International Journal of Management, IT & Engineering
Vol. 8 Issue 5, May,2018
ISSN: 2249-0558
Impact Factor: 7.119Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

SO-CAL for Sentiment Analysis-a Natural Language Processing tool

By

Karthik Trichur Sundaram, karthikts@hotmail.com

1. Introduction

Speech recognition comprises two parts, meaning comprehension and sentiment analysis. With opinion mining gaining more and more relevance every day, sentiment analysis has become relevant to researchers, businesses, and governments alike Sentiment intensity is a more sophisticated understanding of sentiment, in which a scaling system is used to assign a numerical value to each kind of sentiment depending on how powerfully sentiment is expressed. Assuming selected scaling system from +x to -x (most positive to most negative) makes it possible to detect certain degree or level of sentiment being expressed by a speaker. Sentiment intensity can adjust the sentiment score of a given term or text relative to concept and degree of intensification or relaxation. Businesses as well as governments can review public opinion about their products or specific events and can gauge intensity of their likes or dislikes. Social media has increased the importance of sentiment strength analysis as it requires constant monitoring of information to deduce important results like customer satisfaction level. Happiest customers are more receptive to upselling. Intense negative comments can help decision makers to improve their service delivery. A study (Piryani, Madhavi, & Singh, 2017) [4] discusses a framework for processing unstructured data to extract views and identify their moods. A lot of approaches are used by researchers to evaluate and enhance sentiment detection, most commonly used are Machine Learning (ML) based, lexicon based (Kundi, Khan, Ahmad, & Asghar, 2014) [3] and hybrid (combination of both) (Balage Filho & Pardo, 2013) [2] approaches. Sentiment analysis is focused on polarity analyses in form of neutral, positive or negative label.

1.1. Natural Language Processing

Natural language processing is a computerized approach to analyze text in any form and in any language to achieve results that are similar to human processing for a wide variety of tasks or applications. The goal of NLP is "to achieve human-like language processing" and it reveals that

NLP is a sub discipline of Artificial Intelligence (AI). Major focus in NLP is to understand how human beings understand and use language and to develop necessary tools to make computer systems understand and manipulate natural languages to perform the desired tasks. NLP has various applications in machine translation, natural language text processing and summarization, multi lingual information retrieval, processing, designing user interfaces, speech recognition, artificial intelligence and expert systems, and so on.

1.2. Sentiment Analysis

Sentiment analysis is a subfield of NLP, which analyzes people's opinions, feelings, thoughts, sentiments, attitudes, evaluations, appraisals and emotions towards certain entities. These entities can be products, services, organizations, individuals, issues, events, topics, and so on. Since opinion is a key influencer of our behavior so applications of sentiment analysis are applicable in almost every domain of life including consumer products, services, healthcare, and financial services to social events and political elections. Sentiment analysis is a good alternative of surveys, opinion polls ad focus groups.

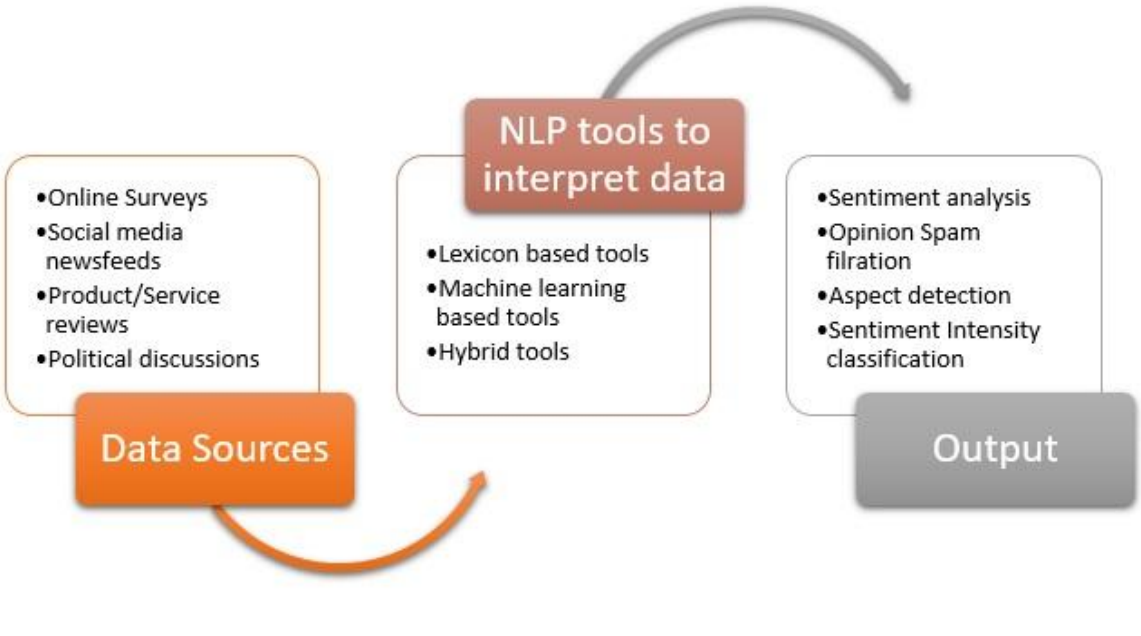

Figure 1: Sentiment Analysis process

2. Different Levels of Sentiment Analysis

Sentiment analysis in text is investigated mainly at different levels discussed below:

2.1. Document level:

Whole opinion document is analyzed to classify and predict its sentiment in the form of positive or negative polarity (Ahire, 2014) [1] Underlying approach includes cumulative sentiment of potential paragraphs to examine an overall tone of the document. It can also be used to categorize chapters or pages of a document as positive, negative, or neutral.

2.2. Sentence level:

Each sentence is analyzed to determine whether it has expressed a positive, negative, or neutral opinion. Neutral usually means no opinion.

2.3. Phrase Level Sentiment Analysis

In phrase level classification, such phrases are extracted from text that contains opinion words. It is preferred in a situation where the exact opinion about an entity can be correctly extracted. Word, being the basic unit of language, defines the ultimate polarity or subjectivity of the sentence (Thet, Na, & Khoo, 2010) [9].

2.4. Entity and Aspect level

Both the document level and the sentence level analysis do not discover what exactly people liked and did not like. To perform finer-grained analysis, entity and aspect level analysis is used in which instead of looking at paragraphs and sentences to determine polarity, directly opinion is searched. It is based on the idea of finding a sentiment (positive or negative) and a target (of opinion) as opinion without its target being identified is of limited use.

3. Important features in Sentiment Detection

Several features of text can facilitate sentiment detection such as emoticons, slang, and frequency of words or punctuation marks. An exclamation mark can hint at an intense emotion whereas a slang or emoticon can guide towards a positive or negative sentiment accordingly. *ROFL, LOL,* for instance,hint at an amusing incident and a sad emoticon will indicate otherwise. Likewise, [this product is very, very useful] hints at an intense positive response.

4. Semantic Orientation SO-CAL

Various tools have been developed to automate sentiment analysis including VADER, SentiWordNet, SenticNet, SentiStrength and SO-CAL. The SO-CAL or Semantic Orientation Calculator (Thelwall, 2017) [8] uses dictionaries in which words are listed along with intensity and polarity of each word. SO-CAL's performance is consistent across domains and on completely unseen data. Text whose semantic orientation needs to be detected is first analyzed to extract semantic words. To calculate semantic orientation of a sentence, adjectives, verbs, nouns, adverbs, intensifiers, down toners, negation, and irrealis, markers are extracted and corresponding value is assigned to each word or part of text depending on its value present in SO dictionary. SO-CAL takes these parts of speech into account, and makes use of more refined methods to determine semantic value of text. Four different datasets are used to test and validate SO-CAL's performance and it showed robust performance across different domains and different data sets

4.1. Advantages

- SO-CAL performance is robust across different domains and dictionaries used with it are reliable as compared to manually designed dictionaries and automatic ones
- It seems easier to construct SO-CAL for some other language as its code written in Python for English language can be reused.
- Contextual valence shifting including negation, parts of speech extraction and intensification has made it an effective approach
- Since, past research shows that automatically created dictionaries are not very stable. In SO-CAL, manual dictionary is used by hand tagging all adjectives. Also instead of using only average to calculate overall score, a lot more sophisticated approaches are used that give finer results.
- In most of lexical approaches, single word entries are used. But one of most effective advantage of SO-CAL is that it allows multi-word entries written in a regular expression–like language like bring about, break through, a pinch of, a little bit etc.

- Sometimes, such terms are used that are actually used to express a sentiment but in current scenario they are used in different context and not expressing any sentiment, SO-CAL blocks such words.
- Those lexicons that are not using negation rules to calculate sentiment are normally poor in performance and luckily SO-CAL is not one of them. In SO-CAL, Negation and intensification together increase performance significantly.
- Average accuracy of SO-CAL is 0.7874 that is quite stable across multiple domains

4.2. Potential Limitations

- Word Sense disambiguation needs to be focused more in SO-CAL. Although POS extraction can help to reduce ambiguity of word sense but still a simple method is needed to get more accurate score (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) [7]. Word Sense or argument structure detection needs more effort in future.
- In SO_CAL, intensifiers are modeled by associating a certain percentage of value with intensifying word. This intensification depends only on keyword that reflects intensification, logically it can be improved if intensification should also depend on term being intensified and already loud item should have a greater overall increase and vice versa. For example there should be different intensification percentage for truly fine and truly awesome.
- SO-CAL performs better on positive reviews but it needs more contribution to get better results on negative reviews.
- SO_CAL assigns equal weight irrespective of whether the text belongs to description, heading or comment. Significant performance can be achieved by assigning lower weights to description and higher weights to headings.
- SO-CAL uses only lexicon-based approach to calculate Semantic Orientation (SO), a machine learning approach can be integrated with it to get better results.
- SO-CAL has limited coverage as its sentiment resource consists of about 5,000 words that are in sufficient as compared to SentiWordNet that has over 38,000 polar word.

• SO-CAL is a lexical-based approach. One of the disadvantages of a lexical based resource is that the resources necessary for a new domain or a new language need to be built from scratch, whereas a machine-learning approach only needs enough data to train.

5. Conclusion

Sentiment detection is crucial to opinion mining in the era of 4th revolution towards IT and technology. With businesses moving online and a significant social media presence creates motivation for organizations to extract feedback from user's opinions to improve their service delivery. In this regard, sentiment analysis techniques need to be robust to extract effective response from the audience. Various tools that exist today have their own peculiarities and serve specific purposes. An analysis of strengths and weaknesses of SO-CAL is presented in detail, keeping in view the rules of NLP. This study will prove effective in developing similar analyses on other tools as well as endeavors to present improvements to other open source tools.

6. References

1. Ahire, Sagar. (2014). A survey of sentiment lexicons. *Computer Science and Engineering IIT Bombay, Bombay*.
2. Balage Filho, Pedro, & Pardo, Thiago. (2013). *NILC_USP: A hybrid system for sentiment analysis in twitter messages.* Paper presented at the Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).
3. Kundi, Fazal Masud, Khan, Aurangzeb, Ahmad, Shakeel, & Asghar, Muhammad Zubair. (2014). Lexicon-based sentiment analysis in the social web. *Journal of Basic and Applied Scientific Research, 4*(6), 238-248.
4. Piryani, Rajesh, Madhavi, D, & Singh, Vivek Kumar. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management, 53*(1), 122-150.
5. Taboada, Maite, Brooke, Julian, Tofiloski, Milan, Voll, Kimberly, & Stede, Manfred. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.
6. Thelwall, Mike. (2017). The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength *Cyberemotions* (pp. 119-134): Springer.
7. Thet, Tun Thura, Na, Jin-Cheon, & Khoo, Christopher SG. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science, 36*(6), 823-848.