International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-5 Issue-7, December 2015

Profiling and Jury Selection Using Sentiment Analysis

Rijuta Wagh, Janvi Shah, Khyati Shah, Sindhu Nair

Abstract*— Jury Selection is the process of selecting 12 jury members from a pool of random people. These selected Jurors attend the trial proceedings and after the closing statements give a verdict on whether the defendant is guilty or not. For a defendant to be pronounced guilty or not guilty the jurors must unanimously vote on it. If there isn't a unanimous vote, then there is a mistrial. A mistrial can mean the whole case being restarted or the case being retired, meaning the case will not be pursued further. Thus the selection of the correct jurors is paramount to a decision in our favor, whichever side we may represent. We aim to develop a model in which the opinion of Twitter users is analyzed to create demographics which the lawyer can use for jury selection. Upon extracting data from Twitter based on hash tags pertaining to a certain case, the data undergoes an extensive cleaning process. We first classify the people according to age, sex, and profession and then plot graphs that can be statistically compared. This helps lawyers to make informed decisions and select a jury favorable to his/her case.

***Keywords*— Maximum Entropy, Naïve Bayes, Neural networks Sentiment Analysis, SVMs**

I. INTRODUCTION

A. Description

Social media enables engagement and conversations of all types among all of the peoples of the world that have access to Internet connections. Social media concerns communities of people who gather online in one form or another to socialize, communicate with one another, and share information and opinions. As such, social media is the ideal online platform for people (potential jurors) to deliberate – i.e. converse – about court cases. These online social communities are made up of hundreds of millions of people and lawyers can tap into them to get an insight into the attitudes and opinions of the general public. However, analyzing each post/tweet and finding patterns manually is a tedious and time consuming task. Also, each lawyer is allowed to ask only one question to the potential jurors who make the task of jury selection all the more difficult [1]. Twitter is one of the most widely used platforms and hence an ideal tool to study the sentiments. An intelligent tweet analyzing system finds trends in the attitudes of people using Data Mining and the graphical results can be an ideal tool for lawyers to gain advantage of the large amount of information social media provides without wasting much time.

Manuscript published on 30 December 2015.
***Correspondence Author(s)**
Rijuta Wagh, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai (Maharashtra), India.
Janvi Shah, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai (Maharashtra), India.
Khyati Shah, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai (Maharashtra), India.
Prof. Sindhu Nair, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai (Maharashtra), India.

B. Problem Formulation

Lawyers are given a pool of people selected randomly and have to form a jury of 12 members [1]. This is done to ensure that the jury is unbiased. However, it is the lawyer's task to select a jury favorable to his/her case. The lawyer is allowed to ask only one question which makes it difficult to gauge the attitude of the person towards the case. Our system displays demographics based on age, sex, and profession that help the lawyer decide what question to ask, and hence, whether to select a person or not.

II. LITERATURE REVIEW

The ability to extract insights from social media using sentiment analysis approach is a practice that is being widely adopted by organisations across the world. Before social media, trial consultants were restricted, on a practical basis, to a finite number of focus groups (termed "jury simulations") comprised of a finite number of participants (termed "surrogate jurors")[1]. Such jury simulations, repeated over and over, could provide excellent qualitative information about the predilections of jurors concerning how they would think about and decide the case. But now, thanks to the availability of online social communities made up sometimes of hundreds of millions of people, along with online social media research that can tap into these vast communities, the number of people trial consultants can test to gauge their attitudes and opinions concerning litigation disputes is virtually limitless. Therefore, trial consultants can now provide legitimate quantitative (as well as qualitative) information about jurors' potential attitudes and opinions. This is a remarkable breakthrough for trial consultants – and for the lawyers they counsel and assist. It provides extraordinary advantages to the attorneys wise enough to take advantage of this remarkable litigation intelligence. Thus, we realised that public sentiment analysis is critical as the results can be strategically used to effectively present demographic evidence to appeal to juror emotions. Technology allows a trial team to "data mine" before, during and after a trial. Through data mining, one can find and address the different sentiments possible among the masses and use that knowledge to drive the demonstrative evidence presented at trial by selecting a jury panel. There are various classification algorithms, which have been employed to aid opinion mining. These algorithms vary from simple probabilistic classifiers such as

Naïve Bayes (probability classifier that assumes all the features are independent and does not use any prior information) to the more advanced classifiers such as maximum entropy (which uses the prior information to a certain extent). Many hyperspace classifiers such as Support Vector Machine (SVM) and Neural Networks (NN) have also been used to correctly classify the sentiments [4]. Between SVM and NN, SVM, in general, works wonders due to the kernel trick. Ren-nie et al discussed the performance of Naïve Bayes on text classification tasks in their 2003 paper. [9] (Pang and Lee 2002) researched the effects of various machine learning techniques (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM)) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model. [9] Researchers have also worked on detecting sentiment in text. (Turney 2002) presents a simple algorithm, called semantic orientation, for detecting sentiment. (Pang and Lee 2004) present a hierarchical scheme in which text is first classified as containing sentiment, and then classified as positive or negative.[6] Work (Read, 2005) has been done in using emoticons as labels for positive and sentiment. This is very relevant to Twitter because many users have emoticons in their tweets.

III. PROPOSED TECHNIQUES

We will look into the following two techniques which we will be using in our project for opinion extraction and determining the sentiment (positive or negative) about a particular case.

A. *Naïve Bayes Classifier*

A Naive Bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Naïve Bayes model involves a simplifying conditional independence assumption. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. In our case, the maximum likelihood probability of a word belonging to a particular class is given by the expression

$$P(x_i|c)=\frac{Count\ of\ x_i\ in\ documents\ of\ class\ c}{Total\ no\ of\ words\ in\ documents\ of\ class\ c} \quad (1)$$

The frequency counts of the words are stored in hash tables during the training phase. According to the Bayes Rule, the probability of a particular document belonging to a class c_i is given by,

$$P(C_i|d) = \frac{P(d|c_i)*P(c_i)}{P(d)} \quad (2)$$

If we use the simplifying conditional independence assumption, that given a class (positive or negative), the words are conditionally independent of each other. Due to this simplifying assumption the model is termed as "naïve".

$$P(C_i|d)=\frac{(\pi P(x_i|c_j))*P(c_j)}{P(d)} \quad (3)$$

Here the *xis* are the individual words of the document. The classifier outputs the class with the maximum posterior probability. We also remove duplicate words from the document, they don't add any additional information; this type of naïve bayes algorithm is called Bernoulli Naïve Bayes. Including just the presence of a word instead of the count has been found to improve performance marginally, when there is a large number of training examples.

B. *Support Vector Machine*

Vapnik [4] proposed Support Vector Machine (SVM) that belongs to supervised learning method which classifies the data into two categories by constructing the N-dimensional hyper plane. SVM uses g(x) as the discriminate function,

$$g(x) = w^T f(x)+b \quad (4)$$

Where w is the weights vector, b is the bias, and f(x) denotes nonlinear mapping from input space to high-dimensional feature space. The parameters w and b are learned automatically on the training dataset following the principle of maximized margin b

$$min \frac{1}{2} W^T W + C \sum_{i=1}^{N} ci \quad (5)$$

Where N denotes the slack variables and C denotes the penalty coefficient. Due to the dimension of feature space is quite large in text classification task, the classification problem is always linearly separable [4] [9] and therefore linear kernel is commonly used.

IV. SYSTEM ARCHITECTURE

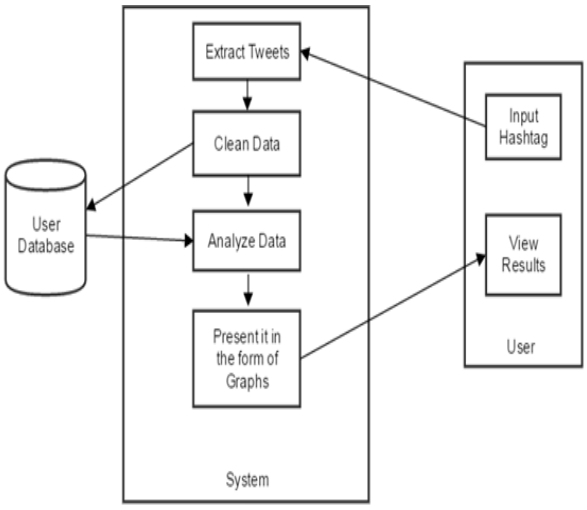

Fig. 1. System Architecture

1. Gaining Social Network Data:

In this module, we gain the data from a social networking site, namely Twitter, including information about the user's profile, inter-links, tweets based on hash tags and activity for a given time period. Over a span of time, the search query is shot and data is collected. The expected outcome is a network of users with detailed information about each, complete with re-tweeters' information.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

2. Cleaning the data obtained:
This step involves cleaning the data in the database. This is important because many times data extracted from Twitter contains discrepancies and minor errors. Therefore, these errors are corrected so that the statistics of opinions is not compromised.

3. Classifying into different categories:
In this step we classify each user into a separate category based on age, sex, location. The categories are male/female/other, 0-20/21-40/40-60/>60.

4. Ascertaining opinions:
The tweets of each category are mined and the opinion expressed in each tweet is ascertained. A percentage ratio for each opinion is calculated for each group. The expected result is the division of each sub-category according to the percentage of people having certain opinion.

5. Mapping data to a graph for analysis:
We subsequently map the statistics obtained in the previous module in a graphical format so as to enable analysis by the lawyer. This allows for easy comparison between different categories.

V. CONCLUSION

We believe that a system like this has been needed for a long time. In this project we aim to present an analysis of online social network content in the context of ongoing lawsuits. We aim to profile users and interactions in a targeted Twitter group using an application of sentiment and social network analysis techniques and based on the results of that we will present a demographic of the beliefs of the public based on various categories like age, sex, etc. This information can be used by lawyers during the jury selection to help them get a sympathetic jury on the panel to judge their case. A system like this is highly helpful since the Lawyers currently rely only on the question they ask each juror and on the influence they have on their social circle to determine the attitude of a person. This approach takes into account the attitude of only a negligible amount of people and may also not be useful since a lawyer may not have any direct/indirect connection with the juror. A better approach would be to consider a larger amount of people which can potentially include the jurors. This is exactly what our project does and gives a more accurate judgement of opinions. Use of electronic media is increasing day by day. Instead of spending time in reading and figuring out the positivity and negativity of text, we can use automated techniques for sentiment analysis. Our project currently takes into account extraction from Twitter using the API over a predefined set of time. Therefore, this leads to a restriction on the number of tweets we can garner. A higher number of tweets extracted will surely yield more definite results. Also, extraction done over a period of time ranging into months can ensure an even higher rate of accuracy. We have also placed a restriction on the number of hash tags we considered. However, an increase in the number of hash tags will yield a more comprehensive list of potential jurors. New algorithms enabling extraction of age, gender, occupation, etc. can also increase the accuracy. Our project can be personalized for each lawyer and take into account the lawyer's strategy for the case.

REFERENCES

1. http://www.thejuryexpert.com/2011/09/trial-strategy-using-social-media-analytics/
2. Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena, Nov 2009.
3. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). *Learning Word Vectors for Sentiment Analysis.* The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).
4. Ms. K. Nirmala1 Devi, Ms. K. Mouthami, Dr. V. MuraliBhaskaran'*Sentiment Analysis and Classification Based on Textual Reviews'*,2012.
5. Kennedy, Alistair, and Diana Inkpen. "*Sentiment classification of movie reviews using contextual valence shifters*." Computational Intelligence 22.2 (2006): 110-125.
6. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86,2002.*
7. A. Esuli and F. Sebastiani. *Sentiwordnet: A publicly available lexical resource for opinion mining*. pages 417–422,2006.
8. Tong. An operational system for detecting and tracking opinions in on-line discussion. In Proceedings *SIGIR 2001 Workshop on Operational Text Classification, 2001*.
9. H. Witten & E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann,2nd edition,2005.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)