An Approach to Convert Non-Grammatical Words to Grammatical Words to Extract Sentiments using Lexicon Analysis

Harmeet kaur, Abhishek Tyagi

Abstract— we use a lexicon based approach for discovering sentiments. Our lexicon is built from tokenization taxonomy consists of positive, negative, neutral phrases. A typical tweet contains word variations, emoticons, hash tags etc. We use preprocessing steps such as stemming, emoticon detection and normalization, exaggerated word shortening and hash tag detection.

Index Terms—hash tags, analysis, emotions, Sentiwordnet, lexicon, stemming, token.

1 System Introduction

Sentiment Analysis is the study of opinions, attitude, and emotions gathered from the people to extract an entity identification .sentiment analysis is also known as opinion mining. The entity can be the form of individual, event or sentence or phrase along with their grammatical meaning in a normalized form. They express a mutual meaning or kind of same meaning words. Some of the experts predicted that sentiment analysis and opinion mining have slightly different notion during representation. Opinion mining can extracts and analyses human opinion in form of text or through words from mouth about a phrases while Sentiment Analysis identifies the sentiment expressed in a text then analysis it and through these emotions also being calculated in the form of negative and positive. Sentimental analysis basically, identify the sentiments they express in the form of sentences they use & expression through their emotions, and then classify their polarity as shown in Fig. 1 [1]

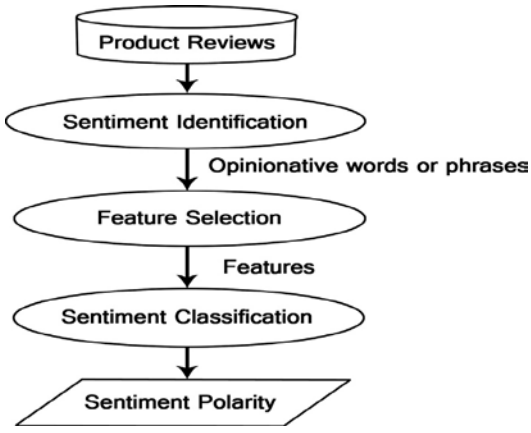


Figure 1: Sentiment analysis process on product reviews

Sentiment analysis is divided into three main classification levels stated as: document level, sentence level, and aspect level sentiment analysis. The main goal of Document level whose goal is to classify an opinion document present in form of text or phrase as expressing a positive or negative opinion or sentiment by finding its root. It considers the whole document a basic information unit in one sentence. Sentence-level SA expects to group assumption communicated in each one sentence to inquiry linguistic uses. On the other hand, there is no basic contrast in the middle of report and sentence level classification in light of the fact that sentences are somewhat short reports. Ordering content at the archive level or at the sentence level does not give the important subtle element required conclusions while substance which is required in numerous applications, to get these points of interest and exact result; we have to go to the subtle element level to figure out its Aspects. Viewpoint in Sentiments expects to arrange the assumption concerning the specific parts of individual elements for discover feeling. The first step is to recognize the substances and their angles and their points. The assessment holders can give distinctive feelings for diverse parts of the same element like this sentence or having same importance "The voice nature of this telephone is bad, however the battery life is excellent' 'sentiment investigation is the undertaking of recognizing positive and negative.

2 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 micro blogs is more challenging compared to longer discourses like reviews. Major challenges for micro blog sentiment analysis are short length status message, informal words, word shortening, spelling variation and emoticons. 2010). We use our lexicon based approach to extract sentiments. The open lexicon such as Sentiwordnet- net (Esuli and Sebastiani, 2006; Baccianella et

- *Er.Harmeet kaur is currently pursuing masters degree program in Inforamtion tchnologin lovely professional University, India, PH-. E-mail:08874009343. harmeet1308@gmail.com*
- *Er.Abhishek Tyagi is currently working in computer science deptt in Lovelyprofessional University, India, PH-09781778027. E-mail: abhishek-tyagi43@gmail.com*

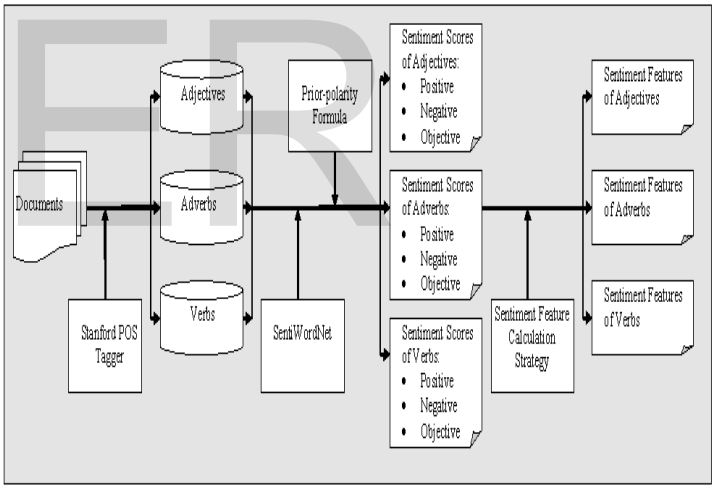
al., 2010), Q-Word Net (Agerri and Garc´ıa-Serrano,2010), Word Net-Affect (Strapparava and Valitutti,2004) are developed for supporting Sentiment Analysis. Studies have been made on preprocessing tweets. . Analyzing Emoticons have been an interesting study. Go et al. (2009) used emoticons to classify the tweets as positive or negative and train standard classifiers such as Naive Bayes,
Maximum Entropy and Support Vector Machines. Hash tag may have some sentiment in it. Davidov et al. (2010) used 50 hash tags 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 Tokenization Approach

Typically, tokenization occurs at the word level. However, it is sometimes difficult to define what is meant by a "word". Often a tokenizer relies on simple heuristics, for example: All contiguous strings of alphabetic characters are part of one token; likewise with numbers. Tokens are separated by whitespace characters, such as a space or line break, or by punctuation characters. Punctuation and whitespace may or may not be included in the resulting list of tokens.
In languages that use inter-word spaces (such as most that use the Latin alphabet, and most programming languages), this approach is fairly straightforward. However, even here there are many edge cases such as contractions, hyphenated words, emoticons, and larger constructs such as URIs (which for some purposes may count as single tokens). A classic example is "New York-based", which a naive tokenizer may break at the space even though the better break is (arguably) at the hyphen.
Some ways to address the more difficult problems include developing more complex heuristics, querying a table of common special-cases, or fitting the tokens to a language model that identifies collocations in a later processing step.

3.1 Creation of lexicon

The lexicon can be created either manually (Taboada et al., 2011; Tong et al., 2001) or expand- ing 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 onetime pass 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 lex icon 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 is 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 several task 2 contains facebook data that can- not 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 are.
• **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 hash tag is a topic or a keyword that is marked with a tweet. Hash- tag is a phrase starting with # with no space between them. Hash tags are identified and

Word	POS*	Sentiment Score	Polarity
Nice	JJ	0.81	Positive
Awesome	JJ	0.75	Positive
Perfect	JJ	0.75	Positive
Best	JJS	0.69	Positive
Amazing	JJ	0.63	Positive
Beautifully	RB	0.63	Positive
Excellent	JJ	0.63	Positive
Unusable	JJ	-0.81	Negative
Terrible	JJ	-0.78	Negative
Worst	JJS	-0.75	Negative

Sentiment Feature Calculation Strategy:

1. Obtain the data from the social network thread **Tr**
2. Extract the list of users **U** from the social networking thread
3. Extract **N** number of Message **M** using dictionary based tokenization
4. Filter message content with STOPWORD list of common English words while Tokenization
5. Load negative and positive sentiment expression word classification information file
6. Calculate word weight score to measure the sentiment **Sn**
7. Count the final sentiment score **S** of each message (Positive/Negative)
8. Calculate sentiment score for tokenized message number **N**
9. Find the sentiment type **St** (Positive/Negative) by validating the sentiment specific word dictionary
10. If **St >0**
    a. Mark the message as positive
    b. Add 1 to **posMsg**
11. If **St<=0**
    a. Mark the message as negative
    b. Add 1 to **negMsg**
    c. Load **Anger** sentiment expression word classification information file
    d. Load **Fear** sentiment expression word classification information file
    e. Calculate word weight score to measure the sentiment A**Sn** for Anger sentiment
    f. Calculate word weight score to measure the sentiment F**Sn** for Fear sentiment
    g. If **ASn>FSn & FSn <=1**
        i. Mark the message spreading Anger
        ii. Add 1 to **anger**
    h. If **ASn<FSn & ASn<=1**
        i. Mark the message spreading Fear
        ii. Add 1 to **fear**
    i. **Else**
        i. Mark the message as Mixed emotion
        ii. Add 1 to **Me**

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

4. 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. In- formal 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 tokenization 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 senseThe taxinomy 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.

5. Conclusion

In this paper we presented a lexicon based method for Sentiment Analysis with facebook data. We provided practical approaches to identifying and extracting sentiments from emoticons and hash tags. We also pro- vided 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.

6. REFERENCES

[1] A. h. K. Walaa Medhat, "Sentimental analysis algorithms and application:A survey," *Ain Shams Engineering Journal*, 2014.

[2] G. Carenini, R.Ng, and A.Pauls. Multi-document summarization of evaluative text. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[3] G. Carenini, R. Ng, and E. Zwart. Extracting knowledge from evaluative text. In Proceedings of the International Conference on Knowledge Capture, 2005.

[4] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005.

[5] M. Dredze, J. Blitzer, and F. Pereira. Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification. In Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL), 2007.

[6] P. K. A. M. Simko, "Sentimental Analysis on Microblog Utilizing appraisal Theory," in *World Wide Web @Springer Science + Buiness social media*, New York, 2014.

[7] k. P. a. H. Lim, "Acquiring lexical knowledge using raw corpora and unsupervised clustering method," in *cluster comput(2014) @springer sceince +business media newyork 2013*, New York, 2014.

[8] B. C. a. C. L. Azevedo, "A Sensitivity - Analysis - Based Approach for the Calibration Of Traffic Simulation Models," in *IEEE TRANSACTIONS ON iNTELLIGENT SYTEMS*, 2014.

[9] A. E. a. F. S. stefano BACCIANANELLA, "Star Track:The Next Generation (Of Product Reviw Management Tools)," in *new generation Computing 31(2013)47-70 Ohmsha Ltd and Springer*, Japan, 2013.

[10] K. L.-C. H.-C. L. a. C. -H. W. Hao-Chiang, "An emotion Recoginition mechanism based on the combination of mutal information and semantic cles," in *J Ambient Intell Human comuput @springer -verlag 2011*, verlag, 2012.

[11] K. H. M. N. A. R. a. J. R. Sasha Blair-Goldensohn, "Buliding a sentment summarizer for local Service Reviews," in *NLPIX 2008*, Beijing,china, 2008.

[12] A. E. a. F. Sebastianai, "Senti WORDNET: A publicly avaliable lexical Resource for opinion mining," in *5rd conference on language resources and evaluation*, Genova,IT, 2006.

[13] V. Y. a. H. E. PrabhU Planisamy, "Serendio:Simple and Practical Lexicon Based approach To sentiments Analysis," 2006.

[14] Z. K. P. N. A. R. S. R. a. V. S. Theresa Wilson, "Sentiment Analysis in Twitter," in *7th international Workshop on semantics Evalutaion for Computing Lingustics*, 2013.