# Is a bunch of words enough to detect disagreement in hateful content?

**Giulia Rizzi**[1,2], **Paolo Rosso**[2,3], **Elisabetta Fersini**[1]
[1]University of Milano-Bicocca, Viale Sarca, 336 - Milan, Italy
[2]Universitat Politècnica de València, Camino de Vera, Valencia, Spain
[3] ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain
g.rizzi10@campus.unimib.it, prosso@dsic.upv.es, elisabetta.fersini@unimib.it

## Abstract

The complexity of the annotation process when adopting crowdsourcing platforms for labeling hateful content can be linked to the presence of textual constituents that can be ambiguous, misinterpreted, or characterized by a reduced surrounding context. In this paper, we address the problem of perspectivism in hateful speech by leveraging contextualized embedding representation of their constituents and weighted probability functions. The effectiveness of the proposed approach is assessed using four datasets provided for the SemEval 2023 Task 11 shared task. The results emphasize that a few elements can serve as a proxy to identify sentences that may be perceived differently by multiple readers, without the need of necessarily exploiting complex Large Language Models. The source code and dataset references related to our approaches are available at https://github.com/MIND-Lab/Hate-Speech-Disagreement-Detection/.
**Warning: This paper contains examples of language that may be offensive.**

## 1 Introduction

In the landscape of social networks, hate speech is a growing concern. However, most of the existing detection methods do not take into account the subjectivity of the task and lack in considering different perspectives, resulting in a critical gap in addressing the inherent subjectivity of this phenomenon when designing hate speech prediction models.

Several psycho-social studies (LaFrance and Roberts, 2019; Huddy and Aarøe, 2019; Sap et al., 2019; Hoskins and Tulloch, 2018) have shown that hate perception is subjective and highly dependent on a range of factors such as preconceptions, stereotypes, cultural background, anonymity of the source, and the specific context in which the speech occurs. Among the possible sources of disagreement, annotators' opinions, beliefs, and knowledge have been identified by several investigations in the state-of-the-art (Sandri et al., 2023; Sap et al., 2022). While disagreement is capturing researchers' attention, the majority of works focus on a posteriori exploiting disagreement information to improve the quality of data (Beigman Klebanov and Beigman, 2009; Sang and Stanton, 2022) or including it in the training phase of machine learning models to improve prediction performance (Lee et al., 2023). Only a few of them address the problem of a priori modeling perspectivism (Sandri et al., 2023; Cabitza et al., 2023) and recognizing potential textual triggers of such a disagreement (Rizzi et al., 2024a).

Detecting disagreement in a hateful sentence and identifying the corresponding disagreement-related constituents could play a fundamental role when creating gold-standard benchmarks to be submitted to crowdsourcing workers. For those contents that could lead to disagreement, specific annotation policies could be adopted (e.g., more annotators to be involved, removal of the sample from the dataset that should be annotated, etc..). Alternatively, specific highlights could be provided to the annotators to focus more on specific constituents that could be perceived differently by the readers (e.g., underlining words, hashtags, or emoji that have been identified as disagreement-related constituents that should be carefully evaluated). In this paper, we propose a novel technique for detecting disagreement in hate speech and identifying sentence features that can suggest a lack of agreement among different readers. The proposed method looks at several textual elements (here referred to as *constituent*), including words, emoticons, and hashtags, to identify the ones that are likely associated with disagreement. Each constituent, opportunely represented in a contextualized embedding space, is evaluated by defining a weighted probability function to account for nuanced perceptions of different elements. Additionally, we investigated if the proposed approach, which is based on the evaluation of a bunch of words, is enough when compared with predictions based on Large Language Models. In order to evaluate the efficiency of our approach, multiple experiments have been performed using hate speech datasets from the SemEval 2023 - Task 11 on Learning With Disagreements (Le-Wi-Di) (Leonardelli et al., 2023). These datasets cover a wide range of features, such as annotation techniques, text kinds, and goals. The diversity of the data allowed us to assess the capability of the proposed approach to identify disagreement at sentence level, by leveraging on selected elements considering the different contexts in which they appear.

In summary, three main contributions are given:

- Contextualized embeddings coupled with weighted probability functions have been proposed to detect disagreement-related constituents in hateful content.
- Several aggregation strategies are investigated to predict the disagreement label associated to each sentence.
- A comparison with a few Large Language Models, opportunely fine-tuned to detect disagreement, has been performed, considering as key elements to evaluate both prediction capabilities and computational requirements.

The paper is organized as follows. In Section 2 an overview of the state of the art is provided, while in Section 3 the proposed approach is detailed. In Section 4, the adopted datasets are presented, while the achieved results are reported in Section 5. In Section 6, conclusion and future research directions are drawn. Finally, in Section 6, the impact of the proposed approach and its current limitations are highlighted.

## 2 Related Work

Various natural language tasks, like sentiment analysis or hate speech detection, have been shown to display ambiguity or subjectivity (Uma et al., 2021). As a consequence, an emerging area of research challenges the assumption that each instance possesses a unique perception and interpretation. Subjectivity is represented in datasets through multiple annotations or the addition of confidence levels to ground truth labels. The general idea is to use several labels to represent the diverse opinions of annotators with different perspectives and understanding (Uma et al., 2021).

The information reflecting annotators' disagreement has primarily been used to improve dataset quality by excluding instances marked by annotator disagreement (Beigman Klebanov and Beigman, 2009; Sang and Stanton, 2022). Alternatively, the annotators' disagreement has been used during training of machine learning models accordingly to two different strategies, i.e., by either assigning weights to instances to prioritize those with higher confidence levels (Dumitrache et al., 2019), or by inducing directly from disagreement without considering aggregated labels (Uma et al., 2021; Fornaciari et al., 2021).

While numerous research papers have been devoted to understanding the reasons behind annotators' disagreement (Han et al., 2020; Sandri et al., 2023; Sang and Stanton, 2022) or to leverage on disagreement when training classification models, less attention has been devoted to explain and a priori recognize disagreement in hateful content (Shahriar and Solorio, 2023; Gajewska, 2023; Sullivan et al., 2023; de Paula et al., 2023; Erbani et al., 2023; Vallecillo-Rodríguez et al., 2023).

In particular, it has been demonstrated how different annotators adopt diverse strategies, involving the adoption of ad-hoc shortcuts and identifying specific patterns, when performing a given task (Han et al., 2020). A significant contribution to the understanding of how humans annotate data is presented by Sang and Stanton (2022), where the authors demonstrate that factors such as age and personality strongly influence annotators' perception of offensive or hateful content. In (Sandri et al., 2023), the authors propose a taxonomy of possible reasons leading to annotators' disagreement and evaluate the impact on classification performance of the different types. Specifically, the authors identify four macro categories of reasons behind disagreement: sloppy annotations, ambiguity, missing information, and subjectivity. Furthermore, methods to examine the annotation quality and consistency have been proposed, aiming at obtaining a clear understanding of users' experience (Lavitas et al., 2021; Sang and Stanton, 2022).

Finally, a few recent works have focussed on explaining and recognizing disagreement. The approach proposed by Astorino et al. (2023) exploits integrated gradients in the definition of a *filtering strategy* aiming at identifying both disagreement and hate speech while identifying textual constituents that contribute in hateful messages explanation. A more recent approach (Rizzi et al., 2024a) proposes a probabilistic semantic approach for the identification of disagreement-related constituents in hateful content. The results achieved in the state of the art suggest that although promising results can be achieved by Large Language Models (LLMs), comparable performances using lower computational resources can be obtained with simpler strategies.

## 3 Proposed Approach

This work represents an extension of the approach proposed by Rizzi et al. (2024a), with the objective of enhancing constituent contextualization and defining a more comprehensive model.

Based on the hypothesis that disagreement can derive from specific constituents within a sentence that can be perceived differently and, therefore, achieve a different interpretation and connotation in relation to the task's label, a score representing the potential for disagreement has been defined.

The proposed approach is characterized by the following steps:

- **POS tagging constituent selection**: for each word in a given sentence, the corresponding lexical term has been identified through Part Of Speech (POS) tagging[1]. The elements corresponding to relevant lexical terms (i.e., adjectives, adverbs, interjections, nouns, pronouns, proper nouns, verbs, and hashtags) have been selected as constituents[2].

- **Constituent Embeddings:** for each constituent $c$ selected from the given sentence, its contextualized embedding representation $\vec{\mathbf{v}}_c$ is obtained by means of the mBERT model.

- **Most similar constituents:** given a constituent $c$ with the corresponding embedding $\vec{\mathbf{v}}_c$, the set $S_c$ of the most similar constituents to $c$ is determined according to:

$$S_c = \bigcup_t \{t | cos(\vec{\mathbf{v}}_t, \vec{\mathbf{v}}_c) \geq \psi\} \qquad (1)$$

  where $cos(\vec{\mathbf{v}}_t, \vec{\mathbf{v}}_c)$ is the cosine similarity between the contextualized embedding representation of element $c$ (i.e., $\vec{\mathbf{v}}_c$) and the contextualized embedding representation of the element $t$ (i.e., $\vec{\mathbf{v}}_t$), where $t \in T$ with T representing the set of constituents identified in the training dataset by performing the previously defined steps. Finally, $\psi$ is a threshold that has been estimated via a grid search approach on the validation dataset.

- **Disagreement Score:** The proposed disagreement score is grounded on probability weighting functions (Prelec, 1998), which are linear and nonlinear functions of probability widely known in behavioral decision theory and behavioral economics. Weighted probabilities denote a probabilistic model wherein individual outcomes are associated with distinct weights, reflecting the differential likelihood of occurrence (Gonzalez and Wu, 1999; Nardon and Pianca, 2015). By assigning appropriate weights to relevant events, **it becomes possible to selectively focus on the subset of events whose occurrence significantly influences the probability of the event under consideration**. This selectivity enhances the precision of analyses and allows for a more targeted understanding of the complex interplay between events within a given system.

  In our case, the weighted probabilities are used to compute the constituent disagreement score by only taking into account the constituents in the selected neighborhood. In particular, given a constituent $c$ with the corresponding set of most similar constituents $S_c$, the weighted probability of the contextualized constituent $s \in S_c$ to be associated with the positive label (+), i.e., the agreement label, can be estimated as:

$$P\left(s^+\right) \frac{cos(\vec{\mathbf{v}}_s, \vec{\mathbf{v}}_c)}{\sum_{a \in S_c} cos(\vec{\mathbf{v}}_a, \vec{\mathbf{v}}_c)} \qquad (2)$$

  Where $P(s^+)$ represents the probability of the constituent $s \in S_c$ to be associated with the positive class label.

  Similarly, given a constituent $c$ with the corresponding set of most similar constituents $S_c$, the weighted probability of the contextualized constituent $s \in S_c$ to be associated with the negative label ($-$), i.e., the disagreement label,

[1] For POS Taggins we used `-core_web_sm` models by spaCy `https://spacy.io` version 3.6

[2] According to the selected spaCy model, the POS tag excluded from the selection are: adposition, auxiliary verb, coordinating conjunction, determiner, numeral, particle, punctuation, and subordinating conjunction.

| Dataset | Language | N. items | Task | Annotators | Pool Ann. | % of items with full agr. |
|---|---|---|---|---|---|---|
| HS-Brexit | En | 1,120 | Hate Speech | 6 | 6 | 69% |
| ArMis | Ar | 943 | Misogyny and sexism detection | 3 | 3 | 86% |
| ConvAbuse | En | 4,050 | Abusive Language detection | 2-7 | 7 | 65% |
| MD-Agreement | En | 10,753 | Offensiveness detection | 5 | >800 | 42% |

Table 1: Datasets characteristics.

can be estimated as:

$$P\left(s^{-}\right) \frac{cos(\vec{\mathbf{v}}_s, \vec{\mathbf{v}}_c)}{\sum_{a \in S_c} cos(\vec{\mathbf{v}}_a, \vec{\mathbf{v}}_c)} \quad (3)$$

Where $P(s^{-})$ represents the probability of the constituent $s \in S_c$ to be associated with the negative class label.

Given the weighted probabilities estimated according to Equation (2) and (3), the Disagreement Score for any constituent $c$ is defined as:

$$DS(c) = \sum_{s \in S_c} P\left(s^{+}\right) \frac{cos(\vec{\mathbf{v}}_s, \vec{\mathbf{v}}_c)}{\sum_{a \in S_c} cos(\vec{\mathbf{v}}_a, \vec{\mathbf{v}}_c)} - P\left(s^{-}\right) \frac{cos(\vec{\mathbf{v}}_s, \vec{\mathbf{v}}_c)}{\sum_{a \in S_c} cos(\vec{\mathbf{v}}_a, \vec{\mathbf{v}}_c)} \quad (4)$$

Equation 4, which can be seen as a difference of all weighted probabilities, ranges in the interval from -1 to 1. The closer the score is to minus one, the more the constituent is related to the disagreement label. The closer the score is to one, the more the constituent is related to the agreement label.

The disagreement scores allow us to estimate the disagreement that may arise between annotators. The Sentence Disagreement Score (SDS) has been estimated by aggregating the scores computed for the single constituents according to the following strategies: **Sum**, **Mean**, **Median**, and **Minimum**. For each aggregation strategy, a threshold $\pi$ has been estimated via a greed search approach to assign the final class label of the sentence.

## 4 Datasets, Baselines and Performance Metrics

In order to evaluate the computational potential of the proposed approach, both from the prediction capabilities and the computational resources needed, 4 benchmark datasets provided by SemEval 2023 Task 11 related to Learning With Disagreement (Leonardelli et al., 2023) have been adopted.

The datasets have different characteristics in terms of types (social media posts and conversations), languages (English and Arabic), goals (misogyny, hate-speech, offensiveness detection), and annotation methods (experts, specific demographics groups, and general crowd). Their characteristics are summarized in Table 1.

- *Hate Speech on Brexit (HS-Brexit)* (Akhtar et al., 2021). This dataset consists of 1,120 English tweets collected with keywords related to immigration and Brexit. The dataset was annotated with hate speech, aggressiveness, offensiveness, and stereotype by six annotators.

- *Arabic Misogyny and Sexism (ArMIS)* (Almanea and Poesio, 2022). The dataset consists of Arabic tweets to study the effect of bias on sexist judgments, focusing on the impact of being conservative or liberal. The data was labeled by three annotators, one conservative male, one moderate female, and one liberal female.

- *ConvAbuse* (Cercas Curry et al., 2021). The dataset contains 4,185 English dialogues between users and two conversational agents. The user dialogues have been annotated by experts in gender studies.

- *Multi-Domain Agreement (MD-Agreement) (Leonardelli et al., 2021).* The dataset consists of 10,000 English tweets from three different domains (*BlackLivesMatter*, *Election2020*, *Covid-19*). Each tweet was annotated as offensive or not by 5 annotators.

All the datasets are characterized by the presence of hard-labels (hateful/non-hateful) and soft-labels (disagreement) for each instance. According to Poletto et al. (2021) all these tasks are under the *hate* umbrella since aggressive, offensive, and abusive language can be triggered by hate, and misogyny is a form of aversion towards a specific target. For this reason, from now on we will refer to *hate* as a comprehensive word embracing all the above-mentioned forms of hostility. Since in this work, disagreement detection is addressed as a binary task, an agreement label has been derived from the available soft-labels. In particular, the agreement label is set equal to $(+)$ when there is a complete agreement among the annotators, while equal to $(-)$ in all the other cases.

Regarding the baseline models, we compare both with the best approach identified by Rizzi et al. (2024a) (i.e., G-minimum) and with widely adopted state-of-the-art AI models: mBERT (Kenton and Toutanova, 2019), Llama-2 (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Llama-3.2 (Dubey et al., 2024), and Phi-3.5 (Haider et al., 2024). In particular, the approach proposed by Rizzi et al. (2024a) comprises a technique for identifying disagreement-related textual constituents and an approach for generalizing towards unseen textual constituents. Additionally, four distinct strategies for identifying disagreement are presented. For what concerns the selected LLMs, instead, they have been fine-tuned using the boolean soft-labels related to the disagreement, adopting the huggingface framework, using default hyperparameters. mBERT (Kenton and Toutanova, 2019) is a well-established and widely recognized transformer-based model trained on more than 100 languages. The use of mBERT allows for results that are easily reproducible without extensive computational power. On the other hand, Llama-based models (Touvron et al., 2023; Dubey et al., 2024) are generative large language models known for their efficiency and scalability. They are designed to handle large-scale language tasks and can be fine-tuned for a variety of classification problems.

Mistral-7B (Jiang et al., 2023) is a further generative language model that is renowned for its efficiency and targeted optimizations. It is designed for high-volume text processing, optimized for multilingual content, and suitable for globalized contexts.

Phi 3.5-mini (Haider et al., 2024) is a lightweight version of the Phi model family that offers robust performance on language tasks while avoiding the high demands of larger models. Its compact structure makes it ideal for constrained environments, with excellent results in multilingual processing and classification.

Each of these models has been proven to be effective on several natural language tasks such as hate speech detection or sentiment analysis. Moreover, a peculiar capability of such models is the ability to process multilingual text and social media content. While all models achieve challenging results on a variety of tasks, the choice among these models usually represents a compromise based on specific requirements of the task, such as the volume of data, the languages involved, and the computational resources available.

Differently from the original Le-Wi-Di challenge (Leonardelli et al., 2023), in this work, disagreement detection is addressed as a binary task, making a comparison with the participants' performances unfeasible. This is mainly motivated by the concerns raised by the organizers (Leonardelli et al., 2023), also recently supported by Rizzi et al. (2024b), where the problem of ranking systems trained on continuous disagreement soft-labels using cross-entropy could be strongly biased by the cross-entropy measure itself. For this reason, we compared the proposed approach with benchmark models.

For what concerns the performance metrics, two main aspects have been considered: (i) prediction capabilities in terms of F1-Measure for both the agreement ($F1^{+}$) and disagreement ($F1^{-}$) labels, together with their average ($F - score$), and (ii) computational requirements in terms of the number of model parameters, RAM, CPU, and GPU. The first evaluation allows for a comparison of the models' capabilities in identifying disagreement among annotators, while the second aspect allows for a comparison of the computational requirements needed to reproduce the whole pipeline (comprehensive of the training phase).

## 5 Results and Discussion

Given the Disagreement Score (DS) of each constituent within a sentence, all the proposed aggregation strategies have been evaluated (i.e., sum, mean, median, and minimum). Table 2 summarizes the results achieved with the best thresholds ($\pi$ and $\psi$) selected through a grid-search approach on the validation set released within the Le-Wi-Di challenge for each dataset. Results are distinguished between agreement $(+)$ and disagreement $(-)$ labels.

A McNemar (McNemar, 1947) test has been adopted to perform a pairwise comparison with each of the proposed approaches (considering a confidence level of 0.95). The McNemar test does not verify if two models have different perfor-

| Approach | ConvAbuse | | | ArMIS | | | HS-Brexit | | | MD-Agreement | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F1^+$ | $F1^-$ | *F1-score* | $F1^+$ | $F1^-$ | *F1-score* | $F1^+$ | $F1^-$ | *F1-score* | $F1^+$ | $F1^-$ | *F1-score* |
| *Sum* | 0.84 | 0.25 | 0.55*$^{\dagger\ddagger\circledast\phi}$ | 0.68 | 0.29 | 0.48* | 0.71 | 0.47 | 0.59 $^{\dagger\ddagger\circledast}$ | 0.50 | 0.69 | 0.59*$^{\dagger\ddagger\circledast\phi}$ |
| *Mean* | 0.80 | 0.36 | 0.58*$^{\dagger\ddagger\circledast\phi}$ | 0.66 | 0.47 | **0.57*** | 0.80 | 0.61 | 0.70 $^{\phi}$ | 0.56 | 0.65 | 0.60*$^{\dagger\ddagger\circledast}$ |
| *Median* | 0.85 | 0.32 | 0.58*$^{\dagger\ddagger\circledast\phi}$ | 0.68 | 0.34 | 0.51* | 0.79 | 0.49 | 0.64$^{\circledast\phi}$ | 0.55 | 0.67 | 0.61*$^{\dagger\ddagger\circledast\phi}$ |
| *Minimum* | 0.86 | $\underline{0.43}$ | **0.65***$^{\dagger\ddagger\circledast}$ | 0.48 | 0.48 | 0.48* | 0.63 | 0.55 | 0.59 *$^{\dagger\ddagger\circledast}$ | 0.48 | 0.71 | 0.60*$^{\dagger\ddagger\circledast\phi}$ |
| *G-Minimum* | 0.85 | 0.33 | 0.59 | 0.59 | 0.48 | 0.54 | 0.84 | $\underline{0.69}$ | **0.77** | 0.54 | 0.64 | 0.59 |
| *mBERT* * | 0.93 | 0.05 | 0.49 | 0.38 | $\underline{0.63}$ | 0.50 | 0.37 | 0.43 | 0.40 | 0.76 | 0.60 | 0.68 |
| *Llama-2-7B* $^{\dagger}$ | 0.92 | 0.13 | 0.53 | 0.71 | 0.37 | 0.54 | 0.84 | 0.63 | 0.74 | 0.59 | $\underline{0.77}$ | **0.68** |
| *Mistral-7B*$^{\ddagger}$ | 0.91 | 0.26 | 0.59 | 0.66 | 0.39 | 0.53 | 0.82 | $\underline{0.69}$ | 0.76 | 0.55 | $\underline{0.77}$ | 0.66 |
| *Llama-3.2-3B*$^{\circledast}$ | 0.92 | 0.17 | 0.54 | 0.67 | 0.25 | 0.46 | 0.85 | 0.59 | 0.72 | 0.59 | 0.75 | 0.67 |
| *Phi-3.5-mini*$^{\phi}$ | 0.89 | 0.23 | 0.56 | 0.67 | 0.35 | 0.51 | 0.71 | 0.36 | 0.54 | 0.53 | 0.64 | 0.59 |

Table 2: Comparison of the different approaches on the test set for disagreement detection. **Bold** denotes the best approach according to the F1-Score, while underline represents the best approach according to the disagreement label. A McNermar test has been performed as a pairwise comparison between the proposed approaches and MBERT (*), Llama-2 ($^{\dagger}$), Mistral ($^{\ddagger}$), Llama-3.2 ($^{\circledast}$) and Phi-3.5 ($^{\phi}$).

| Approach | Parameters | RAM | CPU | GPU |
|---|---|---|---|---|
| *Sum*<br>*Mean*<br>*Median*<br>*Minimum* | 179M | 16 GB | 2-4 CPU cores | Non-necessary |
| *G-Minimum* (Rizzi et al., 2024a) | 179M | 16 GB | 2-4 CPU cores | Non-necessary |
| *mBERT* (Kenton and Toutanova, 2019) | 179M | 16 GB | 4-8 CPU cores | Non-necessary |
| *Llama-2-7B* (Touvron et al., 2023) | 6.74B | 160GB | 6-12 CPU cores* | 100GB |
| *Mistral-7B* (Jiang et al., 2023) | 7.25B | 160GB | 8-16 CPU cores* | 110GB |
| *Llama-3.2-3B* (Dubey et al., 2024) | 3.21B | 90GB | 4-8 CPU cores* | 60GB |
| *Phi-3.5-mini*(Haider et al., 2024) | 3.82B | 90GB | 4-8 CPU cores* | 60GB |

Table 3: Computational requirements of the proposed approaches. Values marked with (*) have been estimated, as the exact information was not provided.

mances, but it tests if there is a significant difference in terms of model prediction by comparing sensitivity and specificity of the two models under analysis.

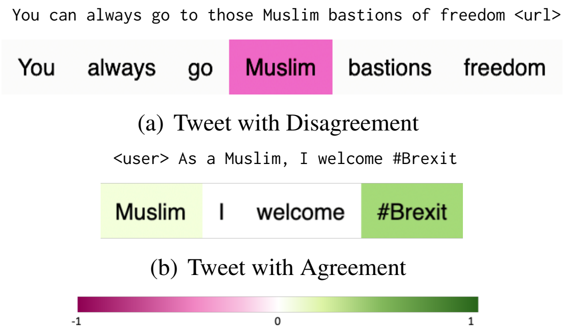
Focusing on the results reported in Table 2, we can observe that all the considered approaches perform better on the majority class, which in general, is related to the complete agreement. Additionally, it is important to note that mBERT is not able to systematically outperform the proposed approach, considering all the aggregation strategies. On the other hand, the proposed approach and the selected LLMs (i.e. Llama-2-7B, Mstral-7B, Llama-3.2-3B, and Phi-3.5-mini) perform in a competitive way: while our strategies work better on ConvAbuse and ArMIS, such models achieve better results on HS-Brexit and MD-Agreement. Although the results of LLMs seem promising on those datasets, such a performance is likely due to the presence of instances on the same topic (e.g., Covid in MD-Agreement, Brexit in HS-Brexit) in the corpora used for training the models. It can be easily noted that the selected LLMs perform worst on the two datasets that are characterized by the underlying lexicon (e.g., ArMIS contains misogynous tweets) or by the type of expressions (e.g., ConvAbuse contains user-bot interactions). An additional consideration relates to the difference in terms of model predictions evaluated through the McNemar test. Although the selected LLMs achieve higher values of F1 score in HS-Brexit, the statistical test shows that in those cases, the behavior of our best approach is analogous and does not highlight any difference in terms of model prediction. On the other hand, on ConvAbuse, our approach outperforms state-of-the-art LLMs, and the statistical test corroborates the hypothesis that our predictions are significantly different. A final consideration refers to the performances achieved on MD-Agreement. As highlighted by Rizzi et al. (2024a), one challenging aspect of the dataset is the inclusion of three main macro-topics of discussion. While the proposed approach performs, on MD-Agreement, poorly with respect to state-of-the-art LLMs, it introduces an improvement in performance with respect to G-Minimum. The primary reason for such behavior seems to be the variety of arguments covered by the dataset, indicating that disagreement may stem not only from differing beliefs or backgrounds but also from the specific topics being discussed.

To provide a complete overview of the models, we report in Table 3 their computational requirements[3]. It can be easily noted that the proposed approach should be preferred: while the number of parameters and RAM are comparable with mBERT, it requires fewer CPU cores. Furthermore, when comparing our approach with the selected LLMs, the necessary resources clearly appear advantageous. Considering both the achieved performances and the computational requirements, we can affirm that simpler models represent a promising alternative to mBERT and other widely adopted LLMs.

A further relevant aspect relates to the usage of the models to highlight disagreement constituents during the annotation phase in the crowdsourcing platforms. While mBERT and the other analyzed LLMs can straightforwardly underline which constituents contribute more to predict disagreement, also the presented approach can be exploited for such a task.

For instance, Integrated Gradients can be used for the identification of such terms from Large Language Models. For the proposed approach, the constituent score can be exploited to evaluate the relationship of each constituent, within the context in which it appears, and the disagreement between annotators (on the hate task).

Figure 1 reports a visual representation of Disagreement Scores (DS) computed for two non-hateful tweets of the Brexit dataset. The first example (Figure 1 (a)) reports a tweet with disagreement, while the second one (Figure 1 (b)) denotes a tweet with agreement. According to the DS score, the proposed approach highlights the word “*Muslim*” as strongly related to disagreement in the first tweet and to agreement in the second one, highlighting the capability to evaluate the constituent with respect to its context. It is important to note that the word “Muslim” was intentionally used by the creators of the dataset as a seed word because it is considered a source of disagreement. The reported example confirms that such a word is correctly identified, according to the context where it appears, as a source of agreement or disagreement in an agnostic way. In fact, the reported tweets - “*As a Muslim , I welcome #Brexit*” and “*You can always go to those Muslim bastions of freedom*” - strongly differ on the connotation of the term “Muslim”. In the first tweet, the term is used as a self-identifier, to identify the religious affiliation of the person expressing a personal opinion on a political issue. The focus is on the individual’s religious identity, and the statement implies that despite being a Muslim, the person supports Brexit. The connotation here is neutral and merely serves to highlight the diversity of opinions within the Muslim community. In the second tweet instead, the term carries a negative connotation since it is used in a stereotypical and possibly derogatory manner. The phrase “*Muslim bastions of freedom*” could be interpreted as sarcastic or mocking, implying that there is a perception that Muslim-majority areas or countries are not associated with freedom.


You can always go to those Muslim bastions of freedom <url>
You
always
go
Muslim
bastions
freedom
<user> As a Muslim, I welcome #Brexit
Muslim
I
welcome
#Brexit
-1
0
1

(a) Tweet with Disagreement

(b) Tweet with Agreement

Figure 1: Visual representation of disagreement scores on sentences from the Brexit dataset. Positive values are represented with green and negative values are represented with pink. The white color is used for constituents with DS values equal or close to zero.

Finally, a more extensive qualitative analysis of the salient constituents for the different datasets has been conducted. Since our approach is based on a contextualized representation of constituents, where the same word can have multiple embed-

[3] The values reported within this table have been estimated according to (Kim et al., 2024) and with the information released by the authors both in the corresponding papers and in the official Hugging Face model-card. All the reported values refer to the original model and do not consider further optimization techniques that might reduce computational requirements at the expense of reduced recognition performance.

ding representations according to its context, we computed the top-scoring words (per dataset) as follows:

- we considered all the scores for each constituent according to its context,
- we computed the percentage of positive and negative scores for each constituent,
- we sorted the estimated percentage to identify the top-k constituents.

| Agreement Constituents | Disagreement Constituents |
|---|---|
| nerdy | compatriots |
| sleepy | throw |
| intelligence | flows |
| greenhouse | reverse |
| sure | Sanders |

Table 4: Top-5 agreement and disagreement constituents for the ConvAbuse dataset

| Agreement Constituents | Disagreement Constituents |
|---|---|
| vote | Obama |
| we | #EURO2016 |
| Duch | #Trump2016 |
| Cameron | France |
| immigrant | invasion |

Table 5: Top-5 agreement and disagreement constituents for the HS-Brexit dataset.

| Agreement Constituents | Disagreement Constituents |
|---|---|
| #blacklivesmatter | Covid |
| Thank | police |
| UK | coronavirus |
| neck | vote |
| blah | President |

Table 6: Top-5 agreement and disagreement constituents for the MD Agreement dataset

| Agreement Constituents | Disagreement Constituents |
|---|---|
| ناقصات (deficiencies) | متسلطات (bossy) |
| المرأة (woman) | صور (photo) |
| دين (religion) | عوانس (spinsters) |
| شوارع (streets) | حلال (halal) |
| النساء (women) | التعامل (dealing) |

Table 7: Top-5 agreement and disagreement constituents for the ArMIS dataset

Tables 4, 5, 6, and 7 list the top-5 agreement and disagreement constituents for each dataset. The elements that show the highest agreement scores are rarely associated with different perceptions, being used frequently in sentences where annotators show a full agreement, while the ones with high disagreement scores are often a proxy of different perspectives.

Since the main goal of this paper is to show the relationship between constituent scores and the agreement/disagreement label, according to the obtained results, the estimated constituent scores can be considered promising because acting as a good proxy of agreement/disagreement. While we acknowledge the potential benefits of post-hoc human evaluation, implementing such a strategy is impractical due to the impossibility of reproducing the exact conditions of the original annotation process. Even by adhering to the dataset creators' approach, obtaining the same annotators is basically not possible (in most cases, anonymous annotators have been involved through crowd-sourcing platforms). On the other hand, introducing additional annotators would imply increasing the variability of potential perspectives without guaranteeing any adherence to the initial annotation and, therefore, to the constituent perception of the original annotators of the dataset.

The proposed solution, contrary to what has been formerly presented in the literature, is able not only to predict if a text can lead to disagreement from different readers' perspectives but also calls attention to those *disagreement-related constituents* in hateful content.

## 6 Conclusions and Future Work

This paper introduces a simple approach for the identification of disagreement-related constituents within the text and exploits them in the prediction of disagreement in hateful texts. By leveraging weighted probabilities, the proposed methodology allows the identification of constituents that not only represent valuable information for a comprehensive understanding of the sources of disagreement within the text but also serve as the foundation for developing an explainable strategy for disagreement detection. The proposed strategies demonstrate a good trade-off between prediction capabilities and computational requirements compared both with G-minimum (Rizzi et al., 2024a) and with well-known state-of-the-art language models:

mBERT, Llama-2, Mistral, Llama-3, and Phi-3.

Future works will consider the adoption of indexing or clustering techniques to reduce the search space of the most similar embeddings by narrowing down the candidates for similarity comparison, resulting in an improvement in efficiency. Moreover, future works might focus on the extension of the proposed approach for the quantification of the level of disagreement in a sentence. Finally, considering the potential of highlighting disagreement-related tokens in the labeling phase, a relevant aspect that will be considered relates to the creation of datasets that include annotators' perceptions at the constituent level.

## Limitations

The proposed approach holds significant promise to improve our comprehension of textual constituents related to disagreement, both in theoretical and practical contexts. By enabling the identification of these constituents, the method contributes to a deeper comprehension of disagreement dynamics within the text. However, it is crucial to acknowledge a current limitation associated with its computational complexity. The comparison within each contextualized constituent representation and every known contextualized constituent represents a significant computational burden, making the approach computationally expensive and difficult to scale. In particular, the time complexity for the computation of the DS scores for a given sentence[4] is $O(n{*}m{*}\textit{time complexity of similarity measure})$, where $n$ represents the number of contextualized constituents in the training data and $m$ the number of contextualized constituents in the given sentence. In our case, the adopted similarity measure is the Cosine similarity that has a time complexity of $O(d)$ where $d$ represents the dimension of the vector to compare. Therefore the overall time complexity is $O(n * m * d)$. This constraint highlights the need for future improvements to improve efficiency while retaining the method's significant insights into textual conflict.

## Ethical Statement

In this research work, we used datasets from the recent literature, and we did not use or infer any sensitive information. The risk of possible abuse of the models and the proposed approach is low.

## Experimental Settings and Setup

We ran the experiments of the proposed methodology on a machine equipped with one Nvidia Testa T4 GPU, CUDA v11.4, 256GB RAM, 2 CPU Xeon Gold. The selected state-of-the-art baselines include generative LLMs. While mBERT has been fine-tuned for the classification task by concatenating a final classification layer, generative LLMs have been instruction-tuned to adapt their generative capabilities for the specific classification task. Further details, along with the code for the reproducibility of the results, are available in the GitHub repository. Regarding mBERT, we used `bert-base-multilingual-cased`.For what concerns the LLMs considered as baselines, we adopted the following: `Llama-2-7b-chat-hf`, `Mistral-7B-Instruct-v0.3`, `Llama-3.2-3B-Instruct`, and `Phi-3.5-mini-instruct`.

## Best Hyperparameters Configurations

The optimal hyperparameter configurations are reported in Table 8.

| | **HS-Brexit** | | **ConvAbuse** | | **MD-Agreement** | | **arMIS** | |
|---|---|---|---|---|---|---|---|---|
| | $\psi$ | $\pi$ | $\psi$ | $\pi$ | $\psi$ | $\pi$ | $\psi$ | $\pi$ |
| **Sum** | 0.85 | 0.5 | 0.95 | 2.2 | 0.7 | 1.5 | 0.85 | 1.30 |
| **Mean** | 0.6 | 0.3 | 0.85 | 0.4 | 0.7 | 0.2 | 0.85 | 0.2 |
| **Median** | 0.7 | 0.3 | 0.7 | 0.6 | 0.75 | 0.4 | 0.8 | 0.3 |
| **Minimum** | 0.7 | -0.4 | 0.5 | 0.5 | 0.7 | -0.4 | 0.85 | -0.4 |

Table 8: Optimal Hyper-parameter Settings

## Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. The work of Elisabetta Fersini has been partially funded by MUR under the grant ReGAInS, Dipartimenti di Eccellenza 2023-2027 of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca. The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project, Grant PID2021-124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU.

[4] The time complexity estimation doesn't consider the time complexity necessary to compute the latent representation of the constituents.

## References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *Preprint*, arXiv:2106.15896.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Alessandro Astorino, Giulia Rizzi, and Elisabetta Fersini. 2023. Integrated gradients as proxy of disagreement in hateful content. In *CEUR Workshop Proceedings*, volume 3596. CEUR-WS. org.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angel Felipe Magnossão de Paula, Giulia Rizzi, Elisabetta Fersini, and Damiano Spina. 2023. Ai-upv at exist 2023–sexism characterization using large language models under the learning with disagreements regime. In *CEUR Workshop Proceedings*, volume 3497, pages 985–999. CEUR-WS.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anca Dumitrache, FD Mediagroep, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of NAACL-HLT*, pages 2164–2170.

Johan Erbani, Előd Egyed-Zsigmond, Diana Nurbakova, and Pierre-Edouard Portier. 2023. When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media with pretrained transformers in a soft label context. *Working Notes of CLEF*.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ewelina Gajewska. 2023. eevvgg at SemEval-2023 task 11: Offensive language classification with rater-based information. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 171–176, Toronto, Canada. Association for Computational Linguistics.

Richard Gonzalez and George Wu. 1999. On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166.

Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, Jianwen Zhang, et al. 2024. Phi-3 safety post-training: Aligning language models with a" break-fix" cycle. *arXiv preprint arXiv:2407.13833*.

Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th international conference on web search and data mining*, pages 241–249.

Andrew Hoskins and John Tulloch. 2018. The construction of hate in online spaces. *International Journal of Communication*, 12:3853–3873.

Leonie Huddy and Lene Aarøe. 2019. The subjectivity of hate speech detection: A study on perceptions of offensiveness. *Political Psychology*, 40(1):3–29.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Taeho Kim, Yanming Wang, Vatshank Chaturvedi, Lokesh Gupta, Seyeon Kim, Yongin Kwon, and Sangtae Ha. 2024. Llmem: Estimating gpu memory usage for fine-tuning pre-trained llms. *arXiv preprint arXiv:2404.10933*.

Marianne D LaFrance and Sarah J Roberts. 2019. The role of bias in hate speech detection. *Journal of Language Aggression and Conflict*, 7(1):1–20.

Liliya Lavitas, Olivia Redfield, Allen Lee, Daniel Fletcher, Matthias Eck, and Sunil Janardhanan. 2021. Annotation quality framework-accuracy, credibility, and consistency. In *NEURIPS 2021 Workshop for Data Centric AI*, volume 3.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *Preprint*, arXiv:2304.14803.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Martina Nardon and Paolo Pianca. 2015. Probability weighting functions. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, 29.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Drazen Prelec. 1998. The probability weighting function. *Econometrica*, 66(3):497–527.

Giulia Rizzi, Alessandro Astorino, Paolo Rosso, and Elisabetta Fersini. 2024a. Unraveling disagreement constituents in hateful speech. In *European Conference on Information Retrieval*, pages 21–29. Springer.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024b. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 84–94.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2420–2433.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. An analysis of annotator disagreement in human interpretations of toxicity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5470–5477.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906.

Sadat Shahriar and Thamar Solorio. 2023. Safewebuh at semeval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation. *arXiv preprint arXiv:2305.01050*.

Michael Sullivan, Mohammed Yasin, and Cassandra L Jacobs. 2023. University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

María Estrella Vallecillo-Rodríguez, FMP del Arco, Luis Alfonso Ureña-López, María Teresa Martín-Valdivia, and Arturo Montejo-Ráez. 2023. Integrating annotator information in transformer fine-tuning for sexism detection. *Working Notes of CLEF*.