Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 549 – 554

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Role of Text Pre-Processing in Twitter Sentiment Analysis

Tajinder Singh and Madhu Kumari

National Institute of Technology, Hamirpur 177 005, India

Abstract

Ubiquitous nature of online social media and ever expending usage of short text messages becomes a potential source of crowd wisdom extraction especially in terms of sentiments therefore sentiment classification and analysis is a significant task of current research purview. Major challenge in this area is to tame the data in terms of noise, relevance, emoticons, folksonomies and slangs. This works is an effort to see the effect of pre-processing on twitter data for the fortification of sentiment classification especially in terms of slang word. The proposed method of pre-processing relies on the bindings of slang words on other coexisting words to check the significance and sentiment translation of the slang word. We have used n-gram to find the bindings and conditional random fields to check the significance of slang word. Experiments were carried out to observe the effect of proposed method on sentiment classification which clearly indicates the improvements in accuracy of classification.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Classification, CRF, n-Gram, Sentiment, Text Pre-Processing.

1. Introduction

Since the early 1990s the use of internet has increased in different forms. People are communicating with each other using various appearances. In the past era the traffic has become almost the double on internet[3]. With this growth of internet traffic different online social networks such as Facebook, Twitter, LinkedIn, etc are also becoming famous. This in the digital world, things are changing in a very small time and become popular and trendy over OSN (Online Social Network). Different practices of sharing and communicating are not based the content but also on the basis of repetition of the content[4]. In the recent era micro-blogging has become very common[21] and popular platform for all online users. Millions/Billions of users are sharing their opinion on various aspects on very popular and trendy websites such as twitter, Facebook, tumbler, flicker, LinkedIn etc.[5] Twitter is a famous micro-blogging and social networking service which provides the facility to users to share, deliver and interpret 140 words' post known as tweet[3,6]. Twitter have 320 M monthly active user. Twitter is accessible through website interface, SMS, or mobile devices. 80% users are active through mobiles[7]. In the micro-blogging services users make spelling mistakes, and use emoticons for expressing their views and emotions[13]. Natural language processing is also playing a big role and can be used according to the opinions expressed[17].

*Corresponding author. Tel.: +91-9882551893.
E-mail address: madhu.jaglan@gmail.com

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.095

Table 1. Twitter's User Distribution.

Twitter Distribution	Total
Monthly Active users	320 M
Active users on mobile	80%
Language Supported	35+
Unique visits monthly to sites with embedded Tweets	1 B

Table 2. Social Text Quality Challenges.

Challenge	Description
Stop List	Common words frequency of occurrence
Lemmatization	Similarity detection of text/words
Text Cleaning	Removal of unwanted from the data
Clarity of Words	To clear the meaning in text
Tagging	Predicting data annotation and its characteristics
Syntax/Grammar	Scope of ambiguity, data dependency
Tokenization	Various methods to tokenize words or phrases
Representation of Text	Various methods and techniques to represent text
Automated Learning	Similarity measures and use of characterization

2. Related Work

Due to irregular, short form of text (hlo, whtsgoin etc.), short length and slang text of tweets it is challenging to predict polarity of sentiment text. In sentiment a mixture of applications are needed to study and these all demands large number of sentiments from sentiment holder. A summary of sentiment is needed, as in polarity disambiguation and analysis; a single sentiment is not adequate for decision. A common form of sentiment analysis is aspect based e.g. phone, quality, voice, battery etc.

Rafael Michal Karampatsis[8] *et al.* described the twitter sentiment analysis for specifying the polarity of messages. They used the two stage pipeline approach for analysis. Authors used the sum classifier at each stage and several features like morphological, POS tagging, lexicon etc are identified.

Joao Leal *et al.*[11] worked to classify polarity of messages by using machine learning approaches. Joachim Wagner *et al.* described work on aspect based polarity classification by using supervised machine learning with Lucie Flekova *et al.*[10] also worked on sentiment polarity prediction in twitter text.

Nathon Aston *et al.*[3] worked on sentiment analysis on OSN. They used a stream algorithm using modified balanced for sentiment analysis. Lifna C.S.[4] puts forward a novel approach where the various topics are grouped together into classes and then assign weight age for each class by using sliding window processing model upon twitter streams. In the similar way Emma Haddi *et al.*[12] discussed the role of text pre-processing for sentiment analysis.

Efthymios Kouloumpis[14] defined and explained three way sentiment analysis in twitter for identify positive, negative and neutral sentiments. Efstratios Kontopoulos[16] proposed a novel approach for analysis of sentiment. The approach is ontology based and it simply find out the sentiment score as well as grade for each distinct notion in the post.

3. Challenges of Social Text Quality

In most of the social media, language used by the users is very informal[15]. Users create their own words and spelling shortcuts and punctuation, misspellings, slang, new words, URLs, and genre specific terminology and abbreviations. Thus such kind of text demands to be corrected. Thus for analysing the text HTML characters, slang words, emoticons[19], stop words, punctuations, urlsetc are needed to be removed. Splitting of attached words are also be noticed for cleansing. Fangxi Zhang *et al.*[9] used Stanford Parser Tools1 for POS tagging and for parsing while the Natural Language Toolkit2 was used for removing stop words and lemmatization.Users who are also rating the product, services and facilities provided by various websites are needed to be addressed. Various systems for analysing users behaviour, views, attitude are needs to be analysed and demands to be normalized. Various shopping and

customer services supporting websites used various scales like star scale system[18] where the highest rating has 5 stars and the lowest rating has only 1 star, binary rating system where 0 and 1 etc. are used which demands to be normalized.

3.1 Text normalization

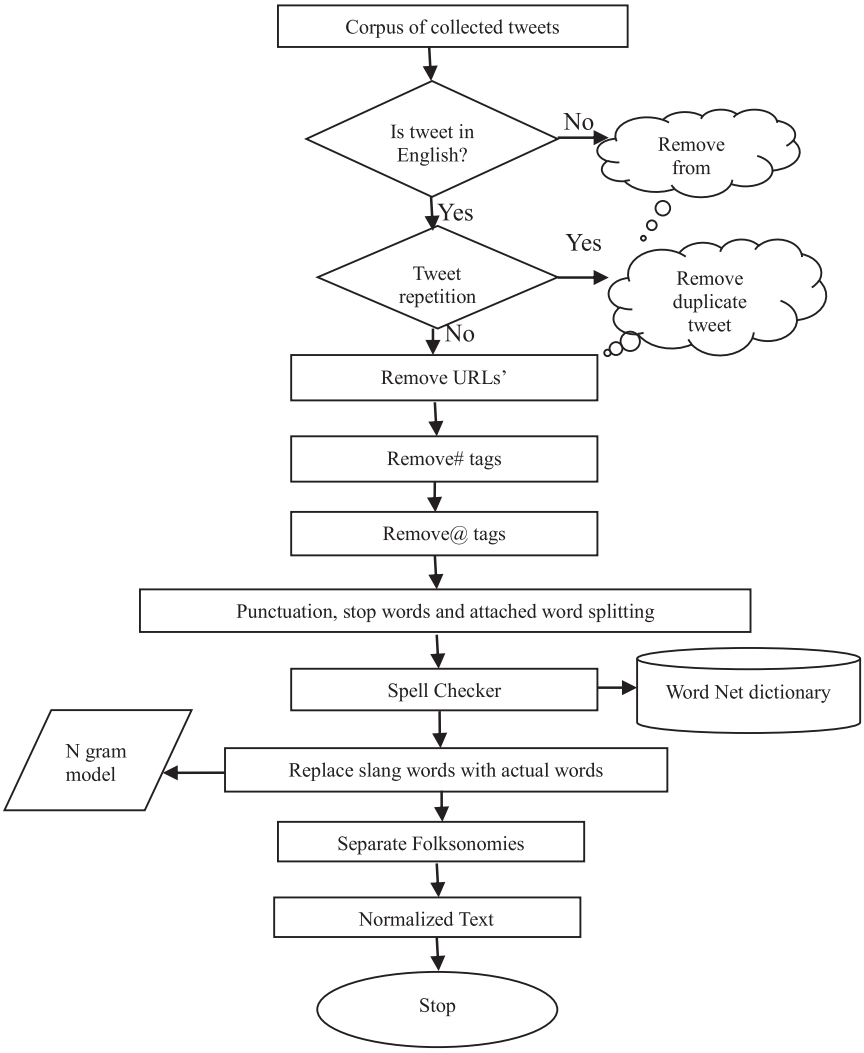

Fig. 1. Text Normalization as Process.

4. Proposed Scheme

The algorithm to deal with slang and identified words of short text massages of twitter used the coexistence of these words with different entities then decides the significance of slang words based on the sentiment strength and probability of co-occurrence of binding words with slang and unidentified words. Various steps involved in the proposed scheme is given below.

Assumption

Two unidentified words cannot be consecutive in a tweet; binding of this word is spread up to maximum two neighboring words.

Input

Tweet which is having unidentified word, slang word (W_s) and Folksonomies (except emoticons).

Output

Insignificance/significance of slang word and if slang is found insignificant then it weeded out from tweet else it is replaced with positive or negative score with respect to hash tag of the tweet.

4.1 Procedure

Step 1: Find the close binding of the slang word with different senses (coexisting) present in collected tweets so far based on bigram and trigrams language models.

Let W_s be unidentified word and $W_x$is word sequence which coexist in collected tweets.

Bigram: If we consider bigram language model then $\mathrm{mod}(W_x) = 1$, then this word can occur. W_x Can occur before or after W_s. Collecting prospective sense binding vector $C(w)$, where w is a ordered pair of words and its associated probabilistic weight

$$C(w) = \{all(W_x, P(W_x, P(W_x, W_s) \text{ where } P(W_x/W_s) \text{ or } P(W_s/W_x) > 0\}$$
$$P(W_x, W_s) = P(W_x)^* P(W_s/W_x) \text{ if } W_x \text{ occurs before } W_s.$$
$$P(W_s, W_x) = P(W_s)^* P(W_x/W_s) \text{ if } W_s \text{ occurs before } W_x.$$

- If W_X occurs before and after W_s, then we can use the following equations to resolve this situation:

$$\mathrm{Max}(P(W_x, W_s), P(W_s, W_x)$$

Trigram: If we consider trigram sense binding vector $C_{(w)}$ where w is a ordered triplet of ordered pair of words and their associated probabilistic weight with Ws.

$$C_{(w)} = \{all((W_{x1}, W_{x2}), P(W_{x1}, W_{x2}, Ws)) \text{ Where } P(W_{x1}/W_s W_{x2}), P(W_s/W_{x1} W_{x2}) \text{ or } P(W_{x2}/W_s W_{x2}) > 0\}$$

- If W_s occurs within W_{x1} and W_{x2} the following combination:

$$W_s W_{x1} W_{x2}, W_s W_{x2} W_{x1}, W_{x1} W_s W_{x2}, W_{x2} W_s W_{x1}, W_{x1} W_{x2} W_s \text{ and } W_{x2} W_{x1} W_s,$$

- Then

$$P(W_{x1}, W_{x2}, W_s) = \max(P(W_s W_{x1} W_{x2}), P(W_s W_{x2} W_{x1}), P(W_{x1} W_s W_{x2}), P(W_{x2} W_s W_{x1}), P(W_{x1} W_{x2} W_s), P(W_{x2} W_{x1} W_s)).$$

At this stage no filtering is done. We try to collect possible bindings.

Step 2: Analysis of these binding of slang word based on fields associated with coexisting words using Conditional Random Fields (CRF) is done at this stage to decide the significance of W_s. Using CRF Part of Speech (POS) tagging of the tweet which contains W_s is done then significance of W_s the measured using following rules:

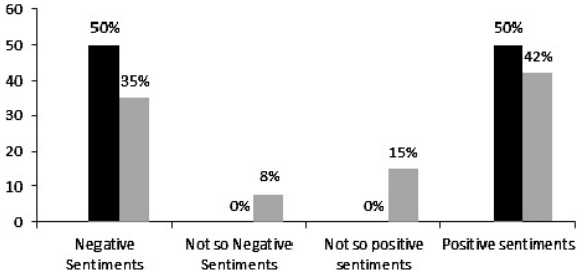

Fig. 2. Distribution of Sentiments over Dataset before and after the Pre-processing.

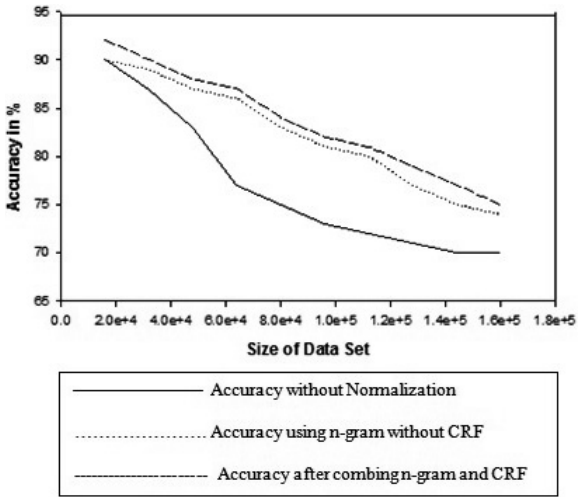

Fig. 3. Accuracy of Proposed Scheme under Different Variants of Preprocessing.

a. If W_s occurs before and after a proper noun then it is significant.
b. If is coexisting with collective noun and has reference to a proper noun then W_s is less significant.
c. Else W_s is insignificant and W_s can be weeded out of tweet.

Step 3: After collecting the significant all W_s these words are replaced by positive/negative sentiment scores with respect to the concept present in the binding set computed in step1. Following procedure is used to compute the sentiment score (Senti_Score) of W_s.

$$\text{Senti\_Score}(W_s) = \max |(P(W_s, W_x) \times \text{senti}(W_x))|$$

$P(W_s, W_x)$ is computed in first step and W_x is a vector, where $\text{senti}(W_x)\ \max[t(x_i)]$. $t(x_i)$ is sentiment of the tweet in which x_i word which is component of W_x is present sentiment of tweet which has W_s is updated as follows:

$$t(W_s) = t_{\text{old}} \pm \text{Senti\_Score}(W_s)$$

t_{old} is earlier sentiment value of tweet which was holding W_s.

5. Experiments and Results

For experimentation we have used twitter corpus data. More description of data can be found in[2], this data comprises of six fields, first field is sentiment class of the tweet which are negative, neutral and positive, represented by 0, 2 and 4 respectively, rest of field are the id of the tweet, the date of the tweet, the query, the user that tweeted, the text of the tweet. In order to evaluate and measure the impact of proposed scheme on the sentiment classification task we have used Support Vector Machine (SVM) based classifier. We carried out experiment in to two phases, in the first phase we applied the proposed scheme of normalization to the tweets' text by ignoring their sentiment class. After the normalization process we consider the sentiment class and class 2 i.e. is resolved in to new classes as: 1 (less negative) and 3 (less positive) based on the sentiments of unidentified (slang) words.

Results of experiments clearly suggest that proposed scheme not only robust to size of data but also perform better in terms of accuracy of sentiment classification.

6. Conclusions and Future Scope

This work is to analyse the impact of pre-processing and normalization on short massages like tweets which are full of information, noise, symbols, abbreviations, folksonomy and unidentified words. Looking at the interestingness to interpret the slang and unidentified words in tweets towards the sentiment, this paper focuses to identify the importance of slang words and to measure their impact on sentiment of the tweet. The proposed scheme used in this paper first gathers the coexisting words with the slang and then exploits characteristics of these binding words to define the significance and sentiment strength of slang word used in the tweet which not only facilitate the better sentiment classification but also ensure the sturdiness of classier as shown in the results. It is yet to be seen the how well the proposed scheme will perform with different classifiers on text streams.

References

[1] http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
[2] Twitter Sentiment Classification using Distant Supervision
[3] N. Aston, T. Munson, J. Liddle, G. Hartshaw, D. Livingston and W. Hu, Sentiment Analysis on the Social Networks Using Stream Algorithms, *Journal of Data Analysis and Information Processing*, vol. 2, pp. 60–66, (2014).
[4] C. S. Lifna and M. Vijayalakshmi, Identifying Concept-Drift in Twitter Streams, *ICACTA-2015*, Elsevier, (2015).
[5] Ayushi Dalmia, Manwitter Sentiment Analysis Thish Gupta, Vasudeva Varma, The Good, the Bad, and the Neutral, Sem Eval (2015).
[6] Santhi Chinthala, Ramesh Mande, Suneetha Manne and Sindhura Vemuri, Sentiment Analysis on Twitter Streaming Data, *Springer International Publishing Switzerland*, (2015).
[7] http://twittercommunity.com
[8] Rafeal Mcheal Karampatsis, John Pavlopoulos and Prodromos Malakasiotis, Sentiment Analysis Two Stage Sentiment Analysis of Social Network Messages, *SemEval*, (2014).
[9] Fangxi Zhang, Zhihua Zhang and Man Lan, ECNU: A Combination Method and Multiple Features for Aspect Extraction and Sentiment Polarity Classification, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 252–258, 23–24 August (2014).
[10] Lucie Flekov, Oliver Ferschk and Iryna Gurevych, UKPDIPF: A Lexical Semantic Approach to Sentiment Polarity Prediction in Twitter Data, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp. 704–710, 23–24 August (2014).
[11] Joao Leal, Sara Pinto, Ana Bento and Hugo Gonçalo Oliveira, Paulo Gomes, CISUC-KIS: Tackling Message Polarity Classification with a Large and Diverse set of Features, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 166–170, 23–24 August (2014).
[12] Emma Haddi, Xiaohui Liu and Yong Shi, The Role of Text Pre-processing in Sentiment Analysis, *First International Conference on Information Technology and Quantitative Management*, Elsevier, (2013).
[13] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, Sentiment Analysis of Twitter Data, Department of Computer Science Columbia University New York, NY 10027 USA.
[14] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, (2011).
[15] Sara Rosenthal, Alan Ritter, Preslav Nakov and Veselin Stoyanov, SemEval-2014 Task 9: Sentiment Analysis in Twitter, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 73–80, 23–24 August (2014).
[16] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades and Nick Bassiliades, Ontology Based Sentiment Analysis of Twitter Posts, *Expert Systems with Applications*, vol. 40, issue 10, pp. 4065–4074, August (2013).
[17] Chetashri Bhadane, Hardi Dalal and Heenal Doshi, Sentiment Analysis-Measuring Opinions, *International Conference on Advanced Computing Technologies and Applications (ICACTA)*, vol. 45, pp. 808–814, (2015).
[18] Xing Fang and Justin Zhan, Sentiment Analysis Using Product Review Data, *Journal of Big Data, 2015 Springer*, (2015).
[19] Xia Hu, Jiliang Tang, Huiji Gao and Huan, Liu, Unsupervised Sentiment Analysis with Emotional Signals, *Proceedings of the 22nd International Conference on World Wide Web, WWW'13, ACM*, (2013).
[20] Ana Mihanović, Hrvoje Gabelica and Živko Krsti, Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media, MIPRO 2014, 26–30 May 2014, Opatija, Croatia, pp. 1463–1468, (2014).
[21] Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece and Irena Spasic, The Role of Idioms in Sentiment Analysis, *Expert Systems with Applications*, Elsevier, (2015).
[22] http://www.iprospect.com/en/ca/blog/10-sentiment-analysis-tools-track-social-marketing-success/
[23] http://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/