TWITTER SENTIMENT ANALYSIS REPORT
BY DEEPANK G

ABSTRACT

Sentiment analysis deals with identifying and classifying opinions or feelings expressed in a source text. Social media generates a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is useful in knowing the opinion of the crowd. Opinion investigation of Twitter data is a field that has been given much attention over the last few years and involves examining "tweets" (comments) and the content of these expressions.

PROBLEM STATEMENT

The task given to us was aimed at predicting the sentiment of the tweets to either positive, negative, or neutral using supervised learning algorithms. The dataset comprised of tweets (text data) scrapped from Twitter and the corresponding sentiment labelled as positive, negative, or neutral. The dataset initially had 2 columns and 27448 rows with a disk size of 2.07 MB.

PRE-PROCESSING

The dataset, in most parts, had to be cleaned for its use in predictions. Rows with missing values was dropped. Using regular expressions, special characters, numbers, punctuations, blank space, twitter handles, hashtags and URLs in 'text' column data was removed. 'sentiment' column data was label encoded for predictions. Next, stop words like the, or, of, and, etc. was removed as these words are generally large in number. Further, texts were normalized using *Lemmatization*. This is known to increase accuracy and learning by the model. Word (text) encodings were later carried out by *pipelining. Tf-Idf and BoW* (Bag of Words) were the techniques employed for text encodings.

VALIDATION STRATEGIES

The dataset was split randomly into training set (train set) and development set (dev set) in the ratio of 70:30. *K-folds strategy* was used for performing cross-validation on the training set. Predictive models and word encoding techniques were combined by building a pipeline. The pipeline was then used for training on the train set data and performing predictions on the dev set data. Accuracy score and Confusion matrix were the two-evaluation metrics used for comparison between the models. The model with the best accuracy was chosen and pickled for predictions on the unseen data (test set). The latter implementation was carried out in *predictions.py.*

MODELS AND ACCURACY

Below are the details of the predictive models implemented and their accuracies.

Model	Accuracy (in %)
Stochastic Gradient Descent Classifier	68.62
Random Forest Classifier	70.50
Support Vector Machine	67.21

Random Forest Classifier was used on the test set and it achieved an accuracy score of **66%**.

CHALLENGES AND IMPROVEMENTS

Accuracy and execution time were some of the challenges faced in this task. Due to a large dataset size and execution on a CPU processor, the running time and learning rate of the model was slow. Notable supervised learning algorithms achieved very less accuracy. These algorithms were clearly *underfitting* the data and consequently its performance on the test set was bad.

The key area for improvement is evidently accuracy. Accuracy can be improved by using techniques like dimensionality reduction and neural networks-based algorithms. Removing abbreviations from the text data and the use of *stemming* may also help. Different word encoding methods like *n-grams, word2vec, hash vectorizer*, etc. can be used. The takeaway through this task was to appreciate the procedures of text mining, natural language processing and sentiment analysis by making use of different programming frameworks.

RECENT TRENDS IN SENTIMENT ANALYSIS

Sentiment analysis is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. Sentiment analysis research tasks and methods have grown with social media channels. In its infancy, sentiment analysis was exclusively about allocating a global, overall polarity label (positive, negative, and sometimes neutral) to English language customer reviews. Sentiment analysis is challenging in its essence, but there is an increasing interest in other related tasks that might be even more difficult. Research on sentiment analysis in the past few years have focused on the following domains:

- Aspect Based Sentiment Analysis
- Emotion Analysis
- Spam and Fake Detection
- Multilingual Sentiment Analysis
- Multimodal Sentiment Analysis
- Real Time Analysis
- Argument Mining
- Opinion Mining

The intrinsic complexity of natural language and new challenging sentiment analysis tasks all with the big data paradigm in the background means that there are now more than ever new fascinating research perspectives in affective language understanding.