Chapter 10
Applications of Feature Engineering Techniques for Text Data

Shashwati Mishra
B.J.B. College (Autonomous), Bhubaneswar, India

Mrutyunjaya Panda
Utkal University, Vani Vihar, Bhubaneswar, India

ABSTRACT

Feature plays a very important role in the analysis and prediction of data as it carries the most valuable information about the data. This data may be in a structured format or in an unstructured format. Feature engineering process is used to extract features from these data. Selection of features is one of the crucial steps in the feature engineering process. This feature selection process can adopt four different approaches. On that basis, it can be classified into four basic categories, namely filter method, wrapper method, embedded method, and hybrid method. This chapter discusses about different techniques coming under these four categories along with the research work on feature selection.

INTRODUCTION

Feature engineering plays a very important role to prepare data for further processing activities. It is one of the vital steps in machine learning, as it helps in extracting appropriate features for predictive analysis (Nargesian et al., 2017). Machine learning algorithms are applied on data values, images, audio etc. for classification, prediction, retrieval and several types of analysis purposes. These algorithms can work if the inputs are in the form a set of feature vectors. So, the feature engineering process is very important and greatly affects the result of further computation and analysis. The feature engineering process involves the extraction of features from the object and selection of the relevant features from the set of extracted features. The final set of selected features is considered for further processing activities.

DOI: 10.4018/978-1-7998-6659-6.ch010

Performing the feature engineering manually, needs a lot of effort. The features selected in the manual process are also specific to the problem and the domain. Automated techniques use a statistical measure for feature selection (Garla & Brandt, 2012). Feature represents an important property, characteristic and attribute of data which is analysed to make some prediction. The data may be from a document or from an image or from a database. Feature engineering techniques help in extracting features from the data to improve the performance of machine learning algorithms. Text data can be structured or unstructured in nature. Structured text data are categorical having structured attributes. But documents contain unstructured data, where there is no specific ordering or arrangement of data. The words in a document vary from sentence to sentence. Word length and sentence length are also not fixed in a document. Words are arranged as per the syntax of the language so that a sentence becomes meaningful.

This unstructured natural language analysis has relevance in medical science to analyse medical text, in analysing public opinion on a particular topic, analysis of sentiment of the people, analysis of any social text etc. (Berry & Kogan, 2010; Aggarwal & Zhai, 2012; Struhl, 2015). J. Mishra (Misra, 2020) discussed about the life cycle of machine learning based solutions used to analyse text. Natural Language Processing Feature Specification Language has several meta elements. These are used by the feature extraction system to interpret the features.

Features from any natural language can be extracted at different levels such as, from a sentence or a group of sentences called paragraph or all the sentences in a document or all the text documents together. The level at which the features are extracted is specified by the analysis unit. The syntactic unit states the component of the linguistic features. It may be a word or a phrase or a N-gram or a regular expression or any combination of these components. Logical unit specifies the logical operators that is to be used between the components. These operators include AND, OR, AND NOT, OR NOT (Misra, 2020).

This chapter will discuss about the different feature engineering techniques for the analysis of such unstructured text data. The proposed chapter will have five sections followed by References. The chapter will first discuss about the need of feature engineering with specific importance to unstructured text data. The feature engineering process and different pre-processing stages will be discussed in section 2 followed by the different techniques of feature selection. The research work on text feature engineering using these techniques will be analysed in section 4. Section 5 will contain the concluding remarks.

FEATURE ENGINEERING PROCESS

Feature engineering process has a vital role in extraction and selection of appropriate features from the input data for further analysis and prediction. The feature engineering process involves deciding the type of features, creating the features, verifying the effectiveness of the features and accordingly improve or accept the features.

Machine learning algorithms are based on various mathematical, statistical and optimization principles. These techniques cannot be directly applied on unstructured text data. Therefore the unstructured text data must be converted to a structured format which will be easy for analysis. The preprocessing stage performs different activities like Tokenization, Noise removal etc..