Chapter 47
Sentiment Analysis with Text Mining in Contexts of Big Data

Carina Sofia Andrade
University of Minho, Guimarães, Portugal

Maribel Yasmina Santos
Centro Algoritmi, DSI – University of Minho, Guimarães, Portugal

ABSTRACT

The evolution of technology, along with the common use of different devices connected to the Internet, provides a vast growth in the volume and variety of data that are daily generated at high velocity, phenomenon commonly denominated as Big Data. Related with this, several Text Mining techniques make possible the extraction of useful insights from that data, benefiting the decision-making process across multiple areas, using the information, models, patterns or tendencies that these techniques are able to identify. With Sentiment Analysis, it is possible to understand which sentiments and opinions are implicit in this data. This paper proposes an architecture for Sentiment Analysis that uses data from the Twitter, which is able to collect, store, process and analyse data on a real-time fashion. To demonstrate its utility, practical applications are developed using real world examples where Sentiment Analysis brings benefits when applied. With the presented demonstration case, it is possible to verify the role of each used technology and the techniques adopted for Sentiment Analysis.

INTRODUCTION

With increasing use of the Internet (social networks, forums, blogs, etc.) grows exponentially the volume of available data (Pang & Lee, 2008). When a user buys online, he usually shares feedback on the item and the store, or when he joins an event or goes to a restaurant, hotel or a movie, he usually also makes a comment about it. All these data can be used by several stakeholders in the decision-making process, considering the opinions that were expressed (Asur & Huberman, 2010).

Organizations show particular interest in these opinions that are freely left by users on the Internet. The world of news is one of the examples that can be used to demonstrate this interest in understanding

DOI: 10.4018/978-1-7998-2460-2.ch047

what people feel when they share their opinion on the Internet. Currently, most newspapers already have an online version. The question to be answered is "Why? Why having an online version if it continues printed?". The answer is simple: only online versions can hold readers' opinions (Gebremeskel, 2011). Associated with the news world is Twitter, a social network that has millions of users sharing the latest news with a personal opinion or sentiment, providing a personal perspective that is interesting to analyse (Gebremeskel, 2011). As the number of Internet users grows, also grows the interest of organizations in retaining users and their opinions. To benefit from this reality, there is the need to analyse these opinions and use the gathered insights in the decision-making process.

Considering the vast amount of available data and the fact that these data may contain implicit peoples' opinions/sentiments in them, there is the opportunity to use sentiment analysis techniques to understand what is mentioned and which feelings are expressed. This work is guided by this research question: Is it possible to define a Big Data architecture able to collect real-time data from Twitter and analyse the collected data for performing sentiment analysis? With this aim, the objectives of this work include the analysis of the current state-of-the-art, the proposal of an architecture that advances the state-of-the-art, and the validation of the proposed architecture with a demonstration case that goes from the collection to the analysis of the data, providing useful insights on data. Although the proposed architecture is, in this case, restricted to the use of data from Twitter, other data sources can be considered, as long as the collection mechanisms are prepared for that. Moreover, the demonstration case uses a particular set of keywords for data collection, only as an example, being possible to define any other set of keywords. In methodological terms, a proof-of-concept is provided through the implementation of the demonstration case, while for data collection, data treatment and data mining, the CRISP-DM methodology (Chapman et al., 2000) is used. This methodology starts by the business understanding, in order to analyse the application domain and the data analysis requirements. Then, data understanding allows the comprehension of the available data for identifying data quality problems, looking for wrong values, missing data, outliers, among others, and defining the appropriate strategies to correct or deal with these situations. After the data treatment, the collected data is ready for further data analysis. In the modelling phase, Sentiment Analysis is applied following a specific set of steps. Finally, the implemented steps are evaluated to test the feasibility of the proposed implementation. With this evaluation, and depending on the obtained results, it can be considered the need of reviewing the first phase (business understanding), and the following ones, to improve the obtained results.

This paper is organised as follows. After the introduction, related work is summarised, providing an overview of the work already undertaken is this area. The architecture is then presented, following an incremental approach that shows the several tested design options. The implementation and validation with a demonstration case shows the usefulness of the proposed architecture for sentiment analysis. This paper ends with some conclusions and proposals for future work.

RELATED WORK

Several works have pointed the impact of the analysis of tweet data in real contexts. The work of (Asur & Huberman, 2010) focused on the analysis of movies premieres based on Twitter data, being able the achieve a model that predicts box-office revenues for movies. During three months, the authors collected tweets related to twenty-four different movies, using the movie title words. The authors did not put aside the fact that before the premiere, producers do marketing campaigns with videos, photos or even actors