Inferring political preference from Twitter tweets

Nisha P. Shetty[1], Daita Ravi Teja[2], TUMMALA SRINAG VINIL[3],
SWATI KANWAL[4], Harsh Mutha [5], AKSHITA BHARGAVA [6]

***Abstract*—Commercial popularity of social media and free availability of vast data has enhanced the interests of the researchers in analyzing its contents. Many businesses make use of such data to harness the mindset and likes of their target audience, thereby improving their profits. Sentiment analysis of Twitter texts have proved to be an effective way of voicing the needs of large masses and is used by many prominent politicians in making better campaigning strategies. Multiple machine learning classifiers are implemented in this study to access the stance of US citizens towards Democrats/ Republicans to deduce which political party a user prefers from his tweets. Performance of a stacked ensemble is compared against a deep neural network for the mentioned problem domain.**

***Index Terms*—sentiment analysis,stacking,LSTM, Ensemble,TF-IDF**

I. Introduction

The recent years marked the proliferation of social media due to easy access of Internet and handy devices such as mobiles. Microblogging platforms like Twitter offer new acumens into thoughts, interests, happenings and opinions of its users. The popularity of these platforms is such that is a routine for masses at a daily basis.

Social media is a directory containing links (friend and follow relationships), exchanges (like, comment, repost), general facts (geographical location, profession, education) and content posted (texts, images, links). These platforms have nowadays become an outlet for all arenas ranging from simple interactions, political and news discussions, brand promotions etc.

As opposed to classical data collection tools like surveys this platform can reach wide masses easily, so social network analysis have garnered the interest of many researchers. Any social online activity (like, comments etc.) is nowadays analyzed thoroughly to gain monetary benefits. One such example is that of online retailers who keep track of search history, comments, likes, general information like age group etc. of customers and give recommendations based on the same.

Popular personalities like Obama have used social network platforms majorly for their campaigns. Politicians use their pages to publicize their political agenda. Thus, we can safely say that such platforms provide a viable source for judging people's political affiliation. Tweets and hashtags from Twitter can be thus evaluated during elections to predict outcome of the election while envisaging people's political affinity.

Composition of the rest of the paper is as follows. Section 1 introduces the problem domain and the need fo it. Section 2 touches upon the key prominent works in the proposed domain. Proposed methodology is exemplified in section 3. Section 4 portrays the obtained results and section 5 concludes the given work while listing out possible future avenues in the domain.

II. Related Works

Michael D. Conover *et al.* (2011) [1] discerned out that SVM trained with hashtags yielded much better results than actual text. Latent Semantic Analysis of the text provided a good intuition on the political affiliation of a person. They concluded that network based analysis outperformed the content based analysis approaches. The authors aim to further their approach to generalize it all platforms and applications.

Conrad *et al.* (2016) [2] experimented on United States Federal Election Commission to predict the political affiliation and donations using popular machine learning classifiers such as C4.5, SVM, Naive Bayes and CNG. However, due to limited size of their data and extracting record labels from disparate data sets a conclusive hypothesis could not be drawn.

Jyoti Ramteke *et al.* (2016) [3] trained Multinomial Naïve Bayes classifier and SVM to predict if the alliance of a person is towards Democratic or Republican. In order to achieve better results the authors proposed an Active learning model which automates the data labelling process.

Chang *et al.* (2017) [4] tried to predict the affiliation of a particular post towards any of the two popular political parties in Taiwan using machine learning models with Term Document Matrix and Partisan prototypical words. The two models suffered from some limitations such as lack of generalization and requirement of more memory.

Rachael Tatman *et al.* (2017) [5] explored the sub lexical features such as capitalization which causes tweeting variation amongst various political groups. However the authors could not garner good accuracy as verbal and graphical features were ignored in this work. They did not take into consideration that writing style can be adapted by any party and have no means to validate any user.

Marco Di Giovanni *et al.* (2018) [6] incorporated many machine learning algorithms with TF-IDF vectorizer. Although outliers could be easily detected, misclassification occurred, which can be decreased by inculcating better content based detection approaches.

Caetano *et al.* (2018) [7] analyzed the political homophily amongst Twitter subscribers via retweets, mentions and follow. The authors plan to explore other features such as gender, age,

and ethnicity while ensuring that sybils are effectively removed from the study.

Marco Pota *et al.* (2018) [8] studied the influence of positive and negative tweets on the public opinion. Authors concluded that combining CNN based approach with the traditional dictionary based approach can yield better results.

Dorle *et al.* (2018) [9] employed LSTM to predict the sentiments of the user towards a political party.

Aziz *et al.* (2019) [11] devised a new method to perform sentiment analysis. The authors constructed a Hierarchical Knowledge Tree (HKT) which performs contextual analysis to identify appropriate words used in various contexts. In order to improve the classification tree similarity and tree difference index was employed in the research which aided in clustering similar words together without using any dictionary and can be extended across many data sets. The authors plan onto apply their work on real time data sets.

Yadav (2020) *et al.* [14] employed packages of R view the results of Delhi election.

Quirós (2020) *et al.* [10] designed an interactive framework which extracted the following groups and opinions of prominent Spanish political parties to draw conclusions such as parties sharing common ideologies, amount of communication between inter-parties and so on.

Himelboim (2016) *et al.* [12] studied the tweets of the users to deduce if they support or criticize a particular political party. NodeXL was used to visualize the network which was created using K-means after employing cluster analysis. Thus, their study effectively proved the concept of homophily which was clearly observed in the interaction scenario. However, the authors failed to explore if various emotions of the same valence can influence choices differently. Data collection, limited to one media space, for the proposed work happened in a biased scenario which caused bias.

Mai (2020)*et al.* [13] introduced a new dimension in sentiment analysis which combined the pros of both aspect level and sentence level analysis of comments. Their method firstly separates product related comments from the rest, finds the opinion, performs sentiment analysis and finally aggregates all the opinion for a product. The proposed method eliminates the need for feature engineering and other linguistic sources. The authors plan onto extend their corpus and incorporate meta-classifiers in their work, which has been done in our study.

Alamanda (2020) [15] created a search engine which collects reviews about a product and analyzes pros and cons of different features of the product from customer review. There are various filtering features such as reviews from a certain time interval. Although this work has been done for textual format it can be extended to other types as well. Future improvements suggested is to find if the user giving the review is genuine or not by exhibiting behavioral analysis so as to determine the authenticity of the reviews.

Wadawadagi *et al.*(2020) [16] explored how various deep neural networks can be employed to perform sentiment analysis. The effective fine tuning of hyper parameters like learning rate, hidden layers etc. discussed helped a lot in our study.Some of the key scope for improvements put are devising a method to automatically tune the hyper parameters. The authors encouraged the extending the research in various avenues and suggested the explore Ensemble learning which is done in our study.

Yadav and Vishwakarma (2020) [17] undertook a study on bio-inspired algorithms and used them to exhibit sentiment analysis. Authors encourage the researchers to implore these algorithms for other modalities like images, videos etc.

Sachin *et al.* (2020) [18] surveyed the various baseline models of RNN like LSTM, GRU and Bi-LSTM and Bi-GRU in the scope of sentiment analysis.

Naresh and Yadav (2020) improved the performance traditional decision tree in [19]

III. Methodology

The Fig. 1 gives a detailed explanation about the proposed methodology.

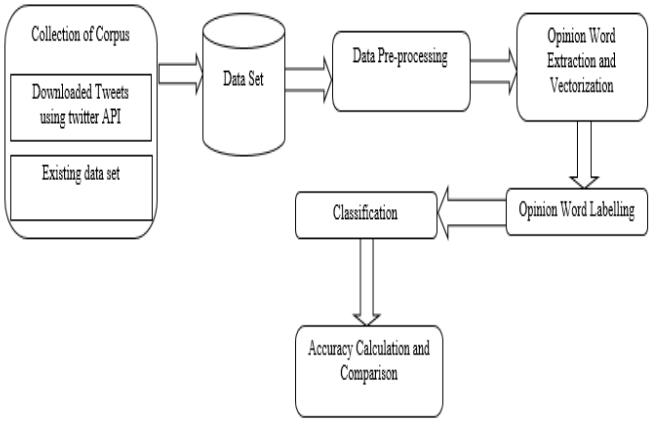

Fig. 1. Overall Methodology

1) Data Collection and Labeling
The collected data set is a combination of primary and secondary research. The secondary data set was found on Kaggle containing multiple tweets from a few users, all labeled into either Democratic or Republican classes. More tweets were collected using Twitter Streaming API These tweets were ranked into levels 1-5 (with 5 being most politically inclined) 'likes' and 're-tweets' for those tweets ranked at a level 5 were obtained. 'Likes' and 're-tweets' were used for collecting further data set (by fetching the corresponding screened).
2) Data Pre-processing
Here tweets were stripped of unwanted punctuations, URLs and special characters like '@'. Stop words were removed and stemming and lemmatization are incorporated to grammatically correct the tweets and get them to apt dictionary form.
   - Obtaining Syntactic words : Parts of speech tagging is done using "Tree-Tagger" to split the sentence into its constituents such as nouns, verbs,

adverbs, adjectives etc.. Adjectives and hash tags give the sentiments expressed in the tweets.

- VECTORIZATION : Obtained list of words is transfigured into array of numbers to aid in classification process.Current methodology incorporates two vectorization methods :
  - *Count Vectorizer*: Counts the number of times a particular word is used.
  - *Term frequency-inverse document frequency (TFIDF)*: Term frequency equals the number of times a word appears in a document divided by the total number of words in the document. Inverse document frequency calculates the weight of rare words in all documents in the corpus, with rare words having a high IDF score, and words that are present in all documents in a corpus having IDF close to zero.

3) CLASSIFICATION OF TWEETS
   - *Opinion Words Orientation:* Using various dictionaries such as SentiWordNet, Opinion-Lexicon and SentiLex to obtain the polarity of the sentence (positive tweets garner +1 while negative tweets obtain -1).
   - *Classifiers Used:* Following classifiers are made to learn features and proclivity of tweets which are further used to categorize new tweets.
     a) Multinomial Naive Bayes(MNB)
     b) Logistic Regression(LR)
     c) Linear Support Vector Classifier (SVC)
     d) Proposed stacked model (Base classifiers : MNB,LR and Linear SVC; Meta-classifier :Support Vector Machine(SVM)) as shown in Fig.2 and Algorithm 1 [20].
     e) Bidirectional Long Short-Term Memory(LSTM)
4) RESULT ANALYSIS AND COMPARISON
   Obtained accuracies were compared and analyzed.

$$Accuracy = tp + tn/tp + tn + fp + fn \quad (1)$$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives and fn is the number of false negatives.

Algorithm 1 Proposed Stacking Ensemble

Input : Training data set D of m tweets
Output: Predicted political party
for $i \leftarrow 1$ *to* b **do**
  learn the base classifier h_i *on D*
end
for $d \leftarrow 1$ *to* m **do**
  construct new data set which has output of base classifiers as features along with original class labels
end
learn the meta classifier on the newly constructed data set.
return the predicted political party

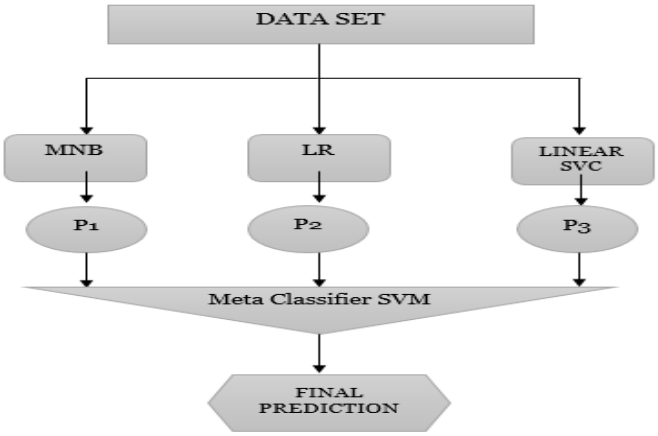

Fig. 2. Proposed Stacking Ensemble

IV. IMPORTANCE OF PROPOSED RESEARCH INVESTIGATION

- The proposed technique can be extended to multiple platforms such as recommendation systems (hotels, movies, products etc.), improving marketing strategies etc. This technique helps the retailers to analyze the product features, adhere to customer needs, pursue trends and keep a finger on the opposition.
- In terms of elections this can serve as a means of voting advice to people as they can analyze the choice of their near and dear ones. Similarly it can apply to many other real world events (government policies etc.) as well.

V. RESULT ANALYSIS

Table I and Fig. 3 exhibits the results of the above work.

TABLE I
RESULTS

Classifier	Accuracy (in percent)
Multinomial Naïve Bayes	85.4
Logistic Regression	84.67
Linear SVC	80.97
Proposed Stacked Model	**87.61**
Bidirectional LSTM	**89.56**

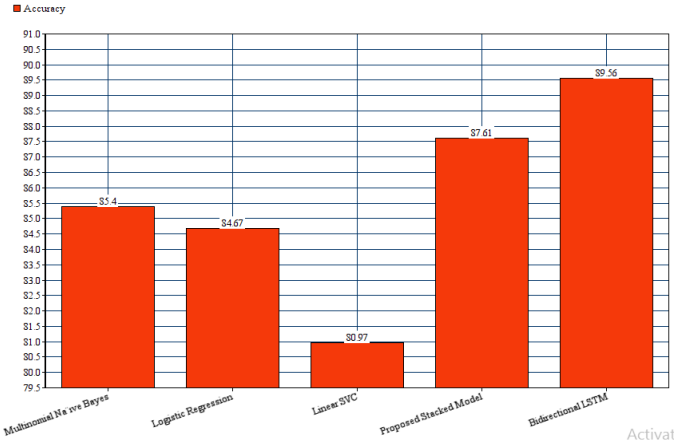

Fig. 3. Results

VI. Conclusion and Future Work

Inference attacks are one of the major toxic effects of social media. The proposed work explores the content of user's tweets and the opinion expressed in them to predict the political party a user prefers. The technique can be extended to multiple platforms such as recommendation systems (hotels, movies, products etc.), improving marketing strategies etc. This technique helps the retailers to analyze the product features, adhere to customer needs, pursue trends and keep a finger on the opposition. In terms of elections this can serve as a means of voting advice to people as they can analyze the choice of their near and dear ones. Similarly it can apply to many other real world events (government policies etc.) as well.

Other machine learning techniques like clustering, reinforcement learning etc. can be explored to further this study. Steps should be taken to analyze the failures (false positives and false negatives) and infer "periphery cases"(profiles which are neutral and show inclination to both the domains). A single system can be developed which works on multiple platforms (Facebook, LinkedIn etc.) and can handle linguistic differences to enforce generality.

References

[1] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Oct 2011, pp. 192–199.

[2] C. Conrad and V. Keˇselj, "Predicting political donations using twitter hashtags and character n-grams," in 2016 IEEE 18th Conference on Business Informatics (CBI), vol. 02, Aug 2016, pp. 1–7.

[3] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using twitter sentiment analysis," in 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 1, Aug 2016, pp. 1–5.

[4] C.-C. Chang, S.-I. Chiu, and K.-W. Hsu, "Predicting political affiliation of posts on facebook," in Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, ser. IMCOM '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3022227.3022283.

[5] R. Tatman, L. Stewart, A. Paullada, and E. Spiro, "Non-lexical features encode political affiliation on twitter," in Proceedings of the Second Workshop on NLP and Computational Social Science. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 63–67. [Online]. Available: https://www.aclweb.org/anthology/W17-2909.

[6] M. Di Giovanni, M. Brambilla, S. Ceri, F. Daniel, and G. Ramponi, "Content-based classification of political inclinations of twitter users," in 2018 IEEE International Conference on Big Data (Big Data), Dec 2018, pp. 4321–4327.

[7] J. A. Caetano, H. S. Lima, M. F. Santos, and H. T. Marques-Neto, "Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 american presidential election," Journal of Internet Services and Applications, vol. 9, no. 1, p. 18, Sep 2018. [Online]. Available: https://doi.org/10.1186/s13174-018-0089-0.

[8] M. Pota, M. Esposito, M. A. Palomino, and G. L. Masala, "A subword-based deep learning approach for sentiment analysis of political tweets," in 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), May 2018, pp. 651–656.

[9] S. Dorle and N. Pise, "Political Sentiment Analysis through Social Media," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 869-873, doi: 10.1109/ICCMC.2018.8487879.

[10] Quirós, Pelayo and Blanca Rosario Campomanes Álvarez. "The new Spanish political scenario: Twitter graph and opinion analysis with an interactive visualisation." EDBT/ICDT Workshops (2020).

[11] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," in IEEE Access, vol. 8, pp. 17722-17733, 2020, doi: 10.1109/ACCESS.2019.2958702.

[12] Himelboim, I., Sweetser, K.D., Tinkham, S.F., Cameron, K., Danelo, M. West, K. (2016). "Valence-based homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election". In New Media Society, 18 (7), 1382– 1400. London: SAGE Publications.

[13] Mai, L., Le, B. Joint sentence and aspect-level sentiment analysis of product comments. Ann Oper Res (2020). https://doi.org/10.1007/s10479-020-03534-7.

[14] Yadav, D., et al. "Political Sentiment Analysis On Delhi Using Machine Learning." Advances in Mathematics: Scientific Journal, vol. 9, no. 3, 2020, pp. 1239–1250., doi:10.37418/amsj.9.3.50.

[15] Alamanda, M.S. Aspect-based sentiment analysis search engine for social media data. CSIT 8, 193–197 (2020). https://doi.org/10.1007/s40012-020-00295-3.

[16] Wadawadagi, R., Pagi, V. Sentiment analysis with deep neural networks: comparative study and performance assessment. Artif Intell Rev (2020). https://doi.org/10.1007/s10462-020-09845-2.

[17] Yadav, A., Vishwakarma, D.K. A comparative study on bio-inspired algorithms for sentiment analysis. Cluster Comput (2020). https://doi.org/10.1007/s10586-020-03062-w.

[18] Sachin, S., Tripathi, A., Mahajan, N. et al. Sentiment Analysis Using Gated Recurrent Neural Networks. SN COMPUT. SCI. 1, 74 (2020). https://doi.org/10.1007/s42979-020-0076-y.

[19] Naresh, A., Venkata Krishna, P. An efficient approach for sentiment analysis using machine learning algorithm. Evol. Intel. (2020). https://doi.org/10.1007/s12065-020-00429-1.

[20] Smitha Rajagopal, Poornima Panduranga Kundapur, Katiganere Siddaramappa Hareesha, "A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets", Security and Communication Networks, vol. 2020, Article ID 4586875, 9 pages, 2020. https://doi.org/10.1155/2020/4586875.