Developing Automatic Hate Speech Detection Methods Based on Text Processing

Yuhan Xia

A thesis submitted for the degree of MSC by Dissertation

School of Computer Science and Electronic Engineering

University of Essex

Oct, 2024

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisor, Dr. Yunfei Long, for his support both in my personal and academic life. As an international student at the onset of my studies abroad, adapting to a new country was full of challenges. Dr. Long was instrumental in helping me navigate these challenges, assisting me with various personal issues which significantly eased my transition. Academically, our collaboration has been both enjoyable and fruitful, sharing similar research interests and exploring scientific questions together. This experience has been truly unforgettable and gratifying. I am also very thankful to Qingqing Zhao from the Chinese Academy of Social Sciences for her support of my master's project throughout this year. Her expertise and enthusiastic encouragement greatly enhanced my academic perspective. Additionally, I would like to acknowledge all the friends I have made during this journey. Your companionship and support have been invaluable to my personal growth and academic endeavors. Lastly, my heartfelt thanks to everyone who supported and encouraged me throughout this academic journey. Your encouragement has been crucial in helping me persevere and complete this significant phase of my academic career.

This thesis closes my master's chapter; my scholarly journey has just begun.

Summary

The paper introduces SensoryT5, a novel neuro-cognitive model developed to integrate sensory information into the T5 (Text-to-Text Transfer Transformer) framework for enhanced hate speech detection. This integration represents a pioneering step in natural language processing (NLP) as it combines cognitive science and computational techniques to improve the effectiveness and sensitivity of language models. SensoryT5 enriches the traditional attention mechanism of the T5 model with sensory cues, which helps achieve a more nuanced understanding of language context and sensory data. This approach has demonstrated superior performance over existing models, including both the foundational T5 and other pre-trained or large language models (PLMs/LLMs), as well as conventional machine learning methods. Through rigorous testing across multiple public datasets dedicated to hate speech detection, SensoryT5 has consistently outperformed contemporary benchmarks in the field. The primary contributions of this research are fourfold. First, it proposes a unique architecture that innovatively merges sensory knowledge with contextual attention mechanisms within a transformer-based model. Second, it showcases significant improvements in model performance for detecting hate speech. Third, by embedding sensory information into word representations, the model enhances the interpretability, providing deeper insights into the interplay between emotion and cognition in linguistic processes. Lastly, SensoryT5's use of sensory knowledge not only aids in refining detection capabilities but also promotes a broader interdisciplinary exchange of ideas between neuro-cognitive science and NLP. Overall, this study proposes a more integrated approach that considers both the cognitive underpinnings and technical aspects of language processing to tackle complex tasks like hate speech detection more effectively.

Outline

Abstract

Traditionally, sensory perception and hate speech detection have been viewed as distinct areas within the field of natural language processing (NLP). However, the profound impact of sensory experiences on cognitive and linguistic processes is undeniable. Previous NLP research has often overlooked the potential of integrating sensory knowledge with advanced language models for hate speech detection. Addressing this oversight, we introduce SensoryT5, a neuro-cognitive model that embeds sensory information into the T5 (Text-to-Text Transfer Transformer) framework, specifically tailored for enhanced hate speech detection. This innovative approach enriches the T5's attention mechanism with sensory cues, fostering a nuanced balance between contextual insights and sensory awareness. Through extensive testing across various datasets, SensoryT5 demonstrates superior performance in identifying hate speech, outperforming both the foundational T5 model and other pre-trained language models (PLMs)/large language models (LLMs) as well as traditional machine learning methods. This advancement not only boosts the model’s efficacy but also underscores the importance of neuro-cognitive data in augmenting the linguistic and emotional sensitivity of machine learning models. Our findings mark a significant evolution in NLP research, advocating for a more integrated approach that considers both cognitive science and

computational techniques in tackling complex tasks such as hate speech detection.

Introduction

In the digital age, social media platforms and chat forums have revolutionized how individuals interact and communicate. These platforms offer a unique opportunity for instant and widespread sharing of personal opinions and experiences. However, this open digital space also fosters the dissemination of harmful content, including hate speech—expressions of animosity or disparagement directed at individuals or groups based on protected characteristics such as race, gender, or religion (Nockleby, 1994). The complexity and rapid evolution of online interactions present significant challenges for the detection and moderation of such content (Alkomah & Ma, 2022; Jahan & Oussalah, 2023).

Hate speech not only violates ethical norms and often legal guidelines but also poses severe social risks by fueling discrimination and social divisions. Recent events, such as the online harassment campaigns during various geopolitical conflicts, underscore the urgent need for effective hate speech detection mechanisms. Given the limitations of manual moderation—such as scalability

and timeliness—there is a substantial push towards developing automatic methods for hate speech detection using advances in Natural Language Processing (NLP) and Machine Learning (ML).

The proliferation of digital platforms has been accompanied by an increase in the linguistic and cultural diversity of the content, which complicates the task of hate speech detection. Automatic detection systems must now cope with subtle linguistic cues and cultural contexts to distinguish between harmful speech and benign communication accurately. The challenges are exacerbated by the diverse legal landscapes across different countries, which influence the definition and thresholds of hate speech. Substantial research efforts have been directed at refining computational techniques to tackle this issue. Traditionally, machine learning models such as support vector machines (SVM), Naive Bayes (NB), Logistic Regression (LR), Decision Trees (DTs) and K-Nearest Neighbor (KNN) have been extensively applied to text classification tasks, including hate speech detection (Alkomah & Ma, 2022). These models rely on feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings to represent textual data in a high-dimensional space. For example, SVM has been particularly popular due to its ability to handle high-dimensional feature spaces and its effectiveness in binary classification tasks, such as identifying hate speech versus non-hate speech.

While these traditional approaches offer advantages in terms of interpretability

and relatively fast training times, they face limitations in capturing complex linguistic patterns, context, and semantic nuances (Yin & Zubiaga, 2021). This has led to the rise of more complex models like Convolutional Neural Networks (CNNs) and transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), which can better understand the contextual relationships between words. However, traditional machine learning models remain a key component in the early stages of hate speech detection research and are often used as baselines for comparing the performance of more advanced methods. Researchers have leveraged a variety of data sources to train both traditional and deep learning models, as highlighted in prominent competitions like SemEval-2019 (Zampieri et al., 2019), SemEval-2020 (Zampieri et al., 2020), and GermEval-2018 (Wiegand et al., 2018), which have significantly advanced the field.

In recent decades, various sensory resources from diverse sources and modalities, such as physical signals (e.g., speech and facial expression) and physiological signals (e.g., electroencephalogram, electrocardiogram, galvanic skin response, and eye tracking), have been extensively used for automatic emotion recognition, which have improved the models for emotion recognition greatly(Fan et al., 2023; Rodriguez et al., 2022; Skaramagkas et al., 2023; Tuncer et al., 2021; Zhong et al., 2022). Recently, Xia et al. (2024) have achieved excellent results in emotion classification by integrating sensory

lexical fusion into the T5 model.

Numerous studies have approached hate speech detection by incorporating techniques from sentiment analysis, reflecting the shared importance of emotional assessments in both fields (Markov et al., 2021; Plaza-Del-Arco et al., 2021; Rana & Jha, 2022). In this light, sentiment analysis does not merely serve to detect feelings but becomes a tool to gauge the intensity and harm potential of online interactions. Some researchers even categorize hate speech detection as a specialized branch of sentiment analysis focused on identifying negative sentiments that could lead to real-world harm (Ali et al., 2021). Inspired by this integrated perspective, this project is founded on the hypothesis that sensory inputs, which significantly aid emotion classification, can similarly enhance hate speech detection. Sensory information enriches the processing capabilities of NLP models by providing them with deeper contextual cues about the emotional state conveyed in text. Thus, by applying these insights, our SensoryT5 model aims to not only identify but also understand the complex emotional layers that underpin hate speech. This exploration is poised to advance our ability to handle the nuanced demands of online communication. Through integrating detailed sensory data into the T5 architecture, we enhance its capacity to discern subtle emotional nuances, thereby improving the precision and reliability of detecting hate speech. Ultimately, this project underscores the value of combining cognitive and sensory data to enrich NLP

applications, pushing the boundaries of what our current technologies can achieve in monitoring and moderating online spaces. This project aims to explore the relationship between sensory input and hate speech based on the assistance sensory input provides in emotion classification tasks. This exploration seeks to enhance the task of hate speech detection. The main contributions of our work can be summarized as follows:

(1) We introduce a novel architecture, SensoryT5, which advances the transformer-based model for hate speech detection by incorporating sensory knowledge. This pioneering effort adapts SensoryT5 to integrate the subtleties of contextual attention with sensory information-based attention seamlessly.

(2) Our experiments across several publicly available datasets for hate speech detection illustrate that our model significantly enhances the performance of existing models and consistently surpasses the contemporary state-of-the-art benchmarks.

(3) Beyond the improved performance and consistency of our SensoryT5 model compared to baseline models, embedding sensory knowledge into word representations augments the interpretability of the findings. This enhancement not only aids in the explainability of the model and its applications but also deepens our comprehension of the complex interplay

between emotion and cognition in human behaviors.

(4) SensoryT5 leverages sensory knowledge within transformer text classification frameworks, contributing to the ongoing efforts to incorporate neuro-cognitive data in NLP tasks. That is, our work illuminates the potential of sensory information in refining hate speech detection, carving fresh prospects for exploration within the realm in NLP. In addition, this study underscores the value of cognition-anchored resources in sculpting attention models, which also encourages continued interdisciplinary dialogue and research between the domains of NLP and neuro-cognitive science.

Related work

Hate speech

Understanding hate speech within the framework of text processing involves grappling with its inherently nebulous boundaries. While the precise definition of hate speech remains a contentious issue without universal consensus, its recognition can significantly streamline annotation processes, thereby enhancing the reliability of hate speech detection systems. This is supported by researchers like Ross et al., who argue that a well-articulated definition could simplify the task of identifying hate speech in textual data (Ross et al., 2016). On the other hand, distinguishing hate speech from permissible free speech often presents a dilemma due to the overlapping nuances and the potential for infringing on freedom of expression. For example, the American Bar Association refrains from endorsing a definitive description of hate speech, suggesting instead that any speech contributing to a criminal offense may be judged under hate crime statutes (Wermiel, 2017).

Consequently, this thesis does not attempt to establish a fixed definition of hate speech. Rather, it explores a variety of definitions put forward by different entities and authors to better understand the common characteristics attributed to hate speech and the various challenges these definitions pose to its detection. This approach not only aids in delineating what constitutes hate speech but also

highlights the complexity and the critical need for nuanced analysis in its identification, reflecting the intricate interplay between language and intergroup dynamics.

1. **Encyclopedia of the American Constitution (Nockleby et al., 2000):** "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity."

2. **Code of Conduct between European Union Commission and companies (Wigand & Voin, 2017):** "All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic."

3. **International minorities associations (ILGA)**[1]**:** "Hate crime is any form of crime targeting people because of their actual or perceived belonging to a particular group. The crimes can manifest in a variety of forms: physical and psychological intimidation, blackmail, property damage, aggression and violence, rape."

4. **Facebook**[2]**:** "We define hate speech as a direct attack against people on

[1] https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech

[2] https://www.facebook.com/communitystandards/hate_speech

the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic."

5. **Twitter**[3]**:** "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."

6. **YouTube**[4]**:** "We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race,

---

[3] https://help.twitter.com/en/rules-and-policies/twitter-rules#hateful-conduct

[4] https://support.google.com/youtube/answer/2801939?hl=en

immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, and veteran Status.”

7. **Academic Perspectives:**

Nobata et al. (2016) view hate speech as acts that "attack or demean a group or individual based on attributes such as race, ethnic origin, religion, disability, gender, age, or sexual orientation". This definition underscores the targeted nature of hate speech, emphasizing the diversity of potential victim categories.

Nockleby (1994) offers a broader definition, describing hate speech as "any communication that disparages a person or a group on the basis of characteristics like race, color, ethnicity, gender, sexual orientation, nationality, religion, or other similar traits". This encapsulates a wider array of discriminatory communications.

Warner & Hirschberg (2012) differentiate hate speech by the speaker's intent to harm, associating it with the presence of language that could incite prejudice or violence.

Waseem & Hovy (2016) specifically categorize racist and sexist remarks as key examples of hate speech, highlighting the common contexts in which such speech occurs.

Waseem & Hovy (2016) define it as "a deliberate attack directed towards a

specific group, motivated by aspects of the group’s identity", pointing out the premeditated nature of such acts.

Fortuna & Nunes (2019) describe hate speech as language that "attacks or diminishes, that incites violence or hate against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, among others". They note the potential subtlety of hate speech, which can even include humor or indirect language.

In exploring the multifaceted definitions of hate speech presented by various entities ranging from constitutional documents to social media policies and scholarly research, it is clear that hate speech encompasses a wide array of expressions driven by prejudice, discrimination, or antagonism. Each definition highlights different aspects of hate speech, from intentional attacks to subtle insinuations, reflecting the complex nature of this social phenomenon.

Given this complexity, the challenge of automatically detecting hate speech in text is significant. The varying definitions imply that any detection system must be highly sophisticated and adaptable, capable of understanding context, nuance, and the subtleties of language that differentiate hate speech from other forms of expression. This is not just a technical challenge

but also a societal imperative, as the impact of hate speech on individuals and communities can be profound.

Automatic hate speech detection

The exploration of automatic hate speech detection in textual data is a crucial aspect of modern computational linguistics and artificial intelligence, aimed at mitigating the pervasive issue of online hate. This section reviews various methodologies and frameworks developed for automatic hate speech detection, emphasizing the evolution from simple lexical approaches to sophisticated machine learning and deep learning models.

A modest share of the studies, about 12%, employ traditional machine learning techniques centered around the Term Frequency-Inverse Document Frequency (TFIDF) approach (Alkomah & Ma, 2022). The TFIDF methodology emphasizes terms that are uncommon in the overall corpus but prevalent in hate speech, thereby aiding in its detection. Significant implementations in this approach include using TFIDF in conjunction with various natural language processing tools to refine detection capabilities, as demonstrated in works incorporating character or word IDF methods (Saha et al., 2018), sophisticated embedding techniques (de Andrade & Gonçalves, 2021), and multiple linguistic features (Davidson et al., 2017; Martins et al., 2018; Nobata et al., 2016; Ombui et al., 2019; Watanabe et al., 2018). Lexicon-based strategies utilize predefined sets

of keywords compiled from scholarly literature or expert recommendations to filter hate speech. An example of effective application is the integration of a hate speech-specific lexicon with SVM (Support Vector Machines) classifiers, which has proven effective in identifying misogynistic content online (Frenda et al., 2019). This method's precision is enhanced by combining traditional TFIDF values with lexicon-based features, increasing the accuracy of predictions (Bauwelinck et al., 2019, p. 5; Chakrabarty, 2020; Davidson et al., 2017; Indurthi et al., 2019; Srivastava & Sharma, 2020).

Beyond lexicon-based approaches, traditional machine learning models have been widely studied. These models include character n-gram Logistic Regression (Gröndahl et al., 2018), Support Vector Machines (Fortuna, Soler-Company & Wanner, 2021; Pamungkas & Patti, 2019; Pamungkas, Basile & Patti, 2020), and shallow networks with pre-trained embeddings, such as MLP with Byte-Pair Encoding (BPE)-based subword embeddings (Heinzerling & Strube, 2018). But these simpler models generally do not perform as well as deep neural networks.

Hate speech detection in NLP has seen substantial evolution, transitioning from basic rule-based methods to sophisticated pre-trained language models (PLMs) and large language models (LLMs). Initially, research in this area largely focused on feature learning through neural networks such as Convolutional Neural Networks (CNN) (Gambäck & Sikdar, 2017), Recurrent Neural Networks

(RNN) (Saksesi et al., 2018), and Long Short-Term Memory networks (LSTM) (Bisht et al., 2020), which primarily addressed syntactic parsing.

Over recent years, the development of PLMs and LLMs like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and ChatGPT (OpenAI et al., 2024) has marked a significant leap forward. These models have been pre-trained on extensive text corpora using self-supervised learning techniques, enabling them to generate complex representations autonomously. This capability has greatly enhanced their performance in various NLP tasks, notably in hate speech detection (Jahan et al., 2024; Y. Jin et al., 2024; Kikkisetti et al., 2024; Shi et al., 2024; Zhang et al., 2024).

Among these, the T5 model is particularly notable for its unique text-to-text transfer methodology, where every NLP task, including hate speech detection, is reformulated as a text-to-text problem. This innovative approach has proven effective in identifying and categorizing hate speech, setting new standards for accuracy and efficiency in the field.

Despite the significant progress made with Pre-trained Language Models (PLMs) and Large Language Models (LLMs), several persistent research gaps remain. Notably, while these models deploy advanced neural architectures capable of parsing extensive text datasets to identify patterns, they often

overlook the potential benefits of integrating cognitive resources to enhance performance. However, recent studies suggest that incorporating sensory resources with LLMs could markedly improve their effectiveness, propelling them towards achieving a level of comprehension comparable to human-like understanding (Khare et al., 2024).

For example, research by Long et al. demonstrated that integrating eye-tracking data significantly improves performance in sentiment analysis, a field closely related to hate speech detection (Long et al., 2019). Likewise, findings by Yan et al. (2024) show that combining electroencephalogram signals and eye-tracking data can greatly enhance Automatic Keyphrase Extraction (AKE) from microblogs, suggesting the vast potential of multimodal data integration. Additionally, the evolution of models like ChatGPT to include capabilities such as voice and image processing represents a significant advancement in making human-machine interactions more intuitive and context-rich. This shift not only facilitates more complex and nuanced interactions but also moves the focus from purely cognitive tasks to those involving emotional and subjective understanding.

Building on these advancements, integrating LLMs with cognitive architectures and sensory inputs emerges as a promising strategy for improving hate speech detection. By constructing and dynamically updating cognitive models that understand complex information about entities and their interactions, such

integration can significantly enhance the ability to detect and interpret hate speech. These developments are essential for creating more sophisticated and effective computational tools that can identify and mitigate hate speech across various communication platforms (Romero et al., 2023).

This study advocates for the integration of sensory resources and cognitive frameworks into large language models such as T5 to significantly boost their hate speech detection capabilities, marking a pivotal advancement in the field. Our SensoryT5 model exemplifies this approach by merging the robust features of T5 with embedded sensory knowledge, aiming for a deeper and more nuanced understanding of hate speech.

By enriching the T5 architecture with sensory insights, SensoryT5 is specifically designed to improve the interpretation of the complex interplay between language and human emotions that often characterizes hate speech. This nuanced understanding is critical because hate speech frequently intertwines overt expressions of animosity with subtler tones of aggression and bias, mirroring the emotional complexity found in typical human interactions.

Moreover, the relationship between sentiment analysis and hate speech detection is intricate and mutually reinforcing. Sentiment analysis, traditionally focused on identifying the emotional tone of a text, provides foundational techniques that are crucial for detecting not just the presence of emotions, but

also their nature—whether they are likely to incite or reflect negative behaviors such as hate speech. By applying sentiment analysis methodologies, researchers can better distinguish between harmful speech and innocuous communication, even when they are contextually similar.

Therefore, by incorporating sensory data—which can provide deeper emotional context—into the analysis process, SensoryT5 offers enhanced capabilities for identifying the often subtle and context-dependent cues that signal hate speech. This innovative integration positions SensoryT5 at the forefront of current efforts to combat hate speech, providing a powerful tool that leverages both cognitive and sensory data to improve detection accuracy and deepen our understanding of the underlying dynamics of hateful discourse.

Sensory resources: Lancaster norms

In recent years, there has been an emergent trend that neuro-cognitive data and computational approaches are synergized in NLP studies. The interdisciplinary synergy unlocks new dimensions in understanding language, perception, and cognition for human beings.

Most of the studies focus on using neuro-cognitive data for metaphor detection. For instance, Chen et al. (2021) incorporated the brain measurement data for modeling word embedding to identify metaphorical usages. Wan et al. (2023) demonstrated the superiority of neural networks for metaphor detection by

leveraging sensorimotor knowledge. These studies collectively underscore a broader shift in the field towards a more integrated approach to NLP. By weaving in neuro-cognitive data, researchers are equipping computational models with a richer and more intricate understanding of language and cognition, which are often overlooked by traditional data-driven methods.

This study utilizes Lynott et al. (2020)'s sensorimotor norms which encompass the metrics of sensorimotor strengths (ranging from 0 to 5) of 39,707 English words spanning six perceptual domains including touch, taste, smell, vision, hearing, and interoception, as well as five action effectors including mouth/throat, hand/arm, foot/leg, head (excluding mouth/throat), and torso. Lynott et al. (2020)'s sensorimotor norms (named “Lancaster norms" hereinafter) were compiled by following the sensory rating task proposed by (Lynott & Connell, 2009, 2013), which asked participants to rate the extent to which the meaning of a lexical item is based on sensory perceptions through the six sensory modalities and the five action effectors. Thus, the norms are language-specific lexical properties representing the correlation between conceptualized lexical meanings and sensory modalities/action effectors. As there is little work reported to show the correlation between emotion and sensory action effectors, this study only exploits the perceptual data in the six sensory modalities. Table 1 shows the perceptual ratings of six sample words in Lancaster norms.

Words	Touch	Taste	Smell	Vision	Hearing	Interoception
soft	4.526	0.368	0.316	1.947	0.684	0.842
flavor	0.219	4.938	1.719	0.344	0.094	1.094
incense	1.412	0.294	5.000	2.824	0.176	0.295
blue	0.150	0.000	0.000	4.450	0.250	0.500
noisy	0.244	0.080	0.090	1.006	4.752	0.920
headache	0.650	0.000	0.000	0.400	0.400	4.900

Table 1: The sensory ratings of six sample words in the Lancaster norms.

This study unveils SensoryT5, a model engineered to construct sensory vectors using the Lancaster norms, effectively enhancing the hate speech detection process. These vectors are seamlessly incorporated into T5's decoder mechanism via an auxiliary attention layer specifically designed for this purpose. Positioned strategically after the decoder, this sensory-centric attention layer works in concert with the decoder's output to forge a comprehensive representation imbued with deep sensory knowledge of the words in the text.

Consequently, SensoryT5 is adept at simultaneously processing contextual cues and sensory information, enabling a powerful convergence of sensory insights with contextual awareness. This innovative integration significantly bolsters the model's ability in detecting hate speech by aligning sensory nuances with the linguistic context. The enhanced capability of SensoryT5 to interpret these complex layers of information not only improves detection accuracy but also aids in understanding the subtleties of hate speech dynamics.

This model promises to advance the field by providing a more nuanced and effective approach to identifying and analyzing hate speech in diverse textual environments.

Our SensoryT5 model

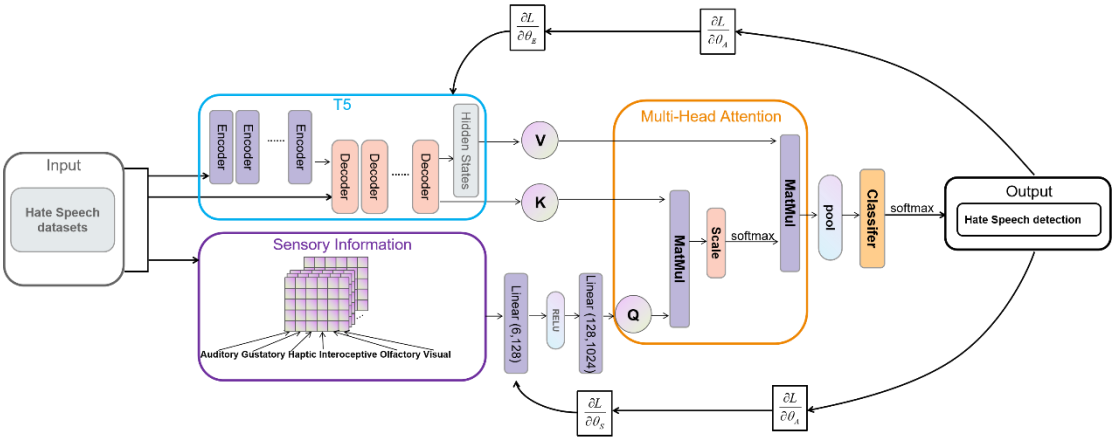

Figure 1: An overview of SensoryT5. The blue box shows a T5 process of deep learning, and the purple box describes sensory information quantified and passed into T5.

The proposed SensoryT5 model, depicted in Figure 1, advances the state-of-the-art in natural language processing by embedding sensory knowledge directly into the architecture of T5 (Raffel et al., 2020). This integration is achieved using an innovative adapter approach, which incorporates attention mechanisms specifically designed to process and synthesize sensory and

contextual information. This fusion is facilitated by a unified loss function that enables joint training, optimizing the model to effectively handle a diverse array of sensory inputs alongside traditional text data.

Choosing the T5 model as our foundation was strategic, reflecting our broader vision for the application of sensory integration across various NLP domains, not limited to hate speech detection. The versatility of T5's text-to-text transfer methodology makes it an ideal candidate for a wide range of NLP tasks including, but not limited to, question-answering, text generation, and more. This approach aligns with the overarching goal of achieving more human-like artificial intelligence by enhancing models' abilities to interpret and generate text that reflects a deeper understanding of human sensory experiences (Duéñez-Guzmán et al., 2023; Khanam et al., 2019).

T5's unique design, which converts every NLP problem into a text-to-text format, provides the flexibility needed to incorporate sensory data effectively. This structural adaptability allows SensoryT5 to excel not only in tasks directly involving sensory data like hate speech detection but also in areas where enriching text with sensory details can be beneficial. For instance: Question-Answering: SensoryT5 can improve the quality of responses in question-answering systems by incorporating sensory nuances that make the answers more detailed and contextually relevant, thus providing a richer user experience. Text Generation: In text generation tasks, the ability of SensoryT5 to integrate

sensory information allows it to produce text outputs that are not only grammatically and contextually accurate but also imbued with sensory details that enhance the narrative depth and emotional resonance of the content. This capability makes it particularly valuable for applications in creative writing, marketing, and other fields where engaging and vivid textual content is crucial. Enhanced User Interaction: By processing sensory inputs, SensoryT5 can better understand and generate language that reflects human-like senses and experiences, thereby improving the interaction between AI systems and users in applications such as virtual assistants and interactive chatbots. Ultimately, the integration of sensory information within the T5 framework through SensoryT5 represents a significant step towards imbuing machines with a more sophisticated understanding of the world as experienced through human senses. This development not only enhances the model's performance across a variety of tasks but also contributes to the broader field of AI by pushing the boundaries of what it means for machines to understand and generate human-like text.

This expanded functionality demonstrates the potential of SensoryT5 to adapt and thrive in a multitude of settings, driving forward the possibilities for more nuanced and intelligent systems in natural language processing and beyond.

Preliminaries

Despite the large size of Lancaster norms, there are still out-of-vocabulary words. Following the method proposed by Li et al. (2017), we use a word embedding model to regressively predict the sensory values of unknown words, aiming to obtain the sensory values for the out-of-vocabulary words.

The objective of hate speech detection is to determine and categorize hate for a piece of text following a defined label schema. Let D denote a collection of documents for hate speech detection. Each document $d \in D$ is first tokenized into a word sequence with the maximum length n, then the word embeddings w_i of these sequences are jointly employed to represent the document $d = w_1, w_2, \ldots, w_i, \ldots, w_n (i \in 1,2,\ldots,n)$.

The core attention mechanism in T5

The word embeddings of these sequences $d = w_1, w_2, \ldots, w_i, \ldots, w_n (i \in 1,2,\ldots,n)$ first enter the T5 model. Each layer of the encoder and decoder has a series of multi-head attention units. The multi-head attention mechanism for the final decoder layer can be represented using the following equation:

$$\begin{aligned} V_d &= \mathrm{MultiHead}(Q_0, K_0, V_0) \\ &= [\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head}_i] W_O \end{aligned} \tag{1}$$

Where each head is computed as:

$$\begin{aligned} \text{head}_i &= \text{Attention}(Q_0 W_i^Q, K_0 W_i^K, V_0 W_i^V) \\ &= \text{softmax}\left(\frac{(Q_0 W_i^Q)(K_0 W_i^K)^T}{\sqrt{d_k}}\right) V_0 W_i^V \end{aligned} \quad (2)$$

W_i^Q, W_i^K, and W_i^V are weight matrices that are learned during the training process. They are used to project the input queries (Q), keys (K), and values (V) to different sub-spaces.(Q_0), (K_0), and (V_0) are derived from the output of the penultimate decoder layer. Additionally, following the common practice for text classification with the T5 model, we employ a zero-padding vector as the sole input for the decoder.

The result (V_d) is the output of the T5 decoder, imbued with context-aware attention. Both (V_d) and (K_0) will be utilized for the integration with sensory knowledge.

Sensory information transformation for T5 integration

We project the perceptual ratings of words in Lancaster norms and the predicted sensory values of the out-of-vocabulary words into a word vector space. Each word is linked with a six-dimensional vector representing sensory scores across six perceptual modalities (haptic, gustatory, olfactory, visual, auditory, and interoceptive dimensions). For a word w, its sensory vector is denoted as $s(w) = [s_1, s_2, \ldots, s_6]$.

To enable effective integration into the T5 model, we use two linear

transformations followed by a ReLU (Rectified Linear Unit) activation function to map the sensory vectors to the same dimension as the T5's word embeddings. Given a T5 model with an embedding dimension of 1024, the transformation process can be formally described as:

$$\mathbf{h}_1 = \mathrm{ReLU}(\mathbf{W}_1 s(w) + \mathbf{b}_1) \tag{3}$$

$$s'(w) = \mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2 \tag{4}$$

Where $W_1 : R^6 \rightarrow R^{128}$ and $W_2 : R^{128} \rightarrow R^{1024}$ are two linear transformation matrices and b_1 and b_2 are the respective bias terms.The shapes of the two weight matrices W_1 and W_2 are $(6,128)$ and $(128,1024)$ respectively. The output h_1 of the first linear layer is a vector of shape $(1,128)$, and the output $s'(w)$ of the second linear layer is a vector of shape $(1,1024)$. After the transformation, the sensory vector $s'(w)$ is projected into the same semantic space as the features generated by the T5 model. The output vector $s'(w)$ with V_d and K_d from the T5 model will be applied for infusing sensory knowledge into the T5 model.

Sensory attention mechanism in SensoryT5

The sensory vector $s'(w)$ generated by the sensory vector transformation is used as the query in the attention mechanism of the sensory adapter, substituting the query vector Q in the T5 model. The sensory adapter performs

the attention calculation as follows:

$$\begin{aligned} A_d &= \text{MultiHead}(s'(w), K_0, V_d) \\ &= [a_1, a_2, ..., a_i]W_d \end{aligned} \tag{5}$$

where each head is computed as:

$$\begin{aligned} a_i &= \text{Attention}(s'(w)W_i^Q, K_0W_i^K, V_dW_i^V) \\ &= \text{Softmax}\left(\frac{(s'(w)W_i^Q)(K_0W_i^K)^T}{\sqrt{d_k}}\right)V_dW_i^V \end{aligned} \tag{6}$$

Once the output $A_d = a_1, a_2, \ldots, a_n$ of the sensory adapter is obtained, we apply dropout and pooling operations to form a final representation P_d, which is then used as the input to the classification layer.

$$P_d = \text{Dropout}(\text{Pool}(A_d)) \tag{7}$$

The pooled representation P_d is then fed into the classifier of the T5 model.

$$C_d = \text{Softmax}(Linear(\text{Dropout}(P_d))) \tag{8}$$

C_d is a probability distribution vector. The class with the highest probability is selected as the predicted label, denoted as y.

The first step of the back-propagation process involves computing the gradient of the loss function with respect to the parameters of the sensory attention adapter. Θ_A represents the parameters of the sensory attention layer, and A_d represents the output of the sensory T5. The computed gradient is used to

update the parameters of the attention layer, enhancing its capacity to integrate sensory information into the T5 model. This is represented as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathrm{A}}} = \frac{\partial \mathcal{L}}{\partial A_d} \cdot \frac{\partial A_d}{\partial \Theta_{\mathrm{A}}} \tag{9}$$

After the gradients for the sensory attention mechanism have been computed, we then compute the gradients for the parameters of the final layer of T5, denoted as Θ_{E}.

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathrm{E}}} = \frac{\partial \mathcal{L}}{\partial V_d} \cdot \frac{\partial V_d}{\partial \Theta_{\mathrm{E}}} \tag{10}$$

Finally, the gradients for the sensory information transformation, denoted as Θ_{S}, are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathrm{S}}} = \frac{\partial \mathcal{L}}{\partial s'(w)} \cdot \frac{\partial s'(w)}{\partial \Theta_{\mathrm{S}}} \tag{11}$$

Here, Θ_S represents the parameters of the sensory information transformation component, which includes the weights and biases of the two linear layers, and $s'(w)$ represents the output of this component. The calculated gradient is used to update the parameters of the sensory information transformation to improve its ability to capture and model the sensory information.

Through these calculations, we are able to update the parameters of the sensory attention mechanism, the T5 model, and the sensory information transformation component.

Experiment

Datasets

Our study employs the HASOC 2020 and HASOC 2021 datasets, meticulously compiled by Mandl et al., which are designed to benchmark natural language processing techniques for identifying and classifying hate speech and offensive content.

HASOC 2020 Dataset:

The 2020 edition of the dataset comprises 3,708 tweets collected for training purposes and 1,592 for testing. We focus primarily on Task 1 (also known as Task A), which targets the identification and classification of text as either hate speech or offensive. This task was initially categorized using a sophisticated combination of SVM classifiers and extensive human judgment to ensure accuracy and reliability (Mandl et al., 2020). The classification under Task A is crucial for developing algorithms that can effectively discern harmful communication in digital conversations, reflecting the ongoing challenges and complexities of moderating online platforms.

HASOC 2021 Dataset:

Following the structure of its predecessor, the HASOC 2021 dataset contains slightly more extensive data, with 3,843 tweets for training and 1,281

designated for testing. Compiled during the COVID-19 pandemic, this dataset uniquely includes tweets that are related to pandemic topics, thereby incorporating the contemporary issues and sentiments prevalent during such a global crisis. As with the previous year, our analysis remains concentrated on Task A, focusing on the detection of hateful or offensive content within these pandemic-contextualized tweets. To enhance our model's performance and adaptability, approximately 10% of the training data is set aside as a development set, which is used to iteratively evaluate and refine our approaches after each training epoch (Mandl et al., 2021).

These datasets not only offer a robust framework for testing the efficacy of various NLP methods but also provide a diverse array of text samples across different contexts and temporal settings. The inclusion of COVID-19 related content in the 2021 dataset, for instance, adds a layer of complexity and relevance to the task of hate speech detection, reflecting how societal issues can influence online discourse. By continuously testing and improving our methods on these rich datasets, we aim to develop more sophisticated and context-aware models capable of handling the nuances of language used in hate speech and offensive communications. Through the careful analysis of these datasets, our study aims to contribute significantly to the evolving field of hate speech detection by providing insights into how different natural language processing techniques can be optimized to detect and categorize offensive

content more effectively. The iterative testing and refinement process enabled by the structure of these datasets ensure that our conclusions are not only based on static assumptions but are also tested against real-world, dynamically changing scenarios. This methodological rigor helps in building algorithms that are not only robust but also adaptable to the shifting paradigms of online communication.

As shown in Table 2, the class distribution in the training sets of HASOC 2020 and HASOC 2021 shows that while the 2020 dataset is balanced, the 2021 dataset has a noticeable imbalance, with the number of "HOF" (hate or offensive) entries being nearly twice that of the "NOT" (non-hate) entries. This imbalance was introduced to give the model more examples of hate speech, which helps it learn better. However, it also creates a slight imbalance in the dataset, which does not fully reflect real-world scenarios where hate speech is usually much less common than non-hate content. In actual hate speech detection tasks, hate speech tends to be rare and datasets are often highly imbalanced. To handle this issue and ensure fair model evaluation, we use both Weighted F1 and Macro F1 as evaluation metrics. These metrics are well-suited for imbalanced data because they provide a clearer picture of how well the model performs on both the more common and less common classes.

Class	HASOC 2020	HASOC 2021
NOT	1,852	1,342
HOF	1,856	2,501
Sum	3,708	3,843

Table 2: Statistical overview of the Training Data. HOF represents hate or offense, and NOT represents non-hate.

Sensory knowledge

Before conducting the hate speech detection experiments, we conducted a preliminary analysis of the sensory lexicon from the perspective of sensory perception value distribution. Figure 2 displays histograms of the six sensory measures across all words in the Lancaster norms. Notably, the distributions of these perceptual measures are unbalanced. Gustatory and olfactory measures predominantly demonstrate a right-skewed distribution[5], with most values ranging between 0 and 1. This suggests that these two sensory perceptions are less frequently represented in the textual context. Thus, it might be challenging to represent gustatory and olfactory perceptions from text.

[5] A right-skewed distribution, also known as a positive skew, is characterized by a tail that extends more significantly to the right, indicating that a majority of data points are concentrated on the left of the peak. This results in the mean being greater than the median, and the median being greater than the mode (Mean > Median > Mode).

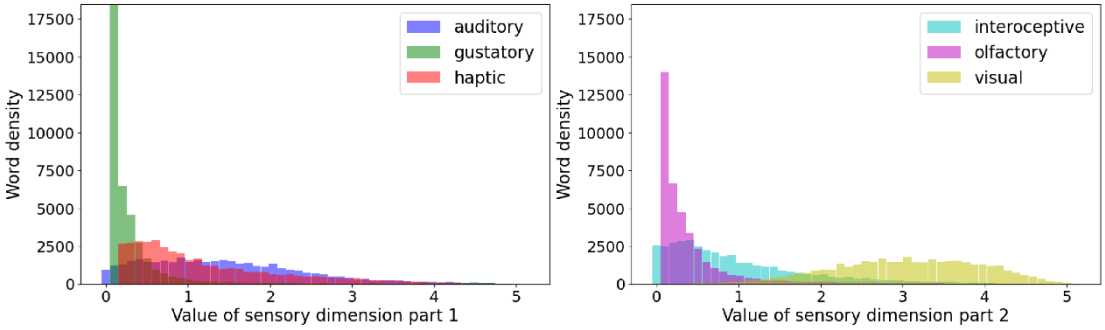

Figure 2: The distribution of six sensory values over words. The X-axis shows the value in a sensory dimension, and the y-axis displays the word density.

In contrast, auditory and visual measures show a relatively uniform distribution. The auditory measure is evenly distributed between 0 and 2.5, while the visual measure ranges between 2 and 4.5. These distributions indicate a higher sensitivity of auditory and visual knowledge to textual information, which suggests that auditory and visual senses may play a significant role within sensory models. Lastly, haptic and interoceptive measures exhibit similar trends, declining from about 2500 to 0 as the values increase from 0 to 5. These declines in the presence of haptic and interoceptive knowledge across the general textual context might suggest that they are less informative sensory dimensions in the majority of cases.

The Lancaster norms are subject to the size limitation, resulting in a significant number of out-of-vocabulary words whose sensory values are unavailable. To

address this challenge, we adopted the method proposed by Li et al. (2017) for predicting sensory values of unknown words through embedding techniques. In our experiments, we utilized both the T5 embedding and the GloVe embedding (Pennington et al., 2014) for this prediction task.

To assess the performance of our predictions, we randomly selected 10% of the Lancaster norms as a validation set and applied the Root Mean Square Error (RMSE) as the evaluation metric. The results of the prediction task, presented in Table 2, demonstrate that the GloVe embedding outperforms the T5 embedding in predicting each sensory dimension. To preserve the original values of the Lancaster norms to a maximal extent, we opted for a smaller version of GloVe with 400,000 data points and 200 dimensions. Following augmentation, the size of our sensory vocabulary has reached 407,572[6].

Although we conducted rigorous analysis and processing, it must be acknowledged that the Lancaster norms have certain limitations. First, the Lancaster norms are static, whereas human understanding of language is constantly evolving. The sensory information and intensity that people associate with the same word may change over time. Additionally, individual differences in sensory perception mean that different people may experience

[6] The whole dataset of the sensory vocabulary can be accessed at: https://osf.io/w8yez/?view_only=0e807dfaa5e6433184e452bfebabd01b.

varying degrees of sensory intensity from the same words. We chose the Lancaster norms because they have been extensively validated and widely applied in experiments across fields such as medicine and cognitive science, which demonstrates the reliability and validity of this knowledge. Furthermore, the Lancaster norms were generated through crowd-sourcing, aggregating data from a large number of participants, which gives this knowledge a certain degree of generalizability and applicability. While we consider the Lancaster norms to be the most reasonable source of sensory knowledge at present, in the future, we plan to explore and develop more effective sensory knowledge sources.

Sensory Name	T5 Embedding	GloVe
Haptic	0.893	**0.698**
Gustatory	0.632	**0.534**
Olfactory	0.572	**0.501**
Visual	0.842	**0.743**
Auditory	0.949	**0.803**
Interoceptive	0.831	**0.662**
Total	0.798	**0.665**

Table 3: Comparison of the accuracy of predictions between T5 embedding and GloVe techniques on different sensory dimensions, as measured by RMSE values. Lower scores indicate higher accuracy in the prediction of sensory values.

Selected baselines

For comprehensive comparisons involving data speech detection, we benchmark our proposed SensoryT5 model against a range of established models.

CNN (Adewumi et al., 2023): Originally renowned in the field of computer vision, CNNs have been effectively adapted for NLP by capturing local text patterns through three convolutional layers, each equipped with 100 filters of increasing size. This adaptation uses ReLU activation and max-pooling techniques to optimize text pattern recognition, supported by a dropout layer for regularization. With a total of 1,386,201 trainable parameters, the CNN model demonstrates high efficiency in processing and classifying text-based data.

Bi-LSTM (Adewumi et al., 2023): A variant of the Recurrent Neural Network, the Bi-LSTM processes input text in both forward and backward directions using two bidirectional layers and pretrained GloVe word embeddings of 100 dimensions, enhancing contextual understanding. This model also incorporates a dropout layer to mitigate overfitting and contains 1,317,721 parameters, making it effective for complex NLP tasks due to its improved handling of context and sequence in text.

BERT (Devlin et al., 2019): A Pretrained Language Model (PLM) that

revolutionized text classification by processing text inputs in the [CLS] text [SEP] format. Its deep semantic understanding makes it highly effective for hate speech detection tasks.

RoBERTa (Liu et al., 2019): As an enhanced version of BERT, RoBERTa refines training processes and hyperparameters to significantly improve performance in NLP tasks.

XLNet (Yang et al., 2019): This model extends BERT's capabilities by learning bidirectional contexts and using an autoregressive formulation, overcoming some of BERT's limitations.

T5 (Raffel et al., 2020): The Text-to-Text Transfer Transformer (T5) adapts all NLP tasks into a text-to-text format, showcasing the versatility and strong performance across various tasks.

Experiment settings and implementation details.

During training, we applied the Adam optimizer in Euclidean space, characterized by its efficiency in navigating the parameter space and adjusting weights to minimize loss. The hyperparameters used are learning rate = $2e-5$, max_seq_length = 100, batch_size = 32.

Results and discussion

Results of experiments

	HASOC 2020 A		HASOC 2021 A	
	Weighted F1	Macro F1	Weighted F1	Macro F1
CNN	-	-	0.776	0.757
Bi-LSTM	-	-	0.784	0.772
$\text{BERT}_{\text{large}}$	0.897	0.894	0.774	0.736
$\text{RoBERTa}_{\text{large}}$	0.900	0.900	0.795	0.780
$\text{XLNet}_{\text{large}}$	0.896	0.896	0.780	0.753
T5_{large}	0.908	0.908	0.799	0.781
SensoryT5	**0.915**	**0.915**	**0.810**	**0.793**

Table 4: Results of the SensoryT5 model in comparison to the baselines across two hate speech detection datasets. The top performances are highlighted in bold.

In the realm of hate speech detection, our proposed SensoryT5 model demonstrates a notable improvement over both traditional machine learning models (such as CNN and Bi-LSTM) and advanced pre-trained language models (PLMs)/large language models (LLMs) like BERT, RoBERTa, XLNet, and T5. We focused our evaluations on the weighted F1 and Macro F1 scores across two major datasets: HASOC 2020 Task A and HASOC 2021 Task A.

In our analysis, T5 was previously identified as the best performing model among all baselines in the HASOC 2020 dataset, achieving a weighted F1 and

Macro F1 score that was 0.8% higher than the second-best model. Similarly, in the HASOC 2021 dataset, T5 outperformed the second-best model by 0.4% in weighted F1 and by 0.1% in Macro F1. Interestingly, traditional machine learning models like the Bi-LSTM performed comparably to PLMs/LLMs, which is unusual for text classification tasks where PLMs/LLMs typically dominate. This anomaly highlights the unique demands of hate speech detection, which relies heavily on lexical foundations and benefits from semantically-based models.

Our SensoryT5 model outshined all baselines, including the high-performing T5. In the HASOC 2020 dataset, SensoryT5 exceeded T5 by 0.7% in both weighted F1 and Macro F1 scores. In the HASOC 2021 dataset, SensoryT5 improved upon T5 by 1.1% in weighted F1 and by 1.2% in Macro F1. These enhancements underscore the effectiveness of integrating sensory information into the model, which significantly boosts its capability to detect hate speech.

The integration of sensory data with the robust text-to-text framework of T5 not only enhances the model's accuracy but also its interpretative power, allowing it to better discern the nuanced contexts and subtleties involved in hate speech. This advanced capability positions SensoryT5 as a superior tool for tackling the complexities of hate speech detection across varied and challenging datasets.

Ablation studies

Evaluating the impact of sensory knowledge

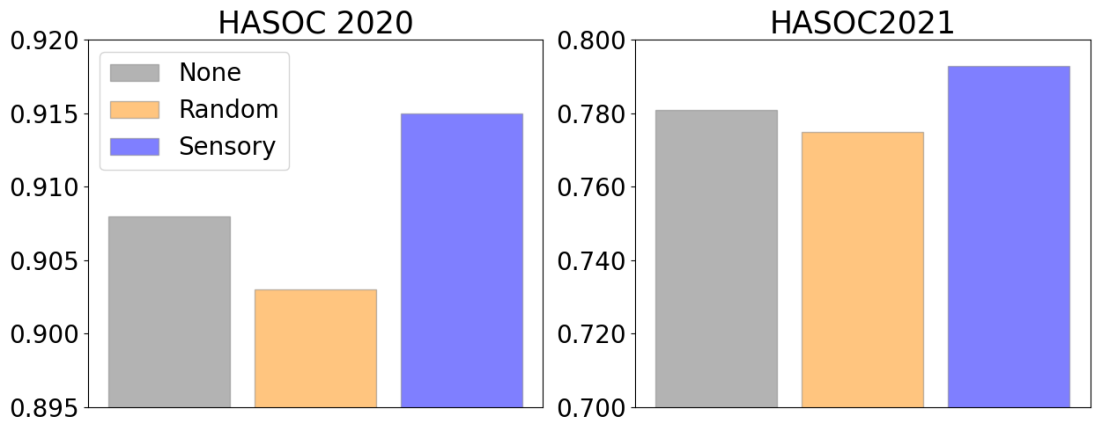

Figure 3: The performances of T5 (None), Random SensoryT5 (with sensory values randomly assigned), and SensoryT5 across two hate speech detection datasets, evaluated using macro F1 as the metric.

To elucidate the impact of different components within the SensoryT5 model, we conducted ablation studies specifically for the task of hate speech detection, using the HASOC 2020 and HASOC 2021 datasets. These studies are instrumental in assessing the efficacy of our novel sensory integration approach. The ablation tests were structured around three primary configurations:

T5 (None): The baseline model without any sensory information, representing the conventional approach in hate speech detection tasks.

Random SensoryT5: A variant where the sensory scores are substituted with random numbers ranging from 0 to 5. This setup preserves the distribution of sensory scores but removes their meaningful correlations with sensory vocabulary data.

SensoryT5: Our proposed model which incorporates cognitively-grounded and linguistically-encoded sensory knowledge.

The results, as depicted in Figure 3, reveal that the SensoryT5 model outperforms the other configurations in terms of Macro F1 scores across both datasets. However, the Random SensoryT5 model exhibits lower performance than the T5 (None) model, particularly highlighting the detriments of arbitrary sensory data integration.

These findings emphasize that the effectiveness of the SensoryT5 model stems not just from the addition of numerical data but from the meaningful integration of sensory knowledge that is both cognitively-grounded and linguistically-encoded (Lynott et al., 2020). Conversely, the poorer performance of the Random SensoryT5 model compared to the T5 (None) setup suggests that indiscriminate inclusion of sensory information can introduce noise, impairing the model's ability to accurately detect and categorize hate speech. This underscores the importance of strategic sensory integration, as random additions could potentially disrupt model performance and prove to be

counterproductive.

Impact of sensory knowledge on different foundation models

To further investigate the versatility and efficacy of sensory knowledge in NLP, we expanded our experiments to include additional foundational transformer models such as BERT, RoBERTa, and XLNet. These models were enriched with cognitively-motivated and linguistically-encoded sensory data, as detailed in Figure 4. Our objective was to assess the impact of sensory integration on these established models, particularly in the context of hate speech detection.

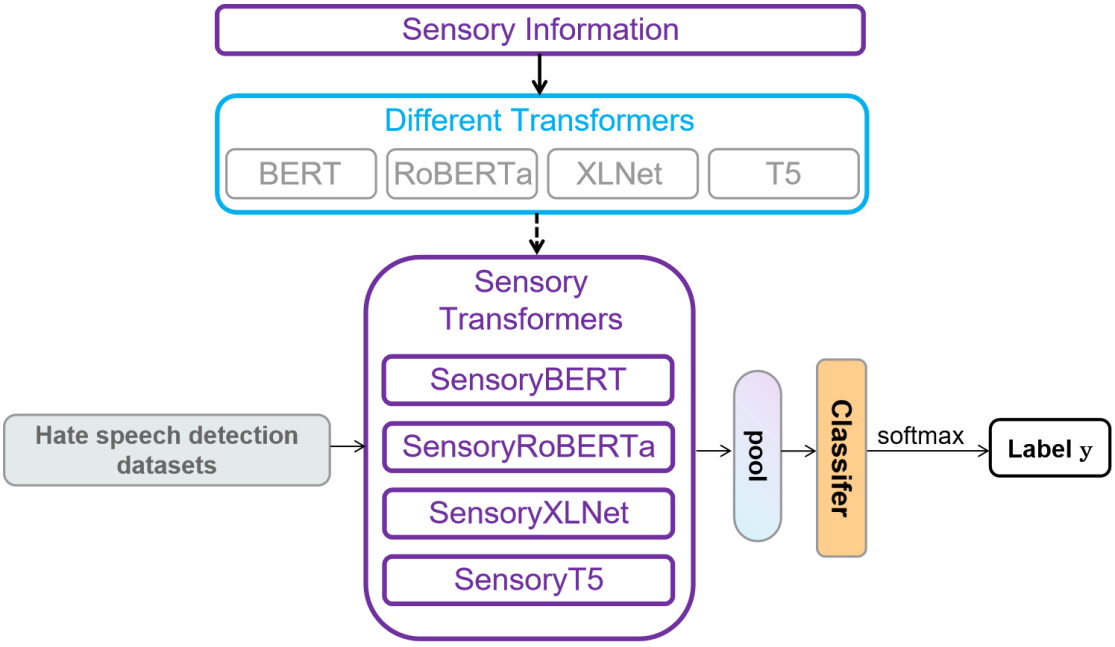

Figure 4: Infusing sensory knowledge into different transformer-based models. In SensoryXLNet, both K (key) and V (value) in the sensory attention layer are constituted by the output of the final hidden layer of XLNet.

The effects of integrating sensory knowledge into various foundational models are summarized in Table 4 and indicate substantial improvements in hate speech detection capabilities across all modified models when compared to their original counterparts.

	HASOC 2020 A		HASOC 2021 A	
	Weighted F1	Macro F1	Weighted F1	Macro F1
$\text{BERT}_{\text{large}}$	0.897	0.894	0.774	0.736
SensoryBERT	**0.900**	**0.899**	**0.778**	**0.741**
$\text{RoBERTa}_{\text{large}}$	0.900	0.900	0.795	0.780
SensoryRoBERTa	**0.905**	**0.905**	**0.808**	**0.788**
$\text{XLNet}_{\text{large}}$	0.896	0.896	0.780	0.753
SensoryXLNet	**0.899**	**0.900**	**0.785**	**0.760**
T5_{large}	0.908	0.908	0.799	0.781
SensoryT5	**0.915**	**0.915**	**0.810**	**0.793**

Table 5: Impact of sensory knowledge on different foundation models.

SensoryBERT: The integration of sensory knowledge with BERT resulted in improvements in hate speech detection F1 scores. For HASOC 2020, SensoryBERT achieved a 0.3% increase in weighted F1 and a 0.4% rise in Macro F1 scores over the original BERT. In HASOC 2021, these gains were even more pronounced, with improvements of 0.4% in weighted F1 and 0.5% in Macro F1.

SensoryRoBERTa: Similar enhancements were observed with RoBERTa upon sensory integration. In HASOC 2020, SensoryRoBERTa outperformed the base

RoBERTa model by 0.5% in both weighted F1 and Macro F1 scores. The 2021 data showed greater improvements, with increases of 0.8% in both weighted F1 and Macro F1 scores.

SensoryXLNet: The SensoryXLNet configuration also demonstrated beneficial outcomes. It surpassed the base XLNet model by 0.3% in weighted F1 and 0.4% in Macro F1 for HASOC 2020, and by 0.5% in weighted F1 and 0.7% in Macro F1 for HASOC 2021.

SensoryT5: The most notable improvements were seen with SensoryT5, which consistently outperformed the base T5 model across both datasets. In HASOC 2020, there was an enhancement of 0.7% in both weighted F1 and Macro F1 scores. The HASOC 2021 results were even more impressive, with SensoryT5 leading by 1.1% in weighted F1 and 1.2% in Macro F1.

These results collectively underscore the significant value of infusing sensory knowledge into transformer-based models for hate speech detection. The consistent improvements across different platforms confirm that sensory integration not only enhances the models' performance but also extends their applicability to more complex NLP tasks. This approach leverages neuro-cognitive data effectively, highlighting the potential of combining computational methods with cognitive science insights to improve NLP systems' understanding and processing capabilities in sensitive areas such as hate

speech detection.

Case study

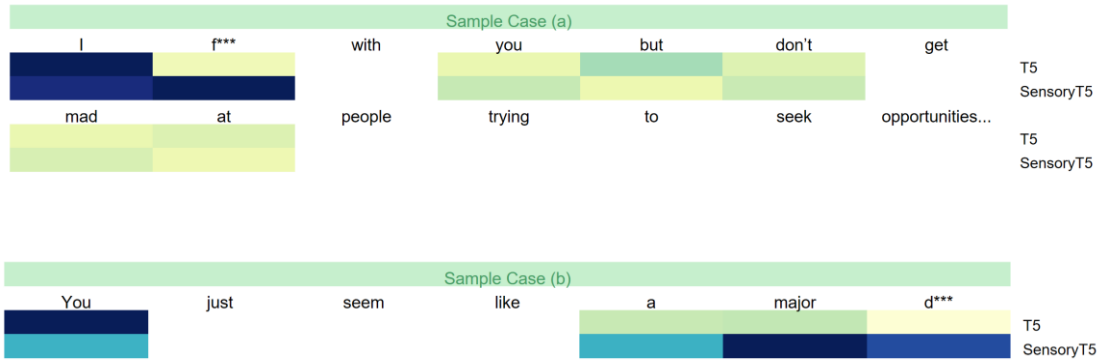

Figure 5: The heatmap visualizes the heat values of the final sensory layer in SensoryT5 and the encoder layer in T5 for two sentences. Darker colors indicate higher attention weights. These sentences are sourced sequentially from the HASOC 2020 and HASOC 2021 datasets.

Case studies are presented in Figure 5 using two sentences from the HASOC 2020 and HASOC 2021 datasets: *"I f*** with you but don't get mad at people trying to seek opportunities..."** (HASOC 2020), *"You just seem like a major d***"* (HASOC 2021). The SensoryT5 heatmaps illustrate the aggregate attention for each token in the sensory layer, while the T5 section compiles and averages attention weights across all encoder layers to reveal the model's overall focus. The SensoryT5 model shows more attention on hate-related phrases: *"f***"* in the first sentence and *"major d***"* in the second sentence. This indicates the model's ability to detect crucial hateful terms that are highly relevant in

identifying offensive content. In contrast, the standard T5's attention is more spread out and less concentrated on these hateful terms. It demonstrates a broader focus on the entire sentence, missing the heightened sensitivity needed to zero in on specific offensive words. These micro-level analyses show SensoryT5's superior capability in recognizing and highlighting hate speech, proving the efficiency of integrating sensory awareness into language models for improved detection of hate and offensive language. By selecting sentences from multiple datasets, we demonstrate SensoryT5's robustness and ability to generalize across different contexts in identifying and focusing on hate-related phrases. These micro-level analyses further validate SensoryT5's advantage in handling offensive language detection by emphasizing key words that signal hateful intent, contributing to its effectiveness in real-world hate speech detection tasks.

Conclusion

This study introduces the SensoryT5 framework, a cutting-edge model designed specifically for enhancing hate speech detection. By integrating sensory knowledge into the established transformer architecture, SensoryT5 excels in identifying subtle contextual cues essential for recognizing hateful language. Utilizing sophisticated attention mechanisms, it meticulously combines sensory and contextual information to provide an in-depth analysis of

text. Through extensive testing across a variety of datasets, SensoryT5 has demonstrated that it exceeds the performance of the foundational T5 model as well as other popular pre-trained language models (PLMs) and large language models (LLMs). Moreover, it also shows improvements over traditional machine learning approaches used in NLP for hate speech detection. This achievement highlights SensoryT5's capability as a bridge that connects sensory perception with linguistic analysis, moving NLP towards integrating neuro-cognitive insights more robustly. The model's effectiveness is underpinned by its use of the well-established neuro-cognitive links between sensory experiences and behavioral outputs to better predict and understand hate speech. By decoding the sensory cues found in texts, SensoryT5 accesses the underlying emotional drivers of hate speech, similar to how humans interpret sensory inputs to form emotional and cognitive responses. This not only improves the accuracy of hate speech detection but also encourages deeper interdisciplinary research and collaboration. SensoryT5 integrates advanced text processing techniques with insights from both sensory perception and cognitive science, showcasing a comprehensive approach in computational linguistics. This strategy emphasizes the importance of understanding language on a nuanced level that goes beyond mere words, incorporating the broad spectrum of human experiences and perceptions that shape communicative intentions.

In summary, SensoryT5 not only refines the accuracy of detecting hate speech

but also enhances the interpretability of detection outcomes, fostering a more profound comprehension of the content and context of hate speech. This comprehensive approach does not just advance the technical finesse of NLP models but also deepens our understanding of the complex interplay between language, emotion, and cognition. As it breaks new ground in methodological approaches, SensoryT5 encourages ongoing research into integrated, human-centric methods in NLP. This pioneering model points towards a future where digital communication spaces are safer and more respectful, propelled by an enhanced understanding of the cognitive and emotional underpinnings of language. This promising direction in NLP research invites further academic exploration and technological innovation, aiming to expand our capabilities in combating online hate speech and enriching our grasp of the nuanced relationship between human cognition and communicative expressions.

Limitation and future plan

In our current research, we have utilized GloVe and T5 embeddings to estimate sensory values for unknown words through a regression-based approach. While this method has proven effective within the scope of our study, it inherently relies on the availability and quality of pre-trained embeddings which may not fully capture the nuanced sensory attributes that are specific to all contexts or languages. Inspired by the work of Chersoni et al. (2020), which

demonstrates the viability of cross-lingual methods in sensory knowledge prediction, we are optimistic about the potential of extending our SensoryT5 model to encompass multiple languages. This would involve adapting the model to utilize multilingual embeddings and developing regression techniques that can operate effectively across different linguistic frameworks. Such advancements could dramatically increase the model’s applicability and effectiveness in a global context, making it a valuable tool for international platforms where multiple languages are used. Another promising direction for future research is to expand the sensory knowledge base used by our model. Currently, the model's performance is partially dependent on the predefined sensory values derived from existing datasets. By incorporating a broader array of sensory inputs and possibly crowdsourcing sensory data from diverse cultures and languages, we could significantly enhance the model's understanding and interpretation of sensory nuances in text. To improve the model’s versatility and its ability to handle a variety of NLP tasks in diverse linguistic environments, we plan to explore more sophisticated machine learning techniques. For instance, employing advanced neural network architectures such as Transformer-XL or GPT-3 could provide a way to better capture long-range dependencies and contextual subtleties that our current T5 setup might miss. These improvements could lead to more accurate predictions of sensory values and, by extension, more effective hate speech detection in

multilingual contexts. We also see substantial value in adopting more interdisciplinary approaches that integrate insights from cognitive science, psychology, and linguistics to enrich the sensory attributes analyzed by our model. This could involve collaborative projects aimed at understanding how sensory perceptions are expressed differently across languages and cultures, which would help in fine-tuning our model to be sensitive to these variations. Looking ahead, our ultimate goal is to develop a robust, adaptable framework that can not only detect hate speech but also provide insights into the emotional and cognitive underpinnings of language use in various cultural and linguistic settings. By continuously refining our SensoryT5 model and expanding its linguistic and cultural reach, we aim to contribute to safer, more understanding online environments worldwide.

By addressing these areas, we aim to overcome the current limitations and significantly expand the scope and efficacy of our SensoryT5 model. This will pave the way for more nuanced and culturally aware NLP applications, potentially revolutionizing how automated systems understand and interact with human language on a global scale.

Reference

Adewumi, T., Sabry, S. S., Abid, N., Liwicki, F., & Liwicki, M. (2023). T5 for Hate Speech, Augmented Data, and Ensemble. *Sci*, *5*(4), Article 4. https://doi.org/10.3390/sci5040037

Ali, M. Z., Ehsan-Ul-Haq, Rauf, S., Javed, K., & Hussain, S. (2021). Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis. *IEEE Access*, *9*, 84296–84305. https://doi.org/10.1109/ACCESS.2021.3087827

Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information*, *13*(6), Article 6. https://doi.org/10.3390/info13060273

Bauwelinck, N., Jacobs, G., Hoste, V., & Lefever, E. (2019). LT3 at SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter (hatEval). *13th International Workshop on Semantic Evaluation (SemEval-2019), Collocated with NAACL 2019.*, 436–440. https://biblio.ugent.be/publication/8604488/file/8640071

Bisht, A., Singh, A., Bhadauria, H. S., Virmani, J., & Kriti. (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In S. Jain & S. Paul (Eds.), *Recent Trends in Image and Signal Processing in Computer Vision* (pp. 243–264). Springer. https://doi.org/10.1007/978-981-15-2740-1_17

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances*

in Neural Information Processing Systems, *33*, 1877–1901.

Chakrabarty, N. (2020). A Machine Learning Approach to Comment Toxicity Classification. In A. K. Das, J. Nayak, B. Naik, S. K. Pati, & D. Pelusi (Eds.), *Computational Intelligence in Pattern Recognition* (Vol. 999, pp. 183–193). Springer Singapore. https://doi.org/10.1007/978-981-13-9042-5_16

Chen, X., Hai, Z., Wang, S., Li, D., Wang, C., & Luan, H. (2021). Metaphor identification: A contextual inconsistency based neural sequence labeling approach. *Neurocomputing*, *428*, 268–279. https://doi.org/10.1016/j.neucom.2020.12.010

Chersoni, E., Xiang, R., Lu, Q., & Huang, C.-R. (2020). Automatic Learning of Modality Exclusivity Norms with Crosslingual Word Embeddings. In I. Gurevych, M. Apidianaki, & M. Faruqui (Eds.), *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics* (pp. 32–38). Association for Computational Linguistics. https://aclanthology.org/2020.starsem-1.4

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., & Gehrmann, S. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

de Andrade, C. M. V., & Gonçalves, M. A. (2021). Profiling Hate Speech Spreaders on Twitter:

Exploiting textual analysis of tweets and combinations of multiple textual representations. *CEUR Workshop Proc*, *2936*, 2186–2192. https://ceur-ws.org/Vol-2936/paper-195.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Duéñez-Guzmán, E. A., Sadedin, S., Wang, J. X., McKee, K. R., & Leibo, J. Z. (2023). A social path to human-like artificial intelligence. *Nature Machine Intelligence*, *5*(11), 1181–1188. https://doi.org/10.1038/s42256-023-00754-x

Fan, T., Qiu, S., Wang, Z., Zhao, H., Jiang, J., Wang, Y., Xu, J., Sun, T., & Jiang, N. (2023). A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition. *Computers in Biology and Medicine*, *159*, 106938. https://doi.org/10.1016/j.compbiomed.2023.106938

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management, 58*(3), Article 102524. https://doi.org/10.1016/j.ipm.2021.102524

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of*

Intelligent & Fuzzy Systems, *36*(5), 4743–4752.

Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. In Z. Waseem, W. H. K. Chung, D. Hovy, & J. Tetreault (Eds.), *Proceedings of the First Workshop on Abusive Language Online* (pp. 85–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3013

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love" evading hate speech detection. *In Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2–12). https://doi.org/10.1145/3270101.3270105

Heinzerling, B., & Strube, M. (2017). BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. arXiv preprint arXiv:1710.02187. https://arxiv.org/abs/1710.02187

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., & Varma, V. (2019). FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 70–74. https://aclanthology.org/S19-2009/

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Jahan, M. S., Oussalah, M., Beddia, D. R., Mim, J. kabir, & Arhab, N. (2024). *A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs* (arXiv:2404.00303). arXiv. https://doi.org/10.48550/arXiv.2404.00303

Jin, Y., Wanner, L., & Shvets, A. (2024). *GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection?* (arXiv:2402.15238). arXiv. https://doi.org/10.48550/arXiv.2402.15238

Khanam, S., Tanweer, S., Khalid, S., & Rosaci, D. (2019). Artificial Intelligence Surpassing Human Intelligence: Factual or Hoax. *The Computer Journal*, *64*(12), 1832–1839. https://doi.org/10.1093/comjnl/bxz156

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

Kikkisetti, D., Mustafa, R. U., Melillo, W., Corizzo, R., Boukouvalas, Z., Gill, J., & Japkowicz, N. (2024). *Using LLMs to discover emerging coded antisemitic hate-speech in extremist social media* (arXiv:2401.10841). arXiv. https://doi.org/10.48550/arXiv.2401.10841

Li, M., Lu, Q., Long, Y., & Gui, L. (2017). Inferring Affective Meanings of Words from Word Embedding. *IEEE Transactions on Affective Computing*, *8*(4), 443–456. https://doi.org/10.1109/TAFFC.2017.2723012

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., & Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective*

Computing, *12*(4), 900–912.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558–564. https://doi.org/10.3758/BRM.41.2.558

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, *45*(2), 516–526. https://doi.org/10.3758/s13428-012-0267-0

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2020). Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. *Forum for Information Retrieval Evaluation*, 29–32. https://doi.org/10.1145/3441501.3441517

Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schaefer, J., Ranasinghe, T., Zampieri, M., Nandini, D., & Jaiswal, A. K. (2021). *Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages* (arXiv:2112.09301). arXiv. http://arxiv.org/abs/2112.09301

Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021). Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. In O. De Clercq,

A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, & V. Hoste (Eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 149–159). Association for Computational Linguistics. https://aclanthology.org/2021.wassa-1.16

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, 61–66. https://ieeexplore.ieee.org/abstract/document/8575590/

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. https://doi.org/10.1145/2872427.2883062

Nockleby, J. T. (1994). Hate Speech in Context: The Case of Verbal Threats. *Buffalo Law Review*, *42*, 653.

Nockleby, J. T., Levy, L. W., Karst, K. L., & Mahoney, D. J. (2000). Encyclopedia of the American constitution. *Hate Speech; Levy, L., Karst, K., Eds*, 1277–1279.

Ombui, E., Muchemi, L., & Wagacha, P. (2019). Hate speech detection in code-switched text messages. *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–6. https://ieeexplore.ieee.org/abstract/document/8932845/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V.,

Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ⋯ Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. *In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 363–370). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2051

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management, 57*(6), Article 102360. https://doi.org/10.1016/j.ipm.2020.102360

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. *IEEE Access*, *9*, 112478–112489. https://doi.org/10.1109/ACCESS.2021.3103697

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J.

(2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rana, A., & Jha, S. (2022). *Emotion Based Hate Speech Detection using Multimodal Learning* (arXiv:2202.06218). arXiv. https://doi.org/10.48550/arXiv.2202.06218

Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., & Roca, F. X. (2022). Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Transactions on Cybernetics*, *52*(5), 3314–3324. https://doi.org/10.1109/TCYB.2017.2662199

Romero, O. J., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2023). Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis. *Proceedings of the AAAI Symposium Series*, *2*(1), 396–405. https://ojs.aaai.org/index.php/AAAI-SS/article/view/27706

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. https://doi.org/10.17185/duepublico/42132

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2018). *Hateminers: Detecting Hate speech against Women* (arXiv:1812.06700). arXiv. https://doi.org/10.48550/arXiv.1812.06700

Saksesi, A. S., Nasrun, M., & Setianingsih, C. (2018). Analysis Text of Hate Speech Detection Using Recurrent Neural Network. *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*, 242–248. https://doi.org/10.1109/ICCEREC.2018.8712104

Shi, X., Liu, J., & Song, Y. (2024). BERT and LLM-Based Multivariate Hate Speech Detection on Twitter: Comparative Analysis and Superior Performance. In H. Jin, Y. Pan, & J. Lu (Eds.), *Artificial Intelligence and Machine Learning* (pp. 85–97). Springer Nature. https://doi.org/10.1007/978-981-97-1277-9_7

Skaramagkas, V., Ktistakis, E., Manousos, D., Kazantzaki, E., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2023). eSEE-d: Emotional State Estimation Based on Eye-Tracking Dataset. *Brain Sciences*, *13*(4), Article 4. https://doi.org/10.3390/brainsci13040589

Srivastava, N. D., & Sharma, Y. (2020). Combating online hate: A comparative study on identification of hate speech and offensive content in social media text. *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 47–52. https://ieeexplore.ieee.org/abstract/document/9332469/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Tuncer, T., Dogan, S., & Acharya, U. R. (2021). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowledge-Based Systems*, *211*, 106547. https://doi.org/10.1016/j.knosys.2020.106547

Wan, M., Su, Q., Ahrens, K., & Huang, C.-R. (2023). Perceptional and actional enrichment for

metaphor detection with sensorimotor norms. *Natural Language Engineering*, 1–29. https://doi.org/10.1017/S135132492300044X

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the Second Workshop on Language in Social Media*, 19–26. https://aclanthology.org/W12-2103.pdf

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://aclanthology.org/N16-2013.pdf

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835.

Wermiel, S. J. (2017). The Ongoing Challenge to Define Free Speech. *Human Rights*, *43*, 82.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. *ISBN*, 1–10.

Wigand, C., & Voin, M. (2017). *Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law*.

Xia, Y., Zhao, Q., Long, Y., Xu, G., & Wang, J. (2024). *SensoryT5: Infusing Sensorimotor Norms into T5 for Enhanced Fine-grained Emotion Classification* (arXiv:2403.15574). arXiv. http://arxiv.org/abs/2403.15574

Yan, X., Zhang, Y., & Zhang, C. (2024). Utilizing cognitive signals generated during human

reading to enhance keyphrase extraction from microblogs. *Information Processing & Management*, *61*(2), 103614.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7, Article e598. https://doi.org/10.7717/peerj-cs.598

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)* (arXiv:1903.08983). arXiv. https://doi.org/10.48550/arXiv.1903.08983

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)* (arXiv:2006.07235). arXiv. https://doi.org/10.48550/arXiv.2006.07235

Zhang, M., He, J., Ji, T., & Lu, C.-T. (2024). *Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection* (arXiv:2402.11406). arXiv. https://doi.org/10.48550/arXiv.2402.11406

Zhong, P., Wang, D., & Miao, C. (2022). EEG-Based Emotion Recognition Using Regularized

Graph Neural Networks. *IEEE Transactions on Affective Computing*, *13*(3), 1290–1301. https://doi.org/10.1109/TAFFC.2020.2994159