The Full Story: Automatic detection of unique news content in Microblogs

Byungkyu Kang
Department of Computer Science
University of California
Santa Barbara, CA 93106
bkang@cs.ucsb.edu

Tobias Höllerer
Department of Computer Science
University of California
Santa Barbara, CA 93106
holl@cs.ucsb.edu

John O'Donovan
Department of Computer Science
University of California
Santa Barbara, CA 93106
jod@cs.ucsb.edu

***Abstract*—In recent years a large portion of news dissemination has shifted from traditional outlets to individual users on platforms such as Twitter and Facebook. Accordingly, methods for detecting newsworthy and otherwise useful information on these platforms have received a lot of research attention. In this paper, we present a novel algorithm to automatically capture core differences in newsworthy content between microblog and traditional news media streams and discuss why it is difficult to capture such information using traditional text-based search mechanisms. We describe an experiment to tune and evaluate the algorithm using a corpus of 35 million Twitter messages and 6,112 New York Times articles on a variety of topics. Finally, we describe an online user study (N=200) to evaluate user perceptions of content recommended by our algorithm. Results show significant differences in user perception of newsworthiness and uniqueness of content from our algorithm.**

I. Introduction

In recent years, microblogs have evolved from an online communication channel for personal conversation to an information hub that curates and widely disseminates a large variety of information. Recent studies revealed that most of today's internet users rely on microblogging platforms such as Twitter and Reddit [1] as a primary source of news information, thus highlighting the need for automated tools that can identify reliable and useful information quickly

Going beyond typical information consumers, professional journalists also admit to relying heavily on social media streams for their news stories [2], [3]. 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 [1]. 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 [4] 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 the 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 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 35 million Twitter posts and 6,112 New York Times articles and attempt to answer the following research questions:

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?

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 [5].

The second experiment samples data in various ways from the highlighted areas of Figure 1 for a range of topics and presents an AB 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

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

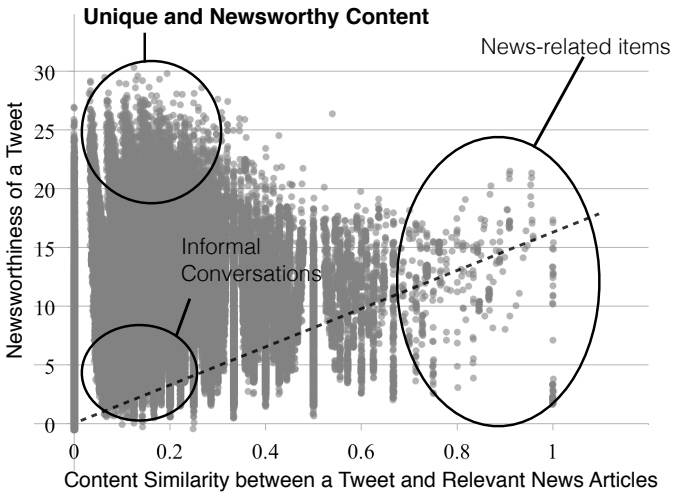


Fig. 1: Overview of approach to filtering unique and newsworthy content. Y-axis tweet newsworthiness is computed from NLTK and from Human Evaluation. X-axis is tweet similarity to mainstream news.

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.

II. 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 [6], [7], [8]. 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. [2] conducted a large-scale online survey and unveiled behaviors of news consumers on social media including microblogs. According to many studies, including [2], microblogs such as Twitter have become a major source of news information for individual consumers and also for professional journalists who rely on they dynamic content for story-hunting and marketing.

A. 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 online 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 [8], [7]. Furthermore, since microblog users not only reproduce and forward original information but sometimes reshape 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 [9] on different types of news contents: soft and hard news.

B. Newsworthiness

Shoemaker [10] 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. [11] 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 [12]. 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 [13] community question/answering platform.

Our research examines several low-level features of microblog posts to arrive at a good classifier. Castillo et al. [14] also explored features that 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.

C. 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], [16] to semantic similarity [17], structural similarity such as stop word n-grams [18] and text expansion mechanisms [19], [20]. In particular, in the context of microblog content analysis, Herdağdelen [21] proposed n-gram based approach to Twitter messages, which we build on in this research.

Our methods apply several content similarity metrics including normalized word n-grams to determine and measure

TABLE I: 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

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 [22]. Bar et al. [22] evaluate different content similarity metrics and report effectiveness and efficiency of the composite of multiple metrics using supervised machine learning approach in their study.

III. 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). Through these 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 I. Before the crawling stage, we selected 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.

IV. Approach

This section describes our approach to filtering unique and newsworthy content from microblog streams based on comparison with mainstream media APIs. Shoemaker [10] argues that newsworthiness is not the only attribute which represents news. However, since it is an important indicator for news contents in general, we assume here that curated news articles are newsworthy. Our 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.

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

[3]Dataset available upon email request

TABLE II: The set of selected metrics analyzed in this study.

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

A. Similarity Computation

A key challenge in this research 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 IV 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.

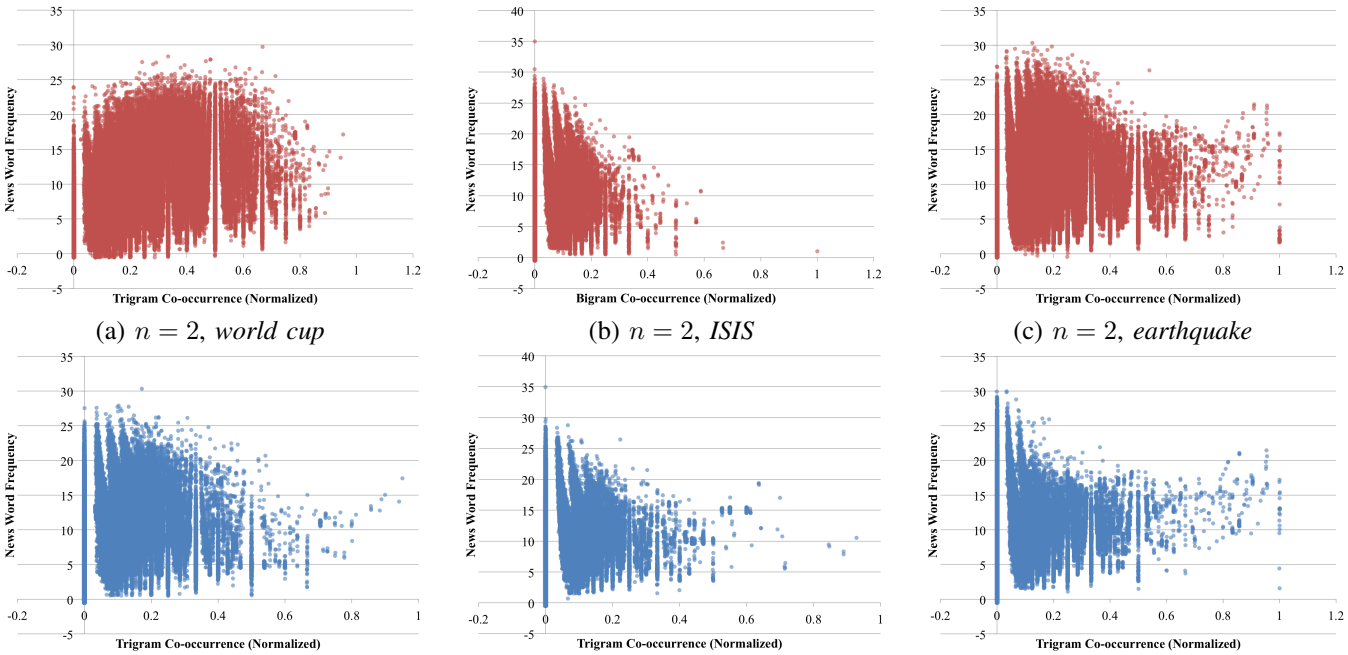
Next, a set of word n-grams as described in [22] were computed, varying n from 1 to 3. Part-of-Speech (POS) tagging was applied to identify potentially useful noun, verb, pronoun and adjective terms. A standard stop-word list was identified and systematically removed as shown in Table IV. A Twitter-specific 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 5 topic-specific collections of twitter posts and NYT news articles. These are detailed in Table IV. Each method computed a co-occurrence score between a *single* microblog post and a larger collection of news articles.

Word n-grams For each event, 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 compute the number of co-occurrences of n-grams from each post with those in the news n-gram corpus. To account for length deviation, this score was normalized by the total number of n-grams in each tweet. As shown in Figure 1, we apply a two dimensional approach to newsworthiness. First, we evaluate similarity of a message to the NYT corpus using the methods in Table IV. Second, and critically, to identfy newsworthy posts that are *not similar* to the news corpus, we apply a content-based newsworthiness score. This is computed using news word frequency from the Reuters news vocabulary corpus[4]. The relationships between n-gram co-occurrence and news word frequency of microblog message across different topics are shown in Figures 1 and 2.

B. Best Feature Selection

We compute different correlation coefficients based on the aforementioned metrics and newsworthiness score annotated by real-world microblog users on individual messages. We also applied composite sets of multiple metrics to model

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



(a) $n = 2$, *world cup* (b) $n = 2$, *ISIS* (c) $n = 2$, *earthquake*

(a) $n = 3$, *world cup* (b) $n = 3$, *ISIS* (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.

correlation between microblog contents and news articles, which is discussed in detail later in this section. Afterwards, we explain our evaluation method and procedure in Section VI

Please note that we have two newsworthiness measures: (1) news word frequency in each tweet ($News_{Reuters}$) and (2) newsworthiness score labeled by real-world microblog users ($News_{User}$) in [0-5] Likert scale.

For $News_{Reuters}$, we compute number of tokens that contain news words using the Reuters news word corpus in NLTK and divide this number by total number of tokens after tokenization of each tweet message.

$News_{User}$ 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 II shows the selected set of similarity metrics that we employ in this study.

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 $m \in T$. Mean correlation coefficient over the entire message collection is represented as $Corr_{Mean}$, whereas correlation coefficient of an individual message (tweet) is $Corr(m)$.

$$Corr(m_i) = \frac{1}{|News(m) - Score_{n-gram}(m)| + 1} \quad (1)$$

Where $News$ is:

$$News(m) = \frac{News_{Reuters}(m) + News_{User}(m)}{2} \quad (2)$$

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

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

Thus, mean of the correlation coefficients between newsworthiness and n-gram co-occurrence score is as follows:

$$Corr_{Mean} = \frac{\sum_{m=1}^{N} \frac{1}{|News(m) - Score_{n-gram}(m)| + 1}}{N} \quad (3)$$

Since we apply a fractional function to the composite correlation metric in Equation 3, intuitively, we maximize gain in highly correlated messages and, likewise, penalize un-correlated messages between $News(m)$ and $Score_{n-gram}(m)$. As briefly mentioned earlier in this section, we believe that both $News_{Reuters}$ and $News_{User}$ represent different aspects of newsworthiness. Unlike the n-gram co-occurrence ($Score_{n-gram}$), which reflects the word-based association on a specific-event, $News_{Reuters}$, which is corpus-based news word frequency, represents topic-independent association between the given data sets. To validate our correlation metric, we performed Pearson and Spearman correlation tests and they are shown in Table III.

As shown in Table IV, *unigram with stemmer only* feature

Avg # Terms in News	Avg # Terms in Tweets	# Co-occurence	# Co-occurence (Normalized)	Stopword Removal	Stemming	Noun Only (POS-tag)	n-gram	Correlation with GT
3,863	17.952	10.509	0.561	N	N	N	1	0.774
12,085	16.965	1.713	0.093	N	N	N	2	0.814
15,246	16.011	0.162	0.009	N	N	N	3	0.689
1,719	7.401	2.678	0.336	N	N	Y	1	0.777
4,596	6.532	0.144	0.018	N	N	Y	2	0.75
5,792	5.762	0.014	0.002	N	N	Y	3	0.714
1,596	17.952	3.868	0.211	N	Y	N	1	**0.96**
4,592	16.965	0.165	0.009	N	Y	N	2	0.758
5,790	16.011	0.006	0.0	N	Y	N	3	0.740
1,564	7.401	2.654	0.333	N	Y	Y	1	0.736
4,509	6.532	0.145	0.018	N	Y	Y	2	0.8
5,678	5.762	0.014	0.002	N	Y	Y	3	0.769
1,557	11.161	1.744	0.146	Y	N	N	1	0.857
4,495	10.171	0.068	0.006	Y	N	N	2	0.714
5,664	9.251	0.006	0.0	Y	N	N	3	0.666
1,557	6.217	1.473	0.216	Y	N	Y	1	0.857
4,495	5.345	0.057	0.008	Y	N	Y	2	0.8
5,664	4.611	0.007	0.001	Y	N	Y	3	0.666
1,557	11.161	2.949	0.25	Y	Y	N	1	0.857
4,495	10.171	0.163	0.015	Y	Y	N	2	0.833
5,664	9.251	0.013	0.001	Y	Y	N	3	0.8
1,557	6.217	1.99	0.293	Y	Y	Y	1	0.857
4,495	5.345	0.136	0.021	Y	Y	Y	2	0.8
5,664	4.611	0.015	0.002	Y	Y	Y	3	0.666

TABLE IV: [n-gram table] Overview 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.

has the highest correlation. Therefore, we select this feature for our user experiment and evaluation.

V. 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. In Section IV, we 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.

A. 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 during this task. For the NYT articles, however, we aggregate them together first before we compute similarity features.

B. 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 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 VI) 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 VI.

C. 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.

VI. 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 IV, 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).

A. 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.

B. 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 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

[5]www.qualtrics.com

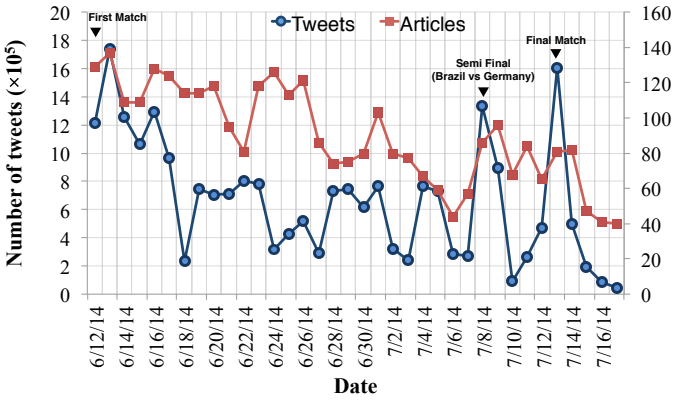


Fig. 3: Temporal distribution of the microblog messages (tweets) and news articles on the topic–*worldcup*. The time period shown in this graph corresponds to the 2014 world cup held in Brazil.

	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
frankenstorm	2012	183	sandy	4330	donate	257
	corrections	8	hurricane	1117	redcross	243
	barron	7	new	423	please	198
	ken	7	help	334	everyone	161
	belson	7	york	254	today	133
	wittenberg	7	nyc	237	video	130
	retail	6	relief	203	got	130
	flegenheimer	6	power	196	due	127
	petroleum	5	victims	176	huracán	126
	estate	5	obama	171	furacão	110
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 VI: Top 10 frequent words extracted from tweets on each topic.

Topic	# of Terms in News			Avg # of n-grams in a Tweet			Avg % of Co-occurrences		
	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%
frankenstorm	2,298	7,242	8,807	18	17	16	59.4%	10.9%	2.1%
occupysandy	3,078	11,865	15,190	18	17	16	60.5%	10.3%	1.0%

TABLE V: Statistics overview across different data sets (stemming only)

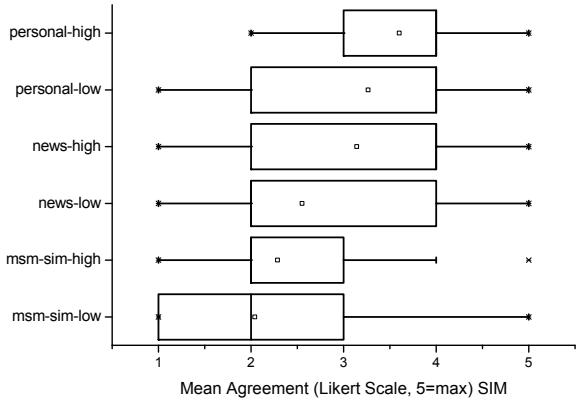


Fig. 4: Mean agreement of the responses from the user study – SIM

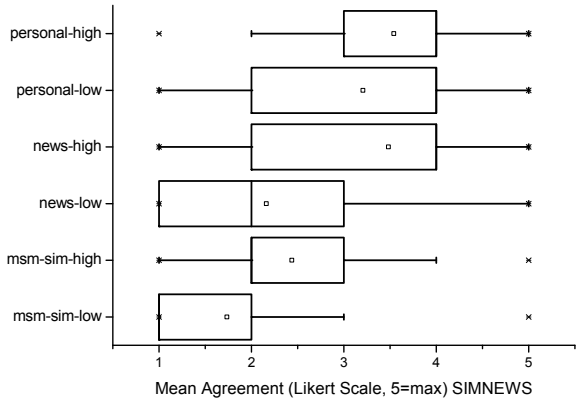


Fig. 5: Mean agreement of the responses from the user study – SIMNEWS

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 IV. 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

1) 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 IV. Two random groups of 10 tweets were sampled from the 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. Partipants 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 VI 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 VI 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 mircoblog column. This is a good example of a global news phenomenon that is covered extensively in most countries, but is largely under-reported in the United States. Now we will discuss our results in the context of the research questions presented earlier.

a): 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 IV. For example, unigram co-occurence for the topic "ISIS" shows 78% overlap witht 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.

b) : 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.

VII. Conclusion

This paper evaluated a novel approach for automatic detection of unique and newsworthy content in microblogs, based on a comparative analysis against a corpus of mainstream and curated news articles. 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. A temporal analysis of topic related posts was presented across the two domains, and a user study was described to evaluate perception of groups of messages sampled from different points on the co-occurrence distribution. Results show that a significant impact on the perception of the information consumer with respect to uniqueness of content could be achieved when the co-occurrence score was used as a filter in tandem with a content-based newsworthiness score. As a next step, topic modeling techniques, keyword and hashtag extraction will be applied to the news corpus and blog stream to gain a different, topic-based perspective on content overlap.

Acknowledgment

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] J. Holcomb, J. Gottfried, and A. Mitchell. (2013) News use across social media platforms. [Online]. Available: http://www.journalism.org/2013/11/14/news-use-across-social-media-platforms

[2] A. Hermida, F. Fletcher, D. Korell, and D. Logan, "Share, like, recommend: Decoding the social media news consumer," *Journalism Studies*, vol. 13, no. 5-6, pp. 815–824, 2012.

[3] L. Willnat and D. H. Weaver, "The american journalist in the digital age," School of Journalism, Indiana University, Tech. Rep., 2014.

[4] 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.

[5] S. Sikdar, B. Kang, J. O', T. Hllerer, and S. Adalı, "Understanding information credibility on twitter." in *SocialCom*, 2013, pp. 19–24.

[6] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting comments on news articles in microblogs," 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6011

[7] 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*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 1173–1182. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398417

[8] 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*. Springer, 2011, pp. 338–349.

[9] T. J. Horan, "softversus hardnews on microblogging networks: Semantic analysis of twitter produsage," *Information, Communication & Society*, vol. 16, no. 1, pp. 43–60, 2013.

[10] P. J. Shoemaker, "News and newsworthiness: A commentary," *Communications*, vol. 31, no. 1, pp. 105–111, 2006.

[11] 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*, ser. CSCW '12. New York, NY, USA: ACM, 2012, pp. 471–474. [Online]. Available: http://doi.acm.org/10.1145/2145204.2145277

[12] J. Hurlock and M. Wilson, "Searching twitter: Separating the tweet from the chaff," 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2819

[13] 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*. ACM, 2008, pp. 183–194.

[14] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[15] D. Gusfield, *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press, 1997.

[16] L. Allison and T. I. Dix, "A bit-string longest-common-subsequence algorithm," *Information Processing Letters*, vol. 23, no. 5, pp. 305–310, 1986.

[17] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[18] E. Stamatatos, "Plagiarism detection using stopword n-grams," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 12, pp. 2512–2527, 2011. [Online]. Available: http://dx.doi.org/10.1002/asi.21630

[19] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005, pp. 79–86.

[20] C. Biemann, "Creating a system for lexical substitutions from scratch using crowdsourcing," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 97–122, 2013.

[21] A. Herdağdelen, "Twitter n-gram corpus with demographic metadata," *Language resources and evaluation*, vol. 47, no. 4, pp. 1127–1147, 2013.

[22] 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*, ser. SemEval '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 435–440.