Towards Understanding How News Coverage Affect Public Perception During Epidemic Outbreaks

Zhishen Pan, Li-Kai Chi, Timothy Dye*, and Jiebo Luo

[1] University of Rochester, Rochester, NY, United States
[2] University of Rochester Medical Center*, Rochester, NY, United States
{zpan3,lchi3}@u.rochester.edu, timothy_dye@urmc.rochester.edu,
jluo@cs.rochester.edu

Abstract. News media is one of the most prominent sources for public to receive information and updates about epidemic outbreak. We choose the 2014 West African Ebola outbreak as the case to investigate how news coverage and public perception are related to the disease activity. Our choices of newspapers take account for the potential delay in coverage related to locations (France, U.S.A., U.K.) and languages (French, English). Cross-correlation analysis is applied to compare the patterns of coverage rates among newspapers. Using sentiment analysis, it identifies factors that affect public perception during Ebola outbreak. The findings are expected to help design proper responses to epidemic outbreaks in the future.

Keywords: News Media,Public Perception,Epidemic Outbreak, Ebola, Cross-Correlation Analysis, Time Delay, Sentiment Analysis

1 Introduction

News media functions as a reliable source where people retrieve first-hand disease information during epidemic outbreak. Exploring the relationship between the coverage rate of disease-related news articles and the level of disease activity could help study whether news media is a useful tool for tracking epidemic outbreaks. Meanwhile, the agenda-setting theory says that if a news item is covered frequently, audience will regard the issue more importantly [1]. we want to know how public perception is influenced by news coverage and the level of disease activity. We choose the 2014 Ebola outbreak as our case study because of the high attention it generated around the globe. With 20,171 confirmed cases and 7,889 deaths by the end of 2014, this epidemic has been widely reported by different newspapers in different countries.

The primary goals of this paper are to better understand how news coverage is related to level of disease activity and how public perception is affected during epidemic outbreaks. This paper first compares the news coverages rate of epidemic outbreak among different newspapers along with the disease situation.

It then tries to identify the potential time delay related to languages and locations. Next, by measuring the sentiment of the comments of the news articles, it investigates how public perception is affected by the news coverage and disease level.

2 Related Works

Social media have been proven to be useful tools to reflect and predict the level of disease activity. Previous study has been shown that Google's search engine query data can be used to track disease activities [2]. Twitter is also proven to be closely correlated to the level of disease activity [3] [4]; Signorini also points out that Twitter can also be used to monitor users' concerns about the epidemic [5]. In the case of the 2010 Haitian cholera outbreak, both social and news media were able to achieve early estimation of epidemiological pattern [6]. However, to our knowledge, no one has compared different news media in different countries during an global epidemic outbreak.

3 Data Collections

New York Times(U.S.A), *The Guardian* (U.K.) and *Le Monde* (France) are chosen to be the three news sources. With different locations and languages, we are able to account for the language and location factor in our analysis (Table 1). The data related to Ebola Infections are provided by WHO and open to public. To collect articles related to Ebola from these newspapers, the Article Search APIs provided by corresponding newspapers are used to collect data. We queried for news articles that contained the word 'Ebola' and was published between March 1st and December 31st, 2014 (Table 2).

Collected Data were then divided into 153 bins with 2-day interval so that the time-series functions will not be too noisy (as in 1-day interval) yet frequent enough to discover subtle differences. The coverage rate(total number of articles related to Ebola / total number of articles published) was calculated as the bin frequency. To analyze public perception, comments of the Ebola-related articles from *The Guardian* were retrieved by crawling the websites[3], and treated as a measurement of public perception.

Table 1. Key Information for Three Newspapers

	Headquarter	**Language**	**API**
New York Times	New York	English	http://developer.nytimes.com
The Guardian	London	English	http://open-platform.theguardian.com
Le Monde	Paris	French	http://www.lemonde.fr/recherche

[3] *New York Times* and *Le Monde* comments are not collected due to access limitation.

Table 2. Data collected through APIs and WHO

Data Description	Number
Total Confirmed Ebola Cases	20171
Total Ebola-related Death	7889
Ebola-related Articles in *The Guardian*	1160
Ebola-related Articles in *New York Times*	4321
Ebola-related Articles in *Le Monde*	503

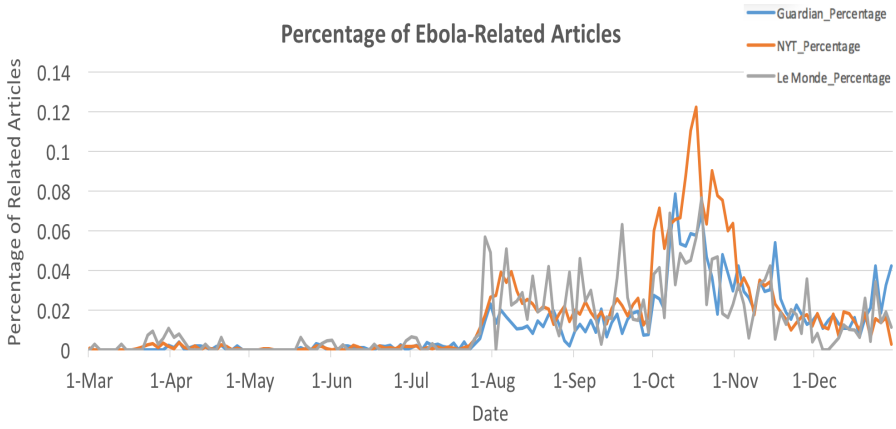

Fig. 1. Percentage of Ebola-Related Articles

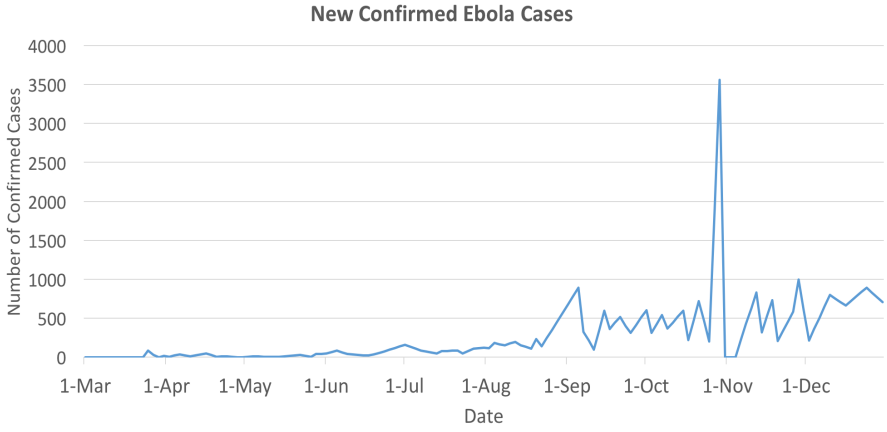

Fig. 2. New Confirmed Ebola Cases according to WHO

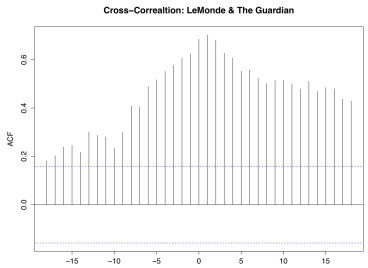

Fig. 3. Cross-Correlation between *The Guardian* and *Le Monde*

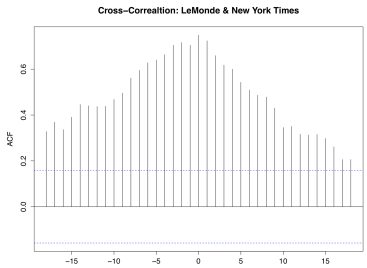

Fig. 4. Cross-Correlation between *New York Times* and *Le Monde*

4 Cross-Correlation Analysis

The coverage rate of Ebola-related articles in all news media (Figure 1) matches closely with the trend of the new confirmed Ebola infected cases (Figure 2). After removing few extreme outliers and negative values[4], the Spearman's correlation test of the Ebola-related articles of *Le Monde* and the new confirmed cases returns a correlation coefficient of 0.707 with a small p-value of 0.001, indicating that there is a strong positive relationship between these them. The more confirmed cases reported, the higher the coverage rate was. We want to further study whether language or location are factors affecting the time delay in news coverage. Cross-correlation analysis is used to accomplish this goal.

Cross-correlation is a measure of similarity of two series as a function of the lag of one relative to the other. It is developed for determining time delay between signals received at two spatially separated sensors in the presence of uncorrelated noise [7]. The maximum value of the cross-correlation function indicates the point in time where two time-series are best aligned. The goal is to identify the leading and lagging media and to identify the magnitude of the lag.

Cognizant of the fact that the Ebola outbreak started in Guinea, a French-speaking country, we assume *Le Monde* as the potential predictor of *The Guardian* and *New York Times*. The highest correlation of *The Guardian* and *Le Monde* occurs at lag = 1, with the coefficient = 0.702 and tapering occurring from both directions from the peak (Figure 3). It is therefore most likely that *Le Monde* reported the Ebola news faster than *The Guardian* by one unit (two days). One possible reason of this lag is that *The Guardian* did not have direct access and required additional time to collect information. Such result supports the assumption that as a French media, *Le Monde* has faster access to the Ebola information. The lag between *Le Monde* and *New York Times* however is most likely to be zero with the coefficient of 0.749 (Figure 4). There is no obvious lag between *Le Monde* and *New York Times*. Therefore, language factor could be a relative advantage in news coverage. However, such advantage is not obvi-

[4] The data points where new cases are large than 1500 or smaller than 0 are removed. 5 data points are removed

ous as events happening in any region of the world could be spread worldwide through Internet in seconds. There should not be an obvious and consistent time lag among different news media for events that are significant enough to catch everyone's attention.

5 Sentiment Analysis of Public Perception

The next goal of this paper is to investigate how the sentiment of the public evolved during Ebola outbreak. Sentiment analysis is a type of opinion mining used to identify the subjective information from the source [8]. Here, we apply Vader, a parsimonious rule-based model for sentiment analysis. Given a paragraph of text, it returns the relative scores of three groups (Positive, Negative, Neutral) based on words with different emotions. We then classify the text into the group that has the highest score. Statistical results have shown that this algorithm has achieved better correlation coefficients and better accuracies than previous works [9].

We use the comments from *The Guardian* as a measurement of public perception, as they are the most direct reflections of peoples' opinions after reading the news. Since *The Guardian* did not start consistently reporting Ebola until July, 2014, we collect comments from July to December. Similar to what Godbole had done in the sentiment analysis for news and blogs [10], we introduce the subjectivity metric to explore how biased people's opinions are through out the disease.

$$\text{Subjectivity} = \frac{\text{Number of Positive and Negative Comments}}{\text{Number of Neutral Comments}} \tag{1}$$

The higher the value is, the more subjective and biased people's comments are. The subjectivity metric fluctuates more at the early stage of the Ebola outbreak (Figure 5). During the massive outbreak period from September to

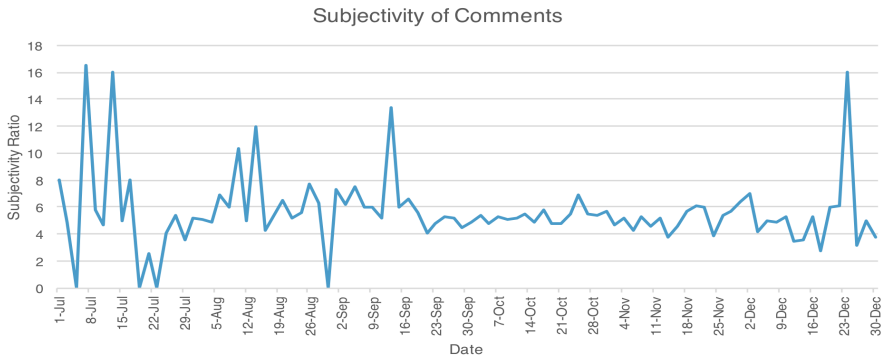

Fig. 5. Subjectivity of *The Guardian* Comments

November, the metric stayed consistent. This pattern shows that with the increasing amount of news coverage, the amount of neutral comments increased as compared to positive and negative comments indicating that people's opinions were less subjective. However, the metric increased dramatically at the end of 2014. Two major events related to Ebola occurred at that time: A technician in Atlanta, USA was potentially exposed to Ebola due to a laboratory error; a worker in Scotland was diagnosed with Ebola and was taken to a treatment center in London. People were expressing huge amount of personal opinions during this period, indicating by the high subjectivity score.

These two events had come into the spotlight for two potential reasons: 1. Ebola was potentially exposed in USA. If the disease was spread rapidly, the outcome would be catastrophic for not only U.S.A., but the entire world; 2. There was potentially Ebola exposure in London, which is close to the majority of *The Guardian* subscribers. A further detailed study of the reasons of this steep increase in subjectivity score should be developed. Our conjecture is that people show stronger incentives to express personal opinions on events that are either affecting the entire world or happening around them.

6 Conclusion

In this paper, we have shown that the trend of news coverage matches closely with the disease trend. The more Ebola cases confirmed, the more news articles were reported. Cross-correlation analysis also shows that *The Guardian* has a small but not obvious time delay compared to *Le Monde* in Ebola news coverage. Through sentiment analysis, we discover that the public could be influenced most by events that could cause significant outcome or events that are happening close by. These findings could potentially help future study to identify potential epidemic outbreak by tracking local news media.

The method we applied could be extending to other news media in different regions. Specifically, in the case of West African Ebola, one should look into newspapers in West Africa. Due to the geographic proximity, local news media is the one that receives the first-hand information. Studying the news coverage of local news media could help identify potential epidemic outbreak and design proper response in advance. To achieve this long-term goal, more comments and articles from each newspaper should be collected and a new sentiment analysis tool (e.g. sentiment analysis tools in Core NLP) that are applicable to different languages should be used in the future.

Acknowledgments

This work was supported by New York State through the Goergen Institute for Data Science at the University of Rochester.

References

1. McCombs, Maxwell E., Donald Lewis Shaw, and David Hugh Weaver: Communication and democracy: Exploring the intellectual frontiers in agenda-setting theory. Psychology Press, (1997)
2. Ginsberg, Jeremy, et al.: Detecting influenza epidemics using search engine query data. Nature 457.7232 (2009)
3. Achrekar, Harshavardhan, et al.: Predicting flu trends using twitter data. Computer Communications Workshops, 2011 IEEE Conference (2011)
4. Culotta, Aron.: Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the first workshop on social media analytics. ACM, (2010)
5. Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen.: The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic.PloS one 6.5 (2011)
6. Chunara, Rumi, Jason R. Andrews, and John S. Brownstein.:Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak.The American journal of tropical medicine and hygiene 86.1 (2012)
7. Knapp, Charles H., and G. Clifford Carter. "The generalized correlation method for estimation of time delay." Acoustics, Speech and Signal Processing, IEEE Transactions on 24.4 (1976)
8. Pang, Bo, and Lillian Lee.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2.1-2 (2008)
9. Hutto, Clayton J., and Eric Gilbert.:Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media. (2014)
10. Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena.: Large-Scale Sentiment Analysis for News and Blogs.ICWSM 7.21 (2007)