Expert Systems With Applications 77 (2017) 236–246

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Enhancing deep learning sentiment analysis with ensemble techniques in social applications

Oscar Araque*, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, Carlos A. Iglesias

Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros de Telecomunicación, Departamento de Ingeniería de Sistemas Telemáticos, Avenida Complutense 30, Madrid, Spain

ARTICLE INFO

Article history:
Received 30 June 2016
Revised 31 January 2017
Accepted 1 February 2017
Available online 3 February 2017

Keywords:
Ensemble
Deep learning
Sentiment analysis
Machine learning
Natural language processing

ABSTRACT

Deep learning techniques for Sentiment Analysis have become very popular. They provide automatic feature extraction and both richer representation capabilities and better performance than traditional feature based techniques (i.e., surface methods). Traditional surface approaches are based on complex manually extracted features, and this extraction process is a fundamental question in feature driven methods. These long-established approaches can yield strong baselines, and their predictive capabilities can be used in conjunction with the arising deep learning methods. In this paper we seek to improve the performance of deep learning techniques integrating them with traditional surface approaches based on manually extracted features. The contributions of this paper are sixfold. First, we develop a deep learning based sentiment classifier using a word embeddings model and a linear machine learning algorithm. This classifier serves as a baseline to compare to subsequent results. Second, we propose two ensemble techniques which aggregate our baseline classifier with other surface classifiers widely used in Sentiment Analysis. Third, we also propose two models for combining both surface and deep features to merge information from several sources. Fourth, we introduce a taxonomy for classifying the different models found in the literature, as well as the ones we propose. Fifth, we conduct several experiments to compare the performance of these models with the deep learning baseline. For this, we use seven public datasets that were extracted from the microblogging and movie reviews domain. Finally, as a result, a statistical study confirms that the performance of these proposed models surpasses that of our original baseline on F1-Score.

1. Introduction

The growth of user-generated content in web sites and social networks, such as Twitter, Amazon, and Trip Advisor, has led to an increasing power of social networks for expressing opinions about services, products or events, among others. This tendency, combined with the fast spreading nature of content online, has turned online opinions into a very valuable asset. In this context, many Natural Language Processing (NLP) tasks are being used in order to analyze this massive information. In particular, Sentiment Analysis (SA) is an increasingly growing task (Liu, 2015), whose goal is the classification of opinions and sentiments expressed in text, generated by a human party.

The dominant approaches in sentiment analysis are based on *machine learning* techniques (Pang, Lee, & Vaithyanathan, 2002; Read, 2005; Wang & Manning, 2012). Traditional approaches frequently use the Bag Of Words (BOW) model, where a document is mapped to a feature vector, and then classified by machine learning techniques. Although the BOW approach is simple and quite efficient, a great deal of the information from the original natural language is lost (Xia & Zong, 2010), e.g., word order is disrupted and syntactic structures are broken. Therefore, various types of features have been exploited, such as higher order *n*-grams (Pak & Paroubek, 2010). Another kind of feature that can be used is Part Of Speech (POS) tagging, which is commonly used during a syntactic analysis process, as described in Gimpel et al. (2011). Some authors refer to this kind of features as *surface* forms, as they consist in lexical and syntactical information that relies on the pattern of the text, rather than on its semantic aspect.

Some prior information about sentiment can also be used in the analysis. For instance, by adding individual word polarity to the previously described features (Pablos, Cuadros, & Rigau, 2016). This

* Corresponding author.
E-mail addresses: o.araque@upm.es (O. Araque), ignacio.cplatas@alumnos.upm.es (I. Corcuera-Platas), jfernando@dit.upm.es (J.F. Sánchez-Rada), cif@gsi.dit.upm.es (C.A. Iglesias).

http://dx.doi.org/10.1016/j.eswa.2017.02.002

prior knowledge usually takes the form of *sentiment lexicons*, which have to be gathered. Sentiment lexicons are used as a source of subjective sentiment knowledge, where this knowledge is added to the previously described features (Cambria, 2016; Kiritchenko, Zhu, & Mohammad, 2014; Melville, Gryc, & Lawrence, 2009; Nasukawa & Yi, 2003).

The use of lexicon-based techniques has a number of advantages (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). First, the linguistic content can be taken into account through mechanisms such as sentiment valence shifting (Polanyi & Zaenen, 2006) considering both intensifiers (e.g. very bad) and negations (e.g. not happy). In addition, sentiment orientation of lexical entities can be differentiated based on their characteristics. Moreover, language-dependent characteristics can be included in these approaches. Nevertheless, lexicon-based approaches have several drawbacks: the need of a lexicon that is consistent and reliable (Taboada et al., 2011), as well as the variability of opinion words across domains (Turney, 2002), contexts (Ding, Liu, & Yu, 2008) and languages (Perez-Rosas, Banea, & Mihalcea, 2012). These dependencies make it hard to maintain domain independent lexicons (Qiu, Liu, Bu, & Chen, 2009).

In general, extracting complex features from text, figuring out which features are relevant, and selecting a classification algorithm are fundamental questions in the machine learning driven methods (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Sharma & Dey, 2012; Wilson, Wiebe, & Hoffmann, 2009). Traditional approaches rely on manual feature engineering, which is time consuming.

On the other hand, deep learning is a promising alternative to traditional methods. It has shown excellent performance in NLP tasks, including Sentiment Analysis (Collobert et al., 2011). The main idea of deep learning techniques is to learn complex features extracted from data with minimum external contribution (Bengio, 2009) using deep neural networks (Alpaydin, 2014). These algorithms do not need to be passed manually crafted features: they automatically learn new complex features. Nevertheless, a characteristic feature of deep learning approaches is that they need large amounts of data to perform well (Mikolov, Chen, Corrado, & Dean, 2013). Both automatic feature extraction and availability of resources are very important when comparing the traditional machine learning approach and deep learning techniques.

However, it is not clear whether the domain specialization capacity of traditional approaches can be surpassed with the generalization capacity of deep learning based models in all NLP tasks, or if it is possible to successfully combine these two techniques in a wide range of applications.

In this paper, we propose a combination of these two main sentiment analysis approaches through several ensemble models in which the information provided by many kinds of features is aggregated. In particular, this work considers an ensemble of classifiers, where several sentiment classifiers trained with different kinds of features are combined, and an ensemble of features, where the combination is made at the feature level. In order to study the complementarity of the proposed models, we use six public test datasets from two different domains: Twitter and movie reviews. Moreover, we performed a statistical study on the results of these ensemble models in comparison to a deep learning baseline we have also developed. We also present the complexity of the proposed ensemble models. Besides, we present a taxonomy that classifies the models found in the literature and the ones proposed in this work.

With our proposal we seek answers to the following questions, using the empirical results we have obtained as basis:

1. Is there a framework for characterizing existing approaches in relation to the ensemble of deep and traditional techniques in sentiment analysis?
2. Can deep learning approaches benefit from their ensemble with surface approaches?
3. How do different deep and surface ensembles compare in terms of performance?

The rest of the paper is organized as follows. Section 2 shows previous work on both ensemble techniques and deep learning approaches. Section 3 describes the proposed taxonomy for classifying ensemble methods that merge surface and deep features, whereas Section 4 addresses the proposed classifier and ensemble models. In Section 5, we describe the designed experimental setup. Experimental results are presented and analyzed in Section 6. Finally, Section 7 draws conclusions from previous results and outlines the future work.

2. Related work

In this section we offer a brief summary of the previous work in the context of ensemble methods and deep learning algorithms for Sentiment Analysis.

2.1. Ensemble methods for sentiment analysis

In the field of ensemble methods, the main idea is to combine a set of models (base classifiers) in order to obtain a more accurate and reliable model in comparison with what a single model can achieve. The methods used for building upon an ensemble approach are many, and a categorization is presented in Rokach (2005). This classification is based on two main dimensions: how predictions are combined (rule based and meta learning), and how the learning process is done (concurrent and sequential).

Regarding the first dimension, on the one hand, in *rule based* approaches predictions from the base classifiers are treated by a rule, with the aim of averaging their predictive performance. Examples of rule based ensembles are the majority voting, where the output prediction per sample is the most common class; and the weighted combination, which linearly aggregates the base classifiers predictions. On the other hand, *meta learning* techniques use predictions from component classifiers as features for a meta-learning model.

As explained in Xia, Zong, and Li (2011), weighted combinations of feature sets can be quite effective in the task of sentiment classification, since the weights of the ensemble represent the relevance of the different feature sets (e.g. n-grams, POS, etc.) to sentiment classification, instead of assigning relevance to each feature individually. The benefits of rule based ensembles were shown also in Fersini, Messina, and Pozzi (2014), where several variants of voting rules are exhaustively studied in a variety of datasets, with an emphasis on the complexity that results from the use of these approaches. In a different work, Fersini, Messina, and Pozzi (2016) have compared the majority voting rule with other approaches, using three types of subjective signals: adjectives, emoticons, emphatic expressions and expressive elongations. They report that adjectives are more impacting that the other considered signals, and that the average rule is able to ensure better performance than other types of rules. Also, in Xia et al. (2011) a meta-classifier ensemble model is evaluated, obtaining performance improvements as well. An adaptive meta-learning model is described in Aue and Gamon (2005), which offers a relatively low adaptation effort to new domains. Besides, both rule based and meta-learning ensemble models can be enriched with extra knowledge, as illustrated in Xia and Zong (2011). These authors propose the use of a number of rule based ensemble models, namely a sum rule and two weighted combination approaches trained with different loss functions. The base classifiers are trained with n-grams and POS features. These models obtain significant results for cross-domain sentiment classification.