ABSTRACT

Online platforms have opened up ample of opportunities for businesses to gain knowledge about the quality of their products and customers feedback through the examination of online reviews of the users. This study implements a sentiment classification and topic extraction system which accurately classifies the Amazon fine food online reviews into binary categories, positive and negative. Sentiment classification is performed via 2 approaches: Vectorized features processed with legacy Machine learning algorithms and deep learning techniques. Various algorithms like logistic regression, Support vector machine classifiers, decision tree classifiers are compared by measuring their performance in terms of how accurately they classify the reviews into positive or negative by learning from a labelled dataset and it is found that logistic regression performs the best on combination of review text and summary text. In this experimental study, the vectorization methods offer the flexibility of choosing or not choosing the order of words through the n-grams concept. In the non- neural network approach, vectorization methods used are countvectorizer and tfidfvectorizer which convert the review texts into document term matrix(numerical form) which is then used as input to drive the algorithms of machine learning. These methods are combined with the balanced and unbalanced datasets along with and without the extracted text meta data to examine the effects on the model accuracy in classifying the reviews. While in the neural network approach, we use the LSTM and GRU networks which take into account the word order or word semantics. Here, we test the role of pertained models like Glove embeddings in the classification of the reviews into positive or negative. Again, these tests are performed on balanced as well as original datasets along with and without the extracted text meta data. It can deduced from the results obtained, that deep learning approach using the LSTM performs the best on the labelled dataset and also that model performance is not much affected by the distribution of prelabelled data or the data skewness. It is also seen that combining the meta data such as review polarity, review length, review character length, etc with the reviews enhances the binary prediction accuracy. As second part of this study, the topic extraction system is implemented which categorizes the reviews into fixed numbers of categories where is each category is a collection of similar and related words which when combined, give an intuition about a topic that the category is speaking about. This task of interpretation is however manual.