Chapter 36
Supervised Sentiment Analysis of Science Topics: Developing a Training Set of Tweets in Spanish

Patricia Sánchez-Holgado
https://orcid.org/0000-0002-6253-7087
University of Salamanca, Salamanca, Spain

Carlos Arcila-Calderón
https://orcid.org/0000-0002-2636-2849
University of Salamanca, Salamanca, Spain

ABSTRACT

Twitter is one of the largest sources of real-time information on the Internet and is continuously fed by millions of users around the world. Each of these users publishes text messages with their opinions, concerns, information, or simply their daily happenings. It is a challenge to address the analysis of massive data in the network, just as it is an objective to look for ways to understand everything that data can offer today in terms of knowledge of society and the market. The sector of science communication is still discovering everything that the web 2.0 and social networks can offer to reach all audiences. This article develops a classification model of messages launched on Twitter, on science topics, in Spanish, with machine learning techniques. The training of this type of models requires the creation of a specific corpus in Spanish for the subject of science, which is one of the most laborious tasks. The classifier is able to predict the sentiment of the message in real time on Twitter, with a confidence interval greater than 80%. The results of its evaluation are at 72% accuracy.

DOI: 10.4018/978-1-6684-6303-1.ch036

INTRODUCTION

Currently, Twitter is one of the largest sources of real-time information on the Internet, which is continuously fed by millions of users from all over the world, whether real or automatic. Each of these users publishes text messages with their opinions, concerns, information or simply their daily evolution. A large amount of data that are mostly public and that arouse the interest of data researchers. It is a challenge to address the management of mass data in the network, and it is a goal to look for ways to understand everything that data can offer today in terms of knowledge of society and the market.

In this work, we focus on the field of science communication, as an area that is still discovering all that web 2.0 and social networks can offer to reach all audiences. Through the media generated by the user, we can access messages that are disseminated and shared without limits, but our motivation is to find out what we can obtain, how far we can get to know the public through their messages of information or opinion, in a social network that has hardly any control. The data will allow us to experiment with different theories, but the results have to mark the next steps to follow.

In times of post-truth, science gains weight and the mechanisms that affect public opinion have been widely studied. In this context, with a media ecosystem where the producer merges with the receiver and the communication of science adapts to change, the importance of scientific knowledge and its dissemination is vital in the construction of reliable social attitudes and trends. However, the analysis of the polarity of opinions on Twitter is a field that takes off and of which there is little previous research in Spanish.

CONTEXT AND MOTIVATION

There is a growing interest in the study of public opinions using large-scale data produced by social media (Bollen, Mao, & Pepe, 2011; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Whitman Cobb, 2015). However, most of these studies are based on manual classification or automated content analysis using dictionaries that label words (for example, giving a negative or positive a priori value to each word) (Feldman, 2013) and other approaches such as supervised machine learning or supervised machine learning (Vinodhini & Chandrasekaran, 2012) derived from artificial intelligence are scarce in communication research (Van Zoonen & Van der Meer, Toni, 2016), in the social sciences and in private consultancies in issues of public opinion, political studies and marketing. Additionally, new technological efforts are dedicated to gather the automated analysis of feelings based on machine learning with streaming or live streaming technologies, which are capable of producing a significant amount of data.

Three billion people around the world express their thoughts and opinions on a regular basis through social networks. Twitter is one of the most outstanding, characterized by being a microblogging service, since it brings together features of blog, instant messaging and social network, growing exponentially since its launch in 2006. Twitter users generate content based on short texts of a maximum of 280 characters (up to November 2017, 140 characters were allowed), on any topic and in real time. Most of the messages are public, although it offers the possibility of sending private messages. A message or tweet can reach a very high audience in a few minutes thanks to the fact that users share the messages again in an endless network.