Chapter 18
Machine Learning in Text Analysis

Neha Garg
https://orcid.org/0000-0003-3806-6964
Manav Rachna International Institute of Research and Studies, India

Kamlesh Sharma
https://orcid.org/0000-0002-6000-5933
Manav Rachna International Institute of Research and Studies, India

ABSTRACT

This chapter provides a basic understanding of processes and models needed to investigate the data posted by users on social networking sites like Facebook, Twitter, Instagram, etc. Often the databases of social networking sites are large and can't be handled using traditional methodology for analysis. Moreover, the data is posted in such a random manner that can't be used directly for the analysis purpose; therefore, a considerable preprocessing is needed to use that data and generate important results that can help in decision making for various areas like sentiment analysis, customer feedback, customer reviews for brand and product, prevention management, risk management, etc. Therefore, this chapter is discussing various aspects of text and its structure, various machine learning algorithms and their types, why machine learning is better for text analysis, the process of text analysis with the help of examples, issues associated with text analysis, and major application areas of text analysis.

INTRODUCTION

Machine learning in the context of text analysis is a set of statistical techniques that are used for identifying a part of speech, entities, sentiments of users etc. Today, we live in a competitive world. This competition is a key to innovation and discovery. Due to this competition, customer is getting desired item at reasonable price and good quality. With increasing competition, to sustain and to progress, it is inevitable for any organization to provide the product to the market before competitors do. Customer feedback plays very important role in this which enables one to establish demand. With decision making

DOI: 10.4018/978-1-5225-9643-1.ch018

analysis based on customer feedback, one can be able to predict the future demand of the market and new product can be introduced accordingly. The accuracy of decision making analysis is what makes any organization be at pole position or doom (Tripathy, Agrawal, & Rath, 2015). As the time has progressed, customer feedback is not the only source of customers' opinion. The use of social media (SM) such as Facebook, Twitter and Instagram etc. by an individual is another mode where personal choices and interests are expressed (James, April 23, 2014). This social media data can be extracted and put to use to decision making(Analytics, 2018). With the increasing usage of social media (SM) by various kinds of users in an unpredictable manner, data is generated at a very rapid pace (Laney, 2001),(Gandomi & Haider, 2015),(Beyer, 2012). Using social media networks, users express their views about the event and occasions in all forms of data(Sagiroglu & Sinanc, 2013). By using this huge amount of information, available on social sites, one can enhance the worth of decision making by improving the outcomes, extracted from databases (Lesser et al., 2000). Hence the various methods to handle all form of data has been introduced, out of them text analysis is still a challenge for machine learning.

BACKGROUND

Machine learning in Text analysis is a process of extracting information from the textual data, based on the historical data. Text analysis process constitutes a series of sub-processes which has mainly unstructured and semi-structured data in nature and each step perform operation to identify the patterns, so that decision making can be done based on these patterns automatically.

In order to deal with unstructured data a process called data preprocessing is performed, to remove unwanted words from the text, which is preceded by a number of process to extract features from the text and based on the problem statement choosing a machine learning model and improvise decision making.

WHAT IS TEXT?

The very first question which arises in mind is what is classified as a text? According to literature "anything that can be read" is called text. It can be an article in newspaper, a sentence in a book, a blog on a social networking site or a tweet etc. it can consist of alphanumeric values, symbols, emoticons etc.

STRUCTURE OF TEXT

According to the situation and usage the text can have any kind of organization (format) like the chapter in a book, article in a newspaper, a blog, tweet etc. everywhere text is organized in an inherited way. Moreover with the advancement of World Wide Web (WWW) many users are getting associated with social media,(Manyika et al., 2011) reading e-newspapers, e-books, shop online, filling e-forms and posting their views, feelings, emotions, thinking and expression in the languages they are familiar most probably their mother tongue, English, Hindi, Telgu, Kannad, Spanish, German etc. further they may use multiple languages in a single sentence to make the sentence more impactful which give rises to code switching. The properties of these categories are summarized in a Table 1 below: