W&G-BERT: A CONCEPT FOR A PRE-TRAINED AUTOMOTIVE WARRANTY AND GOODWILL LANGUAGE REPRESENTATION MODEL FOR WARRANTY AND GOODWILL TEXT MINING

Lukas Jonathan Weber[1], Alice Kirchheim[2], Axel Zimmermann[3]

[1]Mercedes-Benz AG, Stuttgart, Germany
[2]Department Mechanical Engineering,
Helmut-Schmidt-University, Hamburg, Germany
[3]esz-partner Eber, Schwarzer, Zimmermann GbR, Kirchheim,
Baden-Württemberg, Germany

ABSTRACT

The request for precise text mining applications to extract information of company based automotive warranty and goodwill (W&G) data is steadily increasing. The progress of the analytical competence of text mining methods for information extraction is among others based on the developments and insights of deep learning techniques applied in natural language processing (NLP). Directly applying NLP based architectures to automotive W&G text mining would wage to a significant performance loss due to different word distributions of general domain and W&G specific corpora. Therefore, labelled W&G training datasets are necessary to transform a general-domain language model in a specific-domain one to increase the performance in W&G text mining tasks.

The article describes a concept for adapting the generally pre-trained language model BERT with the popular two-stage language model training approach in the automotive W&G context. We plan to use the common metrics recall, precision and F1-score for performance evaluation.

KEYWORDS

Natural language processing, Domain-specific language models, BERT, Labelled domain-specific datasets, Automotive warranty and goodwill.

1. INTRODUCTION

In the automotive W&G sector, an unprecedented number of written feedback texts are generated by global workshops and customer studies every day, making the manual analysis of these texts and the extraction of actionable insights from them an extremely difficult task for individuals. Various of such customer feedbacks are used to analyze and predict current and future product weaknesses. Both for the resulting failure elimination and for internal reporting, the mining of customer feedback plays a crucial role in product warranty. A large part of the written feedbacks are “not in a [directly] machine-processable [...] data [form]" available, which can be made usable for decision-making by text mining approaches [2]. Due to this, the demand for accurate text mining tools to extract information of company based W&G data is steadily increasing. The

David C. Wyld et al. (Eds): CCNET, NLTM, CICS - 2022
pp. 29-35, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120304

analytical competence of text mining methods for information extraction is among others based on the developments and insights of deep learning techniques applied in *natural language processing* (NLP) [3, 4]. Current deep-learning language models are trained on general text corpora (e.g. Wikipedia). Especially in domains such as medicine [5], biology [4, 6], finance [7], common science [8] and W&G, word distributions of general and specific corpora are different, which can be a significant performance problem for NLP-based text mining models [4]. Therefore, the application of generally trained language models to W&G tasks is only suitable to a limited extent [9]. Labelled W&G training datasets are necessary to transform a general-domain language model in a specific-domain one to increase the performance in W&G text mining tasks. To the best of our knowledge, there are not any existing labelled training datasets in the W&G domain public available.

In this paper, we describe the concept of creating a pre- and fine-tuned W&G-specific language model *BERT* [1]. Therefore, in Section 2 we briefly describe the already published articles in the area of language models with a deep dive in their domain specific adaptations. In section 3, we describe the creation of suitable training datasets for the following pre-training and fine-tuning activities of a W&G-specific context-dependent language model. Finally, in section 4 we give an outlook on the overall text mining architecture in which a W&G specific language model can be integrated.

2. RELATED WORK

Silvestri *et al.* [10] describe in their publication the significant performance improvement of NLP tasks through the use of the word embedding language models *word2vec* and *GloVe*. The basis for the *word2vec* language model is a continuous skip-grams (SG) or a continuous Bag-of-Words (CBOW) architecture. The SG architecture predicts the context of an entity based on the entity, whereas the CBOW approach predicts a missing entity based on the immediate context [11, 12]. In contrast to Mikolov *et al.* [12], Jeffrey Pennington *et al.* [13] developed *GloVe* as a global log-bilinear regression model for unsupervised word representation learning that outperforms SG- and CBOW- models on word analogy, word similarity and named-entity recognition tasks. Onan [14] increases the potential of the *word2vec* language model in the topic extraction domain by developing a two-stage architecture using an improved word embedding model and a cluster ensemble framework. The improved word embedding model consists of the fusion of conventional word embedding architectures (*word2vec*, *pos2vec*, *word-position2vec* and *LDA2vec*). In addition, the well-known clustering algorithms *K-means* [15], *k-modes* [16], *k++* [17], *self-organising maps* [18] and *DIANA* [19] are combined specifically unweighted into a cluster ensemble framework. In the *word2vec* and *GloVe* language models, emergent entities in a different context are characterised with the same word vector (context independent language models). Since the same entities take on a different meaning in a different context, the results of the already discussed language models can be improved. The context-based consideration of entities is taken into account in the model *Bidirectional Encoder Representations from Transformers* (BERT) [1]. The pre-training of the language model *BERT* is realised with the help of the static *Masked Language Model* (MLM) [20] and the *Next-Sentence-Prediction* (NSP). In contrast to Devlin *et al.* [1], Liu *et al.* [21] chose a dynamic masking approach in their robustness-optimised language model *RoBERTa* without implementing the NSP task. *BERT* by Devlin *et al.* [1] provides the basis for multi-layered applications in the fields of medicine, biology and finance. The simple adaptation of the *word2vec*, *ELMo* [22] or *BERT* language models would yield weak results in domain specific *Named Entity Recognition* (NER) and *Relation Extraction* (RE) metrics in domain-specific application case, as they were pre-trained on general input data. The authors Lee *et al.* [4] developed the *BioBERT* language model which was pre-trained on bio-medical data. The superiority of a specifically pre-trained *xBERT* language model over conventional language models was also confirmed by Peng *et al.* [23]. Furthermore, Beltagy *et al.* [8] present the potential of their pre-trained *SciBERT* language model in their publication. By implementing a sentence-piece library based on unsupervised tokenisation of the scientific corpus using WordPiece, this generates a performance advantage in the F1 score compared to the *BioBERT*. In addition, Zhuang Liu *et al.* [7] designed the *FinBERT* language model. *FinBERT* is characterised by a parallelised pre-training approach modelled on the Horovod architecture [24] using a mixed-precision training approach [25] on general and finance-specific corpora. The language model developed by Zhuang Liu *et al.* shows not only an improved performance compared to *BERT*, but also considerable success in the common quality validation metrics for a small pre-training corpus (20% of the original corpus). In addition to the context-dependent architecture *BERT* and the specific language models derived from it by Lee *et al.*, Peng *et al.*, Beltagy *et al.* and Zhuang Liu *et al.*, the language model *ERNIE 1.0* developed by Sun *et al.* [26] is based on a masked language models approach. However, Sun *et al.* differentiates itself from the general *BERT* approach by extending the MLM approach in terms of basic-level, phrase-level and entity-level masking. Furthermore, *ERNIE 1.0* differs from *BERT* by a five-stage pre-training phase and the application of the Dialogue Language Model (DLM) to improve the learning ability of semantic representations. In contrast to *ERNIE 1.0*, its extension *ERNIE 2.0* [27] is based on a continuous multi-task learning pre-training architecture. Here, lexical, syntactic and semantic information are learned by using several word-aware, structure-aware and semantic-aware tasks. This allows Sun *et al.* [27] to ensure that the learned parameters encode the previously learned knowledge. The evaluation of the continuous multi-task training method using defined training tasks which demonstrates the potential of this approach. Furthermore, the *ERNIE 2.0* language model represents a considerable increase in performance in the common metrics of the GLUE benchmark compared to *XLNet* [28] and *ERNIE 1.0.*

3. Approach

As shown in Figure 1, we develop our W&G-BERT model based on the architecture of *BERT*, which was elaborated as one of the state-of-the-art language representation models in section 2. Furthermore, we use the popular two-stage language model training approach, which consists of a pre-training and fine-tuning phase to increase the performance of its application on domain specific tasks. Subsequently in this section, we describe the proposed *BERT* model and the planned pre-training and fine-tuning activities of W&G-BERT.

3.1. BERT: Bidirectional Encoder Representations from Transformers

The process of learning word representations in an unsupervised way from a great amount of unannotated semi and unstructured data is a well-known and long-established method. Developed models in the past such as *word2vec* and *GloVe* are focused on learning context independent word representations [4]. To improve the performance of language models, more recent ones like *ELMo*, *XLM* [30], *XLNet*, *ERNIE1.0* and *ERNIE 2.0* are focused on learning context dependent word representations.

BERT is a context dependent word representation language model which is based on a masked language model and a next sentence prediction pre-training architecture by using bidirectional transformer encoder [31]. *BERT* uses the MLM approach by predicting randomly masked tokens in given sentences, while the NSP approach guarantees the sentence-relationship understanding between different sentences in the same corpus. This allows Devlin *et al.* to ensure that the inputs are represented efficiently as the sum of token-, segmentation- and position embeddings.

3.2. Pre-Training W&G-BERT

Authors like Alsentzer *et al.*, Beltagy *et al.*, Gu *et al.*[6], Lee *et al.*, Peng *et al.* and Zhuang Liu *et al.* provided in their results that general pre-trained models usually achieve poor results in common performance metrics on domain specific text mining tasks. Similar to biomedical or scientific texts, automotive W&G texts contain of a huge amount of domain specific terms and expressions, which requires expert knowledge of the corresponding researcher (e.g. *condensation water drain hose*, *center fill* or *open pore wood trim*). In the planned work, we pre-train W&G-BERT on English automotive specific *customer surveys* and *workshop complaints* in an unsupervised way with the MLM-approach in addition to the already pre-trained general corpora and initially weighted *BERT* model (see Table 1).

Table 1. Pre-training corpora

Domain	**Corpus type**	**English (approx. tokens)**
General[1]	Wikipedia	2500M
	BookCorpus	800M
Auto. warranty and goodwill	Customer survey	~50M
	Workshop complaints	~200M

3.3. Fine-Tuning W&G-BERT

With manageable architectural modification, W&G-BERT can be successfully applied to several text mining tasks. Therefore, we plan to fine-tune W&G-BERT on two self-constructed supervised automotive W&G specific datasets and their related text mining tasks: *NER* and *RE*.

W&G Named entity recognition is one of the most important W&G text mining tasks, which deals with the identification of domain-specific automotive W&G expressions and terms. *BERT* and its specifications [4–8] are build up on a multi-layer transformer encoder. Therefore, the simple specification process will be adapted to generate a *BERT* based W&G-BERT-language model. To the best of our knowledge, no specific fine-tuning datasets exist in the NER area for the English automotive W&G sector. Therefore, it is up to us to create a suitable annotated dataset, which is in the range of already annotated domain specific datasets [4, 7]. For performance evaluation, we will use the common metrics *recall*, *precision* and *F1-score*.

W&G Relation Extraction is a task to predict and classify the relationship between named entities in a corpus. Similar to the named entity recognition task, there are no suitable fine-tuning datasets available. Therefore, we will annotate on ourselves English automotive W&G datasets to fine-tune our language model in the specific context. Furthermore, we are planning to anonymize target entities in a sentence by using pre-defined tags like “@failure location#” or “@failure type#” [4]. For example, an annotated sentence with two target entities is represented by *“The customer claims that the @failure location# has @failure type#.”* Similar to the utilisation of the planned performance metrics for the NER task, we will use the same performance metrics for the RE task.

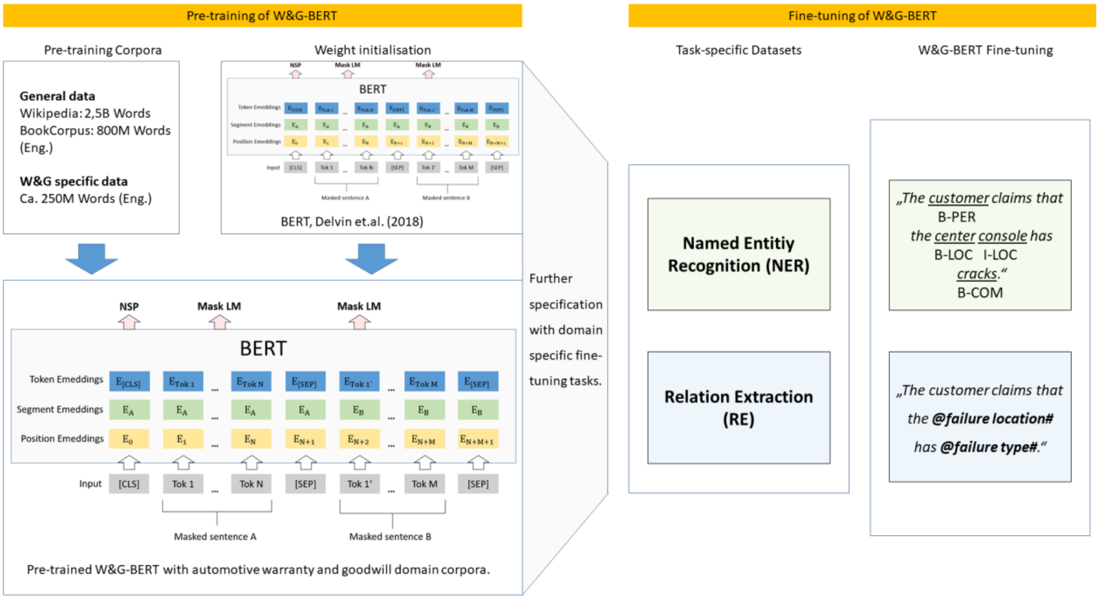

Figure 1. Pre-training and fine-tuning procedure of W&G-BERT

4. OUTLOOK

Based on the planned scientific contributions in the field of language models (W&G-BERT) and in agnostic sensitivity-based explanatory models (xLIME survey), we design a holistic approach to identify similar unstructured and semi-structured datasets in large domain specific corpora (see Figure 2). Therefore, we plan to map given data sets into vector representations using the context-dependent domain-specific language model W&G-BERT. After vectorization, the data will be presented in a high-dimensional numeric format. Due to this, the implementation of a dimension reduction algorithm may be necessary (this will be evaluated during the overall implementation). With the help of the context-based vector representations, semantic similarity values are to be calculated with the help of the cosine distance. Based on these, a state-of-the-art community detection algorithm can be selected and used to identify homogeneous network patterns in a heterogeneous data environment. After detecting communities in the network, we use the automatically labelled data set in means of the detected communities and the manually annotated clusters. Finally, a classifier will be trained with the help of the (un-) supervised clusters to recognise the identified clusters in future datasets. To ensure the acceptability with regulators and the general need for human understandable Blackbox predictions in sensitive areas, we plan to apply the best performing LIME-based explainability method to the domain-specific language model based classification model.

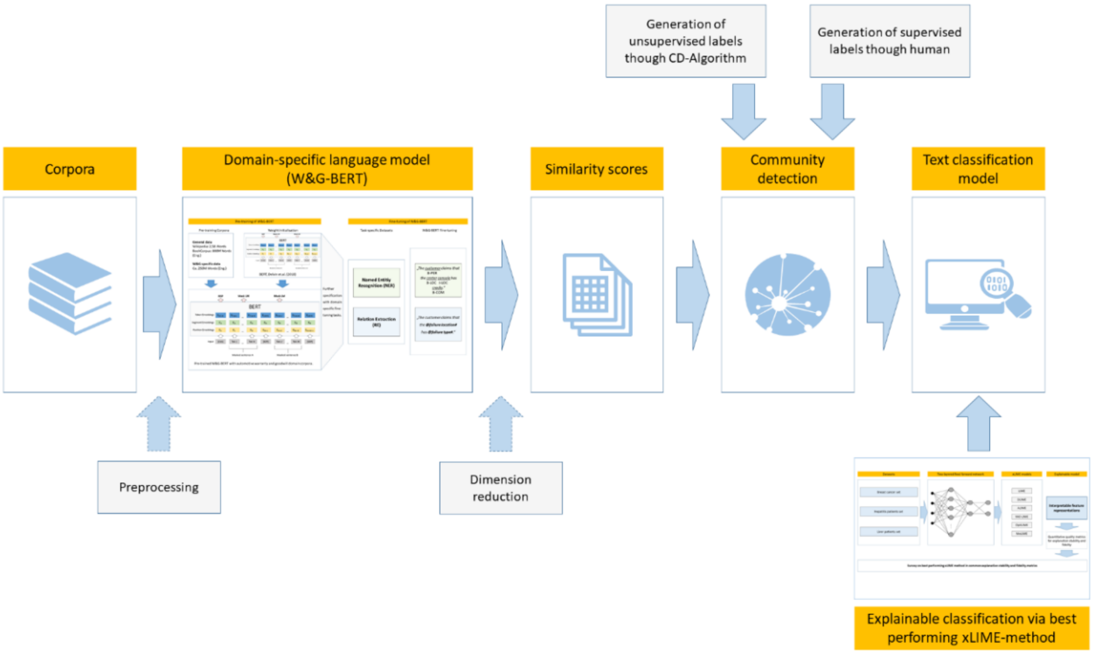
Generation of unsupervised labels though CD-Algorithm
Generation of supervised labels though human
Corpora
Domain-specific language model (W&G-BERT)
Similarity scores
Community detection
Text classification model
Preprocessing
Dimension reduction
Explainable classification via best performing xLIME-method

Figure 2. Overall approach with contributions in domain specific language model and performance based xLIME survey

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018. [Online]. Available: http://arxiv.org/pdf/1810.04805v2

[2] C. Felden, “Extraktion, Qualitätssicherung und Klassifikation unstrukturierter Daten,” HMD Praxis der Wirtschaftsinformatik, no. 247, pp. 54–62, 2006.

[3] M. Allahyari et al., “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” Jul. 2017. [Online]. Available: http://arxiv.org/pdf/1707.02919v2

[4] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics (Oxford, England), vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[5] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” 2019. [Online]. Available: https://arxiv.org/pdf/1904.03323.pdf

[6] Y. Gu et al., “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” 2021. [Online]. Available: https://arxiv.org/pdf/2007.15779.pdf

[7] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao, “FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining,” 2020. [Online]. Available: https://www.ijcai.org/proceedings/2020/0622.pdf

[8] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” Mar. 2019. [Online]. Available: http://arxiv.org/pdf/1903.10676v3

[9] R. Zhu, X. Tu, and J. X. Huang, “Utilizing BERT for biomedical and clinical text mining,” in Data Analytics in Biomedical Engineering and Healthcare: Elsevier, 2021, pp. 73–103. [Online]. Available: https://doi.org/10.1016/B978-0-12-819314-3.00005-7

[10] S. Silvestri, F. Gargiulo, and M. Ciampi, “Improving Biomedical Information Extraction with Word Embeddings Trained on Closed-Domain Corpora,” in 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, Jun. 2019 - Jul. 2019, pp. 1129–1134.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Oct. 2013. [Online]. Available: http://arxiv.org/pdf/1310.4546v1

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Jan. 2013. [Online]. Available: http://arxiv.org/pdf/1301.3781v3

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning, “Glove: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014. [Online]. Available: https://nlp.stanford.edu/pubs/glove.pdf

[14] A. Onan, “Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering,” IEEE Access, vol. 7, pp. 145614–145633, 2019, doi: 10.1109/ACCESS.2019.2945911.

[15] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

[16] Z. Huang, Ed., A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, 1997. [Online]. Available: http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=08889727BE3ABD82AEC0A6888F74EED4?doi=10.1.1.6.4718&rep=rep1&type=pdf

[17] D. Arthur and S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding: Arthur, David & Vassilvitskii, Sergei. (2007). K-Means++: The Advantages of Careful Seeding. Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms. 8. 1027-1035. 10.1145/1283383.1283494.,” no. 8, pp. 1027–1035, 2007. [Online]. Available: https://dl.acm.org/doi/10.5555/1283383.1283494

[18] T. Kohonen, “The self-organizing map,” no. 21, pp. 1–6, 1990. [Online]. Available: https://doi.org/10.1016/S0925-2312(98)00030-7

[19] H. Chipman and R. Tibshirani, “Hybrid hierarchical clustering with applications to microarray data,” Biostatistics (Oxford, England), vol. 7, no. 2, pp. 286–301, 2006, doi: 10.1093/biostatistics/kxj007.

[20] W. L. Taylor, ““Cloze Procedure”: A New Tool for Measuring Readability,” Journalism & Mass Communication Quarterly, no. 30, pp. 415–433, 1953, doi: 10.1177/107769905303000401.

[21] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019. [Online]. Available: http://arxiv.org/pdf/1907.11692v1

[22] M. E. Peters et al., “Deep contextualized word representations,” Feb. 2018. [Online]. Available: http://arxiv.org/pdf/1802.05365v2

[23] Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” Jun. 2019. [Online]. Available: http://arxiv.org/pdf/1906.05474v2

[24] A. Sergeev and M. Del Balso, “Horovod: fast and easy distributed deep learning in TensorFlow,” Feb. 2018. [Online]. Available: http://arxiv.org/pdf/1802.05799v3

[25] P. Micikevicius et al., “Mixed Precision Training,” Oct. 2017. [Online]. Available: http://arxiv.org/pdf/1710.03740v3

[26] Y. Sun et al., “ERNIE: Enhanced Representation through Knowledge Integration,” 2019a. [Online]. Available: http://arxiv.org/pdf/1904.09223v1

[27] Y. Sun et al., “ERNIE 2.0: A Continual Pre-training Framework for Language Understanding,” 2019b. [Online]. Available: https://doi.org/10.1609/aaai.v34i05.6428

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le V, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Jun. 2019. [Online]. Available: http://arxiv.org/pdf/1906.08237v2

[29] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu, and T. Sun, “Application of Pre-training Models in Named Entity Recognition,” Feb. 2020. [Online]. Available: http://arxiv.org/pdf/2002.08902v1

[30] G. Lample and A. Conneau, “Cross-lingual Language Model Pretraining,” 2019. [Online]. Available: https://arxiv.org/abs/1901.07291

[31] A. Vaswani et al., “Attention is All you Need,” 2017. [Online]. Available: https://arxiv.org/abs/1706.