SENTIMENT ANALYSIS WITH MACHINE LEARNING TECHNIQUES

Abstract

Sentiment Analysis is one of Nature Language Processing to identify the attitudes and viewpoints in short text content, such as Twitter, reviews, etc. (Bhavitha et al., 2017). In this report, to investigate the polarity in the reviews of restaurants, (1) the preliminary selection and comparison of the most well-known sentiment analysis supervised machine learning models: Naive Bayes (NB), Supported Vector machine (SVM), Decision tree (DT), K Nearest Neighbors (KNN), Logistic Regression (LR) were made. (2) The contrasting feature engineering methods have been evaluated based on the selected model Logistic Regression LR and Linear SVM. (3) Optimized parameters on features selected and hyperparameters tuning on the LR model have been implemented with 86% accuracy. The model generalisation has been minimised with random shuffled 10 folds cross validation (4) Error analysis of major models were critically analysed, including model bias and variances, optimised split train and validate set with learning curves.

Literature survey

- Preprocessing text data by cleaning and formatting the unstructured texts to make effectiveness in utilising the sentiments from feature extractions. (Pradha et al., 2019). The author offered selections of preprocessing techniques, including lower text, removal of specific symbols, punctuations, stop word, hash bag, and white space, then tokenisation and lemmatisation/ or stemming was also effective on Nature Language Preprocessing (NLP) for sentiment analysis. By comparing the sentiments before preprocessing and manually preprocessing trained by multiped ML models, the SVM performs best with 90.3%, suggesting the preprocessing for sentiments improves accuracy and speed.

- The results from (Laoh et al., 2019)(Widiyaningtyas et al., 2019)(Kaur et al., 2018) proved that n-grams are effectively to be adopted in sentimental classification on short text content like review or Twitter by many machine learning models, such as SVM, KNN, NB LR and etc. The research of classification on movie review has shown a 1.908% improvement in SVM classifiers using a combination of unigram and bigram compared with unigram only, and a 2.352% increase in accuracy on Naïve Bayes model with a combination of unigram and bigram compared with unigram itself(Tripathy et al., 2016).

- This paper (Dhanani et al., 2018) offered conceptualised comparisons from traditional word embedding, including the one-hot-encoding, words of bags and new

word embedding like word2vec, which can be used for utilising the vector for sentiment analysis. In another paper(Katić & Milićević, 2018), besides the traditional word embedding methods, authors also provided bag-of-n-grams and their Term Frequency – Inverse Document Frequency (tf-idf) transformation, word2vec and tf-idf word2vec embedding techniques. A comparison was made between these techniques and combined with different classifiers. This research claims the Bag of n-grams features performed best with the Logistic regression (LR) model with 0.9231% accuracy compared to LSVM and MNB. On the other hand, a bag of n-grams + tf-idf performs best with SVM with 0.929 accuracy compared to LR and NB. The word2vec and word2vec+tf-idf only got 0.8437 and 0.8222 in accuracy on the LR model. However, the word2vec embedding technique achieved an accuracy of 0.9381 on the Convoluted Neural Network (CNN).

- The feature selection, like chi-square, improved the classifiers' NB performance by 20% in accuracy for sentiment analysis (Nurhayati et al., 2019). In another paper (Setiyaningrum et al., 2019), feature selection increases the efficiency of classification time cost in the KNN model when the accuracy remains still.

- As a couple of researchers demonstrated (Katić & Milićević, 2018)(Dey et al., 2020) (Petrescu et al., 2019)(Gupta et al., 2019), the NB, SVM and LR are commonly outperformed in the simple machine learning sentiment analysis. The other machine learning models like KNN and D-tree are sometimes adopted. Other advanced models like CNN and Long, Short-Term Memory (LSTM), Random Forest (RF), and ensembled learning methods are also outstanding in sentiment analysis.

Methodology

Preprocess

To enable the extraction of features from the raw text data, the cleaning and formatting processes need to be performed on the short text contents. The processes contain lowercase conversion, inferring space, removal of specific symbols, punctuations, stop word, hash bag, white space, and the following with tokenisation and lemmatisation/ or stemming.

- ***Inferring space***: Some words are connected due to typos or formatting errors. For example, 'Thedinnerwasdelicous!' There Should be a space in the middle to separate words like this, 'The dinner was delicious!'
- ***Lower text***: Converting 'Sky' to 'sky' helps to unify the text, which helps reduce the redundancy for the same word.
- ***Removal of symbols, punctuations, hash bag, and white space***: Since all these contents do not have an actual meaning as a word, they should be removed from the dictionary.
- ***Removal of the stop word***: This helps to reduce the words in English like 'a', 'an', 'this', etc., which do not contain a polarization degree but do have a higher frequency that should not appear in the features.
- ***Tokenization***: Separate sentences into word strings and store them into lists. Which is preparation for counting the appearance frequency or behaviour as base features.