Sentiment Analysis using Feature Based Support Vector Machine – A Proposed Method

Prakash P. Rokade, Aruna Kumari D

ABSTRACT*: *Business decisions for any service or product depend on sentiments by the people. The mood of people towards any event, service and product are expressed in sentiments. The text sentiment contains different linguistic features of sentence. A sentiment sentence also contains other features which are playing a vital role in deciding the polarity of sentiments.The features like duplication of sentiment, unknown emotics may change the polarity of sentiment.If features selection is proper one can extract better sentiments for decision making. A directed preprocessing will feed filtered input to any machine learning approach. Support vector machine proved as a good tool of machine learning for better sentiment analysis.Better use of parts os speech (POS) folled by guided preprocessing and evaluation will provide less errorus polarity of sentiments

***Index Terms*: feature weighting, n gram model, parts of speech, sentimentanalysis, support vector machine,**

I. INTRODUCTION

Sentiment analysis of twitter moods is playing a vital role for business decisions to plan a good business strategy. A customer can view the sentiments online to decide for purchasing any product. The sentiments posted by customers are thoroughly studied by business people to know the status of the service the industry is providing. These sentiments are posted by a group, a person, and industry also. Hopefully we may expect that all the users are legal and loyal. But there may be some fake posts which are posted by fake users, competitors. As the input data for sentiments analysis is big data, this should be preprocessed before using for design any model. Duplicates sentiments, fake sentiments, blind sentiments must be removed by preprocessing input big data. After preprocessing this data are transfers into well known structure form. Latter on features like noun, adjective adverb, capital letters, numbers are used for sentiments selection. Dictionary approach can be used to find positive, negative, neutral sentiments based on the score of the whole document decision can be made by business intelligence. evaluation. Support vector machine (SVM) is proved better machine learning strategy for sentiments classification.

II. LITERATURE SURVEY

A remarkable work is carried out for sentiment classification. The main focus of this work is on classifying larger pieces of text, like reviews of product or event [1]. Tweets are different from reviews as they have different purpose. Reviews are summary of author’s thoughts. Tweets are limited to 140 characters of text. Tweets represent general mood of people through various reactions based on experience or as an impression for news articles [2]. Hu and Liu have given a technique for Feature Based Summarization system (FBS) of customer reviews of products. It also generated sentiment based summary as either positive or negative opinion using adjective words in reviews [3]. Chaovalit and Zhou compared supervised and unsupervised algorithm for classification and got 83.54% of accuracy for supervised method and 77% of accuracy for unsupervised method [4]. Pang O Keefe and Koprinska have given technique to select features using attribute weights and applied Navie Bayes and SVM classifiers for classification of moods [5, 6]. Linguisticfeatures are used to find twitter sentiment. Author used hash tagged data set (HASH) and emoticon data set. By tokenization, normalization and parts of speech tagging.Classification is done. Results are evaluated by using unigrams and bigrams [7, 8].The study by Hassan introduces classification method for query term sentiment analysis. Here classifier and feature extractor are considering two different components [9]. Each token is assigning a sentiments score called total sentiment index. Base upon the classification algorithm the sentiments are classified as positive or negative polarity sentiments [10]. Political future can be analyzed real time monitoring and analyzing public conversation on social sites [11]. Feature vectors and tagged content of datais used to design model by using machine learning approach. This model further can be used to classify untagged datain text document [13]. For language consistency twitter is more informal. Emoticons are used express the opinion. Many tweets are ambiguous and these are maximizing the opinion for readers; but deflect the opinion to a machine learning algorithm [14]. Sentiment classification algorithm (SCA) and SVM are used to examine the performance of the approach,accuracy, recall, precision are some parameters on which sentiment analysis performance is evaluated [15, 16].

Revised Version Manuscript Received on 16 September, 2019.

***** Correspondence Author

Prakash P.Rokade Email: Prakashrokade2005@gmail.com

Dr Aruna Kumari Professor at Department of CSE,VJIT, Hyderabad, Telangana India.. Email: arunakumari@vjit.ac.in

Retrieval Number: B14630982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1463.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. PROPOSED APPROACH

A. MathematicalModel

LetM bethemodel which describesthe extraction,preprocessing,lebling and evaluating the sentiments .

S= {Tw, Pt, Sl, Se}

Where

- Tw = Twitter sentiments.
- Pt=PreprocessingofTweets
- Sl=Labling the sentiments as positive,negative or neutral
  Sl= {Pv, Nv, Ne}
  Pv={P1,P2,…,Pn}=PositiveClass
  Nv={N1,N2,…,Nn}=NegativeClass
  Ne={Ne1,Ne2,…,Nen}=NeutralClass
- Se=Sentimentevaluation

IV. RESEARCH DESIGN

A proposed research design for sentiments analysis using feature base support vector machine techniques is given below.

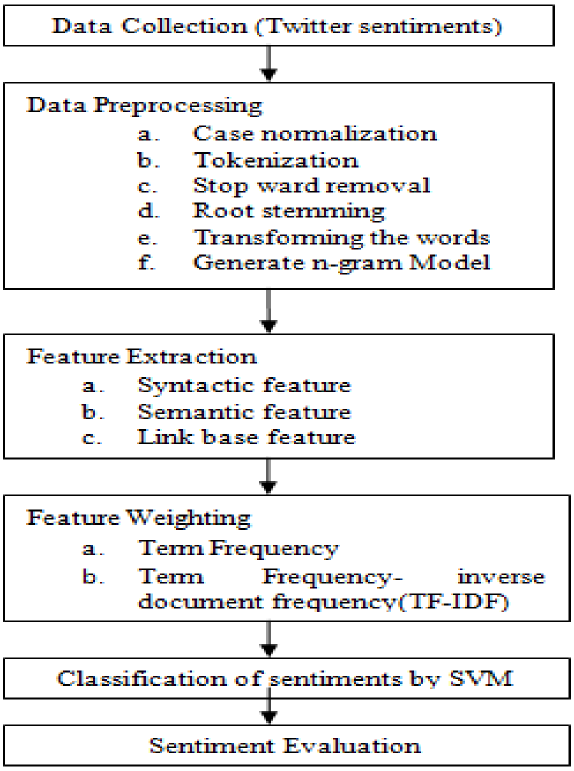

Fig. 1 Flow of Proposed Sentiment Analysis Approach

A. Data Collection

A correct input may leads us to get a correct output. Sentiments data is available on twitter website.

B. Data Preprocessing

a. Case normalization

The tweets are available in combined case that is it may contain upper and lower case characters. In casenormalization the entire document or sentence is converted in to lower case pattern generally.

b. Tokenization

A document is split in to sentences. Sentences may be divide in to words. By removing certain characters like punctuation marks these words are now tokens.

c. Stop ward removal

A set of stop words list is provided to remove them from sentiments. The frequently used stop words are “A, AN, THE, IS, THAT, SHALL, WILL”.

d. Root stemming

In this process derived words are reduced to their stem. The disadvantage of stemming is that all inflected forms must be listed in table. For example “Careful”, “careless”, “carefully” are reduced to “care”.

e. Transforming the words

A set of define rules are used to transform the word to a specific form. For example a word clarifies can be replaced by clarify.

Table 1. Words With Their Equivalent Stem

Words	Stem
Equality, Equally	Equal
Engineering,Engineer,Engineered	Engineer
Manually,Manual,Man	Man

f. Generate n-gram Model

N-gram models sequences, natural language using its statistical properties. The probability of world to be consider if it proves the condition on some number of previous words. For example one word in bigram model, two words in trigram model. Words are model such that every n gram will compose n words. Consider three gram sequence of characters generated from “good evening “ are “goo”, “ood”, “od”, ”ev”, “eve” and so on. It can also be used for good approximate matching. However the value of n can be extended to higherlevel grams. The n-gram model can be better explained with the following examples:
Text: “Do exercise daily and eat healthy food..”
Unigrams: “Do”, “exercise”, “daily”, “and”, “eat”,” healthy”,” food”.
Bigrams: “Do exercise”, “exercise daily”, “daily and”, “eat healthy”.
Trigrams: “Doexercise daily”, “exercise daily and”, “daily and eat”. Unigrams presents the simplest model for then-gram approach. It consists of all the individualwords present in the text. The bigram model definesa pair of adjacent words. Each pair of words forms asingle bigram. The higher order grams can beformed in the similar way by taking together the nadjacent words. Higher order n-grams are moreefficient in capturing the context as they providebetter understanding of the word position.

g. Removal of handles like # etc.

Users include Twitter usernames in their tweets in order to direct their messages.

A de facto standard is
to include the @ symbol before the username (e.g.@alecmgo). An equivalence class token (USERNAME) replaces all words that start with the
@ symbol.

C. Feature Extraction

In this relevant features are extracted the noun, adjective, adverb and their combinations are used as sentiment features. The feature words are extracted and scoring for sentiment is calculated. There are four categories of feature.

Syntactic feature

Related to traditional parts of speech like noun, verb, and prepositions.

Semantic feature

It is the study of meaning. It focuses on meaning base properties of noun. A semantic properties is specified in square brackets and plus or minus sign indicating the presence or absence of that property

Boy is [+human], [+male], [-adult]

A girl is [+human], [+female], [-adult]

Link base feature

Stylistic elements are give an idea or feeling to the literature. The text will be made distinctive in some way using this technique. It improves performance of sentiment classification.

D.Feature Weighting

Term Frequency

The number of times any important terms occurring in a sentiment tweet is called term frequency for that term.

Term Frequency- inverse document frequency (TF-IDF)

There are two rating of TF-IDF phase regularity and
inverse papers regularity.

Support Vector Machine

SVM is supervised machine learning approach which analyzes input data gives classification. Given set of input sentiments, it analyzes them and group in to positive, negative or neutral sentiments. Like linear classification SVM can efficiently perform nonlinear classification also this is called as kernels trick that map inputs into high features space.

SVM constructs one or more hyper planes which can be used for classification, regression. If one hyper plane is not sufficient to classify the data additional hyper planes are considered for input sentiments classification. The width of the hyper plane shows its strength for classification. This width is called as margin. Less margin leads to poor classification where as big margin gives better classification. All input sentiments data points are called as vectors and the vector points which are very close to the margin from both side are called as support vectors.

SVM benefits

- SVM has regularization parameter which is useful to avoiding over fitting
- It can model nonlinear decision boundaries.
- It is robust for high dimensional space.
- It can model real world problems like text classification bio-informatics analysis.
- The generalization quality and training to SVM is better than many traditional methods.

The performance of SVM is decided by the classification rate or error rate. We may consider the time performance to indicate how fast it provides result for given set of data.
New sentiments are they mapped in to the same space and category in which they may fall is predicted.

Working of SVM
(Identify correct hyper plane)

Case 1:

We have three hyper planes P Q R now sentiments can be classified using star and circle as follows

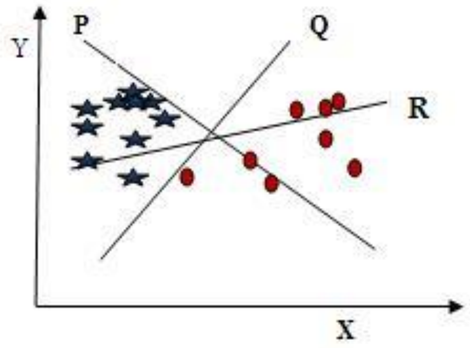

Fig. 2(a) SVM Based Classification Using Three Hyper Planes

Here we will select the hyper plane which will devide to classes of sentiment better. Hyper plane Q is excellent for this moto.

Case 2:

We have three hyper planes P, Q and R. all these are dividing the classes for input sentiments

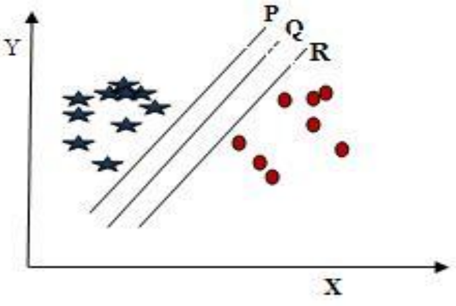

Fig. 2(b) SVM Based Classification Using Three Hyper Planes

We can maximize the distances between nearest sentiments data points to decide correct hyper plane for sentiment classification. These distance is called as

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

margin as shown in following figure

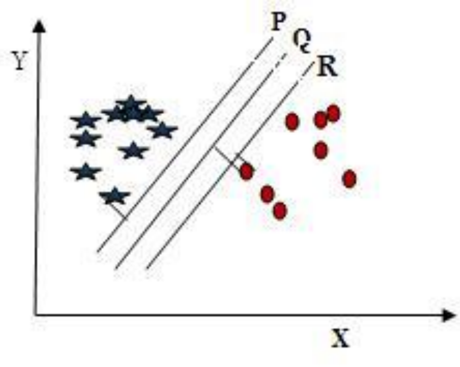

Fig. 2(c) SVM Based Classification Using Three Hyper Planes

In above figure hyper plane Q has high margin as compared to hyper plane P and R. so hyper plane Q is correct for better classification of sentiments. If we select low margin hyper plane, there is more changes of miss classification of sentiments.

Case 3:

Consider there are two hyper planes P and Q used for classification of sentiments here hyper plane P has better margin than hyper plane Q. but SVM gives preference to accurate classification of sentiments than maximizing margin between data points. Hyper plane P has classification error where as hyper plane Q classifies sentiment correctly.

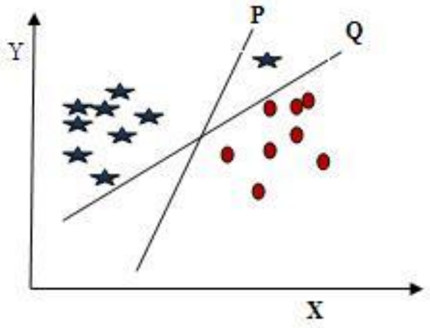

Fig. 2(d) SVM Based Classification Using Three Hyper Planes

Case 4:

Sometimes we may not have linear hyper plane between classes for classifying sentiments.

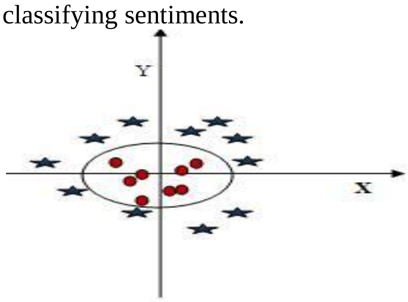

Fig. 2(e) SVM Based Classification Using Three Hyper Planes

SVM will introduce additional feature $z = x^2 + y^2$ .The new plot for sentiments data points will be on axis x and z.Or values for z will be positive as $z = x^2 + y^2$.So red color sentiments data point will be close to the origin of x and y axis and star shaped sentiment data points will be away from the origin with higher value of z.

Sentiment Evaluation

The performance of different text classifier is evaluated by using some performance measures. Precision, recall gives us the relevant and retrieved sentiments. True (TP), true negative (TN), false positive (FP) and false negative (FN) are important performance measures.

- Precision=TP/(TP+FP)
- Recall=TP/(TP+FN)

The most important averages are micro average, which count search document equally important, and macro average, which count search category equally important.

Case Study of Ice-cream or Chornato Using SVM

The Challenge

Classify recipes as ice-cream or chornato.When a new recipe is given; determine if it is ice-cream or chornato.

Steps

1. Find data
2. Apply data science model
3. Review the results

1. Find Data

We can find top 10 recipes for chornato and icecream by firing query in search engine. This will give us the proper input to build a model for prediction.

Table 2.Recorded Top 10 Icecream And Chornato Recipes

Recipe	Flour	Sugar	Butter	Vanilla
Chornato	2 cups	½ cup	¼ cup	1 tsp
Ice-cream	2 cups	¾ cup	½ cup	1 tsp
…	…	…	…	…

Table 2 contents can be explored as amount based recipes.

Table 3. Amount Based Recipes

Recipe	Flour	Sugar	Other
Chornato	2 cup	½ cup	…
Ice-cream	2cup	¾ cup	…

Table 2 can also be explored as percentage based recipes

Table 4. Percentage Based Recipes

Recipe	Flour	Sugar	Other	Total volume
Chornato	47%	24%	…	100%
Ice-cream	42%	21%	…	100%

Table 5.Collection of Training Data For Model Design

Type	Flour	Sugar	Butter
Chornato	55	3	7
Chornato	47	12	6
Chornato	47	18	6
Chornato	50	12	6
Chornato	55	3	7
Chornato	54	7	5
Chornato	47	10	10
Chornato	50	17	8
Chornato	50	17	11
Ice-cream	39	26	19
Ice-cream	34	20	20
Ice-cream	39	17	19
Ice-cream	38	23	15
Ice-cream	42	25	9
Ice-cream	36	21	14
Ice-cream	38	31	8
Ice-cream	36	24	12
Ice-cream	34	23	11

In above Table 5, different combination of contents in percentage for existence of chornato and icecream are shown.

2. Apply Data Science Model

a) Import libraries
The Pandas, numpy and sklearn are imported to allow chart to appear in the note book

b) Import data
Recipe data .csv file imported which contains type of item with detailed contents.

c) Prepare data
Plot two ingredients Sugar at X-axis and Butter at Y-axis.

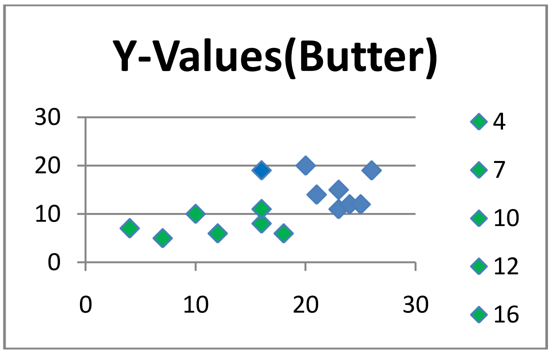

Fig. 3 Data Points Description for Butter against Sugar

d) Fit the model
From the sugar verses butter matrix, by comparing the values we can build a model which will be helpful us to predict item type for any new content.

e) Visualize result
Here we will draw the hyper plane to separate two classes. Parallel to this hyper plane more hyper planes are found. A hyper plane with maximum margin is preferred and its support vectors are located.

f) Predict new case
A new case with item contents will be raised as challenge. We will plot a point for new case in our data scatter graph. Now hyper plane will play an important role to predict the class label for the new case raised.

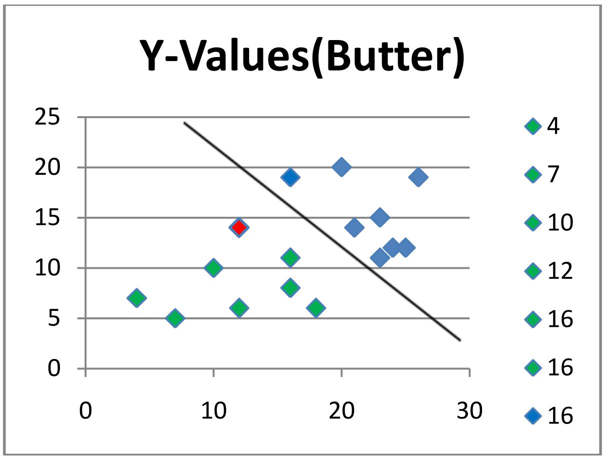

Fig. 4 A New Data Point Laid In Lower Part Of Hyperplane Data Points Description for Butter against Sugar

From the above figure 4, we can predict that the content value for which red colored data point id shown gives the result that it is chornato**.**

V. DISCUSSION AND CONCLUSION

We have compared priority based features for the given input sentiment data. Based on the manual priority we have selected, the sentiment analysis result will vary. Also we can compare the output of SVM for n-grams. . For each case, the time required for getting the polarity for a given set will vary. The proposed technique will give a better analysis of the sentiments which will be proved fruitful for business decisions.

In future one can go for multidimensional SVM planes to improve the results.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79{86, 2002}.)

[2] Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.

[3] M. Hu and B. Liu,"Mining and Summarizing Customer Reviews", Proceedings of the 10th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, (2004).

[4] P. Chaovalit and L. Zhou, "Movie Review Mining: A Comparison between Supervised and
Unsupervised Classification Approaches", In System Sciences, HICSS'05, Proceedings of the 38th Annual Hawaii International Conference on IEEE, **(2005)**, pp. 112c- 112c.

[5] T. O'Keefe and I. Koprinska, "Feature Selection and Weighting in Sentiment Analysis", Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia. (2009).

[6] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, pp. 1320-1326

[7] Efthymios Koulompis, Theresa Wilson, Johanna Moore: Twitter Sentiment Analysis: The Good the Bad and the OMG!.In: Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media, 2011

[8] F. M. F. Wong, S. Sen, and M. Chiang, "Why Watching Movie Tweets Won't Tell the Whole Story?," *Arxiv preprint arXiv:1203.4642*, no. c, p. 6, Mar. 2012.

[9] Hassan Saif, Yulan He and Harith Alani: Semantic Sentiment Analysis of Twitter. In: Proceedings of the 11th International Semantic Web Conference, 2012.

[10] Gann W-JK, Day J, Zhou S. (2014). Twitter analytics for insider trading fraud detection system. Presented at second ASE international conference on Big Data.

[11] Jensen MJ, Jorba L, Anduiza E. Introduction. In E.Anduiza, M. Jensen, & L. Jorba (Eds.), Digital media and political engagement worldwide: A comparative study. New York, NY: Cambridge University Press,2012, 1-15

[12] Kothari, A. A., & Patel, W. D. (2015). A Novel Approach Towards Context Based Recommendations Using Support Vector Machine Methodology.Procedia Computer Science, 57, 1171-1178.

[13] Abinash Tripathy, Ankit Agrawal, Santanu Kumar, "Classification of Sentimental Reviews Using Machine Learning Techniques," *3rd International Conference on Recent Trends in Computing, 2015*

[14] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE) , Issue 1, Volume 2 , January 2015

[15] Github:https://github.com/aababu/sentiment-analysis/blob/master/wordstrength.txt

[16] Sourceforge:https://sourceforge.net/projects/java-ml/files/java-ml/

AUTHORS PROFILE

Prakash P.Rokade has received hisB.E.degree in Computer Pune University, Maharashtra;Indiain 2005.He has received his M.Tech. degree Computer Engineering from Bharti Vidyapeerth, Pune, Maharashtra,India in 2011 and presently pursuing his Ph.D. in Computer Science andEngineering Koneru Lakshmaiah Education Foundation, formerly K L University, Vaddeswaram , Andhra Pradesh, India.His research interest includes Sentiment Analysis, Opinion Mining, and Machine Learning.

Dr. ArunaKumari D Professor at Department of CSE,VJIT, Hyderabad, Telangana India . She has received her Ph.D. degree in Computer Science and Engineering from the K L University, Vaddeswaram, AndhraPradesh, India. Currently, She is Professor Koneru Lakshmaiah Education Foundation, formerly K L University. Her teaching and research areas include in data Mining, Machine Learning and has published more than 50 papers in many National, International journals. She is honored by DST Young Scientist Award (Govt. of India).