tweetbusters@GermEval Shared Task 2025: Comprehensive Feature Engineering for Call to Action detection

Anastasia Schwichtenberg
Mittweida University of Applied Sciences
Technikumplatz 17
09648 Mittweida

Abstract

In previous GermEval Shared Tasks, deep learning methods, particularly the fine-tuning of transformer models, were frequently employed to detect harmful content in social media. However, this trend should not overshadow the fact that traditional approaches, especially when combined with comprehensive feature engineering, remain promising. Against this backdrop, we report on our participation in the GermEval Shared Task 2025, where we are pursuing an approach based on traditional machine learning for the first subtask, the detection of calls to action. Our system utilises a variety of linguistic, word list-based, and statistical features, as well as pre-trained sentence embeddings. A soft voting ensemble of established machine learning methods, including random forest, gradient boosting, logistic regression and a support vector machine (SVM), was used as the classifier. On the official test data set, our system achieved a macro-F_1 score of 66.01%.

1 Introduction

The dissemination of harmful content via social media poses a significant risk. Calls for violence or other harmful acts are particularly worrying. Recent examples illustrate the potential danger: in July 2025, video messages called for violence against Druze in the run-up to an Islamist demonstration (Meischen, 2025); in the United Kingdom, the term 'keyboard warriors' has even been coined to describe people who incite violence or hatred on social media, such as blowing up mosques (Riecke, 2024).

Early detection of such content is therefore crucial for security. Since manual analysis is not feasible on a large scale, automated procedures are required. The first subtask of the GermEval Shared Task on Harmful Content Detection (Felser et al., 2025), which involves identifying calls to action (CTA), addresses this problem. The aim is to capture all forms of linguistic appeals intended to elicit specific reactions from the audience – from harmless examples, such as "Share this post!" or "Support us!", to more serious calls, such as "Shoot!". Detecting these calls enables moderators to be supported by an early warning system for prioritising potentially dangerous content. It allows security authorities to plan resources for protests and demonstrations with potential for escalation.

This paper describes our approach to detecting CTA in the context of the first subtask. Our focus is on constructing a precise feature representation and using a soft-voting ensemble classifier to achieve good classification performance.

The remainder of the paper is structured as follows: section 2 provides an overview of existing literature on CTA and harmful content detection, highlighting the specific research gap addressed by our work. Section 3 details our proposed pipeline, covering data preprocessing, extensive feature engineering, and the ensemble model architecture. Section 4 presents a thorough evaluation of our system's performance, including a detailed analysis of results on both training and validation sets. Section 5 discusses the constraints and potential biases of our approach. Finally, section 5 summarises our findings and outlines promising directions for future research.

2 Related Work

The problem of detecting CTA has only been investigated in a few studies to date, with Achmann-Denkler et al. (2024) being the only one to focus on German-language social media posts. As part of the 2021 German federal election campaign, they analysed around 1,400 Instagram posts from candidate accounts. When comparing Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Generative Pre-trained Transformer (GPT-4) (Achiam et al., 2022) in the detection of CTA in Instagram posts, both models

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 375–384, 2025

achieved macro-F_1 scores above 0.9, with BERT performing slightly better.

BERT and its variants have also proven successful with other types of CTA, such as detecting CTAs for protest in Russian-language social media posts (Rogers et al., 2019), requests for help in tweets after a hurricane (Zhou et al., 2022), and distinguishing whether Spanish-language hate speech tweets contained explicit calls for violence (Pérez et al., 2023). However, it should also be noted that Pérez et al. (2023) in particular had access to contextual information in the form of newspaper articles related to the posts to be classified. The lack of such information for the first task of the shared task makes it more difficult.

Although BERT models were also particularly popular in previous GermEval Shared Tasks in the area of detecting harmful content (Bornheim et al., 2021; Gawron and Schmidt, 2021), such as toxic language (Risch et al., 2021), the potential of traditional classification approaches should not be overlooked, especially when computational resources are scarce. For example, as part of the GerMS-Detect task in GermEval 2024 (Gross et al., 2024), Donabauer (2024) found that traditional methods, such as XGBoost, even outperformed BERT-based models in the binary detection of sexism in German comments on online forums. This result already indicates that state-of-the-art standard language models do not necessarily outperform traditional classification approaches in every case.

Even more relevant to our work are approaches focusing on traditional learning and/or feature engineering in the detection of CTA (e.g., Siskou et al., 2022; Ullah et al., 2021). Lampert et al. (2010) aimed to identify requests in emails using the Enron dataset (Klimt and Yang, 2004) to assign them a higher priority. Lampert et al. (2010) employed a Support Vector Machine (SVM) with extensive feature engineering, including bag-of-words (BoW) features as well as structural features such as message length. Although their results were promising, some of the features used are highly email-specific – such as a binary feature for the presence of sender or recipient information – so the approach cannot be directly transferred to subtask 1.

In contrast, Siskou et al. (2022) examined CTA in the social media context of Latin American elections using tweets and Facebook posts. They developed a rule-based system that classifies posts as CTA using keywords and grammatical structures such as imperative forms. However, generalisability is limited because many rules are specifically tailored to the election context. Nevertheless, some word- and grammar-based features could also be of interest for machine learning.

Mathur et al. (2018) demonstrated that machine learning, combined with comprehensive feature engineering, can achieve high performance in detecting blood donation appeals on Twitter, even though the features are based on metadata such as user follower counts, which were not available in the dataset for subtask 1. Finally, Ullah et al. (2021) focused less on feature engineering and more on the classification algorithm when detecting requests for help in English-language disaster tweets. Of the seven methods tested, logistic regression with N-grams and regular expressions performed best. However, like the studies presented so far, this study also refrains from using ensemble learning. Studies on the detection of hate speech in social media (e.g., Raza and Chatrath, 2024) and the GermEval workshops on the detection of offensive tweets (Wiegand et al., 2018; Struß et al., 2019; Risch et al., 2021) have both found that ensemble methods achieve better results overall than their individual components.

To close this research gap, we are investigating the use of a soft voting classifier for detecting CTAs. Moreover, the particular challenge compared to previous work lies in developing features that are fundamentally suitable for any CTA and are not limited to calls in the context of particular events.

3 Methodology

We utilised the annotated dataset provided by the organisers of the Shared Task for subtask 1 (Felser et al., 2025). This dataset comprised German-language tweets from the context of a right-wing extremist group as the primary data source, with each tweet labelled as either CTA (TRUE) or non-CTA (FALSE). The organisers had previously split the dataset into training and test data, with the training dataset containing 6,840 tweets and the test dataset containing 2,982 tweets. The training data was highly unbalanced, where the number of tweets per category can be taken from the training data in Table 1. For a more detailed description, please refer to the overview paper by Felser et al. (2025).

Before feature extraction, the tweets were subjected to a standardised pre-processing pipeline using spaCy (Honnibal et al., 2020) for the reduction to content-bearing words and for normalisation

Class Label	Freq	%
TRUE	663	9.69
FALSE	6177	90.31
Total	6840	100.00

Table 1: Absolute and percentage number of tweets in the categories TRUE (CTA) and FALSE (no CTA) in the training data set.

purposes.

Lowercasing All tweets were converted to lowercase to reduce vocabulary size and treat words identically regardless of their capitalisation.

Punctuation Removal Punctuation marks were removed to focus on the lexical content.

Lemmatisation Words were reduced to their dictionary form (lemma) to group inflected forms (e.g., "gehen", "ging", "geht" all become "gehen").

Stopword Filtering Common function words (e.g., "und", "der", "ein") were removed as they typically carry little semantic information for classification.

3.1 Feature Engineering

To reliably detect CTA, we designed an extensive feature engineering process that extracts a diverse set of linguistic, semantic, and surface-level features from each text. The features are grouped as follows:

Linguistic Features These features capture grammatical structures that can be considered as indications of commands or requests. To extract these features, part of speech (POS) tagging was first performed. To this end, the tweets were tokenised according to defined rules, taking into account exceptions such as contractions and abbreviations, as described by Honnibal et al. (2020). POS tagging was performed using a model trained on the TIGER corpus (Brants et al., 2004) and provided by the Python library spaCy (Honnibal et al., 2020).

The following features were then extracted:

- Number of verbs, nouns and adjectives in the tweet to capture characteristic patterns such as missing nouns in elliptical exclamations, which, according to Blühdorn (2023), can serve as CTA (e.g. "Jetzt mal aufgepasst!", engl. "Now pay attention!").
- Ratio of verbs, nouns and adjectives to the total number of words to obtain normalised measures.
- Binary indicator of whether the tweet begins with a verb – as is typical for German imperative sentences (e.g. "Geh!", engl. "Go!"). According to Siskou et al. (2022), imperative sentences often express CTA.
- Fragment detection: Binary indicator of whether the tweet is very short (i.e. less than four words) and either begins with a verb or consists only of characters such as exclamation marks – often characteristic of direct requests (e.g., "Ergreift die Initiative!", engl. "Take the initiative!")

Semantic and Content-based Features As in previous work on the detection of CTA (e.g. Siskou et al., 2022; Ullah et al., 2021), we defined features based on specific words and patterns that are frequently associated with CTAs. For this purpose, we created two lists: *CTA-KW* with CTA character-specific keywords and *CTA-PAT* with regular expressions that describe phrases typical for CTA. To generate both lists, Perplexity AI (Perplexity AI, Inc., 2024) was prompted in a zero-shot scenario and the resulting lists were then supplemented manually. The CTA-KW list contained a total of 327 words, such as:

- imperative form, e.g., "geh wählen" (engl. "go vote")
- words that express urgency, e.g., "unbedingt" (engl. "absolutely")
- words associated with resistance, e.g., "boykottieren" (engl. "boycott") or "kämpfen" (engl. "fight")

The CTA-PAT list comprised 35 regular expressions that include common imperative forms and phrases to activate readers, such as:

- `komm\s+\w+` to capture expressions such as "komm mit!" (engl. "come on!")
- `mach\s+dich\s+\w+` to describe expressions such as "mach dich stark" (engl. "Be strong!")

Based on the lists of keywords and regular expressions, the following features were implemented:

- Binary indicator indicating whether the tweet contains any word from the CTA-KW list
- Number of words in the tweet that appear in the CTA-KW list
- Binary indicator whether any regular expression from the CTA-PAT list matches the tweet
- Number of regular expressions from CTA-PAT that match the tweet
- A binary feature *cta_boost* that indicates whether at least two different CTA patterns have been detected, which indicates a higher probability of CTA.
- An indicator if the text starts or ends with a CTA keyword, which can emphasise urgency

Additionally, we added two more features to ensure that most forms of CTA were covered:

- Binary feature indicating whether one of the word stems "geh" (engl. "go"), "mach" (engl. "do"), "sollt" (engl. "should"), "müss" (engl. "must"), "komm" (engl. "come"), "stell" (engl. "put") or "nehm" (engl. "take") is part of a word in the tweet. In particular, according to Siskou et al. (2022), modal verbs such as "müssen" and "sollen" imply obligations, which is why they can be part of a CTA.
- Binary feature indicating whether the tweet contains the common informal construction "mal" + "verb" (e.g., "schau mal", engl. "just look") and "probier mal", engl. "just try"), which, by experience, are often used for casual suggestions or soft commands in German.

Structural and Surface Features These features capture non-linguistic but informative text characteristics.

- Text length (in characters) and word count, as it is obvious that calls such as "Wake up!" can be considerably shorter than factual, explanatory statements.
- Average word length, which can be related to the informativeness of words (Levshina, 2022), making it likely that CTAs will tend to use simple language with short words such as "Los!" (engl. "go")
- Number of exclamation and question marks, which can indicate strong emotions or direct address.
- Number of emojis (using Unicode detection), relevant in social media context for expressing intent or tone.
- A binary feature indicating whether a hashtag appears in the tweet, as a brief analysis of the tweets in the training data set showed that these alone can constitute a CTA (e.g., #allesdichtmachen, #niewiederaufmachen).
- Four binary features, each indicating whether a specific mention type occurred (i.e. [@PRE], [@POL], [@GRP], [@IND]), as it seems likely that mentions are used to direct a call to a specific person/group.
- An indicator for very short tweets (fewer than four words), which might be direct, imperative statements.

All of these features are either numerical (e.g., numbers, ratios) or categorical (e.g., binary indicators). This comprehensive set of features forms the basis for the subsequent modelling and ensemble steps, enabling the model to capture the diverse expressions of CTAs in German texts.

3.2 Text Representation

The tweet was represented in two ways to capture its textual content.

- **TF-IDF Features**: Term Frequency-Inverse Document Frequency (TF-IDF) was applied to extract up to 500 features, considering 1-gram and 2-gram word combinations. TF-IDF assigns weights to words based on their importance in a document relative to the corpus (Sparck Jones, 1972).
- **Sentence Embeddings:** We utilised pretrained SentenceTransformer models, specifically paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019), to generate dense, fixed-size vector representations for each sentence. These embeddings captured the semantic meaning and context of the text, providing a powerful complement to the handcrafted features.

To comprehensively describe the characteristics of a tweet, the linguistic, content-related and superficial text features described in subsection 3.1 were first extracted from the pre-processed tweets. Numerical features (e.g., numbers, ratios) were standardised using Z-score normalisation, and categorical features (e.g., binary indicators) were converted using one-hot encoding (Pedregosa et al., 2011). All features were combined into a feature vector.

In addition, the pre-processed tweets were converted into TF-IDF vectors and sentence embeddings. These three vector representations were ultimately combined into a single feature vector that represents the tweet.

3.3 Data Splitting and Balancing

The dataset was partitioned into training and validation sets with an 80/20 split, respectively. To ensure that both sets retain the original class distribution, stratified sampling was employed. To address the inherent class imbalance (where non-CTA samples typically outnumber CTA samples), we applied the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to the training data. The minority class (CTA) was oversampled to achieve 50% of the majority class's size, thereby providing a more balanced training environment without generating redundant exact duplicates. The number of tweets per category in the training dataset (excluding validation data) before and after applying SMOTE is presented in Table 2.

	Class Label	Freq	%
Before SMOTE	TRUE	530	9.69
	FALSE	4942	90.31
	Total	5472	100.0
After SMOTE	TRUE	2471	33.33
	FALSE	4942	66.67
	Total	7413	100.0

Table 2: Absolute and relative number of tweets in the two classes in the training data set (without validation data) before and after applying SMOTE.

3.4 Model Architecture

Our final classification model is a soft-voting ensemble, which combines the predictions of four diverse machine learning algorithms. Each individual model is chosen for its complementary strengths, aiming to capture different patterns in the data:

- **Random Forest:** An ensemble method that builds multiple decision trees and merges their predictions (Breiman, 2001). Random Forest was configured with 100 trees. The weights associated with the classes were chosen inversely proportional to the class frequencies in the training data to deal with the problem of high imbalance, as described by Pedregosa et al. (2011).
- **Gradient Boosting:** A powerful ensemble technique that builds models sequentially, with each new model correcting errors made by previous ones (Friedman, 2001). One hundred boosting stages were carried out.
- **Logistic Regression:** A linear model widely used for binary classification, providing interpretable probability estimates (Yu et al., 2011). The maximum number of iterations that a solver can take to converge to a solution has been limited to 100 iterations. Similar to Random Forest, the class weights were again set inversely proportional to the class frequency.
- **SVM:** A non-linear model with a Radial Basis Function (RBF) kernel, capable of capturing complex relationships (Cortes and Vapnik, 1995). The class weights were specified in the same way as in Random Forest and logistic regression to minimise the problem of imbalance. To make the SVM output suitable for subsequent soft voting, the class membership probabilities were estimated using Platt scaling (Platt, 1999).

The predicted probabilities from these four models were combined using a soft voting procedure, whereby the predicted probabilities of the four classifiers are averaged for each class (Kyriakides, 2019). This approach aggregates the strengths of individual classifiers, typically leading to improved generalisation performance compared to any single model.

3.5 Threshold Optimization

The default probability threshold for binary classification is typically 0.5 (Nikolov and Radivchev, 2019). However, especially with unbalanced data sets, this threshold value may not necessarily be

the best choice (Iguazio Ltd., 2025). Therefore, we explicitly optimised the probability value for the F_1 measure for the first two runs submitted on the validation data set. The third run differed from these only in that it was optimised according to the macro-F_1 measure instead of the F_1 measure. This process involved searching a range of thresholds (from 0.1 to 0.9 with a step of 0.02) and selecting the one that yields the highest F_1 and macro-F_1 scores, respectively. That ensured that our model's final predictions are optimally tuned for the evaluation metric. We refer to the determined optimal threshold value as θ_{opt} in the following.

3.6 Evaluation and Submission

We evaluated our approach on the validation dataset using macro-F_1, the official competition metric. This metric weights precision and recall equally for both classes and is therefore robust against imbalanced datasets (Zhou, 2021). The probability threshold for classification was explicitly optimised on the validation data, as described in subsection 3.5.

To submit predictions on the test data in the competition phase, we then trained the model on the entire training dataset provided, i.e. on all 6,840 tweets (including the data previously used for validation). To obtain the binary label for the test data after the trained ensemble model made its prediction, the previously optimised threshold θ_{opt} was applied: if the mean probability of the four classifiers was above θ_{opt}, the tweet was assigned to the positive class (*TRUE*).

4 Results and Discussion

This section describes and discusses the results achieved by the proposed ensemble approach on the validation and test data. In addition, the performance of the ensemble approach is compared with that of the individual models.

4.1 Results of the proposed ensemble approach

Hyperparameter tuning yielded a result showing that the best macro-F_1 score of 64.3% could be achieved on the validation data when the threshold for classifying a tweet as CTA was set to 0.48.

Using this threshold, the model trained on the complete training data achieved a slightly improved macro-F_1 score of 66.01% on the test data. The results of our other two runs were slightly worse, as noted in all references to Felser et al. (2025).

Compared to the results of the other participants, our best run yielded poor results. For better comparability, Table 3 lists the macro-F_1 scores of our best run on the test data, together with the result of the best performing system of the SuperGLEBER team for subtask 1. The baseline is a gradient boosting classifier based on polarity features and pre-trained sentence embeddings, as described in detail by (Felser et al., 2025). As can be seen from Table 3, our system outperformed this by more than 6%. Whether this improvement is primarily due to the extended feature engineering or the use of the ensemble approach cannot be clearly determined without an ablation study. However, such an investigation is beyond the scope of this work.

Team	**Run**	**Macro-F_1**
SuperGLEBer	3	86.98
tweetbusters	**3**	**66.01**
baseline	-	59.13

Table 3: Macro-F_1 scores achieved by the best system, our approach (tweetbusters) and the baseline system on the test data.

4.2 Discussion and Error Analysis

The low ranking of the submitted run illustrates that the approach presented needs to be further improved. One possible reason for the limited performance could be redundant features. For example, it is questionable whether the parallel use of the same feature as a binary indicator and as a frequency feature actually provides additional information for the classifier (e.g. whether a word from the CTA keyword list occurs versus how often it occurs). Here, the use of feature selection methods that were developed not only to select distinctive features for a category, but also specifically to avoid redundancy, as proposed by Hussain et al. (2020), for example, appears promising, but this must remain the subject of future work.

To better understand which examples were particularly challenging for the classification approach, the misclassified tweets were analysed. Several recurring patterns can be identified that the system has problems with:

- **Very Short Texts or Fragments:** Short, ambiguous phrases (e.g., "Feuer frei!", engl. "Fire at will!") often lack sufficient context for accurate classification without prior knowledge.

- **Ironically Phrased or Ambiguous CTAs:** Sarcasm, irony, or subtle implications are challenging for automated systems, leading to misinterpretations (e.g., "Go on, make my day", which is an imperative but might be ironic).

- **Rare or Creative CTA Formulations:** CTA patterns that deviate significantly from common templates or established keyword lists (e.g., newly emerging slang, highly metaphorical language) are difficult to capture solely through rule-based or frequency-based features.

- **Contextual Dependency:** Some phrases only become a CTA in a specific dialogue or discourse context, which is not fully captured by analysing isolated sentences.

While expanding keyword and pattern lists during development improved recall for some of these challenging cases, these categories remain areas for further improvement.

4.3 Performance of the individual models

After completion of the competition phase, further analysis was conducted to determine the extent to which the ensemble approach is actually superior to the individual models. To this end, the optimal threshold value was determined for each of the probabilistic classifiers using the macro-F_1 measure on the validation data, analogous to the procedure described in subsection 3.5. Table 4 presents the determined threshold values and the resulting macro-F_1 scores on the validation data.

Method	**Threshold**	**Macro** F_1
Random Forest	0.16	58.15
Gradient Boosting	0.18	62.68
Logistic Regression	0.76	**67.08**
SVM	0.24	62.64

Table 4: Individual performance of the four classification methods combined for the soft voting classifier on the validation data, with the best result highlighted in bold.

When analysing the results, it is immediately apparent that all methods – except logistic regression – favour very low thresholds below 0.25. However, the problem is that the classifiers in question then exhibit a high degree of uncertainty when assigning to the positive class, which argues against the robustness of the system. For future investigations, it is therefore essential to calculate additional evaluation metrics, such as the precision and recall of the positive class. It cannot be ruled out that a higher macro-F_1 measure is primarily due to high recall but low precision, i.e. many tweets incorrectly classified as CTA.

In contrast, logistic regression prefers a significantly higher threshold and also slightly outperforms the ensemble classifier by about 3.3%. The fact that the ensemble approach performs poorly suggests that logistic regression should be combined with other, possibly more advanced methods and that a weighted voting approach should be used to assign a higher weight to high-performance models.

5 Conclusion

This paper presents a soft voting ensemble approach combined with comprehensive feature engineering for detecting calls to action (CTAs) in German social media texts as part of the first subtask of the GermEval 2025 Shared Tasks for detecting harmful content. SMOTE was used to address the severe class imbalance in the provided training dataset. In addition, the threshold at which the voting classifier classifies a tweet as a CTA was optimised using hyperparameter tuning. Our best submitted system achieved a macro-F_1 score of 66.01%. This result is below the performance of other participants and may be due, in particular, to a high degree of feature redundancy.

Future work should therefore focus more on the use of feature selection methods to reduce redundancy and systematically investigate the contribution of individual features and their combinations to overall performance. It is also necessary to question whether optimising the threshold based solely on the macro-F_1 score actually produces a reliable system. A targeted analysis of different threshold values could help to achieve a more balanced compromise between precision and recall for the positive CTA class. Furthermore, the question arises as to what extent the integration of deep learning methods into the ensemble model can further improve classification performance compared to the traditional approaches used to date.

Limitations

Our current approach has several limitations:

- **Language and Domain Specificity:** The extensive keyword and pattern lists are tailored explicitly for German social media texts. Adapting the system to other languages or very different domains would require significant re-engineering of these handcrafted features.
- **Dependency on spaCy:** The preprocessing pipeline relies heavily on spaCy's linguistic capabilities for German (lemmatisation, POS tagging). Changes or limitations in spaCy's performance could directly impact the system's accuracy.
- **Lack of End-to-End Deep Learning Exploration:** This research focuses on combining traditional features with neural sentence embeddings and classical ML ensembles as classifiers. Pure deep learning approaches, such as BERT, have not been systematically investigated or evaluated, despite their potential to enable further performance improvements and eliminate the need for complex feature engineering.
- **No Multi-label/Intent Classification:** The system is designed for binary CTA detection. It does not distinguish between different types of CTAs (e.g. call to vote vs. call to share) or other intents.

Ethical Statement

This work aims to support research on mobilisation and intent detection, contributing to understanding communication patterns in social media. We are fully aware of the potential for misuse, such as aiding surveillance, targeted manipulation, or censorship, and strongly recommend the responsible, transparent, and ethical application of such technologies. Our goal is to empower researchers and platform providers to identify and mitigate harmful content, rather than facilitating its creation or suppressing legitimate speech.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 261 others. 2022. GPT-4 technical report. Technical report, OpenAI, San Francisco.

Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting calls to action in multimodal content: Analysis of the 2021 German federal election campaign on Instagram. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and Short Papers*, pages 1–13, Vienna, Austria. Association for Computational Linguistics.

Hardarik Blühdorn. 2023. Imperative und Aufforderungssätze im Deutschen. *Deutsche Sprache*, 51(2):120–149.

Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2021. FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 105–111, Duesseldorf, Germany. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pia Donabauer. 2024. Pd2904 at GermEval2024 (Shared Task 1: GerMS-Detect): Exploring the Effectiveness of Multi-Task Transformers vs. Traditional Models for Sexism Detection (Closed Tracks of Subtasks 1 and 2). In *Proceedings of GermEval 2024 Task 1 GerMS-detect Workshop on Sexism Detection in German Online News Fora (GerMS-detect 2024)*, pages 39–47, Vienna, Austria. Association for Computational Lingustics.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the GermEval 2025 Shared Task on Harmful Content Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany.

Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

Christian Gawron and Sebastian Schmidt. 2021. FH-SWF SG at GermEval 2021: Using transformer-based language models to identify toxic, engaging, & fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 19–24, Duesseldorf, Germany. Association for Computational Linguistics.

Stephanie Gross, Johann Petrak, Louisa Venhoff, and Brigitte Krenn. 2024. GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora. In *Proceedings of GermEval 2024 Task 1 GerMS-detect Workshop on Sexism Detection in German Online News Fora (GerMS-detect 2024)*, pages 1–9, Vienna, Austria. Association for Computational Lingustics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Syed Fawad Hussain, Hafiz Zaheer-Ud-Din Babar, Akhtar Khalil, Rashad M. Jillani, Muhammad Hanif, and Khurram Khurshid. 2020. A Fast Non-Redundant Feature Selection Technique for Text Data. *IEEE Access*, 8:181763–181781.

Iguazio Ltd. 2025. What is Classification Threshold. Retrieved 29 August 2025.

Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML 2004*, volume 3201, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

George Kyriakides. 2019. *Hands-On Ensemble Learning with Python: Build Highly Optimized Ensemble Machine Learning Models Using Scikit-Learn and Keras*, 1 edition. Packt Publishing Limited, Birmingham.

Andrew Lampert, Robert Dale, and Cecile Paris. 2010. Detecting emails containing requests for action. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 984–992, Los Angeles, California. Association for Computational Linguistics.

Natalia Levshina. 2022. Frequency, Informativity and Word Length: Insights from Typologically Diverse Corpora. *Entropy*, 24(2):280.

Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah. 2018. Identification of emergency blood donation request on Twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–31, Brussels, Belgium. Association for Computational Linguistics.

Dennis Meischen. 2025. Islamistische Demo vor Rotem Rathaus: Polizei ohne Dolmetscher. *Berliner Morgenpost*.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-Radivchev at SemEval-2019 Task 6: Offensive Tweet Classification with BERT and Ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Fabian Pedregosa, Fabian Pedregosa, Gael Varoquaux, Gael Varoquaux, Normalesup Org, Alexandre Gramfort, Alexandre Gramfort, Vincent Michel, Vincent Michel, Logilab Fr, Bertrand Thirion, Bertrand Thirion, Olivier Grisel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 5 others. 2011. Scikit-learn: Machine Learning in Python. *Machine Learning in Python*, 10(11):2825–2830.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE access : practical innovations, open solutions*, 11:30575–30590.

Perplexity AI, Inc. 2024. Perplexity. Retrieved 29 August 2025.

John Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.

Shaina Raza and Veronica Chatrath. 2024. HarmonyNet: Navigating hate speech detection. *Natural Language Processing Journal*, 8:100098.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Torsten Riecke. 2024. Regierung will härter gegen Online-Hass und Fake News vorgehen. *Handelsblatt*. Retrieved August 29, 2025.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to Action on Social Media: Potential for Censorship and Social Impact. In *Proceedings*

of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 36–44, Stroudsburg, PA. Association for Computational Linguistics.

Wassiliki Siskou, Clara Giralt Mirón, Sarah Molina-Raith, and Miriam Butt. 2022. Automatized detection and annotation for calls to action in Latin-American social media postings. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 65–69, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, pages 354–365, Nürnberg/ Erlangen, Deutschland. German Society for Computational Linguistics.

Irfan Ullah, Sharifullah Khan, Muhammad Imran, and Young-Koo Lee. 2021. RweetMiner: Automatic identification and categorization of help requests on twitter during disasters. *Expert Systems with Applications*, 176:114787.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–8, Vienna, Austria. German Society for Computational Linguistics.

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75.

Bing Zhou, Lei Zou, Ali Mostafavi, Binbin Lin, Mingzheng Yang, Nasir Gharaibeh, Heng Cai, Joynal Abedin, and Debayan Mandal. 2022. VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. *Computers, Environment and Urban Systems*, 95:101824.

Zhi-Hua Zhou. 2021. *Machine Learning*. Springer, Singapore.