Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?

Md Main Uddin Rony [§], Naeemul Hassan [§], Mohammad Yousuf [♮]
[§]Department of Computer and Information Science, [♮]Gaylord College of Journalism and Mass Communication
[§]The University of Mississippi, [♮]The University of Oklahoma

***Abstract*—The use of alluring headlines (clickbait) to tempt the readers has become a growing practice nowadays. For the sake of existence in the highly competitive media industry, most of the on-line media including the mainstream ones, have started following this practice. Although the wide-spread practice of clickbait makes the reader's reliability on media vulnerable, a large scale analysis to reveal this fact is still absent. In this paper, we analyze 1.67 million Facebook posts created by 153 media organizations to understand the extent of clickbait practice, its impact and user engagement by using our own developed clickbait detection model. The model uses distributed sub-word embeddings learned from a large corpus. The accuracy of the model is 98.3%. Powered with this model, we further study the distribution of topics in clickbait and non-clickbait contents.**

I. INTRODUCTION

The term *clickbait* refers to a form of web content that employs writing formulas and linguistic techniques in headlines to trick readers into clicking links [1], [2], but does not deliver on promises [1]. Media scholars and pundits consistently show clickbait content in a bad light, but the industry based on this type of content has been rapidly growing and reaching more and more people across the world [3], [4]. *Taboola*, one of the key providers of clickbait content, claims [2] to have doubled its monthly reach from 500 million unique users to 1 billion in a single year from March 2015. The growth of clickbait industry appears to have clear impact on the media ecosystem, as many traditional media organizations have started to use clickbait techniques to attract readers and generate revenue. However, media analysts suggest that news media risk losing readers' trust and depleting brand value by using clickbait techniques that may boost advertising revenue only temporarily. According to a study performed by Facebook [3], 80% users "preferred headlines that helped them decide if they wanted to read the full article before they had to click through". [5] shows that clickbait headlines lead to negative reactions among media users.

Compared to the reach of clickbait content and its impact on the online media ecosystem, the amount of research done on this topic is very small. No large scale study has been conducted to examine the extent to which different types of media use clickbait techniques. Little is known about the extent to which clickbait headlines contribute to user engagement on social networking platforms – major distributors of web content. This study seeks to fill this gap by examining uses of clickbait techniques in headlines by mainstream and unreliable media organizations on the social network. Some of the questions we answer in this paper are– (i) to what extent, mainstream and unreliable media organizations use clickbait? (ii) does the topic distribution of the contents vary in clickbaity contents? (iii) which type of headlines – clickbait or non-clickbait - generates more user engagement (e.g., shares, comments, reactions)?

We first create a set of supervised clickbait classification models to identify clickbait headlines. Instead of following the traditional bag-of-words and hand-crafted feature set approaches, we take a more recent deep learning path that does not require feature engineering. Specifically, we use distributed subword embedding technique [6], [7] to transform the words in the corpus to 300 dimensional embeddings. These embeddings are used to map sentences to a vector space over which a softmax function is applied as a classifier. Our best performing model achieves 98.3% accuracy on a labeled dataset. We use this model to analyze a larger dataset which is a collection of approximately 1.67 million Facebook posts created during 2014–2016 by 68 mainstream media and 85 unreliable media organizations. In addition to identifying the clickbait headlines in the corpus, we also use the embeddings to measure the distance between the headline and the first paragraph, known as intro, of a news article. We use a word co-occurrence based topic model that learns topics by modeling word-word co-occurrences patterns (e.g., bi-terms) to understand the distribution of topics in the clickbait and non-clickbait contents of each media. Finally, using the data on Facebook reactions, comments, and shares, we analyzed the role clickbaits play in user engagement and information spread. The main contributions of this paper are–

•We collect a large data corpus of 1.67 million Facebook posts by over 150 U.S. based media organizations. Details of the corpus is explained in Section II. We make the corpus available to use for research purpose [4].

•We prepare distributed subword based embeddings for the words present in the corpus. In Section III, we provide a comparison between these word embeddings and the *word2vec* [8], [9] embeddings created from Google News dataset with

[1]https://www.wired.com/2015/12/psychology-of-clickbait/

[2]https://www.taboola.com/press-release/taboola-crosses-one-billion-user-mark-second-only-facebook-worlds-largest-discovery

[3]https://www.nytimes.com/2014/08/26/business/media/facebook-takes-steps-against-click-bait-articles.html

[4]URL will be added after acceptance

respect to clickbait detection. We plan to make these embeddings publicly available upon acceptance of the paper.

•We perform detailed analysis of the clickbait practice in the social network from multiple perspectives. Section IV presents qualitative, quantitative and impact analysis of clickbait and non-clickbait contents.

II. Dataset

We use two datasets in this paper. Below, we provide description of the datasets and explain the collection process.

Headlines: This dataset is curated by Chakraborty et al. [2]. It contains $32,000$ headlines of news articles which appeared in 'WikiNews', 'New York Times', 'The Guardian', 'The Hindu', 'BuzzFeed', 'Upworthy', 'ViralNova', 'Thatscoop', 'Scoopwhoop', and 'ViralStories'. [5] Each of these headlines is manually labeled either as a clickbait or a non-clickbait by at least three volunteers. There are $15,999$ clickbait headlines and $16,001$ non-clickbait headlines in this dataset. We used this labeled dataset to develop an automatic clickbait classification model (details in Section III). An earlier version of this dataset was used in [2], [10]. It had $15,000$ manually labeled headlines with an even distribution of $7,500$ clickbait and $7,500$ non-clickbait headlines.

Media Corpus: For large scale analysis, using Facebook Graph API [6], we accumulated all the Facebook posts created by a set of mainstream and unreliable media within January 1^{st}, 2014 – December 31^{st}, 2016. The mainstream set consists of the 25 most circulated print media [7] and the 43 most-watched broadcast media [8] (according to Nielson rating [11]). The unreliable set is a collection of 85 conspiracy, clickbait, satire and junk science based media organizations. The category of each unreliable media is cross-checked by two sources [12], [13]. Figure 2 shows the number of media organizations in each category in the dataset along with the percentage. Overall, we collected more than 2 million Facebook posts. A Facebook post may contain a photo or a video or a link to an external source. In this paper, we limit ourselves to the link and video type posts only. This reduces the corpus size to 1.67 million. For each post, we collect the headline (title of a video or headline of an article) and the status message. For a collection of $191,540$ link type posts, we also collected the bodies of the corresponding news articles. All these contents (headlines, messages, bodies) were used to train a domain specific word embeddings (details in Section III). We also gather the Facebook reaction (Like, Love, Haha, Wow, Sad, Angry) statistics of each post. Table I shows distribution of the corpus.

III. Clickbait Detection

The key purpose of this study is to systematically quantify the extents to which traditional print and broadcast media as well as "alternative" media – often portrayed as unreliable – use clickbait properties in contents published on the web. The first step towards that goal is to identify clickbait and non-clickbait headlines.

[5]https://github.com/bhargaviparanjape/clickbait/tree/master/dataset

[6]https://developers.facebook.com/docs/graph-api

[7]https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States

[8]www.indiewire.com/2016/12/cnn-fox-news-msnbc-nbc-ratings-2016-winners-losers-1201762864/

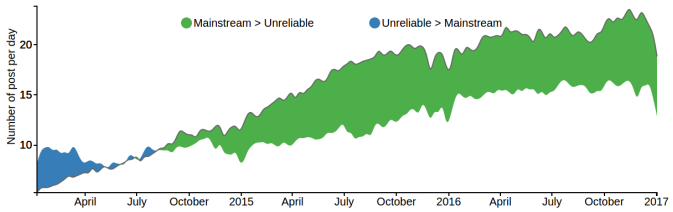

Fig. 1. This figure shows the difference between the number of posts per day from an average mainstream (print, broadcast) media and the same from an average unreliable media during January 1^{st}, 2014 – December 31^{st}, 2016. The green areas indicate that during these time periods, on average, a mainstream media posted more Facebook contents per day than an unreliable media. The blue areas indicate the opposite. General observation is, media organizations are sharing contents in the Facebook more actively now than they did earlier.

TABLE I
Distribution of the Media Corpus

Media	Category	Link	Video	Total
Mainstream	Broadcast	324028	32924	356952
	Print	516713	14129	530842
Unreliable	Clickbait	371834	4099	375933
	Conspiracy	309122	5841	314963
	Junk Science	51923	649	52572
	Satire	41046	151	41197
Total		1614666	57793	1672459

A. Problem Definition

We define the clickbait identification task as a supervised binary classification problem where the set of classes, $\mathcal{C} = \{clickbait, non\_clickbait\}$. Formally, given $\mathcal{X}$, a set of all sentences, and a training set $\mathcal{S}$ of labeled sentences $\langle s, c\rangle$, where $\langle s, c\rangle \in \mathcal{X} \times \mathcal{C}$, we want to learn a function γ such that $\gamma : \mathcal{X} \rightarrow \mathcal{C}$, in other words, it maps sentences to $\{clickbait, non\_clickbait\}$. In the following sections, we describe modeling of the problem and compare performances of multiple learning techniques.

B. Problem Modeling

In text classification, a traditional approach is to use *bag-of-words* (BOW) model to transform text into feature vectors before applying learning algorithms. [2] followed this approach and used BOW model along with a collection of hand-crafted rules to prepare the feature set. However, inspired by the recent success of deep learning methods in text classification, we use distributed subword embeddings as features instead of applying BOW model. Specifically, we use an extension of the continuous *skip-gram* model [8], which takes into account subword (substring of a word) information [6]. We call this model as Skip-Gram$_{sw}$. Below, we explain how Skip-Gram$_{sw}$is used to generate word embeddings.

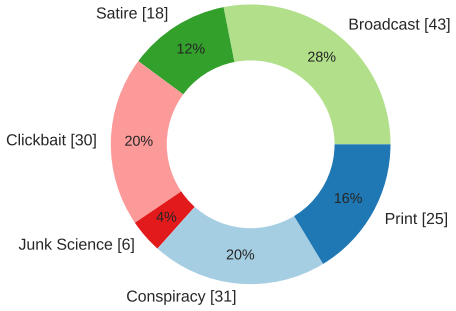

Fig. 2. Category distribution of the Media Corpus

1) Skip-Gram$_{sw}$: Given a large corpus $\mathcal{W}$, represented as a sequence of words, $\mathcal{W} = w_1, \ldots, w_T$, the objective of the skip-gram model is to maximize the log-likelihood

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c | w_t) \tag{1}$$

where the context $\mathcal{C}_t$ is the set of indices of words surrounding w_t. In other words, given a word w_t, the model wants to maximize the correct prediction of its context w_c. The probability of observing a context word w_c given w_t is parametrized using the word vectors. The output of the model is an embedding for each word which captures semantic and contextual information of the word. Skip-Gram$_{sw}$works in a slightly different way. Rather than treating each word as a unit, it breaks down words into subwords and wants to correctly predict the context subwords of a given subword. This extension allows sharing the representations across words, thus allowing to learn reliable representation for rare words. Consider the following example.

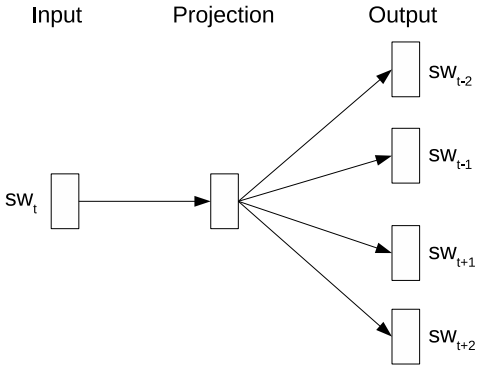

Fig. 3. The Skip-Gram$_{sw}$model architecture. The training objective is to learn subword vector representations that are good at predicting the nearby subwords.

Example 1. "the quick brown fox jumped over the lazy dog"- *take the word "quick" as an example. Assuming subword length as three, the subwords are-* $\{qui, uic, ick\}$. *Skip-Gram$_{sw}$model learns to predict* qui, ick *in the context given* uic *as the input.*

Figure 3 shows the architecture of the Skip-Gram$_{sw}$model. Using neural network, the model learns the mapping between the output and the input. The weights to the hidden layer form the vector representations of the subwords. The embedding of a word is formed by the sum of the vector representations of its subwords. Formally, given a word w and its set of subwords $\mathcal{SW}_w$, we can calculate the embedding of w using the following equation-

$$\mathbf{u}_w = \sum_{sw \in \mathcal{SW}_w} \mathbf{v}_{sw} \tag{2}$$

where $\mathbf{u}_w$ is the embedding of w and $\mathbf{v}_{sw}$ is the vector representation of sw. Further details of the Skip-Gram$_{sw}$model can be found in [6].

2) Pre-trained Vectors: Note that Skip-Gram$_{sw}$does not require $\mathcal{C}$ to learn the embeddings of words in corpus $\mathcal{W}$. It means that one can use the model on any large corpus of text to learn the word embeddings irrespective of whether the corpus is labeled or not. This technique of learning from large text corpus helps having richer word embeddings which capture a lot of semantic, conceptual and contextual information. We use the texts (headlines, messages, bodies) from Media Corpus to learn word embeddings using this model. In Section III-C, we present comparison between our pre-trained vectors and word vectors which were trained on about 100 billion words [9] from the Google News dataset.

3) Classification: For a labeled sentence $\langle s, c \rangle$, we average the embeddings of words present in s to form the hidden representation of s. These sentence representations are used to train a linear classifier. Specifically, we use the softmax function to compute the probability distribution over the classes in $\mathcal{C}$. [7] describes the classification process in detail.

C. Evaluation

We use the Headlines dataset to evaluate our classification model. Section II provides the description of the dataset. We perform 10-fold cross-validation to evaluate various methods with respect to accuracy, precision, recall, f-measure, area under the ROC curve (ROC-AUC) and Cohen's κ. Table II shows performances of the methods. To avoid randomness effect, we perform each experiment 5 times and present the average. There are in total seven methods. We categorize them based on the use of pre-trained vectors. Note that we report performances of Chakroborty et al. [2] and Anand et al. [10] in the table. We keep Anand et al. with the methods which use pre-trained vectors. Because Anand et al. used word embeddings trained on about 100 billion words from the Google News dataset using the Continuous Bag of Words architecture [9]. Each word embedding has 300 dimensions. Both of these works [2], [10] used a smaller and earlier version of the Headlines dataset. Moreover, the training and test sets of the earlier dataset are not available. So, we could not compare our methods with them using the same test bed.

The Skip-Gram$_{sw}$model, even without pre-trained vectors, significantly outperforms the BOW based Chakroborty et al. It achieves a f-measure score of 0.975 (2.5% higher than

Chakroborty et al.) and a κ score of 0.952. Powered with the pre-trained vectors, Skip-Gram$_{sw}$performed even better. We used the same word embeddings provided by [9] as well as our own Media Corpus. Regarding the later, we experimented with three combinations- pre-trained vectors learned from the content headlines only, from headlines and messages, and from headline, bodies and messages. We set embedding size to 300 dimensions while learning from these combinations. For the methods which were applied on the full Headlines dataset, we highlight the top performance in each column. Skip-Gram$_{sw}$along with pre-trained vectors from headlines, bodies and messages performed the best among all the variations. We realize that the differences of the measure values among the methods are small. However, we understand that making a small improvement while working above the 0.95 range, is significant.

Media Corpus has $477,236$ unique embeddings where Google News dataset provided 100 billion embeddings. One interesting observation is, even though the size of our Media Corpus is significantly smaller than the Google News dataset, it contributes more to the clickbait classification task. It can be rationalized as, the embeddings from Media Corpus have more domain specific knowledge than the Google News dataset. We plan to extend this corpus with more Facebook posts and release it along with the pre-trained vectors for research purpose upon acceptance of the paper.

With this powerful clickbait classification model [Skip-Gram$_{sw}$+(Headline+Body+Message)], we move forward and perform large scale study on the clickbait practice by a range of media on social network (Facebook).

IV. Practice of using clickbait in Social Network

We analyze the clickbait practice in Facebook using the Media Corpus from three perspectives.

A. Quantitative Analysis

To understand the extent of clickbait practice by different media and their categories, we applied the clickbait detection model on their contents; particularly on the headline/title of the link/video type posts. From now onward, we will use the term *headline* to denote both the headline of a link content (article) and the title of a video content. Table III shows amounts of clickbaits and non-clickbaits in the headlines of mainstream and unreliable media. Out of $887,794$ posts by mainstream media, $297,774$ (33.54%) have clickbait headlines. In unreliable media, the ratio is 39.26% $(308,095$ clickbait headlines out of $784,665)$. Based on these statistics, the percentage appears to be surprisingly high for the mainstream. We zoom into the categories of these two media to analyze the primary proponents of the clickbait practice. We find that between the two categories of mainstream media, broadcast uses clickbait 47.56% of the times whereas print only uses 24.12%. We further zoom in to understand the high percentage in the broadcast category. The Media Corpus has 43 broadcast media. We manually categorize them into news oriented broadcast media (e.g. *CNN*, *NBC*, etc.) and non-news (lifestyle, entertainment, sports, etc.) broadcast media (e.g. *HGTV*, *E!*, etc.). There are 6 news oriented broadcast media and 37 non-news broadcast media. We find that the ratio of clickbait and non-clickbait is 61.64% in non-news type broadcast media whereas it is only 22.32% (close to print media) in news oriented media. Figure 5 shows kernel density estimation of the clickbait percentage both for news and non-news broadcast media. It clearly shows the difference in clickbait practice in these two sub-categories. Most of the news type broadcast media has about 25% clickbait contents. On the other hand, the percentage of clickbait for non-news type broadcast media has a wider range with peak at about 60%. In case of unreliable media, unsurprisingly all the categories have high percentage of clickbaits in their headlines. In Figure 4, we show the percentage of clickbait in video and link type posts for each of the media categories. Satire is leading in both link and video type posts. Print and conspiracy media have the lowest clickbait practice among all the media categories in link and video type posts, respectively. Table V shows the top-5 clickbait proponents in each media category.

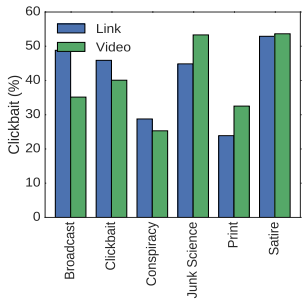

Fig. 4. Percentage of clickbaits in link and video headlines.

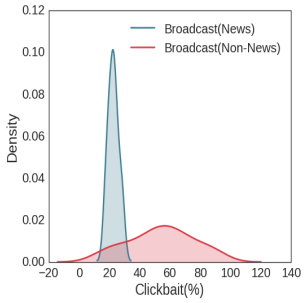

Fig. 5. Broadcast (News) vs. Broadcast (Non-news).

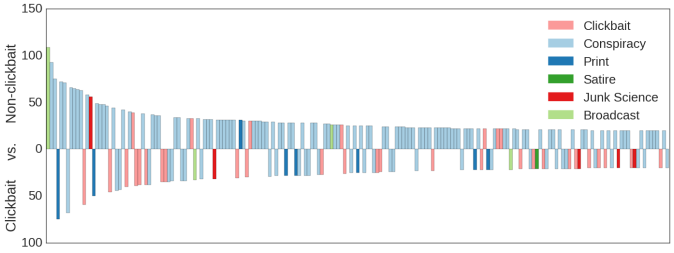

Fig. 6. Frequency of link re-post by different media.

B. Qualitative

Topic distribution: To understand the topics in the clickbait and non-clickbait contents, we applied topic modeling on all the headlines of each category. One concern about applying the traditional topic modeling algorithms (e.g. Latent Dirichlet Allocation, Latent Semantic Analysis) on our corpus is, they focus on document-level word co-occurrence patterns to discover the topics of a document. So, they may struggle with the high word co-occurrence patterns sparsity which becomes a dominant factor in case of shorter context. That is why we

TABLE II
PERFORMANCE OF THE METHODS ON THE HEADLINES DATASET

Method		Precision	Recall	F-measure	Accuracy	Cohen's κ	ROC-AUC
Without Pre-trained Vectors	*Chakroborty et al. [2]	0.95	0.90	0.93	0.93	NA	0.97
	Skip-Gram$_{sw}$	0.976	0.975	0.975	0.976	0.952	0.976
With Pre-trained Vectors	*Anand et al. [10]	0.984	0.978	0.982	0.982	NA	0.998
	Skip-Gram$_{sw}$+ Google_word2vec	0.977	0.977	0.977	0.976	0.951	0.976
	Skip-Gram$_{sw}$+ (Headline)	0.981	0.981	0.981	0.981	0.962	0.981
	Skip-Gram$_{sw}$+ (Headline + Message)	0.982	0.982	0.982	0.982	0.964	0.982
	Skip-Gram$_{sw}$+ (Headline + Body + Message)	**0.983**	**0.983**	**0.983**	**0.983**	**0.965**	**0.983**

* Their experiments were performed on a smaller and earlier version of the Headlines dataset.

TABLE III
AMOUNT OF CLICKBAITS IN VARIOUS MEDIA

Media	Category	Clickbait	Non-clickbait	Clickbait (%)
Mainstream	Broadcast	169752	187200	47.56
	Print	128022	402820	24.12
Unreliable	Clickbait	172271	203662	45.82
	Conspiracy	90389	224574	28.7
	Junk Science	23637	28935	44.96
	Satire	21798	19399	52.91

use Biterm Topic Modeling (BTM) [14] which generates the topics by directly modeling the aggregated word co-occurrence patterns of a short document.

Table IV shows 5 topics in clickbait and non-clickbait contents for each media category. Each topic is represented by a set of 10 words. The words are ordered by their significance in the corresponding topic. The modeling indicates that clickbait headlines in print and broadcast media vary in tones and subject matters from their non-clickbait headlines to a great extent. Clickbait headlines in these media represent more personalized, sensationalized and entertaining topics, while non-clickbait headlines highlight topics of collective problems such as public policies and civic affairs. But this variation is not much evident in unreliable media that use clickbait headlines indiscriminately across all topics.

The model highlights some differences in clickbait topics between print and broadcast media. Most clickbait topics in print media, four out of five, are about U.S. President Donald Trumps views on women. Each of these four topics include all of these four words: Trump, woman, make, new. A manual search shows that print news media often used clickbait techniques (e.g., question based headline) in stories about Trump and women. For instance, "Did Donald Trump really say those things?" was the headline of a Washington Post article dated July 25, 2016. The headline of a New York Times story from May 14, 2016, reads; "Crossing the Line: How Donald Trump Behaved With Women in Private."

Most clickbait topics in broadcast media are about entertainment (e.g., Taylor Swifts new music video; Kardashians new baby) and lifestyle (e.g., food and health). Two topics appeared to touch Donald Trump and his opponent Hillary Clinton. Clickbait topics in unreliable media, however, range from politics to lifestyle. At least three topics appeared to be about politics in which key words include, Trump, Hillary, Obama, Muslim, Cop, and Woman. One topic is about food and health while another is unclear.

Non-clickbait topics remain similar across all three media types, which primarily focus on law and order, and U.S. presidential election campaign. Twelve out of 15 topics – all five in print, three in broadcast, and four in unreliable – are about these two areas. One broadcast topic appears to be about sports and one is unclear. One unreliable topic is about food and health.

Headline-Body similarity: One limitation of Skip-Gram$_{sw}$is, it only considers the headline to determine whether it is a clickbait or not. The body of the news, is not considered as a factor in defining the headline. An attractive headline can be highly relevant to the content/body of a news or it can be very loosely related to the news. Our model is not capable of making the distinction. A metric is required to measure the similarity between the headline and the content to determine if the headline fairly represents the content. In future, we want to systematically incorporate the headline-body similarity in defining the clickbaitiness. Nonetheless, here we measure how similar the clickbait and non-clickbait headlines are to the corresponding bodies using a simple approach. We assume that the first para of an article represents the summary of the whole news [15] and use cosine similarity to measure the similarity between the headline and the sentences in the first para. We use bag-of-words model to transform the sentences into vectors before applying cosine similarity. In future, we plan to use our word embeddings to create the vectors instead. Figure 7 shows the kernel density estimation of the headline-body similarity in clickbait and non-clickbait contents posted by different media. One observation is, in print media non-clickbait headlines are closer to their summary than clickbait headlines. In broadcast media, the difference is less clear and in unreliable media the difference is almost absent.

C. Impact

To measure the reachability and user engagement of clickbait and non-clickbait contents, we use Facebook reactions, comments and shares as metrices. Figure 8 shows number of comments, shares and reactions (summation of like, haha, wow, sad, angry, happy, love) of an average clickbait and non-clickbait post in each media category. Blue areas indicate that

TABLE IV
TOPIC MODEL OF CLICKBAIT AND NON-CLICKBAIT HEADLINES IN DIFFERENT MEDIA

Media	Clickbait	Non-Clickbait
Print	T_1: best, thing, day, new, 2015, cleveland, la, 2016, know, year T_2: trump, woman, donald, new, get, say, make, people, thing, know T_3: trump, new, get, woman, donald, make, star, say, man, chicago T_4: new, best, thing, year, get, kid, day, woman, make, trump T_5: boston, trump, donald, new, say, make, clinton, woman, get, 2016	T_1: new, san, la, jose, police, county, vega, get, bay, school T_2: police, man, cleveland, new, killed, woman, la, shooting, shot, get T_3: news, trump, new, man, say, york, woman, hawaii, police, killed T_4: trump, new, u, clinton, say, state, win, donald, take, world T_5: boston, new, say, trump, sox, chronicle, win, red, get, state
Broadcast	T_1: new, movie, star, make, swift, time, video, best, get, like T_2: new, get, baby, kardashian, jenner, star, first, make, love, say T_3: woman, episode, new, trump, man, black, get, video, full, girl T_4: trump, history, know, thing, donald, clinton, get, make, best, say T_5: day, photo, national, way, best, like, food, dog, thing, geographic	T_1: police, man, new, found, woman, killed, arrested, say, shooting, death T_2: trump, clinton, say, new, obama, u, gop, news, campaign, hillary, T_3: new, u, say, police, found, killed, dead, nbc, year, dy T_4: win, new, say, game, first, get, team, player, take, back T_5: national, geographic, photo, new, shark, day, classic, fs1undisputed, home, found
Unreliable	T_1: trump, hillary, donald, clinton, obama, get, make, say, one, news T_2: video, people, american, black, obama, muslim, u, america, cop, white T_3: chick, trump, eagle, right, woman, hillary, say, get, people, make T_4: man, people, thing, woman, make, year, like, get, way, new T_5: day, reunionfather, human, food, way, health, thing, reason, life, make	T_1: obama, eagle, muslim, police, say, gun, u, cop, man, patriot T_2: trump, hillary, clinton, obama, new, say, campaign, news, donald, republican T_3: u, obama, video, war, isi, new, military, american, world, muslim T_4: new, truth, obama, say, u, republican, police, broadcast, man, american T_5: human, cancer, health, new, vaccine, u, study, food, found, world

TABLE V
TOP-5 CLICKBAIT PROPONENTS IN EACH MEDIA

Media	Name	Clickbait	Non-clickbait	Clickbait (%)
Overall	VH1	13760	1339	91.13
	AmplifyingGlass	692	71	90.69
	MTV	42313	4492	90.4
	ClickHole	8250	930	89.87
	Reductress	3984	484	89.17
Broadcast	VH1	13760	1339	91.13
	MTV	42313	4492	90.4
	Bravo TV	8263	1242	86.93
	Food Network	2990	492	85.87
	OWN	474	118	80.07
Print	Washington Post	13905	15158	47.84
	New York Post	11977	13910	46.27
	Dallas Morning News	3982	8232	32.6
	USA Today	8538	20282	29.63
	Houston Chronicle	8481	21618	28.18
Unreliable	AmplifyingGlass	692	71	90.69
	ClickHole	8250	930	89.87
	Reductress	3984	484	89.17
	Food Babe	2387	638	78.91
	Chicks on the Right	14185	4977	74.03

on average, a clickbait post (link or video) receives more attention (reactions/shares/comments) than a non-clickbait post. Green areas indicate the opposite. Clickbait contents receive more attention and reach to more users in general. One exception is the broadcast media.

We also analyze how often a news article is re-posted in Facebook. Figure 6 shows number of times a link is re-posted by a media. Each bar represents a news link. The height indicates how many times this link was posted in Facebook by the colored media category. We only consider the links

TABLE VI
PRESENCE OF CLICKBAIT IN THE STATUS

Media	Category	Clickbait Status	Non-clickbait Link	Clickbait Status (%)
Mainstream	Broadcast	84192	176177	32.34
	Print	164669	379504	30.26
Unreliable	Clickbait	91747	157886	36.75
	Conspiracy	46851	190477	19.74
	Junk Science	12764	28349	31.05
	Satire	7425	14453	33.94

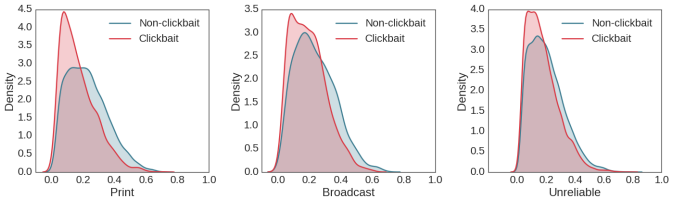

Fig. 7. Headline-Body similarity in clickbait and non-clickbait contents.

which were re-posted at least 20 times. Compare to others, conspiracy media organizations repeat the same link more. This is observed both for clickbait and non-clickbait. Clickbait media seem to repeatedly posting same clickbait links more than others.

Other than headlines, the media organizations also practice using clickbait in the Facebook status message itself. Table VI shows use of clickbait status for non-clickbait articles by different media. A general observation is, the practice is there to allure the readers by giving clickbaity message posts even for non-clickbaity news contents. Unsurprisingly, the clickbait media category is leading in this practice.

V. RELATED WORK

Even though *clickbait* is a relatively nascent term, its traces can be found in several journalistic concepts such as *tabloidization* and *content trivialization*. The linguistic techniques and presentation styles, employed typically in clickbait headlines and articles, derived from the tabloid press that baits readers with sensational language and appealing topics such as celebrity gossip, humor, fear and sex [1]. Clickbait articles are also similar to tabloid press articles in terms of story focus, which puts emphasis on the entertaining elements of an event rather than the informative elements. The Internet and especially the social media have made it easier for the clickbait practitioners to create, publish in a larger scale and reach to a broader audience with a higher speed than before [16]. In the last several years, academicians and media studied this phenomenon from several perspectives.

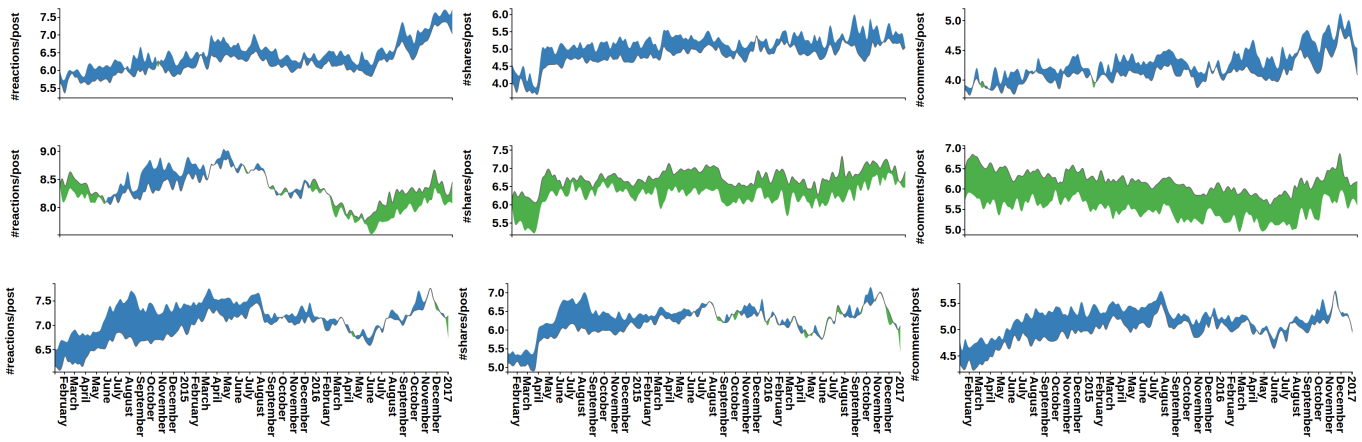

Fig. 8. Top: Print media, Middle: Broadcast media, Bottom: Unreliable media. Blue areas indicate that on average, a clickbait post (link or video) receives more attention (reactions/shares/comments) than a non-clickbait post. Green areas indicate the opposite.

Clickbait– Properties, Practice and Effects: There have been a small number of studies–some conducted by academic researchers and others by media firms–which examined correlations between headline attributes and degree of user engagement with content. Some media market analysts and commentators [17] discussed various aspects of this practice. However, no research has been found, which gauges the extents of clickbait practices by mainstream and alternative media outlets on the web. Nor have we found any study that examined if clickbait techniques help increase user engagement on social media.

A journalism professor [1] manually examined content of four online sections of the Spanish newspaper *El Pais* [9], which apparently used clickbait features to capture attention. The corpus included only 151 articles published in June, 2015. The articles in the corpus appeared to emphasize anecdotal aspects, or issues with little value, and curiosities. The study identified various linguistic techniques used in headlines of these articles such as orality markers and interaction (e.g., direct appeal to the reader), vocabulary and word games (e.g., informal language, generic or buzzwords), and morphosyntax (e.g., simple structures).

Researchers at the University of Texas's *Engaging News Project* [5] conducted an experiment on 2,057 U.S. adults to examine their reactions to clickbait (e.g., question-based headlines) and traditional news headlines in political articles. They found that clickbait headlines led to more negative reactions among users than non-clickbait headlines. Interestingly, the same users were slightly more engaged with non-traditional media that tend to use clickbait techniques more often. This finding questions the conventional belief that user reactions may predict user engagement, and warrants large-scale investigations.

Chartbeat, an analytics firm that provides market intelligence to media organizations, tested 10,000 headlines from over 100 websites for their effectiveness in engaging users with content [18]. The study examined 12 'common tropes' in headlines– a majority of them are considered clickbait techniques – and found that some of these tropes are more effective than others. Some media pundits interpreted the findings of this study as clickbaits being detrimental to traditional news brands.

HubSpot and *Outbrain*, two content marketing platforms that distribute clickbait contents across the web, examined millions of headlines to identify attributes that contribute to traffic growth, increased engagement, and conversion of readers into subscribers. The study suggested that clickbait techniques may increase temporary engagement [19], but an article must deliver on its promises made in headline for users to return and convert.

Automated Clickbait Detection: [2], [10], [20], [21] study automated detection of clickbait headlines using natural language processing and machine learning. [21] collects 10,000 headlines from *Buzzfeed*, *Clickhole*, and *The New York Times (NYT)* and uses Logistic Regression to create a supervised clickbait detection model. It assumes all *Buzzfeed* and *Clickhole* headlines as clickbait and all *NYT* headlines as non-clickbait. We would like to argue that it makes the model susceptible to personal bias as it overlooks the fact that many *Buzzfeed* contents are original, non-clickbaity and there are clickbait practice in *NYT* [22]. Moreover, *BuzzFeed*, and *NYT* usually write on very different topics. The model might have been trained merely as a topic classifier. [20] attempts to detect clickbaity Tweets in Twitter by using common words occurring in clickbaits, and by extracting some tweet specific features. [2] uses a dataset of 15,000 manually labeled headlines to train several supervised models for clickbait detection. These methods heavily depend on a rich set of handcrafted features which take good amount of time to engineer and sometimes are specific to the domain (for example, tweet related features are specific to Twitter data and inapplicable to other domains). [10] presents clickbait detection model which uses word embeddings and Recurrent Neural Network

[9]http://elpais.com

(RNN). These works consider the structure and semantic of a headline to determine whether it is a clickbait or not. However, one important aspect, the body of the news, is not considered as a factor in these works at all. We would like to argue that only the headline itself does not fully represent whether an article is a clickbait or not. If a headline represents the body fairly, it should not be considered as a clickbait. Consider the title as an example, *'The Top 10 Mistakes Of Entrepreneurs'*[10]. It is as clickbait of a headline as it can be. However, the body actually contains reasonably decent materials, which might be interesting to many people.

Clickbait Generation [23]–[25] present automated clickbait generation tools. [23] trains an RNN model using 2 million headlines collected from *Buzzfeed*, *Gawker*, *Jezebel*, *Huffington Post* and *Upworthy*. The model is then used to produce new clickbait headlines

VI. Conclusion

In this paper, we introduce a word-embedding based clickbait detection system which is built on our own collected corpus of news headlines and contents. We showed that our model performs better than the Google news dataset based embeddings. Our analysis also reveals how mainstream media are getting involved into clickbait practicing increasingly. Close scrutiny of the social media posts also reveals that broadcast type media has higher percentage of usage of clickbait practice than the print media and non-news type broadcast media mostly contributes to it. Our study also brings forth another fact of using higher percentage of clickbait practice by unreliable media which is quite obvious. Moreover, results from our topic modeling indicates that clickbait practice is prevalent in personalized and entertaining areas. In future, we want to incorporate the content of the news in defining the clickbaitiness of a headline. We believe, such a system would help social networking platforms to curb the problem of clickbait and provide a better using experience.

References

[1] D. Palau-Sampio, "Reference press metamorphosis in the digital context: clickbait and tabloid strategies in elpais. com." *Communication & Society*, vol. 29, no. 2, 2016.

[2] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 9–16.

[3] M. S. LUCKIE, "Adele and the death of clickbait," http://www.niemanlab.org/2015/12/adele-and-the-death-of-clickbait/, 2015.

[4] C. Sutcliffe, "Can publishers step away from the brink of peak content?" https://www.themediabriefing.com/article/can-publishers-step-away-from-the-brink-of-peak-content, 2016.

[5] J. M. Scacco and A. Muddiman, "Investigating the influence of "clickbait" news headlines," https://engagingnewsproject.org/wp-content/uploads/2016/08/ENP-Investigating-the-Influence-of-Clickbait-News-Headlines.pdf, 2016.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[7] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[10] A. Anand, T. Chakraborty, and N. Park, "We used neural networks to detect clickbaits: You won't believe what happened next!" *arXiv preprint arXiv:1612.01340*, 2016.

[11] J. Eggerton, "Fcc: Nielsen dmas still best definition of tv market," Broadcasting & Cable, 2016.

[12] informationisbeautiful.net, "Unreliable/fake news sites & sources," https://docs.google.com/spreadsheets/d/1xDDmbr54qzzG8wUrRdxQl_C1dixJSIYqQUaXVZBqsJs, 2016.

[13] M. Zimdars, "My 'fake news list' went viral. but made-up stories are only part of the problem," https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem, 2016.

[14] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1445–1456.

[15] R. S. Izard, H. M. Culbertson, and D. A. Lambert, *Fundamentals of news reporting*. Kendall/Hunt Publishing Company, 1994.

[16] M. Ingram, "The internet didnt invent viral content or clickbait journalism – theres just more of it now, and it happens faster," https://gigaom.com/2014/04/01/the-internet-didnt-invent-viral-content-or-clickbait-journalism-theres-just-more-of-it-now-and-it-happens-faster, 2014.

[17] F. Filloux, "Clickbait is devouring journalism but there are ways out," https://qz.com/648845/clickbait-is-devouring-journalism-but-there-are-ways-out/, 2016.

[18] C. Breaux, "You'll never guess how chartbeat's data scientists came up with the single greatest headline," http://blog.chartbeat.com/2015/11/20/youll-never-guess-how-chartbeats-data-scientists-came-up-with-the-single-greatest-headline, 2015.

[19] Hubspot and Outbrain, "Data driven strategies for writing effective titles & headlines," http://cdn2.hubspot.net/hub/53/file-2505556912-pdf/Data_Driven_Strategies_For_Writing_Effective_Titles_and_Headlines.pdf.

[20] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *European Conference on Information Retrieval*. Springer, 2016, pp. 810–817.

[21] A. Thakur, "Identifying clickbaits using machine learning," https://www.linkedin.com/pulse/identifying-clickbaits-using-machine-learning-abhishek-thakur, 2016.

[22] N. Hurst, "To clickbait or not to clickbait? an examination of clickbait headline effects on source credibility," Ph.D. dissertation, University of Missouri–Columbia, 2016.

[23] L. Eidnes, "Auto-generating clickbait with recurrent neural networks," https://larseidnes.com/2015/10/13/auto-generating-clickbait-with-recurrent-neural-networks, 2015.

[24] C. Cha, "clickbait generator," http://www.thisisreallyreal.com/, 2016.

[25] "Linkbait title generator," http://www.contentrow.com/tools/link-bait-title-generator.

[10] www.forbes.com/sites/roberthof/2016/02/23/guy-kawasaki-the-top-10-mistakes-of-entrepreneurs