Polarity Detection Of Online News Articles Based On Sentence Structure And Dynamic Dictionary

Muhammad Usama Islam
Department of Computer Science & Engineering
Asian University of Bangladesh
Dhaka -1230, Bangladesh
usamaislam@iut-dhaka.edu

Faisal Bin Ashraf
Department of Computer Science & Engineering
Uttara University
Dhaka -1230, Bangladesh
faisalashraf@iut-dhaka.edu

Ali Imam Abir
Department of Computer Science & Engineering
Uttara University
Dhaka -1230, Bangladesh
aliimamabir@iut-dhaka.edu

M. A. Mottalib
Department of Computer Science & Engineering
Brac University
Dhaka -1212, Bangladesh
mottalib@bracu.ac.bd

***Abstract*—The importance of online news article has evolved notably with the advancement of information and technology.However, some of the news are violent as well as obnoxious.So, identifying and categorizing online news article automatically is important as well as remains challenging. Using opinion mining and sentiment analysis, we propose an intuitive approach of detecting positive or negative news from an online news article. Our approach consists of a sentence identification phase, followed by a dynamic library of predefined negative and positive strings and at last marking whether the paragraph is positive, negative or neutral. Our approach detects the polarity of online news articles with around 91% accuracy rate. Sentence type identification before using dynamic dictionary of positive and negative words is the key factor which resolves the issue of finding out the part of the sentence which holds the polarity of the sentence.**

***Keywords*—Opinion mining, Sentence analysis, Phrase-level Sentiment analysis, Data mining**

I. Introduction

Opinion mining holds an important position in research with the emergence of web and media [1] [2]. Now a days information became one of the most important things in the whole world [3]. Online news articles are articles from news portal that shares various types of news events [4]. The gradual development of web brought new possibilities of online newspapers and articles. Specific news can be perceived as good, bad or neutral. However the news article itself is not categorized while being read by common pupil. We would like to categorize the news into good, bad or neutral on the basis of its nature. Here an instance of nature will simplify the definition. A news being found violent is perceived to be bad whereas a news that will bring joy is defined to be good. Though neutral news are found very less in general life we have added this in our categorization because in case of polarity being fifty percent for both positive and negative case we would have a different option to mark the news [5].

Opinion mining can be considered as classification process where three level of classification can be followed which are Document-level, sentence-level and aspect-level. Document level analysis classifies the whole document as positive or negative considering the whole document as a basic information unit. Sentence level analysis aims to classify each sentence which is similar to document level classification over a shorter size of data as a sentence actually is a small document. On the contrary, aspect level analysis aims to classify with respect to specific aspects of entities.

Many sentiment analysis approaches have been found in literature [6] [7] [8] [9].Opinion mining is done by detecting the contextual polarity of documents using semantic orientation technique or machine learning approach [10]. The most common approach is based on statistical feature of positive or negative semantic orientation of word counts. Chi-square test, DF etc. are some methods of traditional text classification [10]. Some work has used standard machine learning techniques to sentiment analysis or Opinion mining [11], such as, Pangs selection of feature sentiment words by artificial test method based on word frequency to make a comparison.Physics method is being used to judge the words emotional tendency where all the words are put as a collection of electronics and are divided like electron into cathode and anode to calculate the average of each electronic polarity [12]. In another study, a train of thoughts are put together which is to use the relationship such as synonym, antonym, fluctuation, etc., inWordNet to calculate the words polarity [11] [13].Another study follow the approach to annotate a word from the dictionary for training and classification to determine other words sentiment tendency [11] [14]. Sentence level analysis decides the primary or comprehensive semantic orientation of a sentence while the primary or comprehensive semantic orientation of the entire document is handled by the document level analysis [15] [16].

In this paper, we propose an intuitive approach to classify online news through sentence level analysis using a dynamic library which facilitates the polarity of the news by defining whether it is positive, negative or neutral.

The rest of the paper is organized as follows. Section II discusses about several aspects of sentence recognition. Section III contains sentence type determination and classification of news article. Section IV describes our proposed approach of news classification. Section V discusses about implementation,performance and result analysis of our implementation and lastly Section VI contains the conclusion and future scope of this work.

II. Several Aspects of Sentence Recognition

In this section, we discuss several aspects of our sentence recognition and analysis system which are the key factors of

9978-1-5386-1150-0/17/$31.00

news classification.

A. Subjectivity and Neutrality

A news segment that contains opinion irrespective of being positive or negative is defined under subjectivity. All other sentence is defined to be neutral. For instance, "the law and order situation is declining day by day". Here the word "decline" defines it as subjective with a negative categorization. Again, "there is a cat". This has no subjectivity. Thus this is a neutral sentence. Sentence can be subdivided into two groups - Positive & Negative

B. Positive Subjectivity and Negative Subjectivity

The positive subjectivity and negative subjectivity can be defined from a defined library of words. For instance, if we maintain a library containing the words that have separate table for predefined positive words and negative words which will be used for determining the polarity of the sentence then it will be rather easier for us. For these certain reasons subjectivity is subdivided into two discipline - Positive subjectivity & Negative subjectivity.

III. Sentence Detection and Classification of News Article

In this section, we will discuss about our approach to extract sentences and words from news article and the structure of the library we are using for classification of news articles.

A. Detection of Sentences

We extract sentences from news article finding an dot(.) as every sentence ends with a dot(.) and using the required grammatical information we determine the sentence type. The types of sentences are - Simple, Compound and Complex. There is also another instance of Complex-Compound sentence which is found very little in news articles.

1) Simple Sentence: A sentence with only one independent clause is known as simple sentence. Albeit a simple sentence doesn't have any subordinate clauses, it is often found to be big sentence. A simple sentence often contains subjects, verbs, modifiers, and it is notable that objects in simple sentences may be coordinated.

2) Compound Sentence: Compound sentences contain at least two independent clauses. These sentences can be formed in three basic ways -

- Using a coordinating conjunction such as - and, but, for, nor, or, so, yet etc. to join the main clauses
- Using a semicolon either with or without a conjunctive adverb
- Using a colon for some specific case

3) Complex sentence and Compound-Complex Sentence: A sentence that contains an independent clause and at least one dependent clause is defined to be a complex sentence. The basic difference between complex sentence and compound-complex sentence is that the sentences are formed with two or more independent clauses and at least one dependent clause in a compound-complex sentence.

B. Defining library and classifying the news article

We define a library with a list of reserved words which are subjective. Two different tables of positive and negative strings are manually entered and dynamically updated. After identifying the types of sentence and its subjectivity we at last define the sentence from series of strings from the predefined library.

C. Extracting words from the sentences

As soon as the sentence is being detected, we detect the words from each of the sentences. Word level analysis is important as we aim to detect the polarity of the sentence from the polarity of each words of the sentence.

D. Assignment of weights

For each word we assign a specific weight. The assignment of weight is needed to obtain the polarity of the sentence when the words are accumulated together.

IV. Our Proposed Approach

Following the sentence detection and classification of news article steps we have designed an approach to classify news articles. Our proposed approach is illustrated in Fig 1 and described below.

- Select an online News article.
- Extract the news articles paragraph to sentence by sentence level and eventually to word and phrase level. Here each end of sentence is determined by a . full stop sign.
- Each sentence is determined whether they are simple, compound, complex or complex-compound. For simple sentence,
    - Search for positive or negative words or phrases in the sentence
    - Determine the polarity of the sentence using multiplication scheme
- For compound sentence
    - Divide the sentence into two or more segments.
    - For each segment determine its polarity.
    - Add the polarity found in different segments within a sentence to obtain sentence polarity.
- A complex sentence consists of combination of an independent clause and a dependent clause. Here we are interested in end result of the sentence. So we find out polarity of both independent and dependent clause but for our result we give priority to independent clause more (assign more weights) as it contribute more to end output.
- For compound-complex sentence, we find polarity for each clause or segments separated by comma and perform additive rule to find polarity of the sentence.
- Combine the output found in each sentence of the paragraph and classify the news article. The output will be one of positive, negative, neutral.

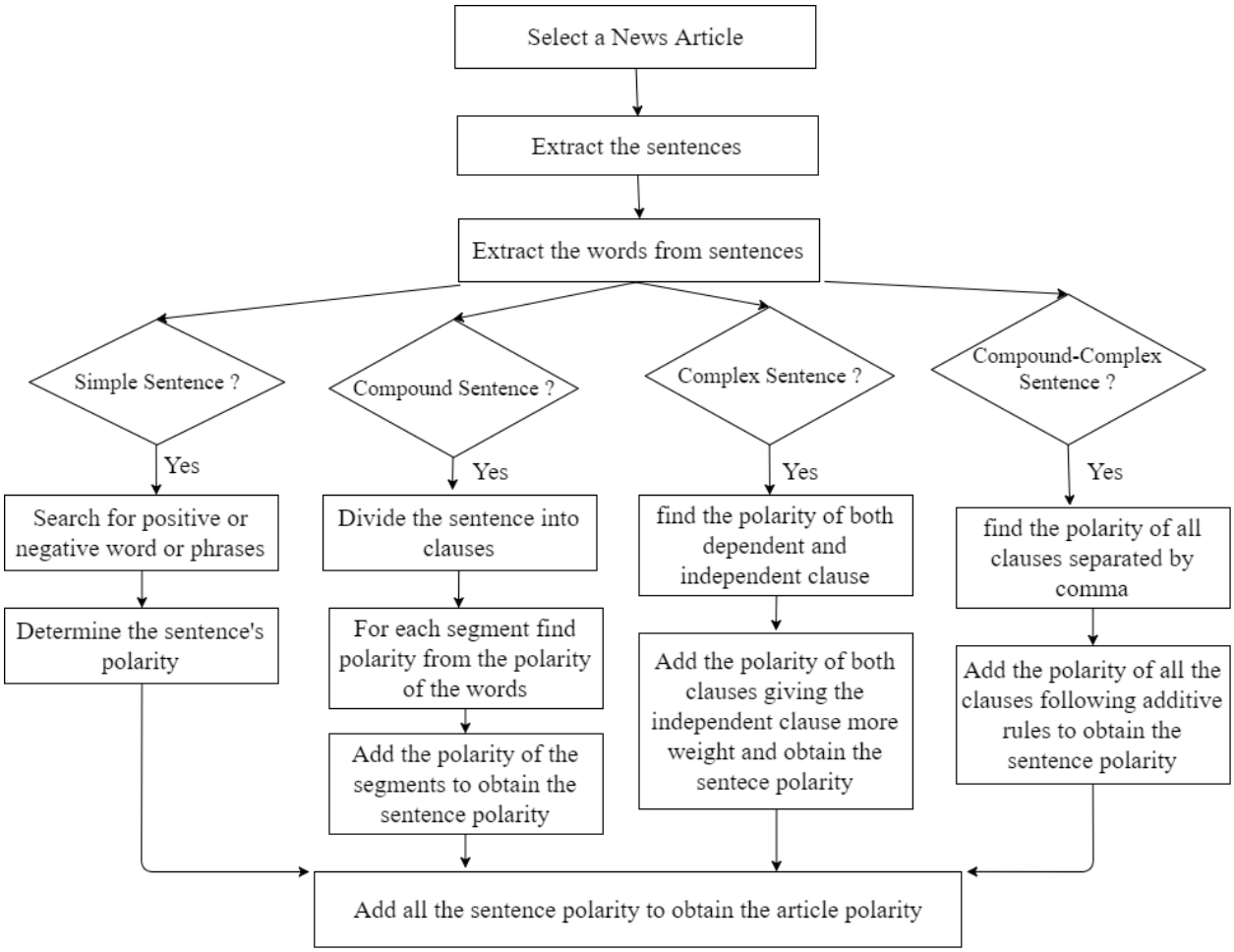

Fig. 1: Illustration of our proposed approach

V. Implementation & Performance Analysis

We have used java programming language and Netbeans IDE in windows environment to code and develop the user interface of the algorithm to obtain the result of our proposed approach to determine the polarity of news article.

A. Description of The interface

The user interface shown in Fig 2 can be divided into some sub regions. The news article of whose polarity we want to determine is copied and pasted on the text box. After successfully pasting the news article the button VIEW RESULT is pressed. It invokes an action performed of searching the whole string and returning the ultimate polarity that includes three strings namely positive, negative and neutral. There is a table that shows a break down analysis of the whole news article. The table includes the type of sentence which means the whether the sentence is complex, compound ,simple or of any mixed kind and shows a result based on weights assigned.

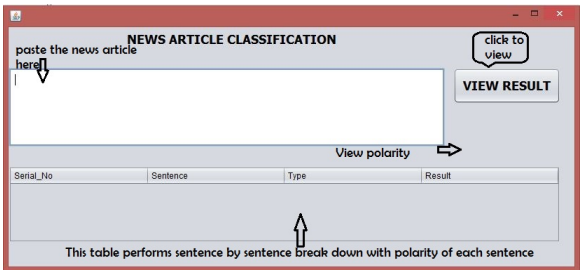

Fig. 2: User Interface for detection of polarity of News article

B. Theory of Performance Analysis

To verify our work we take the help of confusion matrix. A confusion matrix is a classification system. It contains the data about real and predicted or perceived classifications. A matrix is deployed that measures as well as evaluates the performance of the systems. The following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study - a is the number of correct predictions that an instance is negative, b is the number of incorrect predictions that an instance is positive, c is the number of incorrect of predictions that an instance is negative and d is the number of correct predictions that an instance is positive.

The accuracy is the proportion of the total number of pre-

TABLE I: Confusion Matrix

Confusion Matrix		Predicted Result	
		Negative	Positive
Real Result	Positive	a	b
	Negative	c	d

dictions that were correct. It is determined using the equation:

$$Accuracy = \frac{TN(a) + TP(d)}{AP(a+b+c+d)}$$

where, TN= True Negative, TP=True Positive, AP=All Predictions.

C. Experimentation

We have selected newspapers from three category of online newspaper according to google analytics of newspaper readership - The Independent, The Telegraph and The Daily Star. We ran our implemented system for 56 random online news article from these newspapers and recorded the polarity decided by the system. We have also recorded the true polarity of those news articles so that we can calculate the efficiency and performance of our propsed approach.

D. Confusion matrix of our experiment

We have generated a confusion matrix from the result of our experiment described before. The confusion matrix shows that out of 56 randomly chosen news article, 19 of them are true positive with 2 being negative and 26 of the news article are true negative. Besides, 6 of them are found true neutral with 2 being positive and 1 being negative. The confusion matrix of our experiment is shown in Table II

TABLE II: Confusion Matrix of our experiment

Confusion Matrix		True Result		
		Positive	Neutral	Negative
Predicted Result	Positive	19	2	0
	Neutral	0	6	0
	Negative	2	1	26

E. Accuracy of Result

The accuracy of our proposed approach can be calculated using the theory of accuracy from confusion matrix.

$$Accuracy = \frac{19+6+26}{19+2+0+0+6+0+2+1+26}$$
$$= 91.07\%$$

F. Discussion

True polarity of 56 experimented news article are depicted in the first graph of Fig 3 and Predicted polarity from our proposed approach is also depicted in the second graph of Fig 3. We have assigned 1 for "Negative", 2 for "Neutral" and 3 for "Positive" polarity and the result of each article is shown in the graph. The third graph shows the difference from the true result to our predicted result and we can see that only 5 out of 56 articles shows different polarity from true polarity which leads to 8.93% margin of errors. From further analysis, we came to the point that whenever the article has less number of sentences our approach sometimes faces difficulty to detect the true polarity. Most of the lines in news articles have neutral polarity. In case of smaller article, only a few sentences decides the polarity of the whole article. As number of shorter articles in online news are very little and for the average size of articles our approach successfully detect the polarity, our approach is good enough to detect the polarity of online news articles.

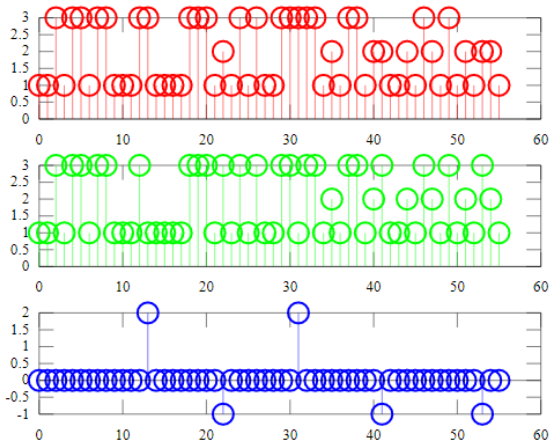

Fig. 3: True Result, Predicted Result and Distortion of prediction from True result graph

VI. Conclusion

The main objective of this paper is to determine the polarity of a news article. The results generated from the algorithm will be helpful for the both the administrator as well as the user to hide or show the output. The theory indicate that the algorithm will perform well in all the domains that are considered whereas practically it shows fair result while depicting positive and negative news . This task contains immense importance because human is accustomed to information. However not all information is for everyone. For example a news of violence will not be appropriate for the mental growth of a child. This is also helpful for large scale companies to detect the pros and cons from the review of their product through analysis of algorithm. The future works primarily includes more feasible process of declaring the library. Also converting voice to text from tv news report and finding more feasible way to determine the polarity of news article is included in the future study.

Acknowledgment

We are indebted to online news content publishing websites from where we have taken our data-sets to measure the performance of our system. Also we are thankful to Islamic University of Technology for their co-operation during the conduction of this research.

References

[1] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 129–136.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[3] S. Allan, *Online news: Journalism and the Internet*. McGraw-Hill Education (UK), 2006.

[4] E. Mitchelstein and P. J. Boczkowski, "Between tradition and change: A review of recent research on online news production," *Journalism*, vol. 10, no. 5, pp. 562–586, 2009.

[5] S. S. Sundar, "Multimedia effects on processing and perception of online news: A study of picture, audio, and video downloads," *Journalism & Mass Communication Quarterly*, vol. 77, no. 3, pp. 480–499, 2000.

[6] S. Padmaja, S. S. Fatima, S. Bandu, P. Kosala, and M. Abhignya, "Comparing and evaluating the sentiment on newspaper articles: A preliminary experiment," in *Science and Information Conference (SAI), 2014*. IEEE, 2014, pp. 789–792.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[8] S. K. Singh, S. Paul, and D. Kumar, "Sentiment analysis approaches on different data set domain: Survey," *International Journal of Database Theory and Application*, vol. 7, no. 5, pp. 39–50, 2014.

[9] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[10] J. Li, S. Fong, Y. Zhuang, and R. Khoury, "Hierarchical classification in text mining for sentiment analysis of online news," *Soft Computing*, vol. 20, no. 9, pp. 3411–3420, 2016.

[11] S. Fong, Y. Zhuang, J. Li, and R. Khoury, "Sentiment analysis of online news using mallet," in *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*. IEEE, 2013, pp. 301–304.

[12] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations of words using spin model," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 133–140.

[13] J. Kamps, M. Marx, R. J. Mokken, M. De Rijke *et al.*, "Using wordnet to measure semantic orientations of adjectives." in *LREC*, vol. 4. Citeseer, 2004, pp. 1115–1118.

[14] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 617–624.

[15] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.