Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Sentiment analysis in twitter data using data analytic techniques for predictive modelling

To cite this article: A Razia Sulthana *et al* 2018 *J. Phys.: Conf. Ser.* **1000** 012130

View the article online for updates and enhancements.

Related content

- A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter
Asniar and B R Aditya
- Sentimental Analysis for Airline Twitter data
Deb Dutta Das, Sharan Sharma, Shubham Natani et al.
- A Peculiar Sentiment Analysis Advancement in Big Data
Manisha Valera and Yash Patel

Sentiment analysis in twitter data using data analytic techniques for predictive modelling

A Razia Sulthana[1], A K Jaithunbi[2], L Sai Ramesh[3]

[1]Department of Information Technology, SRM University, TamilNadu, India

[2]Department of Computer Science and Engineering, RMD Engineering College, TamilNadu, India

[3]Department of Information Science and Technology, Anna University, TamilNadu, India

Abstract. Sentiment analysis refers to the task of natural language processing to determine whether a piece of text contains subjective information and the kind of subjective information it expresses. The subjective information represents the attitude behind the text: positive, negative or neutral. Understanding the opinions behind user-generated content automatically is of great concern. We have made data analysis with huge amount of tweets taken as big data and thereby classifying the polarity of words, sentences or entire documents. We use linear regression for modelling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted X. We conduct a series of experiments to test the performance of the system.

1. Introduction

Big data [1] analytics is collection, organizing and analyzing of large sets of. It helps organizations to understand the information contained within the data. Itidentifies the data that is most important to the business.Data Analytics [2] are used in handling big data.Application of data analytics to structured and unstructured data is important for business planning. Data analytics influences the business decision influenced by behavioral biases. Sentiment Analysis represents another valuable source of information that helps business decisions and performance evaluation.

Twitter tweets were analyzed based on the time and location in [3]. The tweets were analyzed based on the user behavior, tweet language, tweet source and the user joining rate for a year period of time interval. This work also identifies the information contained in tweet as audio file, video file, smileys etc. A pilot study in [4] measures the twitter tweets and identifies the similarity between them on six measures. It obtains the e-cigarette tweets from two online markets and finalizes the one with better measures.

Machine Learning Classifier is used to identify the suicide-related phrases [5] posted in twitter. It identified the strong concerning tweets and possibly concerning tweets were identified. It developed a classified and identified 80% of the tweets correctly. The support vector machines (SVM) and Logistic regression approaches were used along with cross validation approaches were used to identify the accuracy of the system in 10 folds. Psychoanalytic profiling of the tweets is discussed in [6] and the users personality behavior is analyzed. It categorizes the tweets as per the following categories: DISC

(Dominance, Influence, Steadiness, and Compliance) approach. RapidMiner tool and R language is used to extract the tweets which falls under these categories. The analysis shows that most of the tweets fall under Dominance category than compliance category.

Topic based feature selection is applied in [7] to categorize the tweets that fall under respective topics. News information is collected from online websites and similarity match is obtained with topics. Part of Speech(POS) is applied on the tweets and the noun context is extracted from the tweets. The noun words are considered as features.Different Machine learning techniques used in sentiment analysis and evaluation of these techniques are discussed in [8]. Naïve bayes classification is used for classifying the tweets into positive, negative and neutral [9]. It develops a Sentimentor tool, which analyses the tweets using Twitter API. And Regular expression identifies the words like ‘OOOMMMGGG’ & ‘OMG’ etc. It identifies unigram and bigrams from the tweets and pronouns, thus building a sentiment classifier

A survey made with Oncological tweets [10] collected from American Society for Radiation Oncology (ASTRO), American Society of Clinical Oncology (ASCO), and Society of Surgical Oncology (SSO) during the year 2014 shows that tweets spread the outbreak of HIN1 swine flu disease. It analyzed that most of the tweets fell into disseminating Information category and few were posted by general media, personal blogs and scholarly articles. Table 1 shows the performance comparison of the related work.

Table 1. Comparison of Performance of Related Work

	Approaches Used	Accuracy
[5]	SVM	67
[7]	Topic Classification	79
[8]	NB	74.56
	SVM	76.68
[9]	NB	52.31
[13]	SVM	83.33
[14]	NB	76.08

Geotagging the location of the tweets were done in [11] for location the place from where the tweets are been posted. SVM approaches were used and density based clustering approaches were used to cluster the similar tweets posted from similar sites. Precision, Recall and F-measure were the performance measures used to test the performance. It identifies the tweets posted from nearest location which are in surroundings wide of 144.8 meter. These spotswere identified to be the maximal number of tourist attraction. Twitter based sentiment analysis in [12] discussed on the techniques for collecting the tweets, finding the friends circle, identifying the location of the tweet, twitter API’s and approaches used for analyzing tweets.

Identification of pornographic content from twitter [13] is identified using Naïve Bayes (NB), Decision Trees (DT), SVM classification approaches. It was tested with Indonesian English and English language tweets. It identified that DT gave better results with Indonesian English and SVM gave better results with English language tweets. The principal of Entropy and Naïve Bayes Classification is applied in [14] for classification of tweets. It used Bigram approach to classify the negations in handling tweets.

2. Proposed Approach

Twitter is an online social networking service that allows users to send and read short 140-character messages called "tweets". Registered users can read and post the tweets. The unregistered users can only read them. Users would access Twitter through the website interface, Short Messaging Service (SMS) or mobile device app. The Twitter micro-blogging service hastwo Application Programming Interface (API): Rest API and Search API. The Twitter Rest API allows the developers to access core Twitter data. It includes update timelines, status data, and user information. The Search API let the developer to interact with Twitter Search and trends data. It supports the following formats: XML, JSON, and the RSS. We have implemented our research work in Tweepy tool. It is an is open-sourced tool hosted on GitHub. It enables Python to communicate with Twitter API platform.OAuth is the new basic authentication approach that supports Tweepy.

Sentiment analysis [15] is the process of using text analytics to understand the polarity of sentences. The basic task in sentiment analysis is classifying the polarity of document / sentence, in a documenas positive, negative, or neutral. The existing data analytics products use advanced Natural Language Processing (NLP) techniques. We extracted the tweets from twitter. Part of Speech (POS) tags were applied to the collected tweets and the nouns were extracted as tweets. The collected tweets were applied to the classifier. The classifier was trained with linear regression approach. The classifier segregates the tweets into positive and negative tweets and the performance of the system was analyzed. The architecture of the system is shown in figure 1.

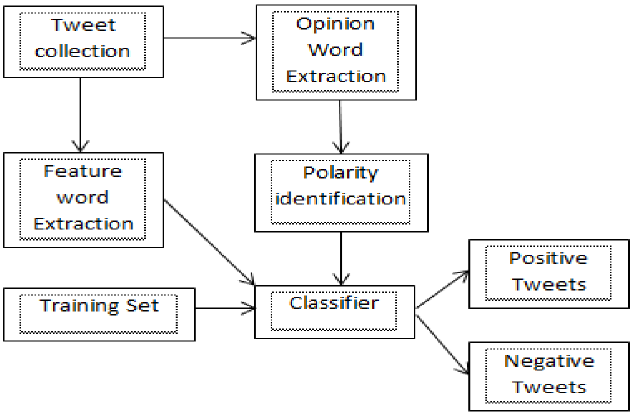


Figure 1. System Architecture

3. Techniques and Methodology Applied

3.1 Getting Data from Twitter Streaming API

APImakes the interaction with computer programs and web services easy. Web services provide APIs to interact with their services. To establish Twitter Streaming API the system need 4 piece of information: API key, API secret, Access token and Access token secret.

The procedure for getting twitter API keys is given below:

1. Create a new twitter account and log in with the twitter credentials.
2. Create anew app and fill in the credentials agreeing to the terms and create the twitter application.
3. Following which extract the "API keys" and "API secret" information
4. Select and copy the "Access token" and "Access token secret".

3.2 Predictive Modelling

We have developed a predictive model (figure 2. Predictive model of our system) for analysis. It is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. The predictive modelling approach uses linear regression with the following parameters: customer's gender, age, purchase history, and future sale. The dependent variables would be: customer's gender, age and independent variables: prediction of future sale. We employ a linear equation and a neural network implemented in python. The data collected from the relevant predictors is analyzed and a statistical model is formulated. The test data is used for making predictions.

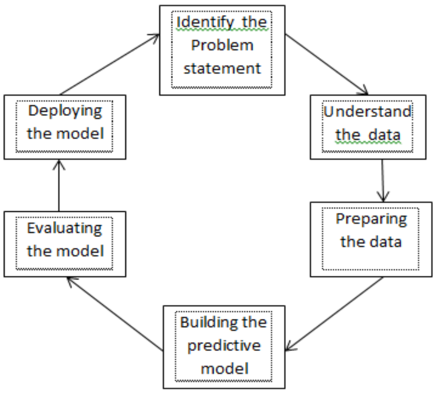


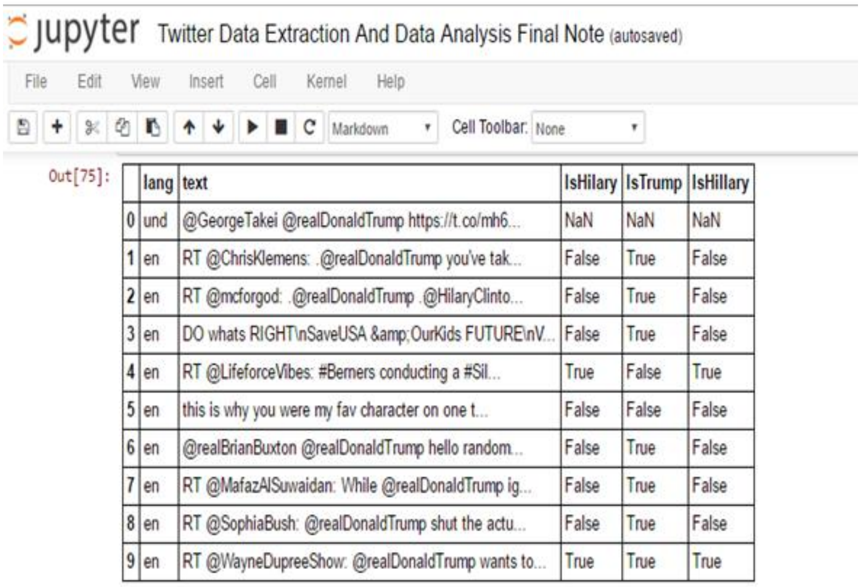
Figure 2. Predictive Model

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimatedfrom the data. Such models are called linear models.There are several ways in which you can implement linear regression in Python. You can do linear regression using numpy, scipy, stats model and scikit learn.

The module implemented here is Scikit. Scikit-learn is a powerful Python module for machine learning. It contains function for regression, classification, clustering, model selection and dimensionality reduction.Scikit-learn is largely written in Python, with some core algorithms written in Cython to achieve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR.The Tools and languages that are used for deployment are JSON, PANDAS, MATPLOTLIB, RE, SCIKIT LEARN, NUMPY

4. Results and Graphical Output of the system

Our system is experimented with 14,000 tweets collected from twitter. The system was executed with linear regression equations. One-third of the dataset was taken for testing and three-fourth of the dataset was taken for training. The system has 10 fold execution pattern and the results are compared with literature results. The tweets were regarding the opinion posted by users towards the election status between Hillary and Trump. The tweets are analyzed to find the peoples opinion over their support towards Hillary or Trump. The screenshot is given in figure 3.

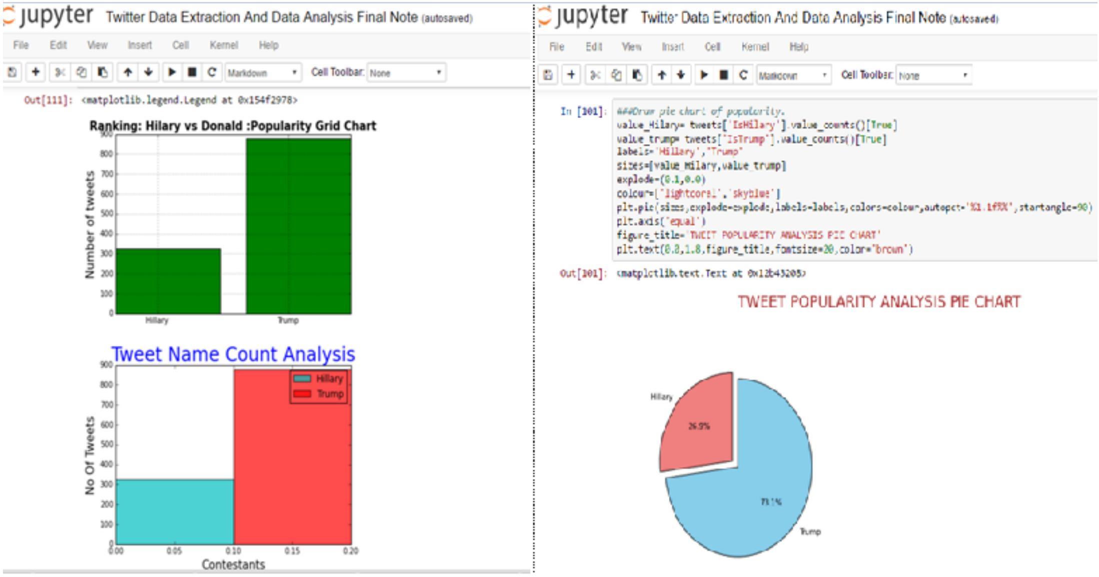

jupyter
Twitter Data Extraction And Data Analysis Final Note (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
Markdown
Cell Toolbar:
None

Out[75]:

	lang	text	IsHilary	IsTrump	IsHillary
0	und	@GeorgeTakei @realDonaldTrump https://t.co/mh6...	NaN	NaN	NaN
1	en	RT @ChrisKlemens: .@realDonaldTrump you've tak...	False	True	False
2	en	RT @mcforgod: .@realDonaldTrump .@HilaryClinto...	False	True	False
3	en	DO whats RIGHT\nSaveUSA &amp;OurKids FUTURE\nV...	False	True	False
4	en	RT @LifeforceVibes: #Berners conducting a #Sil...	True	False	True
5	en	this is why you were my fav character on one t...	False	False	False
6	en	@realBrianBuxton @realDonaldTrump hello random...	False	True	False
7	en	RT @MafazAlSuwaidan: While @realDonaldTrump ig...	False	True	False
8	en	RT @SophiaBush: @realDonaldTrump shut the actu...	False	True	False
9	en	RT @WayneDupreeShow: @realDonaldTrump wants to...	True	True	True

Figure. 3. Result of Analyzed tweets between Hilary and Trump

The count of the respective tweets between Hillary and Trump is given in figure 4.



```
Out[111]: <matplotlib.legend.Legend at 0x154f2978>
```

```
In [101]: ##Draw pie chart of popularity.
          value_Hilary= tweets['IsHilary'].value_counts()[True]
          value_trump= tweets['IsTrump'].value_counts()[True]
          labels='Hillary',"Trump"
          sizes=[value_Hilary,value_trump]
          explode=(0.1,0.0)
          colour=['lightcoral','skyblue']
          plt.pie(sizes,explode=explode,labels=labels,colors=colour,autopct='%1.1f%%',startangle=90)
          plt.axis('equal')
          figure_title='TWEET POPULARITY ANALYSIS PIE CHART'
          plt.text(0.2,1.8,figure_title,fontsize=20,color='brown')

Out[101]: <matplotlib.text.Text at 0x12b43205>
```

Figure 4. Result of Count of tweets between Hilary and Trump

The results of 10 fold execution on linear regression pattern are in Table 2. The accuracy of our system is 81.51%. figure 4.represents the count of tweets collected from Jan 2016 to Jan 2017.

Table 2 Result of 10-fold execution

Fold	Accuracy
2	83.09
3	82.89
4	82.11
5	84.31
6	86.60
7	85.30
8	85.45
9	84.11
10	88.91
Average	85.23

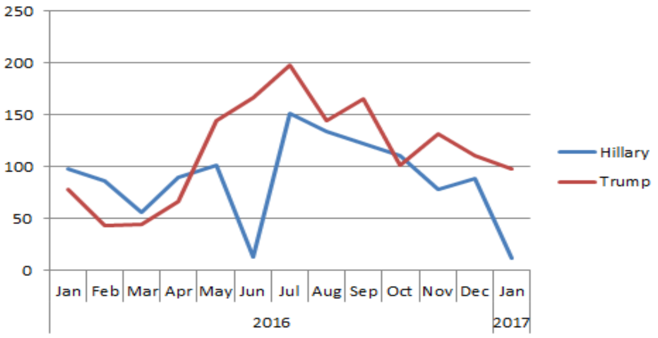


Figure 5. Tweet count between Hillary and Trump

5. Conclusion

Sentiment Analysis of twitter tweets is made to find the users interest over Hillary and Trump. We have used linear regression approach to predict the polarity of the tweets. This data analytic approach performs better than support vector machine and naïve bayes approach. The accuracy of our approach is 85.23% which is shows a significant increase than SVM applied in [13]. We have applied 10 fold cross validation to improve the accuracy of our system. Linear regression is suitable for predictive analysis than other data analytic approaches. The system can be extended for prediction in topic classification.

References

[1] Xia F, Wang W, Bekele TM, Liu H. Big Scholarly Data: A Survey. IEEE Transactions on Big Data. 2017 Mar 1;3(1):18-35.

[2] Sun J, Shen L, Ding G, Li R, Wu Q. Predictability Analysis of Spectrum State Evolution: Performance Bounds and Real-World Data Analytics. IEEE Access. 2017;5:22760-74.

[3] Borruto G. Analysis of tweets in Twitter. Webology. 2015 Jun 1;12(1):1..

[4] Burke-Garcia A, Stanton CA. A tale of two tools: Reliability and feasibility of social media measurement tools examining e-cigarette twitter mentions. Informatics in Medicine Unlocked. 2017 Dec 31;8:8-12.

[5] O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015 May 31;2(2):183-8.

[6] Ahmad N, Siddique J. Personality Assessment using Twitter Tweets. Procedia Computer Science. 2017 Jan 1;112:1964-73.

[7] Weilin L, Hoon GK. Personalization of trending tweets using like-dislike category Model.Procedia Computer Science. 2015 Jan 1;60:236-45.

[8] Kharde V, Sonawane P. Sentiment analysis of twitter data: A survey of techniques. arXiv preprint arXiv:1601.06971. 2016 Jan 26.

[9] Spencer J, Uchyigit G. Sentimentor: Sentiment analysis of twitter data. InProceedings of European conference on machine learning and principles and practice of knowledge discovery in databases 2012 (pp. 56-66).

[10] Jhawar SR, Prabhu V, Katz MS, Motwani SB. Tweet for the cure: a Snapshot of Twitter Usage by Three US Oncologic Professional Societies. Advances in Radiation Oncology.2017 Jun 13.

[11] Oku K, Hattori F, Kawagoe K. Tweet-mapping method for tourist spots based on now-tweets and spot-photos. Procedia Computer Science. 2015 Jan 1;60:1318-27.

[12] Kumar S, Morstatter F, Liu H. Twitter data analytics. New York: Springer; 2014.

[13] Barfian E, Iswanto BH, Isa SM. Twitter Pornography Multilingual Content Identification Based on Machine Learning. Procedia Computer Science. 2017 Dec 31;116:129-36.