applied sciences

MDPI

Article

Regret and Hope on Transformers: An Analysis of Transformers on Regret and Hope Speech Detection Datasets

Grigori Sidorov †, Fazlourrahman Balouchzahi *,†, Sabur Butt and Alexander Gelbukh

Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City 07738, Mexico; sidorov@cic.ipn.mx (G.S.); sbutt2021@cic.ipn.mx (S.B.); gelbukh@cic.ipn.mx (A.G.)
* Correspondence: fbalouchzahi2021@cic.ipn.mx
† These authors contributed equally to this work.

Abstract: In this paper, we analyzed the performance of different transformer models for regret and hope speech detection on two novel datasets. For the regret detection task, we compared the averaged macro-scores of the transformer models to the previous state-of-the-art results. We found that the transformer models outperformed the previous approaches. Specifically, the roberta-based model achieved the highest averaged macro F1-score of 0.83, beating the previous state-of-the-art score of 0.76. For the hope speech detection task, the bert-based, uncased model achieved the highest averaged-macro F1-score of 0.72 among the transformer models. However, the specific performance of each model varied slightly depending on the task and dataset. Our findings highlight the effectiveness of transformer models for hope speech and regret detection tasks, and the importance of considering the effects of context, specific transformer architectures, and pre-training on their performance.

Keywords: regret detection; hope speech detection; transformers; text classification; contextual embedding

Citation: Sidorov, G.; Balouchzahi, F.; Butt, S.; Gelbukh, A. Regret and Hope on Transformers: An Analysis of Transformers on Regret and Hope Speech Detection Datasets. *Appl. Sci.* **2023**, *13*, 3983. https://doi.org/10.3390/app13063983

Academic Editor: Yu-Dong Zhang

Received: 7 February 2023
Revised: 3 March 2023
Accepted: 16 March 2023
Published: 21 March 2023

1. Introduction

Regret is often experienced when a decision's outcome does not match the desired or expected result. This emotion is often sparked by past actions or inaction taken or not taken by an individual. According to [1,2], the source of regret can determine its intensity and duration. They argue that regret resulting from an action can cause immediate pain, but may not have long-term effects, while failing to take action can lead to prolonged and problematic regret. However, ref. [3] suggests that regret can lead to a reevaluation of the decision and a desire to make a different choice if given the opportunity. Therefore, understanding the cause of regret is crucial in examining its effects on behavior and decision-making.

In a recent study, Balouchzahi et al. [4] introduced the task of regret detection and domain identification in English texts from Reddit posts. They created an annotated dataset for this task and modeled it by categorizing posts into three classes: "Action" (regret caused by action), "Inaction" (regret caused by inaction), and "No regret". Additionally, they identified the domain of the text into five predefined categories: "Education", "Health", "Career and Finance", "Romance and Relationships", and "Other domains". Using machine learning and deep learning models, they found that Neural Networks (NNs) built with GloVe performed best in all the experiments and obtained state-of-the-art results, with averaged macro F1-scores of 0.72 for regret detection and 0.63 for domain identification subtasks.

Therefore, in [4], the effectiveness of GloVe embedding for representing the meaning and context of words was demonstrated in regret detection and domain identification. While GloVe performed well in this study, it is worth considering the use of transformer-based models for this task, as they have shown strong performances in a wide range of natural language-processing tasks.

Unlike GloVe, which generates fixed-length vector representations for words, based on their overall frequency in a large text corpus [5], transformers generate contextualized word embeddings, which means that the representation of a word can vary, depending on the context in which it appears [6]. This can be particularly useful for tasks such as regret detection, where the meaning of words can rely heavily on the context in which they appear.

In addition to generating contextualized word embeddings, transformers are highly scalable and can handle long input sequences, enabling them to consider a large amount of context when classifying texts. They also use attention mechanisms, which allow the model to weigh different input words differently based on their relevance to the task [7]. Overall, the use of transformers, with their ability to generate contextualized word embeddings and utilize attention mechanisms, may provide insight into how contextual embeddings can improve the state-of-the-art results for regret detection.

Since no experiments and results were available on transformer methods for fine-grained hope and regret detection in the text, our research revolved around the question of how well transformer methods perform in detecting fine-grained expressions of regret and hope. In this paper, we explored the use of transformer-based models for the task of regret detection, building on the experiments presented in Balouchzahi et al. [4]. We conducted experiments with different transformer architectures and compared their results to the baseline data, as well as state-of-the-art results presented in [4].

The ReDDIT [4] dataset consists of two subtasks, (i) regret detection and (ii) domain identification, however, in this paper, we only tested models on the first subtask (regret detection). Additionally, we compared the performance of these models with their results reported on a multi-class hope speech detection dataset called PolyHope [7]. This allows us to evaluate the effectiveness of different transformer architectures on various context-dependent text classification tasks, and to identify factors that contribute to the best performance. Our experiments provide new insights into using pre-trained language models and transformers for hope speech and regret detection, and demonstrate their potential for these and related tasks.

This comparison showed that the transformer models generally perform well for both tasks. The roberta-based model achieved the highest scores in the regret detection task and the bert-based, uncased model achieved the highest scores in the hope speech detection task. However, the performance of transformers was higher in the ReDDIT dataset than the PolyHope dataset, because the task of hope speech detection was more challenging due to more classes, and also shorter texts that lack sufficient context when compared to long texts in the ReDDIT dataset. The results also highlighted the importance of context, specific transformer architectures, and pre-training in the models' performance of tasks involving text classification.

The main contributions of this paper are as follows:

- Exploration of transformer-based models for the task of regret detection, and comparison of their performances to the baseline data and previous state-of-the-art results;
- Comparison of the performance of transformer-based models on the ReDDIT dataset, utilizing a multi-class hope speech detection dataset to evaluate their effectiveness, in terms of different context-dependent text classification tasks;
- Provision of new insights into the use of pre-trained language models and transformers for hope speech and regret detection tasks;
- A discussion of the limitations of the present study.

2. Related Work

Recent advances in transformer-based models have led to significant improvements in the performance of many natural language-processing tasks, including multi-class text classification. Ref. [8] introduced the transformer model, a NN architecture that uses self-attention mechanisms to process input sequences. The transformer model has been applied to a wide range of NLP tasks, including language translation, language modeling, and

summarization, and has been shown to outperform previous state-of-the-art models in multiple benchmark datasets.

In their paper, Vaswani et al. [9] introduced Bidirectional Encoder Representations from Transformers (BERT), a transformer-based model for natural language-processing tasks. Through a series of experiments on various benchmark datasets, the authors demonstrated the effectiveness of BERT in improving the performance of NLP tasks. The model was pre-trained on the BooksCorpus [10] and English Wikipedia datasets and was then fine-tuned on various NLP tasks, which resulted in significant improvements, including improving the GLUE [11] score to 80.5%, increasing the Multi-Genre Natural Language Inference (MultiNLI) accuracy to 86.7%, and achieving a Test F1 of 93.2 on SQuAD v1.1 and 83.1 on SQuAD v2.0. [12]. This demonstrates the effectiveness of BERT in tackling a variety of NLP tasks, and highlights its potential as a powerful tool in the field.

Köhler et al. [13] conducted the CheckThat! Lab 2022 Task 3, a workshop focused on evaluating technologies for determining the veracity of news articles. The task was designed as a multi-class classification problem. It was conducted in English and German, with the German subtask requiring a cross-lingual approach, and the English subtask being a monolingual task. The datasets for the task included approximately 12,000 English articles and 3000 German articles, which were expertly evaluated in order to ensure high data quality. A total of 25 teams submitted successful runs for the English subtask, with a total of 8 teams for the German subtask.

In both subtasks, many teams used transformer-based models, particularly BERT and its variants, with 12 teams using BERT, 6 teams using RoBERTa, and 8 teams using other BERT versions. A significant number of teams also used feature-based supervised learning approaches, such as SVMs (used by 10 teams), logistic regression (used by 9 teams), random forests (used by 8 teams), and Naive Bayes (used by 7 teams). Some teams employed additional data or ensemble methods in their approaches as well. In addition, most teams used a multilingual language model in the cross-lingual subtask, or had the data translated into another language. Overall, transformer-based models were among the top performers in the task, with the best system in terms of the monolingual subtask (English) achieving a macro F1-score of 0.339, and the best system in terms of the cross-lingual task (German) achieving a macro F1-score of 0.242.

While transformers have done well with high-resource languages, we have also seen much progress in low-resource languages. Low-resource languages have limited data available for natural language-processing tasks. These languages often have smaller speaking communities, fewer written resources, and less research focused on them than on high-resource languages, such as English, Chinese, and Spanish. This lack of data and resources can make it challenging to develop accurate and efficient NLP models for low-resource languages [14]. However, with the advent of pre-trained transformer models, it has become increasingly feasible to train NLP models on low-resource languages.

The HASOC (Hate Speech and Offensive Content Identification) shared task [15] aimed to develop and evaluate tools for detecting hate speech and other offensive content in social media. The subtasks for English, Hindi, and Marathi include two classification tasks: a binary classification task, in which systems classify tweets as containing Hateful or Offensive content (HOF) or Not (NOT), as well as a fine-grained classification task for English and Hindi, with three classes: HATE (hate speech), OFFENSIVE (offensive content), and PROFANITY (profanity). The data for this competition was gathered from Twitter and annotated by experts. The participating teams utilized a range of technologies in this shared task. Several teams used transformer-based models, including BERT, XLM-R [16], or Multilingual-BERT, and these models generally achieved high scores in the classification tasks. Additionally, some teams utilized machine learning models and obtained good results. However, transformer-based models appear to be particularly successful when used for this task, with the best-performing algorithms being mainly transformer variants. The overall highest F1 scores for the classification of hate speech in Marathi, Hindi, and English were 0.91, 0.78, and 0.83, respectively.

For instance, one team employed a GCN-based approach in the Hindi subtask, which involves defining tweets and words as nodes and connecting them with sliding window relationships. They obtained a Macro F1 score of 0.8215 for Task A. Another team used a soft-voting ensemble of four different models in the English subtask, and obtained a Macro F1 score of 0.8215 for Task A. In general, the results of the HASOC competition indicate that both transformer-based and machine-learning models can effectively detect hate speech and other offensive content in social media.

Despite this progress, significant work must be done, in order to improve NLP models for low-resource languages. For example, current pre-trained transformer models are typically trained on data from the internet and can be biased towards particular dialects or cultural groups. Additionally, many low-resource languages have unique characteristics, such as complex morphological structures and different writing systems, making it more difficult to develop models that accurately process and understand the language. An example of this can be seen in the shared task (https://sites.google.com/view/multi-label-emotionsfire-task/ (accessed on 6 February 2023)) conducted on "Emotions and Threat detection in the Urdu language" [17,18]. Most of the submitted models used multilingual transformers, yet had much room for improvement due to confusion surrounding multiple emotions, i.e., sadness and surprise. Hence, producing quality datasets and analyses for understanding the range of emotions is necessary, in order to develop transformer methods further.

To better understand the contribution of transformers, specifically in terms of the sentiment-related tasks, we present Table 1, which identifies the transformer method, F1-score, task, and dataset dimensions used in the study. The table results also reaffirm our hypothesis regarding the transformer method's detection of fine-grained sentiments in English. With different dataset dimensions for several tasks, we see that transformers dominate the results. However, since transformer methods have not been tested on hope and regret in a multi-class setting, such as that proposed in this paper, it is important to see if they can outperform existing state-of-the-art datasets, as well as if they can be compared to the performance of other related emotion-based tasks.

Table 1. The table reports the F1 scores achieved by different transformer methods performing different sentiment-related tasks.

Task	Dataset Dimensions	Transformer	F1_Score
Irony Detection [19]	Binary	RCNN-RoBERTa	0.80–0.90
Hate speech [20]	Multi-class	DistilBERT	0.75
Emotion Detection [21]	Multi-label	RoBERTa-MA	0.49
Hope Detection [22]	Multi-class	BERT	0.92

3. Methodology

Six transformer-based models, namely, BERT [9], ALBERT [23], RoBERTa [24], DistilBERT [25], XLNet [26], and ELECTRA [27], were trained for 15 epochs per fold in a five-fold cross-validation manner. These models were fine-tuned using SimpleTransformers (https://simpletransformers.ai/, accessed on 6 February 2023), a well-designed library used to load pre-trained models from HuggingFace. The SimpleTransformers library is a black box framework that accepts texts and corresponding labels, in order to fine-tune the pre-trained model's performance of the target task.

Similar to the experiments on transformers' performances in multi-class hope speech detection tasks [7], we trained the models with default parameters, and only set a learning rate of 3×10^{-5}. However, since the maximum length of text in the ReDDIT [4] dataset was limited to 500 in the dataset creation, we set the maximum length parameter to this number. On the other hand, because tweets are often shorter than ReDDIT posts, the maximum length parameter for multi-class hope speech detection was set to 100. This provides insight

into the comparison of the performances of different transformers, in terms of short and long sentences.

A general structure of the experiments can be explained through the methodology Figure 1 that explains the complete workflow.

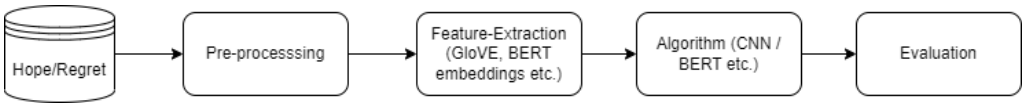

Figure 1. Flowchart of the methodology presented in the article.

4. Experiments and Results

4.1. Pre-Processing

Pre-processing is essential to natural language-processing applications, especially regarding social media data. We removed all the punctuation, URLs, usernames, and stop words that did not contribute to emotion-related tasks. We did not conduct lemmatization or stemming, in order to keep the text as close to the human-written text as possible. Some features of human-related texts, i.e., elongated words, show up as keywords. These features are essential for recording emotions such as hope and regret, which are more fine-grained.

4.2. Features

For our deep learning model, we used GloVe vectors with 300 dimensions and 6 B tokens [5]. GloVe vectors can be a powerful tool in text classification, helping to improve semantic understanding, reduce dimensionality, improve generalization, and increase accuracy. We used character n-grams for the machine learning experiments, as they are language-independent and particularly useful for tasks where misspellings and out-of-vocabulary words are common, or when dealing with short texts.

4.3. Algorithms

The BERT, ALBERT, RoBERTa, DistilBERT, XLNet, and ELECTRA models are all based on transformer architecture, and are trained to perform natural language-processing tasks. However, there are some differences between them, i.e., in terms of the training data, architecture, etc.

BERT was trained on a large corpus of text from Wikipedia and the BookCorpus dataset. The corpus was pre-processed by masking some of the tokens in the text, and the model was trained to predict the masked tokens, based on the surrounding context. ALBERT was also trained on the same corpus of text as BERT, but with a modified training objective that encouraged the model to share parameters across layers, as well as reduce the number of parameters. RoBERTa was trained on a larger corpus of text than BERT, including additional data from CommonCrawl and the WebText dataset. The training objectives were also modified, in order to improve the quality of the representations. DistilBERT was trained on a smaller corpus of text than BERT, but with a modified architecture that is faster and more memory-efficient. The training objective is also simplified compared to BERT. XLNet was trained on a large corpus of text from Wikipedia and the CommonCrawl dataset. The training objective was based on predicting the probability of each token in the text given its context, similar to language modeling. ELECTRA was trained using a pre-training method, where a generator network corrupts the input text, and a discriminator network is trained to distinguish between the original and the corrupted text. The model was trained on a large corpus of text from Wikipedia and the Common Crawl dataset.

For deep learning, CNN is a powerful tool that can extract relevant features from input data, in this case, textual data, through convolutional layers. By capturing local patterns in the input text, CNNs can identify critical features and relationships between words, leading to better performances in sentiment analysis, topic classification, and language translation tasks. Similarly, logistic regression is a simple, yet robust, statistical model that

can handle large volumes of data. Logistic regression is easy to interpret, and its coefficients can provide insight into the impact of each feature on the output.

Hence, all these algorithms contribute to extracting different dimensions of hope and regret, which are later discussed in Section 5.

4.4. Dataset

The ReDDIT dataset [4] and PolyHope dataset [7] were used in this paper. The former dataset was created from Reddit posts, and this task aimed to identify the expressions of regret in the texts and classify them into "regret by action" (Action), "regret by inaction" (Inaction), and, if the post does not convey any regret, "No regret".

Table 2 was borrowed from [4], and indicates some samples of the ReDDIT dataset. The statistics of the dataset include a total of 3425 Reddit posts that are distributed as presented in Table 3.

Table 2. Sample texts from ReDDIT dataset.

Text	Regret Detection
Ok so I've been falling hard for this girl recently and decided that today is the day, do I asked her if she likes me via text. She said I'm a great guy but she's not interested right now. Ugh, I feel like I don't know what to do anymore :(just had to post this somewhere.	Action
I've lost so much weight, and everyone keeps saying how great I look. The medication also makes me mildly euphoric. I know it's wrong, but I tried going off the meds before, and I went through a withdrawal and started gaining weight.	No Regret
"I did not really understand that credit cards are fine as long as you pay it all off on time. So I never applied for one back then because I heard that they are bad for young people. Big mistake. I am trying to decrease the regret/self hate for this oversight"	Inaction
I've made some shitty mistakes, like some really shitty ones. Sometimes I think I'm a terrible person, but that would indicate that guilt makes me not a terrible person. But thinking that excuses my problems is something that my mind can't comprehend. I really just struggle with my thoughts sometimes.	Action
I cheated on ten geometry lessons and two tests in last year. I feel awful and never want to cheat on anything again, ever. That feels better.	Action
Yep that's pretty much it. If they're jewish/asian mix it's all over. I will want to marry her on the spot.	No Regret

Table 3. Statistics of ReDDIT and PolyHope datasets.

Task	Classes	Count
Regret Detection	No Regret	1570
	Regret by Action	1503
	Regret by Inaction	352
Hope Speech Detection	Not Hope	4081
	Generalized Hope	2335
	Realistic Hope	982
	Unrealistic Hope	858

In addition to the ReDDIT dataset [4], we compared the results of the transformers with their performances reported in the PolyHope dataset [7]. The PolyHope dataset was collected from tweets to build a two-level hope speech detection task; in the first level, each post was classified as "Hope" or "Not Hope", and in the second level, hope speech detection was modeled as a multi-class classification task, in which each tweet was classified as "Generalized Hope", "Realistic Hope", "Unrealistic Hope" or "Not Hope" [7]. The PolyHope dataset's statistics are presented in Table 3.

In this study, we used averaged weighted and macro scores to evaluate the performance of our learning models. Since Reddit posts consist of a title and main text, we trained each model using only the main text, and then again using both the text and title, in order to compare the title's contribution to the classification task.

4.5. Results

In the first part of Table 4, the performance of various transformer models is evaluated, where only the text of Reddit posts is used. The results show that the roberta-based model alone achieved the highest averaged macro precision. In contrast, the client-based models had the highest averaged macro recall and accuracy, along with the roberta-based model. The google/electra-based model had the lowest averaged macro precision and recall, with an averaged macro F1-score of 0.70. The bert-based, uncased model performed similarly to the roberta-based and xlnet-based, cased model, with slightly less precision and recall. The albert v2-based and distilbert-based, uncased models also had identical and promising performances.

In the second part of Table 4, the performance of the transformer models improved significantly when the titles of the Reddit posts were included. The roberta-based model again had the highest scores, with an averaged macro F1-score of 0.83. The xlnet-based, cased model, the second-best performer, had a similar performance to the roberta-based model; however, it had a slightly lower precision and recall, and obtained an averaged-macro F1-score of 0.82. The bert-based, uncased, distilbert-based, uncased, and albert v2-based models had similar performances, with slight differences in their scores. These models had lower precision and recall than the roberta-based and xlnet-based, cased models, resulting in lower F1 scores. The google/electra-based model again had the lowest precision and recall, but still performed relatively well, with an F1-score of 0.76, indicating an improvement of 10% when compared to previous experiments utilizing only text.

Table 4. Results regarding regret detection using transformers.

Transformers	Avg. Weighted Scores			Avg. Macro Scores			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
Regret Detection (only text)							
bert-base-uncased	0.78	0.78	0.78	0.75	0.71	0.73	0.78
roberta-base	0.81	0.81	0.81	0.78	0.75	0.76	0.81
albert-base-v2	0.76	0.76	0.76	0.73	0.69	0.70	0.76
xlnet-base-cased	0.80	0.80	0.80	0.77	0.75	0.76	0.80
google/electra-base	0.73	0.72	0.72	0.68	0.66	0.66	0.72
distilbert-base-uncased	0.76	0.76	0.76	0.72	0.69	0.70	0.76
Regret Detection (text + title)							
bert-base-uncased	0.82	0.82	0.82	0.80	0.78	0.79	0.82
roberta-base	0.86	0.86	0.85	0.84	0.83	0.83	0.85
albert-base-v2	0.81	0.81	0.81	0.80	0.77	0.78	0.81
xlnet-base-cased	0.84	0.84	0.84	0.83	0.81	0.82	0.84
google/electra-base	0.79	0.79	0.79	0.77	0.75	0.76	0.79
distilbert-base-uncased	0.81	0.77	0.78	0.79	0.77	0.78	0.81

When comparing the results in Table 4 with the previous state-of-the-art results in Table 5, it can be seen that the transformer models generally outperform the machine learning and deep learning models. In the first part of Table 4, where only the text of the Reddit posts is used, the transformer models have averaged macro F1-scores, ranging from 0.70 to 0.76, while the best-performing machine learning model (LR) has an averaged macro F1-score of 0.56 and the deep learning model (CNN with GloVe) hs an averaged macro F1-score of 0.61. In the second part of Table 4, where the text and title of the Reddit posts are used, the transformer models have averaged macro F1-scores ranging from 0.78 to 0.83, while the machine learning model (LR) has an averaged macro F1-score of 0.55 and the deep learning model (BiLSTM with GloVe) has an averaged macro F1-score of 0.72; this was the previous state-of-the-art model.

Overall, it can be concluded that the transformer models have significantly higher scores when compared to the previous state-of-the-art machine learning and deep learning models for regret detection. This suggests that the transformer models are more effective at classifying Reddit posts in terms of the expression of different types of regret or no regret. The transformer models' ability to capture long-range dependencies and handle large amounts of data may allow them to better analyze the texts and titles of the Reddit posts and accurately classify them. A graphical representation of this comparison between the best-performing models for the regret detection task is presented in Figure 2.

Table 5. Previous state-of-the-art results on regret detection.

Model	Learning Approach	Averaged-Weighted F1	Averaged-Macro F1
only text			
LR	Machine learning	0.64	0.56
CNN with GloVe	Deep learning	0.71	0.61
text + title			
LR	Machine learning	0.63	0.55
BiLSTM with GloVe	Deep learning	0.76	0.72

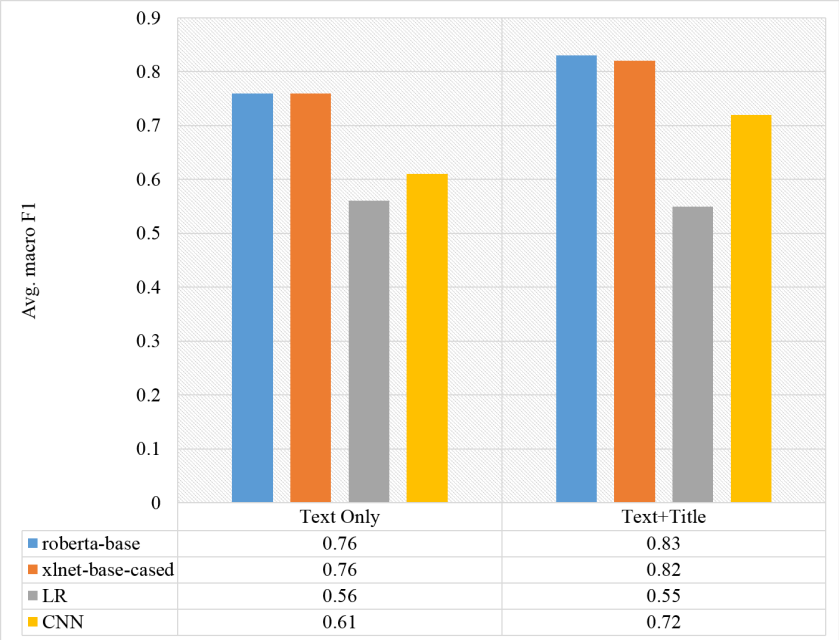

Figure 2. Comparison of best performing models for regret detection.

Similarly, in Table 6, the performances of the various transformer models is evaluated for multi-class hope speech detection from tweets. The bert-based, uncased model has the highest scores, with an averaged macro F1-score of 0.72. The roberta-based and distilbert-based, uncased models have a slightly lower precision and recall, but still perform well, with an averaged macro F1-score of 0.70. The albert v2-based, xlnet-based, cased, and google/electra-based models have similar performances, with slight differences in precision and recall, and averaged macro F1-scores ranging from 0.68 to 0.69.

Table 6. Results of hope speech detection using transformers.

Transformers	Avg. Weighted Scores			Avg. Macro Scores			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
bert-base-uncased	0.78	0.77	0.77	0.71	0.72	0.72	0.77
roberta-base	0.77	0.76	0.76	0.69	0.71	0.70	0.76
albert-base-v2	0.75	0.75	0.75	0.69	0.67	0.68	0.75
xlnet-base-cased	0.76	0.75	0.76	0.69	0.71	0.69	0.74
google/electra-base	0.75	0.74	0.74	0.67	0.69	0.68	0.74
distilbert-base-uncased	0.76	0.76	0.76	0.70	0.70	0.70	0.76

Comparing the results of the ReDDIT and PolyHope datasets represented in Figure 3 reveals that the transformers perform well for both tasks. However, the specific performance of each model may vary depending on the task and dataset. The roberta-based model achieved the highest performance for regret detection, with an averaged macro F1-score of 0.83. On the other hand, for the task of hope speech detection, the bert-based, uncased model outperformed the rest of the models, with an averaged macro F1-score of 0.72. It is worth noting that the regret detection from Reddit posts may provide more context, due to the larger amount of data and the inclusion of post titles, which could explain the higher scores achieved by the transformer models in the ReDDIT dataset when compared to the PolyHope dataset.

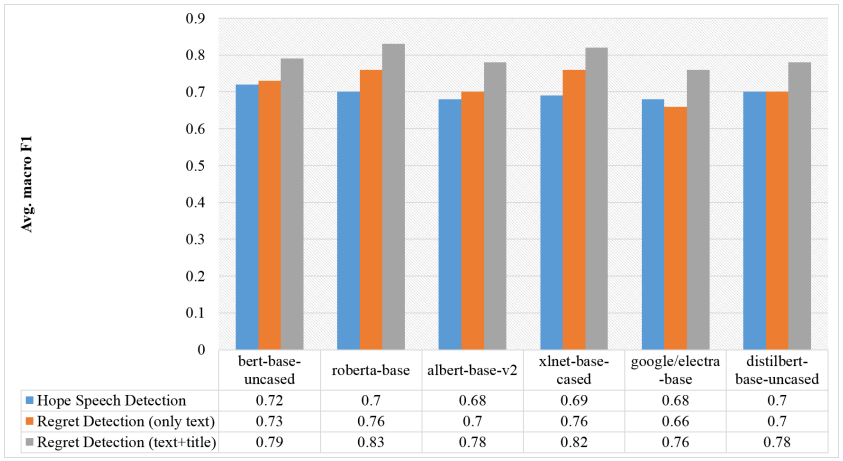

Figure 3. Comparison of the performance of different transformers between regret detection and multi-class hope speech detection tasks.

5. Discussion

5.1. Regret Detection

Several potential insights can be drawn from the comparison of different models for regret detection tasks:

First, the transformer models are more effective at classifying Reddit posts in terms of either expressing a different type of regret, or no regret, when compared to the previous state-of-the-art machine learning and deep learning models. This suggests that the the transformer models' abilities to capture long-range dependencies and handle large amounts of data allow them to better analyze the text and title of the Reddit posts, and to accurately classify them.

Second, including the post titles in the regret detection task significantly improves the performance of the transformer models. This may be because the post titles contain valuable information that helps the models better understand the context of the posts, and classify them accurately.

Third, the roberta-based model consistently performed the best among the transformer models, with the highest scores in both parts of Table 4. This suggests that the RoBERTa

architecture and pre-training on a large dataset using the BERT architecture is beneficial for this task.

Fourth, the machine learning model (LR) performed poorly when compared to the transformer and deep learning models (CNN with GloVe). This suggests that more complex models, such as transformers or deep learning models utilizing context, are more suitable for this task.

Finally, the deep learning model (CNN with GloVe) performed relatively poorly when compared to the transformer models using only text (in the first part of Table 4), but performed relatively well in the second part of the table. This shows that adding a title, not only for the transformer models but also for the CNN models, has significant positive effects. It proves the effectiveness of the CNN model, in terms of capturing the contextual information in the text and title of the Reddit posts.

5.2. Hope and Regret on Transformers

Several insights drawn from the comparisons between the performances of different transformers, of the tasks of regret detection and hope speech detection, are presented as follows:

- Generally, the transformers were expected to perform better on longer texts, such as Reddit posts, than shorter texts, such as tweets. This is because transformers are designed to capture long-range dependencies in the text and can handle large amounts of context. On the other hand, including the post titles in the regret detection task may provide additional context and information that helps the transformer models better understand the content of the Reddit posts. This could explain the higher scores achieved by the transformer models in the regret detection task, compared to the hope speech detection task, where only the text of the tweets is used.
- In terms of the complexity of the tasks, the task of regret detection from Reddit posts comprises a three-class classification task, while the task of hope speech detection from tweets is a four-class classification task. Including an additional class in the hope speech detection task may make it more challenging for the models to classify the tweets accurately. This could explain the lower scores achieved by the transformer models in this task when compared to the regret detection task.
- The dataset size can impact the performance of different transformers, in terms of hope speech detection and regret detection tasks. A larger dataset may provide more diverse and representative samples of the underlying distribution, which can help the transformer learn more accurate and generalizable patterns in the data. This can lead to an improved performance of the task. However, in this comparison, longer texts and more context (ReDDIT dataset) resulted in a better improvement than the number of samples (PolyHope dataset).
- Results in Tables 4 and 6, revealed that transformers were effective in both datasets, with slight differences. Based on the results in Table 4, the roberta-based model performs exceptionally well in terms of the regret detection task, with an averaged macro F1-score of 0.83 when the text and title of the Reddit posts are used. The xlnet-based, cased model also performs well, with an averaged macro F1-score of 0.82. These results suggest that the roberta-based and xlnet-based, cased models may be well-suited for tasks involving longer texts. Regarding the hope speech detection task, as shown in Table 6, the bert-based, uncased model performs particularly well, with an averaged macro F1-score of 0.72. The roberta-based and distilbert-based, uncased models also perform well, with an averaged macro F1-score of 0.70. These results suggest that the bert-based, uncased, and roberta-based models may be well-suited for tasks involving shorter texts. Overall, these results suggest that the transformer models perform well on tasks involving long and short texts. Overall, the roberta-based model was the best-performing transformer, with high performance in both datasets.

6. Limitations

Several limitations to the present work are left for future work and further analysis.

- The study is limited to English texts, which may not be representative of the entire population or all types of communication;
- The study only compares the performance of transformer-based models with GloVe embeddings (as the previous state-of-the-art performing model) and does not consider other types of embeddings or language models;
- The study does not investigate the impact of other factors, such as the choice of hyperparameters, on the performance of the models;
- The study does not address the limitations of the machine learning and deep learning models used, such as the need for large amounts of data and the potential for overfitting;
- The study does not consider the potential impact of the choice of pre-processing techniques, such as stemming, lemmatization, or feature engineering and feature selection techniques, on the performance of the models;
- The study does not explore the use of psycholinguistics, sentiments, or other emotions embedded in the text for the tasks of regret detection and hope speech detection;
- The study does not investigate the effects of fine-tuning the models for more extended periods on their performance.

Validity Threats

We acknowledge that the annotation of the texts for training data may be subject to human bias, leading one to create inaccurate or inconsistent labels. This is especially true in the case of hope and regret, where language constraints and insufficient context can misconstrue the user's true intentions. We might also have selection bias, as the social media texts used to train the model may not represent the population or sample of interest, leading to biased results.

7. Conclusions and Future Work

In this paper, we analyzed the performance of different transformer models on two novel datasets, called ReDDIT and PolyHope, in terms of regret detection from Reddit posts and multi-class hope speech detection from tweets. For regret detection, we compared the averaged macro scores of the transformer models to the previous state-of-the-art results. We found that the transformer models outperformed the previous approaches. Specifically, the roberta-based model achieved the highest averaged macro F1-score of 0.83, beating the previous state-of-the-art score of 0.76. For the task of hope speech detection, the bert-based, uncased model achieved the highest averaged macro F1-score of 0.72 among the transformer models. In addition, we found that the specific performance of each model varied slightly depending on the task and dataset.

Overall, our findings demonstrate the effectiveness of the use of context in transformer models for text classification tasks. The inclusion of the post titles in the regret detection task significantly improves the performance of the transformer models. The task's complexity, the dataset's size, and the specific transformer architecture and pre-training all influenced the performance of the models on both tasks. In future work, we would like to explore the mentioned limitations for these tasks.

Author Contributions: Conceptualization, G.S. and F.B.; Methodology, G.S., F.B., S.B. and A.G.; Software, A.G.; Validation, G.S. and A.G.; Formal analysis, F.B. and S.B.; Investigation, G.S., F.B., S.B. and A.G.; Resources, G.S. and F.B.; Data curation, G.S. and F.B.; Writing—original draft, F.B. and S.B.; Writing—review & editing, G.S. and A.G.; Visualization, A.G.; Supervision, G.S. and A.G.; Project administration, G.S., F.B. and S.B.; Funding acquisition, G.S. and A.G. and G.S.; All authors have read and agreed to the published version of the manuscript.

Funding: The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politecnico Nacional, Mexico. The authors thank the CONACYT for the

computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and the script used in this paper will be available on a request from corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gilovich, T.; Medvec, V.H. The Experience of Regret: What, When, and Why. *Psychol. Rev.* **1995**, *102*, 379. [CrossRef] [PubMed]
2. Hattiangadi, N.; Medvec, V.H.; Gilovich, T. Failing to act: Regrets of Terman's geniuses. *Int. J. Aging Hum. Dev.* **1995**, *40*, 175–185. [CrossRef]
3. Diecidue, E.; Somasundaram, J. Regret theory: A new foundation. *J. Econ. Theory* **2017**, *172*, 88–119. [CrossRef]
4. Balouchzahi, F.; Butt, S.; Sidorov, G.; Gelbukh, A. ReDDIT: Regret Detection and Domain Identification from Text. *arXiv* **2022**, arXiv:2212.07549. https://doi.org/10.48550/ARXIV.2212.07549.
5. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
6. Fazlourrahman, B.; Aparna, B.; Shashirekha, H. CoFFiTT-COVID-19 Fake News Detection Using Fine-Tuned Transfer Learning Approaches. In Proceedings of the Congress on Intelligent Systems, 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 879–890.
7. Balouchzahi, F.; Sidorov, G.; Gelbukh, A. PolyHope: Dataset Creation for a Two-Level Hope Speech Detection Task from Tweets. *arXiv* ***20222022***, arXiv:2210.14136.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**; *30*, 5998–6008.
9. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
10. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 19–27.
11. Wang, W.; Yan, M.; Wu, C. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–28 July 2018; pp. 1705–1714.
12. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 2383–2392.
13. Köhler, J.; Shahi, G.K.; Struß, J.M.; Wiegand, M.; Siegel, M.; Mandl, T.; Schütz, M. Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection. In Proceedings of the Working Notes of CLEF, Bologna, Italy, 5–8 September 2022.
14. Shashirekha, H.; Balouchzahi, F.; Anusha, M.; Sidorov, G. CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts. *arXiv* **2022**, arXiv:2211.09847.
15. Modha, S.; Mandl, T.; Shahi, G.K.; Madhu, H.; Satapara, S.; Ranasinghe, T.; Zampieri, M. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In Proceedings of the Forum for Information Retrieval Evaluation, Virtual, 13–17 December 2021; pp. 1–3.
16. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
17. Butt, S.; Amjad, M.; Balouchzahi, F.; Ashraf, N.; Sharma, R.; Sidorov, G.; Gelbukh, A. Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022. In Proceedings of the CEUR Workshop Proceedings, Chennai, India, 22–24 April 2022.
18. Butt, S.; Amjad, M.; Balouchzahi, F.; Ashraf, N.; Sharma, R.; Sidorov, G.; Gelbukh, A. EmoThreat@FIRE2022: Shared Track on Emotions and Threat Detection in Urdu. In Proceedings of the Forum for Information Retrieval Evaluation, Madrid, Spain, 11–15 July 2022; Association for Computing Machinery: New York, NY, USA, 2022; FIRE 2022.
19. Potamias, R.A.; Siolas, G.; Stafylopatis, A.G. A transformer-based approach to irony and sarcasm detection. *Neural Comput. Appl.* **2020**, *32*, 17309–17320. [CrossRef]
20. Mutanga, R.T.; Naicker, N.; Olugbara, O.O. Hate speech detection in twitter using transformer methods. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 9. [CrossRef]
21. Ameer, I.; Bölücü, N.; Siddiqui, M.H.F.; Can, B.; Sidorov, G.; Gelbukh, A. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.* **2023**, *213*, 118534. [CrossRef]

22. Arunima, S.; Ramakrishnan, A.; Balaji, A.; Thenmozhi, D. ssn_diBERTsity@ LT-EDI-EACL2021: Hope speech detection on multilingual YouTube comments via transformer based approach. In Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Online, 19 April 2021; pp. 92–97.
23. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
24. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
25. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108.
26. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
27. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.