Tracking and analyzing dynamics of news-cycles during global pandemics: a historical perspective

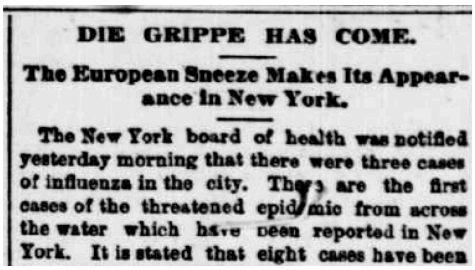
Sorour E. Amiri*, Anika Tabassum*, E. Thomas Ewing+, B. Aditya Prakash*
*Department of Computer Science, Virginia Tech
+Department of History, Virginia Tech
Email: {esorour, anikat1, etewing, badityap}@vt.edu

ABSTRACT

How does the tone of reporting during a disease outbreak change in relation to the number of cases, categories of victims, and accumulating deaths? How do newspapers and medical journals contribute to the narrative of a historical pandemic? Can data mining experts help history scholars to scale up the process of examining articles, extracting new insights and understanding the public opinion of a pandemic? We explore these problems in this paper, using the 19th-century Russian Flu epidemic as an example. We study two different types of historical data sources: the US medical discussion and popular reporting during the epidemic, from its outbreak in late 1889 through the successive waves that lasted through 1893. We analyze and compare these articles and reports to answer three major questions. First, we analyze how newspapers and medical journals report the Russian flu and describe the situation. Next, we help historians in understanding the tone of related reports and how they vary across data sources. We also examine the temporal changes in the discussion to get an in-depth understanding of how public opinion changed about the pandemic. Finally, we aggregate all of the algorithms in an easy to use framework GrippeStory to help history scholars investigate historical pandemic data in general, across chronological periods and locations. Our extensive experiments and analysis on a large number of historical articles show that GrippeStory gives meaningful and useful results for historians and it outperforms the baselines.

1. INTRODUCTION

Tracking a piece of information such as an idea or memes on a network is an important task with many interesting applications such as in marketing [32], opinion formation [24], and anomaly/event detection [30]. In the context of the humanities, for a history scholar, the study of information transmission during a global epidemic across different media is similarly very compelling [9]. For instance, such a study makes it possible to compare news reporting across time, by tracking the day to day reporting in newspapers while also examining the tone of reporting as evidence of how participants in this historical event understood a disease. Furthermore, such a study makes it possible to detect influential newspapers or journals and interpret the ways that an epidemic reporting may have affected behavior, attitudes, and beliefs.



DIE GRIPPE HAS COME.

The European Sneeze Makes Its Appearance in New York.

The New York board of health was notified yesterday morning that there were three cases of influenza in the city. They are the first cases of the threatened epidemic from across the water which have been reported in New York. It is stated that eight cases have been

Figure 1: **A digital image snippet from the Evening Star newspaper, December 17, 1889, p. 9 reporting on the Russian Flu.**

Despite its importance, not much work has looked into news-cycles in historical newspapers or scientific articles during global historical events of intense interest. In contrast, much work has gone into studying the dynamics of information diffusion in the context of online media such as blogs, Twitter and Facebook [3]. As a result, most works in humanities follow a 'traditional' approach and do a manual analysis. Hence in this paper, a collaboration between computer scientists and historians, we highlight an attractive application for data miners in digital humanities, discuss the important challenges in this topic and design a tool GrippeStory for historical analysis and digital humanities which are particularly appropriate for dealing with these issues.

1.1 The Russian Flu Pandemic

We ground the rest of the article using the 1889-90 Russian flu epidemic, as it is an especially appropriate example for an approach that integrates the digital humanities and computational analysis. The Russian flu spread across Europe and then the US in late 1889 and early 1890, causing widespread illness and prompting extensive reporting about the disease globally as well as locally. With the establishment of the global telegraph network, for the first time in world history [34], news about a disease could spread across long distances faster than the disease itself, which was limited by the speed of human travel. In this context of relative international calm, at least among the great powers, transnational communication, was facilitated by both the increased speed of electronic communications and a shared perception of the advantages of sharing scholarly insights. Popular daily newspapers and professional medical journals reported extensively on this epidemic in terms of common symptoms, the extent of illness, and the potential threat to public health (See Fig. 1 as an example). At this same time, medical discoveries were transforming both scholarly and

public opinion about disease origins, transmission, and prevention. Finally, the Russian flu is an excellent case because although it had a relatively low mortality rate, it spread quickly and infected high proportions of the population in each region it reached, thus allowing for mapping of the spread of disease using popular and medical reporting. Using tools from data mining, we can digitize and study these sources, to better understand this pandemic.

Newspapers and medical journals in this era can be interpreted using tools now used to analyze social media, although with important qualifications and reservations. Newspapers were intended to transmit information in a timely manner to wide audiences within a specified range, thus serving similar functions to the ways that social media now serves to transmit information quickly. Newspapers often reprinted information from other newspapers, which makes it possible to track how news spread across time and space, while also allowing for consideration of local responses to global or national developments. Medical journals in this era were read primarily by specialists, but just as medical organizations now use social media to convey important and authoritative information, journals in the 1890s also used newspapers to both gather and disseminate information. The most important dissimilarities, of course, were the speed of information transmission and the fact that social media is driven by user-created content, which appeared in newspapers only indirectly in the form of letters or paid advertisements.

1.2 Challenges

To study historical news cycles, particularly the Russian flu, effectively and efficiently we face several challenges: (1) Working with digitized historical data is challenging due to errors such as missing data and misspelled words. So it needs an extensive data cleaning step. (2) The data mining tasks have to be designed such that they are meaningful to historians, e.g. current news-cycles are far more compressed, while historical ones are elongated and slow-moving. As a result, there is a greater need for subtle analysis of tones. Also, it makes it difficult to detect the changes in tone using the current popular algorithms. (3) Need to handle multiple kinds of data sources and compare their dynamics: in our case, newspapers and medical journals. So, the challenge of interpreting just one newspaper for just one month is compounded by considering dozens of titles across several years, from different genres, such as medical periodicals and newspapers. (4) The framework needs to be extensible and scalable as we need to handle a large body of text data (of around ten years) of daily newspapers and medical journals.

1.3 Contributions

Our work allows for the following distinct contributions:

– *New applications for data miners.* We highlight an attractive application for data scientists in digital humanities, discuss some important challenges in this topic and design and develop an effective and efficient tool GrippeStory. It visualizes historical data, identifies the tone of news articles about the disease and its changes over time to analyze the different stages of public opinion about the epidemic.

– *Enhance research on public health.* We bring attention to the Russian Flu pandemic era, where we had one of the earliest epidemics which was captured by media globally. We connect computational epidemiology with the digital humanities around issues of scope, severity, and impact of Russian Flu pandemic. We contribute to the historical and epidemiological understanding of a major flu outbreak by integrating resources from multiple digital platforms of popular newspapers and medical journals.

– *Scaling up humanities analysis.* While the process of close reading of source materials is an analytical method familiar to humanities scholars, we scale up this analysis to look at a much larger collection of textual data using data mining approaches.

The rest of the paper is organized as follows: we first give related work and our focus questions from a historical perspective; we then describe our datasets, the three modules comprising our system, and the conclusions.

2. RELATED WORK

Epidemiology and Humanities. Existing scholarship on the Russian flu has focused on the popular press as evidence of growing confidence about medical understanding as well as underlying anxieties about the dangers of modernity [14; 15; 28]. More recent work has explored the ways that newspapers and medical journals disseminated the ideas of prominent American medical experts [10]. The Russian flu has also been examined in relation to the more deadly Spanish flu in 1918 by showing how popular and expert understanding of this disease developed over the decades [8]. However, unlike our approach, they follow the traditional humanities methods and do not study a large collection of data. For example, they only track the opinion of one doctor about the disease.

Data mining. As mentioned before, past work has looked into social networks and popular online media to map news cycles. The most related work to our study comes from [21], who cluster distinctive phrases as memes and track their evolution and propagation on blogs and news websites. We discuss some of these works next. Unlike these studies, we explore archival sources – both popular (newspapers) and academic/scientific (medical journals) – to study a major global historical event.

Analyzing use of terms and vocabularies: Several papers analyze vocabulary co-occurrence in news [23] and health vocabulary usage from social media data [16]. Becker et al. [7] propose an online method to distinguish between event-related and non-event tweets. Hamilton et al. [11] detect the semantics of words and show that sentiment varies across time and between communities. Lavrenko et al. [19] build a language model from the co-occurrence of words in news and detect the fluctuation or trend in stock prices based on the language model.

Identifying tone and sentiments in newspapers: Kam et al. [17] developed a tool to identify the difference of contents and tone of arguments in a newspaper. To analyze the difference between contents over time, they built a network from the co-occurrence of keywords used in different paragraphs of content. For analyzing tone of arguments, they used Bayesian classifier to classify the positive and negative tone of each paragraph in a content.

Storytelling using data mining: A lot of text mining and data mining research has analyzed opinion-forming over social media and news [29]. Word2vec [26] has performed very well in representing words and their similarities in vector space. With this technique, several papers have proposed

to analyze semantic and topical changes of words over time both on structured and unstructured data [35; 5]. Some papers also identify, track, and visualize topics over time in documents and online social media [12; 1; 13].

Information flow in social networks: Several papers have studied tracking information flow and diffusion in online social media through networks [25; 36; 33]. Matsubara et al. [25] developed a model to explain the rise and fall patterns of information flow in online media. Romero et al. [31] developed models to detect and analyze the difference in the spreading mechanism of various topics on Twitter.

3. FOCUS QUESTIONS

We particularly explore the following research questions from a historical perspective. They connect themes central to humanities inquiry with the opportunities and challenges presented by the availability of digitized texts and advances in computational analysis.

(Question 1) Vocabulary usage: How do newspapers and journals report the Russian flu pandemic? And how do they use flu-related terms to describe the situation?

(Question 2) Identifying tones: Can we help historians in understanding the tone and sentiment behind reports about Russian flu and investigate the popular opinion about the pandemic?

(Question 3) Storytelling: Finally, can we automatically examine the reports during the incident time and detect the important time segments? This in turn, helps history scholars get an in-depth understanding of how public opinion changed about the pandemic.

We tackle these questions using various data mining techniques, ranging from network analysis, text mining to time-series and sequence analysis. Answering them offers unique insights into a historical era when transnational medical research and global news reporting were important parts of the collective human experience.

We integrate our methods into a unified easy-to-use framework GrippeStory (see Fig 2 for an overview), which helps history scholars to visualize, interpret and analyze historical flu pandemic data. Although the Russian Flu is our major focus, we will take special care to develop work-flows and methods that are of general interest for disease tracking across chronological periods and geographical locations.

4. DATASET DESCRIPTION

We use raw digitized newspapers and medical journals from the United States which provide a comprehensive searchable documentation of the Russian flu pandemic. For each module in GrippeStory, we do some prepossessing which we explain in Sections 5, 6, and 7.

Sources. We collected a large amount of digitized textual data of US newspapers from the Library of Congress[1] and medical journals from three sources: Medical Heritage Library[2], Internet Archive[3], and Hathi Trust[4] between 1889 and 1893. Fig.'1 is a snapshot of Evening star newspaper reporting about the appearance of Russian flu in 1889. A total of 12,345 pages contained the word "influenza" among titles in the Chronicling America collection. One-third of these pages were published between November 1889 and April 1890, the peak months of the Russian flu pandemic. A keyword search for just ten days, December 18-28, 1889, locates more than four hundred pages with reporting on the Russian flu from more than one hundred newspaper titles located in nearly thirty states. We focus on eleven newspaper titles from different regions in the USA: Omaha Daily Bee, The Sun, Pittsburgh Dispatch, Evening World, Evening Star, Record Union, Los-Angeles Herald, St Paul Daily Globe, Wheeling Daily Intelligencer, Rock Island Daily Argus, and Salt Lake Herald. We also focus on seven medical journals: Boston Medical and Surgical Journal, Medical record, Medical age, Medical News, New York medical journal, Indianapolis Medical Journal.

Keywords. To do the term analysis (Question 1) and storytelling (Question 3) we analyze flu-related terms and their changes over the time. Hence, we must identify a set of important terms as keywords to focus our study on them. We used two months of data to extract keywords and phrases. In the initial phase, a team of humanities professionals generated a list of keywords based on close reading of selected texts about the Russian influenza in medical journals and newspapers. These keywords were grouped by categories, such as news reporting or types of symptoms, and then compared between newspapers and journals. Considerable overlap existed in the keyword lists for newspapers and medical journals, although the latter revealed a higher proportion of expert medical terminology. The second method of generating keyword and phrases followed the extraction of collocations around the term "influenza". As the humanities scholars classified phrases, they also identified manually the phrases and words that explained these classifications.

5. VOCABULARY USAGE

In this section, we analyze how different terms or vocabularies have been used in various newspapers/medical journals cover the pandemic.

5.1 Methodology

Goal. We want to understand how different sources describe the same event (i.e. Russian Flu pandemic) using different language. For example, Evening Star newspaper reported the appearance of Russian flu in New York city as "*Die grippe has come. The European sneeze makes its appearance in New York*". While New York medical journal reported that: "*The more severe cases are marked by a decided rigor, alternating with heat and flushing of skin, and the fever ...*". In the above example, medical journals use more advanced scientific terms to describe the pandemic while newspapers use a more informal way to describe similar event. We try to understand this difference by looking at the terms used in documents of various newspapers and medical journals and study their co-appearance in articles.

Methods. The idea is to understand the relationship between flu related terms in different documents. A graph is a perfect data structure for this purpose as it is designed to model relational data. Hence, we are using a graph for our setting to capture the relations between keywords/terms. Additionally a graph can easily handle the sparse nature of the data: this is important because the relations between keywords/terms can be sparse.

[1]http://chroniclingamerica.loc.gov/
[2]http://www.medicalheritage.org/
[3]http://archive.org/
[4]http://www.hathitrust.org/

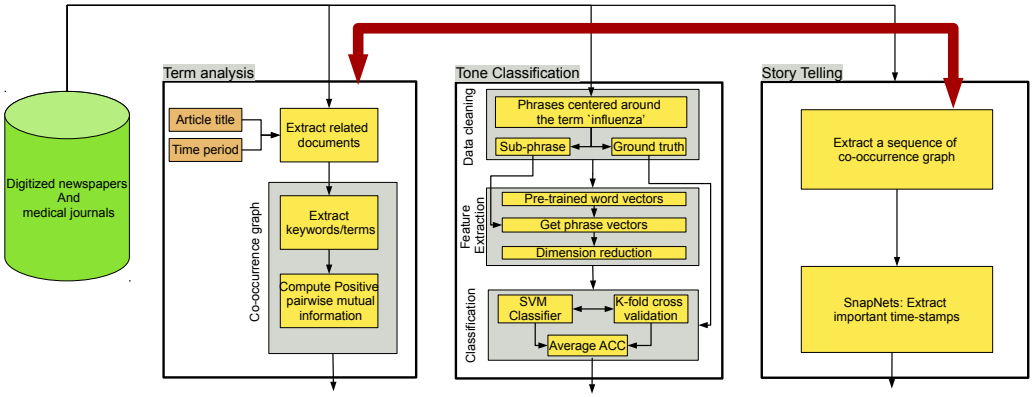
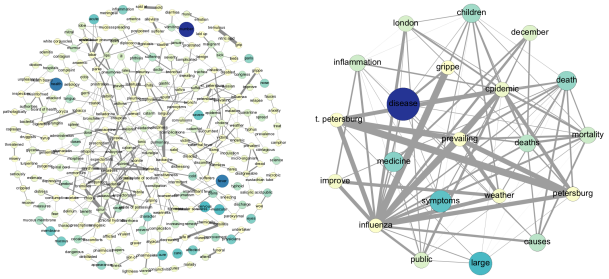


Figure 2: **An overview of GrippeStory. We use three different modules for tracking and analyzing disease from a historical perspective.**



(a) (b)

Figure 3: **The visualization of New York medical journal: (a) Co-occurrence graph (b) Influenza ego-net. The ego-net highlights the most related terms to "influenza".**

Therefore, we design a weighted and undirected so called *co-occurrence* graph $G(V, E, W)$. Nodes V represent given terms/keywords (see Sec. 4) in the documents and edges E show the co-occurrence relation between the end nodes. We say nodes/terms i and j co-occur together if they are nearby in the text of the documents. Edge weights W represent the strengths of this co-occurrence relation. Inspired by natural language processing literature such as [11], we compute the positive pairwise mutual information as edge weights as follows,

$$W_{i,j} = \max\left\{\log\left(\frac{\hat{p}(i,j)}{\hat{p}(i)\cdot\hat{p}(j)}\right), 0\right\} \quad (1)$$

We consider a fixed-size window of text sliding over the entire document. In Eq. 1, $\hat{p}(i)$ denotes empirical probabilities of a term i that is the number of sliding windows the term appear divided by the total number of sliding windows. $\hat{p}(i,j)$ is empirical joint probability of terms i and j co-occurring within the same sliding window of text.

Finally, we further study the difference between newspapers and medical journals by comparing the frequency distribution of the set of keywords in various articles.

5.2 Observations

We build the co-occurrence graph for medical journals and newspapers in 1890 using the keywords extracted with the help of history domain experts. Here we give some observations and analysis of the co-occurrence graphs.

1. Graph and ego-nets visualizations. The co-occurrence graph G helps us to visualize the dataset and gives us a better insight about how newspapers and medical journals report the Russian flu pandemic. For example, we visualized the corresponding graph of New York medical journal and its corresponding ego-net of "influenza" in Fig. 3a and 3b. The ego-net highlights the most related terms to "influenza" such as "symptoms", "children", "death", etc. It indicates that the reports usually talk about the symptoms of influenza and its danger for "children" and "death" cases of it. (Other visualizations are omitted due to lack of space).

2. Heavy Nodes. Using the empirical probability $\hat{p}(i)$ we compute the frequency of terms in different medical journals and newspapers. Tab. 1 shows the overall most frequent terms and the most frequent terms in the ego-net of "influenza". According to Tab. 1, the newspapers share common vocabulary to explain the Russian flu (as do medical journals). However, there is little similarity between the vocabulary of newspapers and medical journals. Comparing the most frequent terms in the entire article with the ones associated with "influenza" shows that in newspapers, the vocabulary usage changes around "influenza". However, it is more consistent in medical journals.

3. Heavy Edges. The weight of an edge in the graph G (Eq. 1) represents the strength of the co-occurring relation of the corresponding terms of its end nodes. Fig. 4 (Top row) shows heaviest edges (i.e., most co-occurring terms) in medical journals and newspapers. Also, Fig. 4 (bottom row) shows most co-occurring terms in the ego-net of "influenza". They indicate that the way of explanation is different from newspaper to newspaper and journal to journal.

4. Degree distribution. Fig. 6 shows the degree distribution of the co-occurrence graph of New York medical journal journal. It shows the *co-occurrence* graphs follow the skewed distribution which means the importance of terms are skewed in the medical journals (Note the graphs of newspapers follows the same degree distribution).

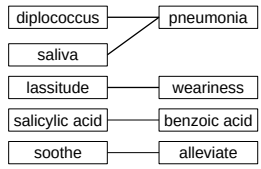
5. Diameter. The average diameter of the *co-occurrence* graphs are 4.75. This relatively small diameter of *co-occurrence* graphs demonstrates that the graphs are dense.

6. Frequency distribution of terms. Fig. 5 shows the frequency distribution across all terms in medical journals and newspapers. Unlike Tab. 1 it shows the usage of *all* terms and not only the most frequent ones. It indicates that the vocabulary usage of medical journals is very similar while in newspapers it changes in different titles. It give an interesting insight regarding the difference between popular and experts viewpoints which we will further explain.

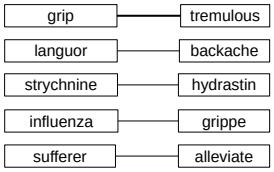
Analysis. Looking at the most frequent terms and most

Type	Title		Overall Rank					Rank in terms associated with "influenza"				
			1st	*2nd*	*3rd*	*4th*	*5th*	*1st*	*2nd*	*3rd*	*4th*	*5th*
Journals	Boston Medical and Surgical Journal	Term	Case	Patient	Treatment	Disease	Hospital	Case	Disease	Hospital	Symptom	Number
		$\hat{p}$	0.098	0.046	0.038	0.037	0.028	0.098	0.037	0.028	0.02	0.017
	Medical record	Term	Case	Patient	Treatment	Disease	Hospital	Case	Disease	Large	Symptom	Fever
		$\hat{p}$	0.089	0.048	0.043	0.041	0.025	0.089	0.041	0.021	0.020	0.018
	New York medical journal	Term	Case	Patient	Disease	Treatment	Large	Disease	Large	Symptom	Medical	death
		$\hat{p}$	0.088	0.050	0.045	0.040	0.021	0.045	0.021	0.019	0.014	0.012
Newspapers	Omaha daily bee	Term	Number	Case	Public	Sold	Price	Number	Case	Public	Ill	Children
		$\hat{p}$	0.009	0.009	0.008	0.0077	0.0077	0.009	0.009	0.008	0.006	0.005
	The Sun	Term	Ill	Large	Number	London	Public	Number	London	Public	Died	Case
		$\hat{p}$	0.019	0.010	0.007	0.005	0.005	0.007	0.005	0.005	0.005	0.004
	The evening world	Term	Ill	Children	Death	Number	Large	Ill	Children	Death	Number	Large
		$\hat{p}$	0.013	0.007	0.006	0.005	0.004	0.013	0.007	0.006	0.005	0.004

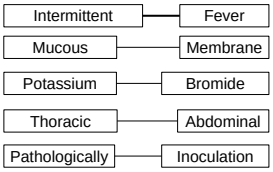
Table 1: **Top five overall frequent terms and Top five frequent terms associated with "Influenza" in medical journals and newspapers. Note the set of frequent terms is almost consistent in different titles of Medical journals while it varies in different newspapers.**



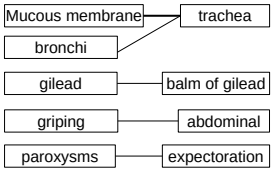
(a) Boston Medical and Surgical Journal



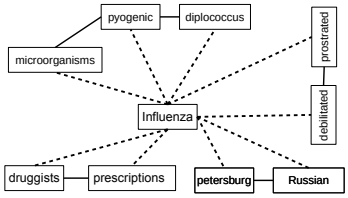
(b) New York medical journal



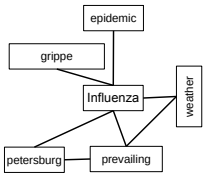
(c) The Sun



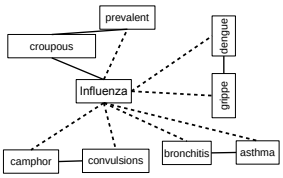
(d) Omaha daily bee



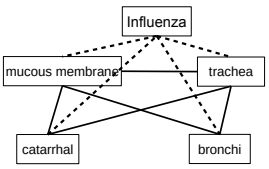
(e) Ego-net: Boston Medical and Surgical Journal



(f) Ego-net: New York medical journal

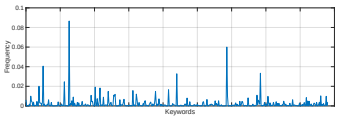


(g) Ego-net: The Sun

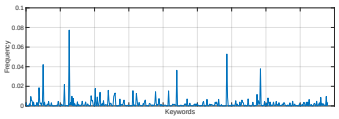


(h) Ego-net: Omaha daily bee

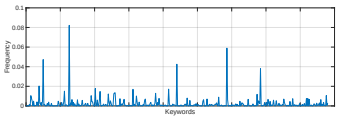
Figure 4: **Top five co-occurring terms in different journals and newspapers in the entire article (Top row) and in the ego-net of "influenza" (Bottom row): (a) and (e) Boston Medical and Surgical journal,(b) and (f) New York Medical journal, (c) and (g) The Sun and (d) and (h) Omaha daily bee newspapers. The edges thickness is proportional to the co-occurrence score of two terms. The dotted edges show the co-occurrence of terms with 'influenza'. These edges are not among the top co-occurring pairs. Note the most co-occurring terms are different between different journals and newspapers.**



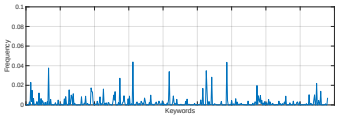
(a) Boston Medical and Surgical Journal



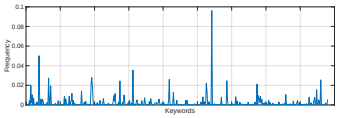
(b) Medical record



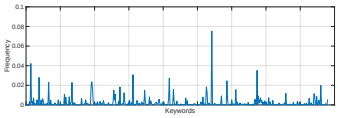
(c) New York medical journal



(d) Omaha daily bee



(e) The sun



(f) The evening world

Figure 5: **Frequency distribution across all terms in medical journals and newspapers. Medical journals: (a) Boston Medical and Surgical Journal (b) Medical record (c) New York medical journal. Newspapers: (d) Omaha daily bee (e) The sun (f) The evening world. The vocabulary usage of different journals are very similar since they explain the situation in a scientific way while the frequency of terms is very diverse in newspaper titles since they reflect the opinions in different areas.**

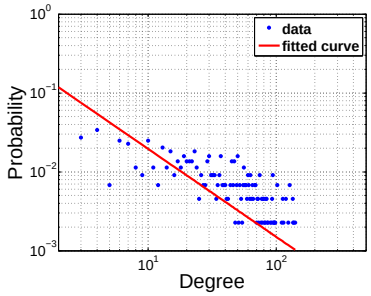


Figure 6: **Degree distribution of co-occurrence graph. Note it follows a skewed distribution. It indicates that few terms are more important and they occur with other terms more frequently.**

co-occurring terms confirm a humanities interpretation that medical journals would use more specialized terms to define diseases while newspapers would use more popular terms to describe the social impact of the disease. The fact that terms such as"treatment" and "hospital" appear more frequently in journals is further evidence of these priorities, while the fact that "cases" appeared at similar frequencies in both sources indicates that this term had both popular and expert relevance. The consistency in frequency distribution across all terms in medical journals indicates that the articles in medical journals invariably explain the pandemic in scientific language which causes the similar term usage. On the other hand, newspaper articles reflect the public opinion which varies regionally throughout the country. In summary, the disparities in word usage of medical journals and newspapers reveal this fact that the aspects the expert community was focusing on, were different from those of the public. For example, high degree nodes in the graph G for medical journals describe the scientific aspect of the disease and its symptoms such as 'Patients', 'Symptoms', and 'Fever', whereas newspapers use more alarmist terms such as 'Death', 'Suffering', and 'Children'.

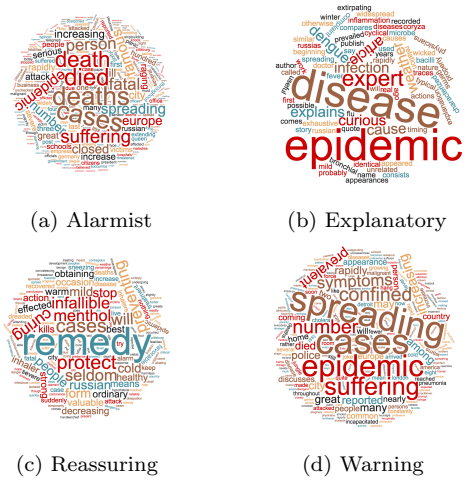
6. IDENTIFYING TONE

The investigation in the previous section helps us to understand the vocabulary usage of newspapers and medical journals. However, it does not give any insight into the reactions to the pandemic in the reports. In this section, we analyze the opinions and sentiments in these reports.

6.1 Methodology

Goal. We need to identify the *tone* of reports in newspaper and medical journals. It helps historians understand the public and experts opinions about the event better. For example, if articles about Russian Flu in medical journals are alarmist, it portrays Russian Flu as a dangerous pandemic threatening the public health. Hence, we would like to go *beyond* merely considering the word usage (as we did in Sec. 5. However, traditional methods to manually detect the tone of sentences is too expensive and not a salable process. Therefore, we want a design a high quality method to automatically identify the tone of phrases related to the Russian flu pandemic in the articles of newspapers and medical journals. Suggested by history domain experts, we plan to classify sentences based on their tone into four classes: (1) **A**larmist: emphasizing the danger of the pandemic and the number of victims; (2) **E**xplanatory: providing information in a neutral manner (3) **R**eassuring: encouraging a sense of optimism by minimizing the danger (4) **W**arning: urging measures to prevent infection and contain the spread of disease. There are very subtle differences between these tones. Different tones convey various opinions.

Methods. The tone classification component in Fig. 2 is an overview of the classification process. First, we extracted the phrases centered around the term "influenza". Domain experts manually classified a few of them into the aforementioned classes as the ground truth. We ask them to mark the important sub-phrases which played a major role in their decision to classify each phrase. We use these phrases for whom we have labels to train the classifier and identify the tone of unlabeled phrases. Fig. 7 and Fig. 8 visualize the word cloud in different tone classes and frequency of each class in newspapers and medical journals. Fig. 8 shows that the tone classes are extremely imbalanced in our datasets. Also, Fig. 7 shows the similarity in word usage in different classes such as Alarmist and Warning. Due to these reasons, the tone classification task is highly challenging in our historical data sources.



(a) Alarmist (b) Explanatory

(c) Reassuring (d) Warning

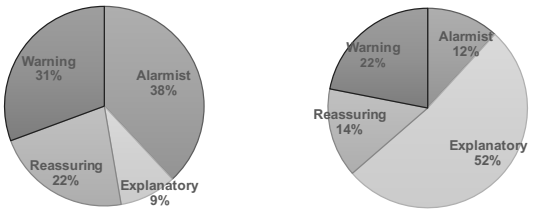
Figure 7: **A word-cloud of sentences of different tone classes. Note the difference of words in each tone.**

Second, we propose to map the phrases to a vector space to be able to classify them. Recent work in Deep learning has been shown to be useful in text classification [37; 18]. Two recently popular ways to extract features from phrases are Doc2Vec method [20] and Vector Space Model [22]. However, they do not perform well in our problem due to the low-quality OCR and lack of data: many words are misspelled in the digitized documents. Therefore, we can not find critical terms in phrases. Also, various spelling of words damages the quality of the extracted features. This is because, the system learns separate feature representation for each spelling which reduces the quality of the feature vectors.

Our approach to reduce the effect of the misspelling of words is as follows: First, we use the Google's pre-trained Word2Vec model[27]. The model includes word vectors for a vocabulary of 3 million words and phrases that they trained on around 100 billion words from a Google News dataset. The vector length is 300 features. Next, we define the feature vector of a phrase as the average vector of all words in the phrase. Note it naturally ignores the words that are not in the pre-trained vocabulary to get the feature vector of

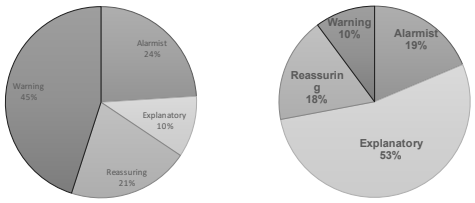
phrases. Finally, we reduce the dimension of feature vectors by using Singular Value Decomposition (SVD) to make the classifier more robust. After extracting features, we leverage SVM classifier and k-fold cross-validation (k = 10) to predict the labels of phrases.

Tone classes in Newspapers and Medical journals are usually extremely imbalanced. For example, in early stages of the pandemic we expect to see large amount of Alarmist articles and a small number of Reassuring ones. Therefore, if we naively classify the input phrases into tone classes, we will get low quality prediction. Our experiments show that in some cases we can not even detect a single phrase from under-represented tone classes (i.e. the $F1$-*score* of the class is 0.0). We propose to improve the quality of the tone classification by training the classifier using oversampling method [4]. In this method, we add copies of instances from the under-represented classes to make the training data more balanced. It keeps the classifier from ignoring the under-represented classes and have a more accurate prediction. Our extensive experiments show that oversampling have a dramatic effect on the performance. In some cases, it improves the $F1$-*score* of under-represented tone classes around 70%.



(a) Newspapers (b) Journals

Figure 8: **(a) and (b) show the frequency of each tone class in newspapers and journals. It indicates that the reports of journals are more explanatory while newspapers are more alarmist.**



(a) Pittsburgh Dispatch (b) Indianapolis Journal

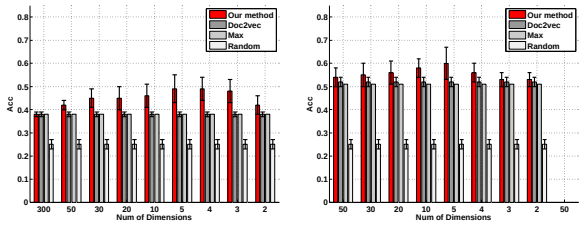
Figure 9: **Distribution of predicted labels for (a) Pittsburgh Dispatch newspaper and (b) Indianapolis medical journal. Note the classifier highlights the difference between the newspapers and journals (i.e., the newspaper articles are more Warning/alarmist while the journal articles are more explanatory.)**

6.2 Observations

We have the following observations,

1. Usage of different tones. Fig. 8a and 8b show an overview of frequency of each class in newspapers and medical journals. In journals, **E**xplanatory is by far the most frequent tone and **A**larmist, by contrast, appeared least frequently. However, in newspapers, **A**larmist and **W**arning are the most frequent tones. These figures indicate that the tone classes are extremely imbalanced which makes the tone classification more challenging.

2. Performance of classifier. We run the classifiers with 10-fold cross-validation 1000 times and compute the average accuracy (i.e. $ACC = \frac{\#\text{Correctly Classified Prases}}{\#\text{Phrases}}$) and its standard deviation to evaluate the quality of the classifier. Also, to further investigate the quality of detecting each tone class we compute the $F1$-*score* and *precision* and *recall* of each class. We compare our method with three baselines: (1) Doc2vec classifier which use the doc2vec method to extract features and use the SVM as the classifier (2) Max classifier which classifies all phrases into the largest class (i.e., Alarmist in newspapers and Explanatory in Journals). (3) Random classifier which randomly classifies nodes to each label. Fig. 10a and 10b show the average Accuracy (ACC) and the standard deviation of the classifier with different numbers of features selected by SVD for newspapers and journals. According to Fig. 10a and 10b, our method gives the best accuracy among other baselines. Also, it shows that we get the best ACC using five features for both journals and newspapers. Tab. 2a and 2b show the *precision*, *recall*, and $F1$-*score* of each class. They confirm the high quality of our classifier and show that oversampling keeps the quality of detecting under-represented classes relatively high.



(a) Newspapers (b) Journals

Figure 10: **The average accuracy of our method and other baselines using 10-fold cross validation. Note despite that the classes are extremely unbalanced our method beats the baselines.**

	Precision	Recall	F1-score
A	0.88	0.41	0.56
E	1.00	0.50	0.67
R	1.00	0.38	0.55
W	0.28	0.88	0.42
avg/total	0.80	0.51	0.55

(a) Newspapers

	Precision	Recall	F1-score
A	0.50	0.50	0.50
E	0.50	1.00	0.67
R	1.00	0.33	0.50
W	1.00	0.43	0.60
avg/total	0.78	0.61	0.59

(b) Journals

Table 2: **Detailed performance of our classifier. It confirms the high quality of our classifier and suggests that more data in each tone class increases the accuracy of detecting it.**

3. Scaling up tone analysis. We use the trained classifier to detect the tone of sentences of new journal and newspaper titles not manually labeled and hence not presented in training set. In this way, our classifier can help historians to scale up the analysis to a much larger set of articles (compared to manual analysis before). Fig. 9a and 9b shows the distribution of detected tones for the new titles. Note the distribution of tone classes in new titles is similar to other titles in Fig. 8a and Fig. 8b. This similarity indicates that the general reaction to Russian flu pandemic was consistent

among experts as well as non-experts. For instance, the tone distribution in Indianapolis journal is similar to other medical journals with a serious yet reassuring tone in their claims that the disease was cyclical and familiar, with causes soon to be discovered, and cure within reach. On the other hand, Pittsburgh dispatch adopts a tone similar to other newspapers reflecting public panic and characterized with alarmist and cautioning rhetoric.

Analysis. The relative distribution of phrases across these four categories demonstrates the complexity of reporting about influenza in both popular and medical texts. For example, there is a tiny difference between the terms used in Alarmist and Warning categories which makes the data mining task more challenging. The majority of sentences in medical journals have an investigative and explanatory tone in accordance with the experts attempts at understanding the situation. On the other hand, newspapers had an alarmist/warning tone reflecting the public fear about the pandemic. Close reading of illustrative text, in our case, allows historians to understand how medical experts and the public explained a disease outbreak. Nevertheless, automatic tone identification of these articles permit interpretations on a larger scale, across a broader range of textual evidence, possibly allowing historians to uncover new angles and illuminating analysis. As an example, an unexpected revelation about the Russian flu coverage in newspapers is that although the word usage is different among various titles (See Sec. 5), we observe a consistent tone in them. In effect, automatic tone identification allows historians to explore sources deeper with new approaches, while enhancing traditional techniques of close reading and layered analysis.

7. STORYTELLING

In previous sections, we focused on static data analysis and did not consider the effect of time in term usage and public opinion. In this section, we want to study the *dynamics* of reporting about the Russian flu pandemic.

7.1 Methodology

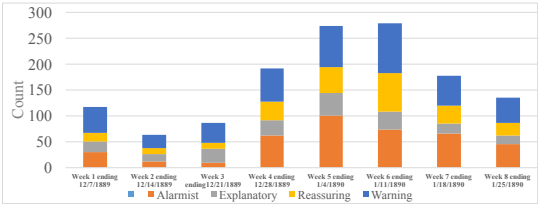
Goal. We would like to figure out change-points in public opinion over time about the pandemic. In other words, we detect how tone/language of newspaper reporting changed regarding the pandemic.

Methods. First, we study the dynamics of the vocabulary usage. We track the co-occurrence graph of newspapers in each week as a representation of how they report about the Russian flu. Hence, the problem of recognizing the change in reporting the Russian flu-related news will be equivalent to finding the best time segment in the sequence of co-occurrence graphs of newspapers. To find the best time segmentation we use SNAPNETS [2] which is a non-parametric and scalable graph sequence segmentation method. SNAPNETS summarizes graphs by grouping similar nodes together and generating a sequence of smaller graphs with super-nodes and super-edges. Next, it extracts important features from these small networks and finds the best time segmentation by maximizing the average distance between adjacent segments.

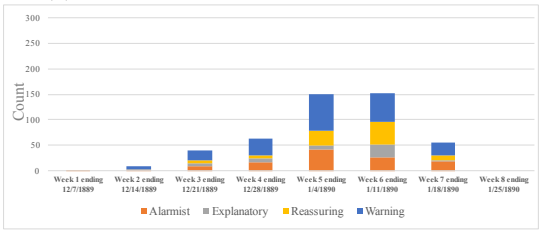
Next, we investigate the change-points in public opinion over time about the pandemic by tracking the frequency of each tone in articles of newspapers. Automatically identifying the tone of phrases gives us the opportunity to detect the tone of articles effectively and efficiently and study the dynamics of articles' tone over time while it is almost impossible by manually classifying the tone of a large number of articles over time. We track the tone of sentences about Russian flu and find the changes in the tone of reports about Russian flu. To get the time segmentation, we adapt SNAPNETS to work with general type data sequences. We force SNAPNETS to directly use the frequency of tone classes as the feature vectors in each time-stamp and find the best segmentation based on them.

7.2 Observations

We find the segmentation considering the tone of reports (i.e. Data sequence approach) with daily and weekly granularity. Also, we extract the segmentation of co-occurrence graph sequence using SNAPNETS method. We use the newspapers data from December 1st 1889 to January 31st 1890. Here we give some observations and analysis of these segmentations.



(a) Newspapers with manually identified tones

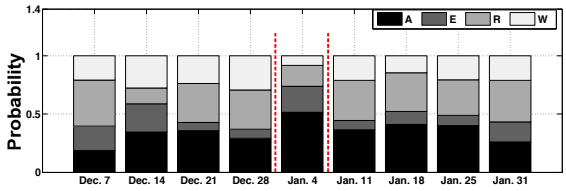


(b) Pittsburgh Dispatch

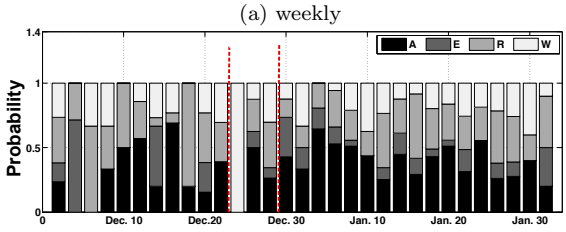
Figure 11: **The temporal distribution of tone classes from December 1889 to January 1890: (a) Newspapers with manually identified tones, (b) The new title (Pittsburgh Dispatch). It shows that our classifier detects the tone of new title overtime and it matches the behavior of the manually labeled sequence.**

1. Dynamics of tone usage. We used our classifier to detect the distribution of tone classes of a new newspaper over time (see Fig. 11b) and compare it with the distribution of newspaper manually labeled (see 11a). In Fig. 11b the increased percentage of **A**larmist reports during the last week of December and the first weeks of January make sense, as these were the weeks of increasing cases and deaths from Russian flu and Fig. 11a confirms it.

2. Language changes. Fig. 13 shows the result of time segmentation using SNAPNETS on co-occurrence graph sequences. It detects the newspapers reports changed in the last week of December. In Fig. 13, the graphs represent a summary of co-occurrence graphs in each segment. According to this figure, the language of newspapers changed over time about the Russian flu: In the early December, many isolated nodes in the summary graph indicates that many flu related terms do not co-occur with others. There were only a few terms popular. However, gradually in late Decem-



(a) weekly



(b) Daily

Figure 12: **The segmentation results: (a) Weekly (b) Daily. Note our approach can give the segmentation with different granularities. See text for more analysis.**

ber and January, more terms were used in reports about Russian flu. Also, looking at the last segment shows one node has a high degree. It indicates that few terms (e.g., Russian influenza) become more popular and co-occur with most terms.

3. Tone changes. Fig. 12b shows the segmentation result on the sequence of tone classes in newspapers. It confirms the result of SNAPNETS and suggests that the newspapers reports changed in the last week of December. According to the segmentation, the tone of the reports changed from more warning (in early December) to more alarmist (in mid-January).

4. Effect of granularity. Our time segmentation approach gives historians the capability to investigate the sequences in different granularities. Hence, we study the effect of different granularity in understanding the story of Russian flu. Fig. 12a and 12b show the segmentation results looking at the tone of weekly and daily reports. Fig. 12a has a lag of one week to detect time segments. This is an indication that in the week of Jan 4 the alarmist reports were the majority tone in newspapers.

Analysis. These results provide a mechanism for comparing the spread of information with the spread of disease during the course of an epidemic. The increased appearance of disease-terms collocated with influenza is suggestive of how the spread of this disease was accompanied by increasingly extensive and complex reporting about the disease. Our storytelling approach is an advance over usual humanities methods: Typically, humanities methods emphasize close reading, but our methods allows researchers to detect a subtle correlation between words and detects more complex patterns automatically.

8. USER INTERFACE

To improve usability, we also designed a user interface which integrates all our algorithms into a unified framework called GRIPPESTORY (based on a common word for flu in that era). Our code is in Python and to visualize graphs we use Gephi layouts [6]. We also parallelized the storytelling telling module and divide its workload into several processors to scale up the process of detecting change points in a large amount of data. Fig. 14 show an example UI screenshot of GRIPPESTORY. It allows users to visualize the vocabulary usage (Module 1), automatically detect the tone of sentences (Module 2), and segment data using daily, weekly, and monthly scales (Module 3).

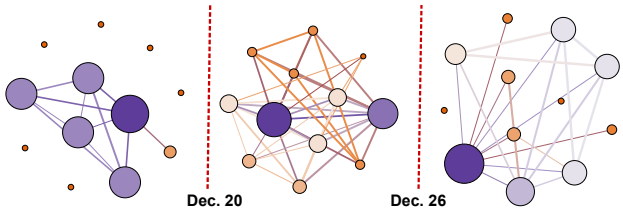
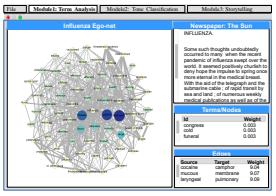
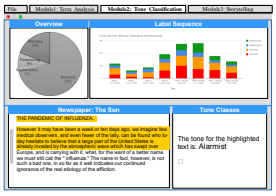


Figure 13: **Co-occurrence graph sequence segmentation of all newspapers on a weekly bases. Note the subtle changes in co-occurrence graphs which show the transformation in word usage over time about the Russian flu.**



(a) Term analysis module



(b) Tone classification module

Figure 14: **Snapshots of the UI of GrippeStory: (a) Users can visualize the influenza ego-net (b) Users automatically classify the tone of selected sentences.**

9. CONCLUSIONS AND DISCUSSION

In this article, we looked into the problem of understanding the dynamics of news-cycles during a global historical pandemic using data mining techniques. We designed an easy to use framework GRIPPESTORY which combines two types of data sources (i.e., newspapers and medical journals) and helps history scholars to visualize, interpret and analyze the popular and experts opinion about the pandemic. We look into three important questions: (I) How do media report the Russian flu pandemic? (II) How can we help historians in understanding the sentiment behind the reports about the pandemic? (III) How can we help history scholars get an in-depth insight of how public opinion changed about the pandemic? We answered these questions by leveraging various data mining techniques, from network analysis and text mining to time-series and sequence analysis. Each of these modules was developed in collaboration with historians, and each enables a task, which would not have been possible without the traditional data mining approaches. For instance, the storytelling module provides a mechanism for them to compare the spread of information with the spread of disease during the pandemic. Our extensive experiments and analysis on various newspaper titles and medical journals between 1889 and 1893 show that GRIPPESTORY gives meaningful results to help history scholars get a deeper understanding about a global pandemic.

Note that the GRIPPESTORY framework is language independent (such as co-occurrence graphs, and segmentation module), or easily extensible (embedding-based tone anal-

ysis). The Russian Flu was a global pandemic and gives a fertile ground to apply our system to other situations. For example we can analyze the newspapers and medical journals beyond the United States and in other languages such as German or Russian. Finally, studying other global pandemics in different time periods, will also be interesting.

Acknowledgments: This article is based on work supported by the NSF CAREER IIS-1750407, the NEH (HG-229283-15), ORNL, Facebook faculty gift, and NSF Urban Computing Fellowship award.

10. REFERENCES

[1] J. Allan. Detection as multi-topic tracking. *Information Retrieval*, 2002.

[2] S. E. Amiri, L. Chen, and B. A. Prakash. Snapnets: Automatic segmentation of network sequences with node labels. 2017.

[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. 2012.

[4] R. Barandela, R. Valdovinos, J. Sanchez, and F. Ferri. The imbalanced training sample problem: Under or over sampling? *Springer*, 2004.

[5] R. C. Barranco, R. F. D. Santos, and M. S. Hossain. Tracking the evolution of words with time-reflective text representations, 2018.

[6] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 2011.

[8] N. Bristow. *American pandemic: The lost worlds of the 1918 influenza epidemic*. Oxford University Press, 2012.

[9] E. T. Ewing, S. Gad, B. L. Hausman, K. Kerr, B. Pencek, and N. Ramakrishnan. Mining coverage of the flu: Big data's insights into an epidemic. *Persp. Hist.*, 2014.

[10] E. T. Ewing, V. Kimmerly, and S. Ewing-Nelson. Look out for 'la grippe': Using digital humanities tools to interpret information dissemination during the russian flu, 1889–90. *Medical history*, 2016.

[11] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. *EMNLP*, 2016.

[12] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. 2000.

[13] A. Hermida, S. C. Lewis, and R. Zamith. Sourcing the arab spring: A case study of andy carvin's sources on twitter during the tunisian and egyptian revolutions. *JCMC*, 2014.

[14] M. Honigsbaum. The great dread: Cultural and psychological impacts and responses to the 'russian'influenza in the united kingdom, 1889–1893. *Social History of Medicine*, 2010.

[15] M. Honigsbaum. *A history of the great influenza pandemics: death, panic and hysteria, 1830-1920*. IB Tauris, 2013.

[16] L. Jiang and C. C. Yang. Using co-occurrence analysis to expand consumer health vocabularies from social media data. 2013.

[17] M. Kam and M. Song. A study on differences of contents and tones of arguments among newspapers using text mining analysis. *JIIS*, 2012.

[18] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. 2016.

[19] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Mining of concurrent text and time series. 2000.

[20] Q. Le and T. Mikolov. Distributed representations of sentences and documents. 2014.

[21] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. 2009.

[22] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 2015.

[23] S. Liu, X. Fan, and J. Chai. A clustering analysis of news text based on co-occurrence matrix. 2017.

[24] A. Matakos, E. Terzi, and P. Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 2017.

[25] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14. ACM, 2012.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, 2013.

[27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[28] J. Mussell. Pandemic in print: the spread of influenza in the fin de siecle. *Endeavour*, 2007.

[29] K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *KBS*, 2015.

[30] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. 2012.

[31] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. 2011.

[32] H. Z. Shuang Hong Yang. Mixture of mutually exciting processes for viral diffusion. 2013.

[33] L. Sorg. Modeling information flow in social media, 2017.

[34] R. Wenzlhuemer. *Connecting the nineteenth-century world: the telegraph and globalization*. Cambridge University Press, 2013.

[35] D. T. Wijaya and R. Yeniterzi. Understanding semantic change of words over centuries. In *Detect*, 2011.

[36] S. Yardi and A. Bruckman. Modeling the flow of information in a social network. *SIGCOMM*, 2008.

[37] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.