Serendio: Simple and Practical lexicon based approach to Sentiment Analysis

Prabu Palanisamy, Vineet Yadav and Harsha Elchuri
Serendio Software Pvt Ltd
Guindy, Chennai 600032, India
{prabu, vineet, harsha}@serendio.com

Abstract

This paper describes the system developed by the Serendio team for the SemEval-2013 Task 2 competition (Task A). We use a lexicon based approach for discovering sentiments. Our lexicon is built from the Serendio taxonomy. The Serendio taxonomy consists of positive, negative, negation, stop words and phrases. A typical tweet contains word variations, emoticons, hashtags etc. We use preprocessing steps such as stemming, emoticon detection and normalization, exaggerated word shortening and hashtag detection. After the preprocessing, the lexicon-based system classifies the tweets as positive or negative based on the contextual sentiment orientation of the words. Our system yields an F-score of 0.8004 on the test dataset.

1 Introduction

Social media websites like Twitter, Facebook etc. are a major hub for users to express their opinions online. On these social media sites, users post comments and opinions on various topics. Hence these sites become rich sources of information to mine for opinions and analyze user behavior and provide insights for:

- User behavior
- Product feedback
- User intentions
- Lead generation

Businesses spend an enormous amount of time and money to understand their customer opinions about their products and services. Thus Sentiment Analysis has become a hot research area since 2002. Sentiment Analysis is used to determine sentiments, emotions and attitudes of the user. The text used for analysis can range from big document (e.g. Product reviews from Amazon, blogs) to small status message (e.g. Tweets, Facebook comments). In this paper, we confine to Twitter data i.e classify a tweet to have a positive, negative or neutral sentiment.

The rest of the paper is organized as follows. In Section 2, we study relevant previous work on Sentiment Analysis on Twitter data. In Section 3, we describe each processing step of our approach in detail. In Section 4, we experiment with the training and the lexicon. In Section 5, we report and evaluate the final result obtained from the test data published by the SemEval team. In Section 6, we present our conclusions and outline our future work.

2 Related Work

Sentiment Analysis on raw text is a well known problem. The Liu (2012) book covers the entire field of Sentiment Analysis. Sentiment Analysis can be done using Machine learning or a Lexicon-based approach. We use our lexicon based approach in our study. The rest of the paper is confined to Lexicon based approach

2.1 Lexicon based approach

The lexicon based approach is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase. Turney (2002) identifies sentiments based on the semantic orientation of reviews. (Taboada et al., 2011; Melville et al., 2011; Ding et al., 2008) use lexicon based approach to extract sentiments.

Sentiment Analysis on microblogs is more challenging compared to longer discourses like reviews. Major challenges for microblog sentiment analysis are short length status message, informal words, word shortening, spelling variation and emoticons. Sentiment Analysis on Twitter data have been researched by (Bifet and Frank, 2010; Bermingham and Smeaton, 2010; Pak and Paroubek, 2010). We use our lexicon based approach to extract sentiments. The open lexicon such as Sentiwordnet (Esuli and Sebastiani, 2006; Baccianella et al., 2010), Q-wordnet (Agerri and García-Serrano, 2010), WordNet-Affect (Strapparava and Valitutti, 2004) are developed for supporting Sentiment Analysis. Studies have been made on preprocessing tweets. Han and Baldwin (2011) used a classifier to detect word variation and match the related word. Kaufmann and Kalita (2010) gives the full preprocessing approach to convert a tweet to normal text. Sentiment Analysis on Twitter data is not confined to raw text. Analyzing Emoticons have been an interesting study. Go et al. (2009) used emoticons to classify the tweets as positive or negative and train standard classifiers such as Naive Bayes, Maximum Entropy, and Support Vector Machines. Hashtag may have some sentiment in it. Davidov et al. (2010) used 50 hashtags and 15 emoticons as sentiment labels for classification to allow diverse sentiment types for the tweet. Negation and intensifier play an important role in Sentiment Analysis. Negation word can reverse the polarity, where as intensifier increases sentiment strength. Taboada et al. (2011) studied role of the intensifier and negation in the lexicon based Sentiment Analysis. Wiegand et al. (2010) survey the role of negation in Sentiment Analysis.

3 Serendio Approach

Serendio sentiment engine extracts and analyzes sentiments for a given product and feature set. Serendio sentiment engine currently works for eight different domains such as banking, tablets, smartphones, televisions, apparel, gaming, automobiles and e-readers. In this section, we will introduce Serendio's Sentiment engine and the enhancements that were made to handle the SemEval Task 2, Task A - Contextual Polarity Disambiguation (Wilson et al., 2013). The main steps of our approach are explained in detail in the subsections.

3.1 Creation of lexicon

The lexicon can be created either manually (Taboada et al., 2011; Tong et al., 2001) or expanding automatically from a seed of words (Kanayama et al., 2006; Kaji and Kitsuregawa, 2007; Turney, 2002; Turney and Littman, 2003). In our study, the lexicon is manually created. It is a one time process. Two types of lexicons are created.

Common lexicon: This contains data that would have the same semantic meaning or sense across different domains and categories.

- **Common or default sentiment words**. Positive and Negative sentiment words that have the same sentiment value or sense across different domains. For e.g. sentiment word "good" always represents a positive sentiment and it is independent of any category. Positive or Negative sentiment words have a sentiment score of +1 or -1 to indicate the respective polarity.

- **Negation Words**. Negation words are the words which reverse the polarity of sentiment. For example, "The battery life is not good" has negative sentiment

- **Blind Negation Words**. In the sentence, "The T.V needs a better remote", "needs" is a blind negation word. Blind negation words operate at a sentence level and points out the absence or presence of some sense that is not desired in a product feature.

- **Split words**. Split words are the words used for splitting sentences into clauses. The split words list consists of conjunctions and punctuation marks. For example the complex sentence, "Camera is good but the battery is bad" is split into two clauses "Camera is good" and "Battery is bad".

Category specific lexicon: Category specific lexicon contains the (1) Product Catalog which identifies all the products that we are interested in. (2) Feature Catalog which is a list of attributes that the product has. This enables the Serendio engine to do analysis at the feature level. (3) Sentiment words (positive and negative) that are specific to the category. For example, for a category such as Televisions, a product would be Samsung TV. The feature would be LCD screen and the word "glare" would be the category specific negative sentiment word.

The semeval task 2 contains Twitter data that cannot be pinned to any specific category. So for this task, only the common lexicon was used.

3.2 Preprocessing

A typical tweet contains word variations, emoticons, hashtags etc. The objective of the preprocessing step is to normalize the text into an appropriate form to extract the sentiments. Below are the preprocessing steps used.

- **POS Tagging**. POS Tagger gives part of speech tag associated with words. POS tagging is done using NLTK (Bird, 2006).

- **Stemming**. Stemmer gives the stem word. Serendio lexicon contains stem words only. So non stem words are stemmed and replaced with stem words. For example, words like 'loved', 'loves', 'loving', 'love' are replaced with 'lov'. This would aid the engine to do the word match from the text to the lexicon. Stemming is done using NLTK

- **Exaggerated word shortening**. Words which have same letter more than two times and not present in the lexicon are reduced to the word with the repeating letter occurring just once (Kouloumpis et al., 2011). For example, the exaggerated word "NOOOOOO" is reduced to "NO".

- **Emoticon detection**. Emoticon has some sentiment associated with it. Twitter NLP (Ritter et. al , 2011; Ritter et. al , 2012) is used to extract emoticons along with the sentiments in the Twitter data.

- **Hashtag detection**. The hashtag is a topic or a keyword that is marked with a tweet. Hashtag is a phrase starting with # with no space between them. Hashtags are identified and sentiments are extracted from them.

3.3 Sentiment calculation

Sentiment calculation is the aggregation of the sum of the sentiment bearing entities of the tweet. Entities can be text, emoticons and hashtags. The sentiment calculation algorithm is shown in Algorithm 1. The sentiment calculation is based on a set of heuristics built on the sentiment orientation of the words. Blind negation words are extracted from the sentence. The presence of the blind negation words indicate negative sentiment. If the sentence contains a blind negation word then other steps are skipped and sentiment is blindly assigned as negative. Next, sentiment words are extracted. The sentiment polarity of the word can be changed due to negation words that occur in proximity (2 word distance). If a sentiment word is not present, then the sentiment negation word becomes additive to the negative sentiment list. The sentence "I can not deal it" has the negation word "not" and it does not contain a sentiment word. So the negation word just gets added to the negative sentiment word. Sentiments from emoticons are extracted with the help of Twitter NLP. Sentiment words within the hashtag are extracted by python regex functions. For example, from the hashtag "#ihateu", the word "hate" is extracted as a sentiment word. The sentiment of the tweet is aggregated as the sum of the sentiments from all the entities.

4 Experimental Data

The training data consist of real time tweets. 9451 subjective expressions are marked from all the tweets and are labeled as positive or negative or neutral. The average number of words of the marked subjective expression is around 2 to 3 words. The common dictionary that is constructed is shown in Table 2. The Serendio sentiment engine is run on the training data set. We identify the correct sentiment of the the phrases which are misclassified as neutral, we include the phrases in our lexicon with their appropriate sentiments.

Algorithm 1: Sentiment Calculation

Data: Preprocessed Twitter data
Result: Output: Positive, Negative, Neutral
Find the list of sentiment words $SentiList$, its position in the sentence;
Find the list of sentiment negation words $SentiNegat$, its position in the sentence;
Find the list of blind negation words $BlindNegat$, its position in the sentence;
if $BlindNegat$ **then**
    return negativity;
else
    if $SentiList$ *and* $SentiNegat$ **then**
        foreach *word in the* $SentiList$ **do**
            if *word is atmost the distance of 2 from* $SentiNegat$ **then**
                Revert the polarity of the word;
            end
        end
    else
        if $SentiNegat$ **then**
            Add the $SentiNegat$ to the negative $SentiList$;
        end
    end
end
$SentiSum$=0;
foreach *word in the* $SentiList$ **do**
    $SentiSum$=$SentiSum$+sentiment of $word$;
end
if $Hashtag$ *is present* **then**
    Find all the sentiment words in hash tag using regex matching and add them to $SentiList$
end
if $Emoticon$ *is present* **then**
    Find sentiment of the emoticon and add emoticon,it's sentiment to $SentiList$
end
$SentiType$="neutral";
if $SentiSum > 0$ **then**
    $SentiType$="positive";
end
if $SentiSum < 0$ **then**
    $SentiType$="negative";
end
return $SentiType$;

Table 1: Training Data

Sentiment type	Expression count
Positive	5865
Negative	3120
Neutral	466

Table 2: Lexicon Details

Data type	Count
Blind Negation word	7
Negation	13
Positive sentiment word	1260
Negative sentiment word	1703
Split word	16

5 Result and Discussion

Our sentiment engine performed reasonably well. Please see Table 3 for Precision and Recall measurements. The recall rates are lower because of our lexicons lack of coverage of all the sentiment words. Informal language of tweets posed another challenge for identifying negative sentiments. For example, swear words such as "sh*t" and "f**k" are generally considered as negative sentiment words. Phrases such as "This sh*t is good" and "F**king awesome" were identified as negative sentiments when in fact they were expressing positive sentiments.

Table 3: Results

	POSITIVE	NEGATIVE
PRECISION	0.9361	0.8884
RECALL	0.7132	0.7912

The Serendio lexicon that we used has sentiment words with a sentiment attached to it. By integrating with a lexical source such as Sentiwordnet, we feel we could get a more nuanced word sense disambiguation. For example, the word "good" is considered to have positive polarity. According to Sentiwordnet 3.0, good as an adjective has 21 different senses with different sentiments. For example, the sentiment word "good" in the phrase "A good mile from here" gives an objective sense, not in a positive sense.

6 Conclusion

In this paper we presented our system that we used for the SemEval-2013 Task 2 for doing Sentiment Analysis for Twitter data. We got an F-score of 0.8004 on the test data set.

We presented a lexicon based method for Sentiment Analysis with Twitter data. We provided practical approaches to identifying and extracting sentiments from emoticons and hashtags. We also provided a method to convert non-grammatical words to grammatical words and normalize non-root to root words to extract sentiments.

A lexicon based approach is a simple, viable and practical approach to Sentiment Analysis of Twitter data without a need for training. A Lexicon based approach is as good as the lexicon it uses. To achieve better results, word sense disambiguation should be combined with the existing lexicon approach.

7 Acknowledgments

We would like to thank the organizers of SemEval 2013. We also would like to express our gratitude to the various reviewers for their encouragement and positive feedback.

References

Rodrigo Agerri and Ana García-Serrano. 2010. Q-WordNet: Extracting polarity from WordNet senses. *Seventh Conference on International Language Resources and Evaluation, Malta*.

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC10), Valletta, Malta, May*.

Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? *Proceedings of the 19th ACM international conference on Information and knowledge management* 1833–1836, ACM.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. *Discovery Science* 1–14, Springer.

Steven Bird. 2006. NLTK: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions* 69–72, Association for Computational Linguistics.

Kushal Dave, Steve Lawrence and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web* 519–528, ACM.

Dmitry Davidov, Oren Tsur and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. *Proceedings of the 23rd International Conference on Computational Linguistics* 241 –249, Association for Computational Linguistics.

Xiaowen Ding, Bing Liu and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. *Proceedings of the international conference on Web search and web data mining* 231–240, ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC Volume 6*, 417–422.

Alec Go, Richa Bhayani and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1–12.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Volume 1* , 368–378.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. *Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* 1075–1083, Association for Computational Linguistics.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* 355–363,Association for Computational Linguistics.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of Twitter messages. *International Conference on Natural Language Processing* Kharagpur, India.

Efthymios Kouloumpis, Theresa Wilson and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* 538–541.

Bing Liu. 2008 Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 1–167,Morgan & Claypool Publishers.

Prem Melville, Wojciech Gryc and Richard D Lawrence. 2011. Sentiment analysis of blogs by combining lexical knowledge with text classification. *Proceedings* *of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1275–1284, ACM.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*, Volume 2010.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval Volume 2 number 1-2*, 1–135, Now Publishers Inc.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing Volume 10*, 79–86, Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* 25–32, Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. *EMNLP*.

Alan Ritter, Mausam, Oren Etzioni, Sam Clark. 2012. Open Domain Event Extraction from Twitter. *KDD*.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. *Proceedings of LREC* 1083–1086.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, volume 37, number2, 267–307, MIT Press.

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

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21(4):315–346.

Peter Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics*, 417–424, Association for Computational Linguistics.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. *Proceedings of the workshop on negation and speculation in natural language processing* 60–68, Association for Computational Linguistics.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Alan Ritter, Sara Rosenthal and Veselin Stoyonov. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Proceedings of the 7th International Workshop on Semantic Evaluation* Association for Computation Linguistics.