Fall 2021

Multimodal Detection of Cyberbullying on Twitter

Jiabao Qiu
San Jose State University

MULTIMODAL DETECTION OF CYBERBULLYING ON TWITTER

A Thesis

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Jiabao Qiu

January 2022

The Designated Thesis Committee Approves the Thesis Titled

MULTIMODAL DETECTION OF CYBERBULLYING ON TWITTER

by

Jiabao Qiu

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

January 2022

Teng Moh, Ph.D.	Department of Computer Science
Melody Moh, Ph.D.	Department of Computer Science
Robert Chun, Ph.D.	Department of Computer Science

ABSTRACT

MULTIMODAL DETECTION OF CYBERBULLYING ON TWITTER

by Jiabao Qiu

Cyberbullying detection is one of the trending topics of research in recent years, due to the popularity of social media and the lack of limitations about using electronic communications. Detection of cyberbullying may prevent some bullying behaviors online. This paper introduced a Multimodal system that makes use of Convolutional Neural Network (CNN), Tensor Fusion Network, VGG-19 Network, and Multi-Layer Perceptron model, for the purpose of cyberbullying detection. This system can not only analyze the messages sent but also the extra information related to the messages (meta-information) and the images contained in the messages. The proposed system was trained and tested on Twitter datasets, achieving accuracy scores of 93%, which was 4% higher than scores of the benchmark text-only model using the same dataset and 6.6% higher than previous work. With the results, we believed that the proposed system performs well and it will provide new ideas for future works.

ACKNOWLEDGMENTS

This paper is a Master's Thesis. Thanks to the people who supported the work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Cyberbullying is a term used for describing a type of bullying behavior that uses electronic communications, such as sending messages via electronic devices and communicating on the Internet. According to Wegge et al. [1], cyberbullying is an extension of traditional bullying and it often occurs among a group of teenagers. However, cyberbullying among adults is also common these days.

According to a study on the statistics of cyberbullying events [2], during the COVID-19 pandemic in 2020, cyberbullying occurred on social media more often than in the past year due to the increment in the number of users online. In the post-pandemic time, paying attention to such events online became important to keep people healthy in both physical and mental ways.

Research on cyberbullying among teenagers in 2021 [3] stated that 21% of teenagers had been cyberbullied and more than a half of these events happened during the pandemic. While many of the victims felt hurt in their feelings, about 12% of victims suffered issues on physical health as consequences. The reported suicide cases also seemed to increase during the pandemic, showing that some public figures chose to end their lives due to cyberbullying, especially in Asia. For example, there was a young Japanese Netflix star who was killed by cyberbullying in May 2020 [4]. Preventing the increment in cyberbullying events was important, but it required lots of effort because of the increasing number of Internet users.

To stop cyberbullying, automatic detection of such behaviors is vital because people can immediately deal with the behaviors once received an alarm. Many researchers are working on automatic cyberbullying detection and many systems have been presented. However, since the usage of language and the form of messages are changing and developing rapidly in recent years, automatic systems have to be updated to follow the trend.

Accuracy in cyberbullying detection became one of the most crucial goals for studies. If the accuracy improves, then the dire needs of resources and efforts will be fulfilled by the machines. People cannot always pay attention to the environment to see if there is cyberbullying behavior that they can stop, but the machines can do the job and set alarms. Improving accuracy while the forms of cyberbullying are developing is one of the main focuses of research.

Moreover, the stage of cyberbullying is changing. Social media platforms like Twitter, Instagram, and Facebook are playing important roles in people's lives. When this project started, Facebook was still the most popular social media platform in the United States; however, in recent days, Tik Tok is playing a more important role among young people.

As a result, multi-media content should be considered in cyberbullying detection, which focused on text classification in the past. Images and videos become important attributes of cyberbullying behaviors, thus Computer Vision may play a part in such detection systems. Natural Language Processing and text classification are not the dominating methods in predicting cyberbullying behaviors.

In this paper, we proposed a Multimodal system that took not only text messages but also other data as inputs, focusing on improving detection accuracy. The system considered meta-information, which was features of the post and the user who posted the post, providing more information to analyze the message and the user. The system also took image inputs, predicting if the post was cyberbullying with the sentiment behind the image.

The remainder of this paper is structured as follows. First, in Section 2, we will discuss the recent works for the task of cyberbullying detection and develop the proposed solution according to the previous works. Second, this solution is introduced in Section 3 with details on each part of its structure, including a Bimodal system for classifying text and meta-information inputs, a linear system for classifying image input, and the

remaining structure that combines the outputs. Third, in Section 4, we will take a look at the datasets used in this paper and the experiments done for this paper. Fourth, the results of the experiments are presented in Section 5. Fifth, Section 6 concludes the experiments and the results, providing an explanation and proposal of questions. Last, in Section 7 we will discuss the possible improvements that can be done based on the observations and discoveries from this paper.

2 LITERATURE REVIEW

In earlier years, cyberbullying occurred in the form of text because the main carrier of messages was plain text. For example, text messages sent via phone, message boards on Bulletin Board Systems (BBSs), and even emails are all considered as possible sources of cyberbullying according to the definition of cyberbullying by Oxford's English dictionary. Hence, most of the early works of cyberbullying detection focused on sentiment analysis. However, because social media played a more important role than plain-text messages in recent years, some works also analyzed social media posts.

Previous works that focused on sentiment analysis primarily used text data from the Internet and performed neural classification. Twitter is one of the most popular data sources for this purpose, thus it is possible to learn from the methods that were used to collect data from Twitter. In [5], researchers collected English data from Twitter. They provided detailed and solid explanations of data preparation, too. Experiments in [6] analyzed non-English Twitter data with Convolutional Neural Network (CNN) and received a high accuracy of prediction at 93.97%. However, due to the differences among languages, this work could not be a baseline for the proposed solution.

There are also previous works focused on multi-media content, such as Twitter or Instagram posts with images or videos. Wang et al. [7] constructed a multi-modal model that took media contents as inputs and achieved a better result than using text content only and achieved an accuracy score of 86.4%. In [8], researchers used self-attention models for classification purposes and proposed a future direction in which "Internet memes," a type of image, can be used in the detection of cyberbullying events.

Moreover, some researchers also noticed other attributes than text and media content and explored a greater possibility in cyberbullying detection. Because bullying events occurred based on people's relationships, some works considered relationships among users to predict cyberbullying behaviors online, though had not proved it was a better

approach yet. In [9], the researchers used meta-information of Tweets such as the number of tags, the network features, and the user profiles to aid cyberbullying detection. This 2020 work achieved 0.9291 area under the curve (AUC) on an Instagram dataset. The work in [7] also made use of some of the meta-information.

In many previous works [5]–[7], [9], researchers used Twitter data as inputs. The posts from Twitter have relatively short text messages, media attachments, and a large amount of meta-information such as hashtags, user information, and external links. Along with the convenient Twitter API which allows developers to collect some information from Twitter, these attributes make Twitter a popular source of data in cyberbullying detection.

In 2021, researchers were still working on improving cyberbullying detection methods. However, many works used Instagram datasets instead of Twitter datasets. While some works were exploring possibilities other than focusing on accuracy, there were still works focused on improving detection accuracy. In August, Cheng et al. [10] proposed a method mitigating bias and achieved a 0.9089 AUC score. In October, researchers proposed another method that made use of social media features, resulting in an average of 78.93% accuracy with the highest as 90.1% [11].

Combining all the significance from previous works, the proposed solution in this paper would use three types of inputs: text, meta-information, and images. Inspired by [6], [7], the proposed solution used CNN for sentiment analysis and constructed a multi-modal model for cyberbullying detection.

Compared to the previous approaches, the proposed system had several improvements. Previous works such as [9], [11] did not make use of media contents, but the proposed system used multiple types of data as input. While another work that took media information as input [7] used the Long Short-Term Memory model for text classification, our system used CNN which was proved to have higher accuracy by other works. Also, the proposed system used the Tensor Fusion method for combining the inputs. The

detailed structure of the proposed system was described in Section 3 and we would see the results in Section 5 to determine if the system that was different from those of previous works worked better.

3 PROPOSED SOLUTION

The system described in this paper consisted of two parts: a model used for the classification of text inputs and meta-information inputs and another model used for sentiment analysis on image inputs. These parts were combined to construct the complete multi-modal system of cyberbullying detection.

The baseline accuracy (86.4%) used for comparison was from a previous work [7], which also constructed a multimodal system for cyberbullying detection. Also, the accuracy score (89%) from the text-only model described below would become the benchmark for the following experiments.

The solution introduced in this session was proposed for the reason that recent approaches only focused on some attributes in a dataset, not considering all possible information. Also, the previous works lacked a method combining the attributes while keeping their characteristics.

3.1 Text and Meta-Information

The inputs of text and other attributes than media were passed to the same CNN model for classification purposes. This was not done by sentiment analysis, because sentiment did not directly related to cyberbullying as stated in the introduction session.

3.1.1 Text

Text data extracted from every single entry in the Twitter dataset was raw and needed preprocessing. It was embedded before being fed as an input. Because this dataset [12] contained data before Twitter increased the maximum length from 140 to 280 characters, each entry used in this paper was less than or equal to 140 characters in length and written in English.

3.1.1.1 Preprocessing Text Data: The text messages extracted contained symbols, special characters, personal information, and so on. To keep it anonymous, the

preprocessing progress was done by a program. Thus we did not have any acknowledgment of the personal information mentioned in the Tweets. Also, the special usage of characters was formalized by the program.

Each text entry was preprocessed by the following steps.

- Replace hyperlinks with the phrase `HTTPLINK`.
- Replace usernames with the phrase `USERNAME`.
- Replace HTML character codes with UTF-8 characters.

  For example:

  - Replace `&` with `&`.
  - Replace `&quot;` with `"`.
  - Replace `&apos;` with `'`.
  - Replace `<` with <.
  - Replace `>` with >.
- Replace a sequence of multiple spaces with a single space.
- Replace a new-line character with a single space.
- Remove special characters, keeping only `A-Z a-z 0-9 $#*(),!?'` `
- Split contraction. For example, replace `I've` with `I 've`.
- Split punctuation in case of a word followed immediately by a punctuation.
- Split hashtags (began with `#`) or mentions (began with `@`) from body of the text.

3.1.1.2 Training on Text Data: To train the model on text data, the data was embedded. Words in each text input were mapped to integers presenting their indices in the vocabulary. Hence, the model received vectors of integers as one of the inputs and this input was named `text_input`. The vocabulary size was 56,028 in this dataset; the maximum input length was 114, representing the maximum number of indices in a single vector. The embedding dimension was set to 128. The input vector was passed into an

Embedding layer, followed by another layer to reshape the input to match the expected shape of the input.

The final shape of `text_input` was `(max_len, embedding_dim, 1)`. Fig. 1 shows the embedding layers of the model.

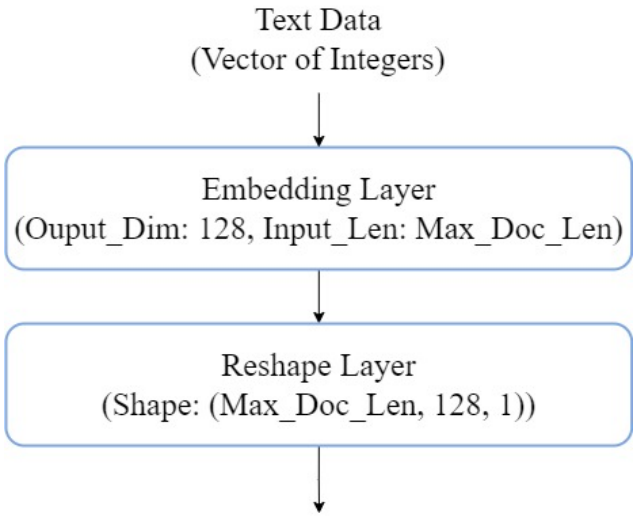

Fig. 1. Embedding layers.

3.1.2 Meta-Information

The meta-information was used to assist the classification of the text input. Therefore, it was added into the text data before being passed to the CNN so that it could add more information to the text input. The original entry of meta-information consisted of many attributes, which included many that were not useful for cyberbullying detection, thus we performed attribute reduction first.

3.1.2.1 Attribute Reduction: To reduce the attributes, we constructed a Decision Table based on the meta-information entries and the labels of each entry. All the attributes except for the key numbers of the entries could be considered as conditions, while the label which defined bullying behaviors was considered as the result. Table 1 was an example of a Decision Table.

Table 1
Example Decision Table from Dataset

	Conditions				Results
key id	**is reply**	**user statuses count**	**user favorites count**	...	**bullying**
0	No	100214	480446	...	TRUE
1	No	47132	21980	...	TRUE
...	...	...	...	...	...
43272	Yes	857	1	...	FALSE

This Decision Table was used for constructing the Discernibility Matrix. A Discernibility Matrix could look like the following:

$$\begin{bmatrix} \{statuses, favorites\} & & & \\ \{statuses, friends, followers\} & \{listed, friends\} & & \\ \{statuses, friends\} & \{statuses, listed\} & \{listed, followers\} & \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

After these steps, Discernible Functions were written from the Discernibility Matrix, using `AND`, `OR`, and parentheses. Then the functions were reduced into Prime Implicants. This process was done by simple programming. Finally, the minimal reductions were shown as follows:

1) {`key_id`, `user_statuses_count`, `user_favorites_count`, `user_listed_count`, `user_mentions`, `retweet_count`, `favorite_count`}
2) {`key_id`, `user_statuses_count`, `user_favorites_count`, `user_friends_count`, `hashtags`, `user_mentions`}
3) {`key_id`, `user_statuses_count`, `user_favorites_count`, `user_friends_count`, `user_mentions`, `retweet_count`}
4) {`key_id`, `user_statuses_count`, `user_favorites_count`, `user_followers_count`}
5) {`key_id`, `user_statuses_count`, `user_favorites_count`, `user_listed_count`, `user_friends_count`}

6) {`key_id`, `user_statuses_count`, `user_followers_count`, `user_listed_count`, `user_friends_count`}

The attributes with names in a format of `user_*_count` was the user information of the person who posted the Tweets, while the others were information about the posted Tweets. The most significant attribute was `key_id` of the entry. However, we noticed that the user information was more important than the attributes of the Tweet itself. For example, `user_statuses_count` was another dominating attribute.

3.1.2.2 Training on Meta-Information Data: We tested every combination from the list of Attribute Reduction with the model and found out that the 4^{th} combination had the best performance among all combinations. It provided a better accuracy score while it had the smallest amount of attributes. The small size of the meta-information input required much less time for training than the other combinations.

The meta-information input was in the form of vectors with a length of three, containing integers from each attribute. However, these vectors were preprocessed before they were added to the vectors of text data. The details of the preprocessing will be discussed in the next section, *Combining Text and Meta-Information*.

3.1.3 Combining Text and Meta-Information

The vectors of text input and meta-information input were combined by calculating their cross-product. Before calculating the cross product, an extra 1 was added to each array as Equation 1. This was a Tensor Fusion technique inspired by [13], a Tensor Fusion Network for sentiment analysis.

$$z = \begin{bmatrix} z_1 \\ 1 \end{bmatrix} \bigotimes \begin{bmatrix} z_2 \\ 1 \end{bmatrix} \tag{1}$$

According to [13], this method helped construct the Tensor Fusion, shown in Fig. 2, so that the matrix could keep the original information from each of the input tensors.

Hence, this method would produce a model that was both bimodal and unimodal. It could enhance the performance by keeping unimodal information.

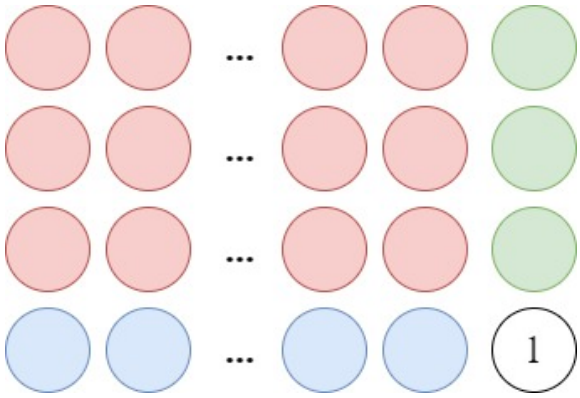

Fig. 2. Tensor fusion, where blue is text data and green is meta-information data (both unimodal), while red is the bimodal matrix.

The combined input for this part of the model was a $\mathbf{129} \times \mathbf{4}$ matrix, with an extra row and an extra column containing the original information from both inputs.

3.2 Images

All media information, no matter it was photo or video, were processed in the form of images. This was because when people were scanning through Twitter, they saw image previews from the Tweets that had media attachments. The way to recognize whether it was a picture or a video clip was whether there was a hyperlink in the body of the Tweet. Therefore, using image input would not lose the information to recognize a video.

These images were extracted from the Tweets for sentiment analysis purposes. Because one could hardly tell if it was bullying or non-bullying from a single image, we decided to do sentiment analysis and use the result as another input attribute.

Since the images extracted from the Tweets did not have sentiment labels, we constructed another model from [14]. This VGG-19 Net model [15] was trained on ILSVRC12 [16] with the original dataset from [14] and finetuned on this dataset. The

model would output vectors of three, indicating the possibilities for the corresponding image to be negative, positive, or neutral in sentiment.

This sentiment analysis result was combined with the classification result from text and meta-information inputs. The details and the complete structure will be discussed in the next session.

3.3 Multimodal Model

The complete model received three inputs: text, meta-information, and images. The former two were combined and passed through a CNN model, while the latter was passed through a VGG-19 Net model. Both models generated classification results, but the former classified the input as bullying or non-bullying and the latter analyzed the sentiment. These results were joined together to produce a final possibility for the particular entry to be a bullying post.

The CNN model took a matrix input and its structure is shown in Fig. 3. The model had three parallel Convolutional layers with different filter sizes, varying from 3, 4, to 5. The Convolutional layers all received the same matrix input and passed them to three different Max-Pooling layers. The Max-Pooling layers had pool sizes corresponding to the filter sizes of their predecessor layers. For example, the Max-Pooling layer connected with the Convolutional layer with a filer size of 3 had a pool size of $\mathbf{114-3+1=112}$, where 114 was the maximum document length.

As shown in Fig. 3, the three output tensors from the Max-Pooling layers were concatenated into a 1-D tensor. This tensor was flattened before being connected to a Dropout layer, and a linear layer was connected to this Dropout layer before it could output the final tensor of this model. The output was through a linear layer with Softmax activation function, shown in Equation 2, where K was the number of classes in classification.

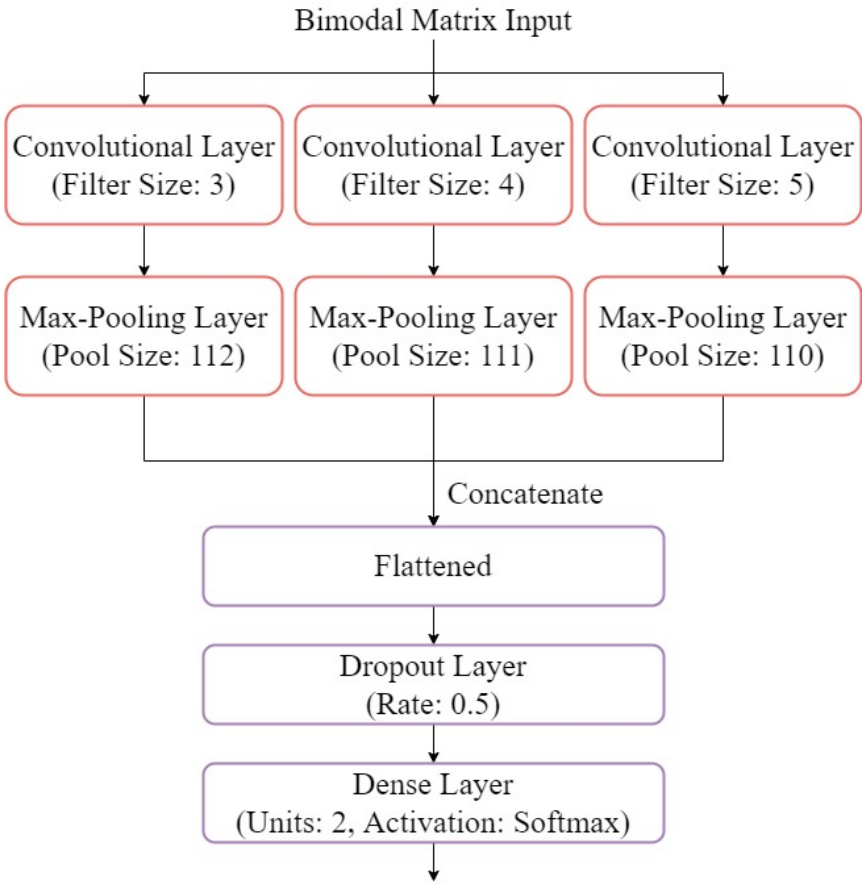

Fig. 3. Covolutional Neural Network.

$$softmax_k(z) = e^{z_j} / \sum_{j=1}^{K} e^{z_j} \tag{2}$$

The VGG-19 Net model [15] was adapted from a 2015 work [16], since the authors of the training dataset [14] used the pre-trained model from the same paper. It was believed to have about 89.6% 5-fold cross-validation accuracy. Basically, this model had a stack of $\mathbf{3} \times \mathbf{3}$ Convolutional layers with $\mathbf{2} \times \mathbf{2}$ Max-Pooling layers. Three Fully-Connected layers were connected to these Convolutional layers, two had 4,096 channels while one had 1,000 channels. A Softmax layer was also added. The structure of this model is shown in Fig. 4.

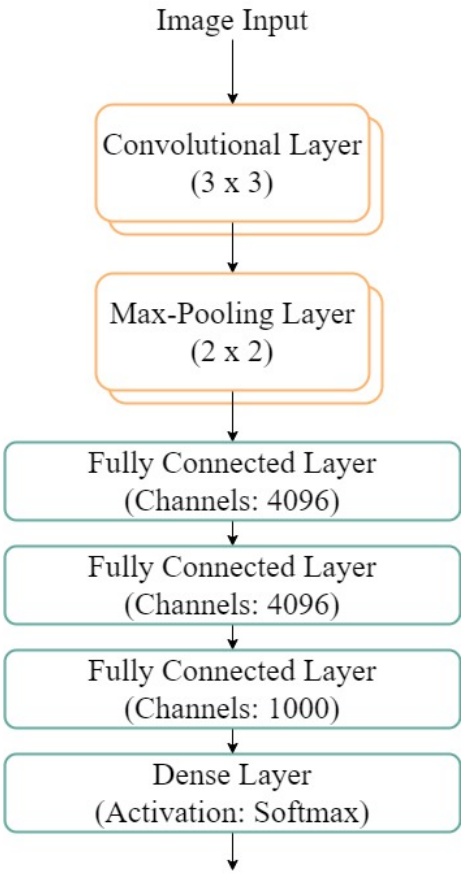

Fig. 4. VGG-19 Net.

The output of this model was a vector with floating-point numbers with the same key IDs as the text and meta-information input. These numbers indicating possibilities in sentiment analysis were passed into a Multi-Layer Perceptron (MLP) model, which was a simple linear model, along with the original label, which was bullying or non-bullying, for the corresponding Tweets according to their key IDs. Thus, the MLP model took the sentiment of image data as input and the original label as its label, then it predicted whether this sentiment was likely to be found in a bullying Tweet or a non-bullying Tweet.

Finally, two different predictions from CNN and MLP were concatenated. The concatenated output was passed through a linear layer with Rectified Linear Unit (ReLU) activation function, shown in Equation 3, followed by a Dropout layer. The final output

was generated through a linear layer with the Sigmoid activation function, shown in Equation 4.

$$ReLU(z) = \max(0, z) \tag{3}$$

$$Sigmoid(z) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{e^{z}+1} = 1 - Sigmoid(-z) \tag{4}$$

Because the model output was in the form of possibilities to be in either category of bullying or non-bullying, to predict the labels, we could pick the category with the higher possibility. The complete structure of this model is shown in Fig. 5.

As shown in Fig. 6, the system took three types of inputs and processed them separately. The text input was cleaned up and embedded, the meta-information input was reduced by feature selection, and the image input was classified with sentiment analysis. The processed text input and meta-information input was combined and classified with a CNN model, while the image sentiment information was classified with an MLP model. The two predictions were combined, thus producing a final prediction for the original Tweet.

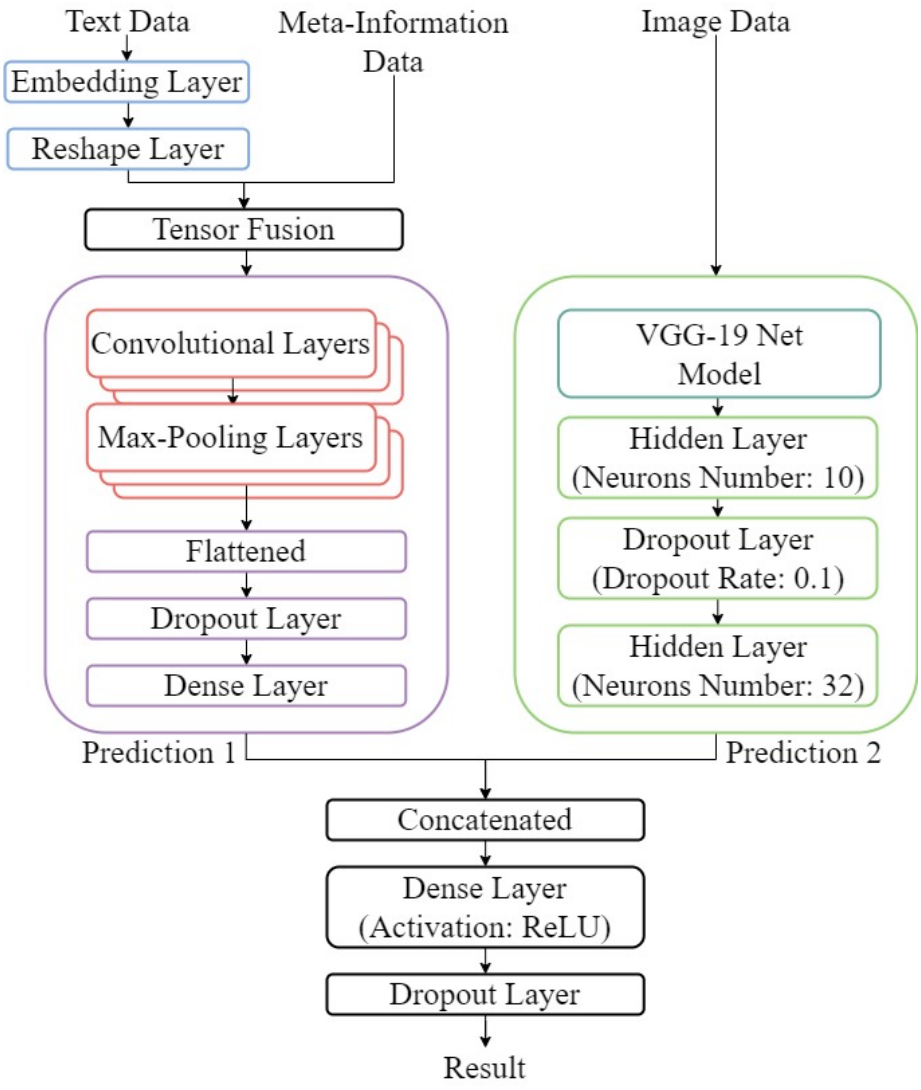

Fig. 5. Complete model structure.

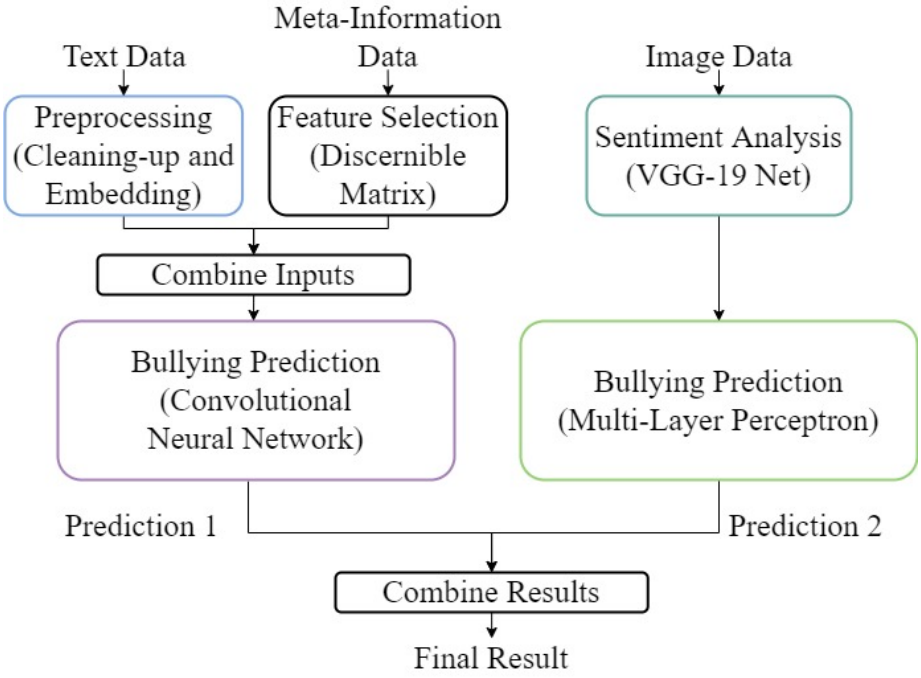

Fig. 6. Top-level architecture of the system.

4 DATA AND EXPERIMENTS

4.1 Data

This paper used two sets of data. One was the basic dataset used in the complete model, the other was used for image sentiment analysis.

The basic dataset was a Twitter dataset from [12]. It contained the unique ID for 100,000 Tweets along with their crowd-sourced annotation. We collected the original information, including text and other meta-information, with Twitter API [17]. Twitter API was a platform for developers to use legal data from Twitter and it had methods collecting information of particular Tweets when ID was provided, thus all Twitter data used in this paper was legal and approved by Twitter and their User Agreements. The collected data from Twitter API was in JSON format, so we extracted the information we needed and stored them in CSV format. The images were in the form of addresses, thus we downloaded all images via the addresses. The annotations, which were either abusive or normal, were added as labels of `TRUE` or `FALSE`.

For example, with the key ID of 42608, the text data was *thanks for this song is f**king amazing, i'm really grateful and proud of you @USERNAME HTTPLINK*, and the meta-information was shown as following: { `is_reply`: No, `favorite_count`: 1, `retweet_count`: 0, `user_followers_count`: 9917, `user_listed_count`: 28, `user_statuses_count`: 78203, `user_friends_count`: 10249, `user_favorites_count`: 54879, `user_mentions`: 1, `hashtags`: 0. It had image attached, and its label was non-bullying, though it had a negative word in the body of text.

The dataset for training the model of image sentiment analysis was from [14], as mentioned in the previous sessions. This dataset was also from Twitter and it had 974,053 entries. These entries were labeled as one of the following: positive, negative, and neutral.

Following are some examples from the image dataset. The sentiment analysis of Fig. 7 showed a very high score in the possibility of being positive, while Fig. 8 was likely to be negative. However, the sentiment had no direct relationship with whether the post was bullying or not.

Fig. 7. Possitive image.

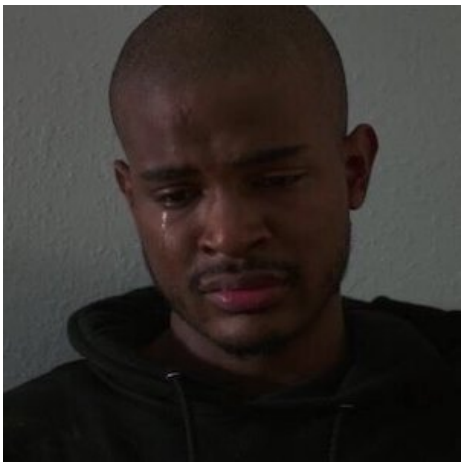

Fig. 8. Negative image.

4.2 Experiments

4.2.1 Experiments on Text Data

At the early stage of experiments, we constructed a CNN model that took text input alone. Its structure was the same as the final version, except that it took a vector input instead of a matrix input. The output tensor passed directly through the linear layer with the Sigmoid activation function and generated prediction.

The preprocessing step was the same at this stage. Hence, the only difference between this version and the later version, described in the next session, was that it did not have the assistance from meta-information and it was a unimodal solution.

The experiments were done as the data was shuffled and cross-validated. The average 5-fold cross-validation accuracy of experiments on text data was about 89.0% with a Standard Deviation of 0.00467.

4.2.2 Experiments on Text Data and Meta-information

Later, the meta-information data was added to the inputs. The process was described in the previous sections. The input for the CNN model was complete at this stage, and the model took matrices as inputs. However, because of the lack of image information, the output tensor was still directly going through the linear layer with Sigmoid activation function, without concatenating with outputs from the MLP model.

This version of the system was bimodal because it took two types of inputs instead of one. It used the Tensor Fusion method to keep unimodal information from both text and meta-information data.

The experiments were done as the data was shuffled and cross-validated. The average 5-fold cross-validation accuracy of experiments on text data was about 91.8% to 92.1%, while the Standard Deviations were 0.00678.

4.2.3 Experiments on Multimodal Model

Finally, the image input was added to the system. Because not all Tweets had images attached, we tried three different approaches.

The first was using only the Tweets with media. This reduced the size of the dataset. Although it eliminated the interference from Tweets without media, it had a smaller size that might lead to poorer performance. The accuracy from a 5-fold cross-validation experiment was 93.0% with a Standard Deviation of about 0.00541.

The second was automatically generating labels for Tweets without media. Vectors of 0's were generated for those Tweets. The accuracy from a 5-fold cross-validation experiment was 91.7% with a Standard Deviation of about 0.01881.

The third was similar to the second, but it chose 1/3 instead of 0. The accuracy from a 5-fold cross-validation experiment was 91.1% with a Standard Deviation of about 0.00005.

Noted that to further compare these results, we performed extra experiments on the dataset. The purpose and result of these experiments will be discussed in Section 5.

In these experiments, we divided the dataset into two separate sets. One set had all Tweets had image attachments and the other had all Tweets that had no image attachment. We trained the model on these datasets separately, without using their image data. Hence, the experiments would only use text and meta-information data so that we could make sure text and meta-information were not the dominating attributes that led to the difference among the three experiments in previous paragraphs.

5 RESULTS

The results from the experiences discussed in Section 4 were compared in two ways: comparing with the baseline model and comparing with each other if they used the same model structure.

5.1 Comparing with Baseline

The comparison of models with the baseline was shown in Table 2. The table indicated whether the model took all types of inputs and their accuracy score in 5-fold cross-validation experiments. The word SD stood for Standard Deviation.

Table 2
Baseline Result Comparison

Model	Text	Meta-Info	Image	5-Fold Accuracy	SD
Baseline (CNN Text Only)	Yes	No	No	89.0%	0.467%
Bimodal (CNN & Tensor Fusion)	Yes	Yes	No	91.8%	0.678%
Multimodal (Bimodal & Sentiment)	Yes	Yes	Yes	93.0%	0.541%

5.2 Comparing with Each Other

The comparison of different data using the same model as shown in Table 3. The entry named **Data Size** showed whether the experiment used all data from the entire dataset or just used the reduced dataset that only contained entries with an image attached. If the dataset included entries without an image, these empty entries were passed into the model with default values. The default values were 0 or 1/3, discussed in Section 4.

Table 3
Data Size Comparison

Model	Data Size	Default Entry	5-Fold Accuracy	SD
Multimodal	Reduced	-	93.0%	0.541%
Multimodal	All	0	91.7%	1.881%
Multimodal	All	0.33... (max length of `float`)	91.1%	0.005%

According to Table 3, the reduced dataset provided the highest accuracy score.

5.3 Final Result

By comparing the results from different experiments, the Multimodal model with a reduced dataset performed better than any other combinations. To verify this observation, we performed extra experiments. These experiments were introduced in Section 4.

Table 4
Image Data Comparison

Model	Data Portion	Text	Meta-Info	Image	5-Fold Accuracy	SD
Bimodal	Has-Image	Yes	Yes	No	90.7%	0.068%
Bimodal	Non-Image	Yes	Yes	No	90.2%	0.174%

The results from these experiments were shown in Table 4. Noted that the entry **Data Portion** stood for whether the experiment used the Tweets that had images attached or it used the Tweets that had no image attachment.

From the results, we found out that the different portions of data had some influence on the results, but the influence was not significant. The difference between the two scores was 0.5% while using all Tweets rather than the reduced dataset resulted in a difference varied from 1.3% to 1.9%.

In conclusion, adding attributes that involved user profiles as inputs and considering the sentiment of images did improve the classification by about 4%. Also, we could see that the Tweets with images were slightly more likely to be bullying posts than the others.

6 CONCLUSION

The paper showed that a Multimodal system can be used in cyberbullying detection. Because bullying behaviors do not solely rely on text, considering other attributes is necessary. Not all attributes are responsible for classifying the data, thus experiments and feature selection methods are required in the progress. The method used in this paper was proved to improve the score by keeping information from each input while also combining them into a new input.

In this paper, according to the feature selected in experiments, the activeness of a user was proved to be relevant to bullying behaviors. Three features of a user were considered most relevant: the total Tweets posted by the user, the total favorite Tweets of the user, the number of followers of the user. These three features showed how active a user might be. It also showed the popularity of a user at some point. Though a higher number did not mean that the user was more likely to be a bully, the bullies were either very active users with high numbers in these features or new users with very low numbers in these features. The possible reason was that the people with a lot of supporters or who spend a lot of time online would be more likely to become a cyberbully. Also, the data of new users showed that some people chose to hide their identity by registering new accounts to fulfill their bullying purpose.

Also, the results showed that it was easier to classify the posts with image attachments than the plain-text posts. This efficiency even increased when the images were considered as one of the attributes. One possible reason was that the posts with images provided more information than plain text, and people were attracted to these posts. Because the data was crowd-sourced, it was easier to annotate these posts than to annotate plain-text posts. Also, pictures could further invoke people's feelings. Therefore, images could help determine whether the post was a bullying behavior or just a joke between friends. Besides emotions, posts that contained media might be more detailed than the plain-text

ones because the users put more effort into editing. A Tweet with only text might be similar to a note or an instant message, lacking much information.

In conclusion, this paper showed the system structure and its results, proving that using text, meta-information, and image data altogether might improve the accuracy of cyberbullying detection. CNNs were suitable for this task according to previous works. Tensor Fusion was also proved to be useful in this case. Although the proposed system was not perfect, it achieved improvements in cyberbullying detection. Some possible directions for future works will be discussed in Section 7.

7 FUTURE WORK

Multimodal approaches might help analyze social media content. Although the experiments were based on Twitter data, this type of approach can be used on any social media and any language.

The researchers in [8] proposed a future direction in which ”Internet Memes” were used in detection. This is different from the image sentiment analysis in this paper. The ”Internet Memes” are generally pictures with text, describing a feeling or an event, and are sometimes considered funny. This type of image is widely used by youngsters these days, but the concept is too new to have any research. Analyzing ”Internet Memes” will not only require analysis of both text and vision on the picture but also crowd-sourcing to collect data.

Moreover, there might be some more efficient or accurate approaches than the proposed solution. Concatenating results from two different models might not be the best solution for this purpose, thus future works can focus on the structure and algorithms used in the detection system.

Last but not least, more user information could be considered. For example, [9] discussed the psychological features of users to study cyberbullying behaviors. Besides learning from users' information, future works can also make use of other media contents, such as profile images, background images, videos, and so on. The information might provide a new point of view, but not guarantee that it can improve the result.

Literature Cited

[1] D. Wegge, H. Vandebosch, and S. Eggermont, "Who bullies whom online: A social network analysis of cyberbullying in a school context," *Communications*, vol. 39, p. 415–433, 11 2014.

[2] S. Karmakar and S. Das, "Understanding the rise of twitter-based cyberbullying due to COVID-19 through comprehensive statistical evaluation," 2021.

[3] S. Team, "Cyberbullying: Twenty crucial statistics for 2021," Aug 2021.

[4] "Hana kimura: Netflix star and japanese wrestler dies at 22," May 2020.

[5] J. Sui, "Understanding and fighting bullying with machine learning," 2015.

[6] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pp. 604–607, 2019.

[7] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

[8] A. Pradhan, V. M. Yatam, and P. Bera, "Self-attention for cyberbullying detection," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–6, 2020.

[9] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," vol. 90, p. 101710, 2020.

[10] L. Cheng, A. Mosallanezhad, Y. Silva, D. Hall, and H. Liu, "Mitigating bias in session-based cyberbullying detection: A non-compromising approach," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2158–2168, Association for Computational Linguistics, 2021.

[11] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," vol. 179, p. 115001, 2021.

[12] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," 2018.

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," *CoRR*, vol. abs/1707.07250, 2017.

[14] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 308–317, Oct 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[17] Twitter, "Twitter api." 2020.