Journal of Computational Science 21 (2017) 316–326

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Extraction of emotions from multilingual text using intelligent text processing and computational linguistics

Vinay Kumar Jain[a,*], Shishir Kumar[a], Steven Lawrence Fernandes[b]

[a] *Department of Computer Science & Engineering, Jaypee University of Engineering & Guna (M.P.) , India*
[b] *Department of Electronics & Communication Engineering, Sahyadri College of Engineering & Management, Mangalore, Karnataka, India*

ARTICLE INFO

Article history:
Received 16 July 2016
Received in revised form 5 December 2016
Accepted 28 January 2017
Available online 1 February 2017

Keywords:
Emotion extraction
Machine learning
Text mining
Twitter
Classification
Natural language processing

ABSTRACT

Extraction of Emotions from Multilingual Text posted on social media by different categories of users is one of the crucial tasks in the field of opining mining and sentiment analysis. Every major event in the world has an online presence and social media. Users use social media platforms to express their sentiments and opinions towards it. In this paper, an advanced framework for detection of emotions of users in Multilanguage text data using emotion theories has been presented, which deals with linguistics and psychology. The emotion extraction system is developed based on multiple features groups for the better understanding of emotion lexicons. Empirical studies of three real-time events in domains like a Political election, healthcare, and sports are performed using proposed framework. The technique used for dynamic keywords collection is based on RSS (Rich Site Summary) feeds of headlines of news articles and trending hashtags from Twitter. An intelligent data collection model has been developed using dynamic keywords. Every word of emotion contained in a tweet is important in decision making and hence to retain the importance of multilingual emotional words, effective pre-processing technique has been used. Naive Bayes algorithm and Support Vector Machine (SVM) are used for fine-grained emotions classification of tweets. Experiments conducted on collected data sets, show that the proposed method performs better in comparison to corpus-driven approach which assign affective orientation or scores to words. The proposed emotion extraction framework performs better on the collected dataset by combining feature sets consisting of words from publicly available lexical resources. Furthermore, the presented work for extraction of emotion from tweets performs better in comparisons of other popular sentiment analysis techniques which are dependent of specific existing affect lexicons.

1. Introduction

Emotion expression plays a vital role in various part of everyday communication. In past, various measures have been used to evaluate it, through a combination of indications such as facial expressions, gestures, and actions etc. Emotions extraction using facial, gestures and action are the part of digital image processing and computer vision [1]. Emotions extraction is more difficult from texts especially from multi-languages texts, like in posts on social media and customers' reviews. This type of data has presence of ambiguity and complexity of words in terms of meaning make them more difficult. Factors such as users writing style, politeness, irony, variability in language is one of the important problems in extraction of emotions [2]. A wide variety of state-of-art work has been carried out in the domain of opinions mining and sentiment analysis but limited research are focused on detection/extraction of emotions in multi-language text.

In English vocabulary, some words express emotion explicitly, whereas other words can be used to get across emotion implicitly depending on the context [3]. Emotion detection in the text has recently attracted the scientific community to explore meaningful inferences hidden in the data and help in decision-making [4]. Many authors classify emotions in multiple classes for a better understanding, like Strapparava and Valitutti have classified emotional words into two classes'direct affective words' and'indirect affective-words [2]. Emotion research is important for building affective interfaces. These affective interfaces provide better user experience in following areas such as Human–Computer Interaction (HCI), Text-to-Speech (TTS) synthesis systems and Computer-Mediated Communication (CMC) [5]. Computational techniques related to emotion extraction present in social media

* Corresponding author.
E-mail addresses: vinay2588@gmail.com (V.K. Jain), dr.shishir@yahoo.com (S. Kumar), steven.ec@sahyadri.edu.in (S.L. Fernandes).

http://dx.doi.org/10.1016/j.jocs.2017.01.010
1877-7503/© 2017 Elsevier B.V. All rights reserved.

Table 1
Basic emotion categories recognized by the different researchers.

Authors	Emotion class	Emotion labels
Tomkins [9]	8	Joy, anguish, fear, surprise disgust, interest, shame, anger
Izard [10]	10	Enjoyment, shame,fear, anger,surprise, interest, sadness,shyness, guilt, disgust
Plutchik [11]	8	Joy, sorrow, anger, fear, disgust, surprise, acceptance, anticipation
Ortony et al.et al. [12]	6	Joy, surprise fear, anger,sadness, disgust
Ekman [8]	6	Happiness, sadness, anger, disgust, surprise, fear
Muni [13]	8	Jugupsa (Disgust), Hasya (Mirth), Krodha (Anger), Rati (Love), Utah (Energy), Bhaya (Terror), Vismaya (Astonishment), Soka (Sorrow)

have paying attention on basis of multiple emotion modalities [6]. However, only limited work has been done in developing automatic emotion recognition system [4,6].

The multilanguage text contains emotional words of different languages and extraction of these emotional words definitely improve emotion identification ratio [7]. In most of the available literature, theses words are treated as stop words in social media data [7]. This paper presented an advanced framework for automatic detection of emotions in Multilanguage text data. The emotion models used for development of proposed framework deals with linguistics and psychology. Proposed framework uses Machine Learning techniques for learning and validation and effective pre-processing Natural Language Processing (NLP) techniques for better extraction of emotions existing in the text.

This paper uses the concept of emotion model given by Ekman [8] as a basis with multiple feature sets to deal with multilingual data. The text under study comprises data collected from Twitter in three different domains such as Political election, Healthcare, and Sports. The first task is to collect real-time data consisting of relevant keywords. Through this paper, a novel technique based on RSS (Rich Site Summary) feeds to collect keyword which has been used for real-time data collection of events, has been introduced. Tweets containing images and emoticons are not considered under the scope of proposed approach. The effective pre-processing technique has been used to filter out irrelevant words and preserving words representing emotion of other languages. The classification of the dataset has been performed using popular machine learning techniques.This work represents the first systematic evaluation of emotion detection in real-time multilingual data in multiple domains. Another key contribution of the presented work is the practical application of emotion models in comparison of corpus-driven approach which assigns affective orientation or scores to words and word frequencies.

The rest of the paper has been organized as follows. State-of-art methods have been presented in Section 2. Proposed data collection methodology has been presented in Section 3. The problem formulation, existing methods, and proposed framework of emotion extraction system have been presented in Section 4. Experimental setup and outcomes with discussions have been presented in Sections 5 and 6. In Section 7 advantages of proposed approach over state of art, methods have been identified. Finally, precise conclusions and scope of future work are mentioned in Section 7.

2. Related work

Nowadays, a lot of research articles have been published for analyzing sentiments in social media data in multiple domains. This literature review section discussed emotion extraction methods and sentiment classification methods related to different domains like election prediction, healthcare, and sports analytics.

2.1. Emotion extraction methods

Researchers have investigated basic human emotions in different categories that are accepted universally. A number of related works in the field of emotion identification, reported in the literature, has been presented in Table 1.

Ekman's emotion theory [8] is the most popular and widely used approaches related to emotion recognition. Human emotions can be recognized using, speech, facial expression, gestures, and writings [4,5]. Research in emotion identification has focused on all these aspects [14]. Finding accurate emotions in the text contains evident in the vast body of research work related to different fields of psychology, social sciences, linguistics, Human–Computer Interaction (HCI) and communication. Table 2 outlines the significance of emotion detection and recognition techniques used in multiple domains.

2.2. Sentiment classification methods

The field of sentiment analysis recently witnessed a large amount of interest from the scientific community [37–40]. It deals for automatically determining the polarity of a textual data based on polarity, whether it is positive, negative, or neutral. More recently, much effort has been invested into the development of sentiment analysis methods in comparison to emotion extraction across multiple domains like movie and product reviews, election result prediction; disease outbreak, sports, stock market etc. [36,41,42]. In this paper, sentiment analysis techniques used in three specific domains such as Political election, Healthcare and Sports has been presented.

Tracking public sentiment during elections is a hot research area and prediction based on these sentiments is effective in comparison to survey-based methods. Nowadays, every election campaign has an online presence and users use social media platforms to express their sentiments and opinions towards political parties, leaders, and important topics during election [43]. Optimistic results regarding the predictive capacity of social media towards the election results in geographical regions are illustrated in Table 3. According to most of the authors, better prediction depends on the quality of data and merits of data collection methodologies [44]. If the data collected is not much relevant towards the event then the outcomes may be inappropriate [45]. Most of the authors focused on corpus-based feature and sentiment orientation technique for predicting the election outcomes, but emotion detection of social media users has not been taken into suitable consideration.

Another popular domain in the field of sentiment analysis is healthcare. Corpus-based features and sentiment based techniques have been providing a rich source of information for detecting and forecasting disease outbreak in all around the world [42]. Chew [71] used specific keywords related to outbreak detection in 2009 H1N1 pandemic.Hu et al. [72] used web services provided by Google related to influenza epidemic using specific keywords. Lampos and Cristianin [73] used content based methods with statistical methods to monitor and measure public perceptions. They also analyzed levels H1N1 pandemic. Chunara et al. [74] detected cholera outbreak using Twitter. Aramaki et al. used Support Vector Machine for predicting influenza rates in Japan [75].Stewart and Diaz [76] developed a real-time data analysis of disease using social media with an early warning system.Bodnar et al. [77] applied various