SENTIMENTAL ANALYSIS FOR BUSINESS INTELLIGENCE USING PRODUCT REVIEWS

[1] G. Janani, [2]N. Ramya Devi
[1,2] PG Student
Department of Information Technology,
PSG College of Technology,
Coimbatore-4
[1]jananiguru18@gmail.com, [2] ramyadevi.nandakumar@gmail.com

Abstract— **Opinion mining also known as sentiment analysis which aims to analyze people's opinions, sentiments, and attitudes towards entities such as products, services, and their attributes. The vast existing system extracts opinion features only from the single domain. In this paper we proposed a method to identify the opinion features from different corpus to flourish a new approach to spawn a new product. Supervised learning technique is tuned to work well in the different corpus reviews. In proposed system market segmentation is used to know what are the things exactly expected from different segments of customers and sales channels and we use data mining concept, apriori algorithm is used to analyze the reviews. To distinguish the reviews Machine learning techniques are used. Experimental results on different corpus show the performance of the well established methods in developing a new product.**

Index Terms— **SENTIMENTAL ANALYSIS Opinion mining.**

I. INTRODUCTION

Opinion mining also known as sentiment analysis aims to analyze people's opinions, sentiments, and attitudes toward entities such as products, services, and their attributes. Sentiments or opinions expressed in textual reviews are typically analyzed at various resolutions. Document-level opinion mining identifies the overall subjectivity or sentiment expressed on an entity in a review document, but it does not associate opinions with specific aspects of the entity. This problem also happens, though to a lesser extent, in sentence-level opinion mining, as sincreasing interest in the evaluation of biometric systems. In opinion mining, an opinion feature, or feature in short, indicates an entity or an attribute of an entity on which users express their opinions. A good many approaches have been proposed to extract opinion features in opinion mining.

Many enterprises devote a significant portion of their budget to new product development (NPD) and marketing to make their products distinctive from those of competitors, and better fit the needs and wants of consumers. Hence, knowledge and feedback on customer demand and consumption experience has become an important information and asset for enterprises. Knowledge of the customers and the product itself reflect the needs of the market. Product design and planning for production lines be integrated with the knowledge of customers and market channels. The knowledge of customers and market channels are transformed into knowledge assets of the enterprises during the stage of NPD. The Apriori algorithm is a methodology of association rule for data mining, which is implemented for mining demand chain knowledge from channels and customers. The polarity of the data provided by the customers and sales person are analysed by using the methodology Support Vector Machine (SVM).Knowledge extraction is illustrated as knowledge patterns and rules in order to propose suggestions and solutions to the case firm for New Product Development (NPD) and marketing. The ordering of new product to the manufacturer can be done. Sales report analyzes process is carried out by the manufacturing company.

II. LITERATURE SURVEY

1. Blei et al, proposed the Topic modelling approaches that can mine coarse-grained and generic topics or aspects, which are actually semantic feature clusters or aspects of the specific features commented on explicitly in reviews.

2. Qiu et al, proposed that Unsupervised natural language processing (NLP) approaches identify opinion features by defining domain-independent syntactic templates or rules that capture the dependence roles and local context of the feature terms. However, rules do not work well on colloquial real-life reviews, which lack formal structure.

3. Zhou et al presented a novel semi supervised learning algorithm called Active Deep Networks (ADN), to address the semi-supervised sentiment classification problem with active learning. First, it is proposed that the semi-supervised learning method of ADN. ADN is constructed by Restricted Boltzmann Machines (RBM) with unsupervised learning using labelled data and abundant of unlabeled data. Then the constructed structure is fine tuned by gradient-descent based supervised learning with an exponential loss function. Second, active learning method is applied in the semi-supervised learning framework to identify reviews that should be labelled as training data. Then ADN architecture is trained by the selected labelled data and all unlabeled data. Experiments on five

sentiment classification datasets show that ADN outperforms the semi-supervised learning algorithm and deep learning techniques applied for sentiment classification.

4. Qu L et al proposed that a regression method based on the bag of opinions model was proposed for review rating prediction from sparse text patterns. Review rating estimation is a much more complicated problem compared to binary sentiment classification. Generally, sentiments are expressed differently in different domains. The sentiment classification methods discussed in the above mentioned papers can be tuned to work very well on a given domain; however, they may fail in classifying sentiments in a different domain.

5. Zhai Zhongwu, et al proposed the method of Clustering product features for opinion mining that focuses on the classic methods based on unsupervised learning using some forms of distributional similarity. However, it is found that these methods do not do well. Then model it as a semi-supervised learning problem. Lexical characteristics of the problem are exploited to automatically identify some labelled examples. Empirical evaluation shows that the proposed method outperforms existing state-of-the-art methods by a large margin.

6. Zirn, et al introduced the field of sentiment analysis and opinion mining and surveyed the current state-of-the-art. It has spread from computer science to management science as opinions about products are closely related to profits.

7. Zhang Y., et al proposed that Implicit feature extraction is a relatively new research field. Whereas previous works focused on finding the correct implicit feature in a sentence, given the fact that one is known to be present, the above research aims at finding the right implicit feature without this pre-knowledge. To distinguish between sentences that have an implicit feature and the ones that do not, a threshold parameter is introduced, filtering out potential features whose score is too low. Using restaurant reviews and product reviews, the threshold-based approach improves the F1-measure by 3.6 and 8.7 percentage points, respectively.

III. SYSTEM ARCHITECTURE

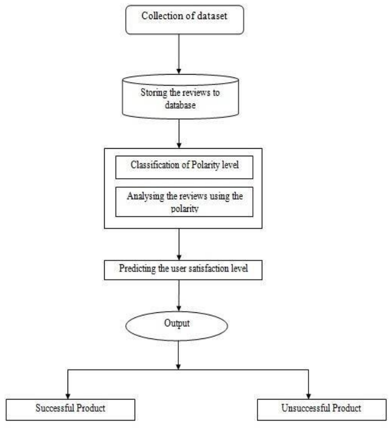

Collection of dataset
Storing the reviews to database
Classification of Polarity level
Analysing the reviews using the polarity
Predicting the user satisfaction level
Output
Successful Product
Unsuccessful Product

Figure 1: System Architecture

A. EXISTING SYSTEM:

Existing System has become common practice for people to communicate or express their opinions and feedbacks on various aspects affecting their daily life through some form of social media. Most of the online interactions are in the form of natural language text. This in turn has led to increased research interest in content-organization and knowledge engineering tasks such as automatic classification, summarization, and opinion mining from web-based data. Due to its high commercial importance, mining and summarizing of user reviews are a widely studied application. All the data only from customers and sales & channels were computerized. There is no proper way to analyze and there is no proper system for decision making to develop a new product. This is the reason for the failure of new products of companies. Another disadvantage is there is no possible ways for marketing strategies without proper analyzing of the data collected from the customers and sales & channels.

B. Proposed system:

In Proposed system market segmentation is used to know that what are the things exactly expected from different segments of customers and sales channels. Data mining concept, apriori algorithm is used to retrieve the frequently occurred complaints and suggestions provided by the customer and sales persons and to define the polarity of the reviews, opinion mining concept Support Vector Machine is used. It classifies the reviews into positive and negative categories based on its polarity level. In this system admin holds the authority to analyse and generate the sales report about the product.

IV. MODULES OF THE SYSTEM

- Extraction of dataset
- Pre-Processing of the dataset
- Applying association rule mining.
- Opinion classification.
- Predicting the customer needs about the particular product.

A. Extraction of dataset:

In new product development each and every reviews of the product is important. The reviews are collected from different corpora. The reviews vary according to the corpora. Customer reviews are mainly based on the product, its spare parts and reviews are based on the internal features of the product. So collecting the reviews from different corpora leads to producing a successful new product.

B. Analysing and clustering the data:

The reviews are clustered according to the frequent item sets. The reviews that are stored in the database are analysed and clustered using the association rule mining such as apriori algorithm. The frequently occurred complaints and suggestions are retrieved for the further development of the product. The frequently occurred suggestions and feedbacks are analysed and stored separately.

C. Opinion classification:

After analysing the reviews of different corpora the polarity of the reviews are defined. To find the polarity, opinion mining technique called Support Vector Machine (SVM) is used. It classifies the reviews into positive and negative categories. The training dataset is provided to SVM to classify the reviews into its respective categories. SVM calculates the marginal value to find the polarity.

D. Predicting the customer needs about the particular product:

In order to predict the customer needs and feedback the polarity of reviews are classified. Many enterprises devote a significant portion of their budget to new product development (NPD) and marketing to make their products distinctive from those of competitors and better fit the needs of consumers. Hence knowledge and feedback on customer demand and consumption experience has become an important information and asset for enterprises.

V. Methodology

A. Data Pre-processing:

All the irrelevant information in the review files are removed which are not necessary for analysis. Data preprocessing include cleaning of data by removing numbers, symbols and the extra useless information e.g. date, name of the reviewer or reference of any third person which are not relevant. It includes tokenization, stop words removal and stemming.

- Tokenization

Text document has a collection of sentences which is split up into terms or tokens by removing white spaces, commas and other symbols.

- Stop words removal

Stop words are words which are filtered out prior to, or after, processing of natural language data. It removes articles like a, an, the, etc. It also removes unwanted words.

- Stemming

Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form.

B. Apriori Algorithm:

In association rule mining, Apriori algorithm is used. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. It uses the level wise search.

```
Algorithm:
Find all large 1-itemsets
For (k = 2 ; while Lk-1 is non-empty; k++)
            {$C_k$ = apriori-gen($L_{k-1}$)
    For each c in $C_k$, initialise c.count to zero
            For all records r in the DB
        {$C_r$ = subset($C_k$, r); For each c in Cr , c.count++ }
    Set $L_k$ := all c in $C_k$ whose count >= minsup
    } /* end -- return all of the $L_k$ sets.
```

Terminologies in Apriori algorithm:

k-item set : a set of k items.

E.g: {mirror ,steering, wheel} is a 3-itemset
{headlight} is a 1-itemset
{seat, indicator} is a 2-itemset

support: an item set has support s% if s% of the records in the DB contain that item set.

minimum support: the Apriori algorithm starts with the specification of a minimum level of support, and will focus on item sets with this level or above.

large item set:. It means one whose support is at least minimum support.

L_k : the set of all large k-item sets in the DB.

C_k : a set of candidate large k-item sets.

4.3 Classification:

To classify the reviews in to two categories, SVM is used. Support Vector Machine is a supervised machine learning technique that analyzes the data and recognize the patterns used for classification. It finds a hyper plane which separates the dimensions into two different classes. SVM classifies the reviews into positive and negative categories.

To calculate the support vector value:

f(x) = Σ α i y i (xiT x) + b

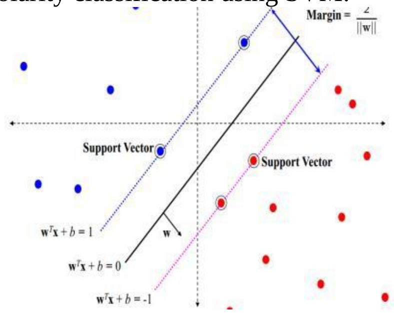
Xi is support vector.

SVM can be formulated as:

maxw 2/||w|| subject to w>xi+b ⩾1 if yi =+1⩽ −1 if yi = −1 for i = 1. . . N

```
Algorithm for Simple SVM:
Simple SVM
Candidate SV = {closest pair from opposite classes}
while there are violating points do
Find a violator candidate SV = candidate SV S U violator
if any αp < 0 due to addition of c to S then
candidate SV = candidate SV \ p
repeat till all such points are pruned
end if
end while.
```

Polarity classification using SVM:



6. Experimental results:

Figure 2: Data Extraction

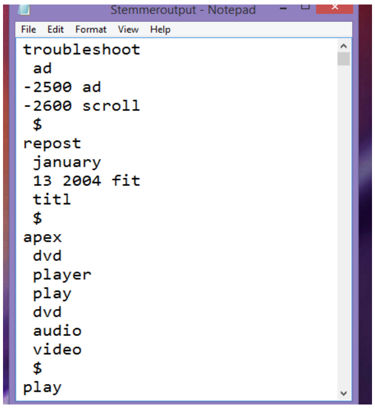
Figure 3: Storing data in DB

1. Wide Rear Tyre give you Awesome Control on road (this one very important for safety)2. Good Looks (Tanks consists of 3 piece so if any damage we can easily replace it at low cost) 3. Low end power gives you less gear shifting 4. Mileage totally depends on maintenance.

Table 1: Sample Data

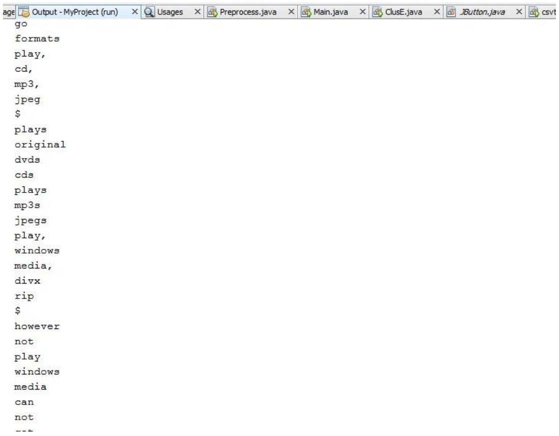
```
Visit www stata com  Datasets for Stata
Press books  Books listed alphabetically by
author   Discovering Structural Equation
Modeling Using Stata  Alan C  Acock  A
Gentle Introduction Stata  Fourth Edition
```

Figure 4: Pre-processing



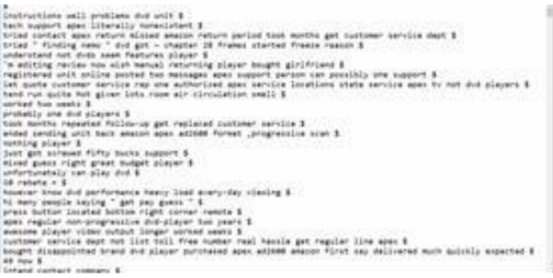
```
troubleshoot
 ad
-2500 ad
-2600 scroll
 $
repost
 january
 13 2004 fit
 titl
 $
apex
 dvd
 player
 play
 dvd
 audio
 video
 $
play
```

Figure 5: Stemmer output



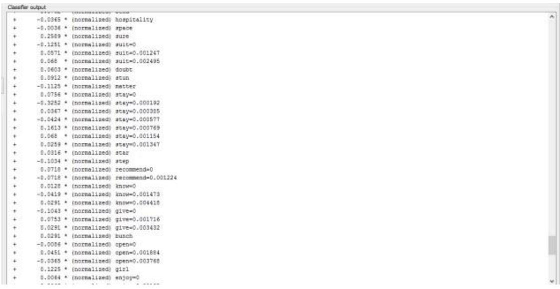
```
go
formats
play,
cd,
mp3,
jpeg
$
plays
original
dvds
cds
plays
mp3s
jpegs
play,
windows
media,
divx
rip
$
however
not
play
windows
media
can
not
```

Figure 6: Stop words Removal



```
$
instructions small problems dvd unit $
tech support apex literally nonexistent $
tried contact apex return missed amazon return period took months get customer service dept $
tried " finding nemo " dvd got - chapter 28 frames started freeze reason $
understand not dvds seem features player $
'm editing review now wish manual returning player bought girlfriend $
registered unit online posted two messages apex support person can possibly one support $
let quote customer service rep one authorized apex service locations state service apex tv not dvd players $
hand run quite hot given lots room air circulation small $
worked two weeks $
probably one dvd players $
took months repeated follow-up got replaced customer service $
ended sending unit back amazon apex ad2600 format ,progressive scan $
nothing player $
just got screwed fifty bucks support $
mixed guess right great budget player $
unfortunately can play dvd $
10 rebate = $
however know dvd performance heavy load every-day viewing $
hi many people saying " get pay guess " $
press button located bottom right corner remote $
apex regular non-progressive dvd-player two years $
awesome player video output longer worked weeks $
customer service dept not list toll free number real hassle get regular line apex $
bought disappointed brand dvd player purchased apex ad2600 amazon first day delivered much quickly expected $
40 now $
intend contact company $
```

Figure 7: After removing duplicate features.



```
 +       -0.0365 * (normalized) hospitality
 +       -0.0036 * (normalized) space
 +        0.2589 * (normalized) sure
 +       -0.1251 * (normalized) suit=0
 +        0.0571 * (normalized) suit=0.001247
 +        0.068  * (normalized) suit=0.002495
 +        0.0603 * (normalized) doubt
 +        0.0912 * (normalized) stun
 +       -0.1125 * (normalized) matter
 +        0.0756 * (normalized) stay=0
 +       -0.3252 * (normalized) stay=0.000192
 +        0.0367 * (normalized) stay=0.000385
 +       -0.0424 * (normalized) stay=0.000577
 +        0.1613 * (normalized) stay=0.000769
 +        0.068  * (normalized) stay=0.001154
 +        0.0259 * (normalized) stay=0.001347
 +        0.0316 * (normalized) star
 +       -0.1034 * (normalized) step
 +        0.0718 * (normalized) recommend=0
 +       -0.0718 * (normalized) recommend=0.001224
 +        0.0128 * (normalized) know=0
 +       -0.0419 * (normalized) know=0.001473
 +        0.0291 * (normalized) know=0.004418
 +       -0.1043 * (normalized) give=0
 +        0.0753 * (normalized) give=0.001716
 +        0.0291 * (normalized) give=0.003432
 +        0.0291 * (normalized) bunch
 +       -0.0086 * (normalized) open=0
 +        0.0451 * (normalized) open=0.001884
 +       -0.0365 * (normalized) open=0.003768
 +        0.1225 * (normalized) girl
 +        0.0064 * (normalized) enjoy=0
```

Figure 8: Classifier Output

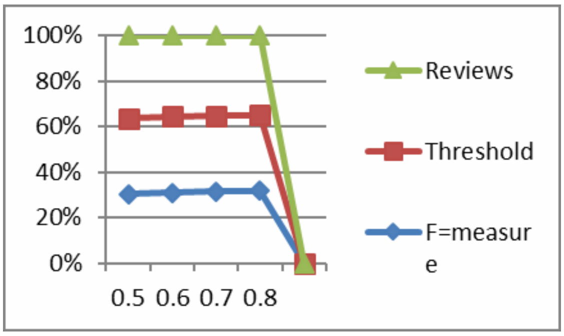


Figure 9: Performance analysis

CONCLUSION

When it is compared with the LDA (Latent Drichlet Allocation) based approach the reviews are gathered together from corpus to analyze the sales report and feedback of a product. But in previous works the single domain reviews are only analyzed and the other corpus reviews are neglected. In this approach, different corpus reviews are analysed with appropriate algorithms and the polarity of the reviews are classified. The proposed system found that using a domain independent corpus of a similar size as but topically different from the given review domain will yield good opinion feature extraction results. According to the frequently occurred polarity the product rating is fixed and the new product is developed according to the customer needs and expectations. Finally a feedback graph is generated to know the current status of the product. This graph helps to identify the satisfaction level of the product among the customers.

Scope of the project

With the help of the proposed system the manufacturer can develop a new product according to the user's suggestions. The algorithm used in this project is efficient and fast. We prove that SVM achieves the highest accuracy in text emotional polarity classification scene based on our experimental results. Thus the reviews are labelled as positive and negative. Using the label we can predict the rating of a product. Further enhancement can be done using other machine learning techniques which would give a better accuracy in terms of performance parameters. In future we can use a enhanced algorithm to develop the product sales.

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167,May 2012.

[2] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, pp. 9-27, 2011.

[3] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.

[4] L. Qu, G. Ifrim, and G. Weikum, "The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 913-921, 2010.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003. I. Titov and R. McDonald, "Modeling Online Reviews with Multi-Grain Topic Models," Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.

[6] I.Titov and R. McDonald, "Modeling Online Reviews with Multi- Grain Topic Models," Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.

[7] R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured Models for Fine-to-Coarse Sentiment Analysis," Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics, pp. 432-439, 2007.

[8] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003.

[9] Erik Cambria, Yangqiu Song, Haixun Wang, Newton Howard, "Semantic Multidimensional Scaling for Open- Domain Sentiment Analysis" IEEE INTELLIGENT SYSTEMS, MARCH/APRIL 2014.

[10] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," Proc. Int'l Conf. Language Resources and Evaluation, European Language Resources Association, 2006, pp. 417–422.

[11] Li S, Zhou G, Wang Z, Lee SYM, Wang R (2011) Imbalanced sentiment classification. In: Proceedings of CIKM-2011, pp 2469–2472.

[12] Tang H, Tan S, Cheng X (2009).A survey on sentiment detection of reviews. Expert Systems with Application 36(7):10760–10773.

[13] Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp 440–447.

[14] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.

[15] W. Jin and H.H. Ho, "A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 465-472, 2009.

[16] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th Int'l Conf. World Wide Web, pp. 959-968, 2008.