American Election 2016 Tweet Analysis

SIMONE MONTI
VITTORIO MAGGIO
GIANMARIA BALDUCCI

	handle	text	is_retweet	original_author	time	entities	extended_entities
0	HillaryClinton	The question in this election: Who can put the...	False	NaN	2016-09-28T00:22:34	{'media': [{'display_url': 'pic.twitter.com/Xr...	{'media': [{'display_url': 'pic.twitter.com/Xr...
1	HillaryClinton	Last night, Donald Trump said not paying taxes...	True	timkaine	2016-09-27T23:45:00	{'media': [{'display_url': 'pic.twitter.com/t0...	{'media': [{'display_url': 'pic.twitter.com/t0...
2	HillaryClinton	Couldn't be more proud of @HillaryClinton. Her...	True	POTUS	2016-09-27T23:26:40	{'user_mentions': [{'id_str': '1536791610', 'n...	NaN
3	HillaryClinton	If we stand together, there's nothing we can't...	False	NaN	2016-09-27T23:08:41	{'media': [{'display_url': 'pic.twitter.com/Q3...	{'media': [{'display_url': 'pic.twitter.com/Q3...
4	HillaryClinton	Both candidates were asked about how they'd co...	False	NaN	2016-09-27T22:30:27	{'user_mentions': [], 'symbols': [], 'urls': [...	NaN
...	...	...	...	...	...	...	...
6439	realDonaldTrump	"@lilredfrmkokomo: @realDonaldTrump My Faceboo...	False	NaN	2016-01-05T03:47:14	{'user_mentions': [{'id_str': '26122621', 'nam...	NaN
6440	realDonaldTrump	"@marybnall01: @realDonaldTrump watched lowell...	False	NaN	2016-01-05T03:44:17	{'user_mentions': [{'id_str': '3477455725', 'n...	NaN
6441	realDonaldTrump	"@ghosthunter_lol: Iowa key endorsement for @r...	False	NaN	2016-01-05T03:42:10	{'media': [{'display_url': 'pic.twitter.com/JB...	{'media': [{'display_url': 'pic.twitter.com/JB...
6442	realDonaldTrump	"@iLoveiDevices: @EdwinRo47796972 @happyjack22...	False	NaN	2016-01-05T03:39:11	{'user_mentions': [{'id_str': '42568997', 'nam...	NaN
6443	realDonaldTrump	"@SalRiccobono: @realDonaldTrump @troyconway D...	False	NaN	2016-01-05T03:36:53	{'user_mentions': [{'id_str': '2358794621', 'n...	NaN

6444 rows × 7 columns

Trump: 3218

Clinton: 3226

Starting Dataset

Columns

- **Handle**
- **Text**
- ~~Is_retweet~~
- ~~Original_author~~
- ~~Time~~
- ~~Entities~~
- ~~Extended_entities~~

Steps

01

Named entity recognition and linking

02

Implementation of a polarity model

03

Tweets Analysis

04

WebApp

Named entity recognition and linking

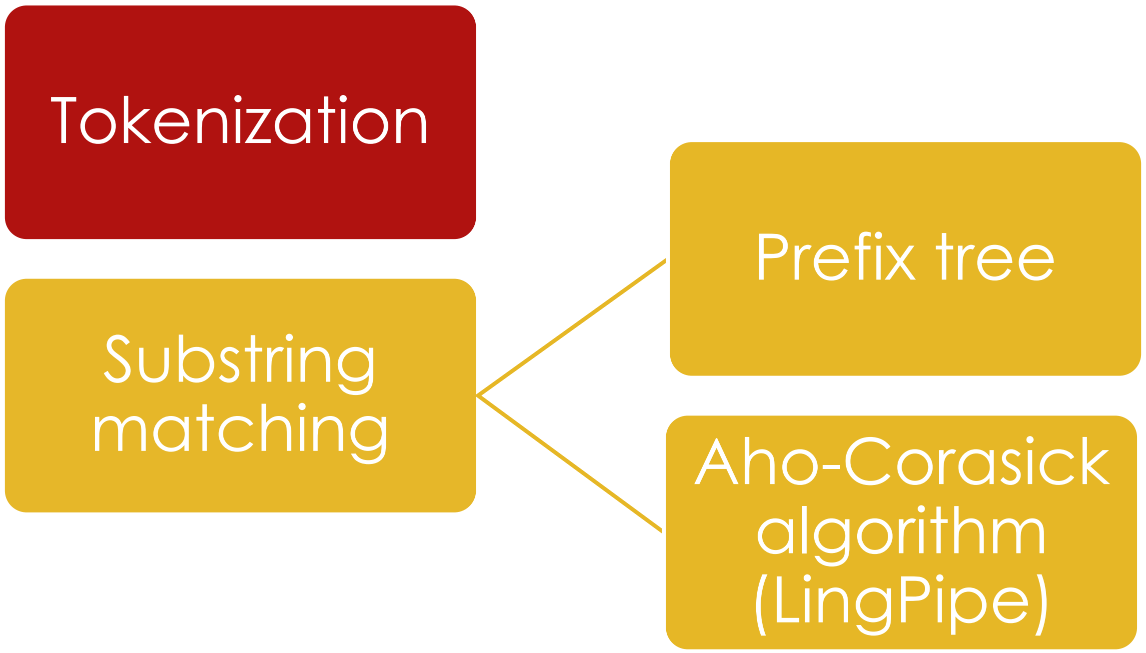
Stages

SPOTTING

CANDIDATE SELECTION

DISAMBIGUATION

Spotting stage



Candidate selection stage

1. Generation of candidates by traversing a finite state automaton encoding all possible sequences of tokens that form known spot candidates
   - OpenNLP
2. Selection of the best candidates
   - overlaps in the candidates are resolved based on a **score** and a **preference-based choice**: PER>ORG>LOC>MISC>NP>MWU>PP>FSA lookup>Capitalized Sequence.
   - Before seeing context – as a «default sense»
   - all candidates that fall below a specified score threshold are removed

Score – candidate selection

$e = entity$
$s = anchor\ text\ (spot)$

- Based on wikilinks
- Annotation probability (prior probability):
  $P(annotation|s) = \sum_e count(e, s) / count(s)$
- Sometimes bad performance (es: with acronyms)

A Generative Entity-Mention Model for Linking Entities with Knowledge Base

Xianpei Han **Le Sun**
Institute of Software, Chinese Academy of Sciences
HaiDian District, Beijing, China.
{xianpei, sunle}@nfs.iscas.ac.cn

Disambiguation

HTTPS://WWW.ACLWEB.ORG/ANTHOLOGY/P11-1095.PDF

Disambiguation

- Generative Probabilistic Model – Entity-mention Model
- Wikipedia dataset:
  - $e \rightarrow entity$
  - $s \rightarrow phrase$
  - $c \rightarrow context$
  - $M \rightarrow article\ links\ with\ their\ anchor\ texts\ and\ textual\ context$

- $P_{LM} \rightarrow$ the smoothed general language model probability of a token that we estimate over all tokens imported to the system as context of an entity mention
  - The Jelinek-Mercer smoothing parameter $\lambda = 0.2$
- $P_{LM} = P(e) * P(s|e) * P(c|e)$
- The hypothesis that the context and phrase were not generated by any known entity - all entity candidates with a lower score than the NIL entity are removed

- $P(e) \rightarrow$ distribution of entities in document
- $P(s \mid e) \rightarrow$ the distribution of possible names of a specific entity
- $P(c \mid e) \rightarrow$ the distribution of possible contexts of a specific entity.

$$P(e) = \frac{\text{count}(e)}{|M|}$$

$$P(s|e) = \frac{\text{count}(e, s)}{\text{count}(e)}$$

$$P(c|e) = P_e(t_1) \cdot P_e(t_2) \cdot P_e(t_3) \cdot \ldots \cdot P_e(t_n)$$

$$P_e(t) = \lambda P_{e\_\text{ML}}(t) + (1 - \lambda) P_{\text{LM}}(t)$$

$$P_{e\_\text{ML}}(t) = \frac{\text{count}_e(t)}{\sum_t \text{count}_e(t)}$$

$$P(\texttt{NIL}) = \frac{1}{|M|}$$

$$P(s|\texttt{NIL}) = \prod_{t \in S} P_{\text{LM}}(t)$$

$$P(c|\texttt{NIL}) = \prod_{t \in C} P_{\text{LM}}(t)$$

Our configuration

- Our confidence: 0.35
  - It will only annotate resources if the **contextual ambiguity** is less than (1−confidence)= 0.65
  - Given a confidence of 0.7, we get the **topical pertinence** threshold that 70% of the wrong test samples are below

Entity

Trump

- Tweet without entities: 9.88%
- Different entities identified: 1856
- Surface-form identified: 2094
- Total entities: 7470

Clinton

- Tweet without entities: 12.43%
- Different entities identified : 1718
- Surface-form identified: 1901
- Total entities: 6615

Most common entities/types

Trump

label	count	types
Donald_Trump_on_social_media	330	[]
Donald_Trump	298	[Person, Agent, Politician]
United_States	185	[PopulatedPlace, Place, Location, Country]
Hillary_Clinton	173	[Person, Agent, Politician]
Make_America_Great_Again	150	[]
Jervy_Cruz	115	[Person, Athlete, Agent, BasketballPlayer]
President_of_the_United_States	113	[]
The_Tonight_Show	111	[Work, TelevisionShow]
CNN	108	[Organisation, Broadcaster, Agent, TelevisionS...
Fox_News	97	[Organisation, Broadcaster, Agent, TelevisionS...

tipo	count
Without_type	2817
Agent	2241
Place	1389
Location	1389
Person	1353
PopulatedPlace	1353
Politician	1009
Organisation	846
Work	831
AdministrativeRegion	721
Region	721

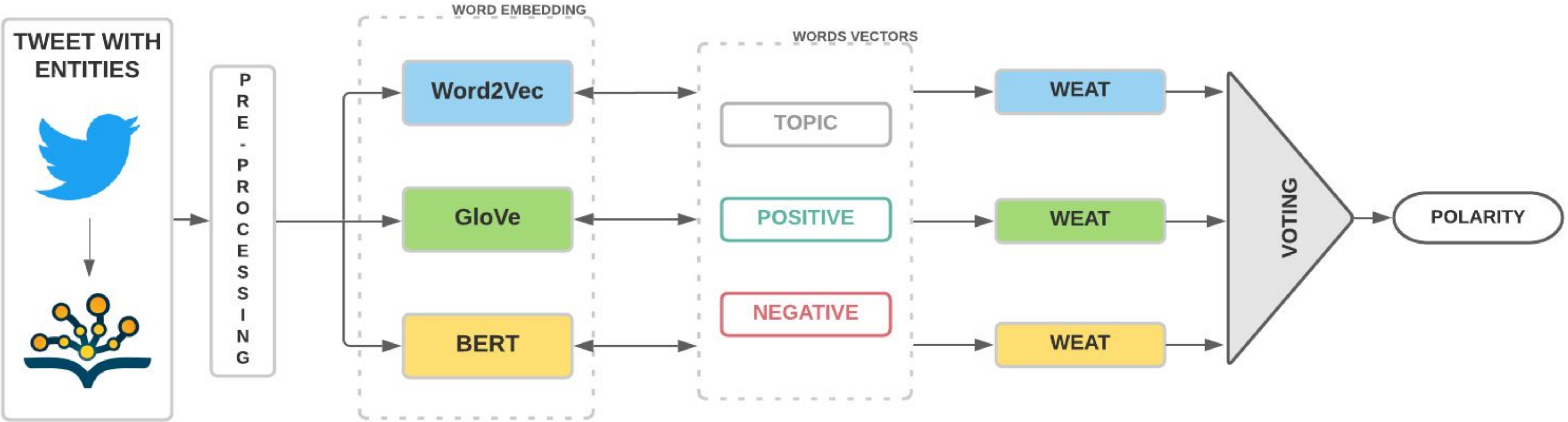
Clinton

label	count	types
Donald_Trump	884	[Person, Agent, Politician]
President_of_the_United_States	396	[]
United_States	336	[PopulatedPlace, Place, Location, Country]
Donald_Trump_on_social_media	116	[]
Hillary_Clinton	108	[Person, Agent, Politician]
Hydrogen	106	[ChemicalSubstance, ChemicalCompound]
Economy	59	[]
Republican_Party_(United_States)	46	[Organisation, Agent, PoliticalParty]
The_Tonight_Show	44	[Work, TelevisionShow]
LGBT	43	[]

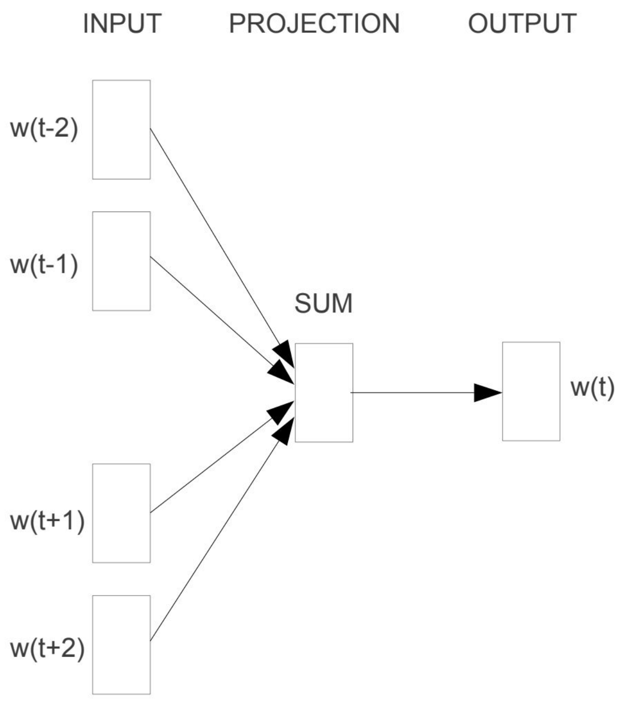
tipo	count
Without_type	3455
Agent	1720
Person	1265
Politician	1180
Location	753
Place	753
PopulatedPlace	700
Organisation	428
Country	422
Work	360

Polarity

Polarity model



Word2Vec



CBOW

Architecture: Continuos Bag-of-Words

- Predict the **current word** based on the **surrounding words** (context words)
- The **objective function** for CBOW is:

$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

- Window : 5

GloVe

- Co-occurrence matrix **X**
- Co-occurrence probabilities $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$
- **Objective**: construct a function $F(w_i, w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- … after a series of steps, we obtain a simplification:

$$w_i^T \widetilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

some parameters to be selected

We are interested in these vectors!

- Then, GloVe builds an **objective function J** that associates word vector to text statistics:

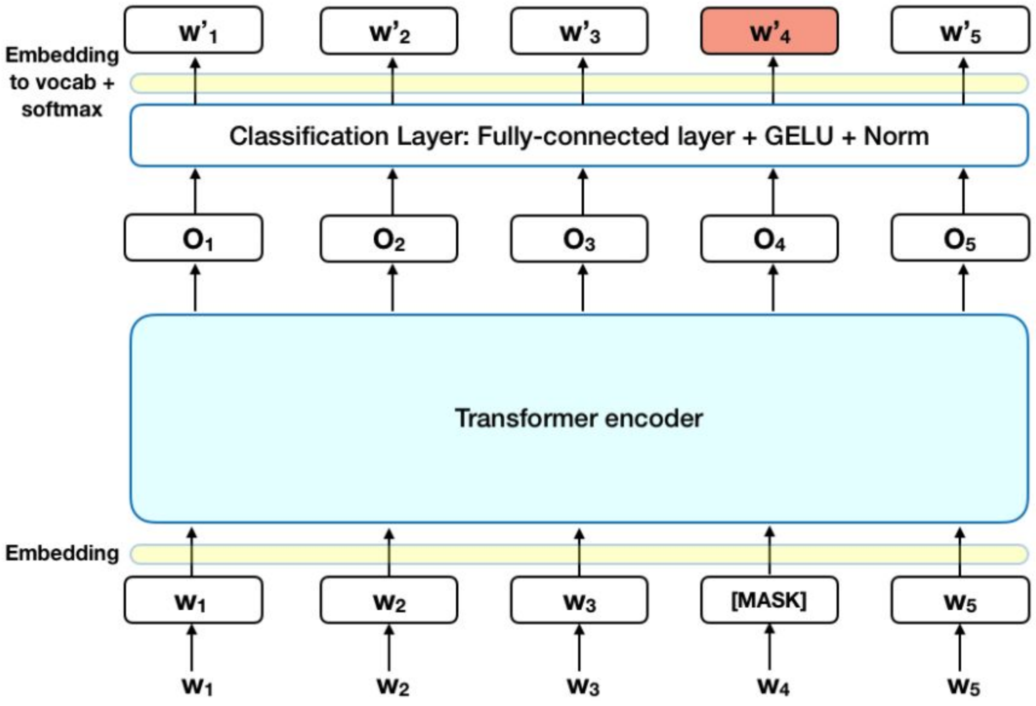
$$J = \sum_{i,k=1}^{V} f(X_{ik})\left(w_i^T \widetilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik})\right)^2$$

(least squares problem)

$$X_{final} = U + V$$

(both capture similar co-occurrence information)

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)



- **Tokenized Input:**
  ['[CLS]', 'I', '[MASK]', 'a', 'police', '##woman', '[SEP]']
- It learns **contextual word representations:**
  - the vector of a word changes with respect to the context
- **Pre-trained model:**
  - **Data:** Wikipedia (2.5B words) + BookCorpus (800M words)
  - **Batch Size:** 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
  - **Optimizer**: Adam, 1e-4 learning rate, linear decay
  - **Architecture**: 12-layer; 768-hidden-layer; 12-head

WEAT - Caliskan

- Single target adaption:

$$S(W, A, B) = \sum_{w \in W} s(w, A, B)$$

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} cosim(w, a) - \frac{1}{|B|} \sum_{b \in B} cosim(w, b)$$

- W = topic, A = positive, B = negative

```
positive_trump = ['good', 'great', 'nice', 'positive', 'love', 'honest' ]
negative_trump = ['bad', 'badly', 'negative', 'false', 'wrong', 'dangerous']

positive_clinton = ['good', 'great', 'nice', 'positive', 'love', 'honest' ]
negative_clinton = ['bad', 'negative', 'hate', 'fail', 'dangerous', 'war']
```

```
# Republican party in Trump
topic = ['donaldtrump',  'donaldtrumponsocialmedia', 'tedcruz', 'cruz', 'marcorubio',
         'mittromney', 'jebbush', 'georgewbush', 'trump', 'rep', 'republican']
```

```
# Democratic party in Trump
topic = ['hillaryclinton', 'hillary', 'barackobama', 'berniesanders',
         'billclinton', 'elizabethwarren', 'timkaine', 'dem', 'democratic']
```

Analysis

Who are the politicians most cited? What they think about them?

Trump

Democratic

Hillary_Clinton	173
Barack_Obama	71
Bernie_Sanders	41
Bill_Clinton	27
Elizabeth_Warren	25
Tim_Kaine	8

Republican

Donald_Trump	298
Ted_Cruz	81
Marco_Rubio	70
Mitt_Romney	27
Jeb_Bush	23
George_W._Bush	10

Clinton

Democratic

Hillary_Clinton	108
Barack_Obama	40
Bill_Clinton	13
Bernie_Sanders	6
Tim_Kaine	6
Franklin_D._Roosevelt	5

Republican

Donald_Trump	884
Mike_Pence	26
Ted_Cruz	6
Abraham_Lincoln	6
David_Duke	5
John_McCain	4

Politician polarity

Leader	Party	W2V	Glove	Bert	Majority
Trump	Republican	-0.00394121	0.45970071	0.30018856	+
	Democratic	-0.00663326	-0.39756773	0.11672534	-
Clinton	Republican	-0.04819707	0.315081534	0.02750225	+
	Democratic	-0.047975619	0.280487134	0.18820126	+

What do the leaders think of each other?

Trump

```
[('hillaryclinton', 1.0),
 ('beat', 0.9996077418327332),
 ('berniesanders', 0.998988151550293),
 ('cant', 0.9989197850227356),
 ('wants', 0.9988921880722046),
 ('barackobama', 0.9987609386444092),
 ('says', 0.9987530708312988),
 ('bernie', 0.9986552596092224),
 ('lyin', 0.998586893081665),
 ('believe', 0.9985146522521973)]
```

Clinton

```
[('donaldtrump', 0.9999998807907104),
 ('donaldtrumps', 0.9999561905860901),
 ('one', 0.999947726726532),
 ('hillary', 0.9999470710754395),
 ('us', 0.9999446868896484),
 ('hillaryclinton', 0.9999420642852783),
 ('families', 0.9999403953552246),
 ('people', 0.9999358654022217),
 ('need', 0.9999339580535889),
 ('hydrogen', 0.9999333024024963)]
```

Which are the non-american countries cited? What are the opinions?

China

Russia

Iraq

Iran

Trump – Entities used with countries

China	• Islamic_State_of_Iraq_and_Levant
Russia	• Crimea
Iraq	• Libya
Iran	• Cash • United States

Ex:

"Crooked Hillary just can't close the deal with Bernie. It will be the same way with ISIS, and China on trade, and Mexico at the border. Bad!"

"Crooked Hillary Clintons foreign interventions unleashed ISIS in Syria, Iraq and Libya. She is reckless and dangerous!"

Trump - Countries polarity

Country	W2V	Glove	Bert	Majority
China	-0.00437261	0.02320465	-0.02463866	–
Russia	-0.00377818	0.15094581	-0.07769103	–
Iraq	-0.00256812	0.15774842	-0.00324684	–
Iran	-0.00410711	0.04681618	-0.00384347	–

Clinton – Entities used with countries

China	• Vietnam
Russia	• Moscow Kremlin
Iraq	• Family
Iran	• Money laundering

Ex:

"Praying for a safe Eid Al-Fitr. My heart breaks for families struck by terror in Turkey, Iraq, Saudi Arabia, and Bangladesh this Ramadan. –H"

"How can you be tough on Iran, given your business partnership with someone connected to Iranian money laundering?"

Clinton - Countries polarity

Country	W2V	Glove	Bert	Majority
China	-0.05291671	-0.01586804	0.00485769	–
Russia	-0.04848618	-0.20686781	-0.00899924	–
Iraq	-0.04357786	-0.19340154	-0.04021230	–
Iran	-0.05132032	0.04703887	-0.01632032	–

Immigration

Trump	• Illegal immigration
Clinton	• Deportation • Donald Trump

	W2V	Glove	Bert	Majority
Trump	-0.00309141	-0.48279724	-0.063714448	–
Clinton	-0.04695621	-0.01348883	-0.037578545	–

Media

Leader	Media cited	W2V	Glove	Bert	Majority
Trump	CNN, Fox News, The New York Times	-0.0040066	0.66897907	0.15018507	+
Clinton	RT, The new York Times	-0.04782281	-0.05068429	0.23628618	-

Limits

- Confidence: 0.35 → Precision

Donald_Trump	298
Hillary_Clinton	173
Jervy_Cruz	115
CNN	108
Fox_News	97
Enjoy_Records	89
Ted_Cruz	81
Barack_Obama	71
Marco_Rubio	70

label	count	plain_text
United_States_Senate	29	Senate, senator, Senate}
RT_(TV_network)	26	{RT}
Orlando_Magic	18	{Orlando}
Portland_Timbers_U23s	17	{por}
Trump_University	17	{Trump University}
Imagine_Software	16	{Imagine}

label	count	plain_text
CNN	108	{cnn, CNN}
Fox_News	97	{Foxnews, Fox News, Fox News Channel, foxnews,...
Enjoy_Records	89	{Enjoy!, ENJOY!, Enjoy, enjoy!}
nited_States_Senate	54	{senator, US Senator, Senate, Senators, Senator}
adcasting_Company	22	{Fox, FOX, Fox Network}
People!	21	{people!}
National_Committee	12	{RNC, Republican National Committee}

Future works

- Expand the dataset with other tweets to improve the word embedding
- Use different algorithms of NER and NEL
- Use Cade to align the different embeddings and improve the analysis

References

- Improving Efficiency and Accuracy in Multilingual Entity Extraction - Joachim Daiber, Max Jakob
- DBpedia Spotlight: Shedding Light on the Web of Documents - Pablo N. Mendes, Max Jakob, Andrés García-Silva, Christian Bizer
- A generative entity-mention model for linking entities with knowledge base - X. Han and L. Sun.
- GloVe: Global Vectors for Word Representation - Jeffrey Pennington, Richard Socher, Christopher D. Manning
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
- Semantics derived automatically from language corpora contain human-like biases - Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan

Thank you for your attention