IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATED CUSTOMER OPINION MINING USING LEXICON BASED APPROACH SENTIMENT ANALYSIS

[1]Letticia Tendai Tagwira, [2]Wellington Makondo,[3] Chiedza Hwata, [4]Trevor Makausi, , [5]Wellington Manjoro, [6]Walter Mambodza

[1]Student, [2, 3, 4, 5, 6] Lecturer

[1, 2]Department of Information Technology, [3]Department of Computer Science, [4, 5]Department of Software Engineering, [6]Department of Information Security and Assurance

[1]Harare Institute of Technology, Harare, Zimbabwe

Abstract: In our daily lives, we take our friends opinions and sometimes influenced by them in our decision making process. Any kind of organization usually makes use of suggestion box to ascertain how its stakeholders feel about their services, but few people make use of it. More people are using internet on a daily basis, the researchers developed an internet based system to enable stakeholders to input their opinions. Sentiment analysis has been used to evaluate the stakeholder's comments. Vadersharp was used as the basis for opinions, which is a common lexicon that contains data that have the same semantic meaning or sense across different domains. The opinions are analyzed and classified into positive, negative or neutral sentiment. Sentiment analysis computationally analyses the public emotions and attitude towards a particular subject. It is useful in monitoring business operations as it allows decision makers to gain an insight of the public opinion behind certain topics.

Index Terms **- Sentiment Analysis, Opinion Mining, Lexicon Based, Semantic, Sentiment Classification.**

I. INTRODUCTION

Well-known tasks for Lexicon Based techniques involve general sentiment analysis for various types of texts and speech. Such models are capable to output the general sentiment of a sentence or piece of text. However, a special subset of tasks such as mining product reviews requires more than high-level sentiment analysis; rather it requires entity level sentiment analysis on the level of review-specific features. Other applications are public opinion predictions, opinion mining or emotion detection. In many applications, the goal is to perform this sentiment analysis over time. Sentiment analysis can be done for other tasks such as summarizing tasks or recommender systems. This project focuses on implementing sentiment analysis to customer feedback within different organizations. The lexicon based approach works on the assumption that the collective polarity of a sentence or document is the sum of polarities of the individual phrases or words. It focuses on emotional research for sentiment analysis dictionaries for each domain. Each domain dictionary is replenished with appraisal words of appropriate training collection that have the highest weight, calculated by the method of Relevance Frequency. The word-modifier increases or decreases the weight of the following appraisal word by a certain percentage. Negation of the word shifts the appraisal word's weight by a certain offset for negative, it increases and for positive words, it decreases.

II. LITERATURE REVIEW

Sentiment analysis (sometimes known as opinion mining) determines the attitude of a writer, speaker, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to an event, document, or interaction. Subsequently, this concept was adopted and enhanced by Turney [1] and Pang et al. [2]. In the following year, the concept was carried on by Nasukawa & Yi [3] and Yi et.

Sentiment analysis is commonly applied to comments of the customer materials such as online and social media reviews, responses from surveys. It is also implemented in applications which range from marketing to customer service as well as to clinical medicine. Sentiment analysis are extensively studied at different levels such as document level, sentence level, and attribute or feature level.

Classification Based on Supervised Learning

Supervised learning techniques can be applied to naïve Bayesian, Support vector machine (SVM) and other algorithms as well. These algorithms are compared in [4] with an ANN-based method in the context of document-level sentiment classification. They adopted classic supervised methods for feature selection and weighting in a traditional bag-of-words model. Experiments were conducted from reviews extracted from amazon.com in GPS, Books, Cameras and Movies reviews dataset. They find that ANN outperformed statistically SVM, especially in the context of unbalanced data in terms of classification accuracy on dataset of Movies Reviews.

Volcani and Fogel [5] described the method which looked specifically at sentiment and identified individual words and phrases in text with respect to different emotional scales. Many other subsequent efforts were less sophisticated using a mere polar view of sentiment, from positive to negative, such as work by Pang et al [6] who applied different methods for detecting the polarity of product reviews and movie reviews respectively. Bo pang et al [7], applied meta algorithm, based on metric labeling formulation of the problem. They considered generalizing to finer grained scales and attempted to apply numerical rating such as numerical rating like three stars or four stars. They also consider what item similarity measure to apply purposing on based on the positive sentences. Dataset collected from internet movies and in English from four authors. The authors examine pairs of reviews attempting to determine whether the review of each pair was either more positive than less positive than or a positive as the second. There are three classes of metric labeling on top of OVA and regression and it shows that employing explicit similarities always improve the result often to a significant degree and yield the overall best accuracies.

Classification Based on Unsupervised Learning

As reported by many researchers in [8, 9, and 10], supervised learning method can also be formulated as unsupervised learning method. In [8] authors presented an unsupervised learning algorithm for classifying a review as recommended or not recommended. For experiment on dataset taken from Epinions, the algorithm attains an average accuracy of 74%. However accuracy on movie reviews was about 66% and for Travel reviews was 80% to 84% accuracy.

Hu and Liu [11] summarized a list of positive words and a list of negative words respectively based on customer reviews. The negative list has 4783 words and the positive list contains 2006 words. Some misspelled words were also included in both lists that were often present in social media content. Sentiment categorization is a classification problem where features that contain opinions should be recognized before the classification. Pang and Lee [5] suggested to remove objective sentences for feature selection by extracting subjective ones.

A text-categorization technique was proposed to identify subjective content using minimum cut. Gann et al. [12] selected 6,799 tokens based on Twitter data, where each token is assigned a sentiment score, namely TSI (Total Sentiment Index), featuring itself as a negative token or a positive token.

Document Level Sentiment Classification

Apart from the document-level sentiment classification, researchers have also studied classification at the sentence level that is classifying each sentence as a subjective or objective sentence and/or as expressing a positive or negative opinion. In [13] Iqu et al. the weakly supervised Multi Expert Model (MEM) was proposed for evaluating the semantic orientation of opinions in natural language reviews. The semantic orientation consists of polarity (positive, negative, or other) and strength. However author uses Base Predictors was used to predict the polarity or single phrase rating which is divided into following four predictors which are: Statistical Polarity Predictor, Heuristic Polarity Predictor, Bag-of-Opinions Rating and SOCAIL Rating Predictor.

In experiments MEM implemented as well as the HCRF classifier. The datasets are taken from various sources like amazon.com on different domains. It was found that base predictors has poor performance across all domains, primarily due to the aforementioned problems related with averaging phrase level predictions. However authors also find that when no fine-grained annotations are available both MEM-Coarse and Majority-Vote outperformed HCRF-Coarse. MEM-Coarse also perform better then Majority-Vote.

Schneider et al., (2009) [14] proposed a novel matrix learning strategy for extending relevance learning vector quantization (RLVQ), in which the correlations between various attributes as well as their significance for classification occurs at the time of training. When contrasted with weighted Euclidean measure utilized in RLVQ as well as its variants, a complete matrix is more powerful for representing the internal structure of data adequately. Huge margin generalization bounds may be transferred to the case resulting in bounds that are not dependent on input dimensionality. This is true for local measures attached to all prototypes that correspond to piece-wise quadratic decision bounds. The protocol was evaluated in contrast to alternate LVQ strategies through usage of artificial dataset, a benchmark multiclass issue from UCI repository, as well as a problem from bio-informatics, the recognition of splice sites for C.

Application of sentiment analysis in recommended system

For the recommended system, sentiment analysis is therefore proven to be a valuable technique and also from the above literature review it is concluded that the use of the Rule Based Approach [15] implementing supervised and unsupervised learning yields a higher performance accuracy at the sentence level. However the efficiency and accuracy depend on the defined rules and this means that a well-defined dataset must be used as the training set.

III. SENTIMENT ANALYSIS METHODOLOGY

In this research, customers posted their comments online and the following techniques were used in order to determine the comments` polarity [16]:

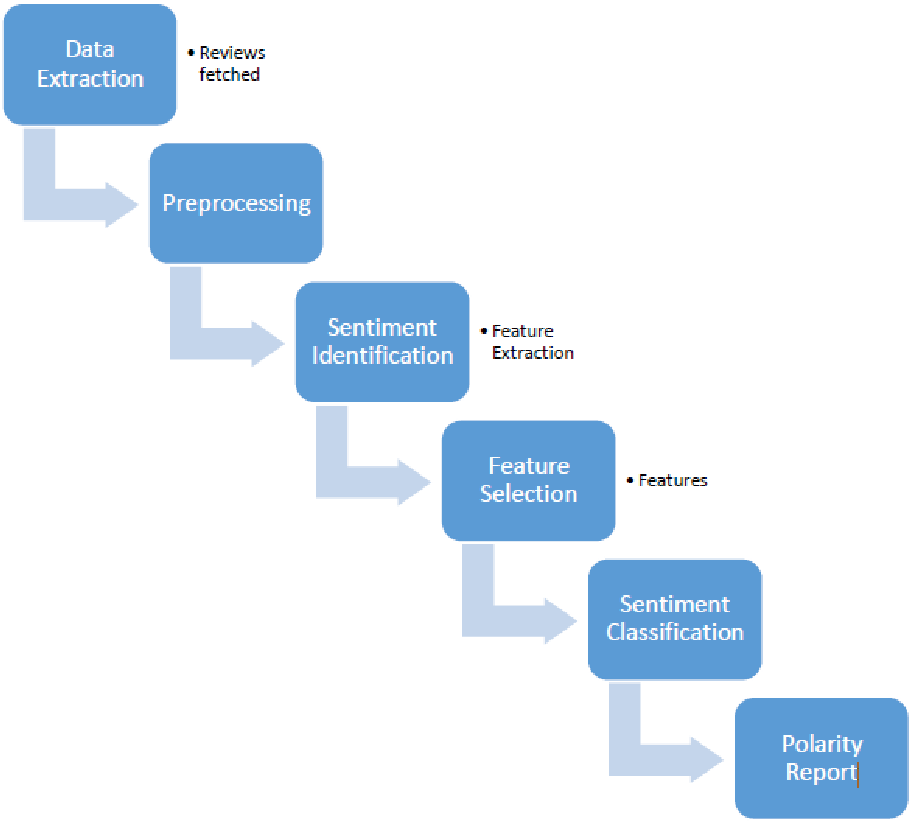

Fig 1. Sentiment Analysis process

Step1: Data Extraction

Consumers post their comments on the feedback page; these comments can either be public or private comments. Opinions and feelings are conveyed in different vocabulary, making the data huge and disorganized, usage of short forms and slang, context of writing. Manual analysis of sentiment data is virtually impossible. Therefore, special languages are used to process and analyze the data.

Step2: Pre-processing

This step involves the cleaning of data, readying the text for classification. Customer feedback usually contain lots of noise and unnecessary parts. Pre-processing the data helps to reduce the noise, which helps to improve the performance of the classifier as well as speeding up the classification process thus helping in real time sentiment analysis. Appropriate text pre-processing includes data transformation and filtering can significantly improve the performance. For instance, the implemented preprocessing strategies include:

- Removing Non-English words - Getting rig of non-English words since I am focusing only calculating the Sentiment Analysis of the English words
- Extra letter removal - Words that have same letter more than two times and not present in the lexicon are reduced to the word with the repeating letter occurring just once. For example, the exaggerated word "happyyyyy" is reduced to "happy".
- Stemming - Replacing words like waiting, waits, and waited with the word wait.

Step3: Sentiment Identification

This stage aims to find out the opinionative words or phrases that best describes the context that we are dealing with and it is a basic technique in many Sentiment Analysis applications. Some existing researches exploit seed words, and lead to low robustness.

Step4: Feature Selection

Decreasing the dimensionality of the feature space is the main goal of this stage. Reducing the feature space cuts down the computational cost. It also reduces the over-fitting of the learning scheme to the training data. Finding a good trade-off between the richness of features and the computational constraints involved when solving the categorization task is important as well at this stage.

Step5: Sentiment Classification

Sentiments are classified into three groups, which are positive, negative and neutral. At this stage, each subjective sentence detected is classified intone of the three groups. It involves spotting sentiment words within a particular sentence. The technique used was to use a dictionary of sentiment terms and their semantic orientations. The dictionary contains the synonyms and antonyms of a word. However, the dictionary-based approach has its own disadvantages associated with it. For instance, in the context of "calories" the sentiment term, "low" might have a positive polarity, whereas in the context of "video resolution" the word "low" is of negative polarity.

IV. CLASSIFICATION: THE LEXICON BASED APPROACH

Classification was done using the lexicon-based approach for this paper. It is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase. The main steps are explained below.

Creation of lexicon

A lexicon is a reference dictionary containing an alphabetic list of words with their respective sentiment values that is either positive, neutral or negative. The lexicon in this project was imported directly from the internet and it is called Vadersharp. Vadersharp is a common lexicon that contains data that would have the same semantic meaning or sense across different domains. For example, sentiment word "good" always represents a positive sentiment and it is independent of any category. Positive or negative sentiment words have a sentiment score of +1 or -1 to indicate the respective polarity whilst neutral words are the ones, which fall on neither the positive nor the negative side, and they have a polarity of 0. It also contains the negation words, which reverse the polarity of sentiment. For example, "The battery life is not good" has negative sentiment [17]. In addition spilt words such as "but" and blind negation words are observed like in the sentence "The T.V needs a better remote", the word needs is a blind negation.

Preprocessing

The objective to normalize the text into an appropriate form to extract the sentiments. The preprocessing steps used are stemming, reducing exaggerated words and removing non English words

Sentiment calculation

Sentiment calculation is the aggregation of the sum of the sentiment bearing entities of the comment. The sentiment calculation algorithm is shown below in Algorithm. Blind negation [18] words are extracted from the sentences and their presence indicate negative sentiment.

```
Data: Preprocessed Customer Comments
Result: Output: Positive, Negative, Neutral
Find the list of sentiment words
SentiList, its position in the sentence;
Find the list of sentiment negation words
SentiNegat, its position in the sentence;
Find the list of blind negation words
BlindNegat, its position in the sentence;
  if BlindNegat then
          return negativity;
  else if SentiList and SentiNegat then
  foreach word in the SentiList do
    if word is atmost the distance of 2 from SentiNegat then
          Revert the polarity of the word;
    end
  end
  else if SentiNegat then
          Add the SentiNegat to the negative SentiList;
  end
 end
end
SentiSum=0;
foreach word in the SentiList do
          SentiSum=SentiSum+sentiment of word;
end
  if Hashtag is present then
          From hash tag, find all the sentiment wordsusing regex matching and add them to SentiList
  end
SentiType="neutral";
   if SentiSum > 0 then
          SentiType="positive";
  end
  if SentiSum < 0 then
          SentiType="negative";
  end
return SentiType;
```

V. Results

The sentiment engine performed quite well, the percentage accuracy is shown by the calculations below. The sentiment engine results are shown on the administration panel I table form exactly as they look like in the database. This table is shown in table 1.2 below. For presentations to the organizational executives, the system generated graphs that show daily comments with their respective quantity as well as sentiment classification as shown in fig 1.3 below. The accuracy rate is quite low due to the informal language of some comments, which ended up being classified as neutral comments. On the positive and negative sentiments, the engine performed really well considering that, the comments were written in formal language.

% Accuracy = (Positive accuracy + Negative accuracy + Neutral accuracy) (100) (1)

$$= \frac{positive\ accurate}{total\ positive} + \frac{negative\ accurate}{total\ negative} + \frac{neutral\ accurate}{total\ neutral} * 100 \qquad (2)$$

$$= \left(\frac{37}{40} + \frac{20}{22} + \frac{15}{18} \right) (100)$$

$$= 90\%$$

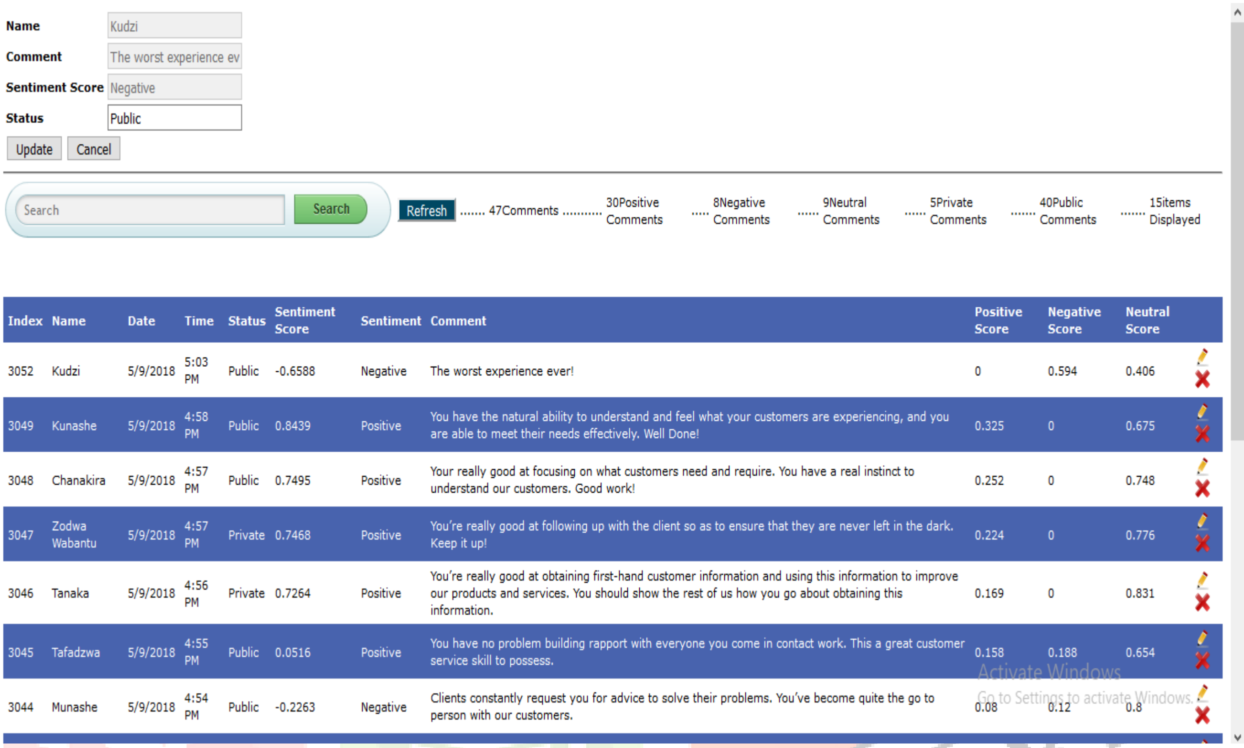

Fig 2. Administrator panel

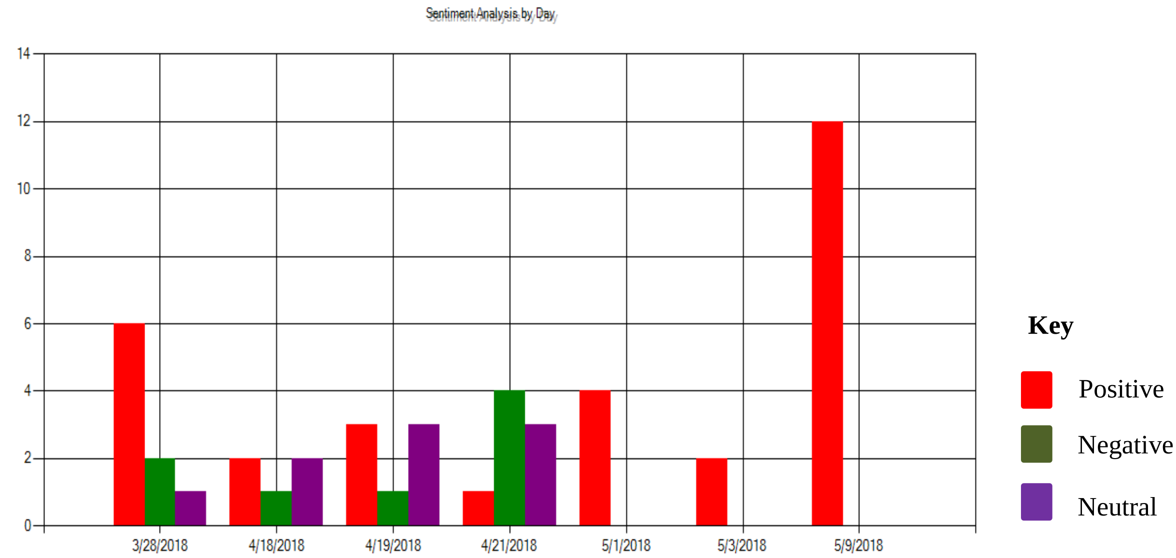

Fig 3. Sentiment Analysis Report

VI. CONCLUSION

This paper presented a system that used the Lexicon based approach to perform sentiment analysis on customer feedback comments. Practical approaches to identify and extract sentiments from customer comments were provided. There was no need for training since a dictionary was used. For better results, a dictionary containing the informal languages such as slang and non-English words must be used to determine their sentiment values as well.

REFERENCES

[1] P. Turney, ―Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews‖, ACL'02, Philadelphia, Pennsylvania, 2002, pp. 417 - 424.
[2] R. Kant, Srinivasan H. Sengamedu, K. Kumar, ―WSDM'12, February 8–12, 2012, Seattle, Washington, USA.
[3] T. Nasukawa and J. Yi, ―Sentiment analysis: Capturing favorability using natural language processing‖, in Proceedings of the Conference on Knowledge Capture (K-CAP), 2003.
[4] Rodrigo Moraes, Joao Francisco Valiati*, Wilson .Gaviao Neto, Document level sentiment classification: An empirical comparison between SVM and ANN‖, Springer 2012.
[5] USA Issued 7,136,877, Volcani, Yanon; & Fogel, David B., "System and method for determining and controlling the impact of text", published June28, 2001.
[6] Bishop, C. M. Pattern recognition and machine learning. Vol. 4. 2006: springer New York.
[7] A. Mukherjee, B. Liu, N. Glance, ―Spotting Fake Reviewer Groups in Consumer Reviews‖, WWW 2012, April 16–20, 2012, Lyon, France , ACM
[8] Turney P, ―Thumbs up or Thumbs down? Semantic Orientation applied to unsupervised classification Reviews‖, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2009), 2009.
[9] X. Glorot, A. Bordes, Y. Bengio,―Lexicon-based methods for sentiment analysis‖, Computational Intelligence, 2010.
[10] Narisawa, K., Bannai, H., Hatano, K., Takeda, and M.: ―Unsupervised spam detection based on string alienness measures‖, In Proc. DS'07, 161-172, LNAI, 2007.
[11] Algorithm" Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.
[12] Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system. In: Proceedings of the second ASE international conference on Big Data. ASE
[13] L. Qu, R. Gemulla, G. Weikum , ―A weakly Supervised Model for Sentence-Level Semantic Orientation Analysis with Multiple Expert‖, In Proceedings of the 2012, JOINT Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 149-159, Jeju Island, Korea.
[14] Mining opinion from text documents: A survey
[15] Swati A. Kawathekar1, Dr. Manali M. Kshirsagar, "Movie Review analysis using Rule-Based &Support Vector Machines methods", IOSR Journal of Engineering Mar. 2012, Vol. 2(3), March. 2012, pp: 389-391.
[16] https://www.edureka.co/blog/sentiment-analysis-methodology/
[17] http://ijcsmc.com/docs/papers/February2016/V5I2201636.pdf
[18] Simple and Practical lexicon based approach to Sentiment Analysis :Prabu Palanisamy, Vineet Yadav and Harsha Elchuri Serendio Software Pvt Ltd Guindy, Chennai 600032, India
[19] Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in twitter: Proceedings of the 19th ACM international conference on Information and knowledge management 1833–1836, ACM.