Social Media Analytics

Team:
Abraham Duplaa
Akshaya Muralidharan
Jiho Yang
Marie Delacourt

Project lead: Dr. Ricardo Acevedo Cabra
Supervisor: Prof. Dr. Massimo Fornasier
Advisor: Matthias Wissel
Dhrubajyoti Mukherjee

26.07.2018

Capgemini

About us

Abraham
MSc CSE

Akshaya
MSc Data Engineering & Analytics

Jiho
MSc CSE

Marie
Msc Mathematics in Data Science

A wee bit of background & motivation

facebook

Over **2.9 billion** monthly active users.

510,000 comments, **293,000** status updates, and **136,000** photos uploaded every **60 seconds.**

twitter

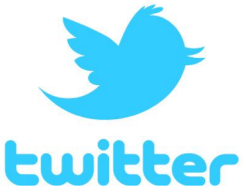
Over **300 million** monthly active users.
500 million tweets sent per day

A wee bit of background & motivation

What are we going to do with all this data?

Questions to address

- What kind of opinion do we want to know and why?
  - At first glance: Market Monitoring
- Where can we get such data?

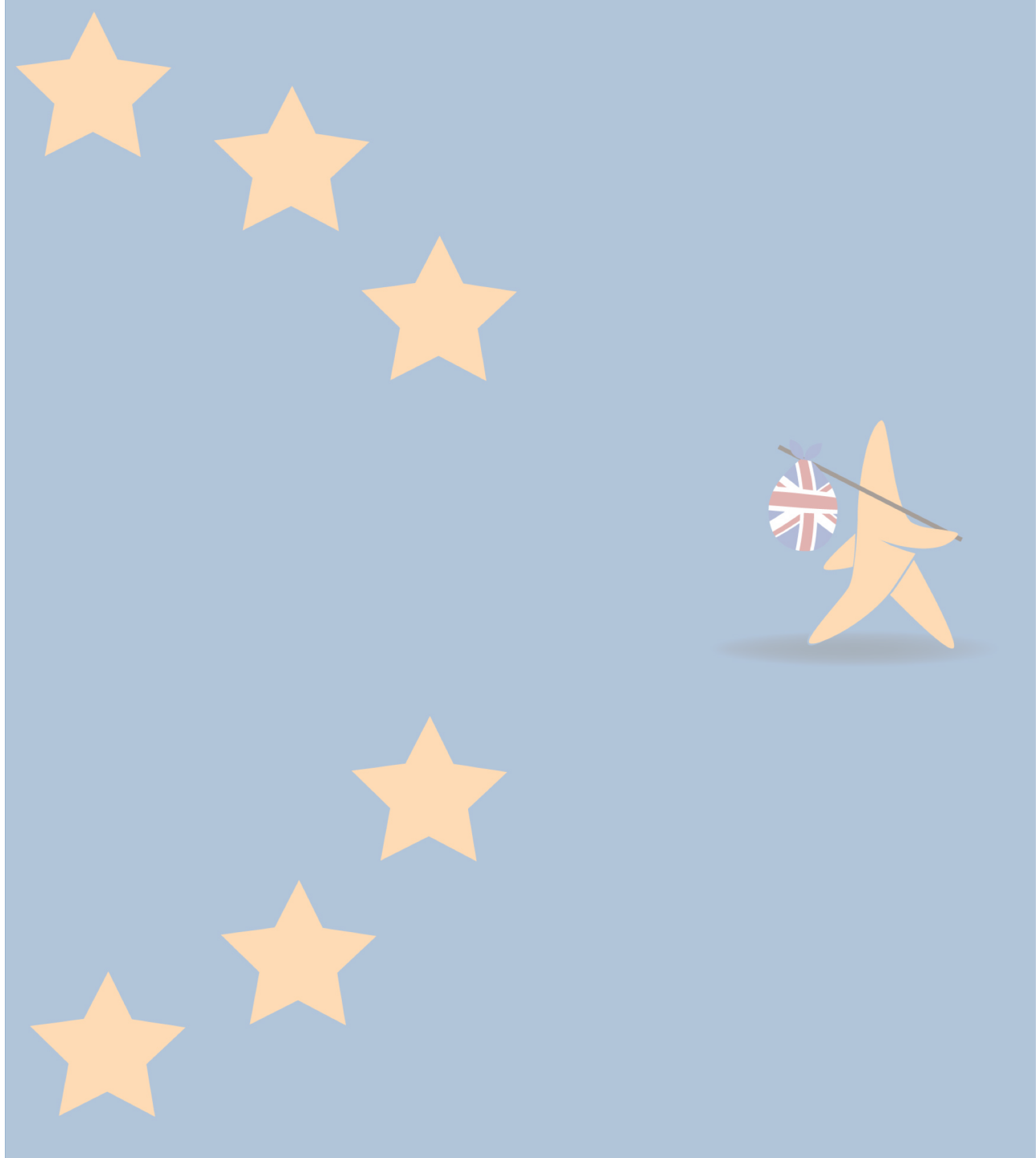


Political data science

United Kingdom European Union membership referendum

- Predict vote results
- Political agenda adjustments
  - Feedback towards policies
  - Brexit: metric for feedback
- Analyse people's opinion towards Brexit before & after the vote
- By no means reflects Capgemini's business directions

Table of contents



Understanding the data

Data from twitter

36 million tweets

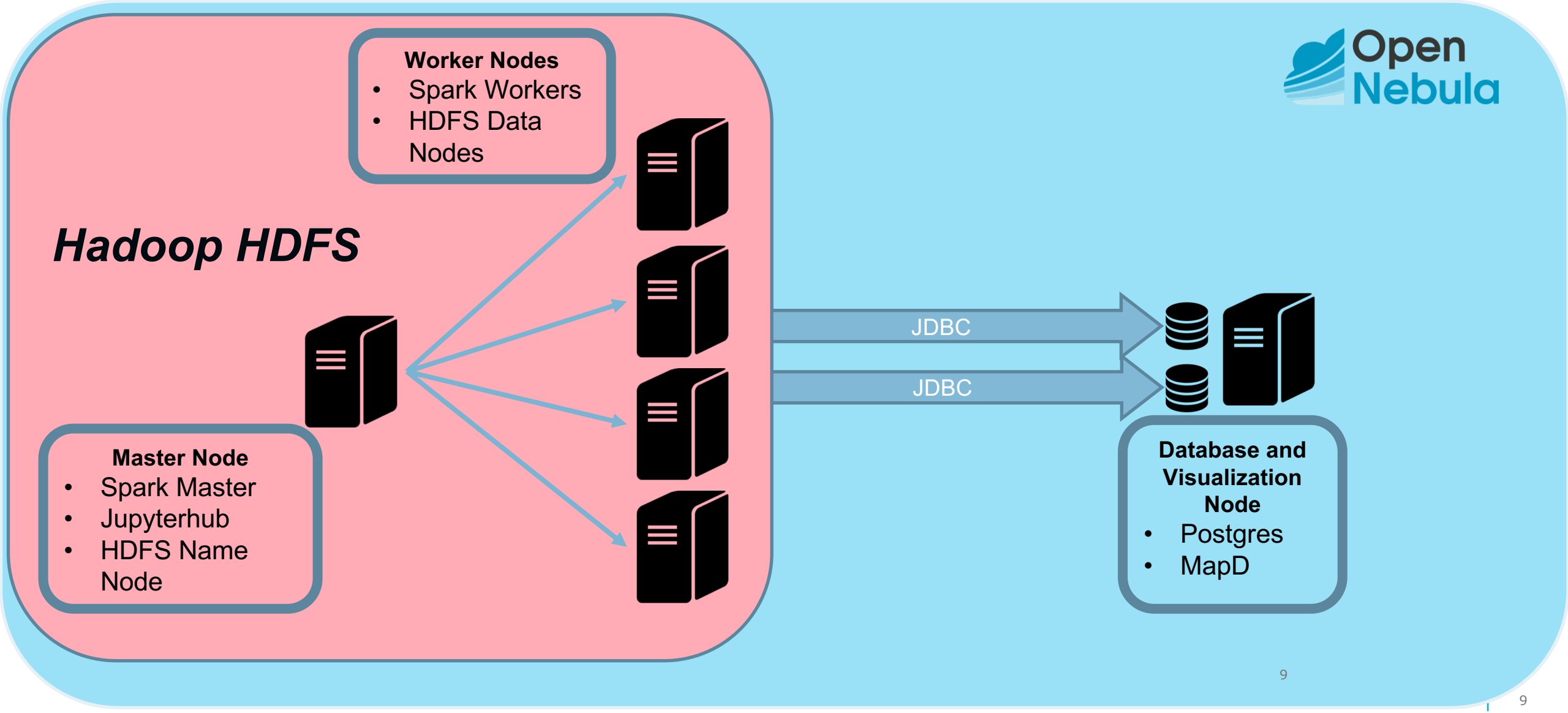
221 countries

4 million users

Tools Used

File systems and databases	Data Processing	Remote UI	Visualization
Hadoop HDFS	Apache Spark	Jupyterhub	Tableau
PostgreSQL	Python	Atom	MAPD
MAPD	PySpark		matplotlib
			plotly

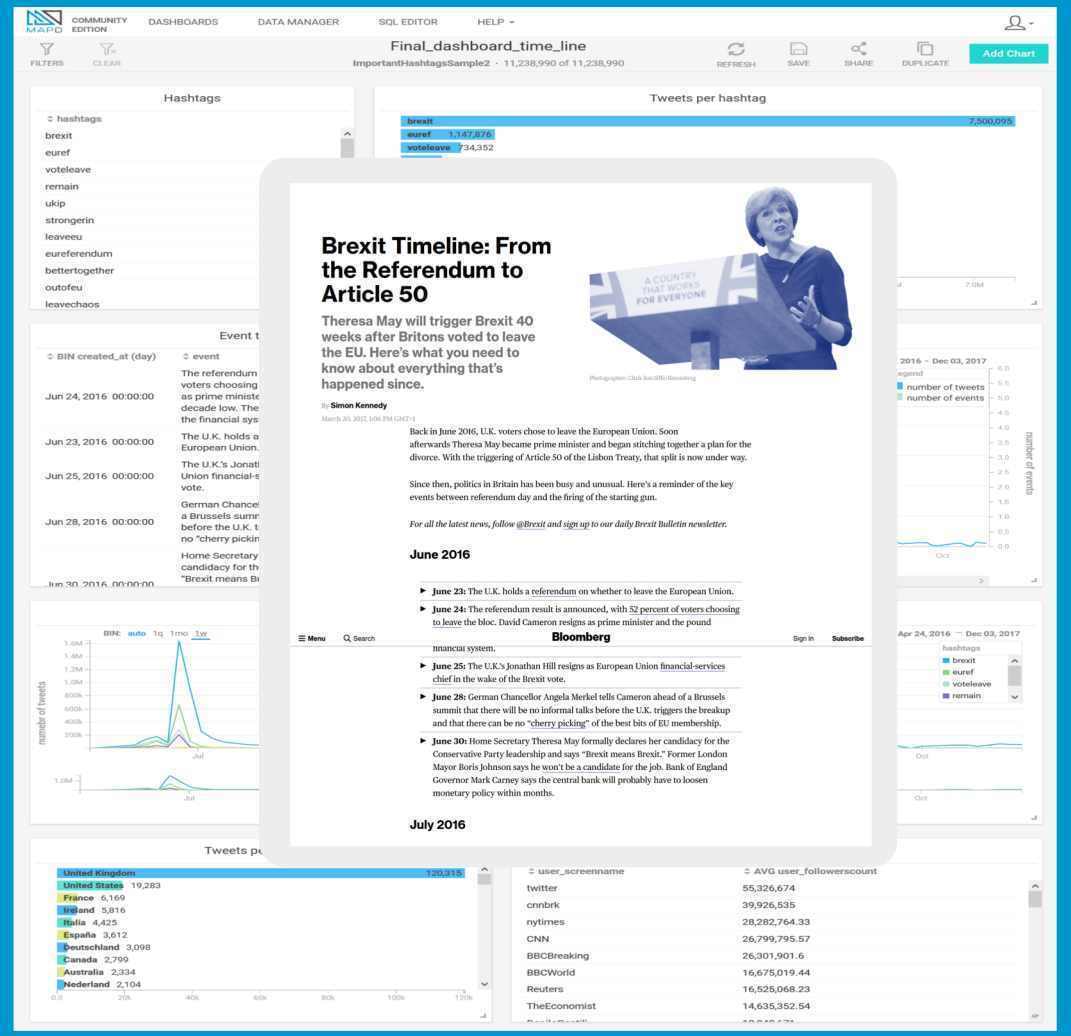
Environment Setup



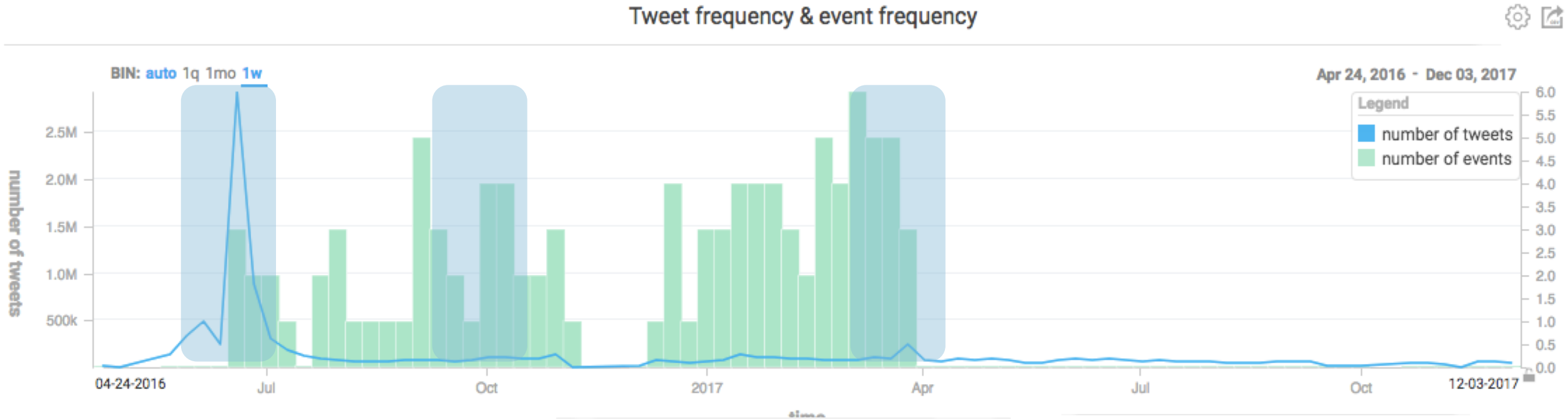
9

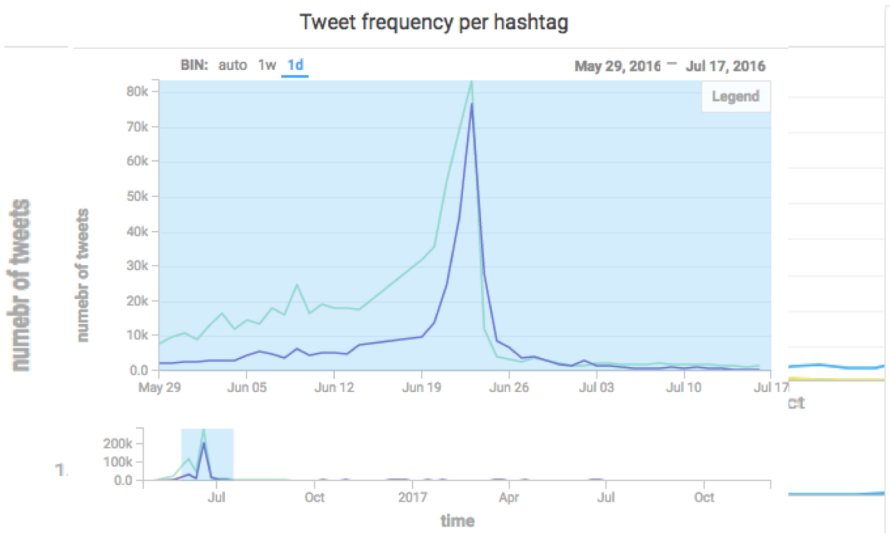
Visualise our data

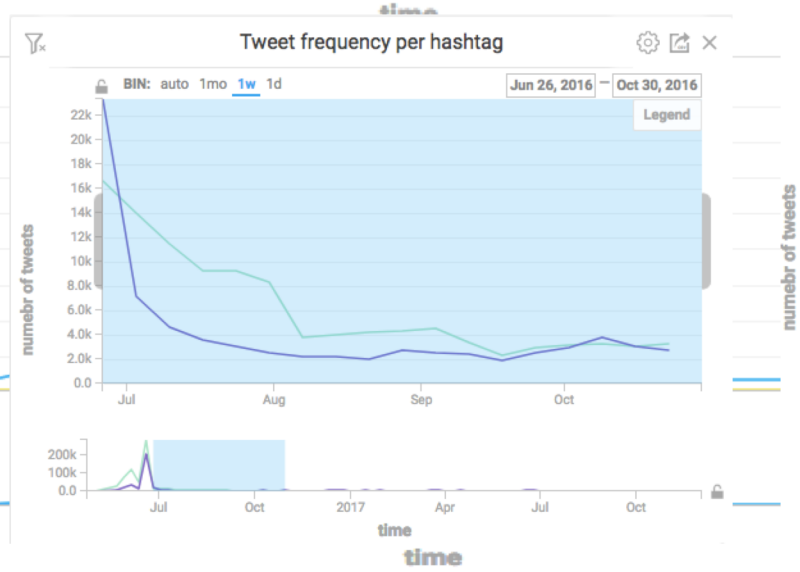
```
root
 |-- user_id: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- tweet_language: string (nullable = true)
 |-- twee
 |-- twee
 |-- twee
 |-- twee
 |-- twee
 |-- twee
 |-- twee
 |-- twee
 |-- user
 |-- user
 |-- frie
 |-- user
 |-- user
 |-- user
 |-- user
 |-- twee
 |-- twee
 |-- twee
 |-- hash
 |-- crea
 |-- created_at_truncated: timestamp (nullable = true)
 |-- is_event: string (nullable = true)
 |-- event: string (nullable = true)
```

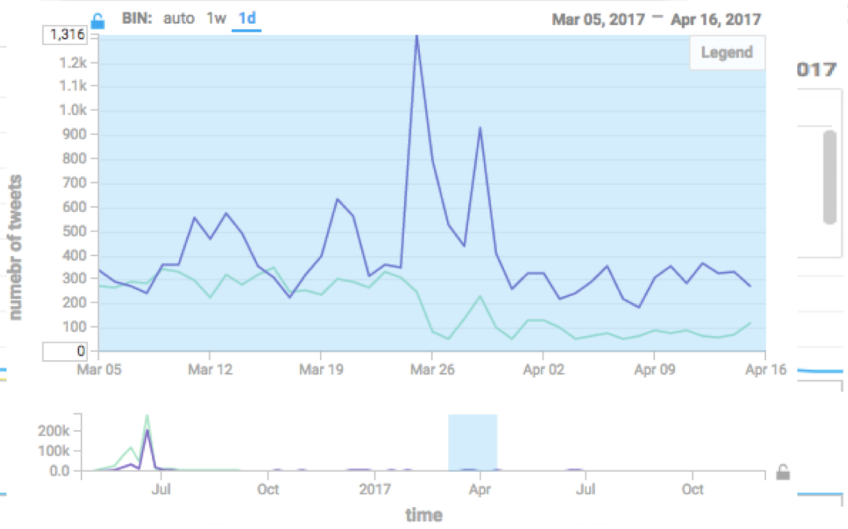


Visualise our data









Getting more insights: Sentiment analysis

Existing sentiment analysis tools

AFINN

Sentiment analysis lexicon

NLTK + TextBlob

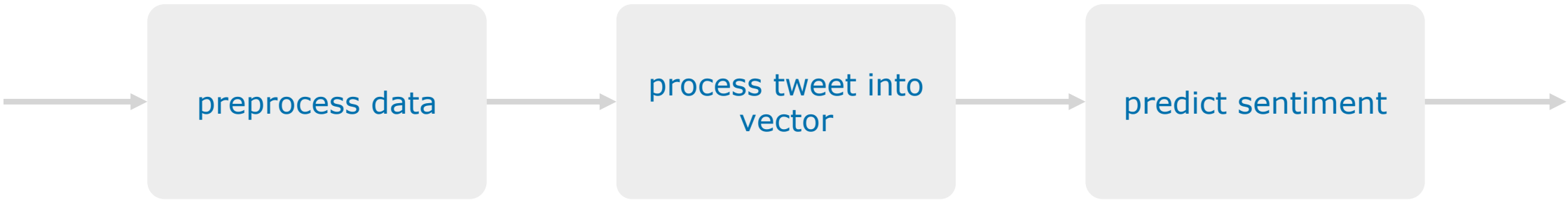
- Tokenizer
- Named entity recogniser
- NaiveBayesClassifier
- Stemmer
- …

Trained on movie reviews

Sentiment analysis API built on top of NLTK

36% accuracy for our dataset

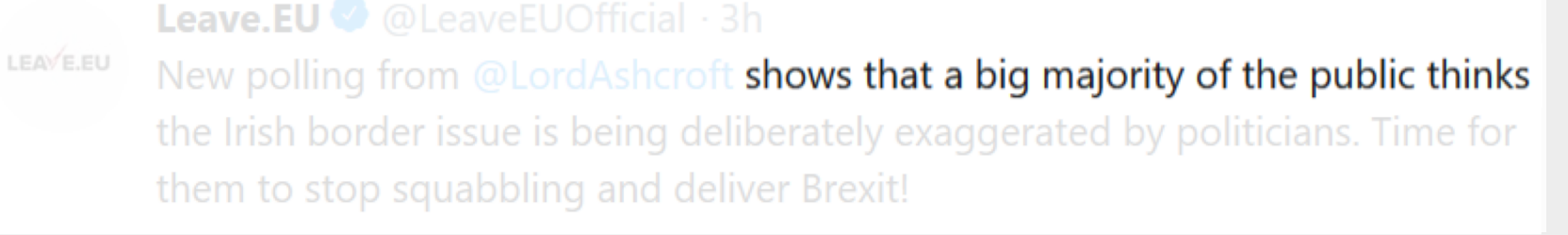
A sentiment analysis pipeline



- Tokenization
- Stop word removal
- Stemming

The principle of word embedding

Select tweet



Tokenize and
Create vocabulary

{ big, border, Brexit, deliberately, deliver, exaggerated, Irish, issue, LordAshcroft, new, politicians, polling, public, show, squabble, stop, think }

Build word vectors

$$\text{majority} = (\underset{\text{big}}{1}\ \underset{\text{borders}}{0}\ \underset{\text{Brexit}}{0}\ \underset{\ldots}{\ldots}\ \underset{\text{show}}{1}\ \underset{\text{squabble}}{0}\ \underset{\text{stop}}{0}\ \underset{\text{think}}{1})$$

Word embedding with Word2Vec

Goal

Feed context into the model

{ big, border, Brexit, deliberately, deliver, exaggerated, Irish, issue, LordAshcroft, new, politicians, polling, public, show, squabble, stop, think }

Get best embedding

$$\text{majority} = (y_1 \quad \cdots \quad y_V)$$

Start simple The bigram model

Leave.EU @LeaveEUOfficial · 3h

New polling from @LordAshcroft shows that a big majority of the public thinks the Irish border issue is being deliberately exaggerated by politicians. Time for them to stop squabbling and deliver Brexit!

Model

Classification problem with V classes.

We use a logistic regression model.

$$y|x \sim multinomial(\sigma(W^T x))$$

Metric Find W' that maximizes likelihood

$$P(output = \text{majority}|input = (1 \ \ldots 0), W)$$

The math behind the neural network

Maximize:

$$P(y|\,x,W) = \prod_{i=1}^{n} P(y_i|\,x_i,W) = \prod_{i=1}^{n}\prod_{v=1}^{V} P(y_i = C_v|\,x_i,W)^{1_{y_i = C_v}}$$

Baye's law & formula of total probabilities

$$P(y_i = C_v|\,x_i,W) = \frac{P(x_i|\,y_i = C_v,W)\;P(y_i = C_v)}{\sum_{j=1}^{V} P(x_i|\,y_i = C_j,W')\;P(y_i = C_j)}$$

logistic regression approach

$$P(y_i = C_v|\,x_i,W') = \frac{\exp(W_v^T\,x_i)}{\sum_{j=1}^{V}\exp(W_j^T\,x_i)}$$

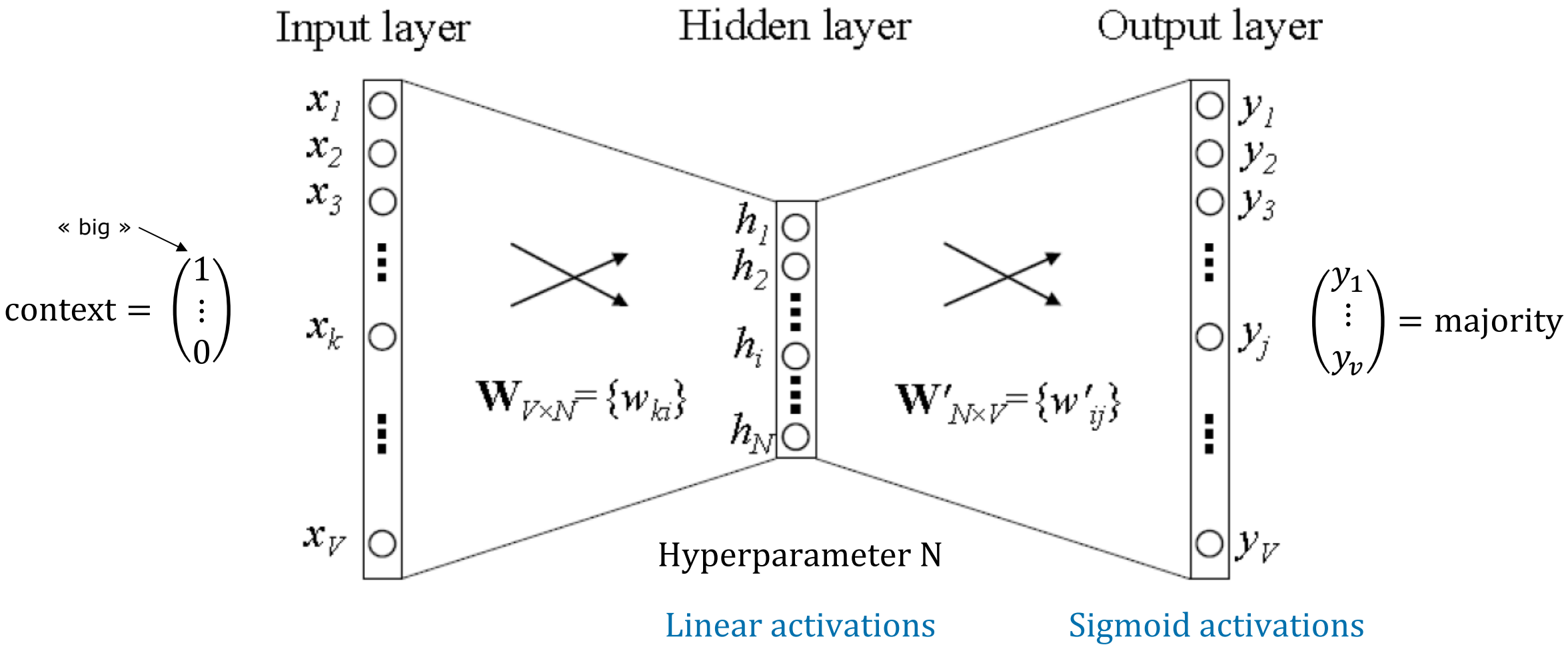
i.e. minimize:

$$E(W) = -ln\,P(y|x\,,W)$$

$$= -\sum_{i=1}^{n}\sum_{v=1}^{V} 1_{y_i = C_v}\ln P(y_i = C_v|\,x_i,W)$$

$$E(W) = -\sum_{i=1}^{n}\sum_{v=1}^{V} 1_{y_i = C_v}\ln\left(\frac{\exp(W_v^T\,x_i)}{\sum_{j=1}^{V}\exp(W_j^T\,x_i)}\right)$$

Word2Vec neural network



Labeling strategies

Try 1 Manual Labeling (Gold Standard Brexit)

- 2000 tweets labeled by professionals
  - pro, against, neutral, irrelevant, undecided/don't care
  - Strength given (1 weak, 5 strong) for pro and against Brexit
- Only considered pro and negative Brexit tweets (1246 tweets)
- 80% used for training and 20% for inference

Try 2 Automated Labeling via Influencers

Hand picked 27 influencers out of 49 (only pro and against Brexit).

11000 tweets from the influencers, labeled based on their view towards Brexit

Account	Occupation	Brexit position
OwenJones84	Author & Guardian columnist	Against
MhairiBlack	Scottish National Party MP	Against
tom_watson	Labour Party deputy leader	Against
GuidoFawkes	Right-wing political blogger	Pro
stellacreasy	Labour MP	Pro

Labeling strategies

Try 3 Automated labelling via hashtags

~6 Million Tweets

Hashtag Used	Pro/Against Brexit
#voteleave	Pro
#leaveeu	Pro
#outofeu	Pro
#ukip	Pro
#remain	Against
#strongerin	Against
#bettertogether	Against
#leavechaos	Against
#voteleavegetchaos	Against

“We are only as strong as we are united, as weak as we are divided." - albus dumbledore #remain

“The eu costs us £350 million every week. let's #voteleave and invest in our priorities instead.” #c4debate

Building a sentiment classifier

What we have

Tokenized tweets

{ big, border, Brexit, deliberately, deliver, exaggerated, Irish, issue, LordAshcroft, new, politicians, polling, public, show, squabble, stop, think }

Output from Word2Vec

Brexit	.565	.565	.565	.565	.565	.565	.565	.565	.565
.	.343	.343	.343	.343	.343	.343	.343	.343	.343
:	.34	.34	.34	.34	.34	.34	.34	.34	.34
Irish	.345	.345	.345	.345	.345	.345	.345	.345	.345
:	.354	.354	.354	.354	.354	.354	.354	.354	.354
:	.642	.642	.642	.642	.642	.642	.642	.642	.642
deliver	.23	.23	.23	.23	.23	.23	.23	.23	.23

What we want

Tweet represented in the form of vector
How do we do this? Term frequency-inverse document frequency (tf-idf)

$$tf - idf_{w,t} = (tf_{w,t}) \cdot \log(1 + \frac{1+n_t}{1+df_{w,t}})$$

Pseudocode

for each word in tweet
*inputVector = inputVector + embedded_vector_from_word2vec[word] * idf[word]*

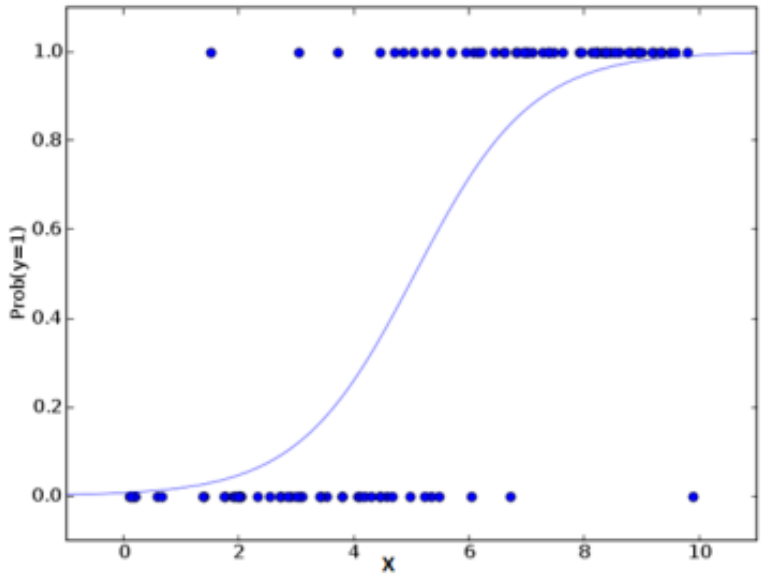
return inputVector

Picking the right model

Try 1 Logistic regression

$$J(\beta) = \frac{1}{N}\sum_{i=1}^{N}[-y_i log(h_\beta(x_i) - (1-y_i)log(1-h_\beta(x_i))]$$

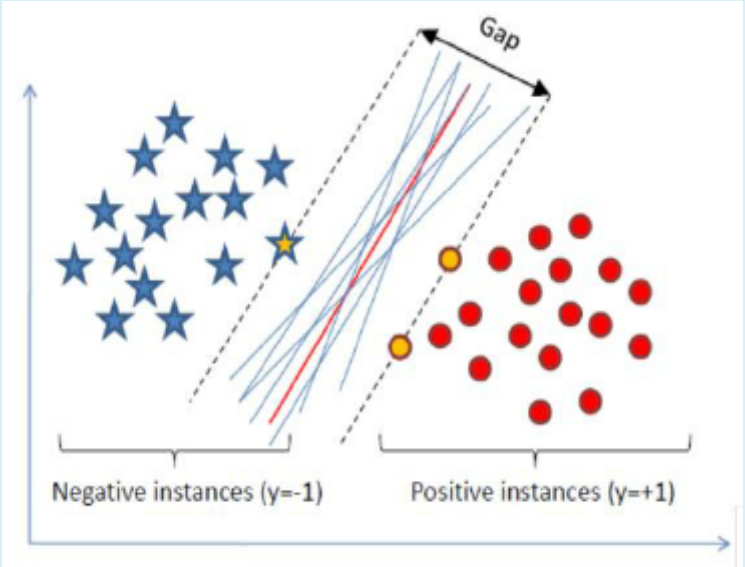
$$h_\beta(x) = \frac{1}{1+exp(-\beta^T x)}$$



Try 2 SVM

$$\min_W \|W\|^2$$

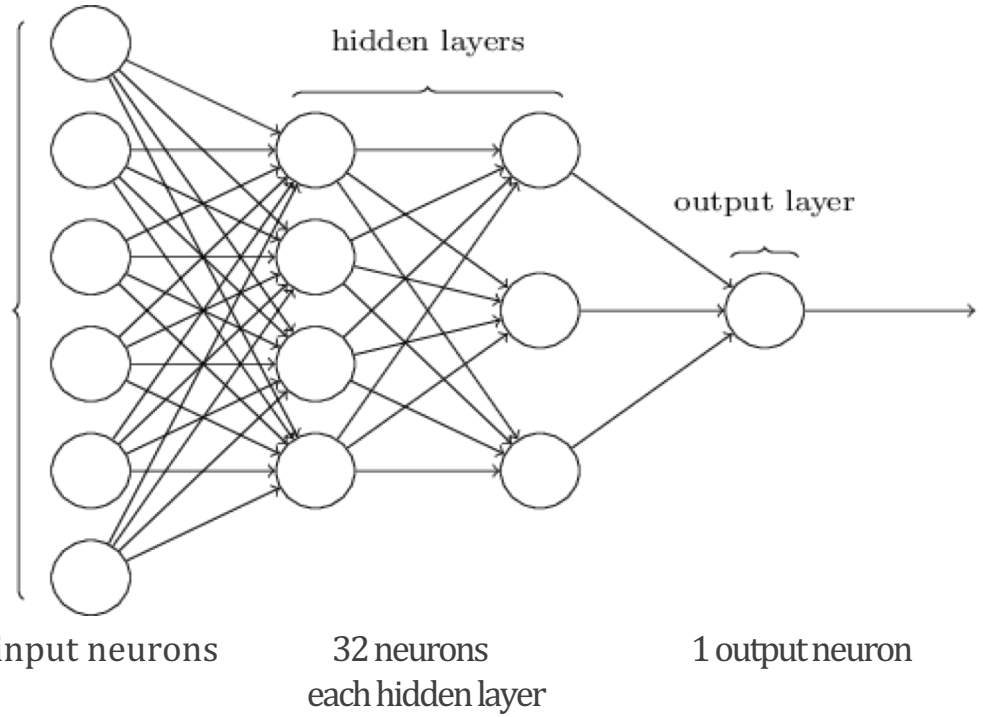
$$y_i(W^T x_i + b) \geq 1 \quad \forall i = 1...N$$



Picking the right model

Try 3 Neural network

Preliminary 2 layer neural network



Loss Function

Binary cross-entropy loss

$$J(\beta) = \frac{1}{N}\sum_{i=1}^{N}[-y_i log(h_\beta(x_i) - (1-y_i)log(1-h_\beta(x_i))]$$

Our choice: a neural network trained on hashtag-labeled data

Results table

According to classification model & labeling method.

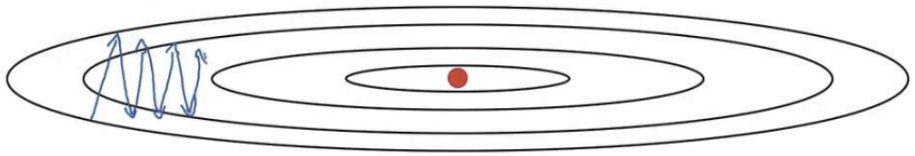
Method	Brexit Gold Standard Data	Influencer Data	Hashtag Data
Neural Network	76%	50%	79%
SVM (linear kernel)	75%	30%	75.6%
SVM (RBC kernel)	62%	12%	75.1%
Logistic Regression	72.5%	30%	75%

Picking an optimiser

Try 1 RMSProp

$$s_{k+1} = \beta s_k + (1-\beta)[\nabla_\theta L \odot \nabla_\theta L]$$

$$\theta_{k+1} = \theta_k - \alpha \frac{\nabla_\theta L}{\sqrt[2]{s_{k+1}} + \epsilon}$$

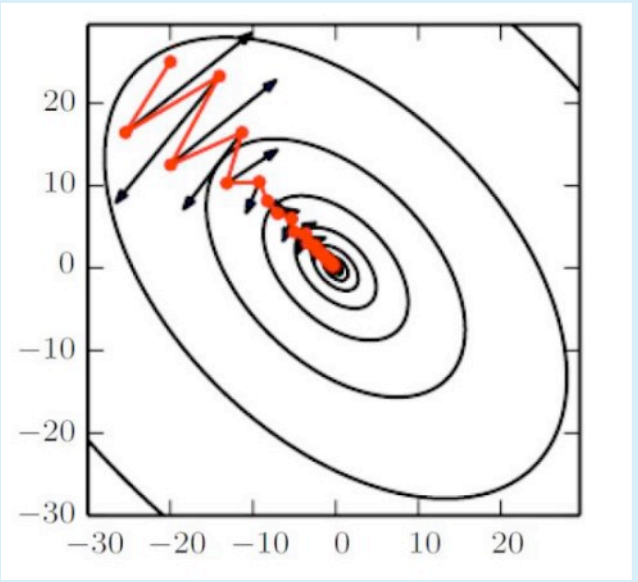


Try 2 Adam

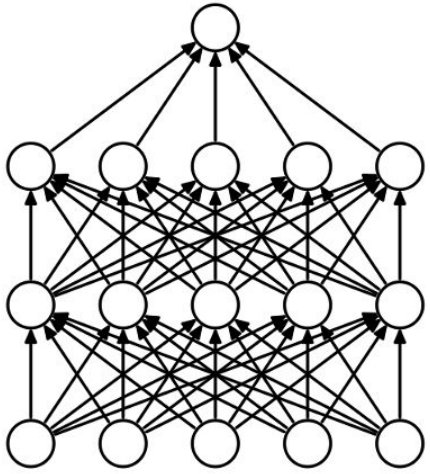
$$s_{k+1} = \beta_1 s_k + (1-\beta_1)[\nabla_\theta L \odot \nabla_\theta L]$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)(\nabla_\theta L)$$

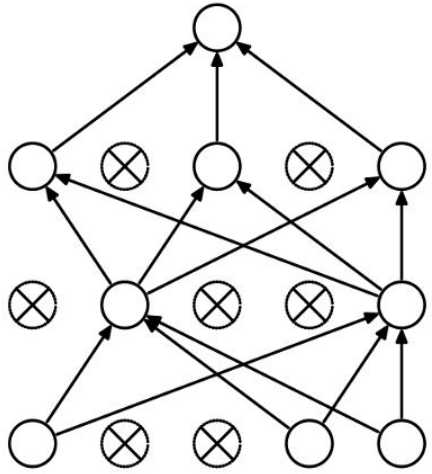
$$\theta_{k+1} = \theta_k - \alpha \frac{v_{k+1}}{\sqrt[2]{s_{k+1}} + \epsilon}$$



Try 3 Dropout



(a) Standard Neural Net



(b) After applying dropout.

Our choice: Adam with 5 layers

Results

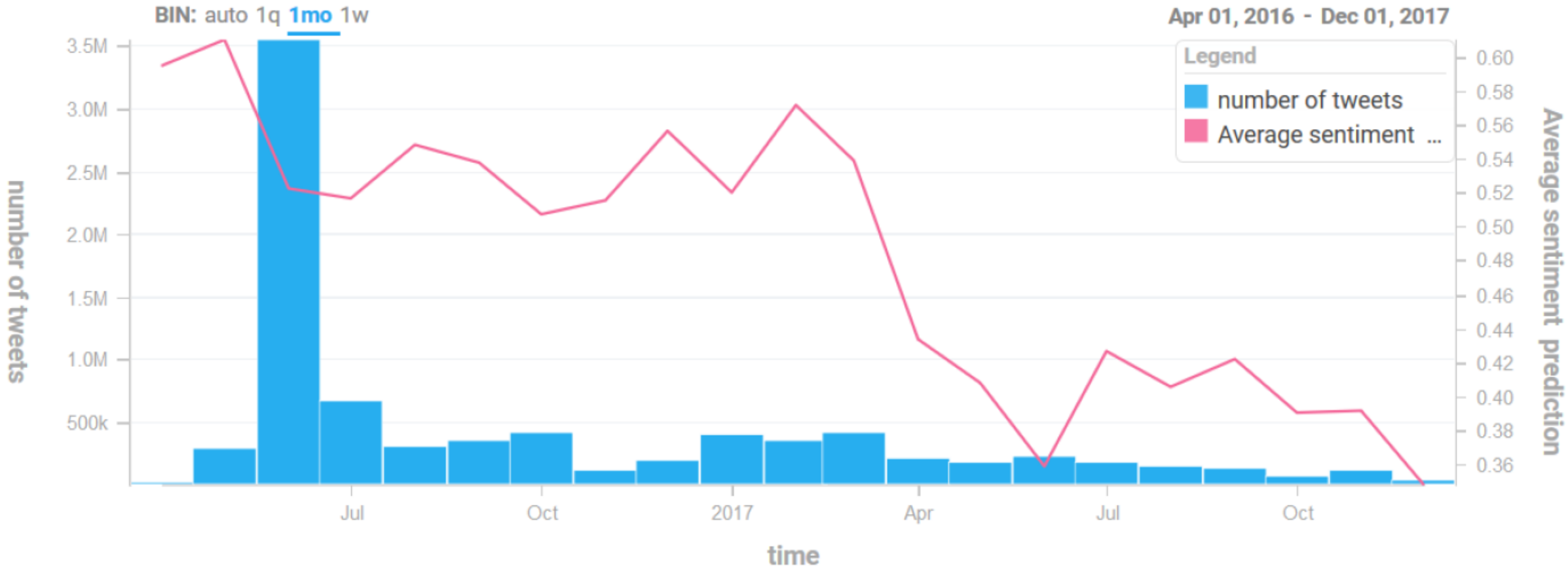
Hyperparameter tuning with hashtag based results.

	Optimiser	SGD		RMSProp		Adam	
	Dropout	0	0.5	0	0.5	0	0.5
Layers	Neurons						
2	32	0.79	0.79	0.82	0.82	0.82	0.82
2	64	0.79	0.79	0.83	0.82	0.82	0.82
3	32	0.80	0.77	0.82	0.81	0.81	0.81
3	64	0.81	0.79	0.82	0.82	0.82	0.82
4	32	0.81	0.78	0.83	0.80	0.82	0.80
4	64	0.82	0.79	0.81	0.82	0.82	0.82
5	32	0.80	0.78	0.81	0.80	0.81	0.80
5	64	0.82	0.78	0.83	0.81	0.83	0.82
6	32	0.82	0.74	0.81	0.79	0.82	0.80
6	64	0.81	0.77	0.82	0.81	0.82	0.80
7	32	0.81	0.57	0.81	0.80	0.82	0.78
7	64	0.82	0.75	0.81	0.81	0.83	0.78

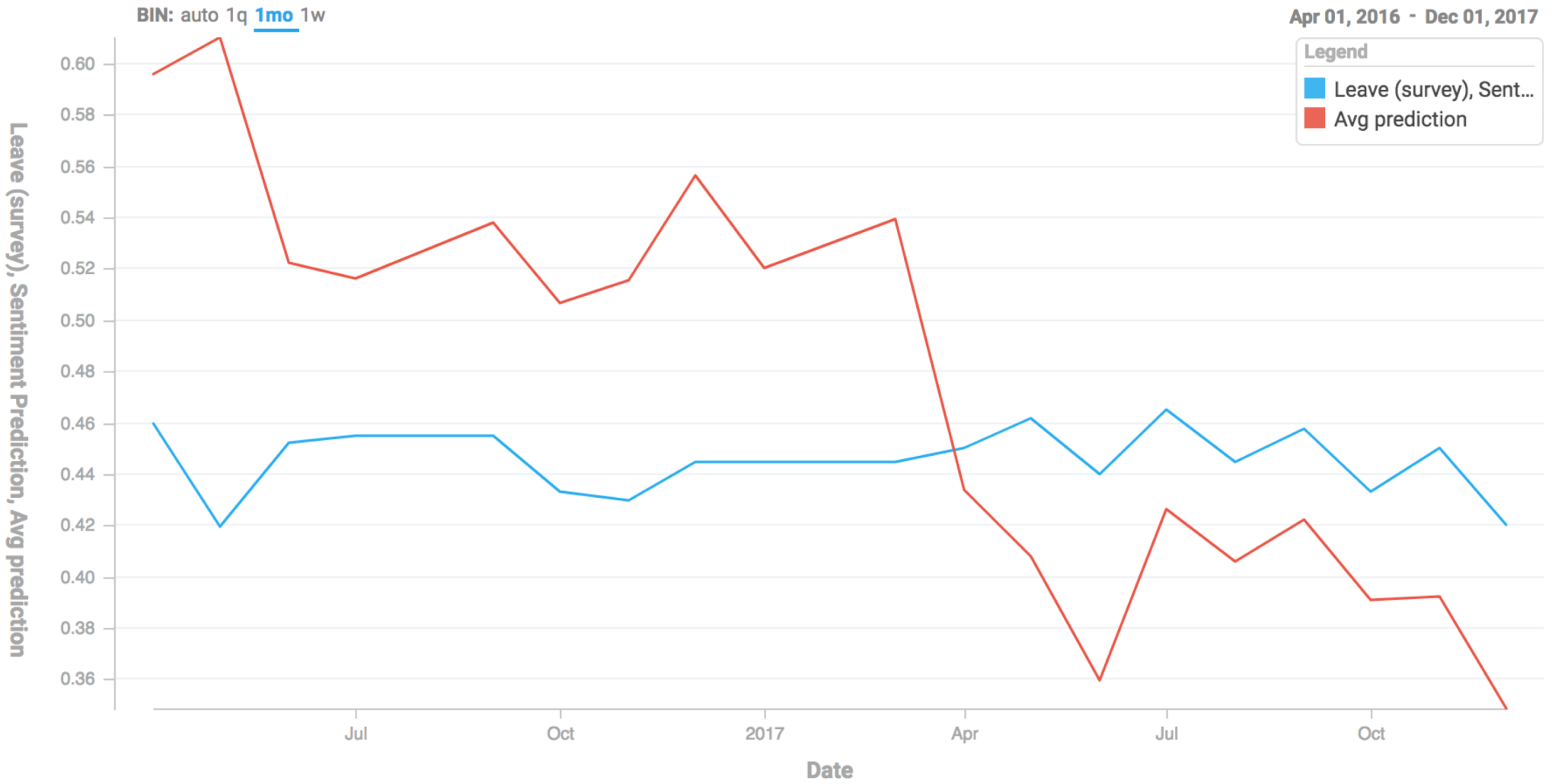
Sentiment over time demo

Tweet frequency with sentiment over time

Tweet frequency with average sentiment over time



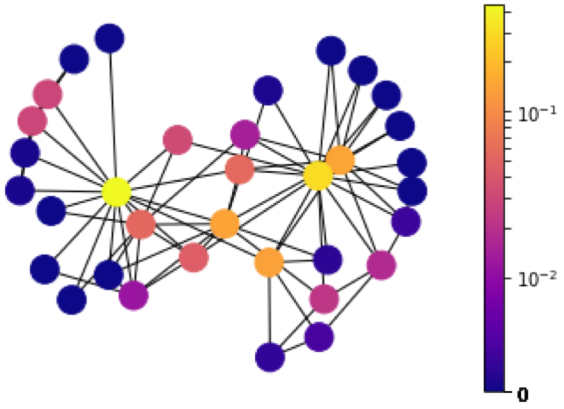
Comparison of survey and sentiment analysis based prediction



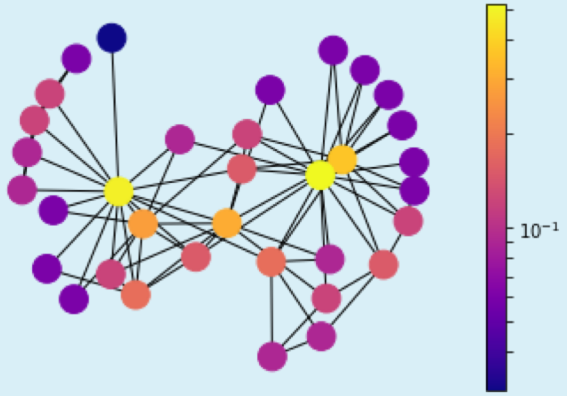
Getting more insights: Network analysis

Influence measures

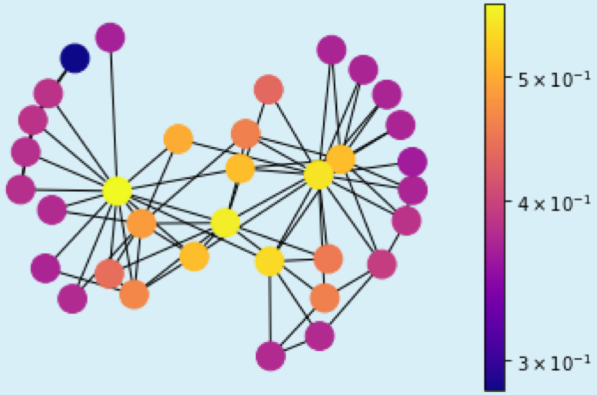
Option 1 Betweenness centrality



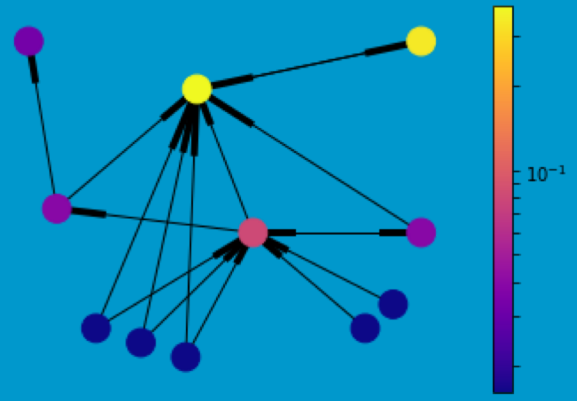
Option 2 Degree centrality



Option 3 Closeness centrality



Option 4 Page rank



Challenges

Data requirements for calculating influence

- List of users in the network (brexit tweeters)
- List of followers for each user

Steps taken

- Twitter Developer Account registration
- list of users from Brexit tweets mining
- Script using python-twitter library to extract followers for each user developing

Challenge – API allows 5000 followers per minute & influential users have ~ million followers each - hours to extract the data of even one user

Conclusion & Future Works

- Social Media serves as a massive opinion pool
- Sentiment analysis is a very powerful method for analysing people's opinions
  - Ability to analyse bigger dataset than surveys
  - May provide more insights than other survey methods
    - Survey predicted remain, sentiment analysis predicted leave
    - Sentiment analysis prediction gave 52% leave (vote result 51.9%)
    - Flexibility
- Big data analytics crucial for real-world applications
- Ethics matters
- Generic models not applicable for specific problems

- Different automated labeling strategies
- Neural network design
- Network analysis

[{Thank}, {you}, {for}, {listening}]

#Questions?

People matter, results count.

About Capgemini

A global leader in consulting, technology services and digital transformation, Capgemini is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms. Building on its strong 50-year heritage and deep industry-specific expertise, Capgemini enables organizations to realize their business ambitions through an array of services from strategy to operations. Capgemini is driven by the conviction that the business value of technology comes from and through people. It is a multicultural company of 200,000 team members in over 40 countries. The Group reported 2016 global revenues of EUR 12.5 billion.

Learn more about us at

www.capgemini.com