Sentiment Analysis and Topic Modeling based on US 2016 Presidential Election Twitter Data

Miltenburg Lino[1 [11136375]], Tsimpoukis Dimitrios[1 [12338281]], Schniepp Michael[1 [12160067]], Symeonidou Anthi[1 [12296082]], and Wilmers Ben[1 [12292117]]

University of Amsterdam, Amsterdam, NL

Abstract. In this work, we study Twitter posts during a monthly period in 2016 related to the US presedential election. Our analysis composed of sentiment analysis towards candidates as well as topic inference from the acquired tweets. We began with performing sentiment analysis based on a Lexicon approach and drew conclusions about the overall as well as state-specific candidate preference. To continue, we performed topic-modeling analysis utilizing the Latent Dirichlet Allocation technique to infer the most discussed topics both on the entire dataset, but also on candidate-targeted posts.

Keywords: sentiment analysis · topic modeling · LDA · social media · twitter · american elections 2016

Introduction

The social media platform Twitter offers millions of individuals the ability to share their thoughts or opinions from anywhere on the planet in the form of a short text. Due to the sheer volume of Tweets being generated every moment, these small texts paired with their meta-data can be evaluated to determine a plethora of information related to what people are currently interested in discussing. We have utilized this platform to gain insight into the 2016 US Presidential Election, to better understand the most discussed topics, the sentiments of the users towards the different candidates and compare them with the final election results. The two main parts of the analysis will be sentiment analysis towards candidate preference and topic-modelling.

Methodology

To initialize our analysis on the 2016 US Presidential election using twitter we utilized a dataset of 657,307 tweets provided by the course. The tweets are geotagged, date from the period of 12.08-12.09, 2016, and were selected according to the following tags/users: `@HillaryClinton`, `#maga`, `#trumppence16`, `#hillaryclinton`, `#hillary`, `#crookedhillary`, `#donaldtrump`, `#dumptrump`, `@realDonaldTrump`, `#nevertrump`, `#imwithher`, `#neverhillary`, `#trump`.

As shown above, the tweets were selected to monitor tweets aimed at the two candidates Donald Trump and Hillary Clinton.

Sentiment Analysis

Before we could begin the Sentiment Analysis task, we extracted and cleaned the text of each tweet in order to optimize the results. The data cleaning procedure for the sentiment analysis involved removing punctuation marks and other non-standard alphabetical characters (including hyperlinks), thus leaving the text as only the constituent words.

In order to analyze the sentiments of our collection of tweets we required a means to classify the sentiments of each tweets. The tweets were classified as: 'positive', 'negative', and 'neutral'. To perform this task, we used a lexicon-based approach which utilized a library of pre-scored words. The scoring was provided by the *pattern.en* library, produced by the Computational Linguistics & Psycholinguistics Research Center (CLiPS). Each word in the tweet was given a sentiment score with the scores ranging from -1 to 1; -1 being negative and 1 being positive. Once the score of each word was evaluated, the total text was averaged to give an overall score. Once the text was scored we classified tweets with scores over 0 as positive and scores below 0 as negative, while scores of 0 were neutral. Classifying exactly 0 as neutral was an appropriate classification as many tweets contain little useful information and were thus scored as 0.

To continue, we also divided the tweets into 6 subcategories to emphasize on the direction of the sentiment. Each tweet could be classified as either pro-trump, pro-Hillary, anti-Trump, anti-Hillary, neutral-Trump, and neutral-Hillary. These classifications were determined by parsing the text for specific key words that would indicate whether the tweet was focused on one of the two candidates examined. No mentions of either or mentions of both candidates resulted in a neutral direction. The rest were categorized purely Trump or Hillary directed. This categorization paired with the sentiment resulted in the sentiment-candidate directional classification.

Topic Modeling using LDA

In this step, we identified abstract topics that occurred in a collection of documents(tweets) based on the primary tweet selection we had done previously. Latent Dirichlet Allocation is an algorithm that uses the bag-of-words approach and treats each tweet as a vector of word counts. Topics are represented as probability distributions over a number of words, with some words falling under multiple topics. Thus, LDA can elicit the topics by calculating the distribution value of topics in each tweet [1]. To perform this task we used `gensim`, a library that contains methods for training the model and identifying topics. The following steps were taken:

Pre-processing. We extracted the "text" attribute of each tweet in our database which we built using the *MongoDB* client and removed all the stop words using the *NLTK* library. We then applied lemmatization using the *wordnet* package and also removed all hashtags(#), mentions(@), links(http(s)), digits and words of fewer than 3 characters. Additionally, we further filtered our data by keeping

only the nouns using the `postag` tool. We then split the tweets into words(tokens) using the *RegexpTokenizer* to arrive at the final corpus.

Preparing document-term matrix. Using the class *Dictionary* from `gensim` we assigned an index to each term of our corpus. This resulted in vectors bearing a word id and word counts.

Run the LDA model. We used the class *LdaMultiCore* to speed up the training process which allows us to train multithreaded. We ran a lot of different configurations, changing the alpha hyperparameter (document-topic density),the number of topics(10, 20, 50) and the number of passes/iterations. We ended up with all multiple models of topics which were mentioned in the tweets and represented by the probabilities of the most contributing words [2].

Results and Discussion

Sentiment Analysis

Our first use of the sentiment analysis was to get an overview of the directional sentiment of the tweets, shown in Table 1.

Sentiment	**Percentage(%)**
neutral	0.419
positive	0.337
negative	0.243

Table 1: Percentage of directional sentiment of the tweets

In Figure 1, the tweet direction, being about Trump or Clinton, and sentiment are combined. We can conclude that there have been significantly more tweets about Trump than Clinton.

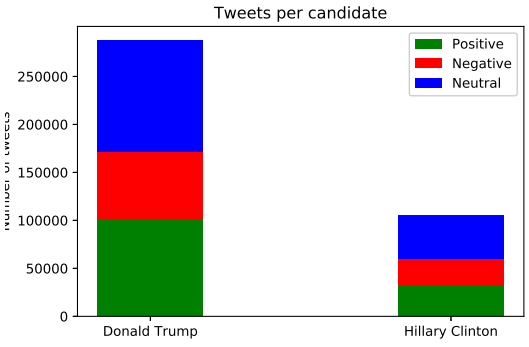
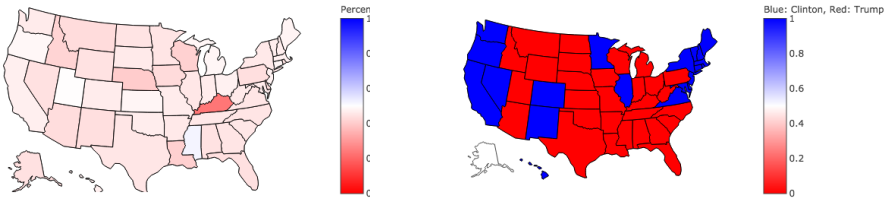


Fig. 1: Sentiment analysis per candidate

Our second use of sentiment classification was to determine if the directional sentiments were correlated to the actual election results on a state-by-state basis. In order to do this we aggregated the total number of favorable tweets for each candidate on a state level. All tweets that were pro-Trump and anti-Hillary were totalled in favor of Trump, while the neutral tweets were not considered. Thus, we divided the totals for each candidate by the total non-neutral tweets for the state to get a percentage of favorable tweets for each candidate. We then declared the 'winner' of each state as the candidate with the higher favourability percentage. These results compared to the actual outcomes (shown below) are not quite accurate, but one interesting insight is that the ratio for each state is very close to the actual ratio of final electoral college votes (Figure 2). The actual result of the electoral college votes was approximately 57 % in favour of trump, whereas the majority of states in our computation showed a ratio of about 55 % in support of Trump [3].



(a) Sentiment analysis results (b) Actual results

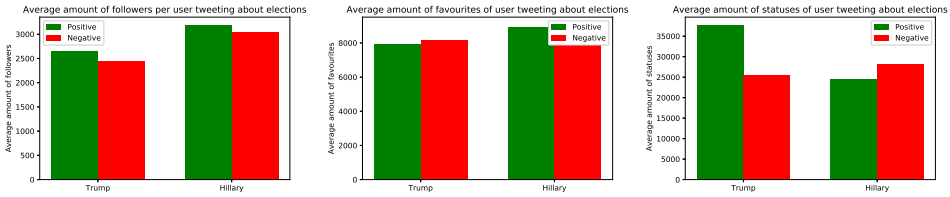
Fig. 2: Map representation

Users Popularity

We further utilized our sentiment analysis results to investigate the correlation between the popularity/activity of the user tweeting about the elections, and the direction of the sentiment going to Trump or Clinton. To investigate the popularity of a user we extracted the number of followers of the user and the total number of the users tweets that have been favourited. To indicate activity of the user we focused on the total amount of tweets posted by a user.
The first graph in Figure 3 shows the average amount of followers of a user according to their candidate preference. The bar chart reveals that the users tweeting about Clinton, both positively and negatively, have on average more followers than users tweeting about Trump. For both candidate preferences, the users with positive tweets seem to have more followers than users with negative tweets.

The total amount of favourites a user has according to its candidate preference is represented in the second bar graph of Figure 3. It shows that the users tweeting about Clinton, both positively and negatively, have on average more favourites than users tweeting about Trump.

The third graph in Figure 3 represents the average amount of statuses a user has according to its candidate preference. The graph shows us that the positive-Trump and negative-Clinton group of users have a higher total amount of statuses compared to the negative-Trump and positive-Clinton tweeters. We could conclude that the pro-Trump group on average is more active on Twitter.



(a) Average amount of followers (b) Average amount of favourites (c) Average amount of statuses

Fig. 3: Tweets' users popularity

Topic Modeling

For the topic modeling analysis, we trained our model on multiple configurations. On the Table 2 and 3, the 10 topics with the most predominant 20 words(highest probability) are represented, using 50 and 100 passes respectively. We can see that by training the model and using 100 passes, which means more iterations, the word ensemble appears to be more coherent than when we are using 50 passes, something that is indicated by the corresponding coherence, which is 0.12 and 0.13, respectively. Although, the second model seems to perform slightly better, the overall performance failed to identify discrete topics, since the assigned words seem to not formulate a clear topic. This is indicated by the use of a data set with narrow margins, since the given data set was filtered based on specific tags related to the elections and the two candidates (see Methodology).

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
0	job	hillary	vote	clinton	president	trump	amp	supporter	america	racist
1	email	tweet	medium	campaign	country	tax	lie	pressure	hrc	truth
2	health	issue	speech	american	woman	nothing	liar	rain	anyone	obama
3	world	debate	support	money	voter	donald	life	crab	problem	anything
4	thank	family	hey	question	party	plan	someone	hpa	words	god
5	putin	idiot	war	state	person	ass	night	orchard	yes	fact
6	everyone	hell	everything	candidate	polls	wall	brain	forecast	comment	press
7	things	gop	child	policy	democrat	something	dems	election	crime	care
8	thanks	law	head	bill	ppl	return	interview	guy	mind	rally
9	record	leader	poll	foundation	wow	lol	chance	show	choice	shit
10	pay	business	days	house	video	potus	doctor	please	school	reason
11	friend	answer	fraud	immigration	corruption	work	place	love	death	hate
12	home	name	attack	pneumonia	word	isis	citizen	race	sign	guess
13	black	history	russia	mexico	wait	let	morning	joke	sound	watch
14	story	security	hope	office	bitch	yeah	vet	weather	control	kid
15	mouth	rest	power	idea	change	course	girl	basket	minority	need
16	general	case	respect	matter	men	bigot	shame	fine	event	deal
17	sorry	fbi	line	part	loser	class	scandal	statement	gun	republican
18	million	month	pres	nation	muslim	immigrant	cause	msm	presidency	others
19	team	con	book	message	twitter	sense	share	fuck	self	stop

Table 2: Topic Analysis results on the entire data set for 50 passes

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	clinton	trump	country	nothing	tax	amp	vote	speech	medium	tweet
1	president	liar	american	donald	racist	campaign	supporter	election	candidate	god
2	america	fact	job	money	email	truth	lie	party	world	family
3	woman	putin	plan	hrc	wall	health	hillary	hate	mexico	problem
4	obama	please	ass	someone	question	guy	pressure	wow	rally	democrat
5	state	polls	life	press	voter	issue	rain	days	child	head
6	policy	everything	lol	thank	anyone	bill	crab	love	leader	business
7	potus	shit	debate	care	anything	hey	hpa	wait	comment	black
8	foundation	video	idiot	everyone	something	person	orchard	home	joke	race
9	war	gop	immigration	things	return	thanks	forecast	russia	story	guess
10	ppl	muslim	reason	words	law	let	support	republican	bigot	night
11	hell	loser	house	work	idea	watch	show	interview	case	history
12	pay	piece	change	record	answer	security	yes	respect	sorry	pneumonia
13	corruption	class	mouth	name	kid	dems	friend	chance	girl	part
14	isis	crowd	immigrant	matter	need	twitter	poll	line	sense	men
15	office	crap	death	fraud	school	team	word	fbi	pres	hope
16	bitch	crook	doctor	yeah	msm	citizen	basket	place	yep	message
17	crime	voting	report	course	border	vet	weather	sound	boy	politician
18	deal	heart	stop	nation	racism	kkk	fine	book	benghazi	general
19	choice	drug	sign	statement	order	fool	mind	group	moron	folks

Table 3: Topic Analysis results on the entire data set for 100 passes

A better visualization of the topic modeling using the `pyLDAvis` library is shown on Figure 4, where the bubbles represent the topics, while their predominant words are on the bar chart on the right. Topic 3 was selected as an example.

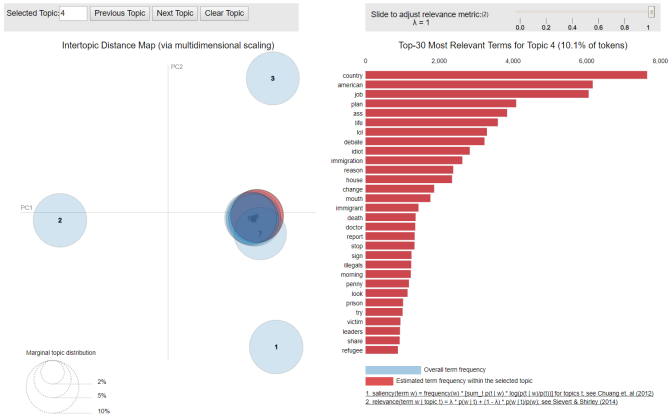


Fig. 4: Representation of Topic No.3 using the pyLDAvis library for LDA modeling

We further performed different LDA modeling analysis, by handling different data sets this time, one for each candidate. The filtering is based on the sentiment analysis we performed above. In this case, we set the number of passes on 100. Likewise, discrete topics cannot be identified, since the given data were filtered based on keywords related with the elections and the two candidates.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	speech	pressure	trump	american	vote	clinton	tax	lie	donald	amp
1	job	crab	president	obama	campaign	country	woman	nothing	plan	hrc
2	state	orchard	medium	god	america	supporter	return	election	money	mexico
3	voter	hpa	tweet	truth	wall	racist	life	something	ass	rally
4	thank	rain	words	candidate	anyone	family	leader	polls	liar	support
5	lol	forecast	care	immigration	guy	person	child	question	putin	please
6	black	show	issue	policy	anything	problem	email	hey	fact	loser
7	gop	fine	thanks	party	world	hate	press	potus	someone	history
8	ppl	weather	word	love	idiot	hell	days	democrat	everyone	home
9	fraud	guess	night	wow	things	pay	record	wait	shit	bill
10	community	fuck	health	law	business	name	kid	joke	everything	church
11	course	hope	mouth	head	work	foundation	comment	bigot	debate	place
12	interview	showery	twitter	isis	war	race	republican	school	reason	brain
13	case	baby	chance	immigrant	poll	corruption	story	need	friend	moron
14	minority	reality	folks	video	yes	yeah	penny	others	house	sign
15	security	self	attack	wife	change	matter	answer	surrogate	idea	side
16	usa	asshole	msm	muslim	let	russia	statement	racism	message	doctor
17	son	dream	report	nation	office	men	crap	sorry	part	basket
18	freedom	amen	sense	team	watch	respect	stop	book	general	try
19	call	forget	debt	border	deal	crowd	followers	play	line	presidency

Table 4: Topic Analysis results on Trump-related tweets (100 passes)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	hillary	president	state	woman	campaign	medium	amp	vote	obama	clinton
1	hrc	health	anything	liar	supporter	press	lie	truth	candidate	email
2	lol	country	world	america	life	question	nothing	job	person	money
3	corruption	american	please	bill	pneumonia	conference	voter	tax	someone	foundation
4	house	anyone	debate	election	video	war	speech	something	plan	work
5	pay	fact	tweet	issue	policy	basket	record	racist	hey	server
6	home	ass	head	family	bitch	comment	guy	care	support	fbi
7	government	law	shit	potus	thanks	answer	everything	party	democrat	benghazi
8	wait	days	word	child	office	need	reason	million	problem	name
9	part	thank	donor	hell	things	girl	doctor	wow	everyone	donation
10	cough	ppl	team	crime	friend	kkk	hate	guess	god	look
11	business	scandal	deal	race	watch	crook	yeah	show	security	fraud
12	jail	prison	stop	power	death	leader	statement	matter	story	body
13	yep	joke	sign	men	brain	deplorables	kid	words	yes	trust
14	choice	idea	phone	criminal	interview	mind	others	polls	rest	conspiracy
15	rally	course	black	husband	lady	piece	morning	dems	let	charity
16	heat	muslim	secretary	sorry	evidence	sound	fire	class	isis	chelsea
17	favor	hope	control	history	love	folks	russia	right	gop	powell
18	play	change	access	politician	case	reporter	voting	corrupt	idiot	dog
19	month	difference	plane	victim	report	proof	decision	republican	nation	bullshit

Table 5: Topic Analysis results on Clinton-related tweets (100 passes)

Conclusion

In this study, we performed sentiment analysis and topic modeling on tweets associated with the US elections of 2016. The given data set was extracted based on keywords and hashtags related to the two candidates, D.Trump and H. Clinton.

For the sentiment analysis, after the preprocessing of the 'text' feature, a lexicon-based approach with pre-scored words was followed and the polarity of each tweet (positive, negative, neutral) was determined. Tweets about D.Trump with all polarities were the most predominant. Moreover the number of tweets with Trump positive sentiment was bigger than tweets with Clinton positive sentiment. The computed ratio of favoured tweets about Trump was about 55%, comparable to the actual result that was 57%. We also examined user's popularity based on the number of followers, favourites and statuses, concluding that pro-Trump users group was more active on Twitter. The data indicated D. Trump was by far more popular from a topical perspective but due to the nosiness of the data and non-domain specific means of sentiment classification, there was no significant evidence to indicate a dominating sentiment. More refined means of classification such as better training using domain-specific data would improve results.

We further performed topic modeling with the LDA algorithm, using the *gensim* library in order to identify the topics. For the model execution, different parameters were used in order to have a more clear view of the most predominant topics, represented by the words with the highest probability. Even though there were a lot of very frequent terms related to the subjects of discrimination, religion, working conditions, economic policy, foreign policy, specific events during the campaign period (e.g Hillary Clinton email scandal), among others, the

LDA results failed to clearly categorize them for further analysis. The reason for this is mostly the fact that the tweets are already filtered based on tags and keywords, hence the margins for exploration for the algorithm are quite narrow, given that the tweets themselves are very small texts.

To conclude, from the analysis in this work we can determine that the vast availability of data on social media today, together with the advancements in Artificial Intelligence enables us to gain very fruitful insight on people's preferences towards different topics, policies and representatives among others. When utilized correctly, accurate opinion-mining can be a very powerful tool used towards a better society.

References

1. Alashri, S., Srivatsav Kandala, S., Bajaj, V., Ravi, R., Smith, K., Desouza, K.: An analysis of sentiments on facebook during the 2016 u.s. presidential election (08 2016)
2. Kharratzadeh, M., Üstebay, D.: US presidential election: What engaged people on facebook. In: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017. pp. 568–571 (2017), https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15582
3. Times, N.Y.: Presidential Election Results: Donald J. Trump Wins, https://www.nytimes.com/elections/results/president, accessed May 25, 2018