SESUG Paper 134-2018

A Case Study of Mining Social Media Data for Disaster Relief: Hurricane Irma

Bogdan Gadidov, Linh Le, Analytics and Data Science Institute, Kennesaw State University

ABSTRACT

In the wake of two recent hurricanes, Harvey and Irma, local, state, and federal governments are trying to provide relief to the millions of affected people. With projected property damage in the hundreds of billions of dollars, these recent natural disasters will have long-lasting effects on their respective areas where recovery could take years. This paper aims to use social media data, specifically Twitter, to analyze how people in the affected areas reacted to these natural disasters in the days leading up to the storm, during the storm, and after the storm. The goal is to see if there are any trends detected in the responses of affected citizens which can be used to help relief efforts in future natural disasters. For the most recent hurricane, Irma, we collected tweets in South Florida and analyze the discussed topics among civilians. Data was collected from Thursday (9/7/2017) to Wednesday (9/13/2017) (with the hurricane making landfall on Sunday morning). We use SAS® Enterprise Miner™ for the analysis of the tweets. Techniques such as stemming and lemmatization of words are used in the pre-processing of the text data. Topic modeling, text clustering, and time series are combined to better understand peoples' reactions throughout a storm event. This analysis is performed at the hourly level.

INTRODUCTION

2017 was the costliest year for natural disasters in the history of the U.S. With hurricanes Harvey and Irma hitting the U.S. just two weeks apart, local communities were left devastated with high numbers of casualties and property damage. In total, it is estimated that all the hurricanes in 2017 will cost the U.S. slightly over 250 billion dollars, which cannot fully account for the loss in economic potential and may take years to recover [1]. The development of social media allows people to express ideologies publicly, including their attitudes toward these kinds of events. In this paper, we aim to analyze the reactions on social networks to forecasted natural disasters from people in the impacted areas throughout the events. While analyzing peoples' reactions to natural disasters cannot stop the destruction of property, it can be used to ascertain whether measures to reduce the number of casualties are effective, and whether people are fully prepared for these massive storms. Specifically, we examine the recent hurricane Irma to form a study case for similar future events.

We collected tweets in areas within a 250-mile radius from Fort Myers in South Florida from Thursday (9/7/2017) to Wednesday (9/13/2017) (with the hurricane making landfall on Sunday morning) which results in a corpus of over five million tweets. Our analysis consists of two main phases: topic analysis and trend analysis. In the first phase, we use text mining techniques to extract topics discussed in the tweets captured during the mentioned period provided in SAS® Enterprise Miner™ (EM). In the preprocessing steps, the tweets are parsed into terms which are then stemmed and lemmatized. We also apply filters to remove Uniform Resource Locators (URLs) and emojis. Text clustering and topic modeling techniques are then used to determine the main theme of discussions in the tweets. Overall, we find three main themes: general comments about Irma, power outages, and hopes or prayers of the Twitter users.

In the second phase of the analysis, we use time series models, namely the Autoregressive Integrated Moving Average (ARIMA) process [2], to analyze the trends of topics of interests over the period of Irma. The main finding is that the number of tweets significantly increase during the 24 hours prior to the landfall of the hurricane. At its peak, approximately 25% of all collected tweets are related to the hurricane in comparison to about 10% during the rest of the week. We also detect a small number of tweets trying to spread important information like evacuation hotlines or price gouging hotlines. These tweets unfortunately are unable to form a separate topic due to their low frequency. Their trend also cannot be modeled since they occur only a few days before the hurricane arrives.

TOPIC ANALYSIS

We use SAS ® Enterprise Miner™ (EM) to analyze the topics in the corpus of tweets. EM provides various nodes for text mining. In this paper, we use these following nodes: Text Parsing, Text Filtering, Text Cluster, and Text Topics.

In short, Text Parsing nodes transform raw texts into terms, then stem and lemmatize them and assign their part-of-speech (PoS) tags. Text Parsing nodes also provide statistics such as the term frequencies in the corpus, or the number of documents which consist of the terms. The Text Filter node computes the weights of the parsed terms (which represent their importance) in the corpus and allow manually dropping uninteresting terms. We consider these two types of nodes' functionality as preprocessing of the corpus.

To determine "topics" discussed in a corpus, the two nodes Text Cluster and Text Topics can be utilized. They both aim to cluster the documents in the corpus into similar groupings, however with different strategies. The former applies a general clustering algorithm on the dimension-reduced documents (typically by Singular Value Decomposition) while the latter utilizes a soft clustering strategy which allows a document to belong to multiple topics. The two techniques use the term-document weights generated from the Text Filtering node.

We first conduct two exploratory topic analyses on the whole corpus, one with a Text Parsing node using the default options, and another has its Text Parsing node filter out all PoS but nouns, verbs, and adjectives. The Text Parsing nodes are followed by Text Filtering nodes which then connect to Text Topic and Text Cluster nodes. Since a Text Cluster node automatically determines the number of groupings, only one is needed. In contrast, Text Topic nodes generate the exact predefined number of topics, therefore we try different number and select the best one based on the meanings of the formed topics. The EM diagram for this part is shown in Figure 1.

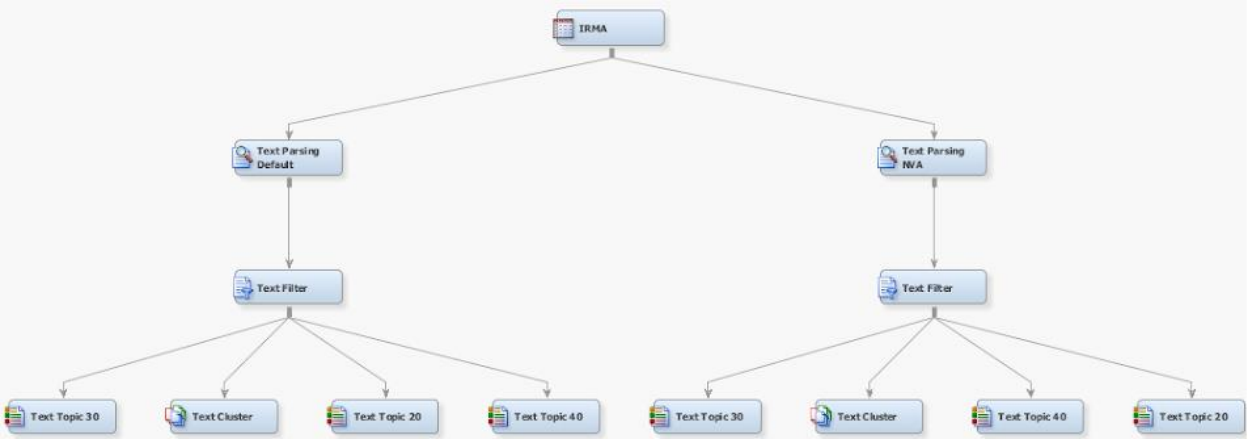


Figure 1. Exploratory Topic Analysis Diagram

The default Text Parsing node is unable to filter out URLs or emojis (in their text forms) which results in noisier processed data and poorer topic models. Some examples of terms parsed from the corpus using each parsing method are show in Table 1. Additionally, we observe that in both analyses, the formed clusters are not as desired. For instance, the key words of the topics are either incoherent or not meaningful (they are URLs or emojis). Moreover, although some topics related to Irma are detected, they have mixed key words from other subjects. Examples of these models' results are shown in Appendix A – Exploratory topic analysis results. Therefore, we decide to perform our topic analysis on a corpus of unique tweets only.

From observations made in the exploratory analysis, we continue our topic analysis on a corpus of over 2.1 million unique tweets, and with the preprocessing step eliminates all types of terms except for nouns, verbs, and adjectives. The EM diagram for this analysis is shown in Appendix B – Topic analysis EM Diagram. As a text cluster node forms hard clusters which allow a document to belong to only one topic which is too strict in general, we do not use any text cluster nodes this time.

Default Parsing				Parsing only Nouns, Verbs, Adjectives			
Terms	**Weights**	**Terms**	**Weights**	**Terms**	**Weights**	**Terms**	**Weights**
rt	0.043	not	0.151	ed	0.146	amp	0.208
ed	0.151	Irma	0.15	hurricane	0.209	evacuate	0.227
bd	0.16	b8	0.192	miami	0.236	time	0.241
amp	0.212	https://...	0.253	accept	0.246	know	0.248
b2	0.272	fe0f	0.267	power	0.256	storm	0.261

Table 1. Examples of Terms Generated in Each Text Parsing Node

The results for the 20-topic model (which generates 20 topics) is shown in Figure 2. We consider this model to be better than others based on its splitting in terms of interpretability of the topics. Overall, there are four topics related to the hurricane detected by the 20-topic model (topic 1, 6, 10, 16). Topic 1 and 16 are relatively general toward Irma, while topic 6 mentions power outages caused by the hurricane, and topic 10 consists of good luck wishes from people. We censor topic 15 as its key words are profanities.

Topic ID	Topic
1	+hurricane,+hit,+update,irma,+category
2	+check,+follow,automatically,+person,unfollowed
3	+ad,+miss,+cry,+girl,+look
4	+job,+hire,+late,+great,+open
5	+know,+thing,wanna,+miami,+feel
6	+power,+back,+lose,+restore,+outage
7	+love,+guy,+man,+girl,+friend
8	+video,+post,+add,+playlist,+photo
9	+good,+morning,+good morning,+thing,+friend
10	+safe,+stay,+stay safe,+hope,+stay
11	+want,+work,+detail,+click,+life
12	+day,today,+few,last,+week
13	+time,first,last,+great,+first time
14	lol,+feel,+right,+watch,+man
15	[redacted]
16	+miami,+storm,+live,+wind,+leave
17	+happy,+birthday,+happy birthday,+love,+hope
18	+people,+thing,+life,+hate,+help
19	+god,+love,+bless,+life,+pray
20	+look,+good,+work,+thing,+man

Figure 2. Topics Detected from the Tweet Corpus

To further determine discussed topics in the IRMA related tweets, we filter the tweets from the four mentioned topics and conduct another topic modeling analysis. Again, multiple text topic nodes with different number of topics (from 20 to 40) are used. The results for the 20 topics are shown in Figure 3. We can observe that there are no new topics detected in this corpus subset; all are resulted from the four topics identified previously. Moreover, some topics get mixed key words that refer to different subjects. As a result, we do not consider this model in further analysis.

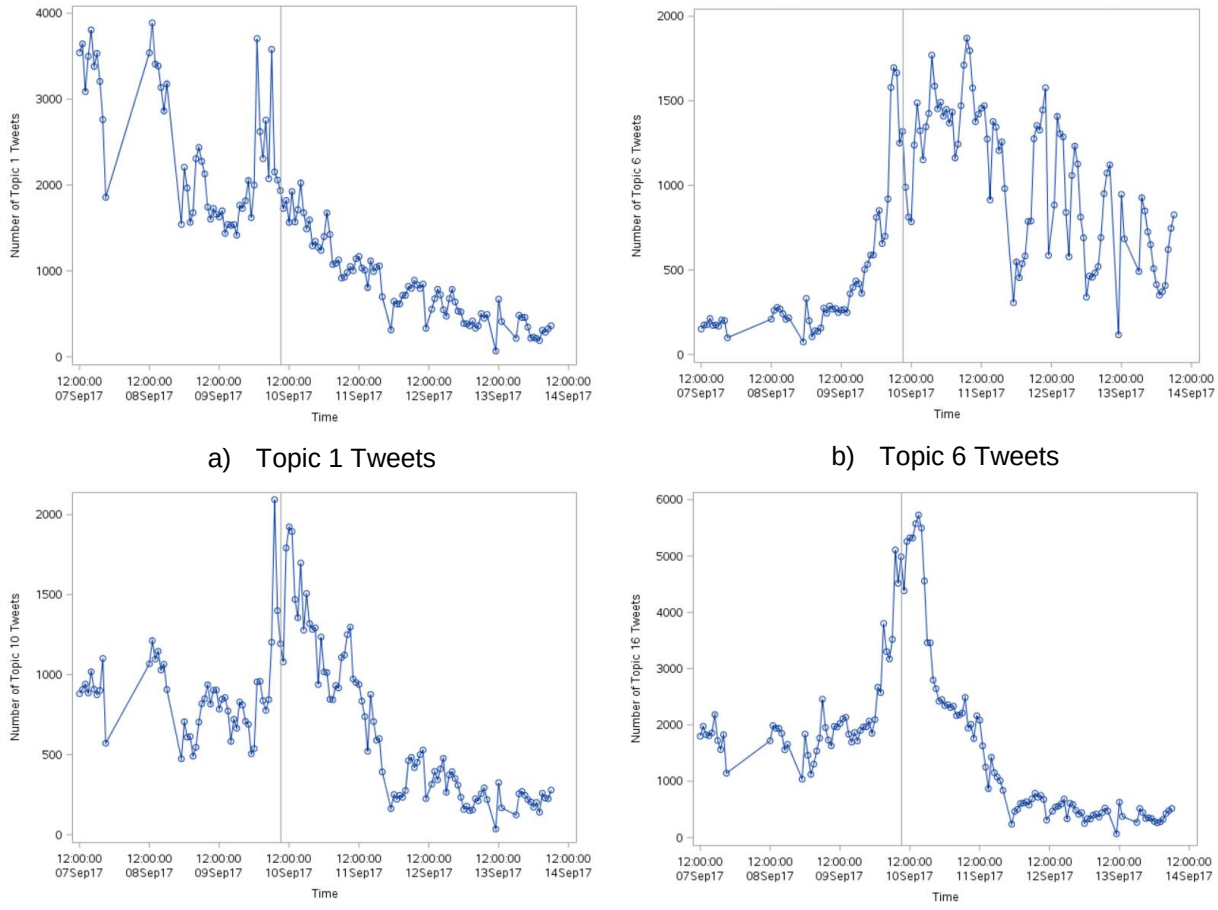
Topic ID	Topic
1	ed,+ad,+power,+miami,+storm
2	+stay,+safe,+stay safe,+close,+update
3	+power,+outage,+power outage,+update,+restore
4	+miami,+job,+hire,+late,+know
5	+stay,+safe,+safe,strong,+want
6	+storm,+surge,+storm surge,tropical,+calm
7	+back,+power back,+power,+want,+work
8	+hurricane,+know,+update,+hit,+eat
9	+hope,best,+safe,+safe,+well
10	+live,+life,+watch,+update,+area
11	+family,+friend,+prayer,+safe,+know
12	+wind,+hour,+gust,+mile,wind
13	+safe,+keep,glad,+god,+know
14	amp,+search,+close,+hurricane,+miami
15	+lose,+power,+house,+watch,+people
16	irma,hurricaneirma,florida,+hurricane,+miami
17	lost,lost power,+power,+house,+god
18	+leave,+home,+house,+hour,+evacuate
19	+pray,+safe,+god,+prepare,+people
20	+day,+hour,+power,+few,next

Figure 3. Topics Detected from Tweets Related to Hurricane Irma

TREND ANALYSIS

EXPLORATORY PLOTS

After identifying topics 1, 6, 10 and 16 as topics which contain hurricane related tweets, hourly plots are created to see the frequency of tweets which pertain to each of these 4 topics and shown in Figure 4. We can see that tweets pertaining to topic 1 (hurricane, hit, update, Irma, category), topic 10 (safe, stay, hope) and topic 16 (Miami, storm, live, wind, leave) follow similar patterns. The vertical line in each figure shows the time of the hurricane hitting the Florida Keys. Shortly after the hurricane hits the mainland in Florida, the frequency of these tweets sharply declines. The number of tweets in Figure 4a seems to have two peaks, one at approximately one day before the hurricane hitting mainland Florida and another as the



a) Topic 1 Tweets

b) Topic 6 Tweets

c) Topic 10 Tweets

d) Topic 16 Tweets

Figure 4. Trends of Irma-Related Tweets

hurricane was approaching. Figure 4b shows that the number of tweets relating to topic 6 (power, back, lose, restore, outage) picks up in the hours as the hurricane approaches the mainland, and remains near its peak for approximately 1 day, before gently declining. This intuitively makes sense, as people will lose power shortly before and during a storm like Hurricane Irma.

Since there are some periods of time where tweets were not collected and other times when the Twitter API rate limit was met so the rate of data gathering was slowed down, it is useful to look at the relative frequency of the generation of Hurricane Irma-related tweets. This standardized series is shown in Figure 5. We see a smoother pattern to the hours than the previous figures where there are relatively large jumps from hour to hour. At its maximum, approximately 25% (approximately 10,000 tweets per hour) of all collected tweets pertained to Irma.

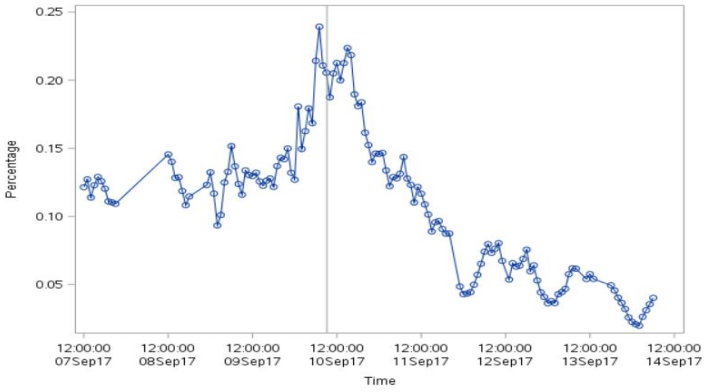


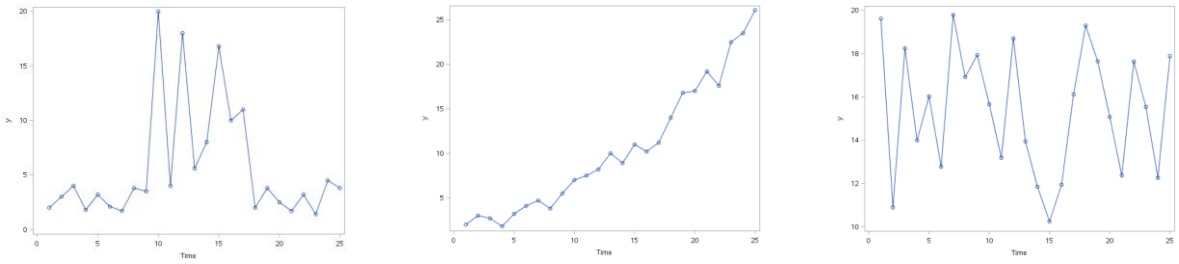
Figure 5. Relative Frequency of all Irma-Related Tweets

TIME SERIES MODEL FOR TWEET TRENDS

After viewing the trends of several of the topics in the data, it is next desired to create a model which can be used for estimation and prediction of how many tweets are generated based on previous time points' activity. In this paper, we use the Autoregressive Integrated Moving Average (ARIMA) model to analyze the tweet trends. An $ARIMA(p,d,q)$ model is an $ARMA(p,q)$ model applied on a time series differenced to the d^{th} order. An $ARMA(p,q)$ process is mathematically expressed as in equation (1):

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_p e_{t-q} + e_t \quad (1)$$

where ϕ_0 is a constant, ϕ_{t-i} and θ_{t-j} are model coefficients,y_t is the value of the series at time point t, $y_{t-\mathrm{i}}$ is value of the series at the i^{th} lag of time point t, $e_{t-j} = y_{t-j} - \hat{y}_{t-j}$ represents the shock at the j^{th} lag of time point t and is assumed to be white noise. An $ARMA(p,q)$ process assumes the time series to be stationary. This means that the series must have constant mean and variance over time, and its autocorrelation function depends only on time differences. Some examples of nonstationary and stationary series are seen in Figure 6. Figures 6a and 6b show two non-stationary time series – with non-constant variance and mean, respectively. Figure 6c shows an example of a stationary series which exhibits both constant variance and a constant mean.



a) Example of a Non-Stationary Series

b) Example of a Non-Stationary Series

c) Example of a Stationary Series

Figure 6. Examples of Stationary and Non-Stationary Time Series

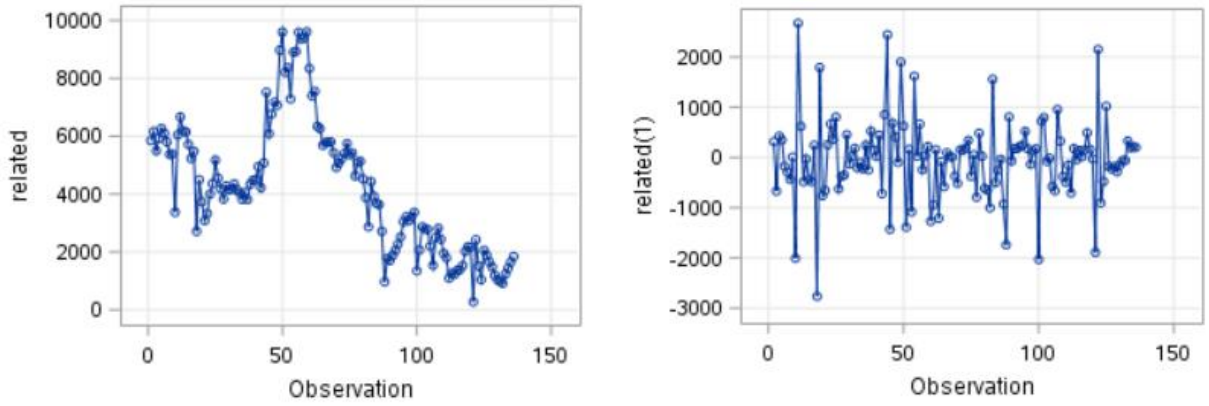
ARIMA models use differencing to fix non-stationarities in the data. The d^{th} order difference of a time series is expressed in equation (2):

$$\nabla^{\mathrm{d}}(y_t) = \nabla^{\mathrm{d}-1}(y_t) - \nabla^{\mathrm{d}-1}(y_{t-1}) \quad (2)$$

with the first order difference as $\nabla(y_t) = y_t - y_{t-1}$. Integrating the d^{th} order difference in a time series model, an $ARIMA(p,d,q)$ can be represented in equation (3):

$$\nabla^{d}(y_t) = \phi_0 + \varphi_1 \nabla^{\mathrm{d}}(y_{t-1}) + \cdots + \varphi_p \nabla^{d}(y_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_p e_{t-q} + e_t \quad (3)$$

The general process of modeling time series would be to check the original series if it is stationary. This can usually be viewed by plotting the series. When the series is not stationary, taking a first order difference will often lead to a stationary one. Figure 7 shows the original series of Irma-related tweets and the differenced series of Irma-related tweets. Figure 7b plots the change in frequency of tweets between consecutive hours. We can see that the original series in Figure 7a does not have a constant mean over time, and that there is a clear pattern in the values for hours around when the hurricane first made landfall in Florida. Figure 7b shows that many of these problems are removed by taking the first order difference. We now have a series with a fairly constant mean around 0, with relatively constant variance, although it exhibits some large jumps at certain time points.



a) Original Series b) First Order Differenced Series

Figure 7. Plots of Original Series and First Order Differenced Series

In SAS, we can use PROC ARIMA to help guide which type of ARIMA model to perform. The SCAN (smallest canonical) and ESACF (extended autocorrelation function) options in PROC ARIMA are methods which can help identify the orders in the ARIMA model. We can give PROC ARIMA the original series as seen in the sample code below:

```
proc arima data=dat;
   identify var=series esacf scan;
run;
```

We can also give PROC ARIMA a differenced series by putting the order of the difference in parentheses for the series values as below:

```
proc arima data=dat;
   identify var=series(1) esacf scan;
run;
```

The output will be tables which recommend the values of p, d and q for an ARIMA model. Figure 8 shows the two tables which are results of running the above code samples. Figure 8a shows the suggested parameters for an ARIMA model on the original series. We can choose the best values for p, d and q by selecting the combination which results in the lowest Bayesian Information Criterian (BIC) [2]. The BIC is similar to the Akaike Information Criterion (AIC) [2] with slightly different penalty term. Here we see that it is recommended to choose an ARIMA model where $p + d = 2$. Figure 8b shows the recommended values for p and q are on the differenced series, which in turns suggests $p + d = 1$. Therefore, an $ARIMA(1,1,0)$ model is a good candidate for modeling this series. We also try an $ARIMA(0,2,0)$ model on the original series which turns out to be a worse model in terms of AIC (note that models with smaller BIC and AIC are generally more favorable).

ARMA(p+d,q) Tentative Order Selection Tests					
SCAN			ESACF		
p+d	q	BIC	p+d	q	BIC
1	1	13.18142	1	1	13.18142
2	0	13.13757	2	1	13.1632
			3	1	13.16343
			5	2	13.26586

a) Recommended Parameters for Original Series

ARMA(p+d,q) Tentative Order Selection Tests					
SCAN			ESACF		
p+d	q	BIC	p+d	q	BIC
1	0	13.11989	0	1	13.14532
0	1	13.14532	1	1	13.15151
			4	2	13.24262
			5	2	13.26629

b) Recommended Parameters for First Order Differenced Series

Figure 8. Choosing the Parameters for ARIMA Model

In addition to the series itself, indicator variables measuring of how close to the hurricane making landfall that the tweet was sent are included. Indicators are created reflecting whether the tweet is more than two days before the hurricane making landfall, between one and two days before landfall, within one day of landfall, within one day after landfall, between one and two days after landfall, and more than two days after landfall. The overall model is estimated and shown in Figure 9 which can be summarized by equation (4):

$$\nabla(y_t) = -75.781 - 0.296 * \nabla(y_{t-1}) + 253.79 * Indicator_{within\_1\_day\_prior} \quad (4)$$

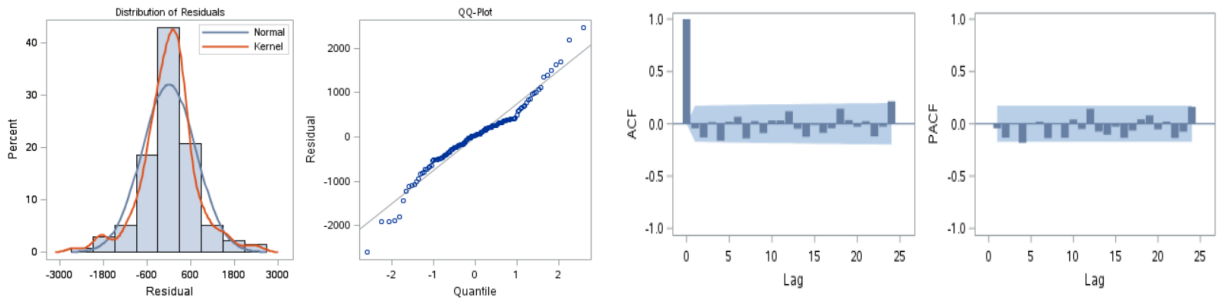
The $AR1{,}1$ term in the model refer to the coefficient of the first autoregressive term. Again, since the first order difference is taken on the data, here we are predicting the difference $\nabla(y_t) = y_t - y_{t-1}$. The coefficient of -0.29588 on the $AR1{,}1$ term shows that $\nabla(y_t)$ is negatively correlated with $\nabla(y_{t-1})$ (the previous time point's difference). This intuitively means that the differenced series oscillates around zero (when the previous time point's difference is negative, this quantity is positive, and when the previous time point's difference is positive, this quantity is negative). The only significant indicator is for tweets created within one day of the hurricane making landfall at 9 am on the morning of 9/10/17. The coefficient of this indicator is approximately 254, signifying that the predicted difference $\nabla(y_t)$ in tweets at time t is 254 higher for an hour which is within one day of the hurricane making landfall. We can see that this estimate is significant at the 0.10 level of significance, indicating that there is a statistically significant correlation between the number of hurricane related tweets and the time directly before the hurricane making landfall.

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	-75.78113	55.18096	-1.37	0.1720	0	related	0
AR1,1	-0.29588	0.08318	-3.56	0.0005	1	related	0
NUM1	253.79098	131.64702	1.93	0.0560	0	within_1_before	0

Figure 9. Final Model Parameters

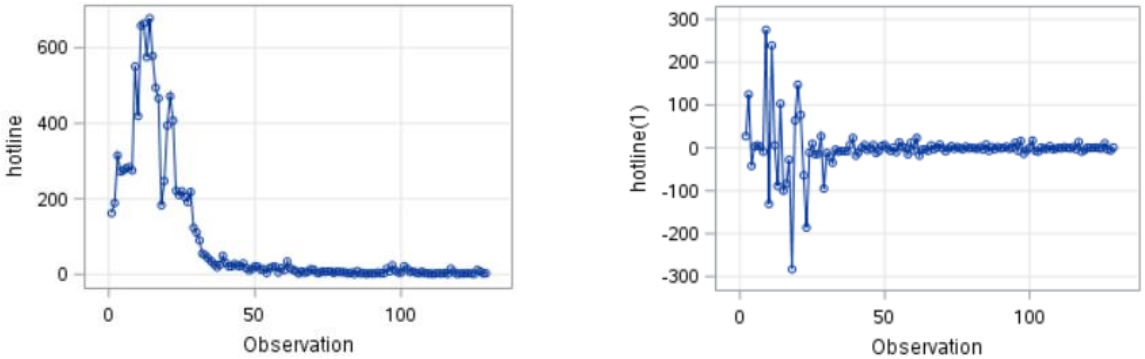
Finally, we perform a residual analysis of the model to verify whether it satisfies the assumptions of a time series model. Figure 10 shows plots checking for normality of the residuals and whether there is correlation among the residuals. Figure 10a shows that the residuals are approximately normally distributed. Figure 10b tests whether any of the lagged time points are correlated with the current time point and indicates no statistically significant autocorrelation is detected. Overall, the assumption of white noise residual is verified, showing that the built model is a good fit for the data.

Additionally, we examine the trend of tweets related to evacuation hotlines. One example of these tweets is "*RT @FLGovScott: If you are concerned that you do not have a way to evacuate please call our transportation hotline at 1-800-955-5504.*". This series and its first order difference is shown in Figure 11. As can be seen, both the original and differenced series are not stationary. Consequently, the number of tweets fluctuates close to zero for about two-thirds of the recorded period. Overall, we believe it is not necessary to have a model for this kind of trends and stop the analysis.



a) Residual Diagnostics b) Test for Correlation among Residuals

Figure 10. Residual Diagnostics for $ARIMA(1,1,0)$ Model



a) Evacuation Hotline Tweet Trend **b) First Order Difference**

Figure 11. The Trend for Evacuation Hotline Tweets and Its First Order Difference

DISCUSSION AND CONCLUSION

Using text mining techniques, we are able to extract tweets related to the hurricane Irma in our corpus. One important thing we want to note is the preprocessing of tweet data. Since tweets are not formal documents and can consist of noises like URLs or emojis, filtering out all types of terms except for nouns, verbs, and adjectives are essential to improve analysis results. Some limitations to the analysis are the gathering of the data. Even though we can use a cost-free API to gather tweets, we frequently exceed the rate limit allowed by Twitter when gathering this much data. Thusly, the data which is collected in this study should be treated as a sample of all the tweets generated.

In the tweets mentioning Irma, we detect four topics, among which, two are general discussion toward the hurricane, one mentions the power outages resulted from the storm, and one consists of hopes and prayers from the Twitter users. We also detect a small number of tweets and retweets sharing important information such as evacuation hotlines and price fraud hotlines, however they do not form a separate topic due to the lack in amount.

In using these results for policy recommendations, it would be useful to track similar natural disasters to create a baseline trend of what is expected from peoples' reactions to a storm. For instance, in this analysis we see that the peak activity of tweets pertaining to the hurricane occur right as the hurricane made landfall in the southern tip of Florida. We would like to know if this is expected behavior or if for other storms people react earlier or later. We also have some trending for a hotline pertaining to evacuating. It would be interesting to see if for other natural disasters there are more hotline related tweets or if they are sent for longer periods of time. This can then be correlated to the number of users discussing topics related to the respective hotline. This analysis is to be used as a case study to assess public attitudes and responses throughout similar events.

REFERENCES

[1] Rice, Doyle. “Natural disasters caused record $306 billion in damage to ...” USA Today, 8 Jan. 2018, www.usatoday.com/story/weather/2018/01/08/natural-disasters-caused-record-306-billion-damage-u-s-2017/1012924001/.

[2] Adhikari, R., & Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting. arXiv preprint arXiv:1302.6613.

[3] Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. Numerische mathematik, 14(5), 403-420.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Bogdan Gadidov
Analytics and Data Science Institute, Kennesaw State University
(678) – 768 – 0318
bgadidov@students.kennesaw.edu

Linh Le
Analytics and Data Science Institute, Kennesaw State University
(678) – 431 – 9581
lle12@students.kennesaw.edu

APPENDIX A – EXPLORATORY TOPIC ANALYSIS RESULTS

Cluster ID	Descriptive Terms
1	rt better +open [illegible] +mother +window shvdmg +[illegible] window' https://t.co/kgixbcatxz +find +back +help jrdquotes ...
2	rt sdbyxx sadieisonfire +mood jdhonda11 windooooooww +miss king_talent +love +episode +look retweet tha_kidd_...
3	+power +help flgovscott +amp fl +stay florida +nurse nigga +safe https +know ananavarro +shelter +people ...
4	+miami hurricaneirma +great +beach +water devinsenaui +arrival +worship +watch +wind +wall +storm +shelter w...
5	rt irma florida +hurricane +accept +offering suplexclty[illegible] https://t.co/fpk3zqzmlz hurricane nwskeywest +key +safe +...
6	ed a0 bd b8 bc a9 b9 bf https://t.co/ucrm7yrn3m b3 niggacommentary fe0f b2 +ad +love ...
7	rt +long +know +shit +relationship +count +distance jackedyotweets +'long distance relationship' https://t.co/eh8bzrgr...
8	rt jackedyotweets +[illegible] +man +job +dog patrick https://t.co/ckm2vbkehz [illegible] [illegible] [illegible]commentary +president ...
9	a0 ed bd b8 bc b4 bf b2 fe0f b9 b1 +ad b7 niggacommentary a5 ...
10	+day +love +want pl +video +look mitchellvii today +work lol +trump +thing youtube +life +feel ...
11	rt +evacuate +time +amp +tree down +blow +fall spearsiobhan https://t.co/gppbeghdu6 +floridian atlanta californiab...
12	+people rt +[illegible] iamdjlive dying single 'single [illegible] +misplace https://t.co/wnr4auggjh +check automatically +follow u...

Figure 12. Text Clustering Results, Default Text Parsing Node

Topic ID	Topic
1	+day,throb,professio,laying,[illegible]
2	tourist,senate speech,https://t.co/kkc1x9ziwb,animal hospital,nugg
3	https://t.co/6ednzcdfs3,_speedyt,downtownmiami,estimation,https://t.co/b15kwklpwu
4	https://t.co/thivhaaaro,sheila_sheiley,complete trash,happy ooonnnneeeee,kittyknits
5	fox35,epic,https://t.co/5ahwac4drg,smybirthday,https://t.co/2lcjvxsv91
6	gamscout,mumblesenrique,luchakli,+upholster,https://t.co/mbqjedacty
7	rescue long term,valemunoza,https://t.co/c1lxq8xfte,valley_witch,+duty
8	+incredible volunteer,https://t.co/7yspq0qv0q,+varsity boy,+re-energize,eco life
9	yes yes yes,https://t.co/gq2fsalpvo,nc_governor,https://t.co/kboxhx3z2v,dragon ball fighter
10	elenacardin43,professio,laying,matthayescfb,eelowan
11	wweperformctr,https://t.co/r7ylhbpsga,+bore hurricane,sunshineleisha,+intern
12	daid,relative safety,https://t.co/zwrn61jrqt,lovely_ltece,quality content
13	countryside_cc,+telepath,full pool,nugg,https://t.co/2icx9fti2o
14	+big venue,https://t.co/kk4uiv386g,https://t.co/uz12zd6qar,https://t.co/3khqhwjjno,legroff
15	audible sigh,politic,iamjlov3,bredren,https://t.co/svgjs3qwhx
16	https://t.co/2sqjo6knwr,https://t.co/frywsyuxcf,nbcmarlon,https://t.co/mru8csjqb,+typical female
17	https://t.co/2icx9fti2o,theeomniscient1,https://t.co/2xqnuboinf,thigh high water,https://t.co/wqipaj
18	https://t.co/ihjldzqt7k,fugly,https://t.co/mxpyoks5qt,https://t.co/mewzaljoeh,sabrybotler
19	wanna sell,postirmabucketlist,https://t.co/trtzazszec,+pac,https://t.co/jjonl5kbxp
20	some serious work,principals,retweet button,gbloskas,minute

Figure 13. 20-Topic Results, Default Text Parsing Node

Cluster ID	Descriptive Terms
1	+tree down +blow +fall +find +back +help +station gas +'gas station' broken +dollar retweet +puppy left ...
2	ed +ad +love +man +scream +end +racism +god +look +[illegible] +baby beautiful +work [illegible] +happy ...
3	+accept +offering vegun refus sydneylai leggedfriends malemalefica uptownlizbrown antwanhoberg bikinatroll malbapalo dima...
4	+evacuate +time +power +floridian +trump +amp ananavarro mitchellvii gt +evacuation +die +traffic +media +white +play ...
5	+real +meet +want +video +thing [illegible] +bad +protect +girl +leave +fate +look +life [illegible] +live ...
6	+hurricane +amp +storm +wind +miami https +down +update +good +eye +move +water +surge +power +area ...
7	+day +love +look +work +year +miss +god +start +week +close +life +want +school +friend +episode ...
8	+miami +long +shit +relationship +count +distance +'long distance relationship' +great https +beach +good +arrival +worship ...
9	ed +know [illegible] +help [illegible] +man +amp +dog +nurse [illegible] +shelter +twitter +text +state +thought ...
10	+open better +window +mother [illegible] shvdmg [illegible] window' +feel +job today first +late +hire +tonight +work ...
11	+watch +water +wall +back +great +miami +job +find +store +help +feeling +open +good +owner +people ...
12	+safe +people +stay +key [illegible] single +shelter dying 'single [illegible] +misplace +follow +safety +animal +check +file ...

Figure 14. Text Clustering Results, Manual Text Parsing Node

Topic ID	Topic
1	+lil baby,correct decision,lowes,spooky sports,eagankemp
2	atonement,+free treat,durin,hour wait,+comfort zone
3	+dodger teammate,automatically,bumanat,+slender arm,damn world
4	hateful,+wax,bringi,pcbearsfootball,real [illegible] bruh
5	mek,emilyboxing,stay focus,+well timeline,soardogg
6	wyldweasil,nicely,beautiful mess,horse [illegible] findom
7	water,spooky sports,tough life,mathewdavid,+absolutely well thing
8	fried oreos,+loud music,lifestyleblogger,+hesitate,+palm
9	heavy ion,sneaky sex,+steel,+dark age,+duty
10	accom,hot knife,swfl emergency response,itsalgee,+undergo
11	haitian hillary,reserved,belle glade,banana rocket launcher,nerette
12	printmag,strong right ear,sewage spill,hate waiting,+source
13	+can veggie,+vent,spea,+win baby,yanatha
14	celebrationfl,church property,calizzzlee,meep,+entire key
15	aggravated fast,fried oreos,ip,+apple jack,+former football team
16	+polar pod,+private account,+pixar movie,green flag,mow
17	+some report,stay positive,water,visitlondon,topanga
18	+psn trophy,live club atrium,lil belly,+presale participant,haitian hillary
19	+waste,+lil baby,chinese ev,+regenerate,batteryjackson
20	austinwolffff,bruh [illegible],fineartameri,trippieredd,good beer

Figure 15. 20-Topic Results, Manual Text Parsing Node

APPENDIX B – TOPIC ANALYSIS EM DIAGRAM

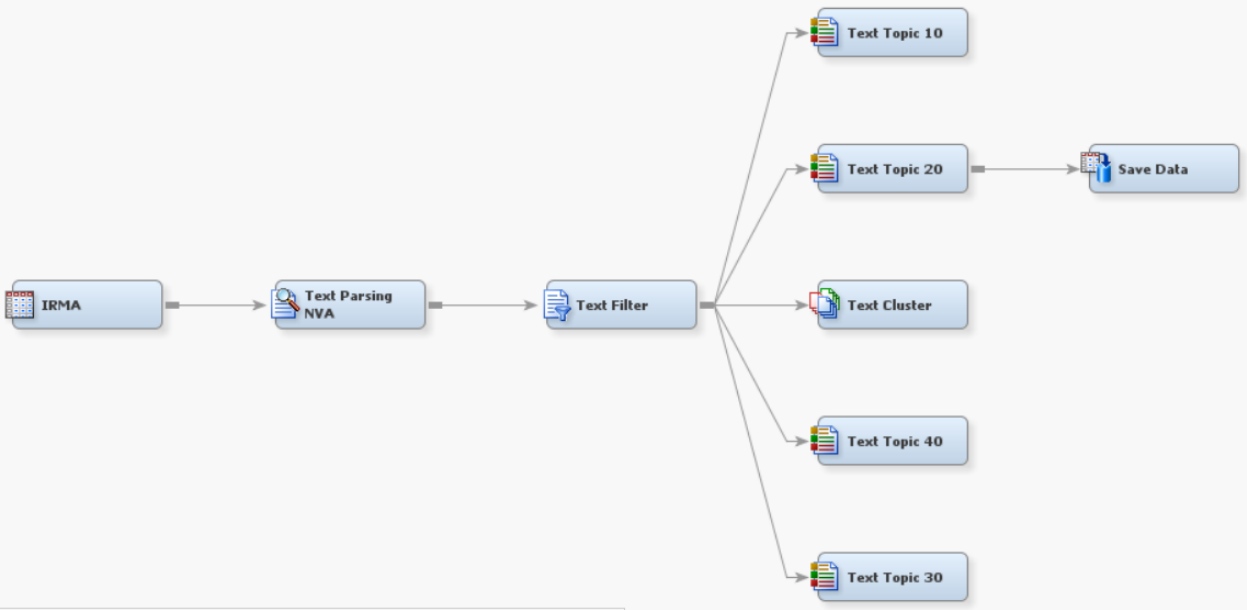


Figure 16. EM Diagram for Topic Analysis