CSCE 5063 Machine Learning Final Project Proposal

Sentiment Analysis by Learning Word Vector in Natural Language Processing

University of Arkansas, Fayetteville

Arpan Poudel

Jared Harris 010792702

Priyanka Ingale 010986084

Problem Identification

In this digital era of metadata – not just storing it but analyzing and processing we can reuse this data to train machine such that trained models can solve few existing problems. Few of such problems are identified below and their potential solutions using Machine Learning are explained.

Problem 1: Due to the limited number of doctors, high hospital charges and excessive amount of time for diagnosing the disease, manual observations of images like X-ray, MRI, CT-Scan leads to untreated patents, incorrect diagnosis, and false treatment.

Solution: Abnormality detection using pattern recognition in images is one potential application of machine learning that can be used to assist in the detection of illnesses and irregularities within a person by analyzing various normal and abnormal cases. Using machine learning and deep learning techniques, an algorithm can analyze the parts images of the patient by finding the pattern in the image to detect abnormalities. Convolution Neural Networks are used to train the model and predict abnormalities with higher accuracy. To be more specific, medical abnormalities such as cancer, diabetic retinopathy, pneumonia can be diagnosed accurately by implementing machine learning algorithms. This includes analysis of x-ray, MRI, CT-Scan, and other images.

Problem 2: In the pool of abundant product market, customers have a variety of choices for products depending upon individual preferences. Still the market is growing with new products, and everyday a few more get added into the pool. The problem with this wide variety of products is that sellers are not finding the appropriate customer for products.

Solution: To solve this problem, ML based Product Recommendation System targets customer based upon their buying history and search pattern and recommends only those products which customer could potentially buy. This is useful saving customer's time in searching for a product from the pool and to avoid confusion about products. This approach also helps the sellers in increasing their sales and gain more profit. A robust algorithm can accurately recommend the product depending on user past behaviors.

Problem 3: Star grading system is used to grade a product, however, sometimes given star rating and written review may not match completely. A positive review may have star rating between three to five stars and reverse is also possible. Which affects correct product analysis. Which can underestimate or overestimate the product.

Solution: Various e-commerce business (for e.g., Amazon, Walmart, IMDB) has reviews and stars ratings of products, available for buyer know about the product. With the use of Natural language processing, we can analyze the sentiment of user's review and categorize whether the user is giving a positive review or negative review. Combining star ratings and review sentiment can better rate the product than alone star rating. Furthermore, an accurate description of the product can be depicted by implementing the sentiment analysis on reviews from the user.

Project Discussion and Proposal

Project Introduction -

After identifying existing biased Star grading system for product reviews, we are planning to build a model which can take's products written reviews as an input to learn the model and identify positive and negative reviews.

With respect to the reviews of certain kind of product (in our case reviews related to movies), even though the reviewers are different, there could be similarity between the kind of words used by them to express the annotations. And there is a probability that set of words may occurs in form of patters. Our project goal is identifying such a different combination of word pattern as a word vector by fiddling reviews as input (individual movie review). Unsupervised learning approach can be used in finding word vectors and then using this word vectors a supervised learning model can identify sentiments associated with them. Finally, these sentiments can be used to predict from test data if the review is a positive review or negative review.

Dataset related to IMDB reviews –

We are going to use 50,000 movie reviews from IMDB. Among this 25,000 will be used as training data and rest 25,000 will be used as test data. This dataset contains not more than 30 reviews from each movie and each dataset has equal number of positive reviews and negative reviews.

Machine Learning and Deep Learning approach –

Data Pre-processing:
The IMDB dataset consists of reviews that are written in formal language, which consists of stop words, and data pre-processing is required to clean the data. Data pre-processing approaches such as tokenization, stemming, and normalization is to be done under the first column of the dataset.

Feature extraction:
To input the data to the machine and deep learning model, we need to convert the reviews to vectors which can be done through word2vec, TFIDF, and other word embedding vectors. Natural Language tool kit can also provide meaningful features from the dataset.

Supervised machine learning approach:
Once the features are obtained from the feature extraction process, we can train the different machine learning models such as Random Forest, SVM, and Naïve Bayes and predict whether the testing dataset is a positive review or negative review.

Deep Learning approach:
The embedding matrix is fed to the different approaches of deep learning like LSTM, a fully connected network that will use an optimizer and hyperparameter to better predict the review.

Expected Output –

The expected output of this is a machine learning model which can be implemented in different online websites and stores which, after implementation, can provide the customer with insight into the product by analysis of users' reviews. This insight the model provides to a customer could be used to recommend specific products that are related to other products that a user either leaves positive for or views with multiple positive reviews from other users. This would also allow the system to avoid recommending other products that a user leaves negative reviews for or views that has multiple negative reviews form other users.

References

[1] Publications Using the Dataset Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).