Prediction of Celebrity Influence during election propaganda using Twitter Sentiment Analysis

Karan Mulage, Rishabh Bassi, Nidhi Vanjare, Amruta Koshe

***Abstract* - With every activity being planned online, data on social networks is increasing many folds. Polarization, Hype, and Influence of these social networks are quite evident during the Election Periods. Seeing the previous elections worldwide, it is clear that the influence and effect caused by these media are powerful more than ever. With some key personalities taking up the positions, they use their followers' mindset or bias to create factions ultimately leading to worldwide influence within seconds. The good thing is due to advancements in Machine Learning and Natural Language Processing, this influence can be measured which can lead to better election propaganda and even maintaining sanity. To demonstrate this, we take the twitter data of various celebrities active during the 2020 US Elections and predict the influence they were able to generate in public. For this, we use Natural Language Processing techniques for the categorization of tweets and then devise a formula for influence prediction using statistics. Our model is able to predict which celebrity was most influential for a particular party during elections. Having this information on hand can smooth-line the entire election propaganda and be useful for different factions to plan their future as well.**

Keywords - Twitter Sentiment Analysis, Machine Learning, NLTK, Web Scraping, Social Influence.

I. INTRODUCTION

Ever since Natural Language Processing has given us the field of Sentiment Analysis, researchers have been utilizing this field's potential and trying to solve real world problems. Extracting people's opinion and mindset while they are writing something-be it a review, a tweet or a suggestion is in itself the solution to many problems. Coming to feasibility point of view, hype of social media over the past years has given a tremendous push to this field. With everything coming up online, we have more than enough data of customers worldwide that we can do sentiment analysis on any topic and solve yet another problem. Sentiment analysis has numerous uses in various fields, such as segment analysis, point of sale review in businesses to obtain feedback for products and in social media to learn about people' comments and reviews. There have been various past researches in the field of Sentiment Analysis to understand crowd reaction during disaster [1] and even extracting information from Twitter or Facebook Data. The method to sentiment analysis has two conceptual frameworks: Rue Based Approach and Machine Learning or Statistical approach. The rule-based method tries to assign relative scores to a complete vocabulary made up of frequently used terms and then add rules based on the circumstance. The Machine Learning or Statistical Approach to this classification problem, on the other hand, is based on a statistical analysis of text similarities and differences for each group. The basic premise of this approach is that "you shall know a word by the company it keeps", which implies that a comparable text will have a similar collection of terms. Because this method employs a Machine Learning algorithm, it needs a big labelled input dataset in order to attain enough accuracy. With enough labelled data, a classifier may be trained to categorize each incoming sentence into a discrete label class, such as "positive", "neutral" or "negative". In this paper, we have identified the problem of predicting celebrities' influence during election propaganda. Especially seeing 2020 US elections and support that celebrities were giving to different parties had a huge role in election results. So we believe if we could find a way that can predict which celebrities caused a greater influence in a particular party, then this information can be of great significance during elections.

Social network analysis is the study of people's interactions and communications on many themes, and it is gaining popularity these days. On social media platforms like Facebook and Twitter, millions of individuals express their thoughts on a regular basis on a variety of issues. It has several applications in a variety of fields of study, ranging from social science to elections. Twitter is now one of the most popular social networking platforms, with over 200 million accounts and around 200 billion tweets per year. For this reason, we found Twitter to be the best fit for our dataset creation as there are over 3,00,000 tweets with US elections 2020 label just during a couple of months around the election dates. Twitter is also considered as a rich resource for sentimental analysis. Some challenges which we encountered were as follows: Some of tweets contained only images which failed to do SA on them and since social networks, especially Twitter, contain small texts and people may use different words and abbreviations, it became difficult to extract their sentiment. In the following sections, we show the high-level abstract of our implementation. We will show data collection and preprocessing of tweets and the methods we have implemented to do sentiment analysis on dataset and also devised a formula to take into account likes and retweets for a particular tweet.

II. LITERATURE REVIEW

This paper is motivated by a paper published by Mohammed and Sunjana [9] in 2019. It talks about the political polarization in the 2019 Indonesian election using sentiment analysis and social network analysis. It uses the naive Bayes sentiment analysis technique to identify the sentiments and classify them as positive, negative and neutral. And then they do the social network analysis based on the calculation of network attributes to arrive at a graph

with nodes, edges, average degree, network diameter, etc. Their results show political leader's accounts which were most popular among others.

Due to the increase in research in the area of sentiment analysis and opinion mining, more and more of its applications are being uncovered recently. These applications span a wide variety of domains like health, education, stock market, business, politics, etc. A paper published on people's perception on Online learning is very fascinating [10]. Research like this, can pave the way towards testing newly introduced products or services by gaining opinion on a huge volume of people. With more and more efficient methods to do sentiment analysis, all our results will also become more accurate.

It is also worth noting that several methods of doing sentiment analysis work even with more accuracy when two or more methods are combined. For example, in the paper Sentiment Analysis of Twitter Data, Agarwal et al. [11], they show how the Kernel plus Senti-feature model outperforms their baseline model, Unigram model. Similarly, they show Unigram plus Senti-feature outperforms Unigram model.

III. Design

With the purpose of finding the amount of influence that celebrities had (whether negative, positive or neutral) on the general public during the US 2020 elections, we have searched and selected 5 different celebrities for each contesting side to be considered for our analysis. These celebrities were active on twitter during the elections and have publicly endorsed either Donald Trump or Joe Biden, making them viable cases for our study.

Celebrities considered for our case study -

1. In support of Donald J Trump

   - Conor McGregor, Jack Nicklaus, Kristie Alley, James Woods, Lil Wayne.

2. In support of Joe Biden

   - Dwayne Johnson, Taylor Swift, Mark Cuban, John Legend, Jennifer Hudson.

Following the selection of five celebrities for each candidate, we extracted the tweets publicly posted by these celebrities wherein they mention the US 2020 elections and their beliefs about any candidate i.e. Donald Trump or Joe Biden. To find the influence of any celebrity on the general public regarding the elections, we need to analyze the sentiments of all comments on that tweet and find out whether they are positive, negative or neutral. Apart from the sentiment of the comment, we need to include the number of likes and retweets on the celebrity's original tweet as well to get an understanding of the degree to which the public agrees/ disagrees with the tweet. The number of likes, retweets and the comments on the tweet were all extracted using Octoparse tool and 4 different pre-trained models (Flair, NLTK, Textblob and XG_Boost) were used to find sentiments of the comments posted on the tweets. The total number of positive, neutral and negative comments for each tweet were gathered and used to compute the total influence of the tweet.

To compute the amount of influence each tweet had, the following formula was used -

$$I.F = R + L + (N*(p-n)) \quad (1)$$

Where,

I.F = Influence factor

R = number of retweets

L = number of likes

N = number of interactions

p = number of positive comments

n = number of negative comments

(p-n) gives the overall public sentiment of the selected tweet. If this is a positive number, it can be inferred that the users agree with the tweet and hence hold a positive opinion towards it, and vice versa. N is an important factor to estimate the degree of influence. Greater user interactions with the tweet counts towards greater influence of the celebrity on the public. Hence, this value is multiplied with the overall sentiment i.e. (p-n). (R+L) denotes the summation of the number of retweets and likes on the tweet. Higher the sum, greater the influence of the tweet. This can be considered as a positive influence since likes and retweets account for a positive opinion of the tweet. Higher the value of I.F, greater is the positive influence.

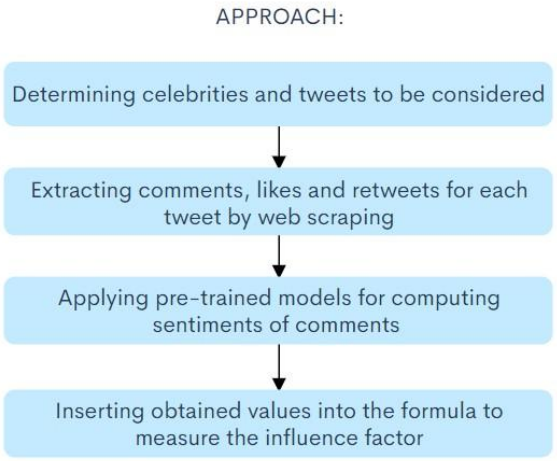

Fig. 1. Research Process

IV. Data pre-processing

A. Data Scraping

Scraping is the act of extracting data or information from websites with or without the consent of the website owner. Scraping can be done manually, but in most cases it's done automatically because of its efficiency. There are many data scraping or web scraping tools available for data extraction like Scrapingbee, Octoparse, Scraping-Bot, Scrapestack etc, we selected Octoparse for our purpose. Octoparse is a tool to collect data easily from social media sites by creating conceptual models and then executing the tasks either in local systems or in the cloud. And then exporting the extracted data in required formats like Excel, CSV, JSON

and HTML or directly exporting them to databases like MySql or SqlServer. This scraping tool will load the URLs given by the users and render the entire website. As a result, we can extract any web data with simple point-and-click and file in a feasible format into our computer without coding. We used a free version of this tool to build and run data extraction models in our local systems.

URLs of the tweets posted by different celebrities in support or endorsement of a candidate in the 2020 US election are considered. From the webpage of these tweets, we iteratively select and extract text of the tweet replies that are posted for a particular tweet by a celebrity, as shown in Fig 2. We were able to consider tweets which were posted publicly as some user's posts were private, we could not consider all of them.

Fig. 2. Data Extraction Model used for data scraping from Twitter Webpage.

B. Data filtering

The text extracted from the tweets is not practical to be used directly for processing because of the fact that these texts include special characters like :, $, #, @, emoticons etc. These texts can also contain URLs or links which do not add anything to the sentiment of the tweet. Also to note that, many tweet replies were just images or short videos or in graphics interchange formats(GIFs), which makes sentiment analysis even more challenging in identifying the opinion of the tweet. We considered tweet replies which only had text content and then filtered that text further by removing the special characters or emoticons. After all this filter, our dataset had many empty rows, which had to be removed. Finally we ended up with a dataset of 1113 tweet replies for our analysis.

V. IMPLEMENTATION

We have used Octoparse to extract the replies on the Tweets of the celebrities as mentioned earlier. The extracted tweets were saved in excel sheets. The platform used for the sentimental analysis is Google Collab. The entire code for this research is available in github repository [8].
The main motive of the analysis is to rank the influence of celebrities on elections via twitter. Aspects like number of likes, retweets, replies etc. on a particular tweet are taken as features which signify the popularity of that tweet. Sentiments for each reply on the tweet were calculated and substituted in place of p and n in the influence formula.

There were total 4 pretrained models used

1) FLAIR

Flair delivers good performance in solving NLP problems such as named entity recognition (NER), part-of-speech tagging (PoS), sense disambiguation and text classification [7]. It's a NLP framework built on top of PyTorch. Flair's sentiment classifier is based on a character-level Long Short Term Memory (LSTM) neural network which takes sequences of letters and words into account when predicting. Flair supports a number of word embeddings used to perform NLP tasks such as FastText, ELMo, GloVe, BERT and its variants, XLM, and Byte Pair Embeddings including Flair Embedding. Using the text classifier, flair predicts the sentiment that is 1 if positive sentiment and 0 if negative sentiment.

2) NLTK

NLTK is a natural language processing tool kit. NLTK's Vader sentiment analysis tool uses a bag of words approach with some simple heuristics [6]. Bag of words is a method of feature extraction with text data. The advantage of this approach is that sentences containing negated positive words (e.g. "not happy", "not good") will still receive a negative sentence sentiment because of the heuristics to flip the sentiment of the word following a negation. The disadvantage of this approach is that Out of Vocab (OOV) words that the sentiment analysis tool has not seen before will not be classified as positive or negative (e.g. typos) [5].

3) TextBlob

TextBlob Sentiment Analysis works in a similar way to NLTK — using a bag of words classifier, but the advantage is that it includes Subjectivity Analysis. TextBlob returns polarity and subjectivity of a sentence.
Sentiment on the reply for the tweet is calculated by the polarity in textblob. TextBlob returns polarity and subjectivity of a sentence. Polarity of textblob sentiment analysis lies between -1 and 1. -1 defines a negative sentiment and 1 defines a positive sentiment. If polarity is 0 then the sentiment is neutral. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

4) Xg_boost Algorithm

XGboost is scalable , which drives fast learning through parallel and distributed computing and offers efficient memory usage. XGBoost is an implementation of gradient boosted decision trees designed for performance [4].

It is an ensemble tree method that applies the principle of boosting weak learners using the gradient descent architecture. However, XGBoost improves upon the base Gradient Boosting Machine (GBM) framework through systems optimization and algorithmic enhancements.

XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems. It is a faster algorithm when compared to other algorithms because of its parallel and distributed computing. It is developed with both deep considerations in terms of systems optimization and principles in machine learning. The goal of this library is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate library. For example, the tweet of Dwayne Johnson in Fig 3.

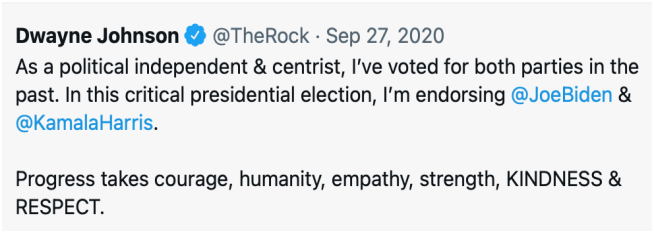

Fig. 3. Dwyane Johns Tweet in support of Biden

This was the tweet for which we scrapped and got 198 replies. After cleaning that data we got 161 replies. After passing all the replies to the tweet in flair classified 91 of these tweets to have positive sentiment and 70 of the rest to have negative sentiment. Total number of likes on the tweet was 372700 and the total number of retweets were 75800.

Formula for calculating the Influence:
Influence_DJ = ((Number of tweets considered * (positive-negative)) + (total_likes+Retweet))

After substituting all the numbers in the formula explained above, we get Dwayne Johnson's influence score with flair model is 451881.Similar method is followed by all the other methods by counting the positive and negative sentiment of each reply on the tweet and then substituting it in the formula for influence score with different models.

For Textblob and NLTK models we have omitted the neutral tweets as they do not contribute to any influence of the celebrity.

VI. RESULTS

Table 1. Sentiment classification of different models (supporting Mr. Biden)

Celebrities	FLAIR		NLTK			Textblob			Xgboost	
	Positive	Negative	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative
Taylor Swift	117	47	27	0	137	95	23	46	160	4
Dwyane Johnson	91	70	13	7	141	74	32	55	148	13
John Legend	67	34	15	1	85	39	10	52	98	3
Mark Cuban	58	67	7	4	114	46	25	54	124	1
Jennifer Hudson	21	6	5	0	22	13	3	11	27	0

Table 2. Sentiment classification of different models(supporting Mr. Trump)

Celebrities	FLAIR		NLTK			Textblob			Xgboost	
	Positive	Negative	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative
Lil Wayne	45	4	12	12	25	2	0	47	49	0
Kristie Alley	59	97	8	3	145	75	35	46	144	12
Conor Mcgregor	42	3	18	18	9	1	0	44	45	0
James Woods	62	73	9	3	123	67	32	36	132	3
Jack Nicklaus	66	84	4	5	141	61	43	46	144	6

We have used four natural language processing pre-trained models namely Flair, NLTK (Natural Language Toolkit), Textblob and XGboost (Algorithma). This was mainly to compare the outcomes of all the four models and how they classify celebrities using the formula (1) based on the influence they were able to create. All the four methods classify the celebrities in consideration in almost the same order. Therefore we can conclude that the prediction is accurate.

Since we considered 5 celebrities for each candidate (Mr. Biden and Mr. Trump), the sentiment classifications by the models are as shown in tables [1] and [2] respectively. The celebrity order when sorted based on influence factor is shown in tables [3] and [4]. It is evident from the results that Taylor Swift proves to be the most influential celebrity in campaigning for Mr. Biden, taking the first place. Second highest influencer is Dwayne Johnson, as classified by all the models. Similarly John Legend is third, Mark Cuban is fourth and Jennifer Hudson is fifth in terms of overall influence in support for Mr. Biden.

It is also observable that Mr. Lil Wayne has the greatest influence among the celebrities considered in support of Mr. Donald Trump. This is followed by Ms.Kristie Alley in second place, James Woods in third , Mr. Conor Mcgregor in fourth and Mr. Jack Nicklaus, with the lowest influence score, in fifth place.

Table 3. Influence score of celebrities for Mr. Biden in order

Influence score of celebrities for Mr. Biden							
Flair		NKTL		Text blob		Xgboost	
Celebrity	Score	Celebrity	Score	Celebrity	Score	Celebrity	Score
Taylor Swift	467480	Taylor Swift	460428	Taylor Swift	467808	Taylor Swift	481584
Dwyane Johnson	451881	Dwyane Johnson	449466	Dwyane Johnson	455423	Dwyane Johnson	470235
John Legend	15202	John Legend	13283	John Legend	14798	John Legend	21464
Mark Cuban	9799	Mark Cuban	12004	Mark Cuban	13549	Mark Cuban	26299
Jennifer Hudson	2974	Jennifer Hudson	11299	Jennifer Hudson	12139	Jennifer Hudson	12598

Table 4. Influence score of celebrities for Mr. Trump in order

Influence score of celebrities for Mr. Trump							
Flair		NKTL		Text blob		Xgboost	
Celebrity	Score	Celebrity	Score	Celebrity	Score	Celebrity	Score
Lil Wayne	654809	Lil Wayne	652800	Lil Wayne	652898	Lil Wayne	655201
Kristie Alley	222930	James Woods	229668	Kristie Alley	235098	Kristie Alley	249762
Conor Mcgregor	128655	Kristie Alley	229638	James Woods	233583	James Woods	246273
James Woods	122915	Conor Mcgregor	126900	Conor Mcgregor	126945	Jack Nicklaus	137000
Jack Nicklaus	113600	Jack Nicklaus	116150	Jack Nicklaus	119000	Conor Mcgregor	128925

We can apply this method to all the celebrities in an election and get the influencers in order of the factor by which they were able to influence positively on the public.

Hence giving the political leadership a tool to select models for their future election propagandas.

VII. Conclusion

With each action being arranged on the web, information on interpersonal organizations is expanding many folds. Polarization, Hype, and Influence of these informal communities are very obvious during the Election Periods. Seeing the past decisions around the world, obviously the impact and impact brought about by these media are incredible like never before. For certain key characters taking up the positions, they utilize their supporters' attitude or bias to make groups, eventually prompting overall impact in no time. The beneficial thing is because of headways in Machine Learning and Natural Language Processing, this impact can be estimated which can prompt better political decision publicity and in any event, keeping up with mental soundness. To exhibit this, we take the 2020 US Elections alongside conspicuous characters/superstars and anticipate the impact they had the option to produce openly. For this, we utilize Natural Language Processing Techniques for the arrangement of tweets and afterward devise a recipe for impact forecast utilizing measurements. Our model can anticipate which superstar caused the most effect on a specific party. Having this data on hand can smooth the whole political decision publicity and be valuable for various groups to design their future too.

Difficulties faced during sentiment analysis were

- Sarcasm: In sarcastic text, people express their negative sentiments using positive words. This fact allows sarcasm to easily cheat sentiment analysis models unless they're specifically designed to take its possibility into account. This is very common on social media platforms.
- Metaphor: It describes an object or action in a way that isn't. This call leads to some disadvantages while analyzing the true sentiment.
- Images and emoticons: Sentiment analysis of images and emoticons as a reply on a tweet is also not part of our project as we focus on text sentiment analysis. This also narrowed our dataset.
- Links: There were some links as the reply to the tweet for which sentiment analysis was not possible.

VIII. References

[1] Beigi G., Hu X., Maciejewski R., Liu H. (2016) An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. In: Pedrycz W., Chen SM. (eds) Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence, vol 639. Springer, Cham. First Online: 23 March 2016. https://doi.org/10.1007/978-3-319-30319-2_13

[2] S. Ao, "Sentiment Analysis Based on Financial Tweets and Market Information," 2018 International Conference on Audio, Language and Image Processing (ICALIP), 2018, pp. 321-326, doi: 10.1109/ICALIP.2018.8455771. https://ieeexplore.ieee.org/document/8455771

[3] M. Chen, Q. Liu, S. Chen, Y. Liu, C. Zhang and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," in IEEE Access, vol. 7, pp. 13149-13158, 2019, doi: 10.1109/ACCESS.2019.2893448. https://ieeexplore.ieee.org/document/8620201

[4] H. Li, Y. Cao, S. Li, J. Zhao and Y. Sun, "XGBoost Model and Its Application to Personal Credit Evaluation," in IEEE Intelligent Systems, vol. 35, no. 3, pp. 52-61, 1 May-June 2020, doi: 10.1109/MIS.2020.2972533. https://ieeexplore.ieee.org/abstract/document/8988224

[5] M. Lobur, A. Romanyuk and M. Romanyshyn, "Using NLTK for educational and scientific purposes," 2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 2011, pp. 426-428. https://ieeexplore.ieee.org/document/5744524

[6] Raphael Kwaku Botchway, Abdul Bashiru Jibril, Zuzana Komínková Oplatková and Miloslava Chovancová, "Deductions from a Sub-Saharan African Bank's Tweets: A sentiment analysis approach", Cogent Economics & Finance, Volume 8, Issue 1(2020), Article 1776006, 08 Jun 2020. https://www.tandfonline.com/doi/full/10.1080/23322039.2020.1776

[7] https://github.com/flairNLP/flair

[8] Github repository for this research paper. https://github.com/nidhivanjare/Twitter_SA

[9] Mohammad Nur Habibi, Sunjana, " Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis ", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.11, pp. 22-30, 2019.DOI: 10.5815/ijmecs.2019.11.04

[10] PERSADA, Satria Fadil et al. Public Perceptions of Online Learning in Developing Countries: A Study Using The ELK Stack for Sentiment Analysis on Twitter. International Journal of Emerging Technologies in Learning (iJET), [S.l.], v. 15, n. 09, p. pp. 94-109, may. 2020. ISSN 1863-0383. Available at: <https://online-journals.org/index.php/i-jet/article/view/11579>. Date accessed: 06 Sep. 2021. doi:http://dx.doi.org/10.3991/ijet.v15i09.11579.

[11] Agarwal, Apoorv & Xie, Boyi & Vovsha, Ilia & Rambow, Owen & Passonneau, Rebecca. (2011). Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media.

[12] Liu B, "sentiment analysis and subjectivity, Handbook of natural language processing", vol.2, pp 627-666, 2010

[13] S Padmaja and Prof. S. Sameen Fatima, "Opinion Mining and Sentiment Analysis -An assessment of People's Belief: A Survey, International Journal of Ad

hoc, Sensor & Ubiquitous Computing (IJASUC) Vol. 4 No. 1, February2013

[14] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," in 2016 international conference on signal processing, communication, power and embedded system (SCOPES), 2016, pp. 1345-1350: IEEE. https://doi.org/10.1109/scopes.2016.7955659

[15] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in 2014 Seventh International Conference on Contemporary Computing (IC3), 2014, pp. 437-442: IEEE. https://doi.org/10.1109/ic3.2014.6897213