BOSTON UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

Thesis

PREDICTING EMOTIONAL INCITEMENT FOR MISINFORMATION DETECTION

by

ZORA CHE

Submitted in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts in Computer Science

2022

ACKNOWLEDGMENTS

From helping me ground my research question to aiding me with experiment design and evaluation, my advisor Bryan Plummer has been integral to this project. I feel fortunate in being given the space to explore a range of new ideas and techniques during this year. His attitude towards research and emphasis on continuous progress has been especially motivating for me.

This project would also not have been possible without Andrea Burns's support and advice, including helping me with data access. Her patience and openness in mentoring me has also given me new perspectives towards research.

Indara Suarez, Tammy Qiu, and Benjamin Pollak have been great role models of researchers for me through my time at Boston University. I feel lucky in calling them friends and I owe my start in research through their influence.

I would like to acknowledge the Kilachand Keystone Program's support of my research, and providing an avenue for me to present my thesis.

Lastly, my ability to attend college and pursue research rests on the organizations and individuals who have lifted me up continuously and restlessly. My work is a reflection of the love, care and resources that others have invested in me.

PREDICTING EMOTIONAL INCITEMENT FOR MISINFORMATION DETECTION

ZORA CHE

ABSTRACT

The rise of social media news sharing has significantly increased the dissemination and consumption of misinformation. Sentiment analysis has been leveraged to aid the detection of misinformation by analyzing the media content using natural language processing for text and computer vision for visual signals. Multi-modal fusion techniques are used to combine image and text features for prediction. Beyond the emotions of the content, the emotions incited in the viewers play a significant role in how misinformation is propagated through social networks. Prior research leveraging incited social emotion assumes the availability of replies data, which may not be possible at desired detection time for preventing misinformation spread. We study predicting emotional incitement for misinformation prediction with only media content. We 1) fine-tune CLIP as an emotional incitement model with collected Twitter Emotional Incitement Dataset, 2) evaluate the efficacy of the predicted emotional incitement feature in a fine-tuned BERT as a misinformation detection model.

On the COVID-19 Rumor Dataset, our emotional incitement model retrieves top incited emotion 73.9% for Recall@1 and 89.13% for Recall@3, demonstrating good transfer ability for tweets. Our method with predicted emotional incitement feature for misinformation detection improved baseline BERT model by 0.2%. Further experiments on an expanded Emotional Incitement Dataset and on a large multi-modal misinformation detection dataset is needed to evaluate the efficacy of predicting emotional incitement for misinformation detection.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The central role of social media in online news sharing has given rise to the dissemination of misinformation. In 2018, one in five Americans says that they get news from social media often (Shearer, 2020). In 2021, 40% of Americans consider social media as an important news source for Covid-19 vaccines (Mitchell & Liedke, 2021). Detecting misinformation with machine learning is an on-going research area which includes analyzing media content, the network effect of the media content, as well as the concepts of the media content (Islam et al., 2020a). Emotion is a high level concept that has been used to discern misinformation, under the assumption that misinformation exhibits different emotional distribution than non-misinformation (Alonso et al., 2021). Most current misinformation detection methods leveraging emotion only consider content emotion, whereas studies with human subjects points to a relationship between the viewer's incited emotion and the spread of misinformation (Han et al., 2020) (Rosenzweig et al., 2021). Existing research using replies data as incited emotion assumes that at test time, replies are available (Zhang et al., 2021). This assumption may not be realistic at early detection time. Therefore, we propose predicting incited emotion with only content data.

1.1 TYPES OF MISINFORMATION

Misinformation is false or misleading information presented as fact, which maybe intentionally or unintentionally created (Asr & Taboada, 2019). Misinformation encompasses a range of contents with different fabrication methods. We broadly cate-

gorize misinformation into algorithmically generated misinformation, and organically generated misinformation. Algorithmically generated misinformation includes methods that employ algorithms which commonly are learning algorithms trained on a significant number of data, producing sophisticated results. Deepfakes of audio, image, or video content aim to impose the likeness of one person onto an existing media content that is imperceptible to an ordinary viewer (Mittal et al., 2020). Neural fake news refers to fake news generated by large language models that mimics the style of real news (Zellers et al., 2019). These methods are data-intensive during fabrication.

We consider organically generated misinformation as media content produced in a low-tech fashion as opposed to algorithmically generated misinformation. Social media posts of rumors and/or conspiracy theories is an example of organically generated misinformation (Asr & Taboada, 2019). Image and videos may be edited through widely available editing applications to mislead the viewer, or be taken out of context (Aneja et al., 2021). For our research, we focus on organically generated misinformation with the presumption that the misinformation content is not created in an algorithmic manner.

1.2 EMOTION AND MISINFORMATION

Misinformation's tendency for virality often contains appeals to emotion, which can increase persuasion (Ecker et al., 2022). Several studies have suggested the relationship between incited emotion and the spread of misinformation through human experiments. Anger has been analyzed as contributing to the spread of COVID-19 misinformation (Han et al., 2020). Happiness and surprise has been associated with worse truth discernment of COVID-19 headlines among social media

users in Nigeria (Rosenzweig et al., 2021). Martel et. al demonstrated that reliance on emotion is correlated with worse truth discernment at the news headline-level for all types of emotion expressed (Martel et al., 2020). Bago et. al conducted further experiments that validated this claim, except for the emotion anger, which they found to increase truth discernment (Bago et al., 2021). Such findings correspond to the analysis of misinformation, categorizing one of its key features as strong emotional incitement in viewers (Molina et al., 2021).

1.3 PREDICTING INCITED EMOTION FOR MISINFORMATION DETECTION

Motivated by misinformation's strong appeals to emotion, we study using predicted incited emotion as a feature for misinformation detection. Existing work incorporating incited emotion relies on the availability of comments data to produce emotion features encompassing signals at a lexical level and a semantic level (Zhang et al., 2021). We consider this assumption to be less realistic for early detection of misinformation on social media platforms. Misinformation could have severe impacts in just a few minutes, making early detection crucial (Guo et al., 2019).

Therefore, our work focus on predicting incited emotion through fine-tuning the large multi-modal model CLIP (Contrastive Language-Image Pre-training). We then incorporate the trained emotional incitement model into a misinformation detection pipeline. This also allows evaluation of the efficacy of incited emotion as a detection feature on a broader range of misinformation datasets since reply data are not always collected.

We consider the benefits of this approach to be threefolds: 1) Learning to pre-

dict incited emotion prediction is a step towards general machine social intelligence via generating emotion models of another, 2) Emotional incitements correlation with viewers change of belief suggests alludes to a higher priority in moderating emotionally inciting content may help to curb possible misinformation, 3) Moving away from relying on a sizable replies data help transition towards early detection.

We summarize our contributions thorough this work be: 1) A multimodal Twitter dataset with tweets, images, and replies to tweets for research on emotional incitement prediction, for the topic of Covid-19. 2) Demonstrating CLIP's transfer ability for predicting incitement emotion. 3) Establishing model pipelines and results on incorporating incited emotion for prediction on the COVID-19 Rumor Dataset (Cheng et al., 2021).

CHAPTER 2
RELATED WORK

We ground our research within machine learning for misinformation detection, specifically content classification with a feature-based approach. We consider related work in machine learning, including those that leveraged either content emotion or incited emotion.

2.1 MACHINE LEARNING FOR MISINFORMATION DETECTION

Expanding on the taxonomy proposed by Potthast et al. categorizing fake news detection methods, we group learning-based approaches for misinformation detection into three categories: knowledge-based, context-based, and content-based (Potthast et al., 2017).

Knowledge-based methods aim to determine whether the information presented is supported by facts. It is closely related to information retrieval, which employs knowledge bases for automatic fact-checking. The detection of misinformation has been formulated as a link prediction task of the possibility of the concepts being connected (Shi & Weninger, 2016). Recent work also factors fact-checks—extract structured information from fact-checking articles—by formulating it as a sequence tagging problem (Jiang et al., 2020).

Context-based methods focus on non-content information such as the propagation pattern and meta-data of the posts or poster. Its focus on context often relates to finding social media bot accounts—accounts controlled in part by software. Bots contribute to a significant amount of misinformation on social media (Himelein-Wachowiak et al., 2021). Detection methods of bots include those regarding net-

work effects, such as synchronized behavior (Mazza et al., 2019), activity stream (Badawy et al., 2018), or analyzing meta-data of the accounts (Yang et al., 2020).

Content-based methods focus on the classification of content, either through style or through detecting artifacts that may present in generated media. Deep learning models, especially large pre-trained language models have demonstrated strong performance in fake news detection as a downstream task (Islam et al., 2020b) (Kaliyar et al., 2021). Generators of media such as GROVER for neural fake news have been found to be a strong detector at the same time (Zellers et al., 2019). Consistency between modalities for multi-modal media has also been found to boost detection performance (Xue et al., 2021) (Mittal et al., 2020).

2.2 EMOTION RECOGNITION FOR MISINFORMATION DETECTION

Most misinformation detection methods utilizing emotion has emphasized content emotion, as compared to incited emotion.

Leveraging sentiment analysis or affective signals in multi-modal content has produced promising results for misinformation detection (Mittal et al., 2020) (Alonso et al., 2021).

Sentiment analysis is a natural language processing technique that determines whether information is presented in a positive, negative or neutral way, as well as how strong the opinion is (Alonso et al., 2021). Expression of sentiment is important in the spread of misinformation. Dickerson et al. has found that sentiment is sufficient for distinguishing between social media bots and real users (Dickerson et al., 2014). This finding corresponds to the observation that misinformation often contains strong appeals to emotion (Ecker et al., 2022).

For Deepfakes detection, consistency of affective signals between audio and

video modality has also been shown to have discriminative power (Mittal et al., 2020).

Zhang et. al introduces dual emotion features—including both content emotion and social emotion features — for detection (Zhang et al., 2021). Their research question of the relationship between emotion in content and emotion incited in crowd as a feature for fake news prediction is most closely related to our research of predicting incited emotion for misinformation detection. Their method assumes the availability of comments, and construct rich features for social emotion and content emotion respectively. Each feature contains both semantic-level signal as well lexical-level signal, including emotional classification, emotional intensity features, and sentiment scores. The content emotion and social emotion is then concatenated to be used as a feature for a misinformation detection model. They demonstrate the effectiveness of the dual emotion feature as an addition to models such as BiGRU (Bidirectional Gated Recurrent Unit) and BERT (Bidirectional Encoder Representations from Transformers). Their method out performed the state-of-the-art task-related emotional features for Weibo-16 Dataset, Weibo-20 Dataset and the RumorEval-19 Dataset (Ma et al., 2016) (Gorrell et al., 2019). We note that improvement on the Chinese datasets (Weibo-16, Weibo-20) is considerably greater than for the English dataset (RumorEval-19), which could due to the difference between rumor and fake news, or the smaller dataset size of RumorEval-19.

Our research contrasts with Zhang et. al's work in that we do not assume the availability of replies data to better simulate an early detection framework. We aim to predict incited emotion from content itself, then utilize it as an additional feature for misinformation detection methods.

CHAPTER 3

METHOD

3.1 EMOTIONAL INCITEMENT DATASET COLLECTION

We consider social media replies to be a proxy for incited emotions in viewers. Manually labeling emotional incitement for multimodal content can be a labor-intensive process. Thus, we consider the emotion expressed in the replies to the social media content to be the supervision for training emotional incitement predictions.

We found that there is few multimodal datasets of social media content with text, images, and replies. Thus, we collected a Twitter dataset during Feburary and March 2022.

The Twitter Emotional Incitement Dataset has 11,900 entries, each with text content, image content, replies to the content, emotion labels for the replies, and meta-data of the tweet. We curated the dataset through Twitter API v2 with Academic Access. We first searched for recent tweets in English that have the keyword "covid," and have image attachments. Out of the queried results, we only consider tweets with 3 or more replies. Using these tweet IDs as a filter, we query for all the replies of tweets satisfying our multimodal criteria. Lastly, we obtained emotion labels for both the original tweet content and its replies using SpanEmo, a distribution based emotion prediction method. We trained the SpanEmo model on the SemEval-2018 Competition dataset E-c (English), a multi-label multi-category emotion benchmark dataset (Mohammad et al., 2018). We trained our model with 11 emotion labels: "anger", "anticipation", "disgust", "fear", "joy", "love", "optimism", "pessimism", "sadness", "surprise", "trust." To aggregate emotion over the

replies, we choose to have two vectors: **mean emo**—the average emotion across the replies (average predicted emotions for all replies), **all emo**—all emotions present across the replies (indicator vector of predicted emotions for all replies).

Each data entry is also accompanied with tweet ID of the content tweet, date of posting, and public metrics of the tweet such as retweet number, like number, and quote number.

3.2 PREDICT INCITED EMOTION FOR MISINFORMATION DETECTION ON WEIBO-16

Building upon Zhang et. al's model pipeline on Weibo-16, we tested misinformation prediction for BiGRU with ground truth incited emotion features, and for BiGRU with predicted incited emotion features to motivate our approach.

BiGRU, or Bidirectional Gated Recurrent Unit, refers to a recurrent network consists of two GRUs, where one takes the input in a forward direction, and the other in a backward direction. GRU (Gated Recurrent Unit) is a recurrent neural network with a gating mechanism that has an update gate and reset gate to solve the vanishing or exploding gradient problem for recurrent network (Cho et al., 2014).

We consider only the 11-category labeled emotions as the emotion feature. The dataset contains labels of 11-category emotion vector from mean pooling and max pooling over replies for each data entry. We test for each type of emotion vector in our experiment.

The incited emotion is trained with BiGRU on the Weibo-16 dataset with the 11-category emotion vectors as ground truths. For both the emotional incitement training and the misinformation training, binary cross entropy was used as an loss

for multi-label prediction.

We also reproduced ablation study of BiGRU with semantic feature only, Bi-GRU with semantic feature and publisher emotion feature for comparison.

3.3 EMOTIONAL INCITEMENT MODEL

We experiment on the Twitter Emotional Incitement Dataset through zero-shot setting of pre-trained CLIP, as well as finetuning CLIP.

3.3.1 CLIP Zero-shot for Emotional Incitement

CLIP (Contrastive Language-Image Pre-training) is a large multi-modal model pretrained on millions of text and image pairs which performs strongly on a range of downstream tasks including under zero-shot settings (Radford et al., 2021).

We perform CLIP zero-shot evaluations to gauge CLIP image and text encoder's ability to infer incited emotion without training on the Twitter Emotional Incitement Dataset. We encode each modality with pre-trained CLIP, then calculate cosine similarity between the encoded content with a list of CLIP encoded text representing all the emotional categories we consider. Thus, for each text or image, CLIP produces a list of probabilities for all the emotions matching with the given content.

We considered different types of filler phrase for the emotion categories to improve performance, including long descriptive filler phrases and no filler phrase.

3.3.2 Fune-tune CLIP for Emotional Incitement

We fine-tune pre-trained CLIP for the emotional incitement task using the emotion labels from replies as ground truth.

We use pre-trained CLIP with frozen parameters to encode image and text, respectively. The feature encoding is then input into two fully connected trainable layers, with the latent dimension of 100, and output dimension of 11 for the emotion category prediction. We use binary cross entropy as a loss for multi-label classification.

We optimized using Adam optimizer, an efficient stochastic optimization that only requires first-order gradients, which it then computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients (Kingma & Ba, 2014).

We compare results for using image feature only, using text feature only, and using both image and text feature (concatenated) for the two fully connected trainable layers.

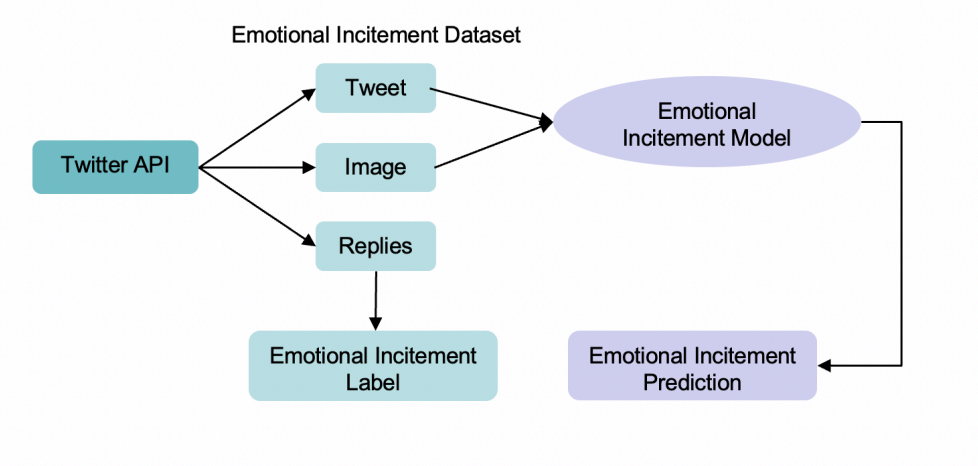

Figure 3.1: Emotional Incitement Model — Pipeline

3.4 MISINFORMATION DETECTION MODEL

We test our proposed direction on COVID-19 Rumor Dataset, a text-only dataset with tweets on Covid-19 fact-checked manually (Cheng et al., 2021). The dataset

has 540 tweets labeled as fake, 1040 tweets labeled as real, and 1125 tweets labeled as unverified (Cheng et al., 2021).

For the baseline model, we fine-tune a pre-trained BERT model for the misinformation detection. BERT, or Bidirectional Encoder Representations from Transformers, is a large language model trained through self-attention that produces strong results on a number of natural language processing tasks, including fake news detection (Devlin et al., 2018)(Kaliyar et al., 2021).

We take the pooling output of BERT, which is average pooled feature of the sequence output of BERT. We consider it to be an aggregate feature of the entire sentence. We append two fully connected trainable layers, which then outputs real or fake.

For our proposed method, we concatenate the predicted emotional incitement feature with the BERT pooling output, before passing through the two fully connected trainable layers . The model is trained with binary cross entropy loss for multi-label output. We also optimized the model using the Adam optimizer.

We vary the latent dimension of the fully connected layers for increasing performance.

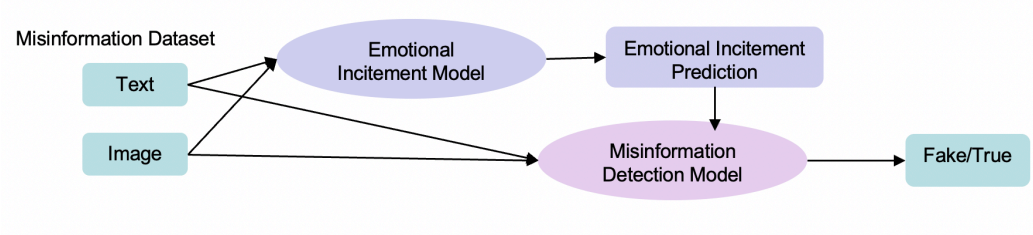

Figure 3.2: Misinformation Detection Model — Pipeline

3.4.1 Upsampling COVID-19 Rumor Dataset

We consider only the labeled real and fake tweets in our training and evaluation process. Since the COVID-19 Rumor dataset is imbalanced with around 34% fake and around 66% real, we create balanced validation and test splits, and upsample the training split (Table 3.1). All data entries in validation and test splits are unique, while the data in train split may be repeated since we sampled with replacement.

Table 3.1: COVID-19 Rumor Dataset Splits

	Total	Upsampled Training	Validation	Test
Fake	540	790	94	95
Real	1040	790	94	95

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 PREDICT INCITED EMOTION FOR MISINFORMATION DETECTION ON WEIBO-16

Table 4.1: Emotional Incitement for Weibo-16 Misinformation Detection

	Accuracy [%]
SEMANTICS	64.0
SEMANTICS+PUBLISHER EMO	68.1
SEMANTICS+DUAL EMO	80.4
SEMANTICS+GROUND TRUTH SOCIAL EMO	73.4
SEMANTICS+PREDICTED SOCIAL EMO	68.8

We motivate the feasibility of social emotion feature for misinformation prediction on the Weibo-16 dataset through training and evaluating the BiGRU model with social emotion features (incited emotion features), as well as comparison with baselines. We find that using ground truth social emotion in the form of **mean emo**—the average emotion category predictions across the replies (average predicted emotions for all replies)— achieves 9.4% increase in accuracy compared to semantics feature only. Using ground truth social emotion also improved upon using a rich publisher emotion feature—with lexical-level and semantic-level signal— by 5.3%. We find that using the predicted social emotion also improved upon methods leveraging only content data: semantics and predicted social emotion improved upon the results of only semantics by 4.8 %, and improved upon the results of semantics with publisher emotion by 0.7%. We also note that the results with ground truth social emotion and predicted social emotion lag behind result using dual emotion. This matches our expectation since our emotion features do

not capture fine-grained lexical-level signal from ground truth replies. However, our results suggest coarse-grained incited emotion as a viable feature for misinformation detection, especially if we can achieve predicted incited emotion close to ground truth.

4.2 EXPERIMENTAL RESULTS ON EMOTIONAL INCITEMENT MODEL

In order to achieve predicted incited emotion close to ground truth, we train and evaluate our emotional incitement model on the test split of the Twitter Emotional Incitement Dataset through calculating mean ranks of the top predicted emotion. Mean ranks calculates the average of the ground truth rank of the predicted emotion. Here, we considered the top predicted emotion category. The closer the mean rank to 1, the more accurate the prediction is. We implemented the mean rank using averages to break ties (1.5 and 1.5 for top two emotions with the same scores). Therefore the theoretical lowest value we may achieve in this section is around 1.2.

4.2.1 CLIP Zero-shot for Emotional Incitement

Table 4.2: Mean Ranks of Top Predicted Incited Emotion by CLIP Zero-shot

	Text Features	Image Features	Random (control)
Top-1 pred of all emotion	5.87	6.12	6.09
Top-1 pred of mean emotion	5.90	6.09	5.99

We find that without tuning, CLIP zero-shot prediction on the Twitter Emotional Incitement Dataset performs poorly. Predictions based on encoded image features perform worse than randomly picking an emotion. Predictions based on encoded text features performed slightly better than random.

Our best zero-shot results are from using a descriptive filler phrase—"This incites the emotion of <emotion>"—as opposed to a less descriptive filler phrase, or no filler phrase.

4.2.2 Fune-tune CLIP for Emotional Incitement

Table 4.3: Mean Ranks of Top Predicted Incited Emotion by Fine-tuned CLIP

	Text Features	Image Features	Text + Image Features
Top-1 pred of all emotion	2.44	2.55	2.41
Top-1 pred of mean emotion	2.41	2.45	2.38

We find fine-tuning CLIP produced significantly better mean rank results as compared to zero-shot CLIP. We trained our models over 30 epochs, and achieved mean rank of around 2.5 for top 1 predicted emotion category, across the modality settings we've considered. We found that using both text and image modality for emotional incitement prediction produced the lowest mean rank for both all emotion, and mean emotion. We found using only text features produced the second best mean ranks performance.

4.3 EVALUATE EMOTIONAL INCITEMENT PREDICTION ON COVID-19 RUMOR DATASET

Table 4.4: Mean Ranks of Predicted Incited Emotion for COVID-19 Rumor Dataset

Replies Ground Truth	Mean Rank of Predicted Emotion
Rank 1	2.50
Rank 2	3.10
Rank 3	4.60
Rank 4	4.86

Table 4.5: Recall@K of Predicted Incited Emotion for COVID-19 Rumor Dataset

	Recall@k of Top Emotion Prediction [%]
Recall@1	73.90
Recall@3	89.13

Evaluating the trained Emotional Incitement Model on the COVID-19 Rumor Dataset, we find it transfers well with achieving low mean ranks for top emotions (Table 4.4), as well as recalling the top incited emotion (Table 4.5). We consider the results with room for improvements both for mean ranks of top emotions, and recall for top-1 emotion. We observe the mean rank results match evaluation results on the test split of the Twitter Emotional Incitement Dataset, which demonstrate good transfer that is likey due to using the same source of data—Twitter.

4.4 EXPERIMENTAL RESULTS ON COVID-19 RUMOR DATASET

Table 4.6: Mean Accuracy for Baseline and Proposed Method for COVID-19 Rumor Dataset

	$N = 8$	
	Accuracy (mean) [%]	Standard Deviation
BERT	70.06	2.328
BERT + Emo Feature	70.26	2.179

The baseline and the proposed method with incited emotion feature are trained over 100 epochs, and evaluated on the test split of COVID-19 Rumor Dataset. We find that the proposed method improves slightly upon the baseline model of fine-tuned BERT, by 0.2%. We also find that across 8 models trained with different seeds for dataset splits, our proposed method has a smaller variance for accuracy by 0.149% .

CHAPTER 5
DISCUSSION

5.1 SUMMARY OF RESULTS

We find that for Weibo-16, using predicted social emotion trained on Weibo-16 improves upon semantic only BiGRU model, as well as BiGRU model with semantic and publisher emotion, by 4.1% and 0.8% respectively. On the Twitter Emotional Incitement Dataset, we found that our predicted mean ranks of emotional incitement is close to the ground truth rank of emotional incitement as labeled by replies. We find that the trained emotional incitement model transfers on the COVID-19 Rumor Dataset, with predicted mean ranks close to ground truth mean ranks. There exists room for improvement for both the recall of the top emotion, as well as the mean ranks for the top predictions. On COVID-19 Rumor Dataset, our misinformation detection method with predicted incited emotion has a 0.2% improvement as compared to fine-tuned pre-trained BERT as a baseline, as well as a smaller variance across 8 trained models with different seeds. We find the improvement for misinformation detection in our experiments is marginal, and that the efficacy of the predicted incited emotion feature is inconclusive on the COVID-19 Rumor Dataset.

5.2 LIMITATIONS

Our Twitter Emotional Incitement Dataset was collected during a short period (two weeks) with a single topic, and we find that the emotional distribution of the dataset to be skewed due to the topic of Covid-19, and the corresponding rise of Omicron during the collection period. Our experiments has only demonstrated

transfer on predicting emotional incitement within the domain of tweets about Covid.

In our misinformation detection experiments with COVID-19 Rumor Dataset, we were only able to consider text features for producing emotional incitement features. Based on the emotional incitement test split results on the Twitter Emotional Incitement Dataset, we expect better emotional incitement prediction on misinformation datasets with image and text modalities.

COVID-19 Rumor Dataset is also a small dataset for misinformation detection, with 1580 unique entries which could have affected tuning and performance.

There exists a concern of the incited emotion feature recognizing polarized and sensational content that may not be false (Molina et al., 2021). Unlike real news that is often produced with a set of prescribed styles and guidelines, real information shared by humans on social media may contain a greater range in sensationalization, with a possible goal of bringing attention to an issue. We consider it an important question to account for strong emotional appeals that may be used by activists.

5.3 FUTURE WORK

To fully evaluate the efficacy of predicted incited emotion for misinformation detection, we plan on improving the emotional incitement feature, and validate on large multi-modal datasets.

To improve the emotional incitement feature, we would like to expand the Twitter Emotional Incitement Dataset to include further topics and a greater range of emotional distributions. We would like to consider including only tweets with 5 or more replies instead of 3 or more replies to compare the results. We would like to

experiment including emotional valence and/or stance (in support, not in support) as a part of the emotion vector, in addition to the current multi-label categories of emotions. We would like to continue model design and tuning for increasing performance in the emotional incitement task. We would also like to have detailed analysis of gap between the ground truth content emotion, and the ground truth social emotion in our Twitter Emotional Incitement Dataset.

To validate our results further, we plan on training and testing on large multi-modal datasets, starting with Fakeddit. We would like to observe the emotional incitement prediction on Fakeddit, which has a different data source. We would also like to evaluate the effect of using predicted emotion feature with a range of misinformation baseline models, to evaluate the method for flexibility.

5.4 CONCLUSION

We collected and labeled the Twitter Emotional Incitement Dataset for studying emotional incitement. Fine-tuning CLIP produces low mean ranks both on the Twitter Covid Emotional Incitement Dataset, and the COVID-19 Rumor Dataset. Using the predicted emotional incitement features, we find improvement upon the baseline fine-tuned BERT model to be marginal. We believe expansion to the dataset by incorporating a greater range of topics and distributions of incited emotions may be needed to improve the emotional incitement prediction. Furthermore, we believe training and evaluating on large multi-modal datasets such as the Fakeddit dataset is appropriate for further evaluation.

BIBLIOGRAPHY

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics*, *10*(11), 1348.

Aneja, S., Bregler, C., & Nießner, M. (2021). Catching out-of-context misinformation with self-supervised learning. *CoRR*, *abs/2101.06278*.

Asr, F. T., & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1), 2053951719843310.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, (pp. 258–265). IEEE.

Bago, B., Rosenzweig, L., Rand, D., et al. (2021). Emotion may predict susceptibility to fake news but emotion regulation does not seem to help.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S., & Bogdan, P. (2021). A covid-19 rumor dataset. *Frontiers in Psychology*, *12*.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dickerson, J. P., Kagan, V., & Subrahmanian, V. (2014). Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, (pp. 620–627). IEEE.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., & Derczynski, L. (2019). Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, (pp. 845–854).

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2019). The future of misinformation detection: new perspectives and trends. *arXiv preprint arXiv:1909.03654*.

Han, J., Cha, M., & Lee, W. (2020). Anger contributes to the spread of covid-19 misinformation. *Harvard Kennedy School Misinformation Review*, *1*(3).

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., Curtis, B., et al. (2021). Bots and misinformation spread on social media: Implications for covid-19. *Journal of Medical Internet Research*, *23*(5), e26933.

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020a). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, *10*.

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020b). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, *10*(1), 1–20.

Jiang, S., Baumgartner, S., Ittycheriah, A., & Yu, C. (2020). Factoring fact-checks: Structured information extraction from fact-checking articles. *Proceedings of The Web Conference 2020*.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765–11788.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, *5*(1), 1–20.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, (pp. 183–192).

Mitchell, A., & Liedke, J. (2021). About four-in-ten americans say social media is an important way of following covid-19 vaccine news. https : / / www . pewresearch . org / fact-tank / 2021 / 08 / 24 / about-four-in-ten-americans-say-social-media-is-an-important-way-of-following-covid-19-v

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: A deepfake detection method using audio-visual affective cues. *CoRR*, *abs/2003.06711*.

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, (pp. 1–17).

Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). fake news is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, *65*(2), 180–212.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *CoRR*, *abs/1702.05638*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, (pp. 8748–8763). PMLR.

Rosenzweig, L. R., Bago, B., Berinsky, A. J., & Rand, D. G. (2021). Happiness and surprise are associated with worse truth discernment of covid-19 headlines among social media users in nigeria. *Harvard Kennedy School Misinformation Review*.

Shearer, E. (2020). Social media outpaces print newspapers in the u.s. as a news source. https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/.

Shi, B., & Weninger, T. (2016). Fact checking in heterogeneous information networks. *Proceedings of the 25th International Conference Companion on World Wide Web*.

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, *58*(5), 102610.

Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, (pp. 1096–1103).

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *CoRR*, *abs/1905.12616*.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, (p. 34653476). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442381.3450004.