Decision Support Systems 89 (2016) 98–112

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

The added value of auxiliary data in sentiment analysis of Facebook posts

Matthijs Meire[a], Michel Ballings[b,*], Dirk Van den Poel[a]

[a]Department of Marketing, Ghent University, Tweekerkenstraat 2, Ghent, 9000 Belgium
[b]Department of Business Analytics and Statistics, The University of Tennessee, 249 Stokely Management Center, 916 Volunteer Blvd, Knoxville, 37996 TN, USA

ARTICLE INFO

Article history:
Received 10 September 2015
Received in revised form 17 June 2016
Accepted 19 June 2016
Available online 27 June 2016

Keywords:
Facebook
Text mining
Sentiment analysis
Machine learning
Social media

ABSTRACT

The purpose of this study is to (1) assess the added value of information available before (i.e., leading) and after (i.e., lagging) the focal post's creation time in sentiment analysis of Facebook posts, (2) determine which predictors are most important, and (3) investigate the relationship between top predictors and sentiment. We build a sentiment prediction model, including leading information, lagging information, and traditional post variables. We benchmark Random Forest and Support Vector Machines using five times twofold cross-validation. The results indicate that both leading and lagging information increase the model's predictive performance. The most important predictors include the number of uppercase letters, the number of likes and the number of negative comments. A higher number of uppercase letters and likes increases the likelihood of a positive post, while a higher number of comments increases the likelihood of a negative post. The main contribution of this study is that it is the first to assess the added value of leading and lagging information in the context of sentiment analysis.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In the beginning of the century, Web 2.0 emerged as an ideological and technical foundation giving rise to the massive production of user generated-content (UGC). Blogging platforms and online retailers are the first examples of this foundation [50]. Today, UGC is still growing rapidly, sparking interest and activity in opinion mining and sentiment analysis [62, 74]. Sentiment analysis is defined as the computational process of extracting sentiment from text [61, 74]. Applications range from the prediction of election outcomes [17, 92], to relating public mood to socio-economic variables [17], to improved e-learning strategies [72].

Early examples of sentiment analysis were mainly based on review data. This type of data rarely contained much more information than the content and the time of posting of the review itself. Models using these data are based on present information, where 'present' refers to the time of posting. This changed with the advent of social networks such as Facebook and Twitter in that much more data became available. On these platforms, not only the focal post's content is available, but, taking into account the time of posting, there is also leading and lagging information. Leading information is available even before content is posted (e.g., user profiles, previous posts) and thus contains information about the past. On the other hand, lagging information is generated a posteriori, after the content was posted (e.g., interactions such as likes or retweets) and thus contains information about the future (seen from the time of posting). Leading information can therefore be included in any sentiment model, while lagging information can be included in tools that do not require real-time sentiment analysis. To the best of our knowledge, there is no study that includes leading and lagging information into sentiment analysis models. However, we believe that we can improve sentiment prediction by including leading and lagging information for several reasons. First, social media suffer from a lot of slang [41, 72] making it harder for traditional methods to achieve satisfactory model performance on text variables alone. Second, leading variables would take into account users' average sentiment, word use, well-being, and mood and demographics, effectively acting as a user-specific informative prior of future sentiment and accounting for heterogeneity among users. Leading variables have been shown to lead to better predictions [10]. Third, extant literature has found significant relationships between post sentiment and lagging information such as likes and comments [87].

To fill this gap in literature, we assess the additional value for sentiment analysis of leading and lagging information over and above information extracted from the focal post. We do this by constructing three models. The first model is the base model that focuses on the present and contains only the focal post (including text and timing of posting). The second model contains both the focal post's content and leading information, and thus contains both present and past

* Corresponding author.
E-mail addresses: Matthijs.Meire@UGent.be (M. Meire), Michel.Ballings@utk.edu (M. Ballings), Dirk.VandenPoel@UGent.be (D. Van den Poel).

http://dx.doi.org/10.1016/j.dss.2016.06.013
0167-9236/© 2016 Elsevier B.V. All rights reserved.

information. Finally, the third model augments the second model with lagging information. This means that the third model takes into account the past, present and future information of a post.

The remainder of this article is structured as follows. First, we provide a literature review focusing on sentiment analysis of social media data and the reasons why leading and lagging information might be valuable in a sentiment prediction model. Second, we detail our methodology including the data, the model description, the predictors, the predictive algorithms and the model evaluation measure. The third section discusses the results. The penultimate section consists of the conclusion and practical implications of this research. In the final section we address the limitations and avenues for future research.

2. Literature review

There are two main approaches to sentiment analysis [72, 88]. The first approach consists of lexicon-based models, which use predefined lexicons of positive, neutral and negative words to assign positivity values to a sentence or text (e.g., [46, 93]). Machine learning-based methods constitute the second approach. These methods use several text features (e.g., syntactic features and lexical features; we refer to McInnes [64] for a complete overview of these features) as input for a training model and predict the sentiment of text using these features [88]. Machine learning methods have been shown to be more accurate than lexicon-based methods in general, but also more time consuming [20, 75]. Lexicon-based methods, however, tend to perform better in less-bounded domains [72]. Recently, the two approaches have been combined by several authors [58, 65, 72, 90, 98], mostly by using the scores from a lexicon-based exercise as input features for the machine learning algorithm. In this study we will adopt such a hybrid approach. The reason is that the approach allows for additional features to be added to the model.

Literature on sentiment analysis can be summarized according to (1) the use of a focal post's features [64] , (2) the use of auxiliary features [10] , and (3) the focal post's source [1] . The focal post's features constitute: (1) lexicon features, which denote either a pure lexicon-based approach or a combination of lexicon and machine learning, (2) lexical features (bag-of-words, n-grams, co-occurrence and collocations), (3) syntactic features (morphology, part-of-speech) and (4) time features. The auxiliary features are divided into leading and lagging features. The former denotes all the information, with regard to a specific user, that is available until the moment of posting. The latter includes information that is available one week after posting (i.e., information on the likes and the comments a post has received). Stated differently, the focal post's features reflect all information of the present, where 'the present' refers to the time of posting, which will be different for every post. Every action that occurred before the present, is referred to as 'the past', while 'the future' indicates all actions that occurred after posting. The leading variables thus originate in the past, while the lagging variables originate in the future.

Table 1 provides a representative overview of literature with a focus on social media applications, as social media contain leading and lagging information. It is apparent that sentiment analysis has been widely applied to a diverse set of social media. Table 1 shows that both the lexicon-based (denoted an x in the column labeled 'Lexicon') and the machine learning approaches have been used, and that plenty of text features have been explored. However, it also shows that there is a large potential source of information for sentiment analysis that remains largely untapped. Indeed, social media do not

Table 1
Literature overview.

	Features of focal post				Auxiliary features		Text source
	Lexicon	Lexical	Syntactic	Time	Leading	Lagging	
Pang et al. [75]		x	x				Reviews
Dave et al. [26]		x	x				Reviews
Yu and Hatzivassiloglou [96]	x	x	x				News Items
Bai et al. [4]		x					Reviews
Gamon [40]		x	x				Customer feedback
Mullen and Collier [68]		x					Reviews
Matsumoto et al. [63]		x					Reviews
Read [80]		x					Reviews
Riloff et al. [81]		x	x				Reviews
Abbasi et al. [1]		x	x				Reviews
Go et al. [41]		x	x				Twitter
Prabowo and Thelwall [78]	x	x	x				Reviews
Melville et al. [65]	x						Reviews
Pak and Paroubek [73]		x					Twitter
Barbosa and Feng [9]	x		x				Twitter
Davidov et al. [27]		x					Twitter
Kouloumpis et al. [53]	x	x	x				Twitter
Taboada et al. [88]	x						Reviews
Agarwal et al. [2]	x	x	x				Twitter
Smeureanu and Bucur [85]		x					Reviews
Wang and Manning [94]		x					Reviews
Neri et al. [69]		x					Facebook
Blamey et al. [15]		x					Twitter
Kumar and Sebastian [56]	x	x					Twitter
Ben Hamouda and El Akaichi [13]		x					Facebook
Troussas et al. [91]		x					Facebook
Tamilselvi and ParveenTaj [89]		x	x				Twitter
Habernal et al. [42]		x	x				Facebook
Ortigosa et al. [72]	x						Facebook
Basiri et al. [10]	x	x			x		Reviews
da Silva et al. [24]		x					Twitter
Fersini et al. [36]		x					Reviews, Twitter
Yu and Wang [97]		x					Twitter
Mohammad and Kiritchenko [67]		x					Twitter
Our study	x	x	x	x	x	x	Facebook

Download English Version:

https://daneshyari.com/en/article/551974

Download Persian Version:

https://daneshyari.com/article/551974

Daneshyari.com