Quantitative Narrative Analysis of US Elections in International News Media

Saatviga Sudhahar, Thomas Lansdall-Welfare, Ilias Flaounas, Nello Cristianini
Intelligent Systems Laboratory
University of Bristol

Abstract

We present a new computational methodology for large scale narrative analysis of news content, with an application to political discourse concerning US presidential elections. News articles concerning the elections are recognised, parsed, and used to generate a list of political actors and a network representing their political relations. That network is in turn analysed to extract information about the role that each actors play in the political narrative. The method is entirely automated and very scalable, and its results can be accessed via the project website. So far our system has analysed 125254 articles in 719 US and international news outlets extracting 31476 actors.

Introduction

Discourse about US Presidential Elections dominates the global news system every four years. Its contents play a major role in shaping voters opinion, and are therefore carefully analysed by commentators, as well as being subjected to sophisticated manipulations by campaign strategists. Candidates are expected to take clear positions about a variety of issues, and many social actors are expected to take sides in this important choice, either endorsing or opposing a candidate.

Over the 10 months separating the first caucus in Iowa from the day of the general election, a vast network of actors is formed, and transformed, in the media arena, while in the primary phase multiple candidates compete to earn the role of candidate, and in the final stage two camps compete for the favours of public opinion.

The amount of news articles devoted to this topic is so large that no exhaustive analysis can be attempted by conventional means. Even if just focusing on the leading English-language outlets, there are hundreds of thousands of articles to analyse just for the primary phase. So any large scale analysis of global coverage will necessarily need to make use of computational methods.

However, most computational approaches to news content analysis are limited to sophisticated forms of keyword counting, be it for sentiment analysis, or topic detection, and relative statistical analysis. This will necessarily miss many aspects of the narration to which voters are exposed, and which may therefore be of interest to analysts.

We are interested in accessing information that is closer to what a human analyst could extract, but still simple enough that can be reliably extracted by computational means in a Big Data setting. In this project, we automated techniques from *Quantitative Narrative Analysis (QNA)* so that they can be applied on a vast scale. This approach is aimed at identifying the actors and the actions that dominate a story, as well as basic units of narration: *Subject-Verb-Object (SVO)* triplets. While still very simple, this information captures a variety of relations that would be missed by classical means, and that are relevant to political discourse.

One of the result is a network whose nodes are actors (represented by noun phrases such as "*the democratic party*") and the edges are actions (represented by transitive verbs such as "*endorsed*"). The domain of US politics is particularly amenable to this type of network analysis, due to the binary nature of the choice, so that all various issues and players need to ultimately fit into a bi-polar playing field. Also the communication is easily analysed, with explicit support or opposition often being stated for the candidates by various actors.

It is therefore possible to automatically detect the relation between these actors, generating a relational network whose topology depends on the political relations between these players. An analysis of the properties of this network can reveal a lot of information about the political landscape, as represented in the news narrative. Another key result of this type of analysis is that we can also identify which actors are more often portrayed as subjects or objects of political discourse, and which of them are more likely to be the subject or the object of positive / negative statements.

Methodology

Data Collection

Our system collects news articles from 719 English language news outlets. We monitor both U.S and International media. A detailed description of the underlying infrastructure has been presented in our previous work (Flaounas, 2011).

In this demo we use only articles related to US Elections. We detect those articles using a topic detector based on Support Vector Machines (Chang, 2011). We trained and validated our classifier using the specialised Election news feed from Yahoo!. The performance of the classifier reached 83.46% precision, 73.29% recall, validated on unseen articles.

While the main focus of the paper is to present Narrative patterns in elections stories, the system presents also timelines and activity maps generated by detected Named Entities associated with the election process.

Text Analysis

We perform a series of methodologies for narrative analysis. Figure 1 illustrates the main components that are used to analyse news and create the website.

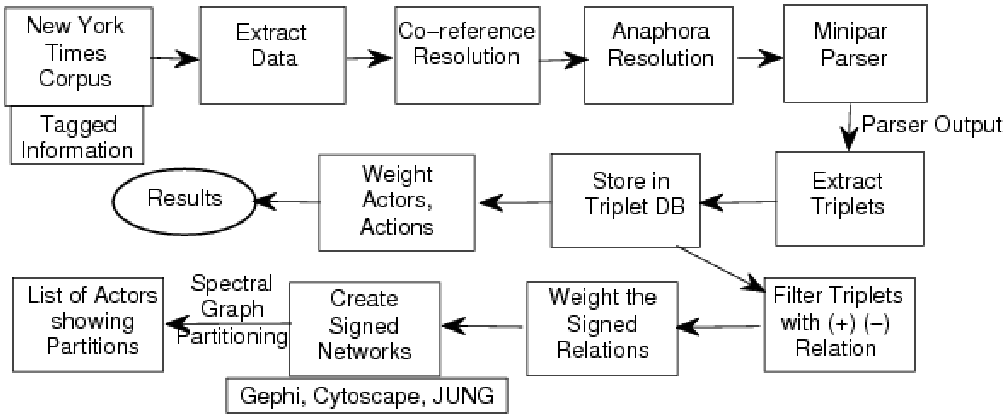


Fig. 1: System Pipeline

Pre-processing. First, we perform co-reference and anaphora resolution on each U.S Election article. This is based on the ANNIE plugin in GATE (Cunningham, 2002). Next, we extract Subject-Verb-Object (SVO) triplets using the Minipar parser output (Lin, 1998). An extracted triplet is denoted for example like, “Obama(S) – Accuse(V) – Republicans(O)”.

We found that news media contains less than 5% of passive sentences and therefore it is ignored. We store each triplet in a database annotated with a reference to the article from which it was extracted. This allows us to track the background information of each triplet in the database.

Key Actors. From triplets extracted, we make a list of actors which are defined as subjects and objects of triplets. We rank actors according to their frequencies and consider the top 50 subjects and objects as the key actors.

Polarity of Actions. The verb element in triplets is defined as actions. We map actions to two specific action types which are endorsement and opposing. We obtained the endorsement/opposing polarity of verbs using Verbnet data (Kipper et al, 2006).

Extraction of Relations. We retain all triplets that have a) the key actors as subjects or objects; and b) an endorse/oppose verb. To extract relations we introduced a weighting scheme. Each endorsement-relation between actors *a*, *b* is weighted by $w_{a,b}$*:*

$$w_{a,b} = \frac{f_{a,b}(+) - f_{a,b}(-)}{f_{a,b}(+) + f_{a,b}(-)} \qquad (1)$$

$f_{a,b}(+)$ and $f_{a,b}(-)$ denote the number of triplets between *a*, *b* that support a positive and negative relation. This way, actors who had equal number of positive and negative relations are eliminated.

Endorsement Network. We generate a triplet network with the weighted relations where actors are the nodes and weights calculated by Eq. 1 are the links. This network reveals endorse/oppose relations between key actors.

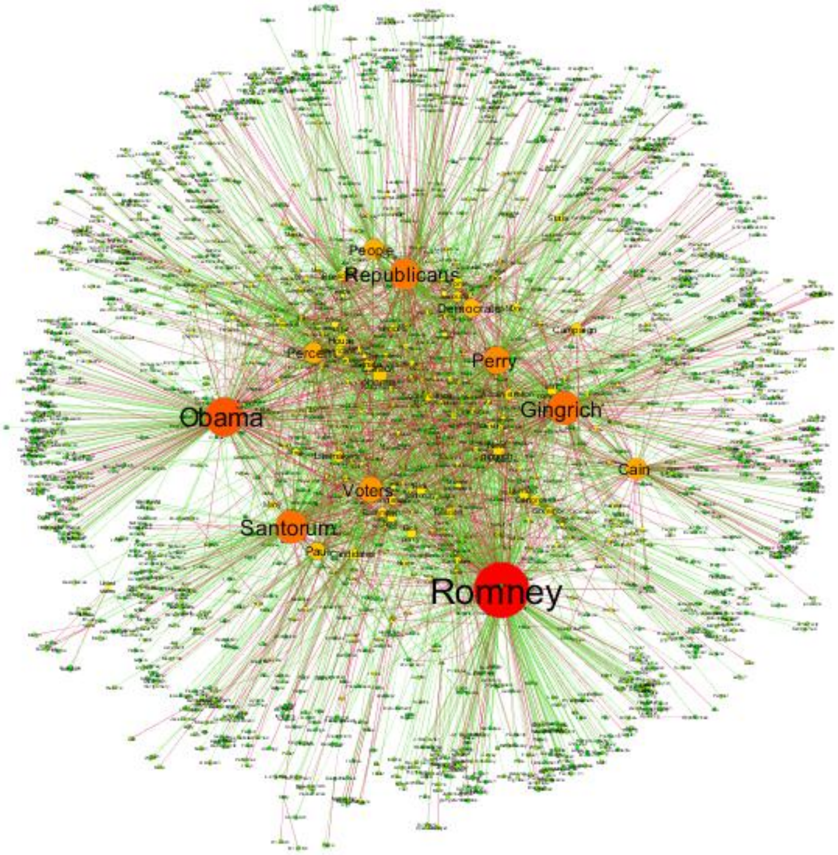


Fig. 2: Endorse/Oppose network of actors in 2012 U.S. Elections

The network in Figure 2, is a typical example of such a network. Node size reveals the frequency of mentions of an actor and the red and green links imply positive and negative relations between actors.

Network Partitioning. By using graph partitioning methods we can analyse the allegiance of actors to a party, and therefore their role in the political discourse. The Endorsement Network is a directed graph. To perform its partitioning we first omit directionality by calculating graph $B = A+A^T$, where A is the adjacency matrix of the Endorsement Network. We computed eigenvectors of B and selected the eigenvector that correspond to the highest eigenvalue. The elements of the eigenvector represent actors. We sort them by their magnitude and we obtain a sorted list of actors.

In the website we display only actors that are very polarised politically in the sides of the list. These two sets of actors correlate well with the left-right political ordering in our experiments on past U.S Elections.

Since in the first phase of the campaign there are more than two sides, we added a scatter plot using the first two eigenvectors. Details of this follow in the next section.

Subject/Object Bias of Actors. The Subject/Object bias S_a of actor a reveals the role it plays in the news narrative. It is computed as:

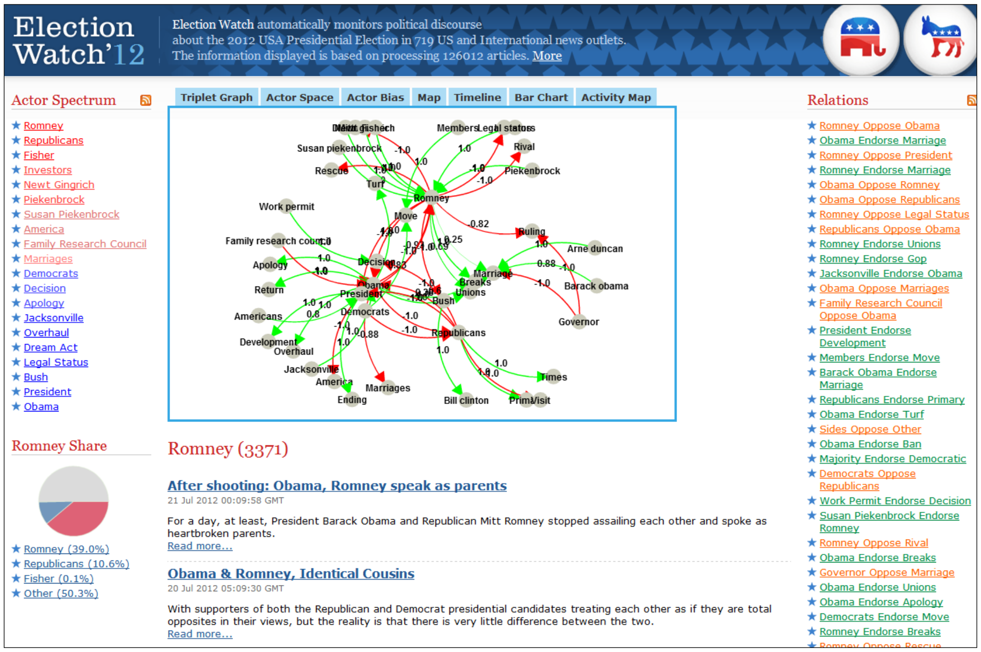
$$S_a = \frac{f_{subj}(a) - f_{obj}(a)}{f_{subj}(a) + f_{obj}(a)} \qquad (2)$$

A positive value of S_a for actor a, indicates that the actor is mentioned more often as a subject and a negative value indicates that the actor is more often an object.

Website

We analyse news related to U.S Elections 2012 every day, automatically, and the results of our analysis are presented integrated under a publicly available website[1]. Fig. 3 illustrates the home-page of ElectionWatch. Here, we list the key features of the site:

[1] ElectionWatch: http://electionwatch.enm.bris.ac.uk



Election Watch automatically monitors political discourse about the 2012 USA Presidential Election in 719 US and International news outlets. The information displayed is based on processing 126012 articles. More

Actor Spectrum

- ★ Romney
- ★ Republicans
- ★ Fisher
- ★ Investors
- ★ Newt Gingrich
- ★ Piekenbrock
- ★ Susan Piekenbrock
- ★ America
- ★ Family Research Council
- ★ Marriages
- ★ Democrats
- ★ Decision
- ★ Apology
- ★ Jacksonville
- ★ Overhaul
- ★ Dream Act
- ★ Legal Status
- ★ Bush
- ★ President
- ★ Obama

Romney Share

- ★ Romney (39.0%)
- ★ Republicans (10.6%)
- ★ Fisher (0.1%)
- ★ Other (50.3%)

Romney (3371)

After shooting: Obama, Romney speak as parents
21 Jul 2012 00:09:58 GMT

For a day, at least, President Barack Obama and Republican Mitt Romney stopped assailing each other and spoke as heartbroken parents.
Read more...

Obama & Romney, Identical Cousins
20 Jul 2012 05:09:30 GMT

With supporters of both the Republican and Democrat presidential candidates treating each other as if they are total opposites in their views, but the reality is that there is very little difference between the two.
Read more...

Relations

- ★ Romney Oppose Obama
- ★ Obama Endorse Marriage
- ★ Romney Oppose President
- ★ Romney Endorse Marriage
- ★ Obama Oppose Romney
- ★ Obama Oppose Republicans
- ★ Romney Oppose Legal Status
- ★ Republicans Oppose Obama
- ★ Romney Endorse Unions
- ★ Romney Endorse Gop
- ★ Jacksonville Endorse Obama
- ★ Obama Oppose Marriages
- ★ Family Research Council Oppose Obama
- ★ President Endorse Development
- ★ Members Endorse Move
- ★ Barack Obama Endorse Marriage
- ★ Republicans Endorse Primary
- ★ Obama Endorse Turf
- ★ Sides Oppose Other
- ★ Obama Endorse Ban
- ★ Majority Endorse Democratic
- ★ Democrats Oppose Republicans
- ★ Work Permit Endorse Decision
- ★ Susan Piekenbrock Endorse Romney
- ★ Romney Oppose Rival
- ★ Obama Endorse Breaks
- ★ Governor Oppose Marriage
- ★ Obama Endorse Unions
- ★ Obama Endorse Apology
- ★ Democrats Endorse Move
- ★ Romney Endorse Breaks
- ★ Romney Oppose Rescue

Fig. 3: Screenshot of ElectionWatch website

Triplet Graph. The main network in Figure 3 is created using the weighted relations. A positive sign for the edge indicates an endorsement relation and a negative sign indicates an opposition relation in the network. By clicking on each edge in the network, we display triplets and articles that support the relation.

Actor Spectrum. The left side of Figure 3 shows the Actor Spectrum, coloured from blue for Democrats to red for Republicans. Actor spectrum was obtained by applying spectral graph partitioning methods to the triplet network. Note, that currently there are more than two campaigns that run in parallel between key actors that dominate the elections news coverage. Nevertheless, we still find that the two main opposing candidates in each party were in either sides of the list.

Relations. On the right hand side of the website we show the endorsement/opposition relations between key actors. For example, "*Republicans Oppose Democrats*". When clicking on a relation the webpage displays the news articles that support the relation.

Actor Space. The tab labelled *"Actor Space"* plots the first and second eigenvector values for all actors in the actor spectrum.

Actor Bias. The tab labelled *"Actor Bias"* plots the subject/object bias of actors against the first eigenvector in a two dimensional space.

Pie Chart. Pie Chart on the left bottom in the webpage shows the share of each actor with regard to the total number of articles mentioning an endorse/oppose relation.

Map. The map geo-locates articles related to US Elections and refer to US locations.

Bar Chart. The bar chart tab, illustrated in Fig. 3, plots the number of articles in which actors were involved in a endorse/oppose relation. The height of each column reveals the frequency of it. The default plot focuses on only the first five actors in the actor spectrum.

Timelines & Activity Map. We track the activity of each named entity in the actor spectrum within the United States and present it in a timeline. The activity map monitors the media attention for Presidential candidates in each state in the Unites States. At present we monitor this activity for "*Mitt Romney"*, "*Rick Perry"*, "*Michele Bachmann"*, "*Herman Cain*" and "*Barack Obama"*.

Experimental Results

As experimental results we present here both experiments on the past 5 election cycles, and up-to date analysis of the 2012 election. The first set will only be based on the New York Times coverage, while the analysis of the 2012 election will be based on more than 719 international outlets, having generated so far more than 125254 articles. So far our system has extracted 408735 triplets which contain 31476 distinct actors. We will concentrate on two classes of results: the properties of the network of political support among actors, which reveals complex party allegiances, and the embedding of actors in a space that reveals their position in the media narrative.

Figure 4 shows a network with positive and negative edges between actors obtained from 2012 elections data from May onwards after filtering for high-confidence relations. The list of actors on the right reveals the party allegiance of actors in an actor spectrum showing Democrats and Republicans on either sides. The thickness of edges corresponds to the frequency of the relation between actors. We analysed the past elections using the same methods during the evolution of the election process.

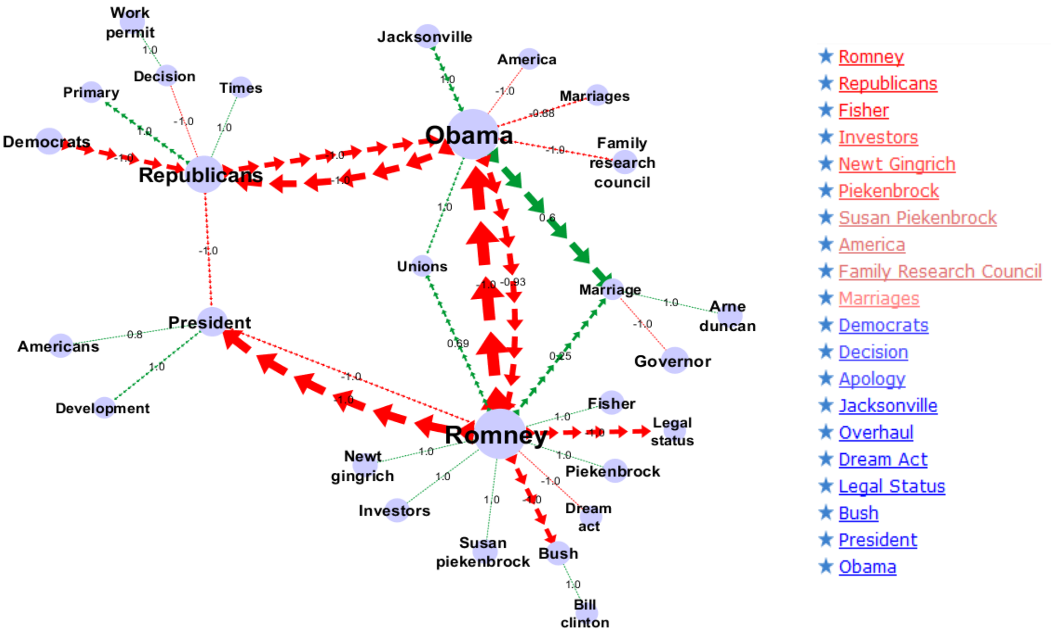


Fig. 4: Network with Positive and Negative edges between Actors - 2012 U.S. Elections (May onwards)

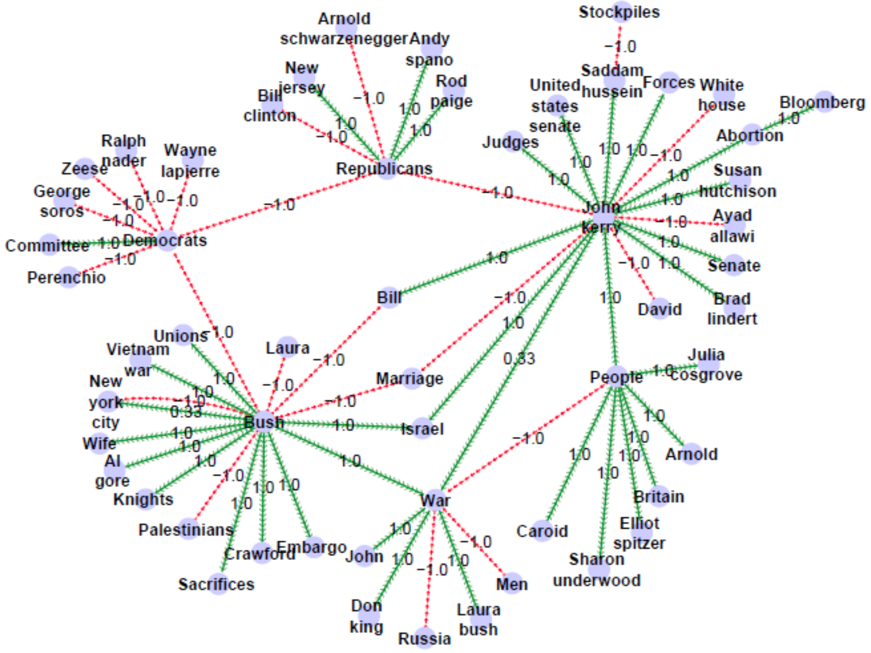


Fig. 5a: Network with Positive and Negative edges between Actors – 2004 U.S. Elections (August to November)

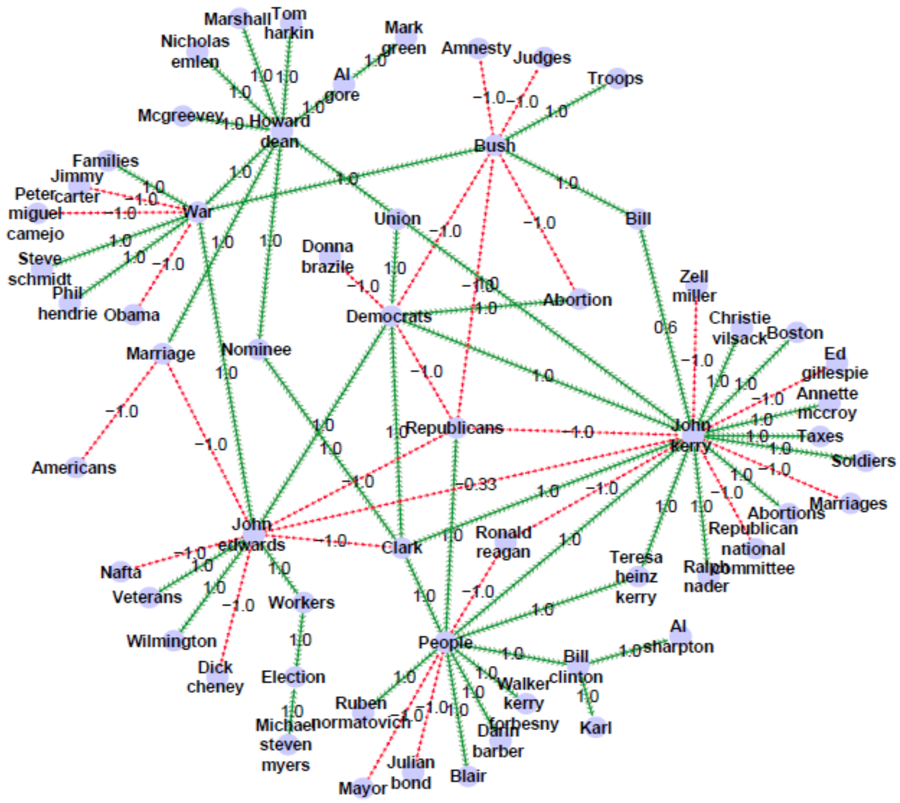
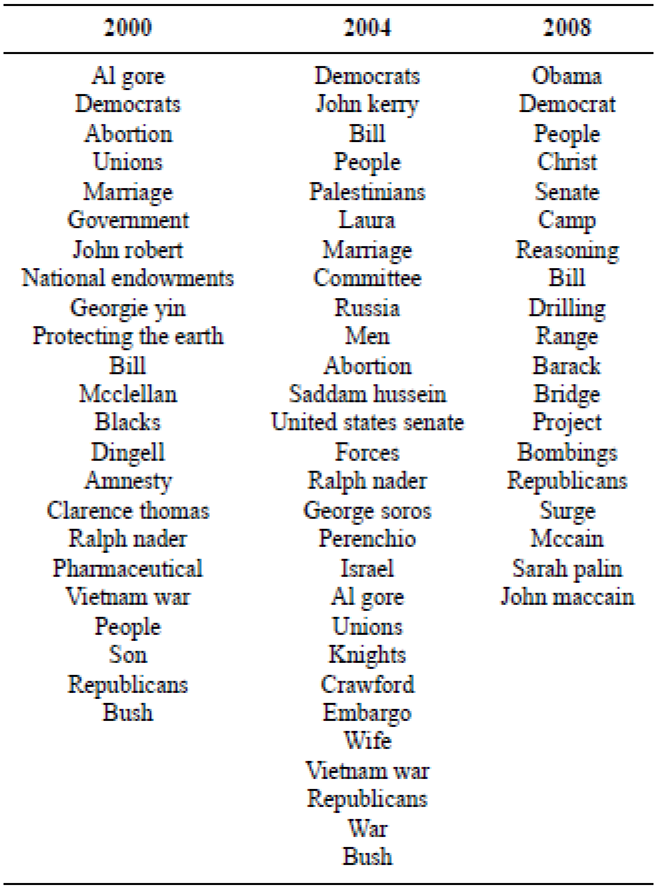


Fig. 5b: Network with Positive and Negative edges between Actors – 2004 U.S. Elections (January to August)

Figure 5a and 5b are examples of an endorse/oppose network obtained from the 2004 U.S. Elections data during the primary and main elections. We observed that in each year there were many hubs representing candidates campaigning in different states in the network for the period of January - August while there were only two main hubs during August-November showing the two main opposing candidates from the Republicans and Democrats.

Figure 6 shows a short version of the actor spectrum obtained from 2000, 2004 and 2008 U.S. Elections Data after removing the actors in the middle of the list. Here we could see the two main opposing candidates in either side of the list representing the Democrats and Republicans.

It is also interesting to see concepts like "*Abortion*" and "*War*" take sides meaning "*Abortion*" being mostly associated with the Democrats and "*War*" with the Republicans.



2000	2004	2008
Al gore	Democrats	Obama
Democrats	John kerry	Democrat
Abortion	Bill	People
Unions	People	Christ
Marriage	Palestinians	Senate
Government	Laura	Camp
John robert	Marriage	Reasoning
National endowments	Committee	Bill
Georgie yin	Russia	Drilling
Protecting the earth	Men	Range
Bill	Abortion	Barack
Mcclellan	Saddam hussein	Bridge
Blacks	United states senate	Project
Dingell	Forces	Bombings
Amnesty	Ralph nader	Republicans
Clarence thomas	George soros	Surge
Ralph nader	Perenchio	Mccain
Pharmaceutical	Israel	Sarah palin
Vietnam war	Al gore	John maccain
People	Unions	
Son	Knights	
Republicans	Crawford	
Bush	Embargo	
	Wife	
	Vietnam war	
	Republicans	
	War	
	Bush	

Fig. 6: Short version of actor spectrums obtained from 2000, 2004 and 2008 U.S. Elections Data

Conclusions

We have demonstrated the system ElectionWatch that presents key actors in U.S election news articles and their role in political discourse. This builds on various recent contributions from the field of Pattern Analysis, such as (Trampus, 2011), augmenting them with multiple analysis tools that respond to the needs of social sciences investigations.

The computational infrastructure is capable of detecting election-related articles, parsing their content, solving co-reference and anaphora, identifying verbs that denote support or opposition, identifying key actors, filtering information that is statistically not reliable, and finally analysing the properties of the resulting relational network. While each step of the extraction phase may be imperfect, the statistical corrections coming from the use of very big datasets deliver a sufficiently clean signal for political observers to monitor the state of play of a complex process such as a US Presidential campaign.

We have tested this system on data from all previous six elections, using the New York Times corpus as well as our own database. We use only support/criticism relations revealing a strong polarisation among actors and this seems to correspond to the left/right political dimension. Results on the past six election cycles on New York Times always separated the two competing candidates along the eigenvector spectrum.

For evaluation we focus on the "*high precision/low recall*" setting that is on the scenario where we rather leave out a triplet, but want to make sure that whatever we extract is correct. To assess the quality of the overall set of triplets generated by our system we have performed two separate experiments, involving comparisons with human annotation. In the first comparison we compared our system with human annotation on a set of documents which were annotated for a separate project (De Fazio, 2012). We measured the probability of a triplet seen only once to be correct, and that has been found to be 62%. This means that there is a probability of 38% of a triplet being wrong if it has been seen once, and 0.38^n when it has been seen n times. By only accepting triplets which have been seen 3 times, we have a 5% probability of error. In the second comparison, we examined by hand a set of 74 triplets that were extracted by our system, which were considered of sufficient quality to be used after applying a high threshold value of 7. These triplets were manually verified by checking the article of origin, to see if these triplets were indeed present in the text. This yielded 96% precision while recall was not evaluated.

Future work will include making better use of the information coming from the parser, which goes well beyond the simple *SVO* structure of sentences, and developing more sophisticated methods for the analysis of large and complex networks that can be inferred with the methodology we have developed.

Acknowledgments I. Flaounas and N. Cristianini are supported by FP7 CompLACS. The members of the Intelligent Systems Laboratory are supported by the 'Pascal2' Network of Excel-lence. Authors would like to thank Omar Ali and Roberto Franzosi.

References

Chang C.C., and Lin C.J. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3):1–27

Cunningham H., Maynard D., Bontcheva K. And Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics 168–175.

De Fazio G (2012) Political radicalization in the making: The civil rights movement in northern ireland,1968-1972. PhD thesis, Department of Sociology, Emory University, Atlanta, Georgia

Earl J., Martin A., McCarthy J.D., Soule S.A. 2004. The Use of Newspaper Data in the Study of Collective Action. Annual Review of Sociology, 30:65–80.

Flaounas I., Ali O., Turchi M., Snowsill T., Nicart F., De Bie T., Cristianini N. 2011. NOAM:News Outlets Analysis and Monitoring system. Proc. of the 2011 ACM SIGMOD international conference on Management of data, 1275–1278.

Franzosi R. 2010. Quantitative Narrative Analysis. Sage Publications Inc, Quantitative Applications in the Social Sciences, 162–200.

Kipper K., Korhonen A., Ryant N., Palmer M. 2006. Extensive Classifications of English verbs. 12th EURALEX International Congress, Turin, Italy.

Lin D. 1998. Dependency-Based Evaluation of Minipar. Text, Speech and Language Technology 20:317–329.

Saatviga Sudhahar, Roberto Franzosi, Nello Cristianini. Automating Quantitative Narrative Analysis of News Data. Proc. of the Journal of Machine Learning Research (JMLR) Vol 17:63-71 in conjunction with the Second Workshop on Applications of Pattern Analysis (WAPA), 19-21 October, 2011, CIEM, Castro Urdiales, Spain.

Saatviga Sudhahar, Thomas Lansdall-Welfare, Ilias Flaounas and Nello Cristianini. ElectionWatch: Detecting Patterns in News Coverage of US Elections. Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 23-27 April, 2012, Avignon, France.

Sandhaus, E. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium

Trampus M., Mladenic D. 2011. Learning Event Patterns from Text. Informatica 35