Knowledge-Based Systems 108 (2016) 110–124

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A hybrid approach to the sentiment analysis problem at the sentence level

Orestes Appel[a], Francisco Chiclana[a,*], Jenny Carter[a], Hamido Fujita[b]

[a] Centre for Computational Intelligence, De Montfort University, Leicester, UK
[b] Intelligent Software Systems Laboratory, Iwate Prefectural University, Takizawa, Iwate, Japan

ARTICLE INFO

Article history:
Received 12 February 2016
Revised 17 May 2016
Accepted 19 May 2016
Available online 20 May 2016

Keywords:
Sentiment analysis
Semantic rules
Fuzzy sets
Unsupervised machine learning
SentiWordNet
Naïve Bayes
Maximum entropy
Computing with sentiments

ABSTRACT

The objective of this article is to present a hybrid approach to the Sentiment Analysis problem at the sentence level. This new method uses natural language processing (NLP) essential techniques, a sentiment lexicon enhanced with the assistance of SentiWordNet, and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences, which provides a foundation for computing with sentiments. The proposed hybrid method is applied to three different data-sets and the results achieved are compared to those obtained using Naïve Bayes and Maximum Entropy techniques. It is demonstrated that the presented hybrid approach is more accurate and precise than both Naïve Bayes and Maximum Entropy techniques, when the latter are utilised in isolation. In addition, it is shown that when applied to datasets containing snippets, the proposed method performs similarly to state of the art techniques.

1. Introduction

The human brain has an inherent ability to detect emotion or sentiment in written or spoken language. Social media and other tools related to today's world have increased the number of sources and volume of information dramatically, and the ability of people to process all that is being seriously compromised. Hence, the ability of having computers to go at high speed through the myriad of data available and extract sentiment and/or opinions would be greatly beneficial. *Sentiment Analysis* (SA) is one of the research areas of fastest growth in the last few years. A number of definitions about it are available. Typically, the main objective of SA is to establish the attitude of a given person with regard to some subject, paragraph or document.

Bing Liu [34] defines an opinion as follows: "In an opinion we find the following items: Opinion targets (entities and their features/aspects), Sentiments (positive or negative), Opinion holders (persons who hold the opinions) and Time (when opinions are expressed). Opinions then can be: (a) Direct opinions, (b) Indirect opinions, or Comparative Opinions. A regular opinion is defined as a quintuple (e_j, a_{jk}, so_{ijkl}, h_i, t_l) where e_j is a target entity, a_{jk} is an aspect/feature of the entity e_j, so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l." Usually, so_{ijkl} (semantic orientation) is *positive, negative* or *neutral.*

Quite often, the most commonly applied techniques to address the SA problem belong either in the category of text classification Supervised Machine Learning (SML) (methods like Naïve Bayes, Maximum Entropy or Support Vector Machine (SVM)) or text classification Unsupervised Machine Learning (UML). However, it seems that fuzzy sets, considering their mathematical properties and their ability to deal with vagueness and uncertainty (characteristics present in Natural Languages) are well-equipped to model sentiment-related problems. As it is well known, fuzzy relations have been extensively used in disciplines as dissimilar as linguistics [15], clustering [46] and decision-making [62], among many others. Thus, a combination of techniques -which would fit the concept of *hybrid*- could be successful at addressing the SA challenges by exploiting the best in each technique. In the next paragraph we will address our motivation for exploring this realm of possibilities.

Dzogang et al. stated in [18] that usually authors refer mainly to psychological models when attacking the SA problem. However, other models may be successful as well in this domain. As per Dzogang et al. "it must be underlined that some appraisal based approaches make use of graduality through fuzzy inference and

* Corresponding author.
E-mail addresses: orestes.appel@email.dmu.ac.uk (O. Appel), chiclana@dmu.ac.uk (F. Chiclana), jennyc@dmu.ac.uk (J. Carter), issam@iwate-pu.ac.jp (H. Fujita).

http://dx.doi.org/10.1016/j.knosys.2016.05.040
0950-7051/© 2016 Elsevier B.V. All rights reserved.

fuzzy aggregation for processing affective mechanisms ambiguity and imprecision." When dealing with SA, Bing Liu [35], one of the main world experts in this area, says that "we probably relied too much on Machine Learning". When it comes to discussing the progress in the SA discipline, Poria et al. [52] introduced a novel idea to concept-level sentiment analysis, which involves combining together linguistics, common-sense computing, and machine learning, aiming to improve precision on polarity detection. This approach of merging techniques is essentially a *hybrid style* of compounding the power of several tools. Considering all of the arguments above, we believe that the following concepts could be applied in combination:

– The concept of *graduality* expressed through fuzzy sets.
– The idea that other tools, together, besides Supervised Machine Learning in isolation, may be viable as well when extracting sentiment from text (especially, if combined with other techniques).
– The positive contribution that NLP tools, semantic rules and a solid opinion lexicon can have in identifying polarity.

Based on these arguments, our research hypothesis can be stated as follows:

Hypothesis 1. A sentiment analysis method at the sentence level, using a combination of sentiment lexicons, NLP essential tools and fuzzy sets techniques, should perform *same or better* than today's accepted text classification supervised machine learning algorithms when the latter are utilised in isolation.

We are establishing the aforementioned hypothesis as we are in search of a sentiment analysis method that closely resembles the way human beings deal with this topic. We expect in the future to be able to expand our method to deal with human-aspects like humour, irony and sarcasm, which most likely will require providing context. However, it is our belief that the sooner we get closer to the way humans process sentiment, the better positioned we will be to take the next step. We call our proposed system a *hybrid*, because of the fact that it uses a combination of methods and techniques that stem from different research disciplines: fuzzy set theory, natural language processing algorithms and linguistic systems.

The rest of this paper is organised as follows: Section 2 addresses related work. Section 3 presents the research methodology with a focus on three main components: the process to follow, the data to be used, and the performance measurement of the SA solution. Section 4 describes in detail the main components of the proposed Hybrid approach to the SA problem at the sentence level. This section culminates with the presentation of both the Hybrid Standard Classification (HSC) and the Hybrid Advanced Classification (HAC) methods, which adds graduality estimation to polarity identification (the latter being performed by HSC). Section 5 shows the experimental results, starting with the outcome of using two well known and accepted SA Machine Learning methods that originate in the text classification field: Naïve Bayes and Maximum Entropy. This section also includes a comparison analysis between the results obtained applying the proposed hybrid method against the aforementioned machine learning techniques. A comparison against the state of the art is shown as well, as a reference. In closing, some concluding remarks and future work plans are presented in Section 6.

For a summarised survey on Sentiment Analysis, please refer to the article by Appel et al. [2]. For a complete review of the evolution of the SA field, please refer to the thorough work of Ravi and Ravi [54]. For a focused account on recent advances in SA techniques, Cambria [7] is recommended.

2. Related work

SA is a discipline that has seen a lot of activity since about 2000, when Rosalind Picard published her important book *Affective Computing* [51], i.e. "computing that relates to, arises from, or deliberately influences emotion or other affective phenomena". When one reviews the most recent trends in the field, of which *Sentic Computing*, led by Erik Cambria [6,10,11] is a good example, it becomes evident the amount of effort that has gone into researching SA. As per their creators [6], sentic computing "relies on the ensemble application of common-sense computing and the psychology of emotions to infer the conceptual and affective information associated with natural language." Other articles worth mentioning explore topics around sentiment lexicon-based techniques, like the contributions of Cho et al. [14] and Huang et al. [31]. The work by Bravo-Márquez et al. [5], on the use of multiple techniques and tools in SA, offers a complete study on how several resources that "are focused on different sentiment scopes" can complement each other. The authors focus the discussion on methods and lexical resources that aid in extracting sentiment indicators from natural languages in general. A comprehensive work on *semantic analysis* is Cambria et al. [8], while Schouten and Frasincar work [56] provides a complete survey specific to aspect-level sentiment analysis.

A number of researchers have explored the application of hybrid approaches by combining various techniques with the aim of achieving better results than a standard approach based on only one tool. Indeed, this has been done by Poria et al. in [52] where a novel framework for concept-level sentiment analysis, Sentic Pattern, is introduced by combining linguistics, common-sense computing, and machine learning for improving the accuracy of tasks such as polarity detection. The authors claim that "by allowing sentiments to flow from concept to concept based on the dependency relation of the input sentence, authors achieve a better understanding of the contextual role of each concept within the sentence and, hence, obtain a polarity detection engine that outperforms state-of-the-art statistical methods". When no matching sentic pattern is found in SenticNet [9] they resort to Supervised Machine Learning. The hybrid approach put forward in the present article uses a dictionary of words frequencies and occurrences instead to address the case when a word is not found in its lexicon. An additional difference with regard to lexicon data centres around polarity ranges. SenticNet enables polarities to be measured in the interval $[-1, 1]$ while SentiWordNet allows polarities to be in the range $[0, 1]$. As it will later be apparent, the hybrid method presented here creates the foundation for the introduction of the concept of *computing with sentiments*, which derives from Zadeh's innovative idea of computing with words [74].

Related to the use of lexicons in SA approaches, it is worth mentioning the following two research efforts. The first one is by Hajmohammadi et al. [23] on a novel learning model based on the combination of uncertainty-based active learning and semi-supervised self-training approaches, which also addressed the challenges associated with the construction of reliable annotated sentiment corpora for a new language. This research provided us with important lessons on the difficulties and potential pitfalls of embracing such a task and how to better deal with it. The other research effort is by Hogenboom et al. [27,28] on the use of emoticons as modifiers of the sentiment expressed in text and as vehicles of sentiment by themselves. According to the findings of the authors, the sentiment associated to emoticons dominates the sentiment conveyed by the text fragment in which these emoticons are embedded. In their work they introduce a sentiment lexicon, which is a point of commonality with the research presented here, as well as a cleverly designed emoticon lexicon.

In [12], Cambria et al. explore how the high generalisation performance of extreme learning machines (feed forward neural