International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

Sentiment Analysis of Product Reviews Using SVM

K. Deepa1, E. Kirubakaran[2]
[1]Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli, India
[2]Head, Department of Computer Science, Karunya University,Coimbatore, India
deepamohan13@gmail.com, ekirubakaran@gmail.com

Abstract: In recent days, Sentiment analysis has gained much attention. Sentiment analysis and opinion mining are important module of NLP (Natural Language Processing).This paper challenges the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. The process of sentiment polarity categorization is offered with detailed process descriptions. Data used in this study are online product reviews collected from Amazon.com. An experiment for review-level categorization is performed with capable outcomes. The different levels of review score are collected for different books using support vector machine.

Keywords: Sentiment Analysis, Sentiment Polarity Categorization, Natural Language Processing, Product Reviews, SVM.

I. INTRODUCTION

Sentiments are derived from emotions. Emotions are relay on common people, studies people's sentiments towards certain entities [5].Internet is a resourceful place with respect to sentiment information. According to user's perspective , people post their own view about the particular product through various social media such as online social networking and micro blogs.From a researcher's perspective, many social media sites release their application programming interfaces (APIs), Prompting data collection and analysis by researchers and developers. For instance, Twitter currently has three different versions available, namely the REST API, the Search API, and the Streaming API. With the REST API, developers are able to gather status data and user information, the Search API allows developers to query specific Twitter Content, whereas the Streaming API is able to collect Twitter content in real time. Moreover, developers can mix those APIs to create their own applications. Hence, sentiment analysis seems to have a strong fundamental with the support of massive online data. The first flaw is that the people can freely post their own content, the quality of their opinions cannot be guaranteed. For example, instead of sharing topic-related opinions, online spammers post spam on the forums.

Some spams are meaningless and others have irrelevant opinions are also known as fake opinions. The second flaw is that ground truth of such online data is not always available.

II. PROBLEM HANDLING

This paper tackles a fundamental problem of sentiment analysis, namely sentiment polarity categorization [5-9]. Fig 2 depicts our proposed process for categorization. Review of particular product aims to classify it as positive sentiment [True positive, False positive] or Negative sentiment [True negative, False negative] and Neutral. Our Contributions mainly fall into different phases 1)A mathematical approach is proposed for sentiment score computation 2)A feature vector generation method is presented about sentiment polarity categorization. 3) Experiments are respectively performed based on review level categorization for different products. 4) The performance of two classification models is evaluated and computed based on their experimental results.

Star Level	General Meaning
★★★★★	Excellent Product.
★★★★	I Like It.
★★★	Not Extremely Happy, But Its Ok.
★★	Very Poor.
★	Pathetic Product.

Fig 1.Rating System for Amazon.Com

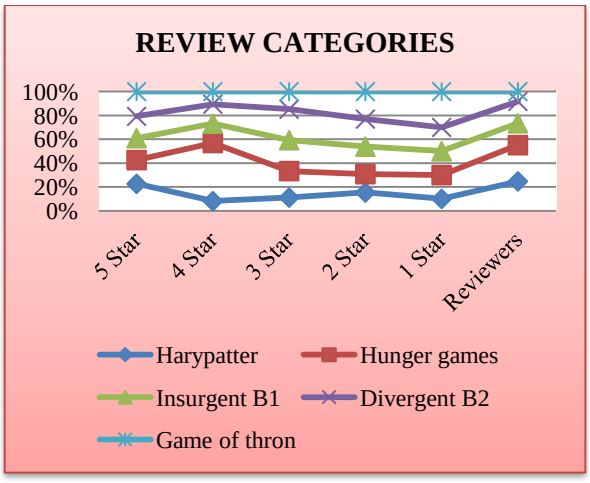
III. RESEARCH DESIGN AND METHODOLOGY

In order to achieve our main goals, it is imperative to do some sentiment analysis on the data set to extract people's opinion about the products they have bought. As far as we know, there is no published work about sentiment analysis in Amazon reviews.

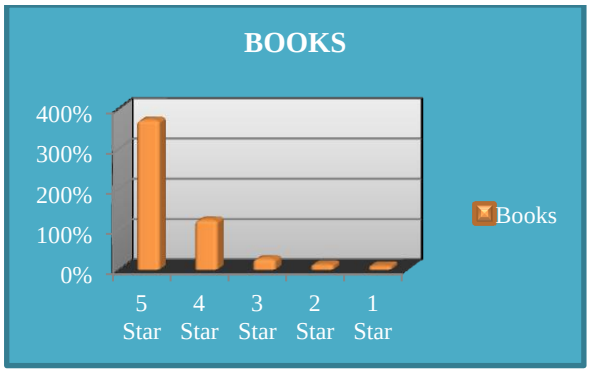
Data Collection:

Data used in this paper is a set of product reviews collected from Amazon.com. From September to November 2017, we collected in total, over 5 lakhs product reviews in which the products belong to categorie: books. Those online reviews were posted by over 5 lakhs of Reviewers towards 1,98, 242 products. Each review includes the following information:1)Reviewer ID 2)Product ID 3)rating4)time of the review 5) helpfulness6)review text. Every rating is based on a 5-star scale (fig 3(b) resulting all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star.

International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com



a) Data Based on Product Categories



b) Data Based on Review Categories

Fig 1. Data collection: a) Data based on product categories b) Data based on review categories

Image Table

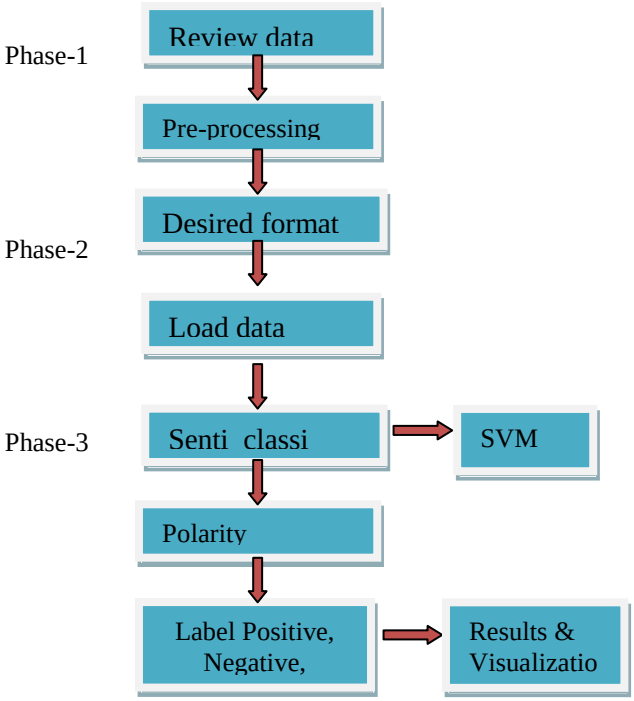


Fig. 2

In the Fig 2. A flowchart of sentiment analysis is represented which gives the general flow of process sentiment analysis.

Data Sets: The data used is a set of product reviews collected from amazon.com on mobile phones.

Data Pre-processing: As the dataset is from Amazon.com, the data are in the form of text. The text data is highly prone to inconsistencies. This step is very important as it extract out unwanted words from tweets. To make the data more relevant for analysis, text reprocessing is performed.

Sentiment Polarity Calculation: Sentiment polarities are divided into four categories like True positive & False positive And True negative & False negative.

Classifiers

Support Vector Machine (SVM): Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data, the algorithm outputs an optimal hyper plane which categorizes new examples. It is a non-probabilistic algorithm which is used to separate data linearly and nonlinearly. Where XT is test tuple, αi and b0 are numeric parameters; yi is the class label of support vector Xi. So If the sign is positive of MMH equation then XT comes in positive category. If the sign is negative of MMH equation then XT comes in the negative category. SVM classifier formula is defined as following.

$f(x)=\sum_{i=1}^{n} ai\ k(x,xi) +b$

IV. RELATED WORK

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g. sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, effect analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia, both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study. Sentiment analysis and opinion mining mainly focus on opinions which express or imply positive or negative sentiments.

In [1] author used data mining techniques for the purpose of classification to perform sentiment analysis on the views people have shared on Twitter. The data was collected from twitter that is in natural language and apply text mining techniques –tokenization,

International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

stemming etc..., to convert them into useful form and then use it for building sentiment classifier that is able to predict happy, sad and neutral sentiments for a particular tweet. The rapid Miner tool is being used, that helps in building the classifier as well as able to apply it to the testing dataset. In [2] paper focused on aspect level opinion mining and proposed a new syntactic based approach for it, which uses syntactic dependency, aggregate score of opinion words, Senti_WordNet and aspect table together for opinion mining. The experimental work was done on restaurant reviews. The dataset of restaurant reviews was collected from the web and tagged manually. This paper [3] Deals with the fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.com are selected as the data used for the study. A sentiment polarity categorization process has been proposed. Experiments for both sentence-level categorization and review-level categorization have been performed.

IN [6, 7-10] deals with the paper One fundamental problem in sentiment analysis is categorization of sentiment polarity [6,7-10]. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral).

Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level [11]. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions.

V. EXPERIMENTAL RESULTS

Procedure: Senti_Classi /* SVM classifier to measure accuracy*/

```
Procedure: Senti_Classi /* SVM classifier to
measure accuracy*/

Input : Labeled positive, negative and neutral
tweets
Output : Accuracy

 1. Begin
2. Train model using classified tweets.
3. Build Classifier Model
4. Test the Model using the newly identified
opinionated tweets
5. Result analysis
6. End
```

Experiments were performed on a dataset obtained by extracting product reviews from Amazon.com. We focused on the mobile phone domain. Considering reviews of one product at a time sentiment of the reviews was classified into four categories namely True positive, False positive, True negative, False negative. Using the proposed algorithm the results obtained for 500 product reviews.

Table 1. Confusion Matrix

	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

The performance matrix is used to calculate classification accuracy.

Accuracy: It is one of the most common performance evaluation parameter and it is calculated as the ratio of the number of correctly predicted reviews to the number of total number of reviews present in the corpus. The formula for calculating accuracy is given as:

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

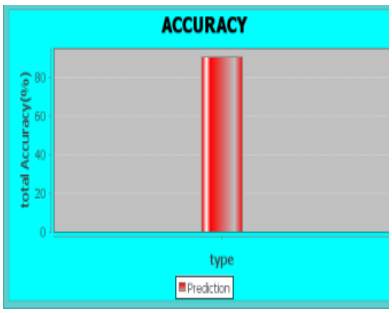


Fig. 4. Accuracy

Accurately represents what percent of prediction were correct. The percentage obtained during precision is 90.47%.Proposed approach gives higher accuracy than existing systems.

F-measure: F-measure is the harmonic mean of precision and recall.

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

Precision: Precision is the fraction of the documents retrieved that are relevant to the user's information need. It is the number of correct results divided by the number of all returned results. The percentage obtained during precision is 87%,

$$\text{Precision} = \frac{Tp}{Tp+Fp}$$

International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

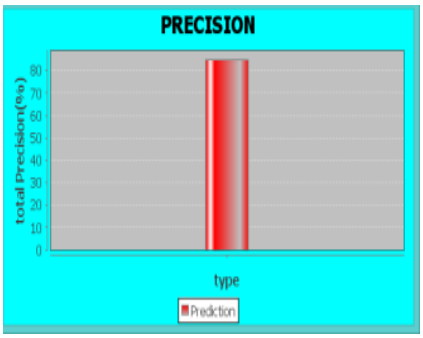


Fig. 5. Precision

Recall: Recall is the fraction of the documents that are relevant to the query, which are successfully retrieved.

Recall = $\frac{Tp}{Tp+Fn}$

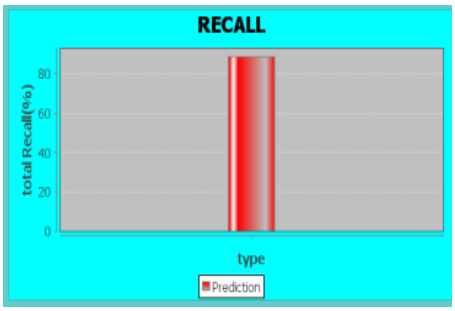


Fig. 6. Recall

It is the number of correct results divided by the number] results that should have been returned. The percentage obtained during recall is 90%.

True positive (TP) is correctly identified.

True negative (TN) is correctly rejected.

False positive (FP) is incorrectly identified.

False negative (FN) is incorrectly rejected.

Graphical Representation:

Table 2. Confusion Matrix for Books Dataset

	Correct Labels	
	Positive	Negative
Positive	371%	123%
Negative	13%	10%

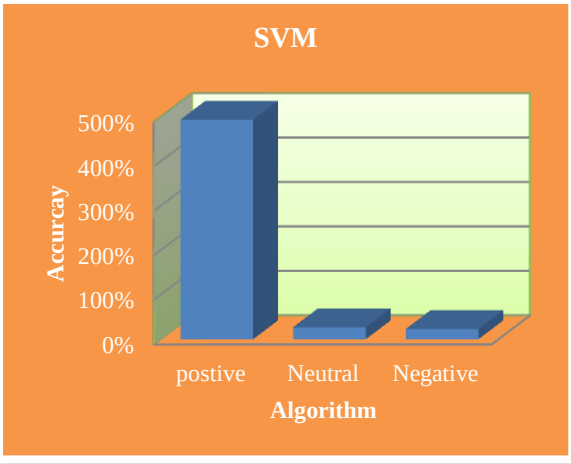
Table 3. Accuracy with Dataset and Classifiers

Data Set	**Positive**	**Neutral**	**Negative**
SVM	494%	27%	23%

Table 4. Table Depicts the Percentage Obtained During Precision, Recall and Accuracy

	Percentage
Recall	90%
Precision	87%
Accuracy	90.47%

Fig. 7. Accuracy Dataset using SVM



VI. CONCLUSION

Categorization of reviews assists the customers to make an informed choice on whether to buy a product or not based on its True positive, False positive, True negative and False negative points by reducing the time that they would have spent reading through a loads of reviews. The proposed approach in this paper tries to predict sentiments from reviews posted by users on the Amazon.com. There are many SVM kernel functions available with many hyper parameters. These values can be modified to improve accuracy.And these extracted features are then added to form feature vector. There are different machine learning classifiers to classify the tweets. From our results, we have shown that Support vector machine performs well and also provide higher accuracy. The results show that we get 98 % accuracy form SVM classifier.

So we can increase the accuracy of classification as we increase the training data. By this project we can say that feature vector performs better for tweets related to FMCG (Fast Moving Consuming Goods) reviews. Therefore, looking forward next step to progress and enhance the review level categorization.

VII. REFERENCES

[1]. P. Tripathi,S. Kr. Vishwakarma, A. Lala, ”Sentiment Analysis of English Tweets Using Rapid Miner,” 2015IEEE International Conference on Computational Intelligence and Communication Networks.

[2]. T.C. Chinsha , S. Joseph " A Syntactic Approach for Aspect Based Opinion Mining," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing.

[3] X. Fang and J. Zhan," Sentiment analysis using product review data" Springer. 2015.

[4] Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA. pp 168–177.

[5] Xing Fang and Justin Zhan "Sentiment analysis using product review data" Fang and Zhan Journal of Big Data (2015) 2:5.

[6] Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1-2):1–135

[7] Chesley P, Vincent B, Xu L, Srihari RK (2006) Using verbs and adjectives to automatically classify blog sentiment. Training 580(263):233.

[8] Choi Y, Cardie C (2009) Adapting a polarity lexicon using integer linear programming for domain-specific sentiment Classification. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, EMNLP '09. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 590–598.

[9] Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th, Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 151–160.

[10] Tan LK-W, Na J-C, Theng Y-L, Chang K (2011) Sentence-level sentiment polarity classification using a linguistic approach. In: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation. Springer, Heidelberg, Germany. pp 77–87.

[11] Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan& Claypool Publishers.