

Unsupervised Content-Based Characterization and Anomaly Detection of Online Community Dynamics

Danelle Shah, Michael Hurley, Jessamyn Liu, Matthew Daggett
MIT Lincoln Laboratory, Lexington, MA
{danelle.shah, hurley, jessamyn.liu, daggett}@ll.mit.edu

Abstract

The structure and behavior of human networks have been investigated and quantitatively modeled by modern social scientists for decades, however the scope of these efforts is often constrained by the labor-intensive curation processes that are required to collect, organize, and analyze network data. The surge in online social media in recent years provides a new source of dynamic, semi-structured data of digital human networks, many of which embody attributes of real-world networks. In this paper we leverage the Reddit social media platform to study social communities whose dynamics indicate they may have experienced a disturbance event. We describe an unsupervised approach to analyzing natural language content for quantifying community similarity, monitoring temporal changes, and detecting anomalies indicative of disturbance events. We demonstrate how this method is able to detect anomalies in a spectrum of Reddit communities and discuss its applicability to unsupervised event detection for a broader class of social media use cases.

1. Introduction

This research is part of a larger effort to explore how human networks respond to disturbances that impact the community's membership, function, or interests. In particular, we are interested in observing network dynamics which may signal unrest, vulnerability, or redefinition in response to a disruptive event. While large-scale events such as natural disasters or political upheaval may be trivial to identify, the detection of localized events impacting specific sub-communities is more challenging. We also seek to study when and why *similar networks behave differently or dissimilar*

networks behave similarly in response to the same event. To these ends, it is necessary to: 1) characterize a community's membership, function, or interests over time; 2) quantify intra- and inter-community similarity; 3) automatically detect events that cause a community disturbance; and 4) model the short- and long-term network effects of an observed disturbance.

A *community* is defined as “a social, religious, occupational, or other group sharing common characteristics or interests and perceived or perceiving itself as distinct in some respect from the larger society within which it exists” [35]. The active association or interaction of individuals within a community to achieve some end (e.g., exchange information, execute a task) is referred to as a *network*.¹ Communities are often hierarchical and overlapping, and an individual may belong to an arbitrary number of communities.

Social media provides a rich, observable data source of dynamic social networks, and it has been extensively studied to model both online and offline human social behavior. Data generated by online social networks have been used to detect significant events in real-time [7, 31], track the spread and sentiment of information [4, 29, 32], monitor population health and safety [13, 33, 34], etc. Reddit (www.reddit.com) is one such social media platform, generating large volumes of content shared by self-organizing, topically-oriented communities. In contrast to the significant body of work dedicated to the definition and detection of communities from individuals and connections [12], the structure of Reddit allows communities to be treated as atomic units without the need to infer membership.

In this paper we describe an unsupervised approach to modeling the content shared within Reddit communities in order to measure inter-community similarity, track volatility over time, and automatically detect anomalous activity corresponding to real-world events. In Section 2, we review related research in

This material is based upon work supported by the Department of Defense under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

¹This distinction is subtle. In the context of this work we assume a 1:1 mapping between the (observable) network and the corresponding (sub)community it represents, and use these terms interchangeably.

the area of event detection in social media. In Section 3, we describe the Reddit dataset used in this work and introduce our model for characterizing subreddit dynamics using natural language processing techniques. Specifically, subreddit post activity is represented in a low-dimensional embedding space, enabling efficient calculation of intra- and inter-community content similarity. We describe an unsupervised anomaly detection method in Section 4 and present two case studies where detected community disturbances correlated with real-world events. We conclude with a discussion of opportunities and future work.

2. Background and Related Work

2.1. Reddit

Reddit is a prominent social media platform² launched in 2005 to capture what's new and popular on the Internet via sharing of hyperlinks to interesting websites. As the user base grew, a community structure organically developed through the use of interest-centric groups called *subreddits*. Users post stories, links, and media relevant to a subreddit's predefined topic, and engage in community-oriented discussions through comments. Due to the encoded structure of user interactions via posts and threads, the heterogeneity of content (text, images, videos, hyperlinks), open and authentic discussion afforded by relative anonymity and minimal censorship, and the observability enabled by its Application Programming Interface (API), Reddit provides a unique opportunity to analyze complex inter- and intra-community dynamics.

While we focus our analysis on Reddit in this paper, the approach presented here can be trivially applied to any online social network in which communities are centralized and explicitly defined. The extension of these techniques to platforms in which users interact in a distributed fashion (e.g., Twitter) or to data of physical (offline) networks remains a future research goal.

2.2. Event Detection in Social Media

Methods for detecting events in online social media has been an active research area since the popular adoption of the internet in the 1990s. A number of excellent surveys have been recently published which provide an overview of approaches used to detect anomalies and behaviors in online social media (e.g. [18, 39, 40]). Using the nomenclature of [18], the technique described in this paper uses an *unsupervised* approach to *dynamic, contextual* anomaly detection.

²As of this writing, Reddit is ranked 5 most visited website in the United States and 16 in the world [2].

Twitter has been leveraged extensively for detecting events in social media using a variety of methods including analysis of social network structures as well as message volume, timing, and content. Applications include detection of events in sports [7, 28, 42], public emergencies and disease outbreaks [3, 34] and natural disasters [8, 33]. Several recent works perform general event detection and trend spotting on Twitter by detecting anomalous or “bursty” phrases [6, 23, 36]. However, Twitter presents many challenges due to the short, unstructured, usually poorly formatted messages that users post, and lacks explicit community structure corresponding to user membership or content relevancy.

Although Reddit has been used as a data source for online social media analysis, relatively little research has been conducted on event detection in Reddit. Instead, the research has more often focused on community analysis [15, 16], topic discovery [13, 37] or on summarizing posts and comments [22]. Most analyses have focused on event summarization as opposed to event detection. Event *summarization* is concerned with collecting and compressing streams of text into a concise summary statement. Event *detection* aims to identify significant changes in the text streams [7, 24].

Hamilton, et al. [14] studied the extent that sentiment varies over time within the content of specific communities on Reddit. They combined domain-specific word embeddings and a label propagation framework to induce sentiment lexicons using a small numbers of seed words from Reddit data with a corpus of 150 years of English words. Lyngbaek's thesis [22] describes an application called SPORK that identifies and summarizes topics in Reddit comment threads. It uses natural language processing based upon term frequency-inverse document frequency (TF-IDF), cosine similarity, and clustering algorithms. Lyngbaek uses a hierarchical clustering algorithm to combine similar comments based upon word-frequency vectors. In this paper we use a similar technique, however rather than clustering individual comments into summarization topics, we model entire subreddit communities over finite time spans in order to detect significant changes in response to real-world events.

3. Modeling Community Content

We seek to model user behavior on Reddit as a means to characterize communities over time and to detect anomalies in community activity caused by consequential events. In this section we describe a content-based activity model, taking advantage of recent advances in natural language processing to characterize the text in subreddit posts. The motivations for this

model are threefold: quantitatively measure subreddit similarity; monitor temporal changes in subreddit activity; and automatically detect anomalies caused by real world events. Note that many of the results in this paper highlight examples that we expect readers may recognize, however the objective is to automatically detect events affecting any community of interest.

3.1. Dataset

The data used in this paper includes over 118 million posts created by over 8 million unique users on Reddit between March 2016 and September 2017, collected from the archive of Reddit made publicly available on Google BigQuery,³ omitting known bots and default subreddits. Bots are computer programs which automatically generate posts and comments; 988 bots (190 active) were identified by manually inspecting abnormally-active users (> 3000 comments per month) and by scraping `/r/botWatcher` for recently-mentioned bots. Default subreddits (those to which all users are automatically subscribed upon account creation) are linked at the top of every Reddit page and are typically among the most active subreddits; 49 default subreddits (as of March 2016) were omitted from this analysis.

3.2. Subreddit Activity Model

The *content* of a subreddit captures the interests, thoughts, questions, complaints, discussions, arguments, and general musings of the members of that community. This content consists of user-contributed posts (and subsequent comments) including articles, videos, and links to other websites, as well as freeform text. By design, posts within each subreddit are typically organized around a predefined theme (sports, politics, video games, etc.), making the characterization of this post content a critical component of understanding the community and detecting change. While posts may contain heterogeneous media, each post is accompanied by a short descriptive title, the text of which is used exclusively to characterize subreddit content in this work.⁴ Specifically, a subreddit's content over a finite time span is represented as a weighted average word embedding of the corresponding post titles.

Word embeddings are a class of natural language processing techniques in which individual words are represented by a vector in a predefined vector space, the dimension of which is significantly smaller than

the size of the vocabulary. The term *word embeddings* was introduced by [5], but it was the release of word2vec [25], a toolkit for training and implementing word embeddings, that greatly expanded the use of the technique. Here, an embedding known as GloVe (Global Vectors for Word Representation) was used. The GloVe algorithm was created by [30] and builds on word2vec to explicitly encode meaning as vector offsets. For example, models created using GloVe should accurately represent semantic relationships such as the analogy “king is to queen as man is to woman,” which would be encoded in the vector space by the vector equation: $king - man + woman = queen$.

A GloVe word embedding was trained on the corpus of over 118M Reddit posts, which included over 1.2M unique words. (Words that appeared less than 5 times in the corpus were omitted.) Note that while one may use pre-trained word embeddings, training on in-domain data is desirable to capture the vocabulary, jargon, and symbols used on Reddit, as well as foreign language use.

The technique of representing a bag of words (such as a sentence or document) as an average of constituent word embeddings has been demonstrated as a successful and efficient approach for many applications, such as detecting semantic similarity of sentences, identifying plagiarism, classifying sentiment from social media, and others [10, 17, 27, 41]. In this work, a weighted average word embedding (i.e., a weekly⁵ subreddit content vector) is calculated by averaging the GloVe embedding vectors for all the words in a subreddit-week, weighted by a term frequency-inverse document frequency (TF-IDF) value:

$$w_{TF-IDF} = (1 + \log f_{t,d}) \cdot (1 + \log (N/n_t)) \quad (1)$$

The term frequency $f_{t,d}$ is given by the number of times a term t occurs in document d , and the inverse document frequency is given by the ratio of the total number of documents N and n_t , the number of documents containing the term t . In this work, a document is the collection of posts from a given subreddit for a given week. This TF-IDF formulation assigns more weight to words that are unique and less to words that are used frequently across many subreddits.

Anecdotally, the weighted average word embedding for subreddit content captures relationships between subreddits in a similar way as the raw GloVe embedding captures relationships between words, e.g.:

$$\begin{aligned} /r/Patriots - /r/boston + /r/Atlanta &= /r/falcons \\ /r/iphone - /r/apple + /r/google &= /r/Android \end{aligned}$$

⁵For the results presented in this paper, a time span of one week was used to smooth out day-to-day volatility, particularly on smaller subreddits with low post activity. Furthermore, many subreddits exhibit a natural weekly periodicity of posting behavior, for example due to higher volumes of activity on weekends and repeating occurrences such as television shows and sporting events.

³<https://bigquery.cloud.google.com/dataset/fh-bigquery>

⁴Although comments typically significantly outnumber posts on any given subreddit, preliminary analysis showed that post titles alone provided a more robust representation, likely because posts are richer in content and less noisy than comments.

3.3. Community Similarity

Although shortcomings of word vectors and the use of word analogies for evaluating vector-space semantic models have been documented [11, 21], this relatively simple and inexpensive calculation performed well in grouping subreddits into topically-similar clusters. To quantify community similarity, the cosine distance between the weighted average word embeddings v for subreddits j and k is used:

$$d_{j,k} = 1 - \cos(\theta) = 1 - \frac{v_j \cdot v_k}{\|v_j\|_2 \|v_k\|_2} \quad (2)$$

The cosine distance of average embeddings is a simple yet effective metric for calculating semantic similarity between documents [27, 41].

Figure 1 illustrates the hierarchical clustering of the content embeddings for the 1000 most active subreddits on the week of September 9-15, 2017. (This is provided for visual inspection only; no analysis on community clusters is subsequently performed.) Here, one can see large clusters of similar subreddits, as well as the presence of individual subreddits that do not cluster well. Most of these outliers are foreign-language subreddits (e.g. **/r/france** in French, **/r/newsokuvip** in Japanese) or have unique vocabulary or posting conventions (e.g. in **/r/ShadowBan** nearly all post titles take the form of “Am I Shadowbanned?”). Figure 2 zooms in on a few of these clusters. Figure 2a shows a large cluster of hobby or personal interest subreddits, with subclusters related to food, fashion, nature, pets, and art. A cluster of sports-focused subreddits is shown in Figure 2b, where baseball, football, boxing/wrestling, and “misc” sports subreddits form distinct subclusters.

While these clusters appear reasonable upon qualitative examination, we can quantify the efficacy of the subreddit embedding similarity metric by using it to predict whether one subreddit explicitly links to another in its community description. For example, at the time of this writing, the subreddit **/r/boston** links to three other Boston-related subreddits: **/r/bostontenants**, **/r/bostonhousing**, and **/r/BostonJobs**.

Table 1 lists the top ten false positives, or the ten subreddit pairs with smallest cosine distance (greatest cosine similarity) that *do not* link to one another. Upon inspection, all ten pairs are objectively similar subreddits with respect to content. Table 2 lists the top ten false negatives, or the ten subreddit pairs with largest cosine distance that *do* link to one another. In this case, nearly all pairs consist of two subreddits in different languages. The pair **/r/sweden** and **/r/cringe** strangely appears in this list, but in fact the subreddit **/r/sweden** contains a link to **/r/Pinsamt** which is described as

Top 10 False Positives	
Subreddit 1	Subreddit 2
askmen	askwomen
answers	nostupidquestions
gametrade	steamgameswap
crazyideas	nostupidquestions
conservative	politic
crappydesign	mildlyinfuriating
grandtheftautov	gtav
india	indianews
answers	crazyideas
offmychest	self

Table 1: Non-linked subreddits with highest similarity.

Top 10 False Negatives	
Source	Target
europe	greece
suomi	ruusia
newsokuvip	newsokur
suomi	denmark
sweden	cringe
europe	sweden
europe	denmark
europe	romania
suomi	sweden
europe	portugal

Table 2: Linked subreddits with lowest similarity.

“Like **/r/cringe**, only a lot more Swedish.” Figure 3 shows the receiver operating characteristic (ROC) curve for all subreddit pairs over 52 weeks. The average area under the ROC curve (AUC) is 0.85.

4. Anomaly Detection

In this section, we leverage the previously-described content model to automatically detect anomalies caused by community disturbances. We consider a disturbance to be an event that has a real or perceived impact on a community’s membership, function, or interests. The motivation for anomaly detection on Reddit is twofold: to leverage the Internet as a sensor for timely detection of real-world events; and to generate a rich dataset of observed community perturbations to facilitate robust analysis of disturbance effects.

4.1. Community Change and Event Detection

Unexpected, controversial, or disruptive events often result in a flurry of discussion about a new topic on relevant affected communities on Reddit. These events

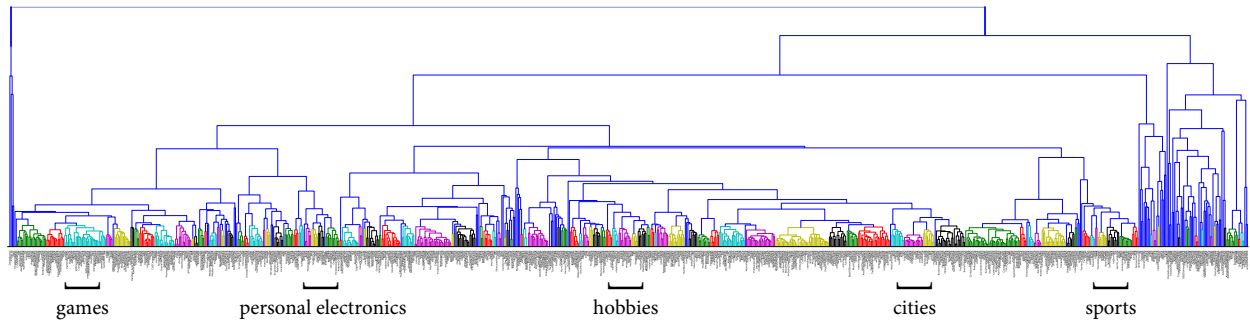


Figure 1: Hierarchical clustering of subreddit embeddings using cosine distance.

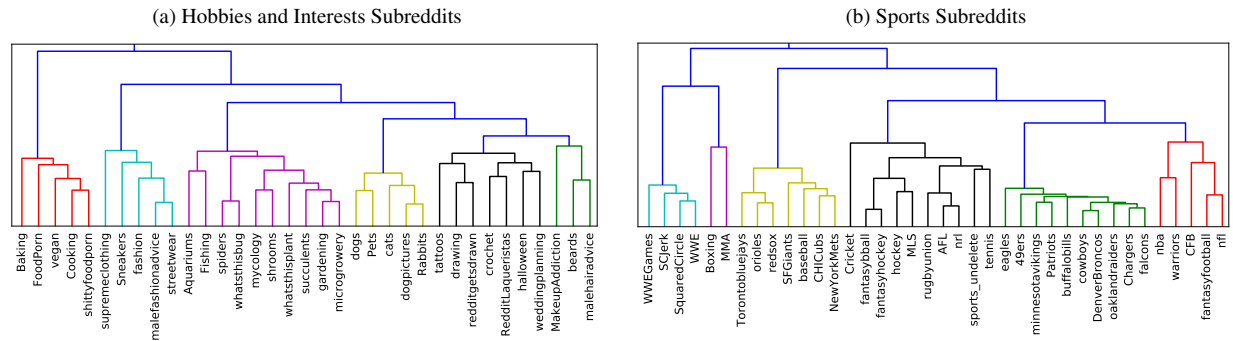


Figure 2: Closeups of two clusters in the subreddit embedding dendrogram from Fig. 1

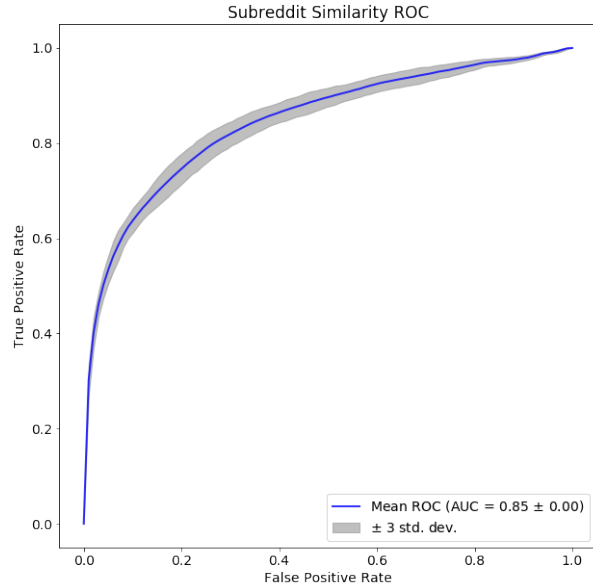


Figure 3: ROC for cosine similarity predicting explicit links between subreddits, averaged over 52 weeks.

may therefore be detectable by anomalous content, anomalous activity volume, or both.

Just as the cosine distance of weighted average word embeddings was used to measure similarity between subreddits (Eqn. 2), this metric can also quantify content change from time t_{k-1} to time t_k for a single subreddit. The amount of change considered “anomalous” for any particular subreddit will depend on its expected (baseline) activity. The content model takes a weighted average over all words in the post titles; popular, broad-topic subreddits will tend to have more stable content week-to-week than smaller, niche subreddits. To model the typical change for a subreddit, the observed week-to-week cosine distances can be approximated by a beta distribution⁶ with support from 0 (no change, i.e., $v_k = v_{k-1}$) to 1 (when $v_k = -v_{k-1}$). By fitting the expected content change distribution, one can estimate the probability of observing a week-to-week change greater than or equal to d_k by evaluating $P(d \geq d_k) = 1 - F(d_k)$, where F is the distribution’s cumulative distribution function (CDF). In the examples presented, a threshold of 0.02 is used to detect anomalies; this threshold is a tuning parameter and corresponds to content changes that are $< 2\%$ likely to occur in the data.

⁶This model was used due to preliminary analysis indicating that week-to-week cosine distances for most subreddits could be well-approximated by a beta distribution.

Calculating a subreddit content vector for any arbitrary time span is computationally inexpensive (the specific complexity depends on the scope of the IDF term in Eqn. 1), therefore content change can be monitored at multiple levels of granularity (day-to-day, week-to-week, etc.). Similarly, the expected change distribution and anomaly calculation is trivially inexpensive to calculate and may be done in an online or rolling fashion. However in practice, one needs enough observations (in terms of total number of time intervals) to sufficiently estimate the content change distribution. In the examples given in this paper the distribution is assumed to be static over all time; in some cases this may not accurately reflect permanent or seasonal changes to subreddit content.

4.2. Anomaly Detection Exemplars

In this section, content models are used to detect anomalies within two distinct community groups. The first case study examines several football-related subreddits to illustrate the efficacy of the presented techniques on an example with which the reader may be familiar. The second looks at subreddits dedicated to various dark net markets (primarily illicit markets which operate via anonymous communication networks such as Tor or I2P) and the anomalies detected on Reddit corresponding to significant disruptions to these anonymous, encrypted websites.

4.2.1. Football. Professional football is considered the most popular American sport, so it is not surprising that a significant amount of football-related content can be found on Reddit. Discussion about the National Football League (NFL) can be found on **/r/nfl**; those interested in NCAA college football can head to **/r/CFB**. Subreddits dedicated to individual teams abound, e.g. **/r/LSUFootball**, **/r/ClemsonTigers** and **/r/Patriots**.

Figure 4 plots the week-to-week content change and post volume for three football-related communities. Detected anomalies reliably correspond to events such as the Super Bowl, draft dates, and dramatic or surprising incidents (anomalies were inspected manually and annotations provided for the reader's convenience). We note several observations:

- Events do not affect all communities, even closely-related ones, in the same way. Although all three subreddits are close in the content embedding space (see Fig. 2b), they exhibit anomalous behavior in response to different events specific to their domain. For example, National Signing Day (Feb. 1, 2017) refers to the day new recruits can commit to their future schools and is coincident with the

largest content-based anomaly on **/r/CFB**, likely caused by users discussing the incoming class of new promising football stars. This event, however, causes no observable change in either **/r/nfl** or **/r/Patriots**.

- Events detected due to anomalous content do not always correlate with spikes in activity volume, and vice-versa. Content-based anomalies often correspond to a surprising or dramatic occurrence, causing the existing community to discuss a new topic. Changes in activity volume correspond to events which would affect the community's popularity (either increase or decrease), such as the start of a new season. Interestingly, Super Bowl LI (Feb. 5, 2017) caused a dramatic increase in post activity on **/r/Patriots** but not a significant change in content, likely because the Super Bowl temporarily attracted more people to the subreddit, but the topics of discussion remained largely the same.
- While the framework described in this paper cannot explain the exact *cause* of an anomaly, word clouds can often provide quick insight into the underlying event(s). For example, word clouds corresponding to the three content anomalies observed on **/r/Patriots** are plotted and described in Figure 5.
- Even for subreddits that correspond to a public and well-defined community such as **/r/Patriots**, the events detected are not always present in other forms of traditional media. For example, 342 news articles containing the word “patriots” were published in the sports section of the New York Times between March 2016-September 2017. Most articles were not relevant to the Patriots football team, and only a small number (25 articles over nearly 19 months) correspond to articles covering non-game related news. These non-game articles represent the type of news stories which would be most likely to shift the content of **/r/Patriots** away from typical, game-related discussions, but only 1 out of 25 (“F.B.I. Recovers Tom Brady’s Missing Super Bowl Jerseys in Mexico”) aligns with the events detected by our method. Other representative non-game news articles include headlines such as “Aaron Hernandez’s Murder Conviction Is Nullified”, “Six New England Patriots Say They Will Skip a White House Visit”, and “Babe Parilli Dies at 87”. The events detected on Reddit and articles published by the New York Times may point to differences between topics deemed relevant by self-identified community members and those recognized by an external editorial staff, illustrating one benefit of using social media as a sensor for detecting community disturbances in addition to more traditional sources.

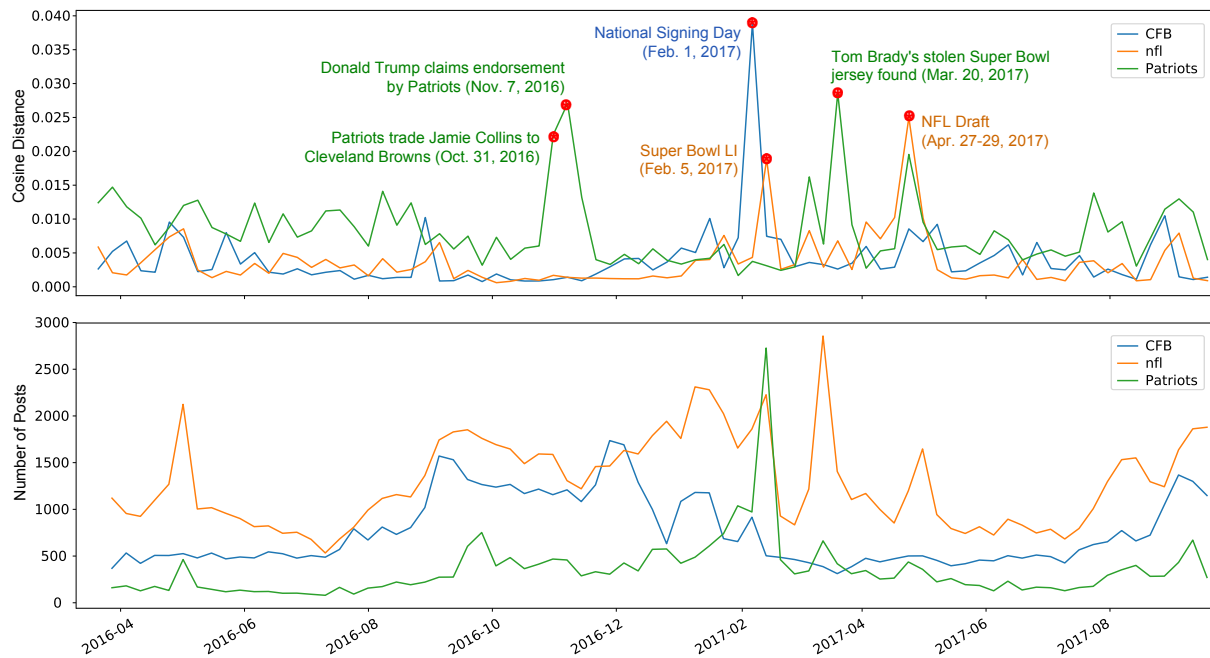
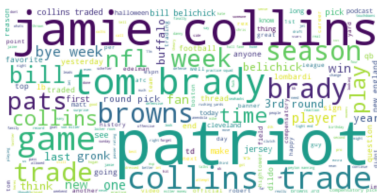


Figure 4: (top) Week-to-week change for three football-related subreddits (more positive indicates greater change). Detected anomalies are indicated by red dots and annotated with relevant events that coincide with the anomalies; (bottom) Subreddit post volume is plotted for comparison.

(a) Week of Oct. 31–Nov. 6, 2016; On Oct. 31, 2016, the Patriots unexpectedly traded top linebacker Jamie Collins to the Cleveland Browns [20].



(b) Week of Nov. 7–13, 2016; On Nov. 7, 2016, Donald Trump gave a speech claiming Patriots coach Bill Belichick and quarterback Tom Brady supported him for president [9].



(c) Week of March 20–26, 2017; On March 20, 2016, Tom Brady's stolen Super Bowl jersey was reported to have been found [1].

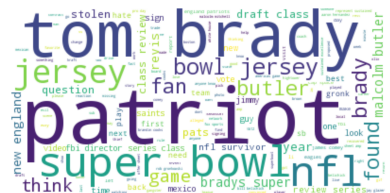


Figure 5: Word clouds corresponding to detected events on */r/Patriots*. Word sizes are proportional to frequency [26].

4.2.2. Dark Net Markets. In July 2017, AlphaBay, an online dark net marketplace, was taken down by law enforcement authorities. With over 240,000 members, AlphaBay was the largest market of its kind and reportedly hosted \$600,000 to \$800,000 in transactions daily. Its takedown, along with the coordinated seizure of Hansa (the second largest dark net market) two weeks later, scattered vendors and buyers to other dark net markets and caused significant disruption to dark net market communities both on and off Reddit.

Prior to its shutdown, AlphaBay users and administrators used the subreddit */r/AlphaBayMarket* to post reviews and updates on site operations. After

its final outage on July 4, 2017, Reddit users speculated that the outage was an exit scam⁷ and exchanged conspiracy theories on */r/AlphaBayMarket* and on */r/DarkNetMarkets*. Federal authorities did not confirm their involvement in the takedown until July 20, and new users flooded to */r/AlphaBayMarket* in the intervening weeks as vendors and buyers tried to determine status of the network.

Vendors and buyers also sought an alternative dark net market on which to continue their transactions. New listings tripled on Hansa and DreamMarket, another

⁷In 2015, administrators of Evolution, another major dark net market, shut down the site and took millions of dollars of users' money in what was known as an exit scam.

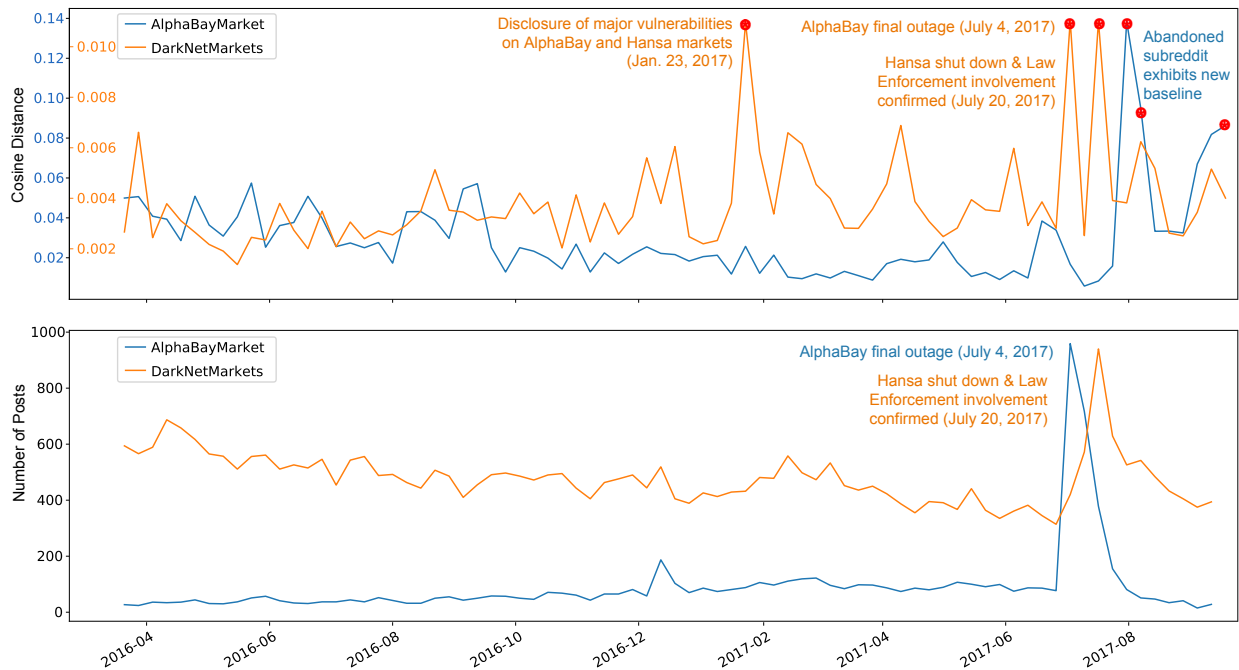


Figure 6: (top) Week-to-week change for two largest Dark Web market subreddits (more positive indicates greater change); (bottom) Subreddit post volume is plotted for comparison.

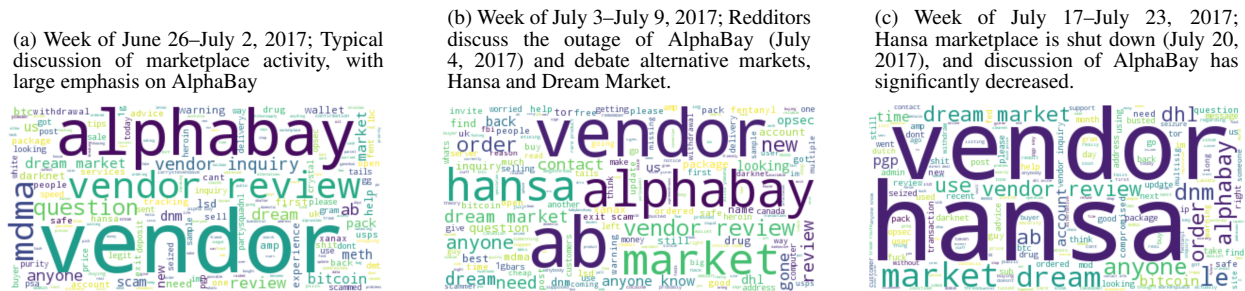


Figure 7: Word Clouds for subreddit /r/DarkNetMarkets before, during, and after the AlphaBay and Hansa dark net markets were shut down in July 2017.

dark net market competitor, when AlphaBay went down [38]. Once it became clear that Hansa was also shut down, DreamMarket became the largest dark net market, although others, such as Trade Route, Tochka, and Wall Street Market, saw similar boosts in traffic [19].

Figure 6 (top) shows anomalous change in /r/DarkNetMarkets content in the weeks of July 3 (when AlphaBay had its final outage) and July 17 (when the federal takedown of AlphaBay and Hansa was formally announced). Word clouds give a sense of how the content of posts on /r/DarkNetMarkets shifted from from AlphaBay to Hansa to DreamMarket as news of the takedown unfolds (see Figure 7). This anomalous content change detected in /r/AlphaBayMarket may

be attributed to a significant reduction in posts after August 2017 (/r/AlphaBayMarket had an average of 92 posts per week from Jan.-June 2017 and an average of 35 posts per week from Aug.-Sep. 2017).

Figure 6 (bottom) shows a large spike in post activity on /r/AlphaBayMarket on the week of July 3, reflecting the large increase in posts on /r/AlphaBayMarket in the confusion following the site's outage (total posts on /r/AlphaBayMarket jumped from less than 80 posts on the week of June 26 to over 950 on the week of July 3). A similar spike on /r/DarkNetMarkets for the week of July 17 likely reflects the increase in posts on /r/DarkNetMarkets regarding the announced government takedown of both

AlphaBay and Hansa; we speculate that posts moved from */r/AlphaBayMarket* to */r/DarkNetMarkets* as it became clear that AlphaBay was completely defunct (total posts on */r/DarkNetMarkets* jumped from 570 posts on the week of July 10 to over 940 on the week of July 17, whereas total posts for the same weeks on */r/AlphaBayMarket* decreased from ~720 to ~380).

The results presented here represent two specific community groups, however these analyses demonstrate the ability to quantitatively characterize subreddit communities and automatically detect disturbances in a wide range of communities in an unsupervised manner. Moreover, these disturbances appear to correspond to events that are unexpected and noteworthy to the members of the community itself, rather than those highlighted by traditional external news sources.

5. Future Work

The methods we have presented show promise in detecting changes to both common and niche communities of interest, a challenge relevant to many academic and commercial fields. It is expected that most of the techniques applied here to Reddit can be generalized to other data sources such as Twitter, although additional work would be required to define community membership and analyze shorter and less structured messages. Verification and analysis of the differences between Reddit and Twitter (and other sources) are subjects for future research.

The analysis and explanation of the *causes* of detected changes in social media is also an area of interest, and the overall efficacy of these approaches likely depends on the ability to correlate and contextualize anomaly detections with auxiliary sources of information. Future work includes automatically generating an interpretation of detected anomalies, detecting events by correlating weak changes across multiple communities, and studying the long term inter- and intra-community effects caused by detected events. A more rigorous utility study must also be performed correlating detected anomalies with real-world events in order to quantify the robustness and reliability of the event detection across a range of communities.

References

- [1] David Agren. “Tom Brady’s stolen Super Bowl jersey found in Mexico”. In: *The Guardian* (Mar. 2017).
- [2] Alexa. <https://www.alexa.com/siteinfo/reddit.com>.
- [3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. “Twitter catches the flu: detecting influenza epidemics using Twitter”. In: *Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 1568–1576.
- [4] Ahmer Arif et al. “How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation”. In: *ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016, pp. 466–477.
- [5] Yoshua Bengio et al. “A neural probabilistic language model”. In: *Journal of Machine Learning Research* 3 (Feb. 2003), pp. 1137–1155.
- [6] James Benhardus and Jugal Kalita. “Streaming trend detection in Twitter”. In: *Intl. Journal of Web Based Communities* 9.1 (2013), pp. 122–139.
- [7] Deepayan Chakrabarti and Kunal Punera. “Event Summarization Using Tweets”. In: *ICWSM* 11 (2011), pp. 66–73.
- [8] Andrew Crooks et al. “#Earthquake: Twitter as a distributed sensor system”. In: *Transactions in GIS* 17.1 (2013), pp. 124–147.
- [9] Jeremy Diamond. “Trump says Patriots’ Tom Brady, Bill Belichick supporting him”. In: *CNN* (Nov. 2016).
- [10] Jérémy Ferrero et al. “Using Word Embedding for Cross-Language Plagiarism Detection”. In: *15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2. 2017, pp. 415–421.
- [11] Gregory Finley, Stephanie Farmer, and Serguei Pakhomov. “What Analogies Reveal about Word Vectors and their Compositionality”. In: *6th Joint Conference on Lexical and Computational Semantics*. 2017, pp. 1–11.
- [12] Santo Fortunato and Darko Hric. “Community detection in networks: A user guide”. In: *Physics Reports* 659 (2016), pp. 1–44.
- [13] Reilly Grant et al. “Discovery of Informal Topics from Post Traumatic Stress Disorder Forums”. In: *Intl. Conference on Data Mining Workshops*. IEEE. 2017, pp. 452–461.
- [14] William L Hamilton et al. “Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora”. In: *2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 595–605.
- [15] Jack Hessel, Chenhao Tan, and Lillian Lee. “Science, AskScience, and BadScience: On the Coexistence of Highly Related Communities.” In: *ICWSM*. 2016, pp. 171–180.
- [16] Jack Hessel et al. “What do Vegans do in their Spare Time? Latent Interest Detection in Multi-Community Networks”. In: *arXiv:1511.03371* (2015).
- [17] Edilson Anselmo Corrêa Júnior, Vanessa Queiroz Marinho, and Leandro Borges dos Santos. “NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis”. In: *Proc. of the 11th Intl. Workshop on Semantic Evaluation*. 2017, pp. 611–615.
- [18] Ravneet Kaur and Sarbjee Singh. “A survey of data mining and social network analysis based anomaly detection techniques”. In: *Egyptian Informatics Journal* 17.2 (2016), pp. 199–216.
- [19] Leo Kelion. “Dark web markets boom after AlphaBay and Hansa busts”. In: *BBC* (Aug. 2017).
- [20] Emmett Knowlton. “The Patriots shocked the NFL world by trading one of their best defenders to the worst team in the NFL”. In: *Business Insider* (Oct. 2016).
- [21] Tal Linzen. “Issues in evaluating semantic spaces using word analogies”. In: (2016), pp. 13–18.
- [22] Steffen Lyngbaek. “SPORK: A Summarization Pipeline for Online Repositories of Knowledge”. MA thesis. CA Polytechnic State University, 2013.

- [23] Carlos Martin, David Corney, and Ayse Goker. "Mining newsworthy topics from social media". In: *Advances in Social Media Analysis*. Springer, 2015, pp. 21–43.
- [24] Sara Melvin et al. "Event Detection and Summarization Using Phrase Network". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 89–101.
- [25] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv:1301.3781* (2013).
- [26] Andreas Mueller et al. *word_cloud: 1.2.1*. <https://doi.org/10.5281/zenodo.49907>. Apr. 2016.
- [27] El Moatez Billah Nagoudi, J  r  my Ferrero, and Didier Schwab. "LIM-LIG at SemEval-2017 Task 1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting". In: *Intl. Workshop on Semantic Evaluations*. 2017, pp. 125–129.
- [28] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. "Summarizing sporting events using Twitter". In: *ACM Intl. Conference on Intelligent User Interfaces*. 2012, pp. 189–198.
- [29] Sen Pei et al. "Searching for Superspreaders of information in real-world social media". In: *Scientific Reports* 4 (2014), p. 5547.
- [30] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global vectors for word representation". In: *Conference on Empirical Methods in Natural Language Processing*. 2014, pp. 1532–1543.
- [31] Swit Phuvipadawat and Tsuyoshi Murata. "Breaking news detection and tracking in Twitter". In: *Web Intelligence and Intelligent Agent Technology*. Vol. 3. 2010, pp. 120–123.
- [32] Jinshan Qi et al. "Discrete time information diffusion in online social networks: micro and macro perspectives". In: *Scientific Reports* 8.1 (2018), p. 11872.
- [33] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors". In: *19th Intl. Conference on World Wide Web*. ACM, 2010, pp. 851–860.
- [34] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic". In: *PloS one* 6.5 (2011).
- [35] *The New Dictionary of Cultural Literacy, Third Edition*. Houghton Mifflin Harcourt Publishing Company, 2005.
- [36] Soroush Vosoughi and Deb Roy. "A Semi-Automatic Method for Efficient Detection of Stories on Social Media." In: *ICWSM*. 2016, pp. 707–710.
- [37] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. "An exploration of discussion threads in social news sites: A case study of the Reddit community". In: *Advances in Social Networks Analysis and Mining*. IEEE. 2013, pp. 579–583.
- [38] Tom Winter. "AlphaBay, Hansa Shut, but Drug Dealers Flock to Dark Web DreamMarket". In: *NBC* (July 2017).
- [39] William H Woodall et al. "An overview and perspective on social network monitoring". In: *IISE Transactions* 49.3 (2017), pp. 354–365.
- [40] Rose Yu et al. "A survey on social media anomaly detection". In: *ACM SIGKDD Explorations Newsletter* 18.1 (2016), pp. 1–14.
- [41] Jiang Zhao, Man Lan, and Jun Feng Tian. "ECNU: using traditional similarity measurements and word embedding for semantic textual similarity estimation". In: *Intl. Workshop on Semantic Evaluation*. 2015, pp. 117–122.
- [42] Siqi Zhao et al. "Sportsense: Real-time detection of NFL game events from Twitter". In: *arXiv:1205.3212* (2012).