Chapter 47
Improvisation of Cleaning Process on Tweets for Opinion Mining

Arpita Grover
https://orcid.org/0000-0001-5273-686X
Kurukshetra University, India

Pardeep Kumar
https://orcid.org/0000-0003-3755-1837
Kurukshetra University, Kurukshetra, India

Kanwal Garg
Kurukshetra University, Kurukshetra, India

ABSTRACT

In the current scenario, high accessibility to computational facilities encourage generation of a large volume of electronic data. Expansion of the data has persuaded researchers towards critical analyzation so as to extract the maximum possible patterns for wiser decisiveness. Such analysis requires curtailing of text to a better structured format by pre-processing. This scrutiny focuses on implementing pre-processing in two major steps for textual data generated by dint of Twitter API. A NoSQL, document-based database named as MongoDB is used for accumulating raw data. Thereafter, cleaning followed by data transformation is executed on accumulated tweets related to Narender Modi, Honorable Prime Minister of India.

DOI: 10.4018/978-1-6684-6303-1.ch047

1. INTRODUCTION

Social media brings people together so that they can generate ideas or share their experiences with each other. The information generated through such sites can be utilized in many ways to discover fruitful patterns. But, accumulation of data via such sources create a huge unstructured textual data with numerous unwanted formats. Henceforth, the first step of text mining involves pre-processing of gathered reviews.

The journey of transforming dataset into a form, an algorithm may digest, takes a complicated road. The task embraces four differentiable phases: Cleaning, Annotation, Normalization and Analysis. The step of cleaning comprehends extrication of worthless text, tackling with capitalization and other similar details. Stop words, Punctuations marks, URLs, numbers are some of the instances which can be discarded at this phase. Annotation is a step of applying some scheme over text. In context to natural language processing, this includes part-of-speech tagging. Normalization demonstrates reduction of linguistic. In other words, it is a process that maps terms to a scheme. Basically, standardization of text through lemmatization and stemming are the part of normalization. Finally, text undergoes manipulation, generalization and statistical probing to interpret features.

For this study, pre-processing is accomplished in three major steps, as signified in Figure 1, keeping process of sentiment analysis in consideration. Foremost step included collection of tweets from Twitter by means of Twitter API. Captured data is then stored in a NoSQL database known to be MongoDB. Thereafter, collected tweets underwent cleaning (Zainol et al., 2018) process. Cleaning phase incorporated removal of user name, URLs, numbers, punctuations, special characters along in addition to lower casing and emoji decoding. The first two phases of data collection and clean ing were demonstrated in previous research. Also, it was shown that application of cleaning process still left data with anomalies and that is why the endmost stage of data transformation is introduced in this research. Data transfor-

Figure 1. Preprocessing steps