CSOHS-Portal

Pre-School Progress System

Sheila Mae Avila
Department of Computer
Science,
Ateneo de Naga University
Ateneo Avenue, Naga City,
Philippines, 4400
0906-466-5454
shavila@gbox.adnu.edu.ph

Stephen Kent Sipalay
Department of Computer
Science,
Ateneo de Naga University
Ateneo Avenue, Naga City,
Philippines, 4400
09165471327
ssipalay@gbox.adnu.edu.ph

Zandra Mae Cu
Department of Computer
Science,
Ateneo de Naga University
Ateneo Avenue, Naga City,
Philippines, 4400
0906-466-5454
zcu@gbox.adnu.edu.ph

Phil Ervie John Teoxont
Department of Computer
Science,
Ateneo de Naga University
Ateneo Avenue, Naga City,
Philippines, 4400
0916-120-3544
pteoxon@gbox.adnu.edu.ph

ABSTRACT
Social media has been an outlet for some. Sentiment Analysis is a way to automate in mining, extracting and processing information to determine the general opinion of the people. The sentiment analysis of the the seven banned countries suggests to know the general opinion what methods are accurate to extract subjective influence from twitter. The goal of this research was to know the sentiments or opinions of the people regarding the banning of six muslim-majority countries. Tweets will be collected, and classify as objective - tweets that has no emotion or subject. To test, the researcher will implement it using rapidminer to be processed and analyzed for accuracy.

Keywords
US, Banned Countries, Muslim, Trump, Sentiment Analysis, LaTeX, text tagging

1. INTRODUCTION
On the 20th day of January, 2017, Donald Trump became the 45th United States President, succeeding Barack Obama[7]. Less than a month after he was elected, the new US President imposed a ban for travellers from seven countries in the world such as Iraq, Syria, Iran, Libya, Somalia, Sudan and Yemen[3]. But after the ban was lifted by the court, US President Donald Trump has signed a new executive order that bans immigration from six Muslim-majority countries, dropping Iraq from previous order. The new travel ban comes six weeks after Trump's original executive order caused chaos at airports nationwide[2]. President Trump temporarily banned of six countries from entering the US. However, the new ban exempts citizen of the banned countries who are legal US permanent citizen or have valid visas to enter the US. The order exempts diplomats and member of international organizations. These restrictions were part of wide ranging immigration controls that also suspended refugee arrivals.[1] The citizens of banned countries, will be blocked from obtaining visas for at least 90 days. The order also suspends admission of refugees into the US for 120 days. One of the reasons why these countries were picked, was because they either had significant terrorist organization presence.

Twitter is a social networking service where users can post and interact with other users. Registered users can able to post restricted 140 characters called as tweet and those who are not registered can only read them. Twitter was created in year 2006 by Jack Dorsey, Noah Glass, Biz Stone and Evan Williams. It was launched in public on July 2006. The concept caught fire and gained popularity worldwide. In year 2013, it was one of the most-visited websites worldwide.[6] The day of the 2016 U.S presidential election, Twitter once again proved to be the largest source of breaking news with millions of tweets on that day. In the last reported quarter, the number of monthly active U.S. Twitter users amounted to 67 million. With more than 310 million monthly active users worldwide as of the first quarter of 2016, Twitter is one of the biggest social networks worldwide.[4] Twitter can be a way to help the researchers to evaluate the polarity sentiments based on the words of the user's tweets. With its restricted 140 words it is easier to perform sentiment analysis for it is easier to read and its content is more concentrated to the issue plus it's hashtags, emoticons and jargons.[5]

Donald Trump's decision made a huge impact and cause people to react on this matter. Regarding this issue, the researchers came up with a research question of what are the sentiments of the people about the controversial banning. This study wants to determine the opinions if they are in favor or not. The researchers want to study the opinions expressed of the people regarding President TrumpâĂŹs banning of countries. In the next chapter the researchers will be showing the answers by using twitter. The twitter will be our main source to support our research.

2. REVIEW OF RELATED LITERATURE

For the related literature, the researchers will examine the literature behind the banning of the six countries. Next the researchers will determine its pros and cons of banning the said countries. Lastly, we look at literature that proposes some experimental methods of the result of the issue.

2.1 Sentiment Analysis

Sentiment analysis is a form of shallow semantic analysis of texts. Its goal is to extract opinions, emotions or attitudes towards different objects of interest. From the first approaches in 2000s, sentiment analysis gained considerable attention with massive growth of the web and social media. Different forms of textual information are becoming easily accessible (e.g., news, blogs, reviews, Facebook comments, Twitter posts, etc.), and different approaches to sentiment analysis were developed MozetiÄ∎, Igor et al (2016). [8] Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources. Sentiment analysis involves classifying opinions in text into categories like âĂIJpositiveâĂİ, âĂIJnegativeâĂİ or âĂIJneutralâĂİ.[9]

2.2 Sentiment Analysis in Twitter

Sentiment analysis in Twitter is a field that has recently attracted research interest Giachanou and Crestani (2016).[10] Sentiment analysis in Twitter tackles the problem of analyzing the tweets in terms of the opinion they express. In addition, we discuss fields related to sentiment analysis in Twitter including Twitter opinion retrieval, tracking sentiments over time, irony detection, emotion detection, and tweet sentiment quantification, tasks that have recently attracted increasing attention. The main form of media in the social network and the main microblog abroad, Twitter attracts people more and more to Zhibo Wang et.al (2016).[11] Tweets contain different tendencies and emotional characteristics; and mining these features is meaningful for public opinion monitoring, marketing, and rumor control. In general, most emotional analysis only divides the text emotion into three categories: neutral, positive, and negative. It is limited to help people listen to the real voice and emotion of society.

2.3 Corpus Collection and Preparation

The researchers collected 1000 tweets consisting 141 positive, 541 objectove and 318 negative tweets.

2.4 Preprocessing

2.4.1 Tokenization

Character that is non-arabic has to be removed. Digit, punctuations, marks, symbols and special character has to be removed.

2.4.2 Normalization

Converting to UTF-8 recording, removing diacritics and Arabic tatweel(-), converting (ØŇ Øĕ Øć ØŇ Øč) to (Alif, Øğ) and changing (Yaa,ÙĽ) to (ÙĽ) and finally converting (Øĺ) to (ÙĞ). The aim of this step is to unify words that is typed differently.

2.4.3 Stop Words Removal

In Arabic words such as "ÙĄÙĽ" (in), "ØźÙĎÙĽ" (on), are words that carry no information. An added additional stop words for the Arabic slang. Stop words for the Arabic slang are like (.... (ØŕÙĞ, ØŕÙĽ, ØğÙĎÙĎÙĽ, ØĺØş) (Da, de, Elly, bas.....). These words are usually removed using stop words list.

2.4.4 Stemming

Process of reducing words to their stem or root from where morphological information is used to match different variants of words.

2.5 Text Classification

2.5.1 Support Vector Machine (SVM)

A set of related supervised learning methods used for classification and regression. In simpler words, a given set of training examples marked as belonging to one of two categories. SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. The operation of the SVM algorithm is based on finding the hyperlane that gives the largest minimum distance to training examples. Mathematically, SVMs learn the sign function: $F(x) = sign (wx + b)$, where w is a weighted vector in Rn an b is known as the bias. SVM find the hyperlane $y = wx + b$ by separating the space Rn into two half spaces with the maximum margin.

2.5.2 Naive Bayes (NB)

Naive Bayes classifier makes the use of the joint probabilities of terms and their categories to figure out the probabilities of categories given as a test data. NB are easy to implement and have better performance algorithm. Given a vector of words w, c the target classes can be estimated as P (w|c)1 + n (w,c)|w| + n (c), where n(w,c) is the number of the word positions that are occupied by w whose class value is c. n(c) is the number of word positions whose class value is c. |w|is the total number of distinct words.

3. METHODOLOGY

This section discusses the actions or methods needed or used by the researchers in order to accomplish the sentimental analysis in the banning of six Muslim-majority countries. This section discusses what program is used to collect tweets that will be used as a training sets, keywords used by the researchers to collect and download twitter data regarding the topic, system design, numbers of tweets collected, as well as the date of when was the tweets collected. Twitter data create a rich source that can be used for capturing information about any possible topic. This data can be used
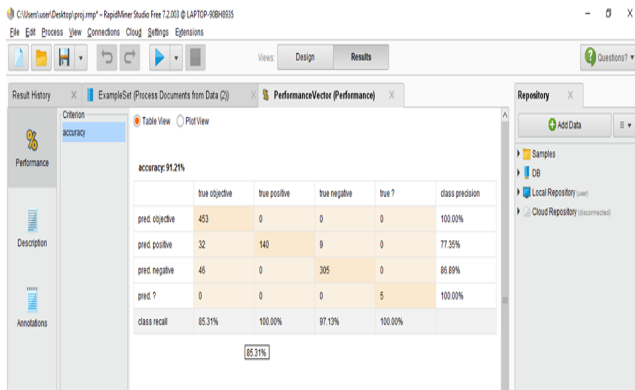
Figure 1: twitter tweepy.py [12]

Figure 2: twitter process.py [12]

Figure 3: Rapid Miner

in different cases such as finding the trends today related to specific keyword/s, gathering feedbacks and analyzing sentiments about a certain topic. In this paper, the researchers used Twitter streaming API and Python program to collect and download tweets. Twitter allows to interact with its data (tweets) using Twitter API(Application Programming Interface). In order to access Twitter streaming API, one must know a server side scripting language such as php, python or ruby and results would be in JSON format to get 4 pieces of information from twitter such as API key, API secret, Access token and Access token secret.

A Python library called "tweepy" is used to connect to Twitter streaming API and download the data. Access token, Access token secret, API key or customer key and API secret or consumer secret contains the user credentials retrieved from twitter to access Twitter API.

The researchers aimed to know the sentiments of the people regarding the banning of the citizens of 6 countries such as Syria, Iran, Libya, Somalia, Sudan and Yemen from entering the US. The researchers are interested in targeting tweeting relating to this topic. The tweets related to keywords such as "USA", "muslimban", "bannedcountries", âĂIJtrumpâĂİ are target tweets by the researchers. Last February 16, 17 and 27, the researchers collected a total of 5354 tweets.

This is a sample of one of the tweets collected, which is in JavaScript Object Notation (JSON) format containing additional information not only the main tweet: "Trump says new version of travel ban coming next week".

To process the sentiment analysis in the study, the researchers have two files of the same tweets such as Labeled data(SC Trump Labeled) and Unlabeled data(SC Trump Unlabeled) and the files must be in csv format.

The operators needed in the diagram:

Two Read CSV(Comma-Separated-Values) are needed, one for labeled data and one for the unlabeled. These operators are used for reading csv files. Values for different attributes are separated by a constant separator. It may have many rows. Each row uses a constant separator for separating attribute values.

The Process Documents from Data, will generate the word vectors from string attributes. Just like the Read CSV operator, there will be needed two operator for, one for Labeled data set and the other for Unlabeled data sets. When setting up the data in Read CSV, Labeled data sets should be in text and attribute while the label part should be polynomial and should be in labed not attribute. While the Unlabeled data, the tweets should be in text and attribute only.

The first Process Documents from Data operator must be connected to Naive Bayes and Process Documents of the unlabeled data. Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. This will help to predict the sentiments for unlabeled data sets. That's why the Process Documents 1(Labeled) must be connected to Process Documents 2(Unlabeled). This will help to determine the accuracy of both data. Naive Bayes should be connected to Multiply and Apply Model 1. Multiply is an operator that copies the objects at its input port to the output ports multiple number of times. As more ports are connected, more copies are generated. When copying DataSet only the references to attributes are copied. It is very important to note here that when attributes are changed or added in one copy of the DataSet, this change has no effect on other copies.

Apply Model is first trained on an Example Set; information related to the DataSet is learnt by the model. Then that model can be applied on another DataSet usually

Figure 4: RapidMiner Result

Figure 5: RapidMiner Result

for prediction. All needed parameters are stored within the model object. It is compulsory that both DataSet should have exactly the same number, order, type and role of attributes. If these properties of meta data of Example Sets are not consistent, it may lead to serious errors. Multiply is connected to Apply Model 1 and Apply Model 2 as well as the Process Documents 2 is also connected to Apply Model 2. The Apply Model 1 is connected to Naive Bayes, because Naive Bayes is the one that classify the Labeled data sets.

Apply Model 1 must be connected to Performance. Performance operator is used for statistical performance evaluation of classification tasks. This operator delivers a list of performance criteria values of the classification task.

After running the process and no errors are detected, then the results such as the prediction to Unlabeled data sets and the accuracy of both data and its matrices will be displayed.

4. RESULTS

The image represents the prediction of Naive Bayes based from the analysis it received from the Labeled data sets. The prediction was almost the same from the Labeled data sets thatâĂŹs why the accuracy was 91.21 percent, although this result was overfitting.

Overfitting fitting a statistical model with too many degrees of freedom in the modelling process.

4.1 Causes of Overfitting:
- Model is too complex (too many predictors).
- Training data too noisy.
- Model being refined over time with ever increasing data inputs.
- Training set too small.
- A very rich hypothesis space.

5. CONCLUSION

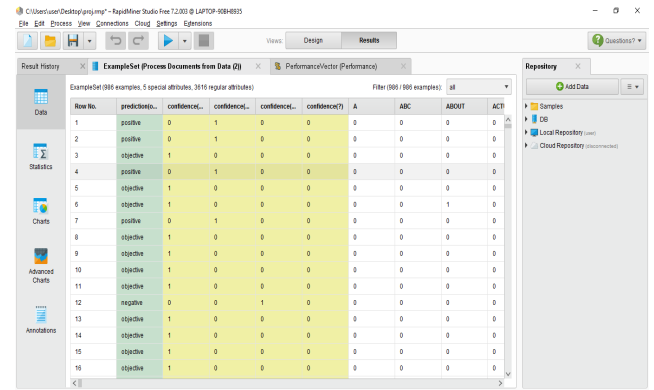We presented results for sentiment analysis on twitter. Based from the results we got, an overall result we got out of 1000 tweets is that people are not happy on President Donald Trump decisions on banning the seven countries because it affects their daily activities of their lives. Some of them might lose their job because of the banning and also some of them might lose their home, family and relatives.

6. REFERENCES

[1] Goodman, Jack. 2017. "US travel ban: Why these seven countries?" BBC News. BBC. January 30.

[2] de Vogue, Ariane , Jeremy Diamond, and Kevin Liptak. 2017. âĂIJTrump signs revised travel ban, Iraq is exempt - CNN Video.âĂİ CNN. Cable News Network. March 7.

[3] Labott, Elise , and Evan Perez. 2017. "Trump administration easing travel ban restrictions for green card holders." CNN. Cable News Network. February 2.

[4] All products require an annual contract. Prices do not include sales tax (New York residents only). 2017. âĂIJTwitter MAU in the United States 2016 | Statistic.âĂİ Statista. Accessed March 22.

[5] "Machine Learning Blog and Software Development News." 2013. Datumbox. September 2.

[6] "Twitter." 2017. Wikipedia. Wikimedia Foundation. March 22.

[7] "Presidency of Donald Trump." 2017. Wikipedia. Wikimedia Foundation. March 22.

[8] Mozetic, Igor, Miha Grcar, and Jasmina Smailovic. "Multilingual Twitter Sentiment Classification: The Role of Human Annotators." Plos ONE 11, no. 5 (May 5, 2016): 1-26. Academic Search Complete,

[9] Tarek Elghazaly et al. âĂIJPolitical Sentiment Analysis Using Twitter DataâĂİ. ACM New York, USA (March 22, 2016)

[10] Giachanou, Anastasia, and Fabio Crestani. "Like It or Not: A Survey of Twitter Sentiment Analysis Methods." ACM Computing Surveys 49, no. 2 (June 2016): 28-

28:41. Business Source Complete.

[11]Wang, Zhibo, et al. "A hybrid model of sentimental entity recognition on mobile social media." EURASIP Journal On Wireless Communications and Networking 2016, no. 1 (October 24, 2016): 1-12. Academic Search Complete.

[12]Moujahid, Adil. 2014. âĂIJAn Introduction to Text Mining using Twitter Streaming API and PythonâĂİ July 21.