Hybrid Data-Driven and Rule-Based Sentiment Analysis on Greek Text

Angela Braoudaki
University of Crete, CSD
antzibraoudaki@gmail.com

Eleni Kanellou
FORTH ICS
kanellou@ics.forth.gr

Christos Kozanitis
FORTH ICS
kozanitis@ics.forth.gr

Panagiota Fatourou
FORTH ICS & University of Crete, CSD
faturu@ics.forth.gr

ABSTRACT

Sentiment analysis is a developing field dealing with the detection of sentiments or opinions expressed in a written text. It commonly relies on methods based on Machine Learning and while some standardized methodologies exist, the sheer volume of data needed in order to train a Deep Learning network makes it important to use designs that are sufficiently complex to accurately detect sentiment but also simple enough to be scalable and performance efficient. In this paper, perform sentiment analysis on a body of hotel review texts collected online and annotated by a linguistics tool with sentiment-defining tags. We propose four different Deep Learning network designs, which we train either on the review texts, the review texts plus some information on the tag annotation, or the annotation of the text alone, and we present the trade-offs between scalability and efficiency that each setup offers.

1 MOTIVATION

Sentiment analysis is a process by which the emotions and opinions codified in a raw text can be detected and classified through a systematic and automated process. The process commonly relies on natural language processing, computational linguistics, and text analysis, among others. Sentiment analysis can help businesses investigate consumer sentiments expressed in online platforms such as forums or social media, and use the conclusions in order to come up with better marketing or improved offered service. However, the volume of data that can be accumulated this way is enormous and processing it can also pose important difficulties, when treating the differences in language, writing style, colloquialisms, and other such characteristics. In view of this, the use of machine learning (ML) comes as a practical solution.

Applying ML on natural language text poses several challenges in practice, such as extensive preprocessing in order to clean up the raw data, ML-friendly codification of words, and the use of complex Recurrent Neural Networks (RNN) networks [2], such as LSTMs [1], in order to properly interpret text content. The larger the data volume to be processed, the more imperative is the need of deploying the ML system on efficient computing platforms, such as distributed setups, or ones containing data-parallel accelerators. However, the more complex the ML network and the larger the amount of its learning parameters, the more difficult it is to efficiently exploit the scalability that these computing platforms have to offer. Thus, it is important to use a text representation and design an ML model both powerful enough to accurately analyze written reviews but also, simple enough to be efficiently deployed and scalable.

2 CONTRIBUTION AND METHODOLOGY

In this work, we deal with a problem of classifying a hotel visitor review into one of five categories expressing customer satisfaction. The goal of our work is to explore how simple a design can be while still offering accurate sentiment prediction. For this, we come up with variants of a Deep Learning (DL) network and experimentally evaluate their performance. The reasoning for aiming for simplicity is that a DL model that has a simple architecture, uses less time and less memory resources and can be more scalable. We propose and compare four DL architectures for the prediction of the score of a review, given a review text. The texts are hotel reviews, written in the Greek language. These hotel reviews are comprised by a piece of text and a numeric score; The score provides an overall grade of the establishment, while the text expresses more details and further nuances of the user's opinion.

Previous work in this area focuses on several aspects of sentiment to be detected, for example in [3], where a fully unsupervised probabilistic modeling framework is used, which detects sentiment and topic simultaneously from the text corpus. Similarly, there are works such as [4], which relies on a lexicon of words and phrases tagged as positive and negative. Then, the contextual polarity of each word is determined by that of the previous word in the text.

The reviews we use have undergone a software pre-processing via a linguistics-based tool, which performs tagging, i.e. which annotates the text with sentiment qualifiers based on words it encounters[1]. Our baseline model is an LSTM network using the entire review text as input, achieving 0.7891 accuracy. When using the tags as additional input, accuracy can reach 0.7896, but this also is the design with most training parameters, longest training time and highest memory consumption. When using an LSTM model with only the review tags as input, we obtain up to 0.6964 accuracy for a drastic reduction of training parameters and training time. Finally, by employing a simple Dense network that trains on the review tags only, we can achieve up to 0.7 accuracy while still requiring significantly lower training time and parameters.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.
[2] L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., USA, 1st edition, 1999.
[3] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, USA, 2009.
[4] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the H*, USA, 2005.

[1] The tagging is performed by a third party and is out of the scope of our work.