Engineering Applications of Artificial Intelligence 51 (2016) 50–57

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets

Basilis Charalampakis, Dimitris Spathis *, Elias Kouslis, Katia Kermanidis

Department of Informatics, Ionian University, Corfu, Greece

ARTICLE INFO

Available online 2 March 2016

Keywords:
Irony detection
Text mining
Twitter
Politics

ABSTRACT

The present work describes a classification schema for irony detection in Greek political tweets. Our hypothesis states that humorous political tweets could predict actual election results. The irony detection concept is based on subjective perceptions, so only relying on human-annotator driven labor might not be the best route. The proposed approach relies on limited labeled training data, thus a semi-supervised approach is followed, where collective-learning algorithms take both labeled and unlabeled data into consideration. We compare the semi-supervised results with the supervised ones from a previous research of ours. The hypothesis is evaluated via a correlation study between the irony that a party receives on Twitter, its respective actual election results during the Greek parliamentary elections of May 2012, and the difference between these results and the ones of the preceding elections of 2009.

1. Introduction

Irony as a para-linguistic element is used to figuratively express a concept with a semantic meaning that is very different from its actual initial purpose. It is a challenging field for computational linguistics and natural language processing due to the high ambiguity and the difficulty to detect it, objectively. Language use is vigorous and creative; there is no pre-defined consensual agreement on how to recognize an ironic expression, due to the high subjectivity involved.

In the last decade, irony expression has been thriving on social networks and particularly Twitter, because of the 140 characters restraint on the status updates, being a perfect fit for good old one-liners. As a public social medium, users realize that their writings may be read and reproduced by potentially everyone, gaining popularity and followers. But this publicity, contrary to Facebook's real name policy, often has no direct consequences to their everyday lives, since the majority participate anonymously, using an avatar and a nickname.

This no-censorship state contributes to the freedom of expressing personal thoughts on tough, taboo, unpopular or controversial issues, part of which contains the political satire. Political satire is a significant part of comedy which specializes in drawing entertainment from politics. Most of the times, it aims just to please. By nature, it does not offer a constructive view by itself; when it is used as part of criticism, it tends to simply pinpoint the unexpected or different.

The high topicality of Twitter, combined with the ephemerality of political news, forms a state which is described as'echo chamber', a group-thinking effect on virtually enclosed spaces, amplified by repetition (Colleoni et al., 2014). As a result, the occasional user might write something political just to 'jump on the bandwagon', without an initial conscious aim to criticize. Adding to that, politics is a topic that almost everybody is familiar with and makes more sense from the engagement and attention side to write about Obama instead of an obscure book you just read.

Studies focus on the simultaneous usage of Twitter and the TV on circumstances like a political debate, where meta-talk tweets reveal critical scrutiny of the agenda or 'the debate about the debate' (Kalsnes et al., 2014).

In the rapidly changing web, there is a plethora of available text, especially from social networks, which is unlabeled, raw or unprocessed. Adding to the traditional supervised methods, there are quite a few techniques that enable us to take these huge unstructured data into account. An insight from our previous work was the subjectivity involved during the tagging of a text as ironic. Three of our authors who took up the tedious task of annotation could not agree on what should be considered as ironic or not. As a result, there cannot be a *gold standard corpus* of ironic tweets. This was our main motivation to explore semi-supervised techniques, since they take

* Corresponding author. Tel.: +30 6946354612.
(D. Spathis).
E-mail addresses: p11char@ionio.gr (B. Charalampakis), p11spat@ionio.gr, sdimitris@csd.auth.gr (D. Spathis), p11kous@ionio.gr (E. Kouslis), kerman@ionio.gr (K. Kermanidis).

http://dx.doi.org/10.1016/j.engappai.2016.01.007
0952-1976/

into account both train and test data. To be specific, the technique we chose is collective classification: a type of semi-supervised learning that presents an interesting method for optimizing the classification of partially-labeled data.

Considering the above, our empirical study tries to detect irony on a corpus of Greek political tweets by training a classifier, using appropriate linguistic features, some of which are proposed for the first time herein for irony detection. Our goal is to find a relation between the ironic tweets that refer to the political parties and leaders in Greece in the pre-election period of May 2012, and their actual election results. We compare the semi-supervised results with the supervised ones from a previous research of ours. Regarding the novelty of our study, this is a first exploration on the field of irony detection with semi-supervised learning and an application in politics.

The remainder of this paper is organized as follows: In Section 2, we present the related literature on the topics of irony detection, Twitter sentiment analysis and political expression. The next Sections 3 and 4 are dedicated to data preprocessing and its representation schema through the set of linguistic features that affect irony detection. The Section 5 describes the training procedure, the evaluation of the algorithms' performance and their test procedure on a large unlabeled dataset. An overview of the study limitations, future research prospects and a summary of the empirical study are described in Section 6.

2. Related work

The greater part of the literature on irony detection in computational linguistics is focused on English, but this is a first attempt to explore this area in the Greek language, to the authors' knowledge.

Reyes et al. (2013) attempt to detect irony by examining the corpus on the following features: signatures (concerning pointedness, counter-factuality, and temporal compression), unexpectedness (concerning temporal imbalance and contextual imbalance), style (as captured by character-grams (c-grams), skip-grams (s-grams), and polarity skip-grams (ps-grams)) and emotional scenarios (concerning activation, imagery, and pleasantness). These features work better when they are used as part of a coherent framework rather than used individually. They used multiple datasets in order to evaluate their hypothesis and achieved a precision of 0.79 at best. Classification is performed by Naïve Bayes and Decision Trees. Also a crisis management case study of the hashtag #Toyota is described.

A study by Rajadesingan and Liu (2014) discovered an interesting aspect of Twitter usage, an 'orientation phase' in which the user is gradually introduced to irony as one gains followers. The threshold of this phase is one's 30 initial tweets. The top features in decreasing order of importance for sarcasm detection are the following: Percentage of emoticons in the tweet, percentage of adjectives in the tweet, percentage of past words with sentiment score, number of polysyllables per word in the tweet, lexical density of the tweet. They evaluate using a J48 decision tree, logistic regression, and SVM to obtain an accuracy of 78.06%, 83.46%, and 83.05%, respectively.

The usual approach on similar irony detection studies on Twitter is to identify the two classes by hashtag analysis. However, this method creates noisy results with low accuracy (González-Ibánez et al., 2011; Liebrecht et al., 2013). Features used by Gonzalez were Lexical (unigrams, affective language, interjections and punctuation) and Pragmatic (positive smileys, negative smileys, and "@toUser" signs if a twitter is directed to another user). Algorithms used were SVM with SMO and Logistic Regression. Overall SMO outperformed LogR, with the best accuracy of 57% being an indication of the difficulty of the task. On the other hand, Liebrecht approached the same problem with the *Balanced Winnow* algorithm for classification. The strongest linguistic markers of sarcastic utterances were markers that can be seen as synonyms for #sarcasm hashtag. Testing the classifier on the top 250 of the tweets it ranked as most likely to be sarcastic, it attains a 30% average precision.

Twitter lexical analysis on Greek tweets has been the main subject of the research by Kermanidis and Maragoudakis (2013), examining the sentimental tagging in a supervised environment. Their hypothesis is focused on the positive / negative distinction, using statistical metrics such as count and frequency distributions. The alignment between actual political results and web sentiment in both directions was investigated and confirmed that there is a relation between political results and web sentiment. We use the same corpus of tweets in our study.

Apart from Twitter, similar techniques have been applied on Amazon reviews as well, making use of structured information of reviews versus the unstructured nature of Twitter. The accuracy results are encouraging due to the *semi-supervised* technique and the huge dataset, requiring human-annotator labor though, Davidov et al. (2010); Tsur et al. (2010). Features used were high-frequency words, content words, sentence length and punctuation. Results on the Twitter dataset are better than those obtained on the Amazon dataset, with accuracy of 0.947 with a k-nearest neighbors implementation.

Semi-supervised techniques on text mining were applied by Fangzhong and Markert (2009). Their approach involves Wordnet, like us, and they propose a subjectivity measure of each Wordnet entry. They suggest a semi-supervised minimum-cut framework that makes use of both WordNet definitions and its relation structure. Minimum-cut is a technique used at graph theory, which uses pairwise relationships between the data points in order to learn from both labeled and unlabeled data. The semi-supervised approach achieves the same results as the supervised framework with less than 20% of the training data.

In the emerging area of *active learning*, where the learning algorithm is able to interactively query the researcher to obtain the desired outputs at new data points, there is some ongoing research. Gokhan et al. (2005) wanted to reduce the labeling effort for spoken language understanding from data gathered at AT&T call centers. The examples that are classified with higher confidence scores (not selected by active learning) are exploited using two semi-supervised learning methods. This enables them to exploit all collected data and alleviates the data imbalance problem caused by employing only active or semi-supervised learning. Their results indicate that it is possible to reduce human labeling effort significantly. Similar technique, namely *collective learning*, was followed by Santos et al. (2011), where they propose a new method that adopts a collective learning approach to detect unknown malware. Their empirical research demonstrates that the labeling efforts are lower than when supervised learning is used, while maintaining high accuracy rates. Collective classification is an approach that uses the relational structure of the combined labeled and unlabeled dataset to enhance classification accuracy (Neville and Jensen, 2003).

Research by de-la-Peña-Sordo et al. (2013) studied the comparison between collective learning and supervised techniques, pretty similar with our methodology. Apart from that, quite similar was their topic of detecting trolling comments on a Spanish platform like Digg or Reddit and their lexical features selection, since irony and trolling may seem indistinguishable in some cases. Their approach obtains nearly the same accuracy than the best supervised learning approaches.

Another study dealing with online opinion and reviews, again by Reyes and Rosso (2011), examined Amazon and Slashdot.com