Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English Language

Aparna Trivedi, Apurva Srivastava, Ingita Singh, Karishma Singh and Suneet Kumar Gupta

Information Technology(B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College
Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College
Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College
Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College
Ghaziabad, 201009, India

Information Technology (Associate Professor), Gautam Buddh Technical University, ABES Engineering College
Ghaziabad, 201009, India

Abstract

This literature work is a survey about Sentiment Analysis of textual data of English language and some of its previous works. Basically sentiment analysis identifies the view point or opinion of a text. For example, classifying a movie review as “Thumbs up” or “Thumbs down”.

Several public opinion surveys from multiple polling organization and people’s aggregate opinion on a topic can be assessed. This survey includes previous works which show how this technique has evolved over the past one and a half decade expanding its horizon and reaching out to almost all areas such as reviews of products, movies etc., travel advice, stock market predictions and in other decision making areas.

Keywords: *Opinion Mining, Sentiment Analysis, Polarity Identification.*

1. Introduction

Natural Language Processing is a domain of computer science and scientific study of human language i.e. linguistics which is related with the interaction or interface between the human (natural) language and computer. Basically NLP commenced as a sub-field of artificial intelligence. Opinion mining or Sentiment analysis refers to a broad area of Natural Language Processing and text mining. It is concern not with the topic a document is about but with opinion it expresses hat is the aim is to determine the attitude (feeling, emotion and subjectivities) of a speaker or writer with respect to some topic to determine opinion polarity. Initially it was applied for classifying a movie as good or bad based on positive or negative opinion. Later it expanded to star rating predictions, product reviews travel advice and other decision making processes.

According to the survey performed by Bo Pang and Lillian Lee, Sentiment analysis identifies the view points of a text. For example, classifying a movie review as thumbs up (recommended) or thumbs down (not recommended). Previous methods focused on selective lexical features (e.g. word "Good"), then classifying document according to the number of such features that occur anywhere within it. But in contrast later following process were followed:

- Identify the sentences in the given input text as subjective or objective.
- Select and apply a standard machine learning classifier to the extracted result.

This could prevent the polarity classifier from considering misleading, ambiguous or irrelevant text. For example, the sentence "The protagonist tries to protect her good name" holds the word "Good", but it reports nothing about author's opinion and could also be implanted in a negative way.

Our work is based on this technique of Sentiment analysis using polarity classification of textual data. In this, we estimate the percentage of positivity or negativity of input text by first tagging all the adjectives, adverbs using a POS (Part of Speech) tagger (Marks words in the input text corresponding to a particular part of speech). Then we estimate the positivity or the negativity of the extracted adjectives corresponding to its value in the SentiWordNet (derived from WordNet, a lexical database, where numerical value indicating polarity sentiment, i.e. positive or negative, information corresponds to each word in it). In order to estimate sentiment orientation we count the positive and negative terms values. Finally, we assign estimated polarity to the given corpus.

2. Evolution

According to the paper titled "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" by Peter D. Turney, presented in the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002), in Philadelphia, Pennsylvania, a major application of sentiment analysis of textual data is in the classification of any review. Example: The semantic orientation of phrases with the help of simple unsupervised learning algorithm. The whole process of classification can be summaries into simple three steps:

1. Identify phrases in the given corpus containing adjectives or adverbs (using a part-of-speech tagger given by Brill in 1994).
2. Approximate the semantic orientation of the identified phrase.
3. Based on the sentiment orientation classify the given input text.

This algorithm makes use of PMI-IR to calculate semantic orientation. Peter D. Turney experimented with 410 reviews of various domains and concluded that the algorithm accomplishes an average accuracy of about 74%. But for movie review its about 66% while 82%-84% for automobiles and banks. The limitations identifier in this work of Peter D. Turney was the time needed for queries which can be eliminated by development in hardware. The level of accuracy can be improved by gelling semantic features with some distinct features of a supervised classification algorithm. This work has its nearness to the work of "Predicting the semantic orientation of adjectives" by Hatzivassiloglow and Mc Keown presented at the Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL in 1997. The use of four step algorithm which:

1. Removes conjunction to isolate adjectives.
2. Uses a supervised learning algorithm to label adjectives into groups of same or different semantic orientation and result in the graph where nodes denote adjectives and links denote similarity or difference in semantic orientation.
3. Using a clustering algorithm, the graph is processed to give two subsets of adjectives: Positive and negative.
4. If the frequency of positive adjective is high, then the text is positive else negative.

But the algorithm overall is complex and improved in various fields.

Another work in this field was R.M Tong's "An operational system for detecting and tracking opinions in on-line discussions" at the ACM SIGIR 2001 Workshop on Operational Text Classification in 2001. The system trails online discussion and gives a graph for positive and negative sentiments looks for phrases like "bad acting", "awesome music", "uneven editing" etc. In This edition of phrases to a special lexicon as tagging of sentiment as positive or negative is done manually. But this work was specific for movies.

Further in this field, a research work on "Sentiment classification of reviews using SentiWordNet" was conducted by Bruno Ohana and Brendan Tierney, of Dublin Institute Of Technology, and presented in the 9th I.T & T Conference, 2009. They used automatic methods for speculating the course of subjective content on textual data. SentiWordNet (opinion lexicon) is basically used to classify automatic sentiment of film reviews and the research done elaborates the results produced. The research goes one step ahead through extending the use of SentiWordNet by building the set of significant features and applying to the machine learning classifier. The set of relevant features provided substantial enhancement over baseline term counting methods. The important conclusion drawn indicated that SentiWordNet has now emerged as an indispensable tool for sentiment classification tasks and further progress can be made in its user and its usage along with other techniques. The research associates words and their synonyms present in Synsets with two numerical number ranging from 0 to 1, each denote the SentiWordNet's positive and negative bias.

SentiWordNet's one of the prominent features is that, in this a term can posses can both positive and negative score to have non-zero values. Stanford part of speech tagger was used in the research to correctly associate scores to terms and then scores for each term was found. The ratio between scores and number of terms was found and overall score was calculated. The document was divided into sections and scoring was performed section-wise and in the end the final sentiment of polarity was analysed.

"From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", was proposed in May 2010, in which Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge and Noah A. Smith analysed several public opinion surveys from multiple polling organisations on consumer confidence and political over 2008 to 2009 period and found that they co-relate to sentiment word frequencies in contemporary twitter messages. The results vary across data sets and sometimes correlations were as high as 80%. For example, if we want to know the extent to which U.S. population likes or dislikes Obama, polling methodology is done. It was extensively developed through 20th century (Krosnick, Judd and Wittenbrink 2005). From text, population's aggregate opinion on a topic can be assessed and then the task can be broken down into two sub problems:

1. Message Retrieval – When we identify the messages relating to topic.
2. Opinion Estimation – Determine whether messages express a positive or negative opinion or just news about the topic.

Tamara Martin-Wanton and Aurora Pons-Porrata gave a paper "Opinion polarity detection" in which an unsupervised algorithm was used for polarity of opinion which uses a word sense disambiguation algorithm to determine the correct sense of the word in the opinion. This proposed method does not depend on the knowledge domain and can be extended to other languages. The resources used by the author for this method is:

1. WordNet (Lexical database).
2. SentiWordNet (Lexical Resource).
3. A subset of General Inquirer (English Dictionary).

The two basic components of this method are:

1. Word sense disambiguation
2. Determination of polarity

Word sense disambiguation identifies the correct senses of the terms and in the determination of the polarity we determine the polarity of the opinion. The uniqueness of this method is using standard external resources along with word sense disambiguation for determining polarity of the opinions. Thus, this method is independent of knowledge domain and can be extended to other languages. Because of wrong annotations of SentiWordNet, there may be failure of method in many cases.

One of the most recent paper is "The Truth About Sentiment and Natural Language Processing" published by Synthesio (a global, multilingual, Social Media Monitoring and Research company) in March'11 which focuses on how brands can ascertain what opinions people have about their brand through sources like social media, blogs, online newspaper and magazines etc. Even if the sentiment analysis is inappropriate and no social media assistance could prove that this technology could accurately access sentiment on a precise topic, but by stalking and leaning it over time we can examine the pattern for changes since we are presumptuous that the in correctness will be constant over time. However there is no proof of this yet.

In 2009 and 2010, Amitava Das and Sivaji Bandyopadhyaya of Jadavpur University presented their research paper "Phrase-level Polarity Identification for Bangla" and "SentiWordNet for Indian Languages", respectively, which emphasize on opinion polarity classification on news texts using Support Vector Machine (SVM) for popular Indian native language Bengali. The contemporary system present directs the course of an opinionated phrase to positive or negative. A pre-requisite for identifying the direction of opinion requires the categorization of texts into subjective or objective. The reason being objective text cannot be predicted by definition. The system uses a combined approach which co-ordinates well with lexicon entities and linguistic syntactic features and the classifier used in the research is rule based subjectivity. The results have accuracy of 70.04% and a recall of 63.02%. The limitation of the research lies on the usage of log-linear functions models like SVM. The major drawback being that a well distinct decision boundary cannot be formed from the conjunction of provided features. The disadvantage can be overcome by providing the conjunctions explicitly as an integrated unit of feature vector, by stating the features as a classical word lattice model. Finally, the post processor gives to the chunk head a polarity value which will be directly proportional to the chunk head's resultant polarity domain.

Now, presently research is done in the advancement of present system in the way of progressive methods for formation of opinions based on their polarity class.

3. Conclusion

This paper represents a lexical unison based measurement of sentiment intensity and polarity in text and its application in various fields. We are further working on how best to analyze the psychological effect of text by exploiting available resources and evaluating polarity of the text.

4. References

1. Tong, R.M. 2001. An operational system for detecting and tracking opinions in on-line discussions. *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (pp. 1-6). New York, NY: ACM.

2. Turney, P.D. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning* (pp. 491-502). Berlin: Springer-Verlag.

3. Hatzivassiloglou, V., & McKeown, K.R. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL* (pp. 174-181). New Brunswick, NJ: ACL.

4. Pang, B., and Lee, L. 2008. Opinion Mining and Sentiment Analysis. Now Publishers Inc.

5. Krosnick, J. A.; Judd, C. M.; and Wittenbrink, B. 2005. The measurement of attitudes. The Handbook of Attitudes 2176.

6. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series : Brendan O'Connory, Ramnath Balasubramanyany, Bryan R. Routledgex, Noah A. Smithy.

7. Amitava Das and Sivaji Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009).

8. Peter Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceeding of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics.

9. Sentiment Classification of Reviews Using SentiWordNet: Bruno Ohana , Brendan Tierney.

10. Synthesio- The Truth About Natural Language Processing, March 2011.

11. Agirre, E., Soroa, A., (2007). Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007),* 7-12.

12. OPINION POLARITY DETECTION: *Using Word Sense Disambiguation to Determine the Polarity of Opinions-* Tamara Martín-Wanton, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, Alexandra Balahur.

13. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79.86

Aparna Trivedi, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Apurva Srivastava, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Ingita Singh, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Karishma Singh, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Suneet Kumar Gupta, currently working as Associate Professor at ABES Engineering College, Ghaziabad and has many years of experience in teaching and research. Currently working on Natural language Processing and Information Retrieval.