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.

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

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