Opinion Mining of Twitter Users Using Computation Techniques.

Minor Project I

Submitted by

Anjali Kumari (9917103248)

Tushit Garg (9917103259)

Abhishek Saxena (9917103246)

Prince Gaur (9917103240)

Under the supervision of

Ms. Anuradha Gupta

Department of CSE/IT

Jaypee Institute of Information Technology University,Noida

OCTOBER 2019

Abstract

In our project we are surveying public opinion from big social media domain-specific textual data to minimize the difficulties associated with modeling public behavior.Machine learning techniques are applied to classify a data corpus of 10,000 tweets.Gun laws is a complex issue and accounts for a large proportion of violent incidents. In this project we set out to predict the pro-gun and anti-gun labels expressed on a social media platform , namely Twitter.

The methodology for analysing public opinion incorporates machine learning and (1) collects, (2) pre-processes, (3) feature extraction (4) topic modelling (5) summarises (6) visualises data. Also, a Temporal Analysis of tweets shows the trend of public opinion over the period of data analysis.

We also compared the accuracy of different machine learning models on our dataset and found out that SVM is producing the highest accuracy on our dataset.

Table of Contents

List of Figures

Abbreviations

TFIDF	*Term Frequency-Inverse Document Frequency*
NER	*Named Entity Recognization*
LDA	*Latent Dirichlet allocation*
POS	*Part-of-speech tagging*

INTRODUCTION

Gun violence is a major issue and accounts for a large proportion of violent incidents in the world. As we know that United States is certainly an exceptional country when it comes to firearms.It's one of the few countries in which the right to bear arms is constitutionally protected, so there is much controversy going on the control of guns.The root of the debate is a single sentence in the U.S. Constitution, where the Second Ammendment says: "A well regulated Militia, being necessary to the security of a free State, the right of the people to keep and bear Arms, shall not be infringed."The interpretation of this line separates those who believe in more regulation of firearms and those who see any such legislation as an infringement on their individual liberties.

Problem Statement

Surveying public opinion in a traditional way is a costly and a time consuming process which can require contacting many people. Twitter being a public opinion platform, is the best source for collecting textual data to minimize the difficulties associated with modeling public behaviour. Our motive is to predict the **stance** of people opinion on **gun control law** expressed on the social media platform Twitter in a ten day period from 19th september to 28th september. We want to capture the trend of public opinion and how it changed over time.

Why it is important?

Understanding public opinion would help in bringing more reformed ammendments in the gun laws. In order to achive this , we are using the modern computation techniques of Machine Learning , Natural Language Processing and Deep learning over a sample of 10000 tweets made by individuals in USA that contains one of predermined relevants keywords. Tweets are dowloaded using the twitter streaming API and labelled using predermined keywords. Data is prepared for analysis through data preprocessing and data cleaning steps.Topic modelling technique is used to get an overview of public opinion on gunlaws in America and Temporal analysis of public opinion over the period of 10 days. Features were extracted using NLP techniques to feed as input to machine learning classifiers for text classification into two stance: progun and antigun. Furthermore deep learning algorithms would we applied for better accuracy of stance prediction.

BackGround Study

Methodology

The methodology for analysing public opinion incorporates machine learning and (1) collects, (2) pre-processes, (3)feature extraction (4) topic modelling (5) summarises (6)visualises data.
Raw data is downloaded from Twitter using Twitter streaming API and stored in **MongoDb** database which helps in eleminating duplicate entries and helps in fast queries. The data is then pre-processed and cleaned into a trimmed data set in JSON format. Feature extraction (Sentiment score, POS tags count,NER count,TFIDF scores, hashtags,hashtags count,mentions,urls,ngrams) is done for text classification. These features serves as an input to machine learning classifiers. Topic modelling technique **LDA** is used to get an overview of the topics public is talking about. Visualization of accuracy scores of different classifiers with increasing training data is done to get the most accurate classifier.

Steps Involved:

3.1. Data collection:

Collection of data is done using the twitter streaming API which provide twitter feed in a machine readable JSON format. We first find out the popular hashtags trending on twitter that are related to anti gun and pro gun . A list of hashtags was prepared seperately for antigun hashtags and progun hashtags.

Pro Gun Hashtags	***Anti Gun Hashtags***
#ProGun	*#AntiGun*
#2ADefender	*#Gunviolence*
#Ar15	*#Guncontrol*
#BetoORouke	*#Gunskillpeople*
#2ndAmendment	*#Marchforlives*
#Libertarian	*#Gunshooting*
#Goodgun	*#Guncontrolnow*
#ProLife	*#safety*
#Freedom	*#tcot*
#Americanmade	*#Gunsense*
#Pewpewlife	*#firearms*
#Gunrights	*#nra*
#FireArmsDaily	*#neveragain*
#GunsDaily	*#NoMoreNRA*
#GunFanatic	*#Guns*
#Gunsmithing	*#Stopgunviolence*
#Gunsmithproud	*#massshooting*
#SecondAmendment	*#MomsDemandAction*

Table 1.1 List of Antigun and Progun hashtags.

These hashtags were used to download the tweets from the streaming api.Data was collected for a period of 10 days , from 19 sep to 28 Sep.About 13000 tweets were collected containing one or more gun related hashtags. The data was uploaded to MongoDb database which is suitable to handle the json formated tweet data.The use of MongoDb not only helps in storing the data is a structured format but also eliminates the duplicate entries and also simplifies different queries.
Now,we extract the id,date and full_text field for each tweet and stored it in an excel file.

3.2 Data PreProcessing

The data collected is messy and full of unnecessary objects which is irrelevant to machine learning classifiers.So it becomes necessary to first clean the data and make it appropriate to feed as input for classifiers. Accuracy of feature extraction also greatly depends on the quality of text data.

- ***Removal of Punctuation marks and symbols***
  Regular expression is used to eliminate the the unnecessary punctuation marks(,;!'".?/*-....etc) and symbols like emojis and emoticons were removed.Urls , extra line feeds were also removed.
- **Tokenization and Removal of Stop Words**
  - *Tokenize*:
    This breaks up the strings into a list of words or pieces based on a specified pattern using Regular Expressions.
  - *Stop Words*:
    Stop words are generally the most common words(such as "the", "a", "an", "in") in a language.These words are of no use because they don't help us to find the context or the true meaning of a sentence.We would not want these words taking up space in our database, or taking up valuable processing time.These words were removed from the previoulsy cleaned tweet text using a famous NLP library **Spacy.**
- **Stemming and Lemmatization**
  Stemming and Lemmatization are Text Normalization techniques in the field of NLP that are used to prepare text, words, and documents for further processing.Lemmatisation in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.
  Lemmatatization of the words is done using the spacy library.

3.3 Feature Extraction:

3.3.1 Sentiment Scores

"The sentiment of a piece of text is its ***positivity*** or ***negativity***."In order to calculate the sentiment of a piece of text, we split it into individual words.We have a database of words, each with a "score" to determine how positive or negative it is.The higher the score, the more positive the word and similarly opposite for negative words.Not every word in a piece of positive text will be positive, and not every word will be negative, but by feeding the number of identified words and their scores into our algorithm, we end up with a score for the sentiment of the text.
Thus we classsify the text as positive, negative and neutral based on these scores. For this we use text.blob library which gives **polarity score** of each text.The polarity score is a float value range:**[-1,1]**.

- Negative sentiment - score<0
- Positive sentiment - score>0
- Neutral sentiment - otherwise

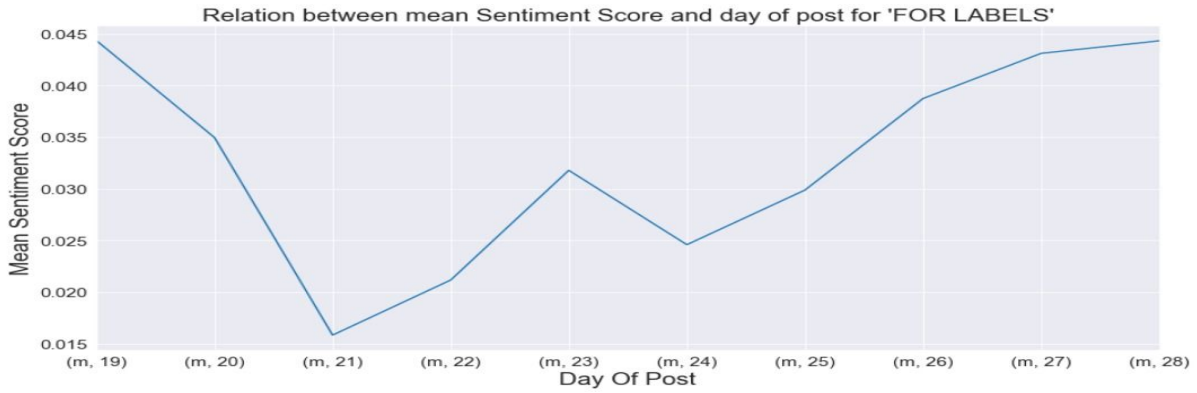

Fig 3.1 Relation between mean Sentiment Score and day of post for 'FOR LABELS'

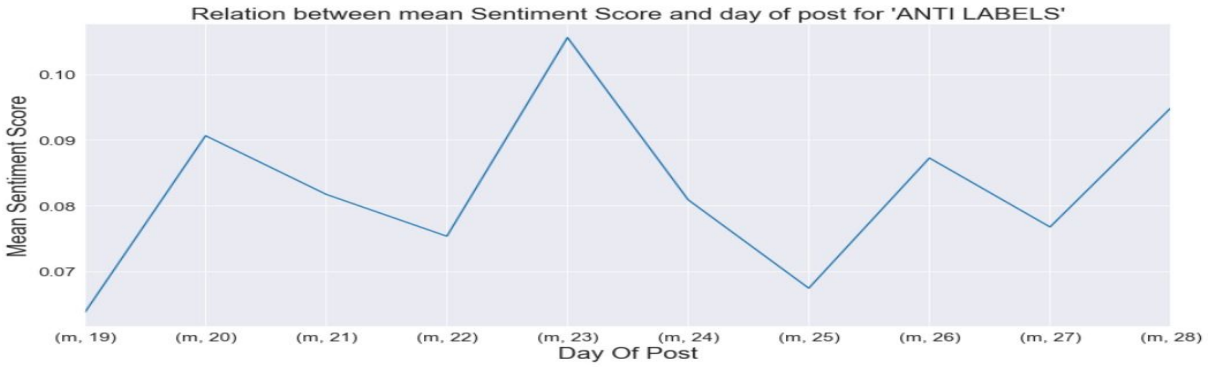

Fig 3.2 Relation between mean Sentiment Score and day of post for 'AGAINST LABELS'

3.3.2 POS Tagging

A POS tag (or part-of-speech tag) is a special label assigned to each token (word) in a text corpus to indicate the part of speech and often also other grammatical categories such as tense, number (plural/singular), case etc. POS tags are used in corpus searches and in text analysis tools and algorithms.

We have used POS tags from Spacy Library to classify the label of our tweet whether it is 'for' or 'against' the topic.

3.3.3 TFIDF

TFIDF, short for **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

- TF(w) = (Number of times term w appears in a document) / (Total number of terms in the document)
- IDF(w) = log_e(Total number of documents / Number of documents with term w in it)

Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.The higher the TF*IDF score (weight), the rarer the term and vice versa.

3.3.4 N-Grams:

N-gram is a contiguous sequence of *n* items from a given sample of text or speech. We have passed n-grams as parameter to tfidf vectorizer.

3.3.5 Extracting Hashtags and Mentions and their count

All the hashtags and mentions are extracted from the text using regular expression and then their count is evaluated.

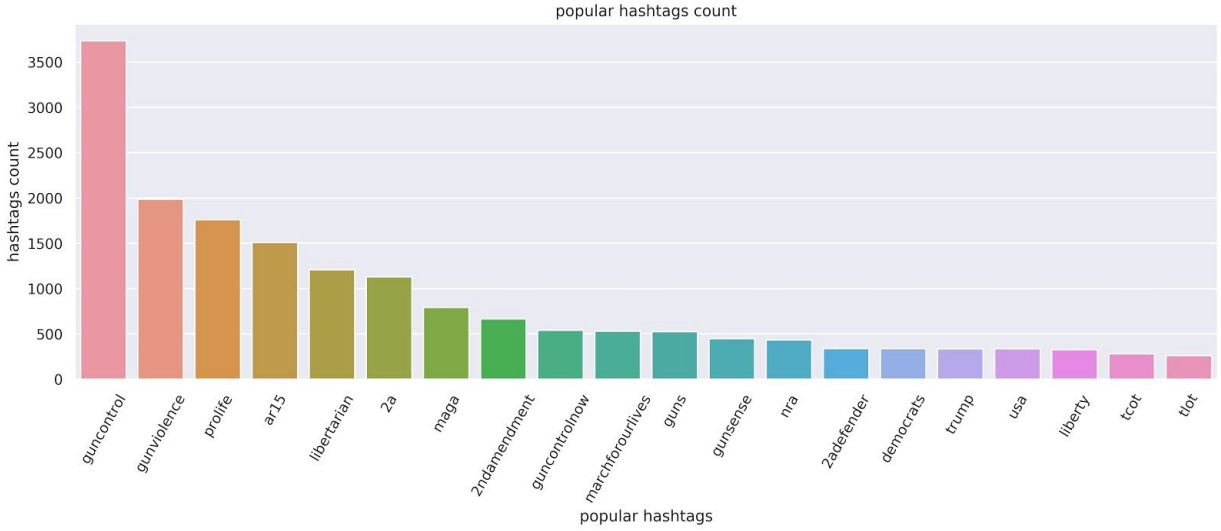

Fig 3.3 Popular Hashtags

3.3.6 Topic Modelling

Topic Modelling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

```
Score: 0.8350743055343628
Topic: 0.011*"guncontrol" + 0.009*"peopl" + 0.009*"shoot" + 0.009*"ndamend" + 0.007*"adefend" + 0.007*"betoorourk" + 0.007*"bri
ng" + 0.007*"gun" + 0.007*"prolif" + 0.006*"kill"

Score: 0.12491016089916229
Topic: 0.014*"libertarian" + 0.010*"gunviol" + 0.010*"conserv" + 0.010*"democrat" + 0.010*"guncontrol" + 0.009*"meme" + 0.009
*"trump" + 0.008*"maga" + 0.007*"great" + 0.007*"protect"
```

3.3.7 NER

Named Entity Recognition is probably the first step towards information extraction that seeks to locate and classifies named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. ,
A count of these entities were extracted from each tweet is extracted using the NLP library **Spacy.**
Example:

Implementation

Temporal study of tweets

Tweets were collected over a time period of 10 days, starting 19 September,2019 to 28 Septempber,2019.An analysis was done to see the trend of public opinion over this period of time. Public opinion is categorised into two categories namely: for and against. People which termed as Anti gun are speaking in favour of gun laws ammendment and the ones which are against any ammendment in gun laws are termed as Pro gun. So, a time series analysis is done to see the trend in opinions in favour and against the gunlaws. A bar chart plotting the tweet count of the two categories is made with the help of pandas group by funtion and the charting and plotting library : Matplotlib and Seaborn.

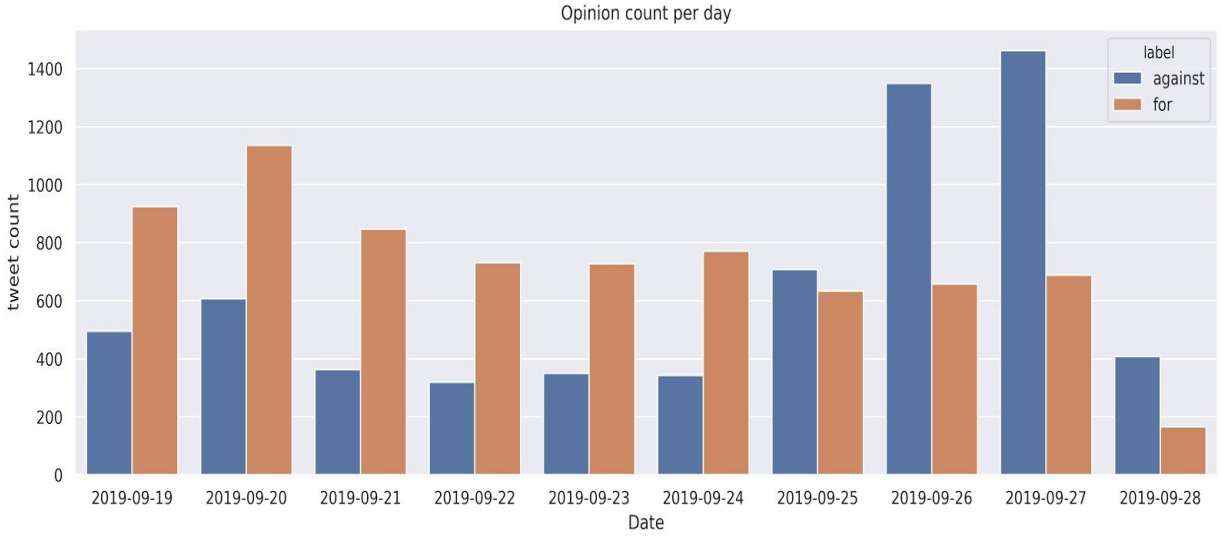

Fig 4.1 Tweet count variation with date

Applying Classifier

After gathering all the features , Sentiment scores, POS, NER,TFIDF score,hashtags count ,we then split the data into training set and testing set.Machine learning Classifiers used: SVM,Random Forest,KNN,Logistic .

Different accuracies were obtained after incorporating different features.

- With PoS and sentimental score using logistic regression 60.1 percent.
- With POS and sentimental score using random forest classifier the accuracy is 65.4 %.
- With TF idf using logistic regression the accuracy is 95.9 percent.
- With TF idf using random forest classifier the accuracy is 95.1 %.

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

```
Results of Random Forest
accuracy 0.9527405602923265
              precision    recall  f1-score   support

         for       0.96      0.93      0.95      1935
     against       0.94      0.97      0.96      2170

    accuracy                           0.95      4105
   macro avg       0.95      0.95      0.95      4105
weighted avg       0.95      0.95      0.95      4105
```

Fig1.7 *Accuracy Results and different parameters when Random Forest Classifier is used*

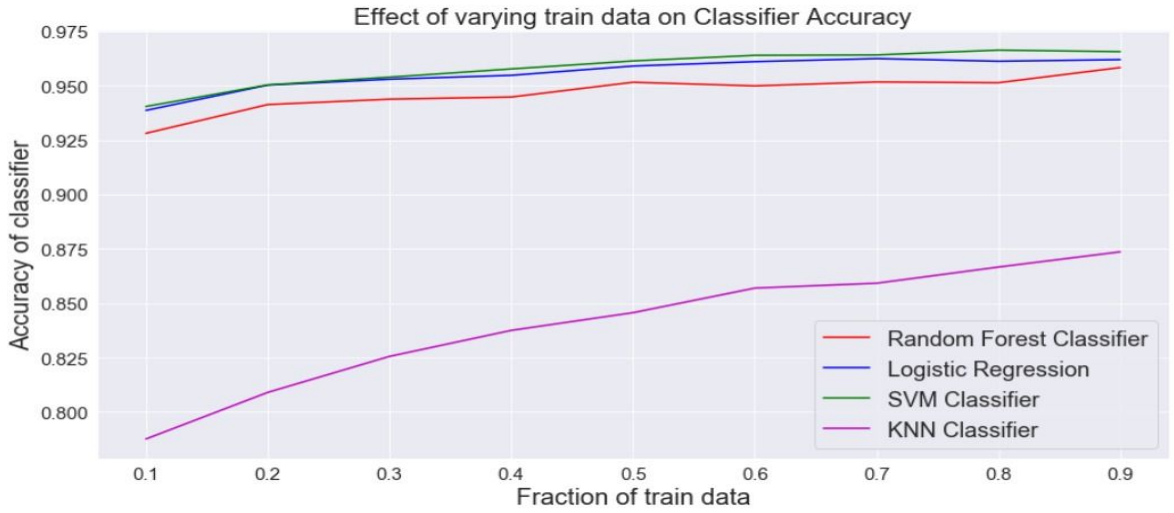

Fig 1.8 *Accuracy score of different Classifiers with increasing training data*

K-Cross Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models. Cross Validation is used to assess the predictive performance of the models and and to judge how they perform outside the sample to a new data set also known as test data. K-Fold CV is where a given data set is split into a ***K*** number of sections/folds where each fold is used as a testing set at some point.

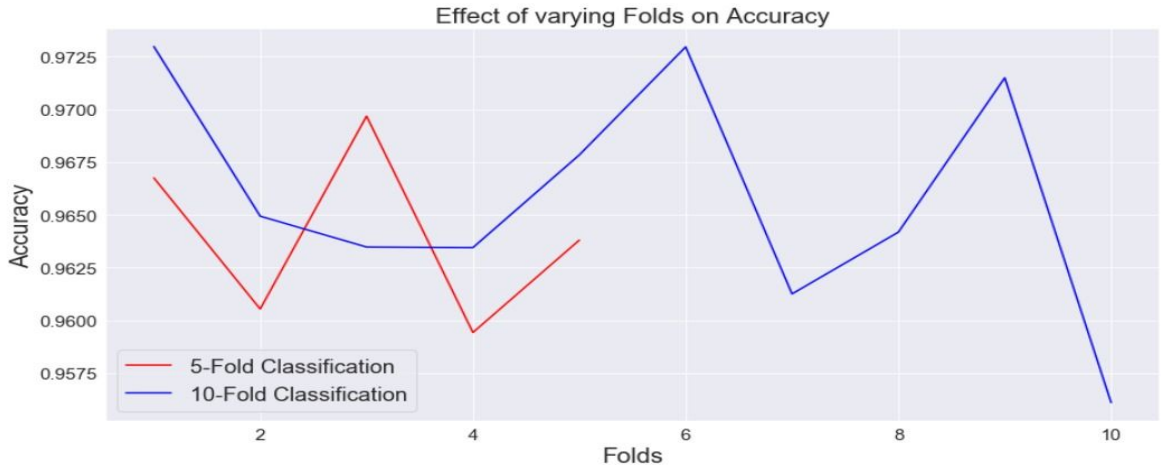

Fig 1.9 *KCross Validation Accuracy Variation*

Conclusion

In this project ,we have evaluated a number of machine learning approaches and identified those most suitable to classifying public opinion. We have shown that it is possible to analyse a large body of social media data using machine learning in a reliable and replicable way by employing a methodology to collect, train and classify tweets.Temporal analysis shows that there was a remarkable shift from antigun to progun tweets over the period of data analysis. We also compared the accuracy of different machine learning models and found out that **Suppport Vector Machine** works best on our dataset.

PROPOSED WORK PLAN FOR FUTURE

We will pursue more sophisticated classifiers, for example, deep architectures that jointly model stance, target of opinion, and sentiment. And also monitor that how the distribution of stance towards a target changes over time. Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including natural language processing, social network filtering so, we will apply the deep learning on our model to increase the accuracy to further level.

References