Advances in Natural and Applied Sciences, 8(16) Special 2014, Pages: 40-43

AENSI Journals

Advances in Natural and Applied Sciences

ISSN:1995-0772 EISSN: 1998-1090

Journal home page: www.aensiweb.com/ANAS

Exhibiting Human Behavior in Transforming Informal Words in to Formal Words on Tweet Messages

[1]Kumaragurubaran Thangavel and [2]IndraDevi M

[1]*Research Scholar, Anna University, Regional Office, Madurai. India.*
[2]*Professor,Department of Computer Science and Engineering,Kamaraj College of Engineering and Technology,Virudhunagar. India.*

ARTICLE INFO

Article history:
Received 3 September 2014
Received in revised form 30 October 2014
Accepted 4 November 2014

Keywords:
Social Networking Sites (SNS), Natural Language Processing, Informal words, Formal words

ABSTRACT

Social Networking Sites (SNS) has drastically changed its color from making friends and maintaining relationships to make reviews and to voice our opinions. This has brought researcher an intention to do their research on this particular area under Natural Language Processing. Though the researcher came with their own ideas, the first and foremost things on their research are to transforming the informal words posted by the users into formal words. This paper expresses a technique that reveals about how the system transforms the informal words in to formal words and enables machines to think or act as intelligent as humans.

To Cite This Article: Kumaragurubaran Thangavel and IndraDevi M., Exhibiting Human Behavior in Transforming Informal Words in to Formal Words on Tweet Messages. **Adv. in Nat. Appl. Sci.,** *8(16): 40-43, 2014*

INTRODUCTION

Two decades past Social Networking Sites were emerged by focusing on bringing people together to interact with each other through chat rooms, and encouraged users to share personal information and ideas via personal WebPages. Some communities took a different approach by simply having people link to each other via email addresses. Later, user profiles became a central feature of social networking sites, allowing users to compile lists of friends and search for other users with similar interests. Thus SNS made pavement to make new friends with peoples of same interest and maintaining relationships with them. In this era the usage of mobile, android phones and tablets gets rapid and their build in features made peoples to update their thoughts within no time. Consequently, the personal websites, blogs, social networking sites, forums and review sites etc has vast information available.

On the same time if a person have to make a decision he/she may go with opinions of his/her friends and relatives. If an organization wants to understand the opinions of public towards their new products or services they may conduct opinion polls and surveys. As years go on, the medium for conducting the surveys and opinion polls were changed but the zealous to make such reviews did not (Akshi Kumar and Teeja Mary Sebastian, 2012). The huge availability of information on personal websites, blogs, social networking sites, forums and review sites etc made marketing intelligence, social psychologists and others to pay their attention on them in extracting and mining views, moods and attitude of the public. The immediate updates on web helps researchers to mine opinions and reviews in fast manner comparatively than other mediums.

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. The pre work can group the existing approaches on Sentiment analysis and Opinion mining is keyword spotting, lexical affinity, statistical methods, and concept-based techniques (Bing Liu, 2012) (Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, 2013). All the approaches to mine reviews and opinions need data in a prescribed format. But the data obtained from SNS, blogs and forum etc may not contain data in prescribed format. As users have liberties to use their own form of languages on SNS, blogs and forums it is essential to convert the data on users' form of languages into prescribed format. Thus the proposed work involves in transforming informal words into formal words before to obtain opinions and reviews from the tweets.

The proposed work is planned to use tweets for mining reviews and opinions. Tweets, the message sent by the users of Twitter which are limited to 140 characters strictly. The choosing of short length tweets for my research will make my job easier where other SNS will not limit their users in accordance with their messages

Corresponding Author: Kumaragurubaran, T., Research Scholar, Anna University, Regional Office, Madurai. India.
Tel: +91 9841237361; E-mail: tkgcse@gmail.com

(Alexander Pak and Patrick Paroubek, 2010) (Barbosa,L and Feng,J., 2010) (O'Connor,B., Balasubramanyan,R., Routledge,B,R., Smith,N,A., 2010) (Go.A, Bhayani.R,Huang.L, 2009).

As the tweets are limited with 140 characters users tend to use shorthand characters or they may use numbers instead of letters or words to comprise their thought within the limit (Sagar Bhutta, Avit Doshi, Uchit Doshi, Meera Narvekar, 2014). Though some users are within limit, due to their exclamation they may use repetition of letters in a word. To continue the research to obtain sentiment, opinion and reviews the researchers have to change the tweets in the above said informal forms to the formal one.

Once the proposed system is failed to resolve the problem with the written procedure a hominoid approach is proposed to rectify the issues. If a user does not understand a particular text while receiving a message he/she will ask the originator of the message to clear the issue and he/she may not get confused if they receive the same text again. They may remember the text from the past incidence. The same idea will be imposed in the proposed system. Though we have many inbuilt files to classify formal and informal words and if the system finds difficult to understand a text, the system will approach human. Once the human have cleared the issue the system is advised to store the actual meaning of the text on its allocated files. If the system faces the same problem again it will search for the file which contains the actual meaning and do use that in run time. Thus a real time elucidation will be followed in the proposed system.

Moreover the tweets will have 'tweet' and 'retweet' symbols and it may have some external links and this may be a barrier in doing the research for obtaining sentiment and opinions (Sagar Bhutta, Avit Doshi, Uchit Doshi, Meera Narvekar, 2014).

MATERIALS AND METHODS

The proposed work on transforming informal words into formal words exhibits three types of informal words that may appear on tweets and they are exaggeration words, shorthand words and alpha-numeric words. On transforming informal words into formal words removal of symbols and links on tweets are too planned.

Purging of symbols and links on tweets:

Before to precede the research the symbols like "?, !, @, #, $, %, &, ', ", /, \" which are present on tweets are planned to remove initially. In the mean time symbols like "tweet, re-tweet" and links that are present on tweets are also planned to remove. In this example, "@whatdelicious Can we push our meeting today back to 4.30?" the term "@Whatdelicious" and in "it's not the Amish you need to worry about; it's the zombies facebook.com/events/3028379…" the link "facebook.com/events/3028379…" are not necessary to mine opinions and reviews. Thus these symbols and links on tweets are extracted before the words in the tweets are transformed from informal words into formal words.

Techniques imposed on proposed system to classify informal and formal words:

To transform informal words in to the formal words the system must equipped with knowledge that to classify the difference between the formal and informal words. To provide the system with that knowledge a dictionary will be chosen and the words in the dictionary are copied to a file. To speed up the system, to scan all the words in the dictionary, separate files are mentioned to store the words that have started with particular alphabet character. By doing so, if the system has to analyze a particular word is a formal one, it is enough for the system to scan a single file that contains words that starts with the starting letter of that particular word. Thus twenty six separate files were created each one for the different twenty six alphabet characters in English language. These twenty six separate files are grouped as single package and this package will get included in the algorithm where opinion mining is performed. The package will help the system to classify a word which will be a formal or informal word.

Meanwhile adding packages to the algorithm which proposed to mine opinions and reviews from the tweets a DB file is planned to add with the algorithm. A Db file which consists of informal words in a column and the appropriate formal words in the next. A detailed survey is made which collects the informal words and its appropriate formal words from the public and those words are depicted in DB file. In this survey it is not possible to collect all the informal words and formal words used by the public and on considering this hominoid approach is made and possibilities of updating the DB file is made on proposed system.

Once found the given word in a tweet is an informal one it is essential to classify what kind of informal word that is. The alpha-numeric word which comprises of alphabets and numbers will pursue the following technique on transforming those informal words into formal words. For example one that have to wish good night on his/her tweet they may use the informal word '9t' instead of the formal word 'night' because both sounds similar. While eliminating this kind of informal words its essential to consider the tweets that contains numeric or numerals text. If eliminate those numeric or numerals text from the tweets it will sure change the meaning of the tweet or the receiver may not receive the proper information via tweets. For example consider the following tweet, "the parcel will arrive by 9am tomorrow", here eliminating 9 from the tweet as considering '9am' as informal text, the tweet turns meaningless.

Advances in Natural and Applied Sciences, 8(16) Special 2014, Pages: 40-43

The proposed system will follow a technique that if the text is a pure numeric the system will leave the numeric as itself. Suppose if the text is a alphanumeric and the numbers is followed by the text like ‘am’, ‘pm’, ‘kg’, ‘gm’, ‘mm’, ‘cm’, and degrees like Fahrenheit or Centigrade, signs like INR and dollars, powers etc, the system will leave the text as itself. If suppose an alphanumeric text comes other than the above discussed formats, the system will check the DB file and replace the particular alphanumeric informal words with the proper informal words mentioned in the file

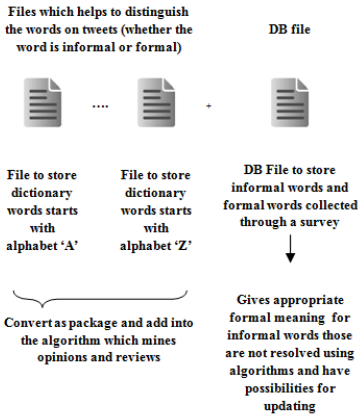

Fig. 1: Creation of Dictionary files and DB files to classify informal words and formal words Renovate alpha-numeric words on tweets in to formal words.

Renovate exaggeration words on tweets in to formal words:

Some times we come across ‘goooood’ instead of the formal word ‘good’ due to the excitement of the users on their tweets. On this case when we predict a particular character comes repeatedly, the repeated character on that word will get deleted from its last one at a time and do check the remaining text with the dictionary. Still the text is an informal one the above process will do continue until the word became formal one.

Renovate shorthand words on tweets in to formal words:

The informal words other than the mix of alpha-numeric and exclamation words are considered as generic informal words. Due to the speed needed in typing the message and to send the messages in hurry public do use short hand characters on their messages. For example the public used to type ‘bt’ for ‘but’ or ‘gr8’ for ‘great’. They usually go with the short hand characters which sounds similar to the formal words. We have made a survey about the informal words used by public and store those words on a DB file along with their actual formal words. Once the system come across these types of informal words it checks for the formal word in the file and uses them in tweets instead.

Though the proposed work has made a survey and having a DB files which comprises of informal words along with its formal words, it is not appropriate that not all surveys are made in full. Thus a hominoid approach is made and if the proposed system cannot resolve informal words into formal words it will ask help from humans by prompting textboxes by which the user can enter appropriate formal words and the system will update that information on its database.

Results:

The proposed system on received the tweets, it will exclude the spaces and links in the tweets. Then it picks a word one at a time from tweets and distinguished that word from whether it is an informal word or a formal word. Once the proposed system found the word is an informal one, it will go through the methods discussed and transforms the informal words into formal words. If the proposed system found the methods discussed so far is not appropriate to resolve the issues it will perform a hominoid approach to resolve the issue. Finally the proposed system transforms every informal word present in the tweets to formal words and the transformed tweets will impose for sentiment analysis and opinion mining.

Discussion:

In past the research papers have discussed about how to mine opinions and reviews from the tweets. But none of the techniques discussed in the past will mine opinions and reviews from the tweets if the tweets are not in prescribed format. The proposed system speaks about how to convert the tweets into prescribed format by transforming informal words into formal words. The proposed system has to verify multiple inbuilt files to distinguish the words on tweets and the system has to come across DB files to substitute formal words for

informal words if necessary. Moreover the hominoid approach was proposed to make the system more reliable. The inbuilt files, DB files and the hominoid approaches may take considerable amount of time in transforming informal words into formal words. When considering the amount of time taken to resolving issues as a major problem with the proposed system it has the tendency to resolve any kind of informal words into formal words on fly. This paper will be a great support for the researcher those who have willingness to do their research on Sentiment analysis and opinion mining in future.

Conclusion:

The proposed system has the features that transforms the tweet which contains the shorthand words, alpha-numeric words, exclamation words, twitter symbols, twitter return symbols, external links in to a tweet which turns comfortable for the researcher to go further with their research to extract the sentiments and opinions from tweets that are posted by the public. Moreover the proposed system is accommodating humanoid behavior that makes the system to learn from its failure during runtime.

The proposed system is only limited with the 'tweet' messages and the 'twitter' social networking site. In future the proposed system is extended to use with other social networking sites like facebook, orkut and bebo etc.

Moreover the proposed system is made ready to work with English language. In future the proposed system will be extended to work as multilingual.

REFERENCES

Akshi Kumar and Teeja Mary Sebastian, 2012. "Sentiment Analysis: A Perspective on its Past, Present and Future", I.J.Intelligent Systems and Applications, 10: 1-14.

Alexander Clark, 2003. Pre-processing Very Noisy Text. ISSCO/TIM, University of Geneva. Proceedings of Workshop on Shallow Processing of Large Corpora.

Alexander Pak and Patrick Paroubek, 2010. "Twitter as a corpus for Sentiment Analysis and Opinion Mining", proceedings of the Seventh Conference on International Language Resources and Evaluation, 1320-1326.

Aqsath Rastis Naradhipa and Ayu Purwarianti, 2011. "Sentiment Classification for Indonesian Message in Social Media", International Conference on Electrical Engineering and Informatics, Indonesia.

Barbosa, L. and J. Feng, 2010."Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING, Poster, 36-44.

Bing Liu, 2012. "Sentiment Analysis and Opinion Mining - Synthesis Lectures on Human Language Technologies", Morgan and Claypool Publishers, 2012.

Chandrakala, S. and C. Sindhu, 2012. "Opinion Mining and Sentiment Classification : A Survey", ICTACT Journal on Soft Computing, 03: 01.

Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, 2013. "New Avenues in Opinion Mining and Sentiment Analysis", Knowledge base approaches to concept-level Sentiment Analysis, IEEE,1541-1672.

Efstratios Kontopoulos, Christos Berberidis, 2013. Theologos Dergiades, Nick Bassiliades, "Ontology-based sentiment analysis on twitter posts", Expert Systems with Applications, Elsevier.

Go, A., R. Bhayani, L. Huang, 2009. "Twitter Sentiment Classification using Distant Supervision". Technical report, Stanford Digital Library Technologies Project.

Liu, B., 2010. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second edition.

O'Connor, B., R. Balasubramanyan, B.R. Routledge, N.A. Smith, 2010. "From Twitter to Polls : Linking Text Sentiment to Public Opinion Time Series. AAAI.

Pang, B. and L. Lee, 2008. "Opinion Mining and Sentiment Analysis", Foundation and Trends in Information Retrieval, 2(1-2): 1-135.

Sagar Bhutta, Avit Doshi, Uchit Doshi, Meera Narvekar, 2014. "A Review of Techniques for Sentiment Analysis of Twitter Data", International Conference on Issues and Challenges in Intelligent Computing Techniques.

Tamilselvi, A., M. ParveenTaj, 2013. "Sentiment Analysis of Micro blogs using Opinion Mining Classification Algorithm", IJSR, 2: 10.