Improvised Sentiment Analysis and spam detection using various learning algorithms

A thesis

submitted to

Jamia Millia Islamia

In partial fulfilment of the requirements of the award of the
Degree of Doctor of Philosophy in
Computer Engineering

by

GUNJAN ANSARI

under the supervision of

PROF. TANVIR AHMAD PROF. M.N. DOJA

Department of Computer Engineering
Faculty of Engineering and Technology
Jamia Millia Islamia
New Delhi

August, 2019

Abstract

Sentiment Analysis is computational study of opinion expressed in a piece of natural language text to identify the writer's attitude towards an entity. The machine learning algorithms have been employed to classify the expressed sentiments in the text as positive, negative or neutral. With explosive growth of users' posts on social media and its effect on business intelligence applications, the sentiment classification tasks have become technically challenging and practically useful in recent years. The key issues targeted in this thesis are spam detection, feature selection and aspect level sentiment analysis.

The positive or negative opinion posted by reviewers can greatly influence the target businesses. This is the main reason that make review sites more susceptible to spam attacks. The identification of spam reviews is a major task in the area of sentiment analysis to provide reliable results to the consumers. In this thesis, we proposed two novel approaches that contribute in the area of spam review detection. The first proposed approach improves content-based spam detection by selecting more relevant textual features to classify a review as spam or ham. The novelty in the approach is that it integrates local with global filter-based feature selection methods to select a more informative feature set from both real and synthetic datasets. We also designed review ranking system for spam recognition that scores each review on the basis of textual and meta-data features scraped from the crawled data of product reviews. The advantage of the proposed system is that it is unsupervised and avoids heavy computation of learning. The result analysis on reviews crawled from e-commerce site show that most of the top-ranked reviews were useful whereas the reviews ranked lower by the algorithm were spam and thus non-helpful to the buyers.

Due to increase of online content on e-commerce sites, feature selection has become a latest challenge for the researchers in the area of sentiment analysis. To overcome the problem of overfitting and high computational time in supervised learning algorithms, we proposed two different techniques for feature selection in this thesis. The first technique aggregates the score of five different feature ranking methods using hesitant fuzzy sets to generate robust feature set. The use of hesitant fuzzy sets for integration of filter methods results in an approach that is algorithmically simpler, less computationally complex and independent of any individual feature ranking method. Another proposed technique to improve feature selection is hybrid of filter and wrapper method for selecting more appropriate and non-redundant features. The features selected by filter-based methods are further optimized in our technique using two wrapper approaches: Recursive Feature Elimination (RFE) and Binary Particle Swarm Optimization (BPSO). The proposed model is suitable for large data size problems as it is able to achieve good performance with significant reduction in the number of features.

Aspect term extraction and its polarity detection is a central problem in the area of sentiment analysis. To address this issue, we also proposed a semi-supervised graph-based learning approach for aspect term extraction. In this approach, every identified token in the review document is classified as aspect or non-aspect term from a small set of labelled tokens using label spreading algorithm. The k-Nearest Neighbor (kNN) algorithm for graph sparsification is employed to make it more time and memory efficient. The proposed work is further extended to determine the polarity of the identified aspect terms from the review sentences to generate visual aspect-based summary of review documents.