ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 274-279
doi: 10.18178/wcse.2019.06.041

Naïve Bayes SentimentAnalysis with Fixed and Variable Length Classes Training Data Sets

Saad Ibrahim Amaya+ and DongYuxin

Harbin Engineering University
Harbin, Heilongjiang, China

Abstract. The tremendous development in technology has led to the increasing number of people that join social networks to share information, opinion and so on. With these developments, the social networks are big targets and easy place to capture many people's opinions about certain things. A lot of works have been done by many researchers on the extraction of sentiments from various data sources. Different works employed differenttechniques and approaches. This particular work investigates the training dataset input. We categorized the training datasets into two (2) and termed them; the Variable Length/size (VS) training dataset and the Fixed Lengths/size (FS) training datasets. In the FS, we took the number of positivedocuments equals the number of negative documents. In the VS, we took the number of positivedocuments greater than the number of negative documents (VS positive) and vice versa (VS negative). Binary Na ïve Bayes algorithm was used to test the FS, VS positive and VS negative training datasets on the test dataset. The results showed that, it is better to use the FS training dataset, and if the numbers of positive and negative texts are going to be unequal, then the ratio number of one class to the other should be very small. We can conclusively say that, the wider the ratio the less accurate results, and the narrower the ratio, the more accurate the results.

Keywords: Sentiment Analysis,Na ïve Bayes', Supervised Machine Learning, Data mining.

1. Introduction

Sentiment analysis is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text [5]. Define Sentiment analysis as the extraction and the analysis of public moods and views [10].

Users of social networks generate huge amount of data expressing their views and opinions. Those huge data are characterized by three computational issues namely; size, noise and dynamism. These issues often make social network data very complex to analyze with the traditional analysis methods, resulting in the pertinent use of computational means of analyzing them [1].Websites contain millions of unprocessed raw data. By analyzing this data new knowledge can be gained[11]. Traditional methods of data analysis require the data to be stored and then processed off-line. Data streams are infinite, and data is generated with high rates and therefore it cannot be stored in main memory [9]. Numerous research works has been already done in field of sentiment analysis. But the informal tone of tweets has always been a challenge for the analysis [4].

This work is going to use a Supervised Machine Learning algorithm, that is; the Binary Na ïve Bayes algorithm. We are going to use this algorithm in two cases: to feed it with two different groups of training datasets. One is to have equal number of positive and negative texts in the training dataset and the other is to have variable number of positive and negative texts in the training dataset. With the Variable number of texts in the training data set, we will conduct the test taking the positive text to be twenty percent (20%) greater

+ Corresponding author. Tel.: +2348138178383
E-mail address: elsaadamaya@yahoo.com

than the negative texts in the training dataset (that is, sixty percent (60%) positive texts and forty percent (40%) negative texts) and vice versa. We will then calculate the accuracies, the positive and negative precisions and the positive and negative recalls.

The rest of the work would have four (4) sections and they are briefly explained as follows:

Section 2 deals with the related work of our study, section 3 presents the proposed work (Datasets and data sources used in our study along with the models and the methodology used), section 4 presents all our experimental results, section 5 presents the conclusion drawn from our survey.

2. Related Work

BhutaSagar., DoshiUchit., and NarvekarMeera [2014]: They worked on number of techniques, both lexicon-based approaches as well as learning based methods for sentiment analysis of text from twitter by addressing many issues [3].

StefanoBaccianella, Andrea Esuli, andFabrizioSebastiani [2010]: They presented SENTIWORDNET3.0, an enhancedlexical resource explicitly devised for supporting sentimentclassification and opinion mining applications, an improved versionof SENTIWORDNET1.0 [2].

Isa Maks and PiekVossen [2012]: worked on Lexicon Model for deep sentiment analysis and opinion mining application. They used lexicon model to describe verbs, nouns and adjectives to sentiment analysis and opinion mining applications. [7].

LifengJia, Clement Yu, and WeiyiMeng [2009]: They investigated the problem ofdetermining the polarity of sentiments when one or more occurrences of a negation term such as “not” appear in a sentence [6].

Prabowo Rudy and Thelwall Mike[2009]: worked on hybrid classification on to improve classification effectiveness [8].

Peter D. Turney [2002]: Worked on semantic orientation applied to unsupervised classification of review using thumb up as positive and thumb down as negative responses.

3. Proposed Work

3.1 Data source and data set

All 120,000 tweets have been extracted for the training and the test datasets for the research work have been downloaded from https://www.kaggle.com/kazanova/sentiment140/version/2

3.2 Methodology

The goal of this research work is to analyze text data for the analysis of sentiments and compare the results accuracy usingFS, VS positive and VS negative as explained earlier. We are going to use the Na ïve Bayes algorithm for this work. Below is the graphical representation of the processes involved in the sentiment analysis or opinion mining.

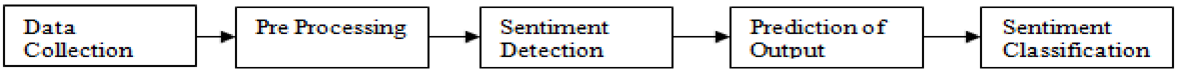

Fig.1: Sentiment Analysis Processes

The task of classification is to take a document **d** in a set of documents**D**as an input, **d** ∈ **D** and give an output by classifying it into a class **c** in one of a class set of discrete classes say **C, c** ∈ **C.**

With Na ïve Bayes, a document ***d***, out of all classes **c** ∈ **C** the classifier returns the class $\hat{c}$ which has the maximum posterior probability given the document. That is, our estimate of the correct class.

$$\hat{c} = argmaxP(c|d)\ for\ c\ in\ C \tag{3.1}$$

Bayes’ rule presents Eq. 3.2; it breaks downs$P(x|y)$ into three other probabilities:

$$P(x|y) = \frac{P(x|y)P(x)}{P(y)} \tag{3.2}$$

We can then substitute Eq. 3.2 into Eq. 3.1 to get Eq. 3.3:

$$ĉ = argmaxP(c|d) = argmax\frac{P(d|c)P(c)}{P(d)}\ for\ c\ in\ C \tag{3.3}$$

We can simplify Eq. 3.3 by dropping the denominator$\boldsymbol{P(d)}$**.** Because we will be computing $\frac{P(d|c)P(c)}{P(d)}$ for each possible class,$\boldsymbol{P(d)}$doesn't change for each class. Thus, we can choose the class that maximizes this simpler formula:

$$ĉ = argmaxP(c|d) = argmaxP(d|c)P(c)\ for\ c\ in\ C \tag{3.4}$$

$P(c)$*is the prior probability*

3.2.1 Training the Naïve Bayes classifier

The Prior probability is the probability of each document that occur in a class and is given by:$\acute{P}(c) = \frac{N_c}{N_{doc}}$. Where N_c is the number of documents in the training data labeled with class c and N_{doc} is the number of documents in all the classes.

We learn the probability of each feature or word $w_{i,}$ in the documents in the training dataset belong to a class $c\acute{P}(w_i|c)$, by computing the number of documents in that class the word appears divide by all the words in the training dataset. This is Binomial Na ïve Bayes.

$$\acute{P}(w_i|c) = \frac{count\ (w_i,c)}{\sum_{w \in V} count(w,c)} \tag{3.12}$$

Here we have the vocabulary V which consists of all the words in the overall documents.

The probability of a feature or word that does not exist in that class will be zero and multiplying all the probability of the features in that document including those whose values are zero, the entire conditional probability of that document belonging to that particular class will be zero. The simplest solution to this problem is to use add-one or Laplace smoothing method.

$$\acute{P}(w_i|c) = \frac{count\ (w_i,c)+1}{\sum_{w \in V}(count(w,c)+1)} = \frac{count\ (w_i,c)+1}{(\sum_{w \in V} count(w,c))+|V|} \tag{3.13}$$

3.2.2 The training datasets

Given a training dataset D of sizeN_D, with each document belonging to either of one of the distinct classes c_1, c_2 ... $c_n$in C of sizeN_c. We then take equal number of documents from each of the classes in C to make the training dataset. This is the mathematical meaning of the Fixed Size (FS) training dataset. For each class in C, we take the number of documents as follows to make the training dataset.

$$|c_i| = |\ \frac{N_D}{N_c}| \tag{3.14}$$

We have all the classes to have the same number of documents.

In theVS training dataset, we are going to have a class to have the most number of documents (***main class***) than the others (***other class)***.The **other class** must be of the same sizes.We choose the difference between the **main class** and the **other class** either by an integer number d, or by percentage d%.

For VS with integer number difference, we obtain the size of the **main class**N_m in the equation 3.15 below:

$$N_m = |\frac{N_D}{N_c}| + d \tag{3.15}$$

And for the other classes size,N_o, we subtract the N_m from N_D and divide by the remaining number classes.

$$N_o = |\frac{N_D - N_m}{N_c - 1}| \tag{3.16}$$

Second, we deal with the VS with percentage difference. Instead of adding d in equation 3.15, we add the d percent of N_D as shown in the equation 3.17 below:

$$N_m = |\frac{N_D}{N_c} + \frac{dxN_D}{100}| \tag{3.17}$$

And for the other class, we can maintain equation 3.16 with the substitution of N_m from equation 3.17.

Now that we have the size of individual class documents in the training dataset, we can run our experiments for different sizes (e.g 10,000, 14,000, 16,000, and so on.) of training dataset on test dataset. First, we will run the experiment with FS and the various VSs.

3.2.3 The Procedures

Below are the procedures for the training and the testing algorithms.

Training Pseudo-code

```
ProcedureTrainAlgorithm(C,File,Cls,N):
   D  < - getDocuments(File,Cls)
   V < - getVocabulary(D)
   F <- getFeatures(D)
   N<-CountAllDocuments(D)
   For each c in C do
      Length[c] <- CountDocuments(D[c])
priorProb[c]<-Length[c]/N
      for each f in F[c] do
         count[f] <- countFreq(f,D[c])
condProb[f][c] <- count[f]+1/count[F]
      end for
   end for
   return V, priorProb, condProb
end Procedure
```

Testing Pseudo-code

```
ProcudureTestAlgorithm(d,V,C,Pp,Cp):
   for each c in C do
cProb[c] <-Pp[c]
      for each f in d do
         if f in V:
cProb[c] <- cProb[c] x Cp[f][c]
         end if
      end for
   end for
   returnargmax$_{c \text{ in } C}$P[c]
end Procedure
```

4. Experimental Result

This chapter is going to analyze the results gotten from several experiments taken for the FS and VS numbers of positive and negative training datasets as input. We are going to evaluate the performances of the tests conducted using Accuracy used for opinion mining. The formula for calculating the Accuracy below with respect to the values in the table 1.

$$\text{Accuracy} = \frac{a + d}{a + b + c + d}$$

Table 1: Confusion Table

	True positive documents	True Negative Documents
Predicted Positive Documents	a	b
Predicted Negative Documents	c	d

We begin by observing some few results of the experiment on the test dataset of 182 positive texts and 177 negative documents. We can see that, with the FS, the positive and negative have little difference in the percentage of texts correctly classified. With the other methods; one class is lopsided.

Table 2: Percentage of correctly classified classes per method

Experiment number	Number of texts in the training sets	Na ïve Bayes FS		Na ïve Bayes VS Positive		Na ïve Bayes VS Negative	
		Positive	Negative	Positive	Negative	Positive	Negative
1.	10,000	69.8	71.8	99.5	19.8	6.0	98.9
2.	14,000	72.5	72.3	99.5	15.8	8.2	98.9
3.	16,000	73.6	69.5	98.4	16.4	10.4	99.4
4.	18,000	75.3	72.3	98.4	16.4	9.3	99.4
5.	20,000	76.9	70.6	98.4	14.1	9.3	98.9
6.	24,000	77.5	70.6	97.8	17.5	11.5	98.9
7.	30,000	75.3	72.9	98.4	15.8	9.3	99.4

The table 3 and the figure 4 show the overall accuracies.

Table 3. Accuracy comparison on Test Datasets.

Experiment number	Number of texts in the training sets	Na ïve Bayes FS	Na ïve Bayes Vs Positive	Na ïve Bayes VS Negative
1.	10,000	70.8	59.7	52.5
2.	14,000	72.4	57.7	53.6
3.	16,000	71.6	57.4	54.9
4.	18,000	73.8	57.4	54.4
5.	20,000	73.8	56.2	54.1
6.	24,000	74.1	57.7	55.2
7.	30,000	74.1	57.1	54.4
8.	34,000	73.0	57.4	54.7
9.	40,000	72.2	57.4	55.2
10.	46,000	71.9	58.2	56.3
11.	50,000	75.5	58.5	55.5
12.	54,000	75.2	58.8	55.8
13.	58,000	74.7	58.5	55.5
14.	60,000	73.3	58.8	55.2

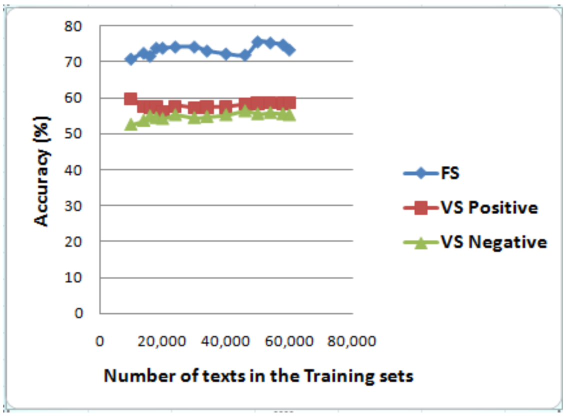

Fig. 4:Diagrammatic representation of the experiment Accuracies

5. Conclusion

This work tests the effectiveness of the Na ïve Bayes Algorithm with different groups of training dataset while maintaining both the training and the testing algorithm. We then evaluate the performance in terms of Accuracy. The result shows that, when we train equal number of positive and negative documents in the training dataset, we get more accurate result than training one class number of documents greater than the other.

6. References

[1] Adedoyin-Olowe M., Gaber M. M. and Stah F. 2014. A Survey of Data Mining Techniques forSocial Network AnalysisSchool of Computing Science and Digital Media. Robert GordonUniversity Aberdeen, AB10 7QB, UK School of Systems Engineering, University of Reading PO Box 225, Whiteknights, Reading, RG6 6AY, UK.

[2] Baccianella S., Esuli A., and Sebastiani F. 2010.SENTIWORDNET3.0: An Enhanced Lexical Resourcefor Sentiment Analysis and Opinion Mining .Istituto di Scienza eTecnologiedell'InformazioneConsiglioNazionaledelleRicercheVia Giuseppe Moruzzi 1, 56124 Pisa, Italy.

[3] Bhuta S.,Doshi U.,and Narvekar M. 2014. A Review of Techniques for Sentiment Analysis of Twitter Data. 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). College of Engineering, Mumbai, India.

[4] Goel A., Gautam J., and Kumar S. 2016. Real Time Sentiment Analysis of Tweets Using Na ïve Bayes. 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016 JSS Academy of Technical Education, Noida, INDIA

[5] Hutto C.J., and Gilbert E. 2014. VADER: A Parsimonious Rule-based Model forSentiment Analysis of Social Media Text. Association for the Advancement of Artificial Intelligence. Georgia Institute of Technology, Atlanta, GA 30032.

[6] Jia L., Yu C.T.,and Meng W.2009. The effect of negation on sentiment analysis and retrievaleffectiveness. In*Proceedings of CIKM 2009, China*.

[7] Maks I., and Vossen P. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems 53 (2012) 680–688. VU University, Faculty of Arts, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.

[8] Prabowo R.,and Thelwall M. 2009. Sentiment Analysis: A Combined Approach. School ofComputing and Information TechnologyUniversity of WolverhamptonWulfruna StreetWV11SB Wolverhampton, UK.

[9] Rahnama, A.H.A 2014. Distributed Real-Time Sentiment Analysis for Big Data Social Streams. Department of Mathematical Information Technology University of Computer Science, Finland.

[10] Ravi K., and Ravi V. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Center of Excellence in CRM and Analytics, Institute for Development and Research in Banking Technology,School of Computer & Information Sciences, University of Hyderabad, India.

[11] Sharma A., Sharma M.K., and Dwivedi R.K. 2017. Literature Review and Challenges of Data Mining Techniques for Social Network Analysis. Advance in Computational Science and Technology. ISSN 0973-6107 volume 10, Number 5, pp. 1337 – 1354.