EAI Endorsed Transactions
on Creative Technologies
Research Article EAI.EU

Capturing Racial & Gender Inequities on Social Media Platforms using Machine Learning

Sonika Malik[1,*], Harshita Chopra[1] and Aniket Vashishtha[1]

[1]Department of Information Technology, Maharaja Surajmal Institute of Technology, Delhi, India

Abstract

Online social media platforms provide a continuously evolving database due to the highly increasing popularity and rapid expansion of its user base. Users share their life experiences towards various inequity incidents faced at the workplace on the basis of their race or gender on these platforms while maintaining their anonymity. We aim at utilising famous social media platforms to perform extensive analysis and classification tasks for posts capturing instances of various types of Inequalities prevalent in today's workplace. We present a framework to mine opinions expressed towards sexual harassment, mental health, racial injustice and gender-based bias in the corporate workplace using NLP techniques on social media data. The documents are represented by semantic similarity to aspect embedding's captured using an attention-based framework for aspect extraction. In addition, we used scores from Empath categories to add information related to emotional facets.

Keywords: Social Media Analytics; Aspect Extraction; Machine Learning; Natural Language Processing.

Received on 27 June 2022, accepted on 06 July 2022, published on 06 July 2022

doi: 10.4108/eetct.v9i31.1879

1. Introduction

In recent years, Big Data has created significant educational research impact on a variety of topics in socio-economic science, finance, and political science among others. For example, Big Data opens up new prospects for efficient development in banking and non-banking financial institutions by having an influence on financial markets through return predictions, valuations of the market and other similar activities. The sharp rise in social media usage and the ever evolving nature of the posts, social media analytics can help us gain valuable insights for creating efficient causal relationships and training ML models for various classification purposes.

Among multiple ways to know employees' opinion of their workplace, such as survey forms, one of them is to understand their views on a thought-expressing platform or community. **Reddit** is a network of communities where users can discuss their interests, passions and views on any specific topic. A **sub-Reddit** is a specific online community, and the posts associated with it. Every post on the sub-Reddit is associated with the context of that particular sub-Reddit. People openly talk and express their views on Reddit and thus Data extracted from Reddit are considered honest and unbiased because people post them without anyone forcing them. For our purpose we took some specific sub-Reddit related to Racism, Sexual Assault, and Mental health and on top of that we filtered the posts based upon some specific words related to corporate life (like work place, co-worker, boss) to get the posts capturing various instances of inequities based on racial and cultural differences.

Opinion mining is a methodology that can have a huge impact in evaluation of policies deployed by a company. In recent times, sentiment analysis has gained a lot of popularity in analysing textual data to understand the opinions associated with it by classifying the text in three categories: Positive, Negative and Neutral. Generally, opinion mining has two approaches: Sentiment Analysis (traditional approaches) and Aspect Extraction. Aspects or

*Corresponding author. Email: sonika.malik@gmail.com

topics discussed in a document are not targeted by the traditional approaches. Aspect-based opinion mining can be broken down into two main stages: Aspect Extraction, and Polarity Classification. The polarity classification's performance depends on the output of aspect extraction. Aspect extraction generates a complete list of objects, aspects and its opinions [1].

In our research work we aim to propose a machine learning pipeline using Aspect Extraction, Empath scoring and word embedding's of Reddit posts where thousands of user express their opinions regarding any particular topic and is a perfect data source to mine opinions and utilise the textual features to understand various aspects associated with the posts and clustering them on the basis of it. We compare our work with widely known algorithms like SVM, AdaBoost, Random Forest which act as baseline for our work and how our proposed pipeline and novel features perform in the classification task of the categories we created for capturing Gender and Racial inequality. We analyse the performance of Hierarchical Attention Network, which are recently being adopted on our proposed feature set capturing contextual and sentiment level information.

In this research, following questions have been investigated:

RQ-1 How can we mine various themes related to racial and gender inequities in social media posts using unsupervised machine learning?

To answer this research question, we have done a literature survey of the existing methodologies for various social media analysis frameworks. We used aspect extraction to study the various themes pertaining to workplace inequities.

RQ-2 How can Supervise machine learning classifiers be trained to identify these themes from the text?

To answer this research question, we created categories using sub-Reddit and trained Machine Learning classifiers on them using both traditional and neural models.

The remaining paper is structured as follows: Section 2 describes the related work; Section 3 discusses the Methodology used; Section 4 gives the Result and the last section concludes the paper.

2. Related Work

Aspect extraction has been extensively worked on in the past for effective opinion mining using various Deep Neural Network Architecture. [2] was one of the first studies to introduce the application of deep learning in the task of aspect extraction. They experimented with a deep CNN and showed its effectiveness for extracting aspects in comparison with existing approaches. Additionally, they used certain linguistic patterns in combination with neural networks for performing the task. [4] follows a Supervised approach for Aspect extraction of reviews of Restaurants where the corpus is for English language. One of the major drawbacks of this work is that the performance of the model depends on how the number of implicit features per sentence are distributed. [5] focusses on a Corpus and dictionary based model aiming for products and restaurant datasets. The model focuses on implicit aspect terms implied only by adjectives which serves as its major drawback. [6] utilises the co-occurrence between the opinions and words for this task on reviews for Restaurants, but this approach serves as a domain specific model. A generalised model which is not constrained to a single domain will be better for real world implementation of aspect based extraction on real world review data. [7] utilised Dependency Parsing for reviewing Hate crime review on Tweets which utilises the output of dependency parsing which gives a head and tail relation in a specific sentence and shows the relation between pairs of words in a sentence. It considered only adjectives as opinion words. [3] followed an unsupervised, rule-based approach but the accuracy of the method depends on the opinion lexicon. [8] used a supervised approach where the datasets having less data but more unique implicit features, the results were not good to be used in practice. [9] used a supervised, SVM based approach but the sentences containing infrequent explicit and implicit aspects were ignored.

[10] followed a supervised approach but the system may incorrectly tag the word in case of ambiguous opinions. [11] suggest a method to analyse user reviews in five aspects. The analysis consists of majorly five stages: data scraping, data pre-processing, retrieval of noun words using Stanford POS Tagger, classification of the noun words into aspects and finally the calculation of an aspect score on the basis of aspect-based sentiment analysis. To address the weaknesses of LDA-based approaches, [12] employed a novel approach using neural networks. The authors utilised the neural word2vec embedding's [13] which generates word vectors such that the words that often co-occur in the same context are close to each other in the vector space. Then, using an attention mechanism, they filtered the word embedding's within a document and created aspect embedding's with the filtered words. Aspect embedding's are trained similar to auto encoders, with dimension reduction employed to remove common factors across embedded sentences, and each sentence reconstructed using a linear combination of aspect embedding's. Words that aren't part of any aspect are deemphasized by the attention mechanism, enabling the model to give more weight to aspect words. The proposed model was called Attention-based Aspect Extraction (ABAE) [12].

[14] presents a method to quantitatively model organisational culture through crowdsourced workplace experiences shared by employees on various review platforms which are anonymized such as Glassdoor to explain the lexical semantics of the organisation's culture.

They reinforced the concepts in organisational behaviour understanding by validating that this model can be used to get a better evaluation of performance by a worker at the workplace. Finally, it contributes to better understanding of workplace experiences, and in improving functioning of an organisation through data-driven technologies.

3. Theoretical Background

In this section, we present the background of the methods used in our framework for classification purposes of social media posts. We give a brief description of N-Gram analysis, Aspect extraction, Empath analysis, Machine Learning models.

3.1. N-grams

N-Gram features have been used in several prediction studies for capturing linguistic expressions and context of the highest occurring terms in a corpus. We have used N-grams for N = 1 to 3 (Uni, Bi and Trigram) to identify the highest occurring single terms, pairs of words and triplets and use these as contextual feature input for our ML classifier. Figure 1 shows the highest occurring N grams of our work used as input features.

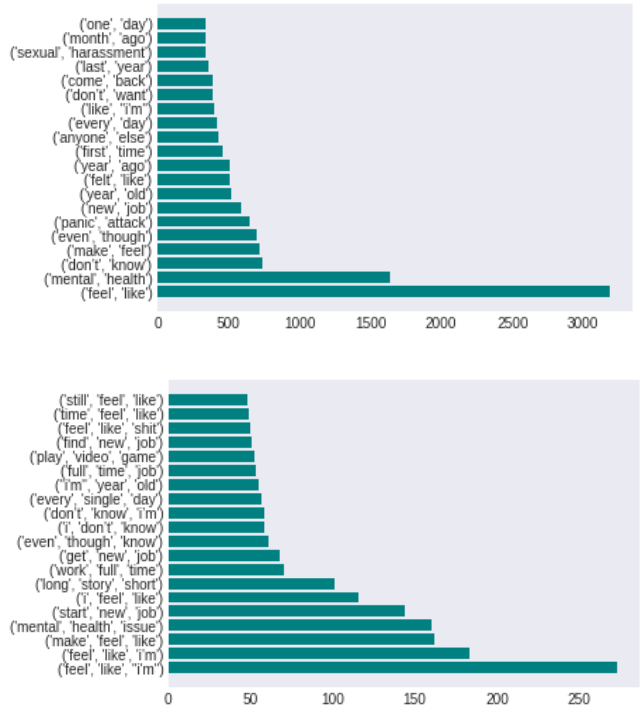

Figure 1. N-gram analysis for our Reddit dataset.

3.2. Empath Analysis

As a set of input features for our model, we have used Empath, a tool for finding category wise normalised word count strength of various inbuilt topics [15], which analyses text across 200 built-in pre-validated categories generated from common topics in a web dataset. The categories whose scores we have calculated for our Reddit dataset are: ['aggression', 'anger', 'contentment', 'confusion', 'disappointment', 'Disgust', 'fear', 'hate', 'horror', 'joy', 'lust', 'negative emotion', 'nervousness', 'optimism', 'positive emotion', 'sadness']

3.3. Aspect Extraction

Aspect extraction is a subtask of sentiment analysis that involves recognising characteristic views or opinions in text, that is, detecting the specific aspects of a review that a user is giving. The ability to record social media users' opinions on social or political movements, corporate strategies, product marketing preferences has piqued the interest of the scientific community and the business world due to the high impact it can have in forming efficient data driven decisions. Today, analysing opinions and effective extraction of topics being discussed on social media platforms have applications in numerous scenarios. Moreover, there are a lot of large and small-scale companies that focus on the sentiments of their employees as a part of their mission.

3.4. Machine Learning Classification

We have trained ML models separately for relevance and valence classification using the above mentioned input features for our scraped data of social media posts capturing gender and cultural inequities in the workplace. We consider multiple classifiers including Random Forest, Support Vector Machine (SVM), Adaboost and HAN.

SVM, Random Forest, Adaboost are famous ML algorithms which we have used as baseline and to understand how they are performing in comparison with Hierarchical Attention Network. The traditional ML models were run upon TF-IDF features containing top 5000 n-grams including unigrams, bigrams and trigrams. Hierarchical Attention Network classified the documents based on word2vec embedding's trained on our dataset. We analyse the performance of these models using the novel set of features which we have proposed for our dataset capturing contextual level information and empath scores of various categories.

3.5. Hierarchical Attention Network

Hierarchical Attention Network (HAN), introduced and proposed by [16] considers the hierarchical structure of documents which can be broken down into document - sentences - words and includes an attention mechanism that is able to give weighted importance to words and sentences in a document. This weighted importance helps the model to focus on parts of the data which hold more value while taking the context of these words and sentences into

consideration. Pre-existing works have focussed on importance or attention on the basis of the previous words in a sentence.

HAN tries to overcome the drawbacks of the previous work such as:

- Every word and sentence in a corpus carry different importance for various NLP tasks.
- The varying context in which a word occurs in a corpus needs to be taken into consideration.

3. Proposed Framework

In this section, we discuss our proposed framework, and the models applied for theme classification tasks on Reddit posts. We describe our methodology for data extraction from Reddit and the working of our pipeline to gain insights from it.

4.1. Design and Dataset

We performed an observational study by curating a dataset by scraping more than 7 thousand posts using the Praw library. The query used for extracting the posts was created using an 'OR' combination of flair and words related to gender inequality toward females, and the persons involved in corporate workspace. Detailed subreddits are mentioned in Table 1.

Table 1. Queries used for scraping posts from each word. Keywords for filtering: workplace, word company, corporate, co-worker, boss, manager, colleague, employer, employee.

Label	Subreddits	Number of Posts
Sexual Harassment	r/SexualHarassment	388
	r/sexualassault	476
	r/meToo	262
Racism	r/racism	577
	r/mixedrace	153
	r/aznidentity	186
Mental Health	r/stress	622
	r/anxiety	2148
	r/mentalhealth	2177
Feminism	r/sexism	80
	r/feminism	1573
	r/feminism	538

title	body	subreddit	search_word	label
Told my boss about what my coworker did to me	Just feeling super sad lately...I told my boss I...	sexualassault	workplace	sexual_harassment
My understanding of stress, anxiety and worrie...	So I have read books on anxiety and worries a...	Stress	manager	mental_health
Company hires men that like to sexually harass...	Not going to say how I know what these guys do...	SexualHarassment	employee	sexual_harassment
Recent anxiety around money	I (20F) have been working on and off since I w...	Anxiety	employer	mental_health
I got a job offer and am freaking out because ...	Hey there.\n\nI'm currently studying at univer...	Anxiety	colleague	mental_health

Figure 2. Overview of five randomly sampled records from the dataset.

Figure 3. Word Cloud analysis for our Reddit database.

An overview of the dataset is shown in Figure 2. Pre-processing of posts was carried out on lowercase converted text by removing white spaces, punctuation, hashtags, mentions, digits, stop words, URLs, and HTML characters. The words present in the text were lemmatized using WordNet Lemmatizer from the NLTK package. Duplicate posts were removed based on identical ID.

Each record contains the following fields: title of the post, body of the post, subreddits where it was posted, search word used, and label assigned. Figure 3 shows the Word Clouds generated from our corpus extracted from Reddit. We can see the major points of discussion in these posts via the figure to get a general idea about the topics discussed.

4.2. Working

Our proposed framework as shown in Figure 4, deals with the issue of capturing Social and Gender inequities through social media analytics. We studied this problem through the lens of both Supervised and Unsupervised Machine Learning. Supervised ML was used for the task of theme classification while we used unsupervised ML concepts for Attention Based Aspect Extraction. Both methodologies and our extensive feature extraction provide us with informative insights about the problem statement and how our analysis framework can help us capture important details for targeted interventions on social media to raise awareness.

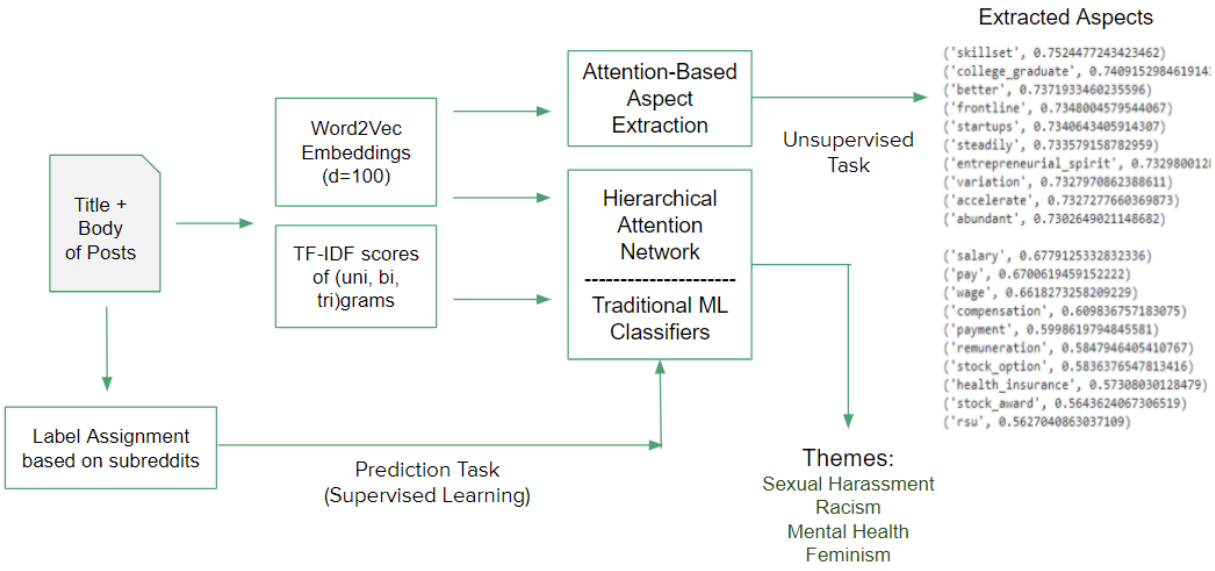

Figure 4. Proposed Framework

Supervised Learning

Our framework takes Reddit data as input and uses TF-IDF scores with its word embedding's from Word2Vec as input features for Traditional ML classifiers (Naive Bayes, Random Forest, SVM, Adaboost) and HAN respectively as depicted in Figure 4. The word embedding's are also used by an Attention Based Aspect Extraction model for generating aspects from the Reddit posts for analysis. We also did an extensive analysis of the social media posts using N-Grams and Empath to understand various emotions associated with the posts and the context involved with it. These findings provide an effective background of the dataset being analysed and associated insights which are important in Feature Engineering and it's better understanding.

Posts were assigned class labels on the basis of the Subreddits these were extracted from during data scraping. These Subreddits names serve as the class names for the posts and are used for training ML classifiers and evaluating their final performance. The predictions performed by our ML models are evaluated using Precision, Recall and F1 Score metrics.

Unsupervised Task

Unsupervised Learning has been widely used in applications where labels are not readily available. It has been used in NLP for finding themes or insights in the data using methods such as topic modelling and aspect extraction. The social media posts in our data represent a huge source of text that can be analysed to extract themes or aspects that people talk about with respect to their workplace. We used the ABAE model to extract 10 aspects or groups of words and demonstrate them with labels interpreted by the authors as shown in Table 1.

Extracted aspects were further analysed for a better understanding of the overview of posts in dataset as a part of the unsupervised ML results as shown in Table 2. Posts scraped were provided labels based on manual inspection and subreddits from which they were taken from and further used for training supervised ML algorithms for class prediction as depicted in Figure 4 flowchart of proposed framework.

5. Results and Discussion

Our findings suggest that users usually post lived experiences about harassment, mental health issues and tend to ask for advice from community members. These posts revealed 10 different aspects derived from Attention-Based Aspect Extraction model and are represented by the terms most similar to the aspect embedding vector (Table 2). We inferred the aspects from top ten terms to assign the most suitable label.

Table 2. List of aspects for Reddit posts, with top representative words for each aspect. Inferred aspect labels were assigned manually.

S.No.	Representative Words	Inferred Label
1.	gad, anxiety, agoraphobia, generalize_anxiety, Zoloft, suicidal_ideation, medicate, suicidal_thought, depression years, situational	Mental Health
2.	schedule_appointment, relay, iPod, FaceTime, skype, note, selfies, checking	Meetings
3.	evening, Saturday, weekend, skip, weekday, work hours, row, Tuesday, Friday, Sunday	Workdays
4.	overwhelming, panicked, intensely, guilt_shame, fragile, paralyzed, chaotic, grief, irrational, loneliness	Negative Emotions
5.	rumor, insult, homophobic, white_passing, fawn, ben, light_skin, proudly, Latino, dismissive	Gender & Racial Bias
6.	job months, hired, jobi, job years, underpaid, promotion, resign, marketing, promoted, transition	Job Satisfaction
7.	foremost, gentle, picky, storyi, wish_luck, headspace, figured, sorry_rant, needless, stress free	Optimism
8.	drinker, cocktail, missed, concert, housemate, subway, steakhouse, grocery_shopping, hostel, filled	Party
9.	repost, tutorial, motivational, summit, extensively, refuge, mobilize, accepting, Sandburg, surveyurl	Motivation
10.	woman, men, women, myth, trump, prostitution, irish, model, asia, descent	Sexism

Our findings showed that negative emotions and nervousness were the most prevalent in the posts. Heat map shows a high correlation between categories like ‘fear’ with ‘sadness’ and ‘nervousness’. ‘Nervousness’ and ‘sadness’ have also formed strong correlation in the posts. Other strong correlations observed in the heat map are between ‘optimism’ and ‘positive_emotion’. The correlations formed on empath scores as shown in Figure 5 help us get a better understanding of the nature of the posts in our dataset and various emotions associated with them. Figure 6 helps us get an overview of the distribution of posts from each category which were used to perform the Empath Analysis.

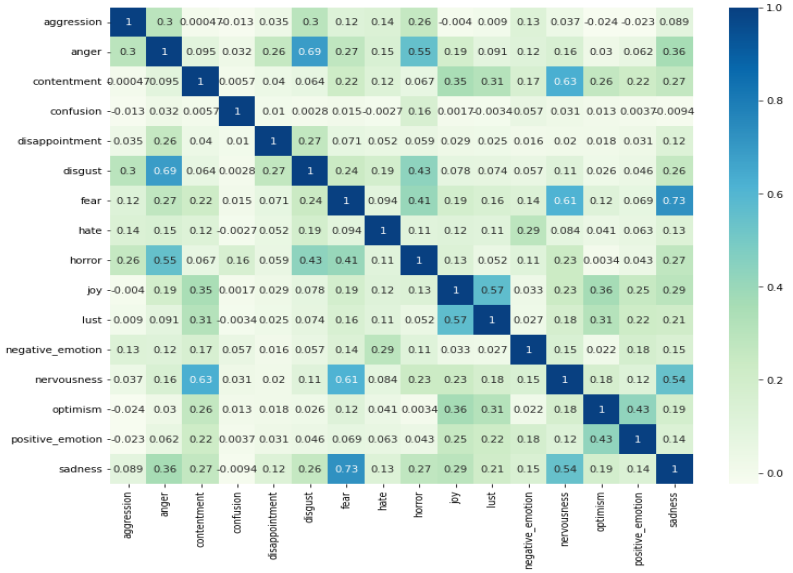

Figure 5. Correlation heat map for scores of empath categories.

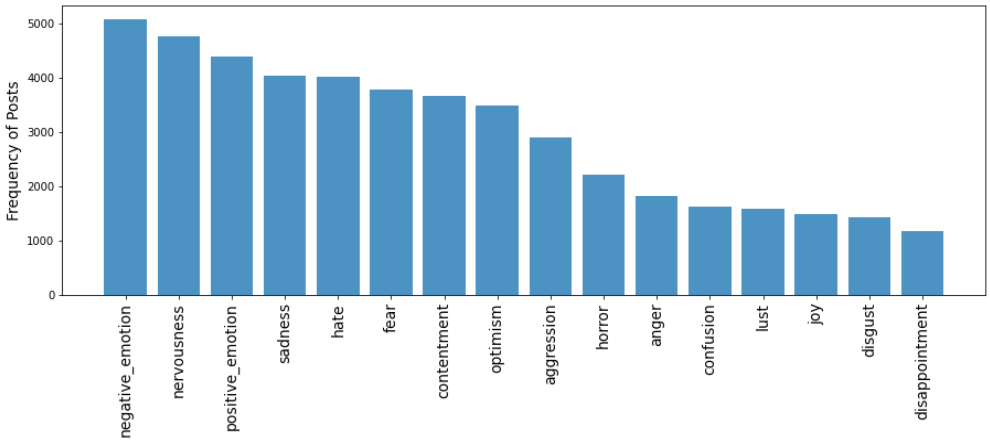

Figure 6. Frequency of posts having positive scores of empath categories.

The performance of multiple machine learning classifiers was evaluated on the Reddit posts. Table 3 shows the results of job satisfaction predictions.

Table 3. Performance of Supervised ML models on independent test set.

Model	Precision	Recall	F1-Score
Naive Bayes	0.843	0.822	0.796
Random Forest	0.866	0.858	0.844
AdaBoost	0.840	0.838	0.835
SVM	0.894	0.896	0.893
HAN	0.901	0.894	0.896

6. Conclusion & Future Work

In this work, we have extracted social media posts from Reddit capturing Social and Gender inequities at the workplace. Social media data was used due to its rich evolving nature of content and the opinions people expressed based on real life experiences. We further extract context rich and sentiment related features from our text corpus with aspects which are inputted into our classification pipeline for improving the pre-existing methodology. We further compare and analyse the performance of baseline Machine Learning algorithms with Attention Network models which have recently gained heavy usage for NLP related tasks. Our work shows an in-depth analysis of social media posts of reddit related to workplace incidents capturing gender & cultural inequities and introduces a methodology of utilising Empath scores and Aspects as features for improving the classification performance. We have extensively evaluated and compared the result of various famous ML algorithms with the Hierarchical Attention Networks model and how it performs on our proposed pipeline of features. A lot of directions can be explored in dealing with this problem statement. Transformer based architectures have shown huge potential in achieving impressive results on such problems. [17] used Transformers for the purpose of detecting racial bias for improving results and model complexity. They also provided a chrome extension for end users to identify racially biased text. Similar techniques can be extended for identification of texts for our problem statement. ML models can be replaced with transformers for better feature understanding and higher accuracy in our framework for possible future directions.

References

[1] Maharani W, Widyantoro DH, Khodra ML. Aspect extraction in customer reviews using syntactic pattern. Procedia Computer Science. 2015 Jan 1;59:244-53.

[2] Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems. 2016 Sep 15;108:42-9.

[3] Poria S, Cambria E, Ku LW, Gui C, Gelbukh A. A rule-based approach to aspect extraction from product reviews. InProceedings of the second workshop on natural language processing for social media (SocialNLP) 2014 Aug (pp. 28-37).

[4] Dosoula N, Griep R, Ridder RD, Slangen R, Schouten K, Frasincar F. Detection of multiple implicit features per sentence in consumer review data. InInternational Baltic Conference on Databases and Information Systems 2016 Jul 4 (pp. 289-303). Springer, Cham.

[5] Jiménez-Zafra SM, Martín-Valdivia MT, Martínez-Cámara E, Ureña-López LA. Combining resources to improve unsupervised sentiment analysis at aspect-level. Journal of Information Science. 2016 Apr;42(2):213-29.

[6] Panchendrarajan R, Ahamed N, Murugaiah B, Sivakumar P, Ranathunga S, Pemasiri A. Implicit aspect detection in restaurant reviews using cooccurence of words. InProceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis 2016 Jun (pp. 128-136).

[7] Fujita, H., and A. Selamat. "Hate Crime on Twitter: Aspect-based Sentiment Analysis Approach." In Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques: Proceedings of the

18th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_19), vol. 318, p. 284. IOS Press, 2019.
[8] Schouten K, Frasincar F. Finding implicit features in consumer reviews for sentiment analysis. InInternational Conference on Web Engineering 2014 Jul 1 (pp. 130-144). Springer, Cham.
[9] Xu H, Zhang F, Wang W. Implicit feature identification in Chinese reviews using explicit topic mining model. Knowledge-Based Systems. 2015 Mar 1;76:166-75.
[10] Chatterji S, Varshney N, Rahul RK. AspectFrameNet: a frameNet extension for analysis of sentiments around product aspects. The Journal of Supercomputing. 2017 Mar;73(3):961-72.
[11] Dina NZ, Juniarta N. Aspect based Sentiment Analysis of Employee's Review Experience. Journal of Information Systems Engineering and Business Intelligence. 2020 Apr 27;6(1):79-88.
[12] He R, Lee WS, Ng HT, Dahlmeier D. An unsupervised neural attention model for aspect extraction. InProceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2017 Jul (pp. 388-397).
[13] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013;26.
[14] Das Swain V, Saha K, Reddy MD, Rajvanshy H, Abowd GD, De Choudhury M. Modeling organizational culture with workplace experiences shared on glassdoor. InProceedings of the 2020 CHI conference on human factors in computing systems 2020 Apr 21 (pp. 1-15).
[15] Fast E, Chen B, Bernstein MS. Empath: Understanding topic signals in large-scale text. InProceedings of the 2016 CHI conference on human factors in computing systems 2016 May 7 (pp. 4647-4657).
[16] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. InProceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies 2016 Jun (pp. 1480-1489).
[17] Onabola O, Ma Z, Xie Y, Akera B, Ibraheem A, Xue J, Liu D, Bengio Y. hBert+ BiasCorp--Fighting Racism on the Web. arXiv preprint arXiv:2104.02242. 2021 Apr 6.