International Journal of Computer Applications (0975 - 8887)
Volume 179 - No.45, May 2018

Brand Analysis using Named Entity Recognition and Sentiment Analysis

Toshal Patel
Student
Shri Ramdeobaba College of Engineering and Management
Nagpur, Maharashtra, India

Megha Gupta
Student
Shri Ramdeobaba College of Engineering and Management
Nagpur, Maharashtra, India

A. J. Agrawal, PhD
Associate Professor
Shri Ramdeobaba College of Engineering and Management
Nagpur, Maharashtra, India

ABSTRACT

Internet has become a platform to host a myriad of services. An individual can make use of any services or resources and share his or her views about it at the same time. Today, social media platforms have become the largest source of accessing global reviews of the public regarding various movies, products, songs, etc. This paper focuses on proposing a method for gathering and analyzing the reviews of the people with respect to a company or different products of the company, and generating a report that will give a sentiment analysis of the reviews of the company's customers. In this paper, we discuss approaches to extract data from Twitter for a particular company or product, and performing Named Entity Recognition to extracted the related tweets. Analysis of the tweets will help in segregating the dataset based on their sentiments, generating a report of positive, negative or neutral customer reviews of a company's products or the brand itself.

General Terms

Twitter, Brand Analysis

Keywords

Named Entity Recognition, Sentiment Analysis

1. INTRODUCTION

The customer reviews are the most important part of analysis of a product or a brand. At each stage of product development, the product manager and his team need to analyze the performance of their product in their target market. The project manager and his team, the marketers, brand strategist, etc. are the key people who involve themselves in the analysis of the customer reviews, and make important strategies depending on the analysis of the data extracted. Also, it is a good way to keep a check on the customer reviews of the competitors. Not only for these people, but also the product analysis important for a customer to make a right decision of investing his or her money in the right product and making a purchase. Brand perception for a company is very important, i.e., the branding manager of the company needs to understand how the companys brand is perceived by the customers, for which they carry out surveys to understand this, and conclude different branding strategies. Knowing and understanding the customers is the most important thing, because today, with the developing technology, customer specific marketing makes the difference in the company's standings. Marketing, promotional and branding strategies are made planned based on the location, gender, age group, race, etc. Hence, they need to research the opinion of the public, of their company and products, and discern customer satisfaction. As a result, there are innumerable companies, large and small, that have opinion mining and sentiment analysis as part of business planning [8].

Apart from these, organizations can use this to gather crucial feedback about the new products that have been released in the market. For example, the company making sports clothing or shoes, like Nike and Adidas. They keep releasing new products with different styles and themes and to get an articulate discernment the brand managers need to analyze the products acceptance in different settings. These settings or features help get a detailed analysis. Thus, to do the brand analysis, various tools are required to evaluate the statistics of customers that are attracted to a particular brand and their promotions, along with what people think about their products.

Social media has now become a prime area of research, and with the increasing users every day, people tend to share their feelings and views on such platforms. We take the customer reviews as tweets from one such social media platform, Twitter. Twitter is a popular blogging platform where users post status messages, called tweets, which express opinions about different topics that can be used to for analysis [4]. We propose to build a model that automates the analysis of the tweets to a greater extend, and gives a thorough output of the sentiments of those tweets, to the user.

We use the Named Entity Recognition, to recognize the entities and extract specific tweets that are related to different products of a company. These tweets are passed on to the sentiment analyzer to determine the sentiment, i.e., positive, neutral or negative sentiments, of the tweet. This results in giving an overview of the performance of the product or company based on these sentiments.

1.1 Twitter Data Extraction

A great magnitude of data is available with the help of Twitter API, which makes it easier to collect millions of tweets for training of machine learning models and analysis. Twitter users post tweets from various devices such as laptops and mobiles phones. The reason for considering Twitter as a platform to get reviews of the customers is that, the users post short messages, restricted to 140 characters, on variety of topics, which helps us retrieve their opinions on various products of different companies. And, due to the character limit, the frequency of misspelt words, jargons, emoticons, slangs, is much higher than in any other text written for a customer review.

The API access to Twitter are based on design and access methods, namely, REST APIs, based on REST architecture, and Streaming API, providing continuous streaming of data. The REST API allows the developers application to make 300 requests within a rate limit window, and up to 180 requests using credentials of a single developer. In the Streaming APIs, each connection allows a developers application to submit up to 5,000 user-ids, and only public tweets published by the user can be retrieved using this API [9].

1.2 Named Entity Recognition

The tweets that are extracted from the Twitter API have high frequency of jargons, misspelt words, irregular capitalization, and various emoticons that are present. Hence, before performing any analysis on the data, we need to pre-process the tweets to take into consideration the irregular capitalizations, jargons, misspelt words and emoticons. For example, the word tomorrow can be written as tomorrow, 2mrw, 2moro, tommorrow, etc. Tweets are filled with these words and hence, pre-processing is a necessary step.

Based on the experimental study by Ritter et al. [3], there may be a case where we need to recognize whether the capitalized word is informative or uninformative, i.e., based whether the classified word is the beginning of the sentence, or either an entity, it will be informative and uninformative otherwise. Sometimes the non-entity words may be capitalized to give emphasis, and in other cases uninformative capitalization is used.

The results of the pre-processing of tweets are passed to the Named Entity Recognizer. Named entity recognition locates and classifies the information into various pre-defined categories such as names of people, different expressions, time, quantities, location, etc. There are various tools available for performing Named Entity Recognition (NER) and extraction from big chunk of data. The tool extracts the tweets for the given product from the millions of tweets that are extracted for a company. It will help identify the tweets that are potential customer reviews for a product. With the help of a pre-defined list of the names of the products of the company or brand, we can, with the help of Named Entity Recognition, extract and classify the tweets into different datasets of tweets for different products. Thus, this will result into different datasets of tweets that will be required to analyze the sentiments of the customers of the products.

1.3 Sentiment Analysis

Sentiment Analysis uses natural language processing to identify and extract one-sided information in the source materials or simply it refers to the process of detecting the polarity of the text [11]. It is performed using various machine learning techniques to determine the sentiments of a huge amount of data, that is in form of text. In the analysis of a brand, sentiment analysis plays a crucial role in determining the reactions of various customers. Sentiment Analysis classifies the text as positive, negative or neutral based on the words that appear in the text. It can help analyze the companys products, services or the brand image, based on what comments or tweets are posted on Twitter. The brand analysis will be based majorly on the sentiment analysis performed on the extracted tweets for a product, which will give the polarity [12] of the tweets, classifying them into positive, negative or neutral sentiments.

2. LITERATURE REVIEW

The brand analysis procedure begins by extracting the data from a social networking site, Twitter, with more than 140 million active users publishing tweets every second [4]. The analysis method explained in the book by Kumar et al. [9] covers the basic procedures for collecting, storing and analyzing the Twitter data. They have explained various ways to collect data using different APIs provided by Twitter, followed by the storage of the data to be used in real-time applications, finally discussing various visualization techniques that help in manual analysis of the result. Their analysis of the topic makes it easier for the developers to give the user of the brand analysis platform, a better approach for extracting, and visualizing the data.

The tweets extracted are noisy, uninformative, having redundancies and texting jargons. For extracting the tweets for various products, named entity recognition needs to be performed on the Twitter data extracted. Ritter et al. [3] explain the procedure to recognize the entities despite their terse and complexly redundant nature. Extracting the tweets for various products assists us to perform sentiment analysis on different products. Their experimental study gives us an insight that the methods used are better than the off the shelf named entity recognizers, such as that of Stanford, and other models trained of news data. The news data is refined and conveying proper context, as a result the models are trained on proper English. But when the same named entity recognizers are applied for recognizing entities in the tweets, they perform poorly, as the tweets contain texting jargons. Therefore, the tools and methods proposed by Ritter et al. [3], trained for handling tweets, giving accurate results, surfaces as better option.

The pre-processing of tweets including word parsing and tokenization, stemming and removal of stop words have been explained by Singh and Kaur [11], for better understanding of preparing the tweets for sentiment analysis. The sentiment analysis using distant supervision proposed by Go et al. [1], and the procedure explained in the internet blog by R. Janardhana [7], classifies the tweets into three different sentiments as positive, negative and neutral. Go et al. [1], have proposed different machine learning classifiers such as Support Vector Machine, Nave Bayes Classifier, and Maximum Entropy for classification purposes. The experimental study by Go et al. [3], gives us the performance of various classifier, coming to the conclusion that bigrams and unigrams both should be used along with Nave Bayes and Maximum Entropy classifier for better accuracy of sentiment classification.

3. PROPOSED METHODOLOGY

To analyze a how a brand is received by the customers, based on the reviews of the customers on the social media platform, we can make use of the techniques in natural language processing and machine learning for various steps of analysis. The flow of the brand analysis is explained in figure 1.

3.1 Twitter Data Extraction

We can extract data using the REST API, that uses the Open Authentication (OAuth) and the developers secret credentials to authenticate the access the Twitter API to get the protected data from Twitter. To get a constant stream of public tweets published by a user, we can create a POST request to the Twitter API and draw the search results as a stream, matching a search query.
Applying a python script to extract the data, making OAuth authentication and HTTP connection we can extract the Tweet object which has various fields including the text of the tweet, which we are mainly interested in, time of creation, retweet count, id of the user, language, place giving the city and the country where the user posted the tweet, and many others.

3.2 Preparing Tweets for NER

The experimental study for Named Entity Recognition by Ritter et al. [3], proposes the methods to preprocess the tweets and use the intermediate results of each step to facilitate Named Entity Recognition.

3.2.1 Part of Speech Tagging.
Part of Speech tagging is used for named entity segmentation and information extraction. The study by Ritter et al. [3], suggests that the POS tagging assigns each word to its most frequent tag and for each Out of Vocabulary (OOV) words, it assigns the most common POS tag. This method lowers the accuracy for the tweets which contains large number of OOV words than in any text for which the POS taggers are trained for. Hence, they developed a special POS tagger, T-POS, to overcome the differences in style of writing between the tweets and the normal text, and the special vocabulary of tweets. Due to their aptness to identify strong dependencies between the neighboring POS tags and use the highly correlated features, such as a words identity along with prefixes and suffixes, it uses Conditional Random Fields [6]. Making use of Brown clusters, POS dictionaries, spelling and contextual features, and leveraging the POS-labelled tokens from Penn Treebank and tokens of annotated IRC chat data, the T-POS gives better classification accuracy.

3.2.2 Shallow Parsing.
Shallow parsing, also known as chunking, is used in information extraction and named entity recognition. It identifies the non-recursive phrases noun, verb and prepositional phrases in a text. As the method suggested by Ritter et al. [3], using the results of T-POS, Conditional Random Fields for inference, and tokens from CoNLL dataset, T-CHUNK shallow parses the tweets with better accuracy.

3.2.3 Capitalization.
For performing named entities, capitalization is a prime feature, and regrettably, we find irregular and unreliable capitalization in tweets. T-CAP capitalization classifier developed by Ritter et al. [3], predicts whether or not the tweets are informatively classified. Support Vector Machines are used for learning, including the features: fraction of words that are capitalized in tweets, the fraction of words that appeared in dictionary of frequently lowercased or capitalized words but are not lowercased or capitalized, the number of times I appears in lowercase, and whether the beginning word of the sentence is capitalized or not.

3.3 Named Entity Recognition

The news-trained Named entity recognizers perform poorly because they rely heavily on the capitalization of the words, which we know is unreliable in the case of tweets. Following the method of Named Entity Recognition by Ritter et al. [3], which divides the work of segmentation and classification of named entities as a separate tasks. To train the models, in-domain data is used, and the features provided by T-CAP. The @ entities which identify the users, are excluded while training as they are ambiguous and trivial. T-SEG is used for segmentation. It models Named Entity Segmentation as a sequence-labeling job using IOB encoding for symbolizing segments. IOB encoding segregates whether the word is either inside, begins or is outside a named entity. T-SEG uses Conditional Random Fields for learning and inference, Brown Clusters and outputs of T-POS, T-CHUNK, and T-CAP in generating features. The results of T-SEG are better than those of other state of the art segmentation techniques.

The work of classifying named entities is difficult, as they are concise, containing incomplete sentences, hence, determining the type of entities becomes difficult. For example, KKTNY in 15mins without any prior knowledge, there is not enough context that will enable the classifier to determine the type of entity, whether the tweeter is referring to a TV show, a movie or a place to visit. Hence, to handle such problems, large lists of entities and their types obtained from open-domain ontology for distant supervision, are used which enable us to use large amounts of unlabeled data in learning. For classifying named entities, distant supervision applying LabeledLDA is used. It constrains each entitys distribution over topics based on its set of possible types. The sharing of information across contexts is beneficial in case of tweets with insufficient text to determine the entitys type. T-CLASS is able to classify the entities mentioned in the context with types as Person, Geo-location, Company, Product, Facility, TV-show, Movie, Sports-team, Band, and Other. The T-NER gives the end-to-end performance on segmentation and classification.

3.4 Sentiment Analysis

Before performing the Sentiment Analysis, the tweets are preprocessed for removing the emoticons and retweets, i.e., the duplicate tweets are removed. The sentiment analysis method proposed by R. Janardhana in his blogpost [7] and Go et al. [1] can be implemented. The proposed method converts the tweets into lower case, eliminate the URLs, usernames starting with @, hashtags, punctuations and additional white spaces.

Additionally, the tweets are filtered, removing the stop words, for example a, the, is, with, etc., and replacing the repetitive letters in a word, for example hungrrryyyy will be replaced by hungry, by finding two or more repetitive letters and replacing by 2 of the same. The tweets that are to be analyzed should always begin with an alphabet.

After filtering the tweets, feature words are extracted from the tweets ignoring the other words. For example, from the sentence, just had some bloodwork done. my arm hurts, the feature words will be bloodwork, arm, and hurts [7]. For each tweet, for the

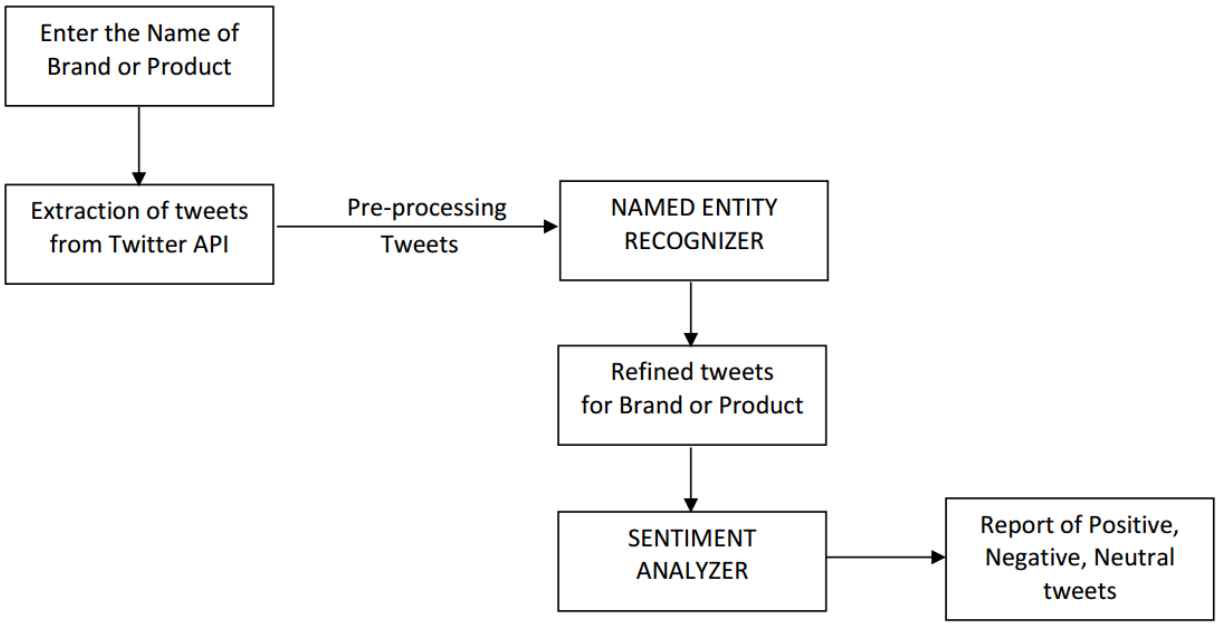

Fig. 1: Flow graph of Brand Analysis.

presence of feature word is marked as 1, and absence is marked as 0. Hence, feature list is formed from feature vector. Feature vector is a combination of such feature words. Distant supervision is used to get a large training dataset, explained by Go et al. [1]. When a new tweet is encountered, the classifier predicts the tweets sentiment by extracting the feature words, and generating a pattern of 0s and 1s. The classifier used for this purpose can be any classifier Nave Bayes classifier, Maximum Entropy classifier, Support Vector Machines.

The output of the entire process can be in the form of a graph that displays the number of positive, negative and neutral tweets of different products or services, or a company as a whole.

4. CONCLUSION FOR FUTURE SCOPE

Brand Analysis is a key factor in determining the progress and growth of a company in the market. It gives an overview to the companys brand strategist, how the products of the company, or the brand is received by the customers, and what is their brands value. Hence, our proposed method helps in basic brand analysis by determining the sentiments of the customers towards particular products.

This analysis can be extended to extracting performing brand analysis for each product of a company, which will help get a detailed brand analysis of the company. It can be made more detailed by adding different features such as categorizing the customers for different products of a company by their geo-graphical location or age group, which can be extracted from the Tweet object fetched from the Twitter API; popularity of the product, by calculating the number of positive tweets or reviews the product have received over a period of time. It can also help in the comparative analysis of various competitors of the company, or between different models or versions of the product series. The number of tweets and retweets can help determine the customers attracted towards a particular promotion or scheme introduced. Also, with the increasing volume of data, several literatures have been exploring the big data issues for sentiment analysis [10], and introduced big data tools for the same [2]. There have also been some improved models for sentiment analysis on big data [5], and regarding the scalability issue of sentiment analysis [5].

5. REFERENCES

[1] R. Bhayani A. Go and L. Huang. Twitter sentiment classification using distant supervision. CS224N, Project Report, Stanford University, Stanford, CA, 2009.

[2] H. Gabelica A. Mihanovic and Z. Krsti. Big data and sentiment analysis using knime: Online reviews vs. social media, 2004. MIPRO.

[3] Mausam A. Ritter, S. Clark and O. Etzioni. Named entity recognition in tweets: an experimental study. In *EMNLP11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, 2011.

[4] K. Sobel B. J. Jansen, M. Zhang and A. Chowdury. Microblogging as online word of mouth branding. In *CHI EA 09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3859–3864, 2009. ACM.

[5] L. Bing and K. C. C. Chan. A fuzzy logic approach for opinion mining on large scale twitter data. In *UCC 14 Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, pages 652–657, December 2014.

[6] A. McCallum J. D. Lafferty and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling

sequence data. In *ICML'01 roceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers, 2001.

[7] R. Janardhana. How to build a twitter sentiment analyzer, May 2012. https://www.ravikiranj.net/posts/2012/code/how-build-twitter-sentiment-analyzer/.

[8] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*, volume 2, pages 1–135. 2008. http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf, unpublished.

[9] F. Morstatter S. Kumar and H. Liu. *Twitter Data Analytics*, pages 21–51. Cambridge University Press, Jan 2015.

[10] H. M. Zin N. M. Sharef and S. Nadali. *Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data*, volume 12, pages 153–168. 2016.

[11] R. Singh and R. Kaur. Sentiment analysis on social media and online review. *International Journal of Computer Applications (0975-8887)*, 121(20):44–48, July 2015.

[12] J. Wiebe T. Wilson and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 2005.