Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2015) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Leveraging topic models to develop metrics for evaluating the quality of narrative threads extracted from news stories

Jason Schlachter[a], Alicia Ruvinsky[a], Luis Asencios Reynoso[a], Sathappan Muthiah[b], Naren Ramakrishnan[b]

[a]*Informatics Laboratory, Lockheed Martin Advanced Technology Laboratories, Kennesaw, GA, USA*
[b]*Discovery Analytics Center, Arlington, VA, USA*

Abstract

Analysts and software systems are increasingly tasked with making sense of a growing amount of data to help their organizations make decisions involving risk and uncertainty. A key enabler of this work is the ability to quickly discover structure in large amounts of text such as news stories and blogs. Recent work in this area has shown it is possible to automatically link documents from a corpus together to build a narrative structure, called a story chain, without the need for prior domain knowledge [1]. This approach is an unsupervised method that discovers large numbers of story chains of variable quality. In this paper, we describe and evaluate methods to identify the most coherent and informative story chains. We explore two types of topic model based analytics. The first type is a measure of representativeness that captures how well a story chain represents the corpus from which it was generated. This is done by comparing the similarity of topics found over time in a story chain against those expressed in the corpus during the same time period. Our hypothesis is that story chains that have similar topic expression to the corpus will convey narratives that are central to the corpus. This type of analytic could help an analyst quickly focus on the key narratives in a large corpus of documents. The second type is a measure of quality of a story chain and is composed of topic consistency and topic persistence measures. Our hypothesis is that high quality chains would be composed of sequences of stories that have clearly defined primary topics that persist across significant portions of the story chain. We used these analytics to predict the clarity of story chains within one of four categories (1) very clear narrative, 2) somewhat clear narrative, 3) somewhat unclear narrative, 4) very unclear narrative, and found we were able to train a data model to label story chains with the same label as human coders 77% of the time. Our dataset was composed of 7,074 English language news stories released during the Brazil Protests of 2013 from which 5,606 story chains were generated. We randomly selected 60 story chains for hand scoring to serve as our gold standard data set for experimentation.

Keywords: Sensemaking; Data analytics; Text analytics; Narrative; Machine learning; Topic modeling

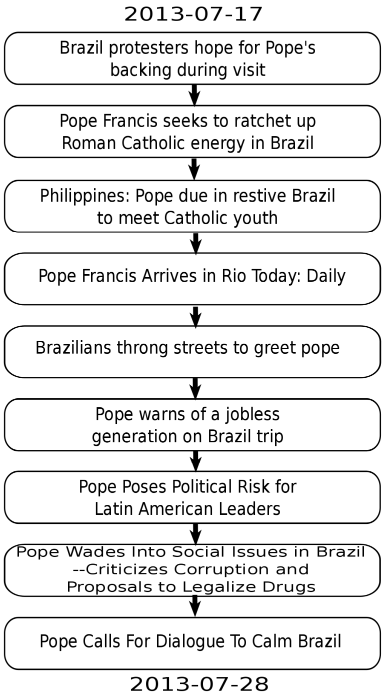
2013-07-17
Brazil protesters hope for Pope's backing during visit
Pope Francis seeks to ratchet up Roman Catholic energy in Brazil
Philippines: Pope due in restive Brazil to meet Catholic youth
Pope Francis Arrives in Rio Today: Daily
Brazilians throng streets to greet pope
Pope warns of a jobless generation on Brazil trip
Pope Poses Political Risk for Latin American Leaders
Pope Wades Into Social Issues in Brazil --Criticizes Corruption and Proposals to Legalize Drugs
Pope Calls For Dialogue To Calm Brazil
2013-07-28

Fig. 1. Example of a story chain, with headlines for each news story in each bubble.

1. Introduction

Making decisions involving risk and uncertainty in today’s global market and military problem spaces requires understanding and monitoring an increasingly large amount of data to build situational awareness. This motivates the need for automated methods of data analysis and sense making to help analysts overcome the vastness of the data. Existing approaches such as software that automatically finds events in text must be configured with domain knowledge [2] and/or labeled training data [3] leading to high costs and long development times. We have developed an unsupervised learning technique called Story Chaining that links related documents in a corpus to build a story or narrative arc [1]. Because it is fully unsupervised, this approach does not require any domain knowledge or training data, making it ideal for new and frequently evolving domains. Figure 1 shows an example story chain generated from a corpus of news stories published in Brazil in 2013.

The story chains generated from this approach can potentially tell a story about what is happening across time and across news articles by focusing on how the same people, organizations, and locations occur between documents. We consider the story chains to be a narrative structure. In this paper we consider ways to evaluate the clarity of the narrative structure contained within the story chain, proposing two different kinds of measures based on our insights from manual inspection of the story chains. The first type is a measure of representativeness that captures how well a story chain represents the corpus from which it was generated. This is done by comparing the similarity of topics found over time in a story chain against those expressed in the corpus during the same time period. Our hypothesis is that story chains with similar topic expression to the corpus will convey narratives that are central to the corpus. This measure assumes the corpus contains dominant topics that are desirable to understand. For example, the story chain in Figure 1. was generated from a corpus of thousands of documents published in Brazil in 2013 and it tells a clear story about the Pope visiting Brazil. The stories in the chain take place over a period of 11 days and fit well with the dominant theme of the corpus during that time period which focuses on social issues and protests. The second type is a measure of quality, which favors story chains that focus on a small number of stable topics, rather than many interleaved or shifting topics. We define a metric called topic persistence to capture how often the primary topic of a chain shifts, and topic consistency to capture the stability of the primary topic across each link in the story chain.

Our data set for this paper is a corpus of 7,074 English language news stories published in Brazil in 2013 which generated 5,606 chains with length 4 or greater from which we randomly select 60 story chains. These story chains are hand scored to indicate their clarity as a narrative, and used as an evaluation set to test our metrics. We apply an unsupervised topic modeling technique to the entire corpus and determine the main topics discussed in each story relative to this model. From this data we calculate the metrics described above and attempt to find correlations between them and story chains scored highly by our human scorers. We learned that based on our analytics alone, a data model can predict the likelihood that a given story chain presents a clear narrative. In this paper, we present background information, a detailed methodology, and the evaluation design, followed by results and conclusions of our analysis and future work.

2. Background

Storytelling as a data-mining concept was introduced by Kumar et.al. in [4]. Storytelling (or "connecting the dots") aims to relate seemingly disjoint objects by uncovering hidden or latent connections and finding a coherent intermediate chain of objects. This problem has been studied in a variety of contexts, such as entity networks [5], social networks [6], cellular networks [7], and document collections [1, 8, 9, 10]. Most existing approaches to storytelling [4, 8] use offline data where a user must specify the start and end documents of the chain, using an algorithm to uncover the sequence of relationships between the two points. This approach relies on building bipartite word-document or word cluster graphs, making it computationally expensive. The story chaining approach uses a real-time, flexible storytelling approach that can be used for streaming (online) data as well as for offline data. This is one of the first approaches to propose a streaming storytelling algorithm. Our models do not require an underlying bipartite graph and are thus computationally less expensive.

Because this is an unsupervised, automated process that generates many results, there is a need to identify the story chains that contain the clearest narratives. Shahriar et.al [1] uses context overlap as a measure to produce stories that stick to one context by extracting context sentences from a document using a Naive Bayes classifier. The authors, for assessing quality, also use dispersion plots and dispersion coefficient to evaluate the overlap of contents of the documents in a chain and thereby quality. Shahaf et.al. in [9, 10] define concepts of chain coherence, coverage, and connectivity that offer more insights into the storytelling process. Our approach differs in that it learns a topic model over the corpus and tries to associate certain types of topic change across a story chain as an indicator of how clear of a narrative structure is contained within a story chain.

Topic models are probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts [11]. They have been applied to a wide range of text to discover patterns of word use, or topics, across a corpus and to connect documents that share similar structure. In this way, topic models provide a way to create a structure from unstructured text in an unsupervised manner. We leverage them in our work primarily for this reason.

3. Methodology

In this section, we describe how our story chains are generated, how we define analytics to correlate changes in the topic distribution between consecutive stories in a chain to absolute characteristics of a chain, and how we learn a data model to predict clarity of the narrative structure for each chain. Our approach considers two data elements, a corpus of news stories and a set of story chains that have been generated from that corpus. From the corpus of news stories, we learn a topic model and assign topic probability vectors to each document in the corpus and measure changes in the topic probability vectors between consecutive stories in a chain to generate analytics.

A labeled training set is developed by having humans score a set of story chains based on the clarity of the narrative they convey. This training set is used to evaluate our analytics and to develop a data model to predict the anticipated value of unseen chains to an analyst.

3.1. Chaining methodology

The algorithm operates incrementally, where every new input article is analyzed as it arrives and is appended to a set of (partially constructed) chains being maintained. This analysis involves a two-step process. In the first step, we compare the incoming article to articles from the last ‘n’ days to identify the most similar articles and then designate candidate chains to attach the current article to. If no similar articles are found, then a new chain is created with this seed article. In our empirical analysis we found the value of n=14 to be most effective. Further, to assess if two documents are referring to the same underlying context, we calculate their similarity scores with respect to three features - textual features, spatial features (geographical coordinates, locations) and actors (person, organizations). The total similarity measure is a weighted sum of similarity scores of individual features. The weights for each individual feature similarity are set manually based on domain knowledge. The overall similarity score is structured as follows

$$sim(D_t, D_{t-k}) = \alpha \times f(D_t, D_{t-k}) + \beta \times f(Locations(D_t), Locations(D_{t-k})) + \\ \eta \times f(Actors(D_t), Actors(D_{t-k})) \tag{1}$$

Where F denotes a similarity metric such as cosine similarity or Jaccard's coefficient, and $\alpha + \beta + \eta = 1$. For our purposes we use cosine similarity. A term frequency–inverse document frequency (TF-IDF) representation is used in calculating the textual similarity between two articles. For extracting the location and actor feature vectors, we use the named entity recognizer from Basis Technology’s RLP suite[1]. We resolve each named entity identified as a location into a <country, state, and city> tuple with help of a geocoder based on probabilistic soft logic [12]. Details of this methodology are described in Muthiah et.al.’s work on planned protest detection [13].

Once a candidate set of chains are found, in the second step, the candidate set is pruned based on the coherence of the article with a story-chain. Here, coherence is calculated as the weighted sum of similarities of two types of frequency vectors - one for the actors and the other for geo-locations mentioned in all elements of the chain. The chain similarity is defined as

$$sim\left(D_t, C_i\right) = \theta \times g(Locations(D_t), Locations(C_i)) + \phi \times g(Actors(D_t), Actors(C_i)) \tag{2}$$

Where g is any similarity measure and $\theta + \phi = 1$. The article D_t is attached at the end of all chains such that $sim(D_t, C_i) \geq \Gamma$. Γ is a threshold parameter and can be used to tune the algorithm to control length and coherence of chains. A higher value of Γ will cause the chains to be shorter but more coherent, and vice versa. If no chain passes the similarity threshold Γ, then a new chain is created with this article. This two-step process is repeated for every new article.

In each iteration, if a story chain is updated with a new article link, the location and actor feature vectors of the chain are also updated. Each element in the these vectors represents the probability of occurrence of a location/actor in the chain and is calculated as following:

$$P(A_j, C_i) = \frac{\#\text{Frequency of } A_j \text{ in documents of } C_i}{\sum_k \text{frequency of } \mathrm{A}_k \text{in documents of } C_i} \tag{3}$$

[1] (2013). Rosette | Basis Technology. Retrieved April 1, 2015, from http://www.basistech.com/text-analytics/rosette/.

3.2. Topic modeling

We use the Machine Learning for Language Toolkit (MALLET) [14] to derive a topic model with 5 topics from the corpus of 7,074 English language news stories from which the story chains were generated. Each topic in the model is composed of a probabilistic vector of words that have been found to co-occur across the corpus. We then use Mallet to assign a topic vector to each document which describes probabilistically how represented each topic is within the document. For example, a story with the headline, "*Brazilians protest spending on the World Cup while poor suffer*", might have a topic vector across the five topics of [*Economic*=0.30; *Federal government*=0. 20; *The world*=0.05; *Internet/media*=0.01; *Police*=0.01] which indicates that the primary topic was *Economic* with *Federal government* as a close second.

We used the standard Mallet implementation of unsupervised topic modeling which requires us to specify the number of topics it should learn. We experimented with learning models with varying numbers of topics from 5 topics up to 20 topics. Our goal was to find a model with the most coherent topics that good coverage of the themes in our corpus. While there are some recent papers that attempt to evaluate coherence of topic models automatically [15], this was not the focus of our effort, so we manually reviewed the models generated and decided that the model with 5 topics provided a nice level of topic granularity. The five sets of topics we learned can be described as primarily representing *federal government*, *global issues*, *economic issues*, *Internet and media*, and *police action*.

3.3. Metric methodology

We calculate two types of measures: (1) *representativeness*, as an assessment of a story chain with respect to its relevance to explaining the corpus, and (2) quality, which consists of *topic persistence* to capture the number of topic changes that occur in the story chain and *topic consistency* to capture the continuity of the primary topic across the story chain. Representativeness quantifies how exemplary the stories within a story chain are of its associated corpus by measuring the similarity between the topics in the story chain to the sequence of topics characterizing the corpus during the same time period.

Considering a corpus of documents, let S_{t_i} be a subset of documents from that corpus such that t_i is a time point defined based on a specific story chain as described next. For an individual story chain, *c*, there is an associated set of time points representing the dates of publication of the documents in story chain *c*, namely, $t_i = \{ t_1, \ldots, t_n \}$. Also, even though *c* is not itself a set of stories, for convenience of notation, we will represent the number of stories that make up story chain *c* by ~~the~~ formalization |*c*|. Finally, every document has a topic vector associated with it represented by T such that $T_{S_{t_i}J}$ is the topic vector for the j^th document in subset S_{t_i} .

$$\text{Represenativeness} = \frac{\sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{\left|S_{t_i}\right|} \overline{\mathrm{T}_{S_{t_i}J}}}{\left|S_{t_i}\right|} - \frac{\sum_{j=1}^{C_{t_i}} \overline{T_{C_{t_i}J}}}{\left|C_{t_i}\right|} \right)}{n} \tag{4}$$

Topic Consistency (TC) is calculated below as the number of times the main topic of the entire chain is also the main topic of a story within that chain, divided by the total number of stories in that chain and is calculated as follows: ~~C~~onsider a topic model that identifies *m* topics over the corpus. For each document, *x*, in the corpus, a topic vector, T_x is produced consisting ~~of~~ *m* values, each assessing ~~h~~ow well represented topic *i* ε {1, …, m} is in document *x*. For each topic vector, T_x, consider the *m* values of T_x as a set *v*. The main topic, *main*, of a document is defined as follows:

$$main_x = \max_{1 \le i \le |v|} \{ v_i \} \tag{5}$$

where *i* is then the index of the topic. This *i* ε {1, …, *m*} value is then assigned as an index into the topic list (See section 3.2). We will identify the reference to the text-based topic associated with *i* as $mainTopic_x$ = i^{th} entry in the topic list.. Within a given story chain, *c*, we count the number of times a particular topic, *j* ε {1, …, *m*}, is the main topic of a document in the story chain, *c*.

$$count(c,j)=\sum_{k=1}^{|c|} f(c_k)=\begin{cases} 1, & j == mainTopic_{c_k} \\ 0, & otherwise \end{cases} \tag{6}$$

The main topic of story chain, *c*, is then calculated to be the most common main topic over all the documents in a story chain in equation 7. Topic Consistency is calculated in equation 8 as the number of times the main topic of the chain is also the main topic of a story within that chain, divided by the total number of stories in that chain.

$$main_c = \max_{1 \le j \le m} \left(count(c,j) \right) \tag{7}$$

$$TC = \frac{main_c}{|c|} \tag{8}$$

Topic Persistence (TP) is defined by how many times the topic of a chain persists across consecutive stories in a chain and is calculated as follows: For each document *x* in the chain *c*, let *mainTopic*$_x$ represents the main topic for *x* as described above. Then count the number of times that the main topic persists across connected documents in the chain and divide by the number of documents in the chain.

$$TP = \frac{\sum_{k=2}^{|c|} f(c_k) = \begin{cases} 1, \text{MainTopics}_k == \text{MainTopic}_{k-1} \\ 0, \qquad \text{Otherwise} \end{cases}}{|c|-1} \tag{9}$$

Consider a chain of 5 stories such that there are 4 connections between consecutive stories in the chain. Each story in the chain has a main topic. We represent this chain as:

$$A_1 \rightarrow B_2 \rightarrow C_3 \rightarrow D_4 \rightarrow D_5 \tag{10}$$

Each story in the chain is represented by X_i where X is the main topic of the story and *i* is the index of this specific story in the story chain. In this example, each story in the chain has a different primary topic, except the last two stories which share topic D, and there are 4 transitions. With this information, we can calculate TC = 0.25 because topic D remains the primary topic from story 4 to story 5, one of four transitions, and thus TC = 1/4. We can also calculate that TP = 0.40 because D is the primary topic for the chain since it is the most common topic, occurring in two of the five stories.

Compare these values to a story chain represented as

$$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow B_4 \rightarrow A_5 \tag{11}$$

Where we see two primary topic transitions, the TC score is 0.5 and four of the five stories have the main topic A, resulting in a TP score of 0.8. By our measures, chain (2) has higher value of topic persistence, indicating that the chain exhibits less fluctuation in main idea between connected stories, i.e., a more fluid narrative. Also note that chain (2) has a higher value for topic consistency indicating that the chain exhibits more of a narrative theme (namely, topic A) than chain (1).

3.4. Hand scoring

The randomly selected 60 story chains were each hand scored by three people and labeled as (1) very clear narrative; (2) somewhat clear narrative; (3) somewhat unclear narrative; (4) very unclear narrative. Chains with at least two of the three reviewers provided the same score were kept and that score was used during training. Chains in which all 3 reviewers provided different scores were removed from the dataset. This resulted in a gold standard evaluate set with 56 story chains. A more rigorous treatment of inter-coder reliability will be left to future work. We investigated how our three metrics of relevance, topic consistency, and topic persistence could distinguish among the levels of the Likert scale.

4. Evaluation

We develop linear regression models for each of our three analytics and as a group against our hand scored story chains in order to evaluate how well they correlate with the hand coded Likert values which we treat as numerical. Highly correlated analytics are good indicators of the clarity of our story chain's narrative.

We also experimented with treating our Likert scores as nominal classes and learned a neural network model using all three analytics as data features to classify story chains. This model was built using the Waikato Environment for Knowledge Analysis (WEKA) and 10-fold cross validation was used to split our dataset into training and testing sets. The Likert scores from our human reviewers were not uniformly distributed. Of the 56 chains, 38 received a majority score of 1, 5 received a majority score of 2, 7 received a score of 3 and 6 received a score of 4.

5. Results

The linear regression model for the relevance analytic is most correlated with the human scores for our chains, achieving a R^2 of 0.61 with a p-value well below 0.01. Linear regression models for our topic consistency analytic and our topic persistence analytic achieved a R^2 value of 0.13 with a p-value less than 0.01 and a R^2 value of 0.08 with a p-value of 0.03, respectively. The linear regression model we fit with all three analytics performed best with a R^2 value of 0.65 with a p-value well below 0.01, although only marginally better than the relevance analytic alone.

The neural network based model correctly classified 43 of 56 instances (76.78% correct). Overall precision for classifying across all four Likert scores was 0.697; however, there was significant variability in precision across classes. When classifying chains with the ground truth score 1, a precision of 0.905 was achieved. No chains with a ground truth score of 2 were scored correctly, resulting in a precision of 0.0. Chains with a ground truth score of 3 and 4 achieved precision of 0.375 and 0.333 respectively. While the model generates lower precision scores for chains with ground truth scores of 2, 3 and 4, there were no instances where a chain with a ground truth score of 3 or 4 was incorrectly scored as a chain of type 1 or 2, and inversely, there were no instances where a chain with a ground truth score of 1 or 2 was classified as a chain of type 3 or 4. Thus our approach was able to distinguish generally good chains (1-2) from generally bad chains (3-4) in all cases across the 56 evaluation chains, despite the drop off in performance in chains with ground truth scores of 2, 3, and 4.

6. Conclusions

The need to build situational awareness from increasingly large sets of textual data means we must have automatic methods to construct narrative structures from text without regard to domain factors such as actors, event types, etc. The metrics presented in this paper provide a means to assess these narrative structures so that only the most useful narrative structures are transformed into narratives. In this work, we define three metrics of relevance, topic persistence and topic consistency to assess narrative structure. We specify and implement these measures with respect to a narrative structure of story chains generated by an unsupervised narrative generation technique presented in [2]. This data is processed to provide analytical evidence for the usefulness of these metrics for identifying high quality story chains.

Our results indicate that using topic model based analytics to predict the quality of a narrative structure may be an interesting avenue of research. We found correlations between all of our analytics and the human scoring of our story chains, but the relevance metric was particularly correlated. Using our neural network based model, we were able to predict significantly better than chance, which of four levels of quality a narrative structure would likely be assigned if human evaluated.

7. Future work

This work is an assessment of initial concepts for measuring quality of narrative structures. Immediate future work will consist of further refining the analysis presented here, and using this analysis to better understand the strengths and weaknesses of the measures proposed. Upon eliciting this understanding, iterations on these metrics will be posed, as well as new measures whose need may be exposed by the analysis.

The measures defined in this work are generalizable measures for any narrative structure from which a narrative will ultimately be generated. We want to implement examples supporting this generalizability claim by translating and applying these measures to other narrative structures.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] Hossain, M Shahriar et al. "Connecting the dots between PubMed abstracts." PloS one 7.1 (2012): e29509.

[2] Van Brackle, D. “Improvements in the Jabari event coder”. 2nd International Conference on Cross-Cultural Decision Making: Focus 2012, 2012.

[3] E. Boschee, R. Weischedel, and A. Zamanian. 2005. Automatic information extraction. In Conference on Intelligence Analysis.

[4] Kumar, D., Ramakrishnan, N., Helm, R. F., & Potts, M. (2008). Algorithms for storytelling. Knowledge and Data Engineering, IEEE Transactions on, 20(6), 736-751.

[5] Hossain, M. S., Butler, P., Boedihardjo, A. P., & Ramakrishnan, N. (2012). Storytelling in entity networks to support intelligence analysts. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

[6] Faloutsos, Christos, Kevin S McCurley, and Andrew Tomkins. "Fast discovery of connection subgraphs." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 22 Aug. 2004: 118-127.

[7] Hossain, M Shahriar, Monika Akbar, and Nicholas F Polys. "Narratives in the network: interactive methods for mining cell signaling networks." Journal of Computational Biology 19.9 (2012): 1043-1059.

[8] Shahaf, Dafna, and Carlos Guestrin. "Connecting the dots between news articles." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining 25 Jul. 2010: 623-632.]

[9] Shahaf, Dafna, Carlos Guestrin, and Eric Horvitz. "Trains of thought: Generating information maps." Proceedings of the 21st international conference on World Wide Web 16 Apr. 2012: 899-908.

[10] Shahaf, Dafna et al. "Information cartography: creating zoomable, large-scale maps of information." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 11 Aug. 2013: 1097-1105.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[12] Kimmig, A., Bach, S., Broecheler, M., Huang, B., & Getoor, L. (2012). A short introduction to probabilistic soft logic. Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications.McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[13] Muthiah, S., Huang, B., Arredondo, J., Mares, D., Getoor, L., Katz, G., et al. (2015). Planned Protest Modeling in News and Social Media. Proceedings of the 27th IAAI conference on Innovative Applications of Artificial Intelligence.

[14] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[15] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In Human Language Technologies: The Annual Conference of the North American Chapte