International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

Sentiment Analysis using Twitter Data

Aanshi Desai[Å], Haranshvir Gujral[Å*] and Sindhu Nair[Å]

[Å]Computer Department, DJSCOE, Vile-Parle (W), Mumbai – 400056, India

Accepted 05 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

Twitter is a social media platform where users post their views of everyday life. Every tweet has a sentiment hidden behind it. Our aim is to use this data which is present in abundance and to make strategic decisions from it. The objective of this paper is to compare the earlier methods: Turney's approach and Pang's approach, study the ongoing technique i.e. NLP, and discuss the various issues and challenges faced while performing sentiment analysis and mining the data obtained from Twitter.

Keywords: *Social media, Twitter, Tweets, Sentiment Analysis, Mining.*

1. Introduction

Social media is the interaction among individuals within which they produce, share, or exchange info and concepts in virtual communities. It differs from ancient or industrial media in many ways, as well as quality, reach, frequency, usability, immediacy, and length. They have provided an open space where individuals are liberal to exchange ideas on corporations, brands, and merchandise.

Twitter may be a social network that enables the user to freely publish short messages, known as Tweets. Twitter is an internet social networking and micro blogging service that permits users to send and browse short 140-character text messages, known as "tweets". Twitter is currently a number one social network for its wealthy however additionally by its content and easy access. With a mean of two hundred million messages sent per day, Twitter is a perfect place to collect instant opinions on varied subjects.

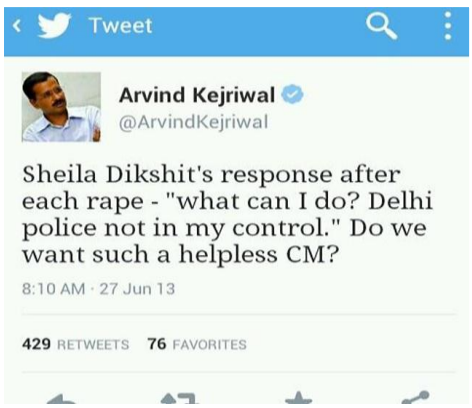

Fig. 1 An example of the content of a Tweet

A tweet may be of at most a hundred and forty characters. This can be a brief message as compared to the other social media. Thus, the short format messages facilitate analysis since short messages barely contain over one idea. Hence, usually a Tweet revolves around a specific thought making it easier to mine the information.

Users can cluster posts along by topic or sort by use of hashtags – words or phrases prefixed with a "#" sign. Similarly, the "@" sign followed by a username is employed for mentioning or replying to different users. One can repost a message from another Twitter user, and share it with one's own follower. The re-tweet function is symbolized by "RT" within the message. Twitter provides its Application programming interface (API) to access these Tweets for the web developers.

Twitter provides with its API's for free. It allows obtaining tweets on basis of geographical location, keywords of hashtags, username, live tweets (streaming data). There are two types of API's used mainly, one is the REST API and the other is the Streaming API. Each has their pros and cons, and was designed for various use cases. The Streaming API was designed preponderantly to run interrupted, capturing everything from never-ending stream of information in real time. REST API was designed to require variety of requests to perform a variety of tasks, beginning with compiling CSDL rules, costing streams, and taking smaller samples of streams.

Micro blogging platforms square measure employed by totally different individuals to express their opinion concerning totally different topics, thus it is a valuable supply of people's opinions. Twitter contains a huge variety of text posts and it grows each day. The collected corpus are often every which way large. Twitter's audience comprises of normal users, celebrities, business representatives, politicians, activists, and even country presidents. Therefore, Twitter has the potential to gather posts of users from totally different social and interests

*Corresponding author: **Haranshvir Gujral**

groups. Twitter's audience is depicted by users from several countries. Though users from United States are prevalent and account for two-thirds of Twitter's revenue, it is possible to gather information in several languages and even in combination of languages. India is expected to surpass United Kingdom as the third largest Twitter population by the end of this year.

2. Sentiment Analysis

Also acknowledged as opinion mining, sentiment analysis refers to the employment of natural language processing, text analysis and computational linguistics to determine and extract subjective data in source. It aims to see the perspective of a speaker or an author with relation to some topic or the contextual polarity of a document.

Tweets are classified into three different opinions: positive, negative or neutral. This is done on basis of few keywords that indicate a certain polarity. Like for example; Positive: yes, happy, excellent, amazing, wow, best, etc. Negative: no, sad, yuck, hate, bad, disgusting, etc. Neutral: okay, least bothered, don't care, etc.

3. Application of Sentiment Analysis

Sentiment classification by using twitter data has varied applications. We could find out the public opinion about any trending topic that is going on. We can classify product review, movie review and other kinds of reviews. It can be used to find out the majority of support of political parties during election times. If a company needs to improve on their existing product, they can use the reviews given by the customers.

4. Problems faced while mining tweets

- Sentiment polarity: Suppose we wish to classify an opinionated text as either positive or negative, according to the overall sentiment expressed by the author within it.
- Multiple language usage: Use of native languages along with English in tweets makes it difficult to determine the sentiment of the tweet.
- Usage of Slang: Social media users are free to use slangs. These slangs are not present in the dictionary. Thus, special database for storing slangs for comparisons are required.
- Abbreviations: Usage of short forms and spelling errors are the most important issues faced in mining. The mining tool must be able to deal with the grammatical errors and short forms.
- Subtle mockery: It is possible that we might associate this as a positive opinion rather than a negative one. For example, "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut." No ostensibly negative words occur.
- Finding overall polarity: It is difficult in determining the overall polarity as every positive/negative word has its own weightage; for example, "wonderful" is more positive than "good". Thus, determining the overall strength is not so easy.
- Multiple Levels of Message content: It is common that tweets contain a link to a webpage, which means to have multiple levels of processing of the message content.

5. Early methods used

5.1 Turney's Approach

The linguistics orientation of words has been elaborate first of all for the adjectives. This observation has typically been thought of because the proof that some adjectives are sensible sentiment indicators. Four stages: initial of all, there's a requirement to create phrase segmentation (part-of-speech). Then, we tend to produce adjectives and adverbs in series of 2 words we tend to apply later SO-PMI so as to calculate the semantic orientation of every detected series. Finally, we feature out a review classification as positive or negative by scheming average of all find orientation. This approach presents a straightforward unsupervised learning algorithmic rule for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is foretold by the common semantic orientation of the phrases within the review that contain adjectives or adverbs. A phrase features a positive linguistics orientation once it's sensible associations (e.g., "subtle nuances") and a negative semantic orientation once it's unhealthy associations (e.g., "very cavalier"). In this, the linguistics orientation of a phrase is calculated as the mutual data between the given phrase and therefore the word "excellent" minus the mutual data between the given phrase and therefore the word "poor". A review is assessed as recommended if the common linguistics orientation of its phrases is positive. The rule achieves a mean accuracy of seventy four once evaluated on 410 reviews from Opinions, sampled from four completely different domains (reviews of cars, banks, movies, and travel destinations). The accuracy ranges from eighty four for automobile reviews to sixty six for film reviews.

5.2 Pang's Approach

Pang and Lee have projected another approach for the polarity classification of cinematographic reviews. This approach consists of 2 goals. The first goal is to find the document's components that are subjective. Then, using an equivalent statistical classifier to determine the polarity solely on subjective fragments detected antecedently. Rather than doing the subjectiveness classification for every phrase individually, they admit that they'll see an explicit degree of continuity within the phrases subjectiveness. They give preferences so as to possess proximity phrases that have constant level of subjectiveness. Each phrase within the document is then labeled as subjective or objective within the method of collective classification.

6. Attempt to solve the problem of Sentiment Analysis (Natural Language Processing)

Sentiment analysis is a complicated field. Since it involves the process and interpretation of linguistic communication,

it should take care of natural language's inherently ambiguous nature, the importance of context, and different complications that don't lend themselves to automation. For example, "Just go read the book." Context is necessary in this case. If mentioned concerning a book, this might be thought of as a recommendation, whereas if it refers to a film adaption of a book, it might appear to advise that the film is not worth watching. Detecting humor or sarcasm could be a drawback that may appear to be on the far side of current technologies. Natural language processing (NLP) is the ability of a computer program to know human speech as it is spoken. Natural language processing is a part of artificial intelligence (AI). Current approaches to natural language processing are based mostly on machine learning, a type of artificial intelligence that examines and uses patterns in information to improve a program's own understanding. The advantage of natural language processing may be seen when considering the subsequent two statements: "Cloud computing insurance should be part of every service level agreement" and "A good SLA ensures an easier night's sleep, even within the cloud." If you employ linguistic communication processing for search, the program will recognize that cloud computing is an entity, that cloud is an abbreviated style of cloud computing and that SLA is an industry form for service level agreement.

7. Method followed for Sentiment Analysis

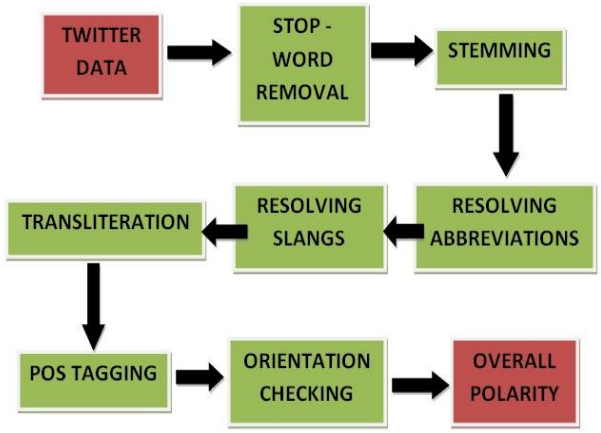

Fig. 2 Flowchart of the process of Sentiment Analysis

- Data: Obtain the data using the required Twitter API.
- Stop word removal: Stop words are those words that need to be removed prior to the mining process. These words carry no meaning hence they are not useful for sentiment analysis. There is no fixed list for stop words. For example: a, an , the , that , they etc these words do not carry any sentiment or opinion hence they are removed so as to increase the efficiency of mining.
- Stemming: It is reducing the words to their root word (stem). That is: is, are, am becomes 'to be'. For Example: "The boy's shoes are dirty" gets reduced to "The boy shoe be dirty".
- Resolving Abbreviations: The data present in social media has a lot of abbreviations/short forms due to the presence of no rules regarding grammar. Thus, finding the correct meaning of an abbreviation is required and is a difficult task. For Example: "I love my hsh" is resolved to "I love my home sweet home".
- Resolving slangs: Social media data contains many slang words, abuses etc. The dictionary fails to detect these words thus, while mining we must make sure we consider slangs also as they are equally important in determining the sentiment of the tweet.
- Transliteration: It is the conversion of one text to another alphabetically but not literally. There is a possibility that a single tweet contains multiple languages. Thus it is important to convert it all into one language for further mining.
- POS Tagging (Part of Speech Tagging): Also called as grammatical tagging. This means marking up the word in the text to its corresponding part of speech. That is, to determine whether the word is a noun, pronoun, adverb, adjective.
- Orientation Checking: Once the POS Tagging is performed, the adverbs are taken into consideration. Adverbs determine the sentiments.
- Overall Polarity: A single tweet can consist of multiple sentiments, thus we need to find the overall polarity of the tweet.

This leads to the overall opinion summarization.

Conclusion

Twitter nowadays became one of the major types of the communication. It is seen as online word-of-mouth branding. The large amount of information available on Twitter makes it an effective source of data for sentiment analysis. In our research, we have seen the method for automatic collection of a corpus that can be used to train a sentiment classifier and observed the difference in distributions among positive, negative and neutral sets. The classifier is able to determine positive, negative and neutral sentiments of tweets. The classifier is based on a variety of methods, which uses stemming and POS-tags as features. As a future work, we plan to collect a multilingual set of Twitter data and compare the characteristics of the data set across different languages. We plan to use the obtained data to design a multilingual sentiment classifier.

Acknowledgment

We would like to thank our honorable principal Dr. Hari Vasudevan of D. J. Sanghvi College of Engineering, and Head of Department of Computer Engineering Dr. Narendra Shekokar for giving us the facilities and providing us with a propitious environment for working in college. We would also like to thank S.V.K.M. for encouraging us in such co-curricular activities.

References

P.Turney, (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proceedings of the Association for Computational Linguistics (ACL)*, pp.417-424.

P.Turney, M.Littman, (2003), Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems*, pp.315–346.

B.Pang, L.Lee, (2004), A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proceedings of the Association for Computational Linguistics (ACL)*, pp.271-278

T.Wilson, J.Wiebe, P.Hoffman, (2005), Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp.347-354

T.Wilson, J.Wiebe, R.Hwa, (2006), Just How Mad Are You? Finding Strong and Weak Opinion Clauses, *Proc. of AAAI*, pp.761-769

V.Hatzivassiloglou, K.McKeown, (1997), Predicting the Semantic Orientation of Adjectives, *Proc. of the Joint ACL/EACL Conference*, pp.174-181.

C.Yang, K.Lin, and H.Chen, (2007), Emotion classification using web blog corpora, *Proceedings of the IEEE/WIC/ACM Interna-tional Conference on Web Intelligence*, pp.275-278.

K.Dave, S.Lawrence, D.Pennock, (2003), Mining the peanut gallery: opinion extraction and semantic classification of product reviews, *Proceedings of the 12th international conference on World Wide Web*, pp.519-528. A.Abbasi, H.Chen, A.Salem, (2008), Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums, *ACM Transactions on Information Systems*, pp.26-28.

R.Y.K. Lau, (2003), Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web, *Web Intelligence and Agent Systems An International Journal*, pp.1(3-4):1–22.