SENTIMENT ANALYSIS BY COMBINING LEXICON-BASED & MACHINE LEARNING METHODS

Satyajit Narayanan, Jagadeesh Hariharan, Riddhiman Sherlekar, Eshan Kirpal, Venkatesh Nayak, Harsh Mehta

Guide: Dr. Ranga Raju Vatsavai

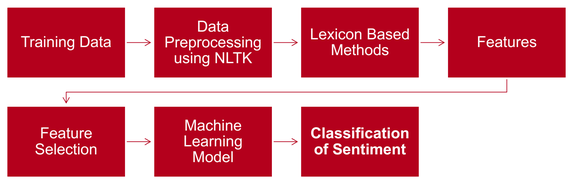
NC STATE UNIVERSITY

DEPARMENT OF COMPUTER SCIENCE

Introduction

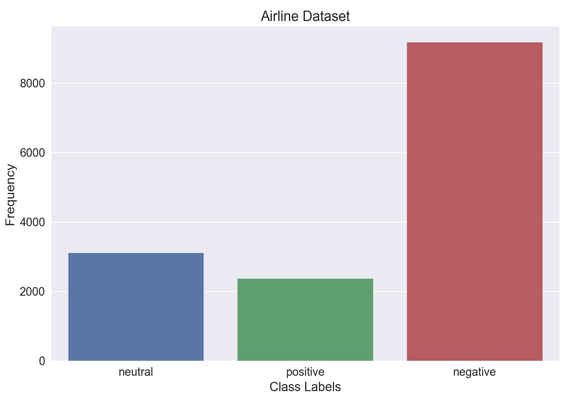
A supervised lexicon-based approach to extracting sentiments from a tweet is presented. A comparative study of existing techniques for opinion mining by combining machine learning and lexicon-based approaches, together with evaluation metrics is provided.

This is done by calculating the Semantic Orientation (SO) using dictionaries of words It incorporates intensification and negation. Using that as a feature along with other features, the algorithm for Decision Tree and Naïve Bayes to predict sentiment of US Airline tweets has been implemented. A representation of our approach is depicted below:



Data Preprocessing

Data used for this study is the Twitter US airline dataset consisting of labelled tweets. Data has three class labels – positive, negative, neutral. A visualization of frequency of each class label is as follows:

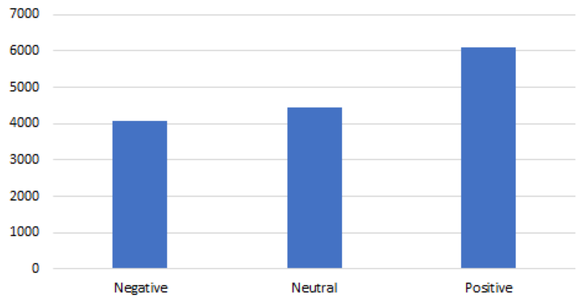


The dataset is preprocessed in NLTK for the following methods:

- Tokenization
- Lemmatization
- Part of speech tagging
- Case conversion

Lexicon Based Method

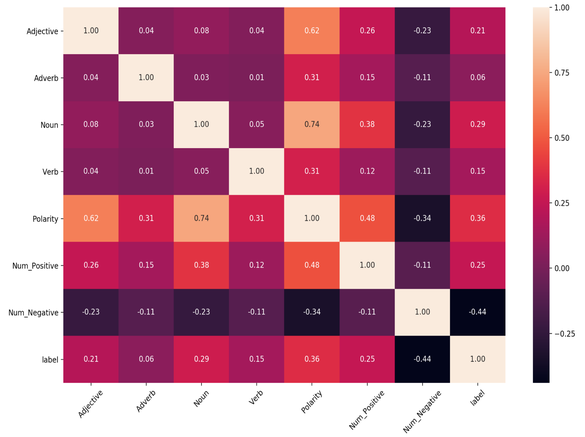
- The lexicon-based approach involves calculating orientation for a sentence from the semantic orientation of words in the tweet using preexisting dictionaries
- We created dictionaries for adjectives, verbs, adverbs and nouns with pre-defined semantic orientations. [1]
- The occurrences of parts of speech associated with the words would be an indicator of the sentiment of the tweet, which can be obtained using NLTK.
- Based on polarities of POS, the class labels are predicted, and the lexicon-based approach is validated.



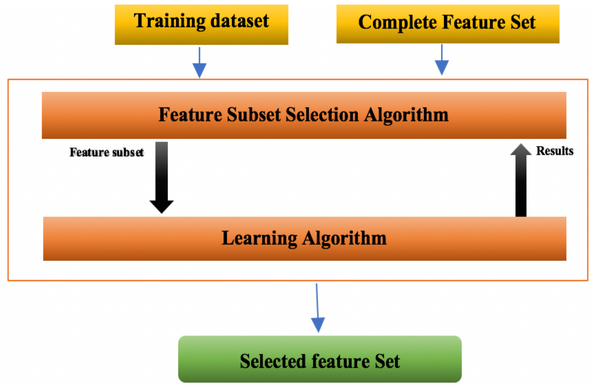
- From the graph it can be inferred that the dictionary-based approach may fail to find opinion words with domain and context specific orientations.
- Hence, in order to improve the model, the output of this method could be used as feature set for machine learning algorithms.

Feature Creation and Selection

- Apart from the four features (noun, verb, adjective, adverb) generated by the lexicon based approach, few more are created, like polarity of the sentence, number of positive words, number of negative and abusive words.
- The performance of any machine learning algorithm depends on the features given to the model.
- The correlation matrix can be formed to find out the relationship between each feature and the class label.
- Univariate and bivariate analysis were performed to understand relationship among features.



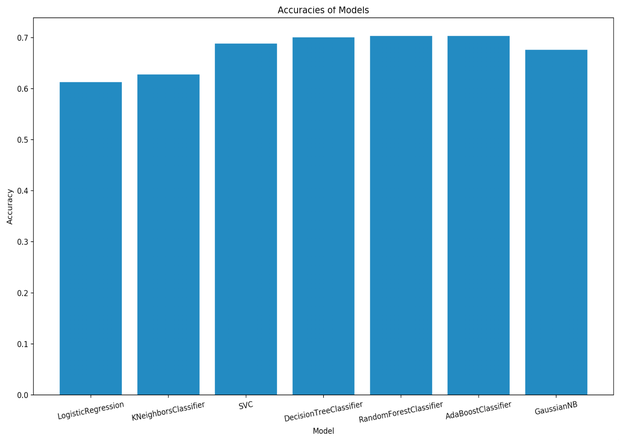
- Then, the wrapper feature selection process was used to narrow down our final list of features.



- The final list of features after performing the wrapper class are polarity, number of positive words & number of negative words.

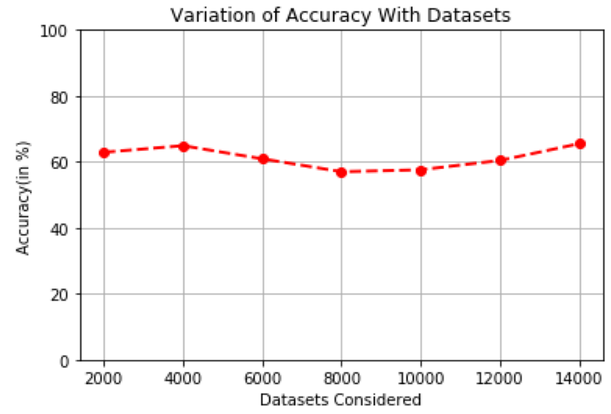
Comparative Study

- A study of the various supervised classifiers is completed to find out the best classifiers, which can be implemented. This is performed by using Python's scikit library.



Implementation of ML Model

- The ID3 algorithm of the Decision Tree has been implemented. The basic idea of this algorithm is to construct the decision tree based on the information gain criteria.
- Starting from the parent node, at each node a property is tested and based on the minimum entropy, the decisions are made. This process is recursively performed to develop a decision tree.
- The dataset is divided into test and train datasets with a proportion of 7:3.
- Entropy is used as the measure to decide the best split.
- The rules are extracted from the object, which are then used to assign class to the records of test data.
- The accuracy obtained through our implementation is 65.70 %.



Conclusion

- The study shows that the sentimental analysis performed using the combination of lexicon & machine learning performed better than suing only lexicon based approach.
- The decision tree implementation performed in this study can be improved by incorporating more features like n-grams, count of exclamation and interrogation marks, type of emoticons.

References

1. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M.. Lexiconbased methods for sentiment analysis. Computational linguistics, 2011.
2.Pavel Blinov, Maria Klekovkina, Eugeny Kotelnikov and Oleg Pestov. 2013. Research of lexical approach and machine learning methods for sentiment analysis. Computational Linguistics and Intellectual Technologies, 2(12):48–58.
3.Link to Project Code: CSC 522 Project Link