Automated Detection of Racial Microaggressions using Machine Learning

Omar Ali
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
omar.ali1@port.ac.uk

Nancy Scheidt
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
nancy.scheidt@port.ac.uk

Alexander Gegov
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
alexander.gegov@port.ac.uk

Ella Haig
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
ella.haig@port.ac.uk

Mo Adda
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
mo.adda@port.ac.uk

Benjamin Aziz
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
benjamin.aziz@port.ac.uk

***Abstract*—Microaggressions describe subtle often offensive comments or actions made by one individual to another. Typically, such comments or actions are made subconsciously with the offender potentially unaware of the impacts on the recipient. Currently, machine learning methods for racial microaggression detection are sparse with no, one, comparable approach to the one we propose further on. Automated detection in this work describes the method of finding m icroaggressions t hrough the use of machine-learning algorithms. Efforts have been made for the detection of hate speech and harassment; providing us with a rather humble place to begin, as we explore further, such methods are proven ineffective. We propose a step forward in solving this problem with the demonstration of an automated racial microaggression detection method. Whilst racial hate-speech detection method provides us with an idea as to where we can start, we find f urther o n, t hat m icroaggressions a nd hate-speech use very different features to portray their sentiment. This work aims to provide a technical review which explores the understanding of the automated racial microaggression detection, outlining the definitions o f m icroaggressions c urrently described in the literature, with the presentation and assessment (through precision, callback and F-measure) of a promising approach in regards to racial microaggression detection. We also intend to analyse a case study in which we detect the presence of racial microaggressions within new reports with the intention of laying a foundation for the potential security-related applications of this work. Our further works begin to discuss this notion in a higher level of detail.**

***Index Terms*—Online Content Moderation, Human Modeling, Behavior, Microaggressions, Emotion, Sentiment Analysis, Hate Speech**

I. Introduction

Microaggressions (MAs) can be defined i n s everal ways. Manifestations of derogatory stereotypes which are 'put downs' by an offender typically subtle, stunning, often automatic and non-verbal exchanges. Brief and commonplace daily verbal, behavioural, and environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults to the target person or group as described in [1], [2] and [3] are also common definitions for MAs. MAs are also interpersonally communicated consisting of "othering" messages related to a person's perceived marginalized status typically unconscious [4]. In lieu of the distinct lack of work around the automated detection of racial MAs (RaMAs), the main contributions this work include a comprehensive review of the literature around this growing topic both from a psychological and technical standpoint – helping us to understand which variables play a role in the automated detection of MAs. We also present a labelled dataset used for the classification tasks we will undertake – derived from a lexicon-approach in which we have hand-crafted a lexicon designed to retrieve subtle racism with valid terms at the time of creation. We then begin to outline our use of both this lexicon and dataset to initially compare 7 algorithms, typically used in sentiment analysis and NLP, to assess the initial task which involves the binary classification of RaMAs amongst other MAs (typically of a sexist orientation). We move forward to our proposed second task involving the use of selected attributes to aid in the classification of RaMAs amongst other MAs; monitoring the impacts this method has in the accuracies of RaMA detection. Finally the presentation of an existing research-approach to this novel dataset for a novel case-study of racial microaggression detection in online news reports to assess potential public-security implications of the transmission of racial microaggressions in said news reports.

The subsequent sections aim to express our motivations more clearly, followed by an overview of the closest previous work around this topic. We then look into the proposed method for the classification of RaMAs amongst general MAs, concluding with our results, applications to our chosen case-study of news-reports and discussion both of the results and work around the literature.

II. Motivation

Automated MA detection offers a challenging new problem in the study of NLP and sentiment analysis. Understanding the nuances of how one constructs a MA together with the classification of MAs in text can help to improve our overall understanding of sentiment in text. The authors of [5] expresses difficulty in the annotation of all types of hate speech. The authors add that hate speech is often expressed without any such hateful terms i.e. sexual or racial slurs. MAs fall directly under this such category of hate without the explicit expressions. From a classification point of view, even though we can classify both racial hate speech and racial MAs as being under the same umbrella as hate-speech, we have to consider both as being two very different classification tasks as they both rely on largely different features to portray a similar sentiment. This means we are unable to employ current hate-speech detection methods in this case. Such difficulty is also expressed by the authors of [6]. The authors suggest future research which focuses on the examination of the role that microaggressions play in setting the stage for hate speech. We believe that microaggressions do set a stage for hate-speech – with one of our main applications of this work (discussed in Further Work) being the detection of more subtle forms of hate. We have chosen the case study of news reports to assess the practicality of our work in regards to its public-security impacts it may have. Our motivation for this case study also stems from the analysis made by the authors of [2] who have also have expressed a need for analysis, in such an area, with their work surrounding the microaggressions in sexist and racist news reports of female athletes following the 2012 and 2016 Olympic games. This gives us a clearer understanding to our the analysis we are carrying out – allowing us to better interpret our results together with grounding our work in current research. The authors of [7] also begin to explore this side of the analysis in their work surrounding the effects hate-speech and fake-news have on the Nigerian election. As the automation of MA detection has not been explored in any comparable way. We look to the case made for such research by the authors of [8], who express the need for the detection of MAs from several categories. The closest comparable work revolves around the automated detection of racism and sexism in online text. An analysis of microaggressions in news media made against women during the 2012 Olympic games is made by the authors of [2]. Another analysis made by the authors of [9] explores the link between racial MAs and racial hate crimes. From a security standpoint, the motivations for the chosen news-related case-study is attributed to potential security-related further-works regarding the moderation of online content in which we could monitor a live-feed of speech, transposed from speech-to-text, to monitor news-reports, tweets, and other broadcasting mediums similar to the works conducted by the authors of [10], whereby, they look into the notion of coded-language that is commonly used to convey subtle sentiments and microaggressions within said news-reports.

III. The Measure of Microaggressions

The over-shadowing problem with the work, as well as any work that surrounds sentiment is the measurement of subjectivity and the impacts that any sentiment-carrying text has. In regards to MAs, the authors of [8] define MAs (See table I) in terms of 4 categories; Attribute, Institutionalized, Forced Teaming and Othering. Each category as several subcategories that further narrow the focus of the MA. These categories help us define particular cases of MAs, however, do not allow us to quantify the MA such that we can measure its causality in terms of racism. On the other hand, these definitions help us to begin searching for features that may be significant in the process of automated detection i.e. lexicon approaches from words derived from Stereotype text; with Stereotype being a subcategory of the Attribute category. A different approach is taken in [1] in their work around MAs. The authors take a different approach to the aforementioned, with the proposition of three categories to MAs:

- microassault: explicit forms of discriminatory verbal or nonverbal attacks intended to victimize a racial or ethnic minority
- microinsult: an unintentional message that carries derogatory connotations to demean a cultural practice or the racial/ethnic identity of people of colour
- microinvalidation: ontological and epistemological erasure through which experiential realities (e.g. history of exclusion) of racial and ethnical minorities are negated and invalidated [11].

A categorical method for quantifying MAs provides us with a jumping-off point for understanding what goes into an MA, however, a more quantitative approach is necessary if we are to measure the causality of an MA in regards to racial hate-speech. The Acceptability of Racial Microaggressions Scale (ARMS) [12] provides us with a deeper understanding of what is considered a *tolerable* MA. The author's motivations for this scale stem from the lack of scales that exist to assess attitudes about the acceptability of MAs. The framework for this scale stems from the authors [1] broader categories and taxonomies of MAs. These categories are implemented to differentiate between different types of racial MAs. The authors of [13] appropriate this scale in their analysis of the acceptability of MAs amongst Asian Americans. Such a scale can provide us with a larger amount of details in regards to measuring the causality of racial MAs – opening the door to large-scale, automated, MA detection.

In regards to work surrounding the case-study, the research conducted in [14] begin to analyse hate-speech amongst potentially fake news expressing their concern for the spread of hate amongst these articles. The authors of [15] aim to analyse, more specifically, the hate spread amongst the comments from these articles.

IV. Automated Microaggression Detection

The subsequent sections describe the methodologies and justifications for our choices. We also describe the data we

Class	Subclasses	Description and Example
Attribute	Attribution of stereotype	Link some attributes to an individual based on their identity. *"Girls just aren't good at math."*
	Alien in own load	Marginalized individuals are foreign. *"But where are you from, originally?"*
	Abnormality	Marginalized individuals are abnormal. *"Why do we need the word cisgender? That's just normal people."*
Institutionalized	Objectification	Diminish the humanity of marginalized individuals. *"If you don't want to get hit on, wear a longer skirt."*
	Criminal Status	Link a persons identity to criminality, danger, or illness. *"You look like a terrorist with that beard."*
	Second-Class Citizen	Marginalized individuals belong to low-status positions in society. *"Oh, you work at an office? I bet you're a secretary."*
Forced Teaming	Myth of Meritocracy	Differences in treatment are due to ones merit. *"They just cast actors who are best for it. Why does it matter if they're all white?"*
	Denial of Lived Experience	Minimize the experiences of a marginalized individuals. *"It was just a joke! You're too sensitive."*
	Ownership	Anyone can have some claim to a marginalized groups experiences. *"Why is it offensive for a white person to wear a bindi? It's just jewelry."*
Othering	Monolith	All members of a marginalized group are identical. *"My gay friend doesn't have a problem with this show. I don't get why you're mad."*
	Erasure	Anyone can claim that an individual does not belong to that group. *"Your mom is white, so it's not like you're really black, though."*

TABLE I
CLASSES OF MICROAGGRESSIONS AS DEFINED IN [8]

used for the study together with a more in-depth analysis of the lexicon and labelled dataset we present. All algorithm implementations used for each classification task are from the open-source machine-learning tool, Weka [16]. We have chosen Weka due to its wide range of implementations available for various machine learning algorithms as well as to simplify the process of exploration – allowing us to focus mainly on the direct analysis of MA detection. Together with automated RaMA detection, we also aim to apply our analysis to a security case study involving the spread of MAs in news reports.

A. Methodology

We firstly present a lexicon designed to categorise MAs into racial MAs (RAMAS) and non-racial MAs (NONRAMAS). The lexicon is used to filter MAs in the compiled dataset on a word-level i.e. if an MA contains that particular phrase; it is labelled as racist. We have chosen this broader, simpler, approach to labelling as we are to make some clear assumptions of what the original corpus represents. As we already know that the corpus only contains MAs we can infer that any of the phrases in the corpus that contain the words present in our lexicon are highly likely to be RaMAs. Such a simplistic approach allows us to quickly and confident label all of the instances in $O(n^2)$ time. This labelling task leaves us with 436 instances of RAMAS and 967 of NONRAMAS. Once our dataset has been labelled, we extract the word-features using trigrams resulting in a total of 2713 word-features for use in the subsequently classification steps. We have chosen Weka's implementations of Logistical Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), J48, MultiClassClassification (MCC) and IBk (k-nearest neighbour) to perform binary classification on the dataset (RAMA and NONRAMA as each class) as they are considered the most conventional ML algorithms to use for text classification, specifically for the use of sentiment analysis and NLP [17], [18] and [19]. We have also made this choice due to the lack of similar work we can compare to. Instead of diving straight into the state-of-the-art ML methods i.e. LSTM networks and deep learning methodologies, we decide to instead cover, the well known, fundamental ML algorithms first to ensure that these methods are accounted for in this field. This also provides a stronger basis for comparison to later works similar to this, which may use more the involved ML methods for their analysis. We proceed with two distinct classification tasks, firstly, using all of the features retrieved from the dataset using trigrams, and secondly extracting the 20 most salient features from the text to feed into each classification algorithm. The features are retrieved using an entropic approach (information gain attribute selection). Finally, for validation, we employ 10-fold cross-validation for each classification task. Below outlines our justification for the aforementioned binary classification, followed by the datasets and lexicons used for the subsequent classification tasks. We have also chosen to not compare our lexicon with that of a hate-speech one. This is due to the lack of labelled datasets available for racial MAs. We would be

unable to fairly compare each lexicon with one another as we do not currently have an accurate ground truth.

B. Justification of Binary Classification

We have only chosen a binary classification model to just detect whether a given comment is a racial MA or not. We have chosen this path to further analyse the specific features present in racial MAs; keeping our scope focused on the use of these specific features for finer-grain analysis of hate-speech in further works.

C. Datasets

1) Dataset Collection: Following the authors of [8] We compiled a corpus from a microaggression oriented Tumblr page[1]. This resource was chosen as it was the only way we could objectively retrieve MAs. The site allows users to post any type (in person or online) of MA they, or someone they might know, have encountered. We focused on only taking the direct quotes from the website rather than focusing on the description. This was the decision was made as we believed the extra comments would provide too much noise as they would include context/background to their encounter with an MA which may have nothing to do with the MA or its features. The dataset is subsequently preprocessed with the removal of stopwords e.g. *the, and, I.*

2) Lexicon: A hand-crafted lexicon is assembled – motivated by the lack of one for use in racial MAs. The lexicon consists of lightly racist words and phrases viewed to potentially be a micro-aggressive. Existing lexicons i.e. a list of racial slurs from the Racial Slurs Database[2] failed to retrieve key-sentences, when we observed the filtering process, that may carry the right type of subtle hate we were looking for. Currently, hate-speech lexicons focus on the elicitation of harsher sentiment. Our lexicon consists of $\approx$ 150 subtly racist terms that we use to filter the dataset into RAMAS and NONRAMAS.

V. RESULTS

As explained earlier, we have chosen to run a wide range of algorithms on the dataset to maximise the coverage of analysis. As there is currently no comparable results, we have chosen to instead compare the performance across the multiple algorithms. The subsequent sections outline the two main classification tasks we performed on the dataset.

A. Preliminary Classification

Explained prior, this first task aims to explore the accuracy of classification using trigram features with 10 fold cross-validation. From our results, we can see a somewhat promising classification across all algorithms except for SVM. SVM has shown, in this classification task, that it was unable to correctly determine the precision and therefore the F1 measure for classification of RaMAs. We have attributed this shortfall to our method of feature extraction – trigrams.

[1]www.microaggressions.com
[2]http://www.rsdb.org/full

TABLE II
RESULTS FROM CLASSIFICATION RACIAL MAS AMONGST OTHER MAS USING THE ORIGINAL 2713 ATTRIBUTES FROM THE TRIGRAM APPROACH

	Class	Prec.	Re.	F1
Log. Reg.	RAMA	0.689	0.652	0.670
	NONRAMA	**0.837**	**0.859**	**0.848**
	We. Avg	0.789	0.792	0.768
SVM	RAMA	-	0.000	-
	NonRaMA	0.675	1.000	0.806
	Weighted Average	-	0.675	-
NB	RAMA	0.626	0.582	0.603
	NONRAMA	**0.806**	**0.833**	**0.819**
	Weighted Average	0.747	0.751	0.749
RF	RAMA	**0.962**	0.491	0.651
	NONRAMA	0.802	**0.991**	**0.887**
	Weighted Average	0.854	0.829	0.810
J48	RAMA	0.708	0.460	0.558
	NONRAMA	**0.779**	**0.909**	**0.839**
	Weighted Average	0.558	0.839	0.748
MCC	RAMA	0.709	0.689	0.699
	NONRAMA	**0.853**	**0.865**	**0.859**
	Weighted Average	0.806	0.808	0.807
IBk	RAMA	0.559	0.268	0.362
	NONRAMA	**0.719**	**0.899**	**0.799**
	Weighted Average	0.667	0.694	0.658

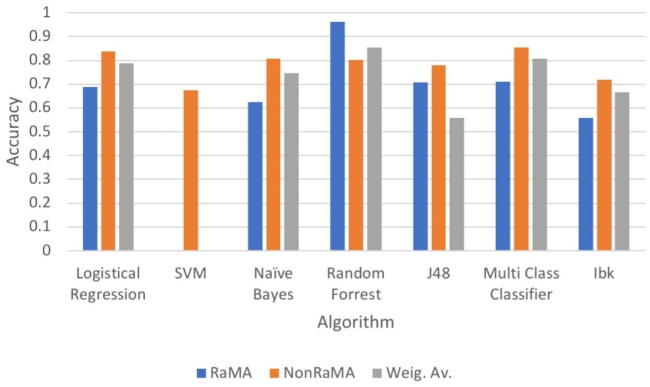

Fig. 1. Precision Accuracy

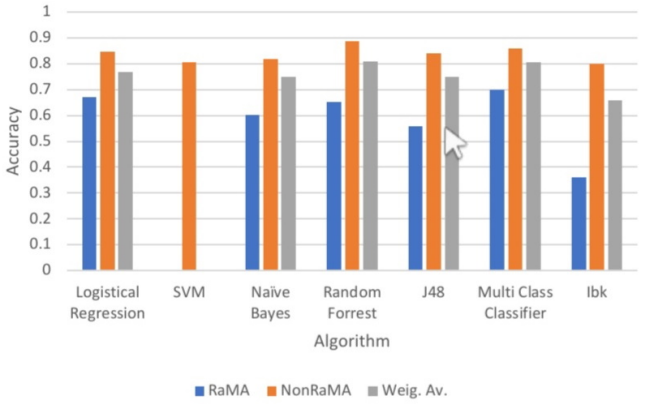

Fig. 2. F1 Scores of Each Algorithm

B. Features Selected Classification

The top 20 attributes amongst the dataset are listed below. These are retrieved using the same entropic approach. As we hypothesised, the key features selected amongst the dataset do not contain any harsh words or sentiments as the hate-speech lexicons used; suggesting that we are unable to use the same approach.

TABLE III
TOP 20 ATTRIBUTES SELECTED FROM THE DATASET USING INFOROMATION GAIN

Attribute
black
white
Asian
Chinese
a black
English
black people
racist
Black
Mexican
Indian
ghetto
a white
skin
dark
American
Native
Indians
lesbian
minority

TABLE IV
RESULTS FROM CLASSIFICATION RACIAL MAS WITH ATTRIBUTE SELECTION

	Class	Precision	Recall	F1
Log. Reg.	RAMA	**0.986**	0.586	0.735
	NONRAMA	0.833	**0.996**	**0.907**
	We. Avg	0.883	0.863	0.851
SVM	RaMA	**0.992**	0.539	0.698
	NONRAMA	0.818	**0.998**	**0.899**
	We. Avg	0.875	0.849	0.834
NB	RAMA	**0.991**	0.455	0.624
	NONRAMA	0.792	**0.998**	**0.883**
	We. Avg	0.857	0.822	0.799
RF	RAMA	**0.986**	0.586	0.735
	NONRAMA	0.833	**0.996**	**0.907**
	We. Avg	0.883	0.863	0.851
J48	RAMA	**1.0**	0.292	0.452
	NONRAMA	0.747	**1.0**	**0.855**
	Weighted Average	0.829	0.771	0.724
MCC	RAMA	**0.985**	0.581	0.731
	NONRAMA	0.832	**0.996**	**0.907**
	Weighted Average	0.882	0.862	0.850
IBk	RAMA	**0.989**	0.587	0.737
	NONRAMA	0.835	**0.997**	**0.909**
	Weighted Average	0.885	0.864	0.853

We can see a considerable improvement to the classification in terms of the attributes used in the second classification task. As stated earlier, we used information gain to select these attributes. We had found that 20 was an ideal number for attribute selection as we had observed no considerable increase or decrease when using 10 and 30 attributes for classification. As explain before by the authors of [6], MAs are particularly hard to identify, with the offenders, themselves, unaware of the offensive nature of what they are saying. This level of subtlety has shown to be challenging for all classification algorithms used.

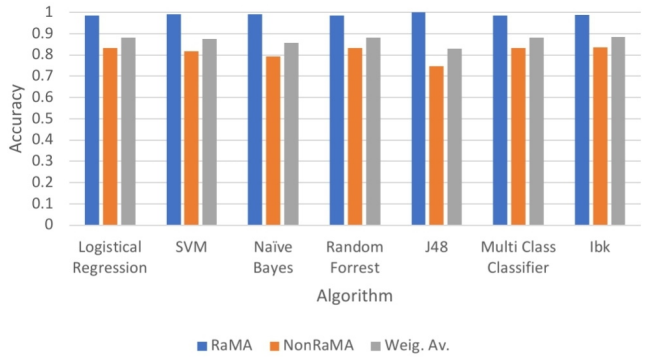

Fig. 3. Precision Accuracy with Attribute Selection

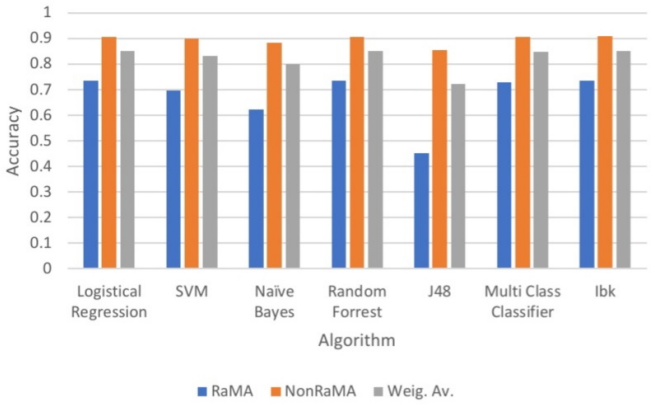

Fig. 4. F1 Scores of Each Algorithms with Attribute Selection

VI. NEWS REPORT ANALYSIS

Following our preliminary analysis of the presented lexicon and dataset, we intend to apply this understanding to news report data retrieved from online BBC reports [20]. The set is pre-split into 5 article types: business, entertainment, politics, sport and technology (tech). In our prior analysis, we did not have much to work off, however, now we have a set of attributes from our labelled MA dataset (see table III) which we can use as an input for our labelling process of the next, BBC dataset. We label our data, in this instance, in a slightly different way. We take each report type as a separate test, with the positive and negative sets being derived in the same labelling manner as before. The labels now follow the convention of the report type followed by either RaMA or NonRaMA. For instance BRAMA and BNONRAMA will constitute the RaMAs and NonRaMAs for the business set. We choose to focus on the algorithms which produced the highest results, in this case, IBk and Random Forest, in our

preliminary test together with 10-fold cross-validation. We performed more tests to provide as much detail to our analysis as possible. The first test involved a multi-class classification between each report type. This involved the comparison of each report-type and their MAs against one another. The second test we performed was a direct binary classification, alike to the preliminary tests we performed. For each report-type, we analyzed the classification of MAs labelled with the 20 attributes gathered previously. Below outlines our results for these classification tasks. With each report type followed by the precision, recall and F1 score metrics.

TABLE V
RESULTS OF NEWS ARTICLE ANALYSIS USING THE TOP 20 ATTRIBUTES FROM THE LABELLED MA DATASET

	Class	Precision	Recall	F1
IBk	BRAMA	0.403	0.240	0.301
	BNONRAMA	0.561	0.519	0.539
	ERAMA	0.718	0.239	0.359
	ENONRAMA	0.576	0.376	0.455
	PRAMA	0.590	0.221	0.322
	PNONRAMA	0.771	0.534	0.631
	SRAMA	0.556	0.349	0.429
	SNONRAMA	0.343	0.918	0.499
	TRAMA	0.923	0.348	0.505
	TNONRAMA	0.931	0.543	0.686
	Weighted Average	0.621	0.516	0.516
Random Forest	BRAMA	1.000	0.033	0.064
	BNONRAMA	0.678	0.956	0.793
	ERAMA	1.000	0.111	0.200
	ENONRAMA	0.698	0.844	0.764
	PRAMA	1.000	0.115	0.207
	PNONRAMA	0.722	0.916	0.807
	SRAMA	1.000	0.147	0.257
	SNONRAMA	0.720	0.995	0.835
	TRAMA	1.000	0.348	0.516
	TNONRAMA	0.848	0.916	0.881
	Weighted Average	0.798	0.736	0.672

The results that are shown in table V, overall, are fairly poor for the first chosen machine learning algorithm with typically a low recall and overall accuracy. Random Forest, however, has shown to perform considerably better, overall, for this particular task with consistently higher precision, however, we do see a drop in the recall, when classing the RaMAs for each report type, which could be caused by the difficulties in detecting these types of MAs using these particular attributes. Results in table VI show larger improvements with binary classification than multi-class with consistently higher results when compared to their multi-class companion. Across both tests, the drop in accuracy could be caused by the formal nature of news reports with the language in such reports being as inoffensive as possible. This implies that our model is not suited for such an analysis as it relies on word-features to extract the sentiment we are looking for. Unlike our previous test, whereby we already could confirm that the text we were analysing what at least an MA, let alone a RaMA, the dataset used in this case provided us with a large level of blindness when it came to labelling using our aforementioned method. This could have lead to several miss-labelled text – leading to lower results. Instance-size differences between the original

TABLE VI
BINARY CLASSIFICATION RESULTS OF NEWS ARTICLE ANALYSIS USING THE TOP 20 ATTRIBUTES FROM THE LABELLED MA DATASET

	Class	Precision	Recall	F1
IBk	BRAMA	0.474	0.223	0.303
	BNONRAMA	0.792	0.923	0.853
	Weighted Average	0.717	0.757	0.722
	ERAMA	0.481	0.316	0.381
	ENONRAMA	0.736	0.848	0.788
	Weighted Average	0.657	0.684	0.663
	PRAMA	0.675	0.260	0.375
	PNONRAMA	0.794	0.958	0.868
	Weighted Average	0.782	0.744	0.319
	SRAMA	0.565	0.372	0.449
	SNONRAMA	0.808	0.902	0.852
	Weighted Average	0.746	0.767	0.749
	TRAMA	0.786	0.319	0.454
	TNONRAMA	0.871	0.981	0.923
	Weighted Average	0.856	0.864	0.840
Random Forest	BRAMA	1.000	0.116	0.207
	BNONRAMA	0.784	1.000	0.879
	Weighted Average	0.835	0.790	0.720
	ERAMA	1.000	0.308	0.471
	ENONRAMA	0.765	1.000	0.867
	Weighted Average	0.837	0.787	0.745
	PRAMA	1.000	0.154	0.267
	PNONRAMA	0.778	1.000	0.875
	Weighted Average	0.834	0.787	0.722
	SRAMA	1.000	0.248	0.398
	SNONRAMA	0.795	1.000	0.886
	Weighted Average	0.848	0.808	0.761
	TRAMA	1.000	0.290	0.449
	TNONRAMA	0.868	1.000	0.929
	Weighted Average	0.891	0.875	0.845

classification task (detection of RaMAs amongst MAs) and this task may have also played a role in the inaccuracy of the results. The MA dataset typically consisted of one-sentence MAs with potentially fewer than 100 words for each instance. In contrast to the BBC news dataset which consisted of full articles ≈10-20 sentences long. Our current labelling technique, in this instance, would have produced a large amount of noise given that our features were word-based. Whilst we do see an improvement in the Random Forest classification, we can observe the decrease in recall when attempting to classify the RaMAs for each report type. We can potentially accredit this to similar issues present in the IBk classification i.e. miss-labelled text.

VII. DISCUSSION

Overall, we have shown that it is certainly possible to extract racial MAs from text using our model, and labelling scheme. Whilst, in an ideal sense, i.e. when we are certain that the text we are analysis is, in fact, an MA, we can quite simply classify it – allowing us to further retrieve the relevant attributes for further classification. This method proved fruitful however when we applied it to our case study we had found that a more nuanced approach may be needed for more-general purpose racial MA detection. A key set back when carrying our this analysis was the lack of data surrounding this topic. This left us with only a few choices for dataset labelling; leading us to create our own lexicon to achieve some level of progress which can be further built upon. A

larger set of data, along with a better classification scheme, would also be useful in mitigating the slight class imbalance we encountered when analysing both racial MAs amongst regular MAs and our news-report analysis. More complex features would prove useful when extracting MAs as we can provide a finer-inspection as to what constitutes an MA beyond words. An example of other features that could be useful, in this case, are context-based features and structure-based features. Context-based features allow us to consider features in regards to surrounding features. For example, the word "black" could most definitely be considered an offensive word; in our case triggering our labelling system to tag that string of text as a racial MA, however, if that world is in the context of something else i.e. *"What colour is my favourite?"* then we should not label that sentence as being offensive. This concept is explored by [21] in their work surrounding context-based sentiment analysis. Structure-based features allow us to look beyond words as being the main carrier of sentiment, and instead, we can peer toward how sentences are structured to aid our labelling method. Authors of [22], [23] and [24] demonstrate a viable solution to the use of structure in text for polarity analysis using Rhetorical structure theory (RST) [25]. RST can allow us to consider different structures of text e.g. contrasting opinions or explanatory points, allowing us to weight these texts more highly in regards to the presence or absence of RaMAs or MAs in general. To this point, a multi-class labelling system may also help us to understand different levels of RaMAs helping us to further label our data more accurately. This can build upon our current, binary, approach of labelling, however, we can take into account the intensity of the RaMA rather than just its presence. A similar approach is applied by the authors of [24] with their approach to RST-based sentiment analysis. The authors consider five classes (very-negative, negative, neutral, positive and very positive) of sentiment rather than two (negative and positive).

VIII. Conclusion

In this work, we have presented a novel lexicon and dataset which have used in a selection of classification tasks to assess their worth in the automated detection of racial microaggressions. We have shown that it may not be viable to use currently available hate-speech data for this type of detection.

A. Successes

The results from our two classification tasks have shown promising results in terms of MA detection. The attributes retrieved from the labelled dataset solidify this fact with phrases/attributes one might suspect as being subtly racist i.e. *dark* and *minority* being selected in the top 20. We also see a distinct lack of overly harsh or profane words/blasphemes in the list. Accuracy improvements when using these attributes for classification have also shown promising results with consistent trends across all algorithms tested i.e. much higher accuracies in classifying NonRaMAs. This could be due to the level subtly present in RaMAs causing each algorithm to classify it as NonRaMAs. The tasks carried out also provides some support in the argument of the validity of our lexicon and labelled dataset. We believe that this lexicon can open up analysis to a finer-level of detection in hate speech with potentially hateful sentences or phrases being wrongly classified as inoffensive. Whilst our work provided a better understanding of MAs and the automated detection of them, we were not without shortfalls. Results regarding the case study chosen to apply our model to, overall, were not as positive as we would have preferred, however, the analysis of news-report data has provided us with a complex application that grounds our results in reality. We have shown, above anything else, that MAs, especially racial MAs, are particularly hard to find. Furthermore, we believe that our results could see drastic improvements should more data become available for this particular case. As we have discovered, through basic observation of our results in the preliminary tests, current material e.g. datasets and lexicons provided by [5] and [26] and Hatebase[3] respectively, will not suffice for such a form of hate-speech as they consist of harsher sentiment which contrasts with the nature of MAs being subtle.

B. Shortfalls

The main issues surrounding our classification task was the lack of material i.e. accurately labelled datasets or lexicons specifically for this task. This required us to produce our own to illicit some valid results in this area. Moreover, we were unable to validate our results with similar works due to the currently, juvenile, state of the field. This indicates that there is a lot to learn from this study in regards to a methodological approach. SVM classification seemed to be inconclusive with trigrams. bigrams and unigrams. This could be due to the inadequacy of n-grams for feature elicitation – preventing the SVM from correctly plotting the hyperplane for accurate classification. We believe this is the case as we begin to see comparable results to the other algorithms when we employ the top 20 attributes in the second classification task. Although the attributes selected for classification helped improve classification particularly with the RaMAs we are still seeing much lower classification scores amongst precision, recall and accuracy. This could be due to the class imbalance present in the data set with $\approx \frac{1}{3}$ of the set being RaMAs and $\approx \frac{2}{3}$ of the set being NonRaMAs. No applicable datasets or lexicons were available for the initial labelling phase; resulting in the manual creation of such lexicon for this step. Such a bottleneck could be a cause for class imbalance as our labelling method may well have missed key MAs. Interestingly, J48 Precision shows worse results than the average, whereas SVM Recall can show a positive average. The seven algorithms show quite similar results overall with a few distinct discrepancies, however, considering and our aim to review which algorithm can detect RaMA the most efficient results would suggest RF. Without as well as with the implementation of the top 20 feature the accuracy of detecting RaMA is relatively high and does not show a big difference. MCC and LR are also

[3]https://hatebase.org/

close runner ups in terms of accurate detection. However, SVM and Ibk are not recommended for the purposes of this paper's research and dataset analysis.

C. Further Works

As this is just a start in a wide range of applications automated MA detection may have; we have outlined further work that may better utilise or apply our presented datasets and lexicons. These applications may, however, rely on material we perhaps were unable to utilise and therefore explore in great depth. As touched on earlier, we see this analysis having an impact on the current hate-speech detection field with nuanced hate-speech, typically undetected by current methods, being successfully classified as hate-speech. The implication this work has in the public security field is also potentially significant. With the large use of online platforms to communicate with the world, we are always at risk of hate-speech typically directed at minorities or the disabled. We believe that such an understanding can help to mitigate the impact these forms of hate-speech (MAs) can have with our work improving the filtering systems on social media platforms. In regards to the further work needed on the technical aspects of the work, we believe that a closer look into the elicitation and nature of features for classification is necessary to see an improvement to the accuracy in regards to general-purpose use of this analysis. Rhetorical structure theory provides a promising approach to the structure-based analysis of MAs together with context-based methods which both may provide the groundworks for a significant step forward in automated RaMA detection. Finally, together with the impacts on nuanced hate speech, we can see immediate applications in the monitoring and moderation of online content, through the analysis of a live-feed, emanating from news reports and Tweets with the applications aiming to mitigate such hate speech from these sources on an automated-basis – analysing the feed and detecting the hate speech as it soon as it is broadcast.

References

[1] 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.

[2] K Allen and CM Frisby. A content analysis of micro aggressions in news stories about female athletes participating in the 2012 and 2016 summer olympics. *Journal of Mass Communication & Journalism*, 7(3):1–9, 2017.

[3] Chester M Pierce, Jean V Carew, Diane Pierce-Gonzalez, and Deborah Wills. An experiment in racism: Tv commercials. *Education and Urban Society*, 10(1):61–87, 1977.

[4] Derald Wing Sue. *Microaggressions and marginality: Manifestation, dynamics, and impact*. John Wiley & Sons, 2010.

[5] 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*, pages 88–93, 2016.

[7] Umaru A Pate and Adamkolo Mohammed Ibrahim. Fake news, hate speech and nigeria's struggle for democratic consolidation: A conceptual review. In *Handbook of research on politics in the computer age*, pages 89–112. IGI Global, 2020.

[6] Emma McClure. 6 escalating linguistic violence. *Microaggressions and Philosophy*, 2020.

[8] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, 2019.

[9] Brendesha M Tynes, Fantasy T Lozada, Naila A Smith, and Ashley M Stewart. From racial microaggressions to hate crimes: A model of online racism based on the lived experiences of adolescents of color. *Microaggression theory: Influence and implications*, pages 194–212, 2018.

[10] Matthew W Hughey and Jessie Daniels. Racist comments at online news sites: a methodological dilemma for discourse analysis. *Media, Culture & Society*, 35(3):332–347, 2013.

[11] Tara Yosso, William Smith, Miguel Ceja, and Daniel Solórzano. Critical race theory, racial microaggressions, and campus racial climate for latina/o undergraduates. *Harvard Educational Review*, 79(4):659–691, 2009.

[12] 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.

[13] 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.

[14] Tarlach McGonagle. "fake news" false fears or real concerns? *Netherlands Quarterly of Human Rights*, 35(4):203–209, 2017.

[15] Karmen Erjavec and Melita Poler Kovačič. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920, 2012.

[16] Frank Eibe, Mark A Hall, and Ian H Witten. The weka workbench. online appendix for data mining: practical machine learning tools and techniques. In *Morgan Kaufmann*. 2016.

[17] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[18] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

[19] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, and A Gupte. Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies*, 5(5):6261–6264, 2014.

[20] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384, 2006.

[21] Gilad Katz, Nir Ofek, and Bracha Shapira. Consent: Context-based sentiment analysis. *Knowledge-Based Systems*, 84:162–178, 2015.

[22] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*, 2015.

[23] Alexander Hogenboom, Flavius Frasincar, Franciska De Jong, and Uzay Kaymak. Using rhetorical structure in sentiment analysis. *Communications of the ACM*, 58(7):69–77, 2015.

[24] Xianghua Fu, Wangwang Liu, Yingying Xu, Chong Yu, and Ting Wang. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In *Asian Conference on Machine Learning*, pages 17–32, 2016.

[25] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[26] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.