Twitsen, A Business Promotional Analyzing Model using Twitter

Heishnam Reeona Devi[1], T. Sudalai Muthu[2]

[1]Department of Computer Sciences & Engineering,
Hindustan University, Chennai, India
reeo.h13@gmail.com

[2]Assistant Professor,
Department of Computer Sciences & Engineering,
Hindustan University, Chennai, India
sudalaimuthut@gmail.com

Abstract: ***Twitter is a microblogging service with more than 500 million users which generates over 340 million tweets daily. This paper aims to harness the data from Twitter and measure the sentiment or opinion associated with the tweet message to assists in the feedback measurement of any product or object. It identifies the polarity of the product as positive, negative or neutral. Part Of Speech tagging is used on the text data for linguistic categorization along with Semantex, an automatic text analyzer which groups the verb classes and Wikipedia's online dictionary, Wikitionary to determine the polarity of adjective in the tweets.***

Keywords: Twitter, POS, Semantex, Wikitionary, Twitsen

1. Introduction

With the increase popularity of social networking/microblogging service, the web consists of enormous quantities of electronic text. The exponential growth of blogging phenomenon plays an increasingly influential role in the global market place. The popular microblogging service Twitter [2] has millions of user and hence has a huge user created data content which can be exploited. Blogger or the Opinionate of Twitter tweets about their views and opinions of a topic or an object. At present, obtaining user feedback about a company product requires bothering the customer to take surveys which is time taking, requires a particular format and also relies on the goodwill of the people to take the survey. This process can be avoided by gathering the necessary information from the web. Twitsen, is design to extract the textual and linguistic feature from the Twitter post to analyze the feedbacks of a product and then a classifier is developed to categorize the blog posts. The blog is classified as subjective Vs objective [4] [7] and the polarity is measure as positive (good), negative (bad) or neutral. It uses the POS tagging for linguistic categorization of the text according to a particular part of speech. To aid in the classification, we use Semantex [3], an automatic text analyzer that groups the verb classes on the text and also the Wikitionary, the Wikipedia's online dictionary, to determine the polarity of adjective in the post/tweets.

1.1 Dataset

Dataset is a collection of data which provides the structure and properties of the data. This paper uses the microblogging site Twitter [2], as the primary source of data. Twitter has a large number of users and hence it provides a rich source of data for opinion mining and sentiment analysis. The user will provide the keyword for the product item against which the sentiment must be provided. Then the twitter search query is executed to fetch the data/tweet message from the Twitter which is relevant to the keyword. The data/tweet message can be obtained for analysis via Jason web syndication feeds from the Twitter database. The sentiment is then calculated for each of the tweet message which is relevant to the keyword provided.

1.2 Part Of Speech Feature

Part-of-speech [1] is a linguistic category of words which is defined by the syntactic (structure or arrangement of any language) or morphological (identification, analysis and description of a structure of a language) behavior of the lexical item. To improve the classification of sentiment in product reviews or feedbacks from customer, POS feature is effectively used in sentiment/opinion classification. Adjectives [9] are important indicator of subjectivities and opinions. So, posts with higher number of adjectives and adverbs are most likely to be subjective. The fundamental function of adjectives and adverbs is to denote the qualities of entities and events.

1.3 Polarity Feature

The polarity feature [10] is used to summarize the content of opinionated text on a topic whether they are positive, negative or neutral. Words that encode a desirable state have a positive orientation while words that represent undesirable states have a negative orientation. The lexical level polarity can be determined by using Semantex and Wikitionary. Thus, the overall polarity which the blogger wants to convey is determined by evaluating the post/tweets as "positive", "negative" or "neutral

2. Related Works

This paper is closely related to the works of Paula Chesley [3] automatic blog classification in which blogs are analyzed at the post level and verb class information are taken into consideration. It uses Semantex and also uses Wikitionary. A classifier is developed to categorize the

blog post with respect to sentiment. Peter D. Turney [1] proposed a semantic orientation used for unsupervised classification of reviews. It uses part-of-speech (POS) tagging to identify the adjectives and adverbs in the phrase and then estimate the orientation. Also, Kevin Gimpel [5] and Olutobi Owoputi [6] design a POS tagging for Twitter which addresses the problem of POS tagging for English data for the microblogging site Twitter.

Tag	Description	Examples	%
Nominal, Nominal + Verbal			
N	common noun (NN, NNS)	books someone	13.7
O	pronoun (personal/WH; not possessive; PRP, WP)	it you u meeee	6.8
S	nominal + possessive	books' someone's	0.1
^	proper noun (NNP, NNPS)	lebron usa iPad	6.4
Z	proper noun + possessive	America's	0.2
L	nominal + verbal	he's book'll iono (= *I don't know*)	1.6
M	proper noun + verbal	Mark'll	0.0
Other open-class words			
V	verb incl. copula, auxiliaries (V*, MD)	might gonna ought couldn't is eats	15.1
A	adjective (J*)	good fav lil	5.1
R	adverb (R*, WRB)	2 (i.e., *too*)	4.6
!	interjection (UH)	lol haha FTW yea right	2.6
Other closed-class words			
D	determiner (WDT, DT, WP$, PRP$)	the teh its it's	6.5
P	pre- or postposition, or subordinating conjunction (IN, TO)	while to for 2 (i.e., *to*) 4 (i.e., *for*)	8.7
&	coordinating conjunction (CC)	and n & + BUT	1.7
T	verb particle (RP)	out off Up UP	0.6
X	existential *there*, predeterminers (EX, PDT)	both	0.1
Y	X + verbal	there's all's	0.0
Twitter/online-specific			
#	hashtag (indicates topic/category for tweet)	#acl	1.0
@	at-mention (indicates another user as a recipient of a tweet)	@BarackObama	4.9
~	discourse marker, indications of continuation of a message across multiple tweets	RT and : in retweet construction RT @user : hello	3.4
U	URL or email address	http://bit.ly/xyz	1.6
E	emoticon	:-) :b (: <3 o_O	1.0
Miscellaneous			
$	numeral (CD)	2010 four 9:30	1.5
,	punctuation (#, $, ' ', (,), , , . , : , ``)	!!! .... ?!?	11.6
G	other abbreviations, foreign words, possessive endings, symbols, garbage (FW, POS, SYM, LS)	ily (*I love you*) wby (*what about you*) 's ♫ --> awesome...I'm	1.1

Table 1: Twitter POS Tags

It used a feature based sequence tagging model with several broad types of orthographic, lexical, and distributional features. It developed a tagset, annotate data, develop features and report tagging. Table 1 is a reference from Kevin Gimpel [5] design of a POS tagging for Twitter which indicates the tagging scheme encompasses 25 tags where each tag is denoted by a single ASCII character(only the Twitter/online specific tags are considered for this paper).

3. Proposed System, Twitsen

The proposed system, Twitsen effectively utilizes and applies the advantages of opinion or sentiment analysis in measuring the feedback of a product from Twitter. It generates the semantic orientation/polarity about the product of a company by analyzing the text data that is available in Twitter. Based on the keyword provided by the user, sufficient data/tweets shared by a Blogger or the Opinionate related to the keyword is fetch from the

Twitter and then analyze it to measure the opinion/sentiment in the post/tweets. In doing so, the customer is no longer bothered in the market survey. Also it reduces the time and helps in the decision making process. It makes use of the POS tagging method for linguistic categorization of the words along with an automatic text analyzer Semantex and Wikitionary in the post/tweet.

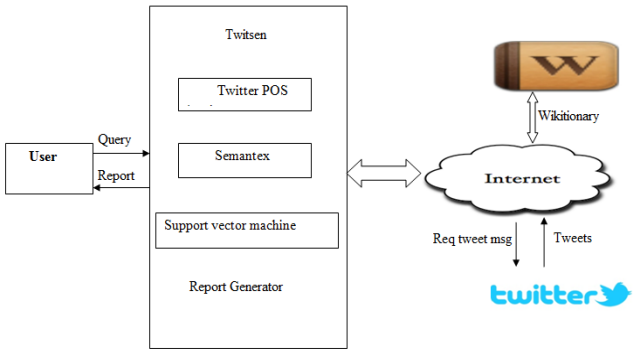

Figure 1: System Architecture

CC	Coordinating conjunction	RP	Particle
CD	Cardinal number	SYM	Symbol
DT	Determiner	TO	to
EX	Existential there	UH	Interjection
FW	Foreign word	VB	Verb, base form
IN	Preposition/subordinate conjunction	VBD	Verb, past tense
JJ	Adjective	VBG	Verb, gerund/present participle
JJR	Adjective, comparative	VBN	Verb, past participle
JJS	Adjective, superlative	VBP	Verb, non-3rd ps. sing. present
LS	List item marker	VBZ	Verb, 3rd ps. sing. present
MD	Modal	WDT	wh-determiner
NN	Noun, singular or mass	WP	wh-pronoun
NNP	Proper noun, singular	WP$	Possessive wh-pronoun
NNPS	Proper noun, plural	WRB	wh-adverb
NNS	Noun, plural	``	Left open double quote
PDT	Predeterminer	,	Comma
POS	Possessive ending	''	Right close double quote
PRP	Personal pronoun	.	Sentence-final punctuation
PRP$	Possessive pronoun	:	Colon, semi-colon
RB	Adverb	$	Dollar sign
RBR	Adverb, comparative	#	Pound sign
RBS	Adverb, superlative	-LRB-	Left parenthesis *
		-RRB-	Right parenthesis *

Table 2: POS Tags

3.2 System Architecture

In the proposed model Twitsen, describe in figure 1 the user is provided with a graphic user interface (GUI) through which the input is given. For the given input the semantic orientation/polarity is calculated. This input is treated as a keyword and the twitter search query is executed to fetch the data/tweet message from the Twitter. The sentiment is then calculated for each of the tweet message which is relevant to the keyword provided. The search can also be narrowed down by filtering the dataset to be processed based on the tweet user information (location, no. of follower, no. of people whom he follows).

3.3 Twitsen

The data fetched is passed to Twitsen where the appropriate algorithm and methodology is applied in order to generate the polarity of the data. Binary classification is used in classifying the blog post/tweet message as subjective Vs objective and then the polarity is measured as positive (good), negative (bad) or neutral. The POS tagging is used for linguistic categorization. The lexical

Advantages

- Advanced filtering mechanism can be applied on twitter dataset.
- Information associated with the Opinionate (location, no. of follower, friends following the user etc) is considered.
- Product survey takes place without the direct involvement of the customer.
- More effective way of extracting knowledge or polarity of a particular product.

level polarity can be determined by using Semantex and Wikitionary.

3.3.1 Tagging Module

In the tagging module, POS feature is applied to all the data/tweet messages from Twitter that consists of the input keyword. The POS tagging is used to split the post/tweet message to classify words into its linguistic category and assign the corresponding part of speech such as noun, pronoun, adjective, adverb etc to the word in a text. Adjectives are important indicator of subjectivities and opinions. So, posts with higher number of adjectives and adverbs are most likely to be subjective. Also twitter has various features and hence it has to be considered. Apart from the linguistic categorization, Twitter POS tokenizer recognizes URLs and emoticons. Table 2 shows the POS tags consist of coded abbreviation.

3.3.2 Classification Module

After the tagging process is complete, the relevant tweet message is classified as subjective Vs objective. To aid in the classification it uses Semantex and Wikitionary. Also, SVM classifier is used for classification task since it is robust and can handle noisy data. Thus, the overall polarity which the blogger wants to convey is determined by evaluating the post/tweets as "positive", "negative" or "neutral".

After calculating the polarity of the post/tweet messages with respect to the user given input, a report is prepared. The report comprises of information such as total number of tweet message processed, total number of positive, negative and neutral feedback with respect to the user input. Finally it is displayed in the form of pie chart or graph to the user.

4. Conclusion and Future Work

The system Twitsen, by utilizing the data from Twitter will help the business analyst in obtaining the overall feedback of a particular product. In doing so, the analyst can detect the growth as well as the cause of product declination in the market. It then enables to carry out necessary strategies for better marketing of the product and also helps in the business decision making process.

At present, the system is focused only with the English language which in the future can be improved to support multiple languages. Also, the system can further be enhanced to excess data from multiple social networking/microblogging services at a particular time.

References

[1] Peter D. Turney (July 2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the ACL pp 1-7.
[2] Pete Cashmore (2009) e-book "The Twitter Guide Book"
[3] Paula Chesley, Bruce Vincent and Li Xu, Rohini K. Srihari (2005), "Using Verbs and Adjectives to Automatically Classify Blog Sentiment", American Association for Artificial Intelligence pp 2-7.
[4] Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens (June 2007),"Automatic Sentiment Analysis in On-line Text ", Proceedings ELPUB2007 Conference on Electronic Publishing pp 1-12.
[5] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith (2011), "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies pp 2-5.
[6] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpely Nathan Schneider (September 2012), "Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances" The research was supported in part by an REU supplement to NSF grant IIS-0915187 and Google's support of the Worldly Knowledge project at CMU pp 3-9.
[7] Bing Liu (2010), "Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing pp 1-38.
[8] Bo Pang and Lillian Lee (2008), "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) pp 1–135.
[9] Rudy Prabowo1, Mike Thelwall (April 2009), "Sentiment analysis: A combined approach", Journal of Informetrics Volume 3, Issue 2, pp 143–157

Author Profile

Heishnam Reeona Devi received the B.Tech degree in Computer Science & Engineering from Dr MGR University, Chennai, India in 2010. And is currently doing M.Tech in Computer Science & Engineering from Hindustan University, Chennai, India.

T.Sudalai Muthu is currently working as an Assistant Professor in Department of Computer Science & Engineering at Hindustan University, Chennai, India.