# Chapter 41
# Sentiment Analysis on Social Media:
## Recent Trends in Machine Learning

**Ramesh S. Wadawadagi**
https://orcid.org/0000-0002-6669-7344
*Basaveshwar Engineering College, Bagalkot, India*

**Veerappa B. Pagi**
*Basaveshwar Engineering College, Bagalkot, India*

## ABSTRACT

*Due to the advent of Web 2.0, the size of social media content (SMC) is growing rapidly and likely to increase faster in the near future. Social media applications such as Instagram, Twitter, Facebook, etc. have become an integral part of our lives, as they prompt the people to give their opinions and share information around the world. Identifying emotions in SMC is important for many aspects of sentiment analysis (SA) and is a top-level agenda of many firms today. SA on social media (SASM) extends an organization's ability to capture and study public sentiments toward social events and activities in real time. This chapter studies recent advances in machine learning (ML) used for SMC analysis and its applications. The framework of SASM consists of several phases, such as data collection, pre-processing, feature representation, model building, and evaluation. This survey presents the basic elements of SASM and its utility. Furthermore, the study reports that ML has a significant contribution to SMC mining. Finally, the research highlights certain issues related to ML used for SMC.*

## OVERVIEW

In recent days, social media applications have emerged as leading mass media, as they allow users to work collaboratively and publish their content (Wadawadagi & Pagi, in press; Anami et al. 2014). Accordingly, large volumetric semantically rich information is being generated and accumulated every day in the form of tweets, posts, blogs, news, comments, reviews, etc. Investigating hidden but potentially useful patterns

DOI: 10.4018/978-1-6684-6303-1.ch041

from a huge collection of SMC is a critical task, due to users struggle with overloaded information (Yang & Rim, 2014). SASM is a practice of collecting data from social networks and automatically identifying whether a phrase comprehends sentiment or opinionative content, and further determines the opinion polarity (Jianqiang & Xiaolin, 2017). However, detecting sentiment in SMC faces several challenges, as they are composed of incomplete, chaotic and unstructured sentences, erratic phrases, ungrammatical expressions, and non-lexical words. Moreover, it is hard to detect correlations among opinion sentences due to the broad range of linguistic issues and drives the SA still more challenging (Choi & Park, 2019). To cope with these challenges a real-time SA system needs to be developed to process a large volume of sentiment data in very little time. Furthermore, knowing the public emotions is very useful in many fields, including marketing, politics, online shopping, and many more (Jianqiang & Xiaolin, 2017). To increase productivity, many business firms encourage their customers to participate in virtual discussions, asking for their feedback, opinions, and suggestions.

SASM is generally operated at different levels-of-granularity varying from coarse-grained to fine-grained levels. The coarse-grained analysis deals with determining the sentiment of a whole phrase, while fine-grained analysis is related to attribute level SA. However, employing the right methodology to any key business will drive SASM as a powerful tool for steering organizations and their individual business units as successful outcomes. After several years of constant development, the methodology of SASM is slowly emerging from a disparate set of tools and technologies to a unified framework. The general framework of SASM is depicted in Figure 1. The framework consists of a series of sub-tasks, in which the first task is data acquisition that acquires sentiment data from different sources and stores using different formats. Soon after data acquisition, the data can either be directly streamed to memory for rapid evaluation of unstructured data (in-memory processing) or can be archived to disk (in-database processing) as messages, files, or any machine-generated content. It is being the case that SMC is generally messed up with inconsistent, incomplete and non-dictionary terms, it needs pre-processing before feature vectors are generated. During pre-processing, a series of techniques (e.g., tokenization, stopwords removal, URL pre-treatment, stemming, replacing emoticons) are employed to decrease the amount of irregularity in the data. Additionally, to facilitate the process of identifying document relevancy, the data need to be transformed from a full-text version to a document vector representation that describes the content of the opinion sentences. Two types of representations are extensively used in the literature, namely feature-based representation and relational representation. Perhaps, the most prevalent feature-based representation technique is a vector space model (VSM) (Salton & Yang, 1975). Deduced from basic VSM, some other representation techniques have been used such as n-gram, key-phrase, and hypernym representations. Recently, an alternative document representation based on distributed representation known as semantic word spaces or word-embeddings has shown great success in capturing fine-grained semantic regularities (Mikolov et al., 2013). These vectors consist of low-dimensional real-valued scores that model syntactic and semantic information of individual words. Eventually, these vectors are used as pre-trained features for many sentiment classification tasks. It is evident from the current research, that the earlier SMA systems were designed to facilitate analysts in writing decision rules, while later systems introduce ML for automatic rule generalization. ML algorithms use an example training set of input data to construct a model to make predictions expressed as outputs. Finally, the business analysts or researchers can make critical decisions based on this rich and high-quality data patterns discovered. The key objective of this chapter is to give a comprehensive survey on recent advances in ML techniques used for SASM and its applications. Investigation and analysis of SMC are potentially useful for many