Abstract

In Micro-blogging web services such as Twitter, the user is often bombarded with tons of information and raw data, with user unable to classify it into right category. The solution to overcome this problem can be derived from automatic text classification process. Social networking websites often limit their users to put up a short text message of length 140 characters only. Hence classifying this raw data continuously on these microblogging websites is a tedious task, as one has to deal with short text. Short text messages are difficult to classify as they have lack of semantic information and they have high risk of getting misclassified. In this research paper, a methodology has been developed that incorporates preparation of semantic database and then employ it to extract the necessary classification features from the database. This prepared database is then used for binary feature extraction from the set of user tweeted database hence the process of extracting features from the available database based on the semantic database approach has been presented. The basic of this paper is mainly focused on extracting nine features and then reducing the features to seven features using logical operations. The process of reducing the features not only reduces the complexity of the written code but also saves the database memory required to save the extracted feature for master training database. The features so extracted are easier to use and operation has less complexity of generation than compared to features generated by other available
Improved Information Filtering and Feature Dimensionality Reduction using Semantic based Feature Dataset for Text Classification: In Context to Social Network

algorithms like Bag-of-Words.

References

- www.twitter.com
- http://www.time.com/time/magazine/article/0,9171,1044658,00.html
- Yang, Lili, et al. &quot;Combining Lexical and Semantic Features for Short Text Classification. &quot; Procedia Computer Science 22 (2013): 78-86.
Index Terms

Computer Science
Information Sciences

Keywords

Text classification  short text  Twitter  semantic  Bag-of-Words