2024

Counteracting Attacks on Science with Social Sentiment Analysis: A Comparison of Approaches for Custom Social Sentiment Analysis Tool

Till Schirrmeister
University Potsdam, Germany, till.schirrmeister@uni-potsdam.de

Lina Goerlich
University Potsdam, Germany, lina.goerlich@uni-potsdam.de

Counteracting Attacks on Science with Social Sentiment Analysis: A Comparison of Approaches for Custom Social Sentiment Analysis Tool

Research Short Paper

Till Schirrmeister[1], Lina Goerlich[1]

[1] University Potsdam, Chair of Business Information Systems and Digital Transformation, Potsdam, Germany

Abstract. Democracy-harming forces in online social networks (OSNs) attack the credibility of scientists aiming to hinder the spread of scientific knowledge. Current sentiment analysis tools are to a large extent inadequate for effectively monitoring attacks on scientists, highlighting the need for custom tools. Our study addresses this by exploring the best techniques for a custom sentiment analysis tool. We manually coded a dataset of tweets appreciating or criticizing scientists during the COVID-19 pandemic and evaluated various supervised machine learning algorithms, ensemble techniques, and zero-shot classification methods. Our findings indicate that stacking is the most effective method for training a custom sentiment analysis tool, while zero-shot classification is unsuitable. These results provide insights for researchers and practitioners to improve their monitoring tools, encouraging scientists to share their knowledge.

Keywords: *Sentiment Analysis, Digital Democracy, Online Social Networks*

1 Introduction

Misinformation has been a tool for propaganda and to advance political agendas for a long time. With the widespread use of the internet, online social networks (OSNs) have become a breeding ground for misinformation (Scheufele & Krause, 2019; Wang et al., 2022). Forces hostile to democracy (e.g., right-wing extremists, climate change deniers, conspiracy theorists) use misleading information to put pressure on democracies (Weinhardt et al., 2024). They influence elections (Cantarella et al., 2023) and undermine our democratic institutions (Jakubik et al., 2023). These hostile forces not only spread misinformation on OSNs, but also attack those who share scientific knowledge. With the intention of undermining the legitimacy of experts, hostile forces to democracy discredit scientists (Egelhofer, 2023). A recent example is COVID-19. During the pandemic, scientists who spoke out in the media were often targeted by online trolls, aiming to intimidate them into silence (Nogrady, 2021; Blümel, 2024). According to a survey conducted by the journal Nature, 58% of scientists who commented on COVID-19 faced attacks on their credibility, and 15% reported receiving death threats (Nogrady, 2021). Since democracies rely on a shared body of knowledge (Lewandowsky et al., 2023), the silencing of scientists in public discourse through attack on OSNs has epistemic consequences that are harmful to democracies e.g., (Dan et al. 2021; Au et al., 2022).

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

To monitor social media attacks on science and ultimately contributing to counteract democracy harming forces on OSN there is a need for Information Systems research that contributes to the resilience of democracies (Weinhardt et. al. 2024). A common approach to this is social sentiment analysis. Social sentiment analysis uses large amounts of OSN data to "provide insights into the factors that contribute to the resilience of digital democracies" (Weinhardt et. al. 2024, p. 6).

So far traditional sentiment tools are used to classify speech on sentiments like positive or negative or to measure the expression of emotions e.g., (Wankhade et al. 2022; Yue et al. 2019). Therefore, machine learning (ML) techniques are mainly used. Existing sentiment analysis tools are not designed to help democracy by detecting attacks on scientists from online social network (OSN) data. During the COVID-19 pandemic, scientists communicating online faced both hostility and praise. To track this criticism and appreciation, there is a need for custom sentiment analysis tools that can detect these sentiments. In this study, we examine different supervised ML methods, ensemble learning and zero shot classification to create custom tools for analyzing social sentiments. We train and test them on a manually coded set of OSN data that classifies appreciation and criticism of researchers during the corona virus crisis. In doing so, we aim to answer the following research question:

RQ: What approaches are suited best to recognize custom sentiments in the given context?

To answer the research question, we collected and manually coded a custom dataset of 1500 Tweets that includes appreciation and criticism of scientists. We then trained and tested seven supervised ML algorithms, two ensemble techniques and a zero shot classification model. Our results indicate that stacking performs best in training a custom social sentiment analysis tool. Zero shot classification on the other hand is not sufficient to be considered for the task. Our results can guide Information Systems researchers who are developing social sentiment analysis tools to strengthen democracy. Our findings have implications for both researchers and OSN operators.

2 Related Work

2.1 Sentiment analysis

Sentiment analysis is the process of using computers to determine and classify opinions or emotions expressed in text (Vashishtha & Susan, 2019). As a field sentiment analysis combines ML, data mining, natural language processing, and computational linguistics (Yue et. al. 2019). To extract sentiment from a text, there are two common approaches that can also be combined: the lexicon-based approach and the ML approach (Wankhade et al., 2022). The lexicon-based approach uses lists of words, where each word is assigned to a sentiment. The ML approach identifies sentiments in text using a prediction model. To train such a prediction model, various ML algorithms are used. The most common ones for this task are Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees, Discriminant Analysis and Neural Networks e.g., (Wankhade et al., 2022; Tan et al., 2023).

In addition to using a single model for sentiment classification, multiple classification models can be also combined called ensemble learning (Wolpert, 1992).

Common techniques include ensemble learning with voting and stacking. Ensemble learning with voting is a method in ML that trains multiple models to complete the same task. A form of "voting" is then used to derive a final prediction from each model's predictions (Wolpert, 1992). Stacking is an advanced ensemble technique in ML where predictions from multiple base models are combined by a meta-model to make the final prediction (Wolpert, 1992).

A more recent approach to classify sentiment is zero-shot classification. Zero-shot classification is a machine learning technique where a model classifies data into unseen categories (e.g., emotions, topics, and events) by using a large pre-trained model (Ye et al., 2020; Yin et al., 2019). This method is useful for sentiment analysis because it allows a model to assign any predefined label to text without specific training for each label. Training includes pre-training and fine-tuning (Howard & Ruder, 2018). In pre-training, the model learns language patterns from various sources such as news articles, books, and blogs, using models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018). Fine-tuning then refines this knowledge on annotated data for specific tasks, enhancing its ability in sentiment analysis.

2.2 Social sentiment analysis

Social contexts play a crucial role in sentiment analysis. Therefore they are integrated in various social sciences including political science, economics, and sociology (Hemmatian and Sohrabi, 2019). For example, politics is the second most popular domain, after marketing for sentiment analysis (Rodríguez-Ibanez et al., 2023) due to the variety of emotional expressions found in political communication (Park et al. 2023). Studies in the domain are mainly conducted in relation to elections, political parties, and specific processes in certain countries (Rodríguez-Ibanez et al., 2023). Main focus on the research includes how language and tone change over time, how communication can appeal to citizens, and how sentiments are reflected in political orientations (Pipal et al., 2024). So far, the analysis focuses on emotional expressions, including feelings like sadness, fear, happiness, as well as positive and negative remarks (Park et al. 2023). Existing sentiment analysis tools are limited in their ability to capture the complex process of public opinion formation. Nonetheless, they hold promise in mitigating the challenges presented by the increasing polarization and radicalization within societies (Weinhardt et al., 2024).

3 Method

3.1 Data collection

To collect the data for our analysis, we used a three-step process. In the first step, we identified active scientists on social media, specifically the platform X (formerly known as Twitter). We reviewed the websites of each virology research institute in Germany. Our inclusion criteria, based on Lavazza and Farina (2020), required that the person had a Ph.D., was affiliated with an official research institution, and had publications in peer-reviewed journals. This resulted in 311 scientists. We then eliminated those who lacked prominence on social media. In particular, we excluded

individuals with less than a predefined threshold of 10,000 followers on X as of August 2021, reducing our list to 7 scientists. By reviewing the accounts followed by these 7 scientists, we identified and added an additional virologist who met all of our previously defined criteria. Of the eight identified scientists, three were female and five were male. In the second step, we collected 583,173 posts related to the scientists identified in the first step using the X API. The posts were collected during six time periods: First Lockdown (20/03/16 to 20/03/29), Second Lockdown (20/10/28 to 20/11/11), Tightening Lockdown (20/12/16 to 21/01/03), Federal "Emergency Brake" (21/03/03 to 21/03/16), Draft Legislation of the new Government (21/11/11 to 21/11/25), and Tightening Measures (21/12/28 to 22/01/25). These periods were chosen to cover different stages of the pandemic, each representing a significant event. In the final step, we randomly selected and annotated 1,500 replies (250 from each period) from the scientists' posts. Criticism was coded if the posts contained negative comments about people, institutions, or policies, including, but not limited to, criticism of scientists' statements, criticism of their expertise, and criticism of their motivation. Appreciation was coded when the posts expressed positive feelings about the scientists or their work. Since a post can be both appreciative and criticizing, it can be assigned to both, one, or neither of these categories. Two researchers manually coded the posts based on a predefined codebook. The coding evaluation showed an intercoder reliability of 0.83 (appreciation) and 0.80 (criticism) based on Krippendorff's Alpha (Krippendorff, 2011) indicating a nearly perfect and substantial agreement (Landis & Koch, 1977). If there was a misalignment in the coding, a third person was consulted, and maturity coding was added to the final dataset. Finally, the tweets were vectorized using TF-IDF in order to train the machine learning model.

3.2 Data evaluation

The manually collected dataset is used to answer the research question. We divided the dataset into 60% training data and 40% testing data. We use the training dataset to train the ML models and the testing dataset to evaluate and compare the models. We evaluated seven ML approaches to find the most suitable one: Multi-layer Perceptron, Support Vector Classification, Logistic Regression, Perceptron, Regression Trees in the form of Gradient Boosting for Classification, Gaussian Naive Bayes and Quadratic Discriminant Analysis. We chose these ML models to cover a range of methods, from simple (e.g., Logistic Regression, Linear Discriminant Analysis) to complex (e.g., Multi-layer Perceptrons, Gradient Boosting). We included discriminative models, which map inputs to outputs directly (e.g., Logistic Regression), and generative models, which model class probabilities (e.g., Gaussian Naive Bayes) (Ng et al., 2001).

After implementing the models individually, we combined them using ensemble learning. For each ensemble technique, we included in the comparison, the combination of models that performed the best, after testing every model combination from the models named above. For the ensemble learning with voting, we combined a Multi-layer Perceptron and Logistic Regression to predict criticism, and used Regression Trees with Gradient Boosting and Logistic Regression to predict appreciation. For stacking, we used Multi-layer Perceptron and Logistic Regression as base models, and Support Vector Classification as the meta-model to predict criticism.

To predict appreciation, we used regression trees with gradient boosting and logistic regression as base models, combined with Perceptron as a meta-model.

We implemented the models using Python and the scikit-learn library (scikit-learn, 2024). The default hyperparameters recommended by scikit-learn were used with a few exceptions. For the Multi-layer Perceptron, we set the solver to 'lbfgs', alpha to 1e-5, hidden_layer_sizes to (5, 2), and random_state to 1. For Support Vector Classification, we set gamma to 'auto'. For Logistic Regression, we set multi_class to 'multinomial', solver to 'lbfgs', max_iter to 1800, warm_start to True, and C to 100. For the Perceptron, we adjusted tol to 1e-3 and random_state to 0. For Gradient Boosting Classification, we set learning_rate to 1.0, max_depth to 1, and random_state to 0. These adjustments were made to prevent underfitting, better detect complex patterns, ensure reproducibility, and avoid overfitting.

Additionally, we used a pre-trained zero-shot classification model implemented with the transformers library in a custom Python script (Laurer et al., 2023). This model was selected for its training on German text and its design for political speech on social media. We selected an open-source model because it allowed us to run it on a local machine, ensuring data privacy. Additionally, this choice makes our results reproducible, as other researchers can also access and use the same model. To classify appreciation, we used the labels “Würdigung” and “keine Würdigung” (German for appreciation and no appreciation). For classifying criticism, we used “Kritik” and “keine Kritik” (German for criticism and no criticism). We classified all the collected data and compared the model's performance with our manual codes.

4 Results

In our results we compare the performance of each model by its accuracy for both categories. In summary, stacking showed the highest overall performance, while QDA and the pre-trained zero-shot model were the least effective for this classification task. When looking at the individual learning models, the Multi-layer Perceptron was the most accurate in classifying criticism and the Regression Trees in the form of Gradient Boosting for Classification was the most accurate in appreciation. The following table summarizes the results.

Table 1. Accuracy by classification technique

Learning model	Accuracy criticism in %	Accuracy appreciation in %
Multi-layer Perceptron	86.7	85.7
Support Vector Classification	86.3	85.7
Logistic Regression	86.2	91.8
Perceptron	83.3	91.8
Regression Trees with Gradient Boosting	79.3	92.2
Gaussian Naive Bayes	72.8	82.7
Quadratic Discriminant Analysis	21.3	21.0
Ensemble learning with voting	85.0	92.2
Stacking	88.0	93.3
Pre-trained zero shot classification model	30.5	35.3

5 Discussion

By promoting appreciation for scientists and minimizing criticism, we can foster a climate where scientists are more likely to express their views, which helps to reduce misinformation. Our research evaluates different machine learning models to create a sentiment analysis tool for classifying appreciation and criticism of scientists.

Our results show that most ML models better predict appreciation than criticism, except for Quadratic Discriminant Analysis (with poor overall performance) and Support Vector Machine (better at predicting criticism). This is likely due to the diverse nature of criticism, which can target the empirical basis of a claim, the scientist, or both, making it harder to identify (Barnes et al., 2018). Overall we suggest using stacking in order to build a custom social sentiment analysis tool.

Our findings have implications for both researchers and practitioners. Social organizations can use the insights to implement better protection measures for scientists. Operators of OSNs can develop more effective strategies to combat the spread of false information and promote scientific content. In addition, researchers can use these results to develop novel social sentiment analysis tools that enhance the ability to assess the formation of public opinion and the impact of scientific communication. This multifaceted approach can ultimately contribute to a more informed and supportive online environment for science.

As with any research, this study has limitations that must be considered. The accuracy of the models discussed may change when applied to different datasets for classification tasks. The quality and diversity of the data used to train the models can greatly affect their accuracy and generalizability. Moreover, there is an imbalance in the dataset towards "no criticism" and "no appreciation," which affects the models' accuracy and introduces bias. Future research should aim to develop tools that provide more generalized sentiment evaluations, not just limited to identifying criticism and appreciation of scientists. Moving forward, we will conclude this research by including more platforms (such as TikTok and YouTube) and different types of crises (such as climate crisis, economic crisis, and ongoing wars) in the training set aiming to provide a more general social sentiment analysis tool for criticism and appreciation of scientists.

6 Conclusion

In conclusion, our study demonstrates the efficacy of various ML methods in developing social sentiment analysis tools to support democratic institutions by protecting scientists and scientific knowledge on OSNs. Among the methods tested, stacking turns out to be the most suitable for social sentiment analysis. Our findings can help researchers and practitioners to refine their monitoring tools. By fostering appreciation for scientists and alleviating criticism, the results can promote an environment in which scientists feel encouraged to speak out, thereby reducing the spread of misinformation.

References

Au, C.H., Ho, K.K.W. & Chiu, D.K. (2022), The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications, in 'Information Systems Frontiers' **24**, pp. 1331–1354.

Barnes, R. M., Johnston, H. M., MacKenzie, N., Tobin, S. J., & Taglang, C. M. (2018), The effect of ad hominem attacks on the evaluation of claims promoted by scientists, in 'PLoS One' **13**(1), Article e0192025.

Blümel, C. (2024), Anfeindungen gegen Forschende: Eine repräsentative Studie des Projektes KAPAZ, Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW), Hannover.

Cantarella, M., Fraccaroli, N., & Volpe, R. (2023), Does fake news affect voting behaviour?, in 'Research Policy' **52**, Article 104628.

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021), Visual Mis- and Disinformation, Social Media, and Democracy, in 'Journalism & Mass Communication Quarterly' **98**(3), pp. 641-664.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in 'North American Chapter of the Association for Computational Linguistics'.

Egelhofer, J. (2023), How Politicians' Attacks on Science Communication Influence Public Perceptions of Journalists and Scientists, in 'Media and Communication' **11**, pp. 361–373.

Hemmatian, F., Sohrabi, M.K. (2019), A survey on classification techniques for opinion mining and sentiment analysis, in 'Artificial Intelligence Review' **52**, pp. 1495–1545.

Howard, J., & Ruder, S. (2018), Universal language model fine-tuning for text classification.

Jakubik, J., Vössing, M., Pröllochs, N., Bär, D., & Feuerriegel, S. (2023), Online emotions during the storming of the US Capitol: evidence from the social media network Parler, in 'Proceedings of the International AAAI Conference on Web and Social Media' **17**, pp. 423–434.

Tan, K.L., Lee, C.P. & Lim, K.M. A (2023), Survey of Sentiment Analysis: Approaches, Datasets, and Future Research, in 'Applied Science' **13**(7), Article 4550.

Krippendorff, K. (2011), Computing Krippendorff's Alpha-Reliability.

Landis, J. R., & Koch, G. G. (1977), The measurement of observer agreement for categorical data, in 'Biometrics' **33**(1), pp. 159-174.

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023), Building Efficient Universal Classifiers with Natural Language Inference.

Lavazza, A., & Farina, M. (2020), The Role of Experts in the Covid-19 Pandemic and the Limits of Their Epistemic Authority in Democracy, in 'Frontiers in Public Health' **8**, Article 356.

Lewandowsky, S., Ecker, U.K.H., Cook, J., van der Linden, S., Roozenbeek, J. & Oreskes, N. (2023), Misinformation and the epistemic integrity of democracy, in 'Current Opinion in Psychology' **54**, Article 101711.

Ng, A., & Jordan, M. (2001), On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, in 'Advances in neural information processing systems' **14**.

Nogrady, B. (2021), 'I hope you die': how the COVID pandemic unleashed attacks on scientists, in 'Nature' **598**, pp. 250–253.

Park, S., Strover, S., Choi, J., & Schnell, M. (2023), Mind games: A temporal sentiment analysis of the political messages of the Internet Research Agency on Facebook and Twitter, in 'new media & society' **25**(3), pp. 463-484.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018), Improving language understanding by generative pre-training.

Pipal, C., Bakker, B. N., Schumacher, G., & van der Velden, M. A. (2024), Tone in politics is not systematically related to macro trends, ideology, or experience, in 'Scientific Reports' **14**(1), Article 3241.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023), A review on sentiment analysis from social media platforms in 'Expert Systems with Applications' **223** , Article 119862.

Scheufele, D. & Krause, N. (2019), Science audiences, misinformation, and fake news, in 'Proceedings of the National Academy of Sciences' **116**, Article 201805871.

scikit-learn (2024). Machine Learning in Python. Retrieved 5 June 2024, from https://scikit-learn.org/stable/

Vashishtha, S. and Susan, S. (2019), Sentiment cognition from words shortlisted by fuzzy entropy, in 'IEEE Transactions on Cognitive and Developmental Systems' **12**(3), pp.541-550.

Wang, X., Zhang, M., Fan, W., & Zhao, K. (2022), Understanding the spread of COVID-19 misinformation on social media: The effects of topics and a political leader's nudge, in 'Journal of the Association for Information Science and Technology' **73**(5), pp. 726–737.

Weinhardt, C., Fegert, J., Hinz, O., & van der Aalst, W. M. (2024), Digital Democracy: A Wake-Up Call: How IS Research Can Contribute to Strengthening the Resilience of Modern Democracies, in 'Business & Information Systems Engineering' **66**(2), pp. 127–134.

Wankhade, M., Rao, A., & Kulkarni, C. (2022), A survey on sentiment analysis methods, applications, and challenges, in 'Artificial Intelligence Review' **55**, pp. 1–50.

Wolpert, D. H. (1992), Stacked generalization, in 'Neural Networks' **5**(2), pp. 241-259.

Ye, Z., Geng, Y., Chen, J., Chen, J., Xu, X., Zheng, S., Wang, F., Zhang, J., & Chen, H. (2020), Zero-shot Text Classification via Reinforced Self-training, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 3014–3024.

Yin, W., Hay, J., & Roth, D. (2019), Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019), A survey of sentiment analysis in social media, in 'Knowledge and Information Systems' **60**, pp. 617–663.