GAttention: Gated Attention for the Detection of Abusive Language

Horacio Jarquín-Vásquez[1,2], Hugo Jair Escalante[2], Manuel Montes-y-Gómez[2]
and **Mario Ezra Aragón[3]**
[1]Dipartimento di Informatica, Universita degli Studi di Torino (UNITO), Italy,
[2]Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico,
[3]Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela (USC), Spain
horaciojesus.jarquinvasquez@unito.it,
{hugojair, mmontesg}@inaoep.mx, ezra.aragon@usc.es

Abstract

Abusive language online creates toxic environments and exacerbates social tensions, underscoring the need for robust NLP models to interpret nuanced linguistic cues. This paper introduces GAttention, a novel Gated Attention mechanism that combines the strengths of Contextual attention and Self-attention mechanisms to address the limitations of existing attention models within the text classification task. GAttention capitalizes on local and global query vectors by integrating the internal relationships within a sequence (Self-attention) and the global relationships among distinct sequences (Contextual attention). This combination allows for a more nuanced understanding and processing of sequence elements, which is particularly beneficial in context-sensitive text classification tasks such as the case of abusive language detection. By applying this mechanism to transformer-based encoder models, we showcase how it enhances the model's ability to discern subtle nuances and contextual clues essential for identifying abusive language, a challenging and increasingly relevant NLP task.

1 Introduction

Abusive language (AL) has become an increasingly urgent concern in today's digitally connected world, where social media and online platforms serve as primary venues for communication (Mandl et al., 2019). Offensive or hateful content can foster a toxic atmosphere, harm vulnerable communities, and exacerbate social tensions (MacAvaney et al., 2019). Consequently, detecting AL has emerged as a pivotal challenge within Natural Language Processing (NLP), While numerous challenges exist in the detection of AL, one of the most critical lies in the need for models capable of interpreting both offensive and non-offensive words, along with the underlying context in which such language is used for the accurate identification of AL (Alkomah and Ma, 2022).

In parallel with this societal need, current encoder–decoder transformer models have gained widespread acceptance in the NLP community, offering a versatile framework for various tasks (Lin et al., 2022). Their adaptability allows for fine-tuning in a wide range of tasks, including Question Answering, Machine Translation, Text Classification, Text Generation, Text Summarization, Sentiment Analysis, and Named Entity Recognition (Durairaj and Chinnalagu, 2021; Ramprasath et al., 2022). A key component of these models is the attention mechanism, which is crucial for capturing dependencies within a sequence and conditioning the model's outputs on these learned relationships (Niu et al., 2021).

Over recent years, numerous attention mechanisms have been proposed to address diverse tasks in NLP and computer vision (Brauwers and Frasincar, 2023). These mechanisms focus on modeling local and global dependencies (Yang et al., 2016; Vaswani et al., 2017), unifying different modalities by projecting features from one modality to another (Patel et al., 2022), integrating various abstraction levels within an input sequence (Vaswani et al., 2017; Zhao and Zhang, 2018), or combining the outputs of distinct attention modules (Huang et al., 2019). Particularly relevant are Contextual Attention (CA) (Yang et al., 2016) and Self-Attention (SA) (Vaswani et al., 2017), which approach the modeling of local and global dependencies in distinct ways. While SA captures local dependencies by deriving its query vector directly from elements within the same sequence, CA models global dependencies by emphasizing specific tokens based on an external query vector, which is updated according to task-specific requirements (Chaudhari et al., 2020).

Both SA and CA have demonstrated strong performance in many applications (Hu, 2019); however, each also presents limitations. CA overlooks the *local* relationships among tokens within a se-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20314–20329
November 4-9, 2025

quence, whereas SA does not explicitly account for the *global* relationships across distinct sequences. This can result in the loss of crucial information, potentially weakening a model's ability to interpret context, a critical element for tasks such as AL detection, where meaning can shift dramatically based on linguistic nuance. For the sake of clarity, we use the term *local relationships* to refer to the connections between the words present within a single message, and the term *global relationships* to denote their interpretation within a broader domain or a task-specific setting. In our application domain, their use in other texts related to AL.

To address these limitations, we introduce GAttention, a novel Gated Attention mechanism that combines the strengths of both *SA* and *CA* to capture local and global dependencies better. By weighting and merging the local (SA) query vectors and the global (CA) query vectors, GAttention preserves the crucial contextual cues needed for complex classification problems. Additionally, we introduce the Multi-Head GAttention mechanism to enhance the model's ability to attend to different aspects of the input simultaneously, improving its capacity to capture intricate relationships within the data.

Although both the GAttention and Multi-Head GAttention variants can be applied to any sequence of word/token encoding features (or multimodal feature sequences), we focus on evaluating their efficacy in context-sensitive text classification tasks, particularly in various tasks related to the detection of AL. These tasks are ideally suited for testing our GAttention mechanism variants since text classified as abusive hinges on the context in which words are used, distinguishing between offensive and non-offensive usage (MacAvaney et al., 2019). The results obtained from the different GAttention configurations were encouraging, achieving performance improvements in five out of six AL detection subtasks when compared to three replicated state-of-the-art approaches, including techniques based on Large Language Models (LLMs). Furthermore, the proposed mechanisms outperformed the best results reported in the shared tasks, while maintaining a favorable trade-off between performance and model size.

In summary, the main contributions of this paper are as follows: 1) The proposal of the GAttention mechanism, which leverages the relevance of the local (instance-dependent) and global (task-dependent) features of any encoding sequence by combining the SA and CA representations. 2) The Multi-headed adaptation of the GAttention mechanism and its integration in contemporary transformer encoding models for the AL detection task. 3) The evaluation and analysis of the GAttention and multi-head GAttention mechanisms across three different datasets dedicated to detecting AL on social media platforms.

2 Related Work

2.1 Abusive Language Detection

Given the well-acknowledged increase of AL across social media platforms, a variety of datasets (Davidson et al., 2017; Marcos et al., 2019) and evaluation campaigns (Fersini et al., 2018; Marcos et al., 2020; Aragón et al., 2020) have emerged to mitigate its effects. AL detection has primarily employed supervised methods (Alkomah and Ma, 2022), using features ranging from bag-of-words to syntactic and linguistic cues (Schmidt and Wiegand, 2017). Moreover, advanced techniques leveraging word embeddings and Transformer-based models, such as GPT-2, BERT, Llama 2, and RoBERTa (Mutanga et al., 2020; Ripoll et al., 2022; Nguyen et al., 2024), have been explored. These approaches have given rise to diverse deep learning architectures, including ensemble methods that integrate various representations (Farooqi et al., 2021), the design of specialized attention mechanisms for AL detection (Jarquín-Vásquez et al., 2021), and the adaptation of Transformer models to the AL detection domain through pretraining tasks specifically designed for this purpose (Jarquín-Vásquez et al., 2024). Within these deep learning architectures, Transformer-based representations have notably excelled in recent years and currently represent the state-of-the-art in AL detection (Alkomah and Ma, 2022; Jahan and Oussalah, 2023; Nguyen et al., 2024).

2.2 Attention mechanisms

Within the realm of attention mechanisms that combine different representations, the use of cross-modal attention holds prominence (Patel et al., 2022). This approach focuses on projecting one modality into another, facilitating enhanced inter-modality interactions. Additionally, nested attention mechanisms have been explored to refine the representation of a sequence of features (Huang et al., 2019), alongside multi-level attention mechanisms, which hierarchically combine different fea-

tures to achieve a more comprehensive representation (You et al., 2022).

Another notable approach involves attention mechanisms designed for the fusion of distinct features and/or modalities (Dai et al., 2020; Jarquín-Vásquez et al., 2021; Wan et al., 2022; Li et al., 2023; Jarquın-Vasquez et al., 2024). These mechanisms integrate features or modalities either through an early fusion approach or by combining feature pairs, extending the use of cross-modal attention through a transversal fusion approach. Specifically, in (Jarquín-Vásquez et al., 2021), the integration of SA and CA mechanisms is exemplified through the Self-Contextualized Attention (SCA) mechanism, which merges both representations via early fusion. However, this approach limits the flexibility of each representation to dynamically adapt to different contexts within the input data, thereby constraining the nuances it can capture.

In contrast to SCA, our proposed GAttention mechanism dynamically integrates SA and CA representations through a weighted approach, adapting to the specific requirements of each instance in the text classification task. This instance-wise adaptation enables a more precise and context-aware interpretation of tokens within an input sequence.

3 Proposed Gated Attention Mechanism

This section is organized into two subsections. The first subsection details the GAttention mechanism, while the second presents its extension into a multi-head perspective. The GAttention mechanism processes inputs from two distinct representations derived from the SA and CA mechanisms, respectively. Figure 1 illustrates the generation of these representations, starting from a sequence of encoded features $X_e \in \mathbb{R}^{d \times n}$, where d represents the number of encoding features, and n is the number of elements in the sequence.

As depicted in Figure 1, the process begins with the encoded features X_e sequence. The representations Q, K, and $V \in \mathbb{R}^{d \times n}$ are computed through a linear projection of the input sequence. This is achieved by multiplying the matrix X_e with the weight matrices W_Q, W_K, and $W_V \in \mathbb{R}^{n \times n}$, respectively. The SA and CA representations are calculated using these matrices, denoted by the matrices X_s and $X_c \in \mathbb{R}^{d \times n}$. For the SA representation, the scaled dot-product attention mechanism, as proposed in (Vaswani et al., 2017), is employed;

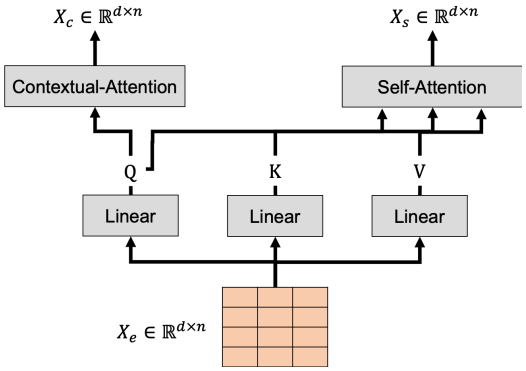

Figure 1: Detailed architecture of the pre-processing phase, to obtain the SA and CA representations.

Equation 1 details the process for computing the matrix X_s.

$$X_s = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

Concerning the CA representation, we use the CA mechanism proposed in (Yang et al., 2016). This mechanism receives the matrix Q as input and utilizes a context vector $u_h \in \mathbb{R}^d$, randomly initialized and subsequently learned during training. The vector u_h is a query vector to compute the attention values $\alpha_c \in \mathbb{R}^n$. This is achieved by measuring the similarity between the elements of the sequence Q and the application domain represented by u_h. Such similarity is calculated as detailed in Equation 2, involving the scalar dot product of u_h^T and Q; the resulting values are then normalized using a softmax function. In contrast to the CA mechanism proposed by (Yang et al., 2016), which employs a weighted sum between each attention value and its corresponding encoded features for the final sequence representation, our representation X_c, as shown in Equation 3, captures all the information from the attention values. This is done by performing an element-wise multiplication $\odot$ between each scalar of α_c and its corresponding encoding features in Q.

$$\alpha_c = softmax(u_h^T \cdot Q) \tag{2}$$

$$X_c = \alpha_c \odot Q \tag{3}$$

3.1 GAttention Unit

The main objective of our proposed GAttention mechanism is to generate a global context-aware representation $G \in \mathbb{R}^{d \times n}$, that combines the CA and SA mechanisms (represented by X_c and X_s

matrices). This integration aims to unify the contextual and internal relationships extracted by both mechanisms. Figure 2 illustrates the overall architecture of the GAttention mechanism. The GAttention mechanism is inspired by the weighted approach of combining different modalities in the Gated Multimodal Unit (GMU) network, as proposed by (Arevalo et al., 2020); our approach differs by conceptualizing the modalities as the CA and SA representations. Moreover, our GAttention mechanism extends the GMU network to handle feature sequences rather than merely feature vectors.

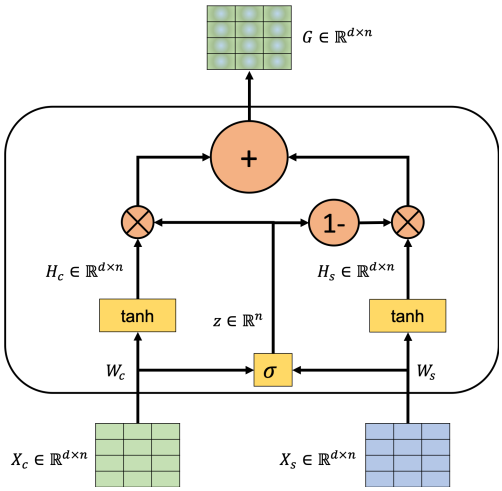

Figure 2: Architecture of the proposed GAttention mechanism; orange circles with the cross sign represent element-wise multiplication, the orange circle with + sign represent the summation of all the inputs, and orange circle with "1−" represents the function $f(s) = 1 - s$.

The GAttention mechanism processes the matrices X_c and X_s as inputs. As a first stage, a hidden representation of both matrices is obtained as follows:

$$H_c = tanh(W_c X_c) \tag{4}$$

$$H_s = tanh(W_s X_s) \tag{5}$$

Where $W_c \in \mathbb{R}^{d \times d}$ and $W_s \in \mathbb{R}^{d \times d}$ are learnable weights, *tanh* is the default activation function and, $H_c \in \mathbb{R}^{d \times n}$ and $H_s \in \mathbb{R}^{d \times n}$ are the resultant hidden representations. To regulate the relevance of both representations, the GAttention mechanism utilizes an internal feature vector $z \in \mathbb{R}^n$. This vector is computed as follows:

$$z = \sigma(W_z[X_c, X_s]) \tag{6}$$

where $[\cdot, \cdot]$ denotes the concatenation operator, $W_z \in \mathbb{R}^{d+d}$ are the learnable weights and σ represents the sigmoid activation function. The final output of the GAttention mechanism, denoted as $G \in \mathbb{R}^{d \times n}$, results from a convex combination of the hidden representations H_c and H_s, weighted by the values of z and $1 - z$ respectively, as delineated in Equation 7. This design enables the GAttention mechanism to determine the influence of each representation on the output selectively. It also implies that the weights in this convex combination will vary for each distinct input, owing to the dependency of z on X_c and X_s. As all these operations are differentiable, this model can be easily coupled with other neural network architectures and trained with stochastic gradient descent.

$$G = z \odot H_s + (1 - z) \odot H_c \tag{7}$$

3.2 Multi-Head GAttention

Building on the successful application of multiple attention mechanisms as demonstrated (Vaswani et al., 2017), we have extended the GAttention mechanism to incorporate a multi-head perspective. This adaptation enables the model to simultaneously attend to information from different representation subspaces at various positions. The mechanism processes k distinct representations from the SA and CA mechanisms, feeding them into k separate GAttention units. The outputs of these units are concatenated to form a new representation $C \in \mathbb{R}^{dk \times n}$, as detailed in Equation 8. Following this, C is subjected to a linear projection layer, which merges the concatenated information and reduces C's dimensions back to those of the original input X_e, as presented in Equation 9, where $W_l \in \mathbb{R}^{d \times dk}$ are learnable weights. The resulting representation, $L \in \mathbb{R}^{d \times n}$, is the culmination of the combined outputs from the k GAttention units. Figure 3 depicts the architecture of the multi-head GAttention mechanism.

$$C = [G_1, G_2, ..., G_k] \tag{8}$$

$$L = W_l C \tag{9}$$

3.3 Classification framework

The classification framework, which independently integrates the proposed GAttention and Multi-Head GAttention mechanisms for AL identification, starts with an encoded sequence X_e, obtained

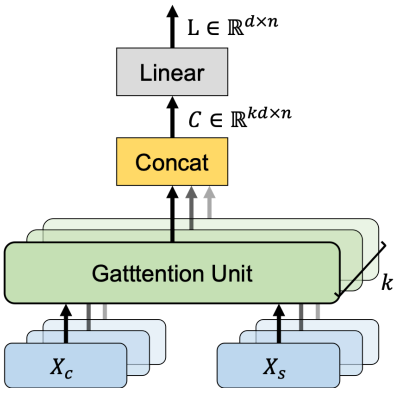

Figure 3: Architecture of the proposed Multi-Head GAttention mechanism, Figure inspired by (Vaswani et al., 2017), k represents the number of GAttention heads.

from either a Transformer model or a recurrent neural network. Specifically, in our experimental configurations, we use the last encoding layer of the Transformer as the encoded input sequence to our proposed GAttention configurations. This choice is motivated by the fact that the last encoding layer provides a comprehensive summary of the input through residual connections that aggregate information from all preceding layers (Clark et al., 2019). This encoded sequence is subsequently processed by the GAttention or Multi-Head GAttention mechanism. The resulting representation is then passed through a Transformer block consisting of add & norm layers and a Position-wise Feed-Forward Network, as described in (Vaswani et al., 2017). Subsequently, this representation is fed into an average pooling layer, which condenses the matrix into a vector. Finally, the vector is fed into the classification layers; two layers handle the final classification: a dense layer with a Rectified Linear Unit (ReLU) activation function and a fully connected softmax layer to obtain the class probabilities and get the final classification.

This classification framework is used for both binary and multi-class classification subtasks. However, one of the selected evaluation subtasks requires detecting multiple types of AL and identifying the corresponding targets. To address this, we adopt a multi-task learning approach in which the vector extracted by the average pooling layer is passed through two independent classification layers, each dedicated to a separate classification task. The loss function in this configuration is defined as the average of the cross-entropy losses computed for each task. The next subsection provides a detailed description of the evaluation subtasks.

4 Experiments

4.1 Evaluation Datasets

AL encompasses various types, with its main divisions categorized by the target and severity of insults (Mandl et al., 2019). Consequently, different collections and evaluation campaigns have explored distinct forms of AL in their study (Schmidt and Wiegand, 2017).

For our evaluation, we employ three English benchmark datasets: *SE 2019 T 6* (Marcos et al., 2019), *AMI 2018* (Fersini et al., 2018), and *HASOC 2019* (Mandl et al., 2019). These datasets were introduced in *SemEval-2019 Task 6*, the *Evalita 2018* Task on Automatic Misogyny Identification (AMI), and the *11th Forum for Information Retrieval Evaluation (FIRE)* as part of the Hate Speech and Offensive Content Identification (HASOC) shared task, respectively. Specifically, the *SE 2019 T 6* dataset addresses the task of offensive tweet classification, the *AMI 2018* dataset focuses on the detection of misogyny in tweets, while the *HASOC 2019* dataset addresses hate speech and offensive content detection in tweets.

These shared tasks comprise various subtasks, including binary and multi-class classification. While some tasks support different evaluation languages, our experiments focus exclusively on English and specifically on the first two subtasks defined in each shared task.

For the *SE 2019 Task 6* dataset, subtask A involves binary classification to detect offensive language, while subtask B focuses on predicting the type of offense, distinguishing between targeted and untargeted insults among the positive instances identified in subtask A. For the *AMI 2018* dataset, subtask A involves binary classification of misogynistic content, and subtask B extends the analysis to identify whether the target is a specific individual (active) or a group (passive), and to classify the type of misogyny among five categories: Stereotype & Objectification, Dominance, Derailing, Sexual Harassment, and Discredit. Lastly, in the *HASOC 2019* dataset, subtask A consists of binary classification of hate speech and offensive language, while subtask B performs a fine-grained categorization of the positive instances from subtask A into three classes: hate speech, offensive, and profane. The details regarding the distribution of the class labels in the training and testing sets of the evaluation datasets are presented in Appendix A.1.

4.2 Proposed Baselines

Evaluations were conducted on four Transformer encoder models across all proposed configurations, integrating the GAttention and Multi-Head GAttention mechanisms into the final encoding layer. For the multi-head variant, results are reported using 6 and 12 attention heads. As a first baseline, each model was fine-tuned on every subtask of the evaluation datasets without incorporating any additional mechanisms. As a second baseline, the SCA mechanism was replicated to assess the effectiveness of the proposed GAttention mechanism against a comparable approach that integrates SA and CA representations. This mechanism combines the SA and CA representations via an early fusion approach (Jarquín-Vásquez et al., 2021); similar to the GAttention and multi-head GAttention mechanisms, the SCA was integrated into the last encoding layer of Transformer models.

To evaluate the performance of the proposed GAttention mechanism against robust and contemporary benchmarks, three different approaches were replicated. The first corresponds to a state-of-the-art (SOTA) method proposed by Jarquín-Vásquez et al. (2024), which relies on re-training a Transformer model with two pre-training tasks specifically designed to constrain the model for AL detection. In particular, the HateBERT model (Caselli et al., 2021) was re-trained, following the two pre-training objectives jointly optimized with a weighted loss function, as indicated in the original paper. The re-training dataset and the experimental setup were strictly replicated to ensure fidelity. Further details can be found in the original publication (Jarquín-Vásquez et al., 2024).

To compare the effectiveness of GAttention configurations against contemporary LLMs, the second and third approaches employ *Llama-3.1-8B-Instruct*[1] as the backbone. Specifically, the second approach follows a Zero-shot (ZS) in-context learning paradigm, while the third approach involves fine-tuning the LLaMA model on each subtask of the evaluation datasets. Fine-tuning was performed using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). This parameter-efficient fine-tuning technique offers a favorable trade-off between performance and the number of trainable parameters. The LLaMA model was loaded in a 4-bit quantized version in both approaches.

These contemporary benchmarks were adopted as SOTA baselines to compare the effectiveness of the proposed GAttention configurations. In the results section, these baselines are referred to as CB_{RHB}, CB_{ZS}, and CB_{QLoRA}, respectively. The source code of the GAttention mechanism and its configurations is publicly available at the following link[2]. Appendix A.2 provides the implementation details, including text preprocessing, hyperparameter settings, the libraries used for developing all configurations, and the specifications of the Transformer models integrated in the proposed approaches.

4.3 Evaluation Metrics

For fair evaluation, all metrics were selected from those defined in the shared-task subtasks. In particular, all subtasks of the evaluation datasets were evaluated using the macro-averaged F_1 score, except for subtasks A and B in the *AMI 2018* dataset. In this case, subtask A was evaluated using accuracy. In contrast, subtask B was assessed based on the average of two macro-averaged F_1 scores: one for misogyny type classification and the other for target type classification. In our experiments, we report the mean results obtained from three randomly initialized training runs for each configuration.

5 Main Results

Table 1 presents the results of the GAttention mechanism and its various configurations across subtasks A and B for the three evaluation datasets. The first section of the table (rows 2–5) reports the performance of the four Transformer models fine-tuned in our experiments. The best results for subtask A were obtained with the ERNIE and DistilBERT models, while ERNIE and RoBERTa yielded the highest performance in subtask B.

To compare the performance of our GAttention mechanism against a similar approach that combines SA and CA representations, the second section (rows 6 - 9) presents the results of incorporating the SCA mechanism (Jarquín-Vásquez et al., 2021) into the final encoding layer of the Transformer models. Integrating the SCA mechanism yields a uniform enhancement across the three evaluation datasets. Overall, the best results of this integration were obtained with the ERNIE model. The third section (rows 10 - 13) shows the perfor-

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2] https://github.com/MasterHoracio/EMNLP-GAttention

#	AM	Model	SE T 6 (A)	SE T 6 (B)	AMI (A)	AMI (B)	HASOC (A)	HASOC (B)	#Params
2	–	$\text{BERT}_{\text{BASE}}$	0.794	0.693	0.695	0.450	0.728	0.529	109M
3	–	DistilBERT	0.808	0.687	0.678	0.456	0.741	0.530	66M
4	–	RoBERTa	0.805	0.717	0.681	0.494	0.737	0.545	124M
5	–	ERNIE	0.807	0.719	0.697	0.505	0.733	0.537	109M
6	SCA	$\text{BERT}_{\text{BASE}}$	0.805	0.708	0.704	0.471	0.735	0.534	114M
7	SCA	DistilBERT	0.811	0.704	0.697	0.468	0.743	0.543	70M
8	SCA	RoBERTa	0.809	0.725	0.701	0.501	0.738	0.547	130M
9	SCA	ERNIE	0.817	0.733	0.703	0.516	0.747	0.549	114M
10	GA	$\text{BERT}_{\text{BASE}}$	0.812	0.713	0.708	0.488	0.741	0.548	120M
11	GA	DistilBERT	0.820	0.721	0.705	0.473	0.752	0.553	77M
12	GA	RoBERTa	0.817	0.733	0.728	0.509	0.759	0.551	135M
13	GA	ERNIE	0.823	0.748	0.725	0.530	0.762	0.557	120M
14	MHG6	$\text{BERT}_{\text{BASE}}$	0.819	0.718	0.720	0.493	0.754	0.552	150M
15	MHG6	DistilBERT	0.827	0.726	0.723	0.482	0.763	0.558	86M
16	MHG6	RoBERTa	0.825	0.741	0.738	0.512	0.771	0.560	162M
17	MHG6	ERNIE	0.832	0.757	0.749	0.538	0.775	0.564	150M
18	MHG12	$\text{BERT}_{\text{BASE}}$	0.825	0.725	0.729	0.497	0.761	0.557	185M
19	MHG12	DistilBERT	0.832	0.734	0.731	0.489	0.770	0.562	106M
20	MHG12	RoBERTa	0.836	0.749	0.752	0.517	0.780	0.567	200M
21	MHG12	ERNIE	**0.843**	**0.769**	**0.757**	**0.544**	0.784	**0.571**	185M
22	–	BSTR	0.829	0.755	-	0.406	0.788	0.544	-
23	–	CB_{RHB}	0.834	0.747	0.720	0.525	**0.791**	0.564	109M
24	–	CB_{ZS}	0.562	0.481	0.617	0.319	0.625	0.347	4.5B
25	–	CB_{QLoRA}	0.748	0.615	0.704	0.445	0.713	0.486	4.7B

Table 1: Comparison results from our four baseline architectures and our proposed GAttention mechanism variants in subtasks A and B of the three evaluation datasets for the AL detection task. The results present the mean from three random training runs. The "AM" column stands for "Attention Mechanism", while "BSTR" and "CB" refer to the Best Shared-Task Result and Contemporary Benchmark, respectively.

mance enhancement in the models upon integrating the GAttention mechanism into the final encoding layer. As indicated by the results, including the GAttention mechanism consistently improved the performance of all models (rows 2 – 5 vs. rows 10 - 13), with a minimum improvement of 1.4% compared to their respective counterparts. Furthermore, when comparing the results of the SCA mechanism with those of the GAttention mechanism (rows 6 - 9 vs. rows 10 - 13), superior performance was achieved across all models in both subtasks of the three evaluation datasets. Overall, the best results of this integration were obtained with the ERNIE and RoBERTa models.

Sections 4 and 5 of Table 1 (rows 14 – 21) display the results of integrating the multi-head GAttention mechanism, utilizing 6 and 12 GAttention units in the final encoding layer of the models, respectively. When comparing the performance of the multi-head GAttention against a single GAttention unit (rows 10 - 13 vs. rows 14 - 17 and rows 10 - 13 vs. rows 18 - 21), a consistent improvement in the performance across all models within the three evaluation datasets is observed, highlighting the benefits of employing multiple attention heads. Additionally, when comparing the performance of 6 versus 12 GAttention units (rows 14 – 17 vs. rows 18 – 21), the results indicate that increasing the number of GAttention units can lead to performance improvements. Overall, the best results of our proposed configurations were obtained with the ERNIE model coupled with a multi-head GAttention mechanism with 12 GAttention units, across all classification scenarios.

Table 1 also compares the performance of our proposed configurations against the top-ranked systems from the shared tasks, as well as three contemporary benchmark approaches (Jarquín-Vásquez et al., 2024). When comparing our best configuration to the top-performing system for each subtask (row 21 vs. row 22), we observe a notable performance gain in subtask A of the *SE 2019 Task 6* dataset and subtask B of the *AMI 2018* dataset, while a slightly lower performance is observed in subtask A of the *HASOC 2019* dataset. Specifically, for subtask A of the *AMI 2018* dataset, the top team achieved an accuracy of 0.704, whereas our configuration obtained a higher accuracy of 0.768. Overall, our approach outperformed the top-ranked systems in five out of the six evaluated subtasks.

When comparing the results obtained from the contemporary benchmarks (rows 23 - 25), it can be observed that the CB_{RHB} approach consistently outperforms the LLM-based benchmarks (row 23 vs. rows 24 and 25) across all subtasks of the three evaluation datasets. Moreover, it achieves this superior performance with significantly fewer parameters, highlighting the advantages of leveraging pre-trained encoding models for AL classification, particularly in fine-grained tasks such as subtask B of the *AMI 2018* dataset. When compar-

ing our best configuration with the best contemporary benchmark (row 21 vs. row 23), our approach again achieves superior performance in five out of six subtasks. The only exception is subtask A of the *HASOC 2019* dataset, where performance is slightly lower. Upon examining the instances in this dataset, we hypothesize that this may be due to the presence of code-switching, which poses additional challenges. It is worth noting that the top-ranked systems and the contemporary benchmarks typically rely on LLMs, ensemble strategies, pre-training of Transformer models on large-scale corpora using pre-training tasks specifically designed for AL detection, and extensive data augmentation techniques (Wang et al., 2019; Saha et al., 2018; Liu et al., 2019a; Wiedemann et al., 2020; Jahan and Oussalah, 2023; Jarquín-Vásquez et al., 2024). In contrast, our approach solely integrates variants of the GAttention mechanism into existing pre-trained encoding models, without requiring additional pre-training or ensemble methods.

Finally, Table 1 compares the number of parameters across different baselines and GAttention configurations. As shown, incorporating GAttention variants increases the number of parameters by up to 69% when comparing the multi-head GAttention mechanism with 12 units to the baselines that do not include any additional mechanisms (rows 18 – 21 vs. rows 2 – 5). This increase translates into an average training time overhead of 5.6 seconds per epoch when using a GPU. Nevertheless, the performance gain in AL detection reaches up to 9%, which is relevant given the harmful impact of this type of content. For a more detailed analysis of the performance of the GAttention mechanism variants, Appendix A.3 presents a statistical significance analysis among the different proposed configurations.

6 Analysis

6.1 Relevance of the SA and CA Representations

NOTE: This subsection contains examples of language that may be offensive to some readers; these do not represent the authors' perspectives.

This subsection will analyze the z and $1 - z$ values, which correspond to the relevance of the SA and CA representations, respectively, to elucidate the proposed GAttention mechanism's adequate performance in AL detection. The examples presented in this analysis were exclusively obtained using the ERNIE model with a single GAttention unit. This analysis is conducted at the dataset level by identifying the most relevant words or expressions for each representation using subtask A across the three evaluation datasets.

As a first step, we calculated the z and $1 - z$ values for all tokens in the test instances post-training. Next, we averaged these values for each token and extracted the top 25 tokens with the highest averages. This process aimed to identify the most relevant words or expressions for the SA and CA representations within the datasets. Figure 4 presents word clouds for these top 25 tokens, showcasing both representations. The word clouds of the CA representation include offensive and harsh terms like 'f**k', 'b**ch', 'h**e', 's**t', etc. Additionally, words frequently associated with targets and stereotypes in hate speech, such as 't**ror**t', 'woman', 'drive', 'liberals', and 'conservatives', are also included. Conversely, the SA representation word clouds predominantly feature reflexive pronouns, verbs, conjunctions, and adjectives. This distinction underscores the CA mechanism's significant role in differentiating contextual and semantic elements in the AL detection task. To further support this analysis, Appendix A.4 contrasts the SA values produced by the GAttention mechanism with the attention weights from the ERNIE model. Additionally, Appendix A.5 provides an instance-level analysis of the SA and CA representation values.

6.2 Error Analysis of Misclassified Instances

To assess the robustness of the proposed GAttention configurations, this subsection presents an error analysis based on misclassified instances by the pre-trained ERNIE model (without additional mechanisms) but correctly classified when integrating the best-performing configuration, the Multi-Head GAttention with 12 heads. For this analysis, ERNIE was selected as the backbone due to its solid performance across all subtasks of the three evaluation datasets. Table 2 reports the analysis of these instances for subtask A across all datasets. A value of 1 in the label column denotes the positive class of each dataset.

As shown in Table 2, integrating Multi-Head GAttention improves the model's ability to capture subtle or context-dependent offensiveness. For example, in instances 1 and 2, the expression "hysterical woman" appears. While the baseline ERNIE model failed to identify its derogatory use in the

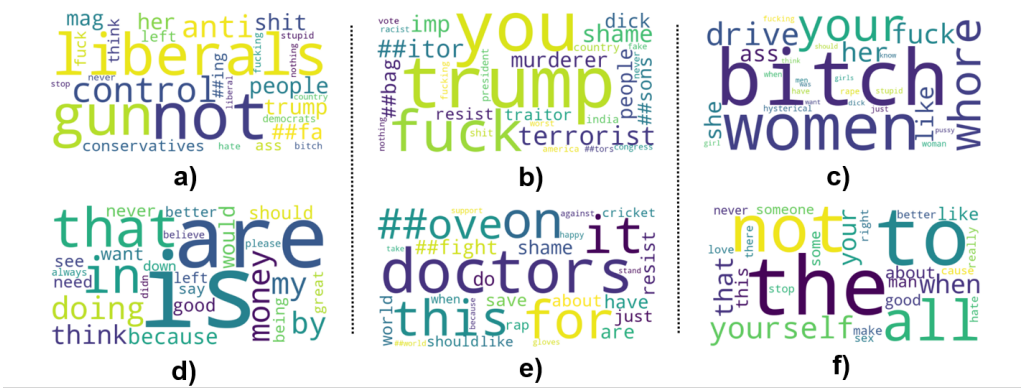

Figure 4: Word clouds of the highest sigma values for CA (a – c) and SA (d – f) mechanisms. Each vertical block corresponds to an evaluation dataset, ordered from left to right: SEM 2019 Task 6, HASOC 2019, and AMI 2018.

first case, the Multi-Head GAttention mechanism successfully captured the offensive intent. Similarly, the mechanism improved the classification of cases where no explicit slurs were present, but the overall context conveyed offensive meaning. In instance 3, the comment disparages a woman by implying she is unbearable to watch or listen to, and in instance 5, the text reflects religious intolerance and homophobia by rejecting interfaith and same-sex marriages. Furthermore, the mechanism also handled instances containing potentially offensive words used in non-offensive contexts. For example, in instance 4, "are you f**king serious?" functions as an intensifier rather than an insult, while in instance 6, the phrase "watching flies f*ck" is metaphorical and intended to emphasize boredom rather than offense. Appendix A.6 presents a complementary error analysis, highlighting areas for improvement by identifying instances that none of the proposed GAttention configurations were able to classify.

#	Label	Text	Dataset
1	1	@user not really. If anything, I would use 'a low IQ, hysterical woman of zero talent'.	AMI
2	0	@user exactly, but I am just a hysterical woman that sasse hates.	AMI
3	1	#person is so full of herself . . . she is just as painful to watch as to listen to . . .	SEM
4	0	@user are you f**king serious? url	SEM
5	1	unless you're marrying someone who's not muslim or someone of the same sex, 99% all parents will accept it at some point	HAS
6	0	watching cricket is like watching flies f*ck.	HAS

Table 2: Examples of misclassifications by the ERNIE model (without GAttention), corrected through the integration of Multi-Head GAttention with 12 heads in the final encoding layer. The column "label" corresponds to the ground truth.

7 Conclusions and Future Work

One of the main challenges in using current attention mechanisms is the loss of contextual or internal information among the tokens in a text sequence. To address this, we proposed the GAttention and the multi-head GAttention mechanisms. These mechanisms combine the SA and CA representations through a weighted approach based on their contributions to the text classification tasks. Specifically, we integrated the GAttention mechanism into four pre-trained Transformer models and evaluated its performance in the AL detection task. This task was chosen due to the highly context-dependent interpretation of words, making it an ideal testbed for assessing the efficacy of the proposed mechanism. The results were encouraging, demonstrating improvements across all models and evaluation datasets, which included both binary and multi-class classification tasks targeting different types and targets of AL. Notably, the best performance was achieved using the multi-head GAttention mechanism with 12 heads integrated into the ERNIE model. Furthermore, our analysis revealed a dynamic adaptation in the relevance of the SA and CA representations, highlighting the mechanism's effectiveness in detecting AL.

As future work, we consider 1) evaluating the GAttention mechanism in tasks where context interpretation is critical for classification, such as in detecting depression, deception, and fake news; 2) extending the application of the GAttention mechanism to AL detection in memes, by integrating attention representations from different modalities; and 3) developing novel loss functions that incorporate the context vector of the CA mechanism into the classification process.

Limitations

Considering the nature of the evaluation datasets, where labels are manually annotated, social biases may be inherent in the annotators' judgments. As a result, the various model configurations could be learning these biases, potentially leading to errors when applied to data of a different nature. Furthermore, the proposed GAttention and multi-head GAttention mechanisms were integrated and evaluated with the BERT, ERNIE, RoBERTa, and DistilBERT models; their integration into alternative pre-trained models may vary their overall performance in the AL detection task. Finally, due to the large amounts of training data required by these types of architectures to perform effectively, limited data access may result in suboptimal learning of the task at hand, thus constraining the model's capabilities.

Acknowledgements

Horacio Jarquín-Vásquez thanks the support obtained from the "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 funded under the NextGenerationEU programme (PI: Viviana Patti). Mario Ezra Aragón thanks the support obtained from MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF), Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF), and the Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and by the FSE+.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Mario Ezra Aragón, Horacio Jesús Jarquín-Vásquez, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Helena Gómez-Adorno, Juan Pablo Posadas-Durán, and Gemma Bel-Enguix. 2020. Overview of MEX-A3T at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In *Proceedings ofI berLEF 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 222–235. CEUR-WS.org.

John Arevalo, Thamar Solorio, Manuel Montes, and Fabio González. 2020. Gated multimodal networks. *Neural Computing and Applications*, pages 1433–3058.

G. Brauwers and F. Frasincar. 2023. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(04):3279–3298.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Sneha Chaudhari, Gungor Polatkan, R. Ramanath, and Varun Mithal. 2020. An attentive survey of attention models. *Association for Computing Machinery*, 37.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 276–286.

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. 2020. Attentional feature fusion. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3559–3568.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. of 11th Intl Conference on Web and Social Media*, pages 512–515.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. *Statistical Significance Tests*, pages 9–21. Springer International Publishing, Cham.

Ashok Kumar Durairaj and Anandan Chinnalagu. 2021. Transformer based contextual model for sentiment analysis of customer reviews: A fine-tuned bert. *International Journal of Advanced Computer Science and Applications*, 12(11).

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 63–74. CEUR-WS.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *Proc. of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263 of *CEUR Workshop Proceedings*, pages 107–114. CEUR-WS.org.

Dichao Hu. 2019. An introductory survey on attention mechanisms in nlp problems. In *IntelliSys*, pages 1–9.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*.

Horacio Jarquín-Vásquez, Hugo Jair Escalante, and Manuel Montes. 2021. Self-contextualized attention for abusive language identification. In *Proc. of 9th International Workshop on Natural Language Processing for Social Media*, pages 103–112.

Horacio Jarquín-Vásquez, Hugo Jair Escalante, and Manuel Montes y Gómez. 2024. Enhancing abusive language detection: A domain-adapted approach leveraging bert pre-training tasks. *Pattern Recognition Letters*, 186:361–368.

Horacio Jarquın-Vasquez, Hugo Jair Escalante, Manuel Montes-y Gomez, and Fabio A. Gonzalez. 2024. Gha: a gated hierarchical attention mechanism for the detection of abusive language in social media. *IEEE Transactions on Affective Computing*, pages 1–14.

Yuanfu Li, Yifan Chen, Haonan Shao, and Huisheng Zhang. 2023. A novel dual attention mechanism combined with knowledge for remaining useful life prediction based on gated recurrent units. *Reliability Engineering & System Safety*, 239:109514.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *ArXiv*, abs/2106.04554.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proc. of the 13th International Workshop on Semantic Evaluation*, pages 87–91.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, page 14–17.

Zampieri Marcos, Nakov Preslav, Rosenthal Sara, Atanasova Pepa, Karadzhov Georgi, Mubarak Hamdy, Derczynski Leon, Pitenis Zeses, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.

Zampieri Marcos, Malmasi Shervin, Nakov Preslav, Rosenthal Sara, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proc. of 13th International Workshop on Semantic Evaluation*, pages 75–86.

Raymond T Mutanga, Nalindren Naicker, and Oludayo O Olugbara. 2020. Hate speech detection in twitter using transformer methods. *International Journal of Advanced Computer Science and Applications*, 11(9).

Thanh Nguyen, Campbell Wilson, and Janis Dalins. 2024. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. In *32th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.*, pages 613–618.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Ashay Patel, Petru-Daniel Tudosiu, Walter Hugo Lopez Pinaya, Gary Cook, Vicky Goh, Sebastien Ourselin, and M. Jorge Cardoso. 2022. Cross attention transformers for multi-modal unsupervised whole-body pet anomaly detection. In *Deep Generative Models*, pages 14–23, Cham. Springer Nature Switzerland.

M. Ramprasath, K. Dhanasekaran, T. Karthick, R. Velumani, and P. Sudhakaran. 2022. An extensive study on pretrained models for natural language processing based on transformers. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 382–389.

Maria Luisa Ripoll, Fadi Hassan, Joseph Attieh, Guillem Collell, and Abdessalam Bouchekif. 2022. Multi-lingual contextual hate speech detection using transformer-based ensembles. In *Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation*, pages 552–562.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. 5th Intl Workshop on Natural Language Processing for Social Media*, pages 1–10.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 1–11.

Boyang Wan, Wenhui Jiang, Yuming Fang, Wenying Wen, and Hantao Liu. 2022. Dual-stream self-attention network for image captioning. *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5.

Bin Wang, Yunxia Ding, Shengyan Liu, and Xiaobing Zhou. 2019. Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. In *Working Notes of FIRE 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 191–198. CEUR-WS.org.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

Ben You, XiaoHong Li, QiXuan Peng, and RuiHong Li. 2022. Using multi-level attention based on concept embedding enrichen short text to classification. In *Intelligent Information Processing XI*, pages 148–155, Cham. Springer International Publishing.

Shenjian Zhao and Zhihua Zhang. 2018. Attention-via-attention neural machine translation. *Proceedings of AAAI*, 32(1).

A Appendix

A.1 Distribution of Class Labels in the Evaluation Datasets

Table 3 shows the class distribution for subtask A in each dataset. As can be observed, all datasets exhibit class imbalance, with the negative class typically having a higher number of instances.

Dataset	Training		Testing	
	OFF	NOT	OFF	NOT
SE 2019 T6	4,400	8,840	240	620
AMI 2018	1,785	2,215	460	540
HASOC 2019	3,591	2,261	865	288

Table 3: Distribution of training and testing data across the evaluation datasets in subtask A. The columns 'OFF' and 'NOT' refer to the positive and negative classes, respectively.

On the other hand, Table 4 presents the class distribution for subtask B in each dataset. Blocks 2 and 3 correspond to the same dataset, *AMI 2018*; however, the distributions are shown separately since subtask B is designed as a multi-task setting. As can be observed, all datasets exhibit class imbalance, which is particularly evident in the fine-grained classification of misogyny types in the *AMI 2018* dataset. For the sake of reproducibility, each training dataset was used to independently train every baseline model as well as each GAttention configuration, and the results reported in Table 1 were consistently obtained using the corresponding test sets.

Dataset	Class	Training	Testing
SE 2019 T6	Targeted	3,876	213
	Untargeted	524	27
AMI 2018	Discredit	1,014	141
	SH	352	44
	Derailing	92	11
	S&O	179	140
	Dominance	148	124
AMI 2018	Active	1,058	401
	Passive	727	59
HASOC 2019	Hate Speech	1,143	124
	Offensive	667	71
	Profane	451	93

Table 4: Distribution of training and testing data across the evaluation datasets in subtask B. The acronyms SH and S&O refer to Sexual Harassment and Stereotype & Objectification, respectively.

A.2 Implementation Details

Various text preprocessing operations were employed to prepare the data. To minimize biases, user mentions and links were replaced with generic tokens <*user*> and <*url*>, respectively. Hashtag segmentation into words was performed using the Ekphrasis library[3]. Additionally, all text was converted to lowercase, and non-alphabetical characters were removed.

[3] https://github.com/cbaziotis/ekphrasis

Four Transformer encoder models were used to evaluate the proposed architectures and configurations: 1) BERT base uncased[4], introduced by Devlin et al. (2019), which consists of 12 encoder layers, each with 12 attention heads and a hidden size of 768; 2) DistilBERT base uncased[5], introduced by Sanh et al. (2019), a distilled version of BERT that preserves much of its language understanding capability while reducing model size and inference time; 3) RoBERTa (Robustly Optimized BERT Approach) base[6], introduced by Liu et al. (2019b), which builds upon BERT with larger mini-batches, more training data, and dynamic masking, removing the next sentence prediction objective and focusing on a more robust masked language modeling approach; and 4) ERNIE 2.0 base in English[7], introduced by Sun et al. (2019), which extends the Transformer architecture by incorporating knowledge masking strategies during pre-training to capture lexical, syntactic, and semantic information more effectively, producing enriched contextual representations.

All architectures and experiments were implemented using PyTorch[8]. The experiments were conducted on a server equipped with an AMD EPYC 7742 64-core CPU, 512 GB of DDR5 RAM, and four NVIDIA A40 GPUs with 48 GB of GDDR6 memory each. All models were trained using the Adam optimizer and the cross-entropy loss function. Table 5 summarizes the hyperparameters used for subtasks A and B in the three evaluation datasets for all the proposed GAttention configurations. These hyperparameters were selected via grid search based on the macro average F_1 score, exploring the following values: number of epochs $\in \{1, 2, \ldots, 5\}$, batch size $\in \{12, 14, 16, \ldots, 32\}$, and learning rate $\alpha \in \{1e{-}5, 1.5e{-}5, 2e{-}5, \ldots, 5e{-}5\}$.

Regarding the replicated contemporary benchmarks, for CB_{RHB}, as reported in Jarquín-Vásquez et al. (2024), both the retraining and fine-tuning phases employed the Adam optimizer with a learning rate of 5e-5. Specifically, the retrained HateBERT model was trained for 2 epochs with a batch size of 32. In contrast, the fine-tuning models for the subtasks of the three evaluation datasets were trained for 3 epochs with a batch size of 16. For the CB_{ZS} and CB_{QLoRA} benchmarks, inference was performed with a temperature of 0.8, a maximum generation length of 8 tokens, and sampling ($p = 0.9$) as the text generation strategy. Concerning the QLoRA-based benchmark, the models were trained for 2 epochs, except for the models corresponding to subtask B of the *AMI 2018* and *HASOC 2019* datasets, which were trained for 3 epochs. In all cases, a batch size of 8 was used, including a rank of 32, $\alpha = 16$, and a dropout rate of 0.05. Table 6 presents the prompt used in both approaches. For each subtask, the corresponding categories were included, and for the training sets of the CB_{QLoRA} benchmark, the respective labels were also added.

Dataset	Subtask	Epochs	Batch	α
SEM	A	2	32	4e-5
SEM	B	4	16	1.5e-5
AMI	A	3	18	2e-5
AMI	B	3	12	2.5e-5
HAS	A	2	24	5e-5
HAS	B	4	24	1.5e-5

Table 5: Summary of the hyperparameter configurations employed for subtasks A and B in the three evaluation datasets for all the GAttention configurations.

```
System: You are an expert in analyzing
abusive and offensive content from
social media platforms.
User: The following message originates
from social media and may contain
abusive or offensive content. For
research purposes, classify the
input text into one of  the
following categories:

[Provide a detailed description of
each category using the format:
- 'category 1': definition
...
- 'category n': definition]

Return only the label as a string:
'category 1', 'category 2', ..., or
'category n'.

text: {ADD INSTANCE TEXT}
label: [ADD LABEL ONLY FOR THE
    TRAINING SETS OF THE QLoRA
    BENCHMARK]
```

Table 6: Prompt template used in our CB_{ZS} and CB_{QLoRA} benchmark approaches.

[4] https://huggingface.co/bert-base-uncased

[5] https://huggingface.co/distilbert-base-uncased

[6] https://huggingface.co/FacebookAI/roberta-base

[7] https://huggingface.co/docs/transformers/model_doc/ernie

[8] https://pytorch.org/

A.3 Significance Testing of GAttention Variants

For a more detailed analysis of the performance of the GAttention mechanism variants, we employed McNemar's statistical test (Dror et al., 2020) to compare the results of our best configuration (12 Multi-Head GAttention mechanism coupled with ERNIE) with those of the proposed configurations using ERNIE for subtask A across the three evaluation datasets. Table 7 illustrates this comparison. Specifically, column 2 compares our best configuration with ERNIE's fine-tuning. In contrast, columns 3 to 5 contrast it with integrating the SCA, GA, and MH6-GA mechanisms coupled with ERNIE. Finally, column 6 compares our best configuration with the best-performing contemporary benchmark. The symbol '=' means not significantly different ($p > 0.05$), '*' means significantly different ($p < 0.05$), '**' means very significantly different ($p < 0.01$), and '***' means highly significantly different ($p < 0.001$). The results indicate that the proposed 12 Multi-Head GAttention mechanism configuration significantly differs from its simpler variants.

Dataset	ERIE	SCA	GA	MH6-GA	CB_{RHB}
SEM	***	***	***	*	*
AMI	***	***	***	*	***
HAS	***	***	***	*	=

Table 7: Pairwise significance differences between the 12 Multi-Head Gattention mechanism coupled with the ERNIE model and the proposed configurations using ERNIE as their backbone, evaluated with McNemar's test based on the macro-average F_1 score across all datasets in subtask A.

A.4 Analysis of Attention Values of the ERNIE Model

To shed light on the effectiveness of the proposed GAttention mechanism, this subsection presents an analysis of the words with the highest attention values from the ERNIE model across the three evaluation datasets without integrating the proposed GAttention mechanism. This comparison aims to contrast the most relevant words identified by analyzing the z and $1 - z$ values corresponding to the SA and CA representations, respectively. For this analysis, we calculated the average attention values from the last encoding layer of the ERNIE model for all tokens in the post-training test instances. Subsequently, we averaged these values for each token and selected the top 25 tokens with the highest averages. This process aimed to identify the most relevant words or expressions identified by the ERNIE model's attention values within the three evaluation datasets.

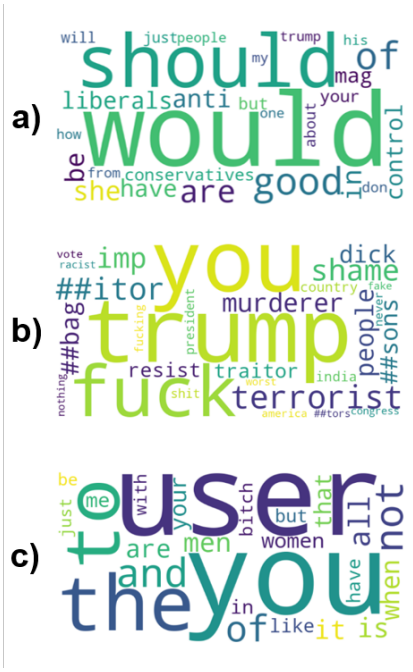

Figure 5: Word clouds of the highest SA values, using the ERNIE model (without integrating the GAttention mechanism). Sections a, b, and c show the word clouds from the SEM 2019 Task 6, HASOC 2019, and AMI 2018 datasets, respectively.

Figure 5 presents word clouds for these top 25 tokens, with sections a, b, and c illustrating the word clouds derived from the SEM 2019 task 6, HASOC 2019, and AMI 2018 datasets, respectively. The word clouds reveal that the most relevant tokens, as determined by the attention values of the ERNIE model, predominantly consist of commonly used words such as pronouns and prepositions. However, the occurrence of potentially offensive words is notable, including terms such as 'f**k', 'b**ch', 'd**k', 'tr**tor', and 'racist'. Additionally, the attention mechanism also captures words commonly targeted in potentially offensive texts, such as 'liberal', 'people', 'conservatives', 'men', and 'women'. In comparison, the words captured by the z and $1 - z$ values of the GAttention mechanism appear to be more segmented. This can be observed in Figure 4, where words related to the AL detection task are distinctly identified by the CA mechanism. In contrast, commonly used words are captured by the SA mechanism. This distinction could significantly aid in discriminating offensive instances and highlights the efficacy

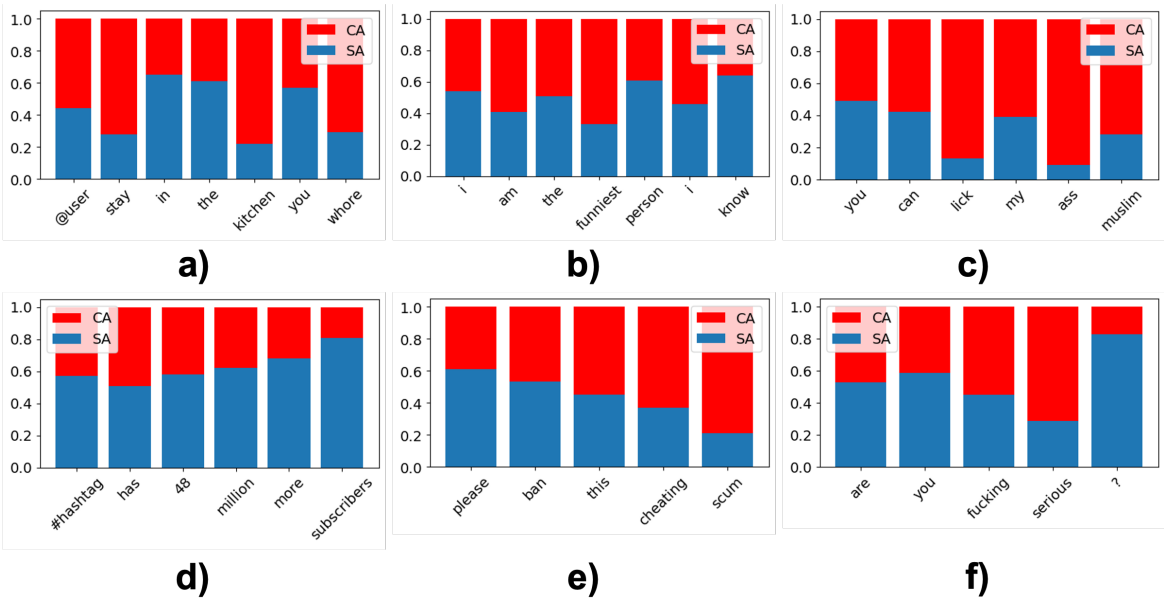

Figure 6: Analysis of sigma values in both offensive and non-offensive instances, which were correctly classified using our best configuration. Each graph illustrates the relevance of CA and SA mechanisms in each word of the instances. Sections a–b, c–d, and e–f correspond to the AMI 2018, HASOC 2019, and SEM 2019 Task 6 datasets, respectively. All instances were randomly selected from the correctly classified instances.

of the proposed GAttention mechanism configurations in AL detection.

A.5 Instance Level Analysis

For this analysis, we plotted the relevance values of the SA and CA representations for all words[9] in a text sequence. This approach allowed us to observe the nuanced behavior of the SA and CA representations within their respective contexts. Figure 6 displays these values across six instances, each correctly classified as either offensive (sections a, c, and e) or non-offensive (sections b, d, and f) in subtask A across the three evaluation datasets. As depicted in the Figure, each instance exhibits a significant variability in its z values, influenced by the context. For instance, in the offensive examples, a higher relevance is noted in the CA representation for words with potentially offensive connotations, such as 'w**re', 'kitchen', '*ss', and 's**m'. Notably, even in non-offensive instances, CA values were elevated for words with positive connotations, like the word 'funniest' in Section b. Conversely, the SA representation predominantly highlighted its relevance in pronouns, verbs, and conjunctions. This pattern underscores the efficacy of the GAttention mechanism, leveraging both representations to improve the detection of AL.

[9]For words segmented into multiple tokens, we manually averaged their values.

A.6 Limitations of the GAttention Mechanism

To gain insights into the limitations of the proposed GAttention configurations, this section presents an error analysis based on three specific settings: the single-head GAttention mechanism and the Multi-Head GAttention mechanism with 6 and 12 heads, all integrated into the final encoding layer of the ERNIE model. For this analysis, we considered an instance to be misclassified when none of the three configurations was able to predict it correctly. The analysis focuses on Subtask A across the three evaluation datasets.

Table 8 provides representative examples of such instances. The results reveal that the GAttention configurations struggle with subtle expressions whose literal meaning may appear non-offensive but are used in a derogatory manner. For example, in instance 1, the word "skank", in instance 5, the phrase "alien head", and in instance 7, the word "fanatic" were all misclassified. Moreover, the mechanisms also fail to detect offensive expressions disguised as acronyms, as illustrated by instance 2, where "stfu" (meaning "shut the f**k up") requires contextual knowledge to be correctly interpreted as offensive.

Additionally, there are cases where non-misogynistic content is incorrectly classified. In instance 3, despite containing expressions such as "b*tch *ss friends", the overall intent is not offen-

sive, as the phrase is used in an emphatic manner, further softened by the introductory "I miss my" phrase. Similarly, in instance 4, the message is offensive, but not misogynistic, since it is not directed at a woman. Finally, we observe instances where the ground-truth labels themselves introduce ambiguity. For example, in instances 6 and 8, the expressions "n*sty" and "s**pid boy" are overtly offensive, yet their labels are marked as non-offensive. Such inconsistencies may reflect annotation biases, which could partly explain the discrepancies between model predictions and the reference labels.

#	Label	Text	Dataset
1	1	@user my little commie skank look at you!!! I'm so proud	AMI
2	1	@user why don't u stfu, that is exactly what's wrong with this country	AMI
3	0	I miss my s**pid pretty stank d*mb wh*re skank trick b*tch *ss friends	AMI
4	0	@user you are just a complete whacko traitor! f*ck you!	AMI
5	1	by reading your comment, it seems to me that you have an alien head.	SEM
6	0	@user she is just n*sty	SEM
7	1	tell that fanatic not to use any car or buses or aeroplane which use muslim petrol diesel	HAS
8	0	these lists of banned substances have been around forever, s**pid boy	HAS

Table 8: Examples of misclassifications by the GAttention configurations using the ERNIE model. The column "label" corresponds to the ground truth.