Rupanshu Banik, IIT Kharagpur

1. Motivation

The motivation is to perform sentiment analysis on IMDB movie reviews to classify them as positive or negative sentiment. This is a common NLP task to evaluate different models like BOW, Word2Vec, TF-IDF and LDA for text classification. Movie reviews have clear sentiment so provide a good benchmark dataset.

2. Abstract

Reviews are preprocessed by cleaning, lowercasing, removing special characters, stopwords and stemming. Feature extraction used is BOW, Word2Vec, TF-IDF and LDA before logistic regression classifier. TF-IDF has best accuracy of 85.59%. Word2Vec gets 85.15% accuracy. BOW has 85.32% accuracy. LDA performs poorly with just 49.14% accuracy. Most positive words are 'excel', 'great', 'perfect' and most negative are 'dull', 'poorly', 'worst'.

3. Introduction

Sentiment analysis is performed on 50K IMDB movie reviews classified as positive or negative sentiment. Goal is to evaluate different NLP techniques like BOW, Word2Vec, TF-IDF and LDA for feature extraction before a logistic regression classifier. Reviews are preprocessed by cleaning, tokenization, lowercase, stopwords removal and stemming. Models compared are BOW, Word2Vec, TF-IDF and LDA pipelines using logistic regression.

4. Data Preprocessing

Reviews are checked for null values and duplicates. HTML tags, special characters and punctuation are removed. Lowercasing, stopwords removal and stemming is done. Word frequency distribution is plotted to check preprocessing. Reviews are tokenized for Word2Vec model creation.

5. Model Architecture

BOW, Word2Vec and TF-IDF are used for feature extraction. LDA is used for topic modeling before the classifier. Logistic regression with default parameters is the classifier model for all pipelines. No changes done to model architecture.

6. Experimental Setup

Sklearn's CountVectorizer and TfidfVectorizer are used for BOW and TF-IDF. Custom Word2Vec and LDA transformers are defined. No parameter tuning done. Logistic regression used with default parameters. Accuracy is the evaluation metric. 80-20 train-test split is used.

7. Hypothesis Tried

Different feature extraction models like BOW, Word2Vec, TF-IDF and LDA tried. No changes to classifier or model architecture. Only preprocessing and feature extraction methods changed.

8. Results

TF-IDF works best with 85.59% accuracy proving word frequency and uniqueness matters. Word2Vec with 85.15% accuracy also works well encoding semantic meaning. BOW has 85.32% accuracy. LDA performs poorly with just 49.14% accuracy.

9. Key Findings

TF-IDF is most effective for sentiment analysis of text. Word2Vec word embeddings also work well. Topic modeling using LDA is not helpful for sentiment analysis. Simple logistic regression is a fast and effective classifier.

10. Future Work

Advanced classifiers like SVM, XGBoost can be tried. Pre-trained word vectors like GloVe can help. CNN and RNN models have been very effective for text classification. Ensembling models can also help improve accuracy further.