IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

Sentiment Analysis Tool

Prof. Ajitkumar Khachane[1], Lakshya Kumar[1], Chandresh Jain[1], Ashay Shah[1]

Department of Information Technology, Vidylankar Institute of Technology, Mumbai[1]

Abstract: Since the advent of Internet, and later on Social Media and E-commerce websites, people have started expressing themselves more and more online instead of on paper. This give businesses and organizations an opportunity to analyze views and interests of people in accordance with a set of provided keywords, time duration, geographic locations, age group and thus. Following the human nature of curiosity, collecting a selected set of opinions and sentiments and then topic-oriented analysis allows extraction of rich information that would help make smarter business decisions, political campaigns and better product consumption.

Keywords: Sentiment Analysis; Natural Language Processing.

I. INTRODUCTION

Sentiment Analysis is a widely-studied application of Natural Language Processing [4] and Machine Learning. This field keeps growing every day and the amount of data to process is ever-increasing. To keep up with the accuracy and efficiency of the sentiment analysis tool, it needs to be updated frequently. There is always a need for new techniques that will help enhance performance.

As the data to deal with is huge and unmarked, supervised methods deem to be time-consuming and inaccurate with time. Sentiment Analysis is carried out in following steps: The subject and object is identified. Sentiments are derived with the help of available datasets. Derived sentiments are polished by referring to previous instances and all possible combinations of the given subject and object.

Sentiment Analysis could be handled at various levels of granularity [3] starting from:

- Document level
- Sentence level
- Phrase level.

II. LITERATURE SURVEY

Sentiment Analysis tool [1] is intended to provide user-specific summarization of huge documents, topic-oriented polarity of public opinions and product-related summarizations. Opinion Mining or Sentiment Analysis refers to an area of Natural Language Processing and also Text Mining. It is not concerned or dependent on the topic of the document/sentence/phrase but solely with the sentiment it expresses.

A survey performed by Lillian Lee and Bo Pang [5] states: Sentiment Analysis identifies the viewpoints of a text. For example, giving a YouTube video 'thumps up (recommended)' or 'thumps down (not recommended).' Former methods used selective lexical features (e.g. the word 'easy' and 'hard'), then giving the document a sentiment or summarization depending on the number of occurrences of such words.

Later, it was found that it is not enough to depend on lexical features completely and the actual sentiment of the document could be very different than the summarization. Hence, more advanced techniques of Machine Learning and Tokenization were adopted. Today, all Sentiment Analysis tools use these techniques to provide more accurate results.

It is also found that the same word could have different meanings depending on the sentence in which it is being used. Consider the following two sentences:

1. The language of the author was crude.
2. Crude oil is extracted from sea beds.

The word 'crude' acts as subject in the first sentence and as object in the second sentence. Old lexical features will splendidly fail to analyze such words in a document. However, with the help of Tokenization and Machine Learning, there is no such effect on the results.
Supervised methods also fail after an extent to determine the sentiment of a document. The word 'unpredictable' could be used in different contexts wherein it could be 'negative,' 'positive' or 'neutral' to the document. This cannot be determined in supervised methods. Unsupervised methods take help of recent descriptive words use in the document and the coherent nature of object and subject to determine in which context such words are being used. It was no news when people started using short forms and slang on the internet as it only made sense to save characters and time while achieving the same impact. So it was only due time to start adding such words in the dictionary and linking it to the actual words or phrases. With keeping all of this in mind: Tokenization, Machine Learning, Unsupervised methods, short forms, slang, and also negation words, a new more effective and accurate design is proposed.

III. TRADITIONAL SYSTEM

Present Sentiment Analysis tools lack the modern way of thinking about sentiments. The same tool works fine in

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

analyzing older documents while it could not keep up with the speed at which language changes on internet. Huge amount of manpower is invested to keep updating the databases. New words have to be added frequently even if it has the same meaning (as in short form or slang form of a word).

A whole lot of phrases have taken form of words or even replaced proper sentences, mostly expressed in short forms. This leads to contingency in exploring sentiment of a sentence. If the tool could not connect the relation between object and subject, it may skip that part of the sentence giving an incomplete result.
Most of the available Sentiment Analysis tool use either Lexicon-based approach or Machine Learning approach. Each method has its own advantages but could not give all advantages of the other. The output shown to the user is in standard textual or percentage/rating form.

Traditional methods fail to know the focal point of every document. This may be due to '#' reading incapability or improper '#' use. Emoticon is also being used more frequently these days to express complex emotions in fewer characters. Present dictionaries do not give emoticons much emphasis or even ignore them altogether in some cases. This, again, gives an incomplete sentiment of the whole document.
Thus discovered incompetence in the traditional system leads to:
1. Relatively more money spent on manpower than maintenance of dictionary.
2. Improper and incomplete sentiment results over time.
3. Data explosion due to huge amount of useless data collection.

The above mentioned drawbacks do not indicate that the traditional system is a faulty one but merely it is outdated. Expression of sentiments over the internet evolves quicker than it can handle, and so a new system is needed to keep up with language and its interpretation.

IV. PROPOSED SYSTEM

The proposed system would use cloud services to handle big data more efficiently with less manpower. This will help utilize more resources for actual maintenance of dictionary rather than error-handling or proofreading. The whole dataset whether a sentence or a document, is taken into account while deriving sentiments so that none of the relationship between an object and a subject is overlooked. This makes sure the derived sentiment is consistent throughout the document and does not give false interpretation when only a part of the result is read.

Both Machine Learning and Lexicon-based approaches are used together in balance to give more polished results. Input texts are marked and prepared for analysis in the first stage as shown in the below architecture of the proposed system.

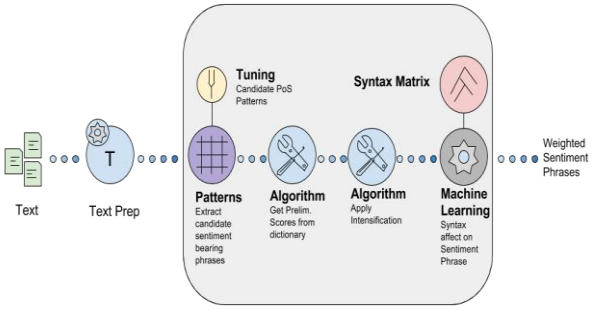

Figure 1: System Architecture

Then various data mining techniques are performed for finding patterns to make further processes easier. Lexicon-based approach is used first to determine the sentiment of the input phrases from an ever-changing heuristic dictionary. This gives a near-accurate result.
Then by comparing it to previous instances of such object-subject relationships, the results are given proper intensity. Next, Machine Learning approach is used to further polish the results by deriving real-time information about the object in question and using probability functions to see whether the author is being sarcastic or not.

The final result derived is:
1. More accurate than if only one approach is used.
2. Complete in sense with respect to both the object and the subject.
3. Faster as the sentiments are not derived using instances of a given word but rather its real-time use on the internet.

Data is replaced more frequently as the language changes on the internet, this prevents data explosion. Descriptive words are marked making them easy to find and update meaning as and when necessary. Short forms and slang words are merged with the actual words saving space and time. Thus, the proposed system overcomes all drawbacks that were mentioned and is made adaptable for changes which might be needed in future.

V. METHODOLOGY

Tokenization is the name given to process of breaking sentences into words and tagging them. The segmentation of tokens could be done with decision trees. Wherein, some of the issues that present are:

1. The choice for the delimiter could be commas or whitespace ("We're going to Peru." >> ["We're," "going," "to," "Peru."]), but the issue arises when there are words with whitespace between them ("We're going to New Jersey." >> ["We're," "going," "to," "New," "Jersey."]).
2. Punctuation marks could not be thrown away without consideration as, for Sentiment Analysis, '!' or '?' could mean emphasis or confusion respectively and so a lot of information could be deduced from punctuation marks.
3. ", ', [], () can mean the words belong together and should be treated accordingly. Also, **bold,** italic, and underlined words have their own importance.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

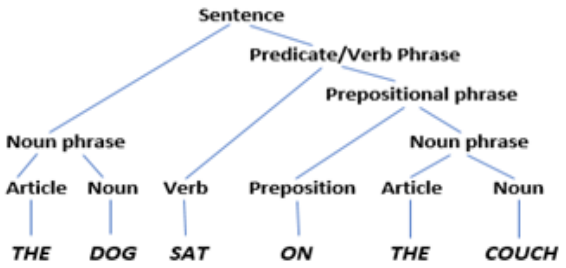

Figure 2: Word analysis of a text

Once the phrase is broken down into tokens, further steps will take place. Each word, part of the object or subject of the phrase, is then mapped with its present meaning in the dictionary. The accuracy of the dictionary depends on how frequently it is updated.

Sentiment Classification techniques are broadly of two types: Lexicon-based approach and Machine Learning approach. Lexicon-based approach gives a word its meaning as present in the dictionary however it may or may not be accurate. Machine Learning approach gives a near-exact meaning of a word however it first needs a dictionary based meaning of the word which it polishes for accuracy. The hybrid approach [2] combines both approaches used one after another in the same process to give more accurate results.
Machine Learning can be supervised or unsupervised. Supervised method makes use of large number of labeled documents to help new words or new meanings derive sentiments. However, it might fail to give object-oriented true meaning of the given sentence. This drawback is overcome in unsupervised method which in turn is always near-exact and not ever perfect.

VI. CONCLUSION

The proposed tool overcomes all the aforementioned drawbacks and further enhances the speed of the process. It only makes use of the resources already available and does not hikes the budget of any organization which is currently using traditional system. It only shifts the focus on the sentiment deriving process from the maintenance of the dictionary.

Unlike previous systems it is not designed only to deal with present situations but is designed to keep changing with the change in language on the internet and its interpretation. It has room for adding more features if required (such as recommender system, artificial intelligence etc.).
The proposed system has applications in several domains: Review-related websites, sub-component technology, business intelligence, sociology, artificial intelligence.

REFERENCES

[1] Jeff Stokes, Larry Franks, Olga Matoula, and Janet Yeilding, "Social media analysis: Real-time Twitter sentiment analysis in Azure Stream Analytics."
[2] Pedro P. Balage Filho, and Thiago A. S. Pardo, "NILC_USP: A hybrid system for sentiment analysis in Twitter messages," Second Joint Conference on Lexical and Computational Semantics (*SEM), vol. 2, pp 568-572, Atlanta, Georgia, June, 2013.
[3] Walaa Medhat, Ahmed Hassan, and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5(4), pp 1093-1113, December 2014.
[4] Steven Bird, Ewan Klein, and Edward Loper, "Natural Language Processing," Natural Language Toolkit, 2009.
[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.