Opinion Mining of News Headlines using SentiWordNet

Apoorv Agarwal
NIT-Jalandhar
Punjab, India
Apoorv.agarwal100@gmail.com

Vivek Sharma
NIT-Jalandhar
Punjab, India
Vivek@crushus.com

Geeta Sikka
NIT-Jalandhar
Punjab, India
sikkag@nitj.ac.in

Renu Dhir
NIT-Jalandhar
Punjab, India
dhirr@nitj.ac.in

Abstract:

Opinion Mining (also known as "Sentiment Analysis") is an area of text classification which continuously gives its contribution in research field. The main objective of Opinion mining is Sentiment Classification i.e. to classify the opinion into positive or negative classes.

SentiWordNet is an opinion lexicon derived from the WordNet database where each term is associated with some numerical scores indicating positive and negative sentiment information. Up until recently most researchers presented opinion mining of online user generated data like reviews, blogs, comments, articles etc. Opinion mining for offline user generated data like newspaper is unconcerned so far despite the fact that it is also explored by many users. As a first step, this paper present opinion mining for newspaper headlines using SentiWordNet.

Further, most of the researchers implement the opinion mining by separating out the adverb-adjective combination present in the statements or classifying the verbs of statements. On the other hand, in this paper we analyze each and every word in the News headline whether it is a noun, verb, adverb, adjective or any other part-of-speech. During experiment, python packages are used to classify words. Then SentiWordNet 3.0 is used to identify the positive and negative score of each word thus evaluating the total positive/negative impact in that news headline.

Keywords: Opinion Mining, sentiment classification News Headlines analysis, SentiWordNet, Positive-Negative Scores.

I. INTRODUCTION:

Opinion mining is a growing field to identify the thoughts and sentiments of people, which they express in form of their feedbacks or reviews on various things. Today due to vast use internet and social platforms, people are having a huge amount of space where they can publically express their opinions. These reviews are present in various forms on web like the feedbacks for products listed on various e-commerce web sites, or the personal posts from Facebook, twitter, bloggers etc. [1]. Some formal reviews are also available in various discussion forums related to products/sites or domains. People also post a lot of personal views in form of movie reviews or the buzz creating news in various articles for magazines and newspapers. These opinions are directly related to how they feel. And this feeling can be classified as being positive, negative or neutral in nature. Positive views have a positive impact on society and negative views creates a negative impact.

Today, media also plays an important role in developing a person's views and thoughts related to any product or scenario. Any news article that a newspaper publishes sometimes creates negative while sometimes positive buzz about any scenario to common public. This buzz directly impacts the society on a large scale. As per Dr. Daniel Dor [2] most of people judge the news contents directly by scanning news headlines only rather than going through complete story and hence they generate a quick thought about it. This shows that even a small headline can also plays a vital role in any judgment.

In this paper we propose an algorithm which classify the given news headlines whether they have positive impact or negative impact on society. This paper is classified into following sections: Section II contains the previous work done in this field. Our proposed work is described in Section III. Section IV consists of the results generated. Conclusions and future scope are discussed in Section V.

978-1-5090-0669-4/16/$31.00 ©2016 IEEE

II. PREVIOUS WORK

Most of the researchers are doing Sentiment analysis by identifying affective words from the statements that are responsible to formulae one's opinion on a particular subject. In past few years researchers have done a good amount of work in the field of sentiment analysis by identifying various combinations of adverb-adjectives and adjective-adverb-verbs [3]. Also there has been work on sentiments analysis of social issues based on verb as most important term in identifying opinions behind reviews. [4]. In other researches like [5], authors have done a clause level sentiment analysis by extracting opinions form independent clauses of statements. To implement their system they used sentiment scores from SentiWordNet [6] which is a lexical resource for sentiment analysis and opinion mining. SentiWordNet 3.0 is an enhanced version of SentiWordNet 1.0 which is publically available for research purposes. It is currently licensed to more than 300 research groups and worldwide a variety or research projects are using it [7]

The domain of analysis of news articles has been traversed before also like [8] but most research uses machine learning techniques to extract sentiments. Researches like [9, 19] have described the way for opinion mining but by analyzing complete articles. There are a few researches like [10] which have analyzed the sentiments by using news headlines only, but by using naïve Bayes classifier technique. In our current paper we tried to analyze the news headlines by using a Part-of-Speech Tagger and globally available resource of SentiWordNet. [11]

Part-of-Speech Tagging is a process by which we assign a suitable part-of-speech to any word in the sentence. Kristina T. et.al. [12] Have presented an enhance POS Tagger with 97.24% accuracy in comparison with previously used taggers. We used this POS Tagging with same abbreviations and details that are general and available online on various links [13]

III. PROPOSED WORK

It is known fact that Media and News plays a crucial role in developing a personal view on any topic. Also while scanning a News on TV, Newspaper or Internet, we are first attracted towards headlines only. In this paper we are evaluating those headlines and classifying them as creating positive or negative buzz about the scenarios.

To evaluate any headline we will be using two algorithms: Algorithm 1 and Algorithm 2. Before running Algorithm 2, Algorithm 1 is run to preprocess the words for passing it into SentiWordNet. This is so because SentiWordNet is not able to handle multi word queries [14].

Algorithm 1 is used to preprocess the words taken from News Headline. While preprocessing it uses *POS-Tagger, Lemmatization, and Stemming* steps which are explained below:

i) POS-Tagger – POS-Tagger or POST is used to find which word is of which part of speech. To get functionality of POST we have used the Natural Language Tool Kit (NLTK) Tagger which contains necessary modules to judge part of speech of a sentence up to 94.0 % and still evolving, which produce fairly accurate results.

Algorithm 1: Preprocessing of each word

1. Input a single headline h.
2. Pass it into POS-Tagger to identify Part-of-speech of each word i.e. noun, verb, adjective, adverb etc.
3. Lemmatize the word to extract its significance that in what reference it is used.
4. Normalize the word by stemming it.
5. Output the resulting word (w) as input for SentiWordNet for further processing.

Algorithm 2: To analyze a news headline

1. Identify all the Headlines H of the News Articles of a day.
2. Preprocess each headline h ϵ H to make it suitable for SentiWordNet 3.0 (Ref. Algorithm 1)
3. Pass each word into SentiWordNet Dictionary to identify its positive, negative and objective scores.
4. Add positive scores (p) and negative scores (n) of all words of h separately.
5. If p>n mark the news headline as having +1 polarity
6. Else if p<n mark the news headline as having -1 polarity
7. Else mark the news headline with 0 polarity.
8. End.

ii) Lemmatization – This is the process of grouping different inflicted form of words so that they can be analyzed as one. We have many words which often comes together, and only makes sense when they come together, if they are broken and used then they will convey different message, Example: 'Good to go' will convey different message when used individually.

iii) Stemming–This is used to reduce inflicted or derived words to their word root or base. Derived words which we come across can't be used for further processing as this contain that form of word which are derived from other words, so we use stemming which converts a word to its root word. For example word 'Doing' & 'Done' are originating from same word 'Do', but in SentiWordNet if we use other form of 'Do' it will show not found error and will only get right results when using Word 'do'. Hence we use stemming for this problem we use Porter Stemmer [15] for this purpose.

For example, the Reuter's news of Dec 1, 2015 states: "*Bill Gates plots a surprise attack on the Energy Sector"*

After applying various steps of above algorithms, following is the analysis of sentiments score of this headline:

Tokenized Word	Classified Part-of-speech	+ve score (P)	-ve score (N)
Bill	Proper Noun	0.0	0.0
Gates	Proper Noun	0.0	0.0
plots	Verb Past tense	0.0	0.0
a ,the	Determiner	0.0	0.0
Surprise	Noun Singular	0.0	0.0
Attack	Noun Singular	0.25	0.125
On	Preposition	0.0	0.0
Energy	Noun Singular	0.0	0.0
Sector	Noun Singular	0.0	0.0
SUM		0.25	0.125

Table 1.

From the Table 1, it can be inferred (from the value of p-n) that this headline will be classified with +1 polarity.

Once we have allotted +1, 0 or -1 polarity to all the news headline of the day, we will then apply following algorithm to find the net score of each day.

Algorithm 3: Analysis of all news headlines of a single day

1. Assign polarity value V, to all news headlines where V→[-1,0,+1]
2. Find sum of all positive polarities into Vp i.e. $Vp = \sum_{V=+1} V$
3. Find sum of all negative polarities into Vn i.e. $Vn = \sum_{V=-1} V$
4. Find total score of a day as: T.S.= $Vp - Vn$
5. Return T.S. as evaluated polarity score of the day.

Using the Algorithm 3, we can easily deduce the total +ve or –ve score that our experiment has allotted to that day.

From above algorithms we have calculated the day wise +ve and –ve scores of news. We also checked the headlines and manually allotted them +ve, –ve or neutral polarity as per their social impact. If our experimented polarity matches the manual allotted value then this will conclude that our algorithm has classified it correctly otherwise our algorithm has returned an error.

Following table shows the computation and evaluation of top 10 world news headlines of 3rd Dec 2015.(http://in.reuters.com/news/archive/worldNews?date=12032015) :

News Headlines (3Dec2015)	Experimental Scores			Manual Scaning	Result
	+ve	-ve	Res		
Headline1	0.5	0.5	0	-1	E
Headline2	0.0	1.0	-1	-1	C
Headline3	0.0	0.375	-1	-1	C
Headline4	0.0	.75	-1	-1	C
Headline5	0.0	0.125	-1	-1	C
Headline6	0.125	0.875	-1	-1	C
Headline7	0.75	0.625	+1	+1	C
Headline8	0.5	0	+1	-1	E
Headline9	0.0	0.125	-1	-1	C
Headline10	0.625	0.25	+1	+1	C

Table 2.

In Table 2, resulting experimental scores (Column 4th) are evaluated from algorithm and compared them with manually provided polarity values to news headlines (Column 5th). If there was any deviation from experimental result to manual result then we have marked them with E(error) in column 6. Otherwise they were marked C (Correct Results).

We have also calculated error percentage using formula:

$$\text{Error \%} = \frac{\Sigma\ Headlines\ classified\ with\ Error}{\Sigma\ Headlines\ Evaluated} * 100$$

For example, in above table we can find error% as,

$$\text{Error\%} = \frac{2}{10} * 100\ \% = 20\%$$

IV. RESULTS

We ran the above algorithms for around 500 news headlines of past 30days and analyzed the deviation of experimented values from expected values of polarity. We summed up the total polarity of the day from both experimented set and manual set separately, evaluated the difference of two values and then found the total average deviation of the data. We are showing a small graph of the results related to this calculation:

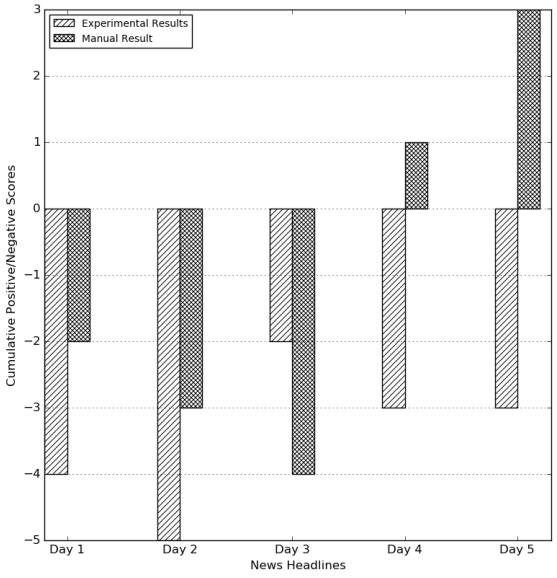

Fig 4.

Days	Exp. Polarities			Manual polarity			Diff
	+ve	-ve	T1	+ve	-ve	T2	|T2-T1|
1	+3	-7	-4	+4	-6	-2	2
2	+7	-12	-5	+8	-11	-3	2
3	+7	-9	-2	+6	-10	-4	2
4	+6	-9	-3	+8	-7	+1	4
5	+7	-10	-3	+10	-7	+3	6
Sum							16

Table 3.

From the above generated results:

Total Deviation for 5 days= 16

Average Deviation=16/5 =3.2

When we ran this for 30days News Headlines, the avg. Deviation found was 2.7. This is so because of either use of sarcasms used in headlines or slightly different processing results from our POS-Tagger then what it should be.

V. CONCLUSION AND FUTURE SCOPE

The Sentiments analysis is a vast field that can be worked from various disciplines including Artificial Intelligence, Natural Language Processing and various other Text mining approaches [16]. We did these opinion mining (sentimental analysis) on set of national/global news headlines but we can again classify these as per different demographics or area. Mining the news of specific area can help a lot of people including government and various decision making bodies as well as common people to analyze the sentiments of that area. Government can analyze the effect of its polices in any region, election candidates can learn about specific requirement and development needs of their zones and business units can learn the sentiments and demands of particular region.

In our experiment we analyzed and tested an algorithm by which any news statement can be classified as containing positive or negative sentiments in it. We tested the algorithm on newspaper headlines but same algorithm can be applied on other contents including Online Reviews, Product and Application reviews on various Play Stores/App Stores, Movie Reviews, and Comments on various social platforms including Twitter/ Facebook and various bloggers. The Experiment and algorithm can be improved to analyze the sarcasm in the statements/Comments. Also we tested the accuracy of the classification by comparing it with manual classifications of news. For a huge data, this manual classification task can be erroneous and a tedious job and can deviate the results. We can classify them using support vector machine technique [17] or apply some other machine learning tasks for the same [18] and hence compare the results of our proposed work with their results.

VI. REFERENCES.

[1] H. Binali, V. Potdar and C. Wu, "A state of the art opinion mining and its application domains", 2009 IEEE International Conference on Industrial Technology, 2009.

[2] Sciencedirect.com, "On newspaper headlines as relevance optimizers", 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378216602001340.

[3] V. Singh, R. Piryani, A. Uddin and P. Waila, "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification", 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013.

[4] M. Karamibekr and A. Ghorbani, "Sentiment Analysis of Social Issues", 2012 International Conference on Social Informatics, 2012.

[5] T. Thet, J. Na, C. Khoo and S. Shakthikumar, "Sentiment analysis of movie reviews on discussion boards using a linguistic approach", Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09, 2009.

[6] Citeseerx.ist.psu.edu, "CiteSeerX â€" Document Not Found", 2015. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0A54D847CA44C482AD2FF613F3B17852?doi=10.1.1.61.7217&re.

[7] S. Baccianella, A. Esuli and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", LREC Conference, 2015. [Online]. Available: http://lrec.elra.info/proceedings/lrec2010/pdf/769_Paper.pdf.

[8] A. Balahur and R. Steinberger, "Rethinking Sentiment Analysis in the News: from Theory to Practice and back", JOINT RESEARCH CENTRE The European Commission's in-house science service, 2015. [Online]. Available: http://langtech.jrc.it/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf.

[9] P. Raina, "Sentiment Analysis in News Articles Using Sentic Computing", 2013 IEEE 13th International Conference on Data Mining Workshops, 2013.

[10] H. kaur and D. Chopra, "Sentiment Analysis of News Headlines using NaÃ¯ve Bayes Classifier", Council For Research And Development Enterprise, 2015. [Online]. Available: http://www.cfrde.com/pages/harpreetPDF.pdf.

[11] Sentiwordnet.isti.cnr.it, "SentiWordNet", 2015. [Online]. Available: http://sentiwordnet.isti.cnr.it/

[12] K. Toutanova, D. Klein, C. D. Manning and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", The Stanford Natural Language Processing Group, 2015. [Online]. Available: http://nlp.stanford.edu/pubs/tagging.pdf.

[13] Computer Science, University of Maryland, "Part-of-Speech Tagging", 2015. [Online]. Available: https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf.

[14] J. Kreutzer and N. Witte, "Opinion Mining Using SentiWordNet", 2015. [Online]. Available: http://stp.lingfil.uu.se/~santinim/sais/Ass1_Essays/Nele_Julia_SentiWordNet_V01.pdf.

[15] M. Porter, "An algorithm for suffix stripping", Department of Computer Science, Old Dominion University, 2015. [Online]. Available: http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf.

[16] H. Binali, Chen Wu and V. Potdar, "A new significant area: Emotion detection in E-learning using opinion mining techniques", 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, 2009.

[17] H. Binali, C. Wu and V. Potdar, "Computational approaches for emotion detection in text", 4th IEEE International Conference on Digital Ecosystems and Technologies, 2010.

[18] C. Troussas, M. Virvou, K. Espinosa, K. Llaguno and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning", IISA 2013, 2013.

[19] Sharma, Vivek et al. "Sentiments Mining And Classification Of Music Lyrics Using Sentiwordnet". 1st IEEE Symposium on Colossal Data Analysis and Networking 2016.

[20] Singh, Deepak, and Dr. Harsh Verma. "A New Framework For Cloud Storage Confidentiality To Ensure Information Security". 1st IEEE Symposium on Colossal Data Analysis and Networking 2016.