A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Spark

Ssd

This paper uses Apache Spark's scalable machine learning library (MLlib)
Also there is three classification techniques from the library are applied.
These techniques are :

- Naive Bayes
- Support vector machine
- Logistic regression

The results are evaluated by performing the accuracy metric

The customer online reviews have been increasing by a huge amount and getting various structures of data (structured, unstructured, semistructured), these changes made it a big data that needs the big data techniques.

Apache Spark, developed at the University of California, Berkeley's in 2009, is an open-source processing framework. It is used for large-scale data because it achieves high performance for both batch and streaming data and has easy-to-use APIs for operating on large datasets.
MLlib is Apache Spark's scalable machine learning library built on top of Spark to deliver both high quality and high speed.

MLlib can be used with Java, Scala, and Python, so that you can include it in complete workflows

This research aims to provide new experiments of sentiment classification on large-scale data using the Spark's MLlib by applying different MLlib

classification algorithms and evaluating their performance.

Related Work

many researchers have used many different methods and algorithms for performing sentiment analysis:

- Nabil et al. interested in testing the performance of different machine learning algorithms on a dataset of more than 10,000 Arabic tweets, they used different algorithms (SVM, MBN, BNB, KNN) for text categorising. They found that SVM is the best classifier
- Duwairi and Qarqaz applied three classifiers (KNN,SVM,NB) on tweets then compared the results using the (Precision and recall rates) to find out that KNN and SVM was better than NB
- Kang et al, proposed an improved Naïve Bayes algorithm to be used for the sentiment analysis of restaurant reviews based on senti-lexicon
- Omar et al, They adopted three classifiers as base-classifiers (Naive Bayes, Rocchio classifier and support vector machines) for Arabic customer reviews.The experimental results show that the ensemble of the classification algorithms with meta-learner improves the classification effectiveness.
- There is researchers that used Hadoop Map-Reduce for sentiment analysis like Liu et al, and Sehgal et al,

- Baltas et al, used MLlib for classification of sentiment analysis of Twitter data they implemented three classifiers (decision tree, Naive Bayes , logistic regression) on real data from Twitter The system is evaluated on different dataset sizes and different features. Naive Bayes were better than the other classifiers, the performance of classifiers could be affected by the size of the dataset.

- Adib et al, Apache spark used to identify malicious users and analyze their behavior to enhance the accuracy of trust. they obtain the distributed environment and speed using spark
- Pranckevičius et al, they implemented apache spark different classifiers and then evaluate it, they compared the performance using different dataset sizes and different number of n-gram. The result of this research is that logistic regression classifier scored higher compared with other classifiers and the classification accuracy increases when using combination of n-gram, and increasing the size of the data set has a positive impact on overall accuracy.

Proposed Approach

The steps of the used approach is:

- data preprocessing
- Feature extraction
- Applying the machine learning classifier (Naïve Bayes, Support vector machine and logistic regression) under Spark environment
- Evaluate the result using the accuracy metrics

DataSet

The data set used for experiments is the Amazon review polarity dataset, This dataset spans for 18 years.
The dataset includes 35 million reviews , it also include class, review title, and review text columns.
It contains 1,800,000 training samples and 200,000 testing samples in each class

Data preprocessing

The preprocessing step including the following phases:

- Removing null reviews (two reviews)
- Tokenization
- Noise removal

- Stop-words removal

Feature extraction

In the feature extraction phase the text is converted into a feature vector to be suitable for the classifier to deal with it, The Term Frequency-Inverse Document Frequency (TF/IDF) is used.
term frequency (TF): is the number of times that a word or term occurs in a text
inverse document frequency (IDF) : is a measure of how much information the word provides

Spark's MLlib Classifiers

- Naive Bayes Classifier (NB) : it is a simple algorithm based on probabilistic Bayes' theorem, NB constructs the model by adjusting the distribution of the number for each feature
- Support Vector Machine(SVM): it is a supervised machine learning used generally in classification problem, it tries to find a hyperplane represented by vectors that split the positive and negative training vectors of documents with maximum margin
- Logistic regression: it is used to predict a binary response based on one or more independent variables or features, the output of logistic regression always lies in [0, 1]

Experimental Results

The purpose of this experiment is to measure the performance of every classifier and compare the results

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

TABLE 2: EVALUATION RESULTS

	Accuracy
NB	85.4%
SVM	86%
Logistic Regression	81.4%