\bar{w}} + \bar{c}_{\bar{w}}) \end{aligned} \tag{5}$$

$$\begin{aligned} &P(w) = (c_w + \bar{c}_w) / N \;\; , \;\; P(\bar{w}) = (c_{\bar{w}} + \bar{c}_{\bar{w}}) / N \\ &P(c) = n_c / N \;\; , \;\; P(\bar{c}) = n_{\bar{c}} / N \end{aligned} \tag{6}$$

Chi-square (CHI) and Variants Chi-square (χ2): Here, the well-known analytical measures are used to define the independence between two alternatives as a feature and a class. These measures are applied to choose the features with superior properties. Also, authors (Ng et al., 1997) suggested a variant of χ2, namely NGL

showed the superior NGL than χ2, in some cases. Besides, authors (Galavotti et al., 2000) displayed a simplified form of χ2, named GSS coefficient. They asserted that GSS produces better results than NGL and χ2. The schema is of concern:

$$GSS_w = c_w \bar{c}_{\bar{w}} - \bar{c}_w c_{\bar{w}}, \quad NGL_w = \frac{\sqrt{N}\, GSS_w}{\sqrt{(c_w + c_{\bar{w}})(\bar{c}_w + \bar{c}_{\bar{w}})(c_w + \bar{c}_w)(c_{\bar{w}} + \bar{c}_{\bar{w}})}}, \quad \chi^2{}_w = (NGL_w)^2 \tag{7}$$

In this article, stratified and bootstrapping sampling methods yield better accuracy. Bootstrapping method performs better stratified method. Therefore, we applied bootstrap sampling in all of the experiments. The reason for using our sampling is simplicity and independence. As can be seen, other methods are more sophisticated and depend on the numbers of words in positive or negative documents whereas, our used sampling method is simple. On the other hand, we used the supervised methods for classification that matched more with our employed sampling method. How do our used sampling techniques work?

These techniques led us to choose the best features through data reduction. In stratified sampling, the folds of the training set are stratified. The class distribution for tuples in a fold is similar to the initial data. It enables the algorithm to preserve the distribution of the training set. Bootstrapping sampling generates a bootstrapped sample from the dataset. This type of sampling may not have all unique examples; hence, it will be different from other sampling techniques. We use both in our model, but bootstrapping with replacement achieves higher performance. We apply bootstrap sampling in all of the experiments. The obtained results showed that bootstrapping gives a high performance on the text input. Here, a sampling structure of scores was applied. Sampling with or without replacement is the first portion. The size of the sample is the second portion. When sampling with replacement happens, the bootstrap is defined as a procedure herein. This bootstrap applies the dimensions of the samples as much as the volume of the original collection demands. In the second portion, a taxonomy is determined by extending the volumes of the possible samples. The distribution of sampling with dimensions will be more extensive the distribution of sampling when the dimension of the original data is not applied. Nevertheless, this event did not consolidate for the samples without replacement. So, this proposal only admits expansion for the bootstrap.

The tuple is selected and and added to the training set again when we choose each tuple of the input set. In each step of iterations, all examples have an equal probability of being selected. When an example is chosen, it remains a candidate for further selection and is ditermined again in the next step. A sample with replacement can certainly have the same examples. Therefore, it is used to create a sample that is greater in volume than the original one. It appears that bootstrapping performs better than the former one. The main idea of utilizing bootstrapping is applying the data as an input set for approximating the sampling distribution. It generates an enormous volume of samples named bootstrap samples. The sample outline is computed for each sample of bootstrap (Han & Micheline, 2006).

Here, some notations for utility are described. Assume a parameter for the population θ is as a target herein. A random sample with a volume *n* provides the data $(x_1, x_2, \ldots, x_n)$. Assume θ is a sample created from the dataset. The distribution of θ with large n can be bell-shaped with a center θ and standard deviation $\sigma/\sqrt{n}$ for each sample, where the positive volume depends on two factors like the population, and the type of statistic θ. There exists technical complexity for standard deviation, when θ is median or correlation of sample. Hence, bootstrap allows a bypass. Assume θ_B is a quantity for presenting the same statistic, which produces on a bootstrap sample of $(x_1, x_2, \ldots, x_n)$. With the limitation of $(n \to \infty)$, the distributions of θ_B were bell-shaped with θ as the center and the corresponding standard deviation $\sigma/\sqrt{n}$. So, the distribution $\theta_B - \theta$ will be the distribution $\theta - \theta$, that is, the bootstrap Central Limit Theorem. It should also be noted that with a limiting distribution of the sampling for a mathematical function that does not include population unknowns, bootstrap distribution

allows a better conjecture than the CLT. If the procedure is $(\hat{\theta}_B - \hat{\theta})/SE$, where SE regarded as a sample estimation of the standard error of $\hat{\theta}$, the limiting sampling distribution will be standard normal.

Here, $\theta = \mu$ is the population means, $\theta = \overline{X}$ is the sample mean, σ is population standard deviation, and s is sample standard deviation reflected, which is generated from the original dataset. Moreover, s_B is the sample standard deviation, which is calculated on a bootstrap sample. Besides, the sampling distribution of $(\overline{X} - \mu)/SE$, with $SE = \sigma/\sqrt{n}$, will be estimated through the bootstrap distribution of $(\overline{X_B} - \overline{X})/SE$ where $\overline{X_B}$ is bootstrap sample means, and $SE = s/\sqrt{n}$. Furthermore, the sampling distribution of $(\overline{X} - \sigma)/SE$ where $SE = s/\sqrt{n}$, will be assessed through the bootstrap distribution of $(\bar{x}_B - \bar{x})/SE_B$ where $SE = s/\sqrt{n}$. Here, the description of the approximating standard error of sample evaluation for utility is of concern.

We assume that the information investigated regarding the population parameter of θ where $\bar{\theta}$ is a sample estimator of θ based on a stochastic sample has size *n*. To estimate the standard error for $\hat{\theta}$, a bootstrap approach is of concern: calculate $(\theta_1^*, \theta_2^*, ..., \theta_N^*)$, through the equivalent relation for $\hat{\theta}$, precisely with N numbers of different bootstrap samples. A primary recommendation for the size N could be $N = n^2$, unless n^2 is too large. In that case, it could be decreased to an acceptable volume, say $n \log_e^n$. So, $SE_B(\hat{\theta})$ defined as:

$$SE_B(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^{N} (\theta_i^* - \hat{\theta})^2}{N}} \tag{8}$$

It showed that more instances could exploit more necessary information about the dataset. Consequently, it may consider a novel example in the dataset, which is outstanding for classification. That is why bootstrapping has become an appropriate tool in our model.

3.3. Classification

Authors (Tu et al., 2012; Dinu & Iuga, 2012; Appel et al., 2016; Tripathy et al., 2016) revealed that supervised methods such as NB and SVM achieved the highest accuracy of the classification task. Consequently, we apply supervised methods to obtain the highest performance. SVM and NB outperformed better than others on the PMD and TSA, respectively. NB is an algorithm that assumes features are independent with equal importance. It is a simple algorithm based on probabilistic theory. SVM is an ML technique depending on the statistical learning concept. The essential role of the training step is to choose a maximum of margin hyperplane. It can produce a maximum separation among classes and examples.

Naïve Bayes (NB): NB received reasonable accuracy. It is simple and assumes to have independent features. Besides, it is used mainly when the volume of the training set is not vast. Here, (1) is utilized to compute the probability of event A in column A, provided that class C holds:

$$P(K = A \mid C) = \frac{1}{\sqrt{2\pi\sigma^2_{K=C}}} e^{-\frac{A - \mu_{K=C}}{2\sigma^2_{K=C}}} \tag{9}$$

where $\mu k = c$ is the column K mean, while the row belongs to the class C and $\sigma^2_{k=c}$ is the variance of the kth therein, and no input classification is required. An example is given to describe the Bayes Continuous Decider, where there exist four features with positive or negative classes.

Support Vector Machine (SVM): SVM is a machine learning technique that is based on the statistical learning concept. This classifier has achieved well in text classification applications. The essential role of the training step is to choose a maximum margin hyper plane which is depicted by vector $\vec{w}$. It departs the document vectors of one class from the others. So, an optimization problem is restrained. Here, $c_j \epsilon \{1, -1\}$ is the correct class of document d_j so that the solution can be depicted as:

$$\vec{w} = \sum_j \alpha_j c_j \vec{d_j} \quad , \quad \alpha_j \geq 0 \tag{10}$$

Also, α_j could be obtained by solving a dual optimization problem (Lagrangian multiclasspliers). The $\vec{d_j}$ as α_j is greater than zero and is named the support vector.

4. RESULTS AND DISCUSSIONS

4.1. *Evaluation Metrics*

The R programming language runs NB and SVM classifiers to conduct experiments. Evaluation measures consist of accuracy, precision, recall, and F1 measure. Here, these measures for evaluating Sentiment Classification have been introduced. P and N are the numbers of positive and negative tuples. TP refers to the positive tuples that have been labeled by the classifier correctly. TN refers to the number of true negatives. FP is the negative tuples that have been labeled incorrectly as positive. FN is the positive tuples that have been mislabeled as a negative tuple. Accuracy is the sum of actual tuples that classified TP and the number of TN relative to the total number of classified instances. The precision state is the percentage of tuples that have been labeled as positive and actual. Recall refers to the percentage of tuples that are labeled positive. F-measure combines precision and recall into a single measure. F-measure comes from a weighted harmonic mean of precision and recall. The main reasons for using these measures are the nature of the dataset, two different domains, and short tweets. Also, in tweets, sentences are short and need more measurement than one. Hence, one measure will not be sufficient and more effective for comparison with other existing works. So, we apply the mentioned measures to have an appropriate comparison. These measures are of concern (Han & Micheline, 2006):

$$Accuracy = (TP + TN) / (P + N) \tag{11}$$

$$\Pr ecision = (TP) / (TP + FP) \tag{12}$$

$$\mathrm{Re}\, call = (TP) / (P) \tag{13}$$

$$F1 = (2.\Pr ecision.\mathrm{Re}\, call) / (\Pr ecision + \mathrm{Re}\, call) \tag{14}$$

4.2. *Datasets*

The first dataset is the movie dataset. Polarity Dataset v2.0 (Pang et al., 2002) was applied for training the classifiers. This corpus has 700 positive and 700 negative movie reviews. The second dataset is the Twitter dataset. There are a few available and free resources for Twitter. None of the existing datasets about Twitter is labeled with the sentiment, except the sanders dataset. This dataset is available at the following link: http://www.sananalytics.com. This dataset has 163 positive and 316 negative tweets.

4.3. *Simulations*

Two experiments show the effect of sampling and n-gram features on the two datasets, respectively. The LibSVM library for training SVM is applied. Epsilon equal to one is considered for SVM. SVM and kernel type are analyzed as C-SVC and RBF, respectively. The residual parameters for SVM remained as default. The number of kernels fixed 10 for NB. The 10-fold stratified cross-validation was employed.

Experiment I. The first experiment investigated the effect of bootstrapping sampling on the classification performance of the two datasets. The obtained results of the datasets are shown in Table 1 and Table 2. It seems that bootstrapping sampling which produced accuracy increases in most cases.

Experiment II. This experiment investigated the effect of n-gram features besides sampling on the two datasets. Dinu and Iuga showed that the best results belong to the feature set consisting of bigrams collocations in the texts (Dinu & Iuga, 2012). However, they could increase the improvement by 1%. Hence, we employ trigrams as a choice besides bigrams to make a difference to our study. The study indicated that if a higher level of n-grams is considered, the result is expected to be better.

Table 1. The accuracy of the NS model with/without sampling on the PMD dataset

Accuracy with sampling	Accuracy without sampling	Classifier	Weighting Schema
92%	83.70%	SVM	TFIDF
92.20%	83.90%	SVM	TF
85.15%	75.55%	NB	TFIDF
84.45%	75.65%	NB	TF

Table 2. The accuracy of the NS model with/without sampling on the TSA dataset

Accuracy with sampling	Accuracy without sampling	Classifier	Weighting Schema
88.11%	81.64%	SVM	TFIDF
85.82%	80.59%	SVM	TF
88.32%	78.28%	NB	TFIDF
85.60%	75.77%	NB	TF

The obtained results are summarized in Table 3 and Table 4. In the PMD dataset, the improvement was more relevant since the data is collected from emoticons. In this type of dataset, there is no specific domain. The other datasets address more specific topics like the technology domain on the TSA.

In Table 3, performance values are presented on the PMD. The highest accuracy, precision, recall, F1 obtained 92.90, 92.76, 92.95, 92.85%, respectively. These results belong to TF schema, SVM, and trigrams. Hence, our pre-processing achieved remarkable results. It showed that the TF schema produced better results in more cases. In this experiment, we revealed that if a higher level of n-grams is considered, the result is expected to be better. On the other hand, due to the essence of the PMD dataset, TF often attains better results. The worst results were obtained through NB and the TFIDF schema. It is not suitable for a large number of examples on the PMD.

Table 3. The performance of the NS model with sampling on the PMD dataset

Classifier	Weighting Schema	N-grams	F1	Recall	Precision	Accuracy
SVM	TFIDF	n = 1	92.02%	92.03%	92.01%	92%
		n = 2	92.35%	92.35%	92.36%	92.35%
		n = 3	92.25%	92.25%	92.26%	92.25%
	TF	n = 1	92.22%	92.25%	92.19%	92.20%
		n = 2	92.74%	92.75%	92.74%	92.75%
		n = 3	92.85%	92.95%	92.76%	92.90%
NB	TFIDF	n = 1	85.17%	85.15%	85.19%	85.15%
		n = 2	86.49%	86.47%	86.52%	86.50%
		n = 3	86.54%	86.52%	86.57%	86.55%
	TF	n = 1	84.47%	84.45%	84.50%	84.45%
		n = 2	86.44%	86.44%	86.45%	86.45%
		n = 3	86.17%	86.15%	86.19%	86.20%

Table 4. The performance of the NS model with sampling on the TSA dataset

Classifier	Weighting Schema	N-grams	F1	Recall	Precision	Accuracy
SVM	TFIDF	n = 1	85.75%	83.63%	87.97%	88.11%
		n = 2	87.89%	84.84%	91.17%	89.79%
		n = 3	86.79%	82.61%	91.41%	88.74%
	TF	n = 1	82.87%	80.12%	85.81%	85.82%
		n = 2	88.22%	84.81%	91.91%	89.99%
		n = 3	87.47%	83.80%	91.48%	89.37%
NB	TFIDF	n = 1	86.71%	87.46%	85.98%	88.32%
		n = 2	88.57%	88.64%	88.50%	90.20%
		n = 3	88.84%	88.98%	88.70%	90.40%
	TF	n = 1	84.36%	85.86%	82.91%	85.60%
		n = 2	85.99%	87.04%	84.96%	87.48%
		n = 3	86.16%	87.04%	84.96%	87.48%

Table 4 shows the performance values on the TSA dataset. The highest accuracy and F1 obtained 90.40 and 88.84%, respectively, and these results belong to TF schema, NB, and trigrams. Also, the highest precision gained 91.91%, which belongs to SVM, bigram, and TF schema. The highest recall is 88.98%, which belongs to the TFIDF schema, NB, and trigrams.

4.4. Discussions

This paper seeks to address the effect of sampling and n-gram features on Sentiment Classification using two base classifiers. Here, we discuss the obtained results. We observe that trigram features can improve evaluation metrics. Also, sampling can try to achieve higher results. Note that NB provides worst results than SVM when TFIDF schema is used. That is because NB is not suitable for a vast number of examples on the PMD dataset. The experimental results produced via SVM are substantially better than those of the results obtained through NB on the PMD dataset.

We have provided further evidence that our model is better than other existing works. Fig. 2 and Fig. 3 show the comparison between the best-obtained results of the NS and the best results in the literature. However, there

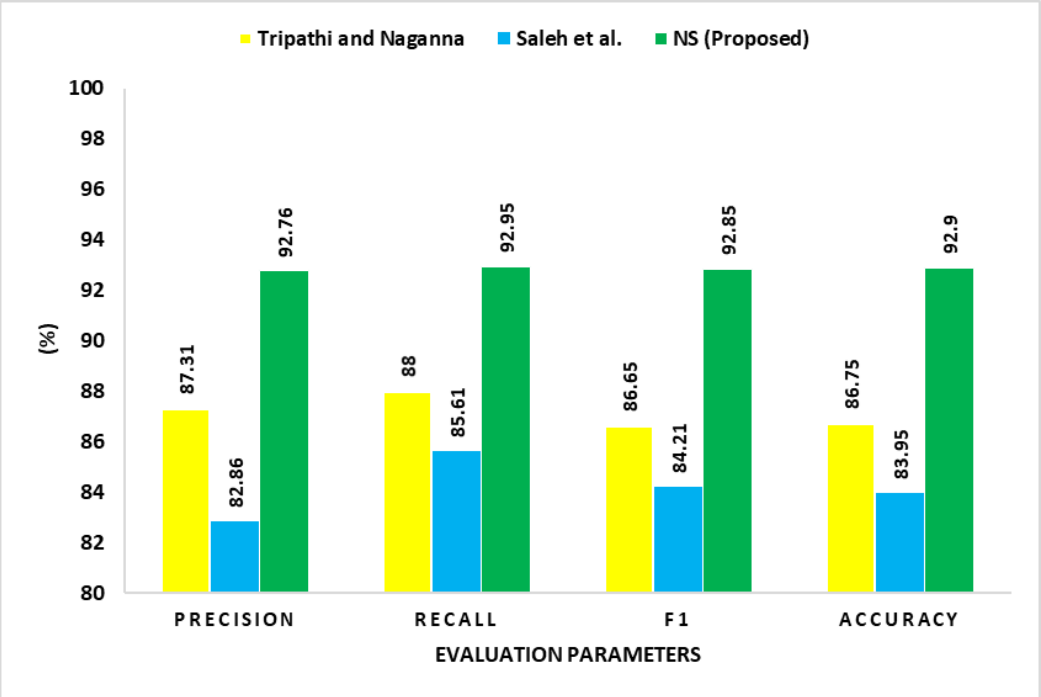

Fig. 2. Overview of our best results and the best results from the literature on the PMD dataset

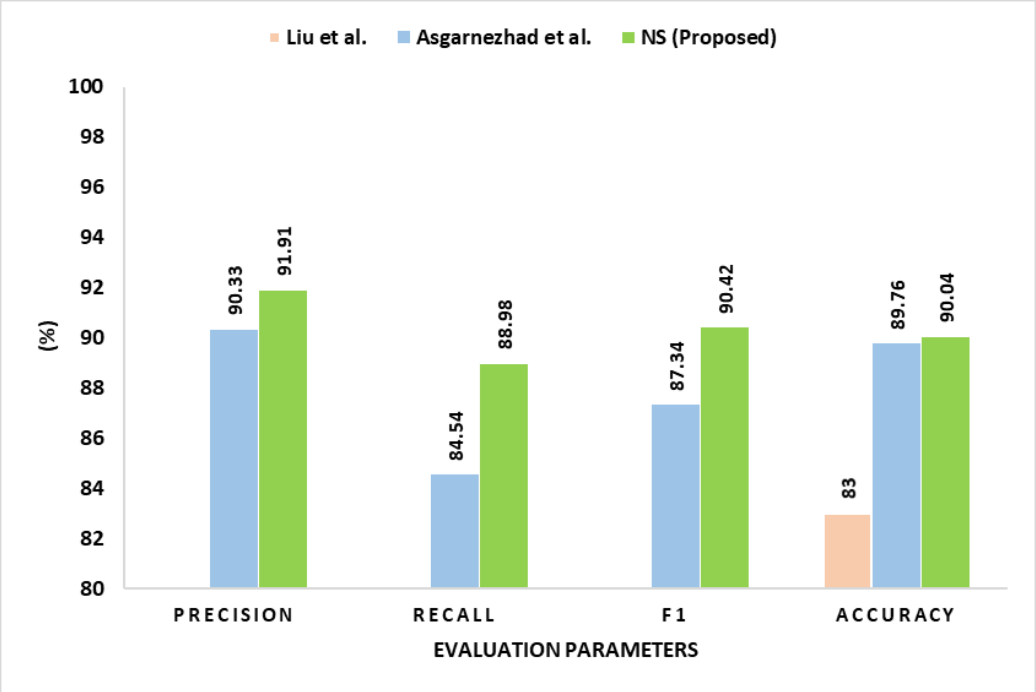

Fig. 3. Overview of our best results and the best results from the literature on the TSA dataset

are few free labeled datasets for Twitter. We believe that we have developed an innovative solution to solve the Sentiment Classification problem on datasets.

According to Fig. 2, we illustrate a comparison among the best results of the NS and the best results in the literature on the PMD dataset (Saleh et al., 2011; Tripathi & Naganna, 2015). It showed that trigram features and sampling gives the highest performance in all cases. The highest accuracy of the NS obtained 90.90% on the PMD dataset, whereas the method proposed by Saleh et al., 2011 and Tripathi & Naganna, 2015 obtained 83.95 and 86%, respectively. The former used SVM as a classifier while the later applied NB besides SVM. Both works created n-gram features. Also, NB, SVM, and n-grams feature were applied to the sample, and this factor could well be responsible for this result. The paper takes a new look at Sentiment Classification for both the PMD and TSA datasets. Also, Fig. 3 demonstrates that our model obtained the highest accuracy than the accuracy obtained by Liu, Li, & Guo, 2012 and our previous model (Asgarnezhad et al., 2018) on the TSA dataset.

5. Conclusions

In this article, we proposed a model for Sentiment Classification on two datasets. We showed that our results were notable for both Twitter and Movie datasets. The use of sampling and n-grams features using supervised methods for Sentiment Classification has been underexplored in the literature. We have demonstrated that our pre-processing stages formed by two base learners can provide satisfactory results. We also compared promising models for the Sentiment Classification task and showed their advantages and drawbacks. When the focus is on accuracy, the best choice is the BOW. According to experimental results, the proposed model is a much better binary Sentiment Classification problem. Besides, the bootstrapping method is used and parameters of SVM are optimized to improve the performance of the models. After two experiments by using different classifiers and combinations, we achieved higher performance on two datasets. All experiments worked notably in two different datasets.

As future work, we are planning to evaluate the ensemble approaches in conjunction with multi-objective algorithms to decrease the number of features. Besides positive and negative classes, we are planning to examine the neutrals in this context.

REFERENCES

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110-124.

Asgarnezhad, R., & Mohebbi, K. (2015). A comparative classification of approaches and applications in opinion mining. International Academic Journal of Science and Engineering, 2 (5), 1-13.

Asgarnezhad, R., Monadjemi, A., & Soltanaghaei, M. (2019). A High-Performance Model based on Ensembles for Twitter Sentiment Classification. Journal of Electrical and Computer Engineering Innovations (JECEI), 8(1), 41-52.

Asgarnezhad, R., Monadjemi, A., & SoltanAghaei, M. (2020). NSE-PSO: Toward an Effective Model Using Optimization Algorithm and Sampling Methods for Text Classification. Journal of Electrical and Computer Engineering Innovations (JECEI), 8(2), 183-192.

Asgarnezhad, R., Monadjemi, S. A., & Soltanaghaei, M. (2020). An application of MOGW optimization for feature selection in text classification. The Journal of Supercomputing, 1-34.

Asgarnezhad, R., Monadjemi, S. A., Soltanaghaei, M., & Bagheri, A. (2018). SFT: A model for sentiment classification using supervised methods in Twitter. Journal of Theoretical & Applied Information Technology, 96(8), 2242-2251.

Dinu, L. P., & Iuga, I. (2012). The Naive Bayes classifier in opinion mining: in search of the best feature set. Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics, 556-567.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3, 1289-1305.

Galavotti, L., Sebastiani, F. & Simi, M. Experiments on the use of feature selection and negative evidence in automated text categorization. International Conference on Theory and Practice of Digital Libraries, 2000. Springer, 59-68.

Habernal, I., Ptacek, T. & Steinberger, J. 2015. Reprint of "Supervised sentiment analysis in Czech social media". Information Processing & Management, 51, 532-546.

Han, J., & Micheline, K. (2006). Data mining: concepts and techniques. Morgan Kaufmann Publishers–An Imprint of Elsevier, 500, 105-150.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence, 22(2), 110-125.

Khan, F. H., Qamar, U., & Bashir, S. (2016). eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. Information Sciences, 367, 862-873.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! Paper presented at the Fifth International AAAI conference on weblogs and social media, 538-541.

Liu, K.-L., Li, W.-J., & Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. Paper presented at the Twenty-sixth aAAI conference on artificial intelligence, 1678-1684.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval: Cambridge university press.

Martineau, J. C., & Finin, T. (2009). Delta tfidf: An improved feature space for sentiment analysis. Paper presented at the Third international AAAI conference on weblogs and social media, 258-261.

Monadjemi, S., Asgarnezhad, R., & Soltanaghaei, M. (2020). FAHPBEP: A fuzzy Analytic Hierarchy Process framework in text classification. Majlesi Journal of Electrical Engineering, 14(3).

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked wordnet graph for sentiment polarity classification in twitter. Computer Speech & Language, 28(1), 93-107.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. Paper presented at the Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining.

Nguyen, D. Q., Nguyen, D. Q., Vu, T., & Pham, S. B. (2014). Sentiment classification on polarity reviews: an empirical study using rating-based features.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications, 62, 1-16.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

Rosenthal, S., Farra, N., & Nakov, P. (2019). SemEval-2017 task 4: Sentiment analysis in Twitter. arXiv preprint arXiv:1912.00741.

Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. (2011). Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38(12), 14799-14804.

Schutze, H., Manning, C. D. & Raghavan, P. 2008. Introduction to information retrieval, Cambridge University Press Cambridge.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34, 1-47.

Seyyedi, S. H., & Minaei-Bidgoli, B. (2017). Enhancing effectiveness of dimension reduction in text classification. International Journal on Artificial Intelligence Tools, 26(03), 1750008.

Tripathi, G., & Naganna, S. (2015). Feature selection and classification approach for sentiment analysis. Machine Learning and Applications: An International Journal, 2(2), 1-16.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57, 117-126.

Tu, Z., He, Y., Foster, J., Van Genabith, J., Liu, Q., & Lin, S. (2012). Identifying high-impact sub-structures for convolution kernels in document-level sentiment classification. Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 338-343.

Uchyigit, G. Experimental evaluation of feature selection methods for text classification. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012. IEEE, 1294-1298.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management, 625-631.

Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Information Processing & Management, 52(1), 36-45.

Zheng, Z., Wu, X. & Srihari, R. 2004. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter, 6, 80-89.