Twitter Sentiment Analysis on E-commerce Websites in India

Devang Jhaveri
Student, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

Aunsh Chaudhari
Student, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

Lakshmi Kurup
Professor, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

ABSTRACT
In today's era, social media has become a valuable source of information, where people express their opinions. Analysis of such opinion-related data can provide productive insights. When these opinions are relevant to a company, accurate analysis can provide them with information like product quality, influencers affecting other customer decisions, early feedback on newly launched products, company news, trends and also knowledge about their competitors. Hence, harnessing and extracting insights from these sentiments is necessary for these companies to implement effective marketing strategies and better customer service. Carrying the same notion forward, we decided to extract sentiments from Twitter relevant to two e-commerce giants in India, Flipkart and Snapdeal. In this paper, various lexicon based approaches are applied and their accuracy is investigated.

General Terms
Natural Language Processing, E-commerce

Keywords
Twitter Analysis, Sentiment Analysis

1. INTRODUCTION
Google defines sentiment analysis as "the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral". In other words, sentiment analysis is a line of research that harnesses people's opinions and attitude in relation to different topics, products, events and attributes. It is an extension of data mining in the NLP (Natural Language Processing) domain. This concept involves studying each opinionated word or phrase in the text and labelling it as positive, negative or neutral.

Today, social media has become an integral part of our lives, where people express their thoughts and opinions about various things. Websites such as Twitter, Facebook, Instagram, Tumblr, and Myspace have become extremely popular.

Recently, on 24th August, 2015, Facebook made a record by counting ONE BILLION people accessing the website in a single day. That is, 1 in 7 people on Earth were connected at the same time. In comparison, the microblogging site, Twitter has a user base of 316 million active users. The content accumulation of so many opinions, comments are useful if extracted and analyzed in the correct way. Our pick for the experimental results was Twitter due to the following reasons:

•The amount of data available online – over 250 million messages per day.

•Real time opinionated messages.

•Efficient analysis of tweets since they are restricted to only 140 characters.

E-commerce (electronic commerce or EC) is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the Internet. [1] In 2013, Asia-Pacific emerged as the strongest business to consumer (B2C) e-commerce region in the world with sales of around 567.3 billion USD. E-commerce in India has grown the fastest in this region with sales touching almost 20 billion USD in 2015, with a growth rate of 700% since its inception in 2009 (2.5 billion). Amongst various e-commerce websites in India, Flipkart and its rival Snapdeal, enjoy a major chunk of the e-commerce market. According to a report by Morgan Stanley, Flipkart, founded as an online book retailer in 2007, tops the list with 44% followed by Snapdeal with 32%. Thanks to the backing of the abovementioned statistics, we decided to collect mentioned tweets related to two of these highly popular e-commerce companies.

The paper is structured as follows. In section II, we review previous research and study conducted in this domain. In section III, we delineate the procedure of collecting and cleaning our dataset extracted from Twitter. The different lexicon-based methods implemented in our experiment are elaborated in section IV. Finally, the results of our experiments are demonstrated in section V while, section VI, presents the conclusion as well as the future work that can be taken up.

2. RELATED WORK
Conventionally, there are two types of approaches taken towards sentiment analysis – the lexicon based as well as the machine learning techniques. All of the related work presented in this section has aided us in getting a better understanding of sentiment analytics.

As a starting point of comprehension, Michelle Annett and Grzegorz Kondrak's study poses as a demonstration of the comparison of different sentiment analysis methods [1]. This was achieved by applying them to an available set of movie reviews. After describing the different lexical and machine learning methods clearly, they proposed an approach based on Support Vector Machines. They report around 50% accuracy for the Baseline approach.

Using Twitter data for the prediction of the US Presidential election, Swathi Chandrasekar, Emmanuel Charon and Alexandre Ginet focused on building a training or testing set [2]. The importance of pre-processing is highlighted. In their case, all words were converted to lower case, punctuation marks were removed and commonly used words which do not

contribute to the sentiment of the tweet were eliminated before analysis. Using the SVM model, they report around 69% accuracy.

Extending the above work forward, Xing Fang and Justin Zhan performed experiments for both sentence level and review level categorization on data (product reviews). Sentiment sentence extraction and POS tagging is implemented [3]. The performance of each model is based on an average F1-score, that being 0.85 for manually labelled sentences. Andrea Esuli and Fabrizio Sebastiani take this research ahead through SENTIWORDNET, a freely available resource in which every word is associated to three numerical scores Objective, Positive and Negative [4].

3. DATA PRE-PROCESSING

3.1 Collection of Twitter Data

From 18th August, 2015 to 25th August, 2015 we collected a total of 7,906 tweets related to "Flipkart" whereas for "Snapdeal" we collected 6,054 tweets. The data gathered using Search API and Streaming API provided officially by Twitter. The Search API allows developers to look up tweets containing a specific word or a phrase. One of the constraints imposed by Twitter is that the Search API produces only 1500 tweets at a time. Hence, to gather more tweets we used to Streaming API which captures tweets in real time.

We used the R programming language to carry out our experiment. The "twitteR" package available for the R environment was used for extract the above mentioned tweets from Twitter.

3.2 Data Cleaning

Data cleaning is an important component of the data mining process. It involves recognition, removal of errors and inconsistency to improve the quality of the dataset prior to the process of analysis [9]. The tweets were cleaned of irrelevant data to improve their quality. In our case, we observed there were certain elements that did not provide any information and hence, had to be removed before processing. The elements were as follows:

1) Links: People generally have a tendency to attach documents (images, blogs, videos, web direction etc.) along with their tweets. These links or URLs had to be eliminated since they were of no use to our analysis.

2) Mentions: Mentions are used in Twitter to reply, acknowledge or start a conversation. Mentions are always written using "@" sign followed by the username. These mentions do not contain any relevant information thus, are removed.

3) Hashtags: A Hashtag ("#") is used to mark keywords or topics in a tweet. Using hashtags, people can search or start a new trend on Twitter. The "#" has been removed from all the tweets of our dataset.

4) Retweets: A retweet involves the re-posting of another user's tweet. It leads to redundancy in data and to avoid this, we eliminated all the retweets from our dataset.

5) Removing Punctuations and other miscellaneous data: Punctuations marks like quotes (""), commas (,) and semicolons (:) do not have any significant role in our analysis and hence, were removed from all the tweets present in our dataset.

After cleaning, there were few tweets in which all characters were eliminated for instance, those which contained only URLs or mentions. We deleted such blank rows from our database. Post the cleaning phase; the dataset included 1,668 tweets of Flipkart and 2,999 tweets of Snapdeal. The cleaned dataset was then passed for further analysis.

4. METHODS

Lexicon based approaches are quite popular in the sentiment analysis domain. These approaches involve tokenization of a particular corpus of text into unigrams, which are then assigned a polarity score [11]. The aggregated sum of these scores determines the sentiment behind the text. It is generally classified as positive, negative or neutral depending on the calculated score. The flowchart describing a general lexicon based approach of Twitter data is shown in fig.1

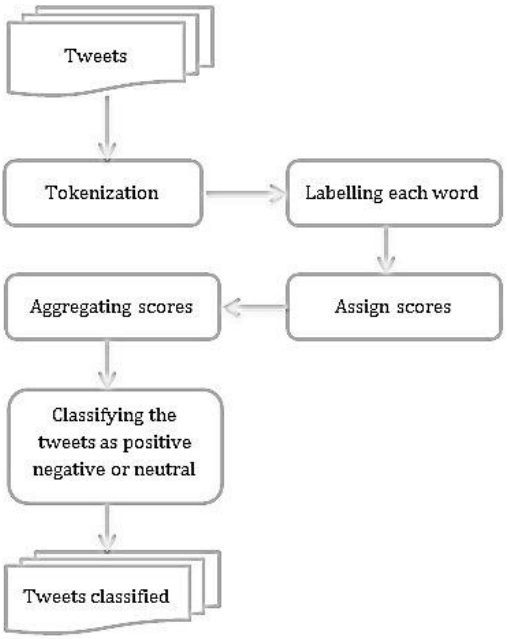

Figure 1. Flowchart of a general lexicon based approach

Our study considers three lexicon based approaches relevant to our domain. They are explained in detail in the following subsections:

3.1 Bag of Words

The simplest and most widely used lexicon based approach is the baseline approach (also called "Bag of Words Approach") [8]. In this method, there are two dictionaries – that of the positively tagged words and negatively tagged words. After tokenization, each individual word of the tweet is searched within those dictionaries, and depending upon the location of the word, it is assigned a polarity score.

Consider a tweet from our dataset: "Great things can be accomplished with ease when you have the best team in the world http://t.co/9xguAoLm4K".

At the end of preprocessing, the text ready for analysis is – "great things can be accomplished with ease when you have the best team in the world".

Following the technique explained above, each of the following words- "great", "accomplished", "ease" and "best" are given a sentiment score of +1 since they are present in the positive words dictionary. On aggregation, the total polarity score of +4 is obtained, indicating that the sentiment behind the tweet is positive.

1) Scoring: If the individual token is found in the positive words dictionary, it is assigned a +1 polarity score value, if present in the negative words dictionary, a score of -1 and lastly, if not present in any of them, a score of 0 is assigned.

2) Aggregation: The total sum of the scores of each word present in the text is calculated and the on the basis of the final polarity value, the tweet can be categorized as positive, neutral or negative.

Consider a tweet from our dataset: "Great things can be accomplished with ease when you have the best team in the world http://t.co/9xguAoLm4K".

At the end of preporocessing, the text ready for analysis is – "great things can be accomplished with ease when you have the best team in the world".

Following the technique explained above, each of the following words- "great", "accomplished", "ease" and "best" are given a sentiment score of +1 since they are present in the positive words dictionary. On aggregation, the total polarity score of +4 is obtained, indicating that the sentiment behind the tweet is positive.

Table 1. Score of the given tweet

Word:	great	accomplished	ease	best
Score:	+1	+1	+1	+1

Total score: +4

3.1 Modified Bag of Words using Afinn – 111

Afinn – 111 is a list of English words rated for valence with an integer between 5.0 (positive) and -5.0 (negative) [7]. The newest version contains 2477 words and phrases each having a corresponding weighted positive or negative score. A few words from the list are given in table 2.

Table 2. Few words and their weights from Afinn -111 dictionary

Abhors	-3
Abilities	2
Ability	2
Aboard	1
absentee	-1

The Afinn – 111 list is most effective for complex sentences. For example, a tweet that contains more than one positive word as well as a negative word having a heavier weight, it is observed that the negative word has a higher influence on the sentiment of the entire tweet.

3.3 POS tagging and scoring using SentiWordNet

One of the important features in a lexicon based approach is the use of a dictionary containing opinionated words or phrases. SentiWordNet is one such lexical resource which has made publicly available for research. SentiWordNet dictionary contains a total of 1, 17,660 inherited words from the WordNet 2.0 database and assigns positive and negative score to each word. WordNet 2.0 is a database developed by the Princeton University which contains thousands of words, organized according to their semantic meanings and classified by their synonyms in groups called synsets. This dictionary classifies all the words or synsets into four categories: Noun (n), Adjective (a), Verb (v) and Adverb (r). The SentiWordNet dictionary assigns three scores, positive, negative and objective to each synset, determining how much positive, negative or objective a particular synset is. The scores take up values between 0.0 and 1.0 and the sum of all three scores will always add up to 1.0. In other words, the dictionary assigns a positive and negative score to each synset and the objective score is always the complement of the sum of the positive and negative scores.

Objective score = 1 – (positive_score + negative_score)

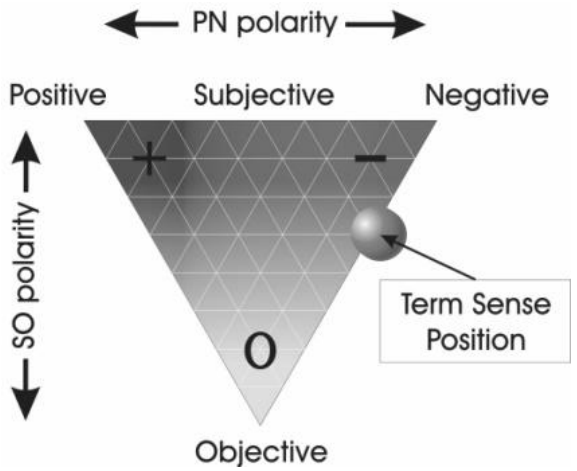

Figure 2. SentiWordNet polarity diagram

1) Classification Phase: Before assigning a score to each word in a particular tweet, the word had to be classified in one of the four mentioned categories (n, a, v, r). We used Part of Speech Tagging (POS Tagging) methodology for the same. POS Tagging is a process of marking each word in a text with an appropriate part of speech like adjective, adverb, verb etc. based on both its definition and position in the text corpus. POS tagging leads to a lot of ambiguity because there are certain words which can take up multiple tags depending on the context of the sentence. Consider two sentences:

1. "The management has refused to back our project."
2. "He lay on his back, starting at the fan."

In sentence 1, the word "back" is used as a verb whereas in sentence 2, the same word is used as a noun. Various techniques like frequency-based tagging, transformation-based tagging etc. are used to tackle this problem. Getting into the details of these techniques is beyond the scope of this paper.

We used the "openNPL" package in R, which uses machine learning techniques to tag each word in the corpus with considerable accuracy. It assigns one of the 36 tags mentioned in the Penn Treebank project to each word present in the corpus. We consider only the first alphabet from each label and classify it into one of the above mentioned categories (n, a, v, r). In other words, our study considers all types of nouns like common singular, common plural, proper singular and proper plural nouns (NN, NNS, NNP and NNPS) as one category that of - nouns (N) [10]. Similarly, the same procedure was applied to verbs, adjectives and adverbs. On the other hand, determiners, pronouns, interjections and other miscellaneous tags are ignored as they have no significant meaning in our analysis.

2) Scoring Phase: The SentiWordNet dictionary assigns a positive and negative score to each synset (array of similar words). The words present in these synsets have certain values assigned to them depending on the training data used to make the dictionary. We call these values as priority values. There are cases where the same word has different priority values in different synsets. For instance, the word short has a priority value of #3 in the synset1 (short#3, little#6) while it has a priority value of #4 in the synset2 (short#4 poor#5 inadequate#2). In such cases, we base our scoring by selecting the synset which contains the word with the least priority value. In the above mentioned example, we would select the score related to synset1 and ignore the scores concerning synset2.

The positive, negative and the objective score for each word in a tweet is calculated. Words with an objective score lesser than a pre-defined threshold value (between 0.0 and 1.0) are discarded, while the ones above the threshold value are added to get an aggregated positive and negative score of a tweet. Finally, a tweet is classified as negative, positive or neutral based on the dominating value.

$$s+ = \sum_{i=0 \; and \; obj_{score} < \alpha}^{n} \text{pos\_score(i)}$$

$$s+ = \sum_{i=0 \; and \; obj_{score} < \alpha}^{n} \text{pos\_score(i)}$$

Where, α is the threshold value.

$$tweet = \begin{cases} positive \; iff \; s+ > s- \\ negative \; iff \; s- > s+ \\ neutral \; otherwise \end{cases}$$

5. EXPERIMENTAL RESULTS

Table 3 shows the results of applying all the methods mentioned in section IV. We applied the same algorithm to all tweets mentioning both Flipkart and Snapdeal. The total accuracy considered in our study is an average of both. The Bag of Words approach gained an average accuracy of 57.65%. Further, we decided to add e- commerce specific words like "cheaper", "sale", "vouchers" etc. in the positive dictionary to improve precision. The addition of words resulted in an increase of accuracy by 3.45% which was comparable to the SentiWordNet approach (increase of 4.91%).

Surprisingly, the accuracy of the baseline approach using the Afinn – 111 drastically decreased by 13.25%. We believe that this drop is due to the fact that the Afinn – 111 dictionary has only 2,477 words or phrases compared to the one used in the Bag or Words approach (6,792 words including positive and negative). A refined analysis using a weighted dictionary containing more words or phrases should give interesting results.

Table 3. The accuracy of Various Lexicon Methods

Approach	Accuracy
Bag of Words	57.65%
Bag of Words + new Words	61.11%
BOW + Afinn – 111	44.44%
SentiWordNet	62.56%

We also noticed that the accuracy obtained after analysing Snapdeal tweets was always more than that of Flipkart tweets, irrespective of the method applied. A closer manual observation of these tweets revealed that Snapdeal tweets had more stereotypical words like "awesome", '"wow", "happy" etc., while Flipkart tweets had words like "contest", "feature", "launches" which have a positive meaning depending on their context.

Given the results, it appears that the accuracy of the lexical schemes depends heavily on the words present in the dictionary. If the dictionary is too sparse, the results might not be accurate even after using an effective scoring scheme. Our results are in accordance with previous findings - it is difficult to achieve accuracy greater than 65% using a lexicon based approach.

6. CONCLUSION AND FUTURE SCOPE

From the lexicon based experiment performed to evaluate the sentiment behind each tweet, we conclude that the SentiWordNet approach has the maximum accuracy relative to other lexicon methods. Further, we also discern the fact that the polarity score in all lexicon based methodologies are largely dependent on the dictionary selected for analysis, both in terms of quantity (number of words) as well as quality (type of words).

The analysis results obtained from Twitter data are susceptible to a topic trending over a specific period of time. For instance, our results were influenced by Snapdeal inaugurating their new office leading to a large number of tweets having an underlying positive sentiment. This susceptibility can be minimized by gathering Twitter data over a longer span of time.

The E – Commerce industry considers social media advertising as an integral parameter for progress. Due to this reason, our dataset involved a lot of advertising tweets which obviously had a positive effect, hence failing to provide the right picture. Excluding these advertising tweets from the dataset would reflect people's opinion in a better way.

We also believe that exploiting methods like stemming in the pre-processing phase and considering emoticons in the analysis phase should improve the precision of our results.

7. REFERENCES

[1] Michelle Annett and Grzegorz Kondrak. A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs in CiteSeerx at Proceedings of the Twenty-First Canadian Conference on Artificial Intelligence, 2008.

[2] Swathi Chandrasekar, Emmanuel Charon, Alexandre Ginet "CS229 Project Predicting The US Presidential Election using Twitter data" in CS229 Machine Learning course at Stanford University, 2012.

[3] Xing Fang and Justin Zhan "Sentiment analysis using product review data" in Journal of Big Data by Springer, 2015.

[4] Andrea Esuli and Fabrizio Sebastiani "SentiWordNet: A publicly Available Lexical Resource for Opinion Mining", 2006.

[5] Jon Tatum and Jonhn Travis Sanchez "Twitter Sentiment Analysis" in CS29 Machine Learning course at Stanford University, 2013.

[6] Sitaram Asur and Bernardo A. Huberman "Predicting Future With Social Media"

[7] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs in Proceedings of the ESWC Workshop on Making Sense of Microposts.

[8] F. Camastra, J. A. Hernandez, P. Kokol, J. Wang, and S. ZhuData cleaning, "Bag-of-Words Representation in Image Annotation: A Review", ISRN Artificial Intelligence, Volume 2012.

[9] Ann Taylor, Mitchell Marcus, Beatrice Santorini, "The Penn Tree Bank: An Overview", 2003.

[10] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, "Lexicon-Based Methods for Sentiment Analysis", Association for Computational Linguistics, 2011.