Online Political Polarization: Assessment on the correlation between people posting about Indian festivals and their political ideologies

Mayank Singh, Kausik Kappaganthula, Heygon Araujo, Riya Dharmesh Damani
University of Koblenz-Landau, Campus Koblenz
{mayank, kausikk, riyadamani, heygon}@uni-koblenz.de

Abstract

In this study we investigate how we can use the machine learning techniques and graphical structure of the users on social media, to find out the political inclination of a user based on the user's reaction to religious festivals in India. Our study's main focus was on the user of Twitter and Reddit. We applied machine learning techniques (Support Vector Machine, Random Forests) on the user's post/tweet to classify them to a particular political party. We found that the Random Forests algorithm produced the best result. In our second approach, we used the network structure of the Twitter and Reddit users. We found that in the case of Twitter we can gain insight into the political inclination of the user using their network structure. We also compared the readability of Twitter and Reddit users using the ARI formula and found that Reddit is more readable. The results represented visually through a network structure depicting the people's inclination towards a particular political party which inline illustrates polarization among people.

Keywords: Online political polarization, Twitter, Reddit, religious festival, Machine learning

1. Introduction

Social media platforms give us a great chance to express our opinion, participate in discussions and react to events happening around the world. In recent years, the rise of social media has given an opportunity for the political parties as well individuals to expand their reach to the people enormously [3]. This in-turn have given rise to conflict between people because of difference in opinion, thus resulting in polarization [8]. Discussion about religion and showing hatred among people belonging to a particular religion can be easily seen on the internet [2]. To express the views, users use hashtags to convey their message. The hashtag is a kind of meta-data that is represented by "#" and is easily searchable in Twitter search. Whenever some event happens, social media platforms erupt with different hashtags in which people following different ideologies attack each other. Similarly, during a festival, some people will warmly wish their followers and on the contrary, some will bash the religious custom. In our context, we are considering India because it is the largest democracy and has various religions. For example, during Diwali, one of the Twitter user says " *Some idiots are bursting crackers early in the morning. The monkeys here are still crying. What is this absurd behavior, 4:03 AM? How's this festivity?*". On the other hand, another Twitter user says " *Let's celebrate Eco-friendly EID this year. Stop killing Animals.!*"
This demeaning of religious belief leaves a vivid devotee to be angry. This research focuses on the polarization of users on social media considering their views about a religious festival. Understanding the dynamics of this information panorama has become critical because social media can strongly influence public opinion[1].

The religious and political spectrum in India is so diverse and there are several prominent festivals that are being celebrated in India. For our research, we considered two important major festivals. Among them Eid-al-Adha (The Muslim festival of sacrifice occurs annually during the Hajj). Animal sacrifices are performed during the festival in recognition of the willingness of Prophet Ibrahim to sacrifice his son Ismail for God's sake. Diwali is one of the major religious festivals of India. It is generally celebrated in the months of October and November and associated with the burning of crackers and sparkles. And coming to the political spectrum India also has many political parties, but the current mainstream party in India is BJP[1] and Congress[2] [8].

Using machine learning techniques and graph theory on social media, we tried to find out if there is a relationship

[1] BJP's full form is "Bhartiya Janta Party". It is mostly called a right-wing party. BJP currently has the majority of seats in the parliament

[2] It is one the oldest party in India, which ruled the country for 70 years after independence but losing the parliament election since the last 2 elections. Congress is been mostly a left-wing party

between the Indian festivals and user's political ideologies. To find out the polarization between the group of people we need to figure out how the users behave in a network and what they post. One of the methods which the researchers are implementing recently is to use a system that can identify the texts and classify them into different classes [4]. Therefore, we used and compared machine learning algorithms such as Support Vector Machine and Random Forests to classify user's tweets to a political ideology.

The other method, which is not being researched much is classifying the users based on their relationships in a graphical network. A user tends to connect to another user if they share the same interests. The users in a social network can be represented by nodes and their relationship by an edge. The relatedness of two nodes can be calculated using the geodesic distance between the nodes. The geodesic distance between two nodes can be calculated by finding out the smallest number of edges connecting the two nodes [5]. So, if two nodes A and B in a network are connected directly, their distance will be 1 which signifies that they are connected closely. Suppose if there are two more nodes in between A and B then their geodesic distance will be 3. If the geodesic distance between two nodes is high then those two nodes are not connected closely. Therefore, to analyze the different types of users, we represented our target users in a graphical network.

2. Related work

We have used two different approaches to find out the extent of the political orientation of our targeted users based on their tweets/Reddit posts.

One of the approaches is based on creating the graphical structure of the user's on Twitter and Reddit. The approach is inspired by the paper by Hutair et al.,2016 [4]. They considered a social network as a network of nodes and edges, where nodes represent the user and edges represents their relationship. The relatedness between the nodes can be seen as the similarity between two nodes based on interests and connections.

The other approach is based on using machine learning algorithms, basically using NLP techniques to classify the users based on their tweets/posts. This approach is closely related to the work done by Prati et al.,2019 [6] where they implemented some of the machine learning algorithms to find out the political inclination of the tweets during the Spanish election, and Sarkar et al.,2019 [7] where they compared multiple machine learning algorithm while classifying Twitter data.

3. Proposed architecture

With the rise of social networks, researchers have concentrated on the study of the methods to find meaningful information from a large amount of unstructured data.

We used two approaches to find out some significant information from the data.

3.1.Machine learning approach

The underlining steps are almost similar for Twitter and Reddit. The only difference comes when collecting the data. This is due to the fact that Reddit users are mostly anonymous by their name. Whereas, on Twitter, political party members can easily be found.

In this approach we collected tweets, then we did prepreprocessing to our data to make sure it can be used in our machine learning algorithm for the feature extraction. We then made two corpora for right-wing and left-wing parties. We converted the texts into machine learning readable format using TF-IDF vector modeling. Two machine learning algorithms are used i.e SVM (Support Vector Machine) and Random Forests to classify the result.

Figure 1 represents the steps followed for our approach using the machine learning technique

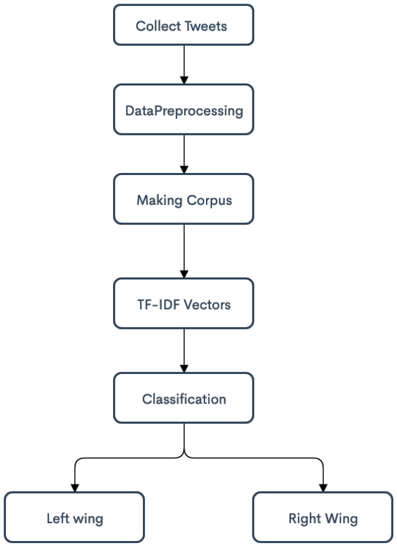

Figure 1: Work flow for machine learning approach

3.2 Graphical representation approach

In this approach for Twitter, we fetched the users and then represented them as nodes and their relationship as edges. We differentiated them with color which can further be useful for our analysis.

On the other hand, for Reddit, we fetched the users based on their comments/posts on the subReddits. We represented their connection with other subReddit to do our analysis.

Figure 2 represents the workflow for our graphical approach.

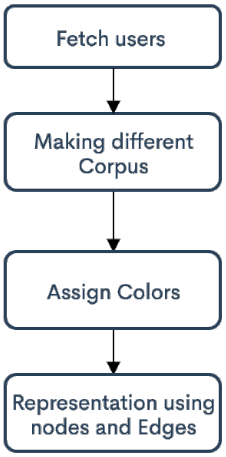

Figure 2: Work flow for graphical representation approach

4. Methodology

On one hand, people celebrate these festivals and spread positive vibes with their posts but on the other hand, some social media users engage in arguments, demeaning the cultural practices of the festival. According to Pierre Beumarchais *”As long as I don’t write about government, religion, politics and other institutions, I am free to print anything”*. This motivated us to find if a person’s dissenting views about a religion or a religious festival can be a key component to find its political inclination.

4.1.1 Machine learning methodology in Twitter

A)Dataset description
The first step was to find out the users and their last 200 tweets who posted derogatory tweets about a religious festival on Twitter. We chose users whose post contains either of the hashtags i.e. #bloodlesseid, #crackerban #veganeid #nocrackersdiwali, #BloodLessBakrdi, #ecoFrienndlyEid, #saynotocrackers. We also manually chose some of the users based on their posts. This means a user can also post derogatory without the use of hashtags.

The second step was to fetch the last 200 tweets from the 5 most popular politicians from the right-wing party(BJP) and left-wing party(Congress) each.

B)Dataset pre-processing
Raw tweets are not suitable to use for machine learning models[4]. All the unimportant features like hashtags, URL, special characters, empty spaces, uneven letter case needs to be removed. We also removed stop words[3] and used lemmatisation[4] to ensure the inflection form of a word can be represented as a single common word. These actions were done using the functions provided by NLTK[5] in Python.

C)Classification techniques
We transformed these texts into numerical vectors so that they can be used for the TF-IDF technique. The TF-IDF technique measures the relevancy of a word towards a particular document in a collection of documents. It is calculated by multiplying the term frequency(number of times a word appears) with the inverse document frequency of the word that occurred in all documents.

TF-IDF = TF x IDF

IDF = $\log_2(N/n_t)$

where TF is the frequency of the term in the tweet. IDF(Inverse Document Frequency) is the logarithm of the total number of tweets (N) divided by the number of documents (tweet) where the word occurs (n$_t$)[6].

The next step was to create a suitable model for our prediction of the user’s political inclination. We used SCIKIT-LEARN[6] package to apply the Random Forests algorithm and the SVC algorithm.

Random Forests is a tree-based structured classifier which is efficient in the large database and can handle thousands of feature variable. During the process of creating the Random Forests, several decision trees are generated randomly during the preprocessing. These decision trees are voted and the unseen dataset is classified depending on the votes overall the decision tree [6]. This method gave us more accuracy.

SVM (Support Vector Machine) is a linear model used for classification by creating a hyperplane between two or more different classes [7]. This method gave less accuracy.

The predictive model mentioned above is then used to identify the political inclination of the user’s tweet.

4.1.2 Machine learning methodology in Reddit

Apart from the data collection part, all the steps are the same as mentioned in subsection 4.1.1.

We manually observed the posts of the two most popular subReddits based in India. They both have more than 100 thousand followers. In some of the posts in both the groups, the user were found to follow a similar pattern as we saw on Twitter i.e. they were making fun of or criticizing the religious custom of a religious festival. Therefore, we took

[3] https://en.wikipedia.org/wiki/Stop$_{word}$

[4] https://en.wikipedia.org/wiki/Lemmatisation

[5] https://www.nltk.org/

[6] https://scikit-learn.org/

the top 1000 users from both of these groups and fetched the 200 posts by them. Here in our approach, we assumed that one of the groups is pro-government i.e. they are right-wing, as most of the posts were about appreciating the government. The other group was sort of anti-government i.e. left-wing as the posts were about criticizing the government.

The rest of the procedure was similar to the methodology used in subsection 4.1.1

4.2.1 Graphical representation methodology in Twitter

The analysis is done on the **Twitter friendship network**[7] and for **Reddit, user's posts in the groups**[8] is considered.

A)Dataset description

Similar to what we did in subsection 4.1.1 (c), we chose the same hashtags. But, instead of collecting all the tweets of a user, we fetched all the followings of that user. We fetched the following of 10 users for each festival i.e. Diwali and Eid.

Furthermore, we fetched the following of the 10 most popular politicians from the right-wing and left-wing each.

TWEEPY[9] which uses API provided by Twitter is used. Using this library, all the users and their followings can be fetched for our data collection. There was a challenge of collecting the usernames of the user as Twitter allows a limited request for a particular time frame. So we fetched the userId instead of username and created a 1 minute limit between each fetch.

B)Dataset pre-processing

We fetched relationships of around 45573 users. For further analysis, we considered only those users who are linked with 3 other users. This was done to make sure that nodes that have less importance are removed. We found a total of 1774 users who has a link with at least 3 other users. We then created a network graph of these users using NetworkX.
NetworkX[10].

C)Classification techniques

During the process of creating a network graph, we assigned colors to each node based on their characteristics.

The blue color is assigned to the followings of the user who posted negatively about the Eid festival. The pink color is assigned to the followings of a left-wing party's politician. Orange is assigned to the followings of a politician of a right-wing party. Similarly, green is assigned to the followings of a user who posted derogatory about the Diwali festival.

The dark green color is assigned to the followings of the user who posted negatively about Eid as well as who is followed by the right-wing politician. Similarly, the dark blue color is assigned to the followings of the user who posted negatively about Diwali as well as who is followed by a left-wing politician.

Black and Red color are assigned to the following who are followed by at least 3 types of our chosen users[11]

4.2.2 Graphical representation methodology in Reddit

For analysis, we selected the same festivals which we used for the data collection on Twitter. We found the two most popular subReddits[12] which does discussion about India. One of the groups was *India* with around 543k followers and the other one was *IndiaSpeaks* 152K followers. Reddit has different ways of fetching data from it compared to Twitter. Reddit API named **PRAW**[13] is used to scrape the data from Reddit such as all posts which are posted by people, details of the author, and their karma score.

A)Dataset description Similar to the steps mentioned in subsection 4.1.2 we fetched the top 1000 posts based on their Karma Score. But here, instead of analyzing the posts, we represented the users who posted that post, and the group they followed as nodes and edges relationship.

B)Dataset pre-processing

The top 1000 posts were from few users, so we took the unique users from our dataset.

C)Classification techniques

We assigned a green color to the subReddit and its followers whose posts were mainly pro-government. The blue color is assigned to the group and its followers who posted mostly negatively about the present government.

Black color is assigned to the users who followed both pro-government and anti-government groups whereas red color is assigned to the subReddit whose followers share the same followers with pro-government subReddit.

5. Result

5.1 Machine learning analysis

5.1.1 Twitter using machine learning and text analysis

With just 10 labeled users and their 822 tweets. The accuracy of our SVM model came out to be 60.377 percent which

[7] By Twitter friendship network we mean the nodes and edges relationship between a user's following

[8] User's posting a content which is bashing a festival customs

[9] https://docs.tweepy.org/en/latest/

[10] https://github.com/networkx/networkx

[11] Total we have 3 types of users. i) a person who posted negatively about Diwali ii) a person who posted negatively about eid iii)a right-wing party politician iv) a left-wing party politician

[12] A subReddit is a niche group/community where people posts, share their ideas about a particular topic.

[13] https://praw.readthedocs.io/en/latest/

means around 60 percent of the tweets are classified correctly to the respective political party. While for Random Forests, the accuracy came out to be 84.9 percent which means 7 84.9 percent of tweets are classified correctly to the respective political party.

The confusion matrix for Random Forests can be referred from Figure 4.

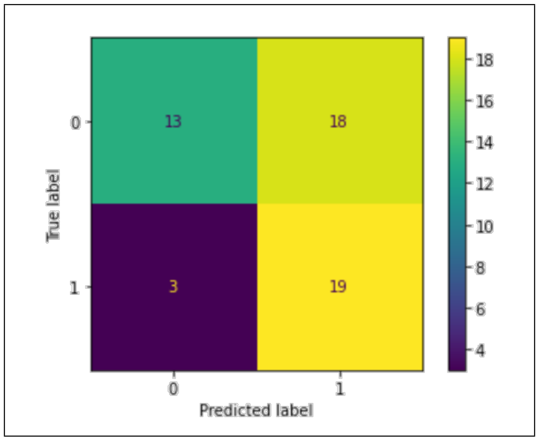

Figure 3: Confusion matrix for Random Forests model for Twitter

5.1.2 Reddit using machine learning and Text analysis

With 25 manually labeled users and their 4392 posts. The accuracy of our SVM model came out to be 50.58 percent. While for Random Forests, the accuracy came out to be 76.17 percent. Both of the model's accuracy dropped when comapred to the classification done on Twitter.

The confusion matrix for Random Forests for Reddit can be referred to from Figure 5 where "0" represents the left-wing party and "1" represents the right-wing party.

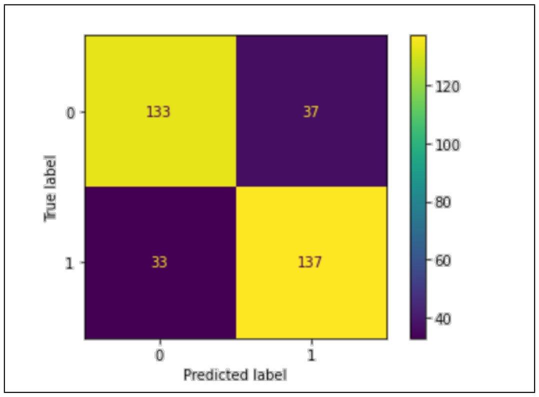

Figure 4: Confusion matrix for Random Forests model for Reddit

5.1.3 Twitter vs Reddit readability index

To compare the readability of the Twitter tweets with the subReddit posts we used ARI(Automated Readability Index) of the posts. Figure 5 represents the automated readability index of tweets for BJP and Left. The majority of the tweet's ARI score is between 7.5 to 12.5 which means that most of the tweets from right-wing and left-wing party supporters can be understood by at least a tenth-grade school student [14]

ARI = 4.71(characters/words) + 0.5(words/sentences)-21.43

The Automated readability index for Reddit can be observed from Figure 6. The majority of posts are in the range of 5-6 which means the Reddit posts are easily readable by a 5th-grade school student.

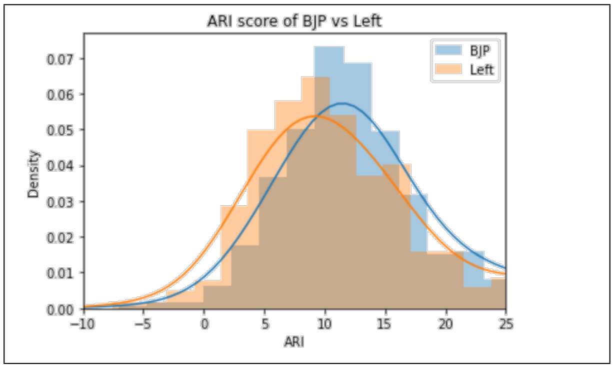

Figure 5: ARI score of the tweets on Twitter

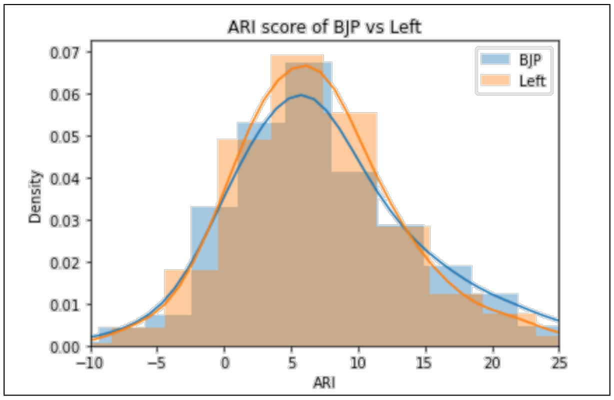

Figure 6: ARI score of the tweets on Reddit

5.2 Graph analysis

5.2.1 Twitter graph analysis

The Twitter graph from Figure 3 reveals that the users who post something offensive for a religious festival often tend

[14] https://en.wikipedia.org/wiki/Automated$_readability_index$

to have a pattern in terms of their political inclination. We can observe that the distance between the orange node and the green node is between either 1 or 2 whereas the distance between the orange and blue node is mostly observed to be more than 2. Similarly, the node distance between pink and blue nodes is either (1 or 2) whereas the distance between pink nodes and green is mostly observed to be more than 2. Had there been no pattern in the preferences of a user, the colors would have been in a mixed form. We came to the conclusion that in the case of the political scenario of India, some of the users who insult the religious custom of Eid are inclined towards the right-wing party. Whereas, some users who insult the religious custom of Diwali are inclined towards the left-wing party.

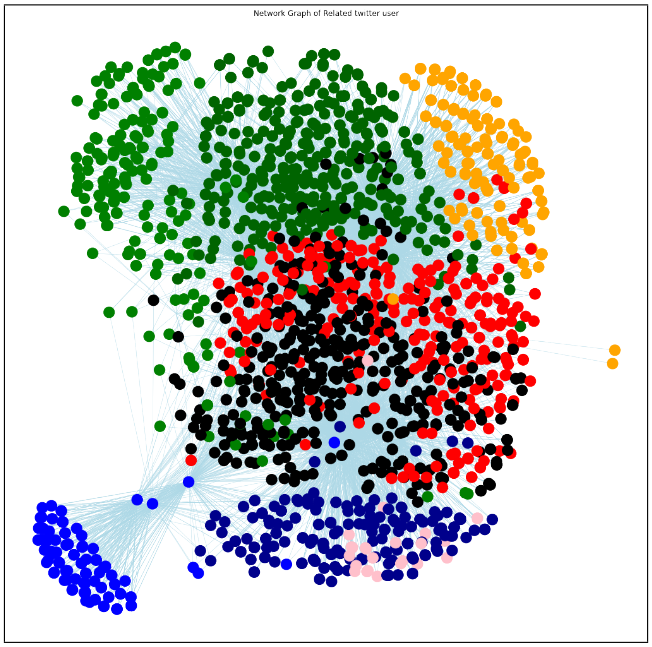

Figure 7: Network Graph of the followings of the users Political Parties included Orange Node - Right Wing Party, Pink Node -Left Wing Party

5.2 Reddit graph analysis

From the network representation of Reddit users, we can see from Figure 8 that people who follow both polarized groups were very few in number and the most active participation is being confined to the closed groups which have a different ideology. Unlike Twitter, there are some limitations in extracting the users information from subreddits. This is due to the fact that PRAW doesn't provide the functionality to see followers of a user. They only allow subReddits followed by a user to be seen.

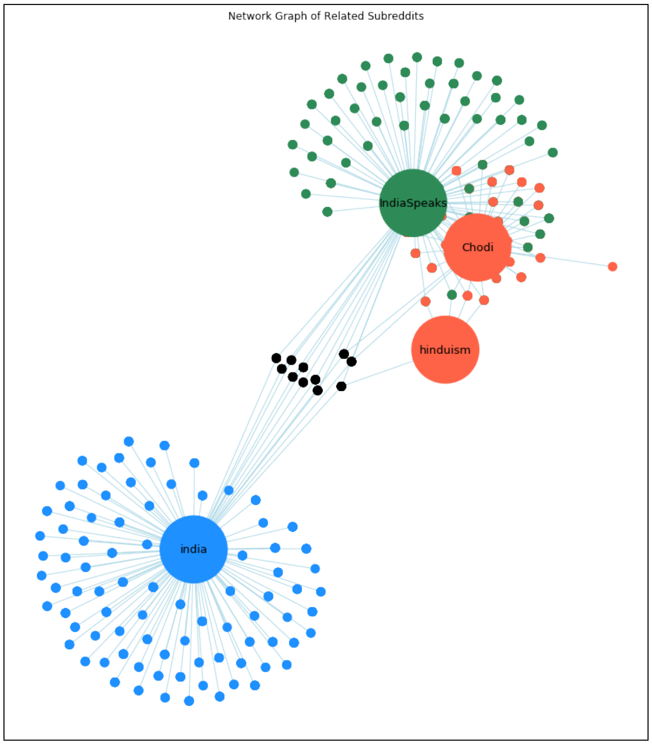

Figure 8: Network graph of Reddit users and the subReddits

6. Conclusion

In this research, we tried to find the political inclination of the user who posted about Indian festivals. Our analysis was mainly concentrated on the hateful comments observed during the festival in India. We observed that SVM is not an appropriate modeling algorithm for classifying text-related data for Reddit. As we observed, in the case of Reddit where there is no limit on the length of text, the model performed poorly whereas, Random Forests can be used to predict the political inclination of a user based on their tweets or post.

From the ARI score, it becomes evident that Reddit is more readable and understandable as compared to Twitter. This might be due to the fact that users express their views very clearly without worrying about the limit on their posts.

Using Graphical analysis, we can come to some conclusion that our initial hypothesis i.e. a person who posts hateful comments about a particular religion is oriented to a particular political ideology.

7. Acknowledgement

We sincerely thank Dr. Oul Han and Mrs.Isabelle Kuhlmann for helpful discussions that contributed to the development of this research work. This work is done in contact with the Institute of Web Science and Technologies.

References

[1] BRADSHAW, S., AND HOWARD, P. N. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project 1* (2018).

[2] EVOLVI, G. Hate in a tweet: Exploring internet-based islamophobic discourses. *Religions 9* (10 2018), 307.

[3] GRACIYAL, G. Freedom of expression in social media: A political perspective.

[4] HUTAIR, M. B., AGHBARI, Z. A., AND KAMEL, I. Social community detection based on node distance and interest. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* (New York, NY, USA, 2016), BDCAT '16, Association for Computing Machinery, p. 274–289.

[5] HUTAIR, M. B., KAMEL, I., AND AL AGBARI, Z. Social community detection based on node distance and interest. In *2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT)* (2016), pp. 274–279.

[6] PRATI, R., AND SAID HUNG, E. Predicting the ideological orientation during the spanish 24m elections in twitter using machine learning. *AI Society 34* (09 2019).

[7] SARKER, A., ZAMAN, M. S., AND SRIZON, A. Twitter data classification by applying and comparing multiple machine learning techniques. *SSRN Electronic Journal 7* (01 2019), 147–152.

[8] TYAGI, A., FIELD, A., LATHWAL, P., TSVETKOV, Y., AND CARLEY, K. M. A computational analysis of polarization on indian and pakistani social media. In *International Conference on Social Informatics* (2020), Springer, pp. 364–379.