Topic Signatures in Political Campaign Speeches

Clément Gautrais[1], Peggy Cellier[2], René Quiniou[3], and Alexandre Termier[1]

[1]University of Rennes 1, IRISA, France
[2]INSA Rennes, IRISA, France
[3]Inria Rennes, IRISA, France

Abstract

Highlighting the recurrence of topics usage in candidates speeches is a key feature to identify the main ideas of each candidate during a political campaign. In this paper, we present a method combining standard topic modeling with *signature mining* for analyzing topic recurrence in speeches of Clinton and Trump during the 2016 American presidential campaign. The results show that the method extracts automatically the main ideas of each candidate and, in addition, provides information about the evolution of these topics during the campaign.

1 Introduction

Political discourse analysis (Van Dijk, 1998) is a branch of discourse analysis that aims at explicitting from speeches or debates the salient features of political discourses. From that point of view, a presidential election provides interesting datasets to study. Indeed, it is a major political event in a country and gives rise to many political meetings where candidates discuss personally selected societal problems and detail their own solutions. In that context, the identification of the favourite topics of candidates as well as how they evolve throughout the campaign is a crucial task.

In (Savoy, 2010), the author presents an analysis of the evolution of topics in political speeches by comparing the words that are overused or underused by Obama and McCain during the 2008 US presidential campaign. The dynamics of these particular words usage is analyzed over monthly periods to identify the underlying dynamics of the campaign topics. A limitation of this approach is that the period is fixed (monthly) whereas predictable (vote, debates) or unpredictable (scandals) events usually give the rhythm to a political campaign. Other work used topic modeling methods, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Latent Semantic Anlayis (LSA) (Landauer et al., 1998) or Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), to study political texts (Prabhakaran et al., 2014; Quinn et al., 2010). For instance in (Quinn et al., 2010), a topic model for legislative speech is defined. However, those works study topics one at a time whereas a set of co-referenced topics is more relevant since it constitutes the core of a candidate's political program. There exist other works about political text analysis (Calvet and Véronis, 2008) but they focus on the use of predefined single words in speeches over time, whereas we aim at finding (without knowing in advance which topic are recurrent) the recurrent topics (multiple topics) usage over time.

In this paper, we propose to identify in political speeches the favourite topics considered by each candidate as well as how and when they evolve throughout the campaign. In our opinion, this gives critical clues to identify and to explain each candidate's main ideas and their evolution. Thus, we describe an approach to extract the *topic signature* of a candidate from her/his political speeches, i.e. the set of topics discussed by some candidate over time. The method associates NMF, a standard topic modeling technique (Lee and Seung, 1999), with *signature mining* (Gautrais et al., 2017) to analyze the speeches of Hillary Clinton and Donald Trump during the 2016 US presidential campaign. The advantages of this approach are twofold. First, the set of campaign speeches is modeled with *topic signatures*, i.e. recurrent topics occurring at a flexible periodicity during the campaign, instead of sets of specific words occurring at a fixed periodicity. The topic signature provides a more abstract view of each candidate's

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2342–2347
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

main ideas and propositions. Second, the *signature mining* technique automatically adapts the periodicity to the campaign rhythms, to provide a better insight of the *campaign dynamics*.

2 Topic Signature Model

To model recurrent topics in a political campaign, we use the signature model (Gautrais et al., 2017). This model was originally developed to capture the recurrent purchase behavior of retail customers. The analogy between politics and retail is that customers' purchases consist of regularly bought products, and, similarly, politicians' speeches contain recurrent topics.

We consider a set of topics ($\mathcal{W}$) and a sequence of speeches (α) such that $\alpha = \langle(t_1, S_1),\ (t_2, S_2) \dots (t_n, S_n)\rangle$ where $\forall i \in [1, n]$, $S_i \subseteq \mathcal{W}$ and t_i gives the timestamp of S_i. For instance, in Figure 1, $\mathcal{W} = \{a, b, c, d, e\}$ and α is a sequence of seven speeches displayed in chronological order. We see that during Speech S3, two topics were addressed: *b* and *d*.

A k*-segmentation* of a sequence of speeches α, $P(\alpha, k) = \langle E_1 \dots E_k\rangle$, is a sequence of k non-overlapping consecutive sub-sequences of α, E_i, called *episodes*, each consisting of consecutive speeches. An example of a 4-segmentation is given in Figure 1, the first episode $E1$ contains 3 speeches ($S1$, $S2$, $S3$), $E2$ contains 2 speeches ($S4$, $S5$), $E3$ contains speech $S6$ and $E4$ contains speech $S7$. This segmentation contains episodes of different sizes, in both number of speeches and time span. This flexibility of the model allows for adapting the episodes size to the sequence rhythm.

A *topic* k*-signature*, $Rec(\alpha, k)$, is defined as a *maximal* set of recurrent topics in a k-segmentation of α. Roughly, given $P(\alpha, k) = \langle E_1 \dots E_k\rangle$ a k-segmentation of α, we have $Rec(P(\alpha, k)) = \bigcap_{E_i \in P(\alpha,k)}(\bigcup_{S_j \in E_i} S_j)$. In other words, $Rec(P(\alpha, k))$ contains the set of all recurrent topics that are present in each episode of $P(\alpha, k)$. $Rec(\alpha, k)$ is *maximal* means that it is obtained from a k-segmentation of α that maximizes the size of the recurrent topics set: $Rec(\alpha, k) = Rec(P_{max}(\alpha, k))$ with $P_{max}(\alpha, k) = \text{argmax}_{\{P(\alpha,k)\}} |Rec(P(\alpha, k))|$. k gives the number of recurrences of the topic signature in sequence α. Thus, given a number of recurrences k, finding the *topic* k*-signature* relies on finding the k-segmentation that maximizes the size of the topic set that appears in each episode of that segmentation. For example, in Figure 1, $\{a, b\}$ is a topic 4-signature, indeed $Rec(\alpha, 4) = E1 \cap E2 \cap E3 \cap E4$ $= (S_1 \cup S_2 \cup S_3) \cap (S_4 \cup S_5) \cap (S_6) \cap (S_7) = \{a, b, c, d\} \cap \{a, b\} \cap \{a, b\} \cap \{a, b, c, e\} = \{a, b\}$. There is no largest set of topics that is repeated in each episode of a 4-segmentation of α. As one can see in this example, episodes can be of different sizes, and speeches are grouped into episodes such that the topic signature is the largest.

The signature model contains two types of information. First, the intersection of all E_j contains the topics that are recurrent. In our case, this reveals the topics that one candidate has been speaking about, throughout the campaign speeches. The second information is temporal, through the episode timestamps. These timestamps reveal the rhythm of the topics usage. The signature actually links both information, to give the recurrent topics and their dynamic.

By varying the value of k, one can explore the main topics (if k is large) or the secondary topics, that are still recurrent (when k is low). Therefore, recurrent topics and their dynamics can be analyzed on different time scales. The difference with some previous approaches (Savoy, 2010) is that the size of each episode E_j is not defined in advance. Instead, the signature adapts the segmentation and episode size to reveal the rhythm of the topics usage.

3 Case Study: 2016 US Presidential Campaign

In this section, the topic signatures of Clinton and Trump during the 2016 US presidential campaign are analyzed.

3.1 Dataset

The dataset contains the transcripts of campaign speeches of both candidates Clinton and Trump, from April, 2015 to November, 2016. The speeches have been extracted from the American Presidency Project (APP)[1]. This yielded a total of 164 speeches: 93 for Clinton and 71 for Trump[2].

3.2 Preprocessing

The dataset was preprocessed as follows. First, the sentences that did not correspond to a candi-

[1] http://www.presidency.ucsb.edu/2016_election.php

[2] Including the 3 presidential debates. Speeches of Clinton prior to April 2015 were discarded

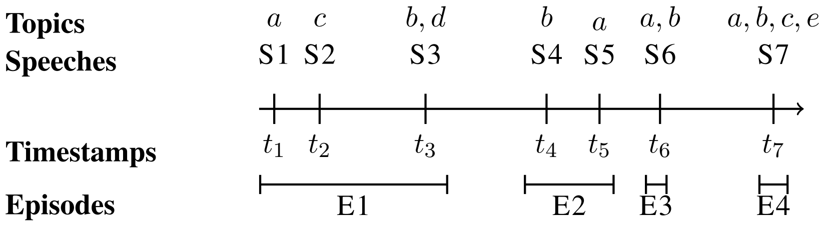

Figure 1: A speech sequence and a 4-segmentation. The recurrent topics are $\{a, b\}$.

date utterance (journalists questions, introduction by another speaker ...) were removed. Next, the sentences were tokenized and the tokens associated with some Part-of-Speech (POS) tags were kept. Precisely, nouns, adjectives and foreign words were kept while verbs and personal nouns were removed. While removing verbs can lead to a loss of semantic information, we found that it resulted in more interpretable topics. This choice of removing verbs has previously been made for topic modeling in political texts (Zirn and Stuckenschmidt, 2014). Personal nouns were discarded to remove all references to interviewers or other politicians. We considered keeping some proper nouns (the ones of both campaigners and of some other politicians) but it added noise in the topic modeling step, without providing additional relevant information. Finally, remaining tokens were lemmatized and stop words were removed. We used the WordNet lemmatizer (Miller and Fellbaum, 1998) and the list of stop words from the nltk library[3] (Bird et al., 2009). The final dataset contained 6240 different lemma.

3.3 Topic Modeling

Even though words could be analyzed directly (Savoy, 2010), we decided to analyze topics. This choice is mainly guided by the fact that we are looking for recurrent topics, and working directly on words gave uninteresting results, as recurrent words are not directly representative of each candidate ideas. Different topic modeling techniques were tested (Stevens et al., 2012) (LDA (Blei et al., 2003) and NMF (Lee and Seung, 1999)) with different parameters, number of topics and settings (with or without verbs for example). As a result, we concluded that using NMF on count vectors with 15 topics produced the most meaningful, diverse, yet non redundant topics. Some of these topics and their top lemma are shown in Table 1.

Table 1: Some topics found by NMF, and their main lemma.

Topic name	**Main topic lemma**
Economic policy	economy, growth, new, business, income, wage
Woman president and voters	woman, election, president, future, young
Illegal immigration	immigration, illegal, law, border, criminal, visa
Climate change	energy, climate, change, clean, future, important

However, it should be noted that other topic modeling techniques ((Greene and Cross, 2017) for example) could be used, and lead to meaningful results. Indeed, as our method is built on top of topics, any technique that provides good enough topics can be used. Any improvements in the topic model can help to draw more precise conclusions (if cleaner topics are available). This remark is also true regarding our choice of removing verbs and personal nouns.

Within NMF, a speech is represented as a numeric weight vector across all topics. However, the signature model works on symbolic data, which means that a set of representative topics for each speech has to be selected. As we want to discriminate the main topics of a speech from the remaining ones, we applied a clustering on the weight vectors of each speech. Two clusters were looked for, the first containing the highest weights i.e. the cluster of the main topics, and the second containing the secondary or absent topics, with lowest weights. We used the spectral clustering technique (Shi and Malik, 2000) from the scikit-learn library[4] (Pedregosa et al., 2011). We did not used techniques based on the Euclidean distance (such as k-means (MacQueen et al., 1967)) as it is not suited to separate main topics from minor ones. Three main topics emerged per speech on

[3] http://www.nltk.org/

[4] http://scikit-learn.org

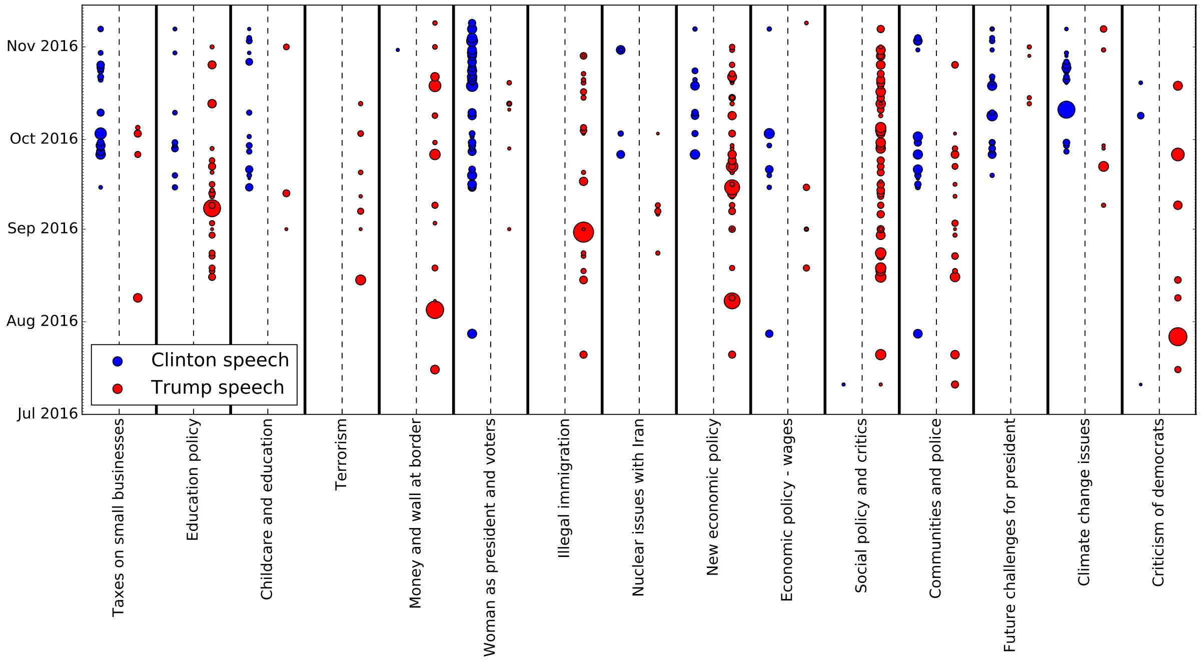

Figure 2: Campaign topics through time for each candidate. Each circle represents the presence of a topic in a speech. The larger the topic, the more present in a speech. Trump speeches are depicted in red, Clinton speeches in blue.

average.

3.4 Topic Signature Extraction

To study the main topics on different time scales, we computed signatures with different values of k. Table 2 displays the results.

3.5 Discussion

About Extracted Topics Figure 2 displays a visualization of all main topics. Only the last months of the campaign are plotted, since both candidates were particularly active during that period and speeches were sparse earlier. The visualization is especially suited to analyze single topics. First, we can see that most topics are discriminative, they appear often in one candidate's speech while being almost absent in the other's. Some topics, like *Communities and police*, are shared but not used on the same time line. Another example is the use of the *Climate change issues* topic. We can see that it is mainly used at the end of the presidential campaign by Clinton.[5]

About Topic Signatures While the previous section shows how individual topics can be analyzed, the signature allows for analyzing the main topics as a whole. Let us look at each candidate's recurring topics in Table 2. The main topics of each candidate are well separated, showing that each candidate has its own targeted voters. Clinton focused on topics related to communities, youth, issues for the next generations, and woman as president. Trump focused on topics such as new economical policies, illegal immigration, new social policies and criticism of the former government.

Table 2: Signature topics in speeches of Clinton (top) and Trump (bottom), for some values of k.

Clinton

No	Recurrences (k)	Signature topics
C1	57	Woman as President
C2	30	C1 + Future challenges for President
C3	16	C2 + Communities and police
C4	12	C3 + Childcare and education

Trump

No	Recurrences (k)	Signature topics
T1	48	Social policy and critics
T2	28	T1 + New economic policy
T3.1	15	T2 + Illegal immigration
T3.2	15	T2 + Education policy
T4.1	9	T3.2 + Illegal immigration (T3.1 + T3.2)
T4.2	9	T3.2 + Money and wall at border

The signature of Clinton is quite simple, as low-

[5] *Climate change issues* became a topic of interest when Clinton attacked Trump on him saying that climate change is a hoax in the first presidential debate (September 26, 2016).

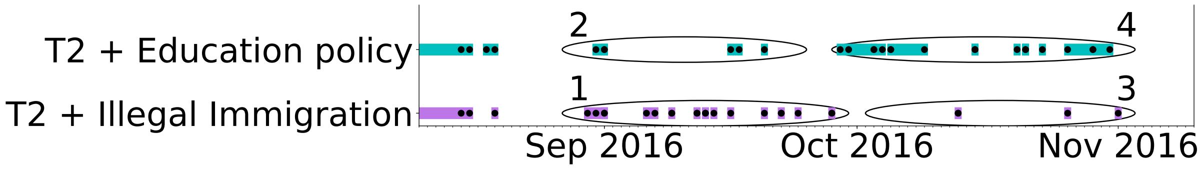

Figure 3: Episodes of two Trump signatures. T3.1: Social policy and critics + New economic policy + *Illegal immigration* ; T3.2: Social policy and critics + New economic policy + *Education policy*. Every rectangle in pink or blue represents an episode, and each black dot represents a speech belonging to an episode. Each numbered ellipse represents a group (annotated by hand) of episodes.

ering the minimal number of occurrences only adds new topics to the signature. This means that Clinton is very stable in her main topics. This observation is also partially true for Trump. Indeed, Trump has sometimes different signatures for a given number of occurrences. For example, with $k = 15$, Trump speeches main topics can include *Illegal immigration* or *Education policy*, but not both together. This is interesting because it shows that Trump is more diverse in his recurrent topics and that some of them rarely occur together.

To further deepen the analysis of the fact that Trump speeches include either *Education policy* or *Illegal Immigration* but rarely both, let us look at the episodes of the related signatures, represented in Figure 3. First, we note that the difference between both signatures episodes began to be apparent by September 2016. Indeed, the signature containing *Illegal immigration* only has three episodes (Group 2), whereas the one with *Education policy* has 11 episodes (Group 1). This large difference shows that, in September, Trump discussed his main topics a lot (*Criticism of former government*, *New social policies* and *New economic policy*) in association with *Education policies*. In October 2016, he switched to *Illegal immigration* while keeping his main topics, as there are 3 episodes for *Education policy* (Group 3) whereas there are 7 episodes for *Illegal immigration* (Group 4). While the fact that Trump stopped talking about *Education policy* at the end of September 2016 is visible in Figure 2, the segmentation performed by the signature brings additional information. Indeed, the signature is changing only one of its topics, so we know that Trump kept talking about his other main topics (*Social policy and critics* and *New economic policy*) while switching from *Education policy* to *Illegal immigration*.

Another important point is that by the beginning of October, when Trump switched from *Education policy* to *Illegal immigration*, the episodes are longer than the remaining ones (Group 4). This means that Trump's main topics are distributed among more speeches than before, which can reflect a change in his strategy. This information is not easily visible in Figure 2, but it is available from a simple analysis of Trump signature.

This case study, based on topic signatures, shows that our method is able to derive each candidate's recurrent topics. Analyzing episodes and related signature topics enables to spot changes in Trump speeches and to explain how some of his recurrent topics are related to each other. This kind of precise analysis is beyond the capabilities of naive regular segmentation techniques.

4 Conclusion

We have presented a new method for analyzing political discourse. It associates standard topic modeling with signature mining and enables the identification of the main topics of politicians during a campaign as well as their dynamics. The 2016 US presidential campaign analysis provides interesting results: though the discourse of H. Clinton was relatively stable, important changes could be identified in the discourse and communication strategy of D. Trump. These specific results on the campaign dynamics were obtained thanks to the temporal flexibility of the model.

In the future, we would like to apply the method to more challenging data, such as political tweets. Preliminary results on the 2016 US campaign tweets show that the topics used by both candidates were different from their speech: the tweets, being shorter than speeches, emphasize more oversimplified criticism of the opponent rather than justified political ideas.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Louis-Jean Calvet and Jean Véronis. 2008. *Les mots de Nicolas Sarkozy*. Seuil París.

Clement Gautrais, René Quiniou, Peggy Cellier, Thomas Guyet, and Alexandre Termier. 2017. Purchase signatures of retail customers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pages 110–121.

Derek Greene and James P. Cross. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis* 25(1):7794. https://doi.org/10.1017/pan.2016.7.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA., volume 1, pages 281–297.

George Miller and Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pages 1481–1486.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1):209–228.

Jacques Savoy. 2010. Lexical analysis of us political speeches. *Journal of Quantitative Linguistics* 17(2):123–141.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8):888–905.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 952–961.

Teun Van Dijk. 1998. What is Political Discourse Analysis? *Belgian Journal of Linguistics* 11:11–52. https://doi.org/10.1075/bjl.11.03dij.

Cäcilia Zirn and Heiner Stuckenschmidt. 2014. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering* 90:38–53.