Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes

Análisis del Sentimiento de mensajes de Twitter con Multinomial Naive Bayes

Alexandre Trilla, Francesc Alías
GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)
atrilla@salle.url.edu, falias@salle.url.edu

Resumen: Este artículo adapta un esquema de Clasificación de Texto basado en Multinomial Naive Bayes para procesar mensajes de Twitter etiquetados con seis clases de sentimiento así como también su tópico. La efectividad de esta estrategia de clasificación de sentimiento se evalúa con el corpus TASS-SEPLN Twitter y se obtiene una tasa máxima de medida F_1 promediada de 36.28%.
Palabras clave: Análisis del sentimiento, Clasificación de Texto, Aprendizaje Automático, Twitter

Abstract: This article adapts a Text Classification scheme based on Multinomial Naive Bayes to deal with Twitter messages labelled with six classes of sentiment as well as with their topic. The effectiveness of this scheme is evaluated using the TASS-SEPLN Twitter dataset and it achieves maximum macroaveraged F_1 measure rate of 36.28%.
Keywords: Sentiment Analysis, Text Classification, Machine Learning, Twitter

1 Introduction

The sentiment classification framework we present in the TASS-SEPLN competition is fundamentally influenced by our previous results in short-text Sentiment Analysis (Trilla y Alías, 2012), which are published in the main stream of the SEPLN 2012 conference. Given a short-text scenario like this one based on Twitter messages, where the amount of available textual instances to train the classifier (e.g., 7219 examples) is much smaller than the dimensionality of the feature space to represent the texts (e.g., 29685 dimensions, considering several feature representations), the most effective scheme (both in accuracy and speed) is buttressed by Multinomial Naive Bayes (MNB) operating on a binary-weighted unigram space (Trilla y Alías, 2012).

In this work, we present a summary of the learning strategy of our approach in Section 2, the preliminary results that we obtained with the target Twitter-based dataset in Section 3, and we explain the conclusions that can be derived from the results provided by the contest organisers in Section 4.

2 Multinomial Naive Bayes

The Multinomial Naive Bayes (MNB) is a probabilistic generative approach that builds a language model assuming conditional independence among the linguistic features. Therefore, no sense of history, sequence nor order is introduced in this model. In reality, this assumption does not hold for textual data (Pang, Lee, y Vaithyanathan, 2002), but even though the probability estimates are of low quality because of this oversimplified model, its classification decisions (based on Bayes' decision rule) are surprisingly good (Manning, Raghavan, y Schütze, 2008). The MNB combines efficiency (it has an optimal time performance) with good accuracy, hence it is often used as a baseline in Text Classification and Sentiment Analysis research (Sebastiani, 2002; Manning, Raghavan, y

Schütze, 2008).

The optimal feature representation for the problem at hand, i.e., a binary-weighted feature space (Trilla y Alías, 2012), displays the presence of each textual feature for each instance (note that all URLs have been converted into a single standard term). Considering that content words are rarely repeated in a 140 character-limited piece of Twitter text (which has 16 words on average), this binary representation seems to be accurate enough to capture the significant traits of the sentiment in this kind of text (Trilla y Alías, 2012).

3 Empirical evaluation

To evaluate the classification effectiveness of the MNB strategy on the Twitter-based corpus and over 5 sentiment levels plus the NONE category label, a 10-fold Cross-Validation (CV) procedure is conducted on the training dataset, yielding a macroaveraged F_1 measure rate of 36.28%. It is to note that the random classification score for six category labels is 16.67%, showing that our strategy based on MNB actually learns successfully.

Regarding the identification of trending topic, it must be stated that MNB has been unable to predict the class with the least generality (Sebastiani, 2002), i.e., the class with the fewest number of training examples. Therefore, the precision score cannot be calculated for the F_1 measure. This result strictly shows that learning 9/10 parts of the training corpus cannot predict the remaining 1/10 part under the 10-fold CV procedure, but the general effectiveness of the MNB strategy for topic-identification can still extrapolate to different texts.

4 Conclusions

The effectiveness results provided by the TASS-SEPLN organisers indicate that our proposal based on MNB is rather effective. We ranked in the 5th place out of 20 in the Sentiment Analysis tasks (both 5 levels plus NONE and 3 levels plus NONE) with a precision score of 57.01% and 61.95% respectively, and in the 3rd place out of 15 in the trending topic coverage task, with a precision score of 60.16%.

In any case, the results show there is some place to improve the effectiveness of the system. Although our approach does capture and model an important amount of useful information, more detailed characteristics still need to be considered through Twitter-specific features.

Bibliografía

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, MA, USA.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proc. of EMNLP'02*, páginas 79–86, Philadelphia, PA, USA, Jul.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47.

Trilla, Alexandre y Francesc Alías. 2012. Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes. En *Procesamiento del Lenguaje Natural*.