Twitter Text Sentiment Analysis: A Comparative Study on Unigram and Bigram Feature Extractions

Biraj Lahkar [1], Dr. Jaibir Singh[2]

[1]*Research Scholar, Department of Computer Science, OPJS University, Churu, Rajasthan, India*
[2]*Associate Professor, Department of Computer Science, OPJS University, Churu, Rajasthan, India*

Abstract - *A new information source for data mining techniques has emerged: Twitter. An important information source for gauging public opinion on topics ranging from politics to current fashion trends is tweets. Citizens can express their ideas on social media platforms without danger. Due to the viral nature of social media, there is an increasing feeling of urgency to grasp public opinion. Making sense of these crowd dialogues is necessary for engaging in meaningful discourse. In a research called sentiment analysis (SA), sentiments are calculated for a conclusion. SA is applicable wherever as the public's opinion on a range of topics may be gauged. Since a few decades ago, a vast quantity of data has been produced online, and it is growing quickly. Manually categorizing brief text documents from both online and offline sources have grown more challenging. In this context, we are examining words with unigrams and bigrams feature extraction for sentiment. In this article, we look at the issue of classifying the emotion of English Twitter posts using machine learning methods. We use various feature selection strategies in this study, including Chi-square (CHI), Information Gain (IG), Symmetric Uncertainty (SU), Correlation-based Feature Selection (CFS) and Gain Ratio (GR). The classification is performed using Naïve Bays Multinomial (NBM), 5-NN,* Sequential Minimal Optimization *(SMO) and REPTree, provided by weka tool. We also look into the finest feature for gleaned thoughts from reviews. Based on each classifier's output, the results for unigram and bigram features were compared. In comparison to unigram features experimental findings indicated that bigram features achieved the highest accuracy of 85.83% with 5-NN algorithm.*

Key Words: *Data Mining; Feature Selection; Unigrams; Bigrams; Feature Extraction.*

1. INTRODUCTION

A significant amount of data is produced each day in the global and digitalized world due to the rising usage of community, social networking, and microblogging websites and portals. Today's world has made using the internet easier. It has fundamentally altered how individuals see and react to the daily happenings and problems. Through online discussion, social media posts, and other means, people may exchange ideas and stay in touch. Deep learning has increased the popularity of twitter sentiment analysis. Microblogging websites are crucial for assembling vast amounts of information. On social networking sites, millions of people express their thoughts on a range of subjects. These microblogging platforms force users to be succinctly expressive in their remarks or ideas because to the 280-character message constraint [1, 12, 7]. Twitter is a social networking platform with approximately 200 million members, of which more than half are active. More over half of twitter users who log in daily send out more than 200 million tweets [6]. Tweets that may be evaluated represent the opinions of the general public. These publicly voiced opinions are crucial to businesses seeking feedback on their goods, to politicians seeking to anticipate election outcomes, and to investors seeking to forecast stock prices. In this studies used unigrams, bigrams feature extraction to categorise attitudes as positive and negative [10]. Along with microblogging traits, they retrieved lexical features and mechanically categorised Twitter sentiments. Positive or negative messages were assigned to the messages. Their architecture comprised two independent parts, classifiers and feature extractors, which used machine learning methods for sentiment analysis to attain improved accuracy. As a result, sentiment analysis of user tweets can be useful in a variety of contexts. It is nearly hard to manually extract such valuable information from this vast amount of data. Positive and negative emotions may all be classified, which aids in determining how the general population feels about certain issues. The objective of this study is to identify feelings from tweets as precisely as feasible [12]. In this study, the training data for each tweet has a class label. Following the application of several classifiers to the training dataset, including Naïve Bayes Multinomial (NBM), 5-NN, SMO, and REPTree, the model is then fed the testing tweets. Thus, with the aid of trained classifiers, the tweets are divided into positive and negative. Our goal is to evaluate various classifiers' performance using the twitter dataset. This work provides a unigrams and bigrams feature selection strategy for categorising text sentiment data that takes these concerns into consideration. Filter-based feature selection approaches, such as Chi Square (CHI), Information Gain (IG), Symmetric Uncertainty (SU), Correlation Based Feature Selection (CFS) and Gain Ratio (GR) have been effectively used because to their simplicity and relatively good performance [3, 4, 5]. The outcomes of the experiments suggest that the unigrams and bigrams may identify resilient and useful features [8, 45].

The main contribution so for this study is highlighted as follows:

i. Use the top-ranked features from the twitter dataset to compare the categorization accuracy by using unigram and bigram feature extraction.

ii. Using the twitter dataset, propose a framework for twitter text sentiment analysis based on unigram and bigram features.

iii. Based on our dataset, examine the effectiveness of four classification algorithms for sentiment analysis of tweets [13].

The following is how the paper is set up: the remainder of this work is divided into the following sections: Section 2 described the Literature Review. The methodology is presented in Section 3. The experimental setup described in Section 4. The Experimental Results and Discussion is presented in Section 5. The conclusion is presented in Section 6.

2. LITERATURE REVIEW

Why feature extraction techniques are necessary: In order to create results for the test data, machine learning algorithms learn from a predefined set of features from the training data. However, the primary issue with language processing is that machine learning techniques cannot be used to directly handle raw text. So, to turn text into a matrix (or vector) of features, we need certain feature extraction algorithms. Among the most often used techniques for feature extraction are: Bag-of-Words and TF-IDF.

Bag of Words: Bag-of-Words is one of the most basic processes for converting tokens into a collection of features is the use of words. Each word is utilised as a feature to train the classifier in the BoW model, which is used to categorise documents. For instance, the presence of terms like "fantastic" and "great" in a sentiment analysis task based on reviews suggests a favourable evaluation, whereas the presence of phrases like "annoying" and "bad" indicates a negative assessment. A BoW model is created in three steps:

a) text-preprocessing

b) create a vocabulary

c) create a matrix of features

The order in which words appear is lost while using this approach since we generate a vector of tokens in a randomized manner. However, by taking into account N-grams (mainly bigrams) rather than individual words, we can fix this issue (i.e. unigrams).

TF-IDF Vectorizer: Term frequency-inverse document frequency is referred to as TF-IDF. It draws attention to a particular problem that might not come up often in our corpus but is really significant. The TF-IFD score rises in direct proportion to the frequency of a word in the document and falls in direct proportion to the number of documents in the corpus that use the term. It is divided into two smaller portions, which are:

a) Term Frequency (TF)

b) Inverse Document Frequency (IDF) [20]

Recent years have seen a dramatic surge in the study of sentiment analysis (SA). The goal of SA is to categorise a text's emotion into positive or negative polarity. The necessity for the industry to understand consumer opinions on their products via internet portals, blogs, discussion forums, and reviews, among other sources, is the driving force behind SA research. For a better sentiment classification method, effective features must be extracted [12].

Sentiments encompass a wide range of emotions, but in this study, they are only specifically referred to as positive and negative. These labels serve as the building blocks for the sentiment analysis field, which broadens to include automatically assigning these labels to texts based on one's knowledge and beliefs. On twitter, for instance, sentiment analysis has been used to ascertain the tone of discussions [2]. However, the terminology used in news story titles and social media articles differs. Unlike news story headlines, which are shorter and more professional, communication on social media is frequently written casually and as lengthier sentences. Texts can be categorised on several textual levels, from single words and brief sentences to whole manuscripts. Unsupervised machine learning methods have shown promise for document level categorization [37, 22]. When they concentrated mainly on the subjective aspects of the texts, supervised ones had fared well [19]. Non-machine learning techniques for sentence level categorization have been developed. Each word can be given a polarity from an enlarged WordNet [17] that was started with a small number of core polarised terms [9] by part-of-speech tagging sentences. Then, the word polarities are joined to create the phrase polarity. The categorization of phrases and words is frequently done using pre-made lists of terms that have been given a polarity [2]. For example the polarities are then modified to meet the situation, taking negations and expletives.

Feature selection is a crucial step in the algorithm training process. Algorithms must be taught based on the features. Feature selection aims to choose an ideal subset of characteristics by removing features that are unnecessary or provide no more information than those in the ideal subset. Forman [16] said that a variety of accessible feature selection strategies may be employed to eliminate superfluous features while enhancing classifier performance. Lee and Pang Pang and Lee [18] effectively categorise papers with accuracy using sentiment data such as "thumbs up" or "thumbs down." Guyon and Elisseeff [21] showed that the decrease of over fitting is a contributing factor in performance gains brought on by feature selection. Word polarity based on previous probability was used by Kouloumpis et al. [8] as an extra feature. In order to outperform unigrams in terms of accuracy, Saif et al. [36] looked into sentiment-topic characteristics and semantic features that may be employed in addition to unigrams. Sentiment categorization also takes into account emotions. Emotional tweets are interpreted as good feelings whereas negatively interpreted tweets are interpreted as positive sentiments. R. Bhayani and L. Huang [23] implement the methods for these. In order to categorise tweets and include sentiment analysis classifier capabilities into web applications, Go, R. Bhayani, and L. Huang [23] looked at the twitter API. The best results for classifying tweets as subjective or objective were obtained, according to Chamlertwate et al. [25], by combining SVM with IG. However, they did not specify how many features were used or which other classifiers were examined. In a comparable area of sentiment classification for movie reviews, Narayanan et al. [26] carried out an experiment illustrating the value of using feature selection; however they only examined a single ranker, mutual information, using Naïve Bayes. Kouloumpis et al. [8] look at the categorisation of emotion on twitter. They employ N-gram features to capture information about the informal and creative language used in microblogging, such as emoticons, abbreviations, and the presence of intensifiers. These features also contain a sentiment lexicon and part of speech characteristics. Their research demonstrates that qualities related to parts of speech really reduce performance. Additionally, they assert that components from an existing sentiment lexicon combined with micro blogging features were relatively beneficial. In this study, we make use of the manually categorised twitter data set. With N grams, we employ the salient feature selection approach. We evaluate the performance of the classification algorithms by presenting various feature selection methods [15].

In earlier research for sentiment categorization using machine learning algorithms, N-gram and tag-based features were often employed [14, 27, 28]. N-gram features are words where the letter “N” denotes the number of words in the feature. Words in tag-based features have their Part-of-Speech (POS) or Sentiwordnet scores assigned to them. The tags may be utilised both in conjunction with features and just for feature selection. Unigrams, bigrams, and adjectives were employed as features in Pang et al. [29] machine learning-based sentiment analysis. On a dataset of movie reviews, the authors utilised SVM, Maximum Entropy, and Naïve Bayes for classification. When employed with unigrams, binary weighting provided more accuracy than term frequency, while SVM provided the highest level of precision. For the purpose of creating the feature vector, Pak and Paroubek [30] employed subgraphs that were taken from the dependency tree of a phrase that had been parsed. They experimented with movie reviews and found that the SVM classifier with subgraph-based features provided the greatest results. On the movie review dataset, Nguyen et al. [31] combined novel rating-based features with unigrams, bigrams, and trigrams to increase the precision of document-level sentiment analysis. Hung and Alfred [61] employed phrase-based features like bigrams and trigrams, as well as word-based features like unigrams, POS-based features and sentiwordnet features. Massive feature sets present a challenge for machine learning-based sentiment categorization. Feature selection techniques are utilised to get rid of the unnecessary and redundant characteristics, which helps sentiment analysis perform better in terms of accuracy and execution time. To decrease the size of the feature vector, several researchers have worked on feature selection techniques as Document Frequency, Mutual Information, and Information Gain [11, 16, 32, 33]. Fisher's discriminant ratio is a brand-new feature selection method that Wang et al. developed [16]. It is based on boolean and frequency data. According to the trials, IG utilising an SVM classifier performs worse than frequency-based Fisher's discriminant ratio. The hybridised IG and rough set-based feature selection strategy for sentiment analysis was put out by Agarwal and Mittal [11, 34].

The most popular method in the literature for handling twitter sentiment analysis is to use machine learning algorithms [8, 23, 35, 36, 38]. Go et al., 2009 [39] groundbreaking work in this area involved the use of emoticons as noisy signals to acquire annotated training data. To create two-class sentiment classifiers, they experimented with SVM, Naïve Bayes, and Maximum Entropy classifiers and employed unigrams, bigrams, and POS tags as features. Their most accurate classifier, Maximum Entropy, had an accuracy of 83.0% and was trained on a mix of unigrams and bigrams. Their findings suggest that mixing various n-gram feature levels improves the performance of the classifiers. The effectiveness of POS characteristics for classifying twitter sentiment has been highly contested in the literature. Pak and Paroubek (2010) [40] demonstrated that the distribution of POS tags is not uniform among different sentiment classes, which suggests that they can be used as discriminating features to train sentiment classifiers. Go et al. (2009) [39] and Kouloumpis et al. (2011) [8] concluded that POS features are not at all useful for classifying twitter sentiment. In the second strategy, POS tags were used to train a Naïve Bayes classifier using n-gram features. They did not, however, look into the extent to which adding POS elements to unigrams improved performance [37].

3. METHODOLOGY

A framework based on unigram and bigram feature extraction is proposed with the goal of improving classification results in order to conduct experiments on twitter dataset. A non-ranked feature collection is examined using 5 feature selection/evaluation procedures. There are hence five sets of ranked feature sets with various rank orderings. Figure-3.1 demonstrates each component and each of the steps that go into the proposed framework for creating an efficient pattern to identify sentiment in tweets. The first stage is gathering twitter data from the publicly accessible website https://www.kaggle.com. For analysis, unigram and bigram features were generated from this data by preprocessing and annotation. Five feature selection techniques are applied to the retrieved unigram and bigram features from the dataset, and the top five feature sets with various rank orderings are chosen. Then, results are produced using 4 distinct classification algorithms. The results are displayed depending on how well the chosen classification algorithms performed.

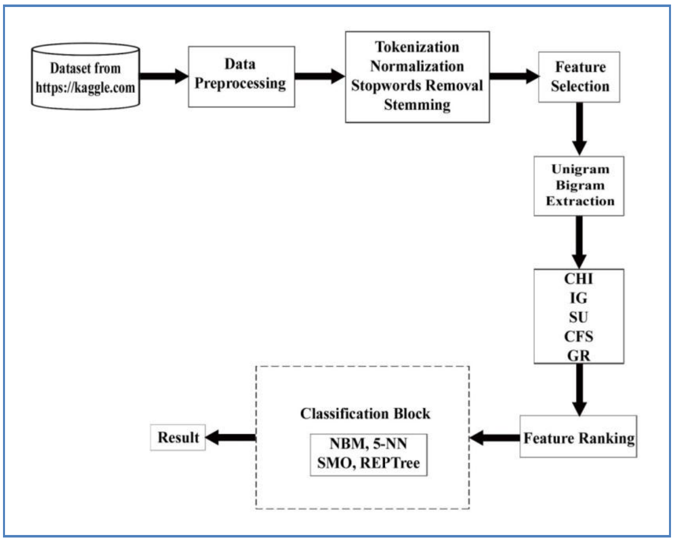


Figure-3.1: Proposed framework for tweets sentiment analysis

3.1 N-grams feature extraction:

N-grams are continuous word, symbol, or token sequences in a text. They may be described technically as the adjacent groups of items in a document. They are relevant for doing NLP (Natural Language Processing) activities on text data. N is merely a variable in N-grams that may take on positive integer values like 1, 2, 3, and so on. Basically, "N" stands for many.

Depending on the value that 'N' takes, the following sorts of N-grams is categorized.

N	Term
1	Unigram
2	Bigram
3	Trigram
N	N-gram

It is referred to as a unigram when N=1, as is shown in the table above. It is referred to as a bigram when N=2, and so on.

For example the sentence: “That’s a nice picture”

Sl. No.	Types of N-gram	Generated N-grams
1.	Unigram	[“that’s”, “a”, “nice”, “picture”]
2.	Bigram	[“that’s a”, “a nice”, “nice picture”]
3.	Trigram	[“that’s a nice”, “a nice picture”]

The terms "unigram" and "bigram" and "trigram" respectively denote taking one word at a time, two words at a time and three words at a time. In this work, we shall only use bigrams up to a certain point.

However, it is crucial to take the time to learn about the ins and outs of this notion rather than passing it by as terminology since it will serve as the basis for comprehending more sophisticated natural language processing tools and procedures.

The number of N-grams for sentence K would be as follows if X=Num of words in a particular sentence K:

$$N_{grams\ K} = X - (N\text{-}1)$$

A string of N words or characters is referred to as an "N-gram" simply. In text mining and activities involving natural language processing, N-grams of texts are frequently employed. They are essentially a collection of words that often appear in a certain window, and while computing the N-grams, we usually advance one word (although we can move X words forward in more advanced scenarios). N-grams are employed for a wide range of purposes. For instance, N-grams are utilised to create bigram and trigram models in addition to unigram models when creating language models. The creation of web scale N-gram models by Google and Microsoft allows for a range of activities, including text summarization, word splitting, and spelling correction. The development of features for supervised Machine Learning models like SVMs, MaxEnt models, Naïve Bayes, etc. is another use of N-grams. Instead of only using unigrams in the feature space, the goal is to incorporate tokens like bigrams [41, 42, 43].

3.2 Feature selection methods:

Feature selection is the process of selecting an acceptable feature subset from a data collection so that classification algorithms can effectively deal with high-dimensional feature spaces. By removing redundant or unnecessary information, feature selection algorithms aim to reduce the training time needed to create a classification model [44]. While filter-based techniques assess the value and utility of features based on heuristics and assessment metrics, wrapper-based approaches choose features based on the performance of a machine learning algorithm to improve prediction performance. The two types of filter-based feature selection strategies are individual feature measures and group feature measures [45]. Individual feature measures evaluate the worth of traits using a particular evaluation metric. Based on the significance of this statistic, a ranking of the traits is established. Group feature measurements are used to evaluate the value of feature subsets. In terms of running time, individual feature measurements are more effective than group-based measures. This section provides a brief description of each of the filter-based measures utilised in the framework [12].

3.2.1 Chi -square (CHI):

Chi Square is a feature selection technique that performs quite well, particularly with multi-class data [46]. The technique has been applied in a variety of contexts, including the classification of tumours detection of network intrusions, text categorization, illness diagnosis, and others [45, 46, 47, 48, 49]. Chi Square calculates the statistical value shown in equation (2) to determine the strength of each feature's association [50]. The formula for Chi-square Test is:

$$X_c^2 = \frac{\sum (O_i - E_i)^2}{E_i} \quad \text{……………. (2)}$$

Where, ***c*** = Degrees of freedom, *O* = Observed Value, *E* = Expected Value

Additionally, using the X_c^2 value in conjunction with the Chi Square distribution table, it is possible to compute the correlation of the significant value. The feature has a strong relevance in the data, or is an important feature, if the signed value is less than a crisis point, which is 0.05 [50].

3.2.2 Information gain (IG):

IG is a filter approach. IG is a classifier agnostic, it may be used with many different classifiers. Based on a certain class, information gain can identify the feature(s) with the greatest information. The likelihood of an event or attribute is used to calculate the entropy, which is a measure of a class's uncertainty. It has a negative relationship to IG. A typical metric for assessing how well a word may be used for classification based on the information it can provide to discriminate across classes is Information Gain (IG) [14]. It serves as a measure of the amount of information that a sentence includes [51]. The formula is shown below [52]. It is an entropy-based method for determining impurity for feature values.

$$I(Y;X) = H(X) + H(Y) - H(X,Y)$$

X and *Y's* combined entropy is *H(X, Y),* where,

$$H(X,Y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} P(X = x_j, Y = y_i)\log_2 P(X = x_j, Y = y_i)$$

When the predictive variable *X* is continuous rather than discrete, the information gain of the corresponding class attribute *Y* is calculated by taking into account all potential binary characteristics, $X\theta$ that originate from *X* when a threshold θ is set on *X* . θ takes values from all of *X*'s values. The information gained is then simply: [43]

$I(Y;X) = \text{argmax}\ X\theta\ I(Y;X\theta)$

3.2.3 Symmetrical uncertainty coefficient:

In order to assess redundancy, symmetrical uncertainty (SU) was defined:

$$IG(X|Y) = E(X) - E(X|Y) \quad \text{………. (3)}$$

$$SU(X,Y) = 2 \times \frac{IG(X|Y)}{E(X) + E(Y)} \quad \text{……………… (4)}$$

Where, IG(X|Y) is the information gained by X after viewing Y, and E(X) and E(Y) are the entropies of features X and Y, respectively. To gauge the correlation between features, C-correlation and F-correlation are defined based on SU. The C-correlation, shown as $SU_{i,c}$, is the SU between any feature F_i and the class C. The SU between any two features F_i and F_j ($i \neq j$), represented by $SU_{i,j}$, is known as the F-correlation [54].

3.2.4 Correlation-based Feature Selecton (CFS):

Since the correlation-based feature selection (CFS) technique is a filter method, it is unrelated to the chosen classification model. As implied by the name, correlations, it exclusively analyses feature subsets based on intrinsic data characteristics. Finding a feature subset with low feature-feature correlation, which prevents redundancy, and high feature-class correlation, which preserves or boosts predictive power, is the objective.

To do so, the method uses the following equation to estimate the worth of a subset *s* with *k* features:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Where, bar r_{ff} is the average feature-feature correlation, bar r_{rf} is the average feature-class correlation and k is the number of features of that subset [55].

3.2.5 Gain Ratio (GR):

Gain Ratio introduces a balancing factor called the Intrinsic Information in an effort to reduce the bias of Information Gain on heavily branching predictors. The entropy of sub-dataset proportions is referred to as the intrinsic information. In other words, it refers to how difficult it is for us to determine which branch a randomly chosen sample is placed in [53].

The Gain Ratio is:

$$\text{Gain (Attribute)} = \frac{\text{Informaton Gain}}{\text{Intrinsic information}}$$

3.3 Classification Algorithms:

Four popular classification algorithms were looked at to gauge the efficiency of FS approaches during the classification process. Naïve Bayes Multinomial (NBM), 5-NN (k-nearest Neighbors classifier with k = 5; referred to as 5-NN in this work), Sequential Minimum Optimization (SMO) and REPTree are all used unless otherwise stated. The WEKA tool was used to create each and every classifier. Support vector machines are frequently trained using SMO, which is implemented by the well-known LIBSVM utility. KNNs are used as instance-based learning classifiers. NBM for Bayes' theorem, and REPTree is tree-based approaches. The goal of the classifier heterogeneity project is to examine the performance of various FS techniques employing unigram and bigram feature extraction on several classifiers with varied feature ranking sizes.

4. EXPERIMENTAL SETUP:

4.1 Dataset used:

For our testing, one tweet dataset, tweets.csv, was gathered from the publically accessible website https://www.kaggle.com. Obtaining data from an open dataset source involves three stages. Start by downloading the data as a CSV file from the aforementioned URL. Next, transfer the whole dataset from one CSV file to another. The dataset then has to be adjusted. Many pre-processing or data cleaning strategies were evaluated on the twitter datasets once they were collected. The tweets.csv dataset contains tweets during the time period that the number of COVID-19 cases in India increased. From the beginning of the epidemic until April 28, 2021, tweets using the hashtag "covidindia" are included in the tweets.csv dataset. Each tweet contains 10 fields: created at, user id, username, name, location, tweet, language, replies count, hashtags, and sentiment. There are 9655 tweets total. The emotion label has two possible values: NEGATIVE and POSITIVE. Unigram (one word inside a tweet's text) and Bigram (two words inside a tweet's text) were retrieved as features. After pre-processing the data and extracting it for our experiment, we had a dataset with 5054 instances and 1956 and 9932 features or attributes for Unigram and Bigram, respectively.

4.2 WEKA Workbench:

WEKA, a machine learning workbench, was utilised to design and evaluate our experiments. The University of Waikato in New Zealand offers the free service known as WEKA, or the "Waikato Environment for Knowledge Analysis." A variety of features and a user-friendly interface are available in this workbench for developing and analysing machine learning models [56]. The automatic evaluation of essays is one of the many uses for these models. All of the research was conducted using a laptop model HP 15-r006TU. The laptop has an Intel(R) Core(TM) i3 - 4010U processor clocked at 1.70 GHz and 4 GB of RAM, however WEKA workbench is only configured to use 1 GB. Windows 7 64 bit is the laptop's operating system [57].

4.3 Evaluation Measure:

We evaluate our algorithm using the following criteria:

4.3.1 Classification accuracy:

There are numerous approaches to calculate a classifier's accuracy, which is the likelihood of accurately guessing the class of an unlabeled instance (Baldi et al., 2000) [58]. Classification accuracy is defined as the proportion of cases that a particular classifier properly classified, or as the number of correctly classified reviews to the total number of reviews. It's expressed as a percentage. Four different classifiers were used to analyse and record the classification accuracy of this feature subset for the evaluation. Here is how classification accuracy is defined:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

In where TP stands for "True Positive," TN for "True Negative," FP for "False Positive," and FN for "False Negative".

4.3.2 k-Fold Cross Validation:

One of the most prominent techniques frequently employed by data scientists is k-fold cross-validation. It is a method of data partitioning that enables us to make the most use of our information to create a more comprehensive model. Any type of machine learning has as its major goal the creation of a broader model that can function well with unknown input. On the training data, a perfect model can be created with 100% accuracy or 0 errors, but it may not generalise to new data. As a result, it is a poor model. The training data are overfit by it. Machine learning is all about generalisation; hence the performance of the model can only be evaluated using data points that were not utilised in the training phase. Because of this, we frequently divide our data into training and test sets. The procedure of data splitting can be carried out more successfully with k-fold cross-validation. We employed 10-Fold Cross Validation in our work [59, 60]. The fitting operation would be carried out ten times using 10-fold cross validation, with each fit being made on a training set made up of 90% of the total training set randomly chosen and the remaining 10% serving as a hold out set for validation.

5. EXPERIMENTAL RESULT AND DISCUSSION:

By contrasting unigram and bigram feature extraction, we assessed the effectiveness of our feature selection method in terms of classification accuracy using five feature selection strategies. Refer to Tables 5.1 and 5.2 for our classification accuracy results using unigram and bigram feature extraction across our dataset, respectively. For each of the feature rankers, we used one of five top ranked feature subset sizes: 10, 50, 100, 200, and 400. The selection of these metrics was made to account for a range of feature subset sizes. Bigram feature extraction has higher classification accuracy with five feature choices in our dataset than unigram feature extraction. The best model for each column's feature subset size in Tables 5.1 and 5.2 is boldfaced. By utilising all 1956 and 9932 features that are accessible in our dataset, for unigram and bigram feature extraction respectively, this makes it possible for unigram and bigram feature extraction to be successful.

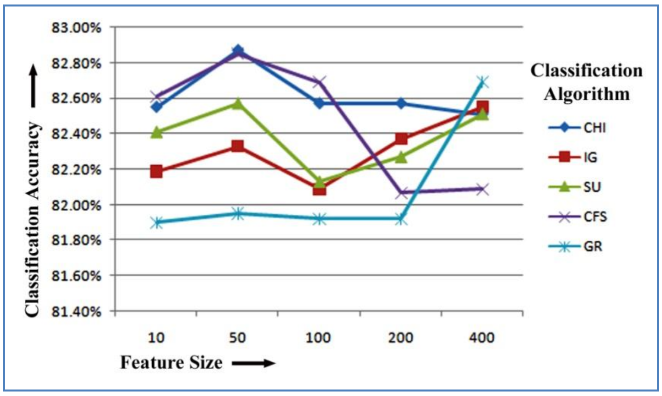
Table-5.1: Classification Accuracy Results for unigram feature extraction

Classifier	Ranker	Feature Subset Size				
		10	50	100	200	400
NaïveBayes Multinomial	CHI	82.55%	**82.87%**	82.57%	**82.57%**	82.51%
	IG	82.19%	82.33%	82.09%	82.37%	82.55%
	SU	82.41%	82.57%	82.13%	82.27%	82.51%
	CFS	**82.61%**	82.85%	**82.69%**	82.07%	82.09%
	GR	81.90%	81.95%	81.92%	81.92%	**82.69%**
5-NN	CHI	82.07%	81.97%	**82.17%**	**82.37%**	**82.53%**
	IG	81.94%	81.80%	81.54%	81.94%	82.35%
	SU	82.15%	81.97%	81.95%	81.90%	82.25%
	CFS	**82.39%**	**82.17%**	81.88%	81.84%	82.21%
	GR	82.01%	81.99%	82.03%	82.13%	82.41%
SMO	CHI	82.56%	82.75%	82.83%	82.67%	82.35%
	IG	82.31%	82.57%	82.77%	**82.77%**	**82.39%**
	SU	82.53%	82.69%	82.81%	82.75%	82.35%
	CFS	**82.59%**	**82.88%**	**82.85%**	82.57%	81.94%
	GR	82.01%	82.21%	81.95%	81.78%	82.27%
REPTree	CHI	82.21%	82.29%	**82.39%**	82.15%	82.13%
	IG	82.17%	82.25%	82.05%	82.21%	82.23%
	SU	82.11%	**82.41%**	82.15%	82.27%	82.07%
	CFS	**82.27%**	82.37%	82.23%	**82.31%**	**82.31%**
	GR	81.92%	81.92%	81.95%	81.86%	82.13%

5.1 Classification Accuracy for unigram feature extraction:

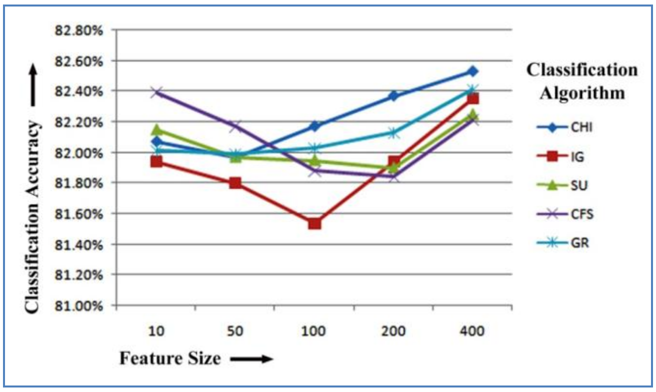
We'll start by looking at the Naïve Bayes Multinomial results. It can be proven that CFS, when trained using NBM, has the highest classification accuracy of 82.61% for the top 10 features and 82.69% for the top 100 features for unigram feature extraction. Once more, while training with NBM on 50 and 200 top features for unigram feature extraction, respectively, CHI has the maximum classification accuracy of 82.87% and 82.57%. For unigram feature extraction train using NBM, GR has the maximum classification accuracy of 82.69% for the top 400 numbers of features. Figure-5.1 displays the Naïve Bayes Multinomial Classifier-based unigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.1: Classification Accuracy comparison of five Feature selection Methods for unigram feature extraction based on Naïve Bayes Multinomial (NBM) Classifier



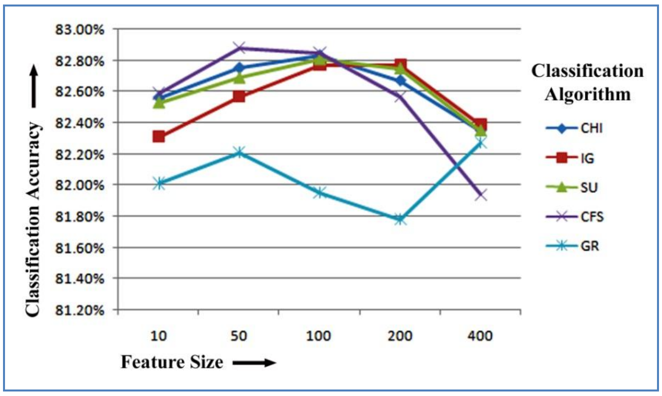
We can see from the 5-NN classifier's results. It can be proven that CFS, when trained using 5-NN, has the highest classification accuracy of 82.39% for the top 10 features and 82.17% for the top 100 features for unigram feature extraction. Once more, while training with 5-NN on 50, 100 and 200 top features for unigram feature extraction, CHI has the maximum classification accuracy of 82.17%, 82.37% and 82.53% respectively. Figure-5.2 displays the 5-NN Classifier-based unigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.2: Classification Accuracy comparison of five Feature selection Methods for unigram feature extraction based on 5-NN Classifier



It is clear from the SMO results that CFS, when trained using SMO, has the highest classification accuracy of 82.59%, 82.88% and 82.85% for the top 10, 50 and 100 features for unigram feature extraction. Once more, while training with SMO on 200 and 400 top features for unigram feature extraction; IG has the maximum classification accuracy of 82.77%, 82.39% respectively. Figure-5.3 displays the SMO Classifier-based unigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.3: Classification Accuracy comparison of five Feature selection Methods for unigram feature extraction based on SMO Classifier



Finally, we can see from the REPTree classifier's results that CFS has the highest classification accuracy of 82.27%, 82.31% and 82.31% for the top 10, 200 and 400 features for unigram feature extraction. Again, while training with REPTree classifier on 50 numbers of top ranked features for unigram feature extraction, SU has the maximum classification accuracy of 82.41%. CHI has the highest classification accuracy of 82.39% when selecting top 100 numbers of features. Figure-5.4 displays the REPTree Classifier-based unigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.4: Classification Accuracy comparison of five Feature selection Methods for unigram feature extraction based on REPTree Classifier

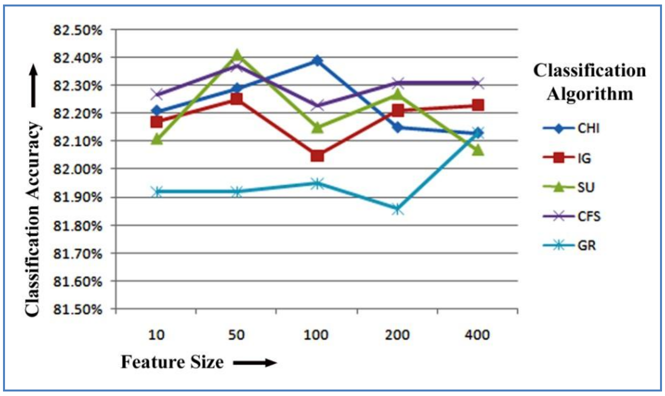


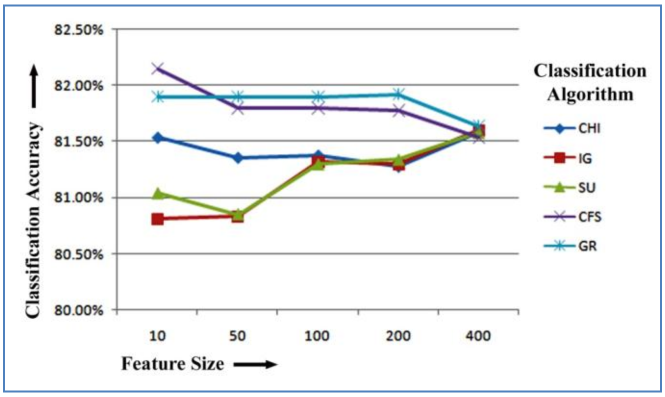
Table-5.2: Classification Accuracy Results for bigram feature extraction

Classifier	Ranker	Feature Subset Size				
		10	50	100	200	400
NaïveBayes Multinomial	CHI	81.54%	81.36%	81.38%	81.28%	81.58%
	IG	80.81%	80.83%	81.32%	81.30%	81.60%
	SU	81.04%	80.85%	81.30%	81.34%	81.58%
	CFS	**82.15%**	81.80%	81.80%	81.78%	81.54%
	GR	81.90%	**81.90%**	**81.90%**	**81.92%**	**81.64%**
5-NN	CHI	84.78%	85.64%	85.71%	**85.77%**	85.56%
	IG	84.80%	**85.83%**	85.75%	85.73%	**85.65%**
	SU	**84.84%**	**85.83%**	**85.77%**	85.73%	**85.65%**
	CFS	83.48%	83.34%	83.72%	83.66%	85.25%
	GR	82.07%	81.95%	81.93%	81.92%	85.58%
SMO	CHI	82.13%	81.94%	81.78%	81.48%	**81.28%**
	IG	81.95%	**82.05%**	81.72%	81.70%	**81.28%**
	SU	**82.15%**	**82.05%**	81.70%	81.70%	**81.28%**
	CFS	82.11%	81.78%	81.78%	**81.76%**	81.24%
	GR	82.07%	81.80%	**81.90%**	81.50%	81.26%
REPTree	CHI	84.45%	85.04%	85.02%	85.04%	85.34%
	IG	84.35%	**85.26%**	85.34%	85.32%	85.32%
	SU	**84.53%**	85.22%	**85.36%**	**85.36%**	**85.36%**
	CFS	83.60%	83.26%	83.38%	83.40%	83.72%
	GR	82.11%	81.94%	81.94%	81.92%	85.34%

5.2 Classification Accuracy for bigram feature extraction:

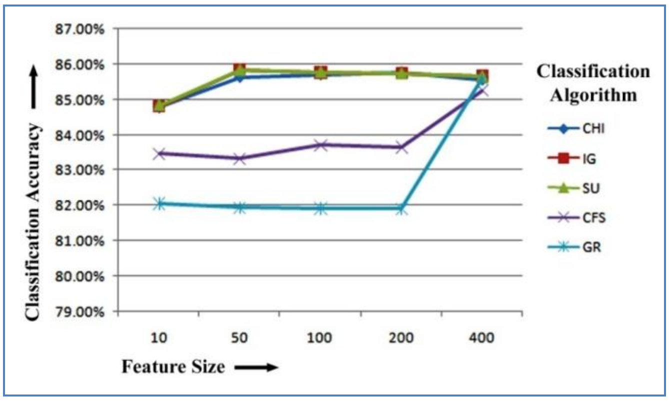
We'll start by looking at the Naïve Bayes Multinomial results. It can be proven that CFS, when trained using NBM, has the highest classification accuracy of 82.15% for the top 10 features for bigram feature extraction. Once more, while training with NBM on 50, 100, 200 and 400 top ranked features for bigram feature extraction, GR has the maximum classification accuracy of 81.90%, 81.90%, 81.92% and 81.64% respectively. Figure-5.5 displays the Naïve Bayes Multinomial Classifier-based bigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.5: Classification Accuracy comparison of five Feature selection Methods for bigram feature extraction based on Naïve Bayes Multinomial (NBM) Classifier



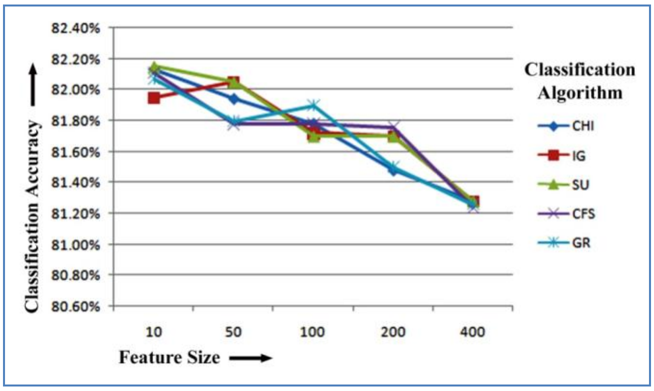
We can see from the 5-NN classifier's results. It can be proven that CFS, when trained with 5-NN, has the highest classification accuracy of 84.84%, 85.83%, 85.77% and 85.65% for the 10, 50, 100 and 400 top ranked features after bigram feature extraction. Again, while training with 5-NN on 50 and 400 top features for bigram feature extraction, IG has the maximum classification accuracy of 85.83% and 85.65% respectively. CHI has the highest classification accuracy of 85.77% when selecting 200 number of features trained with 5-NN classifier. Figure-5.6 displays the 5-NN Classifier-based bigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.6: Classification Accuracy comparison of five Feature selection Methods for bigram feature extraction based on 5-NN Classifier



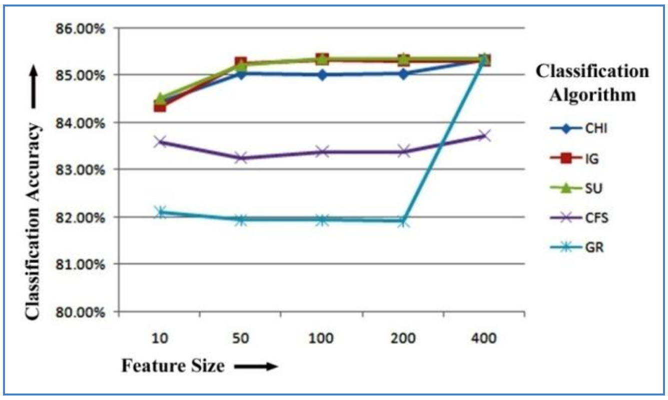
It is clear from the SMO results that CFS, when trained using SMO, has the highest classification accuracy of 82.15%, 82.05% and 81.28% for the top 10, 50 and 400 features for bigram feature extraction. Once more, while training with SMO on 50 and 400 top features for bigram feature extraction, IG has the maximum classification accuracy of 82.05%, 81.28% respectively. For top ranked 100 features GR has the highest classification accuracy of 81.90% when trained with SMO classifier. CFS has the highest classification accuracy of 81.76% when picked 200 top ranked features. CHI has the highest classification accuracy of 81.28% when picked 400 top ranked features. Figure-5.7 displays the SMO Classifier-based bigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.7: Classification Accuracy comparison of five Feature selection Methods for bigram feature extraction based on SMO Classifier



Finally, we can see from the REPTree classifier's results that CU has the highest classification accuracy of 84.53%, 85.36%, 85.36% and 85.36% for the top 10, 100, 200 and 400 features for bigram feature extraction. Again, while training with REPTree classifier on 50 numbers of top ranked features for bigram feature extraction, IG has the maximum classification accuracy of 85.26%. Figure-5.8 displays the REPTree Classifier-based bigram feature extraction results of five feature selection methods for classification accuracy.

Figure-5.8: Classification Accuracy comparison of five Feature selection Methods for bigram feature extraction based on REPTree Classifier



6. CONCLUSION:

The rise of social media, like twitter, has given individuals a free platform to communicate their thoughts and sentiments. The tremendous volumes of opinionated tweets that are produced on twitter encompass every facet of our everyday life. Effective sentiment analysis of tweets can provide high-quality information on the public's worries and preferences. However, this effort is more difficult than in other areas where the content is well-edited because of the informal and slang language used in twitter as well as the high frequency of misspellings. The vast majority of features are produced via feature engineering techniques for categorising tweet sentiment. Furthermore, training classifiers on a sizable dataset is computationally challenging. The technique of feature selection, which has received little attention in tweet sentiment classification research, chooses the best set of features, which reduces the dimensionality of the dataset, lowers computational costs, and may even improve classification. This study examined five filter-based feature selection algorithms using four distinct learners. Five different feature subsets are chosen using these methods for a twitter dataset obtained from https://www.kaggle.com.

In this study, we carefully assessed how well NBM, 5-NN, SMO and REPTree classifiers performed when given the features extracted from unigrams and bigrams. According to the findings of our experiments, adding bigrams feature extraction consistently enhances the performance of the classier feature extraction when compared to unigrams feature extraction. The highest classification accuracy for unigram feature extraction is 82.88% when trained with SMO classifier for the top 50 features. Again, when using our dataset and the REPTree classifier trained with 100, 200, and 400 top ranked features, the highest classification accuracy was 85.36%. We also looked into how bigram properties affected the effectiveness of the classification. As a result, we draw the conclusion that utilizing a bigrams feature extraction is an excellent but straightforward way to enhance twitter sentiment classifier performance, particularly if the training data is sparse. Therefore, to improve twitter sentiment performance, we recommend bigram feature extraction with REPTree classifier.

The results of our work are encouraging, and other feature selection strategies as well as the usage of more than 400 characteristics should be investigated in future research. To determine whether the patterns discovered in this study are also prevalent in other datasets, this research should be expanded to include other datasets. We'd want to incorporate stacking strategies to improve categorization performance. Bigrams and trigrams are examples of the n-gram format, whereas Ranking Aggregation methods that use different classifiers are examples of the suggested feature selection technique.

REFERENCES

[1] Xia, R., Zong, C. and Li, S., 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6), pp.1138-1152.

[2] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011. Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on Languages in Social Media*, pp.30-38.

[3] Kumar, S., Singh, P. and Rani, S., 2016. Sentimental analysis of social media using R language and Hadoop: Rhadoop. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*,.

[4] Liu, B., Blasch, E., Chen, Y., Shen, D. and Chen, G., 2013. Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier. *2013 IEEE International Conference on Big Data*,.

[5] Kotsiantis, S., 2007. Supervised Machine Learning: A Review of Classification Techniques. *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pp.3-24.

[6] Parr, B., 2011. *Twit ter Has 100 Million Monthly Act ive Users; 50% Log In Everyday*. [online] Available at: <http://mashable.com/2011/10/17/twit tercostolo-stats> [Accessed 1 August 2022].

[7] Sumaya Ishrat Moyeen, Md. Sadiqur Rahman Mabud, Zannatun Nayem and Md. Al Mamun, 2021. SENTIMENT ANALYSIS OF ENGLISH TWEETS USING BIGRAM COLLOCATION. *EPRA International Journal of Research & Development (IJRD)*, pp.220-227.

[8] Kouloumpis, E., Wilson, T. and Moore, J. (2011) "Twitter Sentiment Analysis: The Good the Bad and the OMG!", *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), pp. 538-541. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14185 (Accessed: 1August2022).

[9] Kim, S. and Hovy, E., 2004. Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*,.

[10] Hatzivassiloglou, V. and McKeown, K., 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting on Association for Computational Linguistics -*,.

[11] Agarwal, B. and Mittal, N., 2016. Prominent Feature Extraction for Sentiment Analysis. *Socio-Affective Computing*,.

[12] Geetha, S. and Kaniezhil, R., 2018. *A SVM Based Sentiment Analysis Method (SBSAM) for Unigram and Bigram Tweets*. [online] Acadpubl.eu. Available at: <http://www.acadpubl.eu/hub/> [Accessed 1 August 2022].

[13] Jimoh, R., Adewole, K., Aderemi, T. and Balogun, A., 2021. Investigative Study of Unigram and Bigram Features for Short Message Spam Detection. *International Conference on Emerging Applications and Technologies for Industry 4.0 (EATI'2020)*, pp.70-81.

[14] KALAIVANI, K. and KUPPUSWAMI, S., 2019. Exploring the use of syntactic dependency features for document-level sentiment classification. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, 67(2), pp.339-347.

[15] Awachate, P. and Kshirsagar, V., 2016. Improved Twitter Sentiment Analysis Using N Gram Feature Selection and Combinations. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(9), pp.154-157.

[16] Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, pp.1289-1305.

[17] Kilgarriff, A. and Fellbaum, C., 2000. WordNet: An Electronic Lexical Database. *Language*, 76(3), p.706.

[18] Pang, B. and Lee, L., 2004. A sentimental education. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*,.

[19] Pang, B. and Lee, L., 2004. A sentimental education. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*,.

[20] GeeksforGeeks. 2022. *Feature Extraction Techniques - NLP - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/feature-extraction-techniques-nlp/> [Accessed 1 August 2022].

[21] Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, pp. 1157–1182.

[22] Turney, P., 2001. Thumbs up or thumbs down?. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*,.

[23] A. Go, R. Bhayani, and L. Huang. (2015) Sentiment140 api. [Online].Available: http://help.sentiment140.com/api

[24] Li, X., Xie, Q., Jiang, J., Zhou, Y. and Huang, L., 2019. Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, 146, pp.687-705.

[25] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T. and Haruechaiyasak, C., 2012. Discovering Consumer Insight from Twitter via Sentiment Analysis. *Journal of Universal Computer Science*, 18(8), pp.973-992.

[26] Narayanan, V., Arora, I. and Bhatia, A., 2013. Fast and Accurate Sentiment Classification Using an Enhanced Naïve Bayes Model. *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, pp.194-201.

[27] Hung, L. and Alfred, R., 2017. A Performance Comparison of Feature Extraction Methods for Sentiment Analysis. *Advanced Topics in Intelligent Information and Database Systems*, pp.379-390.

[28] Sharma, A. and Dey, S., 2012. Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *International Journal of Computer Applications 3*, pp.975-8887.

[29] Pang, B. and Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), pp.1-135.

[30] Pak, A. and Paroubek, P., 2011. Text Representation Using Dependency Tree Subgraphs for Sentiment Analysis. *Database Systems for Adanced Applications*, pp.323-332.

[31] Nguyen, D., Nguyen, D., Vu†, T. and Pham, S., 2014. Sentiment Classification on Polarity Reviews: An Empirical Study Using Rating-based Features. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp.128-135.

[32] O'Keefe, T. and Koprinska, I., 2009. Feature selection and weighting methods in sentiment analysis. *Proceedings of the 14th Australasian document computing symposium, Sydney*, pp.67-74.

[33] TAN, S. and ZHANG, J., 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4), pp.2622-2629.

[34] Kalaivani, K., Kuppuswami, S. and Kanimozhiselvi, C., 2019. Use of NLP Based Combined Features for Sentiment Classification. *International Journal of*

Engineering and Advanced Technology (IJEAT), 9(1), pp.621-626.

[35] Asiaee T., A., Tepper, M., Banerjee, A. and Sapiro, G., 2012. If you are happy and you know it... tweet. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*,.

[36] Hassan, S., Yulan, H. and Harith, A., 2012. Alleviating data sparsity for Twitter sentiment analysis. *CEUR Workshop Proceedings (CEUR-WS. org).*, pp.2-9.

[37] Mohammed Badr, B. and Fatima, S., 2015. Using Skipgrams, Bigrams, and Part of Speech Features for Sentiment Classification of Twitter Messages. *In Proceedings of the 12th International Conference on Natural Language Processing*, pp.268-275.

[38] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V. and Wilson, T., 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM) Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2, pp.312-320.

[39] Go, A.; Bhayani, R. & Huang, L. (2009), 'Twitter Sentiment Classification using Distant Supervision', *Processing* , 1--6.

[40] Pak, A. and Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA)*, 10, pp.1320-1326.

[41] Analytics Vidhya. 2022. *What Are n-grams and How to Implement Them in Python?*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/09/what-are-n-grams-and-how-to-implement-them-in-python/> [Accessed 3 August 2022].

[42] Kavita Ganesan, PhD. 2022. *What are N-Grams? - Kavita Ganesan, PhD*. [online] Available at: <https://kavita-ganesan.com/what-are-n-grams/#.Ytwi6b3MLIU> [Accessed 3 August 2022].

[43] Medium. 2022. *Leveraging N-grams to Extract Context From Text*. [online] Available at: <https://towardsdatascience.com/leveraging-n-grams-to-extract-context-from-text-bdc576b47049> [Accessed 3 August 2022].

[44] DASH, M. and LIU, H., 1997. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4), pp.131-156.

[45] Ren Diao and Qiang Shen, 2012. Feature Selection With Harmony Search. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6), pp.1509-1523.

[46] Sumaiya Thaseen, I. and Aswani Kumar, C., 2017. Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences*, 29(4), pp.462-472.

[47] Zhang, H., Li, L., Luo, C., Sun, C., Chen, Y., Dai, Z. and Yuan, Z., 2014. Informative Gene Selection and Direct Classification of Tumor Based on Chi-Square Test of Pairwise Gene Interactions. *BioMed Research International*, 2014, pp.1-9.

[48] Mesleh, A.M., 2007. Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *International Journal of Computer Science*, *3*(6), pp.430-435.

[49] Dharmendra Kumar singh, Pragya Patel, Anjali Karsh, Dr.A.S.Zadgaonkar, Analysis of Generated Harmonics Due To CFL Load On Power System Using Artificial Neural Network, Volume 5, Issue 3, March (2014), pp. 56-68, International Journal of Electrical Engineering and Technology (IJEET).

[50] Bachri, O., Hatta, K. and Nurhayati, O., 2017. FEATURE SELECTION BASED ON CHI SQUARE IN ARTIFICIAL NEURAL NETWORK TO PREDICT THE ACCURACY OF STUDENT STUDY PERIOD. *International Journal of Civil Engineering and Technology (IJCIET)*, 8(8), pp.731-739.

[51] Lei, S., 2012. A Feature Selection Method Based on Information Gain and Genetic Algorithm. *2012 International Conference on Computer Science and Electronics Engineering*,.

[52] Hung, L., Alfred, R. and Ahmad Hijazi, M., 2015. A Review on Feature Selection Methods for Sentiment Analysis. *Advanced Science Letters*, 21(10), pp.2952-2956.

[53] Tung M Phung's Blog. 2022. *Information Gain, Gain Ratio and Gini Index*. [online] Available at: <https://tungmphung.com/information-gain-gain-ratio-and-gini-index/> [Accessed 3 August 2022].

[54] Piao, M., Piao, Y. and Lee, J., 2019. Symmetrical Uncertainty-Based Feature Subset Generation and

Ensemble Learning for Electricity Customer Classification. *Symmetry*, 11(4), p.498.

[55] Johfischer.com. 2022. *Correlation-based Feature Selection in Python from Scratch – Johannes S. Fischer*. [online] Available at: <https://johfischer.com/2021/08/06/correlation-based-feature-selection-in-python-from-scratch/> [Accessed 3 August 2022].

[56] Javed, K., Maruf, S. and Babri, H., 2015. A two-stage Markov blanket based feature selection algorithm for text classification. *Neurocomputing*, 157, pp.91-104.

[57] Prusa, J., Khoshgoftaar, T. and Dittman, D., 2015. Impact of Feature Selection Techniques for Tweet Sentiment Classification. FLAIRS Conference 2015, pp.1-6.

[58] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. and Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), pp.412-424.

[59] Marcot, B. and Hanea, A., 2020. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. *Computational Statistics*, 36(3), pp.2009-2031.

[60] Manorathna, R., 2020. k-fold cross-validation explained in plain English (For evaluating a model's performance and hyperparameter tuning). [online] Available at: <https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0> [Accessed 3 August 2022].

[61] Hung, L. and Alfred, R., 2017. A Performance Comparison of Feature Extraction Methods for Sentiment Analysis. *Advanced Topics in Intelligent Information and Database Systems*, pp.379-390.