A Comparison of Two Methods for the Topical Clustering of Social Media Posts

Elizabeth Williams*, Jeff Gray†, Edwin Morris‡, Ben Bradshaw,‡, Keegan Williams‡, and Brandon Dixon†
*University of Alabama, Tuscaloosa, Alabama
eawilliams2@crimson.ua.edu
†University of Alabama, Tuscaloosa, Alabama
{gray,dixon}@cs.ua.edu
‡Carnegie Mellon Software Engineering Institute, Pittsburgh, Pennsylvania
{ejm,bwbradshaw,kmwilliams}@sei.cmu.edu

***Abstract*—While traditional search engines can often provide useful knowledge, some information that is more temporal and contextual can be found more easily or exclusively on social media. However, the number of posts on many social media platforms is too large to manage without a method for clustering posts into distinct topics. In this paper, we compare two approaches for the topical clustering of Twitter posts to discover various subjects that users are talking about. The first approach is GeoContext, our tool for clustering a social media stream into topics that uses a unique approach for calculating the similarity between posts. We compare GeoContext against Latent Dirichlet Allocation (LDA), a commonly used topic modeling algorithm. LDA has been used as a basis in several different approaches for clustering social media posts.**

I. Introduction

Social media platforms, especially those intended for microblogging, provide information to a large number of people very quickly. For this reason, many users rely on microblogging platforms such as Twitter for knowledge of major events. For example, Twitter was used for the real-time dissemination of information during situations such as the Iran elections, tsunami in Samoa, and earthquakes in Haiti [1]. Twitter was also used in the spread of important data regarding potential suspects after the 2013 Boston Marathon explosions [2].

Early work in this area focused on analyzing the characteristics of microblogging services such as Twitter. For example, Java et al. [3] studied the types of information about which Twitter users are posting, as well user's intentions when utilizing Twitter. Newer research has focused on applying topic models to streams of posts from social media sources. Discovering topics within social media can assist in recommending content to users based on their topical interests. Topical clustering can also be useful in finding major events and occurrences, often before they are reported by other forms of traditional media. Sakaki et al. [4] explored the use of Twitter as a method for detecting earthquake activity.

Identifying topics on Twitter using traditional natural language processing techniques can be challenging due to the short allowed length of tweets (140 characters or less). However, the short character limit means that tweets are often limited to a single topic, which makes them a good candidate for topic categorization. Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) [5] can provide a solution for analyzing the content of tweets. In addition, other methods may be needed to improve the classification of tweets beyond what is possible with traditional topic models.

978-1-5090-1496-5/16/$31.00

Existing methods for topic discovery do not focus on the relevant words out of social media posts within a stream. This means that topics may have extraneous words that do not contribute to the overall meaning of the topic. In this paper, we evaluate whether extracting important keywords and assigning values to terms based on relevancy improves the overall topics discovered. The contribution of this paper is as follows:

1) We outline two approaches for topic discovery of a social media stream. The first approach, LDA, is a common topic modeling algorithm. The second approach, GeoContext, is a topical clustering method based on semantic text analysis.
2) We evaluate both approaches using two different datasets. We were able to compare the resulting extracted topics from both methods.

In this paper, we evaluate the accuracy of two methods in categorizing tweets into distinct topics: GeoContext (our system for calculating similarity between tweets to perform topical clustering) and LDA. In Section II, we describe existing work in the area of discovering topics within social media. In Section III, we present an overview of the two approaches, and in Section IV, we outline the evaluation process and describe the results of the evaluation. Finally, Section V contains future work and concludes the paper.

II. Related Work

In this section, we describe the existing state-of-the-art work related to topic discovery in social media. Many approaches utilize Twitter as the chosen social media platform for research.

Vosecky et al.'s approach [6], the Multi-Faceted Topic Model, captures all facets of information from a Twitter stream, including entities present and temporal information. The model incorporates all these types of data in order to discover latent topics from the stream. Their approach is similar to GeoContext in that not all terms within a tweet are treated equally; entities such as people and organizations are extracted separately.

Hong et al. [7] built their Content Model in order to detect bursty events within social media. The Model is based on Binomial Logistic Regression. The authors also include their Pseudo Relevance Feedback module, which assists the system in adapting to content drift, or the change in a single topic over time.

Sakaki et al. [4] used social media as a tool to detect earthquake and seismic activity. They classified tweets based on keywords within the tweet, the number of words in the tweet, and the words other than the keywords (the context) in the tweet. The system was able to detect a high probability of earthquakes at a faster rate than many warning systems, showing the potential value of topic discovery and event detection within social media.

Kim et al. [8] normalized high frequency words within a Twitter stream over time in order to reveal words that dramatically increased in frequency very quickly. Their approach was able to discover terms related to bursty events such as major holidays.

The goal of Ramage et al.'s method [9] was to map the Twitter stream into topics such that users are able to receive updates of topics that are relevant to them, rather than only updates from people that they follow. They used Labeled LDA to map individual tweets into latent topics.

Aiello et al. [10] investigated several methods for topic detection within social media, including LDA, document-pivot topic detection, graph-based feature-pivot topic detection, frequent pattern mining, soft frequent pattern mining, and BNgram. They were able to discover topics from three different datasets.

Yuan et al. [11] described their model, EW4 that uses a generative process to model tweets along with their day, time, words, and location. Their method incorporates four aspects of user behavior in order to improve contextual search and suggestions.

There are several limitations with the existing approaches for topic discovery. First, existing methods mainly treat all text within the social media stream the same. In contrast, GeoContext extracts only words from a tweet that are important and then creates topics based on how relevant the words are to the tweet. This method ensures that topics include only terms that are highly relevant to the topic. Second, several of the existing approaches require a training set in order to train the model. GeoContext works dynamically with no prior initialization or training. Unlike LDA, GeoContext also does not require a fixed set of topics. Rather, GeoContext is able to create new topics as they appear in the stream dynamically.

III. Overview of Two Methods for Topic Discovery

In this section, we describe the two methods for discovering topics within a social media stream: GeoContext and LDA.

A. GeoContext

Although existing topic modeling algorithms such as LDA often remove stop words (words such as a, an, or the that do not add meaning to a piece of text), these methods initially treat all other words in the text as having the same value. LDA then determines topics based on words that often appear together. In contrast, GeoContext takes the approach that terms are not all equal. GeoContext ranks words based on the relevance to the search criteria.

TABLE I
Concept and Keyword Extraction

Tweet	Concepts	Keywords
Pipeline Spills More than 40,000 Gallons of Crude Oil into Yellowstone River, Possibly Contaminating Water Supply http://t.co/OqwntBuSFo	Petroleum (0.949839)	Possibly Contaminating Water (0.93474)
	Water (0.64515)	Crude Oil (0.509798)
		Yellowstone River (0.491219)
		Gallons (0.313184)
		Pipeline (0.258474)

To extract meaningful terms from a tweet, GeoContext utilizes AlchemyAPI's Concept Tagging API[1] and Keyword Extraction API[2]. The Concept Tagging API takes a piece of text as input and abstracts the text into higher-level concepts. For example, as shown in Table I, the tweet shown in the leftmost column is talking about an oil spill into the Yellowstone River. The concepts returned from the Concept Tagging API, shown in the middle column of Table I are *petroleum* and *water*, which represent some of the abstracted topics of the tweet. The Keyword Extraction API returns important and meaningful words extracted directly from the text. The keywords extracted are shown in the rightmost column of Table I. As displayed, the extracted keywords represent the terms from the tweet that give the most meaning.

Each returned concept and keywords is also associated with a relevance score, shown in parentheses in Table I. The relevance score is a value of how important the extracted concept or keyword is to the text. For example, the concept *petroleum* has a higher value of importance to the tweet than the concept *water*.

GeoContext calculates a similarity score between tweets to determine the relatedness of the topics of various tweets. Tweets that have high similarity scores are clustered together into the same topic. The similarity score calculation is shown in Formula 1.

$$tweetSimilarity = max(hashtagsMatch(t1, t2), \prod_{a=0}^{b} avg(relevanceScore(t1_a), relevanceScore(t2_a))) \quad (1)$$

[1]http://www.alchemyapi.com/api/concept-tagging
[2]http://www.alchemyapi.com/api/keyword-extraction

The similarity score can range from 0 to 1, 1 meaning that the tweets are the most similar and 0 meaning that the tweets are the least similar in topic. Given the content of two tweets, *t1* and *t2*, the similarity score is calculated by first determining whether the tweets have any hashtags in common. A hashtag is a device used on some social media platforms to express a specific or popular topic [12]. The *hashtagsMatch* method returns 1 if there are matching hashtags in the two tweets and 0 otherwise.

If two tweets do not have any hashtags in common, GeoContext looks at the extracted concepts and keywords. For any concepts or keywords that match between the two tweets (case-insensitive), GeoContext determines the product of the average of the relevance scores of each matching concept or keyword. In Formula 1, given that *t1* and *t2* are two tweets with matching concepts or keywords, and *a* is the index of the matching concept or keyword, $t1_a$ and $t2_a$ are the a^{th} concepts or keywords that match. Given that *b* is the number of concepts and keywords that match between *t1* and *t2*, GeoContext calculates the average of the relevance scores of all concepts and keywords from *a* to *b*.

By extracting important keywords and using their relevance scores, GeoContext has several advantages. As mentioned previously, all words are not treated equally within the tweet, which results in topics being formed of terms that are more relevant to the topic of the tweet. Also, using keyword extraction eliminates the need for stop word lists. Second, by utilizing relevance scores, GeoContext is able to match tweets into topics that contain terms that are more related. Keywords with low relevance scores do not necessarily end up in the same topic even if they match, which also results topics with more meaningful terms.

Unlike LDA or other topic models, GeoContext does not require a fixed number of topics to be determined beforehand. Rather, topics are generated dynamically based on similarity between social media posts. This allows GeoContext to be used without prior training on a social media stream.

GeoContext has two possible outputs:

1) *tweet clusters*: a set of clusters of tweets grouped together into topics based on the semantic relatedness between tweets
2) *a topic model*: a model, similar to the output of LDA and other topic models, in which topics contain a non-ordered mixture of terms that describe the cluster of tweets. Because traditional topic models are sets of terms directly from the text, GeoContext utilizes only the AlchemyAPI Keyword Extraction API when creating a topic model.

B. LDA

Latent Dirichlet Allocation (LDA) is a commonly-used topic modeling algorithm. LDA takes as input a corpus, which is a set of documents that contain words not considered by LDA to be in any sorted order. In the algorithm, each document can be a mixture of various topics. The distribution of topics among the documents is assumed to have a Dirichlet prior distribution.

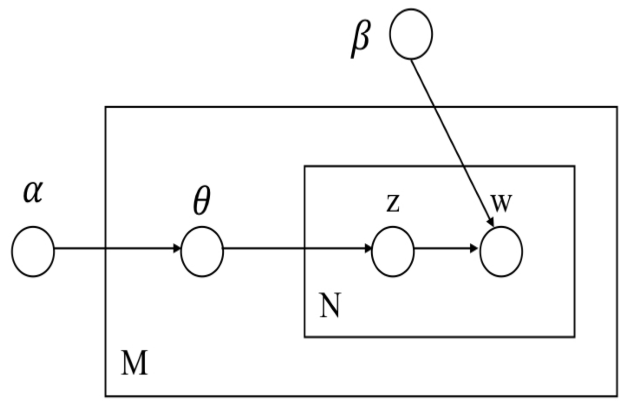


Fig. 1. Plate notation representation of LDA (adapted from [5]).

TABLE II
EXAMPLE ELECTIONS DATASET TOPICS

bernie sanders win wins won call called calling projecting project projects projection hold held senate senator vermont vt
sc carolina romney mittromney mitt wins call projects called held won calling project projection win projecting hold
ma massachusetts wins call projects called held won calling project projection win projecting hold elizabeth warren
barackobama barack obama best come yet

LDA produces topics that consist of terms that often appear together in the text of the documents. The number of produced topics is fixed prior to execution of the algorithm. Because LDA assumes that each document is a bag-of-words, where terms within the documents are unordered, all terms are initially treated equally.

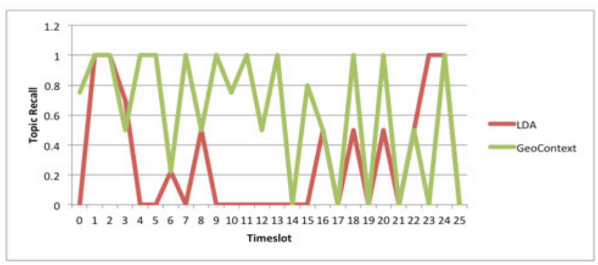
LDA uses a sampling method to calculate both the topic distribution over documents and the topic distribution over the terms. A plate notation representation is shown in Figure 1. *M* represents the number of documents and *N* represents the number of words within a document. The variable *w* represents a single word within a document, *z* represents the topic of the word, and θ represents the topic distribution over a document. α and β are the parameters of the assumed Dirichlet distribution.

LDA produces two types of output:

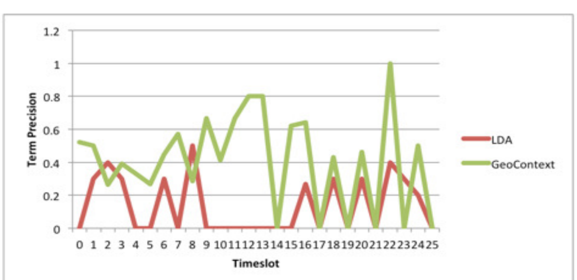
1) *Topic distribution over documents*: the mixture of topics found in each document. LDA calculates the percentage of each topic that the document contains.
2) *Topic distribution over words*: the mixture of words within each topic. LDA calculates the percentage each word contributes semantically to each topic.

TABLE III
SAMPLE DISCOVERED TOPICS OVER TIMESLOTS

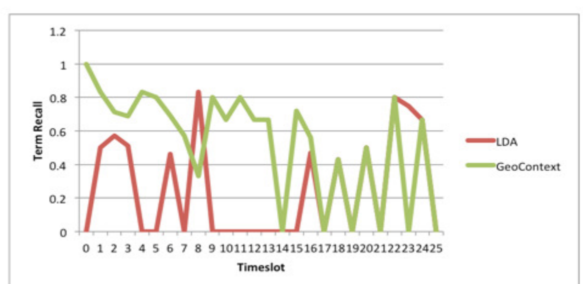
	Timeslot 0	**Timeslot 9**	**Timeslot 12**	**Timeslot 18**	**Timeslot 25**
LDA	cnn election luck night obama good win results early lead	wins wisconsin ryan michigan amp romney obama ohio win home	mccaskill wins akin projects romney electoral obama votes state mitt	elected obama congratulations black voted term win years president america	barack election speech victory obama gt chicago live president supporters
GeoContext	#romney, surprises, Indiana, #Romney, #Obama, #USelection, indiana, kentucky, #obama, #kentucky, #ROMNEY, #Obama2012, RACE, Vermont, #ElectionDay2012, Kentucky, #romney, #romney #romney #romney, votes	vote, america, RT, Romney, Obama, BBCNewsUS, #Florida, LIVE, t.co, rt, romney, #election2012, florida, Florida waiting, line, election, #obama, guy, Ohio, Michigan, auto, job, votes, #FLORIDA #election2012, Vote, country, #Obama, VOTE, #stayinline, #OBAMA, #obama2012, #OBAMA #FORWARD, stay, obama	line, #TEAMOBAMA #Obama2012 #Forward, LINE, FLORIDA, #stayinline, #Obama2012, #Election2012, #stayinline, polls, votes	president obama, Washington power structure, President Obama, Dems, GOP, Senate, change, House, t.co, you., heart, #Election2012, White House, RT, chance, leader, #election2012, Retweet, #FourMoreYears, NBC, election	president, #obama, President, #Obama, campaign headquarters, t.co 5L0Y7c4H, apps, Chicago, victory speech, chicago, t.co MT73sKxJ, stage, Chicago, live, way, long voting lines, issues, CNNelection, chicago, RT, VP MartinSchulz, EU, USA, America, Congrats, admiration, respect, speeches, ChelseaMFineArt, ObamaWon



(a) Topic Recall



(b) Term Precision



(c) Term Recall

Fig. 2. Metrics across all timeslots.

TABLE IV
TOTAL METRIC RESULTS

	Topic Recall	Term Precision	Term Recall
LDA	0.312	0.31	0.539
Geo-Context	0.562	0.468	0.675

IV. EVALUATION

We evaluated both GeoContext and LDA as methods for discovering topics within a social media stream. We utilized two datasets for evaluation. The first dataset from [10] contains tweets from the 2012 United States presidential elections. The second dataset from [13] contains tweets that are categorized into various rumors and truths. These datasets were chosen as a representation of topics likely to be of interest to users.

For both datasets, we compared results from GeoContext and LDA against ground truth topics that are included with both datasets. The ground truth topics for the first dataset are keywords and headlines that were extracted from mainstream media reports about the elections. The ground truth topics for the second dataset are tweets clustered into the rumor and truth topics.

A. Elections Dataset

The Elections dataset consists of tweets from the November 11, 2012, U.S. presidential election. The entire set of tweets is partitioned into timeslots. Each ground truth topic extracted based on media reports is assigned to one time slot. A time slot can have more than one ground truth topic. 64 ground truth topics are present in the dataset.

The dataset consists of 524,886 tweets. The tweets are broken into 26 individual timeslots, where each timeslot is ten minutes long. Example topics are shown in Table II.

Topics include the re-election of Barack Obama and his running mate, Joe Biden, over nominee Mitt Romney. Later timeslots contain topics indicating Obama's victory speech. The dataset also include elections to the United States Senate and House of Representatives, as well as some state governors.

Consistent with [10], we first calculated topics using LDA and GeoContext for each timeslot of the dataset. The number of topics calculated by LDA was 10 for each timeslot. Topics discovered using both LDA and GeoContext for some sample timeslots are shown in Table III.

We then calculated three metrics for the evaluation of the discovered topics: topic recall, term precision, and term recall. Definitions of these metrics are:

1) *topic recall*: topic recall is the total number of topics detected out of the ground truth topics. A topic is considered to be detected if all terms in the ground truth topic are present in the detected set of keywords.
2) *term precision*: for a detected topic and some matching ground truth topic, term precision is the number of correctly detected terms in a topic out of the total number of terms in the detected topic.
3) *term recall*: for a detected topic and some matching ground truth topic, term recall is the number of correctly detected terms in a topic out of the total number of terms in the ground truth topic.

The total topic recall, term precision, and term recall is computed by taking the microaverage of the individual topic recall, term precision, and term recall for each timeslot. The total values for each of the three metrics are shown in Table IV. We also show the topic recall, term precision, and term recall across all timeslots for both LDA and GeoContext in Figure 2.

A limitation exists with this dataset in the evaluation that results from the method of calculating ground truth topics, as described in [10]. Because the ground truth topics were not extracted directly from the dataset tweets, but rather from news stories that described the timeslots, it is not guaranteed that tweets exist with the terms in the ground truth topics. For example, a ground truth topic in timeslot 0 is "bernie, sanders, senator, senate, vermont, vt, wins, call, projects, called, held, won, calling, project, projection, win, projecting, hold." However, there exists only one tweet from timeslot 0 in the dataset that contains the term "bernie" and none that contain the term "sanders." Furthermore, the tweet that contains the term "bernie" does not contain the term "senator", so it is highly improbable that these terms would occur in the same topic produced by any topic discovery method, making this ground truth topic impossible to reproduce.

B. Rumors and Truths Dataset

The Rumors and Truths dataset consists of sets of tweets broken up into various topics. Some topics are popular rumors that circulated throughout Twitter, while others are topics that are true. For this dataset, we decided to evaluate how well the topics discovered by LDA and GeoContext matched the dataset topics, which are human-labeled.

The dataset consists of 41,952 individual tweets. The tweets are categorized into 152 distinct topics. Example truth topics are "Mold inside Capri Sun drinks," "Storm hitting Bay Area," and "iPod classic discontinued." Example rumor topics are "NASA warns of six day blackout," "Actor Macaulay Culkin found dead," and "Malia Obama is pregnant."

Table V shows five example topics from the dataset, LDA, and GeoContext. For each topic, the leftmost column shows a description of the topic and an example tweet from the ground truth dataset. The table also indicates whether the topic is a rumor or a truth. The middle column contains the corresponding topics found by LDA, and the rightmost column contains the corresponding topics found by GeoContext from the topic.

For this dataset, LDA produced topics in several instances that were mixtures of more than one ground truth topic. For example, the topmost topic in Table V contains terms about both the actor David Ryall's death as well as a rumor about the restaurant chain McDonald's stopping service for overweight customers. Also, LDA produced more than one

TABLE V
RUMORS AND TRUTHS EVALUATION

Dataset Topics	LDA Topics	GeoContext Topics
Truth: David Ryall died "@eonline: RIP David Ryall. The Harry Potter actor has died at age 79."	ryall; david; died; actor; harry; potter; stop; customers; overweight; serving	david ryall, elphias doge, #harrypotter star, peace, harry potter, outnumbered, David Ryall, excellent actor, good films, tv progs, #harrypotter, harry potter, actor, age, t.co
Truth: North Korea Sony attack "North Korea AINT PLAYING! RT @necolebitchie Sony Hackers Threaten 9/11-Type Attack On Theaters (cont)"	**1:**north; korea; sony; internet; outage; hack; attack; service; restored; report **2:**north; korea; time; capsule; sony; internet; outage; paul; attack; boston	north korea, pr win, Sony investigators, attack probe, North Korea, links, source, Sony attack, supporters, attack probe-source, suspect, denial, sony pictures, house intel, hacks, sony, information, internet outage, Internet services, outage, restoration, night, nkorea outage, online uncertainties, microscopic corner, case study, internet, Internet outage, dispute, U.S., experts, i4u news, NKorea outage, Internet ha, AP, LONDON, wifi password, South Korea, Dec, access, torontostar, tit, tat, fingers, Sony movie, Internet service, tensions, hack, internet outages, attack, t.co, t.co qijAQlHmA0, #dynresearch #lesleywroughton, dispu, apparent attack, web outage, Web outage, t.co klrSfFKqFX, Internet
Truth: West Virginia train derailed "Fireball erupts into sky as derailment sends tanker into river: A CSW train derailed, pouring crude oil into a..."	**1:**train; west; derailed; oil; virginia; crude; freight; carrying; fire; derailment **2:** train; oil; derailed; virginia; west; crude; carrying; news; fire; freight	Train Derailment, freight train, explosion th, Oil Spill, crude oil, West Virginia, Monday
Rumor: Bobby Shmurda stabbed "Bobby Shmurda stabbed to death in prison ladies and gents. Guess you could say he was alive about a week ago #toosoon"	**1:**shmurda; stabbed; death; bobby; jail; fake; cell; mate; news; rikers **2:**bobby; shmurda; stabbed; death; jail; killed; rice; tamir; cell; mate	jail, bobby shmurda, death l0l, death, prison, cell mate, jail tho, yea, man, jail, way
Rumor: Obama lowers drinking age to 18 "Effective 6/4/2015 President Obama Signs Amendment To Lower The Legal Drinking Age To 18"	obama; age; drinking; lower; legal; lowering; june; law; lowered; signed	legal drinking age, obama, obama signs amendment, president, obama bout

TABLE VI
TWEET PRECISION AND TWEET RECALL

	Tweet Precision	Tweet Recall
Keywords Only	0.927	0.209
Keywords and Concepts	0.814	0.170

topic for several of the ground truth topics. These instances are indicated by numbering in the table.

Because the Rumors and Truths dataset does not include ground truth topics, but rather ground truth tweet clusters, we did not compute the topic recall, term precision, and term recall metrics. Instead, we calculated the tweet precision and tweet recall produced by GeoContext for this dataset. Because the Rumors and Truths dataset consists of clusters of tweets as ground truth, it is well-suited for these metrics. GeoContext can be used for clustering tweets in addition to discovering topics, so it can be useful to determine how well the clusters are formed.

We define the tweet precision and tweet recall as follows:

1) *Tweet precision:* the percentage of correctly clustered tweets out of all tweets in a cluster. We calculated the total tweet precision as the average of the tweet precision for each cluster of tweets.
2) *Tweet recall:* the percentage of correctly clustered tweets out of the total number of tweets in the ground truth cluster. As with tweet precision, we calculated the total tweet recall as the average of the tweet recall for each cluster of tweets.

We noticed during manual evaluations that the concepts extracted by AlchemyAPI's Concept Tagging API was not always accurate in describing the topic of the tweet. Because of this observation, we decided to calculate the metrics both with and without GeoContext's concepts. The total tweet precision and tweet recall over all tweet clusters are shown in Table VI.

As displayed in the table, both metrics are higher for GeoContext using keywords only. Also, the tweet precision is high, indicating that tweets are correctly clustered together. However, the tweet recall is somewhat low, indicating that the ground truth clusters are split apart into multiple clusters by GeoContext. In future work, we plan to investigate GeoContext's algorithm to determine the cause of this splitting.

C. Discussion of Results

The results from this evaluation process clearly show the benefits of using keyword relevance over traditional topic modeling approaches for topic discovery within social media. As shown in Table IV, GeoContext was able to identify more ground truth topics than LDA. The term precision and topic recall metrics were also higher for GeoContext than LDA, showing that GeoContext was able to create more topics that have more related terms than LDA and more topics that contain the terms from the ground truth topics. The high values for these metrics indicate that GeoContext is able to better discover individual events in a social media stream that do not contain mixed topics.

GeoContext contains a drawback in that it can create a dynamic number of topics, which can affect processing time and readability for users if the number of topics is too large. To address this issue, GeoContext also includes a pruning module, which prunes topics that have not had any new tweets added in a certain amount of time. However, because we wanted to evaluate GeoContext's results directly against LDA's results, which does not consider time, we elected not to use this module.

V. Future Work and Conclusion

In this paper, we compared two approaches for topical discovery within a social media stream. We evaluated the performance of GeoContext, which calculates a similarity score between tweets and uses keyword extraction to cluster tweets and determine topics, and LDA, which is a commonly used topic modeling algorithm, on two different datasets. For the first dataset, we compared the topics produced by both methods against ground truth topics extracted from the dataset. For the second dataset, we produced topics using both GeoContext and LDA and matched those topics against ground truth clusters of tweets in the dataset. We also evaluated the performance of GeoContext in clustering tweets against the ground truth clusters.

For future work, we plan to improve the algorithm of GeoContext by improving the method used to calculate similarity scores between tweets. In addition, we plan to produce an extension of LDA in order to further increase the performance as evaluated by the defined metrics in this paper. We also plan to evaluate GeoContext and LDA on more datasets.

Topic discovery in a social media stream can be an invaluable tool for identifying major events around the world. Using social media to gather information can allow us to utilize the opinions and data of millions of people, rather than only a few traditional media outlets.

Acknowledgment

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. [Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution (DM-0003899).

References

[1] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 80–88. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964870

[2] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD '13. New York, NY, USA: ACM, 2013, pp. 8:1–8:9. [Online]. Available: http://doi.acm.org/10.1145/2501025.2501027

[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, ser. WebKDD/SNA-KDD '07. New York, NY, USA: ACM, 2007, pp. 56–65. [Online]. Available: http://doi.acm.org/10.1145/1348549.1348556

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 919–931, April 2013.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[6] J. Vosecky, D. Jiang, K. W.-T. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in twitter," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 879–884. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505593

[7] Y. Hong, Y. Fei, and J. Yang, "Exploiting topic tracking in real-time tweet streams," in *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, ser. UnstructureNLP '13. New York, NY, USA: ACM, 2013, pp. 31–38. [Online]. Available: http://doi.acm.org/10.1145/2513549.2513555

[8] H.-G. Kim, S. Lee, and S. Kyeong, "Discovering hot topics using twitter streaming data: Social topic detection and geographic clustering," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 1215–1220. [Online]. Available: http://doi.acm.org/10.1145/2492517.2500286

[9] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, ser. ICWSM '10, 2010, pp. 130–137. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1528

[10] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Gker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268–1282, Oct 2013.

[11] Q. Yuan, G. Cong, K. Zhao, Z. Ma, and A. Sun, "Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users," *ACM Trans. Inf. Syst.*, vol. 33, no. 1, pp. 2:1–2:33, Feb. 2015. [Online]. Available: http://doi.acm.org/10.1145/2699667

[12] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 1794–1798. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398519

[13] J. Zou, F. Fekri, and S. W. McLaughlin, "Mining streaming tweets for real-time event credibility prediction in twitter," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 1586–1589.