ELECTION RESULT PREDICTION AND ANALYSIS BASED ON TWITTER DATASET

DR R KAVITA[1], MAYANK KUMAR[2], SHUBHAM.K[3], ASHISH SRIVASTAVA[4]

[1]Assistant Professor, Dept of Information Technology, SRM Institute of Science and Technology, Ramapuram, Chennai, India.

[2,3,4]UG Student, Dept of Information Technology, SRM Institute of Science and Technology, Ramapuram, Chennai, India

ABSTRACT-

Sentiment Analysis have influence on user generated content such as blogs, ecommerce sites etc. The aftereffects of Sentiment Analysis are standing out enough to be noticed with advertisers that they can assess the accomplishment of a promoting effort or the demeanor of individuals on another item dispatch. Entrepreneurs and promoting organizations are utilizing Sentiment Analysis to begin new business procedures and to recognize openings for new item improvement. The various sources collected tweets were meant to be classified as positive, negative and neutral, based on polarity. The machine learning classifier algorithms cross validation were applied on the dataset and the results were tabulated for comparing and estimating which classifier algorithm yields the best accuracy. Other execution metric qualities like F Score, Precision, Recall were additionally determined for correlation of different classifier exhibitions on Sentiment Analysis Our system along with machine learning methods and Random forest algorithms proves to be giving better accuracy score compared to earlier methods. We have used the concept of AUC (or Area Under the Curve) which highlights important aspects based on the graphs and helps us reach the optimum possible accuracy with the prediction of results which takes into consideration of various aspects of individuals attitude and behavior as they project on social media.

Keywords : **Random Forest Algorithm, Sentiment Analysis, Area Under the Curve, Classification Algorithm, Polinomial Feature Extraction.**

1- INTRODUCTION

Social media platforms such as twitter is very effective in generating information which targets various individuals such as location,time-stamps, and number of followers, etc., which has been utilized over the years to provide application-centric results.With the increase in availability of real-time online data streams, event detection methods have become the primary and fast mode of finding various trends which could be found in other media platforms.

The two primary challenges that require attention in times of events through social media data have been summarized below.

To store context based relationed data :

The real-time Twitter data collection mayresult in ahuge size. However, the maximum allowed length of atweet is 280 characters, where the most common length of 33 characters in total. Due to the limited length,it becomes a challenge to generate the most relatetweets. In thiresearch work, this is referred to as uncertainty in capturing the contextual relationship among thetweets.

- Increase in computation cost: The Twitter data are available inmassive amounts, and the number of tweets belongingto a particular event is generally massive in number.

Thus, processing all the data to its extremes may leadto redundant computations and incur high computationalcost.

The main focus of project revolves around the steps needed to improve upon a accuracy score, so that the system can predict with much precision.

The project can collect the data from any source but for our convenience we have taken a data from twitter, where people put their opinion and based upon their reaction toward a candidate can speak about their attitude toward them, and with systematically collecting this data and analyzing them can help us predict election result.

2 Existing System-

Existing techniques basically cast the issue as an administered archive order task. These can be separated into two classes: one depends on manual element designing that are then devoured by calculations, for example, SVM, Naive Bayes, and Logistic Regression the other addresses the later profound learning worldview that utilizes neural organizations to consequently learn multi-facets of theoretical highlights from crude information. The current framework was assemble utilizing the information mining strategies. In this technique they manage little level dataset. Here this framework will have a low exactness score like 60 to 70 % as it were.

Drawbacks of existing system-

1- Previous work failed to handle high level data.

2- Low accuracy score.

3. Proposed System –

Our system uses machine learning systems to analyze the data collected from twitter and use the following to predict the election results. Our system uses libraries namely numpy, pandas, sklearn, matplotlib, and also uses the following algorithm for classification and feature extraction namely , Random forest algorithm and Polinomial Feature extraction algorithms. Our system predicts the result based on following : ‘County', 'Population', 'Population-growth', 'Population-density', 'Income-per-capita', 'Percent-white', 'Percent-in-poverty','Bachelors-degree-or-higher', and formulates graphs and finally takes Area Under the Curve (AUC), to predict which candidates are winning from which region. Our system with the help of Random Forest Algo. Is able to predict the result with an accuracy score of above 90%.

Advantage of proposed system-

1-Improve the accuracy score.

2-Deal with large amount dataset.

3- System architecture diagram-

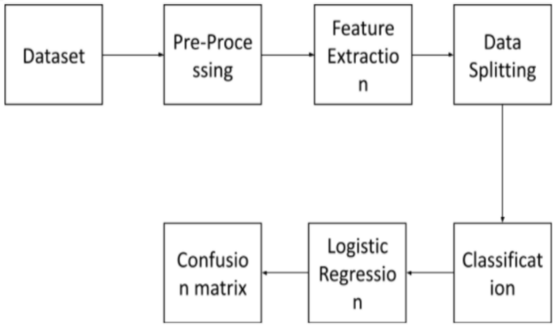
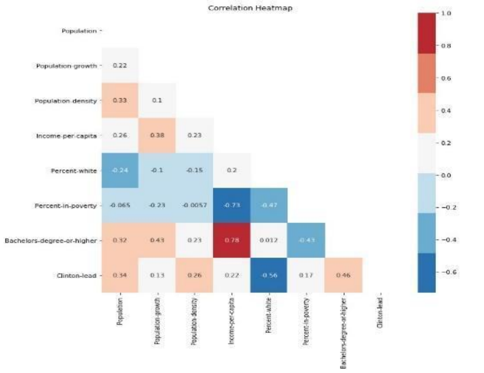
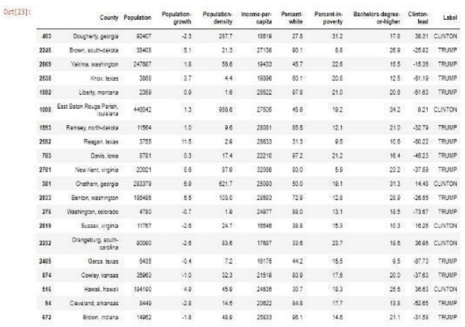


Fig no. 1 – Architecture diagram

4- Result and discussion

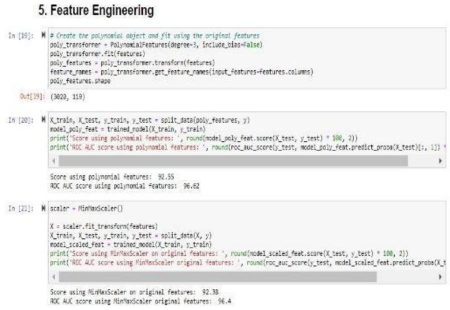
The graph shows various aspects which helps us determine the results, the graph is based on population growth and density along with the characteristics of people living in particular areas, the profession and education history also used in areas to properly map the insights which helps us build on accuracy score, our classification algorithm helps differential important data from stream of texts.





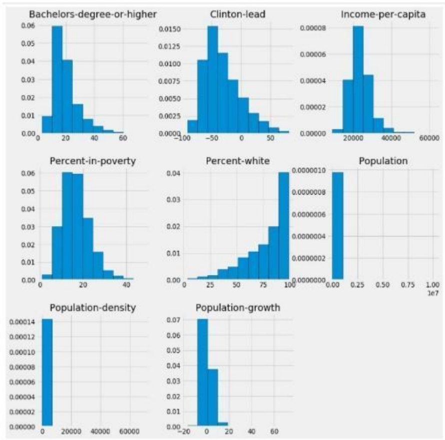
Out[23]:

	County	Population	Population-growth	Population-density	Income-per-capita	Percent-white	Percent-in-poverty	Bachelors-degree-or-higher	Clinton-lead	Label
403	Dougherty, georgia	92407	-2.3	287.7	18819	27.8	31.2	17.8	38.31	CLINTON
2246	Brown, south-dakota	38408	5.1	21.3	27136	90.1	8.6	26.9	-25.82	TRUMP
2869	Yakima, washington	247687	1.8	56.6	19433	45.7	22.6	15.5	-15.36	TRUMP
2538	Knox, texas	3858	3.7	4.4	19396	60.1	20.8	12.5	-61.19	TRUMP
1592	Liberty, montana	2369	0.9	1.6	25622	97.8	21.0	20.8	-51.63	TRUMP
1098	East Baton Rouge Parish, louisiana	446042	1.3	958.6	27526	45.9	19.2	34.2	9.21	CLINTON
1993	Ramsey, north-dakota	11564	1.0	9.6	28081	85.6	12.1	21.0	-32.79	TRUMP
2592	Reagan, texas	3755	11.5	2.9	28833	31.3	9.5	10.6	-80.22	TRUMP
783	Davis, iowa	8781	0.3	17.4	22210	97.2	21.2	16.4	-46.23	TRUMP
2781	New Kent, virginia	20021	8.6	87.9	32068	80.0	5.9	23.2	-37.59	TRUMP
381	Chatham, georgia	283379	6.9	621.7	25093	50.0	19.1	31.3	14.48	CLINTON
2833	Benton, washington	186486	6.5	103.0	28683	72.9	12.8	28.9	-26.65	TRUMP
276	Washington, colorado	4780	-0.7	1.9	24977	88.0	13.1	18.5	-73.67	TRUMP
2819	Sussex, virginia	11767	-2.6	24.7	16546	38.8	15.3	10.3	16.26	CLINTON
2232	Orangeburg, south-carolina	90090	-2.6	83.6	17687	33.6	23.7	18.6	36.96	CLINTON
2485	Garza, texas	6435	-0.4	7.2	16176	44.2	15.5	9.5	-87.73	TRUMP
874	Cowley, kansas	35963	-1.0	32.3	21519	80.9	17.6	20.0	-37.63	TRUMP
516	Hawaii, hawaii	194190	4.9	45.9	24636	30.7	18.3	25.6	36.63	CLINTON
94	Cleveland, arkansas	8449	-2.8	14.6	20622	84.8	17.7	13.8	-52.65	TRUMP
672	Brown, indiana	14962	-1.8	48.9	25833	96.1	14.6	21.1	-31.59	TRUMP



5. Feature Engineering

```
In [19]:  # Create the polynomial object and fit using the original features
          poly_transformer = PolynomialFeatures(degree=3, include_bias=False)
          poly_transformer.fit(features)
          poly_features = poly_transformer.transform(features)
          feature_names = poly_transformer.get_feature_names(input_features=features.columns)
          poly_features.shape

Out[19]:  (3020, 119)

In [20]:  X_train, X_test, y_train, y_test = split_data(poly_features, y)
          model_poly_feat = trained_model(X_train, y_train)
          print('Score using polynomial features: ', round(model_poly_feat.score(X_test, y_test) * 100, 2))
          print('ROC AUC score using polynomial features: ', round(roc_auc_score(y_test, model_poly_feat.predict_proba(X_test)[:, 1]) *

          Score using polynomial features:  92.55
          ROC AUC score using polynomial features:  96.62

In [21]:  scaler = MinMaxScaler()

          X = scaler.fit_transform(features)
          X_train, X_test, y_train, y_test = split_data(X, y)
          model_scaled_feat = trained_model(X_train, y_train)
          print('Score using MinMaxScaler on original features: ', round(model_scaled_feat.score(X_test, y_test) * 100, 2))
          print('ROC AUC score using MinMaxScaler original features: ', round(roc_auc_score(y_test, model_scaled_feat.predict_proba(X_t

          Score using MinMaxScaler on original features:  92.38
          ROC AUC score using MinMaxScaler original features:  96.4
```



5- Modules-

Module 1: Data Pre-processing:

The data pre-processing is nothing but filtering and clean the data in the following steps such that data is ready for further processing. Data pre-processing is done to obtain cleaner data which in turn will provide ease of processing the data further to obtain meaningful results, steps pre-process the data.

Module 2: Sentiment Detection:

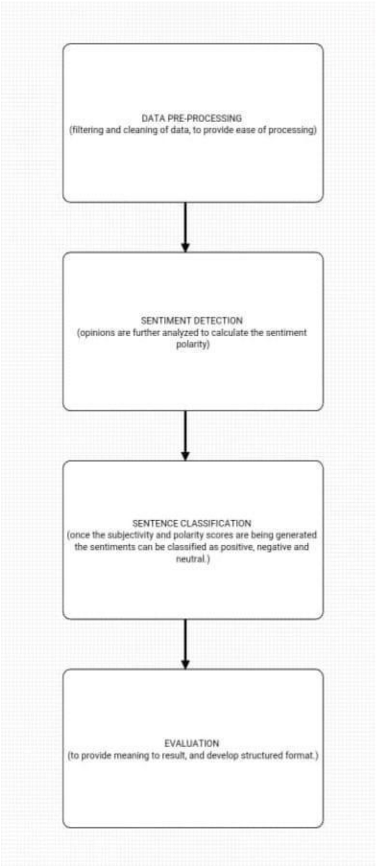
In this Phase, the reviews and opinions are further analyzed to calculate the sentiment polarity and subjectivity.

Module 3: Sentence Classification:

Once the subjectivity and polarity scores are being generated the sentiments can be classified as positive, negative, and neutral.

Module 4: Evaluation:

The reason output presentation is an essential task is to provide meaning to our results. The information gained from the analysis is required to be presented in such a manner that it reflects meaningful information from the unstructured data by visualizing and plotting the results.


DATA PRE-PROCESSING
(filtering and cleaning of data, to provide ease of processing)
SENTIMENT DETECTION
(opinions are further analyzed to calculate the sentiment polarity)
SENTENCE CLASSIFICATION
(once the subjectivity and polarity scores are being generated the sentiments can be classified as positive, negative and neutral.)
EVALUATION
(to provide meaning to result, and develop structured format.)

6- Conclusion

While using the data from different platform for our project it poses challenged at different level. In this paper, we first tackle the shortage of preparing information for text grouping by giving a two phase system. At last we propose our model for prediction of election result which utilize the information by our framework. While our model alone may not be adequate to foresee the outcomes, anyway it turns into a vital segment when joined with other statistical modelling and offline techsWe carried out the proposed model based on a dataset functions which was made by digging Twitter for 3 days. Be that as it may, this model can be stretched out later on to make a mechanized system which mines information for quite a long time since political result expectation is a consistent cycle and requires investigation over significant stretches of time. Highlights ought to be extricated from recently mined information furthermore, contrasted and existing arrangement of highlights. Some comparability metric can be utilized to look at the new and old highlights. As it were in situations where the measurement esteem passes a boundary, the recently mined information ought to be named utilizing the two phase structure. In this manner we suggest making an Active learning model wherein the actual model suggests what information ought to be named. This would limit the endeavors for naming while ensuring that there is no trade off on logical pertinence.

8 -REFERENCES

The format will number references sequentially inside sections

[1] J. Kaur, S. S. Sehra, and S. K. Sehra, “A systematic literature review of sentiment analysis,” Int. J. Comput. Sci. Eng., vol. 5, no. 4, pp. 22–28, 2017.

[2] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” Image Vis. Comput., vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.

[3] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, “Supervised sentiment analysis in multilingual environments,” Inf. Process. Manage., vol. 53, no. 3, pp. 595–607, May 2017, doi: 10.1016/j.ipm.2017.01.004.

[4] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias,

“Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017, doi: 10.1016/j.eswa.2017.02.002.

[5] S. Wang and C. D. Manning, “Baselines and bigrams: Simple, good sentiment and topic classification,” in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, vol. 2, Jeju Islan