ISSN 2319-2720

Volume 2, No.1, January – March 2013
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/2013/ijccn08212013.pdf

Analysis and Implementation of Sentiment Classification Using Lexical POS Markers

Arvind Kumar Jain[1], Yogadhar Pandey [2]
[1]M.Tech, SIRT Bhopal (MP),India, arvindjain98@gmail.com
[2] SIRT Bhopal (MP), India, p_yogadhar@yahoo.co.in

ABSTRACT
Natural language processing has attracted many researchers as the amount of information available on the internet and other media is increasing exponentially. So we need computer interference to process that enormous data.NLP does exactly the same by converting the natural language into something to be understood by computer to process. Though it has a lot of branches I am working in sentiment analysis of text. By this we mean a method to find what the reader is thinking when he/she was writing a particular text. This helps in tracking possible suicidal case, terrorist attacks and also in finding the orientation of a particular product and changing ourselves according to customer reviews. Polarity analysis is a subset of sentiment analysis when we find the polarity or orientation of a particular word and thus finding the overall polarity of a sentence, this is useful in finding whether a particular review of a product/movie/hotel or anything else is good or not [1]. With the ever increasing number of websites available free to post reviews today's customer are smart to follow the reviews before purchasing a product. While many review sites, such as Epinions, CNet and Amazon, help reviewers quantify the positivity of their comments, sentiment classification can still play an important role in classifying documents that do not have explicit ratings.
Key words: analysis, negative, polarity, positive, Sentiment

INTRODUCTION

The recent success of data-driven approaches in NLP has raised important questions as to what role linguistics must now seek to play in further advancing the field. Perhaps, it is also time to pose the same question from the other direction: As to how NLP techniques can help linguists make informed decisions? And how can the advances made in one field be applied to the other? Although, there has been some work on incorporating NLP techniques for linguistic field-work and language documentation (Bird, 2009), the wider use of NLP in linguistic studies is still fairly limited. However, it is possible to deepen the engagement between the two fields in a number of possible areas, and gain new insights even during the formulation of linguistic theories and frame-works [1, 2]. Languages can be represented by their two principal components, a lexicon and a grammar. A language's lexicon contains the legal combination of letters or symbols which represent a meaning. Lexicons are not static; new words are added with new meanings, old words may change or take on a new meaning, and words may become obsolete and eventually work their way out of the lexicon altogether. The basic unit of a lexicon is a "lexeme", a sequence of one or more words. Lexemes have syntactic properties, which define their structural relation to the grammar of the language, and semantic properties which define the 'meaning' of the lexeme symbolically[2]. It has been shown that syntax and semantics are closely related in technical and scientific sublanguages. The way in which a word combines syntactically and semantically with other words can be used as a basis for a classification system.

Sentiment Lexicons

The semantic orientation of a term indicates its capacity for carrying positive or negative evaluative value. It is possible for this information to exist a priori with relative independence from the context it may appear, as seen on words such as "excellent" or "terrible". For this reason knowledge of such terms can be a useful when identifying sentiment and is a motivation for the development of collections of opinionated terms into a sentiment lexicon. Sentiment lexicons exist as manually annotated databases such as the General Enquirer mapping terms in the English language into semantic categories, including sentiment orientation. Initially compiled to assist research on social studies, it has proven useful on opinion mining research and is regarded as a highly accurate lexicon used as a baseline for comparisons [3]. Other ad-hoc manual resources were generated for specific research. However the collection and annotation of a large sized lexicon is a expensive and time consuming task and has motivated research in automated methods that leverage existing language resources to build or expand existing lexicons.

Sentiment analysis

The task of identifying positive and negative opinions, emotions and evaluations. Most work on

sentiment analysis has been done at the document level, for example distinguishing positive from negative reviews [4]. Sentiment analysis for the last few years. It has attracted a great deal of attention because of its challenging research problems and the wide range of applications for both academia and industry. It needs a computational study for extracting knowledge from the people's opinions, appraisals and emotions toward entities, events and their attributes. In today's international global world market and highly growing internet usage, people prefer online shopping, banking, ticket reservation, hotel booking, etc[13,18]. so sentiment analysis from online customer reviews is becoming a requirement of an organization, customer and also manufacturer. The existing work on sentiment analysis can be categorized into document, sentence and word/feature level classification. Word or feature level sentiment analysis gets much importance by applying the natural level processing and statistical methods. Several researchers have worked on extraction of features and opinion-oriented words using a predefined seed word list for extracting semantic orientation and opinion classification. However, tasks such as multi-perspective question answering and summarization, opinion-oriented information extraction, and mining product reviews require sentence-level or even phrase-level sentiment analysis.

The problem of sentiment analysis

The research in the field started with sentiment and subjectivity classification, which treated the problem as a text classification problem. Sentiment classification classifies whether an opinionated document (e.g., product reviews) or sentence expresses a positive or negative however, require more detailed analysis because the user often wants to know what the opinions have been expressed on. For example, from the review of a product, one wants to know what features of the product have been praised and criticized by consumers As for any scientific problem, before solving it we need to define or to formalize the problem[10]. The formulation will introduce the basic definitions, core concepts and issues, sub-problems and target objectives. It also serves as a common framework to unify different research directions. From an application point of view, it tells practitioners what the main tasks are, their inputs and outputs, and how the resulting outputs may be used in practice Opinion. Subjectivity classification determines whether a sentence is subjective or objective many real-life applications.

Decision Trees

Recall that a text can be represented as a binary vector, each position corresponding to an attribute. In this simplified example, we will consider instances represented as 5-dimensional vectors. The first component will be equal to 1 if the text contains the word \wasted", and will be equal to 0 otherwise. The other 4 components depend on the presence/absence of the words \worst", \bad", \awful", \stupid". Thus, the text \This is the worst movie I have ever seen. My time was wasted" can be represented as the vector (1; 1; 0; 0; 0). The positive text \I was expecting this movie to be awful, but it turned out not to be bad at all" is represented as the vector (0; 0; 1; 1; 0). One of the classifiers that I apply to financial news is a decision tree[11]. Figure below is an example of a decision tree obtained when learning to classify movie reviews based on the five above words. To classify an instance, we start at the root. The root attribute in this case is the presence/absence of the word \wasted". Instances where this word is absent are classified by following the left branch of the tree. Instances where this word is present are classified by following the right branch of the tree. In this tree, the left branch leads to another rule, while the right branch leads to an immediate classification: the movie review is negative[7]. Once we follow enough branches, we get to a terminal node, which is called a leaf. Each leaf specifies how the text should be classified. The numbers printed in the leaf say how many texts that reached that leaf were classified correctly and how many incorrectly.

If we try to classify the instance

"This is the worst movie I have ever seen. My time was wasted,"

We start at the root and follow the right branch. This leads to an immediate negative classification, which seems accurate in this case. 77 texts that followed this path were classified correctly. 7 were classified incorrectly.

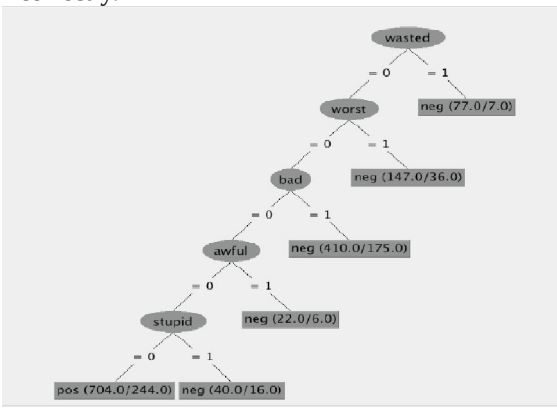

Figure 1: A Decision tree example

Polarity

The term polarity has a number of different uses, but in this dissertation it is used primarily to refer to the

positive or negative sentiment being expressed by a word. However, there is an important distinction between the prior polarity of a word and its contextual polarity. [2],[1] The prior polarity of a word refers to whether a word typically evokes something positive or something negative when taken out of context. For example, the word beautiful has a positive prior polarity, and the word horrid has a negative prior polarity.

Polarity Influencers

Phrase-level sentiment analysis is not a simple problem. Many things besides negation can influence contextual polarity, and even negation is not always straightforward. Negation may be local (e.g., not good), or involve longer-distance dependencies such as the negation of the proposition (e.g., does not look very good) or the negation of the subject (e.g., no one thinks that it's good). In addition, certain phrases that contain negation words intensify rather than change polarity (e.g., not only good but amazing)[14,17]. Contextual polarity may also be influenced by modality (e.g., whether the proposition is asserted to be real (realis) or not real (irrealis) – no reason at all to believe is irrealis, for example); word sense (e.g., Environmental Trust versus He has won the people's trust); the syntactic role of a word in the sentence (e.g., whether the word is in the subject or objective of a copular verb, consider polluters are versus they are polluters); and diminishers such as little (e.g., little truth, little threat).

Proposed Technique

In this research work we try to better evaluate the polarity of a sentence by the use of "Sentiwordnet",a lexical resource for sentiment analysis. We take the help of Stanford POS (part of speech) tagger to tokenize our sentence. We then select only that part of speech which could affect the polarity of a sentence or in other words polar words. We then propose our algorithm to find the overall polarity of the sentence also including those words which could improve, inverse or decrease the polarity of the corresponding word.

Sentiworldnet: sentiwordnet is the lexical resource for opinion mining. In SentiWordNet (http://sentiwordnet.isti.cnr.it/), to each synset of WordNet, a triple of polarity scores is assigned i.e., a positivity, negativity and objectivity score. and "neutrality". Each synset s is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative, and "objective" (i.e., neutral) the terms contained in the synset are. Different senses of the same term may thus have different opinion-related properties. For example the triple {0, 1, 0} (positivity, negativity, objectivity) is assigned to the synset of the term "bad". The sum of all scores of this synset is 1. SENTIWORDNET 1.0 the synset [estimable(J,3)] corresponding to the sense "may be computed or estimated" of the adjective estimable, has an Obj score of 1:0 (and P os and Neg scores of 0.0), while the synset [estimable(J,1)] corresponding to the sense "deserving of respect or high regard" has a P os score of 0:75, a Neg score of 0:0, and an Obj score of 0:25. Each of the three scores ranges in the interval [0:0; 1:0], and their sum is 1:0 for each synset [13]. This means that a synset may have nonzero scores for all the three categories, which would indicate that the corresponding terms have, in the sense indicated by the synset, each of the three opinions related properties to a certain degree. Table 1 and Table 2 below shows the POS tags related to sentiwordnet.

Table 1: POS type

S.NO	First Word	Second Word	Third Word
1	JJ	NN or NNS	Anything
2	RB,RBR, or RBS	JJ	not NN n or NNS
3	JJ	JJ	not NN n or NNS
4	NN or NNS	JJ	not NN n or NNS
5	RB,RBR, or RBS	VB,VBD,VBN, or VBG	Anything

Table 2: Tag set Description

S.NO	Tag	Description
1	NNP	Noun, proper, singular
2	NNPS	Noun, common , plural
3	RB	Adverb
4	JJ	Adjective or numeral, ordinal
5	JJR	Adjective ,superlative
6	NN	Noun, common, singular
7	RBR	Adverb, comparative
8	VB	Verb, base, form
9	VBD	Verb, present participle, or gerund
10	WDT	WH-determiner
11	CC	Conjunction, coordinating
12	CD	Numeral ,cardinal
13	DT	determiner

Algorithm Used

- Step 1 Create an input file "sample-input.txt" containing one or more than one sentences to check the polarities.

- Step 2 Read the file with a file reader object.
- Step 3 Parse each sentence token by token with the help of POS tagger.
- Step 4 POS tagger will assign a tag to each token.
- Step 5 Check the tag of each token if the tag is "JJ" or "JJS" (i.e the tagged token is an adjective/opinion word) then pass this word in SentiWordnet to check the score as well as polarity of that particular word.
- Step 6 SentiWordnet will return the sentiment-type of that word (eg Positive, Weak_Positive, Strong_Positive, negative, strong_negative, Weak _negative,neutral on score)
- Step 7 Count the no of positive(pos_count) and no of negative (neg_count) adjectives for each sentence.
- Step 8 If the neg_count is an ODD number then the sentence is considered "Negative" as a whole.(-)+(+)=(-) else goto step 9.
- Step 9 If the neg_count is an EVEN number(consider zero as even) then the sentence is considered "positive" as a whole.(-)+(-)=(+) or (+)+(+)=(+).
- Step 10 End

Conclusion

We have proposed a sentiwordnet based algorithm to more efficiently find the polarity of a given sentence. The use of part of speech tagger to tag words and search only those words with polarity (adjectives and adverbs) has increased the performance and we don't need to remove stop words. Sentiwordnet on the other hand is the main tool for calculating the score of a particular word in our sentence, besides that we also have to consider those words that in a way effect the polarity(inverse or enhance) of a particular. We have added in our algorithm these words. The algorithm perform quite good on normal input sentences chosen at random We evaluate our work few types of customer review datasets. From the results, it is clear that the proposed system achieved an average accuracy of 69.1% at the sentence level.

Future work for this can be an addition of module to check for spelling mistakes beforehand, also the sentiwordnet itself is not enough. Many words are not listed in the database so we need a new lexical resource or create or own as an extension to this. Online reviews also adds smiley's to them which is a language of its own. A single smiley can easily convey the polarity of the whole sentence without bothering, so we can also add a module for inclusion of smiley evaluation.

REFERENCES

[1] Aurangzeb Khan, Baharum Baharudin, and Khairullah Khan- **Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure**- J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 317–331, 2011.© Springer-Verlag Berlin Heidelberg 2011

[2] Yan Dang, Yulei Zhang, and Hsinchun Chen, University of Arizona- **A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews** -Published by the IEEE Computer Society july/august 2010

[3] Liddy, E. D.In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.- **Natural Language Processing**

[4] N. Indurkhya and F. J. Damerau- **Sentiment Analysis and Subjectivity**-Handbook of Natural Language Processing, Second Edition 2010

[5] Dongjoo Lee, Ok-Ran Jeong, Sang-goo Lee- **Opinion Mining of Customer Feedback Data on the Web.**

[6] TheresaWilson, JanyceWiebe, Paul Hoffmann-**Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis**-Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005. c 2005 Association for Computational Linguistics

[7] Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka-**SentiFul: Generating a Reliable Lexicon for Sentiment Analysis**-978-1-4244-4799-2©2009 IEEE

[8] Bing Liu -**Sentiment Analysis: A Multi-Faceted Problem**-To appear in IEEE Intelligent Systems, 2010.

[9] Yejin Choi and Claire Cardie -**Learning with Compositional Semantics as Structural Inference for Sub sentential Sentiment Analysis**-Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 793–801, Honolulu, October 2008. c 2008 Association for Computational Linguistics

[10] Min Wang, Hanxiao Shi-**Research on Sentiment Analysis Technology and Polarity Computation of Sentiment words**-978-1-4244-6789-1110/$26.00 ©2010 IEEE

[11] Ryan McDonald_ Kerry Hannan Tyler Neylon MikeWells Jeff Reynar-**Structured Models for Fine-to-Coarse Sentiment Analysis** -Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 432–439,Prague, Czech Republic, June 2007. c 2007 **Association for Computational Linguistics**

[12] Jun Zhao, Kang Liu, Gen Wang-Adding Redundant Features for **CRFs-based Sentence Sentiment**-Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 117–126,Honolulu, October 2008. c 2008 Association for Computational Linguistics

[13] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang and Zheng Chen-**Cross-Domain Sentiment Classification via Spectral Feature Alignment**-WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA. ACM 978-1-60558-799-8/10/04.

[14] TheresaWilson, JanyceWiebe, Paul Hoffmann-**Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis**-Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005. c 2005 Association for Computational Linguistics

[15] Jeonghee ,Yi Tetsuya Nasukawa, Razvan Bunescu , Wayne Niblack-**Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques**-Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) 0-7695-1978-4/03 © 2003 IEEE

[16] Xiaojun Li, Shanshan He, Hanxiao Shi-**Construction and Quantization for a Basic Sentiment Lexicon**-978-1-61284-181-6/11/$26.00 ©2011 IEEE

[17] Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka-**SentiFul: Generating a Reliable Lexicon for Sentiment Analysis**-978-1-4244-4799-2©2009 IEEE.