Sosyal Medyada Nefret Söylemi Tespiti: Spor Alanında Bir Vaka Çalışması
Hate Speech Detection on Social Media: A Case Study on the Sports Domain

Bonne Keehnen
Utrecht University
Utrecht, The Netherlands

Senem Kumova Metin
İzmir University of Economics
İzmir, Türkiye

Çiğdem Kentmen-Cin
İzmir University of Economics
İzmir, Türkiye

Hande Aka Uymaz
Muğla Sıtkı Koçman University, Muğla, Türkiye

Albert Ali Salah
Utrecht University, Utrecht, The Netherlands

***Özetçe* —Sosyal medyada belli olaylara ilişkin nefret söyleminin otomatik tespiti zor bir problemdir. Verideki gürültü, bağlamın önemi, eğitim kümelerindeki tutarsız etiketler bu alandaki problemlerin bazılarıdır. Bu çalışmada özellikle spor alanında nefret söylemini tespit etmek için büyük dil modellerine dayanan bir sınıflandırıcının nasıl uyarlanabileceğini inceledik. Veriyi alt-uzaylarda kodlayan farklı modelleri çapraz geçerleme kullanarak MetaHate ve HateCheck veri kümesleri üzerinde karşılaştırdık. Veri kümesi kalitesini analiz ederek gürültüyü azaltmak için ön işleme ve dil filtrelemesi modüllerinin kullanımını inceledik. Spor alanından seçilmiş bir bağlam ile olay odaklı Bluesky ve YouTube veri kümeleri oluşturduk ve bunların bir kısmını elle etiketledik. Sonuçlarımız, veri temizlemenin mütevazı bir katkı yaptığını ve sistemlerin belli bir yüzde ile çalışmakla birlikte daha örtülü hakaretlerde henüz başarılı olmadığını gösterdi. Bluesky verisinde çok az sayıda nefret söylemi olması, problemin dengesiz sınıf dağılımından kaynaklanan zorlukları ve platform farklılıklarını da ortaya koydu.**

Anahtar Kelimeler*—*Nefret söylemi tespiti, doğal dil işleme, MetaHate, HateCheck, Bluesky, YouTube

***Abstract*—Event-specific hate speech monitoring is difficult because social media data are noisy, context dependent, and training datasets may contain inconsistent labels. This paper investigates how an encoder-based classifier can be adapted for detecting hate speech particularly in sports domain. We benchmark encoder models on the MetaHate and HateCheck corpora using stratified cross validation, analyze dataset quality, apply preprocessing and language filtering to reduce noise. We build event-focused Bluesky and YouTube datasets using context-aware collection, topic filtering, and manually annotate a subset to assess agreement. Results show modest gains from data cleaning and reveal systematic failure modes such as negated hate and obfuscated slurs. The Bluesky case study with only a small number of hateful posts highlights platform differences and the need for stronger event-driven retrieval and annotation strategies.**[1]

Keywords*—*Hate speech detection, natural language processing, MetaHate, HateCheck, Bluesky, YouTube

I. Introduction

Hate speech normalizes discriminatory behavior, silences opposing voices, and mobilizes organized hate [1]. Major sports events often intensify identity-based antagonism as large audiences react in real time. Automatically detecting hate speech in such contexts is challenging because hate speech definitions are highly context-dependent. Event-specific processing of such data must separate relevant conversation from background noise, and training corpora assembled from heterogeneous sources can contain label inconsistencies and language contamination.

This paper studies hate speech detection on Bluesky during the UEFA EURO 2024 tournament and asks:

1) which encoder model performs best for binary hate speech classification,
2) how dataset-quality interventions affect performance,
3) to what extent hate speech is observed on the Bluesky and YouTube platforms during EURO 2024.

We analyzed three different BERT-models for hate speech detection; HateBERT, DistilBERT, and DeBERTa, respectively. Apart from experiments on the largest benchmark datasets, we collected and annotated new event-specific datasets from Bluesky and YouTube platforms. As expected, very few genuine hate speech instances were found in such data, as they are actively removed. We present our findings about how well the commonly available tools can address this challenging problem. Our conclusions are that the reported accuracies in the literature are too optimistic, and realistic data collection settings reveal the issues with the current approaches.

[1]This is the uncorrected author proof. Copyright with IEEE. Please cite as: Keehnen, B., S. Kumova Metin, Ç. Kentmen-Cin, H. Aka Uymaz, A.A. Salah, "Hate Speech Detection on Social Media: A Case Study on the Sports Domain," IEEE 34th Signal Processing and Communications Applications Conference (SIU), Istanbul, 2026.

II. Related Work

Hate speech is "language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used" [2]. Machine learning based analysis of hate speech from social platforms is researched extensively, with many public datasets being made available after 2017 [3]. The nature of hate speech in such online environments differs from traditional forms in several ways, including its speed of distribution, the anonymity of its authors, and the challenges associated with moderating it at scale [4], [5]. A general observation is that deep learning models tend to be more accurate when dealing with binary classification for hate speech. One challenge is that models often perform worse on unseen and out of distribution data [6]. Furthermore, platforms have their own language biases, and typically, models are developed in a platform-specific way.

Past studies have highlighted the prevalence of hate speech spikes during major sporting events, particularly those with international participation [7]. While there is extensive research on the detection of hate speech using natural language processing [8], [2], few studies focused on event-specific toxicity, particularly during globally significant events such as EURO 2024. This is especially relevant, as offline events often lead to higher online toxicity [9].

Studies on the X social platform (formerly Twitter) demonstrated significant progress in classifying tweets into categories such as hate speech, offensive language, or neutral content, with some models reporting very optimistic accuracies. While explicit hate speech is relatively rare, offensive language is far more common [10]. Many users engage in harmful discourse without necessarily using obvious hateful words, complicating detection efforts. Context plays an important role in determining whether a statement is hateful or just offensive, and annotators often disagree on classifications.

After Twitter changed ownership, data access became more difficult, and researchers shifted to Bluesky and Mastodon for developing new methodologies that require less data, or pooling resources to maintain API access. Bluesky stands out among the decentralized social media platforms because of its growth in recent times. Since its private beta launch in February 2023, Bluesky reached 40 million registered users by October 2025. In our research, we targeted Bluesky primarily, and validated our findings with additional data from YouTube. In doing so, the paper also contributes to the growing literature on hate speech in YouTube comment sections. As the world's leading video sharing platform and the second most visited website globally, YouTube facilitates extensive user interaction through its comment sections. These spaces enable direct exchanges among users but can also become arenas where hostile interactions emerge. As a result, YouTube provides a large repository of data for identifying and observing discriminatory behavior.

Hate speech research highlights cross-dataset robustness challenges and the frequent conflation of hate speech with general offensiveness [10], [11]. Domain-adapted encoders such as HateBERT retrain BERT on abusive Reddit communities to improve downstream detection [12]. Functional test suites like HateCheck probe model behavior under controlled linguistic transformations and help diagnose systematic weaknesses beyond aggregate accuracy [13]. Newer studies focus on LLM-generated hate speech, and find that newer generation LLMs generate hateful text more difficult to detect [14].

III. Methodology

We investigate encoder-based transformer methods for hate speech detection in social media text, particularly for the sports domain and examine whether task-specific fine-tuning can improve classification performance. The overall workflow consists of dataset preparation, model selection and cross-validated analysis, model refinement via fine-tuning, and evaluation.

A. Datasets

MetaHate [15] is used as the primary dataset for model training and for in-distribution evaluation, providing a broad hate-speech benchmark aggregated from 36 English hate-speech datasets collected from diverse online platforms (e.g. Twitter, Facebook, Reddit, Whisper, Wikipedia). Second, HateCheck [13] is included for evaluation as an external, functional test set designed to probe generalization. It is a controlled evaluation benchmark for hate speech detection: it contains carefully designed test cases that systematically target specific linguistic phenomena (e.g., negation, quotations, group references), allowing researchers to probe whether a model generalizes beyond surface cues. Third, to match the main objective of hate speech detection in sports-related discourse, we construct a Bluesky sports-domain dataset to evaluate how effectively models trained on general hate-speech data transfer to the target sports domain and to quantify the performance gap between general benchmarks and sports-specific content. Finally, we test the models on a domain-specific YouTube dataset we collected to test cross-platform performance.

The Bluesky sports-domain dataset was created by collecting posts from the Bluesky API during UEFA EURO 2024 (14 June–14 July 2024) with the language parameter set to English. Because Bluesky search requires a query, we used hashtag co-occurrence starting from the official #EURO2024 hashtag with additional co-occurring event hashtags (e.g. #CZETUR), complemented by posts containing football players' full names (e.g. Bukayo Saka, Michael Ballack) and participating country names (player names were sourced from a public EURO 2024 player list). The raw scrape yielded 17,221 posts, which became 16,700 unique posts after de-duplication; then BERTopic [16] was used to identify and remove off-topic clusters, producing a final dataset of 10,826 posts. A part of the dataset is annotated by two annotators with an overlapped portion to compute Cohen's kappa (0.857, high agreement) for inter-annotator agreement. Within this set, five clear instances of hate speech were identified. Two out of these were directed at a commentator, one was directed at transgender people, one was racist (about white people), and the last one was a general slur aimed at multiple football players.

For additional experiments and cross-platform comparison, we also collected user comments from YouTube on videos related to UEFA EURO 2024. The dataset was constructed by retrieving globally popular videos using the search term

"Euro 2024" through the YouTube Data API. Videos were collected within a three-month time window covering the month of the tournament as well as one month before and after the event to capture pre-event anticipation, peak engagement during the tournament, and post-event reactions. This search yielded 482 videos and 183,710 user comments. For each video and comment, publicly available metadata were extracted. All non-English comments were translated into English using the Google Translate API to enable cross-linguistic analysis across approximately 60 languages. After collecting the comments, each entry was labeled using automated natural language processing techniques. Specifically, we employed the TweetNLP hate speech detection library, which identifies language that targets individuals or groups based on social identities such as race, ethnicity, nationality, religion, gender, or sexual orientation [17], [18]. The model classified each comment as either hateful or non-hateful. In addition, the Offensive Language Identification component of TweetNLP was used to detect comments containing "any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct" [19, p.75]. Comments labeled as either hateful and/or offensive were combined to form a subset referred to as the YouTube Hate Dataset, resulting in 10,083 hate/offensive comments and 173,627 non-abusive comments.

B. Model Selection

To measure the capabilities of the currently popular tools, we tested three approaches from the BERT-family:

- HateBERT (HB) [12]: a domain-adapted BERT model retrained on hate-related English text.
- DistilBERT (DiB) [20]: a compact, distilled variant of BERT intended to reduce inference cost.
- DeBERTa (DeB) [21]: an encoder model that improves token representation through architectural changes.

All models are treated as text encoders followed by a binary classification head, and trained on the MetaHate [15] dataset. Before training, both non-English instances and platform-specific artifacts such as newline markers, usernames, hashtags, emojis, links, and Unicode decimal codes were removed to finetune models on cleaner training data. To compare base models under consistent conditions, we perform 10-fold stratified cross-validation. The best-performing model is selected based on standard classification metrics (precision, recall, macro-F1, binary cross entropy BCE) computed across folds. Additionally, two complementary analyses were conducted to understand model behavior and data quality. We used confidence learning [22] to identify potentially problematic or inconsistent labels that may affect training quality and measured performance (label quality analysis), and Shapley Additive Explanations (SHAP) [23] based token attribution to examine which tokens contribute most to model predictions.

C. Evaluation

To align with the main objective of building a hate speech detection model for the sports domain, we evaluated performance of models (base, filtered, preprocessed) on both domain-specific and general-purpose test sets. Bluesky sports-related samples represent the target domain and are used to assess how well the model transfers to sports discourse in a realistic setting. But it is not a comprehensive dataset, with few hate speech instances. Therefore, we use MetaHate [15] and HateCheck [13] datasets (as a controlled functional benchmark) to provide general-domain reference points. This three-way evaluation enables a direct comparison between target-domain robustness and generalization, making it possible to quantify the performance gap between sports-specific content and more generic hate speech benchmarks. Finally, we report additional results on the YouTube Hate dataset.

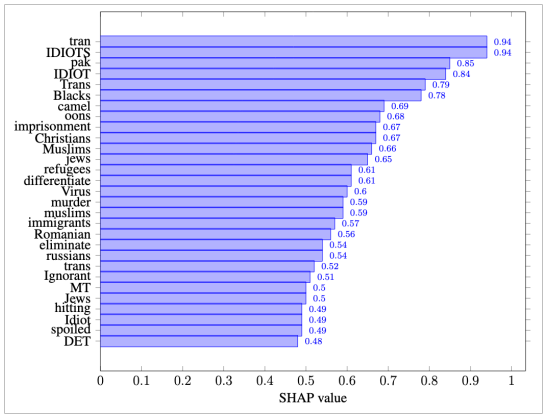

Figure 1: Largest SHAP values of words for the top 1000 TPs

IV. Experimental Results

A. Comparisons on MetaHate

We first report the results on MetaHate in Table I. All three models have similar accuracies on this benchmark.

Table I: 10-fold cross-validation results on MetaHATE.

Model	Loss	Accuracy	Precision	Recall	Macro-F1	Conf. mean
HateBERT	0.314 ± 0.006	**0.905** ± 0.001	0.861 ± 0.002	**0.849** ± 0.001	**0.855** ± 0.001	**0.976** ± 0.060
DistilBERT	**0.303** ± 0.004	0.903 ± 0.001	**0.862** ± 0.001	0.842 ± 0.001	0.851 ± 0.001	0.972 ± 0.064
DeBERTa	0.312 ± 0.025	0.895 ± 0.006	0.848 ± 0.008	0.832 ± 0.014	0.839 ± 0.011	0.960 ± 0.088

We noted that many samples were consistently misclassified by all three models. To gain more insight into potential labeling issues within the dataset, Confident Learning (CL) [22] is used to compute an estimate of potential label issues based on the confidence scores of the predictions. When run on the cross validation predictions of HateBERT, CL finds 75,483 potential label issues, with 39,670 false positives. We retrieved a stratified random sample from the cross validation dataset. Using a 95% confidence level and a 4% margin of error, 600 samples (473 negative) were manually checked using the same proportion of hate speech as in the full dataset. We found that 76 were mislabeled, which is 12.7% of the sampled dataset. We then retrained HateBERT with the top 25% and top 50% of automatically identified problematic samples removed. The accuracy increases from 89.7% to 92.1% (93.9%) when the top 25% (50%) of label issues were removed. The loss decreases from 0.314 to 0.269 (0.232) after the removal of the top 25% (50%).

The SHAP analysis (Figure 1) shows that among the most important tokens that lead to hate speech, five refer directly

to religion, four to race or ethnicity, three to gender, two to (im)migration status. Other tokens are more general abusive words. Some tokens are unclear, but after reviewing the data, the *oons* is part of the word *troons*, which is a slur against transgender women. There are also some tokens that do not directly imply hate speech, such as *virus* or *hitting*, but the context reveals ethnic or nationalistic hate speech.

B. Evaluation on HateCheck, Bluesky and YouTube Hate

The HateCheck and Bluesky evaluations are summarized on Table II. On the YouTube Hate dataset, our best model, HB_{filt} obtains a macro-F1 of 0.53, with precision of 0.20 and recall of 0.32. Detailed inspection on HateCheck reveals that preprocessing improves hate expressed using negated hateful statements but abuse targeted at individuals is difficult to identify, unless identity factors are explicitly named. Different spelling variations also trick the models in making wrong predictions (e.g. "I haet you"). On Bluesky, the selection of the posts that are precisely on topic for targeted analysis is difficult, and the classification problem is severely unbalanced. The latter enables human oversight after automatic screening, which means that false positives (incorrectly tagged normal speech) are less important than false negatives (missed hate instances). This also causes a very low precision. The base model catches two of five hate instances, with 131 false positives, whereas the preprocessed model catches three, while increasing false positives to 249. HB_{filt} reduces false positives to 104, with three false negatives.

Data	Model	Loss	Accuracy	Precision	Recall	Macro-F1	Conf. mean
HateCheck	HB_{base}	**0.431**	**0.881**	**0.940**	**0.884**	**0.867**	0.712 $\pm$ 0.031
HateCheck	HB_{prep}	0.444	0.867	**0.940**	0.863	0.853	**0.714** $\pm$ 0.028
HateCheck	HB_{filt}	0.450	0.859	0.930	0.860	0.842	0.712 $\pm$ 0.035
Bluesky	HB_{base}	0.327	0.988	0.015	0.400	0.511	**0.728** $\pm$ 0.012
Bluesky	HB_{prep}	0.337	0.977	0.012	**0.600**	0.506	0.727 $\pm$ 0.015
Bluesky	HB_{filt}	**0.325**	**0.990**	**0.019**	0.400	**0.516**	**0.728** $\pm$ 0.012

Table II: Model performances on HateCheck and Bluesky.

V. Discussion and Conclusions

We examined hate speech detection in sports-related social media data. Our results have revealed that current systems still struggle with linguistically difficult cases as models rely on explicit identity related tokens, while negation, obfuscated slurs, irony, humor, and other indirect forms of abuse remain challenging. Despite its recent growth, the Bluesky platform contained few hateful instances in our event-specific dataset. Possibly, the users who migrated to Bluesky from X might hold more pluralistic and liberal political orientations, which could shape their online discourse. Platform dynamics and user profiles shape the prevalence of hate speech and the practical usefulness of automated detection tools.

Beyond its empirical contribution, our study also has broader social and policy implications. Online hate speech during major sports events can reinforce structural racism, xenophobia, and misogyny by normalizing stigmatizing language and legitimizing exclusionary ideas about which identities belong in certain sports, discouraging minority fans from following sports or pursuing athletic careers. Online abuse can also reduce athletes' performance, harm their mental wellbeing, and lead them to close their social media accounts and avoid fan interaction [24], [25]. It is reported in [26] that for toxic content classifiers, approximately one third of moderation decisions differ if the training random seed is changed, but different sub-groups (such as racial content or anti-LGBTQ content) are differentially affected. Our findings highlight the need for stronger platform moderation and greater cooperation between social media platforms, sports organizations, and relevant stakeholders to reduce the spread of hateful content during high-profile events.

Acknowledgment

This work is supported by the European Union's Horizon-Europe Research and Innovation Programme under grant agreement No. 101094684.

References

[1] C. Kentmen-Cin, "Hate speech on social media: A systemic narrative review of political science contributions," *Social Sciences*, vol. 14, no. 10, p. 610, 2025.

[2] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM CSUR*, vol. 51, no. 4, pp. 1–30, 2018.

[3] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13(6), 2022.

[4] K. Balci and A. A. Salah, "Automatic analysis and identification of verbal aggression and abusive behaviors for online social games," *Computers in Human Behavior*, vol. 53, pp. 517–526, 2015.

[5] K. Crawford and T. Gillespie, "What is a flag for? social media reporting tools and the vocabulary of complaint," *New Media & Society*, vol. 18, no. 3, pp. 410–428, 2016.

[6] L. Yuan and M.-A. Rizoiu, "Detect hate speech in unseen domains using multi-task learning: A case study of political public figures," *arXiv preprint arXiv:2208.10598*, 2022.

[7] C. Kearns *et al.*, "A scoping review of research on online hate and sport," *Communication & Sport*, vol. 11, no. 2, pp. 402–430, 2023.

[8] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. NLPSM*, 2017.

[9] Y. Lupu *et al.*, "Offline events and online hate," *PLoS one*, vol. 18, no. 1, p. e0278511, 2023.

[10] T. Davidson *et al.*, "Automated hate speech detection and the problem of offensive language," in *Proc. AAAI-WSM*, 2017.

[11] D. Antypas and J. Camacho-Collados, "Robust hate speech detection in social media: A cross-dataset empirical evaluation," *arXiv preprint arXiv:2307.01680*, 2023.

[12] T. Caselli *et al.*, "HateBERT: Retraining BERT for abusive language detection in English," in *Proc. WOAH*, 2021.

[13] P. Röttger *et al.*, "Hatecheck: Functional tests for hate speech detection models," *arXiv preprint arXiv:2012.15606*, 2020.

[14] X. Shen *et al.*, "Hatebench: Benchmarking hate speech detectors on llm-generated content and hate campaigns," in *USENIX Security*, 2025.

[15] P. Piot, P. Martín-Rodilla, and J. Parapar, "Metahate: A dataset for unifying efforts on hate speech detection," in *Proc. AAAI CWSM*, vol. 18, 2024, pp. 2025–2039.

[16] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[17] J. Camacho-Collados *et al.*, "TweetNLP: Cutting-edge natural language processing for social media," in *Proc. EMNLP*, 2022, pp. 38–49.

[18] V. Basile *et al.*, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. SemEval*, 2019, pp. 54–63.

[19] M. Zampieri *et al.*, "Predicting the type and target of offensive posts in social media," in *Proc. NA-ACL*, 2019, pp. 1415–1420.

[20] V. Sanh *et al.*, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[21] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *ICLR*, 2021.

[22] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *JAIR*, vol. 70, pp. 1373–1411, 2021.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeuRIPS*, vol. 30, 2017.

[24] K. Toffoletti *et al.*, "Not just trolls: The experiences and effects of online harm on elite women's sport athletes," *Comm. & Sport*, 2026.

[25] E. Kavanagh and I. Jones, "Understanding cyber-enabled abuse in sport," in *Digital Leisure Cultures*. Routledge, 2016, pp. 132–146.

[26] J. F. Gomez, C. Machado, L. M. Paes, and F. Calmon, "Algorithmic arbitrariness in content moderation," in *Proc. FAccT*, 2024.