Chapter 57
Evaluation of Topic Models as a Preprocessing Engine for the Knowledge Discovery in Twitter Datasets

Stefan Sommer
Telekom Deutschland GmbH, Germany

Tom Miller
T-Systems Multimedia Solutions GmbH, Germany

Andreas Hilbert
Technische Universität Dresden, Germany

ABSTRACT

In the World Wide Web, users are an important information source for companies or institutions. People use the communication platforms of Web 2.0, for example Twitter, in order to express their sentiments of products, politics, society, or even private situations. In 2014, the Twitter users worldwide submitted 582 million messages (tweets) per day. To process the mass of Web 2.0's data (e.g. Twitter data) is a key functionality in modern IT landscapes of companies or institutions, because sentiments of users can be very valuable for the development of products, the enhancement of marketing strategies, or the prediction of political elections. This chapter's aim is to provide a framework for extracting, preprocessing, and analyzing customer sentiments in Twitter in all different areas.

INTRODUCTION

Twitter is one of the fastest growing social network platforms in the world. In 2014 the social network consists of over two billion users, which submit 582 million messages per day (Twopcharts, 2014). In comparison to 2011 the number of users increased by factor four and the amount of tweets per day increased by factor 5 (Stieglitz, Krüger, Eschmeier, 2011). By using Twitter people share news, information

DOI: 10.4018/978-1-4666-9840-6.ch057

or sentiments in a short message, which is limited to 140 characters, named tweet. Twitter is a so-called microblog, a special kind of a blog that combines an ordinary blog with features of social networks. The communication platforms of Web 2.0 gain in importance, as interaction increases between users through these media (Stephen & Toubia, 2010). Due to the positive development of microblogs and Twitter in particular, these services become a valuable source for companies or institutions (Pak & Paroubek, 2010; Barnes & Böhringer, 2011).

Today users are considered to be key communication partners for companies: providing relevant feedback, requests, and testimonials to the company's performance, a political party or an institution (Richter, Koch, Krisch, 2007; Tumasjan et al., 2010; Sommer et al., 2012). They share their sentiments with other users through the communication platforms of Web 2.0 (Jansen, Zhang, Sobel, Chowdury, 2009). By spreading their thoughts through platforms, such as blogs, communities, or social networks, users influence other users in the process of their own sentiment creation (O'Connor, Balasubramanyan, Routledge, Smith, 2010). But how to deal with this huge amount of unstructured or semi-structured data of social networks within a complex IT infrastructure?

Our research aim is to provide a preprocessing framework for Twitter data, which extracts, transforms and supplies the relevant tweets of a huge amount of data in order to make the data applicable for sentiment analysis. The framework can be used for different approaches and areas. On the one hand to enhance the company's products e.g. by adding mandatory features and finally involving users in the product development as so-called 'prosumers'. On the other hand to optimizing the company's marketing activities by analyzing which advertising is most discussed or referenced in the Web 2.0 platforms, especially in the case of viral marketing (Jansen et al., 2009; Lee, Jeong, Lee, 2008; Liu, Hu, Cheng, 2005). Other examples show the use of Twitter to predict the voters' opinion in political elections (Tumasjan et al., 2010). In this chapter, we focus on the evaluation of our framework, which is based on topic models. Our Twitter dataset for the evaluation contains tweets covering the TV debate between the candidates for the election of the Deutsche Bundestag in Germany.

RELATED WORK

In the last five years many articles have been published in the area of Sentiment Analysis. Liu (2007) gives a broad overview of characteristics, tasks and methods of Sentiment Analysis and places them into the context of Web Data Mining. Next to the theoretical descriptions of Liu (2007) you can find various articles covering different existing Sentiment Analysis systems, which are systematically presented by Lee et al. (2008), as Table 1 shows.

A direct comparison of the performance of these systems is difficult, because the test datasets are not equal and the extraction methods are very different. Nevertheless, Lee et al. (2008) state that systems, which are using syntactic analysis tend to obtain better results regarding the extraction of sentiment expressions. Further information about the systems can be found in Lee et al. (2008) or in the research of authors of the different systems (Dave, Lawrence, & Pennock, 2003; Liu et al., 2005; Yi, Niblack, 2005; Popescu & Etzioni, 2007; Scaffidi et al., 2007).

Due to the popularity of microblogs and Twitter, in particular, there are many papers covering this research area. Böhringer & Gluchowski (2009) describe the microblogging service Twitter, and how users are able to communicate with each other by using this platform. Tweets can contain different content, for