Santos *et al. Journal of Big Data* (2026) 13:53
https://doi.org/10.1186/s40537-026-01379-8

Journal of Big Data

RESEARCH Open Access

All together against hate: ensemble based LLMs for multi class hate speech classification in the football context

Guto Leoni Santos[1], Vitor Gaboardi dos Santos[1], Colm Kearns[1], Gary Sinclair[1], Jack Black[2], Mark Doidge[3], Thomas E. Fletcher[4], Daniel Kilvington[4], Katie Liston[5], Patricia Endo[6] and Theo Lynn[1*]

*Correspondence:
Theo Lynn
theo.lynn@dcu.ie
[1]DCU Institute for Business and Society, Dublin City University, Dublin, Ireland
[2]Sheffield Hallam University, Sheffield, United Kingdom
[3]Loughborough University, Loughborough, United Kingdom
[4]Leeds Beckett University, Leeds, United Kingdom
[5]Ulster University, Belfast, United Kingdom
[6]Universidade de Pernambuco, Caruaru, Brazil

Abstract
The rise of social media platforms like Twitter has transformed communication, fostering community engagement and knowledge sharing across diverse groups. However, it has also provided a stage for toxic content, including hate speech, which can manifest in harmful ways within specific contexts, such as discussions surrounding football. Hate speech in this domain often targets individuals or groups based on attributes such as race, ethnicity, or nationality, and is exacerbated by the emotionally charged nature of sports discourse. While binary classification models have traditionally been employed to detect hate speech, they struggle to address nuanced and context-specific forms of abuse, including microaggressions and intersectional hate speech. Multi-class classification enables a more detailed understanding by distinguishing between various types of hate speech, but these models face challenges such as lexico-semantic variability and rapidly evolving norms within the football community. In this paper, we propose an ensemble technique leveraging BERT-based transformers to improve hate speech detection in football-related discussions on Twitter. Our method integrates manually-annotated datasets and multiple classifiers within ensemble frameworks to enhance accuracy and robustness. The results demonstrate that our approach significantly improves the identification of diverse forms of hate speech in the football context, contributing to more effective content moderation and fostering safer online communities.

Keywords BERT, Ensemble methods, Euros, Hate speech detection, Multi-class classification

Introduction

Hate speech is a form of communication that expresses and promotes hatred towards others, often using derogatory terms based on ethnic or national origin, religion, gender, disability, sexual orientation, or political conviction [1, 2].

It includes attacks against individuals and groups based on a wide range of characteristics including race, ethnicity, religion, country of origin, sexual orientation, and disability, amongst others.

The emergence, ubiquitous adoption and use of social media transformed how we communicate, interact, share content with each other [3, 4]. While social media facilitates community building and knowledge sharing across diverse groups [4], it can also foster division [5]. Toxic content is a form of online abuse that is particularly prevalent on social media, particularly in the sports discourse [6, 7]. It covers a wide range of online attacks involving media that targets an individual or group including hate speech, profane or offensive speech, incitement, purposeful embarrassment, incitement, sexual harassment and unwanted explicit content [8]. Extant research suggests significant portions of society have been exposed to online abuse either directly or indirectly. In the UK, the Alan Turing Institute reports that up to 40% have seen online abuse and 20% have experienced online abuse [9], while the Pew Internet & America Life Project reports nearly half (46%) of US teenagers have experienced cyberbullying [10].

Combating hate speech on social media has gained prominence in the media and public discourse with the introduction of the European Union Digital Services Act, which obliges online platforms to address hate speech and other forms of online abuse or face significant penalties [11], and the Brazilian government's temporary suspension of X in 2024 for failing to comply with regulations on illegal content, including hate speech and misinformation [12]. Social media platforms largely rely on the content moderation ecosystem to prevent and mitigate hate speech [5]. This ecosystem comprises volunteer reporters, computational techniques, and directly employed or outsourced human moderators to detect content and enforce rules [5].

Major sporting events often trigger a surge in abusive messages directed at players, teams, and supporters, with racial abuse being a recurrent pattern. This is particularly the case in football. For example, England players faced large volumes of online abuse after the Euro 2020 final, highlighting how sports rivalries can amplify discriminatory behaviour on social media. Recent studies have begun to address this issue more systematically. Mullah et al. [13] investigated racial abuse toward football players on Twitter during Euro 2020, framing the problem as a multi-class classification task. By leveraging transfer learning and fine-tuning pre-trained models, their work achieved high accuracy in identifying negative sentiments and proposed robust transformer-based models for monitoring online abuse in football contexts. This line of research underscores not only the urgency of combating hate speech in sports, but also the promise of advanced NLP methods in building more effective tools for detection and moderation.

Motivation

Solutions to detect hate speech online can focus on a single category or multiple hate speech categories. Binary classification hate speech typically involves distinguishing between hate speech and non-hate speech and reducing the task to a single harmfulness threshold [14]. While this approach can be useful for detecting extreme or explicit forms of hate speech, e.g., slurs or common hate phrases, it struggles with more nuanced or subtle variations, such as coded language or microaggressions, which consequently results in a high number of false negatives [15, 16], as well as context-specific and intersectional hate speech [17]. Binary classifiers are often too simplistic to capture the cornucopia of hate speech that exists and spans a wide range of protected classes including race, gender, sexual orientation, disability, amongst others. These different categories of hate speech suffer from lexico-semantic challenges such as polysemy, neologisms, and

spelling variations, which further complicate classification [18]. Furthermore, some categories of hate speech are under-represented in research and datasets and therefore binary classifiers may fail to detect subtle or less overtly harmful expressions [17].

Multi-class classification overcomes many of these drawbacks by capturing the semantic diversity of hate speech more effectively than binary classifiers [19]. For example, racist hate speech often involves explicit racial slurs or subtle implications of racial superiority, while sexist hate speech may take the form of derogatory comments about gender roles [20]. In many contexts, e.g., sports, these may surface simultaneously. By labelling different types of hate speech, multi-class classifiers offer greater flexibility and accuracy in detecting various forms of hate speech and handling intersectionality i.e., hate speech that may target multiple identity groups simultaneously [17]. Despite its advantages, multi-class classification presents new and different challenges to binary classification. As discussed earlier, different types of hate speech language evolve in their own ways, and their classification as hate speech can be impacted by different levels of social acceptance, reclamation, and other societal norms, thereby making it difficult for a single model to handle all categories equally well [15]. Multi-class classifiers are particularly prone to class imbalance. This can occur at multiple levels. For example, some high level categories may be over-represented (e.g. racism vs sexism) and some slurs may be more common than others, resulting in underperformance in less frequent categories or commonly used slurs or words [16]. More recently, researchers have sought to leverage Bidirectional Encoder Representations from Transformers (BERT) for hate speech detection [21, 22]. While BERT-based transformers are powerful tools for multi-class hate speech detection, they also struggle with semantic variability and intersectionality [21, 23]. Overcoming such challenges requires a range of techniques, including fine-tuning models with larger and more diverse datasets that capture the different variations within each category of hate speech [16, 24], and leveraging ensemble methods to further enhance accuracy by combining the strengths of multiple models [25].

Despite the recent emergence of more advanced Large Language Models (LLMs), such as the GPT and LLaMA families, BERT-based models continue to be highly competitive for text classification tasks [26, 27]. Their relatively smaller size, faster inference times, and lower computational requirements make them especially well-suited for deployment in production environments, where efficiency and scalability are critical [28]. Furthermore, fine-tuned variants of BERTs frequently achieve state-of-the-art or near state-of-the-art performance on a wide range of text classification benchmarks [29], without necessitating the extensive infrastructure typically associated with more recent LLMs.

In this article, we propose a novel pipeline for automatically detecting and classifying multi-class hate speech content in the Twitter discourse. To achieve this, we fine-tune multiple BERT-based transformer models on a new manually curated and annotated dataset and propose a novel ensemble technique that leverages these models to improve overall accuracy. By combining multiple classifiers, ensemble models can offer greater flexibility and accuracy in identifying different types of hate speech. Each classifier in the ensemble may specialise in detecting specific hate speech categories (e.g., racism, sexism), which increases the overall robustness of the model [20].

We evaluate the effectiveness of our approach using a dataset we built focused on football-related content, a popular and widely discussed topic on social media platforms[30]. Football fans frequently turn to social media for its convenient and accessible means

of engaging in real-time discussions during matches [31, 32]. Unfortunately, much like the incidents of hate speech that occur in stadiums [33, 34], the virtual space has seen a troubling rise in hate speech within sports-related discussions [6, 7]. Consequently, detecting hate speech within football discourse presents a highly relevant and critical context for advancing the study of hate speech in online environments.

Contribution

In summary, we make the following contributions:

- Build a new manually annotated dataset of 9,374 tweets, focusing on three types of hate speech: racism, sexism, and ableism. This dataset spans a 15-year period and is related to the context of the UEFA European Football Championships.
- Collate a comprehensive dictionary of terms, word stems, and phrases for detecting potential hate speech related to racism, sexism, and ableism. This dictionary serves as a valuable tool for researchers, aiding in the identification of hate speech across various platforms and contexts, beyond just the soccer domain.
- Fine-tune and evaluate five different multi-class BERT-based LLMs for detecting hate speech within the soccer context.
- Propose a novel ensemble technique called *Dynamic Weighted Average (DWA)*, which adjusts model predictions by assigning them weights based on their accuracy in a test dataset. We also compare the performance of our proposed DWA ensemble method with a standard voting-based ensemble approach and other LLMs.

Related works

In this section, we review related works on detecting hate speech, focusing on strategies involving Machine Learning (ML) models. Specifically, we discuss approaches that address (i) binary classification for determining whether a text constitutes hate speech or not, (ii) multi-class classification to identify different types of hate speech within a given text; and (iii) ensemble methods designed to combine multiple models to improve classification performance.

Binary hate speech classification

Saleh et al. [35] explore transfer learning with BERT [36] and the combination of domain-specific word embeddings with a bidirectional Long Short-term Memory LSTM model for binary hate speech detection. Their results emphasise the importance of large pre-trained models and domain relevance in improving detection accuracy. Siino et al. [37] take a slightly different approach, introducing a CNN-based deep learning model to profile Hate Speech Spreaders (HSS). By employing a single convolutional layer for binary classification (HSS vs. non-HSS), their model achieved 80% accuracy on a multilingual dataset comprising English and Spanish texts. Ghosh and Senapati [38] investigate transformer-based models, such as BERT, RoBERTa, ALBERT, and DistilBERT, on Indian hate speech datasets. They highlight the superior ability of transformers to capture context, particularly when hate speech is embedded in complex language. Their comparisons of multilingual models (e.g., MuRILBERT, XLM-RoBERTa) versus monolingual models (e.g., MahaBERT, NeuralSpaceBERT) reveal that MahaBERT performs best on Marathi data, while MuRILBERT excels in Hindi and Bengali datasets. As noted earlier, binary classification serves as a foundational step in hate speech moderation.

However, for more effective mitigation, a more robust approach is needed–one that goes beyond binary classification to identify specific types of hate speech.

Saleh et al. [35] and Siino et al. [37] focused on binary hate speech detection, with the former using BERT combined with domain-specific embeddings, and the latter employing a CNN-based model for profiling hate speech spreaders. Both approaches emphasise the relevance of model architecture and pre-trained embeddings but limit their scope to binary classification. In contrast, our work expands by fine-tuning multiple BERT-based models for multi-class hate speech detection, covering racism, sexism, and ableism. Additionally, Ghosh and Senapati [38] investigate transformer-based models on Indian datasets, underscoring the importance of multilingual transformers, while our study focuses on multi-class detection in the specific context of European soccer.

Pookpanich et al. [39] investigated binary hate speech detection in YouTube live streaming chats of football news in Thailand. They employed multiple BERT-based models, including multilingual and Thai-specific models, combining automated and manual annotation strategies. Their results demonstrate the effectiveness of transformer-based models for multilingual and domain-specific hate speech detection, with XLM-RoBERTa achieving the highest recall and F1 scores. This study underscores the importance of tailored approaches for sports-related content and multilingual settings, complementing our focus on European football and English-language tweets. Although this study demonstrates strong performance in detecting hate speech within the football context, it is limited to binary classification–a simpler task compared to multi-class classification, which requires capturing the nuanced distinctions between different types of hate speech.

Multi-class hate speech classification

Benítez-Andrades et al. [40] address the challenge of moderating social media content focused on detecting both racist and xenophobic messages in Spanish tweets. The authors evaluate deep learning models, including Convolutional Neural Network (CNN), LSTM, a hybrid CNN + LSTM, and two BERT-based models. Their findings demonstrate that BETO [41], a BERT model pre-trained on Spanish texts, outperformed the other models, achieving the highest precision in detecting hate speech. Yigezu et al. [42] focus on detecting between five different categories of LGBT+ phobias in Mexican-Spanish messages. They employ different traditional ML models and LLMs, including BERT and RoBERTa. They found that RoBERTa outperformed other models with the highest F1-score.

Ahmad et al. [43] introduced the UA-HSD-2025 dataset for hate speech detection in Arabic and Urdu tweets, covering both binary and multi-class classification. They compare a translation-based approach with a joint multilingual model. Using pre-trained transformers such as XLM-R and various embeddings, they achieved high accuracy, around 95%, particularly in multi-class classification, highlighting the potential of robust multilingual models.

While both works emphasise the effectiveness of BERT-based models for specific languages and hate speech categories, our work takes a broader approach by detecting three types of hate speech (racism, sexism, ableism) in a multilingual dataset related to European soccer, fine-tuning multiple BERT-based models, and introducing the DWA

ensemble technique to optimise multi-class classification performance across different contexts.

Although the work proposed by Ahmad et al. [43] is closely related to ours, there are fundamental differences. While they focus on classifying hate speech considering aspects such as sarcasm, exclusionary language, direct, and disguised expressions, our study targets different hate speech categories that affect distinct social groups (racism, sexism, and ableism). Moreover, our work is situated in the sports context, which introduces specific terms and nuanced language, making the task of hate speech classification even more challenging.

Ensemble methods

Khandaya et al. [44] applied ML models and ensemble techniques to detect hate speech in COVID-19-related tweets. They extracted features such as TF-IDF, bag of words, sentence length, and emphatic features as input for models like logistic regression, Naive Bayes, SVM, and decision trees. Ensemble methods including Bagging, AdaBoost, Random Forest, and Gradient Boosting were used for classification. The decision tree classifier performed best among traditional models, while Gradient Boosting outperformed all algorithms. Mnassri et al. [45] explored hate speech classification in tweets using BERT alongside various deep neural network architectures, demonstrating that ensemble approaches can enhance classification performance. Specifically, they integrated BERT with three models: CNN, LSTM, and Multilayer Perceptron (MLP). To further boost performance, the authors applied ensemble methods – including maximum voting, hard voting, and stacking – using these combined architectures. The results showed that ensemble methods outperformed the classifications compared with the traditional BERT and the combination of BERT and other deep neural architectures. Garcia et al. [46] focused on detecting homophobia in Mexican Spanish through two key tasks: identifying whether a tweet is homophobic; and distinguishing between different types of LGBTX+ phobia. They fine-tuned several monolingual and multilingual LLMs, extracting sentence embeddings from each model to input into a multi-input neural network. Their findings revealed that ensemble methods yielded the best performance overall.

While these studies make valuable contributions to the field, further advances can be made. Khandaya et al. [44] used traditional ML models and ensemble techniques to analyse COVID-19-related tweets, relying on hand-crafted features such as TF-IDF and bag of words. In contrast, we focus on fine-tuning BERT-based models for multi-class hate speech detection, specifically addressing racism, sexism, and ableism, by leveraging end-to-end language model features rather than relying on manual feature extraction. Similarly, Mnassri et al. [45] and Garcia et al. [46] employ BERT employ BERT alongside deep learning models and ensemble methods, demonstrating the advantages of these techniques for hate speech detection. Although both studies highlight the importance of ensemble approaches, our innovative DWA ensemble method takes a step further by adjusting predictions based on accuracy, resulting in a more nuanced approach to multi-class hate speech detection compared to their static strategies. Furthermore, our dataset and dictionary specifically target hate speech within a sporting context, enhancing the real-world applicability of our findings.

Dataset

For the purposes of this study, we have selected the international football tournament discourse on Twitter (now known as X) as the empirical context. This context is particularly relevant due to the prevalence of hate speech in football-related content [6, 7] and the intense emotional responses that international tournaments often evoke. Additionally, these events feature idiosyncratic language, including chants, slogans, and unique metaphors, which pose challenges for accurate language analysis.

We also focus on three specific types of hate speech - racism, sexism and ableism. For the purposes of this study, we use the following definitions:

- Racist hate speech involves discrimination based on race, skin colour, or ethnic origin, and seeks [47].
- Sexism refers to prejudice based on an individual's gender, typically targeting women through degrading or stereotypical remarks [48].
- Ableism devalues individuals with disabilities, fostering negative attitudes, stereotypes, and stigma toward those with physical or mental impairments [49].

While racism and sexism feature in existing research [22, 50], ableism is a relatively new area for hate speech classification. In Figure 1, we present our proposed pipeline for creating a dataset with different types of hate speech. Initially, we use the Twitter API to collect data in the context of football by employing specific keywords and terms associated with the Euros. Then, we create a dataset containing three categories of hate-speech: racism, sexism, and ableism. In this process, we first filter tweets using specific hate speech terms, followed by manual labelling. We describe this pipeline in more details within the following sections.

Data collection

The initial step involves identifying and compiling a list of terms and hashtags related to the Euros. These terms and hashtags serve as keywords to search and filter relevant tweets from Twitter. This step is important to ensure that the dataset is focused on Euros-related content, increasing the relevance of the collected data for the task.

We collected tweets spanning one week before and one week after each tournament, covering a total of eight men's and women's tournaments from 2008 to 2022. Using the Twitter API, we defined a query with a list of terms and a specified time frame. The Twitter API then returned all tweets matching our query criteria.

Our collection criteria included tweets featuring the official tournament hashtag, references to the tournament name, and mentions of tournament-specific usernames, UEFA, and FIFA. Additionally, we included hashtags related to all matches in the

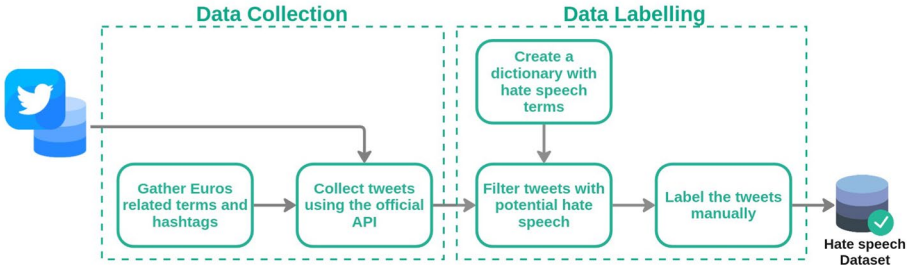


Fig. 1 Pipeline to create dataset of tweets containing different types of hate speech

tournaments and abbreviations for team names (e.g., the match between England and Russia was tagged as #ENGvRUS and #ENGRUS). We also incorporated official championship hashtags (e.g., #euro2016 and 'euro 2016'), and official Twitter accounts (e.g., @euro2016). All the collected tweets were stored in a local database for further analysis. In this work, we limited to tweets that are written in English.

Table 1 summarises the data for each Euro used in this paper. The data is divided into three types: original tweets, representing a general posts made on Twitter; replies, direct responses to a specific tweets; and retweets, when a user shares another user's tweet.

Interesting observations can be made about this dataset. First, the number of tweets increases considerably over the years, due to the widespread adoption and growth of social media platforms, especially Twitter, that have provided fans with a convenient way to share their thoughts and engage in real-time discussions during the tournaments [51]. There is also a noticeable increase in the number of tweets during the Euros 2020 and 2022. Both tournaments took place during the COVID-19 pandemic (with Euro 2020 held in 2021), when fans were unable to attend stadiums. As a result, social media became the primary platform for cheering and discussing the games, explaining the tweets increase.

Furthermore, there is a substantial difference in the number of tweets between the men's and women's Euros, indicating that the men's tournament is much more discussed on social media. For instance, the 2016 men's Euros generated over five times more tweets compared to the women's Euros. This disparity highlights the greater social media engagement and public interest in the men's tournament. Factors contributing to this difference may include broader media coverage, larger fan bases, and historical prominence of men's football compared to women's football. Lastly, it is noteworthy that the volume of replies and retweets has increased over the years. This trend indicates a shift in how people use the platform, with a growing tendency for users to engage in discussions and share content from others. The rise in replies suggests that users are more actively participating in conversations, while the increase in retweets reflects a higher level of content dissemination and interaction within the Twitter community.

Data Labelling

As shown in Table 1, we collected over 16 million tweets related to the Euro. To effectively train machine learning models to detect hate speech, we need training samples that include all types of hate speech we aim to detect, as well as samples that do not contain any hate speech. However, since our data collection was based on keywords related

Table 1 Number of tweets collected per Euros

Euro	Original Tweets	Replies	Retweets	Total
2008	7,281	313	0	7,594
2009	15,665	1,668	756	18,089
2012	1,177,370	54,783	398,259	1,630,412
2013	86,599	10,888	51,847	149,334
2016	867,640	37,576	1,405,363	2,310,579
2017	157,256	22,562	244,531	424,349
2020	2,319,620	984,821	6,201,577	9,506,018
2022	675,700	311,263	1,365,250	2,352,213
Total	5,307,131	1,423,874	9,667,583	16,398,588

to the Euros rather than hate speech, only a small portion of these tweets are likely to contain hate speech content.

To identify potential tweets containing hate speech from the dataset and train our models, we employ a dictionary-based approach. This involves using a predefined list of words to filter tweets that include at least one of these terms. For each type of hate speech, we compile a list of words commonly associated with online abuse and hate speech. We begin by sourcing these terms from the Hatebase project[1], an online collaborative repository designed to assist organisations in moderating online discourse and identifying hate speech. While Hatebase offers a valuable starting point for identifying hate speech terms, it may not include more contemporary terms, neologisms, slang, emojis or other variations prevalent in the general or sports discourse on social media. To address this limitation, we expanded the dictionary by incorporating additional hate speech terms identified in existing literature and references. This expansion includes commonly used terms, phrases, word combinations, plurals, hyphenated forms, and common misspellings. For instance, when identifying racist terms, researchers need to consider variations such as “blacks" (plural), “black-owned" (hyphenated), and alternative spellings like “nigga".

This comprehensive dictionary, composed of terms the hate speech types considered in this study, is available on a GitHub repository[2], allowing other researchers to utilise and further extend it.

Each type of hate speech has its own dictionary with specific words. We executed SQL queries to search in our dataset for potential hate speech tweets. Each query searches for tweets that mention at least one term from the dictionary. Nevertheless, it is essential to acknowledge that certain terms and word stems might produce false positives because of overlaps in general and domain-specific vocabulary and semantic ambiguity. For example, “black" is a term commonly used in racist speech but is also a common word in phrases like “Black Friday" or “the black car". Similarly, “girl" is a term that can be used in a sexist manner in phrases like “You shoot like a girl" but is also commonly used in neutral or positive contexts such as “girl power" or “girl group". For ableism, terms like “blind" may appear in derogatory contexts but are also used in everyday language, such as “blind spot". Therefore, a manual review by three human annotators specialised in hate speech was necessary to identify hate speech tweets. The final label was determined based on the majority agreement among the annotators. Due to resource and time constraints, we only labelled a subset of the potential hate speech tweets. In total, we identified 5,460 tweets related to racism, 2,037 tweets related to sexism, and 1,877 tweets related to ableism.

It is also necessary to obtain samples of non-hate speech tweets to enable the LLMs to distinguish between hate and non-hate content. This involved selecting tweets that were not classified as hate speech but included the previously defined hate speech terms, as well as general tweets that did not contain any offensive words. Human annotators evaluated these tweets to ensure the absence of hate speech connotations and to validate their relevance to soccer-related discussions. To maintain dataset balance, we selected the same number of non-hate tweets as the largest hate-speech class (i.e., racist tweets), resulting in 5,460 non-hate tweets.

[1] https://hatebase.org/

[2] https://github.com/GutoL/H-DICT

Although this dataset provides a valuable and domain-specific benchmark for hate speech detection in football, it has some inherent limitations due to practical constraints. First, it is restricted to English-language tweets, and second, it focuses on three hate speech categories (racism, sexism, and ableism). These choices were driven by the high cost and effort required for manual annotation: labelling a substantial number of tweets for a single category is resource-intensive, and expanding to additional categories or languages would multiply this effort. For multilingual datasets, a considerable number of tweets in each target language would need to be manually annotated to ensure sufficient training data for fine-tuning LLMs. While these limitations restrict generalisability, they allowed us to maintain high annotation quality and ensure the reliability of the resulting dataset, since we wanted to ensure that the tweets were reviewed by human annotators, instead of adopting an automatic labelling approach. We explicitly acknowledge these constraints in the Discussion and highlight directions for future work, including extending the dataset to other languages and incorporating multimodal content such as images, memes, and emojis, which are often present in online hate speech in sports contexts.

Ensemble LLMs for hate speech detection

In Figure 2, we present our proposed pipeline for classifying hate speech content by ensembling results from LLMs. First, we focus on preprocessing the hate speech dataset detailed in Section 3. This step involves cleaning the data to ensure consistency and quality, as well as defining the training and testing datasets. Next, we fine-tune different state-of-the-art LLMs using these datasets to detect hate speech in tweets and propose ensembling strategies to aggregate the results of LLMs to enhance hate speech classification. The rationale behind ensembling is to capture diverse aspects of the data and address individual model limitations [52]. We detail these steps in the following sections.

Dataset preprocessing

Prior to training and evaluating the models, we applied a preprocessing step to clean the raw textual data and ensure its suitability for machine learning algorithms [53]. Our approach is based on the methodology proposed by Hegde et al. [54], with adaptations to better align with recent best practices in social media text classification [55]. Specifically, we removed elements such as user mentions (e.g., @username), hashtags, non-alphanumeric characters, URLs, and stop words. These elements are often domain-specific, noisy, or semantically redundant, and their removal has been shown to improve classification accuracy in similar tasks [56]. Regular expressions were used for pattern-based

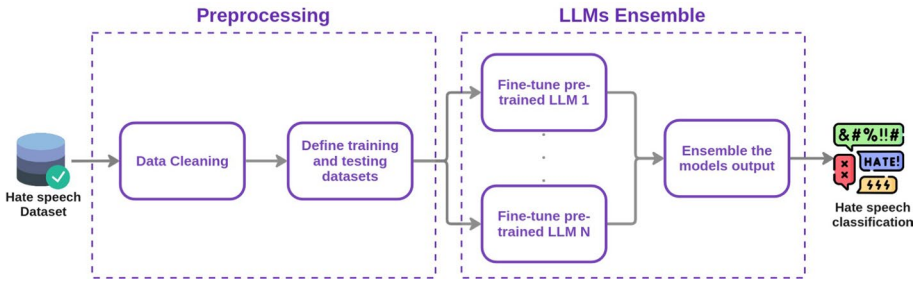


Fig. 2 High-level pipeline to preprocess the dataset, train the LLMs, and perform the ensemble for hate speech classification

text cleaning, and stop words were filtered using the NLTK library[3]. This preprocessing pipeline helps reduce dimensionality and noise, enabling models to focus on more informative linguistic features. An important step in our preprocessing pipeline involves handling emojis. Emojis are standardised pictorial symbols used to represent a wide range of concepts, including emotions, objects, and actions [57]. On social media platforms, they are frequently employed to express nuanced sentiment or to supplement textual content with visual context. Importantly, emojis are not merely decorative; they can carry substantial semantic weight and even alter the perceived meaning of a message [58]. This is especially relevant in tasks such as hate speech detection, where the presence or absence of a specific emoji may determine whether a message is flagged as offensive. For instance, a tweet like "*Can't believe that* 🐒 *scored again*" uses the monkey emoji in a racially charged manner, and without the emoji, the tweet might not be considered hate speech at all. This example illustrates how emojis can be integral to understanding the true intent behind a message. However, despite recent advances, current LLMs often struggle to interpret emojis effectively due to limitations in tokenisation and pretraining datasets that may not fully represent these symbols in meaningful ways. To address this challenge and better preserve semantic context, we used the Python emoji library[4] to convert each emoji into a corresponding textual description. For example, 👍is converted to *:thumbs_up:*, allowing language models to process the emoji as a word-like token. This transformation enables more accurate downstream interpretation and helps retain the original intent behind the message, especially in tasks requiring nuanced sentiment or semantic understanding.

Before initiating the fine-tuning process, we established distinct training and testing datasets with the goal of ensuring robust evaluation and class balance. To construct the test set, we randomly selected 375 samples from each class, including all hate speech categories and the non-hate category. This number corresponds to approximately 20% of the samples from the least represented class (ableism, with 1,877 tweets), and was chosen to avoid over-representation of any class in the evaluation phase.

Balancing the test set is particularly important in text classification tasks involving social data, where class imbalance is common and can lead to skewed performance metrics [59]. Without a balanced test set, models might appear to perform well simply by favouring dominant classes, masking weaknesses in detecting underrepresented types of hate speech. Therefore, by allocating an equal number of test samples per class, we ensure a fair assessment of model performance across all categories, thereby mitigating bias toward majority classes during evaluation.

The remaining samples were allocated to the training set, resulting in a training dataset of 13,334 tweets. To enhance the evaluation's realism and rigour, we also manually added 100 non-hate speech tweets to the test set that contain offensive or provocative language commonly found in hate speech. These challenging cases are critical for evaluating the model's ability to distinguish between harmful and non-harmful usage of such language. For example, the tweet "*Fuck thought this was a resignation letter ffs, we go again*" includes strong language but is not inherently hateful. Including such examples helps test whether models are capable of interpreting linguistic context rather than simply detecting profanity.

[3] https://www.nltk.org/api/nltk.html

[4] https://carpedm20.github.io/emoji/docs/index.html

After finalising the split, the resulting dataset comprises 13,334 tweets for training and 1,600 tweets for testing, ensuring a controlled, balanced, and semantically rich foundation for fine-tuning and evaluation.

Fine-tuning LLMs

We fine-tuned five pre-trained LLMs to classify hate speech. Fine-tuning large language models has proven effective in achieving state-of-the-art performance across a variety of downstream tasks, including classification of topics such as patent [60], racism [22], cycling [61], fake news [62], and public safety incidents [63].

Three of the LLMs in our pipeline are general models that were not previously fine-tuned for hate speech detection: BERT [64], RoBERTa [23], and BERTweet [65]. BERT is a pre-trained LLMs that employs transformers to understand word context in a bidirectional manner. BERT has set new benchmarks across many NLP tasks given its ability to both comprehend human-like text [66]. RoBERTa is an enhanced version of BERT trained using larger datasets, changing the training procedure, using longer input sequences, and adjusting hyperparameters like batch size and learning rate [23]. It achieves superior performance than BERT in many NLP benchmarks, excelling in tasks such as text classification, question answering, and language inference. BERTweet is another LLMs that has the BERT architecture, but fine-tuned on a corpus of 850M English tweets using the RoBERTa [23] pre-training procedure. Fine-tuned on a large corpus of tweets, BERTweet processes the informal language, slangs, and unique characteristics common to the platform, making it useful for tasks that include texts similar to tweets. It has proven effective in detecting hate, offensive and profane content [67] and health misinformation [68].

The other LLMs were fine-tuned using datasets specifically focused on hate speech. Vidgen et al. [69] developed a novel process that involves both human input and model interaction to create dynamic datasets for enhancing hate detection models. Their method includes generating and labelling a dataset of approximately 40,000 entries, which contains around 15,000 challenging perturbations and detailed hate labels. The resulting dataset consists of 54% hateful entries and leads to improvements in model accuracy. The authors fine-tuned the RoBERTa model using this dataset, which is referred to as RoBERTa Hate[5] in this paper. Barbieri et al. [70] propose a framework for classifying tweets across seven different tasks, including sentiment analysis, irony detection, and offensive language identification. The task of identifying offensive language involves determining whether a tweet contains any form of offensive content. For this purpose, they fine-tuned a RoBERTa model using a dataset of 11,916 tweets. We refer to this model as RoBERTa Offensive[6] for the remainder of this paper.

We fine-tuned all five selected LLMs–BERT, RoBERTa, BERTweet, RoBERTa Hate, and RoBERTa Offensive–using the proposed training dataset described in Section 3. The fine-tuning process employed the following hyperparameters: 20 training epochs, a learning rate of 5×10^{-6}, the AdamW optimizer [71], a batch size of 8, and early stopping with patience of one epoch based on validation accuracy.

These hyperparameter values were defined empirically through preliminary experiments, guided by common practices in transformer-based fine-tuning for text

[5] https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

classification tasks [64, 72]. The learning rate was selected from the typical range of 1×10^{-5} to 5×10^{-5}, which offers a good balance between convergence speed and stability when fine-tuning pretrained language models [73]. A relatively small batch size of 8 was used due to GPU memory constraints, which is also consistent with findings that smaller batches can still yield effective generalisation in fine-tuning scenarios [74]. Early stopping was employed to mitigate overfitting and reduce unnecessary computation, especially given the limited size of the training dataset.

The training and experiments were performed using a computer with Intel(R) Core(TM) i7-12700 CPU at 2.10GHz, 32 GB RAM, and Nvidia GeForce GTX 1660 SUPER.

All experiments were executed with a fixed global random seed (seed = 42) to ensure deterministic behaviour in data splitting, model initialisation and training procedures where applicable. We employed stratified train/test splits to preserve the class distribution during hyperparameter definition and final evaluation.

Ensemble methods

We employed two ensemble methods in our experiments. The first method, referred to as *Ensemble Voting*, is a standard majority voting approach used as a baseline. This method combines the predictions of multiple models by selecting the most frequent class label among them. Despite its simplicity, voting-based ensembles are widely regarded as strong baselines in classification tasks [75], as they can effectively reduce variance and benefit from the complementary strengths of individual models. The second method, *Ensemble Dynamic Weighted Average* (DWA), is our proposed approach, which improves upon standard ensembling by weighting each model's prediction according to its performance, leading to more informed aggregation.

In the *Ensemble Voting* method, we input the same tweet into each of the five fine-tuned LLMs and get a predicted class from each model based on the highest probability. The final prediction is determined by the class that is selected most frequently across all models. In the case that two or more classes receives the same number of votes, we chose the output of the model with the best accuracy.

The Ensemble DWA method adjusts model predictions by weighting them according to the models' accuracy on a test dataset. Figure 3 demonstrates this approach using two fine-tuned models as examples. The process involves four key steps. First, we compute

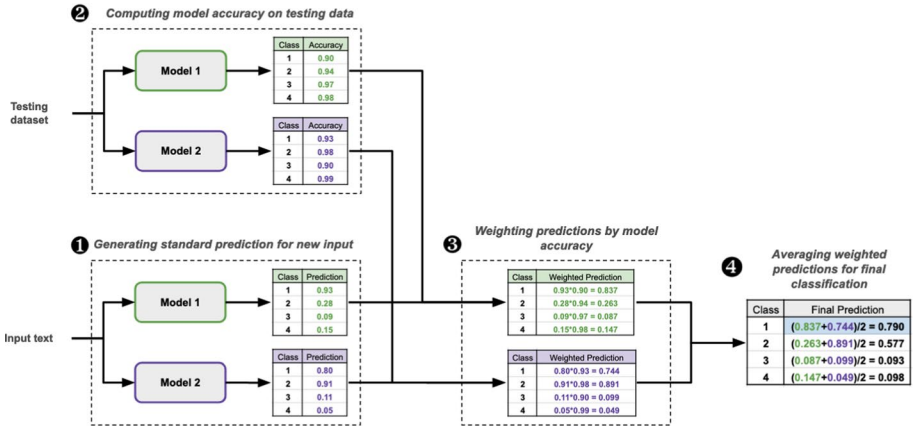


Fig. 3 Example of *Ensemble Dynamic Weighted Average* pipeline

the predictions for a new text input using all fine-tuned LLMs. Second, we calculate the accuracy of each class for all models using a separate testing dataset. Third, we adjust each class prediction by multiplying it with the corresponding model accuracy for that class, resulting in a weighted prediction for each model. Fourth, we average these weighted predictions to obtain a final prediction, selecting the class with the highest score (in Figure 3, this is class 1). This method ensures that predictions from models with higher accuracy for a class are given more influence, while predictions from less accurate models are down weighted.

Algorithm 1 details the Ensemble DWA method. It receives as input the text to be classified, the testing dataset (composed of the text and the respective labels), and the list of fine-tuned LLMs. The algorithm output is the ensemble hate speech classification. In Line 2, we initialise an empty list to store predictions adjusted by model-specific accuracy weights. We then use a for loop (Line 3) to iterate over each fine-tuned LLM. For each model, we first generate predictions for the entire testing dataset (Line 4) and calculate the model's accuracy for each class (Line 5), which serves as a weight reflecting the model's reliability for that class. Next, the model predicts class probabilities for the input text (Line 6), and in Line 7, we store the product of these predictions and the corresponding accuracy weights. This adjusts the predictions based on the model's accuracy per class, increasing the influence of accurate predictions and reducing that of less reliable ones. After processing all models, we calculate the average of these weighted predictions in Line 9 to aggregate them into a single prediction. Finally, in Line 10, we return this final prediction as the output of the ensemble method.

Algorithm 1 Pseudocode of Ensemble DWA method

```
   Input  : text, testing_dataset, trained_models_list
   Output: ensemble_prediction
 1 begin
 2   weighted_predictions ← []
 3   for model in trained_models_list do
 4     model_predictions ← generate_model_prediction
         (testing_dataset.texts, model)
 5     accuracy_models_per_class ← accuracy_per_class(models_predictions,
         testing_dataset.labels)
 6     prediction ← generate_model_prediction (text, model)
 7     weighted_predictions.append(prediction * accuracy_models_per_class)
 8   end
 9   ensemble_prediction ←
       sum(weighted_predictions)/size(weighted_predictions)
10   return ensemble_prediction
11 end
```

Results

In this section, we present the results of fine-tuning different LLMs to detect three categories of hate speech: racism, sexism, and ableism. We begin by analysing the classification results, focusing on metrics such as accuracy, precision, recall, and F1-score, along with an examination of the confusion matrix. Next, we provide examples of tweets from our test dataset, along with their corresponding classification outcomes from the

top-performing models. This will offer insights into the characteristics of our test data and the effectiveness of the models.

Models performance

Table 2 shows the performance of the LLMs and the ensemble methods. BERT, RoBERTa, and RoBERTa Offensive Speech achieved the highest metrics, circa. 88%, outperforming BERTweet and RoBERTa Hate Speech.

Surprisingly, RoBERTa Hate Speech had the worst performance, despite being already fine-tuned for hate speech classification. We believe this happened because it was trained on data that does not align well with the language and context of football-related tweets, resulting in poorer performance and less accurate weight adaptation to the nuances of the new dataset.

BERTweet showed the second-worst performance, even though it was fine-tuned on Twitter data. However, since its pre-training was focused primarily on COVID-19 content, this may have led to weaker weight adaptation for detecting hate speech, as the model was not optimised for the language and context of football-related tweets. In contrast, both traditional BERT and RoBERTa, which performed better, were pre-trained on a larger, more diverse corpus that included general language patterns as well as hate speech, making them more adaptable to different domains.

The RoBERTa Offensive Speech model, initially fine-tuned for offensive speech classification and further fine-tuned for hate speech using our dataset in the context of football, demonstrated strong performance. It achieved an accuracy of 88.63%, a precision of 88.98%, a recall of 88.63%, and an F1-score of 88.54%. These results indicate that the model is able to maintain a balanced performance across precision and recall metrics. The high precision suggests that the model is adept at correctly identifying instances of hate speech, while the high recall indicates its ability to capture most hate speech occurrences.

The vanilla versions of BERT and RoBERTa delivered the best results. BERT demonstrated the second-best performance with an accuracy of 88.57%, a precision of 88.83%, a recall of 88.57%, and an F1-score of 88.50%. RoBERTa slightly outperformed BERT, achieving the best performance among LLMs with an accuracy of 88.94%, precision of 89.28%, recall of 88.94%, and an F1-score of 88.86%. These results demonstrate RoBERTa's superior ability to distinguish hate speech from non-hate speech content, making it the most reliable model for this classification task.

The ensemble methods outperformed all individual LLMs. The voting-based ensemble technique demonstrated robust performance with an accuracy of 89.19%, a precision of 89.45%, a recall of 89.19%, and an F1-score of 89.09%. These metrics suggest that the ensemble effectively leverages the strengths of individual models, resulting in

Table 2 Models performance comparison

Model	Accuracy	Precision	Recall	F1-score
BERT	88.5696	88.8252	88.5696	88.4993
BERTweet	87.3204	87.6369	87.3204	87.2447
RoBERTa	88.9444	89.2844	88.9444	88.8567
RoBERTa Hate Speech	86.6334	86.8253	86.6334	86.5623
RoBERTa Offensive Speech	88.6321	88.9783	88.6321	88.5406
Ensemble Voting	89.1943	89.4485	89.1943	89.0909
Ensemble DWA	89.5690	89.9167	89.5690	89.4675

improved overall performance. The balanced precision and recall values highlight its ability to accurately detect hate speech while maintaining a high rate of correctly classified instances, making it a reliable approach for this task.

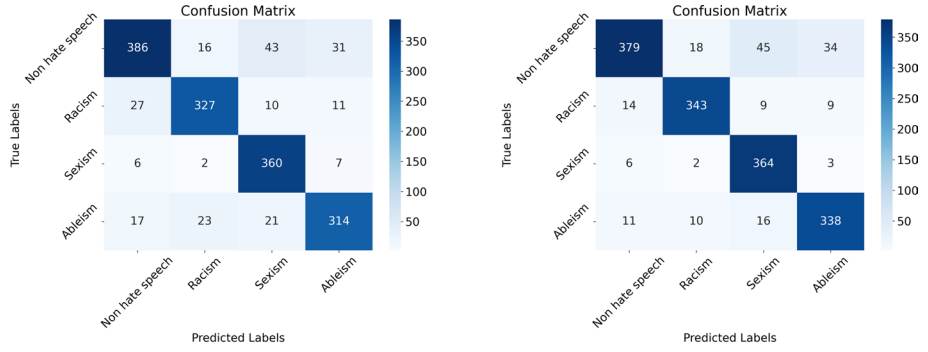
The Ensemble DWA technique demonstrated the best performance, surpassing all the other approaches. It achieved an accuracy of 89.57%, a precision of 89.92%, a recall of 89.57%, and an F1-score of 89.47%. The high precision and F1-score highlight the method's strength in reducing false positives while enhancing the accurate detection of hate speech. These results show that Ensemble DWA effectively harnesses the complementary strengths of individual models, applying dynamic weighting to optimise predictions, which leads to more accurate and consistent hate speech classification. Its ability to finely balance precision and recall ensures that the model not only identifies hate speech instances correctly but also maintains a low rate of incorrect classifications.

To understand better the models' performance per class, Figure 4 shows the confusion matrices for the LLMs with the best (RoBERTa) and worst performances (RoBERTa Hate).

Overall, RoBERTa Hate outperformed RoBERTa in correctly classifying non-hate speech tweets. Specifically, RoBERTa Hate accurately identified 386 tweets as non-hate speech, while RoBERTa identified 379, resulting in a difference of seven tweets. Among these seven non-hate tweets that the RoBERTa model misclassified, two were incorrectly labeled as racism, two as sexism, and three as ableism.

On the other hand, RoBERTa demonstrated superior performance in classifying tweets that contain hate speech, as well as in identifying the specific types of hate speech. RoBERTa Hate accurately classified 327 tweets as racist, whereas RoBERTa classified 343, representing an increase of 4.9%. For sexism, the improvement was smaller, with RoBERTa classifying 364 tweets correctly compared to 360 by RoBERTa Hate, resulting in an increase of just 1.1%. The most significant improvement was observed in the classification of ableism, where RoBERTa Hate correctly identified 314 tweets, while RoBERTa accurately classified 338, reflecting an improvement of 7.6%.

These results indicate that RoBERTa's pre-training strategies enhance its nuanced understanding of the diverse language and context associated with different types of hate speech. The significant improvement in classifying ableism, with a 7.6% increase, underscores RoBERTa's capacity to identify more subtle and specific patterns of discrimination that other models might overlook. Although the increases in detecting racism and sexism were lower – 4.9% and 1.1%, respectively – these results still demonstrate



(a) RoBERTa Hate. (b) RoBERTa.

Fig. 4 Confusion matrix of the models with the worst and best performance

that RoBERTa's fine-tuning process enables it to handle the diverse expressions of hate speech more effectively.

Figure 5 presents the confusion matrices for the ensemble methods employed in this study. As indicated in Table 2, these methods achieved a similar performance, a trend that is further illustrated in the confusion matrices.

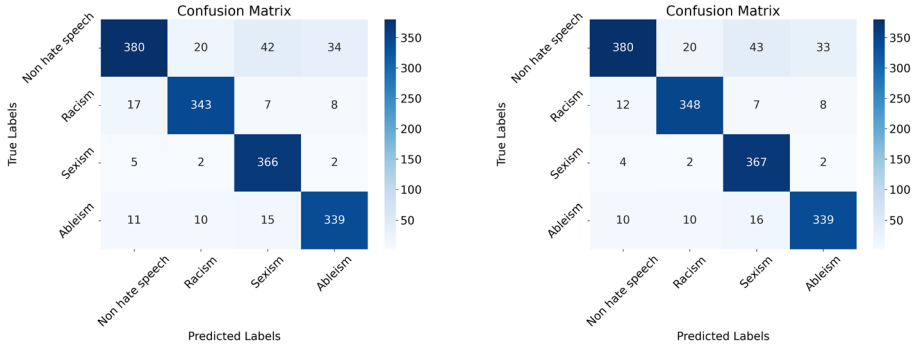
Both ensemble methods correctly classified the same number of non-hate speech tweets, totaling 380. However, the key distinction lies in their ability to detect specific types of hate speech. Unlike the LLMs shown in Figure 4, the improvement between the ensemble methods is more evident when focusing on particular categories. For ableism classification, both ensembles correctly identified 339 tweets, and sexist tweet detection improved only marginally (from 366 to 367 tweets). In contrast, racism detection showed a more pronounced gain: the Ensemble DWA increased correct classifications from 343 to 348 tweets, a 1.5% improvement.

While this numerical difference may appear modest, it carries substantial implications in real-world scenarios. In the football domain, where racist expressions are particularly harmful and often subtle, even a small increase in detection can prevent a significant number of harmful messages from circulating unchecked. In this regard, the Ensemble DWA's 1.5% improvement compared with Ensemble Voting in racism detection is particularly meaningful: it highlights the method's potential advantage in capturing nuanced cases of discriminatory language that single models or simpler ensembles may overlook. The contrast becomes even more evident when compared to the weakest-performing individual model, RoBERTa Hate as shown in Figure 4a, which correctly classified only 327 racist tweets. This means that Ensemble DWA achieves 21 additional correct classifications in this critical category, representing a relative improvement of over 6.42%. Such improvements, although sometimes numerically small, can be decisive in practical moderation systems where the cost of misclassification is especially high.

Classification analysis and discussion

Table 3 presents examples of tweets classified by the top-performing methods. It includes the predictions from both ensemble methods (Voting and DWA) as well as RoBERTa.

Tweet 1 provides an example where the ensemble methods correctly classified the tweet, whereas the top-performing LLMs misclassified it. The text shows a sexist content, since it objectifies women by reducing them to their physical appearance ("fit birds")



(a) Ensemble voting. (b) Ensemble DWA.

Fig. 5 Confusion matrix of the ensemble methods

Table 3 Classification examples. Usernames were removed to preserve the privacy

#	Tweet	RoBERTa	Ensemble Voting	Ensemble DWA	True Label
1	Bbc and ITV Always shows the fit birds watching football... Makes the shit games worth watching! #Euro2012	Non-hate speech	Sexism	Sexism	Sexism
2	Hope Germany gas them #Euro2012	Non-hate speech	Non-hate speech	Racism	Racism
3	The #dwarfs are paying for biting this #ItalianPizza #EURO2012	Racism	Racism	Racism	Ableism
4	This ref is a fucking joke #fix #euro2012	Non-hate speech	Non-hate speech	Non-hate speech	Non-hate speech
5	Come on you lionesses!!!! Let's fucking smash the Irish!!!!	Racism	Racism	Racism	Non-hate speech

and suggests that their presence is the only thing that makes certain football games worth watching. This kind of language perpetuates harmful stereotypes about women, valuing them primarily for their looks rather than their skills, knowledge, or other attributes. While the ensemble methods successfully detected the sexist connotation of the term 'bird' in the text, the RoBERTa model misclassified it as non-hate speech. This discrepancy arises because models like BERT, BERTweet, and BERT Hate were able to recognise the dual meaning of 'bird,' which is challenging, as it is more commonly referred to the animal. The challenge lies in discerning the term's context, a task that some models, particularly the ensemble methods, handle more effectively than others.

Tweet 2 shows an example that the Ensemble Voting and RoBERTa misclassified, while Ensemble DWA classified correctly. The term "gas them", mentioned in the tweet, is a reference to the Holocaust, where millions of Jews and other minority groups were killed in gas chambers by Nazi Germany. Using this expression in any context, especially in reference to Germany and in a competitive event like the Euro 2012 evokes a traumatic historical atrocity. Therefore, the tweet dehumanises a group of people by implying they should be subjected to the same violent and genocidal treatment, which is deeply rooted in racial and ethnic hatred. Although this tweet presents a racist content, some models failed to recognise the historical and cultural significance of the phrase "gas them". Without this context, the model might treat it as a generic violent expression rather than one tied to hate speech, which was the case of most of the models that classified this tweet as non-hate speech. This explains why the Ensemble Voting method also classified this tweet as non-hate speech. However, Ensemble DWA successfully detected the racist intention behind the message. This example highlights Ensemble DWA's advantage over the voting method in classifying hate speech, as most models struggle to consistently interpret subtle language cues and nuanced, historically loaded connotations.

Ableism poses unique challenges for hate speech classification due to its pervasive normalisation in society and the relative lack of focused research, making it difficult for models to distinguish subtly discriminatory language and socially accepted biases. This is particularly the case in the use of ableist language in able-bodied settings, such as the Euros. In these cases, the use of ableist terms perpetuates harmful stereotypes of those who have physical and mental disabilities. Tweet 3 illustrates some of the challenges in detecting ableism. It is an ableist tweet that RoBERTa and both ensemble methods classified incorrectly. The ableist content in this tweet relates to the usage of the hashtag *#dwarfs*, that in this context is inappropriate and offensive, as it refers to people with

dwarfism. Using this term in a derogatory or mocking way contributes to ableism, which involves discrimination or prejudice against people with disabilities. However, all models misclassified the tweet as racist. Although it is challenging to determine the exact reason for this misclassification, it may have stemmed from the hashtag *#ItalianPizza*, leading the models to interpret the reference to Italian nationality as a racist remark. This example illustrates the complexities of hate speech classification, where ambiguous expressions and language nuances can make accurate automated detection difficult.

Tweets 4 and 5 illustrate tweets that do not contain hate speech, with the RoBERTa model and both ensemble strategies correctly classifying one and misclassifying the other. These examples are challenging for the models to interpret, as both include the word "fucking", a term frequently associated with hate speech. However, its presence alone does not necessarily indicate hate speech.

Tweet 4 expresses strong frustration and uses profanity, but it does not contain hate speech. While hate speech typically involves abusive, derogatory, or threatening language targeting an individual or group [76], this tweet criticises a referee, likely within a sporting context, but it does not target anyone based on identity or affiliation. Although the language is offensive and inappropriate, it does not meet the criteria for hate speech. The ensemble methods and RoBERTa classified this tweet correctly as non-hate speech.

Tweet 5 expresses intensive support for the Lionesses (England's women's national football team) using strong language. However, it does not contain racism, since the phrase "Let's fucking smash the Irish", refers to the Irish football team, and not about the national or ethnic group. Nevertheless, the phrase "smash the Irish" might come across as derogatory or inflammatory tone, which can contribute to a negative or hostile atmosphere. It's important to be cautious with language that targets national or ethnic groups, as it can potentially be seen as promoting negative stereotypes or fostering divisiveness. Despite the absence of hate speech, the ensemble methods and RoBERTa classified this tweet as racist. Although the tweet appears to be spirited support in a sports context, language directed at an entire nationality may inadvertently reinforce negative stereotypes or xenophobic undertones, regardless of intent.

Automatically classifying hate speech is a challenging task due to the nuanced and context-dependent nature of language. For instance, in Tweet 5 expresses a message about someone with passionate sports enthusiasm, but it could be also interpreted as aggressive language directed at a nationality raises concerns about xenophobia and racism. Similarly, the use of terms like "dwarfs" in Tweet 3 involves ableism, but understanding this requires models to detect both literal and derogatory meanings, and as discussed, be capable of distinguishing subtle discriminatory language and socially accepted biases.. The complexity arises from the need for models to grasp cultural references, double meanings, and intent behind words, which varies significantly across different contexts.

Limitations and future works

Our research is not without limitations which may provide fruitful avenues for research. First, our study is limited to text data in one language, English, one sport, soccer, and one social media platform, Twitter. Sports discourses are international and multi-lingual and therefore the detection of hate speech needs to be sufficiently sophisticated to cater for the complexity of the sports discourse. Restricting the dataset to Twitter, where posts are generally short, limits the model's exposure to extended discourse. Longer post

formats or transcripts of posts, such as those on Facebook or TikTok or in blogs, introduce greater complexity due to the increased number of tokens and potential for multi-layered contextual shifts. Such data could challenge LLMs' token-processing capacities, revealing limitations in their ability to handle extended, nuanced content accurately. Addressing these issues in future datasets would enable more robust hate speech detection across varying platforms and text lengths. Relatedly, a significant portion of hate speech posts include multi-modal content, e.g., images and videos. Indeed, in some cases, the posts only become hate speech due to that addition of these additional media. Combining text with both the visual and audio context Incorporating multi-modal content, including images and videos, is a critical next step to capture the full spectrum of online hate speech and to provide a more comprehensive understanding of complex hate expressions.

Second, our dictionary and the model's current architecture only classifies racism, sexism, and ableism. We intend to add additional classes such as homophobia, Islamophobia, and antisemitism, amongst others, to address intersectional hate speech and improve nuanced classification across overlapping hate speech categories. As noted earlier, building larger, more diverse, and transparently annotated datasets is central to this goal. Expanding the dataset to include multiple hate speech categories, languages, and socio-cultural contexts would enhance model generalisability and fairness. However, sourcing and annotating hate speech data at scale pose considerable challenges [77]. Hate speech is inherently context-dependent, often under-represented, and evolves rapidly through new slang, coded expressions, and platform-specific conventions. Furthermore, annotation is resource-intensive and subjective, requiring careful management of inter-annotator agreement, ethical considerations for annotator well-being, and privacy protections for users whose content is analysed. Addressing these challenges will require future research to adhere to rigorous annotation protocols, expert supervision, and systematic bias auditing to ensure that expanded datasets remain representative and equitable.

Third, a significant limitation of this work is our focus on exclusively BERT-based models. More advanced architectures such as GPT, Llama, Mixtral, Gemini and others may offer performance advantages although also introducing reproducibility and repeatability challenges. Furthermore, using in-context examples within prompts could help provide a more specific football-related context, potentially improving model performance.

Fourth, a number of opportunities are available to researchers to address a key challenge, namely, distinguishing illegal hate speech from offensive content. Integrating neurosymbolic logic frameworks, which leverage symbolic reasoning capabilities of models like GPTs with neural networks, show significant promise in advancing the state of the art in hate speech detection. As suggested by recent studies [78, 79], neurosymbolic logic holds promise for augmenting LLMs with formal reasoning, potentially improving the model's sensitivity to contextually complex or borderline cases of hate speech. This approach could help refine distinctions between legally actionable and offensive language by providing structured reasoning layers alongside traditional classification techniques. Future work should also explore advanced ensemble techniques, such as stacking and neurosymbolic augmentation, which may offer more adaptable decision-making in varied online environments. Lastly, further research that expands model interpretability and creates clearer benchmarks for performance metrics will ensure the model's

ethical and effective deployment in real-world applications, reflecting wider emphasis on responsible and transparent AI development.

Fifth, while our ensemble approach improves accuracy, it also inherits and may amplify biases present in training data and pre-trained models. Such biases may manifest in uneven detection rates across identity groups or contexts, leading to bias amplification in practice. Furthermore, false positives, where non-hateful content is incorrectly flagged, may result in over-censorship, disproportionately silencing certain communities or legitimate discourse. Conversely, false negatives (i.e., where hateful content is missed) can enable continued harm and reinforce systemic discrimination. To mitigate these risks, future work will include bias auditing using demographic-specific evaluation subsets, fairness-aware loss functions, and human-in-the-loop moderation to contextualise model outputs. Transparency in model design, ensemble weighting, and dataset composition is also crucial to ensure accountability and prevent misuse of the technology in automated moderation systems.

Finally, a promising avenue for future work is the integration of explainable AI techniques into our classification pipeline. While our current approach focused on improving accuracy through ensemble strategies, it comes at the cost of interpretability, since understanding the contribution of each individual model in the ensemble is not straightforward. Incorporating explainable AI would allow us to better understand why specific tweets are flagged as hate speech, increasing transparency and trust in the system.

Conclusion

In this article, we highlighted the critical challenges associated with hate speech detection on social media, particularly the limitations of binary classification models in capturing the complexity and nuance of hate speech. Our exploration of multi-class classification, supported by BERT-based transformers and ensemble methods, demonstrates a promising path toward more effective detection systems.

Our results indicated that among the BERT-based models, RoBERTa achieved the highest metrics, outperforming all others. Furthermore, our ensemble method, DWA, surpassed all the compared techniques, including the traditional majority voting. The results demonstrated a consistent trend of models misclassifying non-hate speech tweets, leading to an inflated number of false positives in the hate speech category. This suggests a potential bias in the models, where certain language patterns are overgeneralised as an indicative of hate speech, likely due to the inherent complexity and ambiguity of language in social media contexts. Additionally, the findings show that sexism was the type of hate speech most accurately classified, while ableism posed the greatest challenge. This difference may be due to the prevalence of explicit sexist language compared to more subtle or context-dependent expressions of ableism, making them harder to detect with conventional text classification methods.

Looking ahead, translating these research findings into practice requires careful consideration of real-world implementation and ethical responsibility. Ensemble models such as the proposed DWA could be integrated into multi-layered content-moderation pipelines, where automated classifiers act as first-line filters that flag potentially harmful content for human verification. This hybrid framework balances efficiency with interpretability and oversight, ensuring that automation supports rather than supplants human judgment. The adoption of confidence thresholds, explainable AI tools, and

transparent reporting mechanisms can further strengthen trust and accountability in practical applications. Ultimately, by combining robust model design, diverse and ethically sourced datasets, and responsible deployment strategies, this work contributes to the development of safer, fairer, and more transparent approaches to moderating online hate speech thereby helping to foster more inclusive and respectful digital communities.

Acknowledgements
The research in this paper was partially funded by the UK Arts and Humanities Research Council and the Irish Research Council (Grant Number AH/W001624/1) and the Federation Internationale de l'Automobile.

Author contributions
All authors contributed equally to the manuscript. TL supervised the research presented in this manuscript.

Data availability
Information on access to data is provided within the manuscript.

Declarations

Ethics approval
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 June 2025 / Accepted: 21 January 2026
Published online: 25 February 2026

References

1. Esau K. Hate speech (hate speech/incivility). DOCA-Database of Variables for Content Analysis 2021.
2. Jovanova D. Hate SPEECH AND MEDIA
3. Aral S, Dellarocas C, Godes D. Introduction to the special issue-social media and business transformation: a framework for research. Inf Syst Res. 2013;24(1):3–13.
4. Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS. Social media? Get serious! understanding the functional building blocks of social media. Bus Horiz. 2011;54(3):241–51.
5. Schoenebeck S, Lampe C, Triêu P. Online harassment: assessing harms and remedies. Soc Media Society. 2023;9(1):20563051231157297.
6. Farrington N, Hall L, Kilvington D, Price J, Saeed A. Sport. London: Racism and Social Media. Routledge; 2017.
7. Kearns C, Sinclair G, Black J, Doidge M, Fletcher T, Kilvington D, et al. A scoping review of research on online hate and sport. Communication & Sport. 2023;11(2):402–30.
8. Thomas K, Akhawe D, Bailey M, Boneh D, Bursztein E, Consolvo S, Dell N, Durumeric Z, Kelley PG, Kumar D *et al.* Sok: Hate, harassment, and the changing landscape of online abuse. In: 2021 IEEE Symposium on Security and Privacy (SP), 2021;247–267. IEEE.
9. Institute TAT. Online abuse prevalence: Summary report. Technical report, The Alan Turing Institute (November 2019). Accessed: 2024-09-14. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_summary_24.11.2019_-_formatted_0.pdf
10. Center PR. Teens and Cyberbullying 2022. Accessed: 2024-09-14 (2022). https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/
11. European Parliament and Council of the European Union: Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX Accessed: 2024-09-12 (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX
12. Silva J, Buschschlüter V. Top brazil court upholds ban of musk's x. BBC News (2024). Accessed: 2024-09-03
13. Mullah NS, Zainon WMNW, Ab Wahab MN. Transfer learning approach for identifying negative sentiment in tweets directed to football players. Eng Appl Artif Intell. 2024;133:108377.
14. Vidgen B, Staton S, Hale S, Kammar O, Margetts H, Melham T, Szymczak M. Recalibrating classifiers for interpretable abusive content detection. In: Fourth Workshop on Natural Language Processing and Computational Social Science 2020, 2020;132–138 . ACL Anthology.
15. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, 2017;11, 512–515
16. Zhang Z, Luo L. Hate speech detection: a solved problem? the challenging case of long tail on twitter. Semantic Web. 2019;10(5):925–45.
17. MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: challenges and solutions. PLoS One. 2019;14(8):0221152.
18. Vidgen B, Harris A, Nguyen D, Tromble R, Hale S, Margetts H. Challenges and frontiers in abusive content detection. In: Proceedings of the Third Workshop on Abusive Language Online (2019). Association for Computational Linguistics.
19. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR). 2018;51(4):1–30.
20. Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA. Machine learning techniques for hate speech classification of twitter data: state-of-the-art, future challenges and research directions. Comput Sci Rev. 2020;38:100311.

21. Mozafari M, Farahbakhsh R, Crespi N. Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS One. 2020;15(8):0237861.
22. Santos GL, Santos VG, Kearns C, Sinclair G, Black J, Doidge M, Fletcher T, Kilvington D, Endo PT, Liston K, *et al.* Kicking prejudice: Large language models for racism classification in soccer discourse on social media. In: International Conference on Advanced Information Systems Engineering, 547–562 (2024). Springer
23. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 2019.
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
25. Lai Z, Zhang X, Chen S. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. arXiv preprint arXiv:2403.13335 2024.
26. Aftan S, Shah H. A survey on bert and its applications. In: 2023 20th Learning and Technology Conference (L&T), 2023;161–166 . IEEE.
27. Wang Y, Qu W, Ye X. Selecting between bert and gpt for text classification in political science research. arXiv preprint arXiv:2411.05050 2024.
28. Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, Gallagher A, Biswas R, Ladhak F, Aarsen T, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663 2024.
29. Joloudari JH, Hussain S, Nematollahi MA, Bagheri R, Fazl F, Alizadehsani R, et al. BERT-deep cnn: state of the art for sentiment analysis of covid-19 tweets. Soc Netw Anal Min. 2023;13(1):99. https://doi.org/10.1007/s13278-023-01102-y.
30. Bruns A, Weller K, Harrington S. Twitter and sports: Football fandom in emerging and established markets. In: Twitter and society [Digital Formations, Volume 89]. 2014. p. 263–80.
31. McDonald H, Biscaia R, Yoshida M, Conduit J, Doyle JP. Customer engagement in sport: an updated review and research agenda. J Sport Manag. 2022;36(3):289–304.
32. Fenton A, Keegan BJ, Parry KD. Understanding sporting social media brand communities, place and social capital: a netnography of football fans. Communication & Sport. 2023;11(2):313–33.
33. Jarvie G. The changing face of football: racism, identity and multiculture in the English game. Sociol Sport J. 2002;19(3):333–4.
34. Kassimeris C, Lawrence S, Pipini M. Racism in football. Soccer & Society. 2022;23(8):824–33.
35. Saleh H, Alhothali A, Moria K. Detection of hate speech using BERT and hate speech word embedding with deep model. Appl Artif Intell. 2023;37(1):2166719.
36. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, 2019;4171–4186
37. Siino M, Di Nuovo E, Tinnirello I, La Cascia M, *et al.* Detection of hate speech spreaders using convolutional neural networks. In: CLEF (Working Notes), 2021;2126–2136
38. Ghosh K, Senapati A. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, 2022;853–865
39. Pookpanich P, Siriborvornratanakul T. Offensive language and hate speech detection using deep learning in football news live streaming chat on youtube in thailand. Soc Netw Anal Min. 2024;14(1):18.
40. Benítez-Andrades JA, González-Jiménez Á, López-Brea Á, Aveleira-Mata J, Alija-Pérez J-M, García-Ordás MT. Detecting racism and xenophobia using deep learning models on twitter data: CNN, LSTM and bert. PeerJ Comput Sci. 2022;8:906.
41. Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J. Spanish pre-trained bert model and evaluation data. arXiv preprint arXiv:2308.02976 2023.
42. Yigezu MG, Kolesnikova O, Sidorov G, Gelbukh AF. Transformer-based hate speech detection for multi-class and multi-label classification. In: IberLEF@ SEPLN 2023.
43. Ahmad M, Waqas M, Hamza A, Usman S, Batyrshin I, Sidorov G. UA-HSD-2025: multi-lingual hate speech detection from tweets using pre-trained transformers. Computers. 2025;14(6):239.
44. Khanday AMUD, Rabani ST, Khan QR, Malik SH. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights. 2022;2(2):100120.
45. Mnassri K, Rajapaksha P, Farahbakhsh R, Crespi N. Bert-based ensemble approaches for hate speech detection. In: GLOBECOM 2022-2022 IEEE Global Communications Conference, 2022;4649–4654 . IEEE.
46. García-Díaz JA, Zafra SMJ, Valencia-García R. Umuteam at homo-mex 2023: Fine-tuning large language models integration for solving hate-speech detection in mexican spanish. In: IberLEF@ SEPLN 2023.
47. UNIES N. International convention on the elimination of all forms of racial discrimination. UN General Assembly (UNGA) 2006.
48. Bigler RS, Liben LS. Developmental intergroup theory: explaining and reducing children's social stereotyping and prejudice. Curr Dir Psychol Sci. 2007;16(3):162–6.
49. Assembly UG. United Nations disability inclusion strategy (UNDIS) 2020.
50. Rodríguez-Sánchez F, Carrillo-de-Albornoz J, Plaza L. Automatic classification of sexism in social networks: an empirical study on twitter data. IEEE Access. 2020;8:219563–76.
51. Yoon J, Pedersen PM. An examination of the public's twitter usage of youth olympic games and olympic games from 2010 to 2016. J Glob Sport Manag. 2022;7(1):71–88.
52. Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Systems with Applications, 2023;122778 .
53. Kumar D, Cohen R, Golab L. Online abuse detection: the value of preprocessing and neural attention models. In: Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2019;16–24
54. Hegde A, Anusha MD, Shashirekha HL. Ensemble based machine learning models for hate speech and offensive content identification. In: FIRE (Working Notes), 2021;132–141
55. Zimbra D, Abbasi A, Zeng D, Chen H. The state-of-the-art in twitter sentiment analysis: a review and benchmark evaluation. ACM Transactions on Management Information Systems (TMIS). 2018;9(2):1–29.

56. Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named entity recognition: fallacies, challenges and opportunities. Comput Stand Interfaces. 2013;35(5):482–9.
57. Guibon G, Ochs M, Bellot P. From emojis to sentiment analysis. In: WACAI 2016, 2016.
58. Mubarak H, Hassan S, Chowdhury SA. Emojis as anchors to detect arabic offensive language and hate speech. Nat Lang Eng. 2023;29(6):1436–57.
59. Srihith D, Sai IV. Training data alchemy: Balancing quality and quantity in machine learning training. Journal of Network Security and Data Mining 2023;6(3) .
60. Lee J-S, Hsiang J. Patent classification by fine-tuning bert language model. World Pat Inf. 2020;61:101965.
61. Santos VG, Santos GL, Lynn T, Benatallah B. Identifying citizen-related issues from social media using llm-based data augmentation. In: International Conference on Advanced Information Systems Engineering, 2024;531–546 . Springer.
62. Pavlov T, Mirceva G. Covid-19 fake news detection by using bert and roberta models. In: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), 2022;312–316 . IEEE.
63. Zahera HM, Elgendy IA, Jalota R, Sherif MA, Voorhees E. Fine-tuned bert model for multi-label tweets classification. In: TREC, 2019;1–7 .
64. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018.
65. Nguyen DQ, Vu T, Nguyen AT. Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200 2020.
66. Ghojogh B, Ghodsi A. Attention mechanism, transformers, bert, and gpt: tutorial and survey 2020.
67. Glazkova A, Kadantsev M, Glazkov M. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. arXiv preprint arXiv:2110.12687 2021.
68. Wahle JP, Ashok N, Ruas T, Meuschke N, Ghosal T, Gipp B. Testing the generalization of neural language models for covid-19 misinformation detection. In: International Conference on Information, 2022;381–392 . Springer.
69. Vidgen B, Thrush T, Waseem Z, Kiela D. Learning from the worst: Dynamically generated datasets to improve online hate detection. arXiv preprint arXiv:2012.15761 2020.
70. Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv:2010.12421 2020.
71. Loshchilov I. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 2017.
72. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, *et al.* Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020;38–45
73. Mosbach M, Andriushchenko M, Klakow D. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. arXiv preprint arXiv:2006.04884 2020.
74. Smith SL, Kindermans P-J, Ying C, Le QV. Don't decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489 2017.
75. Kuncheva LI. That elusive diversity in classifier ensembles. In: Iberian Conference on Pattern Recognition and Image Analysis, 2003;1126–1138 . Springer.
76. Council of Europe: Combating Hate Speech. Council of Europe 2022.
77. Kovács G, Alonso P, Saini R. Challenges of hate speech detection in social media: data scarcity, and leveraging external resources. SN Comput Sci. 2021;2(2):95.
78. Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229 2024.
79. Boix-Adsera E, Saremi O, Abbe E, Bengio S, Littwin E, Susskind J. When can transformers reason with abstract symbols? arXiv preprint arXiv:2310.09753 2023.

Publisher's Note