Explainable NLP for Monitoring the German Automotive Industry

Andrea Zielinski (dr.andrea.zielinski@gmail.com), Fraunhofer
Henning Kroll (Henning.Kroll@isi.fraunhofer.de), Fraunhofer ISI
Anna Grimm (anna.grimm@isi.fraunhofer.de), Fraunhofer ISI

Introduction

The latest developments and trends in the automotive industry, including topics such as the advance of electric cars, are of public interest and widely discussed in the news. Several online news providers have established a platform for sharing such information, and web scraping provides free access in nearly real-real time.

An area with great interest is classifying large volumes of such news articles to study how the press perceives the automotive sector from a national perspective, i.e. how sentiment relates to survey studies like VDA or if it can serve as a relevant innovation indicator.

However, sentiment analysis on this data is challenging, because even if the author of an article explicitly states his/her opinion, this might not be unique across the text, but refer to some specific aspect only. Moreover, it might also involve interpretation or the use of world knowledge (Balhahur et al., 2009). In addition, to avoid getting irrelevant search results, a proper categorization of news articles into thematic categories is required.

For our application, both *accuracy* and *interpretability* are a critical concern. Generally, there is a trade-off between the benefits of interpretability associated with lexicons and the higher levels of accuracy associated with machine learning. Our goal in this study is to compare SOTA models for our task w.r.t their performance and interpretability.

Research Questions are:

- How well can we automatically classify a document as expressing a *positive, negative or neutral* opinion at the document-level
- How well can we automatically classify a document by one of the labels *Diesel, Autonomous Driving, Robotaxi. Hydrogen, Synthetic Fuel, and other.*
- Which model delivers interpretable results, highlighting words or phrases that have lead to the class, and check, if this is also reflected by human rationales.

Dataset

We compile a dataset, focusing on websites of national automotive channels, a valuable source for opinionated news. The data collection currently comprises 5K German news articles, based on search queries related to the keywords *e-mobility, autonomous driving, hydrogen, robotaxi and Diesel* along with synonyms and German translations. The queries resulted in a high recall but low precision, so that the thematic category needs re-assessment.

From this data, we create a benchmark and so far annotated 350 German news articles manually w.r.t. positive/negative sentiment words as well as the overall document-level sentiment, and assessed inter-annotator agreement (IAA) using the *tagtog* platform. We achieved an IAA of 80.34% for overall sentiment classification and 88.98% for text classification.

Methodology

For the tasks *sentiment analysis* and *news categorization* we identify three conceptually distinct groups of methods: (1) lexicons, (2) traditional machine learning, and (3) transfer learning. The use of contextualized representations of words, as exemplified by the transformer model BERT (Devlin et al. 2018; Guhr et al., 2020) have advanced the SOTA performance across many tasks, including sentiment analysis (Poria et al., 2020), and news categorization (Ha, 2021) because they can capture semantic variability. In our experiments, we compare BERT, CNN, and LSTM and traditional ML approaches like SVM, and rule-based system. Moreover, we experiment with different XAI algorithms implemented in captum (Kokhlikyan et al., 2020), that seek to explain model predictions.

Preliminary Results

Document-level Sentiment Analysis: We compare different SOTA classifiers and evaluate the models on our annotated data, split into train, test and development. For example, the ML method SVM$_{\text{linear}}$ reaches an accuracy of 53% (see Fig. 1), outperforming GerVADER (Tymann et al., 2019) with an accuracy of 50%.

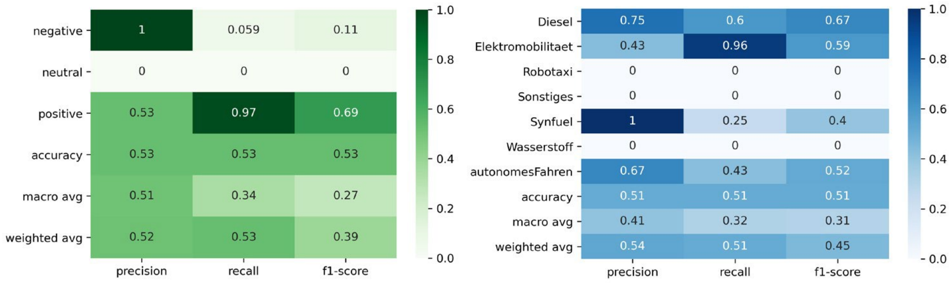
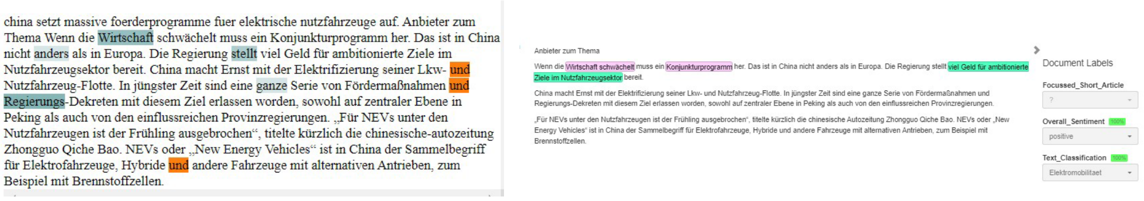


Figure 1 The evaluation results of document-level sentiment classification (left) and news categorization (right) for SVM. Since the datasets are imbalanced, we also present macro-averaged precision, recall, and F1 measure. (Note: performance results are only preliminary, since work on the benchmark corpus is still in progress.)

Explainable NLP: Deep learning models are black-box models with little explainability. Therefore, we apply various interpretation techniques to our prediction models that seek to explain their analysis results post-hoc, such as LIME and Integrated Gradients (ITO et al., 2021). To this aim, we first created a benchmark of human annotations, where salient features are highlighted within the news article. An example for a single news article of the corpus is shown below (see Fig.2, right). Then, we check for agreement between human annotations and the model's extracted rationales (see Fig2, left).



china setzt massive foerderprogramme fuer elektrische nutzfahrzeuge auf. Anbieter zum Thema Wenn die Wirtschaft schwächelt muss ein Konjunkturprogramm her. Das ist in China nicht anders als in Europa. Die Regierung stellt viel Geld für ambitionierte Ziele im Nutzfahrzeugsektor bereit. China macht Ernst mit der Elektrifizierung seiner Lkw- und Nutzfahrzeug-Flotte. In jüngster Zeit sind eine ganze Serie von Fördermaßnahmen und Regierungs-Dekreten mit diesem Ziel erlassen worden, sowohl auf zentraler Ebene in Peking als auch von den einflussreichen Provinzregierungen. „Für NEVs unter den Nutzfahrzeugen ist der Frühling ausgebrochen", titelte kürzlich die chinesische-autozeitung Zhongguo Qiche Bao. NEVs oder „New Energy Vehicles" ist in China der Sammelbegriff für Elektrofahrzeuge, Hybride und andere Fahrzeuge mit alternativen Antrieben, zum Beispiel mit Brennstoffzellen.

Anbieter zum Thema

Wenn die Wirtschaft schwächelt muss ein Konjunkturprogramm her. Das ist in China nicht anders als in Europa. Die Regierung stellt viel Geld für ambitionierte Ziele im Nutzfahrzeugsektor bereit.

China macht Ernst mit der Elektrifizierung seiner Lkw- und Nutzfahrzeug-Flotte. In jüngster Zeit sind eine ganze Serie von Fördermaßnahmen und Regierungs-Dekreten mit diesem Ziel erlassen worden, sowohl auf zentraler Ebene in Peking als auch von den einflussreichen Provinzregierungen.

„Für NEVs unter den Nutzfahrzeugen ist der Frühling ausgebrochen", titelte kürzlich die chinesische Autozeitung Zhongguo Qiche Bao. NEVs oder „New Energy Vehicles" ist in China der Sammelbegriff für Elektrofahrzeuge, Hybride und andere Fahrzeuge mit alternativen Antrieben, zum Beispiel mit Brennstoffzellen.

Document Labels

Focussed_Short_Article: ?

Overall_Sentiment 100%: positive

Text_Classification 100%: Elektromobilitaet

Figure 2 shows the content of the news article entitled ' *China announces massive promotion programs for electric commercial vehicles*' on **e-mobility** with overall **positive sentiment**. The human rationale is annotated in *tagtog* on relevant entities on the right (blue=positive, red= negative). This is compared to the output of the machine learning model (here **Random Forest**) on the left.

Conclusion and Outlook

In this work, we report on our methodology for creating labeled German news data, as well as applying sentiment analysis, news categorization and explainable NLP interpretation techniques. Our main contributions focus on a discussion on the differences between SOTA sentiment detection and news categorization models for our German benchmark data w.r.t. the criteria performance and interpretability. We believe that this will help select and improve a model and/or its explanations in the long run.

References

Balahur, A., & Steinberger, R. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA*, *9*, 1-12.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis.

Devlin J, Chang M, Lee K, Toutanova K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv

Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques.

Guhr, Oliver, et al. "Training a broad-coverage German sentiment classification model for dialog systems." *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020.

Ha, S., Marchetto, D. J., Dharur, S., & Asensio, O. I. (2021). Topic classification of electric vehicle consumer experiences with transformer-based deep learning. *Patterns*, *2*(2), 100195.

Ito, Tonnoki (2021) Word-Level Contextual Sentiment Analysis with Interpretability. AAAI.

Kokhlikyan, Narine, et al. "Captum: A unified and generic model interpretability library for pytorch." *arXiv preprint arXiv:2009.07896* (2020).

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., & Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proc. of the 2020 Conf. on EMNLP: System Demonstrations* (pp. 107-118).

Tymann, K.. Lutz, M., Palsbroker, P., Gips.C. (2019). Gervader - a German adaptation of the vader sentiment analysis tool for social media texts. In LWDA, pages 178–189

VDA https://home.kpmg/de/de/home/themen/2021/11/global-automotive-executive-survey-2021.html