

Accurate Information Extraction from Customer Comments Posted Online



Mishu Jain, Rajni Jindal

Abstract—Customer comments form an integral part for identification of failures and success of a product. Buying patterns of a customer greatly depends on the pattern of comments posted online. Online review/comments can be broadly classified into positive, negative and neutral. Many tools available in market can be used for their classification. However, there are various flaws in classifying methods that can tweak the result of these comments such as “Unidentified/Hidden information in neutral comments”, “Wrong keyword extraction while splitting words”, “fake comments based on frequency of duplicate comment or reviewer”. This paper addresses this problem based on online product comments posted on Amazon website and proposes an ideal flow chart and algorithm to address these problems.

Keywords: polarity, customer comments, drill down, opinion mining

I. INTRODUCTION

Companies across globe use information from online comments to understand their customer's VOC (Voice of concern). Companies extract opinions about their products by extracting relevant information, emotions, and issues from these online comments. The success or failure of a product greatly depends on customer comments because it has direct influence on future prospective buyers who can openly read these comments on Social Media. Hence, this method of mining opinion which includes extracting information from textual comments drill down, provides opportunities for future business growth [1]. The problem is to identify accurate information using some tool. Various tools have been proposed in the past such as RapidMiner [2] but these tools have some underlying problems as far as polarity distribution is concerned. Each algorithm devised so far will assign “Positive”, “Negative” or “Neutral” category to the comments. However, it is known that a positive review might have some negative information or vice versa.

Researches in sentiment analysis have been done in different languages, for example, Arabic [4], Thai [7] etc. These researches mainly focused on calculating score from emotion word and marking them as positive, neutral, or

negative into sentiment polarity [8][7][10][9]. However, the question about accuracy arises from the fact that a neutral review may have hidden emotion which might lie under positive or negative category. Another question arises from the fact that English language has contextual meanings in a sentence. Any review that might be categorized as positive could have an overall negative impact due to context-based sentence usage.

This paper will have sections as following: section II will show the related work, Section III will be used to describe various underlying problems and proposed method to solve one of these problems. Section IV will be used to show experimental result comparison (before vs after). Conclusion will be summarized in section V.

II. RELATED WORKS

Opinion mining is amongst one of many popular research areas. One major obstruction in path of opinion mining is that the data is not well structured and has a lot of outliers and hidden information. Opinion mining involves Natural Language processing (NLP) for sentiment segregation and analysis of text as corpus, specific dictionary for language and lexicon [1],[7],[4],[11],[8].

Use of NLP to assign sentiment score on twitter comments [5] and scoring based on sentiment in lexicon [6] describes how scoring of words can be done to attain a positive, negative or neutral score for the entire statement. Both the papers [5],[6] aims to find our polarity based on emotional content by extracting opinion words, assigning score and summing individual scores to get the whole sentiment.

III. PROPOSED METHODOLOGY

We gathered representative data of more than 5000 comments from Amazon and scored them by our tool. The tool focusses on gathering data, removing stop words, reduction of data and preprocessing it. We identify key symptoms which are used by majority of customers and calculate a computation score.

When the score was reviewed, we observed that conventional approaches mentioned in previous researches [5][6] did not focus on 3 major problems i.e. Wrong keyword extraction and categorization, Neutral review drill down, and removal of duplicate comments which biased overall score.

The solution to these problems is addressed in our algorithm.

Manuscript published on November 30, 2019.

* Correspondence Author

Mishu Jain*, Chief Engineer/Senior Professional working in Samsung Electronics.

Dr. Rajni Jindal, Head of Computer engineering Department in Delhi Technological University.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

A. Problem 1: Wrong Keyword extraction and categorization

Many comments posted online can fall into false category due to wrong keyword selection or splitting. A word that can be split into 2 meaning full words can be misunderstood by the tool. Consider mentioned review posted online by a customer in Fig. 1.

Symptoms	Reviews	Polarity Old	Polarity New
Network,Battery,Camera,Heat,Display,Sensor	excellent, weight, display quality super excellent battery life and quality: i did a video conference call using mobile hot spot from this phone and the battery was barely discharged, plus the phone did not heat up at all the speed is excellent too i have not used the fingerprint reader, and only used the basic features of the camera so cannot make a proper assessment the price is a little high	Neutral	Positive

Fig. 1. Wrong symptom and polarity due to split word

This review talks about “Hotspot” which is a feature in mobile technology. However, tools that are used to split the keywords will break it into Hot + Spot. This will infer that the mobile device is getting heated up and eventually the tool will place it in wrong category (neutral polarity in this case).

B. Problem 2: Neutral scored comments are not really neutral

Due to lexicon base scoring algorithm [5][6], it is observed that many comments might fall into neutral category as their overall score is ‘0’. However, such comments have hidden information which is tilted toward either positive or negative category.

Consider below example

Symptoms	Reviews	Polarity Old	Polarity New
Sound,Battery,Camera,Sensor	battery life is good in camera wide angle is most effective sound quality is good but there are no vibration on fingerprint camera in day light is amazing but low light in camera is notgood	Neutral	Positive

Fig. 2.Extracting information from Neutral comments

In the above review, user wants to express Positive emotions however Lexicon based scoring algorithm [5] calculates +1 score for positive word and -1 score for negative word. In the above case, the overall score is calculated as 0 despite the fact that the review is tilted towards positive emotion.

C. Problem 3: Fake or duplicate single word comments

The comments posted online has single word comments that do not give relevant information but these comments tilt the plotted polarity graph. Comments mentioned in Fig. 3 shows how irrelevant comments can bias the overall polarity of a product.

Symptoms	Reviews	Polarity Old	Polarity New
	nothing	Neutral	NA
	nothing	Neutral	NA
	plz call me	Neutral	NA

Fig. 3.Duplicate and meaningless comments affecting overall polarity

Fig 6. On the other hand, shows that if these comments are removed or neglected by applying a duplicity check, the graph changes.

D. Solution: Proposed Algorithm and Flow of events

Fig.1 Is basic flowchart that describes key events and method that governs the proposed polarity scoring system.

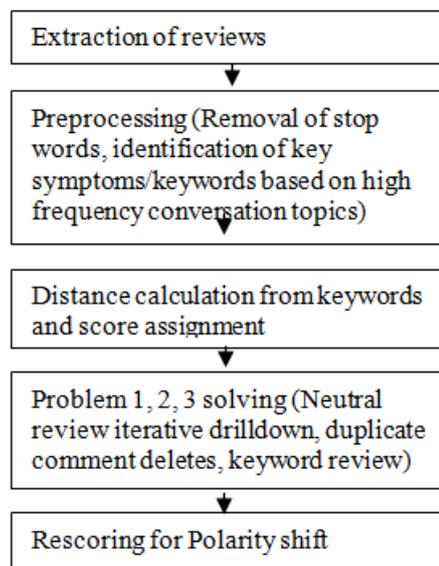


Fig. 4.Scoring Work Flow

TABLE I. ALGORITHM

Algorithm	Polarity Scoring of Online comments
Input Output	<i>Extracted online comments</i> <i>Emotional Polarity</i>
Step 1	Extract keyword K_i based on high frequency conversations
Step 2	Calculate distance of adverb/adjectives from keyword and assign score S_i based on closest distance D_i and emotional intensity E_i
Step 3	Sum S_i to Identify Polarity P For $S_i < 0$: Assign Negative as N For $S_i > 0$: Assign Positive as P For $S_i = 0$: Assign Neutral as N_u Call function Neutral Drill Down ()
Step 4	Neutral Drill Down () { Calculate frequency of Negative N_i and positive words P_i Link Emphasis words E_i to N_i & P_i and rescore S' If $N_i > P_i$ && $S' < 0$, Assign New Polarity $P' = N$ else if $N_i < P_i$ && $S' > 0$, Assign New $P' = P$ else Assign $P' = N_u$ } Print New Polarity P'

IV. EXPERIMENTAL RESULTS

This experiment was used to analyze a product quality from Amazon website. We gathered five thousand comments from a product and used conventional approach [5][6] to score the comments in positive, negative & neutral category. Later this approach was further modified by the proposed algorithm, to extract information from neutral comments.

Unnecessary word like “a, an, the” present in the review are removed as they do not hold any emotion. Identified words are stored and processed using dictionary.

We used approach used in [3], to design a dictionary consisting of negative and positive meaning words. Positive words are given positive score and negative words are given negative score. Total cumulative score determines the polarity of the review.

We also combined the approach mentioned in [5] to give higher score to emphasis words. For example, "Device is very hot", then multiple emphasis on "Very" indicates high emotional intensity E_i and hence more score is assigned.

However, the problems identified in previous sections were visible in Fig. 5 since neutral review count was too high. Hence, proposed algorithm was used to calculate new polarity as represented in Fig. 6.

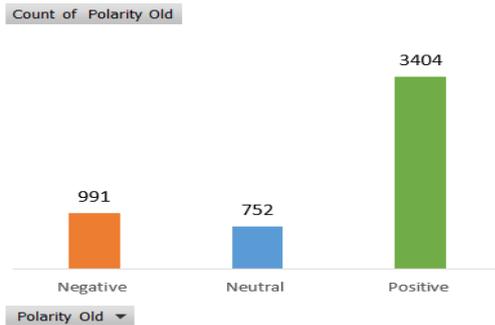


Fig. 5. Neutral comments high count from conventional approach

Table. 1 shows proposed algorithm to further drill down neutral comments. The idea is that neutral comments are not really neutral in nature but they have underlying information. This information can be extracted with this algorithm to place these comments in positive or negative category.

Whenever a review is categorized as neutral, drill down function is called. This function further checks frequency of positive words, negative words and emphasis words and compare the frequency and scores again. This helps to move earlier marked neutral comments to correct category.

The experimental result on 5000 Amazon comments is shown in fig 6. On comparing with Fig. 5, it is observed that neutral review count has dropped. This clearly depicts that initially these neutral comments were wrongly placed and have improved by 70% (on comparing neutral review count in Fig.5 and Fig.6).

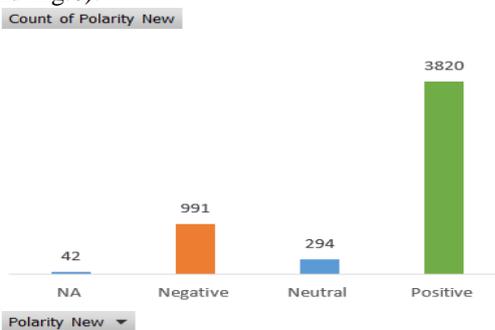


Fig. 6. Reduced neutral review count

V. CONCLUSION

In this paper, neutral review further drill down approach is explained. The idea is to clarify that neutral comments have hidden information which gets missed due to mathematical

score-based calculation scheme. These comments can be re-categorized into positive or negative polarity in order to gain clearer information about the product being reviewed by consumers.

VI. FUTURE WORK

There is tremendous scope of research in this field. One such area is to drill down comments posted by fake consumers or competitors. These competitors might try to destroy a company's brand value by posting repeated negative comments.

Further there can be development related to categorizing comments written in slang language that do not really exist in dictionaries

REFERENCES

1. Wararat Songpan, "The analysis and prediction of Customer review rating using opinion mining", Khon Kaen University Thailand, IEEE SERA, June 2017.
2. Pragya Tripathi, Santosh Kr Vishwakarna, and Ajay Lala, Sentiment Analysis of English Tweets Using Rapid Miner, International conference on computational intelligence and communication networks, 2015.
3. <https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/>
4. K. Shaalan and S. Atia, Increasing the accuracy of opinion mining in Arabic, Proceeding of the 1st International conference on Arabic computing linguistics, 2015, pp.106-113.
5. Srinidhi Bhat, Saksham Garg and Poornalatha G, "Assigning sentiment score for Twitter tweets" 2018, IEEE.
6. Yi Ding, Bing Li, Yuqi Zhao & Can Cheng, Scoring Tourist Attractions Based on Sentiment Lexicon, 2017, IEEE
7. T. Chumwatana, Using sentiment analysis technique for analyzing Thai customer satisfaction from social media, 5th International conference on computing and Informatics, 2015, pp.659-664.
8. S.Ahmed and A.Danti, A novel approach for sentiment analysis and opinion mining based on sentiwordnet using web data, International conference on trends in Automation, Communications and computing technology, 2015, pp. 1-5
9. R.K. Bakshi, R. Kaur, N. Kaur and G.Kaur, Opinion mining and sentiment analysis, 3rd International conference on computing for sustainable Globl development, 2016
10. P. Ghaffari, J.G. Breslin, P.Barnaghim, Opinion mining and sentiment polarity on twitter and correlation between events and sentiment, 2nd international conference on Big Data Computing service and application, 2016, pp.52-57.
11. L. Lin, J. Li, R. Zhang, W. Yu and C. Sun, Opinion mining and sentiment analysis in Social Networks: A retweeting strucutue-aware approach. Proceeding of the 7th International conference on utility and cloud computing, 2014, pp.890-895.

AUTHORS PROFILE



Mishu Jain, Chief Engineer/Senior Professional working in Samsung Electronics. Work experience of more than 8 years. Masters of Technology from Delhi Technological University, India and Bachelors of Technology from Uttar Pradesh Technical University. Working in project management, data representation, effort estimation and marketing strategies.



Dr. Rajni Jindal, Head of Computer engineering Department in Delhi Technological University. Work experience of more than 29 years in research and academics. Authored around 80 research papers and is a senior member of IEEE. M.E from Delhi college of engineering and Ph.D. from Delhi

university.

