The goal of this paper is to evaluate specific event-related Twitter posts, called tweets, for their usability as intelligence data. Using text mining software for Natural Language Processing to attempt to establish the best keyword search strategy for actionable information, 2000 tweets documenting 2 independent events were analyzed. Additionally, the Twitter users who provided the tweets were evaluated for source reliability to determine if their information was likely to be credible.

The results of this study suggest that the best information retrieval strategy for tweets featuring reliable and actionable information, is to use hashtagged topics and/or location details as search keywords. Not only is “location” frequently a facet of determining actionability, the co-location of event and user increases the likelihood of a high-reliability source.

Headings:

Microblogs

Relevance ranking (Information science)

Text mining (Information retrieval)
BETTER INTELLIGENCE THROUGH MICROBLOGGING: TEXT MINING TWEETS FOR ACTIONABLE INFORMATION

by
Meghan Croll

A Master's paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.

Chapel Hill, North Carolina
May, 2013

Approved by:

________________________
Stephanie W. Haas
# Table of Contents

- Table of Contents ................................................................. 1
- Introduction ................................................................. 2
- Literature Review ......................................................... 5
- Methods ................................................................. 16
- Results ................................................................. 28
- Analysis & Discussion ................................................. 34
- Conclusions ................................................................. 43
- References ................................................................. 45
- Appendix ................................................................. 48
INTRODUCTION

How much information can be communicated in 140 characters? The answer appears to be: enough to change the world. In just over seven years since its introduction, Twitter has gone from being one popular social media startup among many to being one of the most trafficked websites on the internet, a major source of breaking news and political expression, and a generally ubiquitous feature of life that produces a staggering volume of variable quality information. Twitter is an online service made for microblogging- a type of information sharing involving an individual or organization writing and releasing information to the public using a web service that formally or informally limits the volume of text one can post at a time. Microblogging users post messages that are very short; the text limit is 140 characters for the two most popular options, Twitter and the Chinese Weibo. In practice, tweets, the name for a Twitter post, average only 11 words per message, but with millions of users making numerous tweets each day the characters add up quickly (O’Connor et al., 2010). In spite of the simultaneous limitation and abundance of tweets or because of it, tweets have gained a reputation as timely sources on a variety of topics, if you can find the right ones.

Public health organizations and businesses use Twitter both to communicate with their audience and to understand it by analyzing tweets relevant to their interests and engaging with their “followers,” other Twitter users who subscribe to another’s Twitter feed. Twitter is credited with enabling the formation and organization of resistance and
protest movements in several Arab nations during the summer of 2009, although some might argue that the effectiveness of Twitter as a tool for political communication or social justice is overstated by popular media commentators and technology evangelists, particularly in Iran (Ulicny & Kokar, 2012). Part of the reason for this resistance to heralding Twitter as a tool of democracy is that democratic movements are not the only category of people who can use Twitter.

Case studies of Twitter use have found that the service can be and has been used to support terrorist and illegal activities; but, through analysis of tweets from both the criminals and civilians experiencing the results of criminal actions, Twitter can also deter or stop its less well-meaning users (Cheong & Lee, 2010). Relatively few tweets have restrictions on who can read them, many include topic-relevant classifications through the “hashtag” feature, the creation and publication of a tweet can be done quickly, and geographic data is often included. Twitter therefore is well suited to use in Open Source Intelligence (OSINT), which analyzes openly available information through a number of methods, including text mining, in order to inform intelligence and counter-intelligence activities. Intelligence entities worldwide monitor and analyze Twitter for insights into domestic and foreign activities. These entities are the presumed users whenever use is mentioned in this study.

In this paper, 2000 tweets selected to represent two independent events were processed using text mining to identify keywords or word categories that are most likely to produce information that will lead to intelligence. The users providing the actionable information were also evaluated for reliability, which informs the credibility of their tweets without requiring each tweet to be fact-checked. The hypothesis was that these
streamlined methods for finding actionable information and reliable sources would reduce the time and effort spent on Twitter-based OSINT projects without sacrificing the quality of data.
LITERATURE REVIEW

Intelligence, for the purposes of this paper, is defined as the product of information gathered from analysis of raw data. In this case raw data is in the form of tweets. Put another way, intelligence is information, but information is not intelligence until it has been analyzed and can be used to inform the actions of the intelligence user, most often a member of law enforcement, a military, or the intelligence community (IC) (Lowenthal, 2009). Getting from information to actionable information (information that is usable for intelligence purposes) to intelligence is increasingly challenging in the age of exponential information growth. Zanasi (2009) wrote that the largest obstacle in the intelligence gathering process is the amount of information constantly being produced; because regardless of the volume, analysts must regularly “retrieve, read, filter and summarize” (p. 59) the information relevant to their work in order for their organization to contain threats and locate opportunities.

Using an example of the interactions between terrorists and the IC, in what has been termed the “Netwar” (p. 54); Zanasi (2009) studied one of the earliest occurrences of widespread use of Twitter to collect and broadcast strategic information during a catastrophic event, a major terrorist attack in Mumbai in 2008. He remarks that a number of studies undertaken since the early 2000s have researched data and text mining in the context of terrorism informatics and they had universally concluded that informational superiority is the best weapon against the asymmetric threat that characterizes terrorism.
(Zanasi, 2009, p. 56). They also agreed, and Zanasi concurred, that the best method of approach in the Netwar is by automating as much of the process as possible, including the use of text mining, due to the proliferation of information about events. For an example of the amount of information analysts have to work with, Twitter users began tweeting information almost immediately after the initial attack on Mumbai at a rate of about 70 tweets every 5 seconds, diminished only slightly by later findings that many of the tweets regarding the attacks contained rumors and falsified information (International Risk Governance Council [IRGC], 2012, p. 7).

In the same terrorist attack on Mumbai, satellite phone conversations between terrorists and their handlers were recorded, giving unprecedented access to the operational communications engaged in by terrorist actors. Terrorist have been known to adapt to emerging technologies, using social media and online forums to recruit members, solicit donors, broadcast propaganda and instructions, combine and compare intelligence, evade scrutiny, and organize attacks. The Mumbai terrorist attack is most well-known for the last and for being one of the earliest events to demonstrate the possible consequences of unchecked citizen reporters.

An analysis of the conversations recorded in the Mumbai attack discovered that the terrorists monitored live new updates and tweets about the civilian response to enhance their level of situational awareness (SA) and inform further actions (Oh, Agrawal, & Rao, 2011). Several major news outlets used tweets from users who had seen and reported the attacks and the aftermath as a significant source of information. Since it has been established that the Mumbai terrorists were following news reports when deciding where and when to strike next, one can conclude that Twitter users had
accidently exposed what Oh et al. (2011) refer to as “operationally sensitive information without any information control” (p. 41) which led to further terrorist incidents.

The so-called “Green Revolution” in Iran following the 2009 election results used social media to communicate and organize, like many other popular movements in the region at that time. Twitter and Facebook were used by protesters to spread information about activities and news updates, but fell victim to the same problem that troubled Mumbai- a lack of information control. Khonsari, Nayeri, Fathalian, A., & Fathalian, (2010) found through an analysis of the social network of the Green Movement that approximately 24% of network members were supporting or affiliated with the Iranian government. This might not have been a problem, except that the social dynamics of the network left them vulnerable. Although there were many links between users, the network was largely decentralized with very few users providing the most information; the loss of most of the users wouldn’t disrupt the information exchange but the small central core that formed the foundation of the network was susceptible to manipulation which would in turn control the informational capacity of the entire network (Khonsari et al., 2010). Oh et al. (2011) recommend that governments adopt a strategy to control digital information during terrorist attacks based on the goals and political ideologies of the terrorist groups. Their research has found that the ideologies inform the choice of targets and the approach to destruction or disturbance, which the authorities can then use to decide which tweets need to be censored.

In the United States, Twitter use in cataclysmic situations has a more positive connotation among the social media-literate population, as it has notably been used for natural disaster information and relief. FEMA manages several Twitter accounts for
regions, specific events or programs, a national account, and one for their administrator. The purpose of these accounts can range from disaster preparedness tips to disaster updates to responding to comments or the use of #smem (social media in emergency management hashtag) in a tweet. The current FEMA Administrator, Craig Fugate has praised Twitter as providing a level of SA that regularly exceeds official information sources, something that even a few years prior wouldn’t have been possible (IRGC, 2012).

Text mining is “the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources” (Hearst, 2003). As mentioned previously, it is an increasingly popular method of intelligence gathering, classification, and analysis. However, text mining alone is often insufficient to deal with the size and complexity of the available data and many advancements in software and programming have been made by government and private entities.

One of the latest and most promising applications designed for text mining Twitter is called, appropriately, Twitcident. Abel, Hauff, Houben, Stronkman, & Tao (2012) were already satisfied with studies confirming Twitter’s use as a news source and a popular target for information science research, but noticed a lack of research in Social Web search strategies and information filtering. Researched, created, and operated by Abel et al. (2012), the Twitcident system is connected to the emergency broadcasting services in the Netherlands and Twitter; it is triggered by the former and then draws information from the latter. Twitcident is automatically triggered by a P2000 message, which is broadcast in the Netherlands in response to events which are a threat to public
safety; the application then begins automated collecting and filtering of tweets the system thinks are relevant to the incident. The program uses weighted facet-value pairs, named entity detection programs to enhance the contextual data, additional data from the RSS feeds of established news sources, and the tweets’ metadata to determine the relevancy of the tweets to the event in question. Twitcident then filters out most foreign languages and “chatter” which includes retweets (a tweet by another user, forwarded to you by someone you follow, often used to spread news or share valuable findings on Twitter) and tweets of less than 100 characters in order to use their analytical processing power on the tweets they estimate to have the most potential for actionable information. Finally, the remaining tweet data is put through analytical programs and the finished product is a report summarizing the current threats which is then made available to Twitcident users. The Twitcident program and others like it focus on using publicly available, frequently updated microblogs (tweets) in order to aid law enforcement and civil authorities in responding to extant and potential threats in real-time.

Figure 1: Twitcident information architecture (Abel et al., 2012, p. 2)
Experts in the field of social media research are divided concerning the value of retweets when evaluating actionable information. The Twitcident team apparently see retweets as functionally irrelevant. Thomson et al. (2012) have the opposite opinion of retweets, as well as links and other information from tweets which originates from a third party, describing them as “socially filtered information [which] constitutes more value to the information ecosystem as a whole when compared with original tweets which were not passed on” (p. 9). Regardless of the validity of their stance on retweets, Twitcident offers two lessons learned: the first is that for natural disasters and finite occurrences (e.g. a fire or royal ceremony), Twitter activity peaks at four hours after the event. The application has not been used for events with an indeterminate lifespan, such as riots and protests which can last as long as participants want and authorities allow. The second concerns the value tradeoff with links. Given the high volume of tweets during an event, the time it takes even an automated program to open and evaluate the contents of links may not be the best use of processing power, particularly if the tweet the link was in already had usable data.

Human beings, like most animals, have evolved senses to warn them of possible dangers, including some ability to judge truthfulness. With Twitter, this ability is hampered by the digital medium that separates the producers and consumers of information; additionally some people may contribute inaccurate information without meaning to deceive or being aware of its inaccuracy. However, human perceptions of accuracy or credibility can still inform the usability of data and certainly affects the dissemination of information.
A study of perceived source reliability following the Fukushima nuclear disaster in 2012 looked at the judgments of several annotators evaluating tweets on the event without referring to formal reliability standards or in depth investigations of the users. Previous studies had shown that public confidence in the reliability of Twitter users and the credibility of information originating from or filtered through Twitter as an accurate news source is lower than that for mainstream media, online news organizations, and even anonymous blogs. Thomson et al. (2012) determined that the nearness of a Twitter user to an event improved the ratio of references to high-credibility sources “suggesting proximity to the disaster mediating the degree of credibility of shared content” (p. 1). This study also found that the closer a user was to an event, the more they tweeted about it, which also improves the sources likelihood of being judged reliable by the more formal STANAG 2022 standards (discussed below). The tweets subject to analysis were compiled in part by searching for the hashtag #fukushima; in almost 70% of retweets, tweets containing links, and other varieties of tweets containing some form of third-party information resulting from this hashtag search, the third party information was analyzed as highly credible (Thomson et al., 2012). The Fukushima study is highly unusual in that it found no intentionally false information or rumors in the tweets studied, which had been an issue after earlier earthquakes in Japan and Chile.

Finally, several computer programs have been made specifically to judge the reliability of sources and data on the internet and many others have incorporated some reliability judgment into larger programs. Google’s PageRank algorithm, for example, evaluates the reliability of the websites brought up by search results without having to independently vet each one (Ulicny & Kokar, 2012). Unfortunately, most companies and
organizations making programs for this purpose zealously guard their trade secrets, to
stay competitive in their field or due to security concerns, making some of the most
successful software and techniques unavailable to researchers.

The standard guide for judging source reliability is STANAG (standardization
agreement) 2022, developed and adopted by NATO. The judgment is graded on an
alphanumeric scale which designates the reliability of the source with a letter and the
credibility of the given information a number, where reliability is defined as the
likelihood that any information from the source is accurate based on past experiences
with the source and credibility as the likelihood that a piece of information is accurate
based on the source and corroborating evidence. The reliability and credibility are
closely linked, as are past instances of providing information. Cholvy and Nimier (2004)
summarize the STANAG 2022 designations in the following graded scale:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Meaning</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>completely reliable</td>
<td>tried and trusted source which can be depended upon with confidence</td>
</tr>
<tr>
<td>B</td>
<td>usually reliable</td>
<td>been successfully used in the past but there is still some element of doubt in particular cases</td>
</tr>
<tr>
<td>C</td>
<td>fairly reliable</td>
<td>occasionally been used in the past and upon which some degree of confidence can be based</td>
</tr>
<tr>
<td>D</td>
<td>not usually reliable</td>
<td>been used in the past but has proved more often than not unreliable</td>
</tr>
<tr>
<td>E</td>
<td>unreliable</td>
<td>been used in the past and has proved unworthy of any confidence</td>
</tr>
<tr>
<td>F</td>
<td>cannot be judged</td>
<td>has not been used in the past</td>
</tr>
</tbody>
</table>

*Table 1: STANAG 2022 source reliability classifications*

This scale relies heavily on past experiences with the source, hindering the
prospect of evaluating new sources as reliable or not and risking undervaluing sources
and their information. The confidence rating for information is dependent on both
previous information and the source or sources that are reporting it, with the highest
rating reserved for that which has several other sources reporting the same or similar information. Unfortunately, in unforeseen events, many of the eyewitnesses tweeting about developments are becoming intelligence sources for the first time and so are automatically rated “F” as the STANAG 2022 scale denies the possibility that the reliability of nascent sources can be evaluated.

\[ \text{Influence}(X) = \sum_{Y \in \text{Followers}(X)} \left( 1 + p \cdot \text{Influence}(Y) \right) / \| \text{Following}(Y) \| \]

Figure 2: TunkRank equation (Ulicny & Kokar, 2012, p. 5)

Ulicny and Kokar (2012) propose determining the reliability of Twitter users by the value of their followers, determined by the “TunkRank” algorithm which combines the number of followers who will see a tweet by the source/user and the probability that the source’s tweets will be retweeted by their followers, eliminating the likelihood that a user’s reliability is wrongly inflated by fake Twitter accounts which often follow but don’t engage in activity. The percentile ranking from the algorithm is then linked to the six STANAG 2022 reliability grades. Letting the popularity of users determine their reliability is an imperfect plan as not all tweets are intended as accurate information. The current third TunkRanked Twitter user is satirical news organization The Onion, whose work has been mistaken for credible information before by actual news providers (Mirkinson, 2011). To attempt to teach a text mining program that information from a source that is ranked as highly reliable should be ignored because it is satirical would be an exercise in futility. With regards to the credibility of a source’s information, the accuracy of what a source reports via a tweet cannot generally be determined by formal reasoning (Ulicny & Kokar, 2012).
For all that there are many standards for reliability, in the IC and beyond, it must be stated that there is no universal judgment on what how to define reliability in any field. (Cholvy & Nimier, 2004). Furthermore, although it is not clear how source reliability is tracked and monitored by human analysts in practice today, it is clear that with the multitude of Twitter users posting messages, it is impossible to individually vet each one. Therefore attempts to establish reliability through automated applications must be made (Ulicny & Kokar, 2012). Cholvy and Nimier (2004) are quick to point out that the STANAG 2022 vocabulary can be broadly interpreted and it neglects to give further recommendations for situations such as conflicting or intentionally false information. Conflict may be intentionally created or it might just be a matter of differing source perspectives. Taking an example from tweets surrounding the London riot (for background and a timeline of the events see: http://www.bbc.co.uk/news/uk-14436499), an ongoing interaction between police and rioters may be reported as “the police have the rioters controlled at Green Square” or “rioters are assaulting police at Green Square.” The descriptions both qualify as actionable information as described in the following paragraph, but the content of the messages are diametrically opposed. Analysts might be able to judge which one is more credible based on the reliability of the original sources; Cholvy and Nimier (2004) propose a weighted majority equation for dealing with multiple sources of various reliability grades providing conflicting information.

For this study, the information used for intelligence needs to be not only relevant and credible, but actionable. The concept of actionable information is central to this research and is defined as follows. Actionable information gives specific, objectively observed information that will allow for usage by the intended audience. For example,
consider the tweet “Sepulveda Boulevard will have a roadblock from 2-4 pm.” This is specific (providing the location, time, and event) and objective (it is not given as an opinion, but a fact, albeit alleged until proven) and can be used to inform the actions of people in that area at that time (drivers may subsequently avoid that road at that time). Actionability is a key feature in separating information from intelligence (Lowenthal, 2009). Considering all of the approaches and drawbacks to finding information on Twitter, there is a need for simple, dependable, and inexpensive ways for OSINT analysts to find actionable, relevant, and credible information. Such methods will reduce the time and effort it takes to get results from this part of the intelligence process.
METHODS

The experiment consisted of four distinct parts: information retrieval, coding tweets, text mining, and source reliability. Information retrieval (IR) involved identifying two events with sufficient volume of tweets for text mining and finding tweets related to the events being studied. The next step was compiling the tweets into corpora and coding the tweets as containing actionable information or not. After that, the tweets were run through a text mining program to gather statistics on the individual features of the tweets and create language models for testing. The final part of the process was investigating the sources (i.e. Twitter users) who wrote the tweets used in the corpora for reliability.

Information Retrieval

The first task of this study was IR. The Occupy Wall Street protest and the London riots were chosen as the events to analyze based on the duration of activities, the well-known use of social media during the events, and the majority of tweets being written in English, the only language the author of this study is fluent in. The riots lasted for a week in various incarnations and the Occupy encampment persisted for 28 days; as
a result, they were both estimated to have enough associated tweets to undertake analysis. The London riots were an outbreak of extremely destructive civil disorder begun after a police shooting of a man in Tottenham, a district in Northern London. The riots started in that area, but other areas of London soon fell to opportunistic looters and violence; there was no perceived leadership or mission. Occupy Wall Street, in contrast, was a planned peaceful protest encampment in Zucotti Park, New York City, against income inequality, instigated in part by the Canadian anti-consumerist magazine *Adbusters*, and enacted as a grassroots movement that expanded into local movements worldwide. The events were similar in demographics, use of social media, and police presence, but differed in their goals, methods, and public support. The differences will hopefully reduce the possibility of drawing conclusions based on only a single event, or events of a single type.

Twitter allows access to only a couple of months’ worth of tweets and both of the selected events occurred almost two years before this research was undertaken. It was therefore not possible to retrieve the tweets from the website where they were originally posted. Older tweets are archived, but accessible through third-party websites like Topsy. Topsy was used as a database of tweets for this study because it allows for searching with keywords, hashtags, and usernames and can be refined by the date and media (i.e., text tweet or picture tweet). Searching was completed for each event by looking for text tweets from the beginning of each event until a week past, which included the entirety of the London riots (August 6-13, 2011) and the detail-rich formative period of the Occupy encampment (September 17-24, 2011). The search was further narrowed to include only tweets including one or more of three elements: event-
relevant hashtags, keywords, and/or official users, the full list of which is shown in Table 2.

<table>
<thead>
<tr>
<th>Event</th>
<th>Hashtagged terms</th>
<th>Non-hashtagged terms</th>
<th>User names</th>
</tr>
</thead>
<tbody>
<tr>
<td>London riot</td>
<td>#londonriot, #londonriots, #tottenham, #tottenhamriot, #riots, #ukriots, #cameron</td>
<td>Duggan, riots, rioters, Tottenham, London, North London, Manchester, Enfield,</td>
<td>@guardian, @bbcnews</td>
</tr>
<tr>
<td>Occupy Wall Street protest</td>
<td>#occupy, #occupywallst, #occupywallstnyc, #sep17, #ows</td>
<td>occupy</td>
<td>@OccupyWallSt, @OccupyWallStNYC, @Adbusters</td>
</tr>
</tbody>
</table>

*Table 2: Search terms used for retrieving event-related tweets*

The search terms for the London riots corpus include descriptions of the event (terms including the word “riot” or “riots”), locations that background reading on the event had indicated were areas of activity, the last names of the British Prime Minister and the victim of the police shooting that ostensibly began the riots, and the official Twitter accounts of two major news outlets covering the event. The search terms for the Occupy Wall Street corpus include various iterations of the movement’s name proceeded by hashtags, the official Twitter user names for the local and national movement, and the user name for the organization that organized the protest. Usage of hashtagged or non-hashtagged search terms depended on how many words were being searched for and the likelihood that a non-hashtagged term might return irrelevant results. Hashtagged terms are frequently more than one word typed without spaces and non-hashtagged terms tend to indicate a singular occurrence as opposed to the trend implied by a hashtag. The Occupy movement actively encouraged the use of certain hashtagged terms by supporters in order to make it easier for other supporters to find information about the movement.
and the protest. In searching for both events, results retrieved by the obvious terms (i.e. #londonriots and #occupywallst) frequently included less obvious terms (i.e. #cameron and #sep17) which were then adopted as search terms themselves.

The faceted searching designed and implemented by Twitcident is unusable in this study due to the time that has passed since the tweets were posted and Twitter’s policy on the availability of archived tweets. Twitcident is intended to be used in real time directly with Twitter, their faceted search approach was not designed or intended to be used with third-party tweet archives. Twitcident’s semantic filtering shows promising results for future applications of text mining Twitter for events in real-time or with historical tweets, but requires greater technical resources than are available for this study. According to Amati et al. (2011) the most important features to search for when retrieving data for text mining tweets are the desired topic (e.g. earthquake, fire, royal visit) and the time the tweet was published, with the most recent tweets having the highest priority for analysis, as information may change as quickly as it is tweeted. Since the analysis in this study is not taking place in real time, however, how recent the information is to the current time is irrelevant as long as the tweets are contemporaneous to the events. Cheong and Lee (2010) gathered tweets for their study by searching for tweets from a given location (Mumbai) and trending topics associated with terrorist activity which are otherwise rarely used (e.g. #explosion, #attack), which is similar to the approach used in this study. However, this study did not use location data when selecting tweets for inclusion.

The retrieved event-related tweets were copied into files for text mining. In general, the more data available for text mining, the more accurate the results will be; due
to being more likely to include the variety of characteristics seen in the overall population. The number of tweets for each event was set arbitrarily at 1000 as being a balance of enough data to draw reasonably reliable conclusions from and a reasonable amount to process in the available time. The number of tweets retrieved using the dates and search terms described above amounted to more than 1000 tweets per event. Tweets were selected for inclusion in the research corpus on the basis of uniqueness; retweets were allowed if the original tweets had not already been included, as were tweets with identical or similar messages sent for separate occasions (e.g. daily announcements of Occupy protest meetings), but multiple tweets using exactly the same wording for a single occurrence were not included as they added no value (i.e., unique features) to the documentation of that occurrence.

**Coding Tweets**

Once 1000 tweets had been collected for each event, 2000 tweets in total, the tweets needed to be annotated by a human to determine which ones contained actionable information before they could be analyzed by text mining for specific features indicating actionability. The purpose of coding is twofold: the text mining program needed both positive and negative instances for training models and analyzing features for actionability, and human judgment functions as a superior standard for assessing the data because a text mining program cannot intrinsically know what constitutes actionable information or not. The program follows the patterns learned from analyzing the human-coded tweets when attempting to categorize features or other tweets as actionable.
information. Sample coding was completed prior to the author’s coding of the entire corpus to establish inter-operator agreement, meaning that the concept of actionability could be agreed upon.

Due to time and financial constraints limiting the resources available for coding tweets, a sample of 50 randomly selected tweets (25 from each event) was coded as actionable or not actionable. This task was completed by two volunteers, experienced but non-professional social media users, whose judgment of each tweet was then compared to each other and that of the author. A set of guidelines was created to assist the coders in making educated judgments. The volunteer coders followed the guidelines to code the tweets; afterwards a discussion of the guidelines between the author and the volunteer coders further refined the guidelines and the tweets were coded again with the revised guidelines. The revised guidelines were then used to direct the coding of the entire corpus by the author alone. The guidelines provided a summary of the events, and definitions and examples of the specificity, objectivity, and utility that determines actionability.

Initial agreement was 90%, the two coders and the author agreed on 45 tweets and the coders disagreed on the remaining five, the author agreed with the judgments of one coder in all of those instances. The instances which were the subject of disagreement were debated. It was determined that the disagreement stemmed from different understandings of information utility and the mindset of the intelligence professional that was the presumed user of the data. The discussion resulted in the editing of the guidelines to specify the independent consideration of each tweet (“Each tweet must be considered separately, without presumed knowledge gained from any other tweet.”) and
to include a specific example for the Occupy corpus about the pizza delivered to the protestors, a frequent topic of Occupy tweets. The subsequent re-coding of tweets raised agreement to 96%; the two tweets that the coders still could not agree on were both from the London riot corpus and involved police activities. The author and one coder believed that reports of police activities (e.g. “riot police in ‘stand off’ with crowd”) had no utility for law enforcement while the other coder disagreed. The 96% achieved with the second coding was high enough agreement to assume that the corpus could be coded accurately by one person. The author then used the revised guidelines (found in the Appendix) to code the rest of the collected tweets which were contained in comma separated value files for text mining.

**Text Mining**

The coded tweets were put into seven .csv files as shown in Table 3. The entire corpus of 2000 tweets was used for initial candidate feature identification.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training file- # of tweets (% of subcorpus)</td>
<td>800 (80%)</td>
<td>800 (80%)</td>
<td>1600 (80%)</td>
<td>N/A</td>
</tr>
<tr>
<td>Testing file- # of tweets (% of subcorpus)</td>
<td>200 (20%)</td>
<td>200 (20%)</td>
<td>400 (20%)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Table 3: Subcorpora of tweets used for experimenting.*
For each subcorpus (London, OWS, Combined), 80% of the tweets were used for training and 20% were used for testing. The short length of tweets present both a challenge and simplification for text mining, there is a limit to how much a small amount of information can be manipulated (Amati et al., 2011). The feature extraction file containing all of the tweets was mined for features revealing the keywords or categories of keywords most likely to be found in tweets that are considered actionable. The three training corpora had features extracted and the tables of extracted features were then used to make models to predict the actionability of the corresponding test corpora (London training was used to create a model to test on London test, etc.). The events were evaluated separately and together because Occupy Wall Street and the London riots involved significantly different actors, actions, and motivations—Occupy participants were largely politically motivated and well-educated while the London riots began as political but were quickly taken over by self-serving individuals characterized as having less privileged backgrounds—and the results of the text mining might be more successful on one than the other. The difference or similarity of results from testing the subcorpora would indicate if the extracted features were universal across event types or if one type of event required more research and refinement of text mining methods than the other.

Text mining was done using the open source program, LightSIDE (http://lightsidelabs.com/). Experimenting was done on training files and the feature extraction file with feature extraction configurations such as n-grams, removal of stop words, binary features, POS bigrams, and stemming alone and in numerous combinations to view the features and their predictive statistics, including the performance of each feature measured by precision, recall, and f-measure.
N-grams define the number of features in a sequence, generally limited to unigrams, bigrams, and trigrams. Using unigrams as a feature configuration creates what is called a “bag of words” approach; each feature is valued individually without context from the features surrounding it. Bigrams judge two consecutive features and trigrams three. Removing stop words eliminates common words with little to no value (e.g. “the” or “and”) from the table of features. When binary features are not used, the program only counts the instances of the word without including any other statistics. POS (an acronym for “part of speech”) bigrams judge two consecutive features classified by their syntactic function. POS tagging classifies words as parts of speech ranging from the simple (NN indicates a noun, FW means foreign word) to the complex (VBZ is a third person singular, present tense verb). Stemming reduces features to their linguistic base.

Precision, recall, and f-measure are all standard measures used in text mining and IR to evaluate different aspects of results; the benefits of each are described in more detail below.

<table>
<thead>
<tr>
<th>Feature configuration</th>
<th>Best recall</th>
<th>Best precision</th>
<th>Best f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram, binary features (LightSIDE default)</td>
<td>0.356</td>
<td>1</td>
<td>0.503</td>
</tr>
<tr>
<td>Bigram, binary features</td>
<td>0.123</td>
<td>1</td>
<td>0.184</td>
</tr>
<tr>
<td>Trigram, binary features</td>
<td>0.123</td>
<td>1</td>
<td>0.184</td>
</tr>
<tr>
<td>Unigram, binary features, remove stop words</td>
<td>0.331</td>
<td>1</td>
<td>0.474</td>
</tr>
<tr>
<td>Unigram, binary features, punctuation</td>
<td>0.9</td>
<td>1</td>
<td>0.578</td>
</tr>
<tr>
<td>Unigram, binary features, stemming</td>
<td>0.356</td>
<td>1</td>
<td>0.503</td>
</tr>
<tr>
<td>Unigram, binary features, POS bigrams</td>
<td>0.582</td>
<td>1</td>
<td>0.676</td>
</tr>
</tbody>
</table>

*Table 4: Various feature configurations and their results at threshold of five*

The feature configurations with the best f-measure and recall were selected to create feature tables for the training models and feature extraction file. The best
precision was one in all configurations, so that wasn’t a deciding factor. These preliminary experiments demonstrated that the combinations “unigram/binary features/punctuation” and “unigram/binary features/POS bigrams” achieved the best results; these were the feature configurations used in identifying the key features of actionability and feature extraction for the training models.

Aside from the uniformity of the best precision, recall was valued higher than precision because it emphasizes sensitivity. Searching for features with a high recall for actionable information may result in the retrieval of data that is not actionable information, but it is less likely to miss actionable information. In the context of searching tweets for terms that indicate data on riots, terrorism, or similar events it is better to include some tweets that aren’t actionable than to risk missing something important. On the other side of the precision/recall trade-off, the features with the highest precision may not always produce all of the tweets with actionable information, but it is likely that the tweets in which these features appear do have actionable information.

The f-measure (also called the F₁ measure) is a compromise to valuing only recall or only precision. The f-measure is the harmonic mean between precision and recall, calculated as:

\[
\frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}
\]

*Figure 4: F-measure calculation (Witten et al, 2011, p. 175)*

The f-measure can also be altered to value either recall or precision more while still taking the other into account. Since recall is more valuable in intelligence projects for the previously mentioned reasons, the formula can be altered to reflect that. The result is called the F₂ measure.
Reliability Assessment

Reliability of the sources was appraised separately, based on the techniques of Ulicny and Kokar (2012) and Thomson et al. (2012). Source reliability for this paper was informed by these theories of reliability discussed earlier in this paper, but evaluated using the judgments of the author. Factors that indicate a higher source reliability are the source’s popularity, in terms of number of followers who also have a large number of followers (evaluated using the program Twiangulate; see: http://www.twiangulate.com), and geographic proximity to the event they are reporting as determined by an analysis of the content of the tweets. It should be noted that content analysis is not the most accurate or efficient method of determining the location of a Twitter user. Twitter is practically ubiquitous on mobile devices.

While geographic location is a feature of Twitter user profiles and more easily determined than inferences drawn from the text of a tweet, it is not required information, nor is it required to be honest and accurate information when given. Additionally, the current location of a user may be different from when the user was tweeting about the event. For these reasons, the content of a tweet was the deciding factor in determining a Twitter user’s proximity. Content that indicates proximity includes tweets such as “Despite everything that has happened, I have never felt more connected to my community in UK. #UKRiots” and “Building in north Tottenham ablaze. Young men in
masks won't let me get closer. #riot” [emphasis author’s]. There were many tweets whose language did not indicate the user’s proximity or distance. Without being able to determine the user’s location, the proximity of these users was categorized as unknown and their reliability was downgraded.

The results of the reliability evaluations were couched in the terms of the STANAG 2022 reliability rankings. Sources with over 100,000 followers (all follower counts taken at time of writing) are automatically classified as A rated sources. Of the hundreds of users who contributed the remaining tweets in the corpus, the ones whose language suggests that they are in the area and personally witnessing the developments that they are tweeting could be judged as B (>10,000 followers) or C (<10,000 followers) rated sources whose information requires additional independent verification in order to be considered credible. The number 10,000 followers was chosen arbitrarily; between users who reciprocally follow followers, individuals with multiple user accounts, and user accounts created by computer programs, getting to 1000 followers or more without contributing to the Twitter environment is fairly easy. The remainder of sources who are clearly outside of the area of the event or do not indicate their location and do not have the large follower base that implies pre-judged reliability by the general public display no criteria that would suggest they are reliable sources for these events, but neither have they been found unreliable. This last group of tweeters has been rated F.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Followers</th>
<th>Proximity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>&gt;100,000</td>
<td>Irrelevant</td>
</tr>
<tr>
<td>B</td>
<td>&gt;10,000</td>
<td>Indicated</td>
</tr>
<tr>
<td>C</td>
<td>&lt;10,000</td>
<td>Indicated</td>
</tr>
<tr>
<td>F</td>
<td>&lt;100,000</td>
<td>Distant or unknown</td>
</tr>
</tbody>
</table>

*Table 5: Summary of reliability grading criteria*
RESULTS

Actionability

Using the feature configurations unigrams, binary features, and punctuation on the coded 2000 tweet corpus with the threshold (minimum feature occurrences) set at 5 resulted in 975 features. The following were found to be the features indicating actionable information with the highest recall:

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$ measure</th>
<th>$F_2$ measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER SIGN</td>
<td>0.222</td>
<td>0.9</td>
<td>0.356</td>
<td>0.446</td>
</tr>
<tr>
<td>PERIOD</td>
<td>0.196</td>
<td>0.769</td>
<td>0.313</td>
<td>0.389</td>
</tr>
<tr>
<td>COMMA</td>
<td>0.214</td>
<td>0.342</td>
<td>0.263</td>
<td>0.285</td>
</tr>
<tr>
<td>in</td>
<td>0.266</td>
<td>0.216</td>
<td>0.270</td>
<td>0.230</td>
</tr>
<tr>
<td>occupywallst</td>
<td>0.239</td>
<td>0.216</td>
<td>0.277</td>
<td>0.223</td>
</tr>
</tbody>
</table>

*Table 6: 5 features with the highest recall value indicating actionability for the unigram/binary features/punctuation configuration*

Three of the top five keywords being punctuation marks was unhelpful for understanding why certain features might be used to find actionable information except for the number sign. The number sign, also known as a hashtag, is a linguistic convention commonly used on Twitter to indicate the key concept or topic of a tweet. It’s also used to tag content so that other users can find tweets; if enough tweets contain a single hashtag, the hashtag topic is “trending” or a popular issue on Twitter. In other words, tweeters hashtag tweets to spread or draw attention to information.
Using the feature configurations unigrams, binary features, and POS bigrams on the coded 2000 tweet corpus with the threshold set at 5 resulted in 1602 features.

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Precision</th>
<th>Recall</th>
<th>F$_1$ measure</th>
<th>F$_2$ measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;NUMBERSIGN&gt; NNP</td>
<td>0.245</td>
<td>0.414</td>
<td>0.308</td>
<td>0.337</td>
</tr>
<tr>
<td>&lt;PERIOD&gt; &lt;NUMBERSIGN&gt;</td>
<td>0.254</td>
<td>0.355</td>
<td>0.296</td>
<td>0.313</td>
</tr>
<tr>
<td>NN &lt;COLON&gt;</td>
<td>0.196</td>
<td>0.342</td>
<td>0.249</td>
<td>0.274</td>
</tr>
<tr>
<td>&lt;NUMBERSIGN&gt; NN</td>
<td>0.250</td>
<td>0.342</td>
<td>0.289</td>
<td>0.305</td>
</tr>
<tr>
<td>in</td>
<td>0.226</td>
<td>0.337</td>
<td>0.270</td>
<td>0.290</td>
</tr>
</tbody>
</table>

*Table 7: 5 features with the highest recall value indicating actionability for the unigram/binary features/POS bigram configuration*

The results with this slightly different configuration differ from the results using punctuation instead of POS bigrams, except for the presence of the hashtag. The hashtag, here in conjunction with nouns and proper nouns (e.g. London, Tottenham, riot, etc.), exhibits a higher recall value than most other features. From these experiments we can conclude that the use of a hashtag when searching for tweets with actionable information increases the likelihood that some of the retrieved tweets will be useful for analysis, although not all of the tweets containing hashtags will have actionable information.

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>3pm</td>
<td>1</td>
<td>0.023</td>
<td>0.045</td>
</tr>
<tr>
<td>tottenhamriot</td>
<td>1</td>
<td>0.013</td>
<td>0.025</td>
</tr>
<tr>
<td>direct</td>
<td>1</td>
<td>0.013</td>
<td>0.025</td>
</tr>
<tr>
<td>davidgraeber</td>
<td>1</td>
<td>0.013</td>
<td>0.025</td>
</tr>
<tr>
<td>corner</td>
<td>0.875</td>
<td>0.018</td>
<td>0.035</td>
</tr>
</tbody>
</table>

*Table 8: 5 features with the highest precision value indicating actionability*

The features with a higher precision value were the same for both of the configurations used previously. The features listed above correlated strongly with tweets containing actionable information. Location and time are both represented in the high precision features and tend to be specific information which fulfill the first criteria for
determining actionability set by the coder guidelines created for this paper. Time is also a primary factor of relevance used by Twitcident.

Prioritizing either recall or precision can have serious consequences when the text mining is being used to identify important information in potentially dangerous and volatile situations. Having to analyze many more tweets than will give the analyst information or missing tweets that contains actionable information are unsatisfactory options for the intelligence professional. Using the F measures is a compromise that balances the benefits and costs of valuing only recall or precision. Once again, hashtags are a key feature.

<table>
<thead>
<tr>
<th>Feature</th>
<th>$F_1$ measure</th>
<th>$F_2$ measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER_SIGN</td>
<td>0.356</td>
<td>0.446</td>
</tr>
<tr>
<td>PERIOD</td>
<td>0.313</td>
<td>0.389</td>
</tr>
<tr>
<td>&lt;NUMBERSIGN&gt;_NNP</td>
<td>0.308</td>
<td>0.337</td>
</tr>
<tr>
<td>&lt;PERIOD&gt;, &lt;NUMBERSIGN&gt;</td>
<td>0.296</td>
<td>0.313</td>
</tr>
<tr>
<td>&lt;NUMBERSIGN&gt; NN</td>
<td>0.289</td>
<td>0.305</td>
</tr>
<tr>
<td>in [occurs in both]</td>
<td>0.27</td>
<td>0.290</td>
</tr>
<tr>
<td>COMMA</td>
<td>0.263</td>
<td>0.285</td>
</tr>
<tr>
<td>&lt;NUMBERSIGN&gt;_CD [cardinal number]</td>
<td>0.256</td>
<td>0.238</td>
</tr>
<tr>
<td>at</td>
<td>0.254</td>
<td>0.234</td>
</tr>
</tbody>
</table>

Table 9: Features with the highest $F$ measures indicating actionability using both of the previous configurations

The trained models had more variable degree of success, suggesting that different types of events have different properties that affect the success of predictive modeling.

The results from the trained models used on test sets are shown in the table below.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>$F_1$ measure</th>
<th>$F_2$ measure</th>
<th>Recall</th>
<th>Precision</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupy (POS config.)</td>
<td>0.595</td>
<td>0.634</td>
<td>0.500</td>
<td>0.733</td>
<td>85% (170)</td>
</tr>
<tr>
<td>Occupy (Punc. config.)</td>
<td>0.565</td>
<td>0.587</td>
<td>0.636</td>
<td>0.509</td>
<td>78.5% (157)</td>
</tr>
<tr>
<td>London (POS config.)</td>
<td>0.108</td>
<td>0.133</td>
<td>0.25</td>
<td>0.069</td>
<td>83.5% (167)</td>
</tr>
<tr>
<td>London (Punc. config.)</td>
<td>0.125</td>
<td>0.125</td>
<td>0.125</td>
<td>0.125</td>
<td>93% (186)</td>
</tr>
<tr>
<td>All collected tweets</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84.25% (337)</td>
</tr>
</tbody>
</table>

Table 10: Selected data from trained models used on test sets
In Table 10, the model that had the best results for the Occupy corpus used the features that were generated with the POS bigrams configuration. The London riots had better results with the model using features generated by the punctuation configuration for $F_1$ measure, but a higher $F_2$ measure from the model with the features generated by the POS bigrams configuration. The results from the combined corpus had the same results for both configurations. As precision and recall (and therefore the f-measures) both use correctly classified positive data in their calculations ($TP/(TP + FP)$ and $TP/(TP + FN)$, respectively) and the Occupy corpus had the higher proportion of tweets with actionable information, the Occupy model still had the better precision, recall, and f-measures even though it had the lower accuracy. The model for the combined corpus simply predicted that every tweet was negative and consequentially contained zero true positives; precision, recall, and the f-measures were thus also zero.

**Reliability Evaluations**

By the standards of source reliability proposed by Ulicny and Kokar (2012) and based on STANAG 2022 conflated with TunkRanks, only established news sources such as the BBC and the Guardian in the London riots and the official OccupyWallStNYC and OccupyWallSt Twitter accounts which have over 100,000 followers each would be considered objectively reliable. They are also prolific Twitter users and users of retweets and sourcing information from other tweets. The opportunity for non-famous individuals to communicate with a large audience is one of the most important aspects of Twitter in this study. Twitter users with a large following are likely to have other available avenues
of communication that are also heavily trafficked, such as dedicated websites. For most Twitter users, their tweets are their most visible communications and they have much smaller numbers of followers. Considering that the TunkRank approach, favors only a small percentage of users, the link between geographic proximity to an event and perceived high-reliability information on that event studied by Thomson et al. (2012) is a better basis for evaluating the source reliability for Twitter users in the events studied when that information is available, but it is also the more difficult factor to determine.

On the basis of reputation, popularity, and proximity, the majority of tweets used in this study came from reliable sources; however, the reliable sources did not always tweet actionable information. Two hundred thirty-eight of the tweets in the Occupy Wall Street corpus came directly from the user @occupywallst, rated A based on the large number of influential followers and the user’s presence at the event, @occupywallstnyc is rated A for the same reasons. Users @bbcnews and @guardian, although less prolific tweeters of the events than the previous two users, are still rated A for the London riots and Occupy Wall Street. “[D]avidgraeber,” one of the features found to have a high precision for actionable tweets, indicates that member of Occupy Wall Street and Twitter user David Graeber provided or was connected to actionable information in almost 34% of his tweets included in the Occupy corpus. The other members of the Occupy movement seem to have identified him as a reliable source for first-hand information, which supports the theories of reliability linked to followers. As Graeber has more than 10,000 followers and demonstrated his proximity to the Occupy protest, he is a B rated source for that event. Twitter user @sabzbrach was also demonstrably an active participant in the protest at Zucotti Park and tweeted actionable information in 30% of
her tweets which were included in the study’s corpus, but has less than 10,000 followers and so is rated C.
ANALYSIS & DISCUSSION

Out of the 2000 tweets which comprise the corpora used in this paper, 1611 tweets, almost 81%, do not contain actionable information. The Occupy corpus had a higher proportion of tweets with actionable information than the London riot corpus, with 225 (22.5%) out of 1000 tweets possessing actionable information for the Occupy corpus, compared with 156 (15.6%) of London riot tweets. Given the informal nature of discourse on Twitter, it is reasonable to think that the relative lack of actionable information reflects the typical use of the medium by a largely non-professional-user base.

Based on the results of this study, it is feasible to retrieve tweets with actionable information through search Twitter for hashtagged topics, particularly locations. However, this strategy only works if the general public in the area and the analysts are aware of the event and if Twitter is used by the population of the location (or participants of the event) being analyzed. If an event does not have immediately obvious effects and the local or national authorities where the event occurred do not want anyone to know, it is highly unlikely that the event will be reported on Twitter in a timely manner, if at all. Likewise if the event occurs in a remote area or in a culture to whom social media is not familiar, information regarding the event will take time to filter through more traditional channels of communication before it reaches Twitter, if it does at all. Even then, organized and promoted events will have a superior presence on Twitter, due to either the
purposeful dissemination of information to the general public or use of Twitter by event participants and coordinators as a promotion tool or both. The London riots, a chaotic and spontaneous event, had fewer tweets with actionable information than the organized Occupy protest. The London riot tweets were also more frequently tweeted by sources who had fewer followers and were therefore judged less reliable. In an unplanned event, the hashtags are less structured (see the diversity in hashtagged terms for the London riots in the Methods section) and sources less vetted.

Searching for sources with fewer than 100,000 followers who had tweeted actionable information without indicating that they are in the area of the event they are describing introduced a question: how would an individual who is not in the proximity of the event be able to provide actionable information? They would either be tweeting false information or they found the information through another source. On Twitter this would most likely be a retweet, but the primary source could also be a phone call from a friend, an overheard conversation, etc. If the tweet in question is a retweet, an effort should be made to identify the original source and evaluate that source for reliability. Neither this study nor any of the studies cited herein have found a way to eliminate misinformation from Twitter. Sources with a higher number of followers are more likely to be reliable and provide credible information, while sources in proximity to an event are more likely to provide actionable information.

Although it is feasible to use this method to retrieve tweets for intelligence analysis, the results of the study are not so encouraging as to be adopted in its current incarnation. Tweets retrieved from third-party sites, such as the one used in this study, omit the location metadata connected to the original tweets. Researchers with greater
technological capacities analyzing Twitter directly would be able to gather the metadata to find the approximate location the tweet originated from. Otherwise, researchers trying to determine Twitter user locations have to rely on the honesty of the Twitter user’s profile or contextual clues in the tweet. Additionally, the order of research taken in this study—to find actionable information and then determine who the reliable sources of information are—is a backwards approach. The logical approach then is to identify the most followed Twitter users for a given area where an event is occurring (or an area of interest that requires monitoring) and mine their tweetstream for hashtags.

The reliability analysis used to produce the claim of source reliability in the results section of this paper doesn’t reflect all of the available reliability theories. Several of the studies referenced have their own approaches to reliability. Twitcident collects the profile information and statistics of Twitter users when it gathers event-relevant tweets from that user; the metadata can be used by the Twitcident consumer as well as the program itself to assess trustworthiness and relevance (Abel et al., 2012). Thomson et al. (2009) warn against over-reliance on their findings correlating proximity and reliability as the Twitter response to Fukushima is considered unusual for its lack of overt spam or misinformation, and the relatively limited amount of users contributing with the hashtag #fukushima, which the researchers used to collect relevant tweets.

Source independence is an important property of STANAG 2022 source reliability criteria and one that is insufficiently addressed in the context of social media by that standardization agreement. Source independence (a lack of connection between sources) can affect information independence (information that has not been affected by or copied directly from another source without independent verification by the first
Source independence avoids one unreliable source spreading information to the point that poor information is reported by several sources, becoming credible by STANAG 2022 standards. Also, independence can be understood in several ways. For websites, having a different domain name is generally enough distance to be considered independent; Twitter, with its network of “followers” and “following” makes independence more difficult to determine. A retweet is obviously not from an independent source, neither are multiple users including links to the same website in their tweets. Tweets whose content is identical, suggests that users may have copied information from a single original source, this is also not independent. In fact, the average path length between any two users on Twitter has been determined empirically to be only 4.12 links (Ulicny & Kokar, 2012), the authors have interpreted this to mean that two sources can be determined to be independent if they have at least four degrees of separation. Paradoxically, the more reliable a source of information is, the more people are likely to follow that user and the path length between users shrinks.

The degrees of separation between several of the sources evaluated in this paper is less than that. @OccupyWallSt and @OccupyWallStNYC are discrete users, but they do follow each other and many of their followers follow both; @OccupyWallSt, @OccupyWallStNYC, and Twitter users who follow one or both of these accounts and retweet, rephrase, or elaborate on someone else’s tweet are capable of creating an information echo chamber almost immediately thanks to the collective reach of the popular accounts followers and the followers’ followers. Counting their followers’ followers both accounts reach millions of people while on their own, neither has had even 200,000 followers. The following chart looks at the number of users reached by the 50
biggest followers (those with the largest number of followers) of the five biggest Twitter users included in this corpus and an individual user selected from the corpus at random for comparison:

<table>
<thead>
<tr>
<th>User 1</th>
<th>Reach of 50 Biggest Followers</th>
<th>User 2</th>
<th>Reach of 50 Biggest Followers</th>
<th># Shared Big Followers</th>
<th>Reach Shared Big Followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>@occupywallst</td>
<td>31,700,866</td>
<td>@occupywallstnyc</td>
<td>23,810,504</td>
<td>24</td>
<td>9,365,949</td>
</tr>
<tr>
<td>@occupywallst</td>
<td>31,700,866</td>
<td>@adbusters</td>
<td>9,523,551</td>
<td>11</td>
<td>2,384,836</td>
</tr>
<tr>
<td>@adbusters</td>
<td>9,523,551</td>
<td>@occupywallstnyc</td>
<td>23,810,504</td>
<td>14</td>
<td>3,419,038</td>
</tr>
<tr>
<td>@bbcnews</td>
<td>89,729,944</td>
<td>@guardian</td>
<td>75,205,599</td>
<td>17</td>
<td>33,431,071</td>
</tr>
<tr>
<td>@occupywallst</td>
<td>31,700,866</td>
<td>@guardian</td>
<td>75,205,599</td>
<td>5</td>
<td>3,006,464</td>
</tr>
<tr>
<td>@jamyerson</td>
<td>3,664,756</td>
<td>@occupywallstnyc</td>
<td>23,810,504</td>
<td>6</td>
<td>1,174,425</td>
</tr>
</tbody>
</table>

Table 11: Reach of the five largest users of the corpus, one non-institutional user, and their overlap as calculated by Twiangulate (http://twiangulate.com/search)

The Twitter account for J.A. Myerson, a reporter for a liberal internet radio program who participated in and reported from the Occupy encampment, has less than 5000 followers. Several of Mr. Myerson’s tweets were retweeted by other users, including the official Occupy Wall Street Twitter accounts. Mr. Myerson may tweet about the number of protesters at the Occupy encampment, information from Myerson’s tweet may be used in a tweet or simply retweeted in its entirety by @occupywallst. The New York Times, a respected newspaper, follows @occupywallst on Twitter and may use Myerson’s information in an article on the protest. A reader of the Times who doesn’t even use Twitter is still only three relationships from a Twitter user whose information they have unknowingly consumed.

Figure 6: Diagram of 3 degrees of separation.
In the London riots, however, even taking into account BBC News and *The Guardian*, more of the sources were independent as the active participants of the riots were not reporting on their own activities. Instead, most of the tweets came from bystanders who acquired the information they tweeted by chance (e.g. coming across a looted shop in the process of a regular commute). The lack of organization in the riots simultaneously reduced the amount of actionable information, while increasing source, and therefore information, independence.

Even with reliable sources and relevant, credible information, using tweets for intelligence data requires a discussion of ethics and practicality. To begin with the latter, “reliable” as determined by standardized criteria or well-informed personal judgment is not necessarily reliable in reality. Almost every community on the internet is familiar with “trolls,” people who are intentionally disruptive. However, in matters of security and intelligence, there are even more malicious threats. Twitter is used by people from every walk of life, including terrorists, who are aware of intelligence gathering operations and may attempt to subvert information collection programs with intentional misinformation, distractions, or even false flag operations.

An informed, responsive information collection, analysis, and evaluation policy is needed to combat the detrimental effect of terrorists, trolls, opposition members, and the well-meaning misinformed. As mentioned in the literature review, Khonsari et al. (2010) estimated that around 24% of Twitter users participating in discussions during the Green Revolution were supporters or possibly even employees of the Iranian government; although part of the blame rests with the infiltrators, the authentic members of the network are responsible for their failure to develop good information policies and being
poor “gatekeepers” (p. 1). Oh et al. (2011) have two suggestions to prevent a situation where publicly reported information is used against the public interest, as in Mumbai:

“First, information control measure needs to be implemented in a way to encourage public accountability, and, at the same time, guarantee freedom of speech. Second, during terrorist attacks, government or security forces should monitor social reporting media such as Twitter and actively be involved in the social reporting process. That is, government and security forces need to make citizens aware what is harmful or desired information for security operation with credentials and authoritative voices by actively using such communication channels as hyperlink, blog, RSS, email, text message, live TV and Retweet etc.” (p. 41).

On an individual level, Twitter users can set privacy settings for their account, but restricting access to information negates the purpose of tweeting relevant information during an event in the first place.

Much has been said about intelligence community work in relation to these areas of study. Militaries use similar tools and concepts, albeit less frequently involving continuous Twitter crawling as a collection technique. Military research and funding has been responsible for several technological advancements and conceptual frameworks used by diverse disciplines. Ulicny, Matheus, Powell, Dionne, and Kokar (2007) stress the value of accurate contextual knowledge for military operations on the first page of their paper; the same could be said for intelligence operations which are often quasi-militaristic in execution if not in approach or mindset of the participants:

“The military commander’s view of the state of the battlespace is constructed from reports generated by both human and non-human sensors in the battlespace. In many cases… the import or credibility of what is being communicated by these reports is determined by the background information. Therefore, unless background conditions are accurately described, an accurate depiction of what is currently going on in the battlespace is not possible. As such, background information is not simply ancillary information in a military environment” (p. 1)
This paper has primarily referenced STANAG 2022 as the standard for source reliability, however the standard of vocabulary and some concepts that STANAG 2022 was built from, JC3IEDM. JC3IEDM was originally created and intended for international military cooperation, it is primarily used to facilitate the exchange of information between information systems (Multilateral Interoperability Program-North Atlantic Treaty Organization Management Board [MNMB], 2004). It is still in use and can be used in conjunction with IR activities, among other possibilities. For example, JC3IEDM advocates tagging information and sources for accuracy and reliability, it also provides standardized vocabulary, enabling a message to be communicated and understood across the Intelligence Community and the military (MNMB, 2004). Its use is expanding into a resource for building C3 (consultation, command, and control) information systems. A primary reason for JC3IEDM’s new role is its depth of expert knowledge and extensive vocabulary consisting of 289 entities, 396 relationships between entities, 1729 entity attributes and nearly 7000 value codes, which might otherwise be needlessly repeated in attempts to design what JC3IEDM already provides (Ulicny et al., 2007).

In the nearly seven years since Twitter was created, an industry of text and data mining has grown up around it, social analytics sites (such as Twiangulate, used in this study) cater to helping users understand Twitter data, researchers develop programs specifically for it. Twitter is an internet phenomenon whose value has not yet begun to peak. Twitter will probably be the subject of many more information innovations before it becomes obsolete. Analytical programs are becoming increasingly beginner-friendly, a
trend which will probably see text and data mining become a mainstream activity within a decade, likely honed on Twitter’s abundant public data.

One application that would be both an analytical challenge and potentially beneficial to public well-being is the development of technologies that can use the copious available data on Twitter to build predictive models. As this paper demonstrates in its reliance on known locations, at their best Twitter mining applications are good at recognizing what is going on, but currently still lack the capacity for predictive intelligence unless the user-analysts already know what location or topic they are looking for. Public health officials have tracked reports of cold and flu symptoms through tweets and have been able to anticipate patterns of impending outbreaks. If indications or “symptoms” can be accurately linked to future events, Twitter could become a technological crystal ball. Much of this paper has focused on the premise of text mining tweets to inform law enforcement and intelligence community understanding of events. Once data collection is completed, analysis is the next phase in the intelligence life cycle. In professional practice, the information would be analyzed with specialized framework and an extensive understanding of the subject matter. The final result, actionable information that has been analyzed, is the product known as “intelligence.”
CONCLUSIONS

The intention of this paper is to demonstrate one possible approach to using text mining to extract useful intelligence data from event-related tweets. Software applications for machine learning can not only work faster than a human when properly programmed, they can identify patterns that may have otherwise gone unrecognized. The analysis presented above demonstrates three things: the multi-faceted usefulness of iterations of location for information retrieval, relevance judgments, and source reliability; the importance of the hashtag, one of the few consistent communication methods used on Twitter; and the fundamental uncertainty that working with Twitter entails.

However, the study done for this paper barely scratches the surface of methods to exploit Twitter. Individuals and organizations with far more technical skill and resources are drawn to Twitter for the possibilities it holds for OSINT, competitive intelligence, epidemiology, and other fields of study and work. This paper presents only one option for the text mining novice to begin a pre-intelligence program using Twitter, a few simple concepts, and readily available resources. The entire corpus being evaluated by a single coder, even with a test for the universality of concept (of actionability) and inter-operator agreement (determined by the sample coding described above), invites the possibility of bias or poor judgment of actionability. This is a valid criticism that can be addressed in future studies.
In addition to the results of the text mining portion of the paper, several current approaches to the analysis of Twitter were presented for reference and debate. The volume of literature and research on the uses and characteristics of Twitter is growing as the application becomes more embedded in modern communication and culture. Twitter is a relevant topic to a diverse selection of disciplines and audiences and this is reflected in the creation of new tools to include Twitter in the existing range of studied media, as well as in the modification and synthesizing of existing benchmarks and analyses.

In spite of the popularity of Twitter as an area of study in text mining, IR, and intelligence research, there are no concrete answers. Being able to find new, reliable sources of actionable information through Twitter is the goal of every OSINT analyst, but part of the nature of Twitter is to repeat information and misinformation is a common feature of the internet as a whole. It is the recommendation of this study that the most practical approach to finding event-specific information on Twitter that is both reliable and actionable is to search the location of the event for the most popular Twitter users and then mine their tweets for hashtags, the feature of tweets found by this study to frequently be linked with actionable information.
REFERENCES


International Risk Governance Council [IRGC] (2012). Addressing the challenges of using social media to improve crisis communication and management (Concept


Oh, O., Agrawal, M., & Rao, H. R. (2011). Information control and terrorism:

Tracking the Mumbai terrorist attack through twitter. Information Systems Frontiers, 13(1), 33-43. doi:10.1007/s10796-010-9275-8


Retrieved from http://www.guardian.co.uk/uk/series/reading-the-riots


APPENDIX

Coding Guidelines

The three key factors of actionable information are specificity, objectivity, and utility. Each tweet must be considered separately, without presumed knowledge gained from any other tweet.

- Specificity- has details. E.g. around midnight, 2 miles from Chapel Hill, an old sedan.
- Objectivity- an assertion of a fact, not an opinion. “The car that caught on fire was blue” is objective, “her hat made me sad” is not.
- Utility- the information can be used to inform actions. This is an area where your best judgment comes into play. For the purposes of this project we will assume that the target audience for this information is law enforcement and/or intelligence personnel. The context of the information can also inform its utility; the context to the tweets is provided below. When reading these tweets ask yourself “How would a police officer law enforcement agent, or intelligence community officer be able to use this information?” If the answer to that is “They could not,” or the answer is absurd or unlikely, then that tweet has no utility. For an example- a tweet stating that leaders are all wearing a
white armband could be used by your audience to determine who to focus surveillance on. A tweet saying that the protest group is having pizza for dinner again might be useful if the police think the group is using the food delivery as an illicit courier, but that seems unlikely in the context of a peaceful and open protest. This tweet has no use for your audience. On the other hand, a tweet giving the number of pizzas delivered to the protest or where they are from might give the police an idea of the support the protest has, the resources available to them, or another location to observe for information. This tweet might be used by your audience.

Context for coding the tweets:

The data in these files revolves around two events- the London Riots and the Occupy Wall Street movement. Later surveys cited a variety of festering problems that contributed to the London riots, but the most direct catalyst was the shooting death of Mark Duggan by the police in Tottenham in the North of London on August 4th, 2011 (The Guardian, 2012). The riots began in that same area on August 6th. Approximately 200 people staged what started as a peaceful protest at a local police station when it suddenly turned chaotic and destructive. There were a number of allegations for the direct cause of the riots, including police brutality of one of the protestors, but no substantial evidence was ever proven for any of the theories. Aside from Tottenham, riots sprang up in several other neighborhoods, apparently spurred on by, but independent of the original event. The first three days of rioting were escalating, but subsided into less formidable and more easily controlled chaos after that. The majority of the violence was directed
primarily at buildings, cars, and other property; however, 5 civilian deaths and 16 civilian
injuries are attributed to the rioters, along with numerous claims of injured law
enforcement. The collected tweets in this dataset are from August 6-10th.

Occupy Wall Street, or OWS, began as a protest in New York’s Wall Street
district against American income equality and irresponsible financial institutions that
grew into an international popular movement. The original protest movement, ostensibly
organized by Canadian magazine Adbusters, began with a gathering in Zuccotti Park on
September 17, 2011. Aside from the encampment, several marches were undertaken
around the city. The NYPD made several arrests over the course of the movement citing
various laws and city codes, allegations were made of police brutality and manipulation.
The protestors themselves were accused of some instances of theft and sexual assault
against other protestors, but remained generally peaceful. The movement was highly
polarizing, opinions of protestors and police alike were the subject of many debates. The
protestors were removed en masse from Zuccoti Park in mid-November, signaling a
decline in the organized movement; however re-encampments have taken place
repeatedly, albeit without the same success or support as the original. The collected
tweets in this dataset are from September 17-24th.