Chapter 34
Develop a Neural Model to Score Bigram of Words Using Bag-of-Words Model for Sentiment Analysis

Anumeera Balamurali
St.Joseph's College of Engineering, India

Balamurali Ananthanarayanan
Tamilnadu Agriculture Department, India

ABSTRACT

A Bag-of-Words model is widely used to extract the features from text, which is given as input to machine learning algorithm like MLP, neural network. The dataset considered is movie reviews with both positive and negative comments further converted to Bag-of-Words model. Then the Bag-of-Word model of the dataset is converted into vector representation which corresponds to a number of words in the vocabulary. Each word in the review documents is assigned with a score and the scores are later represented in vector representation which is later fed as input to neural model. In the Kera's deep learning library, the neural models will be simple feedforward network models with fully connected layers called 'Dense'. Bigram language models are developed to classify encoded documents as either positive or negative. At first, reviews are converted to lines of token and then encoded to bag-of-words model. Finally, a neural model is developed to score bigram of words with word scoring modes.

INTRODUCTION: KNOW THE BASIC TERMS?

Natural Language Processing or NLP is generally defined as the automatic understanding of natural language, like speech and text. The study of natural language processing has been popular around for more than fifty years and grew out of the field of linguistics with the evolutions of computers. Current end applications and research includes information extraction, machine translation, summarization,

DOI: 10.4018/978-1-6684-6303-1.ch034

search and human computer interfaces. While complete semantic understanding remains a way still far from distant goal, researchers have studied a divide and conquer approach and identified several subtasks and methods needed for application development and analysis. These ranges varies from the syntactic methods, such as part-of-speech tagging, chunking and parsing, to the semantic method, such as word sense disambiguation, semantic-role labelling, named entity extraction and anaphora resolution. The field of Natural Language Processing (NLP) aims to convert human language into a formal representation which makes easy for computers to manipulate.

As Internet services for movies has increased in popularity, more and more languages are able to make their way online. In such a world, a need exist for the rapid organizing of ever expanding online reviews. A well-trained movie reviews can easily improves the quality of movies provided through online platform: there are so many different reviews other than movies like product review or feedbacks in so many different languages and most of them cannot be parsed immediately with a glance eye. Thus, an automatic language identification system is needed to analyse the reviews so the system is built to take this task. Because of the sheer volume of reviews in online to be handled, the categorization must be efficient, consuming as small storage and little processing time as possible.

N-gram models are the most widely used models for statistical language modelling and sentiment analysis, which is implemented by artificial neural networks (NN). NN is the powerful technique that is widely used in various fields of computer science. Most of the current NLP systems and techniques use words as atomic units which defines that there is no notion of similarity between words, as these are represented as indices in a vocabulary. The observation so far tells that the simple language models trained on huge amounts of data which outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modelling and text categorization in google, amazon etc...

Text categorization addresses the problem of splitting a given passage of text (or a document) to one or more predefined classes. This is an important area of sentiment analysis research that has been heavily investigated. The goal of text categorization is to classify the given reviews into a fixed number of pre-defined categories which is then listed as result to data analytics companies (Barry, 2016).

Deep learning architectures and algorithms have already created spectacular advances in fields like computer vision and pattern recognition (Brownlee, 2017).

Following this trend, the recent natural language processing is currently more and more specialized in the field of recent deep learning strategies. (Collobert et al., 2011) Deep learning algorithms is found to use the unknown structure for the input distribution to give good representations, usually at multiple levels, with higher-level learned features stated in terms of lower-level features. Deep learning strategies aim at learning feature hierarchies with features from higher levels of the hierarchy with which it is created by the composition of lower level features. Automatic learning features at multiple levels of abstraction permit a system to learn complex functions mapping the input to the output directly from data, while not relying fully on human-crafted features (Youngy et al.,, 2018).

Text Pre-Processing

Tokenization is a way of breaking a text into words or sentences. Tokenization is the method by which huge amount of text is splitted into smaller parts, referred as tokens. Natural language processing is employed for building applications like Text classification, intelligent chatbot, sentimental analysis, language translation, etc. It becomes important to grasp the pattern within the text to attain the above-stated