Exploiting Linked Open Data and Natural Language Processing for Classification of Political Speech

Exploiting Linked Open Data and Natural Language Processing for Classification of Political Speech

Giuseppe Futia*, Federico Cairo**, Federico Morando**+, Luca Leschiutta****

*Nexa Center for Internet & Society, DAUIN - Politecnico di Torino, giuseppe.futia@polito.it
**Nexa Center for Internet & Society, DAUIN - Politecnico di Torino, federico.cairo@polito.it
***Nexa Center for Internet & Society, DAUIN - Politecnico di Torino, federico.morando@polito.it
****Nexa Center for Internet & Society, DAUIN - Politecnico di Torino, luca.leschiutta@polito.it

Abstract: *This paper shows the effectiveness of a DBpedia-based approach for text categorization in the e-government field. Our use case is the analysis of all the speech transcripts of current White House members. This task is performed by means of TellMeFirst, an open-source software that leverages the DBpedia knowledge base and the English Wikipedia linguistic corpus for topic extraction. Analysis results allow to identify the main political trends addressed by the White House, increasing the citizens' awareness to issues discussed by politicians. Unlike methods based on string recognition, TellMeFirst semantically classifies documents through DBpedia URIs, gathering all the words that belong to a similar area of meaning (such as synonyms, hypernyms and hyponyms of a lemma) under the same unambiguous concept.*

Keywords: *DBpedia, Natural Language Processing, e-democracy, Text Categorization, White House Speeches*

Acknowledgement: *This paper was drafted in the context of the Network of Excellence in Internet Science EINS (GA n°288021), and, in particular, in relation with the activities concerning Evidence and Experimentation (JRA3). The authors acknowledge the support of the European Commission and are grateful to the network members for their support.*

Introduction

Text has been described as "arguably the most pervasive--and certainly the most persistent--artifact of political behavior" (Monroe & Schrodt, 2008). Therefore, the systematic analysis of official texts is traditionally one of the instruments in the toolkit of those who want to make sense of political processes. Technology has greatly expanded the potential of such analysis, both by making tedious activities (e.g., looking for and counting keywords) much quicker and less error prone, and by greatly expanding the availability of texts to be analyzed (e.g., the Web is making virtually any relevant political text available to anybody in the world, mostly without charge). Automatic speech recognition is expanding even more the rich set of available documents

to analyze, transforming any recorded speech into a text. The growing popularity of blogging and then social network and micro-blogging platforms expanded further the potential of a systematic analysis of political texts, encompassing not only the analysis of texts produced by politicians and journalists, but also the automatic analysis of the viewpoint of ordinary citizens.

A more recent and relatively less explored strand of literature built on the previous ones, to explore whether democratic deliberation could be supported by natural language processing (NLP) tools, in order to enable citizens to pre-process an input to take more informed decisions (e.g., Muhlberger, Stromer-Galley & Webb, 2012; Jensen & Bang, 2013). The paper at hand fits in this last strand of literature, developing a framework to use NLP techniques to assist anyone interested in categorizing political speeches, including citizens who are forming their own political opinions. In particular, we will describe a first and preliminary attempt to do so using TellMeFirst (TMF), a tool for classifying and enriching textual documents leveraging the DBpedia knowledge base[1] and the English Wikipedia linguistic corpus.

Section 2 of this paper describes related approaches to the content analysis concerning political texts. Section 3 provides the reasons for using DBpedia as knowledge base for text classification in the political domain. Section 4 explains the TMF approach to text categorization. Section 5 reports the results of the text analysis with TMF. Section 6 draws the conclusions, and outlines some future developments.

Related Works

Automated content analysis concerning political texts progressed at a fast pace since Benoit and Laver's seminal works (Benoit & Laver, 2003; Laver, Benoit, & Garry, 2003) focusing on *wordscores*. (Wordscores is a procedure, still widely used, to infer policy positions --i.e., the scores-- associated with a document, on the basis of a training set of pre-classified documents. See Lowe, 2008 for a detailed description of this approach.) Similar techniques, with different statistical assumptions, have also been proposed by Slapin & Proksch (2008), also leading to the production of the Wordfish software[2]. Following these and other works, e.g., the one of Simon & Xenos (2004), more complex semantic analysis techniques are also becoming tools to assist and partly substitute the human coding of political content. For a recent and extensive survey of methods for the automatic analysis of political texts (which would be outside of the scope of this paper), we remand to Grimmer & Stewart (2013).

Sentiment analysis techniques, originally developed for marketing purposes, are more and more used to infer the political implications of the big flow of data exchanged on social networks and micro-blogging platforms (e.g., Tumasjan, Sprenger, Sandner, & Welpe, 2010 or Ringsquandl & Petković, 2013).

This paper fits in a third and relatively less explored domain, focusing on the use of natural language processing (NLP) tools to support and inform the political participation of citizens. In this specific domain, for instance, Muhlberger, Stromer-Galley & Webb (2012) discuss how NLP tools can empower participants to provide more informed input into public comment processes related to federal and state agency rulemakings (in the US). To our knowledge, the paper at hand

[1] http://dbpedia.org/.

[2] See http://www.wordfish.org/publications.html for a list of related publications (and applications to various cases).

may also be the first one using Linked Open Data, and in particular the unique URIs exposed by DBpedia, to unambiguously identify the categories of political texts. (Related works specifically connected with our TMF NLP technology are mentioned in Section 5.)

DBpedia as a Knowledge Base to Enable the Classification of Political Texts

As described by Grimmer & Stewart (2013), "assigning texts to categories is the most common use of content analysis methods in political science. For example, researchers may ask if campaigns issue positive or negative advertisements [...], if legislation is about the environment or some other issue area [...]. In each instance, the goal is to infer either the category of each document, the overall distribution of documents across categories, or both." Text classification consists in assigning a document to one or more categories or classes among a number of possible classes. When categorization is performed on the basis of documents' topics (often called "semantic" categorization), the set of possible categories is part of a taxonomy, an ontology or knowledge based where nodes are "concepts" or "topics". Since in most machine learning-based classification systems, such as TMF, categorization is accomplished by calculating a similarity score between target document and all possible categories, classification process works the more successfully the greater is the coverage of the domain of interest in the knowledge base.

DBpedia has proven to be a very suitable knowledge base for text classification, according to both technical reasons and more theoretical considerations (Mendes et al., 2012; Hellmann et al., 2013; Steinmetz et al., 2013). DBpedia is directly linked to the arguably largest multilingual annotated corpus ever created, which is Wikipedia: thus, it is technically perfect for automated tasks in the fields of Natural Language Processing and Text Mining. As lately noticed, "DBpedia has the potential to create an upward knowledge acquisition spiral as it provides a small amount of general knowledge allowing to process text, derive more knowledge, validate this knowledge and improve text processing methods." (Hellmann et al., 2013). Besides, concepts within DBpedia (called "entities" and identified by URIs[3]) are the result of a semantic consensus collaboratively reached by a wide community of Internet users (the "Wikipedians"). An effective criterion for classifying documents on the Web, in fact, should not be imposed from above, but it should follow the same principles of freedom and transparency that have always been the essence of the Internet itself.

The uneven coverage of different topics in Wikipedia is reflected in the DBpedia knowledge base with a greater or lesser presence of entities and relationships between entities. If a Wikipedia article is particularly full-bodied and rich in information, it will be characterized by numerous inbound links, and will have a very rich and structured infobox: accordingly, the profile of the corresponding DBpedia entity will be more complex. This has a deep impact on a DBpedia-based classification software, because documents about some topics will be classified more accurately than others.

As explained by Brown (2011), the political coverage in Wikipedia is "often very good for recent or prominent topics but is lacking on older or more obscure topics". Assessing the accuracy of Wikipedia in reporting gubernatorial candidate biographies who ran between 1998 and 2008 in US,

[3] For example: http://dbpedia.org/resource/Barack_Obama.

and the accuracy of the US gubernatorial election results reported on Wikipedia, Brown also notices that Wikipedia's greater flaws are the omissions rather than inaccuracies.

As an indicator of the coverage of a topic in Wikipedia, we detected the Wikipedia category[4] that seemed to describe more accurately that topic and we count how many Wikipedia articles fall into that Wikipedia category.

In order to compare the coverage of US politics with the coverage of politics of other countries, we identified three main areas of political domain, selecting in each area three Wikipedia categories for the countries of interest (i.e, United States, United Kingdom, and France). These main areas, which correspond to the graphs below, are: (i) conduct, practice, and doctrine of politics of a country (see Figure 1, in orange); (ii) official government institutions and offices of a country (see Figure 1, in blue); (iii) politicians involved in the politics of a country (see Figure 1, in green)[5].

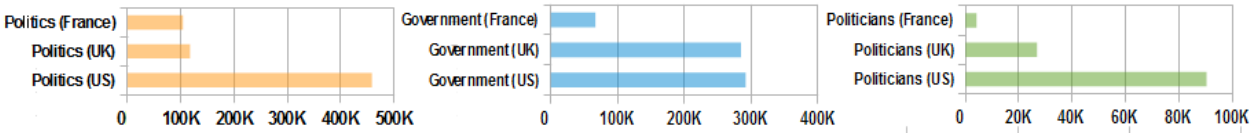

Figure 1: Comparison between the coverage of US politics and the coverage of politics of other countries

TellMeFirst Approach to Text Categorization

TellMeFirst is an open-source software for classifying and enriching textual documents via Linked Open Data[6]. TMF leverages Natural Language Processing and Semantic Web technologies to extract main topics from texts in the form of DBpedia resources. Every DBpedia resource (for instance http://dbpedia.org/resource/Barack_Obama) has a corrisponding article in Wikipedia (http://en.wikipedia.org/wiki/Barack_Obama), therefore TMF output is a list of Wikipedia topics intended to be the main subjects of the input text.

Like other software of the same kind (e.g., DBpedia Spotlight[7], Apache Stanbol[8], TAGME[9], etc.), TMF exploits DBpedia as a knowledge base for topic extraction and word sense disambiguation. DBpedia is a suitable training set for any machine learning-based approach, because it is directly linked to the wide, cross-domain linguistic corpus of Wikipedia. In order to accomplish document categorization, TMF adopts a memory-based learning approach, which is a subcategory of instance-based learning, also known as "lazy learning". Its distinctive feature is that the system doesn't deal with creating an abstract model of classification categories (aka "profiles") before the actual text categorization process, but it assigns target documents to classes based on a local

[4] http://en.wikipedia.org/wiki/Help:Category.

[5] A query on DBpedia (stored locally in a Virtuoso triplestore) traverses chains of skos:broader relations, using SPARQL 1.1 Property Paths, in order to obtain all members of each subcategory of a specific Wikipedia category: select distinct ?member where { ?member http://purl.org/dc/terms/subject ?cat ?cat skos:broader* http://dbpedia.org/resource/Catgory:Name_of_Category }.

[6] http://tellmefirst.polito.it/.

[7] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki.

[8] http://stanbol.apache.org/.

[9] http://tagme.di.unipi.it/.

comparison between a set of pre-classified documents and the target document itself (Cheng et al., 2009). This means that the classifier needs to hold in memory all the instances of the training set and calculate, during classification stage, the vector distance between training documents and target documents. Specifically, the algorithm used by TMF is k-Nearest Neighbor (kNN), a type of memory-based approach which selects the categories for a target document on the basis of the k most similar documents within the vector space. The variable k in TMF is always equal to 1, thus the winning category is the one which has higher similarity with the target document.

TMF training set consists of all the Wikipedia paragraphs where a wikilink[10] occurs. These textual fragments are stored in an Apache Lucene[11] index, as fields of documents which represent DBpedia resources. In the TMF index each DBpedia resource becomes a Lucene Document that has as many Lucene Fields as the paragraphs where a link to that resource occurs[12]. At classification time (following the "lazy learning" approach) the target document is transformed into a Lucene boolean query on all index fields, in order to discover conceptual similarity between the document and all textual fragments surrounding a wikilink in Wikipedia. To calculate similarity, TMF uses the Lucene Default Similarity, combining Boolean Model of Information Retrieval with Vector Space Model: documents "approved" by Boolean Model are scored by Vector Space Model. The similarity between two documents can be viewed geometrically as the distance between two vectors that represent the documents in a n-dimensional vector space, where n is the number of features of the entire training corpus.

In a Lucene query, both the target document and the training set become weighed terms vectors, where terms are weighted by means of the TF-IDF algorithm. The query returns a list of documents in the form of DBpedia URIs, ordered by similarity score. Scoring formula is:

$$\text{cosine-similarity}(q,d) = \frac{V(q) \cdot V(d)}{|V(q)|\,|V(d)|}$$

where q is the query, d is the training document, V(q) is the query weighted vector, and V(d) is the document weighed vector. The above equation can be viewed as the dot product of the normalized weighted vectors, in the sense that dividing V(q) by its euclidean norm is normalizing it to a unit vector. Once we got the sorted list of results, we can apply RCut thresholding to keep only the first n topics and discard others.

TellMeFirst Effectiveness in the Domain of Interest

In order to verify the effectiveness of the TMF classification process, we used as test set the profiles of the US Presidents published on The White House website[13]. We run TMF on these documents performing two test suites. In the first test suite we submitted the US Presidents profiles to TMF and we collected the classification results. For each profile TMF provided as output the seven most relevant topics (in the form of DBpedia URI) of the document sorted by relevance. On the basis of our evaluation criterion, a topic detection result is correct if the first DBpedia URI refers to the US

[10] http://en.wikipedia.org/wiki/Wikilink#Hyperlinks_in_wikis.
[11] http://lucene.apache.org/.
[12] This technique has been borrowed from the DBpedia Spotlight project (Mendes at al., 2011).
[13] http://www.whitehouse.gov/about/presidents.

President described in the profile[14]. Results show that TMF managed to identify the first topic of a text with precision of 95.4%. In the second test suite, slightly more challenging, we automatically removed all the strings referring to the main topic's label (e.g. label "Barack Obama" for the topic http://dbpedia.org/resource/Barack_Obama), nevertheless TMF also identified the first topic (just on the basis of its linguistic context) with precision of 45.4%. Below an overview of TMF success rate on US presidents profiles. Full results are available for download on TMF website[15].

Table 1: Success rate (%) of the TellMeFirst classification process on the Us Presidents profiles

	1st topic	**Within the first 2 topics**	**Within the first 7 topics**
Full text of the Presidents profiles	95.4%	100%	100%
President profiles without name and surname	45.4%	61.3%	90.9%

Furthermore, the results obtained classifying White House speech transcripts demonstrate that TMF is far more suitable to make clear the subject of a political speech compared with a simpler bag-of-words based text analysis tool. The example[16] in Figure 2 shows how TMF identifies as the main topic the unambigous concept "Patient Protection and Affordable Care Act"[17] while a popular tag cloud tool[18] gives as result a set of often redundant or inconsistent strings.

Figure 2: Results obtained with TMF (on the left) and with TagCrowd (on the right)

Results

3173 videos in English were available on the White House website on the 24th of November 2013. These videos are part of the political communication of the White House and are categorized according to a taxonomy not related to the subject of the speeches. These categories are instead

[14] For the profile "Abraham Lincoln" available at http://www.whitehouse.gov/about/presidents/abrahamlincoln, the first result provided by TMF should be http://dbpedia.org/resource/Abraham_Lincoln.

[15] See note 6.

[16] "President Obama Speaks on the Affordable Care Act": http://www.whitehouse.gov/photos-and-video/video/2013/09/26/president-obama-speaks-affordable-care-act#transcript.

[17] http://dbpedia.org/resource/Patient_Protection_and_Affordable_Care_Act.

[18] http://tagcrowd.com/.

related to the place of the event ("Press Briefings", "West Wing Week"), and to the person who delivered the speech ("The First Lady", "The Vice President").

TMF tries to add a semantic layer that point out the content of the speeches, so that questions such as "what is the First Lady talking about?" could be automatically answered (see Section 5.2) and/or people interested in specific issues could easily find related videos.

This section reports the results obtained extracting the topics of speech transcripts published on the White House website. Table 2 shows the top 20 topics on the total number of occurrences and the value in percentage of each topic at the overall level. Furthermore, Table 2 reports the values in percentage of each topic, considering each year from 2009 to 2013. The values highlighted in red indicate a number of occurrences greater than 1% while the values highlighted in green indicate a number of occurrences greater than 0.5% (and lower than 1%). An interesting result with a high number of occurrences (141) is New Deal, probably used as a metaphor within the political speeches of President Obama[19]. Apart from these results, that give an overall view of the topics treated by the White House, there are some outliers that provide cases that can be further investigated.

The entity "Libya" (in the 61st place for number of occurrences) has a value corresponding to 1.00% in 2011, while is less than 0.2% in 2012 and in 2013, and it is not available for 2010 and 2009. This result can be related to the full-scale revolt beginning on 17 February 2011 in Libya and concluded on 23 October 2011.

A similar behaviour occurs with the entity "Deepwater Horizon oil spill". In 2010 it reaches the 1.05% of the occurrences, while it does not occur in 2013 and in 2012. This result is probably related to the marine oil spill which took place in the Gulf of Mexico that began on 20 april 2010 and concluded on 15 July 2010.

Table 2: Amount and percentage of topic occurrences extracted with TellMeFirst

Topic	**Occ.**	**% overall**	**% 2013**	**% 2012**	**% 2011**	**% 2010**	**% 2009**
Barack Obama	607	4.88%	5.68%	4.52%	5.51%	4.45%	3.88%
White House	381	3.06%	2.75%	2.91%	3.32%	2.94%	3.38%
Patient Protection and Affordable Care Act	286	2.30%	3.06%	1.35%	1.91%	2.47%	2.71%
American Recovery and Reinvestment Act of 2009	278	2.23%	1.09%	1.82%	2.88%	2.84%	1.88%
Social Security	272	2.19%	2.58%	1.77%	3.54%	1.61%	0.78%
Medicare	183	1.47%	2.10%	0.52%	1.19%	1.58%	1.99%
New Deal	141	1.13%	1.00%	1.25%	1.79%	0.90%	0.44%
Health insurance	131	1.05%	1.62%	0.31%	0.47%	1.14%	1.99%

[19] Obamacare vs. The New Deal Historical Comparison, New Republic, 24 October 2013 http://www.newrepublic.com/article/115326/obamacare-vs-new-deal-historical-comparison.

Economic growth	127	1.02%	0.96%	0.73%	1.00%	1.08%	1.33%
George W. Bush	116	0.93%	1.18%	0.83%	0.60%	1.21%	0.83%
Renewable energy	114	0.92%	0.52%	0.99%	1.00%	1.08%	0.89%
Unemployment	113	0.91%	0.61%	0.83%	0.94%	1.08%	1.00%
Iraq War	106	0.85%	0.57%	1.09%	0.97%	0.56%	1.27%
Bush tax cuts	98	0.79%	0.52%	1.19%	1.25%	0.68%	0.06%
United States Congress	98	0.79%	1.22%	1.09%	1.19%	0.31%	0.06%
Income tax	97	0.78%	0.17%	1.56%	1.00%	0.93%	0.06%
Robert Gibbs	88	0.71%	#N/D	#N/D	0.38%	1.67%	1.22%
Sales tax	87	0.70%	0.09%	1.30%	1.00%	0.83%	0.06%
Economic development	85	0.68%	0.66%	0.57%	0.69%	0.71%	0.78%

Correlations Among Topics in the Political Speeches

As explained in Section 4.1, the TMF text categorization process extracts the seven most relevant topics of a text. Exploiting this feature it is possible to quantify the correlation among the topics addressed in a political speech.

In Figure 3, for example, we noticed that the "War" is often associated to topics such as "Veteran", "United States Department of Veterans Affairs", "Veterans of Foreign Wars", "Vietnam veteran", likely very sensitive issues for the US electorate. Among other topics there are "Al-Qaeda" "September 11", "Terrorism", "Osama Bin Laden", a sign that probably this concept is often linked to the war on terrorism.

Focus on the First Lady Speeches

The Wikipedia page "First Lady of the United States"[20] represents the shared consensus among wikipedians (and a good proxy of the consensus amongst Internet users) about the role of the US First Lady. According to this view, the First Lady is "first and foremost, the hostess of the White House", she "often plays a role in social activism" and "organizes and attends official ceremonies and functions of state". Moreover, "[o]ver the course of the 20th century it became increasingly common for first ladies to select specific causes to promote, usually ones that are not politically divisive."

[20] http://en.wikipedia.org/wiki/First_Lady_of_the_United_States.

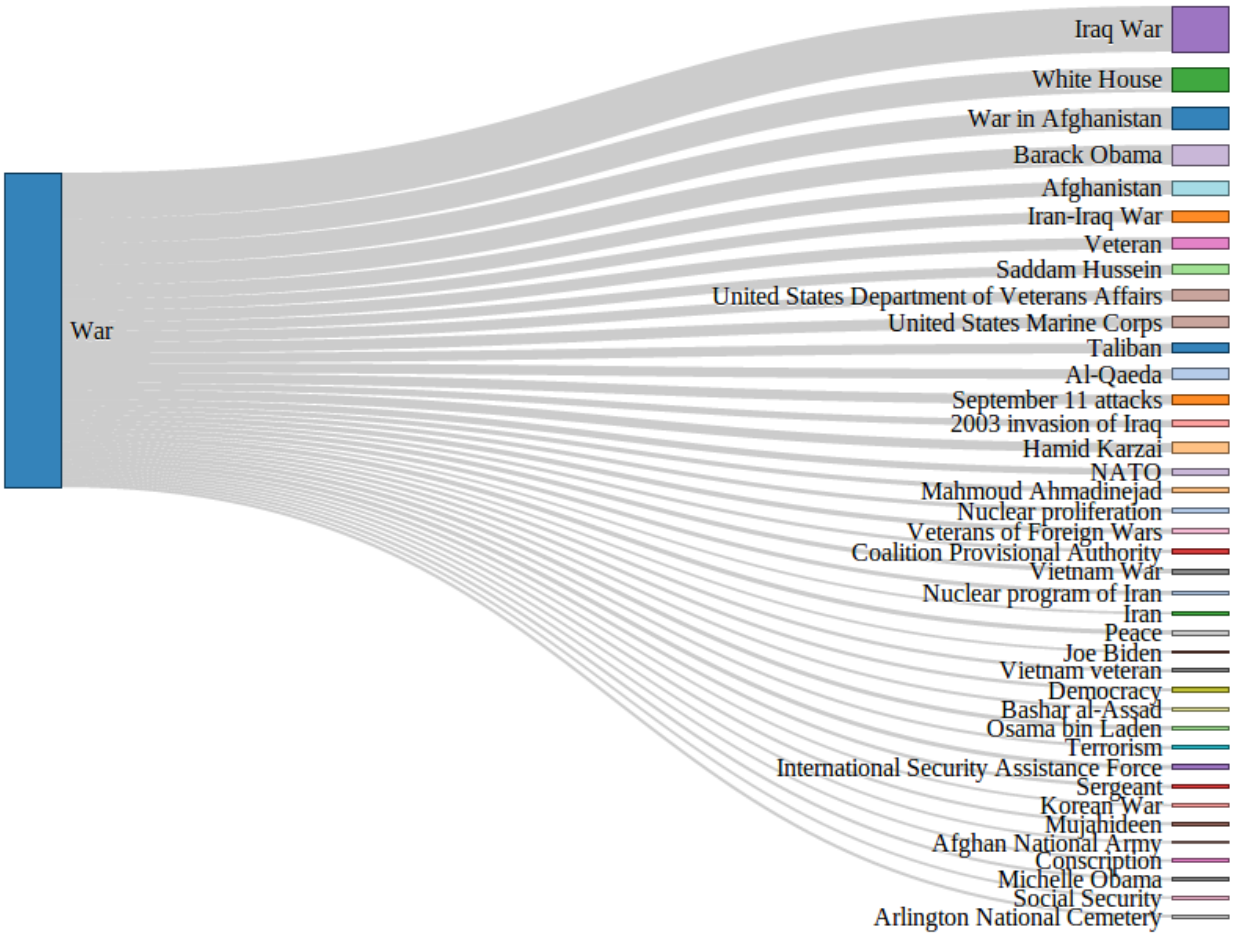

Figure 3: The most mentioned topics related to "War", sorted by decreasing number of occurrences.

According to Michelle Obama's page on the White House website, in her case these causes are in particular[21]: "supporting military families, helping working women balance career and family, encouraging national service, promoting the arts and arts education, and fostering healthy eating and healthy living for children and families across the country."

We tested whether TMF confirms or not these impressions and claims, manually selecting nine Wikipedia categories which seemed to be related to the aforementioned issues[22]. We then interrogated the SPARQL end-point of DBpedia with a query to collect all the topics of these categories and of their sub-categories until the third level. This is the kind of query we used:

```
select distinct ?member where {
    { ?member dc:subject Category:NAME-OF-CATEGORY .
    } union { ?member dc:subject [ skos:broader Category:NAME-OF-CATEGORY ] .
    } union { ?member dc:subject [ skos:broader [ skos:broader Category:NAME-OF-CATEGORY ] ] .
    } union { ?member dc:subject [ skos:broader [ skos:broader [ skos:broader Category:NAME-OF-CATEGORY ] ] ] . }}
```

[21] http://www.whitehouse.gov/about/first-ladies/michelleobama.

[22] We are perfectly aware of the fact that different categories could have been selected, leading to significantly different results. Here we just want to highlight a promising path, which has to be followed starting from the definition of a sound (e.g., statistically or otherwise empirically grounded) methodology.

We then associated each topic to one or more of the nine high-level categories (notice that one topic may fit in two or more categories and that some categories may even be sub-categories of another one, e.g., Gender equality is a second level sub-category of Social issues).

Table 3 shows that these nine categories encompassed almost 75% of the topics.

We then routinely tested the use of less categories, showing that four categories selected to maximize coverage still encompass more than 60% of the topics. (Because of the significant overlaps between, e.g., Education and Nutrition or Government of the United States and Barack Obama, we did not simply eliminate the smallest categories: for instance, Arts included much less topics than Nutrition, however it did not overlap significantly with other categories and it was therefore kept amongst the final four categories with the highest coverage.)

Table 3: Wikipedia categories addressed in the White House speeches with a focus on the First Lady

Wikipedia Category	**First Lady sp. 9 categories**	**First Lady sp. 4 categories**	**All speeches 9 categories**
Government of the United States	26.68%	26.68%	32.68%
Education	21.64%	21.64%	5.40%
Nutrition	19.96%	excluded	1.61%
Social issues	14.71%	14.71%	28.38%
Barack Obama	13.66%	excluded	14.00%
Health care	11.34%	excluded	7.57%
Arts	8.61%	8.61%	1.11%
Military personnel	3.99%	excluded	3.16%
Gender equality	2.73%	excluded	0.84%
Others (unclassified topics)	25.63%	37.61%	38.34%

Conclusions and Future Works

This paper shows the effectiveness of a DBpedia/Wikipedia-based approach for document classification in the e-government field, showing as use case the analysis of speech transcripts of the White House political members.

The ability for citizens to easily retrieve the content of political speeches and decisions is a crucial factor in e-participation. This is not guaranteed by a traditional keywords search, as in most of the public administration websites. The White House online portal, for example, offers a textual-search interface and minimal categories, which only allow users to find keywords in the video's title. By typing the word "education", for instance, users get as result only videos that have the word education in their title. But all the terms that belong to the semantic area of education (such as "university", "school", "students", "teachers", "curriculum", etc.) are omitted. When documents

are semantically classified through DBpedia URIs, instead, all synonyms, hypernyms and hyponyms of lemmas are traced to the same concept (in this example, all the listed words are gathered under the entity http://dbpedia.org/resource/Education), making user search more effective. Besides, leveraging Wikipedia categories would allow to go even a step further, taking advantage of the links between concepts as designed by the Wikipedia community.

The main future development of our project is therefore to build around the scraping / classification module a software layer of semantic search and navigation of the contents. There are many advantages of using a knowledge base to increase the "intelligence" of a document search engine: semantic indexing, faceted browsing, graphical conceptual navigation, search recommendation, related concepts, integration with other Linked Open Data repositories on the Web. This kind of user experience certainly increases the citizens' awareness to the issues discussed by politicians in their country.

The entire TellMeFirst code, including the algorithm which computes document similarity for assigning a classification, is open source. In future developments, one or more online communities can customize and improve the default classification algorithm according to their goals of political participation.

References

Benoit, K., & Laver, M. (2003). Estimating Irish party policy positions using computer wordscoring: The 2002 election–a research note. *Irish Political Studies*, 18(1), 97-107.

Brown, A.R. (2011). Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage. *PS: Political Science & Politics*, (44) 339-343.

Cheng, W., & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. (W. Buntine, M. Grobelnik, D. Mladenic, & J. Shawe-Taylor, Eds.) *Machine Learning*, 76(2-3), 211-225. Springer.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.

Hellmann, S., Filipowska, A., Barriere, C., Mendes, P. N. & Kontokostas, D. (2013). NLP & DBpedia - An Upward Knowledge Acquisition Spiral. *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, October, Sydney, Australia: CEUR Workshop Proceedings.

Jensen, M. J. & Bang H. P. (2013). Occupy Wall Street: A New Political Form of Movement and Community? *Journal of Information Technology & Politics*, 10(4), 444-461.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02), 311-331.

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, 16(4), 356-371.

Mendes, P.N., Jakob, M., Bizer, C. (2012). DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2012, 21-27 May 2012, Istanbul.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia Spotlight : Shedding Light on the Web of Documents. Text, 95(2), 1-8. Facultad de Informática (UPM).

Monroe, B. L., & Schrodt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. Political Analysis, 16(4), 351-355.

Muhlberger, P., Webb, N., & Stromer-Galley, J. (2008). The Deliberative E-Rulemaking project (DeER): improving federal agency rulemaking via natural language processing and citizen dialogue. *Proceedings of the 2008 international conference on Digital government research* (pp. 403-404). Digital Government Society of North America.

Muhlberger, P., Stromer-Galley, J., & Webb, N. (2012). An Experiment in E-Rulemaking with Natural Language Processing and Democratic Deliberation. In K. Kloby, & M. D'Agostino (Eds.) *Citizen 2.0: Public and Governmental Interaction through Web 2.0 Technologies* (pp. 23-40). Hershey, PA: Information Science Reference.

Simon, A. F., & Xenos, M. (2004). Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis. *Political Analysis*, 12(1), 63-75.

Steinmetz, N., Knuth, M., & Sack, H. (2013). Statistical Analyses of Named Entity Disambiguation Benchmarks. *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia, October, Sydney, Australia*: CEUR Workshop Proceedings.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705-722.

Ringsquandl, M., & Petković, D. (2013, March). Analyzing Political Sentiment on Twitter. In 2013 AAAI *Spring Symposium Series*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.Chadwick, A. (2006*). Internet Politics: States, Citizens, and New Communication Technologies*. Oxford: Oxford University Press.

About the Authors

Giuseppe Futia

Giuseppe Futia is a Research Fellow and the Communication Manager at the Nexa Center for Internet & Society at Politecnico di Torino, since February 2011. He holds a Master Degree in Media Engineering from Politecnico di Torino. Giuseppe has expertise in data analysis and data visualization, useful to both sustain the outreach of some of the Nexa projects, and to support research in the field of Open Data.

Federico Cairo

Federico Cairo, Ph.D., is a project manager at Expert System SpA and a fellow at the Nexa Center for Internet & Society. His main research interests are Linked Data technologies and natural language processing. He is the technical lead of TellMeFirst, an open-source software for classifying and enriching textual documents via Linked Open Data.

Federico Morando

Federico Morando is an economist, with interdisciplinary research interests at the intersection between law, economics and technology. He holds a Ph.D. in Institutions, Economics and Law from the Univ. of Turin and Ghent. He is the Director of Research and Policy of the Nexa Center for Internet & Society. From Dec. 2012, he leads the Creative Commons Italy project.

Luca Leschiutta

Luca Leschiutta is the IT manager of the Nexa Center for Internet & Society and of the Human Genetics Foundation of Torino. He has an MSE in Electronic and a Ph.D. in Information Technology pursued at the Internet Media Group of the Politecnico di Torino. He also taught programming courses at the a.m. Politecnico. In the past he worked as a reliability engineer at Alenia Spazio in the ISS project.