# Investigating The Effect of Social Media Campaign on German 2017 Elections

**Arshad, Jani Anbarasi.L, Modigari Narendra, Pushbarani.S, Dhanya.D**

***Abstract:*** ***Social media usage has witnessed a big surge in recent years. Having started as a socialization platform at a very small scale, it has now helped itself into every domain possible. Of them include, advertising products, sharing news, driving businesses and much more. In this work, the influence of social media on the German 2017 Bundestag elections is examined and analyses were performed to show successful predictions. The 10 GB twitter dataset consists of more than 1,200,000 tweets which cover more than 120,000 users. The data set included members of the party and those users who retweeted their tweets inclusive of the tweets with "#BTW17" hashtag. Moreover, twitter API was used to separately get the list of followers of each member of the party. Using the above data set, graph modelling is done using Neo4j. User-user follow relationship and User-user retweet relationships are modelled for better analysis of the data. Various centrality measures were calculated to determine influential persons that drove the election campaign. The results are then compared with the ground-truth data to better understand whether social media has any significant effect on election results.***

## I. LITERATURE REVIEW

Social network analysis has always been a sought after area of research. With online socializing being on the increase day by day, a lot of information and data is being created. The amount of information is just not small and there have also been many problems created due to Big data. Moreover, such data also opens an opportunity for researchers in various fields. One such field is social network analysis.With social network analysis, one can gain huge insights into how users interact and socialize with each other. User's behavior can be captured with such social data. Liking a post, sharing a post, following a person, adding a friend to a network are all the ways through which users' interact with each other on social networks. It is through these ways – liking patterns, following patterns etc by which a social network analysis researcher can capture user behavior and the ways in which people interact with each other.

Although, socialization is healthy, there have also been many negative effects of the use of social media on people's lives. Of them, is people spreading and using aggressive language on social media. This very act is called aggression and it includes different types of aggression such as overt aggression, covert aggression, racial aggression, communal aggression, physical threat/aggression. One such work was done by Ritesh[1].

It includes a facebook and twitter dataset from Indian context, and using Natural Language processing, they build a classifier that could categorise a given text into one of the various types of aggressions they came up with.Another such work[2] analysed the interactions within and between extreme right communities in social media. There work included selecting known right community groups on twitter and analyzing their activity patterns on social media and answering questions like, are they differentor is there something unique that separates these people from other normal users of social media. They also visualized networks across countries and how right communities were clustered using a reciprocal-follower relationship graph. Their work also included the application of community detection methods and text processing to analyse the emotions attached to the most frequently used words. It was concluded that the existence of stable communities and associated topics was stronger in such groups.With continuing research in the same field of hatespeech, a paper titled "Surfacing contextual hate speech words within social media" [3] came with a novel way of detecting hatespeech by trying to address the difficulty of detecting hate speech when users intently try to use particular patterns in order to evade automatic detection and hatespeech filters.Another interesting research was of Ben David[4] which monitored the Facebook Pages of Extreme-Right Political Parties in Spain. They argue that hate speech and discriminatory practices are not only explained by users' motivations and actions, but are also formed by a network of ties between the platform's policy, its technological affordances, and the communicative acts of its users. They conclude that in the case of the Spanish extreme-right political parties, Facebook hosts an increasing volume of covert discriminatory practices that not only circulate data and content, but also trigger overt hate speech by the parties' followers.

A more recent interesting work titled "Identifying Key-Players in Online Activist Groups on the Facebook Social Network" [5] actually used a lot of social network analysis techniques to identify the most influential users in the spread of information on social media. Their work was focussed on select UK-based activist groups from facebook.

**Revised Manuscript Received on 30 May 2019.**
***** Correspondence Author

**Arshad**, VIT University Chennai (Tamil Nadu), India.
**Jani Anbarasi,** VIT University Chennai (Tamil Nadu), India.
**Modigari Narendra,** VFSTR Deemed to be University, Guntur (Andhra Pradesh), India.
**Pushbarani S,** Meenakshi College of Engineering, Chennai (Tamil Nadu), India.
**Dhanya D,** Mar Ephraem College of Engineering and Technology, Marthandam (Tamil Nadu), India.

They modelled a facebook comments graph and determined degree centrality and the frequency of communication between nodes depicted by edge thickness.

Moreover, a user-post graph was also included. It was concluded that the results arrived at actually matched with the ground-truth.

Excluding hatespeech, aggression, another work included analysing political campaign of the General Election 2017 of the UK. [6] They analysed how twitter posts varied over time and how they increased as the election drew closer.

Moreover, hashtags and most mentioned were also analysed. The topics most associated with a particular politician were also put forth. The paper then ended with the conclusion that although twitter is not representative of the voting public, it can be used to gauge the mood of those who are motivated enough to comment on social media.

Data-Driven political campaigns were successful for Obama 2012 campaign whereas after four years it was not successful for Hillary Clinton. Both campaign targeted users based on the sociodemographic and psychographic data. But it was interesting to find Trump campaign was not only to mobilize his supporters but also to demobilize Hillary's Supporters. So these insights of the mined data can be used for data analysis and can understand various factors like

a) Influencers in the political crowd
b) Are they influential
c) Are these influencers aware or unaware in the political research
d) How robust these are against demobilizing effect.
e) Identification of social trends
f) Measuring the feeling of being politically penalized etc.

Twitter provides samples of these data using streaming API's. Barberá and Rivero emphasize "the opportunities offered by Twitterfor the analysis of public opinion: messages are exchanged by numerous users in a public forum andthey may contain valuable information about individual preferences and reactions to different politicalevents in an environment that is fully accessible to the researcher" [9].Twitter data sets has been studied for various events like a)Understanding the political representativenessrepresentativeness [10 ], b)Updates on the Live TV Context [11], c)Special Relationship between the tweets and the votes [12 ] etc. The dataset collected for German campaign of 2017 published by [7] is analysed using social network analysis to predict the success of Tweets.

## II. PROPOSED WORK

### 2.1 Description of dataset

1. Twitter API was used to gather follower accounts of party members of the various parties.
2. This was done by first submitting a review to Twitter and after much interrogations, permission was granted for the use of Twitter API
3. A python program was written for the same using the Tweepy library as an interface between the programmer and the twitter API
4. For each party, a separate CSV file is maintained that contains a list of members and their followers

### 2.2 Data Preprocessing

The downloaded dataset was a whopping 10 GB dataset that could not be directly loaded by any program for any kind of processing. The dataset comprises more than 1,200,000 tweets from 120,000 users. The dataset contains names of the politicians and their party members, recorded tweets, user information, user mentions, hashtags, media, reply to tweets, retweets, timestamps for all status.

1. This is because, normal PCs are not powerful enough to handle such Big data.
2. One solution could have been to use a distributed environment for storing data, however, that would bring me some unnecessary complexities in the process and would make us loose track from our aim.
3. Therefore, another solution was required.
4. The 10 GB dataset consisted of a 1000 JSON files. It would have become an headache to manually load all the JSON files into the database.
5. Therefore, as a primary step to reduce processing time, all the 1000 JSON files were first minified. This required a python program to be written to do the same.
6. The next step was to reduce the no. of JSON files such that they could be manually uploaded to the database easily and swiftly.
7. This required another python program to iterate through a folder and concatenate JSON files in batches of 100. Essentially, all the files could have been concatenated to one, but a huge RAM would be needed to actually load such a huge file into the database directly.
8. Therefore, a total of 14 JSON files were created from the around 1000 JSON files.
9. Moreover, a couple of invalid JSON files had to be removed or thrown out.
10. After having gone through all the above steps, the JSON files which contained tweet objects were now ready to be imported into the database.

### 2.3 Data Importing and Graph Modelling

1. First, a followed-accounts JSON file was imported and a "MEMBER_OF" relationship was created between the party member and the party itself
2. Next, multiple CSV files are imported that contains members and their follower Ids. A relationship called "FOLLOWS" is created between the follower and the member.
3. Tweet objects are also imported from minified JSON files.
4. Using the tweet object a user-posts-tweet, tweet-retweet-tweet relationships are modeled

Graph modeling was performed to show the relationship between the member and the party, user and the user, user and the user retweet, 2.3.1 Member – Party

This graph is modelled by the relationship "MEMBER_OF". It relates a member of a party to the party. Visualisation of this graph depicts the varied connections between various parties.

Some parties have more members than the others and some members have more followers than the others.

The no. of followers for each member of a party could show how dominating a party is.

2.3.2 User – User follow

This graph is modelled by the relationship "FOLLOWS". If one user has followed the account of another user, then a "FOLLOWS" relationship is created. Using this graph, we can identify party members with the highest follower count. The most important members across different parties can be compared.

2.3.3 User – User retweet graph

This graph is modelled by the relationship "RETWEETS". This is an inferred relationship from the graph. A "RETWEETS" relationship is created between two users when a user retweets the tweet shared/posted by another user. Using this graph, we can detect amplification of information. A user retweeting another user means that that message is getting amplified and reaching a wider audience. Moreover, we can also try to investigate whether a specific tweet is getting amplified and if they are from the same community.

Similarly centrality measures were identified to generate the dominant user and the party.

1. Centrality measures

a. Page rank

Page rank algorithm measures how important a given node is in a graph. Nodes with higher in-degree are ranked very high.

b. Closeness centrality

Closeness centrality detects nodes that are able to spread information very efficiently through a graph. The closer or the shorter the path a node has to any other node, higher will be the closeness centrality.

c. Betweenness centrality

Betweenness centrality is the measure of how often a node has to pass through along the shortest path to reach any other node. It detects the amount of influence a node has over the flow of information in a graph. It is often used to find nodes that serve as a bridge from one part of a graph to another.

2. Community detection.

a. Label propagation algorithm

Every node is initialized with a unique label (anientifier). These labels propagate through the network.At every iteration of propagation, each node updates its label to the one that the maximum numbers of its neighbours belong to. Ties are brokenuniformly and randomly.LPA reaches convergence when each node has the majority label of its neighbours.

b. Clustering coefficient algorithm

Triangle counting is a community detection graph algorithm that is used to determine the number of triangles passing through each node in the graph. A triangle is a set of three nodes, where each node has a relationship to all other nodes. This can be used to determine how closely related are a group of nodes.Applicability of various Social Network Analysis methods in context of the Political dataset

1. Page rank

Page rank can be used to detect key players in the network that can easily spread messages. Higher page rank would usually mean that the given member of a Political party has more followers and hence the information that is being conveyed can easily reach out to a larger audience and can also have a greater impact.

2. Label propagation algorithm

This algorithm is used to detect communities. This is run on both the follower relationship as well as the retweet relationship. With this algorithm, the whole political dataset can be segregated and partitioned into different communities. The result can then be evaluated against the ground truth – whether the real-life party groups fall under the same community/partition from the algorithm.

3. Harmonic centrality

Harmonic centrality can be used to detect people who can readily spread information to a wider group of communities. Their information is not likely just to say within their community, but it is likely to spread across communities given that they have a higher degree of closeness centrality. This is particularly useful when parties want their ideas to reach a variety of audience.

4. Average clustering coefficient

After having detected communities, a way to evaluate how closely knit are those communities can be done by calculating the average clustering coefficient for each of those respective communities. Using this information, one can gather which political party was more connected within themselves and how this high clustering within their party could probably lead to more people indirectly getting involved in their social media campaigns.

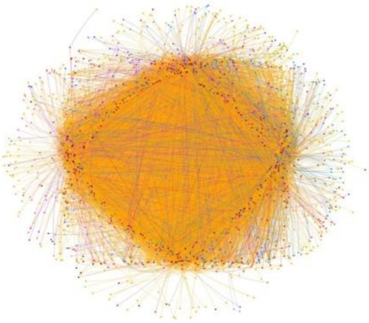
5. Betweenness centrality

Betweenness centrality can give an insight into which users are better at connecting two different communities or groups. In a political dataset context, this could enable insights into users who can connect different political parties. This is particularly useful for parties who would like to spread their views/posts to people of other parties so that they could turn other party followers into their followers.

## 3. Results and Discussion

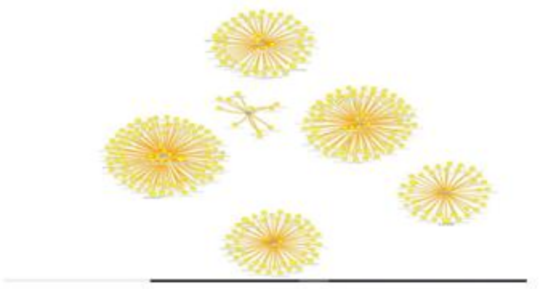
Social network analysis does give some insight into this dataset. The analysis showed that the amount of influence each party member has was different with various parties. Applying different analysis resulted in different parties topping the list. The common similarity that was observed amongst the more influential users is that they were mostly from a particular place. Hence, the membership of a party did not seem to explicitly act as a differentiator, but the geographic location did.Nonetheless, a key insight was captured by time analysis of the no. of tweets leading into the election. Except for the month of July, there is gradual increase in the no. of tweets from May-September. This shows how parties started campaigning more vigorously leading into the election.Moreover, doing a time analysis of the number of twitter accounts created show that the month of July witnessed the largest number of newly created twitter accounts. Just a month leading into the election! There is a possibility that the new accounts might have been created to further amplify or voice a party's campaign.

### 3.1Graph modeling



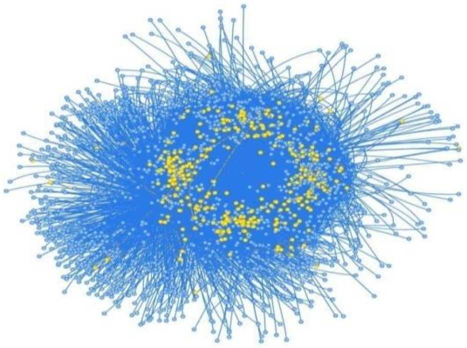
a)Member-party graph

The above graph depicts 6 parties and their party-members.
b)User-User Follow graph

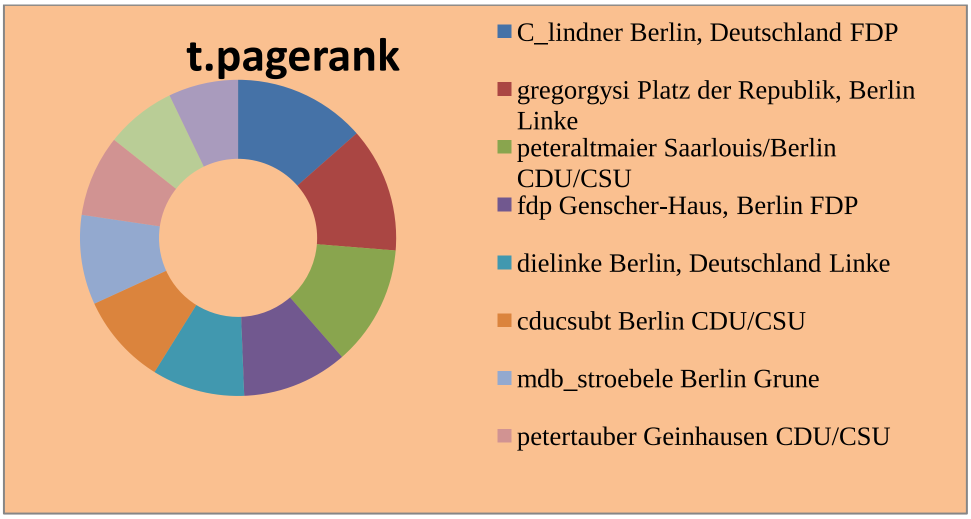


The above graph depicts the follows relationship. Two nodes(users) are connected if one user follows the other. Moreover, the relationship becomes bidirectional if both follow each other.

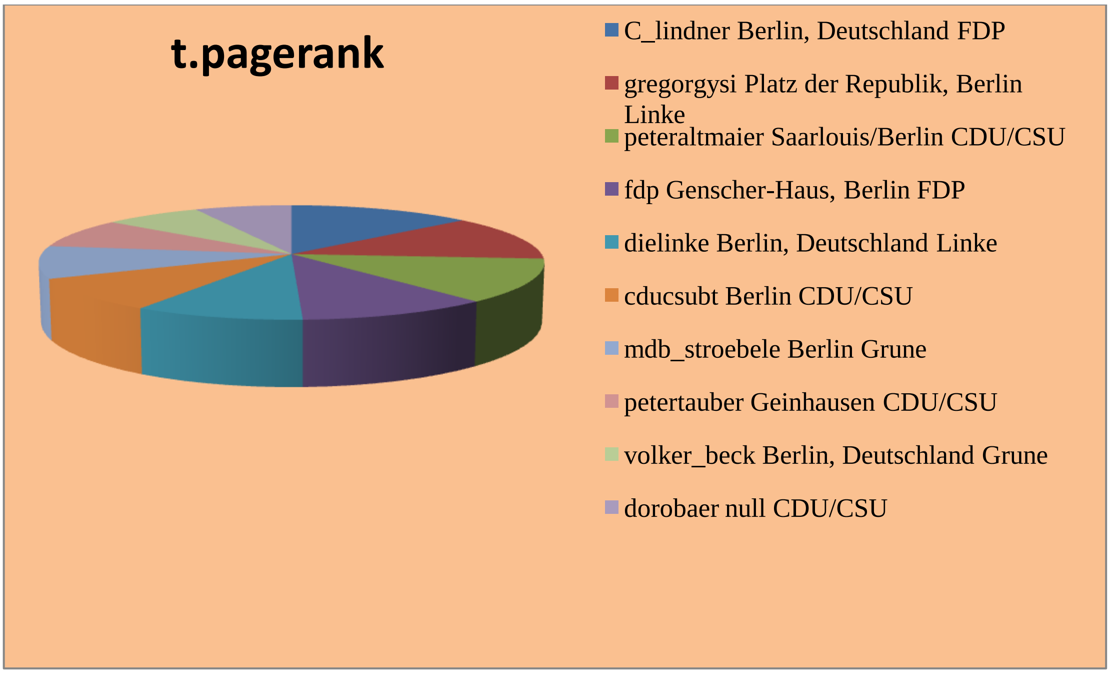
c)User-User Retweet graph



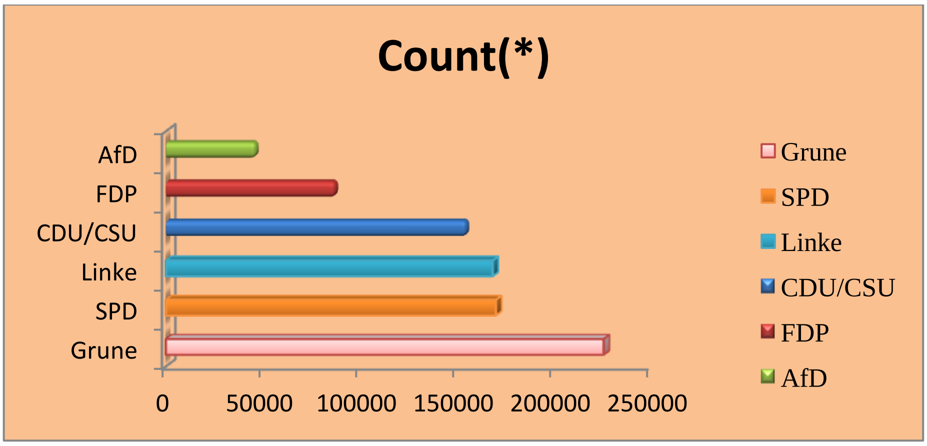
The above graph depicts the retweets relationship. Two nodes(users) are connected if one user has retweeted any of the posts of another user. Moreover, the strength of the relationship is denoted by the no.of times a user has retweeted content posted by another. This is shows by the thickening of edge connecting the two users.Based on the Page rank,Label propagation,Harmonic centrality and the Betweenness centrality few insights that has been achieved is given astop 10 users who have the most number of follower



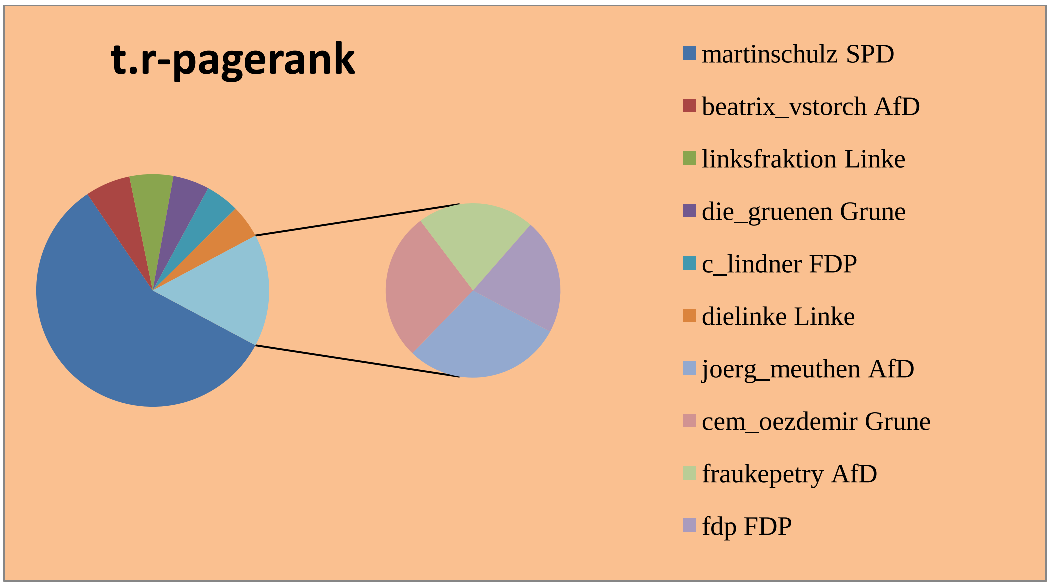
d)Location commonalities between the users who were followed the most



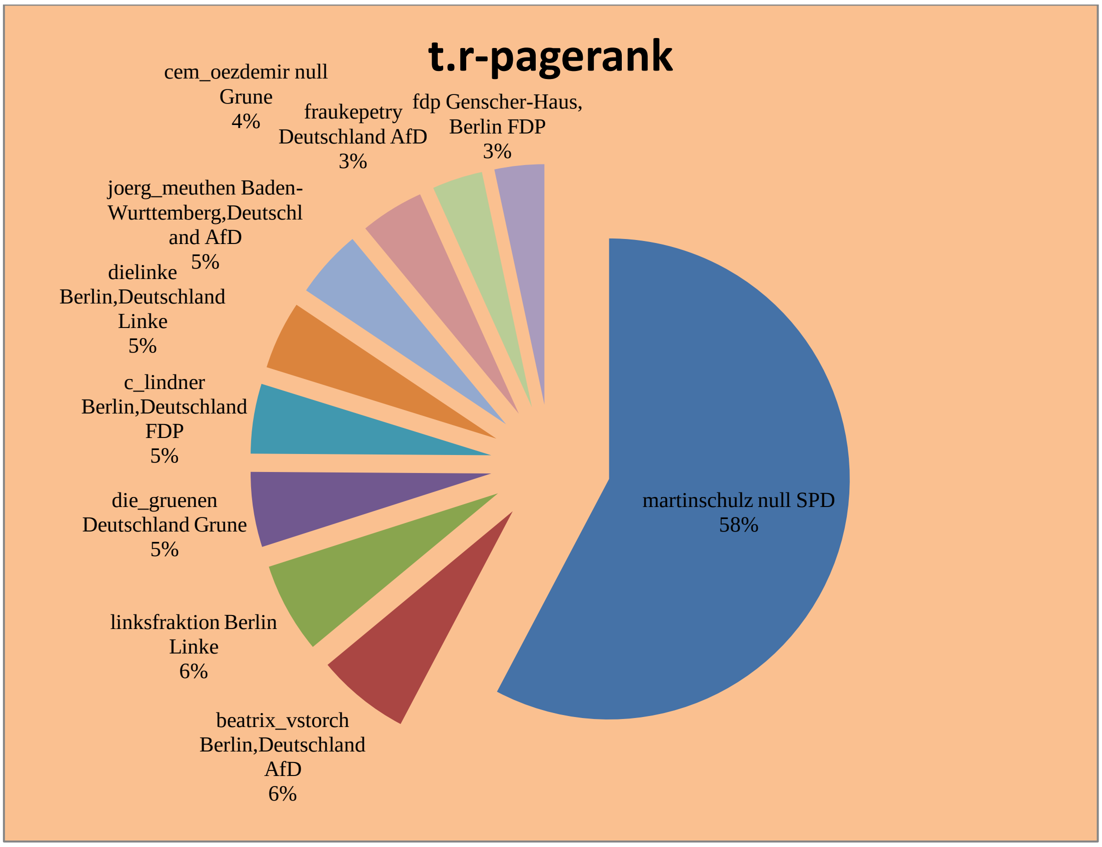
e)The total number of followers who are directly/indirectly related to each party



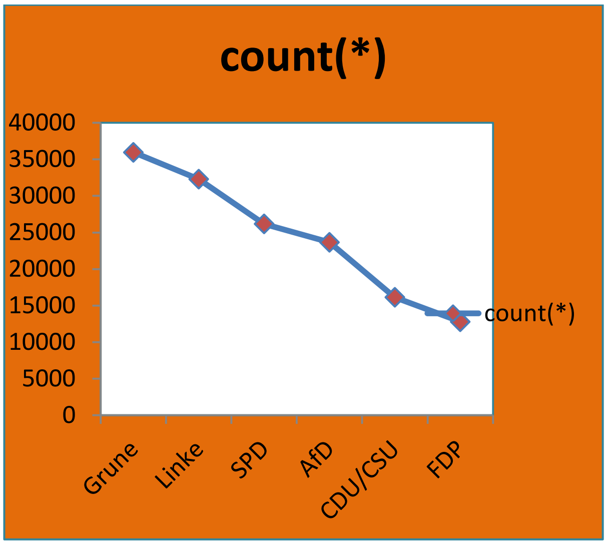
f)Get top 10 users whose posts were retweeted the most



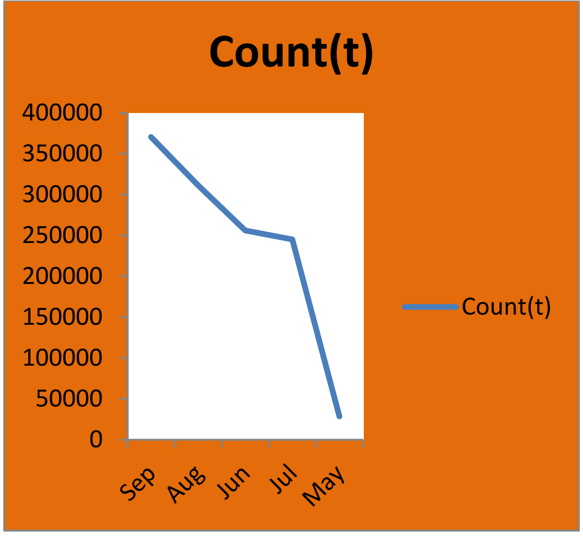
g) Location commonality between the users whose posts were retweeted the post



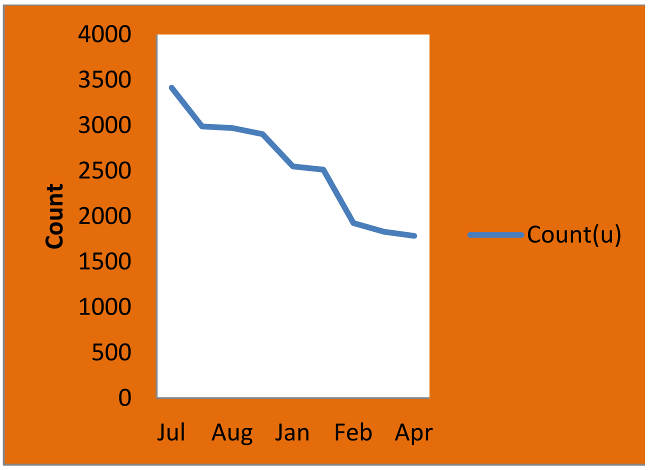
h). Get the count of the no. of retweet relationships for each community/party



i) Compare the no. of tweets generated in the last 5 months leading into the election – also group by party



h. Compare the no. of newly created twitter accounts every month from some months prior to election till the election date.



## IV. CONCLUSION AND FUTURE WORK

To conclude, it could be said that, although twitter data might not be able to predict election results to a high degree of accuracy, it could still aid in deriving overall conclusions on the influences of various users on the elections. Nonetheless, more advanced techniques on data such as natural language processing might aid in giving a better idea about the direction of an election.

## REFERENCES

1. Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, TusharMaheshwari, “Aggression-annotated Corpus of Hindi-English Code-mixed Data”, 11th edition of the Language Resources and Evaluation Conference (LREC - 2018), 7-12 May 2018, Miyazaki, 2018
2. Derek O’Callaghan, Derek Greene, Maura Conway, Joe Carthy, P´adraig Cunningham, “Uncovering the Wider Structure of Extreme Right Communities Spanning Popular Online Networks”,Proceedings of the 5th Annual ACM Web Science Conference, 2014
3. Cherries Taylor, Melvyn Peignon, Yi-Shin Chen, “Surfacing contextual hate speech words within social media”,CoRR,2017.
4. ANAT BEN-DAVID1, ARIADNA MATAMOROS-FERNÁNDEZ, “ Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain”,InternationalJounral of Communication 10, 1167-1193, 2016
5. Mariam Nouh, Jason R. C. Nurse, “Identifying Key-Players in Online Activist Groups on the Facebook Social Network “,IEEE 15th International Conference on Data Mining Workshops, 2015
6. Laura Cram, Clare Llewellyn, Robin Hill, “ UK General Election 2017: a Twitter Analysis “WalidMagdy, 2017.
7. NaneKratzke” The #BTW17 Twitter Dataset–Recorded Tweets of the Federal Election Campaigns of 2017 for the 19th German Bundestag” Center of Excellence for Communication, Systems and Applications (CoSA), Lübeck University of Applied Sciences, 23562 Lübeck, Germany, 2017
8. PriteeSalunkhe, SachinDeshmukh, International Research Journal of Engineering and Technology (IRJET), Volume 4, 2017
9. Barberá, P.; Rivero, G. Understanding the Political Representativeness of Twitter Users. Soc. Sci. Comput. Rev.2015, 33, 712–729.
10. Barberá, P.; Rivero, G. Understanding the Political Representativeness of Twitter Users. Soc. Sci. Comput. Rev. 2015, 33, 712–729.
11. Abreu, J.; Almeida, P.; Silva, T. From Live TV Events to Twitter Status Updates—A Study on Delays. In Applications and Usability of Interactive TV; Springer International Publishing: Cham, Switzerland, 2016; Volume 605, pp. 105–120.
12. Jungherr, A. Tweets and Votes, a Special Relationship: The 2009 Federal Election in Germany. In Proceedings of the 2nd Workshop on Politics, Elections and Data, PLEAD ’13, San Francisco, CA, USA, 28 October 2013; pp. 5–14.