HICE: Hate Group Identification with Community Embeddings

Jayjay Jeongjin Park
jpark3141@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Marcos Grillo
mgrillo3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Cedrique Shum-Tim
cshumtim@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Reddit is an online platform well known for its relatively lax moderation [2]. This feature of the platform can lead to the forming of large communities focused on topics such as misogyny, discrimination against certain cultures and/or ethnic groups, and communities generally centered around hatred. However, previous research addressing such issues were mostly based on hate speech detection models, which do not consider community-based characteristics. To this end, we propose HICE, a classification model which detects hate groups on Reddit based on three different features that represent subreddit groups. Specifically, two different embeddings are trained and concatenated with one of the pre-trained community embedding using entity-based concatenation model [20]. Through experimentation, we observe that HICE with all of three embeddings performs the best compared to other baselines.

KEYWORDS

reddit, neural networks, community embeddings, hate groups

ACM Reference Format:
Jayjay Jeongjin Park, Marcos Grillo, and Cedrique Shum-Tim. 2023. HICE: Hate Group Identification with Community Embeddings. In *Proceedings of CSE-8803-DSN*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As a platform, Reddit is well known for being an exceptionally large social media network as well as being one where administrative level moderation is quite lax, particularly when compared with other mainstream networks such as Facebook or Twitter [2]. As such, and thanks to this hands-off approach, Reddit is particularly vulnerable to the rise of communities centered around topics like misogyny, racism, and political extremism. These communities have proven to be particularly harmful.

As an example, r/GenderCritical was a subreddit with a supposed focus on feminist ideology that ended up harboring and promoting transphobic viewpoints. Despite occasional efforts from the Reddit administration team to ban controversial subreddits, such as the aforementioned r/GenderCritical, or, most famously, r/The_Donald, a subreddit that began as a parodical Donald Trump support subreddit and rapidly devolved into an alt-right hub, rife with extremist views regularly being expressed. Left unchecked, this could prove problematic as having a mainstream social media site such as Reddit vulnerable to these sorts of communities could provide an easy and reliable entryway to what has been commonly referred to as the alt-right pipeline [19], which can be defined as a gradual process of radicalization in which online users are gradually exposed to increasingly polarized worldviews. with the aim of eventually redefining the individuals' views.

To better address such complex social challenge, we aim to build a classifier, HICE, which detects hate group from Reddit. Especially, we train our model with dataset from PushShift[1] from which we obtained comments and posts of 65,475 subreddits (from Jan 2014 to April 2017). Specifically, through our model, we try to answer research questions on ground truth geneartion for hate subreddit, which kind of features can possibly contribute in detecting hate subreddit and if we can develop an effective classifier for hate group identification. Unlike previous research which emphasized keywords to detect hate groups on Facebook [24] or other research which only focused on individual users rather than groups of user [9], we take community feature into account through pre-trained community-user embedding [14, 15] and creating node embedding for each subreddit based on subreddit hyperlink network [14]. Through experiment, we were able to observe that our proposed method, HICE with all of 3 embeddings combined performs the best in terms of all of metrics. Finally, our research has following implications: first, unlike other method used for creating community embedding such as community2vec [16] and word2vec [18], our vectorization method is specifically designed and developed for hate groups; secondly, our model uses rich social information, such as community-user interaction and community-community interaction, along with textual information to represent subreddit.

2 PREVIOUS WORK

Other studies have shown [8] that users on social media present tendencies towards isolating themselves in polarized groups which continuously reflect and reinforce their ideas. Reddit, as a particularly community focused network, is especially vulnerable to this problem. This same study also found that users who are more active in social media platforms present more rapid shifts towards expressing negative sentiment in their comments, seemingly independently of the topic in discussion. Users who were more exclusively active in these echo chambers also showed increased levels of negative sentiment in their expression online. These echo chambers also have a tendency towards exponential levels of growth in their early stages, before plateauing and stabilizing. This makes a rapid response to these groups especially important, since by the time more

CSE-8803-DSN, Fall 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://github.com/pushshift/api)

conventional detection methods come into action, these groups have already gone through most of their growth.

These hate groups can be problematic not only due to their impact on online spaces, but also due to their real-life consequences [10], where they can be used to promote and plan real life activities which can put others' lives in jeopardy. This is one of the main reasons why it is so important to address this issue in a rapid and effective manner. These communities could also be havens for recruiting of real life hate groups, as well as promoting a public image for public hate organizations, with the Ku-Klux-Klan as an example. Being able to improve early detection of these communities might hinder these organizations recruiting processes.

When it comes to a formal definition of what Hateful communities entail, we draw upon a modified version of a definition used by a previous study by Chau et al.[6]. to define hate groups in terms of groups of users who attack, hate and abuse other entities, which can range from groups of people, companies, ethnicity's, cultural or sexual identities and ideologies. We find this definition propitious towards our goals due to it encompassing several of the different groups that hate subreddits on Reddit have historically targeted, with r/coontown, as an example, being one that targeted black ethnic groups.

Early attempts at hate group detection have focused on a purely keyword-based approach. In a study by Ting et al., content from Facebook groups was analyzed for keyword prevalence based on a pre-defined list of hate group terms [24]. These early approaches, though a good start, were limited in their capability to detect these groups due to the limited scope of their methods.

Previous tools, such as https://subredditstats.com/subreddit-user-overlaps, have focused on building networks of subreddits based on what users are commonly active between them. These tools could provide a rudimentary method of detecting additional hate subreddits once an initial one has been identified. However, this method is heavily weighted towards correlating subreddits of relatively large size with one another, thus not providing utility when it comes to the task of early detection of hate subreddits at their early stages, before they hit the mainstream.

Though there have been concerted efforts focused on the detection of hate group subreddits [23] [21], these have mainly focused on detecting user-level features for embedding, leaving a noticeable gap of neural approaches to community-level embeddings.

Other studies [5] have demonstrated that the banning of subreddits has significant positive effects on the level of hate speech expressed on Reddit as a whole. Chandrasekharan et al. have found that the mass banning Reddit performed on multiple hate speech focused subreddits in 2015 led to significant reductions in hate speech expression on the site, even among those users who stayed on the site but migrated onto other subreddits. Although some "migrants" from these banned subreddits formed alternative new subreddits with the same purposes and goals as a means of dodging these bans, these were quickly banned from the site as well. After this wave, most of the participants of these former communities either left the site entirely or significantly reduced their levels of vitriol.

Waller et al. [25] have attempted to build networks of subreddits in order to classify subreddits as "generalists" or "specialists" according to their cosine similarity with other subreddits. This constitutes evidence of an application of community embeddings in order to achieve an experimental task on Reddit. Kumar et al. [14] also create community embeddings in order cluster subreddits into groups according to categories such as "gaming" or "controversial topics".

3 METHOD

Inspired by the work done by Waller et al. [25] and Kumar et al. [14], we propose a community-based graph embedding, and a classification model, HICE, which uses 3 different embeddings to detect whether or not a given subreddit is what we define as a hate subreddit, adhering to the definition of a hate group provided earlier. We define a binary classification problem, where a subreddit can only ever be classified as deserving of the hate group tag or undeserving of it. With our method, we try to address our main research questions:

(1) How can we generate ground truth for hate subreddit?
(2) How effective or useful is our generated network embedding in hate group detection?
(3) Is it possible to construct a community-based classifier to determine whether or not a given subreddit is a hate group?

3.1 Dataset

In this section, we will explore our process for data acquisition and processing.

We utilized the pushshift dataset [1] as a source for our data. This dataset is publicly available and allows both querying via an api as well as accessing monthly dumps of all comments and submissions created on Reddit during that time period. Compressed files containing all comments and posts were acquired from the dates of January 2014 to April 2017 (aligning with Kumar et al.'s work as we have planned to use their pre-trained embedding [14, 15]. Comments and submissions were sampled at regular intervals from these compressed files, resulting in a final pull comprised of 11,810,605 comments and 17,971,800 posts, originating from 65,475 subreddits. Since our extraction was done with regular intervals between both posts and comments, a nearly identical amount of submissions and comments were extracted from each month (roughly 295,265 comments and 449,295 posts per month). This was done in order to ensure our sample equally represented the different time periods in question. Comments sampled per subreddit ranged between 1 and 902,582, with posts sampled per subreddit ranging between 1 and 513,187. This is reflective of the wide range in variability in size between different communities on Reddit.

3.2 Ground Truth Generation

Although some previous research have conducted case studies on hate group in social media platform, to the best of our knowledge, no ground truth for hate groups on Reddit existed [22]. Therefore, to experiment with our embedding and classifier, it was only natural to generate ground truth first. We create ground truth in 3 steps and evaluate it with manual annotation.

First, a list of potential hate subreddit was obtained by using hate keywords from previous research on Twitter hate speech dataset [11]. Specifically, we utilized 51 hate keywords created by ElSherief et al. [11], which include hate key words for various categories of

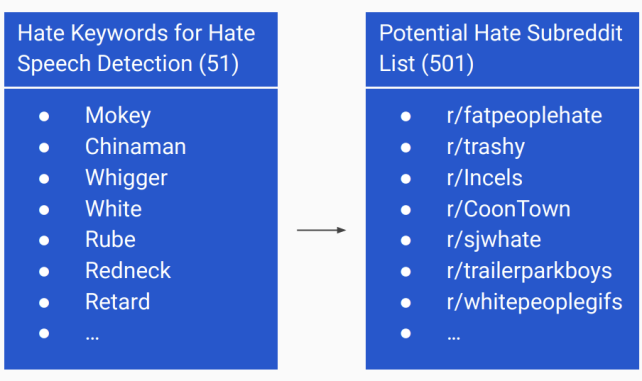

Figure 1: List of Hate Keywords and obtained List of Potential Hate Subreddit

Label	Non-Hate Group	Hate Group
Weighted Jaccard Index	score < 0.14	score >= 0.14
Number of Subreddit	471 (94%)	30 (6%)

Table 1: Result of Labeling using Weighted Jaccard Index

hate speech target such as ethnicity, gender, nationality, religion and sexual orientation. By selecting subreddits that contained those hate keywords from the entire dataset, we obtained 501 potential hate subreddits.

Then, to look deeper into what kind of content these potential hate groups discuss, we conduct topic modeling using Latent Dirichlet Allocation [3]. As each subreddit is of different size; therefore, have different number of topics discussed, rather than using one fixed number of topics, we optimized coherence score by finding the best performing number of topics. The resulting topics included target-relevant words such as woman, guy, child, girl, and hate keywords such as kill or rape.

Next, we computed weighted Jaccard Index score to see how each subreddit's topics are relevant to hate group. Like it was mentioned in previous research, Jaccard Index has limitations in that it does not reflect to which level one set is in to the other set [7]. In our case, as original Jaccard Index score does not take frequency of hate keyword appearing and size of subreddit into account, we created weighted Jaccard index score, to complement the weakness of the original Jaccard Index Score [13].

$$Weighted\ Jaccard\ Index = Jaccard\ Index \times \frac{Frequency\ of\ Hate\ Keyword}{\#\ of\ Topic\ Obtained\ from\ LDA}$$

In order to decide the threshold value for Weighted Jaccard Index to classify hate group from non-hate group, through observation of the result, we agreed upon a threshold of 0.14. This is because 0.14 was the lowest score that not only includes most number of hate groups but also most effectively distinguishes publicly known 8 hate groups from non-hate group. To obtain those definite 8 hate groups, we combined identified hate groups from previous research and publicly identified hate groups from The Washington Post article[2] [23]. Using this 8 hate groups as our baseline for Weighted Jaccard Score Index, we observed that the list of subreddit groups that has Weighted Jaccard Index higher than equal to 0.14 holds 7 out of 8 definite hate groups. Hate subreddit that was not inlcuded in the list, NEO_FAG, had Weighted Jaccard Score of 0.08, which is indistinguishable value from non-hate group. One of the possible reason NEO_FAG had such a low Weighted Jaccard Index although it is a hate group is because comments that contained hate keyword are already deleted and removed.

As a result of automatic labeling using Weighted Jaccard Index shown above, we obtained the following result, Table 1, obtaining 30 hate group. However, through inspection, we realized that some false negatives exist, which means that Weighted Jaccard Index is still missing some possible hate groups as possibly because of the dataset we have. Regardless of its activeness, some hate group's comment was 1 or 2 as most of it was deleted.

Therefore, to reduce the number of false negatives, we manually annotated 501 potential hate groups. Each subreddit's top scoring posts were assessed. When no comments or only few comments were available and subreddit was already banned, we used waybackmachine[3] for more information. Decision on whether or not a certain subreddit is a hate group or not was based on our definition of hate group mentioned prior. Disagreements were settled by group discussion. As a result of the annotation process, we found 87 hate groups. Inter-rater's agreement, Kappa Score, was 0.701, which implies a substantial agreement between the raters [17].

After gaining the final ground truth from annotation, we wanted to assess by what degree Weighted Jaccard Index is a useful metric. When calculating independent t-test score, we obtained two-tailed p value less than 0.0001, which means that Weighted Jaccard Index between hate group and non-hate group was statistically significant. At the same time, when calculating accuracy of automatically labeled result by comparing it with our annotated ground truth, we obtained 0.89. These results imply that actually our Weighted Jaccard Index is helpful in detecting hate group. A possible future research can be done regarding in-depth threshold evaluation to find the most effective value in classifying hate group from non-hate group using Weighted Jaccard Index as metric.

3.3 Proposed Method: HICE

By observing our result from topic modelling from the reddit dataset, we noticed that quite a lot of hate keywords are shared between hate group and non-hate group. For example, gaming communities' topic words contained "kill" and other abusive words. Therefore, our method considers network features in addition to content features.

[2]https://www.washingtonpost.com/news/the-intersect/wp/2015/06/10/these-are-the-5-subreddits-reddit-banned-under-its-game-changing-anti-harassment-policy-and-why-it-banned-them/

[3]https://archive.org/web/

Table 2: Comparison between multiple HICE models with different Hyperparameter

Model	Optimizer	Train-Test Split	Precision	Recall	AUC
HICE 1	Adam	80-20	1.00	0.06	0.78
HICE 2	SGD	80-20	1.00	0.06	0.75
HICE 3	Adam	KFold	1.00	0.07	0.81
HICE 4	SGD	KFold	1.00	0.11	0.85

Table 3: Comparison between HICE and baseline models on Precision, Recall and AUC

Model	Feature included	Precision	Recall	AUC	F1-Score
Baseline: RF	User-Com	0.00	0.00	0.5	0.00
Baseline: SVM	User-Com	0.08	0.25	0.5	0.12
RF w/ all 3 embeddings	Content + User-Com + Com-Com	0.00	0.00	0.5	0.00
SVM w/ all 3 embeddings	Content + User-Com + Com-Com	0.50	0.39	0.65	0.44
HICE: w/o graph embedding	Content + User-Com	1.00	0.06	0.77	0.53
HICE: w/ all of 3 embeddings	Content + User-Com + Com-Com	1.00	0.11	0.85	0.56

User-Com is abbreviation for Community-User embedding, Com-Com is abbreviation for Community-Community embedding which was created using GraphSAGE [12]. Finally, Content refers to context embedding which was created with Universal Sentence Encoder [4]

As a result, for our model we consider 3 different types of embedding: content embedding, community-user embedding, community-community hyperlink embedding. After generating different types of embedding, which contribute in representing subreddit group in different aspects, we concatenate all of 3 embeddings using entity-based concatenation model to obtain final representation of one subreddit.

$$Subreddit\ Group = \begin{bmatrix} Content\ Embedding, \\ Community - User\ Embedding, \\ Community - Community\ Embedding \end{bmatrix} \quad (1)$$

For contextual representation of a subreddit, We build content embedding using Universal Sentence Encoder [4], which will capture semantic meaning when that is not possible when only relying on hate keywords detection. Universal Sentence Encoder version 4 has pre-trained embeddings from Google that encodes sentences or words as a vector of 512 features. Sentences and words that are semantically similar will have similar embedding vectors.

For community embeddings, we re-utilize the pretrained embedding and hyperlink network of subreddits created by Kumar et al. [14, 15], which is accessible from Stanford Network Analysis Project[4,5]. These embedding and dataset are based on Reddit data ranging from January 2014 to April 2017. Specifically, Subreddit embeddings are built based on user-to-subreddit posting network using obejctive function that is similar to Word2Vec [18]. The size of embeddings are 300 dimensions each, and in total this dataset contains 118,381 users and 51,278 different subreddits [14, 15]. For the rest of the paper, we will call this Subreddit embedding created by Kumar et al. as Community-User embedding, as it captures subreddit's user base and for the sake of distinguishing embedding that will be generated using Reddit Hyperlink Network [14].

To create Community-Community Embedding, which captures network or interaction feature of different subreddit groups, we utilized Reddit Hyperlink Network, which is built based on post that contains hyperlinks from one subreddit to another subreddit [14]. Specifically, the network we used was generated based on hyperlinks in the title of the posts, containing 55,863 different subreddits and 858,490 edges [14]. Additionally, each edge has weights (-1 or 1) with attributes, text property vectors [14].

Using this network data, we create node embedding for each subreddit using GraphSAGE [12]. GraphSAGE takes a community network dataframe of source and target subreddits as vertices and a community feature dataframe for node characterization. The subreddit interactions were provided by the paper [14] and subreddit network were taken from the community-users interactions embedding.

To concatenate graph embedding with contextualized word representation into one, we use entity-based concatenation model proposed by Polignano et al. [20]. To be more precise, we concatenate all 3 embeddings into a single 1-dimensional array to use for binary classification. For the content embeddings, each subreddit has a single 1-dimensional for every comment in the subreddit, creating a 2-dimensional feature matrix that is representative of the subreddit's content. This 2-D matrix is thus flattened in order to become a single dimension to be concatenated with the other embeddings. For the Community-User interactions embedding, each subreddit has a vector with the size of 300 features that embed the community's interactions with users, taken from [14]. Since this vector is already 1-dimensional, no data preprocessing is needed

[4]http://snap.stanford.edu/data/soc-RedditHyperlinks.html
[5]http://snap.stanford.edu/data/web-RedditEmbeddings.html

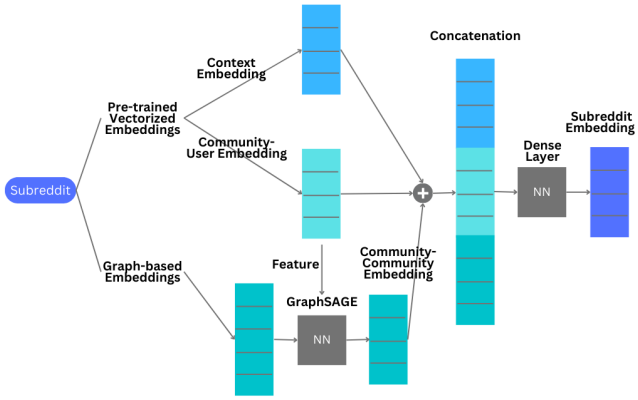

Figure 2: Entity-Based Concatenation Model

before concatenation. For the Community-Community interactions embedding, a combination of source-target subreddit data from [14] and the Community-Users interactions network will be used to create a GraphSAGE [12] node embedding that will be in turn concatenated with the other 2 embeddings.

3.4 Experiment Setting and Baselines

For the purposes of this project, the required task is the binary classification of Reddit communities into two categories: hate or non-hate groups. As stated in our methodology section, the hypothesized model takes 3 types of embeddings (content, community-user interactions, and community-community interactions), concatenates them, and performs binary classification on this data. In order to demonstrate the advantages of including all 3 types of embeddings, the classification task was run with a concatenation of 2 embeddings (content and community-user interactions) and with a concatenation of 3 embeddings for performance comparisons.

For our experiments, we chose the following evaluation metrics: precision, recall, AUC, and F1-score. The reason we chose these evaluation metrics is because these metrics are recommended for the evaluating performance on a dataset that is imbalanced, which ours is.

The parameters off our experiments are described ahead. When experimenting with the binary classification model, we ran a Feed-forward Neural Network 4 times as shown in Table 2. Each experimental run had a different combination of optimizer (Adam or Stochastic Gradient Descent) and train-test data split methods (normal random 80-20 split or KFold with 5 number of splits). Each iteration of the Neural Network is a TensorFlow Keras Sequential model with a Dense layer of 30 neurons with ReLU activation function and a Dense layer of 1 neuron with Sigmoid activation function for binary classification. The default TensorFlow Keras Sequential model cross-validation was used. Through this experiment setting, we found that settings for HICE 4 yielded the best performing model with highest precision, recall, and AUC scores. The system settings for these experiments are: 2.3 GHz Quad-Core Intel Core i5 CPU, 8 GB 2133 MHz LPDDR3 RAM, and Intel Iris Plus Graphics 655 1536 MB GPU.

For baselines, we decided to use both an SVM model and a Random Forest model as a base. For both baselines, we performed an 80-20 split, and trained both models with no parameter tuning methods. We use precision, recall, F1-scores and AUC as performance metrics for the evaluation of these baselines.

Our Random Forest baseline shows good precision, recall and F1-scores for the classification of non-hate subreddits (precision: 0.95, recall: 1.00, f1: .98). However, the model performed extremely poorly when it came to the classification of hate subreddits (precision: 0.00, recall: 0.00, F1: 0.00). The overall statistics for the model are reflective of this imbalance in performance (precision: 0.48, recall: 0.50, F1: 0.49, AUC: 0.50).

Our SVM baseline performed remarkably similarly. Like the previously mentioned baseline, this one performed well in the task of classifying non-hate subreddits (precision: 0.95, recall: 1.00, F1: 0.98), but poorly in the task of hate subreddit detection (precision: 0.00, recall: 0.00, F1: 0.00). Overall this baseline once again presents performance metrics reflective of this imbalance (precision: 0.48, recall: 0.50, F1: 0.49, AUC: 0.50).

We hypothesize that the poor performance in the detection of hate subreddits is due to the imbalance present in our dataset. For the test set used here to establish baselines, there were 100 non hate subreddits, and 5 hate subreddits present.

4 EXPERIMENT RESULT

When comparing the results that were obtained from our baselines and our new approach in Table 3, we can see that the use of the 3 concatenated embeddings with a Feed-forward neural network performs best, as this yields best precision, AUC, and F1-scores. This shows that the addition of the Community-Community interactions GraphSAGE[12] embedding to the other 2 embeddings improves upon its performance in a significant manner, as well. However, it is also possible to see that just by using the concatenation of the 3 embeddings improves the performance of one of our baselines, the Support Vector Machine classifier, where its precision goes from 0.08 to 0.50, its recall goes from 0.25 to 0.39, it's AUC score increases from 0.5 to 0.65, and its F1-score improves from 0.12 to 0.44. The performance of the Random Forest classifier doesn't change and stays very poor for this task. This might be the case since Random Forest models are an ensemble of Decision Tree models, which are very sensitive to dataset balance because of the process of pruning features.

5 CONCLUSION

5.1 Limitations and Future Work

We encountered several roadblocks throughout our project. Here, we explore some of the limitations of our work.

Firstly, our research was conducted with relatively old Reddit data (ranging from January 2014 to April 2017). This would potentially diminish the effectiveness of a hate group detection model if it were to be used with more current data.

Secondly, our final dataset used for modelling contained a very small number of hate subreddits. This is due to the fact that the community embeddings provided by Kumar et al. and the data extracted by our team for analysis do not fully align in terms of the included subreddits. Reddit has an extremely large number of

subreddits, and we have found that most of the hate groups detected by our analysis are relatively small, thus most of them were not captured by the group embedding analysis. This could potentially harm the applicability of our classification model, as it runs the risk of being overfit to the relatively small corpus of hate groups we worked with.

Both of these issues can be remediated by re-running our extraction process on modern Reddit data. Cloud computing resources could be used to achieve this task, and our entire method process, including community embedding generation, could be performed on the same raw data pull. With these resources, we could pull a larger corpus of data, and perform a more unified analysis which could then be fed into a classifier, in hopes of working with a more modern dataset, as well as one with, ideally, more hate subreddits included.

5.2 Contributions

Jayjay designed the entire architecture of the proposed model HICE, all parts related to ground truth generation including topic modeling using LDA, formulating and calculating Weighted Jaccard Index, annotation and analysis using independent t-test; wrote method part and other parts for final report including creating figures and tables, and assisted in graphSAGE embedding generation, HICE implementation, and experimentation of classifiers and HICE.

Marcos was the initial proposer of the project idea. He created the pipeline for data sourcing from the Pushshift dataset, as well as retrieving the data used in this work. He also performed extensive bibliographical review in order to retrieve methodologies for previous work methodologies. Marcos also created the baseline models and assisted in manual labelling of the dataset. He provided support in explaining the nuances of how the Reddit community works, as well as identifying potential roadblocks for the project.

Cedrique generated Universal Sentence Encoder content embedding, graphSAGE node embedding, and developed and trained 4 different versions of HICE. He assisted in the writing and implementation of Weighted Jaccard Index score. He also annotated subreddit dataset. He contributed in the refinement of the proposed HICE method with fine-tuning of micro ideas. He was responsible for the entire data pre-processing pipeline that was required for getting the data ready for embedding generation and model design, after being pulled with PushShift API. He wrote "Experiment Setting and Baselines" and "Experiment Result" sections of final report, and also assisted in writing method section.

REFERENCES

[1] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[2] Iris Birman. 2018. Moderation in different communities on Reddit–A qualitative analysis study. (2018).

[3] Andrew Y. Ng Blei, David M. and Michael I. Jordan. 2003. Latent dirichlet allocation. In *Journal of machine Learning research*.

[4] Yang Y. Kong S. Y. Hua N. Limtiaco N. John R. S. Kurzweil R Cer, D. 2018. Universal Sentence Encoder. (2018).

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[6] Michael Chau and Jennifer Xu. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies* 65, 1 (2007), 57–70.

[7] L. D. F. Costa. 2021. Further generalizations of the Jaccard index. In *arXiv*.

[8] Vivaldo G. Bessi A. et al. Del Vicario, M. 2016. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Sci Rep* 6 (Dec. 2016). https://doi.org/10.1038/srep37825

[9] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[10] Karen M. Douglas. 2007. Psychology, discrimination and hate groups online. *The Oxford handbook of internet psychology* (Dec. 2007), 155–164.

[11] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Instigators and Their Targets. In *Proceedings of the 12th International AAAI Conference on Web and Social Media* (Stanford, California) *(ICWSM '18)*.

[12] Ying Z. Leskovec J. Hamilton, W. 2017. Inductive representation learning on large graphs. (2017).

[13] P. Jaccard. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines.. In *Bull Soc Vaudoise Sci Nat*.

[14] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*. 933–943.

[15] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1269–1278.

[16] Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 27–31.

[17] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[18] Chen K. Corrado G. Dean J. Mikolov, T. 2013. Efficient estimation of word representations in vector space. arXiv.

[19] Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday* 24, 6 (June 2019). https://doi.org/10.5210/fm.v24i6.10108

[20] Musto C. de Gemmis M. Lops P. Semeraro G. Polignano, M. 2021. Together is Better: Hybrid Recommendations Combining Graph Embeddings and Contextualized Word Representations. (2021).

[21] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society* 7, 4 (2021), 20563051211052906.

[22] Kümpel A. S. Wich M. Kiening T. Groh G. Rieger, D. 2021. Assessing the exten and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. Social Media+ Society.

[23] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159* (2017).

[24] I Ting, Hsing-Miao Chi, Jyun-Sing Wu, Shyue-Liang Wang, et al. 2013. An approach for hate groups detection in facebook. In *The 3rd International Workshop on Intelligent Data Analysis and Management*. Springer, 101–106.

[25] Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600, 7888 (2021), 264–268.