International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
573 – 577

Sentiment Classification Using Supervised and Unsupervised Approach

Manju Bala
I. P. College for Women,
Delhi University, Delhi
manjugpm@gmail.com

Abstract: In past few years, the data available on internet has multiplied at an alarming rate. Tweets, reviews, blogs and comments on social media have been a huge factor which has resulted in such a huge amount of increase in the available data. Because of this datasets being highly unstructured and of high dimensionality, sentiment classification becomes a very tiresome task. Sentiment Analysis is used to estimate the user opinion on various issues. It consequently mines states of mind and perspectives of clients on particular issues. It's a multistep preparation where choosing and extracting elements is an indispensable stride that controls execution of sentiment classifier. In this paper we have used three supervised techniques namely SVM, Decision Tree and Nave Bays Algorithm and three unsupervised techniques called DE, PSO and K-Means The results are validated using different three benchmark labeled datasets data sets and on the different feature sets We have also performed feature selection using genetic algorithm and validated results using the features selected by the GA Experimental results shows that supervised techniques have outperformed supervised techniques on one dataset while for the two datasets supervised techniques have outperformed unsupervised techniques

Keywords: *Sentimental Analysis, Feature Extraction, Feature Selection, Swarm Intelligence.*

*****

I. INTRODUCTION

With each passing day, a huge amount of data is collected through social networking sites, blogs and other media. Now, this data may contain some opinion related information that can be used very efficiently by many departments. For example, the government can use this information to find out how a particular scheme made by them is received by the people. Manually, it's not quite possible to extract the required information out of such a huge amount of data. Here comes the requirement of **Sentiment analysis[5]**. As it is quite clear from the name itself, Sentiment Analysis is basically a technique for extracting user sentiments or opinions from reviews over a particular subject, area, product or an item on web. It is an application of Natural Language Processing (NLP), computational semantics and machine learning figuring out how to recognize helpful data from the given information. The assessments are divided into two classes like "Positive" and "Negative" [1]. Henceforth, it decides the state of mind or assessment of the client over a specific theme, whether the client is supportive of it or against it.

Reviews from any social networking site, e-commerce sites or any other media are collected along with their polarity which serves as the training dataset for the algorithm. On the basis of this dataset, our algorithm further classifies the reviews as positive or negative. To obtain subjective and factual response from the gathered information, public opinions are extracted by features extractor.

In machine learning, there are two approaches for handing every situation one is we are providing corresponding target value with every input value (supervised learning) and other is we are just training our data with inputs only (i.e. Unsupervised learning).

Scenario 1

You are a kid, you see different types of fruits, yours father tells you that this particular fruit is a mango after him giving you tips few times, you see a new type of fruit that you never saw before - you identify it as that it's not a mango.

Scenario2

You go bag-packing to a new country, you did not know much about it - their food, culture, language etc. However from day 1, you start making sense there, learning to eat new cuisines including what not to eat, find a way to that beach etc.

Scenario1 is an example of supervised classification, where you have a mentor to guide you and learn concepts, such that when a new data comes your way that you have not seen before, you may still be able to identify it.

Scenario2 is an example of unsupervised classification, where you have lots of information but you did not know what to do with it initially. A major distinction is that, there is no teacher to guide you and you have to find a way out on your own.

Firstly, we have used supervised classification for extracting sentiments out of user reviews and find their accuracy. After getting the results, we enhance our research by extracting more features and used Binary Genetic Algorithm and some nature inspired algorithms like ABC (Ant Colony Optimization), PSO[13] (Particle Swarm Optimization, and more) for feature selection and optimization. After feature optimization, we have applied some more techniques like Differential Evolution (DE) **[15]** and K-Means**[16]** Clustering, and compared the results thus obtained.

II. PROPOSED METHOD

Figure 1 displays the flow of our proposed method.

A. Datasets/ Input data

Here we have used 3 datasets where one has reviews about movies; second have reviews about apple phones collected

 ISSN: 2454-4248
573 – 577

from twitter and third has reviews about amazon food products. All of them are leveled are labeled datasets. Table 1 shows the information about datasets.

B. Pre-processing

The raw tweets and reviews collected from twitter and other online e-commerce sites have unwanted, fuzzy , meaningless words, stop words , URLs , extra spaces etc. which are required to be removed before feature extraction. Hence the methodology uses following preprocessing steps before feature extraction.

- Convert all the words of reviews into lowercase.
- Remove punctuation from reviews (like @,!).
- Remove any numbers from reviews (1,2,3)
- Remove all the stops like a, an, the etc. from the reviews.
- Convert all the words into stemming words.
- Finally remove extra white spaces from the reviews.

Table 1: Considered Datasets

DATSETS	NUMBER OF REVIEWS	
	Positive	Negative
1. Twitter Sanders Apple (479)[10]	163	316
2. Movie Reviews (8544)[11]	3998	4546
3. Amazon Food Review (4950)[12]	3708	1242

C. Unigram and bigram

For supervised learning algorithm, we divide each datasets into two parts one for training and other for testing.

Then we converted them into document term matrix using unigram and bigram. Text mining and natural language processing tasks used N-gram techniques. They are group of words within a given sentence and when calculating the n-grams we move one word forward after each round. For example, for the sentence *"My name is John Mark"*. If N=2 (called as bigrams), then the n-grams are:

- My name
- Name is
- Is John
- John Mark

Therefore here we have 3 2-grams.

D. Supervised learning algorithm.

After extracting features, we have applied different supervised learning algorithm for our training and testing datasets.

1. Naïve Bayes Classification

Naïve Bayes classifiers[17] are simply probabilistic classifiers based on Bayes Theorem with strong relation between different features.

If there are *n* numbers of features, using Bayes' theorem, the conditional probability[2] can be decomposed as

$$P(Ck|X) = \frac{P(Ck)P(X|Ck)}{P(X)}$$

Here, x is x1, x2, x3….xn.

2. Support Vector Machine. [14]

Support vector machine creates hyper-plane in infinite-dimensional space, which can be used for classification. Intuitively, hyper-plane that has the largest distance to the nearest point to the clusters creates a good separation, since the generalization error of the classifier will be less if there is a large margin.

3. Decision Tree.

Decision tree is decision support machine tool in which we make a decision depending on the condition. It works a test on attribute and each branch represents outcome of the test.

Figure 1: FLOW CHART

E. Feature Extraction[6][7]

By using LIWC software we extract 93 features for each review from each datasets. List of features given in figure 2.

review	Label	WC	Analytic	Clout	female	body
Authentic	Tone	WPS	Sixltr	Dic	male	health
function	pronoun	ppron	i	we	cogproc	sexual
you	shehe	they	ipron	article	insight	ingest
prep	auxverb	adverb	conj	negate	cause	drives
verb	adj	compare	interrog	number	discrep	affiliation
quant	affect	posemo	negemo	anx	tentat	achieve
anger	sad	social	family	friend	certain	power
differ	percept	see	hear	feel	bio	reward
risk	focuspast	focuspresent	focusfuture	relativ	motion	space
Exclam	time	work	leisure	home	money	relig
Dash	death	informal	swear	netspeak	assent	nonflu
filler	AllPunc	Period	Comma	Colon	SemiC	QMark
Quote	Apostro	Parenth	OtherP			

Figure 2 List of features

F. Feature Selection

After feature extraction, we select those features which give us maximum accuracy and optimum results. So here we used Binary Genetic Algorithm for feature selection.

 ISSN: 2454-4248
573 – 577

G. Unsupervised Learning Algorithm.[3]

- *Particle Swarm Optimization[4]*

Particle swarm optimization (PSO) is a population based and inspired by behavior of bird flocking. This technique developed by Dr. Eberhart and Dr. Kennedy in 1995.
There are group of random particles which are initially initialized with random values and then searches for optima by after updating so many generations. In every generation, each particle is changed by two "best" values. The first one is the best solution it has achieved so far i.e. pbest value. Another "best" value that is tracked by PSO is the best value, obtained by any particle in the population i.e. gbest value. Every particle has its own local best values i.e. known as lbest.

After calculating these best values, the particle updates its velocity and position by these formulae.

$$v(t+1) = v(t) + c1 * rand * (pbest - present) + c2 * rand * (gbest - present)$$
$$present(t+1) = v(t+1) + present(t)$$

v[] is the particle velocity, persent[] is the current particle position. rand () is a random number between (0,1). c1, c2 are learning factors.

- *Differential Evolution[4]*

DE is an optimisation technique which iteratively modifies a population of candidate solutions to make it converge to an optimum of your function.
It is similar to genetic algorithm (GA) except that the candidate solutions are not considered as binary strings (chromosome) but (usually) as real vectors

- K-means[9][4]

***K-means* clustering** is a method of creating clusters in data mining. *K-means* clustering aims to distribute *n* reviews into *k* clusters in which each observation belongs to the cluster with the nearest Euclidean distance.

II. EXPERIMENTAL RESULTS

The accuracy of proposed supervised and non-supervised techniques has been tested on 3 different datasets. A brief description of the datasets has been explained in table x.

3.1 Twitter-sanders-apple

Sanders Analytics have collected this dataset for Apple Corp. on four separate topics: Apple, Microsoft, Twitter and Google. It consists of a total of 479 reviews, out of which 163 are positive and 316 are negative.

3.2 Amazon Movie Reviews

This dataset contains movie reviews collected from amazon website. Positive reviews are labeled as 'pos' and negative reviews as 'neg'. There are a total of 8544 reviews. 3998 are labeled as positive and remaining 4546 as negative.

3.3 Amazon Food Reviews

Food product reviews have been collected from amazon website and after being classified as negative or positive have been saved under this dataset. Out of a total of 4950 reviews, 3708 are positive and 1242 negative.

The datasets have been preprocessed to eliminate undesired words such as hash tags, urls, stop words, etc.

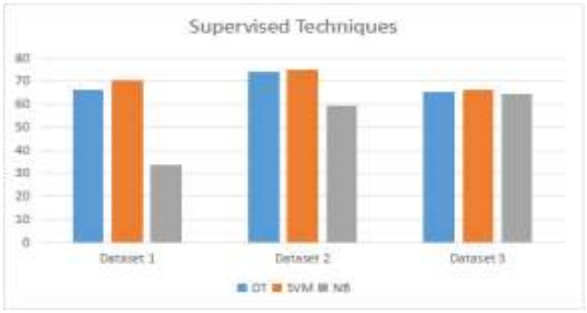

Figure 3: Result of Supervised Techniques

Table 2: Result using Supervised Techniques

Algorithm	N-gram	Accuracy		
		DS1	DS2	DS3
Decision Tree	Unigram	65.83	73.81	65.6
	Bigram	65.83	73.81	64.2
SVM	Unigram	70.00	74.22	66.1
	Bigram	70.58	74.54	66.3
Naïve Bayes	Unigram	33.16	59.18	64.0
	Bigram	33.61	59.44	64.8

Figure 3 represents the pictorial representation of results acquired from supervised techniques. Here, average of unigram and bigram accuracies is taken to represent the graph. Table x shows the values obtained for both unigram and bigrams. It is quite clear from the results that SVM provides the best result in both unigram and bigram features. Moreover, it is quite clear that bigram provide a better result than the unigrams.

For non-supervised techniques, a total of 93 features were extracted using the LIWC dictionary. Genetic Algorithm was then applied to reduce the features and get the best features out of the total of 93 features to maximize the accuracy. Further, a set of 73, 53 and 33 features were randomly selected and non supervised techniques were applied on the same to find some relation among the accuracy obtained and the number of features selected.

The following results were obtained after applying non supervised techniques:

Table 3: Accuracy obtained using PSO, K Means and DE (for Dataset 1):

	PSO	DE	K Means
93 features	75.36	75.36	52.6
73 features	65.97	65.97	75.36
53 features	65.97	65.13	66.17
33 features	65.34	68.47	65.35
After GA	75.36	75.36	50.93

List of features selected by GA :

 ISSN: 2454-4248
573 – 577

Clout, Authentic, Tone, Function, Shehe, Auxverb, Adj, Number, Anx, Social, Friend, Female, Male, Insight, Discrep, See, Affiliation, Achieve, Focuspast, Relativ, SemiC

Table 4: Accuracy obtained using PSO, K Means and DE (for Dataset 2):

	PSO	DE	K Means
93 features	70.04	73.61	68.08
73 features	67.71	74.52	68.82
53 features	69.76	70.66	68.72
33 features	65.05	65.43	68.82
After GA	58.5	54.56	54.38

List of features selected by GA :
Clout, Quant, Affect, Negemo, See, Bio, Focuspast, Motion, Death, Informal

Table 5: Accuracy obtained using PSO, K Means and DE (for Dataset 3):

	PSO	DE	K Means
93 features	57.99	57.95	57.55
73 features	58.05	58.09	57.59
53 features	58.05	58.06	57.99
33 features	58.05	58.05	57.55
After GA	63.94	63.94	64.2

List of features selected by GA :
Clout, Quant, Affect, Negemo, See, Bio, Focuspast, Motion, Death, Informal.

Figure 4: Result of Non Supervised Techniques on DS 1

Figure 5: Result of Non Supervised Techniques on DS 2

Figure 6: Result of Non Supervised Techniques on DS 3

It's clear from the results that there is no relation between the numbers of features used to the output accuracy using an appropriate technique. Also, features selected by GA give best result only for third dataset, which comprised of the Amazon movie reviews. Moreover, all the 3 algorithms used give different result on different dataset. Unlike SVM which gave the best result in each case for supervised methods, there is no clear winner in this case.

III. Conclusion

In this paper, we have used 3 supervised techniques, namely, Decision Tree, Support Vector Machine and Naive Bayes Classifier, and 3 non supervised techniques, namely, K Mean Clustering, Particle Swarm Optimization and Differential Evolution and compared the accuracy of these techniques on 3 different datasets.

Out of the Supervised Techniques, SVM was clearly the best method, which gave the best accuracy among all three datasets. When it came to non supervised techniques, there was no such technique which gave the best result in each case. Here, the techniques used were dependent on the type of dataset, which they were applied on. For Dataset 1, best accuracy obtained was 75.36% which was obtained by all 3 techniques for different set of features. For Dataset 2, Differential Evolution proved to be the best when used with the set of 73 features. For the final dataset, K Mean marginally overshadowed the other 2 techniques to give the best result.

Moreover, there was no such pattern noticed which could persist between the amount of features used for sentiment analysis and the accuracy obtained w.r.t those features. Genetic Algorithm, which was also used for feature reduction didn't prove to be the best when it comes to selecting the best available features, except for the last dataset, where accuracy was the best when features selected by GA were used.

References

[1] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada, "Survey of review spam detection using machine learning techniques",2011

[2] Hamad Alhammady. "Weighted Naive Bayesian Classifier", 2007

[3] Luiz F. S. Coletta, Nadia F. F. da Silva, Eduardo R. Hruschka,Estevam R.Hruschka Jr." Combining Classification and Clustering for Tweet Sentiment Analysis",2014.

 ISSN: 2454-4248
573 – 577

[4] Akshi Kumar, Renu Khorwal* and Shweta Chaudhary: "A survey on Sentiment Analysis using Swarm Intelligence", 2016
[5] Bo Pang and Lillian Lee: "Thumbs Up? Sentiment Classification using Machine Learning Techniques", 2002
[6] Henrique Siqueria and Favia Barros: A feature extraction process for Sentiment Analysis of Opinions on services
[7] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Fazal Masud Kundi: "A review of feature extraction in sentiment analysis", 2014
[8] Bingwei Liu*, Erik Blasch, Yu Chen, Dan Shen*, and Genshe Chen*: "Scalable Sentiment classification for Big Data Analysis using Naive Bayes Classifier, 2013
[9] Avinash Chandra Pandey ∗, Dharmveer Singh Rajpoot, Mukesh Saraswat: "Twitter sentiment analysis using hybrid cuckoo search method"
[10] Twitter-sanders-apple:(2015). http://boston.lti.cs.cmu.edu/ classes/95- 865- K/HW/HW3/ .
[11] movie_pang: http:// www.cs.cornell.edu/ people/ pabo/ movie-review-data/
[12] amazon fine food review https:// www.kaggle.com/ snap/ amazon-fine-food-reviews
[13] PSO http://www.swarmintelligence.org/tutorials.php
[14] SVM:https:// en.wikipedia.org/ wiki/ Support_vector_ machine
[15] DE_algorithm: https://en.wikipedia.org/ wiki/ Differential _evolution
[16] K-means:https://sites.google.com/site/dataclustering algorithms/k-means-clustering-algorithm
[17] Naïve Bayes Classifier: https://en.wikipedia.org/wiki/ Naïve_Bayes_classifier