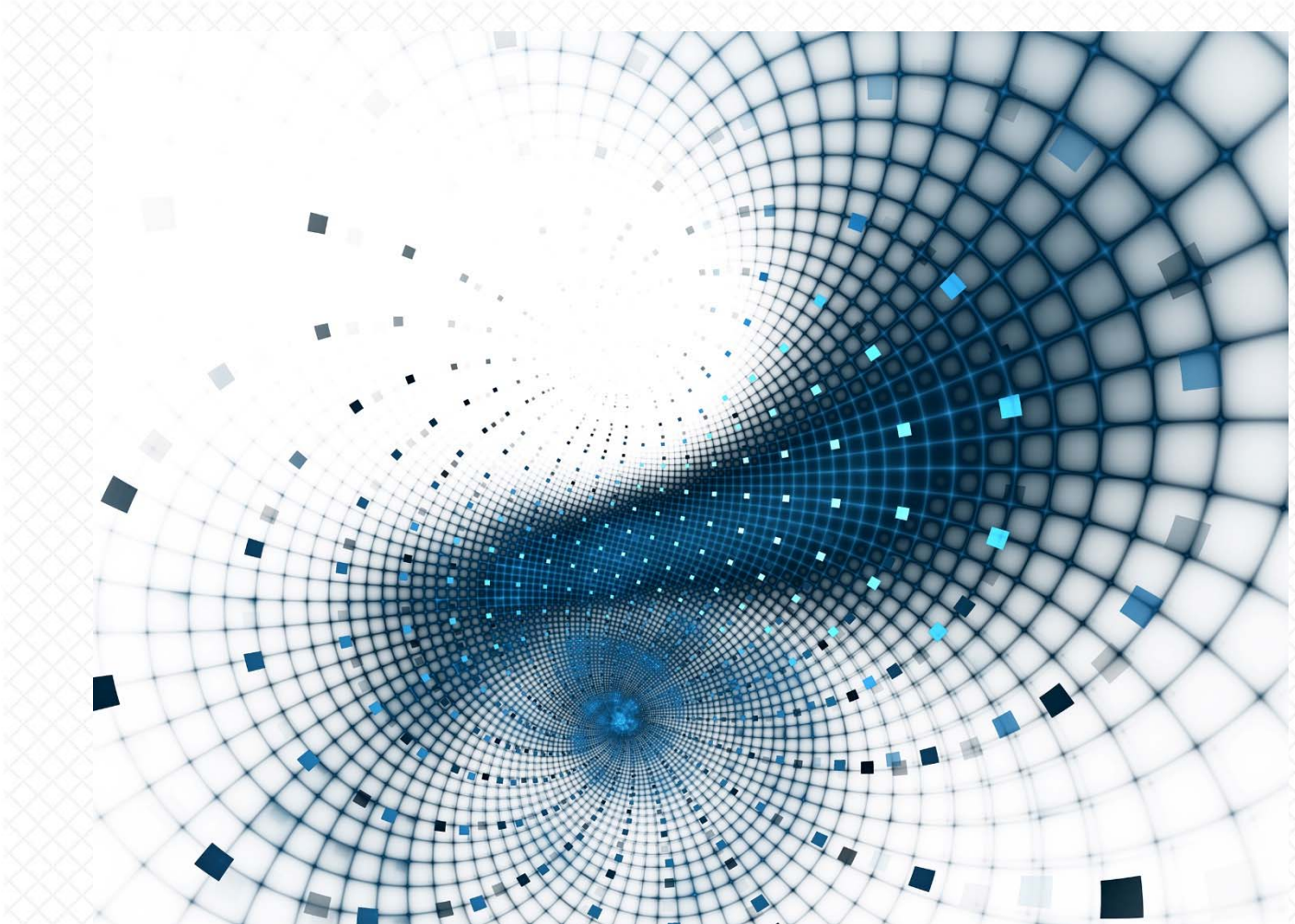




Social Media Analysis of Catastrophic Response: Twitter Data Analysis of Tornadoes





Social Media Analysis of Catastrophic Response: Twitter Data Analysis of Tornadoes

AUTHORS

R. Dale Hall, FSA, MAAA, CERA, CFA
Steven Siegel, ASA, MAAA
Kailan Shang, FSA, CFA, PRM

Caveat and Disclaimer

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information.

Copyright ©2016 All rights reserved by the Society of Actuaries

TABLE OF CONTENTS

Section 1: Tornado Twitter Data	4
1.1 Data Source	4
1.2 Data Processing	5
1.3 Word Cloud	6
Section 2: Catastrophe Communication Signature.....	8
Section 3: Geolocation Analysis	12
Section 4: Insurance Related Twitter Data	13
4.1 Author Category Distribution.....	13
4.2 Key Messages	17
4.3 Key Influencers.....	17
Section 5: Sentiment Analysis	19
Section 6: Suggestions for Future Research	20
Section 7: Conclusion	21
References.....	22
About The Society of Actuaries	23

Social Media Analysis of Catastrophic Response:

Twitter Data Analysis of Tornadoes

Social media has grown in the past decade to become an important channel for information dissemination and communication. With user-generated content, social media can provide more localized and personalized information. Social media data has been widely used for marketing, customer relationship management and custom sentiment analysis.

Growing utilization of social media in the insurance industry has been seen in the past few years. Many companies have set up official accounts in social networks to communicate with customers and investors. Most of the communications focus on company news, financial results, insurance education, and social responsibility. Some also use social media to promote healthy lifestyles and better customer behaviors that could eventually turn into reduced risks and cost of insurance.

Additionally, social media has been widely used in catastrophic events for warning, information gathering and sharing, and post-event recovery. Understanding the social media response can help the insurance industry better understand the effectiveness of early warning systems, information flow and potential loss control measures. Insurers may use this information to design better warning systems on natural disasters, pandemic and other insurance events to clients and the public. They may also use it to increase public awareness of loss prevention methods, improve customer services and monitor the risk on a real-time basis during a catastrophe.

This research report studies the social media response to three extreme tornado outbreaks: May 22, 2011 in Joplin, MO, November 17, 2013 in Washington, IL and April 9, 2015 in Rochelle, IL. It uses Twitter data with a 14-day period for each tornado event with a focus on a 24-hour period surrounding the event. It starts with an overview of the data followed by analysis of geolocation, insurance related user influence, social network, and sentiment.

Section 1: Tornado Twitter Data

Three severe tornado events in the Midwest United States were chosen for analysis. Unlike some other severe weather categories such as hurricanes, tornadoes are usually less predictable, short-lived but very destructive. The formation of the tornadoes is less understood due to their short durations. Effective social media can assist on warning and real-time status monitoring because of its quick and broad reach.

1.1 Data Source

Twitter data were chosen for study considering the large number of users, public accessibility of data and affordability.

Table 1
Three Tornado Events

Location	Joplin, MO	Washington, IL	Rochelle, IL
Date	Sunday May 22 2011	Sunday November 17 2013	Thursday April 9 2015
Start Time	5:34 PM	10:59 AM	6:40 PM
Duration	38 minutes	10 hours, 48 minutes	1 hour, 14 minutes
Severity*	EF-5	EF-4	EF-4
Casualties	158 deaths 1000+ injuries	8 deaths 190 injuries	2 deaths 30 injuries

*Severity is measured using enhanced Fujita (EF) scale.

The data were obtained from a vendor called Sifter. Sifter is one of a number of vendors that provide extracts of Twitter data for a fee based on criteria provided by a user.

Data was collected from historical Twitter data based on consistent queries. Historical Twitter data include all tweets that have not been deleted by the users at the time of querying. A 14-day period for data was used for each tornado event, starting from 3 days prior to the outbreak of the tornado. The queries used are given below.

Query for 2011 Joplin, MO Tornado

(contains:tornado OR contains:EF OR contains:hail OR contains:storm OR contains:damage OR contains:injur OR contains:kill OR contains:wind OR contains:thunder OR contains:water OR contains:flood OR contains:rain OR contains:wound OR contains:insur OR contains:loss OR contains:help OR contains:save)) AND ((MO OR contains:Missouri OR contains:joplin)

Query for 2013 Washington, IL Tornado

((contains:tornado OR contains:EF OR contains:hail OR contains:storm OR contains:damage OR contains:injur OR contains:kill OR contains:wind OR dead OR contains:thunder OR contains:water OR contains:flood OR contains:rain OR contains:wound OR contains:insur OR contains:loss OR contains:help OR contains:save) AND (contains:illinois OR IL))

Query for 2015 Rochelle, IL Tornado

((contains:tornado OR contains:EF OR contains:hail OR contains:storm OR contains:damage OR contains:injur OR contains:kill OR contains:wind OR dead OR contains:thunder OR contains:water OR contains:flood OR contains:rain OR contains:wound OR contains:insur OR contains:loss OR contains:help OR contains:save) AND (contains:illinois OR IL OR contains:Rochelle OR contains:Fairdale))

Each data record represents a tweet with meta information. Some key data fields used in the study are explained below.

Table 2
Key Twitter Data Fields

Field Name	Explanation
text	Content of the tweet
username	User name
user_bio_summary	Bio summary of the use. This can be helpful for identifying the user’s profession or industry
real_name	The “real” name that the user choose to show
posted_time	The time the tweet was posted
favorties_count	The favorite count of the tweet. It indicates the influence of the tweet.
followers_count	The number of Twitter followers of the user. It indicates the influence of the user.
friends_count	The number of Twitter friends of the user. It indicates the influence of the user.

1.2 Data Processing

The retrieved Twitter data were further filtered and cleaned to improve their relevance to the tornado events.

1. Language Filtration. Historical Twitter data querying does not provide language detection. The raw data contains significant amount of irrelevant tweets in foreign languages. For example, “IL” is used in the query when searching for Illinois tornado data. “IL” was frequently used as the location information in tweets. However, “il” means “he” in French and “the” in Italian. Most tweets in foreign languages were removed from the datasets used for analyzing

social media response. Table 3 shows the tweet count before and after the filtration. The impact of filtration varies much by query mainly because the frequency of “il” is much higher than that of “mo” in foreign languages.

Table 3
Twitter Data Language Filtration

Location	Joplin, MO	Washington, IL	Rochelle, IL
Date	Sunday May 22 2011	Sunday November 17 2013	Thursday April 9 2015
Tweet Count in Raw Data	489,652	853,173	863,391
Tweet Count after Filtration	227,268	96,853	95,174

The tweet count after filtration decreased significantly from the 2011 Joplin tornado event to 2013 Washington tornado event and followed with little change for 2015 Rochelle tornado event. Contrarily, active Twitter users increased significantly during the time period, with 85 million users in Q2 2011 and 241 million in Q4 2013 and 304 million in Q2 2015¹. A possible explanation of the reverse trend is that the tornado Twitter data volume was extremely heavy, given the severity, casualties damages and national coverage of the event. The 2011 Joplin tornado outbreak was more severe and has much higher casualties than the other two tornado events.

2. Removing stop words. Stop words are the most common words such as “is”, “at”, “who”, and “that”. They are frequently used in the data but do not provide useful information of contents of tweets. Therefore, they are removed before data analysis. The stop words in the Python package “nltk” are used to identify the stop words in tweets.
3. Stemming. Stemming is the process of removing morphological affixes from words. For example, a stemming algorithm may reduce words “rains”, “raining” and “rained” to “rain”, the root of the derived words. This helps to reduce the number of distinct words to be counted across the Twitter data and help find the most popular words and topics.

1.3 Word Cloud

A basic way of extract the key information from tornado Twitter data is to identify the most popular words based on the frequency of appearance in the tweets. The outcome is usually presented as a word cloud with the size of each word set according to its frequency. Figure 1 shows the word cloud of 2013 Washington, IL tornado Twitter data. Word frequency is listed in Table 4.

¹ Data Source: Statista (<http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users>).

Table 4
2013 Washington IL Tornado Twitter Data Popular Words

Word	Count	Word	Count	Word	Count
tornado	90,738	@pzfeed	3,151	EF	2,084
IL	50,677	Reported	3,126	East	2,082
Illinois	47,097	ground	3,067	@foreverchasin	2,036
Washington	26,217	town	2,970	destruction	1,995
near	10,748	help	2,941	@severestudios	1,990
damage	10,068	water	2,868	Path	1,961
#ilwx	9,123	aerial	2,660	@stormcoker	1,899
Peoria	7,252	new	2,621	Severe	1,881
storm	6,857	adam	2,620	Wow	1,840
warning	5,864	lucio	2,603	MO	1,833
county	7,687	outbreak	2,590	rain	1,833
video	4,527	dog	2,576	picture	8,441
Chicago	4,322	wind	2,559	view	1,757
hit	4,160	home	2,503	news	1,749
Central	4,032	relief	2,431	warnings	1,742
Roanoke	3,903	Midwest	2,339	Indiana	1,731
moving	3,766	mph	2,319	destroyed	1,713
@twcbreaking	3,493	weather	2,292	city	1,705
breaking	3,456	Pekin	2,289	rubble	1,671
victims	3,445	dangerous	2,221	touches	1,667
watch	7,491	please	2,214	house	1,646
confirmed	5,139	like	2,185	prayers	1,635
large	3,263	issued	2,167		

Popular words usually include words used in the query and useful information such as locations, most influential users in the social network and words describing the impact of tornadoes. Word clouds for the other two tornado events contain similar popular words.

Section 2: Catastrophe Communication Signature

A Catastrophe Communication Signature (CCS) is a potential way to measure the social media response to any catastrophic event by counting the tweets in 5-minute intervals over a 24-hour period around the tornado event. It starts with a 6-hour period prior to the outbreak, and completes 18 hours following the event. Additionally, it may be further broken down to consider communication volume by looking at the 6 hours “prior to” the event, a 4-hour period “during” the event, and the 14-hour period “after” the event. Figures 2 to 4 shows the CCS of the three tornado events studied. The grey columns are the period of tornadoes with darker color indicating higher severity.

Figure 2
2011 Joplin MO CCS

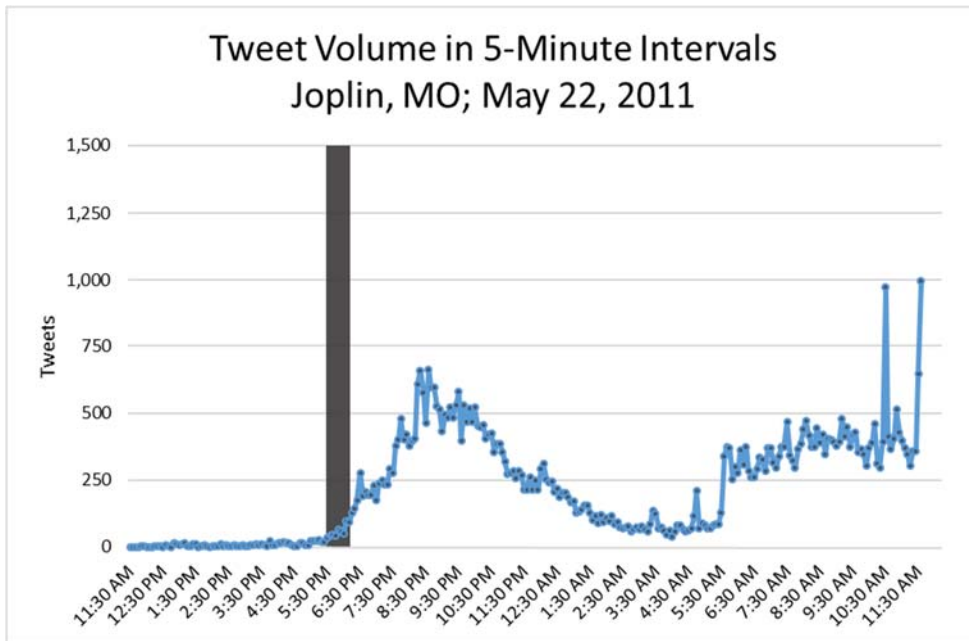


Figure 3
2013 Washington IL CCS

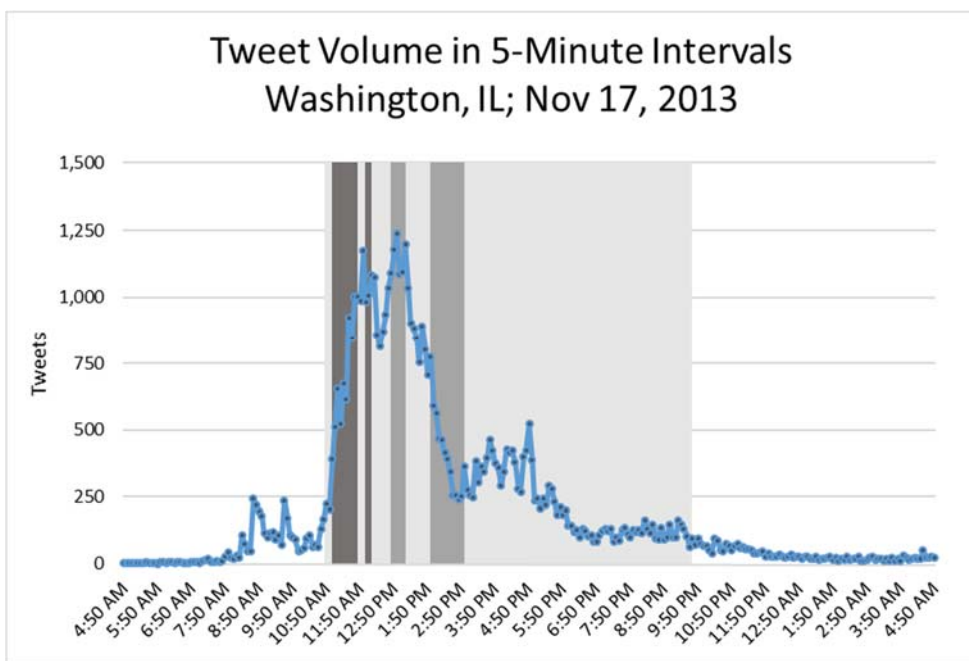
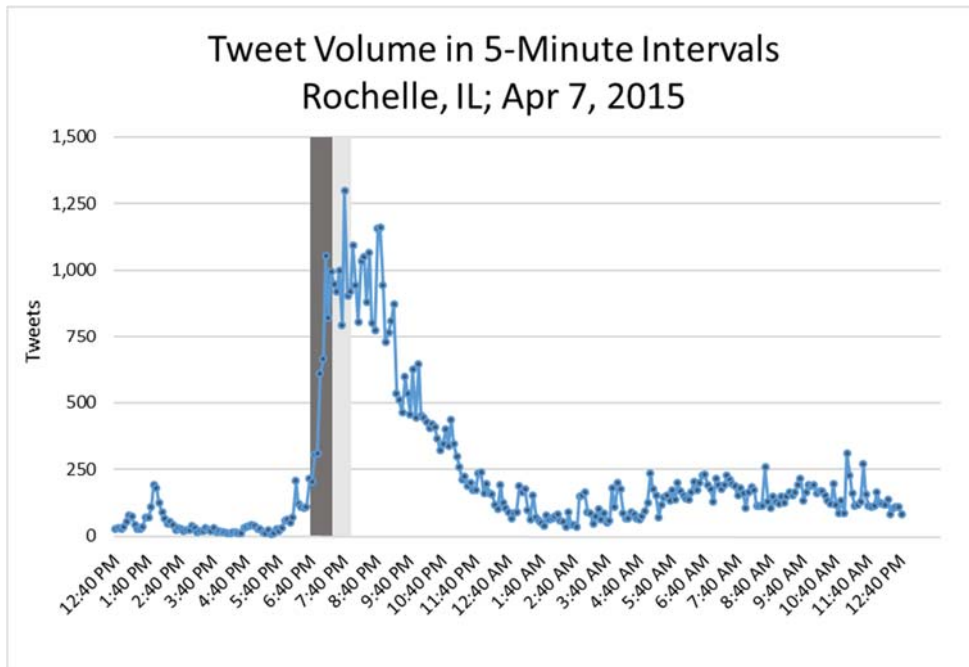
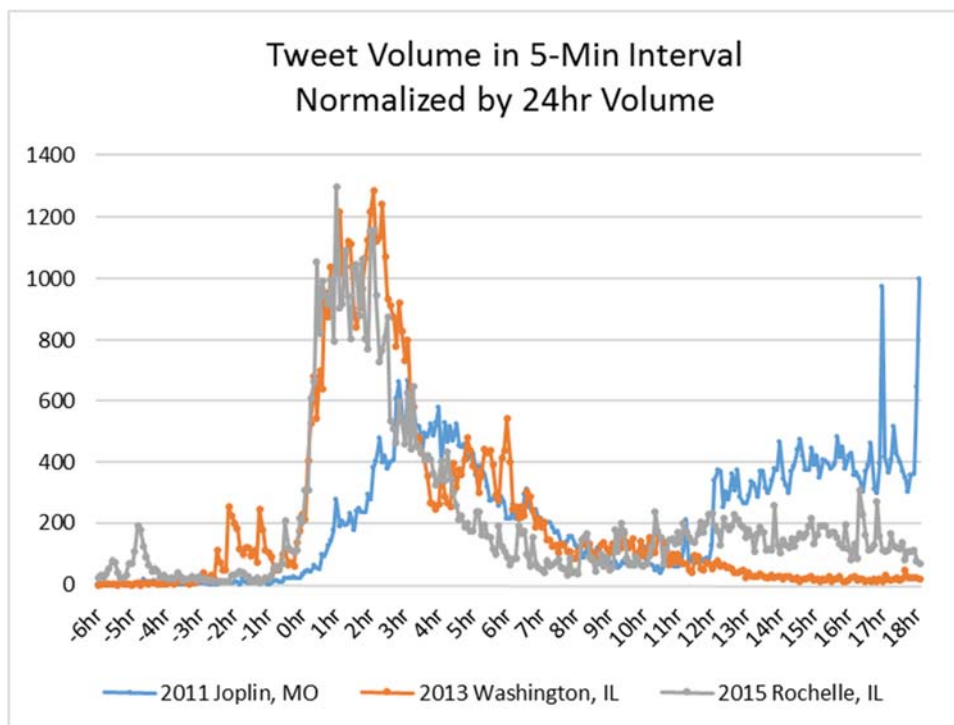


Figure 4
2015 Rochelle IL CCS



The CCS of the 2011 Joplin tornado event shows a sharp increase of tweet count after the outbreak followed by a decrease at night and a sharp increase in the next morning. The patterns spotted from 2013 Washington tornado and 2015 Rochelle tornado are similar with a spike caused by warnings few hours before the outbreak, a sharp increase during the tornado and then returned to pre-event level. Figure 5 shows a normalized comparison of the CCS of the three tornadoes. The normalization is based on the 24-hour total tweet count.

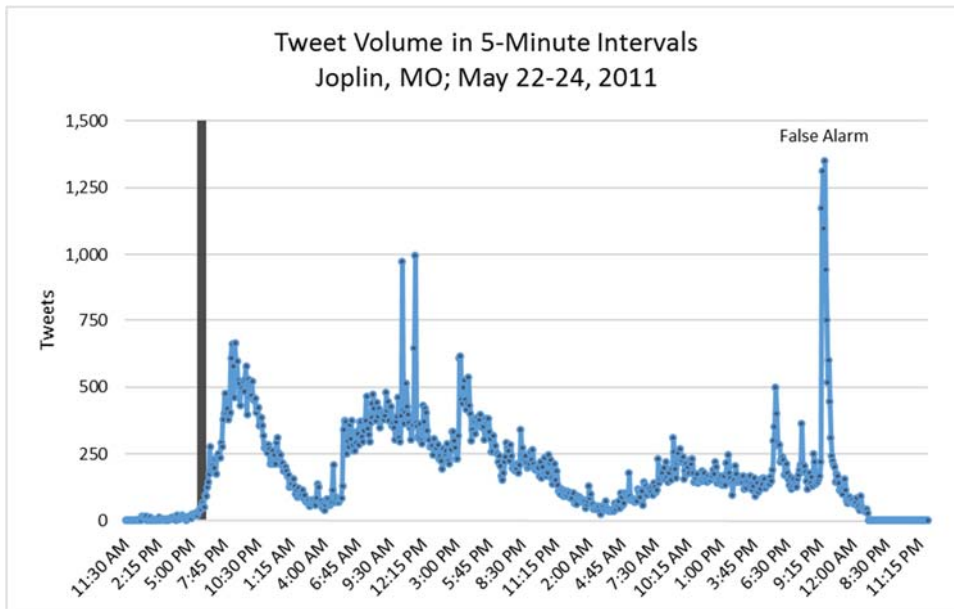
Figure 5
Normalized CCS of Three Tornado Events



One noticeable change in the CCS of the storms studied was the growing tendency of social media interaction prior to the start of the event. Today, many US federal, state and local government authorities have the ability to originate text-like messages called Wireless Emergency Alerts. Technology used ensures these alerts are not encumbered by wireless network congestion. Alerts are broadcast from cell towers whose coverage areas best match the zone of an emergency, and alert users of imminent threats to safety in their area. The National Oceanic and Atmospheric Administration’s (NOAA) National Weather Service began delivering Wireless Emergency Alerts on June 28, 2012. While the Joplin 2011 tornado had other notifications of the storm, it occurred prior to the National Weather Service’s use of Wireless Emergency Alerts. A helpful NOAA review of forward-looking ways to improve notification after the Joplin 2011 event can be found at http://www.noaanews.noaa.gov/stories2011/20110920_joplin.html .

With a longer time period in focus, 2011 Joplin tornado Twitter data shows a spike around 9:30 PM, May 24, 2011, two days after the tornado outbreak. The spike was caused by a false alarm of a new tornado.

Figure 6
2011 Joplin MO CCS over 2.5 days



Section 3: Geolocation Analysis

Geolocation analysis of tornado Twitter data can help identify the potential impact of tornado severity on social media response. However, location data sharing is voluntary for Twitter users and less than 1% of the data records contain location information. As a workaround, location information contained in the tweet texts are used as a proxy for geolocation analysis. Tweet count by location over the 14-day period for 2013 Washington IL tornado event is listed in Table 5. The tweets are divided into three categories: before the tornado, during the tornado and after the tornado.

Table 5
2013 Washington IL Tornado Tweet Count by Location

Location*	Latitude	Longitude	# of Tweets			
			Total	Before	During	After
Peoria-EF2	40.57	-89.63	7355	36	45	7274
Woodford,LaSalle,Livingston,Tazewell (Washington City)-EF4	40.62	-89.55	26761	1655	889	24217
Macoupin,Montgomery-EF0	39.23	-89.72	454	338	8	108
Clinton-EF1	38.7	-89.51	235	3	168	64
Washington County-EF4	38.42	-89.45	405	27	25	353
Christian-EF1	39.39	-89.07	228	201	1	26
Fayette,Effingham-EF2	38.99	-88.92	413	99	65	249
Grundy-EF2	41.24	-88.3	351	200	75	76
Will-EF2	41.41	-87.93	2502	459	326	1717
Jasper-EF1	38.85	-88.08	169	7	41	121
Moultrie-EF1	39.77	-88.56	11	7	0	4
Douglas-EF3	39.84	-88.28	96	7	74	15
Champaign,Vermillion,Iroquois-EF3	40.24	-88.06	725	152	383	190
Kankakee-EF1	41.3	-87.55	121	100	0	21
Jefferson-EF1	38.27	-88.78	172	86	0	86
Wayne,Edwards-EF2	38.45	-88.15	506	209	196	101
Wabash-EF2	38.38	-87.91	197	123	16	58
Massac,Pope-EF3	37.1	-88.74	311	56	169	86

* Location includes the county name and the severity of the tornado based on enhanced Fujita scale. The source coordinates information is Wikipedia (https://en.wikipedia.org/wiki/November_17,_2013_tornado_outbreak).

The correlation coefficients of the EF scale and the number of tweets by time are calculated. The correlation during the tornado is relatively high compared to other periods. Other factors such as population density and urban/suburban area could also have a material impact on the tweet count by location.

Table 6
2013 Washington IL Tornado Twitter Data Geolocation and Severity Correlation

	Total	Before the Tornado	During the Tornado	After the Tornado
Correlation	47%	38%	57%	46%

Section 4: Insurance Related Twitter Data

The involvement of the insurance industry in social media response to tornadoes is studied focusing on author categories, typical shared information and key influencers. Subsets of the tornado Twitter data were created for insurance related analysis by only including data that contain “insur” in tweet text and/or user bio summary.

4.1 Author Category Distribution

Authors are categorized into the following types based on bio summary information.

- Agent Professional Association
- Agents
- Appraisers
- Attorneys
- Claim Adjusters
- Communications Professionals (Radio, etc.)
- Contractors
- Health Care Professionals
- Individual
- Insurance Companies
- Insurance Company Association
- Insurance Department
- National Underwriter
- News Organization/Publication/TV
- Other
- Politicians
- Real Estate Closers/Notaries
- Risk Managers
- Scientific Consultants
- Travel Insurance

Figure 7 and Figure 8 shows the 2011 and 2013 tornado events tweet count by author category for five periods: More than 6 hours before the tornado, 6 hours before, 4 hours during, 14 hours after and more than 14 hours after the tornado. Most insurance related tweets were posted more than 14 hours after the 2015 tornado event and were not shown here.

Figure 7
 2011 Joplin MO Tornado Insurance Tweet Count by Author Category

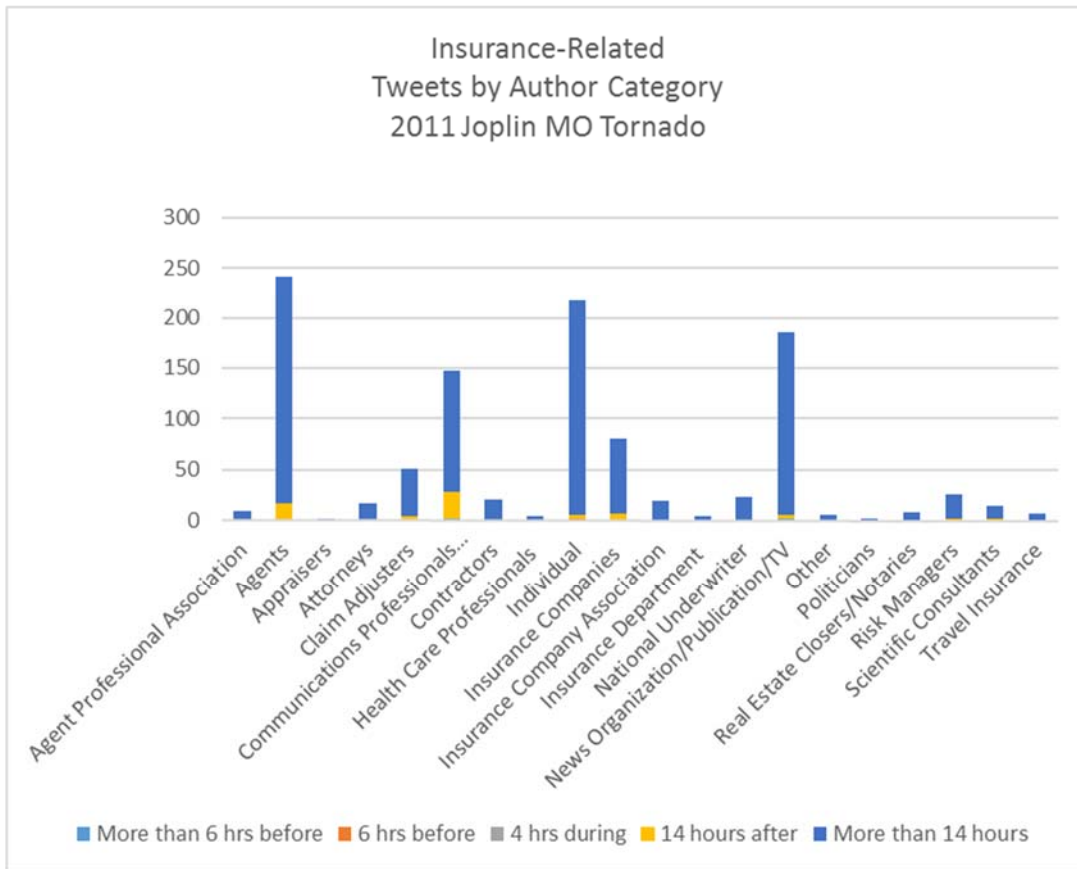
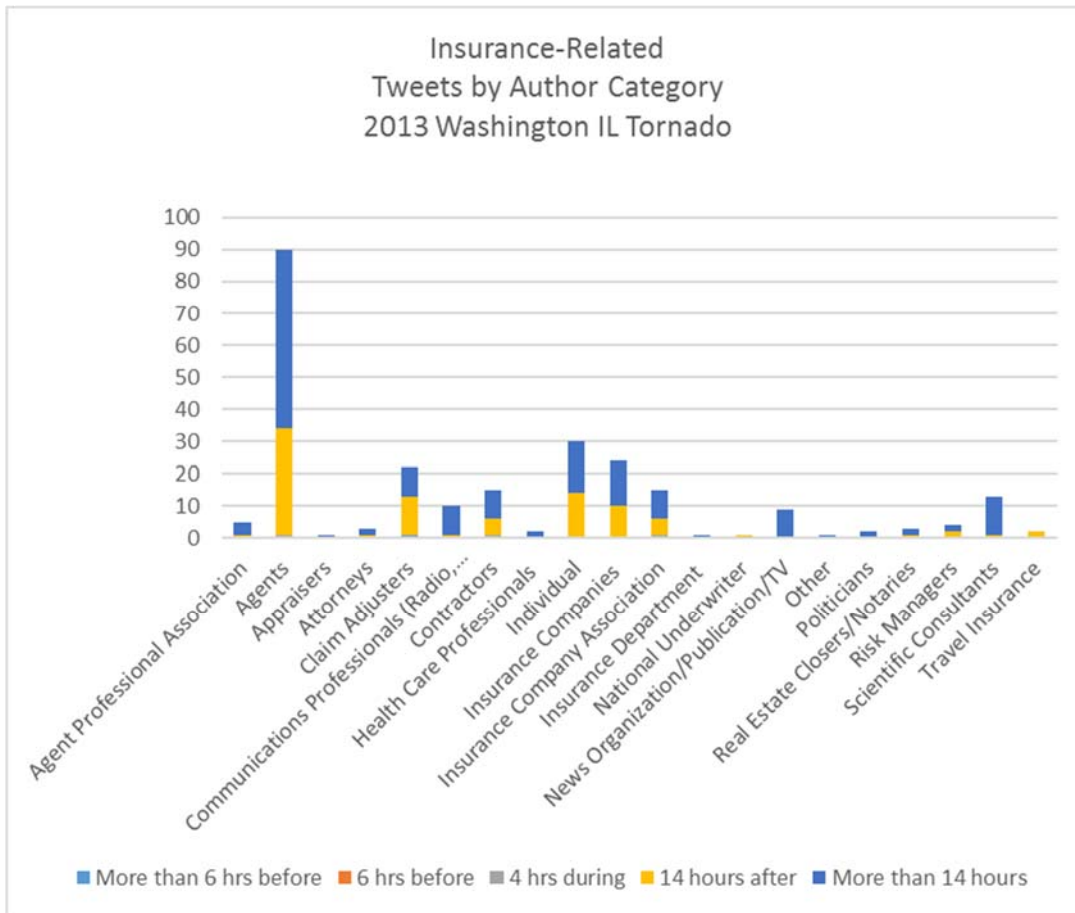
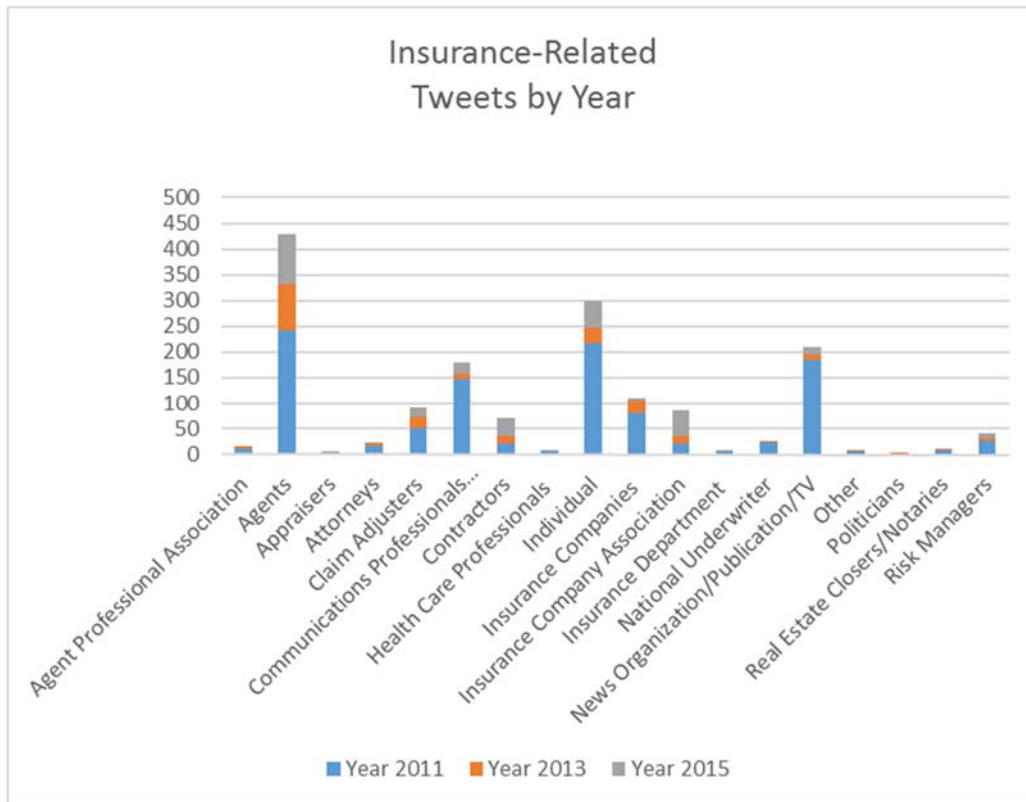


Figure 8
2013 Washington IL Tornado Insurance Tweet Count by Author Category



Agents are the largest group of insurance-related tweeters and most insurance tweets were posted more than 14 hours after the end of tornadoes. Figure 9 shows the insurance tweet count for the three tornadoes by author category. Again the 2011 tornado event has the largest amount of insurance tweets which could be impacted by the level of severity, casualties and damages.

Figure 9
Insurance Tweet Count by Author Category and Year



4.2 Key Messages

Given the small number of insurance related tweets, the use of Twitter among insurance companies and insurance professionals was limited in the past three tornado outbreak. Based on the 2013 tornado Twitter data, most of the tweets before and during a tornado outbreak are retweeting warnings of the tornado event. After event tweets have more diversified contents with the following categories:

1. Information about damages by tornadoes including deaths, destroyed houses, power outage, pictures, videos and so on.
2. Insurance counseling services.
3. Advices on insurance coverages for natural disasters such as flood insurance.
4. Insurance damage assessment.
5. Help and donation for tornado relief.

More efforts can be made by the insurance industry... to use social media for sharing information on loss prevention before and during the tornado and loss control and communication after the tornado.

4.3 Key Influencers

Social networks are used to share information quickly and widely. The Twitter network has only 5.2 degrees of separation, which means that on average everyone is only about five steps or less away from each other based on a 2010 study². The large number of users in tornado Twitter Data also indicates that tornado related information were

² TWITTER FRIENDSHIP DATA, SIX DEGREES OF SEPARATION, TWITTER STYLE, April 2010 (<https://sysomos.com/inside-twitter/twitter-friendship-data>).

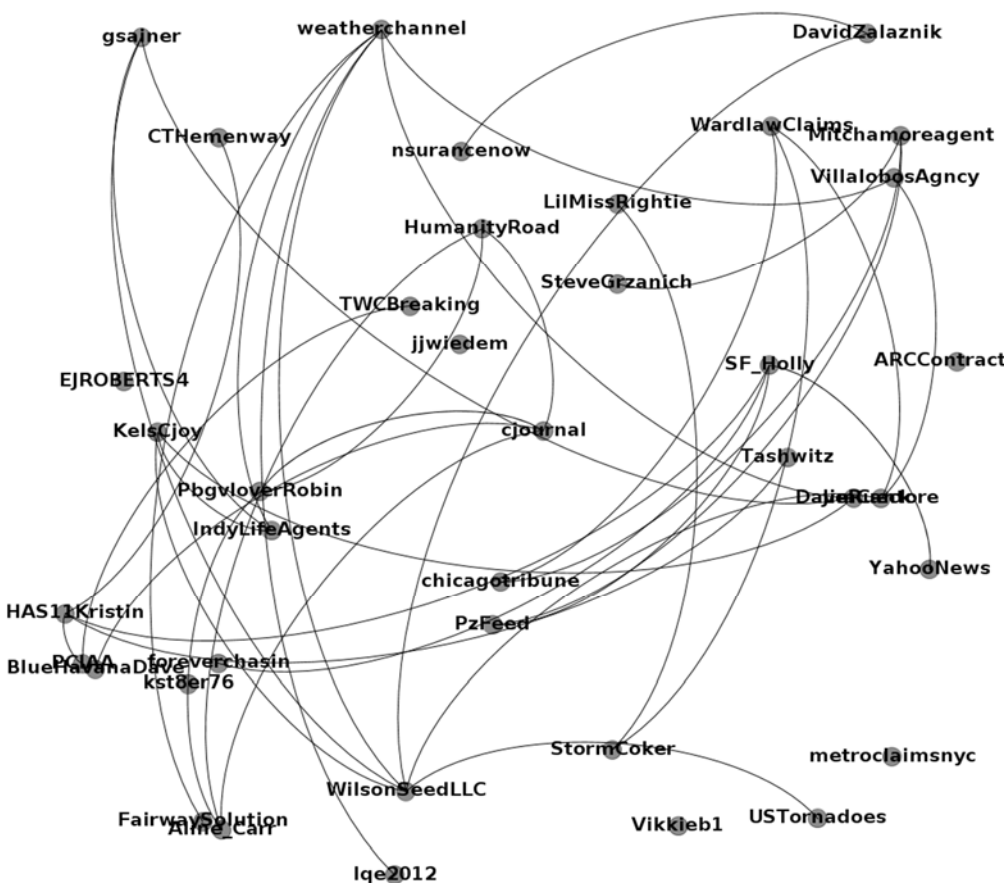
widely spread via social media. Table 7 lists the top 20 users based on the number of insurance tweets posted in the 2013 Washington IL tornado Twitter data. It also shows the number of followers and friends which indicates the wide reaching of the tweet.

Table 7
2013 Washington IL Tornado Top Insurance Tweeters

User Name	No. of Tweets	No. of Followers	No. of Friends
WardlawClaims	12	583	472
PCIAA	8	6,840	1,170
Ayodejiojetola	7	501	675
SF_Holly	6	248	556
IIAofIL	6	595	533
kst8er76	5	2,989	3,296
LilMissRightie	4	10,576	1,053
WilsonSeedLLC	4	200	185
BlueHavanaDave	4	183	87
Mitchamoreagent	3	358	487
metroclaimsny	3	133	231
lqe2012	3	1	18
ins4contractors	2	121	25
Ke11yRoss	2	225	815
EJROBERTS4	2	66	239
INSURANCE_NY	2	97	25
TravellnsDirect	2	1,665	1,674
JeffreyLBrewer	2	802	1,149
ARCCContracting	2	150	277
jjwiedem	2	58	141
Top 20 Total	81	26,391	13,108
Total	253	209,958	144,671

In addition, Twitter allows users to @ other users so that the tweets are showed as notifications and less likely to be neglected. The network of users and mentioned users in the insurance-related tornado Twitter data can be analyzed to identify key influencers. Figure 10 shows the key influencers of the 2013 Washington IL tornado insurance related Twitter network. Users with no less than 3 edges are included which means that the total times that user was @ed by other users or @ed other users in the insurance network is more than 3.

Figure 10
2013 Washington IL Tornado Insurance Network Key Influencers



Section 5: Sentiment Analysis

Sentiment analysis of social media data can help understand the change of aggregate sentiment of the users over time. For tornado Twitter data, it is expected to have a low sentiment on average because of the disastrous event. Stanford CoreNLP³ was used to conduct the sentiment analysis which evaluated the sentiment of each tweet based on the following scoring system:

- 0: Very Negative
- 1: Negative
- 2: Neural
- 3: Positive
- 4: Very Positive

³ Details about the natural language analysis tool can be found at <http://stanfordnlp.github.io/CoreNLP>. Given the unique features of tornado Twitter data, further improvement can be made by providing new training examples to adjust the tool for Twitter data sentiment analysis.

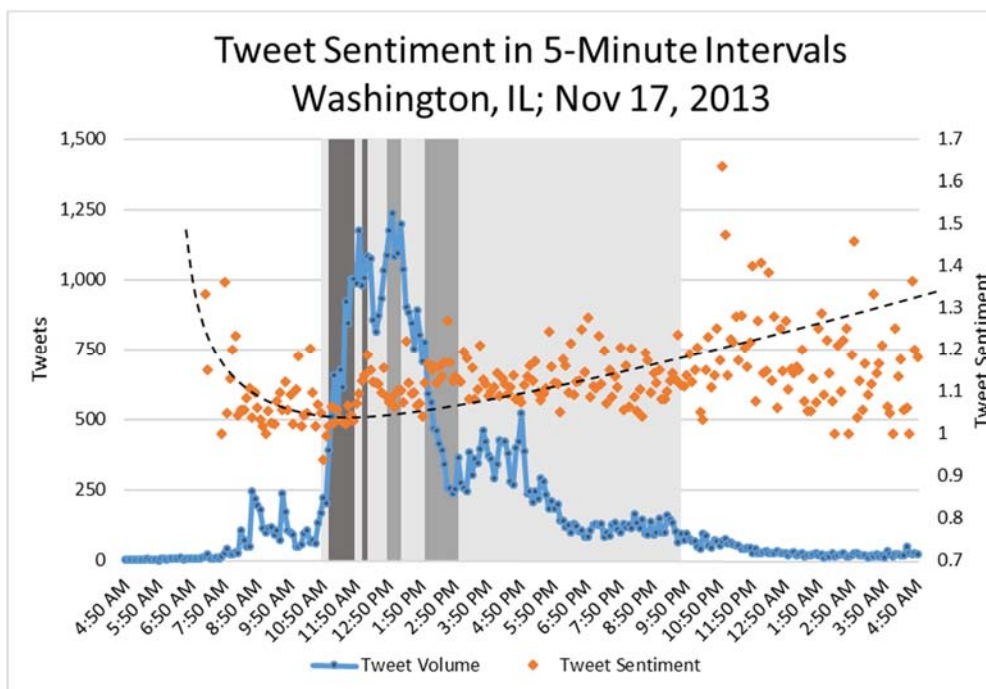
Table 8 shows the aggregate sentiment of Twitter data over the 24-hour period for each of the three tornado events. All events have a sentiment level of about 1.1 which means negative.

Table 8
Tornado Twitter Data Aggregate Sentiment

	2011 Joplin, MO Tornado	2013 Washington, IL Tornado	2015 Rochelle, IL Tornado
Aggregate Sentiment	1.133	1.132	1.097

Figure 11 shows the sentiment in 5-minute intervals of the 2013 Washington IL tornado Twitter data. The sentiment decreased when approaching the tornado and recovered slowly after the tornado.

Figure 11
2013 Washington IL Tornado Twitter Data Sentiment



This pattern is intuitively what would be expected during the progression of the tornado.

Section 6: Suggestions for Future Research

The following are suggestions for future research and enhanced analyses:

- Analysis of other social media data patterns associated with other catastrophes such as forest fires, snowstorms, and earthquakes.
- Development of a predictive model that uses the volume and type of tweets to predict the duration and severity of a catastrophe.

- Enhanced Twitter analysis of additional variables based on recently announced expansion of Twitter capabilities.
- Expanded sentiment analyses comparing different populations (intra-nationally and internationally) in the face of catastrophes.
- Development of social media strategies to enhance warning and loss control in catastrophic events.

Section 7: Conclusion

Social media have been used for warning, information sharing and communication in catastrophic events. Given its fast speed of spreading and far reaching, it could be helpful for enhancing warning systems, provide real-time risk monitoring support, communicate insurance knowledge and loss prevention measures with the public and customers, and facilitate claims. The participation of the insurance industry in social media is still limited and may be improved with a better understanding of social media data and analytics.

References

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

About The Society of Actuaries

The Society of Actuaries (SOA), formed in 1949, is one of the largest actuarial professional organizations in the world dedicated to serving 24,000 actuarial members and the public in the United States, Canada and worldwide. In line with the SOA Vision Statement, actuaries act as business leaders who develop and use mathematical models to measure and manage risk in support of financial security for individuals, organizations and the public.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement, and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and non-actuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org