Journal of King Saud University – Computer and Information Sciences (2016) **28**, 330–344

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques

Muhammad Bilal *, **Huma Israr, Muhammad Shahid, Amin Khan**

CS/IT Department, IBMS, University of Agriculture, Peshawar, Pakistan

Received 31 July 2015; revised 14 October 2015; accepted 4 November 2015
Available online 12 December 2015

KEYWORDS

Roman Urdu;
Opinion mining;
Bag of words;
Naïve Bayes;
Decision Tree;
k-Nearest Neighbor

Abstract Sentiment mining is a field of text mining to determine the attitude of people about a particular product, topic, politician in newsgroup posts, review sites, comments on facebook posts twitter, etc. There are many issues involved in opinion mining. One important issue is that opinions could be in different languages (English, Urdu, Arabic, etc.). To tackle each language according to its orientation is a challenging task. Most of the research work in sentiment mining has been done in English language. Currently, limited research is being carried out on sentiment classification of other languages like Arabic, Italian, Urdu and Hindi. In this paper, three classification models are used for text classification using Waikato Environment for Knowledge Analysis (WEKA). Opinions written in Roman-Urdu and English are extracted from a blog. These extracted opinions are documented in text files to prepare a training dataset containing 150 positive and 150 negative opinions, as labeled examples. Testing data set is supplied to three different models and the results in each case are analyzed. The results show that Naïve Bayesian outperformed Decision Tree and KNN in terms of more accuracy, precision, recall and F-measure.

1. Introduction

Due to extensive use of computers, smartphones and high speed internet, people are now using web for social contacts, business correspondence, e-marketing, e-commerce, e-surveys, etc. People share their ideas, suggestions, comments and opinions about a particular product, service, political entity and current affairs. There are so many user-generated opinions available on the web. From all those opinions, it is difficult to judge the number of positive and negative opinions (Khushboo et al., 2012). It makes it difficult for people to take the right decision about purchasing a particular product. On the

* Corresponding author.
E-mail addresses: qec_mbilal@aup.edu.pk (M. Bilal), huma.israr@gmail.com (H. Israr), shahid_swabi@yahoo.com (M. Shahid), amin-khan@aup.edu.pk (A. Khan).
Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.11.003
1319-1578

other hand, it is also difficult for manufacturers or service providers to keep the track of the public opinions about their products or service and to manage the opinions. Similarly, an analyst wants to conduct a survey to get feedback of public on a specific topic. He/She will post the topic on a blog to analyze the sentiment of people about that topic. There will be so many opinions on that post. For all these opinions, it will be difficult to know how many opinions are positive and negative. So a computer machine may be trained to take such decisions in a quick and accurate manner.

The important thing in opinion mining is to extract and analyze the feedback of people in order to discover their sentiments. Growing availability of opinion-rich resources like online blogs, social media, review sites; raised new opportunities and challenges (Pang and Lee, 2008). People now can actively use information technologies to search the opinions of others.

There are many issues involved in opinion mining. The first is some words in opinion are representing a positive sense in one situation and negative in the other. For example consider an opinion "the size of this mobile is **small**". Here the word **small** comes in positive sense. On other hand, consider another opinion, "The battery time of this mobile is **small**". Here the word **small** is interpreted negatively (Rashid et al., 2013). Another issue in opinion mining is that most of the text processing system depends on the fact that a small difference in two sentences does not change the meaning very much. In sentiment analysis, the text "the movie was great" is different from "the movie was not great". People may have contradiction in their statements. Most of the reviews have both positive and negative comments, which is a bit manageable by analyzing sentences one at a time. However in more informal medium like facebook, twitter and blogs, lack of context makes it difficult for the people to understand what someone thought based on a short piece of text. One important issue in opinion mining is that product reviews, comments and feedback could be in different languages (English, Urdu, Arabic, etc.), therefore to tackle each language according to its orientation is a challenging task (Rashid et al., 2013).

Most of the research work in sentiment mining has been done in English and Chinese languages. Currently, limited research is conducted on sentiment classification for other languages like Arabic, Italian, Urdu and Hindi, etc. Urdu is an Indo-Aryan language which uses extended Persian and Arabic script. Roman script for Urdu does not have any standard for the spelling of the word. A word can be written in different forms with different spellings not only by distinct people but also by the same person at different occasions. Specially, there is no one to one mapping between Urdu letters for vowel sounds and the corresponding roman letters (Ahmed, 2009). There is no major difference in the pronunciation of Urdu and Hindi, therefore the roman version of Urdu and Hindi are written almost the same. Hence, this research is conducted in Roman Urdu and could be applicable in Roman Hindi. These are the most spoken languages in Pakistan, India, Bangladesh and among the people of these areas living in different parts of the world.

Previous work (Daud et al., 2014) conducted Roman Urdu opinion mining by using the key matching method. Adjectives of the opinions were matched with a manually designed dictionary to find polarity of that opinion. It was found that the accuracy of that work was low because the adjective alone cannot determine the polarity of an opinion. For example, consider a comment "I really **like** Iphone" here adjective is **Like** which has positive sense but on the other hand, consider another comment "I didn't **like** Iphone" here adjective is again **Like** which gives a positive sense but the comment interprets negative sentiments about Iphone. So it shows that all words of the opinions are equally important to indicate a comment either positive or negative. Thus the proposed model will use Bag of Words Model and three different classification techniques to improve the accuracy of Roman-Urdu sentiment classification.

The objectives of this research are to mine the polarity of public opinions written in Roman-Urdu with blend of English and Urdu extracted from a blog, to train the machine using a training data set, and to build Naïve Bayesian, Decision Tree and KNN classification models and to predict the polarity of new opinions by using these classification models.

This paper is organized into five sections. In the first and second sections the introduction and previous related work is briefly described. In the third section, the methodology adopted to perform different experiments is explained. In the fourth section, calculation and evaluation of experiments are performed to get various results and discussion on these results is conducted. In the last section, certain conclusions are drawn on the basis of outcomes.

2. Related work

In 2015, Daud et al. proposed a system called Roman Urdu Opinion Mining System (RUoMiS) which uses natural language processing technique to find the polarity of the opinion. In this study, the adjectives in the opinions were compared with a manually designed dictionary to find the polarity of the opinions. The results of the experiment were recorded with a precision of 27.1%, however, RUoMiS categorized about 21.1% opinions falsely. In 2014, Kaur et al. used a hybrid technique for Punjabi text classification (Kaur et al., 2014). In this research the combination of Naïve Bayesian and N-gram techniques were used. The features of the N-gram model were extracted and then used as training dataset to train Naïve Bayes. The model was then tested by supplying testing data. It was found that by comparing results from already existing methods, the accuracy of the proposed method was effective. Ashari et al. in 2013, used Naïve Bayes, Decision Tree, and *k*-Nearest Neighbor in searching for the alternative design by using WEKA as a data mining tool and developed three classification models (Ashari et al., 2013). Their experiments showed that the Decision Tree is fastest and KNN is the slowest classification technique. The reason they mentioned is that, in the Decision Tree, there is no calculation involved. The classification by following the tree rules is faster than the ones that need calculation in the Naïve Bayes and KNN. Moreover, KNN is the slowest classifier because the classification time is directly related to the number of data. If the data size is bigger, larger distance calculation must be performed and this makes KNN extremely slow. They concluded that Naive Bayes outperformed Decision Tree and KNN in terms of accuracy, precision, recall and F-measure. Jebaseeli and Kirubakaran in 2012 investigated the use of three classifiers namely Naïve Bayes, KNN and random forest for prediction of opinions as positive or negative about the M learning system for the