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

Web search model for Automatic Annotation of Tweets in Opinion Mining

H. Mohamed Zakir, S. Vinila Jinny

Abstract: In sentiment analysis, annotation is the crucial step of data processing to label the review or sentences as positive, negative, or neutral. An annotation process is usually performed by three key approaches: (i) manual, (ii) crowdsourcing, and (iii) automated annotation. Manual annotation is preferred in most of the literature's and crowdsourcing tools are used in some of the works. This indicates that there is a scarce of automatic annotation and its service is highly essential to support more systematic research in sentiment analysis. Manual procedures mostly depends on external annotators, resulting in costly and time-consuming processes. Thus, we propose a method for automatic annotation using web search model to dynamically label reviews (positive, negative, or neutral) that are not available in the dictionary. Some research works consider the product opinions from e-commerce sites where most of the texts may not be available in the dictionary, in such cases, the web search model can be used instead of manual annotation. A large-scale opinosis dataset is used to evaluate the accuracy of the algorithms and feasibility of the model. The experimental results indicate that this model outperforms conventional methodologies and therefore we firmly believe it will be useful for current researchers in the field of opinion mining.

Keywords : annotation, opinion mining, websearch

I. INTRODUCTION

Social media is a powerful platform for people to share their ideas, thoughts, and sentiments about a product, individuals, etc. [1]. One of the known techniques for mining and analyzing the content of social media is sentiment analysis (SA). The aim of Sentiment Analysis is to gather consumer feelings about many items, such as identifying brand desires and guiding business strategies [2]. Inorder to perform SA, an essential and highly expensive step is required, which is corpus annotation. Annotation can be achieved at various levels, such as sentence level, word level or both. The importance of the annotation process results from making a machine-readable version of the meta-data to train a machine-learning classifier by annotating the corpus [3]. There are different levels of corpus annotation available, such as syntactic annotation, which translates each sentence in the corpus. To label each word a suitable POS tag in the corpus, POS tagging is applied and for sentiment mining, semantic annotation is used and during the annotation method different labels are involved such as positive, negative, and neutral at the semantic level. Within the same tweet, positive and negative sentiments will occur, which is considered a challenge to the annotation process [4].

The three major techniques for corpus annotation are (i) individual manual annotation (ii) crowdsourcing approach and (iii) automated method [1]. In the first method, the annotation is performed by a small group of people. A large number of people are involved in the second method, to handle the annotation process with the help of annotation tools. The annotation is determined from the database itself as reviews star rating in the automated approach. The main drawback of the manual method is, it generally depends on the number of native speakers. The disadvantage of the crowdsourcing method is, it involves freelancers available on the web. Both approaches, rely on external annotators which leads to an expensive and time-consuming process. Such factors clearly indicate that there is a need to develop an automatic annotation method either at sentence or word level to support more systematic research in the field of sentiment analysis area. Thus, we propose an automatic annotation method that will extract the words from the tweet and compare them with the data available in the dataset. If the words not found in the dataset, through the proposed algorithm it automatically label (positive, negative, or neutral) the extracted word and insert the word into the dataset. This paper is structured according to the following. Section 2 describes the related work. The proposed algorithm is explained in section 3 and section 4 discusses the experimental results. Discussions are made in section 5. Some applications are briefed in section 6. The conclusions are made in section 7.

II. RELATED WORK

A. Practical Guide to Sentiment Annotation: Challenges and Solutions

The author Saif M. Mohamed [5] proposes two manual sentiment annotation schemes called (i) simple sentiment annotation questionnaire (ii) semantic role-based questionnaire for annotating different types of words or sentences as 'positive', 'negative', or 'neutral'. In this work, the author lists the most notable challenging types of sentences to be used for annotation using the above schemes. The first scheme aims to determine the dominant sentiment inferable from the sentence.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
H.Mohamed Zakir, Department of Computer Science, Noorul Islam Centre for Higher Education, Kumaracoil, India. Email: hmohamedzakir@gmail.com
Dr.S.Vinila Jinny, Department of Computer Science & Engineering, Noorul Islam Centre for Higher Education, Kumaracoil, India. Email: vinijini@gmail.com

The second scheme using a detailed questionnaire provides a rich cross-section of information which can be used by many downstream applications. The aspects of the proposed questionnaire may not be appropriate for all applications. These schemes are not experimentally evaluated. Hence, no accuracy has been charted in this model.

B. Annotation Technique for Health-Related Tweets Sentiment Analysis

The author Asma Baccouche et al. [6] suggests an automatic annotation scheme based on emojis and semantic labeling for health-related tweets. The author assumes that happy emoji is found in the positive tweets and the sad emojis exists in the negative tweets. This assumption may not be true for all nature of tweets, e.g. sarcastic nature of tweets may contain positive emojis in negative tweets and vice versa. Due to the quality of the annotated data sets, some misclassification is noted in the accuracy of the final prediction. This scheme works for a particular domain.

C. Minimum Annotation Identification of facial affects for video advertisement

The author Gaurav Goyal et al. [7] suggests a method to automate an advertising product review. In this model, neurometric information of customers against an advertisement is recorded and the classification algorithm is applied. In this model, no annotation process is involved. This model detects emotions like happy, sad, anger, neutral, and surprise.

D. An Annotated Corpus for Turkish Sentiment Analysis at Sentence Level

The author Sevinc Ilhan Omurca et al. [8] builds a turkish sentiment corpus which are semi-automatically annotated inorder to enable the information management for aspect-based sentiment analysis. The aspects & sentiments are taken from the phrases and these phrases are tested for annotation by two separate human experts. Turkish hotel reviews dataset is used.

III. EXPERIMENT EVALUATION

Our proposed approach for automatic annotation is explained in Fig 1. The proposed model consists of three stages which include (i) extracting phrases from the review which contains adjectives or adverbs (ii) calculate the SO of each extracted phrase and (iii) label the review based on the calculated SO of the phrase. In this model, opinion lexicon by Bing Liu [9] has been used to automatically label the tweets. İt consists of 2006 positive words and 4794 negative words. However, there may be cases that some words may not available in the given lexicon therefore, the web search model as explained in section D has been used to label those words automatically.

We decreasingly ranked the POS tags as adverbs, adjecti-ves, verbs, and nouns for calculating the scores. This is because adverbs and adjectives have a stronger effect in a sentence than verbs and nouns [6]. The negations exist in the text like 'not', 'no', 'don't', 'can't', 'won't', 'didn't', and 'never', are also considered to correct the polarity of unigrams and bigrams semantically.

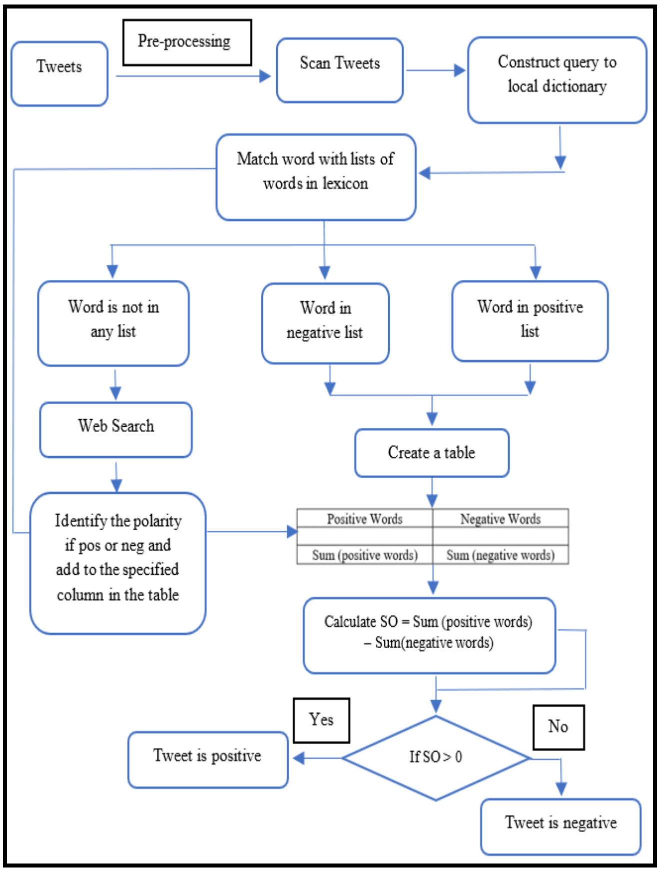

Fig. 1. Proposed Approach for automatic annotation

A. Datasets Description

Opinosis Opinion dataset sentences are extracted from reviews on a given topic. Examples of such topics are "performance of Toyota Camry" and "sound quality of iPod nano", etc. There are 51 such topics in this dataset and the opinions are obtained from Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics). There are approximately 100 sentences per topic. This dataset is used by many researchers for their experiments [10].

B. Pre-processing

The methods of pre-processing evaluated in this paper are as follows: In our previous comparative study paper, we analyzed various data cleaning approaches by experimentally comparing different algorithms. The results revealed that the DySNI and Brushing algorithm can be combined to achieve a better accuracy in reduced execution time, along with applying negation replacement and acronym expansion.

Negation replacement: Tweets are made up of different conceptions of negation. In general, negation plays a major role in determining the twee t's sentiment. Here, the process of negation is transforming 'not', 'no', 'don't', 'can't', 'won't', 'didn't', and 'never', respectively.

Acronyms Expansion: Acronyms and slang are popular in tweets without form. It is highly essential to expand it to its original words. Thus, in this work, the acronyms and slang are expanded using the acronym dictionary Internet Slang Dict [11]. It is made up of slang and acronyms created by users to save keystrokes.

Retrieval Number: B4662129219/2019©BEIESP
DOI: 10.35940/ijeat.B4662.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Terms have originated from different sources such as Bulletin Boards, Yahoo, Chat Rooms, Email, and Cell Phone Text Messaging. Each acronym is an explanation, for example, "2maro" is "tomorrow", "fyi is For your information", "asap is As soon as possible".

C. Steps for automatic annotation if phrase exists in lexicon

- The pre-processing output is used as an input to this phase, where each tweet is treated as bag of words.
- Scan each tweet in the document to extract each word and issue queries to the opinion lexicon dictionary to determine the polarity of each word in the tweet.
- If the word is found in the positive list in the dictionary allocate its occurrence in the positive column as in Table I, if the word is found in the negative list allocate its occurrence in the negative column.
- If the positive word comes after any negation like ('not', 'no', 'don't', 'can't', 'won't', 'didn't', and 'never') then, it will be considered as negative and allocated in the negative column. E.g. " the service is not exciting" should be considered as negative because of not.

Table I: Allocating entries in the table

Tweet	Positive Words	Negative Words
	Word1	Word2
	Word3	Wordn
Total	Sum(No. of positive words)	Sum(No. Of negative words)

- Find the sum of both positive words & negative words and calculate the difference between them to find the semantic orientation (SO) of the tweet$_i$ as below, SO(tweet$_i$) = ∑(positive words in the tweet) - ∑(negative words in the tweet) (1)
- If SO is greater than 0, (SO > 0), then the polarity of the tweet is positive, otherwise, if the result of SO is less than 0, (SO < 0) then the polarity of the tweet is negative. For the sample tweet, "Bus journey is always hectic and tedious", the SO is calculated in Table II.

Table II: Semantic Orientation Calculation

Tweet	Positive Words	Negative Words
	NA	hectic
	NA	tedious
Total	0	2

So, as per formula (1), *SO(tweet$_i$)* = (0 – 2). Hence, the polarity of the tweet is negative.

- If the extracted word is not available in the opinion lexicon dictionary, our web search model is used to find the polarity of the words.

D. Web search model for automatic annotation if phrase not exists in lexicon

Web search model is used to automatically label the word to positive or negative if the word is not found in the existing dictionary. Yahoo search is used in performing the search query and the words "excellent" and "poor" are used as reference words. The comparison terms "excellent" and "poor" have been chosen because one star is commonly defined as "poor" and five stars as "excellent" in the five-star review rating system. Semantic orientation is positive when a phrase is more closely linked to "excellent" and negative when a phrase is more closely linked to "poor" [12].

- Scan the tweet and identify the adjective or adverb word by POS tagger (Brill, 1994[1]). Two consecutive phrases are retrieved from the tweet only if their tags matches any of the patterns specified in Table III.
- Using yahoo search, find the number of documents that contain the phrase and "excellent" word -> (phrase and excellent). E.g. if the phrase is "intolerant", Hitspos = hits (phrase and "excellent")
- Using yahoo search, find the number of documents that contain the phrase and "poor" word -> (phrase and poor). E.g. if the phrase is "intolerant", Hitsneg = hits (phrase and "poor")
- Calculate ratio by:

Ratio(phrase) = Hitspos / Hitsneg (2)

- Find semantic orientation

SO(ratio) = log2(SO(phrase)) (3)

- If SO(ratio) > 0, the phrase is positive, otherwise the phrase is negative.
- After calculating the SO(ratio), as mentioned in section C, add the phrase to the suitable column in the table (positive or negative) and complete the process to determine the polarity of the tweet.
- Insert the phrase to the existing dictionary in a suitable column.

Table III: Tags pattern to extract two word phrase from tweets

No	First Word	Second Word
1	RB, RBR, or RBS	VB, VBD, VBN, or VBG
2	NN or NNS	JJ
3	JJ	NN or NNS
4	JJ	JJ
5	RB, RBR, or RBS	JJ

IV. EXPERIMENTAL RESULTS

Table IV describes 416 reviews from opinosis opinion dataset that have been used in the experiment. 176 (42.30%) of the reviews are labeled as negative and the remaining 240 (57.70%) reviews are labeled as positive. The third column in the table shows the average number of characters extracted for each tweet. Some examples of positive reviews are shown in Table V and negative reviews are shown in Table VI. The algorithm has been evaluated using visual studio 2017 and the whole experiment is conducted in Windows 10 Home 64-bit Operating System with Intel® Core™ i7-8550U CPU @ 1.80GHz 1.99 GHz processor and 8.00 GB of RAM.

Table IV: Summary of reviews

No	Topic Name	No. Reviews	Average No. of characters per Review
1	Battery Life (ipod)	69	92.28
2	Comfort Honda Accord	166	90.34

[1] http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z

No	Topic Name	No. Reviews	Average No. of characters per Review
3	Comfort Toyota Camry	119	88.23
4	Windows 7 Features	62	114.40

Table V: Sample review processing labeled as positive

No	Extracted Sentence	POS Tags	Semantic Orientation
1	Local branch	JJ NN	0.421
2	Online experience	JJ NN	2.252
3	Low fees	JJ NNS	0.332
4	True service	JJ NN	-0.732

Table VI: Sample review processing labeled as negative

No	Extracted Sentence	POS Tags	Semantic Orientation
1	Clever tricks	JJ NNS	-0.040
2	Unethical practices	JJ NNS	-8.484
3	Extra day	JJ NN	-0.286
4	Lesser evil	RBR JJ	-2.288

Table VII: Automatic annotation using web search model

	Accurate	Nice	Ugly
Hits (Excellent & Phrase)	37,700,000	19,300,000	20,000,000
Hits (Phrase, Poor)	25,700,000	17,500,000	24,600,000
Ratio	1.46692607	1.10285714	0.81300813
log(ratio)	**0.552796164**(Positive)	**0.14124593** (Positive)	**-0.29865832** (Negative)

V. DISCUSSION

Table VIII shows the experimental results and an average accuracy of 85.37% has been achieved from our model. The results of the classification algorithm has been compared with VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

Table VIII: Accuracy Analysis

No.	Topic Name	Accuracy
1	Battery Life	88.00%
2	Comfort Honda Accord	82.12%
3	Comfort Toyota Camry	84.83%
4	Windows 7 Features	86.53%

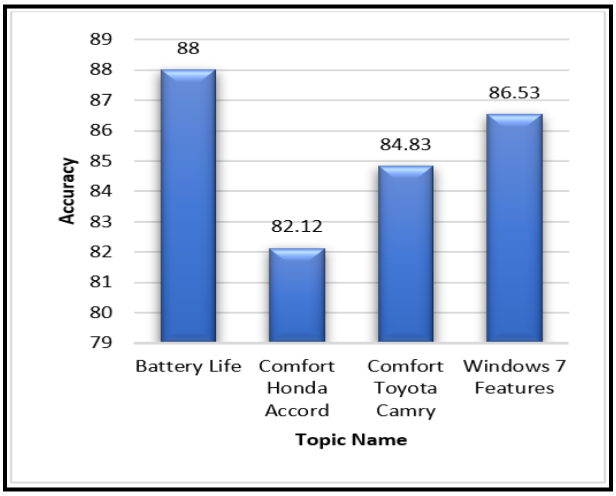

Fig. 2. Accuracy Analysis Chart

VI. APPLICATIONS

There are a number of possible automated review rating systems. Preliminary studies indicate that semantic orientation is helpful in summing up the reviews. A positive review could be summarized by choosing the sentence with the highest positive semantic orientation, and by extracting the sentence with the lowest negative semantic orientation, a negative review could be summarized. It also helps advertisers track ad campaigns, politicians are able to track public opinion, reporters are able to track public response to current events.

VII. CONCLUSIONS

This paper describes a method for automatic annotation using web search model. In general, annotation is performed by manual methods that are highly dependent on human annotators. In the related work section, most of the literature's use manual procedures or no annotation is involved. In the pre-processing phase, data cleaning is performed to remove duplicates along with applying negation replacement and acronym expansion. The experimental section describes the steps of automatic annotation in two segments (i) if the phrase from opinion exists in the dictionary and (ii) using web search model, if the phrase from opinion not exist in the dictionary. The web search model extracts the phrase that includes adjectives or adverbs, then the semantic orientation is calculated to label the reviews. For experimental evaluation, a total of 416 reviews from four topics has been collected from opinosis opinion dataset and the model achieves an average accuracy of 85.37%. The results have been compared with VADER, which is a lexicon and rule-based sentiment analysis tool. The proposed model replaces the manual annotation procedures which mostly depends on external individuals resulting in exensive and time-consuming process and the data annotated using web search method can be used to train machine learning models to identify repeated patterns. Thus, we believe, this model will help the current researchers to support more systematic research in opinion mining.

Retrieval Number: B4662129219/2019©BEIESP
DOI: 10.35940/ijeat.B4662.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. E. Refaee, "Sentiment Analysis for Micro-blogging," vol. 2, pp. 275–294, 2017.
2. A. Kaur and V. Gupta, "A survey on sentiment analysis and opinion mining techniques," J. Emerg. Technol. Web Intell., vol. 5, no. 4, pp. 367–371, 2013.
3. N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: Trends and challenges," Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2014, no. June, pp. 148–155, 2014.
4. C. Aggarwal, C.C., Zhai, "Mining Text Data," Springer Science & Business Media,New York, 2012.
5. S. Mohammad, "A Practical Guide to Sentiment Annotation: Challenges and Solutions," pp. 174–179, 2016.
6. A. Baccouche, B. Garcia-Zapirain, and A. Elmaghraby, "Annotation Technique for Health-Related Tweets Sentiment Analysis," 2018 IEEE Int. Symp. Signal Process. Inf. Technol. ISSPIT 2018, pp. 382–387, 2019.
7. G. Goyal and J. Singh, "Minimum annotation identification of facial affects for video advertisement," Proc. - 2nd Int. Conf. Intell. Circuits Syst. ICICS 2018, pp. 306–309, 2018.
8. S. I. Omurca, E. Ekinci, and H. Türkmen, "An annotated corpus for Turkish sentiment analysis at sentence level," IDAP 2017 - Int. Artif. Intell. Data Process. Symp., 2017.
9. B. Liu, "Opinion Mining, Sentiment Analysis, and Opinion Spam Detection." [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
10. K. Ganesan, C. X. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," Coling 2010 - 23rd Int. Conf. Comput. Linguist. Proc. Conf., vol. 2, no. August, pp. 340–348, 2010.
11. "Internet & Text Slang Dictionary & Translator." [Online]. Available: https://www.noslang.com/dictionary. [Accessed: 11-Nov-2019].
12. P. D. Turney, "Turney-Acl02-Final," p. 8, 2006.

AUTHORS PROFILE

The first author holds a master's degree in computer science from Annamalai University in 2013. He received his Bachelor of Engineering degree in computer science & engineering from Anna University, Chennai in the year 2006. Currently pursuing his Ph.D. degree in computer science from Noorul Islam Centre for Higher Education, Kumaracoil, India. His research area is sentiment analysis and the research's main goal is to extract the essential attribute from the set of customer opinions that can help make perfect decisions on the side of both manufacturer and e-commerce firms. He has 5 publications in international journals and 2 papers in international conferences.

The Second author is working as Associate Professor, Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil. She received her B.E. degree in Computer Science and Engineering from M.S. University, Tirunelveli in 2004 and M.E degree in Computer Science and Engineering from Anna University, Chennai in 2006. She completed her Doctorate degree in 2016 from Noorul Islam Centre for Higher Education, Kumaracoil, India. She has 12 International publications in Journals and 30 papers in International Conferences to her credit. She is a Member of IEEE. Her field of interest includes data mining, Network Security, image processing. Have teaching experience of 13 years.