International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Gradual Weight Updating for Sentiment Mining

Sugandha Nandedkar, Jayantrao Patil, Sunil Kawale

Abstract*: Nowadays, many people prefer the use of social media for communicating and exchanging opinions with each other over face to face communication. This has lead to a generation of a tremendous amount of textual opinioned data. Understanding this opinioned data is useful from all perspectives. But the major challenge exists here is how to extract the exact sentiment hidden behind this huge data. To solve this problem, keyword spotting or dictionary-based approaches are followed. In this paper, we present a Gradual Weight Updating for sentiment mining. It not only considers the polarity of each word similar to the unigram methodology but, it also focuses on the entire cluster of words that contains the unigram. The different steps it follows for sentiment extraction of the word are polarity fetching, cluster marking, weight tagging, valence shifter, adversative conjunction handling, and final score generation. The paper contributions in the area of domain independent opinioned word extraction and accurate polarity mining with the help of context marking approach. We used the various opinionated datasets to compare and illustrate the performance of our proposed system.

***Keyword*: Natural Language Processing, Opinion Mining, Sentiment Analysis, Text Mining**

I. INTRODUCTION

The development of Web 2.0 has completely changed the way people use to express their opinion on any topic. Instead of face to face communication, people prefer to exchange their opinion using different social media platforms. Day by day increasing number of users of social media such as Facebook, Twitter, WhatsApp, blogs, etc. represents the same. Understanding this opinion shared on social media is very important. It helps the service provider to understand consumers' opinions and sentiments about their product or service. Thus the information can be used to improve the product quality and sales. At the same time, this information is helpful for the new consumer to understand the opinion and sentiment of existing consumers and to select the right product or service as per their requirements. In all the existing consumers' opinion shared on the social media platform is playing a key role in sales market management [1], [2]. The era of sentiment mining includes natural language processing, text data mining, computational linguistics, etc.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Sugandha Nandedkar, Department of Computer Science and Engineering, Dr. B. A. M. University, Aurangabad (MS), India.
Prof. Dr. Jayantrao Patil, Department of Computer Science and Engineering, R. C. Patel College of Engineering, Shirpur (MS), India.
Prof. Dr. Sunil Kawale, Post Graduate Department of Statistics, Dr. B. A. M. University, Aurangabad (MS), India.

Sentiment mining plays a vital role in analyzing customer reviews and survey responses for a variety of entities on online and social media platforms. The real challenge that arises at the time of this opinion sentiment analysis is its tremendous volume. In this process, we have to find out the exact emotion tone behind words [3]. For a human analyzer, it is not possible to analyze millions of tweets and comments arriving in a day. On the other side, it is very difficult for a machine to grab the exact sentiment behind the opinion text proposed by the consumer [4].

In this paper, we initially propose the study of existing machine learning techniques for consumers' sentiment mining. Keywords-spotting is one of the simplest techniques used for it. In normal texts, it performs well but it drastically fails for mixed sentiment sentences such as "The food was tasty but they took so much time to prepare and serve it." Similarly, it also faces trouble for the deceptive type of opinions [5]. Initially, we studied keyword spotting and attaching polarity scores from dictionaries such as AFFIN and BING. Their working is simply based on the concept of n-gram terms. Instead of the simple n-gram terms sentiment score summation we present here another methodology to calculate sentiment polarity. It is not only based on the polarity of the word itself but also considers the words surrounding it. Thus we can improve the sentiment score generation task by considering this additional information.

The paper is structured in sections as follows: introduction, the study of the related work, proposed system architecture, experimental set-up and results, conclusion and future scope.

II. RELATED WORK

Sentiment mining is applied at different levels such as document level, sentence level and aspect level. The document level sentiment mining approach detects sentiment polarity for the whole document as positive / negative / neutral. Sentence level sentiment mining detects sentiment polarity for each sentence. Some researchers have tried for more fine level - aspect level sentiment mining. Our Work is closely related to Hu and Bing work in [1] on mining and summarizing customer reviews at sentence level. Their work is divided into three stages. During the first stage, the feature word from the opinioned sentence is extracted. The process follows the syntactic rule based approach. In the second stage, the adjective present in the sentence is searched and the polarity score in the range of -3 to +3 is attached to it. In the last stage document summary is generated. The approach performs well for feature based summary of product reviews. Here the coherent feature and opinion word pair is present in the sentence. The approach lags for handling the interdependence of pronouns, verbs, adverbs, etc. present in opinioned sentences.

This thread of research is continued by Yu and Weixiang in [4] on extracting implicit features for opinion mining. They called implicit feature for those sentences where the feature/target word and opinion word pair is unclear. Their work focused on the co-occurrence matrix and modification matrix. These matrices are formed based on the corpus. By following the bootstrapping approach the author tries to collect the hidden target word for the opinion word. The concept of co-occurrence of feature opinion pair is extended by K. Lui et al. as Partially Supervised Word Alignment Model (PSWAM) in [3] and by Hybrid Approach in [2], [11]. This approach is the mere keyword spotting with the help of a predefined dictionary. Our approach differs from it by considering the entire context to determine the sentiment polarity.

Considering the context for polarity determination is the other thread of research in sentiment mining. In line with it, Y. Wu and F. Ren implemented a model for learning the influence of others' opinions on the sentiments of a user. The work focused on finding the influence probability model for twitter sentiment analysis. The study contributed for finding propagation of sentiments in social media. The same thread is continued by S. Tan et al. in [7] as a generalized LDA based model for interpreting public sentiment variation on twitter. The impact of foreground topic and background topic on a user's opinion is analyzed here. R. Xia et al. proposed the use of sentiment revered corpus along with the original corpus as a Dual Sentiment Analysis model. Such dual training helps to improve the performance of sentiment classifier for a deceptive type of opinion but degrades for others. To attend this, L. Yu. et al. [9] adopted a novel strategy to learn sentiment embeddings. Instead of creating a new word embedding from labeled corpora, they have proposed a word vector refinement model. In simple words, the word vector is converted into real values sentiment intensity score. The proposed refinement model improves each word vector such that it can be closure in the lexicon to both semantically and sentimentally similar words. It refers to the words with similar intensity scores. It further considers the words with dissimilar intensity scores also. Further G. Xu et al. proposed an improved word representation method in [10]. They integrated the contribution of sentiment information into the traditional TF-IDF algorithm and generated weighted word vectors. Bidirectional Long Short Term Memory (BiLSTM) inputs weighted word vectors to capture the context information effectively. The study also focused on building comment vectors. The sentiment tendency of the comment is obtained by a feed-forward neural network classifier.

The immediate surrounding words such as polarity valence shifter, adversative conjunctions, etc. are yet not considered properly for polarity determination. Our research paper focuses on this research gap. The following section illustrates the proposed system architecture.

III. PRPOPOSED SYSTEM ARCHITECTURE

As per the discussion in the above section, we identified that the surrounding context should also be considered for determining the final polarity score of the sentence. For illustration purpose, we used the dictionary based approach along with the proposed method of Gradual Weight Updating (GWU) for sentiment polarity calculation.

A. Dictionary based approach

Standard dictionaries such as AFFIN, BING, NRC are used here. The steps followed are,

1. Break the entire corpus in sentences. Further, break these sentences in tokens.
2. Pre-process the corpus by removing punctuation marks, stop words, stemming, etc.
3. Fetch the polarity value for each token of the sentence from the standard dictionary.
4. Calculate summation for the entire sentence. Depending upon the numeric score determine the polarity as positive, negative or neutral.

B. Gradual Weight Updating (GWU) approach

GWU approximates the sentiment polarity (positive/negative/neutral) of the sentence in the text. In this process, the sentiment polarity weights are updated gradually based on the words present in the cluster. The surrounding words may be intensifier or detoner words which may either increase or decrease the intensity of the target opinion word. e.g. in the sentence "The governing system of the machine is very accurate" and "The governing system of the machine is less accurate". The word 'very' is the polarity intensifier and the word 'less' is detoner for the opinion word 'accurate'. The proposed GWU architecture follows the steps mentioned in fig. 1and illustrated below. The basic steps involved are polarity fetching, cluster marking, weight tagging, valence shifter, adversative conjunction handling, and final score calculation. Our attempt is to find the exact sentiment with the help of all these stages. It can also be called as context analysis for sentiment mining.

- **Polarity Fetching:** In this step, document (D) is divided into paragraphs (p_i). Each paragraph (p_i) is subdivided into sentences (s_{ij}) and each (s_{ij}) is further split into words as w_{ijk}. On the basis of polarized words dictionary, polarity (positive, negative or neutral) is attached with each word as (pw_{ijk}). Here the initial polarity value of each word is estimated by using standard dictionaries. The initial polarity value (pw_{ijk}) is revised by further steps.
- **Cluster Marking:** In this step, the peripheral of the word is marked. This is the most crucial step, as it determines the words we will be considering for final score calculation of the token. It is also known as polarized context cluster marking for the word (pw_{ijk}). The cluster of (pw_{ijk}) is represented by C= {(pw_{ijk-b}), …… , (pw_{ijk}), ……. (pw_{ijk+a})}. It consists of a few words before the opinioned word and a few words after it. We can choose our values for 'a' and 'b'. Thus the cluster size is flexible but, it is restricted by punctuation marks like comma, colon, semicolon, etc. The cluster words may be the intensifier or down toner or neutral. In the further step, we will calculate the impact of these words on the polarity of the opinioned word.

Retrieval Number: B4099129219/2019©BEIESP
DOI: 10.35940/ijeat.B4099.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- **Weight Tagging:** The weight for a polarized word (pw_{ijk}) is decided cumulatively by the polarized word dictionary and the words within its cluster. Depending upon the surrounding words its polarity weight may be intensified or may be degraded or kept as it is. It highly depends upon the part of speech tag of the word and its relationship with the words in the cluster.
- **Valence Shifter:** Despite intensifier and down-toner words, the polarity cluster may also consist of flip or valence shifter words such as *'not'*, *'mis'*, *'dis'* etc. For example in the sentence, *"This is not a good product"*, *'not'* is a valence shifter of the word *'good'*. Such negation flips the sign of polarized words (pw_{ijk}). A simple principle of 2^i is used to determine the final polarity shift present in the sentence.
- **Adversative Conjunction Handling:** Adversative conjunction words such as 'but', 'however', 'although' etc. also affect the polarity weight of the word. If it appears before the word in the cluster then it increases the weight value of the word. If it appears after the word in the cluster then it decreases the weight value of the word. For example in the sentence, *"This is a good product but it has many bugs"*, the word *'but'* is going to reduce the weight of the word *'good'*.
- **Final Score Calculation:** Based on the above steps B, C, D and E final score is calculated for each sentence. Depending upon this score polarity of the sentence is decided. The results are discussed and compared in the next section.

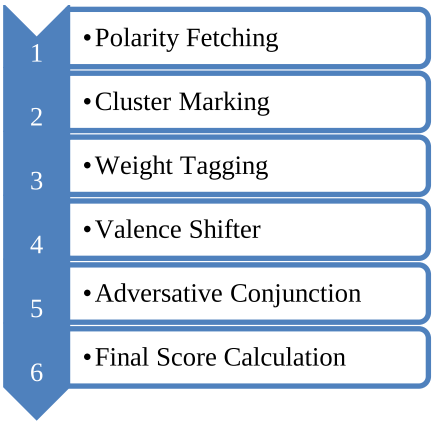


Fig. 1: Steps for GWU

IV. EXPERIMENTAL SETUP AND RESULTS

We evaluate the dictionary based approach and the GWU on various datasets. Dataset 1 is from https://archive.ics.uci. edu /ml/datasets and dataset 2 is from https://www.cs.uic.edu /~liub/FBS/sentiment-analysis.html. It contains sentences labeled with positive or negative sentiment, extracted from reviews of products, movies, and restaurants. The labeling task is done manually. This pre-labeling we will be using to calculate how accurate our results are. It follows the format as

sentence \t score \n

The score is either 1 (for positive) or 0 (for negative). Table 1 represents statistical information about the data sets. The dataset 1 come from three different websites imdb.co, amazon.com, yelp.com. For each website, there exist 500 positive and 500 negative sentences. Those were selected randomly from larger datasets of reviews. The attempt was to select sentences that have a positive or negative annotation. The goal was for no neutral sentences to be selected. Dataset 2 represents reviews for four different products. Dataset 1 has an equal number of positive and negative sentences whereas dataset 2 has an unequal number of positive and negative sentences.

On the basis of the outputs obtained for different datasets, we prepared the confusion matrix. It contains TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). We used the confusion matrix to measure the system performance on the basis of the equations (1) – (4).

The accuracy factor is insufficient to analyze the performance of sentiment mining work. We need to consider other parameters like precision, recall and f-score for comparing performances of the systems.

Table 2 and Table 3 represent the statistical analysis of the obtained results on different datasets. The obtained results from figure 2 and figure 3 clearly show that a mere keyword spotting for a bag of n-grams and polarity calculations is insufficient to get the exact context of consumers' opinions. Considering the entire cluster's polarity, valence shifter and adversative conjunctions handling improve sentiment polarity results.

Table I: Statistical Analysis of Dataset

Dataset Name		No. of Negative Sentences	No. of Positive Sentences
Dataset 1	imdb_labelled	505	495
	Yelp	500	500
	amazon sale	500	500
	deceptive	800	800
Dataset 2	Digital Camera 1	50	185
	Digital Camera 2	30	127
	Mobile Phone	77	188
	DVD Player	194	149

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})} \quad (1)$$

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \quad (2)$$

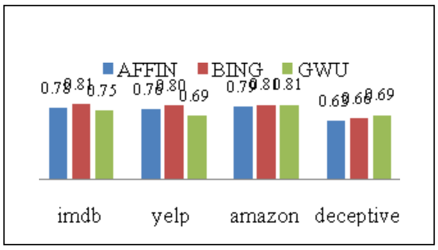
$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \quad (3)$$

$$F\_Score = \frac{2 * (\mathrm{Precision} * \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})} \quad (4)$$

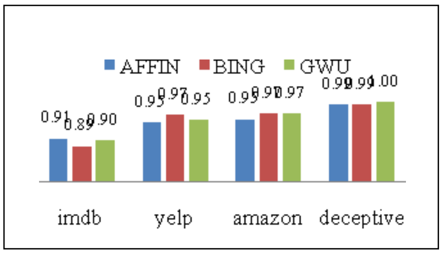
Table II: Statistical Analysis of Result Obtained for Dataset 1

Dataset Name	Method	Precision	Recall	F_Score

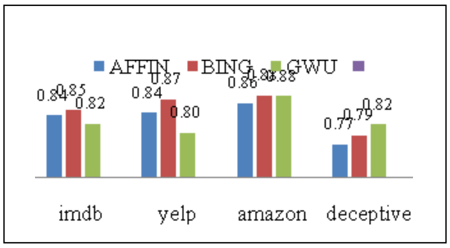
imdb	AFFIN	0.78	0.91	0.84
	BING	0.81	0.89	0.85
	GWU	0.75	0.90	0.82
yelp	AFFIN	0.76	0.95	0.84
	BING	0.80	0.97	0.87
	GWU	0.69	0.95	0.80
amazon sale	AFFIN	0.79	0.95	0.86
	BING	0.81	0.97	0.88
	GWU	0.81	0.97	0.88
deceptive opinion	AFFIN	0.63	0.99	0.77
	BING	0.66	0.99	0.79
	GWU	0.69	1.00	0.82



(a) Precision



(b) Recall

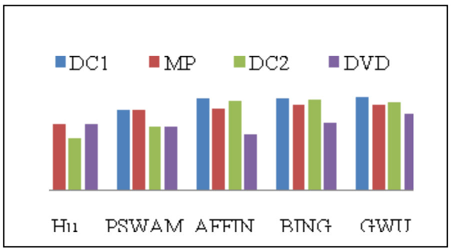


(c) F Score

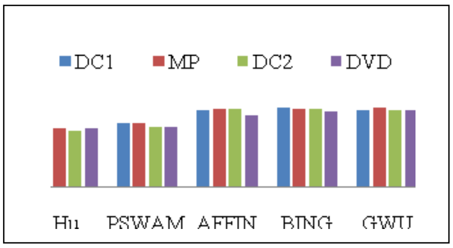
Fig. 2. Statistical Analysis of Result Obtained for Dataset 1

Table III: Statistical Analysis of Result Obtained for Dataset 2

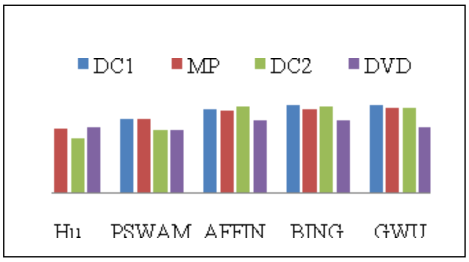
Dataset Name	Method	Precision	Recall	F1_Score
Digital Camera 1	Hu	0.72	0.74	0.73
	PSWAM	0.84	0.77	0.8
	AFFIN	0.95	0.93	0.91
	BING	0.96	0.96	0.95
	GWU	0.97	0.93	0.96
Mobile Phone	Hu	0.69	0.71	0.7
	PSWAM	0.83	0.77	0.8
	AFFIN	0.85	0.95	0.90
	BING	0.88	0.94	0.91
	GWU	0.89	0.96	0.92
Digital Camera 1	Hu	0.54	0.68	0.6
	DP	0.66	0.73	0.69
	PSWAM	0.92	0.94	0.93
	BING	0.94	0.94	0.94
	GWU	0.92	0.93	0.93
DVD Player	Hu	0.69	0.72	0.71
	PSWAM	0.66	0.73	0.69
	AFFIN	0.59	0.87	0.79
	BING	0.70	0.92	0.79
	GWU	0.79	0.93	0.72



(a) Precision



(b) Recall



(c) F Score

Fig. 3: Statistical Analysis of Result Obtained for Dataset 2

V. CONCLUSION AND FUTURE SCOPE

The paper presents a gradual weight updating approach for sentiment mining. The proposed model focuses on the opinioned word along with its context. The context is marked in the forward and backward direction for the accurate sentiment mining of the sentence. We considered the cumulative impact of the valence shifter and adversative conjunction on the sentiment polarity. The obtained results represents that, the system works fine for the regular type of opinion.

At the same time, the performance is a little bit decayed for the deceptive type of opinion. In future, we will try to improve system performance by differently handling these deceptive types of opinions.

REFERENCES

1. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD*, Washington, DC, USA, 2004.
2. S. Nandedkar and J. Patil, "Co-Extracting Feature and Opinion Pairs from Customer Reviews using Hybrid Approach" in *Proc.* IEEE Int. Conf. for Convergence in Technology, Apr. 2018.
3. K. Liu, L. Xu, and J. Zhao, "Co-extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model", *IEEE Trans. Knowledge and Data Engineering*, vol. 6, no. 1, January 2013.
4. Y. Zhang and W. Zhu, "Extracting Implicit Features in Online Customer Reviews for Opinion Mining", in Proc. 22nd Int. Conf. on World Wide Web Companion, pp. 103–104, 2013.
5. M. Kinge, S Nandedkar, and G. Narkhede, "A Survey on Traffic Sentiment Analysis", in IJARETS, Vol. 2, Issue 6, June 2015.
6. Y. Wu and F. Ren, "Learning Sentimental Influence in Twitter", in Proc. IEEE Int. Conf. on Future Computer Sciences and Application, pp. 119 – 122, 2011.
7. S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, "Interpreting the Public Sentiment Variations on Twitter", *IEEE Trans. on Knowledge and Data Engineering*, vol. 6, no. 1, pp. 1158 – 1170, September 2012.
8. R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, Issue 8 , pp. 2120 – 2133, Aug. 2015.
9. L. Yu, J. Wang , K. Lai, and X. Zhang, "Refining Word Embeddings using Intensity Scores for Sentiment Analysis", IEEE/ACM Trans. Audio Speech Lang. Process. vol. 26 no. 3 pp. 671-681 Mar. 2018.
10. G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM", IEEE Access vol. 7, pp. 51522-51532, 2019.
11. S. Nandedkar and J. Patil, "Feature Opinion Co-Extraction based upon Genuine Score Analysis" in Springer edition Advances in Intelligent System and Processing, Vol.1025, pp. 771 – 781, Nov. 2019.

AUTHORS PROFILE

Sugandha Nandedkar received the Master of Engineering degree in computer science from Dr. B. A. M. University, Aurangabad (MS), India, in 2010, where she is currently working toward her Ph.D. degree in the department of Computer Science. Her research interests include opinion mining, sentiment analysis and topic modeling. In particular her research focuses on sentiment analysis associated with text mining in the applications of customer reviews extraction, filtering and analyzing. Her work focuses on co-extraction of feature –opinion word pair and assigning aspect label for aspect based sentiment summarization. Her papers have been published in various conferences and journals such as IEEE, Springer, IJSEM, IJIRSE, and IJAR.

Jayantrao Patil received the Bachelor of Engineering in Electronics Engineering from Marathwada University, Aurabgabad (MS), India, M. Tech. Computer Science and Data Processing from IIT Kharagpur, WB, India and the Ph.D. degree in Computer Engineering from N. M. University, Jalgaon (MS), India. Currently he is the Principal of R. C. Patel Institute of Technology, Shirpur (MS), India and a university recognized Ph.D. guide. His research interest includes web catching, web mining, text watermarking, web personalization, and web security. He has published more than 50 research papers in renowned journals like IEEE, Springer, Elsewire etc. He got many awards and recognition for his contribution in the field of engineering education. He has been a session chair and key note speaker for various conferences. He has life membership of different international organizations like IEEE, ASEE, ISTE, CSI etc. He received a research grant of Rs. 500000+. He has written a book on "An Integrated Prefetching and Caching Approach – for Web Browsers / Web Proxies". He got a patent for his research on "Method of ROI Extraction for Palm print".

Sunil Kawale is professor and head in P.G. department of Statistics, Dr. B. A. M. U, Aurangabad. He has published around 25 national/international standard research articles in *various* national and international journals of repute. 3 scholars have already been awarded Ph.D. under his guidance and presently 2 scholars are working under him for Ph.D. degree. His research area includes Operations Research, Stochastic Modeling, Sentiment Analysis and Data mining. He also has delivered key notes and invited talks in various international and national conferences. He is working as member of editorial board and reviewer for four journals.