2016 president election prediction using twitter

Weifeng Li#1

# Department of Computer Science and Engineering, University of Bridgeport  
Bridgeport, CT, US

1 weifli@my.bridgeport.edu

Abstract—In data mining area, social media data mining is getting more and more hot with the improvement of software and hardware as well as the newer achievements of related research area have been accomplished. Sentimental data mining is one of the branches in social media data mining. To do sentimental data mining using social media, several important things should be considered. First of all, how to design and implement a complete data-mining system including data collection, data-cleaning, data-store, data-mining, model-evaluation etc. is one of the most important parts. Second, as we know, several classification algorithms can be used to perform sentimental data mining such as Naïve Bayes, Support vector machines, Decision trees, Neural networks etc. How to choose the right algorithms and using the appropriate parameters are among the most important parts. Last but not least, finding a good way to visualize the result is also very important. In this paper, all these contents mentioned above will be talked about in detail.

1. Introduction

Nowadays, data of social medial websites are getting more and more popular to be used as one of the most important data source for the data mining from which we can find the useful and interesting patterns. In this project, base on twitter data set that I collect using twitter API, I performed the sentimental mining and topic modeling. In the data collection phase, I used keywords such as the candidates’ name to filter the related data decreasing the noise to the most extend. To accomplish the sentimental mining, I chose Naïve Bayes algorithm and Support vector machine Model(SVM) two of the most commonly used algorithms that can be used as the classifier in the sentimental analysis. Then I trained these classifiers using a data set which was also from twitter and was related to 2016 presidential election from Kaggle and made the predication using twitter data set that I collected. Besides, Latent Dirichlet Allocation model was used to fulfil the topic modeling analysis finding the most frequent topics from the data of presidential election related tweets. At last, I evaluated the performance of each classification algorithm.

1. Dataset

Two data sources were chosen by us as the input of our data-mining project. First data-set we used are from twitter. To collect twitter data, I used tweepy one python model that wraps twitter API for its easiness for using and its open source. As the 2016 presidential election is getting more and more clear, other candidates in GOP will hardly threaten Donald Trump’s front-runner status. At the same time, in democratic, only Hillary Clinton and Bernie Sanders will compete the presidential candidacy. Hence, I just chose to collect twitter related these three persons by key of ‘hillary clinton’, ‘bernie sanders’, and ‘donald trump’. During a period of one week, I collected about 100,000 tweets. After the phase of pre-processing there were about 10,000 records left. The second dataset is from <https://www.kaggle.com>. It is the result of analysis of tweets on the first 2016 GOP Presidential Debate. we Part of the most import attributes of the second dataset are shown as below.

TABLE I  
Attributes for Dataset

|  |  |  |
| --- | --- | --- |
| **ID** | NAME | Description |
| 1 | ID | Tweets ID |
| 2 | Candidate | Candidate name |
| 3 | Text | Content of Tweets |
| 4 | Sentiment | Sentiment of tweet |
| 5 | Confident | Confident of sentiment |

Our goal is to train our model using this dataset and then do the prediction using the trained-model for the first data set.

1. Problem Definition

To achieve our goal, we try to solve four problems in our project.

* Firstly, using sentimental analysis, based on the twitter data, the popularities of each candidate will be come up with. In this phase, two calcific algorithms Naïve Bayes and SVM will be used as the classifier and performance of them will be evaluated at the end.
* Secondly, I will do the parameter tuning for the algorithms chosen to find the best parameters for each algorithms.
* Third, I will evaluate the performance of each algorithm. To do this, Confusion Matrix was used here .
* Besides, I will use Latent Dirichlet allocation(LDA) model extracting frequent topics related to 2016 presidential election using twitter data set. For better illustration of the results, I will visualize the results in World Cloud.

1. ALGORITHMS Introduction
2. Naïve Bayes

Naive Bayes classifier is a very simple classification algorithm, call it naive Bayes classifier is because the idea of this method is really very simple, naive Bayes ideological foundation is this: For the classification to be given items, solving the probability of each category appear under the conditions of this appears, the largest of which, it considers this to be sorted items belong to which category. Popular, like such a reason, you see a black man in the street, I ask you, where do you think this man came, you probably guessed Africa. why? Because the highest rate of black Africans, of course, people may be American or Asian, but no other information is available, we will choose the largest category of conditional probability, which is the ideological foundation Naive Bayes.

Suppose an individual to have n items feature (Feature), respectively, F1, F2, ..., Fn. M existing categories (Category), respectively, C1, C2, ..., Cm. Bayesian classifier is to calculate the probability that the largest category, which is seeking the maximum of the following formula:

P(C|F1F2…..Fn) = P(F1F2….Fn|C) / P(F1F2…Fn)

Since P (F1F2 ... Fn) are the same for all categories can be omitted, the question becomes seeking the max value of P(F1F2….Fn|C).

Assuming that all the features are independent, we can get

P(F1F2….Fn|C)P(C) = P(F1|C)P(F2|C) … P(Fn|C)P(C)

Each item on the right-hand side of the equation, can be obtained from the statistics, which can calculate the probability of each category, and to find the maximum probability of that class.

Although "all the characteristics independently of each other," this assumption is unlikely to set up in the real world, but it can greatly simplify the calculation, and studies have shown little effect on the accuracy of the classification results.

1. Support Vector Machines

Support Vector Machines(SVM) are a serial of supervised data mining methods that be used for classification.

The advantages of SVM:

* The relatively effectiveness in high dimensional spaces.
* Higher effectiveness in cases where number of dimensions is greater than the number of samples.
* Memory efficient.
* Versatile: different Kernel functions can be used for the decision function.

The disadvantages of SVM:

* In the scenario of high features density, the performance is likely to be poor.
* SVMs do not directly provide probability estimates, these are calculated using an expensive file-fold cross-validation.

.

1. Design and Implementation

To address the four problems described above, there are two main steps: A) data-collection B) pre-processing C) model training D) parameter tuning E) prediction

1. Data Collection

In this Phase, I used tweepy stream API collecting totally 100,000 tweets which are in format of JSON as shown below.

|  |
| --- |
| ../../../../../../Desktop/Screen%20Shot%202016-04-17%20at%204.26. |

For convenient, I converted the JSON files into CSV format and removed all the fields that cannot be used, as shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Id** | **candidate** | **text** | **Time** | **Place** |
| 710277894539513857 | Hilary Clinton | RT @NancyLeeGrahn: How did everyone feel about the Climate Change question last night? Exactly. #GOPDebate | Thu Mar 17 01:33:41 +0000 2016 | "Atlanta, GA" |
| 710277896791859200 | Donald Trump | RT @DanScavino: #GOPDebate w/ @realDonaldTrump delivered the highest ratings in the history of presidential debates. #Trump2016 http://t.coâ€¦ | Thu Mar 17 01:33:41 +0000 2016 | "Atlanta, GA" |

1. Pre-processing

As show in the architecture of the system, in the pre-processing phase, Hyperlink removal, special characters filtering, and sentence splitting will be conducted for the text field whose result will be the input of the classifier.

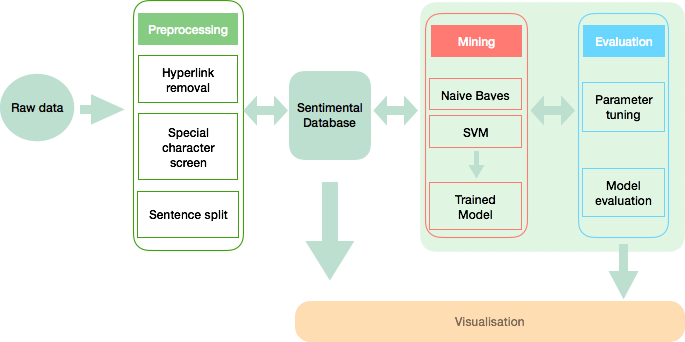


Fig. . System Architecture

1. Model training

During this phase, ‘the result of analysis of tweets on the first 2016 GOP Presidential Debate’ dataset from Kaggle were used as the training data set to conduct the model training. At first, I used all the data set as the training data set. For both Naïve Bayes and SVM, I got the confident not more than 65%. To improve the confident, I only chose the data which sentiment was 100%. The results were improved greatly. The confident of Naïve Bayes and SVM respectively reached to about 75% and 80%.

1. Parameter tunning

Using parameter tunning technologies, I completed parameter tunning in this phase and found the best parameters for each algorithms as shown below.

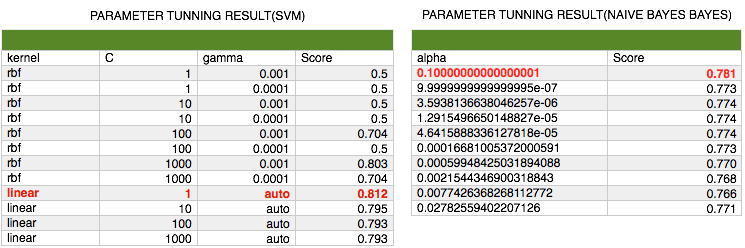


Fig. . System Architecture

To find the best parameters for these classification algorithms, I used cross-validation method for the parameter tuning. As can be seen below, throw this way, the best kernel parameter for SVM and the best alpha parameter for Naïve Bayes were found.

1. Prediction

In this phase, I used the trained classifier to predict the sentiment of the input twitter data set.

1. Result

Using visualization tools we visualized the result as shown below:

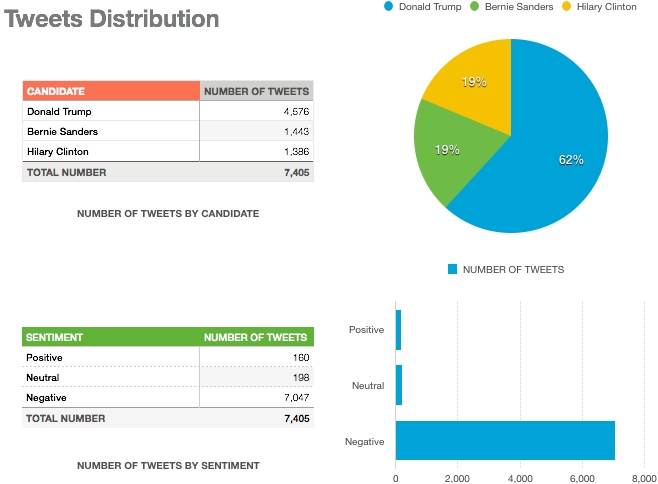


Fig. 3. Total Number of Crimes by Year

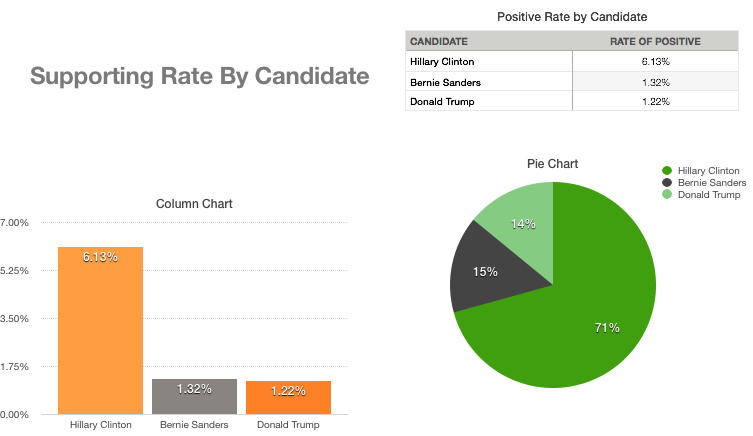


Fig. 4. Number of Crimes in Different Time Period

We can see that several interesting things from Fig. 3 and Fig. 4. Firstly, the overwhelming number of tweets are Negatives. Secondly, Donald Trump was more popular than the other candidates. Last, although Trump was more popular, Hilary was more possible win the election.



Fig. 5. Word Cloud for 2016 presidential election

Using Latent Dirichlet allocation(LDA) model algorithm, I extracted the most frequent topics related to 2016 presidential election using twitter data set. For better illustration of the results, I will visualize the results in World Cloud.

1. Performance Evaluation

To evaluate the accurate performance, Confusion Matrix was used here. As shown below, the accuracy of Linear SVC classifier was better than Naïve Bayes.

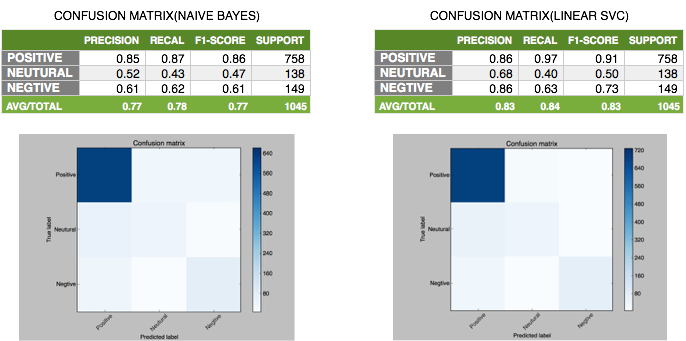


Fig. 6. Proportions of different Income level

1. Conclusion

As social media such as twitter has not been systematically studied yet, with the development of science and technology especially in big data and machine learning area, more and more new methodologies and algorithms can be used to find the interesting patterns from social media. I use semantic analysis and topic modeling methods mining the twitter data set and find the similar result comparing with the result from other data set such as polls and final contribution data set. The result show that, we can find more useful and interesting pattern and information from social media data mining.

.

Acknowledgment

We would like to thank our Professor Lee for stimulating discussions on this class project and for helpful comments on how to conduct good work.

References

1. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [Online]. Available: <http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fbb21d-4b90-81d4-d909fdb96b87.pdf>
2. Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [Online]. Available at <http://cordis.europa.eu/fp7/ict/einfrastructure/docs/hlg-sdi-report.pdf>
3. Y.Demchenko, P.Membrey, P.Grosso, C. de Laat, “Addressing Big Data Issues in Scientific Data Infrastructure,” in First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.
4. NIST Big Data Working Group (NBD-WG). [Online]. Available: <http://bigdatawg.nist.gov/home.php>
5. Definting Big Data Architetcure Framework: Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013. Presented at NBD-WG, 24 July 2013 [Online]. Available: <http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf>
6. Reflections on Big Data, Data Science and Related Subjects. Blog by Irving Wladawsky-Berger. [online] Available <http://blog.irvingwb.com/blog/2013/01/reflections-on-big-data-datascience-and-related-subjects.html>
7. E.Dumbill, What is big data? An introduction to the big data landscape. [Online]. Available: <http://strata.oreilly.com/2012/01/what-is-bigdata.html>
8. The Big Data Long Tail. Blog post by Jason Bloomberg,January 17, 2013. [Online]. Available: <http://www.devx.com/blog/the-big-data-longtail.html>
9. J.Gantz and David Reinsel, Extracting Value from Chaos, IDC IVIEW, June 2011. [Online]. Available: <http://www.emc.com/collateral/analystreports/idc-extracting-value-from-chaos-ar.pdf>
10. The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>
11. Big Data defintion, Gartner, Inc. [Online]. Available: <http://www.gartner.com/it-glossary/big-data/>
12. S.Sicular, “Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s”, Gartner, Inc. 27 March 2013. [Online]. Available: http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners-big-data-definition-consists-of-three-parts-not-to-be-confusedwith-three-vs/
13. J.Layton, “The Top of the Big Data Stack: Database Applications”, July 27, 2012. [Online]. Available: <http://www.enterprisestorageforum.com/storage-management/the-top-ofthe-big-data-stack-database-applications.html>
14. Explore big data analytics and Hadoop. [Online]. Available: <http://www.ibm.com/developerworks/training/kp/os-kp-hadoop/>
15. A.Bloom, 7 Myths on Big Data: Avoiding Bad Hadoop and Cloud Analytics Decisions, April 22, 2013. [Online]. Available: <http://blogs.vmware.com/vfabric/2013/04/myths-about-running-hadoopin-a-virtualized-environment.html>
16. European Union. A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe. Brussels : European Commission, 2012. Final Report. Contributing author. Internal identification SMARTNr 2011/0056. [Online]. Available: Available at <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-finalreport.pdf>
17. Y.Demchenko P.Membrey, C.Ngo, C. de Laat, D.Gordijenko., Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, Proc. Secure Data Management (SDM’13) Workshop. Part of VLDB2013 conference, 26-30 August 213, Trento, Italy
18. NIST SP 500-292, Cloud Computing Reference Architecture, v1.0. [Online]. Available: [http://collaborate.nist.gov/twiki-cloudcomputing/pub/CloudComputing/ReferenceArchitectureTaxonomy/NIS T\_SP\_500-292\_-\_090611.pdf](http://collaborate.nist.gov/twiki-cloudcomputing/pub/CloudComputing/ReferenceArchitectureTaxonomy/NIS%20T_SP_500-292_-_090611.pdf)
19. Y.Demchenko, M. Makkes, R.Strijkers, C.Ngo, C. de Laat, Intercloud Architecture Framework for Heterogeneous Multi-Provider Cloud based Infrastructure Services Provisioning, The International Journal of NextGeneration Computing (IJNGC), Volume 4, Issue 2, July 2013
20. NIST Big Data Reference Architecture. NBD-WG, NIST [Online]. Available: <http://bigdatawg.nist.gov/_uploadfiles/M0226_v10_1554566513.docx>
21. NIST Big Data Technology Roadmap. NBD-WG [Online]. Available: <http://bigdatawg.nist.gov/_uploadfiles/M0087_v8_1456721868.docx>
22. Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0. [Online]. Available: http://www.opendatacenteralliance.org/docs/ Information\_as\_a\_Service\_Master\_Usage\_Model\_Rev1.0.pdf