Language Models as Boosters for Sentiment Analysis

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Ziegler Vorname: Steven

[redacted] [redacted]

Prüfer: Prof. Dr. Michael H. Breitner
Betreuer: Rouven Wiegard

Hannover, den 29.09.2019

List of Contents

1 Introduction

1.1 Motivation and Relevance of the Topic

"The world's most valuable resource is no longer oil, but data" (The Economist, 2017).

A lot of resources are precious in their raw form. But real value creation starts by processing and aligning them with activates, resulting in products and services (Woiceshyn & Falkenberg, 2008). In terms of the resource data, it is roughly estimated that 20% of the useful company data are pre-processed or structured (Gharehchopogh & Khalifelu, 2011). That means they are stored in databases, aligned to rows and columns and can be used as an input for predictions, targeted advertising, service or product development and decision support (Masona & Rainardi, 2008; Qiu et al., 2010). The other 80% of the data are unstructured. It contains no meta-data or in other words: "There is no data about the data" (Ibm, 2013).

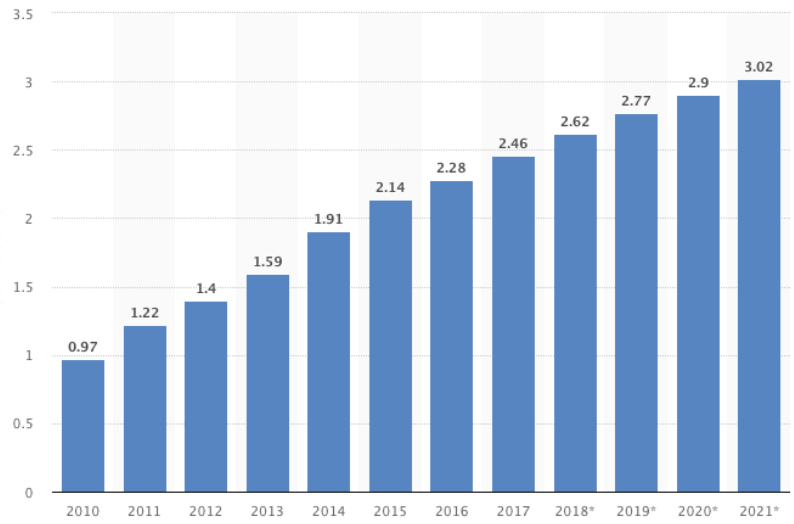

Figure 1: Number of Social Media Users Worldwide from 2010 to 2021 in Billions (Statista, 2019)

Consequently, these valuable and powerful resources are untapped. Unstructured data are, for example documents, images, audio, text, emails and Internet websites (Gharehchopogh & Khalifelu, 2011). The most frequent type of unstructured data is text data. That enhances a wide range of forms: from word documents to content of blogs and social media (Inmon & Linstedt, 2014). One main driver of this massively

amount and continuously increase of these unstructured text data has been the development of the web 2.0, which has allowed users to create their own content (Mason & Rennie, 2007). The number of social media users is constantly increasing (Statisa, 2019). Hence, the amount of unstructured created text data, as well.

This enables a great opportunity to make use of these unused resources and create value by structuring them in the first place, before using them as inputs for the application domains mentioned above. It is not surprising that the five most valuable companies in the world: Apple, Alphabet, Microsoft, Amazon and Facebook, that all set up their business model around the key resource data, make major efforts in the research of making use of these text data (Statista, 2018).

One of the most important research domains processing text data is sentiment analysis. Sentiment analysis has focused on this task of analyzing an incoming message and classifies the underlying sentiment from positive to negative in different ranges (TDW, 2018). It is mainly used for getting insights into unstructured text that comes from web chats, social media emails, forums and comments (Gharehchopogh & Khalifelu, 2011). It is a sub-domain of Natural Language Processing (NLP), which targets to analyze and represent the human languages by using computational techniques (TDW, 2018; Young et al., 2018).

Sentiment Analysis is one of the fastest-growing research areas in computer science. As a consequence, there have been developed good performing models and techniques to classify the underlying sentiment of text data (Mäntylä et al., 2018). The Sentiment analysis of a text can be conducted on different levels. Usual sentiment analysis classifies the sentiments' polarity on a sentence level. This method leads to a problem if the sentiment polarity of a sentence is not only determined by the content but also by certain considered aspects. But that is often the case, especially in terms of reviews. A real information gain for the review "The *display* is ok, but the *hard disk* is slow." does not arise by classifying the sentiment of the whole sentence then rather classify it regarding the aspect *display* and *hard disk* (Wang et al., 2016). To handle this issue, the Aspect-based sentiment analysis is applied, where the sentiment polarity gets classified on an aspect level. This leads to specific information from an analyzed text or review about a product or service and their aspects that contain highly valuable insights for users and companies (Liu, 2012; Pontiki et al., 2016).

1.2 Research Question

A major issue in Aspect-based sentiment analysis, and concerns NLP in general, is the conversion of text into a form the machine can “understand” it. The common approach for this is the use of word embeddings, which describe high dimensional vectors containing semantic information of words (Wang et al., 2016). These vectors come from a specially designed model that was trained to capture the semantic meanings of words. Commonly the vectors come from the word2vec or GLOVE model (Kamath et al., 2019). These externally generated embeddings then get inserted as an extra layer between the inputted text and the actual model that performs the down-streaming aspect-based sentiment analysis task (Wang et al., 2016).

Since the year 2018, a paradigm change takes place in handling NLP tasks. Instead of focusing on models that perform down-streaming task, the spotlight has shifted to place special emphasis on the conversion of text data using language models (Howard & Ruder, 2018). These language models overcome many limitations of the usual applied word embeddings and have the ability to “understand” a language more accurate and comprehensive. The language understanding is not just compromised in a fixed-sized word embedding layer than instead rooted in the whole architecture of the model, learned in a pre-training stage (Kamath et al., 2019). Based on that highly improved language understanding the NLP task to be accomplished gets performed within the same architecture, where only the models’ last layer gets modified regarding the actual down-streaming task. This is called the fine-tuning stage. These models are tremendously large and have to be trained computably expensive on an enormous corpus to capture the language. Therefore, the pre-trained models, containing the stored language understanding, are available open-source (Devlin et al., 2018). Looking at the latest performance result of these models performing several NLP tasks with these models as backbones a senior scientist of Google Brain called it not less than “*A new era of NLP...”*. (Twitter, 2018)

This new language model approach, with the greatly improved understanding of language, can especially represent a booster for the performance of fine-grained NLP tasks as aspect-based sentiment analysis. The success of mastering this task lies in the detailed understanding of the text input, as it requires the classification of the sentiment polarity regarding a particular aspect. Aspect sentiment analysis is commonly the generic term of two sub-tasks: the aspect extraction and the aspect sentiment polarity classification (Tsytsarau & Palpanas, 2012). The aspect extraction deals with the challenge of extract and detect features from text data. The aspect

sentiment analysis then classifies the polarity of the sentence regarding the given aspects. Both sub-tasks represent their own domain of research. This work concentrates on the aspect sentiment polarity classification. Song et al. (2019) presented incredible results for aspect-based sentiment classification. The architecture relies on the award-wining language model BERT, where just the output layer got modified regarding the aspect sentiment classification task. This approach opens new opportunities in the field of aspect-based sentiment analysis. The validation of these remarkable results in the aspect sentiment analysis domain, achieved by language models is the aim of this work that results in the research question (RQ):

How to design a topology of a language model performing Aspect-based Sentiment Classification? And how well does is it perform compared to the common approach?

To answer this research question, two neuronal network models have been developed regarding the task of aspect sentiment classification. One follows the common approach of handling the aspect sentiment classification, the other one in the langue model manner. On the one hand, the BERT language model, following Devlin et al. (2018) and Song et al. (2019), represents the language model approach. On the other hand, the ATAE-LSTM model, following Wang et al. (2016), represents the common approach. Both models are state of the art architectures in terms of aspect sentiment classification. To evaluate their performance, they get applied to the SemEval 2014 Task 4^2 Subtask 2, which is an accepted benchmark in assessing models' capability regarding the classification of a reviews sentiment for a given aspect. Additionally, both models had to prove their generalization ability in terms of predicting several reviews from the test dataset as well as in a short case study.

The next section (2) gives a brief overview of the relevant theoretical background. The section includes a subsumption of aspect-based sentiment classification and neural networks as well as their intersections. The section also describes the development of language representation towards language models. Section (3) describes the development of the ATAE-LSTM and the BERT model in detail, including data pre-processing. Section (4) explains the underlying dataset and the task to perform as well as the training regularization techniques. In section (5), the results regarding the SemEval 2014 Task 4^2 are presented. Moreover, the models were assessed with regard to predict the polarity of several reviews from the test dataset and a case study. The obtained results get set in context and explained in section (6). Finally, section (7) comes up with a short conclusion and an outlook.

7 Conclusion

The aim of this paper is the evaluation of language models in terms of their performance regarding aspect-based sentiment classification. In order to obtain conclusive results, two models have been developed: the BERT language model according to Delvin et al. (2018) and Song et al. (2019) as well as the ATAE-LSTM model according to Wang et al. (2016). Both models describe a state of the art approach handling the aspect-based sentiment analysis of the language method or the common neural network approach. In this work, both models could achieve or exceed the results from the underlying papers regarding the benchmark aspect-based sentiment analysis classification SemEval 2014 Task 4^2 Subtask 2. Comparing the two models' results, attest to the BERT model n large performance improvement compared to the ATAE-LSTM model with regard to the fine-grained classification task. Both models performed best on the restaurant dataset with an accuracy of 0.8571 and 0.7857. The superiority of the BERT model can be summarized in three reasons: architectural elements, the special two-stage training procedure, as well as the size of the model. Concerning the architecture, the transformer encoder block and the tokenization have to be mentioned. Both elements are outlined in detail in this work. The same applies to the language model particular, two-stage pre-training and classification procedure. Combined with the enormous size of the language models, it is capable of learning a deep and comprehensive language understanding in the pre-training. That boosts the performance of the following aspect-based sentiment analysis classification compared to the ATAE-LSTM model. Several conducted predictions of reviews considering an aspect confirm that.

Transferring these high accuracy values into the practice has to be considered with reservation since the underlying test and training data have shortages as an imbalanced distribution among the polarity classes. Moreover, the models perform of varying quality the number of aspects and polarity classes implied in a review. These points show that language models are on the right way in a comprehensive language understanding, boosting the following down-streaming task, but still with air up. The approach of rooting them into the whole models' architecture instead of fixed-sized vectors is the right way must be expanded further. Special effort should be put in a perfect understanding and optimization of the multi-head attention mechanism of the transformer encoder block that represents the heart of the architecture. As this well-performing technique for sequential data has just released recently, a lot of examinations are still ahead. For example, Kovaleva et al. (2019) showed that *"there is a limited set of attention pat-terns that are repeated across different heads"*; hence

a lot of redundancy (Kovaleva et al., 2019: P. 1). The enhancing of the language understanding than, results in a better performance of the down-streaming tasks as aspect-based sentiment analysis as this work showed.