# Twitter Based Brexit Analysis

Surya Vajjhala  Zhi Dou

BOSTON UNIVERSITY

## Introduction

**Problem :**

- Analysis general attitude and trend towards Brexit
- Analysis main topic of both support and against Brexit classes

**Approach :**

- Scrape data from twitter
- Label data by sentiment analysis
- Classify unlabeled data and analysis trend
- Analysis main topic using clustering

## Dataset

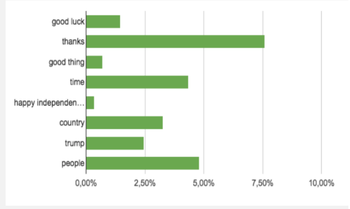
**Tweets :**

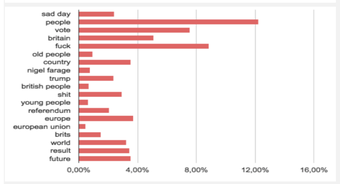
- Each tweet contains attributes: text, location, created time, user ID, tweets ID and hashtag
- **120050** labeled tweets (**23.7M**) from March to September, 2016 with hashtag like: yes2eu, yestoeu, betteroffin, votein, for remain and no2eu, notoeu, betteroffout, voteout, for leave
- **71275** unlabeled tweets (**12.7M)** from March to September, 2016 with hashtag Brexit

## Related Work

**Sentiment Analysis based :**

- Performed sentiment analysis on these tweets to get polarity of each tweets, and get positive, negative classes
- Extract key words of tweets of different sentiments to get different different point of view


**Potential Problem :**

- Sentiment of tweets cannot fully reflect users' attitude to Brexit
- Key words extraction basted on Sentiment might be unrelated to Brexit

## Model

**Work Flow**

- Scrape data and label
- Train classifier and predict unlabeled data
- Clustering inside each class to find main topic

## Model

**Scrape and Label**

- Implement scraper in python
- Scrape data with remain and leave hashtag separately and apply sentiment analysis on this data, then decide the label.

**Classification**

$$w = argmin||wx - y|| \quad s.t.\ y = wx$$

- Train linear regression as classifier
- Fine tune threshold for accuracy

**Clustering**

- Tokenize texts and reduce dimension by SVD
- Implement Kmeans and GMM clustering

## Result

**Classification**

- Compare different classifier




Figure 1 Comparison of Classifiers

Three classifier:

- Random Forest
- Logistic Regression
- Linear Regression

- Apply Linear Regression on unlabeled data set
  Show trend of remain and leave

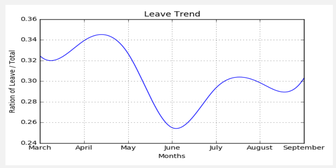



Figure 2 Ratio of Remain to Leave

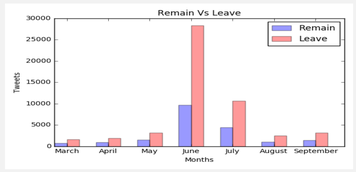
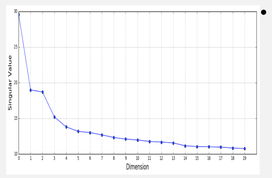



Figure 3 Trend of Remain to Leave

**SVD Decomposition**

- Compare different classifier




- Top5 Corresponding Topic

| Remain | Leave |
|---|---|
| Economy/Trade | Freedom/Sovereignty/Great |
| Jobs/Market/Risk | Tax |
| Law/Govern | Economy/Trade |
| Security | Migrants |
| Nation/Leader | Lie/Brussels |

## Result

- **Kmeans Clustering**

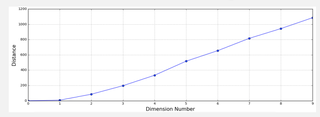



Figure 4 Distance, Dimension Number

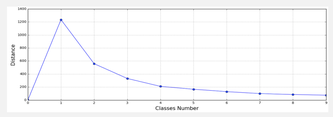



Figure 5 Distance, Classes Number

## Result

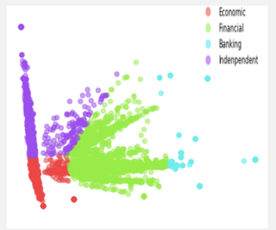



Figure 6 Remain, Kmeans Cluster

Figure 7 Word Cloud

- Label for remain: Economy, Trade; Financial; Banking; Independent
- Word Cloud: Trade, Economic; Dangerous, Risk, Critics; Football


Figure 8 Leave, Kmeans Cluster

Figure 9 Word Cloud

- Label for leave: Refugee, Expense, Trade, Employment, Political Change
- Word Cloud: Migration, Laws, Tax, Britain, Control, Lies
- **GMM**

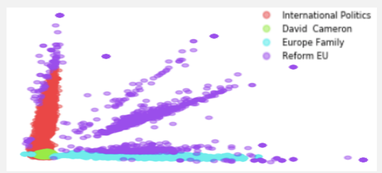



Figure 11 Remain, GMM Cluster

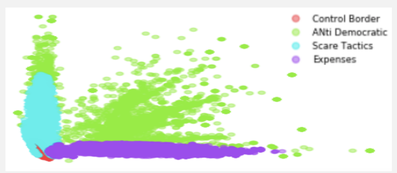
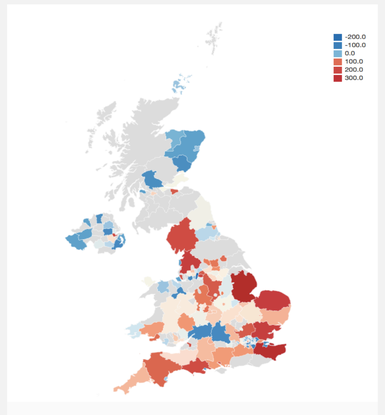
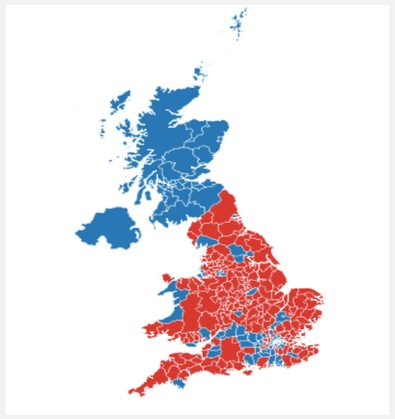



Figure 12 Leave, GMM Cluster

- Label for leave: Control Border, Anti Democratic, Scare Tactics, Expense
- Label for remain: International Politics, David Cameron, Europe Family, Reform EU

- **Distribution**



Both graph show people's attitude towards Brexit. Red means support Brexit, and blue means against. Left figure is what we get from our model, while right is from news

**Citation** :

[1] http://www.investopedia.com/terms/b/brexit.asp

[2] http://www.networkworld.com/article/3088617/analytics/amazing-analysis-of-the-brexit-with-machine-learning.html