Chapter 76
Experimenting Language Identification for Sentiment Analysis of English Punjabi Code Mixed Social Media Text

Neetika Bansal
College of Engineering & Management, India

Vishal Goyal
Punjabi University, India

Simpel Rani
Yadavindra College of Engineering, India

ABSTRACT

People do not always use Unicode, rather, they mix multiple languages. The processing of codemixed data becomes challenging due to the linguistic complexities. The noisy text increases the complexities of language identification. The dataset used in this article contains Facebook and Twitter messages collected through Facebook graph API and twitter API. The annotated English Punjabi code mixed dataset has been trained using a pipeline Dictionary Vectorizer, N-gram approach with some features. Furthermore, classifiers used are Logistic Regression, Decision Tree Classifier and Gaussian Naïve Bayes are used to perform language identification at word level. The results show that Logistic Regression performs best with an accuracy of 86.63 with an F-1 measure of 0.88. The success of machine learning approaches depends on the quality of labeled corpora.

DOI: 10.4018/978-1-6684-6303-1.ch076

INTRODUCTION

According to statistical data reports, there are 525.3 million internet users currently in India. The use of social networking media has gained popularity since decades. Facebook is emerging as the most popular social networking site in the country. Other Social media networks include WhatsApp, Google+, and Skype. India ranks third with the most Instagram users with 69 million users. As per July 2019 statistical data, Instagram is one of the most popular social networks worldwide, especially where youngsters share selfies or other photographic content such as travel pictures, and moreover they try to keep up with favorite athletes and celebrities. Facebook and YouTube accounted for the largest penetration, both at 30 percent each as of the third quarter of 2017. T-Series is reported as the most subscribed YouTube channel till April 2019, with 92.38 million subscribers.

The number of people using social media as a part of their daily life is noticeably high and they use these platforms merely to share their experiences, preferences and opinions regarding the products or services of some brands. With analytics of data the buyers as well as sellers can enhance future decisions and actions. Electronic Word of Mouth (EWOM) has become important source of information for consumers and website owners. The analysis of data gives insight into consumer choice, brands and products. Sentiment analysis here emerges as a field for the digital world and helps website owners to design marketing strategies to increase the revenues.

Customer reviews available online have become valuable for customers and firms. The product reviews by users act as a valuable source of information for buyers in making product choices. On the contrary, reviews posted online act as feedback for firms as it requires huge advertising and huge investments. These firms attain publicity at large that too with no additional cost.

The post-visit brand image and the pre-visit image of a destination through EWOM has become important source for destination planning. Interested consumers use websites when planning their trips to a destination. Thus, businesses owners can utilize make the best use of such data. Furthermore, online reviews left by guests have business value in terms of understanding customer perceptions of hotel products and services attributes. Hoteliers can use this information to set priority rules for making improvements and use the generated electronic word of mouth effect from online customer reviews to enhance their performance. Online customer reviews have strong information. Online customer ratings and their likes and dislikes can be considered as indication of customer satisfaction.

Sentiment Analysis is used for analyzing trends, evaluation of public opinions, identifying ideological bias, targeting advertisements, analyzing reviews of a product and services, opinions and reactions to ideas. English being the dominant language around the world, therefore the work in the field of Sentiment Analysis is predominantly in English. There is absence of large volume of datasets, linguistic and lexical resources for Indian languages; posing challenges ahead. With the use of advanced Artificial Intelligence Techniques and advancements in deep learning algorithms, there has been considerable improvement in the field of sentiment analysis of textual data. Sentiment analysis has achieved higher accuracies with deep learning algorithms and the language identification work performed by authors can be used as base for Sentiment analysis.