Information Processing and Management 51 (2015) 532–546

Contents lists available at ScienceDirect

Information Processing and Management

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

Reprint of “Supervised sentiment analysis in Czech social media” [☆,☆☆]

Ivan Habernal [a,b,*], Tomáš Ptáček [b], Josef Steinberger [a,b]

[a] Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
[b] NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 20 December 2013
Accepted 7 May 2014
Available online 29 May 2015

Keywords:
Sentiment analysis
Czech language
Social media
Machine learning
Feature selection

ABSTRACT

This article describes in-depth research on machine learning methods for sentiment analysis of Czech social media. Whereas in English, Chinese, or Spanish this field has a long history and evaluation datasets for various domains are widely available, in the case of the Czech language no systematic research has yet been conducted. We tackle this issue and establish a common ground for further research by providing a large human-annotated Czech social media corpus. Furthermore, we evaluate state-of-the-art supervised machine learning methods for sentiment analysis. We explore different pre-processing techniques and employ various features and classifiers. We also experiment with five different feature selection algorithms and investigate the influence of named entity recognition and preprocessing on sentiment classification performance. Moreover, in addition to our newly created social media dataset, we also report results for other popular domains, such as movie and product reviews. We believe that this article will not only extend the current sentiment analysis research to another family of languages, but will also encourage competition, potentially leading to the production of high-end commercial solutions.

1. Introduction

Sentiment analysis has become a mainstream research field since the early 2000s. Its impact can be seen in many practical applications, ranging from analyzing product reviews (Stepanov & Riccardi, 2011) to predicting sales and stock markets using social media monitoring (Yu, Wu, Chang, & Chu, 2013). The users’ opinions are mostly extracted either on a certain polarity scale, or binary (positive, negative); various levels of granularity are also taken into account, e.g., document-level, sentence-level, or aspect-based sentiment (Hajmohammadi, Ibrahim, & Othman, 2012).

Most of the research in automatic sentiment analysis of social media has been performed in English and Chinese, as shown by several recent surveys, i.e., (Liu & Zhang, 2012; Tsytsarau & Palpanas, 2012). In Czech, there have been very few attempts, although the importance of sentiment analysis of social media became apparent, for example, during the

DOI of original article: http://dx.doi.org/10.1016/j.ipm.2014.05.001

[☆] This article is an extended version of our conference paper (Habernal, Ptáček, & Steinberger, 2013).

[☆☆] A publisher’s error resulted in this article appearing in the wrong issue. The article is reprinted here for the reader’s convenience and for the continuity of the special issue. For citation purposes, please use the original publication details; Information Processing and Management, 50/5, pp. 693–707

[*] Corresponding author at: Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic. Tel.: +420 377632456.

E-mail addresses: habernal@kiv.zcu.cz (I. Habernal), tigi@kiv.zcu.cz (T. Ptáček), jstein@kiv.zcu.cz (J. Steinberger).

http://dx.doi.org/10.1016/j.ipm.2015.05.006
0306-4573/© 2014 Elsevier Ltd. All rights reserved.

recent presidential elections.[1] Many Czech companies also discovered a huge potential in social media marketing and started launching campaigns, contests, and even customer support on Facebook—the dominant social network of the Czech online community with approximately 3.6 million users.[2] One aspect still eludes many of them: automatic analysis of customer sentiment of products, services, or even a brand or a company name. In many cases, sentiment is still labeled manually, according to one of the leading Czech companies for social media monitoring.

Automatic sentiment analysis in the Czech environment has not yet been thoroughly targeted by the research community. Therefore it is necessary to create a publicly available labeled dataset as well as to evaluate the current state of the art for two reasons. First, many NLP methods must deal with high flection and rich syntax when processing the Czech language. Dealing with these issues may lead to novel approaches to sentiment analysis as well. Second, freely accessible and well-documented datasets, as known from many shared NLP tasks, may stimulate competition, which usually leads to the production of cutting-edge solutions.[3]

This article focuses on document-level[4] sentiment analysis performed on three different Czech datasets using supervised machine learning. For the first dataset, we created a Facebook corpus consisting of 10,000 posts. The dataset was manually labeled by two annotators. The other two datasets come from online databases of movie and product reviews, whose sentiment labels were derived from the accompanying star ratings from users of the databases. We provide all these labeled datasets under Creative Commons BY-NC-SA licence[5] at http://liks.fav.zcu.cz/sentiment.

The rest of this article is organized as follows. Section 2 examines the related work with a focus on Czech research and social media. Section 3 thoroughly describes the datasets and the annotation process. In Section 4, we list the employed features and describe our approach to classification. Section 5 contains the results and provides a thorough discussion. Finally, Section 5.3 explores the influence of feature selection methods.

2. Related work

There are two basic approaches to sentiment analysis: dictionary-based and machine learning-based. Whereas dictionary-based methods usually depend on a sentiment dictionary (or a polarity lexicon) and a set of handcrafted rules (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), machine learning-based methods require labeled training data that are later represented as features and fed into a classifier. Recent attempts have also investigated semi-supervised methods that incorporate auxiliary unlabeled data (Zhang, Si, & Rego, 2012).

2.1. Supervised machine learning for sentiment analysis

The key point of using machine learning for sentiment analysis lies in engineering a representative set of features. Pang, Lee, and Vaithyanathan (2002) experimented with unigrams (presence of a certain word, frequencies of words), bigrams, part-of-speech (POS) tags, and adjectives on a movie review dataset. Martineau and Finin (2009) tested various weighting schemes for unigrams based on the TFIDF model (Manning, Raghavan, & Schütze, 2008) and proposed delta weighting for a binary scenario (positive, negative). Their approach was later extended by Paltoglou and Thelwall (2010) who proposed further improvements in delta TFIDF weighting.

The focus of current sentiment analysis research is shifting towards social media, mainly targeting Twitter (Kouloumpis, Wilson, & Moore, 2011; Pak & Paroubek, 2010) and Facebook (Go et al., 2009; Ahkter & Soria, 2010; Zhang et al., 2011; López, Tejada, & Thelwall, 2012). Analyzing media with a very informal language benefits from involving novel features, such as emoticons (Pak & Paroubek, 2010; Montejo-Ráez, Martínez-Cámara, Martín-Valdivia, & Ureña López, 2012), character *n*-grams (Blamey, Crick, & Oatley, 2012), POS and POS ratio (Ahkter & Soria, 2010; Kouloumpis et al., 2011), or word shape (Go et al., 2009; Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011).

In many cases, the gold data for training and testing the classifiers are created semi-automatically (Kouloumpis et al., 2011; Go et al., 2009; Pak & Paroubek, 2010). In the first step, random samples from a large dataset are drawn according to the presence of emoticons (usually positive and negative) and are then filtered manually. Although large high-quality collections can be created very quickly with this approach, it makes a strong assumption that every positive or negative post must contain an emoticon.

Balahur and Tanev (2012) performed experiments with Twitter posts as part of the CLEF 2012 RepLab.[6] They classified English and Spanish tweets with a small but precise lexicon, which also contained slang, combined with a set of rules that captured the manner in which sentiment is expressed in social media.

Finally, we would like to direct the reader to an in-depth survey by Tsytsarau and Palpanas (2012) for actual results obtained from the above-mentioned methods.

[1] http://www.mediaguru.cz/2013/01/analyza-facebook-rozhodne-o-volbe-prezidenta/ [in Czech].

[2] http://www.m-journal.cz/cs/jaky-je-skutecny-pocet-ceskych-uzivatelu-facebooku__s288x9161.html [in Czech].

[3] E.g., named entity recognition based on Conditional Random Fields emerged from CoNLL-2003 named entity recognition shared task.

[4] Or *post-level*, as documents correspond to *posts* in social media.

[5] http://creativecommons.org/licenses/by-nc-sa/3.0/.

[6] http://www.limosine-project.eu/events/replab2012.