TOWARDS IDENTIFICATION OF MICROAGGRESSIONS IN REAL LIFE AND SCRIPTED CONVERSATIONS USING CONTEXT-AWARE MACHINE LEARNING TECHNIQUES.

Anonymous authors
Paper under double-blind review

ABSTRACT

The advent and rapid proliferation of social media have brought with it an exponential growth in hate speech and overt offensive language, with one of the most subtle yet pervasive subcategories of hate speech being Microaggressions (MA). MAs are unintentional, hostile, derogatory, or negative prejudicial slights and insults toward any group, particularly culturally marginalized communities and growing bodies of research are linking long-term MA exposure to serious health problems. The scarcity of studies leveraging AI techniques to identify MAs in text and in spoken conversations, coupled with the lack of investigative analysis on the impact of context on the performance of algorithms used for this task, makes this a relevant topic for the AI community. In this paper, we explore the degree of effectiveness of MAs detection often found in spoken human communications across various contexts (e.g., workplace, social media, conversations) using Machine Learning models. We further examine the extent that art may imitate life, by comparing the ability of these models trained on real-life conversations to infer MAs, occurring in scripted Television shows. We apply a Support Vector Machine (SVM) classifier using N-grams and contextual modeling representation, using the Robustly Optimized Bidirectional Encoder Representation for Transformer (RoBERTa) model, whose performance is evaluated based on its pretraining size and ability to accurately detect hate speeches, with comparative results from BERT based-uncased and the HateBERT model respectively. Overall, the results show that contextual transformer models outperform simpler context-free approaches to classifying MAs collected from surveys and online blogs. We also found that these models trained on real-life conversations could infer MAs in scripted TV settings, though at reduced levels, and equal rates, suggesting there may be a disconnect between contexts of MA found in art and those from real life.

1 RESEARCH OVERVIEW AND BACKGROUND

Text classification of offensive speech explores the relationship between expressed language and subjective perceptions. This topic is emerging as an important sub-field of Natural Language Processing(NLP) (Fortuna et al., 2021), with applications to domains like negative sentiment (Taboada, 2016), hate speech (Jacobs & Potter, 1998; Walker, 1994), and toxicity (Kolhatkar et al., 2020). One less examined area is the topic of MAs, where the speech is offensive to an individual's identity, and often manifests in subtle ways that may not have even been intended by the speaker. The current project seeks not only to examine the ability of NLP methods to detect this type of speech in texts but also, to understand the nature of how this speech unfolds. Because prejudice and bias might be reflected in society's media, we examine how well models trained on real-world examples of MAs generalize to drawing inferences on MAs expressed on television, assessing how well "Art imitates Life" in the context of MAs. In recent years, network science and NLP research has focused on identifying offensive speech extracted from online social media platforms such as Twitter. Identifying offensive speech, however, is quite challenging (MacAvaney et al., 2019), and in some cases, biased (Davidson et al., 2019). One difficulty is that not all the prior work agrees on what constitutes "offensive speech," using a myriad of terminology, such as "abusive language," "toxicity," "online harassment," "cyberbullying," "damaging speech," and "hate speech." To address these

issues, researchers examining offensive speech often focus on narrower definitions. We adopt this perspective to the current research project, focusing specifically on MAs. MAs are defined as brief and commonplace daily verbal, behavioral, or environmental indignities whether intentional or unintentional, that communicate hostile, derogatory, or negative prejudicial slights and insults toward any group, particularly culturally marginalized groups (Sue et al., 2007). They can be behavioral but are often discussed in a racial context however. Moreover, these aggressions can be expressed verbally, i.e., through comments or questions that are hurtful or stigmatizing. MAs are especially important to examine in the context of NLP, mainly because MAs are sometimes unintentional, and therefore, there may be a more implicit process guiding their verbal construction. This implicit nature highlights the role of NLP via ML in identifying these statements, in the absence of explicitly defined rules. Additionally, MAs are often subtle, versus overt, slights, suggesting that the language is similar to that found in daily conversions. Whereas bag-of-words Machine Learning (ML) models may perform well when words are different from one category to another, the contextual nature of MAs makes them more suited to models that are more adept at processing sequences (e.g., N-gram models), and the interrelationships between words (e.g., transformers). MAs may also have a reciprocal relationship between life and society. While media may try to avoid certain themes or present an unrealistically flattering portrayal of society, MAs, because of their subtle nature, may be more pervasive in these mediums. Art may imitate life, and therefore we would expect that the information learned from MAs in real-life contexts could help classify them in media. The rest of the paper is structured as follows: Section 2 summarizes literature pertaining to toxic, hate and multimodal offensive language while in section 3, we dive into the research methodology, which involves a more detailed account of the datasets and models used, as well as the text preprocessing and feature engineering techniques implemented. Section 4 describes the results obtained from experimentation and section 5 concludes our work, highlighting our research limitations and suggesting future directions.

2 RELATED WORKS

This section summarizes some research related to modelling hate speeches, toxic, and multimodal offensive comments implemented using ML and NLP methods.

2.1 TOXIC COMMENTS

"Toxicity" is an umbrella term that represents general offense and different types of "aggression" (Kolhatkar et al., 2020). Toxic comments are rude, unreasonable, and disrespectful remarks that are likely to make someone uncomfortable, and leave a discussion (Van Aken et al., 2018). Toxicity is often considered a multidimensional construct capturing elements of personal attacks (Wulczyn et al., 2017), abuse (Nobata et al., 2016), harassment (Bretschneider et al., 2014), threats (Spitzberg & Gawron, 2016), profane, obscene or derogatory language (Sood et al., 2012; Wang et al., 2014; Davidson et al., 2017), inflammatory language (Wiebe et al., 2001), and hate speech (Warner & Hirschberg, 2012; Djuric et al., 2015; Waseem & Hovy, 2016). The Kaggle toxic comment classification challenge dataset [1], defines six classes of toxic speech namely: toxic, severe toxic, obscene, threat, insult, and identity hate, and has often served as a benchmark for automated classification of toxic speech. Over the years, researchers have applied various ML and Deep Learning techniques to help improve toxic speech detection in texts, and hence design sophisticated real-time toxic speech detectors. Early work that dichotomized toxic speech into toxic and non-toxic classes was implemented using Convolution Neural Networks(CNN) and achieved an F1 score of 0.92 (Georgakopoulos et al., 2018). The issue of class imbalances which complicated the categorization of these toxic speeches resulted in the creation of more robust datasets (Van Aken et al., 2018; Juuti et al., 2020), and the researchers applied more sophisticated models such as LSTM, RNN, CNN, and bidirectional GRUs (Saif et al., 2018; Zaheri et al., 2020) and an ensemble of them all (Ibrahim et al., 2018), for modeling and testing the approaches. Currently, State-Of-The-Art (SOTA) transformer models (Yang et al., 2019) have achieved the highest F1 scores (Ghosh & Kumar, 2021). Part of the difficulty with examining toxic speech is its multifaceted nature. As stated by, (Gilda et al., 2022), "Detecting patronizing and condescending language is still an open research problem because, amongst many reasons, condescension is often shrouded under 'flowery words' ", we believe the same can be said about MAs. In their research, the authors utilize a CNN-LSTM to classify

[1] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview

comments into eight different classes and the results reported, show neutral sentiments predictions towards utterance with condescending tones. Moreover, a joint "Perspective" project by Google and Jigsaw uses ML to automatically detect toxic online comments from adversarial examples, by attributing a toxicity score to each comment, with the prime objective of lowering a comment's toxicity score while still retaining its negative context. As eloquently put, the researchers believe "Detecting subtler forms of toxicity requires idiosyncratic knowledge, familiarity with the conversation context, or familiarity with the cultural tropes". Therefore, in the future, we hope more consideration is placed on researching toxic comments, subtle or not by analyzing the speech context to better grasp their nuanced nature.

2.2 HATE SPEECH

The challenges of automatically detecting MAs from text stem from the nuances associated with MAs, hence making its collection and annotation quite difficult. In (Waseem & Hovy, 2016), the authors categorize MAs as a class of hate speech and weigh the importance of specific extra-linguistic features associated with character N-grams. To determine their impacts on hate speech detection, they create a list of criteria based on critical race theory, which can be used as a guideline for labeling hateful slurs. Similarly, (Davidson et al., 2017) implemented unigram, bigram, and trigram features on multiple ML algorithms including SVMs to automatically classify hateful texts using a 5-fold Cross Validation(CV) model structure. Furthermore, to help capture all different aspects of the hate speeches during modeling, a multiview SVM classifier approach for hate speech categorization, designed and built from a combination of multiple view classifiers each comprising word TF-IDF from Unigrams to 5-grams, fitted with a Linear SVM model and a transformer BERT model, is proposed and presented in (MacAvaney et al., 2019). Also, authors like (Burnap & Williams, 2016) endeavor to automate the detection and classification of hateful text from a range of protected characteristics such as race, disability, and sexual orientation. Their approach involves building several different models and employing N-gram feature engineering techniques to generate relevant numeric vectors from the predefined dictionary of hateful words gathered from social media posts, then modeling them using an SVM classifier. Another method for automatically detecting racial MAs from text is proposed by (Ali et al., 2020). Here, the authors design a lexicon to help filter these MAs from texts and then classify them as either Racial or Non-Racial. The experiment utilizes the trigram features engineering approach and assesses the accuracy of 7 ML algorithms including the SVM classifier. An unsupervised experimental approach for identifying MAs in text by leveraging the inherent flaws of pretrained language embeddings is presented in (Sabri et al., 2021). Primarily, the study uses racial and gender MAs extracted from the `https://www.microaggressions.com` Tumblr website, and a novel unsupervised pretrained word embedding algorithm, designed from Fastext, Word2Vec, and GloVe word embeddings. The unsupervised model developed is trained with the most optimal parameters selected through grid searching and tested on unseen data from diverse backgrounds, revealing some promising results. In the same vain, (Caselli et al., 2020) present a novel implementation of the BERT transformer models for detecting abusive, offensive, and hateful languages by utilizing the HateBERT designed from 3 datasets, and comparing the results obtained against that of the parent BERT base model. The results reported show HateBERT consistently outperforms BERT, on all criteria considered. A BERT-based transfer learning approach for hate speech detection is proposed by (Mozafari et al., 2019) . The implementation involves two publicly available Twitter-based datasets and several different fine-tuning strategies on the BERT algorithms to determine their impact on the overall model's performance. In the end, the results validate the efficiency of the BERT-based model at detecting hate speech as well as infiltrated biases from the data annotation process. This prior findings particularly important as they highlight the relevance of N-grams and SVM models in hate speech, MA and offensive language detection and classification. Additionally, we also see the opportunities to better understand MAs provided by more sophisticated, contextual transformers models.

2.3 MULTIMODAL OFFENSIVE LANGUAGE

According to (Poria et al., 2021), language is inherently multimodal and there are characteristics found in the speakers' communication styles, that can provide contradictory information on how a listener perceives the information. In their research, the authors examine multiple multimodal features such as facial expressions, emojis, and acoustics, which could be used to enhance the effectiveness of NLP algorithms in addressing specific tasks. We believe some of these multimodal

features, implemented using NLP could be relevant to our research, as other demographics such as the victims' race, ethnicity, and gender could also contribute to distorting the initial intention of the message. Additionally, the offender's body language including utterances, and gestures, could be key to deciphering the intention behind the MA instance. This work could prove scientifically relevant as we plan to also examine how contradicting facial expressions and acoustic patterns manifest during a micro aggressive encounter. Future use of multimodal criteria may also be instrumental in providing more information about the context of the message, through the identification of gesturing, nodding, and other features that textual information just cannot provide. For this same purpose, other researchers like (Sharma et al., 2022) are investigating and analyzing memes, especially ones with harmful messaging found on social media, to determine their effects in providing contextual information from an exchange. For this experiment, the authors released a 10K human-annotated dataset containing internet memes with the sentiment, type of humor (sarcastic, humorous, offensive, or motivative), and the intensity of the humor, labeled. It is worth noting that this particular dataset has been instrumental in creating multimodal models to complete the three sub-tasks of the Memotion Analysis Challenge 2021 Task 8 namely, sentiment analysis; humor classification; and the scales of sentiment. Overall, we believe this work is quite important as it helps develop methods for evaluating texts with visual components because despite them not being verbal, they could often be characterized as micro aggressive. Furthermore, the image characteristics including the scenery and background images could provide context for the future study of MAs.

3 METHODOLOGY

In this section, we provide a detailed description of the dataset and models used, an overview of the modelling criteria, parameters and proposed method.

3.1 DATASET DESCRIPTION

This research utilizes 3 different data sources: 1) web scraped, 2) student surveys and interviews, and 3) data collected from American TV shows. The first text corpus comprising MA extracted from the Tumblr Website Microaggression.com was manually annotated by a diverse group of 5 labelers, trained on spotting and documenting MAs from research from (Sue et al., 2007; Law et al., 2019; Mekawi & Todd, 2018), and pop-culture examples from social media, and articles like "What Exactly Is a Microaggression." (Desmond-Harris, 2015). These data collection and extraction approaches were undertaken while being fully cognizant of the fact data from these crowd-sourcing websites are often not vetted for authenticity, we assume that the user experiences recorded are genuine and hence valid since more reflexive of real-life scenarios. However, for quality control, each instance of MA collected underwent a cyclic review and inter-raterratability process where in the end, the MA retained were those with an average acceptance threshold of at least 70%. The next phase of this process involved matching the MA sampled, with related non-MA instances from the same shows using the cosine similarity to get the best match. This involved scrapping random collections of blogs that had engaged with any posts or content from the Tumblr website, stripping away all comments of length less than 2, and vectorizing the remaining comments using the TD-IDF document term vectorizer. Overall, this process resulted in about 51 425 comments, which were then matched against the initial 1713 MA quotes scraped earlier, hence resulting in a balanced set comprising 1713 instances of MA and non-MA respectively, each at a 70% threshold cosine similarity score. This dataset contains a total of 51 041 words with 5823 unique words, and on average, 14.91 words per statement, which corresponds to a Standard Deviation (SD) of 13.52. Our Second data source was gathered from surveying and interviewing a sample of 105 students from an American Polytechnic State University prior to a class lesson on aggression (IRB #21-915) and required each participant to reflect on their past experiences (in their workplace, school ...) and document 20 instances where someone's comment about their identity bothered/irritated them. They were also asked to supplement these, with an equal number of examples statements that a co-worker has said or could say that did/would not offend them. For transparency, the process ensured the students had no prior knowledge of the size/demographics of the survey population sample, and the responses were kept confidential. In the end, we collected 2170 and 2168 instances of MA and non-MA classes respectively, with the following data statistics: 28 473 total words, 2314 unique words, and, on average, 6.56 words per statement (SD=2.85). Television plays a vital role in society, and it is also responsible for influencing human behavior, communication and interac-

Table 1: Examples of MAs and Non MA texts from our three data Sources

Data source	MA Setting	MA instance	Non MA instance
TV show	Sitcom "All in the family"	There is huh then how come we don't have a black president; I mean some of our black people are just as dumb as Nixon	Black as the ace of spades
Website	Workplace	Oh, i don't know how to pronounce those names.	Intelligence is knowing how to pronounce pikachu
Survey	Social	I'm Surprised you do not have an accent.	I like the customers here

tion with their surroundings (Myrtek et al., 1996; Washington et al., 2021). Television has always been a great educational and entertainment tool that can transcend generations and give people the opportunity to experience life from different eras. However, some TV shows could contribute in exacerbating societal stereotypes, prejudices and biases. As Art often imitates life, it is not uncommon for people to copy and replicate utterances like MAs from these broadcasting media, in real life communication settings. Our third dataset was built by watching, extracting, and documenting MA instances from the following TV shows: -Blackish, Martin, Golden Girls, The Office, All in The Family, Everybody Hates Chris, It's Always Sunny in Philadelphia, and That 70's Show. Despite the scripted nature of their texts and their entertainment aspect, we believe the text collected through this media could help provide context for real-life scenarios that resemble some of the characters or personas played or seen on TV. The same annotation quality and inter-rater reliability check outlined for the web scraped process was also applied here. Moreover, for better transcription accuracy, each of the MA text collected was also compared against the TV Shows' transcripts extracted online from sites such as `https://transcripts.foreverdreaming.org` and `https://www.simplyscripts.com`. The non-MA instances considered through this process were collected from the same TV shows from where the corresponding MA examples originated, and the MA non-MA pairs were to be within 30% length of each other, and on a 70% cosine similarity threshold. The 256 pairs of MA and non-MA classes gathers consisted of 6568 total words, 1686 unique words, and, on average, 13.08 words per statement (SD=9.63). Some examples of MAs and non MA instances from our three data sources can be seen in (Table 1)

3.2 Dataset creation and preprocessing

For this experiment, we join the data collected from the web scraped and survey data source thus resulting in 7,764 examples each categorized as either a MA (1) or a non-MA (0), splitting the data on a 9:1 ratio (i.e. 6987, 777) for training and validation and using 512 scripted data instances for testing. The data preprocessing phase includes lower-casing, punctuation removal, digits encoding with "number", decoding and representing all characters into ASCII format, and removing all HTML tags. Text decontraction is also implemented, and the Spacy library is used to tokenize and lemmatize each instance. For the transformer model, we apply no preprocessing and use the text in its raw state. All random number generators for train-test splits used 999 as the seed value.

3.3 Models description

We chose two algorithms for modeling namely: - the SVM classifier using bag-of-word vectorizations, and the transformer model RoBERTa. Although transformer models are the current SOTA for NLP classification tasks, we included SVMs as a way to better understand the nature of MAs and the extent to which varying amounts of context, through N-gram, could impact the overall models' classification performance. Furthermore, including a method that does not explicitly account for word order or sequence when used with bag-of-word, helps explore whether the contextual awareness offered by transformers is necessary (Clavié & Alphonsus, 2021). When using the SVM model, we use a document-term with a unigram and an N-gram vectorization of texts. Just as it is important to compare transformers to bag-of-word approaches to understand context, these N-gram approaches were so chosen to better understand how much contextual clues within the text mattered. For the unigram representations, the sequence of the text is lost and all that is known is word presence while

for N-gram models, the sequence of the words is preserved at a local level, though the greater context is potentially lost for longer texts. Transformer models use self-attention mechanisms and position embeddings to account for the complete interrelationships between words in a text and because transformer models also include word embeddings in addition to contextual attention mechanisms, we also isolated the role of context by examining an SVM with word embeddings of the sentences. We hypothesized that, if this model performs equally to the N-gram approaches, and worse than the transformer model, then context likely has a large role in inferring MAs in text. Therefore, this approach not only allows us to understand how well subtle statements of offense can be classified but also what model properties are important to preserve when detecting MAs. One alternative to N-gram models is the word embedding approach, which is more robust for tasks that involve terms occurring at low base rates. This approach encodes words in terms of their semantic meaning, in a fixed dimension space (Mikolov et al., 2013). Therefore, the model does not require the exact word used in the training corpus to draw similar inferences about related words it encounters later. If the model sees a sentence like, "This is amazing" but has only seen sentences such as, "This is great" and "This is awesome," the model will still draw similar instances about the unseen sentence because the average of its word embeddings for synonymous words are roughly the same. We use the Glove embeddings (Pennington et al., 2014) trained on 840B tokens from the Common Crawl, which offers 300-dimensional representations of English words, and averaged the embeddings for each word in the text, which was preprocessed as described earlier. Our motivation for utilizing these data sources stems from our interest to find the relationship (if any) between accounts of MAs expressed from real humans, and those extracted from scripted TV shows, where the combined instances of utterances gathered for the student survey and the Tumbler website would serve for building a model for detecting MA in text, while the MAs from scripts would be used to test our model performance on unseen text, and hence evaluate, and validate our hypothesis on whether "Art imitates Life".

4 MODELLING

For the SVM model trained with the TD-IDF document-term matrix, we explored ranges of hyperparameter values using the Grid Searching Cross Validation approach, to determine the most optimal combination for the task. The hyperparameters we varied were the inverse regularization strength, C, and the kernel flexibility parameter, gamma. The parameter C, common to all SVM kernels, trades off misclassification of training examples against how simple the decision surface is and because C is an inverse regularization parameter, a lower value of C makes the decision surface smoother, while a higher C is more able to classify all training examples correctly. Gamma, however, defines how much influence a single training example has and as such, the larger its coefficient, the closer other examples must be to be affected. We also varied the kernel function with the option of either a linear or a Radial Basis Function (RBF) kernel, which has more predictive flexibility to capture non-linear relationships, but at the expense of potentially overfitting. Our best performing model used the RBF kernel, C = 1.7, and gamma =1.3. After modeling on the training sample of human's real-life reports of offenses, we examine the model's CV performance on the held-out sample of data and calculate the precision, recall, and F1 score of the model in classifying these statements. ultimately, to test if the trained model is generalizable, we test it on the left-out unseen test set of data collected from sitcom TV shows computing the same predictions as done in the CV on this set of data. All models include lower-order terms as well (e.g., unigrams and bigrams for the trigram model). We use the same 90/10 data split done with the unigram trained model and performed a 3-fold CV grid search for each model. We observed that stronger regularization was required as the N-gram size increases (increasing the dimensionality of the feature matrix; bigram C = 1.3; trigram C = 1.0; 4-gram C= 1.0; 5-gram C=1.0). All models performed best with an RBF kernel, and this kernel generally required less influence from nearby points as the N-gram size increased (bigram gamma = 1.3; trigram gamma = 1.0; 4-gram gamma = 0.8; 5-gram gamma= 1.3). We train an SVM model that used the average word embeddings of the same text provided to the unigram model, also on a similar 90/10 data split as done with the unigram trained model, using a 3-fold CV grid search for the best parameter, hence resulting in a C = 1.3 on a Linear Kernel. We use the RoBERTa architecture, which is like the BERT model, though with different pretraining. Whereas BERT was trained using a language masking strategy, wherein the system learns to predict intentionally hidden sections of text, RoBERTa modifies key hyperparameters in BERT, including removing BERT's next-sentence pretraining objective, and training with much larger mini-batches and learning rates. This modification allows RoBERTa to improve on the masked language mod-

eling objective compared with BERT and leads to better downstream task performance. Similar to its BERT parent model, RoBERTa is pretrained using 16GB of text from Wikipedia and Google books but also with 144 GB of additional data from the OpenWebText (Liu et al., 2019) and CommonCrawl. These additional data sources contain text from the web, which may be more relevant to capturing MAs. We utilize the pretrained version on HuggingFace Transformers library (Wolf et al., 2019) to train our model. The base RoBERTa model available under the MIT license has 12 layers, whose hidden dimension side is 768, and with 12 self-attention heads comprising 125 million total parameters that can be fine-tuned. We also used the BERT-base-uncased model and a version of it fine-tuned to detect overt Hate speech, called HateBERT (Caselli et al., 2020). These models provide additional, transformer-based comparison to the bag-of-words based methods, and also as a comparison to RoBERTa. The base BERT model was trained with less data, illustrating the role that larger training sizes can play. HateBERT has the same architecture as BERT but is optimized for overt hate speech, showing the potential overlap between MAs and overt hate speech. To train the model, we use the Graphic Processing Unit (GPU) type Tesla p100PCIE16GB 65 after determining that our system supports CUDA. We use the Huggingface transformers library to access the model and run the model for 10 epochs and a training batch size of 8, doing our implementation using the Adam with Warmup optimizer and padding every BERT models to a maximum length of 128. We replicated the trained using learning rates of 1e-04, 1e-5,1e-6, and 1e-7. For all models, 1e-5 served as the best performing learning rate.

5 Evaluation Metrics

In the end of the modeling process, we compare the models' performance on the following metrics: (a) Precision, the percentage of positively predicted outcomes that were actually positive. Because microaggression detection is the focal category, as it has the most capacity to cause interpersonal harm, (b) Recall; the percentage of actual positives that were predicted positively. we emphasize the precision and recall of "MAs". (c) F1 score which is the harmonic mean of precision and recall, representing the balance between the two when a single metric is desired to describe overall performance at classifying the focal category. This analysis provides evidence of how well, MAs can be inferred using a variety of NLP methods. The (d) Confusion matrix which is a technique for summarizing the classification aptitude of algorithms, was also used to visualize the congruence between predicted labels and accurate labels. By comparing performance across models that incorporate different amounts of contexts in the analysis, we can examine how much context matters in determining whether a statement is a microaggression. Additionally, to test if art imitates life, we used the sitcom data as the evaluation set and compared the classification metrics.

6 Results

6.1 Real-world MAs classification performance from Real-life Settings

The modeling results seen in (Table 2), show that the unigram and multigram models performed similarly at detecting microaggressions within real-life statements across all metrics of interest. The model precision ranges from .72 (5-gram model) to .74 (unigram model), the recall from .75 (unigram model) to .76 (multigram models) while the F1 scores ranges from .74 (multigram models) to .75 (unigram model). The results reveal that there appear to be words that distinguish MAs from non-MAs. However, there do not appear to be distinguishing patterns of N-grams, as these models perform equally to the unigram models. Interestingly, the average word embedding model performed similarly to the token-based models with precision = .70, recall = .74, and F1 score = .72. This model also does not capture word sequence but should perform better when the synonyms for distinguishing terms are informative. Because the performance is similar, the terms that are indicative of MAs are likely specific to the text. The contextual model, RoBERTa, performed the best out of all of the approaches across all metrics (Precision = .79, Recall = .87, F1 score = .82). One alternative explanation is that RoBERTa performs so well, not because of its contextual awareness, but because of the amount of pretraining data it has. In a follow-up analysis, we examined the performance of the BERT base uncased model, whose training data is 1/10th the size of RoBERTa's (16GB vs 160GB) and actually less pretraining data (3.3 billion tokens) as the embedding models (840 billion tokens). The performance of BERT was similar to RoBERTA, if not higher (Precision = .87, Recall = .83, F1 score = .85), suggesting that the results cannot simply be explained due to

Table 2: Cross-validation Performance at Classifying Microaggressions from Real-life Settings

Model	Precision	Recall	F1 score
Unigram +SVM	.74	.75	.75
Bigram + SVM	.73	.76	.74
3-gram + SVM	.73	.76	.74
4-gram + SVM	.72	.76	.74
5-gram + SVM	.72	.76	.74
Av word embedding + SVM	.70	.74	.72
RoBERTa	.79	.87	.82
BERT base uncased	.87	.83	.85
HateBERT	.83	.90	.86

pretraining data size. These results are some of the first to show that not only are MAs inferable using NLP but also that this inference is dependent on incorporating contextual clues. Another alternative explanation is that MAs are not any different from hate speech and that existing models would do just as well. We applied the HateBERT model to the real-world MA data, and its performance across most metrics was equally similar to the BERT approaches (Precision = .83, Recall = .90, F1 score = .86). Therefore, knowledge of overt hate speech structure, does not offer major improvements in the detection of MAs, and they appear to be different domains where the domain-specific pretraining did not assist performance.

6.2 Generalizability of MAs classification to scripted TV data

We applied the models trained on the real-life data, as described in the previous section, to the MAs extracted from TV Shows and found that all models performed equally well with precision ranging from .51 to .59, recall ranging from .65 to .67, and F1 scores ranging from .57 to .64. See (Table 3). These values, however, are all lower than the classification metrics found in the real-life data, suggesting that MAs occur in similar, but not identical ways across domains. The equivalence of bag-of-words approaches with more contextually sensitive N-gram and and transformer models suggests words can be microaggressive across contexts, but the contextual usage is likely different. While the RoBERTa transformer model showed superior classification performance when applied to its training context, it performed roughly equally to the other models when applied to a new context with a precision = .51, recall = .77, and F1 score = .62. The HateBERT model, fine-tuned to MA data, did show a slight improvement compared to BERT, but not compared to RoBERTa. This finding like the real-life data analysis highlights how knowledge of overt speech structure does not offer major benefits compared to models lacking that pretraining. Overall, the similarity all of the models, which vary in terms of the contextual awareness, suggests how MAs may generalize across domains. While the terms used on MAs in real-life and in art may be similar. The contextual information learned by the transformer model for real-life domains does not translate to provide additional information for classifying MAs in the media.

Collectively, these results help inform and validate the following observations (1) that microaggressions in real life are contextually dependent as the models' performed strongly with increasing contextual variation (2) that art has similarity to life to some extent, given that the model used were able to consistently detected and classified microaggressions above base rate levels. However, we notice that there may be a disconnect between the contexts of microaggressions found in Art and those from real-life, as the models performed almost equally on both the humans and arts data, with the testing set benefiting more from the information learned and transferred during training but not really from the inherent contexts conveyed.

7 Discussion

7.1 Limitations

The current project is not without limitations. Our model treats offensiveness dichotomously and ignores the severity of the offense. One future possibility is asking people to rank the severity of

Table 3: Cross-validation Performance at Classifying Microaggressions from TV scripts data

Model	Precision	Recall	F1 score
Unigram +SVM	.59	.66	.62
Bigram + SVM	.57	.67	.62
3-gram + SVM	.57	.65	.61
4-gram + SVM	.58	.66	.62
5-gram + SVM	.58	.67	.62
Av word embedding + SVM	.60	.67	.64
RoBERTa	.51	.77	.62
BERT Base Uncased	.51	.65	.57
HateBERT	.53	.74	.62

MAs or to provide a continuous rating of severity on a visual analog scale. This improvement would provide a more nuanced description of the interpersonal harm experienced by others, and perhaps the long-term effects of these statements. Additionally, it is unclear how much the model captures the variability in experienced offense due to differences in the offendee's personality and life experiences or group (gender, ethnicity, socioeconomic status, etc.) differences. Despite having a large sample of statements collected from Microaggressions.com and students, the demographic information of all individuals is unknown. A larger, diverse sample could provide insight into the differences or similarities among types of offensive statements said to specific groups of people. Further, the model does not capture the physical or cultural environment where the dialog occurred. Our results highlight the importance of context, within the text alone, when inferring the presence of a microaggression. A missing contextual factor that is potentially important when experiencing or interpreting offense is the source of the message. The identity of the perpetrator and the similarity of their identity to the target of the statement can influence whether the offense is experienced as the target interprets intent. Other situational variables that text-based MAs fail to include are the information gained from tone, facial expressions, and body language.

7.2 Future Directions

Drawing from the limitations described in the previous section, future research could incorporate a multimodal approach to capture the nuances of spoken language. Rich affective information can be conveyed through spoken and non-verbal language. Statements' meaning and intent can shift with a subtle modulation of tone, word emphasis, speed, or pause frequency. Training a model that includes text, visual, and auditory information could provide more contextual information about the interaction in which an offense takes place. In the hopes to achieve complete natural language understanding and test if art truly imitates life, training a model on a database that reflects the complexities of speech and context in the real world would be profound.

It is important to consider the potential negative societal impacts of the research as well. This research can potentially be used to identify areas that are most sensitive to underrepresented and vulnerable populations. This knowledge could inform those who are unaware of these issues and vulnerabilities. However, they could potentially be exploited to created further division in public discourse and create bad-faith actors. Therefore, not only is it important to study the nature of MAs, but future research may desire to find effective ways to counteract and reduce their experienced harm.

8 Conclusion

From our results, we infer that MAs in real-life are very complex and contextually dependent. The more contextually aware transformer model performed much more strongly at detecting real-life MAs than context free approaches. We found that the MAs expressed in art are similar to those expressed within life to a degree because the MAs could be detected consistently above base rate levels, but that context learned from more sophisticated models in real-life domains did not translate to increased performance in artistic domains.

REFERENCES

Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. Automated detection of racial microaggressions using machine learning. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2477–2484. IEEE, 2020.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. Detecting online harassment in social networks. 2014.

Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15, 2016.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

Benjamin Clavié and Marc Alphonsus. The unreasonable effectiveness of the baseline: Discussing svms in legal text classification. *arXiv preprint arXiv:2109.07234*, 2021.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 512–515, 2017.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

J Desmond-Harris. What exactly is a microaggression? *Vox. Accessed October*, 27, 2015.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pp. 29–30, 2015.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pp. 1–6, 2018.

Sreyan Ghosh and Sonal Kumar. Cisco at semeval-2021 task 5: What's toxic?: Leveraging transformers for multiple toxic span extraction from online comments. *arXiv preprint arXiv:2105.13959*, 2021.

Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, 198:360–366, 2022. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.12.254. URL https://www.sciencedirect.com/science/article/pii/S1877050921024935. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pp. 875–878. IEEE, 2018.

James B Jacobs and Kimberly Potter. *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand, 1998.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N Asokan. A little goes a long way: Improving toxic language classification despite data scarcity. *arXiv preprint arXiv:2009.12344*, 2020.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190, 2020.

Josephine P Law, Paul Youngbin Kim, Jamie H Lee, and Katharine E Bau. Acceptability of racial microaggressions among asian american college students: Internalized model minority myth, individualism, and social conscience as correlates. *Mental Health, Religion & Culture*, 22(9):943–955, 2019.

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

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

Yara Mekawi and Nathan R Todd. Okay to say?: Initial validation of the acceptability of racial microaggressions scale. *Cultural diversity and ethnic minority psychology*, 24(3):346, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pp. 928–940. Springer, 2019.

Michael Myrtek, Christian Scharff, Georg Brügner, and Wolfgang Muller. Physiological, behavioral, and psychological effects associated with television viewing in schoolboys: An exploratory study. *The Journal of Early Adolescence*, 16(3):301–323, 1996.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Soujanya Poria, Ong Yew Soon, Bing Liu, and Lidong Bing. Affect recognition for multimodal natural language processing. *Cognitive Computation*, 13(2):229–230, 2021.

Nazanin Sabri, Valerio Basile, Tommaso Caselli, et al. Leveraging bias in pre-trained word embeddings for unsupervised microaggression detection. In *CLiC-it*, 2021.

Mujahed A Saif, Alexander N Medvedev, Maxim A Medvedev, and Todorka Atanasova. Classification of online toxic comments using the logistic regression and neural networks models. In *AIP conference proceedings*, pp. 060011. AIP Publishing LLC, 2018.

Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*, 2022.

Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285, 2012.

Brian H Spitzberg and Jean Mark Gawron. Toward online linguistic surveillance of threatening messages. *Journal of Digital Forensics, Security and Law*, 11(3):7, 2016.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271, 2007.

Maite Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2: 325–347, 2016.

Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

Samuel Walker. *Hate speech: The history of an American controversy*. U of Nebraska Press, 1994.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 415–425, 2014.

William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pp. 19–26, 2012.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

Gloria J Washington, Gishawn Mance, Saurav K Aryal, and Mikel Kengni. Abl-micro: Opportunities for affective ai built using a multimodal microaggression dataset. In *AffCon@ AAAI*, pp. 23–29, 2021.

Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A corpus study of evaluative and speculative language. In *Proceedings of the second SIGdial workshop on discourse and dialogue*, 2001.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399, 2017.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Sara Zaheri, Jeff Leath, and David Stroud. Toxic comment classification. *SMU Data Science Review*, 3(1):13, 2020.