Disinformation Detection: An Explainable Transfer Learning Approach

Mina Schütz[1,2], Alexander Schindler[2], and Melanie Siegel[1]

[1] Darmstadt University for Applied Sciences, 64295 Darmstadt, Germany
melanie.siegel@h-da.de
http://www.h-da.de

[2] Austrian Institute of Technology GmbH, 1210 Vienna, Austria
{mina.schuetz, alexander.schindler}@ait.ac.at
http://www.ait.ac.at

Abstract. Facilitated by new technologies, disinformation can be spread much more easily. Recent reports of targeted disinformation campaigns show the impact it has on democracy, journalism, and opinion mining. In this work, we propose a novel approach to gain new insights into the automatic detection of fake news with a focus on German and English content. Claims, narratives, and entities are extracted from news articles and social media content using transfer learning approaches from natural language processing and Transformer models. This extracted information and its relationships among them can be used to improve classification models through knowledge infusion during training (interpretability) and for final prediction results (explainability). To ensure high understandability for our given approach, we will implement a final dashboard using visual analytics with respect to different types of end-users. In addition, we will propose a model that generalizes and uses other emerging concepts that are considered as indicators of fake news, such as hate speech, propaganda and conspiracy theories, as well as emotions, sentiment, aggression and toxicity. These should help identify potentially misleading content, as well as provide relevant information to derive and collect intelligence from disinformation campaigns that could provide relevant basis for possible countermeasures.

Keywords: disinformation · fake news · transfer learning · multilingual · natural language processing · explainable artificial intelligence

1 Introduction

An increased news consumption and fast dissemination of content on social media has lead to an information overload in the web, which makes it complicated for readers to distinguish between deceptive and real news articles [12]. The term *fake news*, which is commonly used since the US presidential election in 2016[11] is usually defined by the intention (unintentional: misinformation; intentional: disinformation) behind an article and its factuality [17]. It is even said that it developed as a threat "[...] to democracy, journalism, and freedom of expression."

[22]. *Fake news* includes misleading biased and persuading propaganda, fabrications, truth hiding hoaxes, click-bait content, not verified rumors and even satire [11,8,17]. In this work we focus on the aspect of intentionally spread content that contains misleading information to detect disinformation campaigns in an early manner, instead of focusing on end-users, who unintentionally spread, for example, satire articles as a fact.

Furthermore, we define the task of fake news detection as a system that helps users to identify misleading information with the usage of machine learning techniques automatically - detecting and explaining to users why this content is considered misleading. The counterpart - manual fact-checking - is expensive and slow, through the extended use of experts and immense amount of online content [22]. Automatic detection with natural language processing (NLP) methods is complex and does not yet yield satisfactory results. Linguistically based approaches are easy to interpret and offer deeper insights into themes and narratives, but suffer from ambiguous and colloquial language, different writing styles, semantics, and abbreviations that are also inconsistent across themes and time. Recent approaches based on deep neural networks, such as Transformers [19], show high detection rates, but are known for their "black box" nature. Derived results therefore contribute little to an overall insight into the problem domain. Additionally, the concept of fake news is highly intertwined with other emerging phenomena on social media and online content, such as hate speech, which mainly includes racism and sexism [7].

Even though there is already an extensive research on the fake news phenomena, the mentioned issues and combination of concepts have not been included in current research enough. Therefore, we propose a novel approach to detect disinformation with focus on multi-lingual transfer learning, with German and English as the main languages. Our approach will be extended through an extensive research on breaking the black-box apart with knowledge infusion through graphs and visualization techniques.

2 Related Work

Disinformation can be detected through different machine learning methods. The most common one is the content-based approach with NLP: news articles are examined regarding linguistic features that can be directly extracted out of the body text or title. This can be on character, word, sentence or document level [12] with focus on the writing style, syntactic and semantic features, such as sentiment [21]. Social-context-based approaches on the other hand use user engagements in social networks. This can either be done by examining the propagation of news in a social network or by analyzing the comments on posts and articles. The latter can be enhanced by detecting the stance regarding a post or article in the network, which can give clues about its validation or falsification [12,5]. In contrast, automated fact-checking helps to construct knowledge-bases [22]. Furthermore, many studies not only focus on solely classification approaches but rather use a hybrid combination of multiple approaches [4]. Especially in fake news detection, the newly invented Transformer models have been applied to many NLP tasks: claim verification, classification, evidence retrieval and stance detection [9,18,1,3].

Research on visualization tools includes exBERT by Hoover et al. [6]. The authors used visualization techniques to analyze the learned representations and attention mechanisms by the Transformer model. Verify2 by Karduni et al. [10] is more focused on the linguistic analysis of misinformation in Tweets, such as moral foundations, subjectivity, bias, emotion and sentiment. The authors created a visual interface with multiple views for users to explore also topics, keywords, named entities and relationships between Twitter accounts. Similarly Yang et al. [20] built an application with decision trees and other visualizations for analyzing the probability that an article is fake. The authors focused mainly on the explanation of key components, predictions of words and phrases, and linguistic features.

3 Main Contributions & Research Questions

In this work we propose a novel approach in the field of automatic detection of disinformation. The goal is to explore disinformation with natural language processing, data visualization techniques, and the use of explainable artificial intelligence (XAI) to 1) improve automatic disinformation detection and 2) provide tools to gain new knowledge about disinformation itself, such as common or trending narratives. Therefore, the following main research question and sub-questions are proposed:

To which extend can we obtain new linguistic knowledge about disinformation with NLP methods and XAI techniques?

1. To which extent can we extract statements and narratives from corresponding content and put them in a common context?
2. To which extend can we classify the content and statements of the textual data with current state-of-the-art models in NLP?
3. How can we model and evaluate complex and abstract semantic concepts such as credibility, truthfulness or factual content without ground truth on disinformation?
4. To which extend are methods of XAI and the use of visualization techniques helpful to gain valuable insights on the decisions by models and to present them in a user-friendly way?

The contributions of this work are two-fold: firstly, we propose a novel approach - that is not solely focused on a simple downstream prediction task as a solution to the detect disinformation - but uses data science techniques and NLP to create new knowledge about disinformation based on the underlying data for the German and English language. Secondly, our approach will use XAI methods with comprehensible visualizations, automatic adaption methods and knowledge infusion.

4 Methodology

In our early work exploring the detection of fake news, we have seen that a stand-alone linguistic binary classification approach for a specific domain does not give us enough valuable information to tackle the problem of disinformation detection, even though Transformer models gain an overall high prediction accuracy [16,14]. However, using similar data for a given task, such as in our work for the detection of sexism [15], can help improve the prediction of classification models. Therefore, we propose using transfer learning for multiple NLP tasks, such as claim extraction and verification [18], entity and relationship extraction [2], and event detection [13]. Those should help understand the narrative of a given news article and can be used for the creation of a knowledge graph. Those graphs can be used in XAI to infuse additional information during the training of a model (interpretability) and plays an important role in understanding the prediction and analysis results of machine learning models (explainability). To further gain more knowledge - besides a thorough linguistic analysis - about the influence of hate speech and propaganda on fake news articles, we also propose to evaluate the content based on high-level features such as the degree of emotion, aggression, toxicity, moral foundations and sentiment.

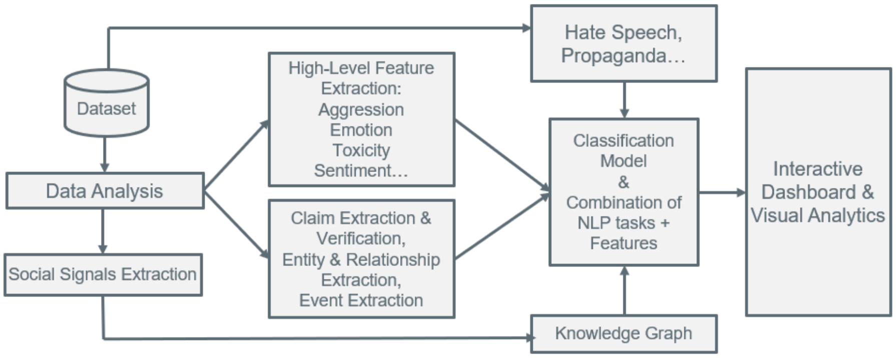
Dataset
Hate Speech, Propaganda…
High-Level Feature Extraction: Aggression Emotion Toxicity Sentiment…
Data Analysis
Classification Model & Combination of NLP tasks + Features
Interactive Dashboard & Visual Analytics
Claim Extraction & Verification, Entity & Relationship Extraction, Event Extraction
Social Signals Extraction
Knowledge Graph

Fig. 1. Methodology overview.

To conduct our research, we will contribute a novel annotated benchmark dataset for disinformation detection in the German language. Pre-training language models can be conducted in an unsupervised manner with additional data from social networks or websites in German or English. Existing benchmark datasets for the different disinformation concepts can be used and evaluated. An interactive dashboard will show the results of the proposed approach with visualizations, to help experts and end-users to not only understand why an article contains potentially critical content, but also to comprehend the decision of the model.

5 Conclusions

In this work we gave an overview of disinformation, related work and already published tools in the research area of NLP. Our proposed approach combines transfer learning strategies, multilingual language understanding with state-of-the-art Transformer models, generalizability over multiple domains and the usage of visualization techniques and knowledge infusion to enhance our approach with interpretability and explainability for end-users as well as experts.

6 Acknowledgments

This work is enhanced by the Darmstadt University of Applied Sciences (h_da) in collaboration with the Austrian Institute of Technology GmbH (AIT). The Darmstadt University supported this work with the research in Information Science (https://sis.h-da.de/) and the Research Group on Human-Computer Interaction and Visual Analytics (https://vis.h-da.de/). Additionally, this contribution has been funded by the FFG Project "Defalsif-AI" (Austrian security research program KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT), grant no. 879670) - as well as the project "DeTox" (Cybersecurity research funding of the Hessian Ministry of the Interior and Sports).

References

1. Antoun, W., Baly, F., Achour, R., Hussein, A., Hajj, H.: State of the art models for fake news detection tasks. In: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). pp. 519–524 (2020)
2. Braşoveanu, A.M.P., Andonie, R.: Semantic fake news detection: A machine learning perspective. In: Rojas, I., Joya, G., Catala, A. (eds.) Advances in Computational Intelligence. pp. 656–667. Springer International Publishing, Cham (2019)
3. Cruz, J.C.B., Tan, J.A., Cheng, C.: Localization of fake news detection via multitask transfer learning. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 2596–2604. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.316
4. Della Vedova, M.L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., de Alfaro, L.: Automatic online fake news detection combining content and social signals. In: 2018 22nd Conference of Open Innovations Association (FRUCT). pp. 272–279 (2018)
5. Graves, L.: Understanding the promise and limits of automated fact-checking (2018)
6. Hoover, B., Strobelt, H., Gehrmann, S.: exBERT: A visual analysis tool to explore learned representations in Transformer models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 187–196. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-demos.22

7. Istaiteh, O., Al-Omoush, R., Tedmori, S.: Racist and sexist hate speech detection: Literature review. In: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA). pp. 95–99 (2020)
8. Jr., E.C.T., Lim, Z.W., Ling, R.: Defining "fake news": A typology of scholarly definitions. Digital Journalism **6**(2), 137–153 (2018)
9. Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H.: exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences **9**, 4062 (2019)
10. Karduni, A., Cho, I., Wesslen, R., Santhanam, S., Volkova, S., Arendt, D., Shaikh, S., Dou, W.: Vulnerable to misinformation? verifi! (2018)
11. Khan, S.A., Alkawaz, M.H., Zangana, H.M.: The use and abuse of social media for spreading fake news. In: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). pp. 145–148 (2019)
12. Mahid, Z.I., Manickam, S., Karuppayah, S.: Fake news on social media: Brief review on detection techniques. In: 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA). pp. 1–5 (2018)
13. Nguyen, D.T., Jung, J.E.: Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems **66**, 137–145 (2017)
14. Schütz, M.: Detection and identification of fake news: Binary content classification with pre-trained language models. In: Information between Data and Knowledge, Schriften zur Informationswissenschaft, vol. 74, pp. 422–431. Werner Hülsbusch, Glückstadt (2021), gerhard Lustig Award Papers
15. Schütz, M., Boeck, J., Daria, L., Slijepčević, D., Kirchknopf, A., Hecht, M., Bogensperger, J., Schlarb, S., Schindler, A., Zeppelzauer, M.: Automatic sexism detection with multilingual transformer models (2021)
16. Schütz, M., Schindler, A., Siegel, M., Nazemi, K.: Automatic fake news detection with pre-trained transformer models. In: Bimbo, D., et al (eds.) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Sciences. vol. 12667. Springer, Cham (2021)
17. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol. **10**(3) (Apr 2019)
18. Soleimani, A., Monz, C., Worring, M.: Bert for evidence retrieval and claim verification. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 359–366. Springer International Publishing, Cham (2020)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
20. Yang, F., Pentyala, S., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E., Ji, S., Hu, X.: Xfake: Explainable fake news detector with visualizations. In: International Word Wide Web Conference Committe (IW3C2). pp. 3600–3604 (05 2019)
21. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: A theory-driven model. Digital Threats: Research and Practice **1**(2) (Jun 2020)
22. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv. **0**(ja) (2018)