Predictive Modeling and Sentiment Analysis: Data Mining Approach

Mr. Vijay D. Chougule[1], Mrs. Anis N. Mulla[2]

[1] Student of M.E.(CSE),CSE Department, ADCET, Maharashtra, India
[2] Assistant Professor, CSE Department, ADCET, Maharashtra, India

---***---

Abstract - ***Data mining technology have widely been applied in various businesses and manufacturing companies. Sharing data has become a trend among business partnerships, as it is supposed to be a mutually beneficial way of increasing productivity. In this proposed work, we use sentiment analysis and prediction modeling to determine future scope of product. For sentiment analysis we take as an example online review of peoples towards the product they bought and services they received. Analysis of different online reviews on large scale will help to produce useful actionable knowledge. Conducting extensive experiments on large data set confirms the effectiveness of the proposed approach.***

Key Words: *Review mining, Sentiment analysis, Prediction modeling.*

1. INTRODUCTION

Posting online reviews has a popular way for people to share with other users their opinions and sentiments toward products and services. It has become a common practice for e-commerce websites to provide the venues and facilities for people to publish their reviews, with a famous example being Amazon (www. amazon.com) or (www. snapdeal.com). Those online reviews provide a wealth of information of the products and services to the user. As a result, review mining has recently got a huge response.

An increasing number of recent studies have focused on the economic values of reviews, searching the relationship between the sales performance of products and their reviews [1], [2]. Since what the public thinks of a product about how well it sells, understanding the opinions and sentiments expressed in the relevant reviews is having more importance, because collectively these reviews reflect the what the general public think and can be a very good indicator of the product's upcoming sales performance. Here we concerned with generating actionable knowledge by developing models and algorithms that can use information mined from reviews. Such models and algorithms can be used to effectively predict the future sales of products.

Prediction of product sales is a highly domain-driven task, for which we use highly recommended product list. In this system the product which is on top of table is declared as a highly recommended product. As user is submitted his reviews about product the list is going to adjust all the time. From this user knows which product is best.

2. LITERATURE SURVEY

The presented methodology lies between the fields of sentiment analysis and predictive modeling. The work that has been presented in this project can be considered as an effort along this direction in that it aims to deliver actionable knowledge by making predictions of sales performance.

Review mining has attracted a great deal of attention. Early work in this area was determining the semantic orientation of reviews. Among them, some of the studies attempt to learn a positive/negative classifier at the file level. Pang et al. [3] employ three machine learning approaches (Naive Bayes, Maximum Entropy, and Support Vector Machine) to tag the polarity of IMDB movie reviews. In follow-up work, they suggest to first extract the subjective portion of text with a graph min-cut algorithm, and then feed them into the sentiment classifier [4]. Instead of applying the straightforward frequency-based bag-of-words feature selection methods, Whitelaw et al. [5] defined the concept of "adjectival appraisal groups" headed by an appraising adjective and optionally customized by words like "not" or "very." Each appraisal group was further assigned four types of features: attitude, orientation, graduation, and polarity. They account good classification accuracy using the appraisal groups. They also show that the classification correctness can be further boosted when they are combined with standard "bag-of-words" features. We use the same words and phrases from the assessment groups to compute the reviews' feature vectors, as we also believe that such adjective appraisal words play a very important role in sentiment mining and need to be distinguished from other words.

Academics have also recognized the impact of online reviews on business intelligence, and have produced some vital results in this area. Among them, some studies attempt to answer the question of whether the polarity and the volume of reviews available online have a quantifiable and significant effect on actual customer purchasing [6], [1]. Various economic functions have been utilized in examining revenue growth, stock trading volume change as well as the order price variation on commercial websites, such as Amazon and eBay.

By studying above, the sentiments are captured by explicit ranking indication such as the number of stars;

few studies have attempted to exploit text mining strategies for emotion classification. To fill in this gap, Ghose and Ipeirotis [2] argue that review texts contain richer information that cannot be easily acquired using simple numerical ratings. In their study, they assign a "dollar value" to a compilation of adjective-noun pairs, as good as adverb-verb pairs, and investigate how they affect the bidding prices of various products at Amazon. Compared to sentiment mining, identifying the quality of online reviews has got relatively less attention. A few modern studies along this direction attempt to detect the spam or low-quality posts that exist in online reviews. Jindal and Liu [7] present a categorization of review spams, and propose some novel techniques to detect different types of spams. Liu et al. [8] propose a classification-based approach to discriminate the low quality reviews from others, in the expect that such a filtering strategy can be incorporated to enhance the task of opinion summarization.

3. PROPOSED SYSTEM

The following are the steps in the proposed system:-

1. Request for Item: In this system, the single user or multiple users will request for items.

2. Dataset (Sale Performance): Dataset can be formed according to sale performance of that product.

3. Recommended Product List: User selects product for purchasing and gives his review or opinion or sentiment about product.

4. Review mining: Then by taking no. of reviews from users, we made review mining.

5. Sentiment wordlist: It contains database of no. of sentiment wordlist which are found in review mining.

6. Weighting scheme: Then by using weighting scheme we assign weights to no. of words.

7. Prediction model: Then by using prediction model we make some prediction and found some results about the product. Then from that we predict future sales performance of that product. And from that we adjust the ratings of the products in recommended product list. Which is depends on sales performance of that product. By doing this users or customer easily choose items that he want to buy.

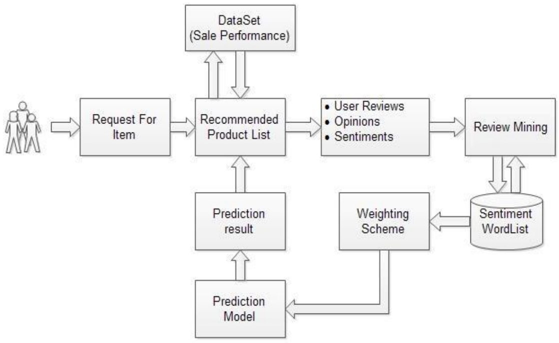

Fig -1: Proposed system architecture

4. METHODOLOGY

This work introduces how to make sentiment analysis on different no. of reviews of peoples. For that different algorithms have implemented in this work. KNN for auto weight generation, Porter stemming algorithm for stemming of words, Naive bayes classifier algorithm for classification of sentimental words.

The proposed six modules are

1. Registration
2. Auto-weight generation
3. POS tagging
4. Preprocessing
5. Extraction of emotions using emotion dictionary
6. Word net
7. Process for making predictions from given reviews

4.1 Registration

In this module customers are login with his username and password. If the customer is new one then he has to fill registration form with some details. Then he can enter into system.

4.2 Auto weight generation

In this work we define training set with emotion category & its weights. Here we are dealing with emotions like Happy, Sad, Fear, Anger and Neutral. Training set is used for weight generation of different word, by using KNN algorithm.

KNN algorithm

The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. Then to determine the distance between the two scenarios, we can simply pass through the data set, one scenario at a time, and compare it to the query scenario.

KNN run in these steps:

1. Store the output values of the nearest neighbors to query scenario in vector r = {r1,….,rm} by repeating the following loop M times:

  a. Go to the next scenario si in the data set, where i is the current iteration within the domain {1,….,P}
  b. If q is not set or q < d (q, si): q ← d (q , si), t ←oi
  c. Loop until we reach the end of the data set (i.e. i = P)
  d. Store q into vector c and t into vector r

2. Calculate the arithmetic mean output across as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=1}^{M} r_i$$

3. Return as $\bar{r}$ the output value for the query scenario q.

4.3 POS Tagging

In this module, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. It makes part of speech tagged with types of words are nouns, verbs, adjectives, adverbs, etc.

4.4 Preprocessing

The data preprocessing is done in order to remove unnecessary content from the text and to find out the root form of the words. In this module, first enter review in textbox and making process on that by using following operations:

1. Stop word removal
2. Stemming
3. Stop word removal

4.4.1 Stop word removal

In emotion computing, stop words are words which are filtered out prior to, or after, processing of natural language data (text). Stop words are common words that carry less important meaning than keyword. These stop words are some of the most common, short function words, such as the, is, at, which and on.

Stop word removal is the process of removing these words. To find out the emotion from a text all unnecessary content must be removed so it is needed to remove the stop words that bear no meaning about Emotion and the text put into an array.

Algorithm:

The following is an algorithm for stop word removal

```
1. Take the Input
2. Declare the dictionary of stop words
3. Split parameter into words
4. Allocate new dictionary to store found words
5. Store results in this String Builder
6. Loop through all words
7. Convert to lowercase
8. If this is a usable word, add it
9. Return string with words removed
10. Display query without stop words
```

4.4.2 Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form-generally a written word form. Here, Porter stemming algorithm is used to make variant forms of a word which are reduced to a common form, for example,

connection
connections
connective ---> connect
connected
connecting

Stemming algorithm:

The algorithm used for stemming is Porter stemmer. The stemmer operations are classified into rules where each of these rules deals with a specific suffix and having certain condition(s) to satisfy. A given word's suffix is checked against each rule in a sequential manner until it matches one, and consequently the conditions in the rule are tested on the stem that may result in a suffix removal or modification. Stemming algorithm consists of different steps of word reductions, applied sequentially. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix.

The algorithm of stemming works as follows:

Rules	Illustrations
S ->	cats -> cat
EED ->	EE agreed->agree
(*v*) ED ->	plastered ->plaster
(*v*) ING ->	motoring ->motor

Rules	Illustration
ATIONAL ->	ATE relational -> relate
TIONAL ->	TION conditional-> condition
IZER ->	IZE digitizer -> digitize
ATION ->	ATE predication -> predicate
IVENESS ->	IVE decisiveness -> decisive

4.5 Extraction of emotion using emotion dictionary

4.5.1 Extraction of emotions

In this module, emotions are extracted from given sentences, which is helpful to assign weights to them.

4.5.2 Classification of emotions

In this module, it gives weights to extracted emotions by using Naive bayes classifier algorithm, also it decides category of given emotion. This is helpful to predict the final conclusion of given paragraph.

Naive Bayes classifier algorithm:

Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. It is used the knowledge of prior events to predict future events.

The Bayes Theorem:

$$P(h\,|\,D) = \frac{P(D\,|\,h)P(h)}{P(D)}$$

P(h) : Prior probability of hypothesis h
P(D) : Prior probability of training data D
P(h/D) : Probability of h given D
P(D/h) : Probability of D given h

This algorithm is used to assign weights to words. From that get category of emotion.

Following are the steps of algorithm that have implemented:

1. Select words from table
2. Check the category from table if it is present otherwise insert into table.
3. Select the max and min from another table where is the training set is available.
4. And then save the weight in table for word.
5. Then read the training set.
6. Set and get the data table for each row and each column.
7. Check the final source and select the max value and name from table
8. Then calculate the average value from source
9. Then calculate sum
10. Then calculate the summation.

Here, get final conclusion of given paragraph by checking probability of category of emotions. By doing this easily predict the sentiment of user.

4.6 WordNet

By using WordNet dictionary, if we don't know the meaning of any word can easily get it. Also it is inserted on table which is helpful for sentiment analysis.

4.7 Process for making predictions from given reviews

Here I take dataset as a product. Given option to user to choose any product. Select any product from given list to see features & prices. Then he gives comment about the product whether it is positive or negative. After submitting the comment the pre processing is done on given comment. i.e. Stop word removing, Stemming, Extraction of emotional features & Classification of emotions. By doing this the exact sentiment has been find out of user about the product. By submitting it, the given product list is going to be adjusted depends upon the comment has given i.e. positive or negative. The product which got good comments is on top of list. After that second, third so on so forth. The prediction has been done based on the count of comments. If it is positive then count is increases otherwise it is decreases. This is helpful to user to choose highly rated product from given list.

5. RESULT DISCUSSION

In this analysis is done by using precision-recall method on given dataset of reviews about products given by user. In this system gives different types of results about reviews i.e. it may be True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN).

1. True Positive (TP) is the proportion of positive cases that were correctly identified.
2. True Negative (TN) is defined as the proportion of negatives cases that were classified correctly.
3. False Positive (FP) is the proportion of negatives cases that were incorrectly classified as positive.
4. False Negative (FN) is the proportion of positives cases that were incorrectly classified as negative.
5. Precision or positive predictive value (PPV) is the proportion of the predicted positive cases that were correct
$$PPV = TP/(TP+FP)$$
6. Recall or true positive rate (TPR) is the proportion of positive cases that were correctly identified.
$$TPR = TP/P = TP/(TP+FN)$$
7. Accuracy (ACC) is the proportion of the total number of predictions that were correct.

Table (1): Precision-Recall table

Dataset	TP	FP	FN	TN	Precision	Recall	Accuracy
Mobile	8	2	0	2	0.8	1	0.833333
Laptop	9	1	2	1	0.9	0.818181	0.769230
car	10	0	2	2	1	0.833333	0.857142

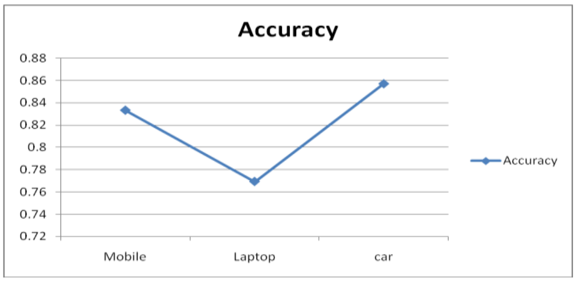

Figure (1): Accuracy Graph

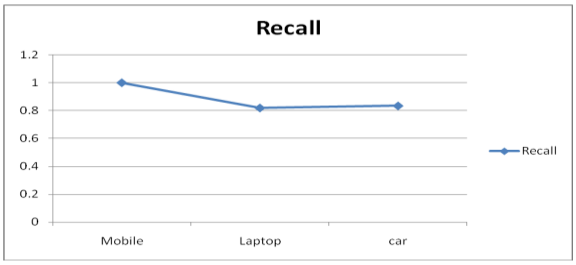

Figure (2): Recall Graph

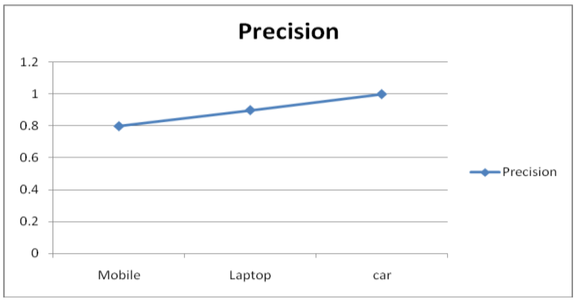

Figure (3): Precision Graph

Final conclusion of analysis is that all the graphs i.e. Figure (1), Figure (2) & Figure (3) show good results. Amongst the several reviews maximum reviews shows true positive results which satisfies our expectations about this work.

6. CONCLUSIONS

In this system, the work is based on the different algorithm used, we can analyze the sentiment of an individual and with the help of analysis done, we predict the emotion that an individual carry while capturing the idea or option. The work is related to the text based emotion mining using different approaches. The analysis of the result obtained will be done by measuring accuracy of emotion prediction. This will satisfy our work to achieve the target.

REFERENCES

[1] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, and Aijun "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", Proc. IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 4, April 2012..

[2] Ghose and P.G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews," Proc. Ninth Int'l Conf. Electronic Commerce (ICEC), pp. 303-310, 2007.

[3] Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002.

[4] Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics (ACL), pp. 271-278, 2004.

[5] Whitelaw, N. Garg, and S. Argamon, "Using Appraisal Groups for Sentiment Analysis," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 625-631, 2005.

[6] N. Archak, A. Ghose, and P.G. Ipeirotis, "Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer reviews,"Proc.13thACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining (KDD), pp. 56-65, 2007.

[7] N. Jindal and B. Liu, "Opinion Spam and Analysis," Proc. Int'l Conf. Web Search and Web Data Mining (WSDM), pp. 219-230, 2008.

[8] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," Proc. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP), pp. 334-342, 2007.

BIOGRAPHIES

Student of M.E.(CSE),CSE
Department, ADCET,
Maharashtra, India

Assistant Professor, CSE
Department, ADCET,
Maharashtra, India