Online News Media Bias Analysis using an LDA-NLP Approach

Sarjoun Doumit and Ali Minai
School of Electronic & Computing Systems, College of Engineering, University of Cincinnati, Ohio 45221-0030, U.S.A. (email: doumitss@mail.uc.edu & ali.minai@uc.edu).

It is widely recognized that every media outlet has its own "spin" on news, and this bias has been described in many ways and at many levels. In political news for example, the bias can be liberal, conservative, moderate, corporate, etc. In addition, recent research has focused on the 'sentiment dimension' to further identify and categorize news bias. This is achieved through analysis of the adjective and adverb terms found in the news texts. The accuracy and generality of these models depend on the evaluation methods used to appraise the intensity and emotional weights of the adjectives and adverbs, thus rendering the results open to controversy. In this paper we propose a unifying system to extract information from political news texts and analyze it within a cognitive network. We view the different news media sources as agents with unique personalities, which we assume are latent within their texts. We use a combination Latent Dirichlet Allocation (LDA) and natural language processing (NLP) methods to identify the different agents' personality traits with respect to various topics or concepts. An agent's personality traits affect its inclination to word a certain event in a specific way. Using the common concepts stored in the cognitive network, our system can compare the different agents on a unified and normalized platform.

1 Introduction

Suppose that there is a finite collection of news sources or media outlets that we get all our daily news from on a regular basis. We can actually describe this phenomenon as a complex system, were the media outlets are agents thriving in

a large semantic environment composed of all types of news, ideas and memes. The media-agents are news producers and sometimes even consumers, and each one occupies a niche in that environment. As with any complex system, every agent receives a stimulus, which in this case originates from the 'real' world in a form of from a factual event. Each agent then proceeds into *manifesting* this event into a *news structure* which gets released into the semantic environment. These generated news stories are mutated forms of the same factual event, each carrying the *genetic* signature of its media source originator. But there exists many media outlets, covering all types of events and producing a large amount of news for all types of events. Understanding and categorizing the information found in these texts makes it an indomitable task for the reasons mentioned.

This is especially problematic for political news analysts and policy makers who try to understand and track the sentiment or *bias* found in these stories which could lead them to the hidden agendas of the groups behind the media outlets. There are many metrics one could propose to measure this bias, which is latent but yet palpable in each media outlet. Defining what bias actually represent when dealing with news of a political nature becomes a critical step for quantitatively measuring the bias and position of any news. There exists many statistical tools for natural texts that might be used in an attempt to make sense and categorize the different texts into groups, but so far these tools fail to give any additional insight into the general news. This fact is due to the complex *semantic* nature of these texts in addition to their existence in great numbers. Also when using statistics, one has to be careful to take into account data skews, because some media outlet are much more active and produce more news than others, and therefore can saturate the semantic environment with its own preferred word.

Even more important, coming up with a unifying and equitable system that measures the bias in the different media outlets on equal footing is critical to obtain any useful results. This is especially useful when complex natural language processing techniques are used. The usage of NLPs allows for the analysis of adjectives and adverbs which are usually associated to *sentimental* bias in conjunction with specific keywords related to a specific event. Such approaches for using adjective and adverb-based sentiments are still subject to the same weaknesses mentioned prior because of the great number and diversity of the political news, especially since in these approaches, the sentiment factor is weighted by a group of experts and are not *emergent* from the system.

We applied *latent Dirichlet allocation* (LDA) [2], a probabilistic topic modeling tool to extract latent topics from our data. In LDA each news document is considered to be generated from a distribution of topics where a topic is a distribution over words. We also employed Antelope[14], a natural-language processing (NLP) tool to analyze the documents' semantic structure, in combination with LDA. Our approach, applying LDA and Antelope together to a semantic framework that incorporates sentiment allows us to achieve relevant insights that neither system can achieve on its own. The aim of this paper is to test our model for establishing media-outlet personality signatures based on

the semantic structure of its news articles which offers its user a more robust framework for comparison and analysis

The rest of this paper is organized as follows: Section II reviews similar systems. Section III gives an overview of the LDA model, followed by a discussion in section IV of our proposed model. Finally in section V simulations, results and conclusions are presented.

2 Background

There exist many research and commercial systems today that analyze and cluster textual news employing methods that range from the purely statistical to graphical models. It is up to the news analyst or user of the system to organize the output according to his or her own specific needs to benefit from the result. To mention a few systems, WEIS[10, 15] and CAMEO[5] are both systems that use *event analysis*, i.e. they rely on expert-generated dictionaries of terms with associated weights then parse the text to match the words from the news event to those in the dictionary. Then they can categorize the information again into a set of expert-predefined categories with respect to the sentiment intensity values. In other systems, such as Oasys2.0 [3], they use another construct called *opinion analysis*, which instead depends on user feedback rather than on experts, in order to determine the value of the intensity of an opinion of a topic. The Oasys2.0 approach is based on aggregation of individual positive and negative references identified which have been evaluated on the individual sentiment [1]. Again other systems, RecordedFuture [4] and Palantir [13], rely on experts and have at hand massive amounts of data, with inference and analysis tools that uses data correlation techniques to produce results in response to specific keywords in user queries. And last, and as recently as this year, topic chain modeling [7, 12, 8] tracks topics across time using a similarity metric based on LDA to identify the general topics and short-term issues. It is important to notice that all the systems mentioned above, except to the latter adopt query-driven approaches to produce results.

3 The Latent Dirichlet Allocation model

There has been a recent interest in Latent Dirichlet Allocation or LDA ever since the publishing of the seminal paper by Blei, Ng and Jordan[2]. It is a machine learning technique (shown in extended version in Fig.1), that evolved from a previous model called *Probabilistic Latent Semantic Analysis* [6] (pLSA) for reducing the dimensionality of a certain textual corpus while preserving its inherent statistical characteristics. The main advantage of LDA is that it assumes that the documents in a corpus have multiple latent topics which, in turn, are distributions over words found in the documents of the corpus. LDA assumes that documents are made of a list words where the order of the words is not important i.e. Bag-Of-Words approach. LDA is a generative model, it can gen-

erate a document from a set of topics but can also be used as an inference tool to extract that topics back from a corpus of documents. To generate a corpus of D documents, where each document has N_d words, and for a total of T topics, LDA's generative algorithm is:

1. Pick a document size N_d
2. Pick a set of topics $\theta \sim Dirichlet(\alpha)$
3. For each of the N_d words w_n found in document d:
   (a) Draw a topic $t_{w_n} \sim Multinomial(\theta)$
   (b) Draw a word $w_n \sim Multinomial(\phi_{t_{w_n}})$, where $\phi \sim Dirichlet(\beta)$.

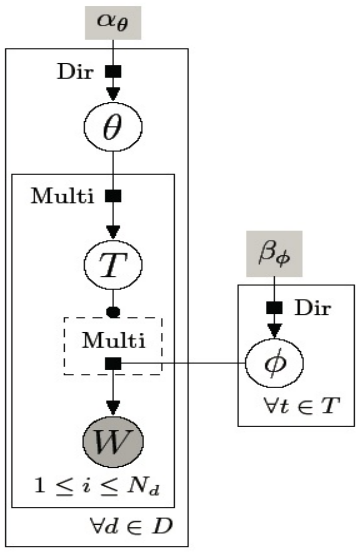

Figure 1: Directed factor graph for LDA. ϕ is the words-topic distribution, θ represents the topics-document distribution, α is the *Dirichlet* hyperparameter for θ, and finally β is the *Dirichlet* hyperparameter for ϕ

The probability equation is the following:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{t=1}^{T} P(\phi_t; \beta) \times \prod_{d=1}^{D} P(\theta_d; \alpha)... \\ \times \prod_{w=1}^{N_d} P(T_{d,w}|\theta_d) P(W_{d,w}|\phi_{T_{d,w}}) \quad (1)$$

The probability $P(T_{d,w}|\theta_d)$ is the probability of drawing topic T for word w from document d, given that that topic's distribution for that specific document d is θ_d. On the other hand, $P(W_{d,w}|\phi_{T_{d,w}})$ is the probability of drawing word W for the wth word from document d assuming it is drawn from the distribution for topic $T_{d,w}$.

4 Proposed Model

We start by defining how we view the process of political news story generation and how bias gets embedded into the very fabric of the original factual news. Let E represent a factual event, then E^{K_c} represents the smallest unbiased text that can summarize event E using K number of factual clauses c. A factual clause contains the basic semantic elements to describe the event or parts of the event without any usage of adjectives or adverbs. Let $\mathbb{O}$ represent the existing media outlets as a population of agents such that $\mathbb{O} = \sum_1^Z O_i$, where O_i is an identifiable media outlet with identity i and Z is the total number of media outlets in the environment. Each O_i is characterized by a set of subject matters $\mathbb{S}$ and a set of biases $\mathbb{B}$, where each bias $B \in \mathbb{B}$ is associated with at least an S i.e. $\forall S \in \mathbb{S}$, $\exists B \rightarrow S$. So the bias is defined by $B^{L_d}_{S_i,O_i}$ where L_d represents a number of L sentences d where d is a dependent clause, which are all associated to the ith subject S for media source O_i. A dependent clause could be either an adjective clause or an adverb clause, an adjective clause is a clause with an adjective in it and an adverb clause is an clause with an adverb in it.

When a reporter who works for a specific media outlet O_i comes across a factual event E, he/she first tries to figure out what subject(s) S this event is relevant to. Then armed with the specific bias B for S, he initiates the process of creating a news story using the L_d of bias B for S of O_i in conjunction to the E's K_c to create a biased version of E which we label as E_b. E_o stands for the original factual event and $E_b^{O_i}$ represents the biased news story of E for media outlet O_i. The average news reader receives on his computer the final product which is the biased event in form of a text, and in this paper we're trying to break apart the biased news story and try to isolate the biases or the L_d which includes the adjectives and adverbs in order to compare them on a deeper semantic/cognitive level instead of a pure lexical-weighted comparison.

The process of *weaving* the bias into the factual event is another process that controls yet three smaller sub-processes, each governing a collection set of text generated, which we call 'Actors', 'Actions' and 'Sentiments'. These processes work together in a preset fashion and seamlessly as if they are one unit, exactly like *synergies* (which we call meme-synergies(Fig.2)). In the domain of motor control, synergies are primitives or smaller subsets of degrees of freedom that are jointly controlled and it is agreed that the control of complex body movements is most likely be organized in terms of synergies. We compare the synergies responsible for moving one's arm to hit a tennis ball, or hands to play a piano, to our meme-synergies for generating coherent textual news products. In our model we consider a media's biased clauses or *embedded memes* to be preconfigured to

trigger or produce text-generation in accordance with those of the original event E, especially amongst those that share a similar construct such as a concept or sentiment.

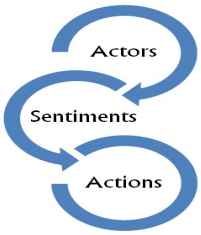

Figure 2: Textual/Meme-Synergies

In our model, the 3 parts of every synergy is a combination of *Actors* represented by proper nouns and nouns or (concepts and features), *Actions* represented by verbs, and *Sentiments* represented the adjectives and adverbs. For example, the proper noun "Obama" is an Actor which is strongly correlated with another noun "president" in most media irrespective of bias. On the other hand, based on context and bias a controversial concept or Actor such as the "IRA" which could be strongly related to the sentiment adjective "terrorist" by one media, or "freedom fighter" in another.

We will briefly describe how we build our semantic framework. We use LDA to extract the set of topics from our database that covers a specific month in order to have a time frame to compare to. The news articles that we treat are usually originating from a single source. We repeat that for all the sources that we have and then do a general LDA topic extraction using all the data that we have across all sources for that month. We generate 10 topics for each news source and we pick the top 10 words for each the 10 topics. These words are then fed to Antelope and with the help of our wrapper, identify the semantic property of these words and all the words in the semantic chunks that they are associated with. These constructs become the the anchors for the formation of more complex *concepts* which, in turn, are organized to form newsMemes. These chunks are identified and labeled to become the semantic constructs which are at the roots for the sentiments, actors and actions. This allows us to create a hierarchical network of coherent and relevant sets of words and phrases which we can then use to build memes. This means that we can solve the context issue where we can identify all the role that an actor can assume in all the news stories and their attached sentiments. For example the rigged "elections" in Ivory Coast that had different sentiment values than those of Egypt after their revolution. Therefore at the very center of a large group of similar topic-oriented news stories is a cluster of memes that are organized based on the distance in terms of phrases. If we visualize this as a network, it would be easy to see different group of equally connected phrases, with the ones with the highest values at the very center of the network and the rest are distributed around the

center.

When organizing our synergies, we simply reinforce the stronger coherent memes and inhibit the weaker ones. This results in forming smaller phrases that truly explain the reigning meme or dominant semantic topic from the LDA generated topics. The ambiguity when viewing the traditional Bag-Of-Words topics from LDA arises because a simple distribution of words representing different topics doesn't explain much and can be hard to discern. In our approach, reinforcing the strongest memes within the same LDA topic and across all topics gives a more robust platform to identify the memes behind the media outlet in a more human-understandable fashion. In many cases we discovered that actually the same semantic topic can have more than one synergy that are related but not the same as will be shown in the next section 5. We will also study the distributions and *shapes* of the media's meme-synergies in a qualitative way to see if we can find any correlation between the bias of the medias and the specific subject-topic .

5 Simulation results and conclusion

We've been collecting and building an extensive database covering about 33 online world-wide news media sources through their RSS feeds [9] to test our assumptions and model, i.e. $\mathbb{O} = \sum_1^{33} O_i$. We collect around the clock, at specific intervals of the day, the news articles ($E_b^{O_i}$) from these that these news media have to offer. Obviously this means that there will be redundancy because an important news article can *stick* on the top list of news longer than others, or might *evolve* with time as more information and analysis becomes available regarding its event. This aspect of news lingering is important to our model and we capture it so that we can measure the intensity and importance that this news represented for its parent media sources. This represents another aspect of bias, which is not covered by the conventional systems that only look for keywords such as adjectives. This aspect of bias captures the preferential alignment towards a certain direction or subject, which is another *latent* way of influencing the readers in accordance to a specific agenda. So repetitiveness is captured as a re-enforcing rule that strengthens the bonds across the synergetic functions. As previously mentioned we used a smoothed-LDA method based on the work of Newman [11] in addition to a wrapper program around Antelope in order to use it efficiently for our NLP analysis. In the following tables and graphs, we will describe the different aspects of bias capturing that our model has to offer with respect to actual events that happened world-wide.

We start by showing in table 1 the smoothed-LDA topic results for a sample of media sources, mainly the NewYork Times, BBC, CNN, USA Today and CBS for the month of December 2010. After getting the 10 topics for each media source, we chose 3 general 'themes' if we can call it that, because judging from the Bag-Of-Words result alone, that's the best a human reader can discern. So the themes were ***China***, ***WikiLeaks*** and the ***Koreas***. As a reminder, in the month

of December there was an international issue with China protesting the winning of a Noble peace prize by one of its citizens, the famous Wikileaks with their unveiling of secret documents and the incidents between North and South Korea with the North firing at the South. Table 1 shows for each theme the resultant topic for those media sources and also indicates the table number where the meme results for that theme are shown. We should state that the comparison of the different news sources will yield more results when multiple media outlets are compared across many themes so that the pattern and differences become more apparent. So after reviewing all the results shown in this paper, one can have a deeper appreciation to the differences between the media.

The results shown in table 2 are from the NY Times, BBC and CNN for the topic involving a story about the China. Note that this theme did not appear in the top 10 global themes that were common to all the 33 media outlets but was a theme that appeared in the top 10 topics for the individual media outlets. What we first notice is that *in general*, the highest ranking meme less number of words than that of lower memes, and that is normal because it is supposed to be more general. We see that for the NY times the important results were *Nobel Peace Prize*, *Liu Xiabo dissident* and *block foreign news* which clearly explains the story which is about a dissident Chinese citizen who won the Nobel Peace prize and that China is blocking the foreign news sources from this subject. The BBC has similar content but did not mention that foreign news where being blocked, but still one can make out the gist of the news. CNN on the other hand seemed to have at least the same information as the BBC but has also connected to so many other concepts that it rendered it more enigmatical. Another interesting point was that the system uncovered 2 memes for the NY Times.

The results shown in table 3 are from the NY Times, BBC, USA Today and CNN for the topic involving the Wikileaks story. Note that this theme did appear in the top 10 global themes that were common to all the 33 media outlets. What we first notice is that CNN happens to have the same pyramid-like shape for its memes, while other media outlets kept the same shape for the previous theme as well. What's interesting is that from all news sources, it was the BBC's meme that mentioned the sexual allegations against the Wikileaks founder Julian Assange. The NY Times was again more objective with clear phrases such as *American whistle-blowers Web site* while CNN meme mentioned all the different concepts that were affected by the leaks, such as Italy's Berlusconi and Afghanistan's Asif Rahimi etc..

Finally the results shown in table 4 are for the NY Times, BBC, USA Today and CBS for the topic involving a story about the 2 Koreas. Note that this theme did appear in the top 10 global themes. We start to see that NY Times memes always start with the main Actors that are involved and then about the Action of what is happening in the event. It is evident from other news sources that the memes have now more sentiments i.e. *defiant*, *major* etc.. What is interesting was that CBS used terms such as *war track*, *remains defiant* and especially *Attack Exercices* vs NY Times's *Island Drills* at the same ranking level. The more provocative nature of the former is evident with respect to

Media	*Theme*	Bag-of-Word Top 10 Topic Words	Table ref.
NY Times	China	china chinese lives peace prize secretary forces russian friday president	Table 2
BBC	China	election china minister host government internet peace presidential prize help	Table 2
CNN	China	peace prize thursday christmas nobel country death chinese christian explosion	Table 2
NY Times	WikiLeaks	wikileak american cables diplomatic continues site update leak reader blower	Table 3
BBC	WikiLeaks	police wikileak afghanistan country president founder court minister high assange	Table 3
	WikiLeaks	wikileak government cables attack manchester united league minister released reveal	Table 3
USA Today	WikiLeaks	wikileak police christmas julian assange country house british friday million	Table 3
CNN	WikiLeaks	wikileak charges court assange authorities julian president police thursday london	Table 3
	WikiLeaks	minister wikileak government prime website sunday diplomatic cables told latest	Table 3
NY Times	Koreas	afghan attack north south leader american tension korea killed killing	Table 4
	Koreas	afghan attack north south leader american tension korea killed killing	Table 4
BBC	Koreas	south korea north according film award children figures suggest tiger	Table 4
CBS	Koreas	president obama north korea report prince attack afghanistan south leader	Table 4
	Koreas	military report china glor jeff korea show north speed drill	Table 4
USA Today	Koreas	korea military attack north south friday sunday month bomb terrorist	Table 4
	Koreas	korea north south island saturday attack korean monday official british	Table 4
	Koreas	afghanistan korea killed killing north attack taliban suicide eastern monday	Table 4

Table 1: LDA Topics for NY Times and BBC for the 3 Stories shown

#	NY Times	BBC	CNN
1.	Friday	China	Chinese
2.	Chinese ceremony	Nobel Peace Prize	dissident China
3.	Nobel Peace Prize	Nobel Chinese Dissident Liu Xiaobo	peace imprisoned Liu Xiaobo prize nobel thursday
4.	Liu Xiaobo dissident	ceremony jailed	award committee Nobel Peace Prize
5.	imprisoned	committee website winner	Pope Benedict XVI Middle East absentia Friday protests condemnation laureate renowned artist human rights advocate rhetoric interference internal affairs first-ever Norwegian choice latest casualty government effort no-fly list prominent guests travel ban ceremony director Nobel Institute winner Peace Prize chair troubled lands
1.	China		
2.	block foreign news		
3.	schedules Skip Ceremony Nobel		

Table 2: Chinese Dissident Noble Prize Memes

#	NY Times	BBC	USA Today	CNN
1.	leak	Assange Julian founder Wikileaks	Julian Assange	cables diplomatic
2.	WikiLeaks	court	WikiLeaks founder	government US Prime Minister parliament WikiLeaks
3.	diplomatic	London	British	sunday Iraqi Nuri al-Maliki latest Tuesday
4.	cables	appeal allegations sexual	judge Tuesday bail	coalition cabinet rival incumbent Monday months-long political stalemate Ayad Allawi series votes leadership dispute ally summer allegations website
5.	American whistleblowers Web site	custody assault	police court appeal	online leaked island nation country deep sectarian tensions throats Zimbabwe elections process view writers minister level Agriculture Asif Rahimi bribery release pages military information United States Wednesday thousands sensitive picture corruption Afghanistan fall Berlusconi confidant cable supporters Australian

Table 3: WikiLeaks

the more descriptive definition of the latter. Finally we noticed that when we compared each media outlet's LDA's 10 top topics to the global LDA's 10 top topics (shown in figure 4) across every day in December 2010, we found out that the topics that had the most match where the same topics with larger memes. In figure 3 we show in a graph how the BBC's Korea's memetic-synergy looks like, with the nodes with the same distance from the central node arranged on a circle and the semantic connections that link each node to the other.

#	NY Times	BBC	USA Today	CBS
1.	North Korea	South Korea	North Korea	president
2.	South Korea	live-fire North Korea north	military	North Korea
3.	South American North Korea island drills	largest force tensions tinderbox provocations planned military drills disputed maritime border major exercises warnings drill Women defectors envoy Russia today sacred war	attack South Korea	South Remains Defiant Attack Exercises
4.	Talks live-fire artillery Yeonpyeong		drills island north shelling	War track
5.	leader korea Sign US-China Tension Delay Taliban Afghanistan Military nuclear		jets live-fire month sunday chief defense south deadly Korea border troops	Drills Heavily Armed Border Planned Last Minute Remains Defiant review strategy troops US North Korea Lee Myung-bak

Table 4: South Korea and North Koreas Tensions

This paper investigated a novel approach to infer information from a large and diverse corpus of political news. Furthermore, the paper suggested that the semantic coherence of an article is a result of it being influenced by a textual synergy and hence different news sources can be evaluated by observing their corresponding memes. By lowering the entropy of the standard LDA's topic distribution we can achieve better undersanding of the news corpus for deeper

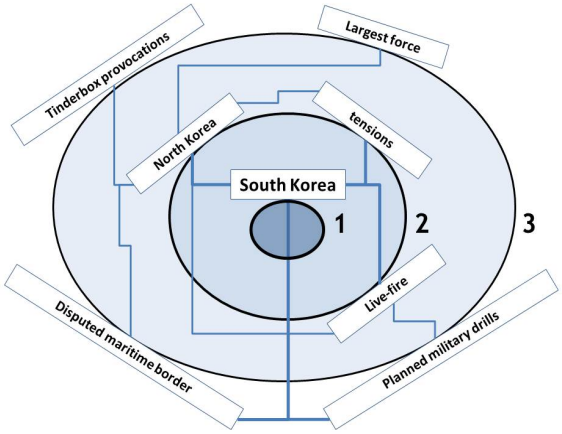

Figure 3: BBC's South and North Koreas synergy for December 2010

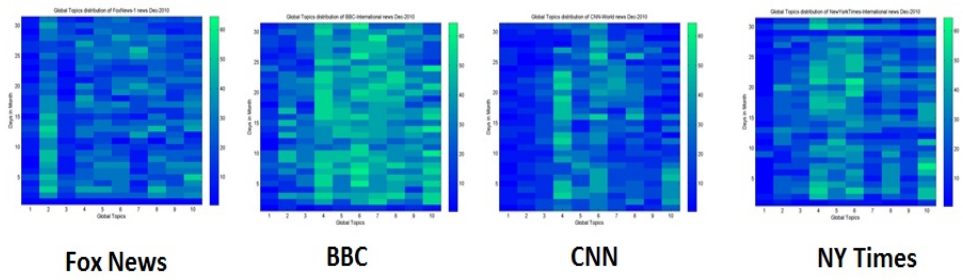

Figure 4: Comparison between 4 Media Sources for Normalized Topics

analysis and comparison.

Bibliography

[1] Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, and V.S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone", *International AAAI Conference on Weblogs and Social Media (ICWSM)* (2007), 203–206.

[2] Blei, D.M., A.Y. Ng, M.I. Jordan, and J. Lafferty, "Latent dirichlet allocation", *Journal of Machine Learning Research* **3** (2003), 993–1022.

[3] Cesarano, C., A. Picariello, D. Reforgiato, and V.S. Subrahmanian, "The oasys 2.0 opinion analysis system.", *International AAAI Conference on Weblogs and Social Media (ICWSM)* (2007), 313–314.

[4] Future, Recorded, "Recorded future - temporal & predictive analytics engine, media analytics & news analysis" (2010), [Online; accessed 22-November-2010].

[5] Gerner, D.J., R. Abu-Jabr, P.A. Schrodt, and . Yilmaz, "Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions", *International Studies Association of Foreign Policy Interactions* (2002).

[6] Hofmann, T., "Probabilistic latent semantic analysis", *Uncertainty in Artificial Intelligence, UAI99* (1999), 289–296.

[7] Kim, D., and A. Oh, "Topic chains for understanding a news corpus", *12th International Conference on Intelligent Text Processing and Computational Linguistics(CICLING 2011)* **12** (2011).

[8] Leskovec, J., L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), 497–506.

[9] Libby, D., "Rss 0.91 spec, revision 3", *Netscape Communications* (1997).

[10] McClelland, C., "World event/interaction survey", *Defense Technical Information Center* (1971).

[11] Newman, D., "Topic modeling scripts and code", *Department of Computer Science, University of California, Irvine* (2010).

[12] Oh, A., H. Lee, and Y. Kim, "User evaluation of a system for classifying and displaying political viewpoints of weblogs", *AAAI Publications, Third International AAAI Conference on Weblogs and Social Media* (2009).

[13] Palantir, "Privacy and civil liberties are in palantirs dna" (2004), [Online; accessed 10-December-2010].

[14] Proxem, "Antelope (Advanced Natural Language Object-oriented Processing Environment)" (2010), [Online; accessed 30-April-2010].

[15] Tomlinson, R.G., "World event/interaction survey (weis) coding manual", *Department of Political Science, United States Naval Academy, Annapolis, MD.* (1993).