Through The Grapevine: A Comparison of News in Microblogs and Traditional Media

Byungkyu Kang, Haleigh Wright,
Tobias Höllerer, Ambuj K. Singh, and John O'Donovan

Department of Computer Science,
University of California, Santa Barbara. CA, 93106-5110, United States
{bkang,wright,holl,ambuj,jod}@cs.ucsb.edu

Abstract. In recent years the greater part of news dissemination has shifted from traditional news media to individual users on microblogs such as Twitter and Reddit. Therefore, there has been increasing research effort on how to automatically detect newsworthy and otherwise useful information on these platforms.

In this paper, we present two novel algorithmic approaches—content-similarity computation and graph analysis—to automatically capture main differences in newsworthy content between microblogs and traditional news media.

For the content-similarity algorithm, we discuss why it is difficult to capture such unique information using traditional text-based search mechanisms. We performed an experiment to evaluate the content-similarity algorithm using a corpus of 35 million topic-specific Twitter messages and 6,112 New York Times articles on a variety of topics. This is followed by an online user study (N=200) to evaluate how users assess the content recommended by the algorithm. The results show significant differences in user perception of newsworthiness and uniqueness of the content returned by our algorithm.

Secondly, we investigate a method for identifying unique content in microblogs by harnessing network structure of the information propagation graphs. In this approach, we study how these two types of information differ from each other in terms of topic and dissemination behavior in the network. The results show that the majority of subgraphs in the traditional group have long retweet chains and exhibit a giant component surrounded by a number of small components, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore, results from LDA and BPR algorithms indicate that strong and dense topic associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

Keywords: Newsworthiness, Newsworthy, Unique, Content Similarity, N-Gram, Topic Modeling, LDA, BPR, Graph Algorithm, Social Network, Information Dissemination, Information Propagation, Twitter, Retweet

1 Introduction

Over the last decade, microblogs have evolved from an online communication channel for personal use to a central hub for information exchange between users. On microblogging platforms, users produce or share information with friends or strangers. Recent studies revealed that the greater part of today's internet users rely on information on microblogs [19] (e.g. Twitter and Reddit) as a primary source of a wide range of information, particularly news. Accordingly, this new paradigm highlights the importance of automated tools that detect reliable and newsworthy information on microblogs.

Going beyond typical information consumers, professional journalists also admit to relying heavily on social media streams for their news stories [18, 32]. During the last decade, microblogs have been studied by researchers in communication and journalism as an essential news gathering tool and several guidelines are proposed[1]. Many users favor to browse microblogs such as Reddit and Twitter on a daily basis since these platforms provide personalized news content based on their previous browsing patterns. Recent research also highlights that traditional news outlets still play an important role in the provision of reliable, well curated news content [19].

However, news outlets are typically biased in some way or other, and do not always act as the best information filters in all cases. A recent study by [11] highlights the polarizing political bias that exists across most of the top US traditional news outlets. Despite the possibility for bias, we believe that curated news from a variety of sources can be leveraged to help identify and classify newsworthy messages in social media streams. In particular, we propose a novel method for identifying *niche* user-provided topics from social media that is a) not reported in traditional curated news, and b) is newsworthy information. Figure 1 shows an overview of our first approach. Each data point represents a Twitter post, located on the x-axis by similarity to a target set of news articles, and on the y-axis by general newsworthiness of the message content. The distribution shows a linear trend indicating the correlation of newsworthiness and similarity to curated content, as we would expect to see. In this case however, we are interested in the highlighted "niche content" section in the top left of the graph, which contains those unique messages that are *not similar* to mainstream media, but do have newsworthy content based on other metrics. This content could be found through a series of text based search queries, but defining relevant keywords is difficult, and may potentially only uncover a given slice of the true overlap between the data sources. To explore this concept, we study a variety of topics from 37 million Twitter posts and 6,112 New York Times articles and attempt to answer the research questions below. The authors would like to note that an earlier version of this study has been published in [22]. The novel contribution in this manuscript includes all of the research on network-based (LDA and BPR) analysis of news in social media networks.

[1] `http://asne.org/Files/pdf/10_Best_Practices_for_Social_Media.pdf`

1. **RQ1** How can we best detect newsworthy information in social media that is not covered by traditional media?
2. **RQ2** How do information consumers perceive the detected information?
3. **RQ3** How do the niche information get propagated differently from traditional news in the network?

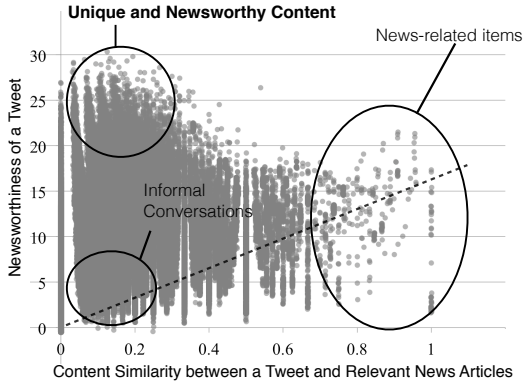


Fig. 1: Overview of approach to filtering unique and newsworthy content. Y-axis tweet newsworthiness is computed from NLTK and from Human Evaluation. X-axis is tweet similarity to mainstream news.

In this paper, we propose two distinct approaches to capture unique news content on microblogs. First approach is based on a variety of content-similarity metrics. Simply put, we compute different content-based similarity metrics on microblog posts and a corpus of traditional news articles. Using these similarity metrics with our newsworthiness scores computed on individual tweets, we can locate the niche (unique and newsworthy) contents and analyze them to find important features that can be utilized for developing automated detection algorithm. Specifically we describe two experiments: first, an automated evaluation is performed to test a variety of mechanisms that predict overlap between a microblog post and a corpus of news articles. These include manipulations on n-grams, part-of-speech tags, stop words and stemming techniques. A co-occurrence score is produced for each message, which is in turn compared to a set of manually annotated newsworthiness scores, combined with a content-based newsworthiness score. The different strategies are ranked by the resulting distance and the best approach is used for experiment 2. Manual annotations of newsworthiness were collected using a crowd-sourced study described in [30]. The second experiment samples data in various ways from the highlighted areas of Figure 1 for a range of topics and presents an A/B style questionnaire about newsworthiness, similarity to traditional media content, and personal focus to 200 participants in an online study.

Results of experiment 1 show that a simple n-gram approach with word-stemming but without stop word removal produced the most accurate approximation of the manual annotations. Results from experiment 2 show that there is a significant difference in reported "similarity to mainstream news content" for messages sampled from the top left area of Figure 1 compared with a random sample from the right side, indicating that the method is capable of automatically identifying newsworthy content that is not covered by mainstream media.

To address **RQ3**, the second approach—network analysis on microblog news contents—has been demonstrated in Section 4. In this approach, we apply a variety of commonly used network measures of structural and functional connectivity to microblog information to unveil unique characteristics that represent both niche and generic news contents on microblogs. Particularly, two experiments are performed on the collection of 2.4M Twitter dataset to find the differences between the two groups (niche and traditional groups) in network topology (`Exp 1`) and topical association across users (`Exp 2`).

Results of `Exp 1` show that the majority of subgraphs in the traditional group have long retweet chains with a giant component surrounded by a number of small components. On the other hand, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore, results from `Exp 2` indicate that strong and dense topic associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

The differences between the unique and traditional news groups that we found in this study will benefit future studies for intelligent and scalable algorithms to automatically classify or predict unique or interesting news in microblogs. We will discuss our future work and possible applications for which our model can be applied in Section 5 and 6.

2 Related Work

With the increasing reliance on user-provided news content from microblogs, recent research has focused on the relationship between microblogging platforms and traditional news outlets [24, 14, 35]. As we briefly discussed in the previous section, news content, including opinions and conversations about news now comprise a significant portion of overall content on microblogs. Hermida et al. [18] conducted a large-scale online survey and unveiled behaviors of news consumers on social media including microblogs. According to many studies, including [18], microblogs such as Twitter have become a major source of news information for individual consumers and also for professional journalists who rely on the dynamic content for story-hunting and marketing.

2.1 Microblogs and Traditional Media

Over their short history, microblogs have been a communication channels upon which users share useful information that they discover elsewhere, such as on-

line news media, blogs or forums. Recent studies have focused on the relationship between microblogs and traditional news outlets since both end-users and journalists rely on microblogs for information. To understand the relation between these sources, researchers investigated association using topic modeling algorithms such as LDA [35, 14]. Furthermore, since microblog users not only reproduce and forward original information but sometimes re-shape content by adding additional value such as personal opinion or on-site images of an event, "produsage" (the hybridization of production and consumption) behavior and its byproducts have been studied [20] on different types of news contents: soft and hard news.

2.2 Newsworthiness

Shoemaker [29] argue that news and newsworthiness have different underlying concepts. However, they also admit that newsworthiness is one of the important components that makes news public. In this study, we assume that newsworthiness is a core information attribute that categorizes a piece of content in terms of usefulness to the general public.

Quality of information in microblogs has been widely studied in the information retrieval community, and remains releveant in this research. André et al. [4] studied microblog content through the first large corpus of follower ratings on Twitter updates collected from real users. They found that 64% of tweets are reported as not worth reading or middling, which implies that users tolerate a large amount of useless information on microblogs. In addition, factors that make microblog content 'useful' and 'not useful' were investigated through a qualitative study in search tasks [21]. We revisit their question about the content value in microblogs with particular focus on their unique role in news consumption. In other words, we examine microblog contents and pan for niche content which only exists on microblogging platforms, not others. Community feedback was also exploited to automatically identify high-quality content in the Yahoo! Answers [1] community question/answering platform.

Our research examines several low-level features of microblog posts to arrive at a good classifier. Castillo et al. [13] also explored features that can be exploited to automatically predict newsworthiness of information on microblogs. Participants of their crowd-sourced online study were asked to label a group of microblog messages with either a "news" or "non-news" category. The tweets labeled with news category were then annotated with newsworthiness score in 5 Likert scale in the subsequent annotation task. This study showed the possibility of automated identification of newsworthy information through machine learning. Moreover, the authors revealed important features which can be directly obtained or processed from microblog contents and metadata, without the need for human-labeled examples.

2.3 Content Similarity

Due to the scale and complexity of microblog and news data, it would require a huge effort for an end user to capture newsworthy content in a microblog that is not covered in traditional media using a series of traditional text-based search queries. Our automated approach to filtering for newsworthy information relies heavily on content matching techniques. A wide range of content similarity metrics have been studied and proposed for many years, ranging from simple string-based measures [15, 2] to semantic similarity [26], structural similarity such as stop word n-grams [31] and text expansion mechanisms [23, 8]. In particular, in the context of microblog content analysis,

Herdağdelen [17] proposed n-gram based approach to Twitter messages, which we build on in this research. Another well known approach by Becker et al. [6] learns useful similarity metrics that can be used for event detection in social media data. This approach contrasts to our work in that we adopt a simpler per-tweet similarity. For future work we will examine automated event extraction algorithms and evaluate our uniqueness approaches at the event rather than message level. Other approaches, such as Anderson et al. [3] and Guy et al. [16] take a user-based approach to similarity for social media content analysis. Our approach contrasts to this work by focusing only on similarity at the content level, but we believe that user-level analysis have significant potential in this area, in particular by supporting discovery of broader, more diverse content, and supporting serendipitous discovery of new, unique new content.

Our methods apply several content similarity metrics including normalized word n-grams to determine and measure how two information sources–*microblog* and *traditional news outlet*–are quantitatively associated. We carefully consider the limited nature of microblog contents: the limited number of characters and embedded items. Our choice of metrics for content were proposed in [5]. Bar et al. [5] evaluate different content similarity metrics and report effectiveness and efficiency of the composite of multiple metrics using supervised machine learning approach in their study.

2.4 Network Measures and Metrics

Recent studies have focused on either network structure or retweet behavior [27] by looking at the characteristic of the information diffusion. For example, [27] has shown that call for action type of retweets generated sparse graphs while tweets sharing information generated a denser network during their propagation. Also, [33] proposed a model that measures speed, scale, and range of information diffusion by analyzing survival of each message in the network.

Both content and network-based features considered, a recent work done by Canini et al. [12] shows a good example of how various features can be used to measure the quality of information on microblogging platforms.

2.5 Topical Similarity and Object Association

In the network analysis, several different approaches are considered to model the topic space of each group (traditional and niche) in the given network. We first apply Latent Dirichlet Allocation (LDA) topic-modeling algorithm [9, 28] in order to generate n words on each group of messages.

Secondly, we compare one from another in terms of structural similarity of the topics extracted from the content. Specifically, we tokenize each message into individual elements and compute stopwords and word n-grams as used in [25].

Afterward, Bipartite Projection via Random Walks (BPR) [34] is applied to construct topic-similarity network for each group. BPR is a method that produces associations between objects, defined in [34]. This method performs random walks on a two-mode bipartite network. In this network, edges exist only between nodes of different modes, and these edges represent an association between these two nodes. The result of this random walk is a one-mode unipartite network that captures the similarity between nodes of the chosen mode. This method takes into account the overall structure of the original bipartite network.

3 Content Similarity based Approach

This section describes our approach to filtering unique and newsworthy content from microblog streams based on comparison with contents from mainstream media. According to the study in [29], Shoemaker claims that newsworthiness is not the only attribute which represents news. However, in this study, we adopt newsworthiness as the central indicator of news contents in general. Basically, we assume here that curated news articles are newsworthy. Our first approach exploits news articles as a reference to identify Twitter postings about a target topic that are newsworthy but are not the focus of curated mainstream news. We begin by exploring a set of mechanisms for computing similarity between a microblog post and a topic-specific corpus of news articles.

3.1 Data Collection

To examine real-world microblog messages and news contents, we choose "Twitter" and "New York Times" as representative examples for microblogging platforms and traditional media outlets. Both provide well documented application program interfaces (APIs)[2] through which we can retrieve microblog messages or news articles as well as a rich set of metadata (e.g. keywords, embedded multimedia items, urls). With the two APIs we collected about 35 million (35,553,515) microblog messages from Twitter and 6,112 news articles from New York Times and other sources such as Reuters and Associated Press (AP). An overview of this data collection is shown in Table 1. Before the crawling stage, we selected

[2] New York Times Article Search API: `http://developer.nytimes.com/docs`
Twitter API `http://dev.twitter.com`

major news events such as natural disasters, world cup and various political issues over the course of 4 years (2012 - 2015) to examine how both media differs from each other and see if there is topic-specific bias across different events. We collected topic-specific data sets[3] using related keywords to retrieve microblog messages and news articles from Twitter and New York Times databases. In particular, for Twitter data, we used the Streaming API to monitor transient bursts in the message stream while we collected regular data about the events.

Table 1: Overview of the data sets collected from New York Times and Twitter.

topic	*world cup*	*ISIS*	*earthquake*	*hurricane sandy*
tweets	22,299,767	8,480,388	921,481	3,851,879
articles	4,097	422	329	1,264
from	6/24/14	1/20/15	1/20/15	10/29/2012
to	7/17/14	3/29/15	3/31/15	12/31/2012
days	24	69	71	64

3.2 Similarity Computation

A key challenge in this approach is to discover meaningful mappings between a short microblog post and a larger corpus of news articles. Since traditional text-matching mechanisms such as TF-IDF or topic modeling do not work well with short messages, a variety of simpler mechanisms were evaluated. Table 4 shows an overview of the mechanisms tested and their performance with respect to manually labeled "ground truth" assessments of newsworthiness. An initial pre-processing was applied to all messages to remove superfluous content such as slang and gibberish terms.

Word n-grams Next, a set of word n-grams as described in [5] were computed, varying n from 1 to 3. Part-of-Speech (POS) tagging was applied to identify potentially useful noun, verb, pronoun and adjective terms. A standard stop-word list was identified and systematically removed as shown in Table 4. A Twitter-specific stop-word list was compiled from a manual analysis of posts. This list contained platform-specific terms such as "twitter", "rt", "retweet", "following" etc., based on a term frequency analysis. In total, 24 combinations of lightweight NLP techniques were applied to 4 topic-specific collections of twitter posts and NYT news articles. These are detailed in Table 4. Each method computed a content-based similarity score between a *single* microblog post and a larger collection of news articles.

For each topic studied, we obtained thousands of n-grams from the NYT article collection and use it as a corpus of news n-grams ($n = 1, 2, 3$). Next, we applied n-gram extraction on the entire tweet collection and computed the number of co-occurrences of n-grams from each post with those in the news n-

[3] Dataset available upon email request

gram corpus. To account for length deviation, this score ($Score$) was normalized by the total number of n-grams in each tweet.

Newsworthiness In this study, we apply a two dimensional approach to newsworthiness: (1) news term frequency in each tweet ($News_{Term}$) and (2) newsworthiness score labeled by real-world microblog users ($News_{User}$) in [0-5] Likert scale.

For $News_{Term}$, we compute number of tokens that contain news terms using the Reuters news word corpus in NLTK[4] and divide this number by total number of tokens in each message.

$News_{User}$ is the human-annotated newsworthiness score, and is also normalized by the maximum score. Normalization is performed on both metrics in order to eliminate bias of different message sizes in tweets and take the average of the two metrics for Equation 1. Table 2 shows the selected set of similarity metrics that we employ in this study.

Table 2: Metrics analyzed in the study.

Metrics	Nomenclature	Description
n-gram Similarity	$Score$	Number of n-grams that co-occur between news article corpus and a tweet
News Word Frequency	$News_{Term}$	News word frequency with NLTK Reuters corpus
Newsworthiness Score	$News_{User}$	Human annotated newsworthiness score [0-5] on a tweet

3.3 Strategy Selection

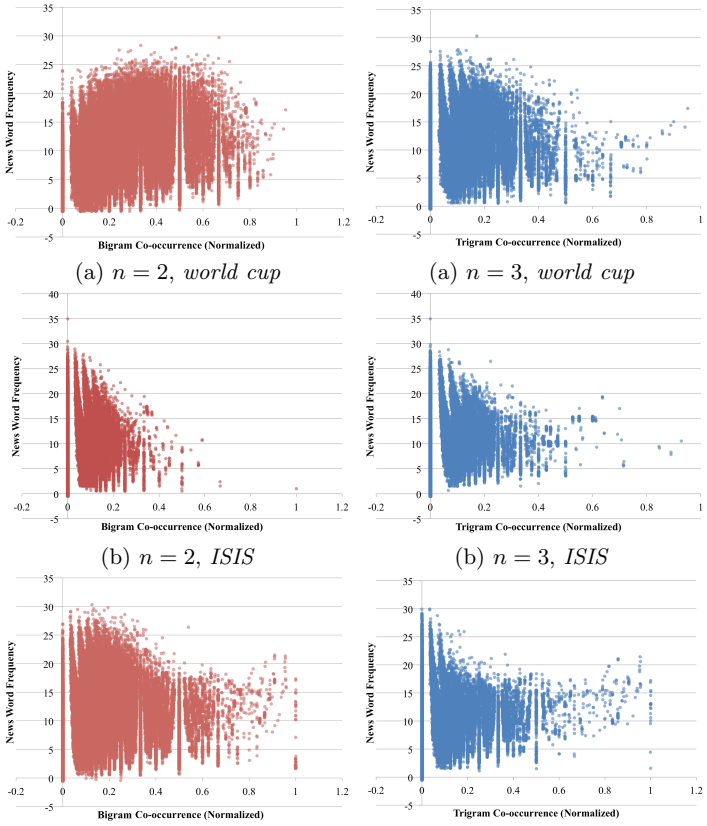
We define a simple inverse distance metric in order to evaluate our content-based similarity measure ($Score$) and select the best performer among 24 candidates. This metric is then applied to the composite sets of multiple metrics to select the best feature based on the linear relationship between the similarity score and newsworthiness of a message. We discuss the procedure in detail in this section. Afterwards, we explain our evaluation method and procedure in Section 3.5.

Definition 1 *Each event-specific data collection T contains N messages where $T = \{m_1, m_2 \dots m_N\}$, and we represent individual message as m where $m \in T$. Inverse distance of a message between newsworthiness and content similarity to news corpus is represented as $InvDist$.*

$$InvDist(m_i, c_N) = \frac{1}{|News(m_i, c_R) - Score(m_i, c_N)| + 1} \tag{1}$$

Where $News(m_i, c_R)$ is:

[4] NLTK Reuters Corpus has 1.3M words, 10k news documents categorized `http://www.nltk.org`



(a) $n = 2$, *world cup* (a) $n = 3$, *world cup*

(b) $n = 2$, *ISIS* (b) $n = 3$, *ISIS*

(c) $n = 2$, *earthquake* (c) $n = 3$, *earthquake*

Fig. 2: News word frequency on tweets and n-gram ($n = 2, 3$) co-occurrence with mainstream news articles (NYT) on different topics.

$$News(m_i, c_R) = \frac{News_{Term}(m_i, c_R) + News_{User}(m_i)}{2} \quad (2)$$

Please note that c_N and c_R are a corpus of news articles on a topic and the Reuters news vocabulary corpus in NLTK, respectively.

Since we compare one strategy against others in the selection procedure, we use the average of inverse distance for a strategy over all messages, computed using Equation 1.

We apply a fractional function to the inverse distance metric in Equation 1. Intuitively, this approach maximizes gain in highly correlated messages and, likewise, penalize un-correlated messages between newsworthiness $News(m)$ and content similarity $Score(m)$. As briefly mentioned earlier in this section, we believe that both $News_{Term}$ and $News_{User}$ represent different aspects of newsworthiness. Unlike the n-gram co-occurrence ($Score$), which reflects the word-based association on a specific-event, $News_{Term}$, which is corpus-based news word frequency, represents topic-independent association between a microblog message and the Reuters news word corpus. To validate our inverse distance metric, we performed Pearson and Spearman correlation tests with the best feature selected and they are shown in Table 3. The best feature selection is summarized in Algorithm 1.

Algorithm 1: n-gram strategy evaluation (Best feature selection)

Result: Best performing strategy
initialization;
for *all n-gram strategies* **do**
  for *all message m where* $m \in T$ **do**
    $nGram \leftarrow$ computeNGramScore(m, *strategy*, $corpusNYT$);
    $newsTerm \leftarrow$ computeNewsTerm(m, $corpusReuters$);
    $news \leftarrow$ mean($newsUser$, $newsTerm$);
    $similarity \leftarrow$ computeSimilarity($news$, $nGram$);
  end
  $\overline{similarity} \leftarrow 1/n \sum_{i=1}^{N}$;
end
$best \leftarrow \underset{strategy}{\text{argmax}}\, \overline{similarity}$;
return *best*

As shown in Table 4, *unigram with stemmer only* feature has the highest correlation. Therefore, we select this feature for our user experiment and evaluation.

Table 3: Correlation coefficients between newsworthiness $News(m)$ (arithmetic mean of news word frequency and user annotated newsworthiness score) and n-gram co-occurrence score $Score(m)$ (all metrics normalized [0,1])

	Correlation Coeff.	2-Tailed Test Significance
Pearson	0.47063	$< 1e-10$
Spearman	0.41414	$< 1e-10$

Avg # Terms in News	Avg # Terms in Tweets	# Co-occurence	Stopword Removal	Stemming	Noun Only (POS-tag)	n-gram	Inverse Distance
3,863	17.952	10.509	N	N	N	1	0.774
12,085	16.965	1.713	N	N	N	2	0.814
15,246	16.011	0.162	N	N	N	3	0.689
1,719	7.401	2.678	N	N	Y	1	0.777
4,596	6.532	0.144	N	N	Y	2	0.75
5,792	5.762	0.014	N	N	Y	3	0.714
1,596	17.952	3.868	N	Y	N	1	**0.96**
4,592	16.965	0.165	N	Y	N	2	0.758
5,790	16.011	0.006	N	Y	N	3	0.740
1,564	7.401	2.654	N	Y	Y	1	0.736
4,509	6.532	0.145	N	Y	Y	2	0.8
5,678	5.762	0.014	N	Y	Y	3	0.769
1,557	11.161	1.744	Y	N	N	1	0.857
4,495	10.171	0.068	Y	N	N	2	0.714
5,664	9.251	0.006	Y	N	N	3	0.666
1,557	6.217	1.473	Y	N	Y	1	0.857
4,495	5.345	0.057	Y	N	Y	2	0.8
5,664	4.611	0.007	Y	N	Y	3	0.666
1,557	11.161	2.949	Y	Y	N	1	0.857
4,495	10.171	0.163	Y	Y	N	2	0.833
5,664	9.251	0.013	Y	Y	N	3	0.8
1,557	6.217	1.99	Y	Y	Y	1	0.857
4,495	5.345	0.136	Y	Y	Y	2	0.8
5,664	4.611	0.015	Y	Y	Y	3	0.666

Table 4: [n-gram table] Comparison of different NLP mechanisms applied to computing co-occurrence between a microblog message and a news corpus (topic:*occupysandy*). Each row in this table represents a different combination of text-matching mechanisms that were evaluated in our study.

3.4 Experimental Setup

In this paper, we aim to identify unique newsworthy contents on microblogs that differs from those in mainstream news media like New York Times. So far we have explored different features based on content similarity metrics and text processing techniques. To validate our approach discussed in the previous section, we conduct an experiment including a crowd-sourced user study.

Random Sampling For the experiment, we randomly sample 10,000 tweets from each collection. This sampling task allows us to avoid possible scalability issue from the high volume of our data sets and fit the experiments and user study. We sampled tweets that are primarily written while events were taking place or shortly thereafter. For the NYT articles, however, we aggregate them together first before we compute similarity features.

Niche Content Extraction Our hypothesis is that, in general, newsworthy contents on microblogs do not completely overlap with mainstream news contents. In this study, the term "niche content" was coined for microblog exclusive (unique) newsworthy information. As the coined term implies, we assume that this type of information has a unique value and, thus, we believe that it is worth to investigate. The aim of this study is to find the unique characteristics of the niche content on microblogs and exploit our findings to provide a guideline to design more effective newsworthy information filtering algorithm in many applications.

We apply both statistical and heuristic approaches, including manual inspection on the contents with semantic relatedness in mind, to the experiment. Specifically, we manually inspect frequently used unigrams (see Table 6) after removing noisy information via stop word removal. Next, we classify these frequent terms into three different groups. Exploratory analysis such as frequency and burst analysis was also performed to scrutinize the data collections and compare contents from different categories with the features. We then sample microblog messages from two different groups: contents with high/low similarity with regard to mainstream news media contents. To perform this second-phase sampling task, we choose 20 and 80 percentile in n-gram feature distribution as the thresholds. We will provide some insights into the distinction that we interpreted from the experiment and discuss limitations later in Section 3.5.

User Study Following our content extraction and comparative analysis, we conduct a crowd-sourced user study to validate our hypothesis. In the user study, the participants were shown two groups of 10 tweet messages. Each group of tweets were randomly sampled from the messages with high similarity and low similarity to main stream news media contents in $News_{n-gram}$ metric, respectively. The participants were then asked to answer 6 different questions regarding (1) similarity to traditional news articles, (2) newsworthiness and (3) how personal the shown content is. They were also asked to answer to general questions

such as demographic information (gender, age, education level, etc.) and their microblog usage.

3.5 Evaluation

We now discuss evaluation of the research questions posed earlier. Using the best performing co-occurrence method from the 24 mechanisms for computing similarity between a short Twitter message and a larger collection of news, showing in 4, we conducted a user experiment to assess perceived differences between messages sampled from the niche areas shown in Figure 1 and a general sampling of messages in the topic. The experiment consisted of two conditions: 1) message sampling along the 20th and 80th percentiles of the x-axis from Figure 1 (I.e.: the co-occurrence score between a tweet and the NYT article corpus), and 2) messages sampled from the top left corner of Figure 1. I.e.: co-occurrence score combined with a content-based newsworthiness score for the message. This area represents messages that are inherently newsworthy but do not frequently occur in the mainstream corpus. In both conditions, the samples were shown alongside randomly sampled messages about the topic and user perception was evaluated. Information consumers can perceive newsworthiness differently over time, so we first examine a sample of temporal distributions of topics across the two domains (NYT and Twitter).

Frequency Analysis Figure 3 shows a frequency analysis of Twitter postings and NYT articles related to the 2014 world cup. Multiple peaks on both line plots show sudden bursts of discussions (on microblogs) or reports (from news outlets) on the corresponding topic (*world cup*). In this representative example, both streams follow a similar trend, but the bursts are more pronounced on Twitter than in traditional news. This trend in bursts is representative of several analyzed topics, so, while Twitter appears to be more reactive to events in terms of bursts, both streams show peaks of interest for critical events (semi-final and final in this case), indicating that newsworthiness of events is similar on both sources.

	# of Terms in News			Avg # of n-grams in a Tweet			Avg % of Co-occurrences		
Topic	unigram	bigram	trigram	unigram	bigram	trigram	unigram	bigram	trigram
world cup	9,274	75,036	122,573	18	17	16	77.7%	25.6%	6.3%
ISIS	2,573	9,764	12,724	19	18	17	63.1%	14.9%	2.4%
earthquake	2,303	7,114	8,772	18	17	16	64.3%	15.9%	4.1%
occupysandy	3,078	11,865	15,190	18	17	16	60.5%	10.3%	1.0%

Table 5: Statistics overview across different data sets (stemming only)

Study Participants and Procedure Participants for the user experiment were recruited though Amazon's Mechanical Turk (MTurk). A total of 200 participants took the study which lasted an average of 8 minutes. 48% of participats

	Article		Common		Tweet	
	word	#	word	#	word	#
worldcup	2014	412	worldcup	4801	fifaworldcup	1011
	thursday	231	world	2492	bra	763
	skiing	86	cup	2363	arg	706
	longman	76	soccer	1161	ned	551
	table	65	brazip	1077	joinin	418
	association	64	germany	873	mesutozil1088	296
	1994	61	ger	656	worldcup2014	294
	golf	60	final	598	gerarg	273
	governing	60	team	580	fra	214
	christopher	59	argentina	509	crc	211
ISIS	8217	33	isis	4872	amp	665
	adeel	16	iraq	445	via	497
	2015	13	syria	370	dress	294
	fahim	12	obama	340	cnn	170
	schmitt	11	islamic	339	isil	162
	1973	10	video	295	share	134
	fackler	8	state	281	foxnews	126
	corrections	6	us	274	bokoharam	119
	badr	6	alive	259	usa	113
	abdurasul	5	jordan	225	daesh	107
earthquake	sniper	31	earthquake	5165	utc	484
	2011	22	magnitude	835	amp	333
	kyle	19	japan	515	breaking	309
	defense	15	tsunami	451	feel	274
	former	14	california	348	via	261
	marine	12	usgs	345	newearthquake	254
	tea	10	new	333	mar	192
	routh	9	ago	295	alert	191
	navy	8	strikes	256	sismo	186
	nations	8	quake	245	map	161
occupysandy	blackouts	49	sandy	641	occupysandy	5867
	andrew	32	help	410	sandyaid	598
	presidential	30	new	343	ows	425
	conn	29	need	298	sandyvolunteer	340
	newtown	28	hurricane	248	please	329
	barack	26	relief	207	occupywallstnyc	310
	education	25	nyc	205	520clintonos	269
	connecticut	24	volunteers	194	today	264
	gasoline	21	occupy	193	info	216
	senate	21	rockaway	182	thanks	210

Table 6: Top 10 frequent words extracted from tweets on each topic.

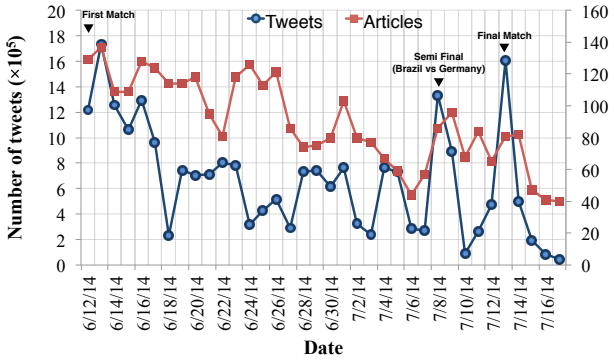


Fig. 3: Temporal distribution of the microblog messages (tweets) and news articles on the topic–*worldcup*. The time period shown in this graph corresponds to the 2014 world cup held in Brazil.

were male and 52% were female. All participants were active microblog users. Age ranged between 18 and 60, with the majority between 25 and 50 (78%). 69% of participants reported having a 4-year college degree or higher. Participants were all located within the United States and had completed a minimum of 50 previous successful tasks on the MTurk platform.

Participants were shown a Qualtrics survey[5] that asked basic demographic questions. Next, they were shown two groups of 10 microblog posts, side by side with random ordering. Two conditions were evaluated. Condition 1 showed groups of messages randomly sampled from within the 20th and 80th percentiles along the x-axis of Figure 1. To recap, this axis represented the co-occurrence score of the best performing mechanism from Table 4. Condition 2 users were shown ten messages that were sampled from the top left portion highlighted in Figure 1 (the 'unique' and 'newsworthy' messages), and ten randomly sampled from within the topic. This selection used both the x-axis similarity and the content-based newsworthiness score described earlier. In each case, participants were asked to rate their agreement with three statements for each group shown (total of 6 ratings):

1. *The messages in group x are similar to what I would find in mainstream news such as the New York Times.*
2. *The messages in group x are newsworthy*
3. *The messages in group x are personal*

Results Results of the experiment are shown as box plots in Figures 4 and 5. Our first task was to assess the effect of the co-occurrence metric chosen from the 24 options in Table 4. Two random groups of 10 tweets were sampled from the

[5] www.qualtrics.com

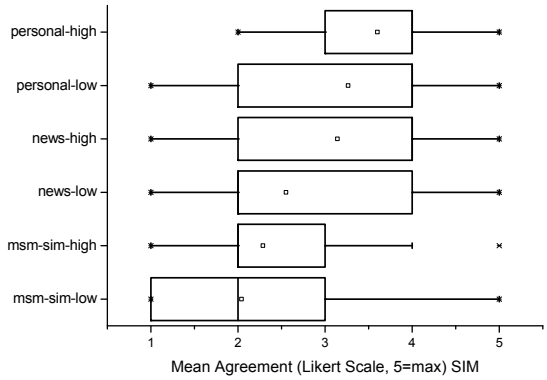


Fig. 4: Mean agreement of the responses from condition 1 – `SIM`

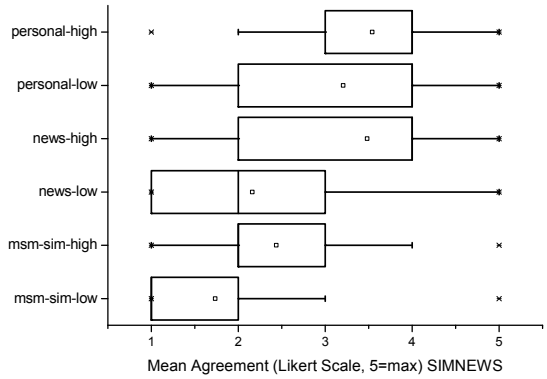


Fig. 5: Condition 2: Mean agreement of the responses from the user study – `SIMNEWS`

poles of this distribution (shown as the x-axis in Figure 1) and displayed side-by-side to participants. Participants were asked to rate their agreement with the questions listed above on a Likert scale of 1-5, with 5 indicating full agreement with the statement. Responses to the above questions are shown in Figure 4. Participants reported that the similarity to mainstream media was higher for messages with high co-occurrence, but, we did not observe a statistical significance for this result. Figure 5 however, does show a significant difference at $p<0.05$ between the sampled messages. So, by augmenting the co-occurrence score with a content-based newsworthiness score, shown in Equation 2, we achieved a significant shift in perception of uniqueness of content. Interestingly, the perception of newsworthiness for these messages was reasonably high and did not change significantly along the x-axis (similarity to NYT), meaning that the approach did find messages that people felt were unique to the microblog domain and were also newsworthy.

Results of a term-based analysis are shown in Table 6 which displays three sample topics ("worldcup", "ISIS" and "Earthquake". The table shows the top n=10 terms from each data set as they overlap with the source data. The left column (Article) shows terms that are mostly unique to news articles. The center column shows combined terms, while the rightmost column shows terms that are popular on Twitter but not overlapping with the mainstream news. From manual inspection, the combined terms in the middle column in Table 6 appear to be a good descriptor of the topic. For example, the "ISIS" topic contains "ISIS"; "IRAQ"; "SYRIA"; "OBAMA"; "ISLAMIC" as the top 5 terms. Terms unique to mainstream media appear to be focused more on official structures and laws, while terms unique to the microblog tend to be more personal and emotional. Interestingly, the term "BOKOHARAM" is listed in the microblog column. This is a good example of a global news phenomenon that is covered extensively in most countries, but is relatively under-reported in the United States. Now we will discuss our results in the context of the research questions presented earlier.

RQ1 How can we best detect newsworthy information in social media that is not covered by traditional media? We have examined 24 mechanisms for computing the similarity between a short microblog post and a corpus of news articles. Our findings show that a simple approach using simple unigram term matching and a porter stemming algorithm provides a better approximation of manually labeled examples than other methods tested, including POS tagging, stop-word removal and matching on bi-grams and tri-grams. Our initial expectations were that bi-gram and tri-gram overlap would produce better matches to the manual labels. Our experimental data showed that single term overlap was a better metric. We assume that since microblog posts have a limited number of terms, overlap in bi and tri-grams was sparse, as highlighted by the statistics in Table 4. For example, unigram co-occurence for the topic "ISIS" shows 78% overlap with the news article database, while bi-gram overlap is 26% and trigram overlap is just 6.3%. For future work we plan to apply a combination of n-gram overlaps to create better mappings between microblog posts and news articles.

RQ2 How do information consumers perceive the detected information? Our online evaluation of 200 paid participants shows us that sampling messages from the distributions created by the co-occurrence computation produces a significant increase in perception of the uniqueness of messages, while not affecting perception of newsworthiness. We believe that this is a promising result for the automated detection of niche and newsworthy content in social media streams.

4 Network based Approach

Following the previous approach, we propose another approach to capturing unique news content on microblogs using structural and functional metrics of network. In this section, we demonstrate our strategies to find differences in network structure and topic association between niche and traditional groups of tweets.

The main idea that guides our two proposed approaches is that there is a unique portion of newsworthy content in microblogs that are not covered by traditional media. The underlying assumption in the second approach is that such unique content travels from a node to its neighbors in a different fashion from those covered by traditional news outlets. Let us assume that a node u_i produces a "newsworthy" content m_i in the network and m_i becomes exposed to u_i's neighbors in a given time Δ_t. Unlike one-to-many propagations for contents directly provided by traditional media (e.g. tweets posted by @BBC), we expect arbitrary one-to-one or one-to-few type of propagations in the unique content group.

In this approach, we apply 1) network and 2) topic association analyses to our microblog datasets. First, we convert the crawled tweets and their associated users into two different graph data structures (network and topic spaces) based on the typical vertex/edge graph structure ($G = (V, E)$). Before analyzing the two spaces, for the network space, we reconstruct a retweet chain graph using our datasets. In this graph structure, every node, or a vertex, i represents a user u_i, and an edge $(i, j) \in E$ ($E \subset V \times V$) that connects nodes i and j becomes a retweet. We can say that $i \sim j$ if $(i, j) \in E$. For the topic space, we apply topic modeling to microblog messages using Latent Dirichlet Allocation (LDA) and extract associated topics from the messages. Using the topics extracted from the tweets, we construct a bipartite graph G_{LDA}. In this graph, we have a set of users $U = \{u_1, u_2, ...u_m\}$ and another set of topics $T = \{t_1, t_2, ...t_m\}$ that are associated with the users $\in U$. Afterwards, we generate the final graph G_{BPR} using Bipartite Projection via Random Walks algorithm proposed by Yildirim and Coscia [34]. The algorithmic detail of the two methods are described in Section 4.4 and 4.5.

4.1 Hypotheses

In this study, inspired by our motivations, we aim to answer the last research question (**RQ3**) we have in Section 1.

– **RQ3:** How do the niche information get propagated differently from traditional news in the network?

As a recap, in this paper, we assume that the unique and newsworthy contents on microblogs do not completely overlap with mainstream news contents. Accordingly, the following hypotheses are derived to further shape the experimental setup for our network-based approach.

Hypothesis 1 *A difference in network structure can be observed between the spread of niche (unique) and traditional media content.*

Hypothesis 2 *A difference in topical association can be observed between the two groups.*

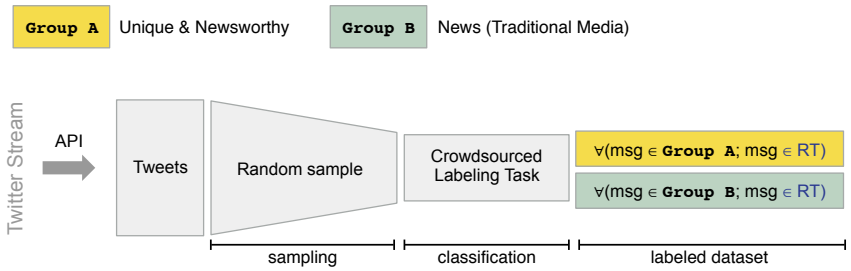


Fig. 6: A diagram that describes crawling and labeling data sets.

4.2 Data Collection and Preprocessing

To utilize real data from the microblogging platform Twitter, microblog posts, or "tweets", were crawled for specific keywords. In this study, we have crawled the total of 2,353,334 tweets using Twitter REST API on three different topics: *#Calais* (86,627), *#prayforparis* (1,431,467), *#paris* (835,240). After examining all datasets, we decided to focus on the *#paris* dataset which covers most news threads and relevant discussions on related subtopics. The datasets were collected during the terrorism in Paris (Nov. 8 $\sim$ Nov. 15.) This crawling process is shown in Figure 6.

Using the crawled datasets, we reconstructed retweet chain graphs in which the nodes represent users and the edges between them represent a retweet. In the data pre-processing task, the content (message text) of each tweet and corresponding metadata such as retweet count, number of friends/followers, user id and screen name, language, self-reported location are extracted using a document-oriented database[6] and parsing scripts.

[6] A NoSQL database (Mongo DB) was used.

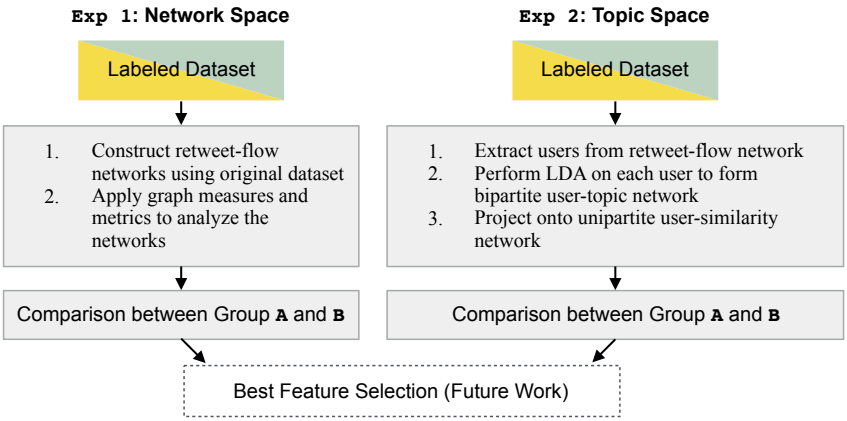


Fig. 7: A diagram that demonstrates how we process data and evaluate the model proposed in the study.

4.3 Labeling Tweets

Before the comparative analysis on the two groups of contents (Group A and B), we need to classify the messages into one of the groups. Since both newsworthiness and uniqueness of content are subjective metrics, we conducted a labeling task on a crowdsourcing platform[7]. Each individual message of the 300 sampled retweets from our data collection is shown to three different participants. During the task, each user was asked to rate *newsworthiness* and *uniqueness* of the given tweet in [1-10] Likert scale and answer the foundation of their judgement on newsworthiness among usefulness, timeliness, novelty (rarity) and interestingness (see Table 7.) We asked multiple participants to label on each message to avoid personal bias towards/against specific topic or information source. Thus, we only use the tweets that have high agreement on both newsworthiness and uniqueness of the content across three participants.

Table 7: Distribution of the foundation of newsworthiness assessment in the labeling task

News type	# Responses	News type	# Responses
Usefulness	314	Timeliness	210
Interestingness	233	Novelty or Rarity	143

[7] Crowdflower (`http://crowdflower.com`) was used for the labeling task.

4.4 Network Analysis (Exp 1)

In this study, we are interested in investigating how "newsworthy and unique" content differs from other generic news contents. In particular, we want to analyze who generally produces this unique content and how this content is structured, i.e. propagated, in the network. Borrowing the perspectives from graph mining and social network analysis, we assume that each node corresponds to a message (or a user who posts/re-posts that message) and each edge to a propagation of a message from a node to its neighboring node. The list of network metrics we use are shown in Table 8.

Table 8: The metrics used for analyzing the network structures of the groups A and B

Metric	Symbol	Description
Node/Edge Count	N/M	Number of nodes and edges of a graph G
Average degree	$< k >$	The mean of number of edges connected to all nodes of the graph G
Closeness Centrality	Cen_C	Inverse average distance to every other vertex
Betweenness Centrality	Cen_B	Fraction of shortest paths that pass through the vertex
Eigenvector Centrality	Cen_E	Importance of a node in a graph approximated by the centrality of its neighbors
Mean Clustering Coefficient	C	The mean clustering coefficient of the graph G

Besides the metrics we use to indicate network structures, in this study, we examine how vertices are associated with their neighbors by looking at the structure of the graphs through graph visualizations. We will discuss our findings in Section 4.6.

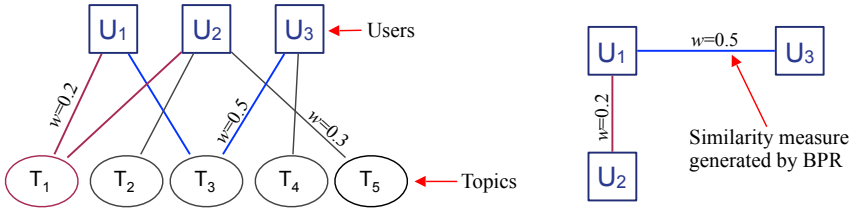
4.5 Topic Association (Exp 2)

The second experiment seeks to explore topological differences in topic-similarity networks of users that are responsible for spreading unique versus non-unique posts. The BPR method from [34] is utilized to create a user-user content similarity network for this purpose.

From the original retweet network, each user is extracted along with their 100 most recent tweets, which are aggregated into a single document. Latent Dirichlet Allocation (LDA) [9] is then performed and a document-topic matrix is produced. From this, a two-modal bipartite graph is constructed. For the *#paris* retweet network, LDA was performed with 25 topics ($K = 25$). If a user u_i's last 100 tweets contain topic t_j, an edge is drawn between i and j. Figure 9 shows that this network is connected and edges exist only between ($u_i \sim t_j$) pairs; requirements for utilization of the BPR method can be found in [34].

Thresholding To construct a unipartite graph G_{BPR}, described in Figure 8, we set the threshold τ, not establishing every edges when two users share at least one topic regardless of the weights between the users. This strategy is considered for the ease of understanding the topology of the graph and scalability of computation. For a given bipartite graph $\mathcal{G}$, let $\theta_{\mathcal{G}} \in [0, 1]$ denote the threshold of weight between the user u_i and the topic t_j such that

$$(u_i, t_j) \begin{cases} \text{exists} & \text{if } \mathsf{weight}(u_i, t_j) \geq \theta_{\mathcal{G}} \\ \text{not exists} & \text{if } \mathsf{weight}(u_i, t_j) < \theta_{\mathcal{G}} \end{cases} \tag{3}$$



(a) Bipartite Graph **(b) Unipartite Graph G_{BPR}**

Fig. 8: Bipartite graph construction using Bipartite Projection via Random Walks. Note that although there is an inherent weight assigned to edges in the (a) by the document-topic matrix, (b) is constructed using a simple binary adjacency matrix.

The BPR [34] projection method, shown in Figure 8, is performed on this user-topic content similarity network, and thus predicts edges in the user-user content similarity network. This technique accounts for the overall structure of the bipartite graph, which helps ensure that topic hubs do not saturate its unipartite projection with unlikely links.

Figure 7 shows the overall process of data processing and evaluation of our approach.

4.6 Results and Discussions

In this section, we will discuss the findings from our two experiments (`Exp 1` and `Exp 2`).

Network Analysis (`Exp 1`) Since our primary interest is how information is produced and propagated along the connections in microblogs, we study how they differ between `Group A` and `Group B` by re-constructing retweet chains from the dataset into undirected graphs and compute the graph metrics in Table 8 on these graphs. These metrics can help us gain some insight into the structure, behavior, and dynamics of the given network. Specifically, for example, we can

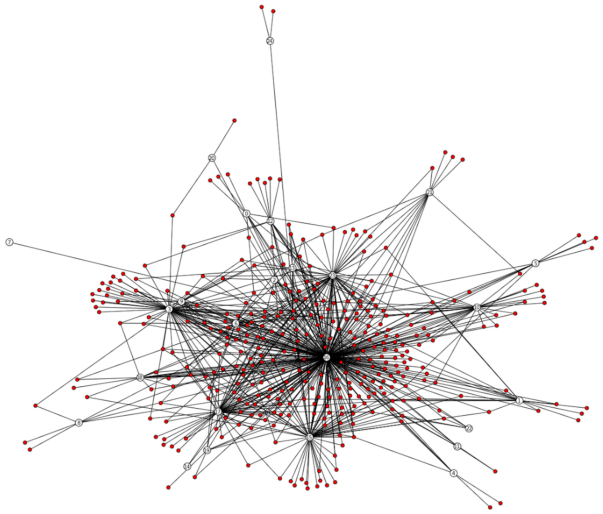


Fig. 9: *User-topic Content Similarity Network for #paris.* Red nodes represent users, and white nodes represent topics.

answer to such questions: 1) what are the dominating nodes in the propagation chain/network; 2) how densely do the nodes connected to each other; 3) can we partition this network into N different components; 4) does a giant component exist in this graph. To evaluate structural characteristic of the graphs in each group, we visualized the landscape of the entire data collection, and this is shown in Figure 10.

Figure 10 shows the network on the topic of *#paris* with a giant component surrounded by many isolated nodes and small components. In this graph, the giant component is loosely connected with many subcomponents via single or a few edges. Intuitively, we can divide the giant component into multiple clusters (or subcomponents) through these low-connectivity edges with high betweenness centrality. Intuitively, this type of structure can be sparsified into a simplified graph structure using sparsifier graph H (d-regular Ramanujan graph). According to Benczur-karger approximation model [7], we can sample low-connectivity edges (with high probability), eliminating high-connectivity edges within densely connected components.

For the comparison of `Group A` and `Group B`, we sampled 2 most representative subgraphs for each group from the dataset. Structural characteristics of each set were then analyzed through visual and computational assessments.

`Group A`*: Unique and Newsworthy Contents* Our labeling task performed on the crowdsourcing platform revealed that the participants favored unique 3rd-party news providers or quotes from celebrity accounts (e.g. *@musicnews*, *@BrianHonan*) and labeled them as niche contents. For example, the tweet "*RT*

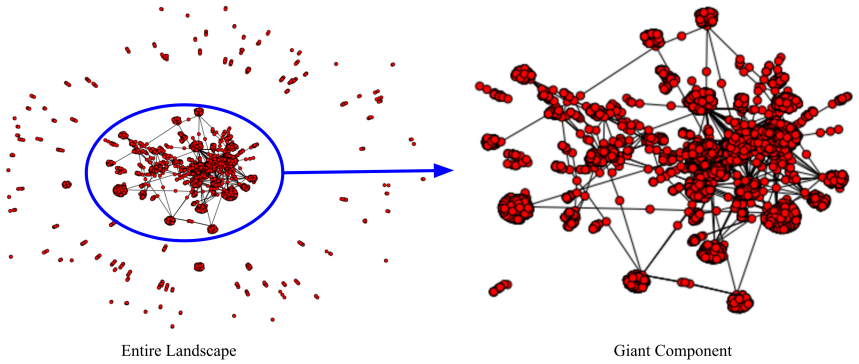


Fig. 10: The landscape of the retweet chain graph reconstructed from the dataset "#paris." One giant component and a number of small components were observed.

@BrianHonan: With the news breaking from Paris it's wise to remember this. https://t.co/bKZP5Vh46n" was rated as highly newsworthy and unique (in other words, less likely to be seen in or covered by traditional news outlets.) Interestingly, many tweets that contain both personal opinion with sentiment and a short news headline (sometimes with a url that directs users to an external source of information) within a tweet received high newsworthy and uniqueness score.

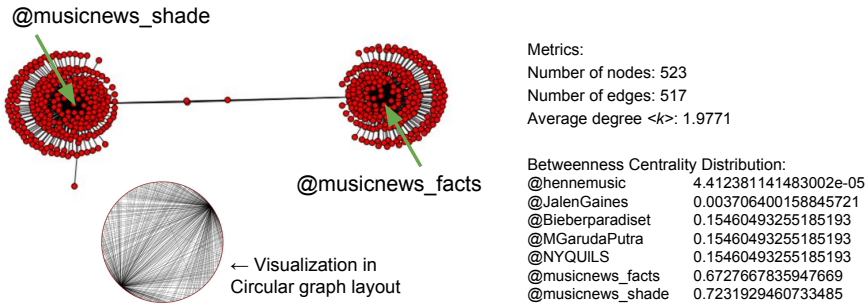
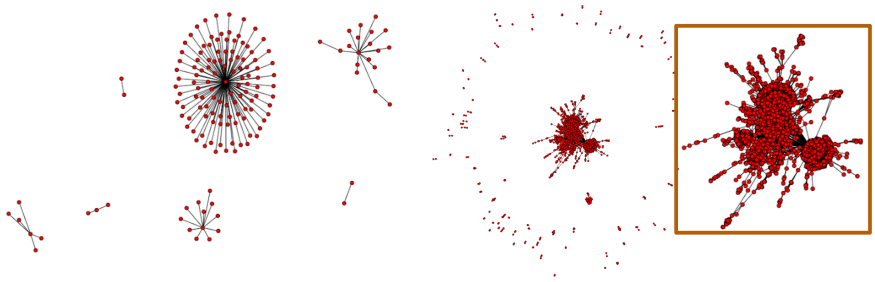


Fig. 11: An example of dumbbell type graph found in *#paris* dataset.

`Group B`*: Traditional News* Most tweets that fall into this category are, expectedly, news headlines or blurbs provided by major news providers or other institutional accounts. Most of the graphs in `Group B` has long retweet chain that either spans across the comparatively big component or connects two neighboring components. In some cases (an example is shown in Figure 11) one or two

nodes exist(s) that bridges two small components in similar size, constructing a dumbbell-shaped graph. An example might be where the New York Times tweets about an event to its many followers, one of which is CNN News, who then retweets to its many followers. Another example of this effect that occurred in the crawled data about the Paris terrorism event involved a popular Dutch journalist who re-tweeted false information about the lights in the Eiffel Tower being turned off as a mark of respect for the victims. This created a dumbbell shaped graph between the Dutch and French communities, that also happened to contain misinformation, since the lights were actually turned off as a matter of routine.



(a) `Group A` (b) `Group B`

Fig. 12: Graph visualization of the unique news group `Group A` and traditional news group `Group B`

Group	# Nodes	# Edges	$< k >$	C
@globalnews (`Grp A`)	159	151	1.8994	∼0.0
@CNN (`Grp B`)	5,245	6,563	2.5026	0.045

Table 9: Basic metrics computed for the representative subgraphs (retweet chain graphs) of group `A` and `B`. ($< k >$: mean degree, C: mean clustering coefficient)

Topic Association (`Exp 2`) The user-user content similarity network generated by the BPR is the network of interest. Figure 14 shows this network for the *#paris* example. For each projection, many possible networks can be formed based on the threshold of similarity τ between users needed to form an edge between them. Continuing the *#paris* example, the power law is reflected in Figure 15, which plots the number of edges in the user-user network versus the

similarity threshold used to form that specific network. This relationship seems to fit a power-law distribution, which would suggest that the BPR method has successfully captured scale-free decay in the number of similarities as the similarity threshold increases. Without any threshold, the giant component does not in fact grow to the entire network; the network remains unconnected. Notably, the unconnected nodes in the user-user network have an average degree of only 1.03 in the user-topic network, which explains why BPR did not predict any edges for these users.

Additionally, some user-user content similarity networks that were generated for #paris are suspected of exhibiting a power-law degree distribution themselves; an example of which is shown in Figure 16. To corroborate this claim we will investigate further into the degree distributions of these networks as a future work.

To fully utilize the power of these user-user similarity networks in comparing unique versus non-unique content spread, the same process was carried out on a set of sampled retweet networks of the groups `A` and `B`.

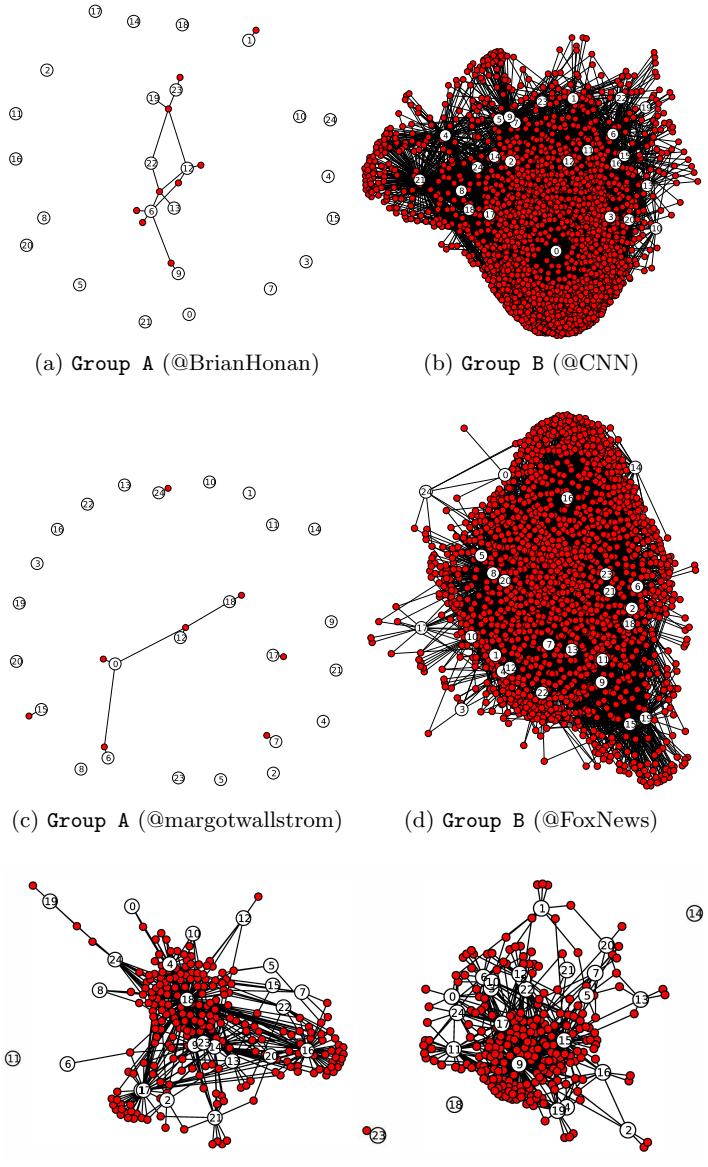
It is suspected that user-user content similarity will differ between users that spread non-unique (`Group B`) posts versus users that spread unique (`Group A`) posts, as these group's corresponding retweet-chain network structures are different. Also, comparing outlying users (users that become unconnected in user-user similarity networks) to those in the giant component of the opposite group could help provide insight to any overlap in users that spread content from both groups.

Group	N	M	$\#CComp$	$< k >$	C	Cen_B	Cen_C	Cen_E
A-BrianHonan	9	8	4	0.889	0.367	0.7	1.0	0.545
A-musicnews_facts	228	6895	2	30.241	0.573	0.033	0.766	0.189
A-margotwallstrom	8	1	7	0.125	0.0	0.0	1.0	0.707
B-CNN	1743	375	1647	0.215	0.029	0.554	0.521	0.287
B-NBCNews	226	7934	7	35.106	0.551	0.025	0.777	0.181
B-FoxNews	1565	226	1494	0.144	0.029	0.382	0.769	0.322

Table 10: Network metrics computed for the bipartite (topic-user) graphs of group `A` and `B`. Please note that all centrality metrics are computed on the max centrality nodes in the main connected component (#CComp: number of connected components. For other symbols, see Table 8).

5 Discussion and Future Work

In this study, a few challenges have been discussed in order to achieve our final goal: developing a reliable and automated detection algorithm for unique news content on microblogs. For future work, we will apply the salient features that



(a) `Group A` (@BrianHonan) (b) `Group B` (@CNN)

(c) `Group A` (@margotwallstrom) (d) `Group B` (@FoxNews)

(e) `Group A` (@musicnews_facts) (f) `Group B` (@RasmusTantholdt)

Fig. 13: Bipartite graphs of user-topic association network. Please note that LDA topic nodes are labeled with index numbers (from 1 to K; K=25). Please note that (e) and (f) are the examples of crossover accounts.

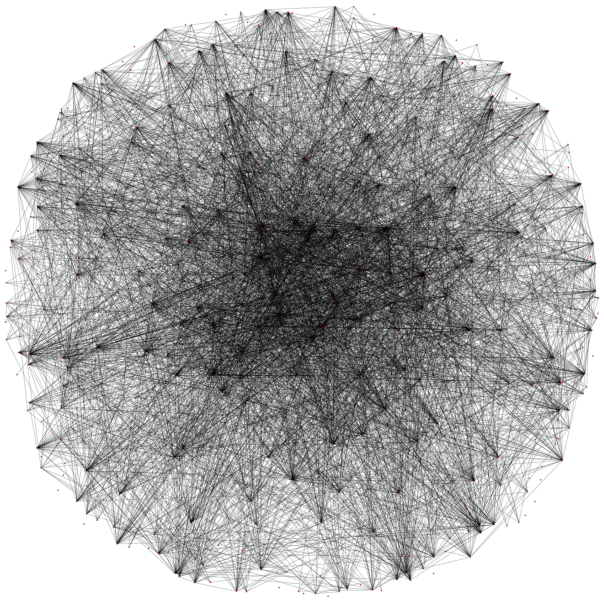


Fig. 14: *User-user Content Similarity Network for #paris.* This specific network was constructed using a similarity-threshold of 0.00001 (4295 edges).

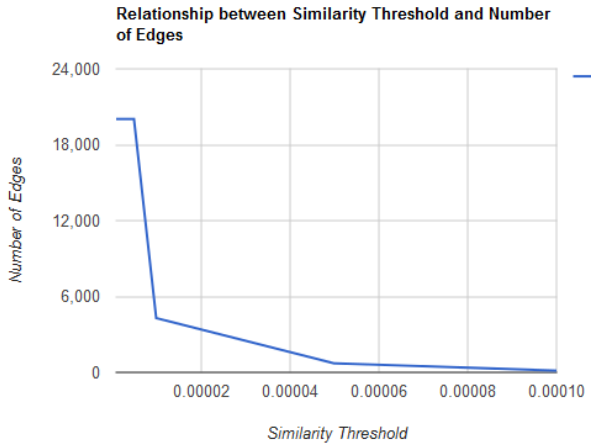
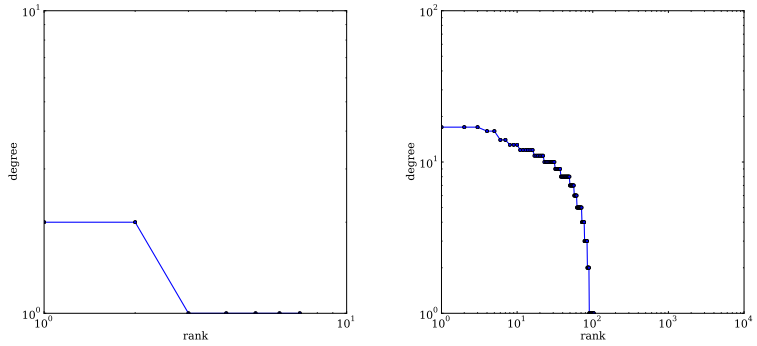
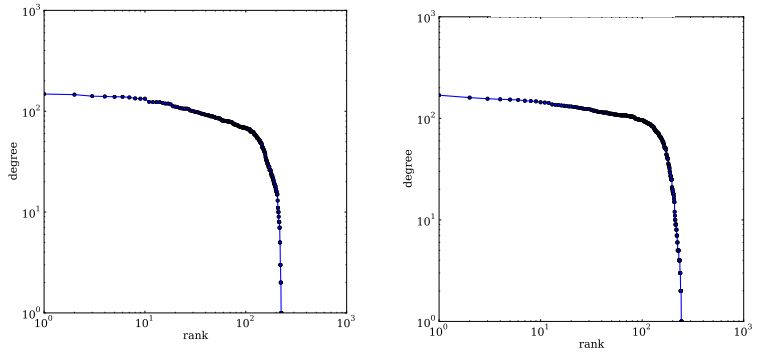


Fig. 15: Plot of similarity threshold versus number of edges generated in user-user content similarity network using the threshold τ. Calculated power-law constants using $\tau \times 1000$: alpha = -1.113, B = 20.358.



(a) `Group A` (@BrianHonan) (b) `Group B` (@CNN)



(c) `Group A` (@musicnews_facts) (d) `Group B` (@RasmusTantholdt)

Fig. 16: Distributions of topic associations of users in group `A` and `B`. X-axis shows users in rank order (log scale) and Y-axis shows number of topic associations, also on a log scale.

we found in this study to different machine learning algorithms and find an effective way to automatically locate niche microblog contents. Moreover, a temporal analysis will be performed on retweet-chain graphs in order to reveal differences in network dynamics between the groups. Any temporal patterns, found by the analysis, may allow online learning algorithms to predict niche content across time. Specifically, by investigating multiple snapshots of each network, we can measure the temporal differences and compute related metrics over the course of development of each network. In this type of analysis, tensor and different decomposition methods such as high order SVD, PARAFAC/CANDECOMP (CP) decompositions can be applied to find out multidimensional characteristics of the given network. When we incorporate the best features into an automated algorithm, however, the algorithm might need to be optimized requiring occasional user feedback due to the ambiguity and subjectivity of newsworthiness. One of the key challenges for this work is the problem of entity detection for news. This study is limited in the sense that we focus on simple messages and compute inherent newsworthiness and similarity scores. This raises two challenges for future research. First, is a problem of understanding word meaning at the right level. Consider the classic problem of a Jaguar being either a cat or a car, for example. User-provided hashtags can help us to disambiguate, however these can lack granularity in some cases. Another challenge is diversity in spelling and form, such as 'Ph.D', 'PHD' 'ph.d' for example. Approaches such as Bostandjiev et als study on LinkedVis [10], use entity resolution methods to resolve these problems. For example, by resolving every mention of PhD to the Wikipedia page for PhD. Such content-based approaches could help for some news event detection and disambiguation, but of course would not guarantee a comprehensive solution to these difficult problems.

5.1 Scalability and Real-Time News Gathering

Many information filtering methods that rely on underlying similarity models, such as automated collaborative filtering, for example, are, or have components that are computationally complex. In contrast to these approaches, which typically employ a modeling phase with complexity of $O(n^2)$ on the number of users, our method for computing inherent newsworthiness can theoretically be applied as a linear time function of number of terms in the message, and further, with efficient indexing structures for n-gram access in the news corpus (an offline process, which, lets say using a quicksort algorithm, would run in $O(nlog(n))$ time), results in a linear time search that is a function of the number of n-grams per message. This allows for a highly scalable, real-time news search experience for the end user. However, aside from the theoretical analysis, there are practical rate limitations on the Twitter end-points that would require commercial agreements to achieve the full potential of our approach.

6 Conclusion

This paper evaluated novel approaches for automatic detection of unique and newsworthy content in microblogs, using a comparative analysis between a corpus of curated news articles from traditional media and collections of "uncurated" microblog posts. Our initial approach examined differences in content similarity between the two. 24 combinations of simple NLP techniques were evaluated to optimize a similarity score between a short Twitter post and a corpus of news articles about a target topic. Next, a user study was described that gathered human annotations of newsworthiness for use as ground truth to evaluate our filtering method. Results showed general agreement between predicted scores from our approach and the human annotations.

We extend our news detection method to include information about the underlying network and dynamics of the information flow within it. LDA and BPR algorithms were used to explore structural and functional network metrics for the purpose of predicting newsworthiness and uniqueness of content. Primarily, we have studied the structure of various subgraphs underlying multiple topic-specific collections of microblog posts. Moreover, we have proposed a method to explore the topical association between different nodes in a graph, i.e. the vertices that tend to belong to either unique or traditional news groups. The results of our empirical analysis show that structural differences are observed between the unique and traditional news groups in microblogs. For example, the majority of subgraphs in the traditional group have long retweet chains and exhibit a giant component surrounded by a number of small components, unique contents typically propagate from a dominating node with only a few multi-hop retweet chains observed. Furthermore, results from LDA and BPR algorithms indicate that strong and dense topic associations between users are frequently observed in the graphs of the traditional group, but not in the unique group.

Acknowledgement

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

2. L. Allison and T. I. Dix. A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23(5):305–310, 1986.
3. A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Effects of user similarity in social media. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 703–712, New York, NY, USA, 2012. ACM.
4. P. André, M. Bernstein, and K. Luther. Who gives a tweet?: Evaluating microblog content value. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 471–474, New York, NY, USA, 2012. ACM.
5. D. Bär, C. Biemann, I. Gurevych, and T. Zesch. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 435–440, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
6. H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 291–300, New York, NY, USA, 2010. ACM.
7. A. A. Benczúr and D. R. Karger. Approximating st minimum cuts in õ (n 2) time. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 47–55. ACM, 1996.
8. C. Biemann. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122, 2013.
9. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
10. S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 35–42, New York, NY, USA, 2012. ACM.
11. C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. In *Proceedings of the Nineth International Conference on Weblogs and Social Media, Oxford, UK*. AAAI, 2015.
12. K. Canini, B. Suh, and P. Pirolli. Finding credible information sources in social networks based on content and social structure. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 1–8, Oct 2011.
13. C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
14. W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1173–1182, New York, NY, USA, 2012. ACM.
15. D. Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press, 1997.
16. I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel. Same places, same things, same people?: mining user similarity on social media. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 41–50. ACM, 2010.
17. A. Herdağdelen. Twitter n-gram corpus with demographic metadata. *Language resources and evaluation*, 47(4):1127–1147, 2013.
18. A. Hermida, F. Fletcher, D. Korell, and D. Logan. Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6):815–824, 2012.

19. J. Holcomb, J. Gottfried, and A. Mitchell. News use across social media platforms, 2013.
20. T. J. Horan. softversus hardnews on microblogging networks: Semantic analysis of twitter produsage. *Information, Communication & Society*, 16(1):43–60, 2013.
21. J. Hurlock and M. Wilson. Searching twitter: Separating the tweet from the chaff, 2011.
22. B. Kang, T. Höllerer, and J. O'Donovan. The full story: Automatic detection of unique news content in microblogs. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, pages 1192–1199, 2015.
23. P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.
24. A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei. Detecting comments on news articles in microblogs, 2013.
25. E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg!, 2011.
26. R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.
27. M. Nagarajan, H. Purohit, and A. P. Sheth. A qualitative examination of topical tweet and retweet practices. *ICWSM*, 2(010), 2010.
28. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.
29. P. J. Shoemaker. News and newsworthiness: A commentary. *Communications*, 31(1):105–111, 2006.
30. S. Sikdar, B. Kang, J. O', T. Hllerer, and S. Adalı. Understanding information credibility on twitter. In *IEEE/ASE SocialCom*, pages 19–24, 2013.
31. E. Stamatatos. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527, 2011.
32. L. Willnat and D. H. Weaver. The american journalist in the digital age. Technical report, School of Journalism, Indiana University, 2014.
33. J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM*, 10:355–358, 2010.
34. M. A. Yildirim and M. Coscia. Using random walks to generate associations between objects. *PLoS ONE*, 9(8), 2014.
35. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.