Detecting Change in News Feeds Using a Context Based Graph

Lenin Mookiah, William Eberle
Department of Computer Science,
Tennessee Technological University,
Cookeville, TN, United States.

Maitrayi Mondal
Sunworks Consultants Private Limited,
Haryana, India.

***Abstract*— News feeds have been utilized as a resource for extracting media context, and in particular for the discovery of unusual information within common news articles. In this paper, we present our research of interesting unusual and useful patterns within the context of different social issues. We build a temporal context-based graph using news articles from different news channels built around actions employed by different entities, as well as resources involved with particular social issues such as accidental fires, kidnappings, human trafficking, road accident victims, mining accidents, ebola virus, swine flu, structure failures, senior citizen and juvenile incidences, migrant boat accidents, and slavery, with particular attention being paid to events such as agitation and the passage of legislation. From each article, we extract information such as the organization name, numbers, etc., and build a temporal graph of news articles. We then use sentiment analysis to measure the *sentiment* of each sentence, which are then used as edge values in our graph. For *context*, we utilize the graph structural connections among verbs, organization names, numbers, etc. For *content*, we use the new word count for each article. We propose a graph-cut model that leverages context, content, and sentiment information, empirically evaluate our proposed method, and present results that improve upon baseline methods in terms of precision, recall, F1, and accuracy.**

Keywords:- Change Detection, News Graph Mining

I. Introduction

Proliferation of news channels on the web has introduced a wide range of diverse data. These news articles actively report on stories involving crimes, terrorist attacks, and security issues relevant to the general population. Communities at risk deal with issues such as children forced into labor, senior citizens traveling in high-risk urban crime zones, and senior citizens not having access to public restrooms. Dark heterogeneous data sources such as news feeds, html, pdf files, and tables provide various statistical information and expert opinion analysis on these issues. For example, an article on child workers reported that "Half of the 5.5 million working children in India are concentrated in five states: Bihar, Uttar Pradesh, Rajasthan, Madhya Pradesh and Maharashtra"(Source:timesofindia.com, 13 Jun 2015). These types of web data provide a rich and complex set of information and knowledge on societal issues, such as policies proposed by the government or the implementation of a new law, that need to be extracted in a meaningful way for knowledge representation.

News mining has been studied in a variety of contexts. Earlier work involved grouping related news items, the fusion of news articles, and the summarization of information from disparate sources. However, news mining within a specific context could be more useful for a *targeted group of users* based upon their interests - what is commonly referred to as *personalized context mining*. Specific targeted groups could be based on age, location, gender, physical attributes, etc., or a variety of aspects based upon one's own interests. For instance, one could be interested in community challenges such as corruption, or perhaps safety in public places such as on a beach, road, or at a railway station.

First, we will define the various types of changes that our approach attempts to discover. Section III presents our definition of what constitutes a news article, or generally, a document. Section IV describes our process of data collection and preparation. Section V presents our proposed graph topology, and the tool used to extract the information, followed by Section VI that discusses the ground truth and evaluation methods for our experiments. Section VII presents our proposed method. Section VIII presents our experimental setup for our proposed method and baseline comparisons, followed by experimental results in Section IX. We then conclude with related work, conclusions, and future work.

Supported by NSF grant IIS-13185957.

II. Change Detected

First, we need to define what we mean by a document where there is "change detected".

Definition of Change Detected. A document (article) might contain one or more sentences in it that catch the attention of policy makers and/or groups interested in a particular social issue. In other words, the document includes one or more sets of information particularly useful to policy-makers and interest groups. In order for a document to be useful, the information should *have mention of a solution or intervention, an account of large resource damages, and/or contextual information explaining the issue*. We will mark articles containing such information as *change detected* articles. In other words, the article has "changed" from being just a typical news story.

Three Types of Change Detected. There are three types of *change detected* articles we are interested in from the perspective of interest groups:

- *Solution-based Change Detected (SBCD):* If an article contains a sentence that mentions an intervention or solution, then the article is marked as change detected. For example, precautions against potholes in order to avoid an accident:

TABLE I
FEATURES FROM A SAMPLE DOCUMENT (ARTICLE).

Name	Value
url	http://timesofindia.indiatimes/city/patna/one-kidnapped-in-vaishali-dist/articleshow/164404.cms
body	HAJIPUR: One Haribansh Rai of Mohanpur village of Vaishali district was kidnapped from his house on Tuesday night allegedly by the Sabal Rai gang. According to police sources, the kidnapping of Haribansh, 45, was due to an old enmity. He was kidnapped when he was fast asleep in his house. Sabal Rai's gang had created a reign of terror in the diara areas of Vaishali district. The gang is believed to be behind the killings of several truck drivers and cleaners.
date	2015/3/5

  - "Potholes rile commuters in the city. But after the Brihanmumbai Municipal Corporation (BMC) took cognizance of the menace and launched its pothole-tracking mobile application for people to complain about potholes."(Source: ndtv.com)

- *Context-based Change Detected (CBCD):* If an article is rich in context, such as getting attention from public and policy makers, then the article is marked as changed. For example, only a few global warming articles mention expert opinion for possible reasons behind the event, and when they do, information provided is very detailed. In other words, rich context articles provide asymmetric information content on the corresponding topic. For example:
  - "As the sea area freezes and melts each year, shrinking to its lowest extent ever recorded, Professor Peter Wadhams of Cambridge University called it a global disaster now unfolding in northern latitudes, the guardian reported." (Source: timesofindia.com)
- *Resource-damage-based Change Detected (RBCD):* If an article contains a number of resource losses, such as through injuries, deaths, or revenue, that are higher than expected, then the article is marked as changed. For example, buildings that collapse due to a conflict or an earthquake, usually result in a significant amount of property and lives lost:
  - "Around 1,300 people have been killed in Ukraine's Separatist conflict."(Source: ndtv.com)

III. DOCUMENT DEFINITION

A news article can be represented as a document, where D represents a collection of documents. Each document (article) is denoted by $d_i \in D$. Each document contains three features: *url*, *datetime* and *body*. Specifically, $d_i = \{url, body, datetime\}$, where url contains a web link to an article, $body$ refers to the *textual content* of the article, and $datetime$ contains the date and time of the article. Table I shows an example of features from a sample document.

The goal is that for each document $d_i \in D$, our algorithm will mark the document as either changed detected or not.

TABLE II
DATA STATISTICS FROM OUR DATA SET

Newspaper	Archive years	Total News Articles	Articles mentioning 12 societal issues
The Hindu	2009-2015 & 2000-2005	36715	358
Times of India	2009-2015	1726674	7527
NDTV	2009-2015	189976	2118

IV. DATA COLLECTION AND PREPARATION

The following is how we collected and prepared the data.

A. Data Collection

First, we crawled the index page of yearly archive pages from three Indian news papers - The Hindu[1], Times of India[2], and NDTV[3] - extracting the list of *url*s from the archives. The urls also contain the title of the article embedded in it, as shown in Table I. We then used grep on each url to filter news articles based on 12 different societal issues using the following keywords: fire, traffic, kidnap, senior citizen, juvenile, mining, ebola, swine, migrant, slavery, collapse, and road accident. It should be noted that while this particular list of keywords was chosen somewhat arbitrarily, it was based upon feedback from experts in India who deemed these particular issues of the most importance. We did not use any lexical expansion on our keywords search filtering, but we did filter out irrelevant articles, which reduced our experimental data set down to 8,433 news articles. Data statistics is shown in Table. II.

B. Data Preparation

Second, we analyzed the news articles related to the 12 societal issues for possible changes. Articles that report uncommon incidents within the context of social issues and policy, are marked for change detection. Each article was read by human annotators and searched for one or more of the change detected types defined previously: solution-based, context-based, and/or resource-based. If the article provides a *solution* or intervention to a social problem, such as the discovery of the ebola virus, it is marked as change detected because of the solution-based impact. If an article mentions experts' opinions, such as expert opinion on the ebola outbreak, that is considered a *context-based* change. If the article mentions huge resource losses, such as mass human casualties due to ebola, it would be considered a *resource-based* change, and would be marked as change detected.

We also discovered a few out of context (noise) articles. For example, an article mentioning a "cease fire" is inappropriate for a fire accident, and thus is removed from consideration. In addition, a few articles might contain information appropriate for one or more of the different change detected types. For example, the article at [12] is marked for change detection based on both resource-based and solution-based impact because 30 people died in the stampede making it eligible for being considered as a resource-based change, and a process was initiated to correct the root cause of the stampede, thereby also providing a solution.

V. GRAPH TOPOLOGY

Input for our approach is a graph. An example of our proposed graph topology of news articles is shown in Figure 1. In order to create this graph, we used openNLP[4] to

[1] $www.thehindu.com$
[2] $www.timesofindia.com$
[3] $www.ndtv.com$
[4] www.opennlp.apache.org

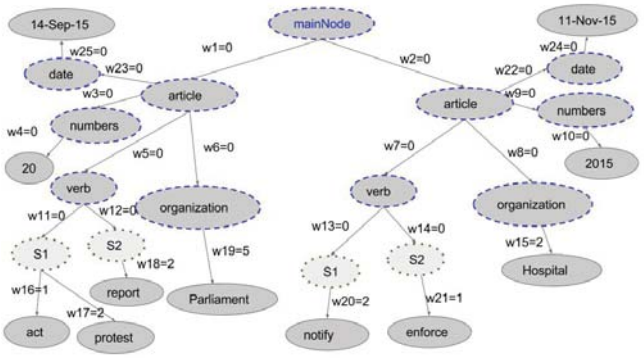

Fig. 1. Example graph topology of news articles.

extract the information doing part-of-speech tagging. For example, if a given sentence is "Cops arrested the murderer in the park", openNLP tags it as "Cops/NNS arrested/VBN the/DT murderer/NN in/IN the/DT park/NN". From this tagged sentence, we extract verbs, numbers, etc., as shown in Figure 1. We also used the Stanford NLP tool [5] to extract organizations as nodes, as shown in Figure 1 , where dashed and dotted lines represent types, such as "organization" and "verb", and nodes with solid lines represent actual values from the articles. For each article, we created an "article" node under a "mainNode" node (used solely for ease of retrieval). Under an "article" node, we created hard-coded number, verb, and organization nodes. For "verb", we further created hard-coded (dotted) nodes named $S1, S2, ..Sn$ where *S1* represents the first sentence, *S2* represents the second sentence, and so on. In other words, we create a tree-graph for each article. The extracted verb, numbers, and organization names, are attached as child nodes under their corresponding articles. The child nodes from one article can then be linked with other child nodes of other articles.

Graph weight. We use the Stanford Sentiment Analysis library[6] to mark the sentiment of each sentence in a news article. For each sentence, Stanford sentiment analysis predicts the following sentiment values: 1-very negative, 2-negative, 3-neutral, 4-positive, and 5-very positive. These values then represent edge weights in our graph. For all other edge relationships, we set it to *zero*. In Figure 1,w15, w16, w17, w18, w20, and w21 contain non-zero sentiment values. For all others that are zero, which we just represent as edge labels. The result is a weighted graph with vertices consisting of verbs, organization names, dates, and numbers. The resulting graph consists of *117,343* vertices and *231,142* edges.

VI. Evaluation

For the ground truth needed for evaluating our approach, each article is examined in detail by two policy making experts working for think-tanks. Annotator 1 is employed by *SunWorks Consultant Private Limited*, leading a team examining news article publications related to Chinese social polices, government activities, foreign policies, china neighborhood relations, and the economy. Annotator 2 also works in the area of Chinese affairs studies, employed at the *Institute of Chinese Studies (ICS)*. ICS is funded by *Ministry of External Affairs, Government of India*. ICS promotes interdisciplinary studies and research on China and the rest of East Asia with a focus on domestic politics, international relations, economy, history, health, education, border studies, language and culture. They crawled documents related to the three different change detected documents described previously, using the approaches described in the data preparation section, marking each article as change detected or not. It is important to note that only articles where both annotators would agree were marked as change detected. If there was a disagreement in terms of the context of a specific article, the article was removed from the data set. This resulted in 74 articles (out of the original 8,433) being marked as irrelevant.

In order to evaluate our approach, we use recall, F1-score, and accuracy, compared against existing standard approaches.

VII. Our Proposed Method

In this section, we propose an algorithm that uses an objective function based upon a greedy Graph-Cut approach. First, we present the important aspects relevant to change detected documents. Articles that are considered as appropriate for detecting this type of change involve one or more of the following aspects:

User-based Impact. An user is a named entity that is an organization, as marked by part-of-speech tagging (as discussed earlier). For example, articles of interest may mention an organization such as the World Health Organization (WHO). We then need to capture attribute information about the mentioned entity (user) particularly organization using *Stanford NLP* library. For example, the following sentence can be parsed to recognize that the entity: "The benefits of globalisation can be directed to reduce rural poverty if national and international economic policies take into account its effect on agriculture, according to Joachim von Braun, director-general, International Food Policy Research Institute (IFPRI)." (Source: thehindu.com, 23 Aug 2007). In this work, an *expert* is defined as the organization that is extracted using the *Stanford NLP* tool.

Action-based Impact. We are interested in capturing verbs that imply changes such as a new law being proposed or implemented. For example: "Delhi government clears Jan Lokpal Bill". The basic idea is to capture interesting verbs, such as "clears" or "passes", in the context of change. In this work, we try to leverage such verbs using graph structure.

Resource-based Impact. In the case of resource damages, lives lost (or hurt), or revenue lost, we need to capture attribute information about the resource affected. For example: "It was said at the time that over 6,000 houses were burnt." So, in this example, we want to capture the value 6000 in the context of houses that were destroyed.

In short, in order to classify an article as change detected or not, our method has to effectively capture and leverage contextual information that mentions user-based information (e.g., organization), action-based verbs (e.g., protest, strike),

[5] www.nlp.stanford.edu/
[6] www.nlp.stanford.edu/sentiment/

and resource-based information (e.g., "6000").

Graph G In this work, news articles are represented as a Graph G. Each verb, organization, and number, are represented as nodes. Edges connect nodes of the same value (verb, name, number) between news articles. D represents the set of articles (documents) in our data set. D_N represents the total number of documents in our data set. The definition of a news article is provided below.

Definition 1. An article. An article is represented as $d_i \in D$. Document d_i has N preceding documents by *datetime*. Each of the neighbors are represented by $d_j \in D$. D_N represents the total number of articles. A document d_i is defined as $d_i = \{d_i^{vb}, d_i^{org}, w_i^{neg}, N_{num}, N_{yr}, d_i^{neig}\}$. d_i^{vb} represents the number of common (action) verbs the article shares with all other articles in the graph G. d_i^{org} represents the number of organization names (tokens) mentioned. w_i^{neg} represents the number of negative sentiment sentences, N_{num} represents the number of times a number is mentioned (excluding numbers that represent a year), N_{yr} represents the number of times a year is mentioned. $d_i^{neig} = \{d_1, .., d_j\}$ represents the set of neighboring (preceding N) articles. N represents the number of preceding articles (neighbors) we wish to compare. We discover that a value of 5 gives an increased F1 score as shown in Table III, and is subsequently used as the minimum neighbor in our experiments.

Definition 2. Smoothing Function. In order to overcome noise while fitting the data for our model, we implement a smoothing function. We examined several smoothing functions, but most were sensitive to outliers such as zero. Thus, we ended up implementing equation 1. In particular, we require smoothing for features such as d_i^{vb} and d_i^{org}. We use fn_i^{vb} and fn_i^{org} as mentioned in equation 1 for d_i^{vb} and d_i^{org} respectively. First, this smoothing function provides smoothed values by converting outliers such as $d_i^{vb} = 0$, $d_i^{org} = 0$ to 1. Second, values are smoothed to be in the range [0, 1]. Precision and recall of the graph-cut with neighbor $N = 5$ are 0.4249 and 0.1675 respectively if we use d_i^{vb} and d_i^{org} instead of fn_i^{vb} and fn_i^{org} in our objective function. There is a decline in recall of 0.1675 from 0.4019 due to smoothing effect on new values that include outliers.

$$\begin{aligned} fn_i^{vb} &= fn(d_i^{vb}) = 1/(1 + (d_i^{vb})^2) \\ fn_i^{org} &= fn(d_i^{org}) = 1/(1 + (d_i^{org})^2) \end{aligned} \tag{1}$$

$$CC = \Big\{ \overbrace{\sqrt{fn_i^{vb}.fn_i^{org}} * w_i^{neg}}^{\text{A}} - \beta \overbrace{\frac{\sum_{j=1}^{N} \sqrt{fn_j^{vb}.fn_j^{org}} * w_j^{neg}}{N}}^{\text{B}} \Big\} * \underbrace{\{ d_i^{nword} - \frac{\sum_{j=1}^{N} d_j^{nword}}{D_N} \}}_{\text{C}} \tag{2}$$

$$\beta = \alpha * \{ \sqrt{N_i^{num} * N_i^{yr}} - \underbrace{\frac{\sum_{j=1}^{N} \sqrt{N_j^{num} * N_j^{yr}}}{N}}_{\text{E}} \} \tag{3}$$

Objective Function. The following discusses the context (graph structure) information, content information, and cut cost (CC) used in our proposed greedy approach. For content and context information, we remove the stop words and perform word-stemming. Our approach greedily calculates the minimum graph-cut in Equation 2 in comparison with its neighbors. We leverage two types of information from each article: Context and Content. Context leverages information such as verb, numbers, and sentiments using graph structure. Content leverages information such as new words and similarity metrics. Our algorithm is based on the strength of context and content of an article *with its neighbor*.

Context. Contextual information involves the mention of common action verbs (social context) and popular social organizations across multiple articles, thus forming edges, or links, across documents. For example, two or more articles might be connected via the node of a common verb such as "act", "notify", or "announce". These common (action) verbs might occur in phrases such as "Law passed" or "New traffic rules announced", indicating a change (per our definition). Similarly, organization names and numbers, such as the year in one document, can form linkages to other documents. For instance, the more numbers are mentioned in an article, the more likely it has a statistical significance. In addition, more numbers in an article likely indicates that resource-based changes are contained in the article. For example: "2100 died because of ebola in year 2015 alone". In short, we use organization names, numbers, and common action verb linkages, tagged using the *openNLP* library and *Stanford NLP* tool, for the contextual (structural) information in the graph.

Content. Content information that carries *new words* is considered to be carrying new information. For example: "New mobile application has been introduced for senior citizens.". Equation 3 is based upon new words. For each document d_i, we calculate the count of new words that have not occurred in *N* preceding documents, which we call *neighbors*. Thus, in terms of context and content, our first intuition is that an article with considerable changes will also mention the organization name, common (action) verbs, numbers, etc. Our second intuition is that this article will have a fewer number of negative sentiments than its neighbors. This is due to fact that these change detected articles might have information such as solutions proposed, statistical information, and guidelines mentioned by experts - all (presumably) positive information to the reader. In other words, interesting change detected articles might have *fewer* sentences with negative sentiments. In other words, an interesting change detected article will have fewer w_i, and is more than likely mentioning numbers, organization names, and action verbs, as compared to its neighbors.

Cut Cost. For each article, we calculate Cut Cost (CC), as

defined in Equation 2. We first calculate the geometric mean of the common verb function fn_i^{vb} and the organization function fn_i^{org} for a given document d_i, which is marked as *Part A* in equation 2. Then, the geometric mean is compared to the average of its neighbors, which is marked as *B* in equation 2. Additionally, the count of the number of negative sentences in the article w_i^{neg} is multiplied against the geometric mean. Similarly, for a given document d_i, we calculate a d_i^{nword} count of new words that have not occurred in the past *N* documents. *Part C* of equation 2 shows where we capture the difference of the calculated new words of d_i with the average number of new words of all documents in the data set. If the cut cost is greater than zero, we mark the article as change detected. Again, the basic intuition is that a change detected article will have mix of statistics, less negative sentiments, and organization names that act as change agents in comparison to their neighbors.

Penalty For articles that contain no relevant information on context, our objective function needs to over-penalize them on the cut cost. To do this, we introduce penalty β as shown in Equation 3. β captures the difference between the geometric mean of the count of numbers and the count of years, against the average of its neighbors. *Part E* of Equation 3 represents the average of its neighbors' geometric mean, count of numbers, and count of years. Parameter α controls the degree of penalization. In our experiments, we evaluated $alpha$ from 0.1 to 40 and found the best $F1$ to be when $\alpha = 0.95$. Figure 2 shows the effect of α on precision, recall, and F1. After $\alpha = 2$, precision, recall and F1 do not improve, and actually maintain the same percentage.

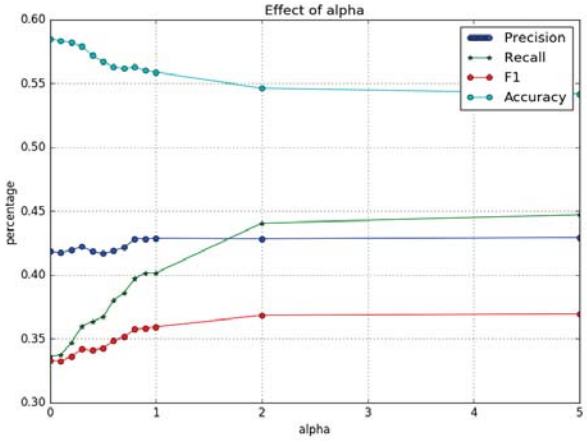

Fig. 2. Effect of penalty parameter alpha.

VIII. Experiment

The four comparison methods we use are cosine similarity, new word count, new word count with threshold, and jaccard coefficient. In our comparison methods, as mentioned earlier, for all *N* preceding document comparisons we use a value of 5 for *N*. All experiments are run on a Mac with 2.8 GHz Intel Core i7, with 16GB of memory, and a 1600 MHz DDR3.

A. Our proposed Method.

We implement our proposed method as an iterative algorithm, as shown in Algorithm 1. The algorithm iterates over one document d_i at a time. It gets all the child nodes such as date, number, verb, and organization under the "article" node of d_i, then calculates the cut cost for d_i as per Equation. 2. If cut cost is greater than 0, it classifies it as "change detected".

Algorithm 1 Graph-Cut Algorithm

```
Input: D documents of our data set.
alpha = 0.9
procedure GRAPH–CUT–ALGORITHM
    for each article d_i ∈ D get "article" node do
        Get all child nodes under "article" node of d_i
        Calculate cut cost CC for d_i as per Equ. 2
        IF calculated CC > 0
            classify d_i as "change detected"
        ELSE classify d_i as "NOT change detected"
    end for
end procedure
```

B. Comparison Methods

Cosine similarity [13] and *Jaccard Coefficient* [14] are the two most popularly used methods. These methods have also been applied to problems such as novelty detection [4]. We use them to calculate TF-IDF similarity between 2 articles. Another approach involves calculating the *new words* count when comparing a document to past documents, which can be used to discover an uncommon document. We also chose this technique as one of our baselines as it has been used repeatedly in other related research [6][7][9]. For our experiments, the articles from all 3 news papers listed in Table II are merged and sorted chronologically. Since several articles from *thehindu* newspaper have missing dates, we only use the month and year for the chronological ordering of news articles. Our basic intuition is that rich (uncommon) contextual documents will have less similarity with their recent past and future documents. For each of the baseline methods, D^q represents a set of documents chronologically ordered containing a topic query q . We prepared a TF-IDF for each document $d_i^q \in D_q$. TF-IDF methods are often used as a fast and effective means of comparison with similarity based methods such as cosine similarity. We also removed stop words and performed word-stemming.

Cosine Similarity For a query q, we iterate through each individual document d_i^q and calculate the cosine similarity for each of the N preceding documents. Then, we average the pair of documents' calculated cosine similarity. The Top n% of documents with the least cosine similarity are marked as change detected. As discussed under *Definition 1*, for *N* preceding documents in the range [1-4]. Thus, we experimented with values of 5, 10, and 15, and found the best precision, recall, and F1-score using $N = 5$.

New Word Count. For a query q, we iterate through each individual document d_i^q and count the new words that have not occurred in each of the *N* preceding documents. Then, we average the new word count for all pairs of documents. The Top n% of documents with the highest new word count are marked as change detected.

New Word Count with Threshold. This method is similar to the *new word count* approach. In this case, we provide a threshold of percentage difference between new words found in an individual document d_i^q with the average new word count of *N* preceding documents. The Top n% of documents with the highest percentage difference are marked as change detected. We experimented with threshold values between 80 and 90, and discovered the best precision using a $threshold = 90$.

TABLE III

RESULTS SHOWING PRECISION (PREC), RECALL, F1-SCORE, AND ACCURACY (ACC) FOR BASELINES METHODS AND OUR GRAPH-CUT APPROACH. FOR BASELINE METHODS, THE TOP 20% OF ARTICLES ARE MARKED AS CHANGE DETECTED. DIFFERENT VALUES FOR *N* REPRESENTING NUMBER OF PRECEDING DOCUMENT (NEIGHBORS) EXPERIMENTS AND RESULTS SHOWN.

Top n%	N	Method	Prec	Recall	F1-score	Acc
20%	5	Cosine	0.3462	0.1869	0.2064	0.5489
20%	5	New word count	0.4207	0.1799	0.2363	0.58
20%	5	New word count-Threshold	0.2363	0.2935	0.1778	0.6414
20%	5	Jaccard	0.355	0.1768	0.2112	0.6648
-	5	Graph-Cut	0.4283	0.4019	0.3583	0.5589
20%	10	Cosine	0.3363	0.1831	0.1979	0.5451
20%	10	New word count	0.429	0.1857	0.2426	0.5832
20%	10	New word count-Threshold	0.2363	0.2935	0.1778	0.6414
20%	10	Jaccard	0.3673	0.1824	0.2173	0.6672
-	10	Graph-Cut	0.4266	0.4	0.3565	0.5576
20%	15	Cosine	0.3231	0.1767	0.1912	0.5399
20%	15	New word count	0.4256	0.1842	0.2407	0.5818
20%	15	New word count-Threshold	0.2363	0.2935	0.1778	0.6414
20%	15	Jaccard	0.3802	0.188	0.2246	0.6698
-	15	Graph-Cut	0.4289	0.4055	0.3611	0.5611

Jaccard Coefficient. For a query q, we iterate through each individual document d_i^q and calculate the jaccard coefficient for each of the *N* preceding documents. Then, we average the pair of documents' calculated jaccard coefficients. The Top n% of documents with the lowest jaccard coefficients are marked as change detected.

In all above methods, we iterate through each document from a chronologically sorted set. We usually take *N* preceding documents. However, for the first few documents, there might not be any preceding documents. Hence, we take either the *N* succeeding documents (if available), or a combination of preceding and succeeding documents. For example, when iterating on the 6^{th} document, we need a total of *N*=15 preceding documents, where as we got only 5 preceding ones. In this case, we then include the succeeding 10 documents.

TABLE IV

RESULTS SHOWING EFFECT OF DIFFERENT *Top n %* VALUES MARKED AS CHANGE DETECTED ON PERFORMANCE OF BASELINES.

Top n%	N	Method	Prec	Recall	F1-score	Acc
10%	5	Cosine	0.3070	0.0842	0.1102	0.5727
10%	5	New word count	0.4692	0.0923	0.1471	0.6040
10%	5	New word count-Threshold	0.2021	0.1640	0.1169	0.6230
10%	5	Jaccard	0.4042	0.0933	0.1384	0.6420
15%	5	Cosine	0.3445	0.1322	0.1566	0.5605
15%	5	New word count	0.4729	0.1371	0.1974	0.5971
15%	5	New word count-Threshold	0.1839	0.2240	0.1349	0.6262
15%	5	Jaccard	0.4103	0.1357	0.1810	0.6540

IX. RESULTS AND DISCUSSION

Table III shows our experimental results, comparing baseline methods to our proposed Graph-Cut approach. We first compare and discuss results when dealing with the Top 20% and $N = 5$. The accuracy of the baseline approaches, except *cosine similarity*, are noticeably better than our Graph-Cut approach. However, Graph-Cut gives the overall better precision (*new word count* comes close), recall, and F1-score. We also varied values of *N* preceding (neighbor) documents, with results for *N* values of 10 and 15 shown in Table III. The results are similar to that of *N* with value 5, albeit the *new word count* approach again is close or even slightly better in terms of precision, but not better when it comes to recall or F1-score.

In addition, we evaluated the baseline methods with different values for the *Top n %* other than the *Top 20%* marked for change detection. It is worth to note that the percentage of change detected articles in our data set is approximately 14%, so we experimented with values of the Top 10% and Top 15% being marked as change detected. We include Table IV to further show that results are only impacted by different values of n used for the *Top n %* marked as change detected documents. For the Top 15 %, performance reduces in precision, recall, and F1, and for the Top 10% performances is even lower. In Fig. 2, graph-cut achieves the best $F1$ when $\alpha = 0.95$. After $\alpha = 2$, all of our measures flatten out. For Top 15% and N=5, student's t-test of our 4 evaluation metrics of New word count with Threshold and Graph-Cut shows significance only at 14 %. However, Graph-Cut has required better F1 and accuracy.

It should also be noted that the running time of the baselines algorithms ranges anywhere from 5-10 seconds. In comparison, our Graph-Cut method takes approximately 1 second to complete.

X. RELATED WORK

Our work is most closely related to the research that is being done on novelty detection, particularly in a temporal setting. Gaughan and Smeaton [3] study novelty detection

using the TREC data set. The NIST TREC[7] data is from 2002-2005, and includes tracks that are divided into event and opinion topics. For example, they use the 2004 data, which uses 3 news feeds from Xinhua, the New York Times and the Associated Press. In their study of novelty detection, they employ a Term Frequency-Inverse Document Frequency (TD-IDF) variant. The authors use *F1-score* for evaluation, and achieve an F-score of 0.622 and 0.807 on the 2004 and 2003 data respectively. Li et al [6] [7] also study the novelty detection problem using TREC. First, their algorithm converts the query into a query and its expected related answer type. The basic idea is that if there is a combination of query words, named entities, etc., available in a sentence, this increases the possibility of answer. The approach uses a concept called "answer patterns". Answer patterns are a list of answer candidates, each with a specific pattern prepared for each question using a belief or heuristic .

Schiffman et al. [8] leverage contextual information for the novelty detection problem. The authors use the context of the sentence along with novel words and named entities. The algorithm tries to find the optimal value for 11 parameters, weights, and thresholds. The algorithm uses a random hill-climbing algorithm with backtracking for learning weights. The algorithm achieves a recall of 0.86 on the average of all runs, in comparison to cosine similarity with 0.81. Karkali et al. [4] study the problem of online novelty detection on news streams. Their work uses two data sets, one from the Google news RSS feed and another from Twitter. Novelty is defined in terms of a predefined window on the past. The proposed algorithm is based on the TF-IDF , and is evaluated using a linearly combined single detection cost [5].

Our work differs from previous efforts as our proposed method is a graph-based approach. Our graph-based approach works better because of advantages in leveraging (1) structure, (2) content information, and (3) context information. We use an aspect-level and fine-grained approach of individually extracting and using dates, numbers, organization, and (4) sentiments in graph. Also, we are currently focused on discovering news articles relevant to policy makers - a problem for those who are tasked with generating relevant societal policies.

XI. Conclusion

In this study, we collected data from 3 different news sources. We extracted common verbs, organization, and numbers, and built a weighted graph using sentiments of each sentence in news articles. We proposed a greedy graph cut algorithm that outperforms baselines in precision, recall, and F1. We also study the penalty parameter α and report the impact on our evaluation metrics. In the future, we will investigate using an external ontology for policy making such that it could help to better leverage the context (structural informal). For example, in wikipedia[8] the "Union Council of Ministers of India" provides different department names within the government. This might enable us to better capture entities such as organization name. In addition, we will examine augmenting the graph with certain important features such as the designation of people names. For example: "Forensic science officials will also be called upon to examine the building quality". One challenge here is to effectively extract the designation (i.e., "forensic science officials"). Due to an overwhelming number of common nouns, these designations (nouns) become underrepresented and hence are not used effectively. Also, we are currently working on extending our temporal graph to graph streaming approaches, thereby exploring related streaming techniques.

[7] www.trec.nist.gov

[8] https://www.wikipedia.org/

Acknowledgment

We sincerely thank Jayshree Borah from the China Studies Centre, Indian Institute of Technology Madras for helping in labelling the articles, and providing useful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 1318957.

References

[1] Yu W, Aggarwal CC, Ma S, Wang H. *On anomalous hotspot discovery in graph streams*, In Data Mining (ICDM), IEEE 13th International Conference (2013), pp. 1271-1276.

[2] Arackaparambil C, Yan G. *Wiki-watchdog: Anomaly detection in Wikipedia through a distributional lens*, In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, (2011), pp. 257-264.

[3] Gaughan G, Smeaton AF. *Finding new news: Novelty detection in broadcast news*, InInformation Retrieval Technology (2005), pp. 583-588.

[4] Karkali M, Rousseau F, Ntoulas A, Vazirgiannis M. *Efficient online novelty detection in news streams*, In Web Information Systems Engineering-WISE (2013), pp. 57-71.

[5] Manmatha R, Feng A, Allan J. *A critical examination of TDT's cost function*, In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002), pp. 403-404).

[6] Li X, Croft WB. *Novelty detection based on sentence level patterns*, In Proceedings of the 14th ACM international conference on Information and knowledge management (2005), pp. 744-751.

[7] Li X, Croft WB. *Improving novelty detection for general topics using sentence level information patterns*, In Proceedings of the 15th ACM international conference on Information and knowledge management (2006), pp. 238-247.

[8] Schiffman B, McKeown KR. *Context and learning in novelty detection*, In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (2005) pp. 716-723.

[9] Li X, Croft WB. *An information-pattern-based approach to novelty detection*, Information Processing & Management, 44.3, (2008), pp. 1159-1188.

[10] Shahaf D, Guestrin C. *Connecting the dots between news articles*, In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, (2010), pp. 623-632.

[11] A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*, New York: Academic Press, 1977.

[12] http://ndtv.com/allahabad-news/allahabad-stampede-not-due-to-railing-collapse-railway-minister-pawan-kumar-bansal-512954

[13] Blank D. *Resource Description and Selection for Similarity Search in Metric Spaces. University of Bamberg Press; 2015 May 7.*

[14] Goodman LA, Thompson KM, Weinfurt K, Corl S, Acker P, Mueser KT, Rosenberg SD. *Reliability of reports of violent victimization and posttraumatic stress disorder among men and women with serious mental illness. Journal of traumatic stress. 1999 Oct 1;12(4):587-99.*