Analysis Of Supervised Machine Learning Classification Algorithms onTextual Data.

Mr Suresh Kumar G[1], Sarojini Devi E[2], S Laxmi Prasanna

Abstract - *Data is divided into a number of classes that are either dependent or independent of one another through the classification process. Different classifier algorithms can be used to accomplish the classification process. However, text classification—the application of classification to textual data (brief tweets)—cannot be done simply. This essay's major goal is to investigate the multi-class text classification task and discover ways to get high classification accuracy while dealing with text. Text classification is dealt with using the Naive Bayes technique. Additionally, it performs better when classifying texts using several labels. Although categorizing text data takes a lot of effort, it is necessary for automatic text classification. the use of conventional classification techniques*

***Key Words*: Multi- class text classification, Multinomial Naïve Bayes, Linear support vector machine, Uni-grams, bi-grams, confusion matrix.**

1. INTRODUCTION

The practice of grouping brief texts into groups or classes is known as text data categorization. It is well known that supervised learning algorithms may "learn" to carry out text-categorization tasks utilizing input training data. For instance, a brief phrase might be categorized as "other" or "spam" in relation to a predetermined class. There is a "multi-class" classification difficulty since many textual data sources, like Internet news feeds, email, and digital libraries, contain many themes or classes. A small paragraph may also be applicable to multiple different classes in multi-class situations. A news story might be pertinent to "politics class" and "business class," for instance. One method for categorizing tiny texts with multiple classes and labels is to

classify all of the data into distinct categories, one

1.1 Feature extraction using TF-IDF:

TF-IDF stands for term frequency-inverse document frequency, and also the tf-idf weight could be a weight usually utilized in data retrieval and text mining. This weight could be an applied math live wont to judge however necessary a word is to a document during an assortment or corpus. The importance will increase proportionately to the quantity of times a word seems within the document however is offset by the frequency of the word withinthe corpus.

Variations of the tf-idf weight theme are usually employed by search engines as a central tool in grading and ranking a document's connection given a user question. Tf-idf are often with success used for stop-words filtering in numerous subject fields together with text report and classification

TF: The acronym for TF is Term Frequency, which measures how frequently a term/word occurs in a document. Since every document is different in length, it is possible that a term/word would appear more times in long documents when compared shorter ones. Thus, the term frequency is often divided by its document length.

ISSN 2454 - 7239

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

IDF: Inverse Document Frequency

This measures the value which tells how important a term is. While computing Term Frequency, all the terms are given equal importance. Most of the cases it is known that some terms, such as "is", "of", and "that", may appear a lot of times and actually have the least importance.Depending on the precise behaviour of the probability model, the naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification. Because variables are considered to be independent variables, the main focus has to be on the variances of the variables for each category and need to be determined and not the entire covariance matrix. Because of the simplified assumptions made by the naïve Bayes classifier, the naïve Bayes algorithm mostly offers better results in many difficult real-world scenarios than an individual might think off. The naïve Bayes classifiers had been recorded to perform surprisingly well for many real-time use cases of classification problems under some specific conditions.

$$idf(w) = log(\frac{N}{df_t})$$

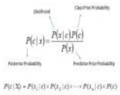
1.2 Feature extraction using N-grams

An N-gram model predicts the occurrence of a word based on the occurrence of its n–1 previous words. So here we are answering the question how far back in f a sequence of words should we will predict the next word? For instance, a bigram model (n = 2) predicts the occurrence of a word given only its previous word (as N – 1 = 1 in this case). Similarly, a trigram model (n = 3) predicts the occurrence of a word based on its previous 2 words (as N – 1 = 2 in this case).

2. Supervised Multiclass classification Algorithms

2.1 Naïve Bayes classifier

Naïve Bayes classification algorithm is a basic probabilistic classifier built using Bayes' Theorem with robust independence assumptions. In other words the underlying probability model can be independent feature model. These independence assumptions of features of the model make the algorithm independent of the order in which features occur and most importantly the presence of one feature does not affect the presence of other feature during the classification process. These assumptions add up to the efficiency of Bayesian classification algorithm, but this assumption may limit the algorithms applicability.


Likelihood
Class Prior Probability
$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

The main advantage of the naïve Bayes classifier is that it requires a small textual data set as training data to estimate the required features which are necessary for classification. Bayesian classification approach arrives at the right classification as long as the correct class is more likely to occur than the others. Category's

ISSN 2454 - 7239

probability values need not be estimated very accurately. To be clearer, the overall classifier is strong enough to ignore serious deficiencies in its underlined base naïve probability model.

The main disadvantage of the naïve Bayes classificatioproach is its relatively low classification performance compare to other discriminative algorithms, such as the SVM with its

outperformed classification effectiveness. Hence, several active types of research have been carried out to show and highlight the reasons that the naïve Bayes classification algorithm fails in some classification tasks and enhance the standard approaches by implementing some effective and more efficient techniques. Where P(Ci)= and P(dj|ci)
= Naïve Bayes has been one of the popular machine learning methods over the years.

Its simplicity makes the framework usable in various tasks and with good performance, results are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there is a number of interesting works of investigating naive Bayes classification algorithm. Recently the researches show considerably good results by selecting Naïve Bayes over SVM for small text classification with less number of training inputs. The author proposes a Poisson Naive Bayes text classification model with weight enhancing method, and shows that the new model assumes that a document is generated by a multivariate Poisson model.

Many of the research results say that per-document term frequency normalization is used to estimate the Poisson parameter, while the standard multinomial classification algorithms estimate their features by considering all the training documents as a whole different huge training data set. Many researchers presented that naive Bayes classification algorithm can perform surprisingly well in the classification process where the probability itself is calculated by the naive Bayes, which is considered not that important.

The most frequent example that we encounter, where they use a naïve Bayes classification algorithm is for spam filtering. Mostly this technique is largely used in email, web contents, and spam classification. Naive Bayes classifier shows good performance on both numeric and textual data and it can be easily implemented when compared with other algorithms. In most cases, conditional independence assumption is violated by the algorithm for real-world data and performance varies when we have highly correlated and it does not give importance to the frequency of word occurrences.

2.2 Support vector machine (SVM):

Support vector machines (SVMs) are one of the discriminative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory. The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM is well-founded that very open to theoretical understanding and analysis.

The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so-called the hyperplane. The document representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data

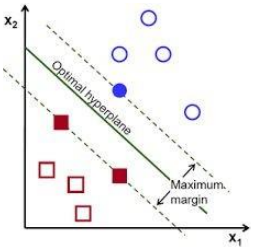


Fig -1: Support vector machine representation

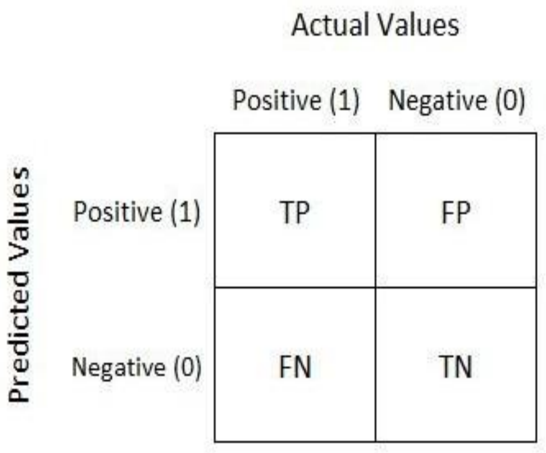
3. Accuracy measure

Confusion Matrix:

It is a performance metric of machine learning classification problem where output can be 2 or 3 more classes. It is a matrix with 4 differentcombinations of predicted and actual values.

Precision: The result of correctly predicted classesout of all classes should be as high as possible.

$$\mathbf{Pr} \qquad \frac{TP}{P + FP}$$



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

It is mostly used for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

True Positive: The interpretation predicted is positive but it is actually false.

True Negative: The interpretation predicted is negative but it is actually true.

False Positive: The interpretation predicted is positive but it is actually false. It is also called as Type 1 error

False Negative: The interpretation predicted is negative but it is actually true. It is also called as Type 2 error

Let's understand confusion matrix through math:

Recall: The result of correctly predicted classes outof all positive classes should be as high as possible.

F-measure: In order to compare 2 models with low precision and high accuracy we need F-score. Recall and Precision can be measured at the same time using F-measure. F-score Harmonic Mean instead of Arithmetic Mean by tuning the extreme values more.

$$\textit{Recall} = \frac{TP}{TP + FN}$$

$$\textit{F - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

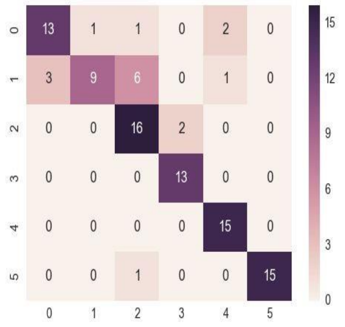


Fig -2: Confusion matrix representation

4. CONCLUSIONS

In the above paper, we choose accuracy estimation as the basic metric for comparing algorithms on our textual data. We have represented that the algorithm accuracy depends on the size of the data set. We have used the box plot representation to visualize our accuracy comparisons among various algorithms. If the size of the data set varies the performance of the algorithm may also vary. Even a slight difference/change in the training data will reflect changes in the total algorithm performance. We all can see that the support vector machine is outperforming over Multinomial naïve Bayes algorithm.

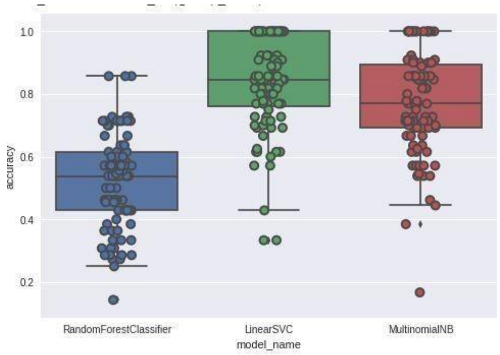


Fig -3: Representation of the performance of various classifiers.

5. Acknowledgement

I would like to thank and express our deep sense of gratitude to our internal project guide Mr.Sravan Kumar G, Assistant Professor, Department of CSE, Sphoorthy Engineering College and our Major project coordinator Dr.Chaitanya Kishore Reddi.M Professor, Department of CSE, Sphoorthy Engineering College for their valuable time and suggestions.

REFERENCES

[1] http://www.tfidf.com/ [2]https://blog.xrds.acm.org/2017/10/introduction- n-grams-need/

[3]https://towardsdatascience.com/understanding-c onfusion-matrix-a9ad42dcfd62

[4]http://www.jait.us/uploadfile/2014/1223/20141 223050800532.pdf

[5] ve-bayes-classifier-for-text-analysis-python-8dd6825 ece67

[6] https://www.saedsayad.com/naive_bayesian.htm [7]https://en.wikipedia.org/wiki/Support-vector_machine