Expert Systems With Applications 110 (2018) 298–310

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis

Symeon Symeonidis*, Dimitrios Effrosynidis, Avi Arampatzis

Database & Information Retrieval Research Unit, Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece

ARTICLE INFO

Article history:
Received 28 December 2017
Revised 21 May 2018
Accepted 8 June 2018

Keywords:
Sentiment analysis
Text pre-processing
Machine learning
Text classification
Ablation study
Combination study

ABSTRACT

Pre-processing is the first step in text classification, and choosing right pre-processing techniques can improve classification effectiveness. We experimentally compare 16 commonly used pre-processing techniques on two Twitter datasets for Sentiment Analysis, employing four popular machine learning algorithms, namely, Linear SVC, Bernoulli Naïve Bayes, Logistic Regression, and Convolutional Neural Networks. We evaluate the pre-processing techniques on their resulting classification accuracy and number of features they produce. We find that techniques like lemmatization, removing numbers, and replacing contractions, improve accuracy, while others like removing punctuation do not. Finally, in order to investigate interactions—desirable or otherwise—between the techniques when they are employed simultaneously in a pipeline fashion, an ablation and combination study is contacted. The results of ablation and combination show the significance of techniques such as replacing numbers and replacing repetitions of punctuation.

1. Introduction

In the last decade, Sentiment Analysis in microblogging has become a very popular research area. People share their daily life through messages on platforms such as Twitter, where posts of users involve various topics. Interesting approaches for classification methods in Sentiment Analysis are presented in many research papers (e.g. Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Mohammad, Zhu, Kiritchenko, & Martin, 2015), and the important role of pre-processing before and during the feature selection process is widely noted.

In this context, pre-processing is the procedure of cleansing and preparing texts that are going to be classified. It is a fact that unstructured texts on the Internet—in our case on Twitter—contain significant amounts of noise. By the term noise, we define data that do not contain any useful information for the analysis at hand, i.e. Sentiment Analysis.

According to Fayyad, Piatetsky-Shapiro, and Uthurusamy (2003), the total percentage of noise in datasets may reach 40%, a fact that causes confusion in machine learning algorithms. Twitter users are prone to spelling and typographical errors and to the use of abbreviations and slang. They may also (over- or mis-) use punctuation marks to emphasize their emotions, like for example many exclamation marks. Usually, it is not necessary to include all terms of the initial form of a text in the machine learning step, and some of them can be ignored, replaced, or merged with others. Thus, the need of cleansing and normalizing the data arises, as their quality is a key factor to the success of the machine learning that follows pre-processing.

The purpose of this study is to gather common pre-processing techniques from previous studies, add a few new ones that have not been used a lot by researchers, such as replacing contractions and replacing negations with antonyms, and examine their significance in feature selection by measuring their accuracy in sentiment classification and their resulting number of features.

In the end, based on the results obtained, the techniques that are more suitable for Twitter Sentiment Analysis and those that have to be avoided are suggested to future researchers. The present study is a comprehensive extension of our previous work (Effrosynidis, Symeonidis, & Arampatzis, 2017), and it also investigates the interactions among pre-processing methods via ablation and combination studies.

The rest of this paper is structured as follows. Section 2 reviews some of the related literature. In Section 3, the pre-processing techniques that will be compared are presented. Section 4 describes the datasets, the machine learning algorithms, and the evaluation methodology, while our results are presented and discussed in Section 5. Conclusions and directions for future research are summarized in Section 6.

[*] Corresponding author.
E-mail addresses: ssymeoni@ee.duth.gr (S. Symeonidis), deffrosy@ee.duth.gr (D. Effrosynidis), avi@ee.duth.gr (A. Arampatzis).

https://doi.org/10.1016/j.eswa.2018.06.022
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

2. Related work

In Sentiment Analysis, especially on microblogging texts, the role of pre-processing techniques is significant as a part of text classification. Many research efforts have been made in order to demonstrate the difference between these techniques and their contribution to the final result of classification.

Singh and Kumari (2016) examine the effects of pre-processing on Twitter data for the fortification of sentiment classification. They focus on tweets which are full of symbols, abbreviations, folksonomy, and unidentified words. By removing URLs, hashtags, user mentions, punctuation, and stopwords, they recognize and accept the importance of slang words and spelling correction. In their experiments, an SVM classifier is employed.

Bao, Quan, Wang, and Ren (2014) studied the impact of preprocessing methods on Twitter sentiment classification, evaluating on Stanford Twitter Sentiment Dataset. The experimental results presented a positive effect on sentiment classification when using the pre-processing techniques of URLs features reservation, negation transformation, and repeated letters normalization, while stemming and lemmatization had a negative impact.

The role of pre-processing is also investigated by Haddi, Liu, and Shi (2013) on movie reviews. They use pre-processing techniques such as expansion of abbreviations, removal of non-alphabetic signs, stopword removal, negation handling with the addition of the prefix 'NOT_', and stemming. An SVM classifier is also employed and the authors correlate the number of features to its accuracy. It is shown that appropriate text pre-processing methods, including data transformation and filtering, can significantly enhance the classifier's performance.

Pre-processing techniques are also explored by Uysal and Günal (2014) for two languages on e-mails and news. They employ stopword removal, lowercase conversion, and stemming, and they evaluate with micro-averaged F_1 score using an SVM classifier. They conclude that there is no unique combination of preprocessing techniques that improves accuracy on any domain or language and the researchers should carefully analyse all possible combinations.

Zhao and Gui (2017), focused on effects of text pre-processing methods and used six pre-processing methods on five Twitter datasets with two feature models and four classifiers. The effectiveness of sentiment classification increased by the methods of expanding acronyms and replacing negations, and decreased by the methods of removing URLs, numbers, and stop words.

The Workshop on Noisy User-generated Text[1], which takes place annually since 2015, focuses on natural language processing applied to noisy user-generated text. In 2015, the workshop introduced a lexical normalization task, aiming at normalizing nonstandard words in English Twitter messages to their canonical forms. In studies of Saloot, Idris, Shuib, Raj, and Aw (2015) and Yamada, Takeda, and Takefuji (2015), new approaches were presented to minimize the noise of Twitter messages, by using the maximum entropy model for normalizing Tweets and Entity Linking which is a method to detect entity mentions for text and resolve them to corresponding entries.

Other studies (Liao, Wang, Yu, Sato, & Cheng, 2017; Severyn & Moschitti, 2015; Tang, Wei, Yang et al., 2014) examined the deep learning approach for Sentiment Classification and used Convolutional Neural Networks and Word Embeddings in order to achieve better results than those obtained through traditional techniques. dos Santos and Gatti (2014) examined the creation of a network which took advantage of different levels of information to perform Sentiment Analysis, and used character-level, word-level, and sentence-level representations and features. The performance of manually-hand extracted features combining with automatically extracted embedding features by using deep learning techniques and integrating them with traditional approaches was examined by Araque, Corcuera-Platas, Sánchez-Rada, and Iglesias (2017). In the work of Tang, Wei, Qin, Liu, and Zhou (2014), the sentiment-specific word embedding features concatenated and annotated to become hand-crafted features for Twitter sentiment classification, and new features tested the latter to a deep neural network.

Despite the fact that many studies have examined the role of pre-processing, generally and specifically in Sentiment Analysis, none of them has gathered in a comparative study a large number of popular techniques as it is done in this work.

3. Common pre-processing techniques

As a first step in pre-processing, most (if not all) studies, e.g. Wang and Manning (2012), Symeonidis, Effrosynidis, Kordonis, and Arampatzis (2017), Pak and Paroubek (2010), Giachanou, Gonzalo, Mele, and Crestani (2017), apply tokenization. According to Balazs and Velásquez (2016) tokenization is "a task for separating the full text string into a list of separate words". Atkinson, Salas, and Figueroa (2015) defined tokenization as "a kind of lexical analysis that breaks a stream of text up into words, phrases, symbols, or other meaningful elements called tokens". At its core, the process of tokenization is a standard method for further Natural Language Processing (NLP) transformation in preprocessing.

The 16 pre-processing techniques we will experiment with are described below. The order that they should be applied is of major importance; we present them in the recommended order which enables combinations of them in the same pre-processing pipeline with as few adverse effects as possible. We briefly describe each technique, why it is applied, give an example, and mention related works that used it before.

3.1. Remove unicode strings and noise

Not all datasets are given clean. So, first of all, using some regular expressions we remove non-english characters and unicode strings like "\u002c" and "\x06" which were remnants of the crawling procedure that created the dataset. This technique is considered a baseline for our experiments.

3.2. Replacing URLs and user mentions

In Twitter texts, the majority of sentences contain a URL, a user mention, and/or a hashtag symbol. Their presence does not contain any sentiment and one approach is to replace them in preprocessing with tags as, e.g. Agarwal et al. (2011) do. In our work, we use the tags 'URL' and 'AT_USER' and removed the hashtag symbol. Some other thoughts could be to either remove only the punctuation signs in user mentions and keep the username or remove it completely (Bermingham & Smeaton, 2011; Khan, Bashir, & Qamar, 2014), but this case was not examined.

This technique is not universal and only applies to Twitter texts. So, it should/could be done before any other technique.

For example, the tweet

> *RT @BoomerLivingNow: Retirement: Don't Run Out of Money Before You Run Out of Time http://ow.ly/15RgI #finances #boomer #retirement*

after this particular pre-processing step is transformed to:

> RT AT_USER Retirement : Don't Run Out of Money Before You Run Out of Time URL finances boomer retirement

[1] http://noisy-text.github.io/.