PROBLEM STATEMENT

Finding the social network user opinion on controversial topics, predicting their stance with temporal analysis.

STATE OF THE ART :

Research papers:

1) A Machine Analysis of twitter sentiments to the Sandy Hook Shootings by Nan Wang, Blesson Varghese and Peter D. Donnell.

Link: https://ieeexplore.ieee.org/abstract/document/7870913/keywords#keywords

Objective: To capture pro-gun and anti-gun sentiment and how they changed over time

Methodology: The methodology for analyzing public sentiment incorporates machine learning and (1) collects, (2) pre-processes, (3)classifies, (4) summarizes and (5) visualizes data.

Result: Sentiment analysis shows a peak of anti-gun feeling on the day of the Sandy Hook school shooting which quickly falls to pre-event levels. More surprisingly the analysis shows a peak of pro-gun sentiment on the day of the shooting that is sustained at an elevated level for a number of days

2) Analyzing polarization of social media users and news sites during political campaigns by Fabrizio Marozzo Alessandro Bessi.

Objective: Analysing polarization of social media users and news sites during political campaigns

Methodology: This paper presents a methodology aimed at discovering the behavior of social network users and how news sites are used during political campaigns characterized by the rivalry of different factions using (1) collects, (2) pre-processes, (3)classifies, (4) summarizes and (5) visualizes data.

Result: According to their study, 48% of Twitter users were polarized toward *no*, 25% toward *yes*, and 27% had a *neutral* behavior

Limitations:

1. Size of twitter corpus.

2. Trade-off between performance and computing power.

Dataset

▶ Topic – Gun Control Law in USA

▶ Real time Tweets – 13685

▶ Hashtag –

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

Methodology

▶ **<u>Data Collection</u>** :

Collection of data is done using the twitter streaming API which provide twitter feed in a machine readable JSON format.

▶ **<u>Text Pre-processing</u>** :

• Removal of Punctuation marks and special characters.

• 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.

▶ Feature Extraction:

1) Tf-idf

2) Topic Modeling

3) Sentimental scores

4) Pos Tagging

5) NER

▶ Training & Testing:

- ▶ train_test_splits: scikit-learn provides a helpful function for partitioning data, train_test_split , which splits out your data into a training set and a test set in the ratio of 75:25.

- ▶ Classifiers used are Logistic regression, random forest classifier, SVM Classifier.

▶ Analysis

- ▶ Temporal Analysis

 - ▶ Over the period of 10 days we saw remarkable shift from pro-gunlaws to anti-gunlaws.

Our objective:

- 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 behavior.

- 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

Work Distribution

► Our work is mainly divided into 3 phases

    ► Collecting and processing data

    ► Applying classification algorithm

    ► Evaluating accuracy and fetching results

► We are using Python as the backbone of our project and our steps are.

1) Data extraction --- Everyone

2) Data Pre-processing --- Tushit &Anjali

3) POS Tagging   --- Abhishek & Prince

4) Classifiers (Random forest, Linear SVM)   -- Everyone

5) Logistic Regression   --- Everyone

6) N grams (Count Vectorizer) --- Prince & Abhishek

7) K Fold cross validation   ----- Abhishek & Prince

8) Tf-idf   ---- Everyone

9) Topic Modeling -----Tushit & Abhishek
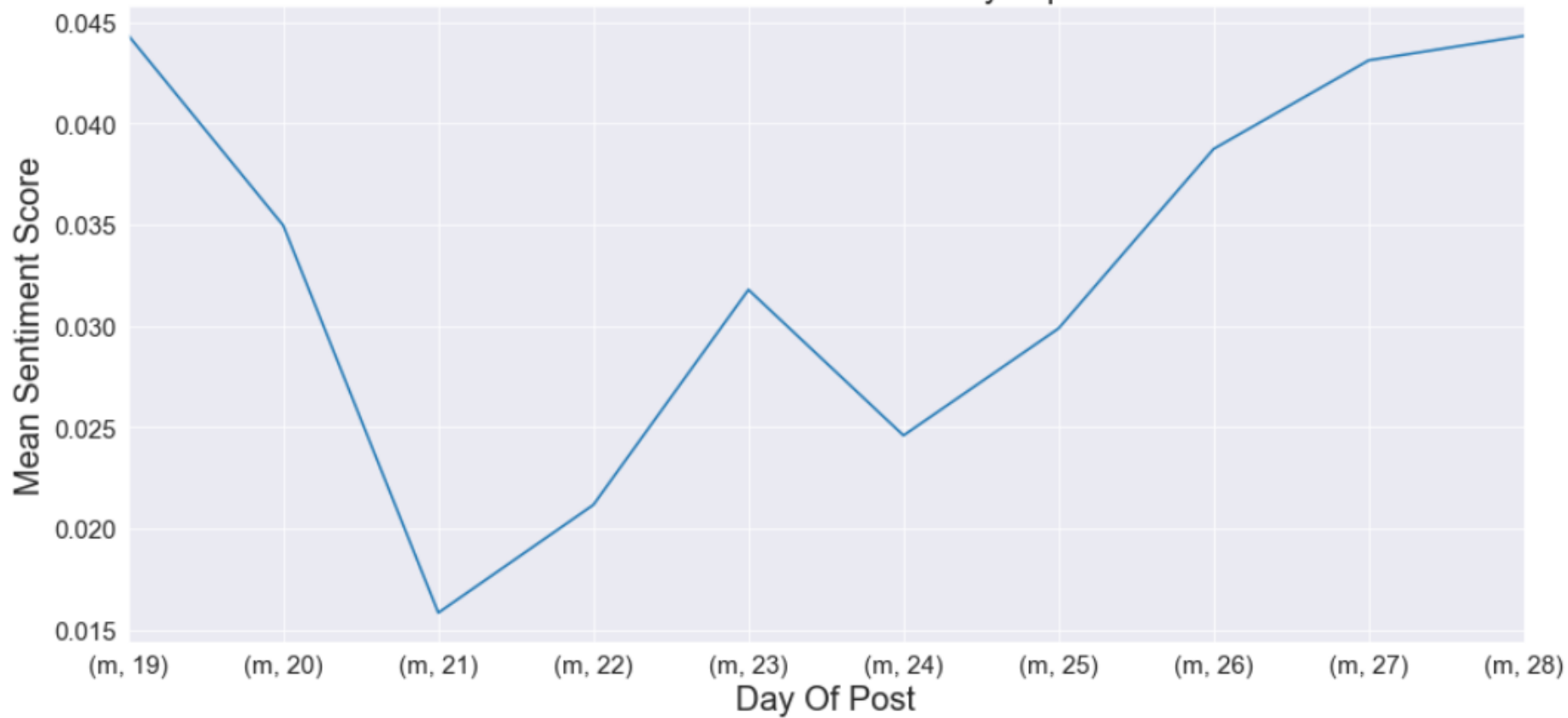
10) Temporal Analysis ----- Abhishek & Prince

Results and Illustrations:

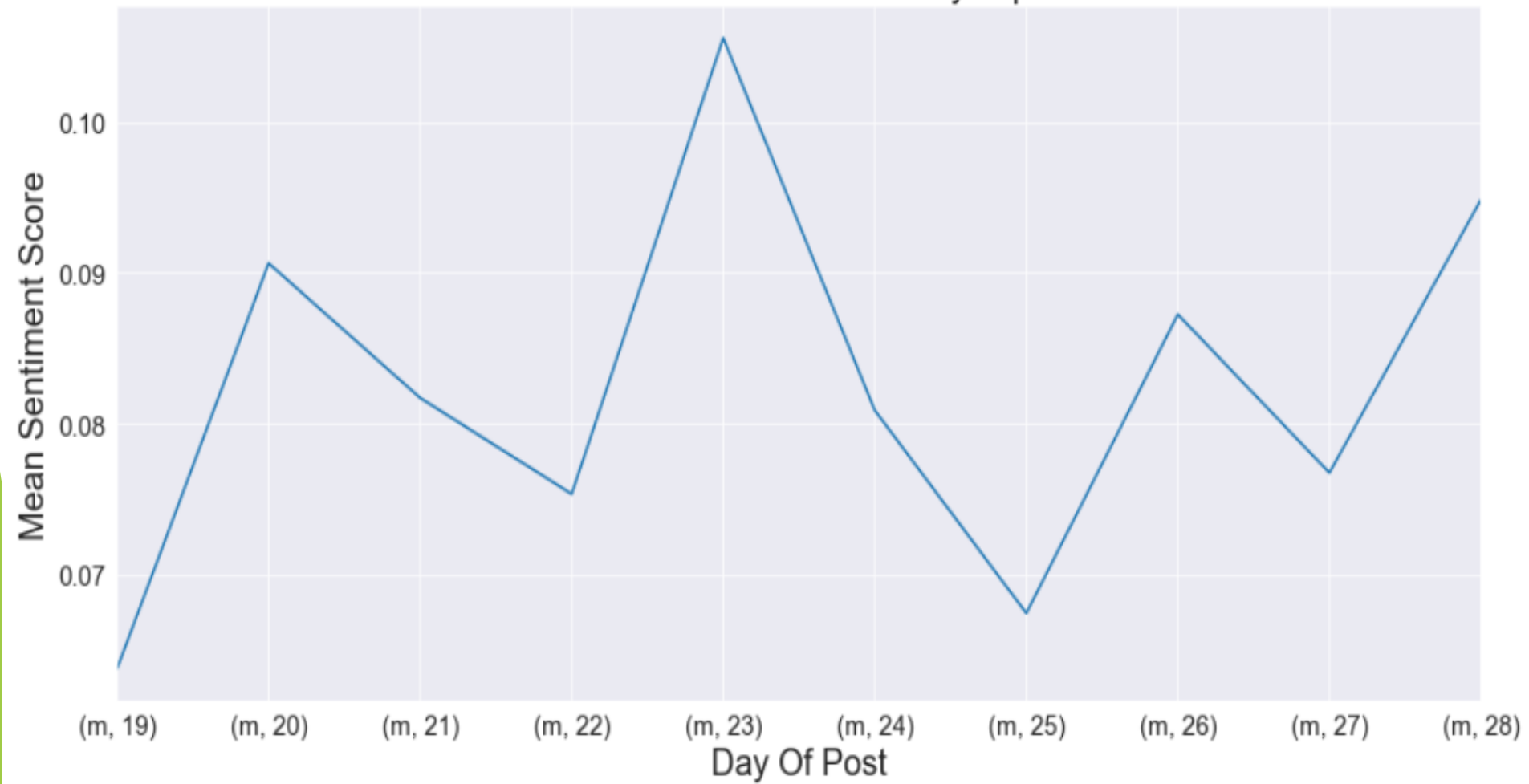Effect of varying test data on Classifier Accuracy
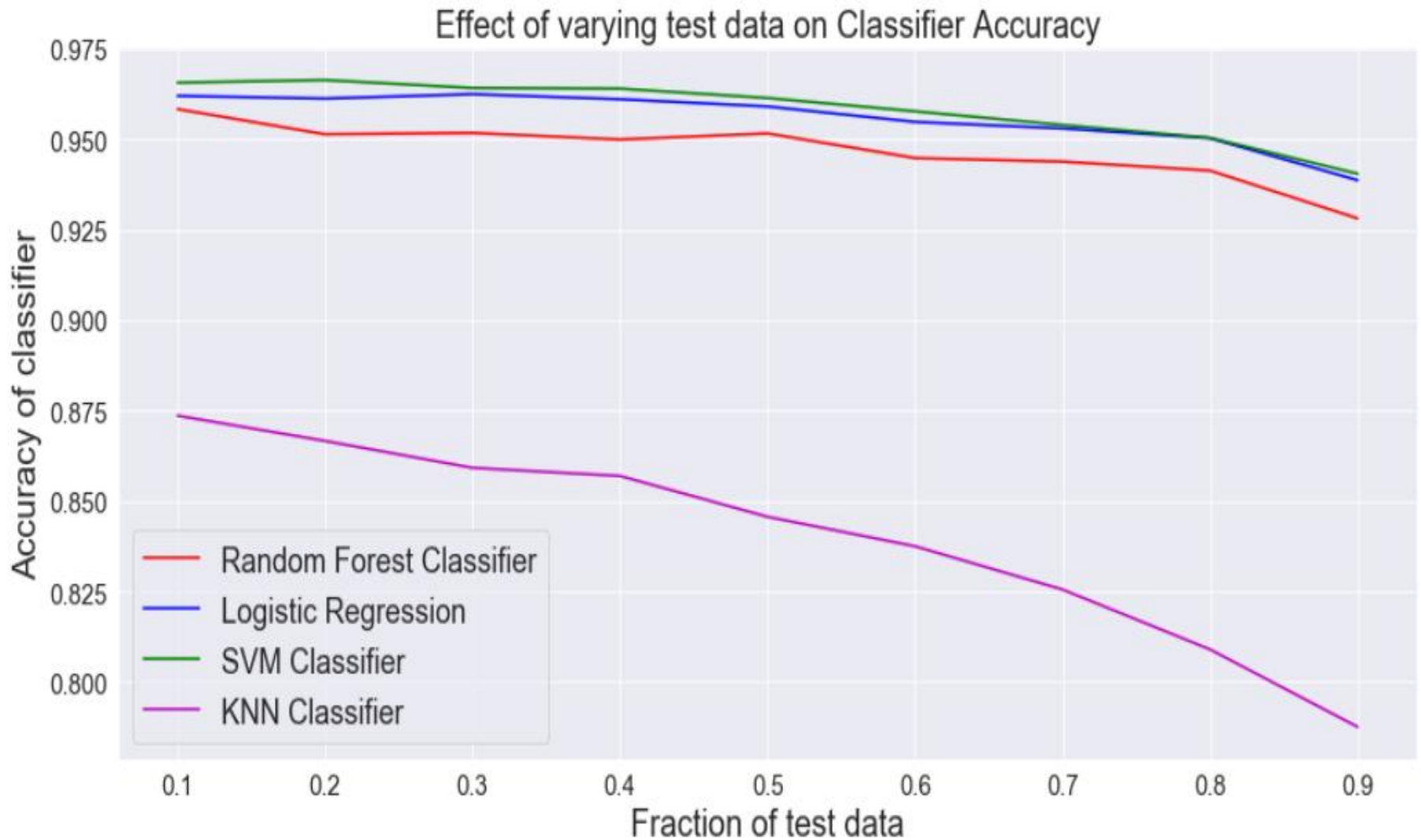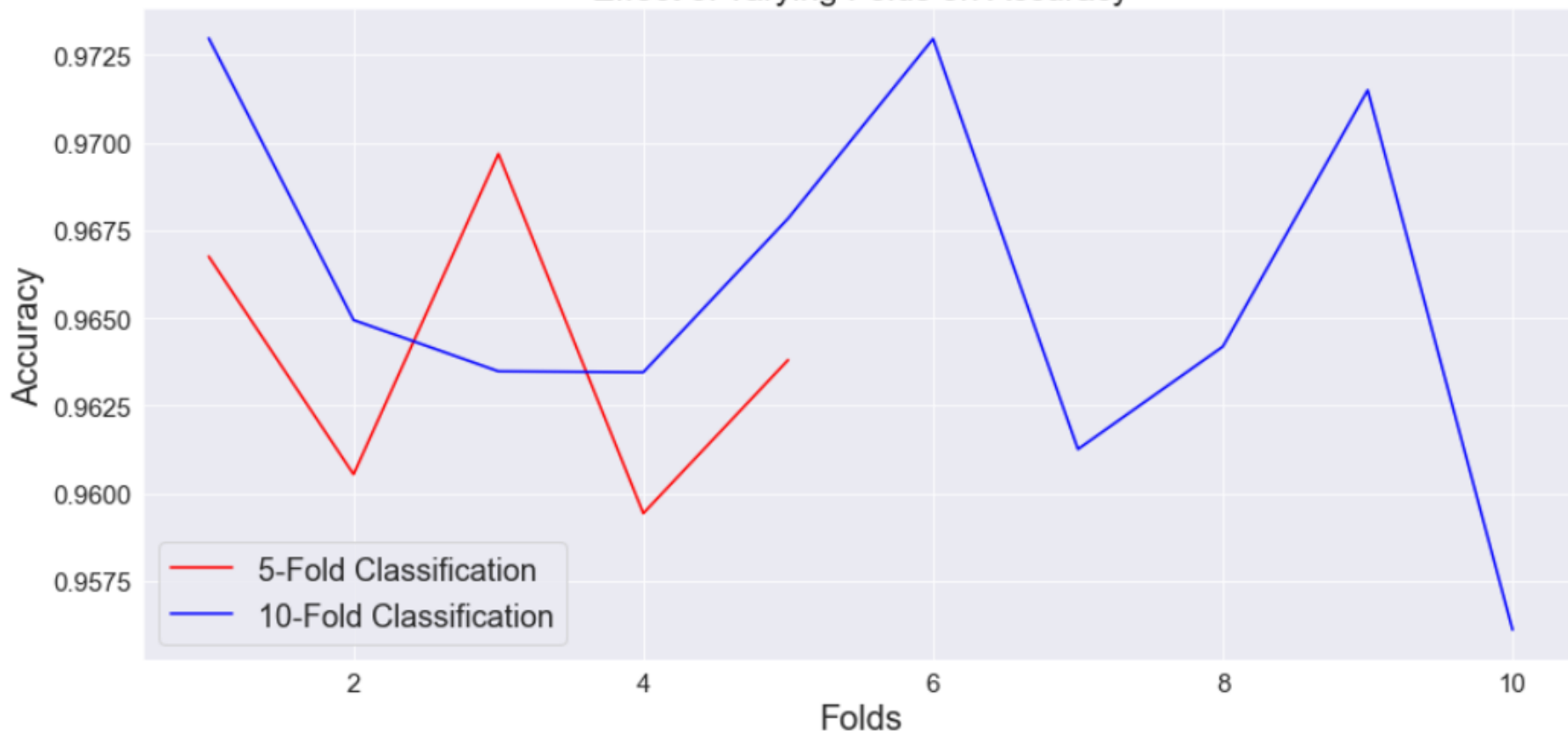
Effect of varying train data on Classifier Accuracy
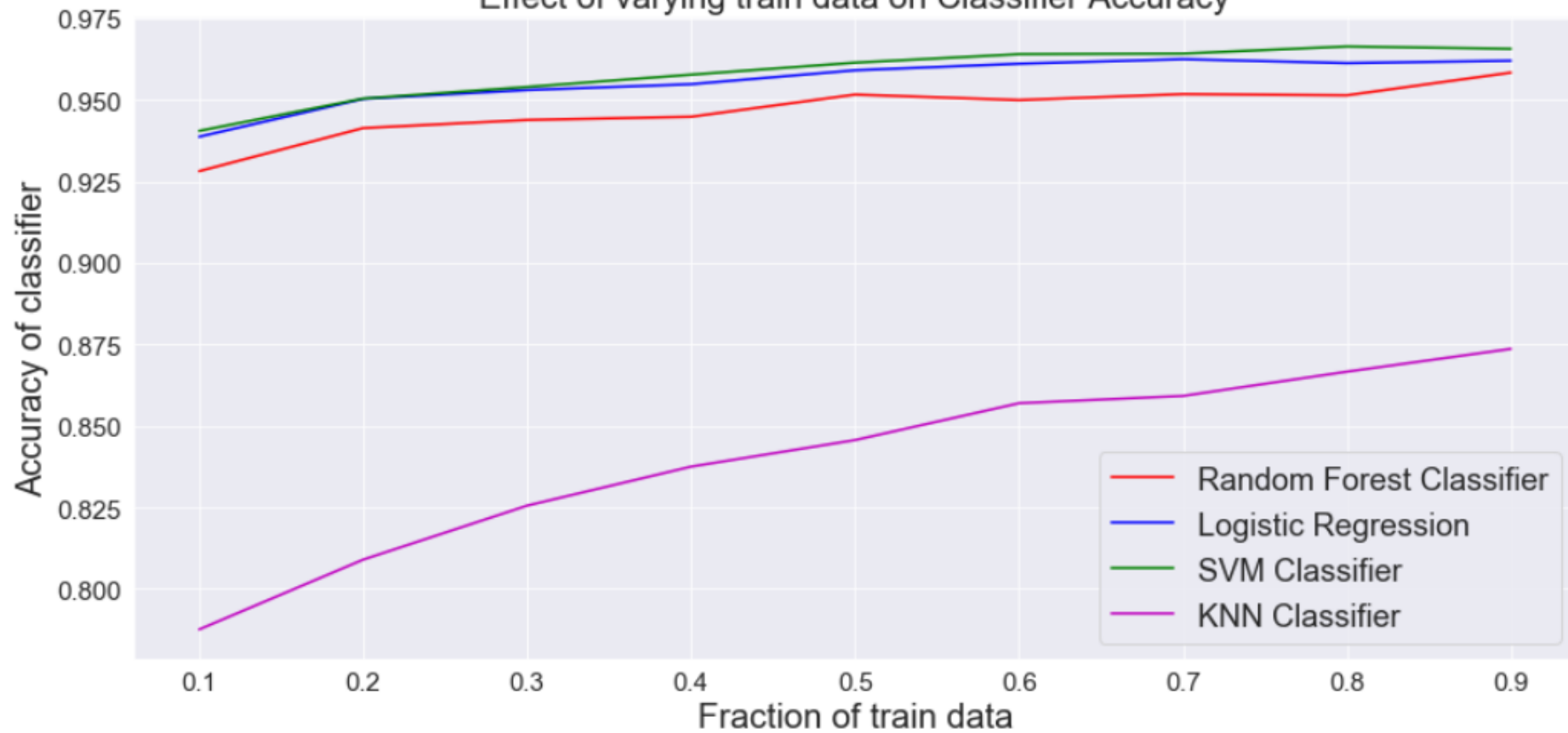
Relation between mean Sentiment Score and day of post for 'FOR LABELS'

Relation between mean Sentiment Score and day of post for 'ANTI LABELS'
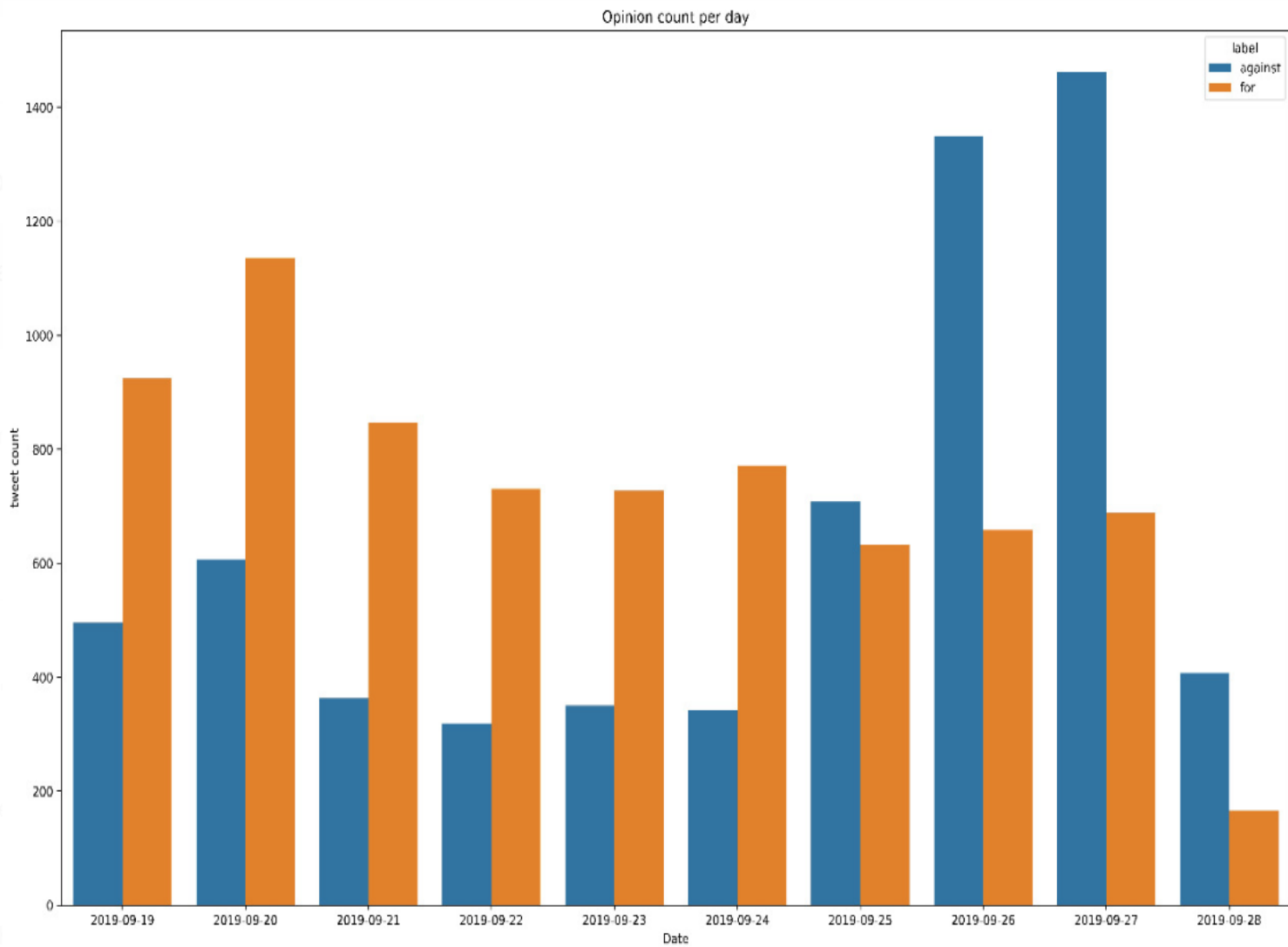
Effect of varying Folds on Accuracy

```
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
```

Opinion count per day

date	label	count
2019-09-19	against	495
2019-09-19	for	924
2019-09-20	against	607
2019-09-20	for	1136
2019-09-21	against	363
2019-09-21	for	847
2019-09-22	against	319
2019-09-22	for	731
2019-09-23	against	350
2019-09-23	for	727
2019-09-24	against	343
2019-09-24	for	771
2019-09-25	against	708
2019-09-25	for	633
2019-09-26	against	1349
2019-09-26	for	658
2019-09-27	against	1462
2019-09-27	for	688
2019-09-28	against	408

popular hashtags count
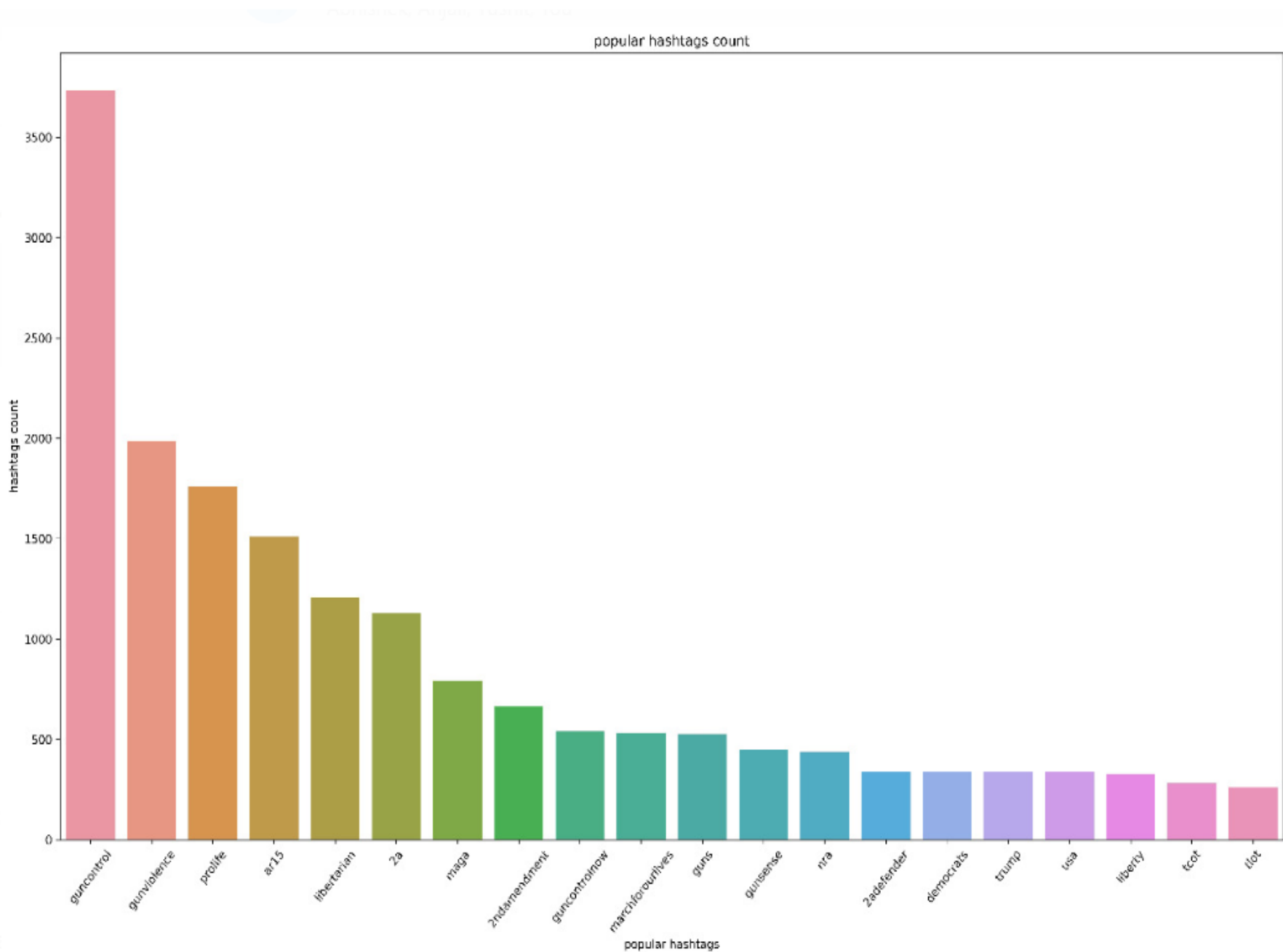
CONCLUSION

▶ Tf-idf renders increased accuracy by whooping 30.5% (from 65.4 to 95.9).

▶ LDA is very useful for manually analyzing the various topics available in document in a collection of corpus or words

▶ We used a Random forest and logistic regression that leveraged word and character n-grams. Notably, we showed that even though sentiment and using all of POS features that are useful for stance detection, they alone are not sufficient.

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 computer vision, speech recognition, natural language processing, audio recognition, social network filtering so, we will apply the deep learning on our model to increase the accuracy to further level.

THANK YOU