Semantic Characterization of Tweets Using Topic Models:
A Use Case in the Entertainment Domain

Andrés García-Silva, Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

Víctor Rodríguez-Doncel, Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

Oscar Corcho, Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

ABSTRACT

In the entertainment domain users tweet about their expectations and opinions regarding upcoming, current and past experiences, while companies advertise and promote the shows. This characterization, important for customers and companies, goes beyond traditional sentiment analysis where the polarity of the sentiments expressed in opinions is usually identified as positive, negative or neutral. The authors investigate different tweet representation models, including bags of words and probabilistic topic models, to shed light on the semantics of the messages. Their experiments show that topic-based models generated with Latent Dirichlet Allocation (LDA) yield, most of the times, better categorizations when compared to TF-IDF based features, particularly when these models are enriched with natural language features and specific Twitter slang.

Keywords: Classification, Latent Dirichlet Allocation (LDA), Latent Topics, Semantics, Twitter

INTRODUCTION

Interest for analyzing social media has gained a large attention in the last few years, as it has proved to be a valid tool to pulse the sentiment of the masses towards commercial brands, political options, public affairs, etc. Profiling the consumers' attitude by scanning social media is a biased measure of the mass opinion addressing a particular stratum of the population, but it is a cheap, easy and fast signal which can be used to timely modify marketing campaigns, pricing policies and the communication strategy of anybody with a public face.

Twitter is an optimal candidate to be studied for its large number of users, ubiquitous availability and messages heterogeneity. Much work has been done in the last few years in the akin fields of opinion mining and sentiment analysis in Twitter. The general motivation of twitters and the very nature of the messages was very well established in Java, Song, Finin, and Tseng (2007) and Krishnamurthy, Gill, and Arlitt (2008), with the relevant focus in the diffusion of word of mouth opinions studied in Jansen, Zhang, Sobel, and Chowdury (2009). Sentiment analysis algorithms proposed by the academia have actually materialized in a bunch of applications now in use by market analysts,

DOI: 10.4018/ijswis.2013070101

community managers and social researchers in general. These tools (TweetFeel[1], Twendz[2], Sentiment140[3], Social Mention[4], Twitometro[5] etc.) usually provide a polarity figure measuring the attitude towards a brand or any other queried topic.

In the sector of plays and musicals (which generates a gross sale of £500M a year[6] in London, selling 250,000 tickets alone in musicals), opinion mining is of particular relevance, as performances are a much communicated act. The whole entertainment industry has a strong dependence on the public opinion, and mouth to mouth advertisement is crucial for the success among the whole entertainment offer. Tweets' analysis have proved to anticipate box-office revenues even by merely observing the number of references to a movie per day (Asur & Huberman, 2010), and it has been related to the ratings given in IMDB (Oghina, Breuss, Tsagkias, & de Rijke, 2012).

Opinions published in Twitter on shows can also be mined to evaluate how it has been received by the public. These messages are precisely dated, but the absolute time is not as important as determining whether it was issued before or after the emitter watched the event. Furthermore, separating true opinions from advertisement—made by community managers or other interested parties, is instrumental to get a reliable opinion assessment. Determining whether a tweet refers to a future event or to a past event, relative to the user's experience, is a field that needs more attention, and not many efforts have been made in sorting out which of the messages actually carry an opinion about a lived event, express an expectation about a future event or convey any form of atemporal advertisement. As an example the tweet "Richard III at the globe was great. Samuel Barnett and Janes Garnon were fantastic but need I say that Mark Rylance was sublime" is a typical opinion, "Going to see the wonderful Mark Rylance at The Globe Thursday, very excited" an exemplary expectation Tweet and "Three last chances to enjoy acclaimed #RichardIII this wkend" an advertisement on the same work.

This classification may benefit customers and companies in this domain. Expectations would help to increase the hype of shows, and opinions can be used to motivate undecided customers, while ads of the competitors can be filtered out so that they do not reach the company customers base or they can be aggregated so that customers can pick the best offer.

This paper evaluates different algorithms to characterize the expectation level of the audience and the aroused opinions after having watched shows, and describes how to separate actual praises from advertisement (whose semantics is too close and prone to be confused as alike). We investigate whether probabilistic topic models are suitable to represent tweets and if they outperform representation schemes, which are based only on the words, in classification tasks for this purpose. Though the value of topic models has been widely investigated in document collections, including scientific reports (Griffths & Steyvers, 2004) and news (Newman, Chemudugunta, Smyth, & Steyvers, 2006), they have just started to be applied to tweets in general contexts (Hong & Davison, 2010 ; Ramage, Dumais, & Liebling, 2010). The main concern regarding topic models is whether they can identify latent topics given the limited length of tweets.

The next section summarizes the state of the art, followed by an enumeration of the possible Twitter information representation models. Next, the experimental work is described, including a qualitative description of the semantics of tweets in this domain and the details of the classification task. Finally, a short vision on the future semantic characterization along with its challenges is presented.

STATE OF THE ART

Determining whether a tweet expresses a personal feeling on a spectacle and whether this sentiment is prior or posterior to the event is a classification problem. The intense research done in the last few years in the area of sentiment analysis, specifically in Twitter, matches partially this problem, as determining whether subjectivity exists is a key task in the classification process. One of the widest surveys

N-Dimensional Matrix-Based Ontology: A Novel Model to Represent Ontologies

Ahmad A. Kardan and Hamed Jafarpour (2018). *International Journal on Semantic Web and Information Systems (pp. 47-69).*
www.igi-global.com/article/n-dimensional-matrix-based-ontology/203692?camid=4v1a

Web X.0: A Road Map

San Murugesan (2010). *Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications (pp. 1-11).*
www.igi-global.com/chapter/web-road-map/39161?camid=4v1a