Diving Deep into Clickbaits:

Who Use Them to What Extent in Which Topics with What Effects?

Md Main Uddin Rony[1], Naeemul Hassan[1], Mohammad Yousuf[2]

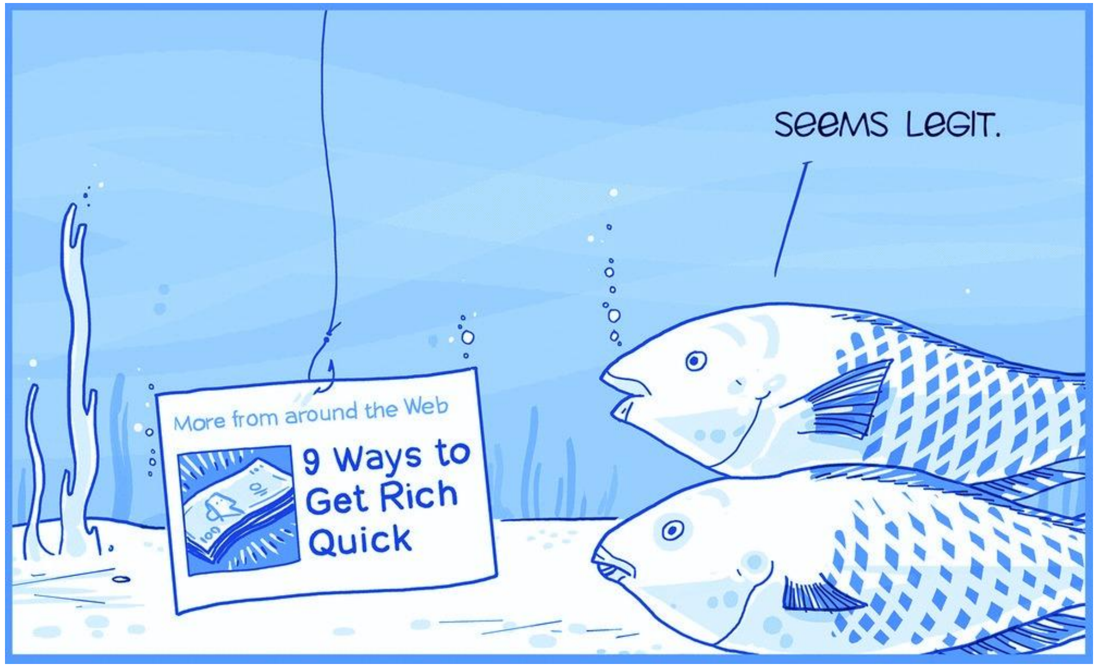
[1]The University of Mississippi, [2]The University of Oklahoma

Familiar with the term "Clickbait"?

- Techniques used in headlines to trick readers into clicking links
- But fails to deliver what users really look for
- Examples-

Eek! What's Lurking in the Shadows?! I Have to Know!

I Left My Daughter And THIS Happened!



Clickbait has become widespread ...

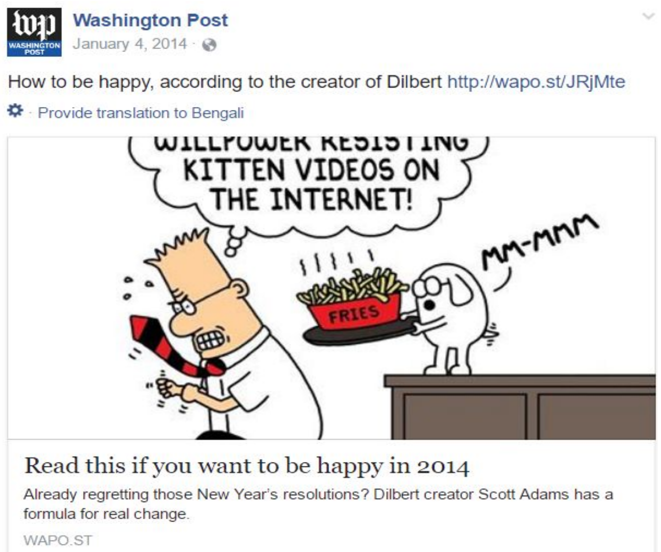
- Both mainstream and unreliable media practice it
  - Reachability is more than ever before
  - Social media has become a practice field

- Has become a source of easy revenue
  - More click means more revenue
  - Competitive media market

Taboola

Claims to have doubled its monthly reach from 500 million unique users to 1 billion in a single year from March 2015

In 2020,The entertainment and media market in the United States is expected to be worth over 720.38 billion U.S (Source: www.statista.com)

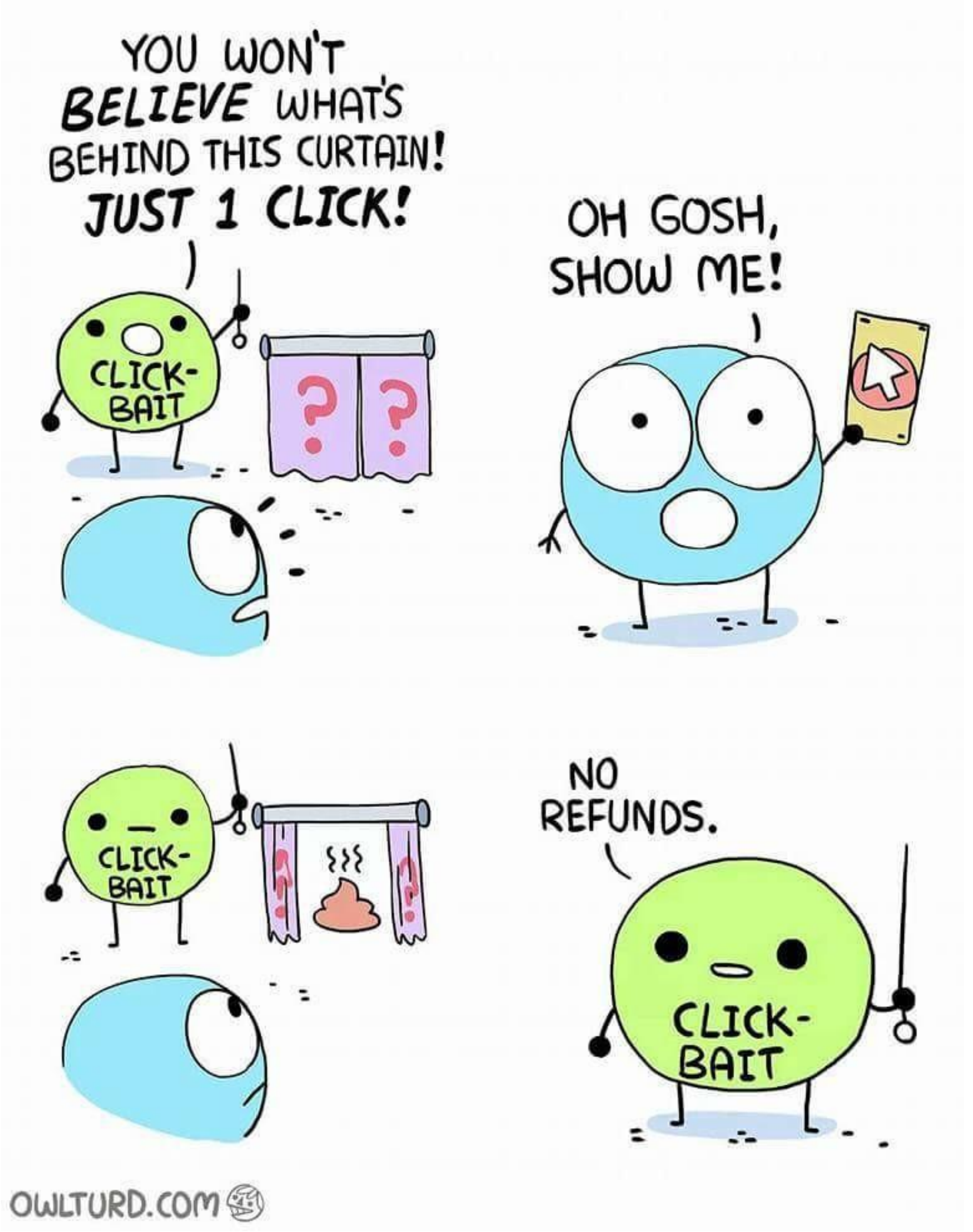


Shocking impact..

- Negative impact on media eco-system
  - Risk user trust
  - Depleting brand value

Flickr

Is Clickbait Content Destroying Your Brand?



Satisfactory research on this???

- No
- Small amount of research compared to its reaching and impact
- No large scale analysis on practice of clickbait by media organizations.
- No study to show its contribution to public engagement on social media.

So, in this work, we answer-

To what extent, mainstream and unreliable media organizations use clickbait?

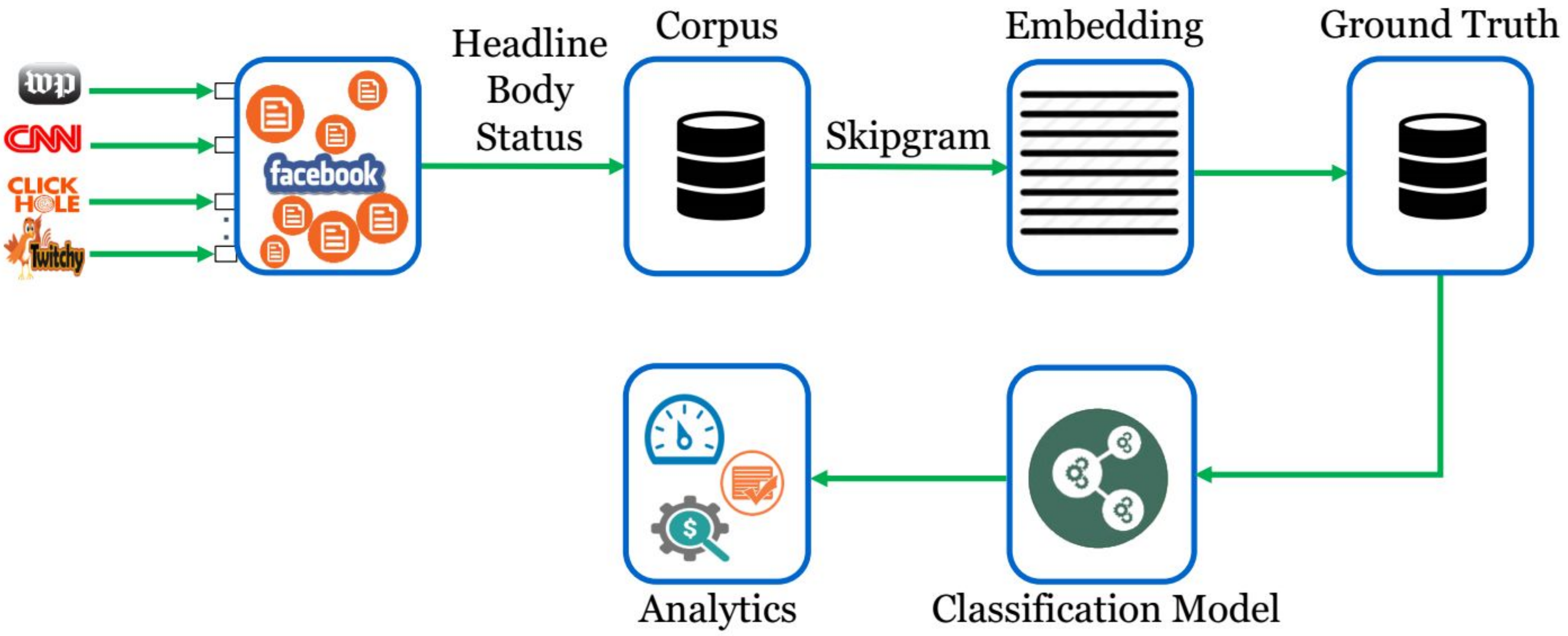
Does the topic distribution of the contents vary in clickbaity contents?

Which type of headlines – clickbait or non-clickbait - generates more user engagement (e.g., shares, comments, reactions)?

Our contributions…

- Developed a supervised clickbait detection model
- Prepared distributed subword based embeddings
- Performed detailed analysis of the clickbait practice in the social network from multiple perspectives

Workflow…



Headline
Body
Status
Corpus
Skipgram
Embedding
Ground Truth
Analytics
Classification Model

Clickbait detection: Problem Definition

- A supervised binary classification problem
- Want to model a function that can categorize sentence into clickbait and non-clickbait

Clickbait detection: Problem Modeling

- Traditional text classification uses bag-of-words(BOW) model to transform text into feature vectors
  - Can't handle the order of words (I eat rice == rice I eat)
  - Can't capture semantics of the sentence (I eat apple / I eat orange)
  - Scalability challenges (Sparse Matrix, One column for each word)

- Solution: Probabilistic language modeling
  Word2Vec
  - Skip-Gram (*predicting the context given a word*)
  - CBOW (*predicting the word given its context)*

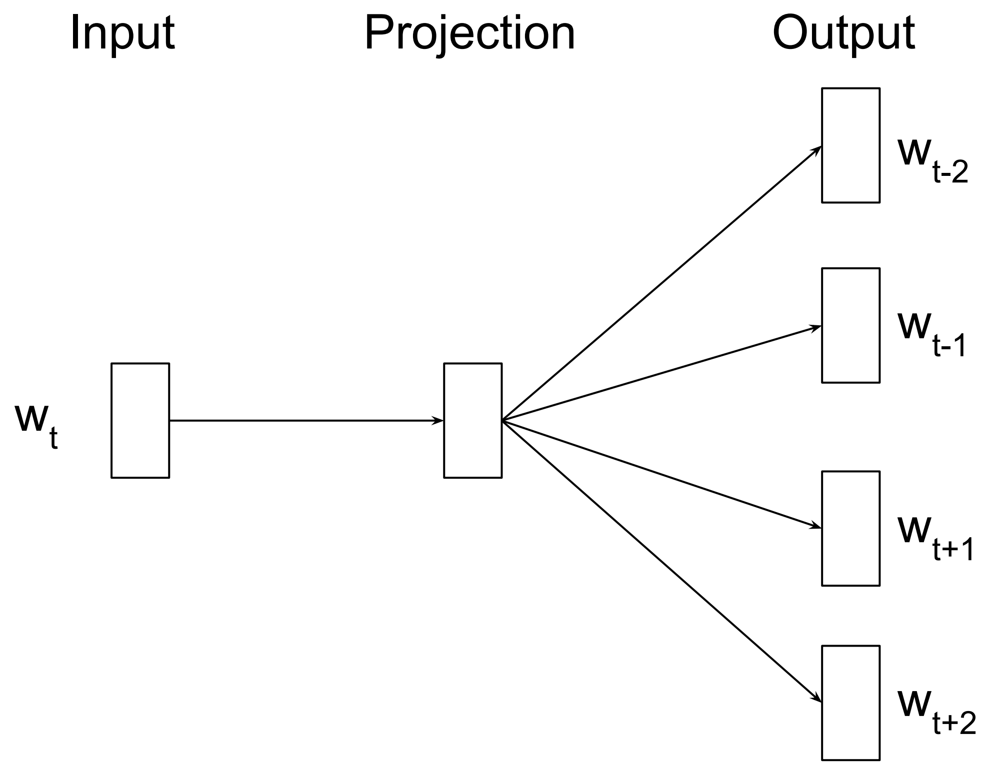
- Why Skip-gram?
  - Able to extract more information when more data is available

Clickbait detection: Skip-Gram

- Formal Definition: 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

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

where the context $\mathcal{C}_t$ is the set of indices of words surrounding w_t

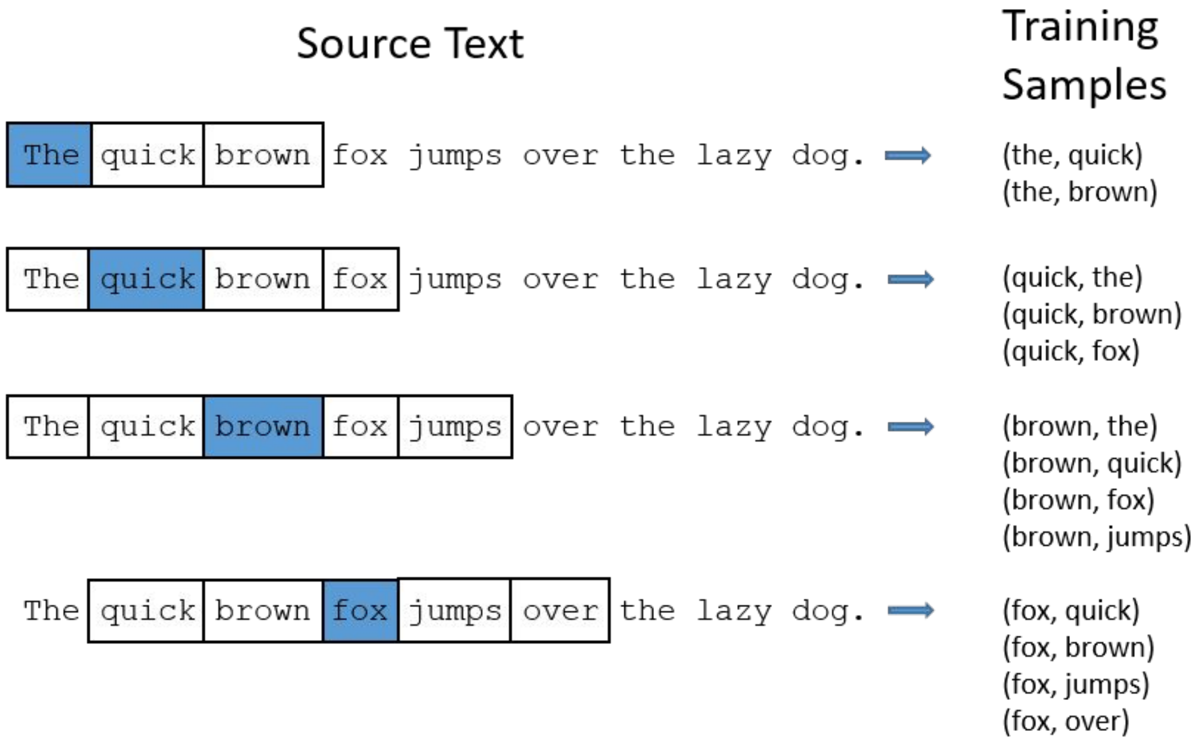


Clickbait detection: Skip-Gram

- It’s a neural network which gives the probability of a word being the “nearby word” that we chose.
- “nearby” $\rightarrow$ “window size”
- For a given word “Soviet”, which will produce more probability?
  - Union?
  - Russia?
  - Watermelon?
  - Kangaroo?

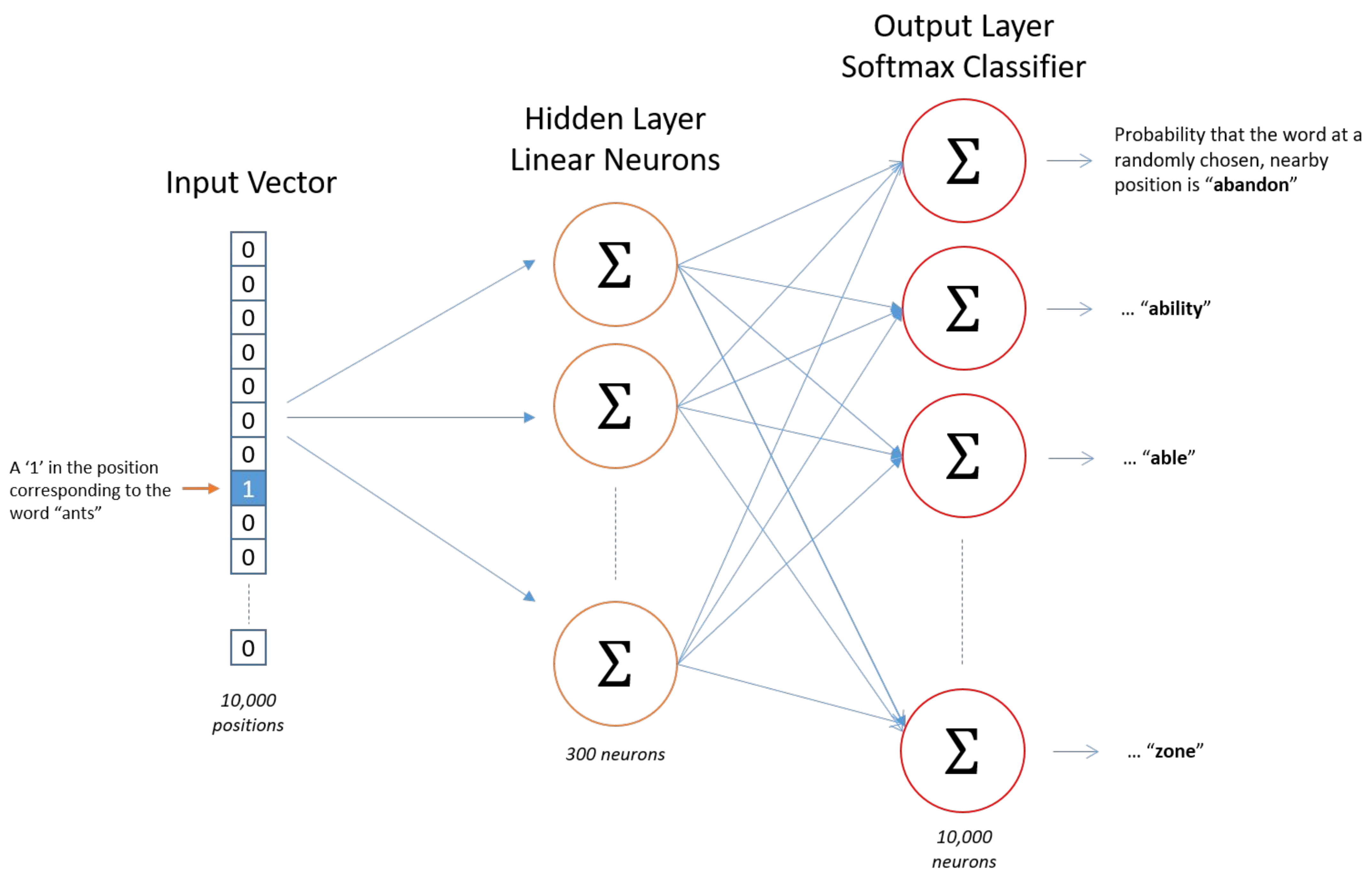
Clickbait detection: Skip-Gram

- Target is to replicate the idea from a set of given word pairs
- Let's look an example:
  "The quick brown fox jumps over the lazy dog."



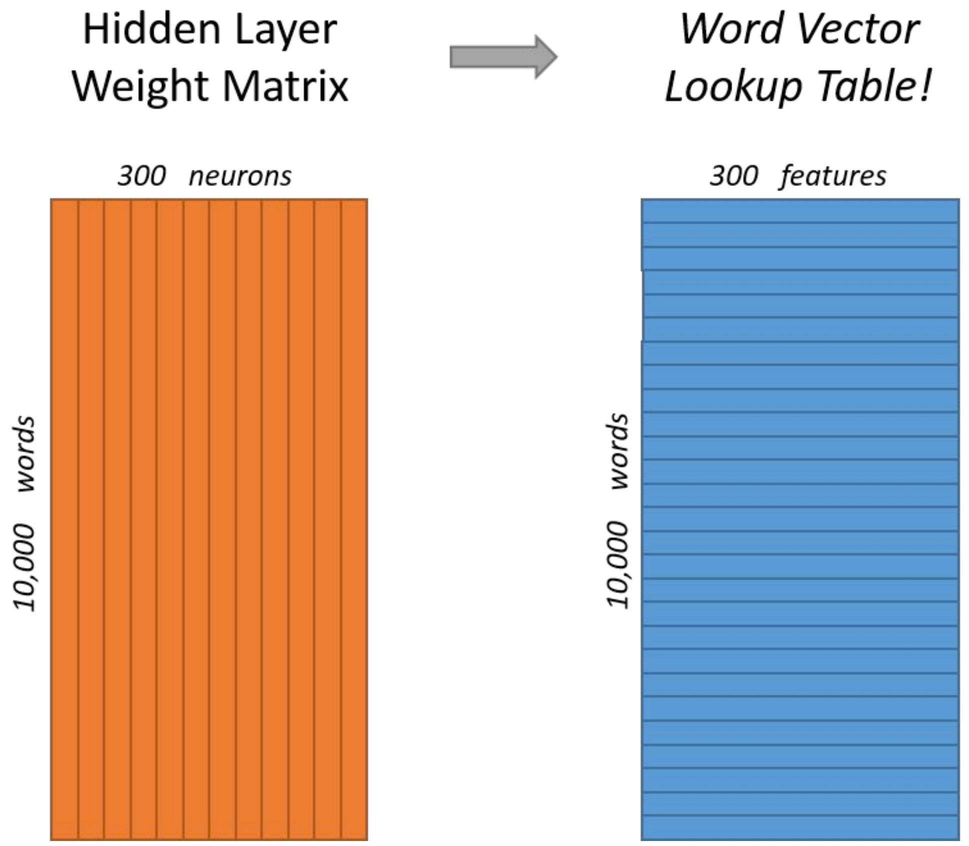
Clickbait detection: Skip-Gram

- So how to model this?
- We need a way to represent the words to the network
- Need a vocabulary of words from our training set (e.g., 10000 words)
- Represent an input word as a one-hot vector(e.g., [0,0,0,…1,0,0])
- Output of the network is a single vector (also with 10,000 components)


Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
0
0
0
0
0
0
0
1
0
0
0
A '1' in the position corresponding to the word "ants"
10,000 positions
300 neurons
10,000 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"

Clickbait detection: Skip-Gram

- Say we're learning word vectors with 300 features
- Hidden layer is going to be represented by a weight matrix with 10,000 rows and 300 columns
- Want to learn this hidden layer weight matrix



Clickbait detection: Skip-Gram

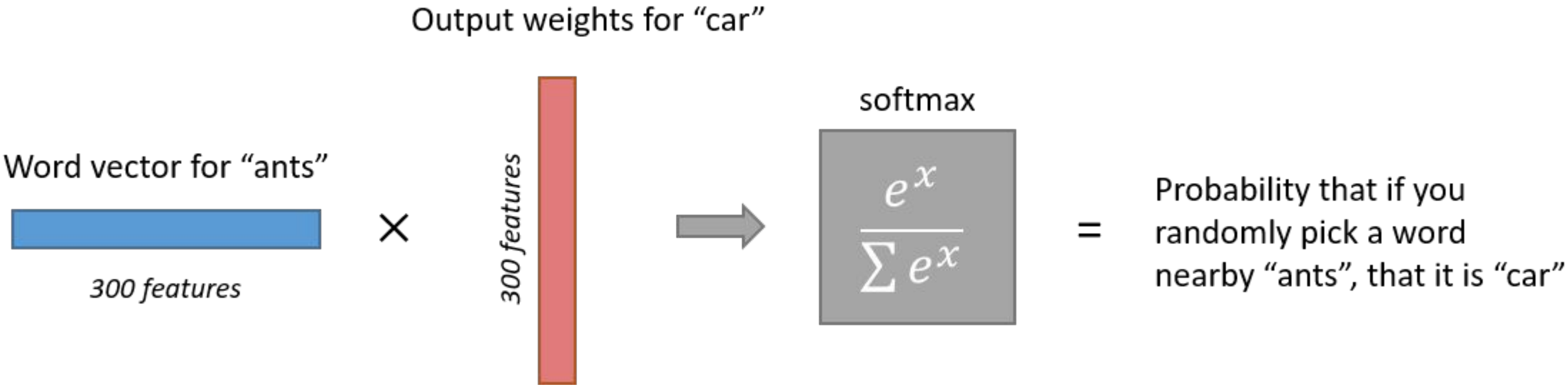
- One-hot vector is almost all zeros… what's the effect of that?

$$\begin{bmatrix} 0 & 0 & 0 & \boxed{1} & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 12 & 19 \end{bmatrix}$$

- Hidden layer → lookup table
- output of the hidden layer is just the "word vector" for the input word

Clickbait detection: Skip-Gram

- Output layer is a Softmax regression classifier
- Each output neuron will produce an output between 0 and 1
- The sum of all these output values will add up to 1



Clickbait detection: Skip-Gram(Extension)

- We use an extension of the continuous *skip-gram* model
- Takes into account subword (substring of a word) information
- Back to previous example, we consider a word e.g., "quick"
  The quick brown fox jumps over the lazy dog.
- Assuming subword length as three, the subwords are- *{qui, uic, ick}*
- This model learns to predict *qui*, *ick* in the context given *uic* as the input.

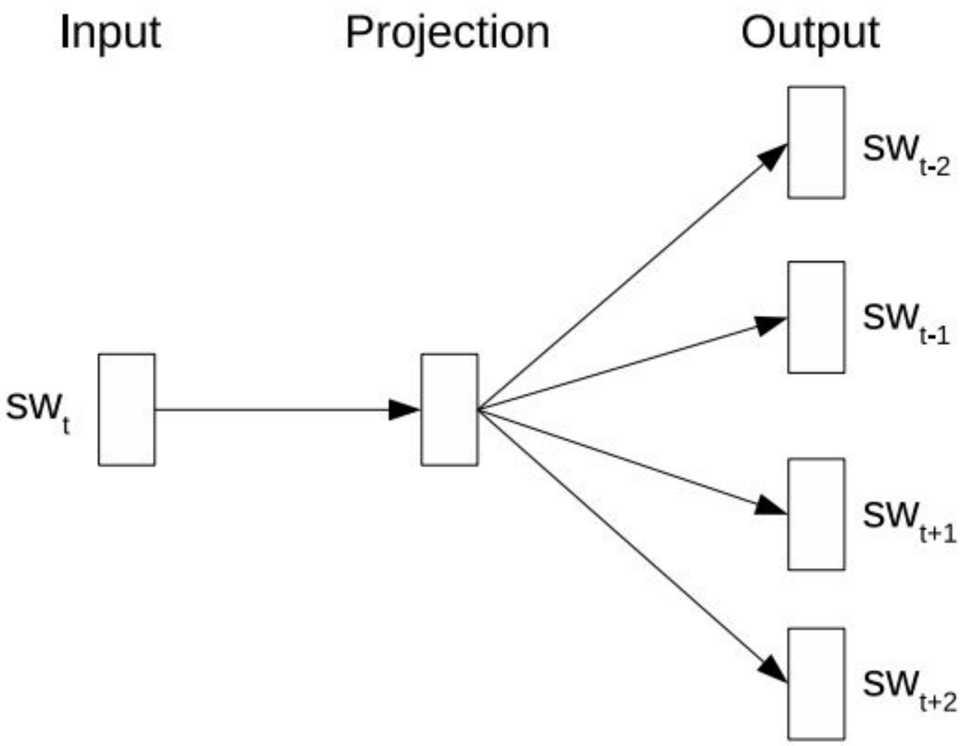
Clickbait detection: Skip-Gram(Extension)

- Embedding of a word is formed by the sum of the vector representations of its subwords
- The equation is:

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

$\mathbf{u}_w$ = embedding of word, w

$\mathbf{v}_{sw}$ = vector representation of sw



Clickbait detection: Skip-Gram(Extension)

- Why we used it?
  - Allows sharing the representations across words (Information of "run" can be passed to "running")

  - Able to learn reliable representation for rare words (An embedding of unknown word can be formed from its subword embedding)

Clickbait detection: Pre-trained Vectors

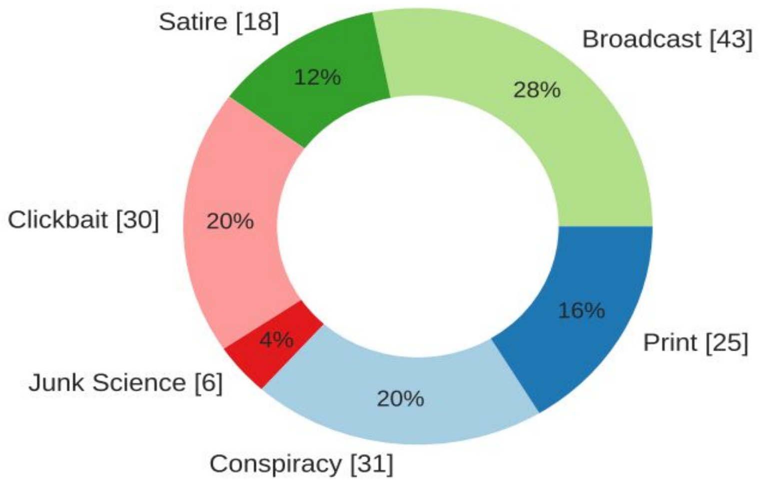
- Great opportunity to use richer word embedding
- Use the texts (headlines, messages, bodies) from our own collected dataset(4,77,236 unique embeddings)

- Why not Google News data?(100 billion unique embeddings)
  - Embeddings from Media Corpus have more domain specific knowledge than the Google News dataset
  - Processing will be faster with smaller dataset

Data Collection

- Ground Truth
  - 32,000 manually labeled headlines curated by Chakraborty et al.*
- Media corpus
  - About 1.7 million Facebook posts
  - Collected from 68 mainstream and 85 unreliable media
  - Data collection period: 2014 -2016

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. 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

Clickbait detection: Classifier

- Use Ground Truth dataset as a training/testing set
- 15, 999 clickbait headlines and 16, 001 non-clickbait headlines
- Train test ratio : 80-20%
- 10 - fold Cross validation
- Repeat 5 times to avoid randomness

Clickbait Detection: Evaluation

	Method	Precision	Recall	F-measure	Accuracy
Without Pre-trained Vectors	*Chakroborty et al. [2]	0.95	0.90	0.93	0.93
	Skip-Gram$_{sw}$	0.976	0.975	0.975	0.976
With Pre-trained Vectors	*Anand et al. [10]	0.984	0.978	0.982	0.982
	Skip-Gram$_{sw}$+ Google_word2vec	0.977	0.977	0.977	0.976
	Skip-Gram$_{sw}$+ (Headline)	0.981	0.981	0.981	0.981
	Skip-Gram$_{sw}$+ (Headline + Message)	0.982	0.982	0.982	0.982
	Skip-Gram$_{sw}$+ (Headline + Body + Message)	**0.983**	**0.983**	**0.983**	**0.983**

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

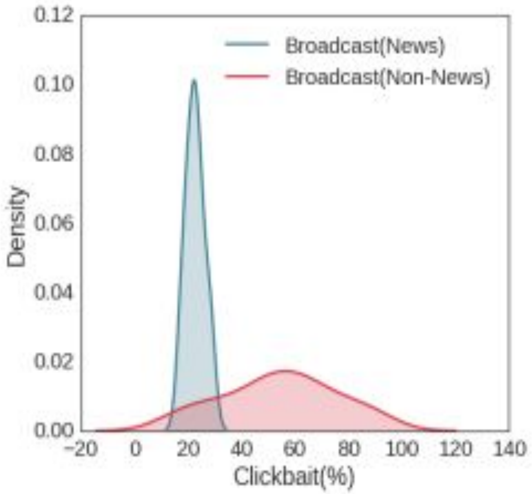
Quantitative Analysis

Media	Category	Clickbait	Non-Clickbait	Clcikbait(%)
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

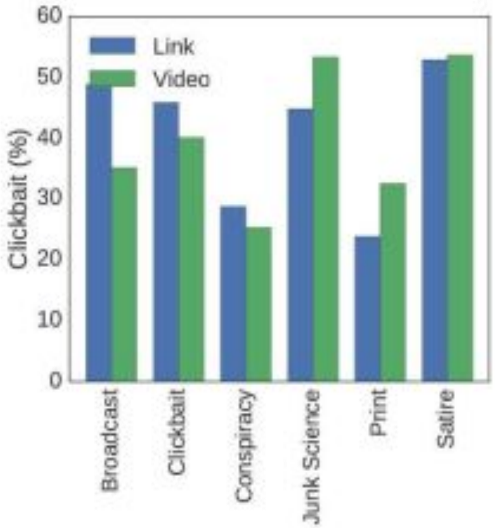
% of clickbaits in various media

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

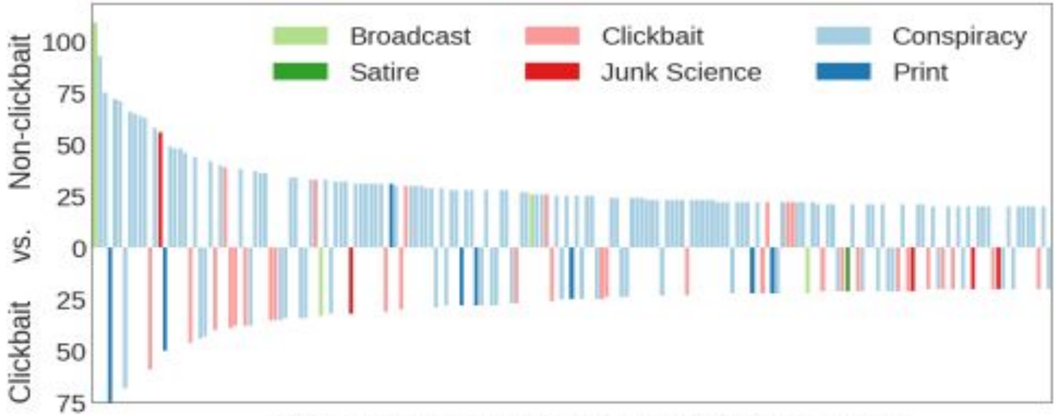
% of clickbait in Facebook Status



% of clickbait in news & non-news



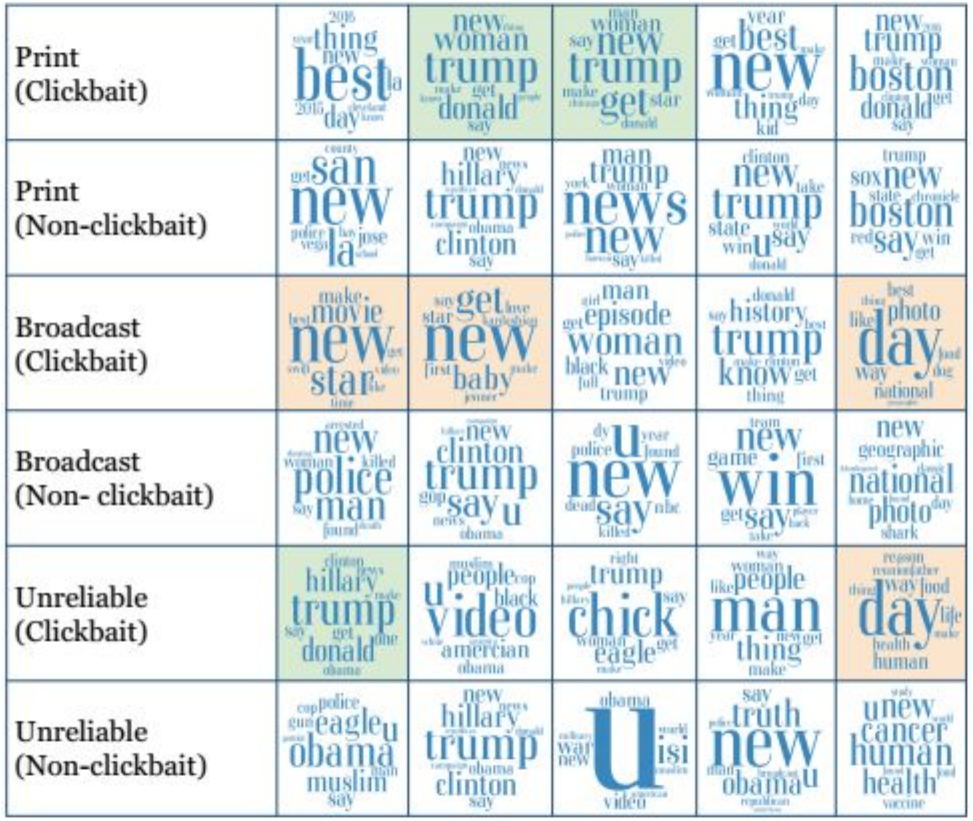
% of clickbait in link & video



Frequency of re-post by different media

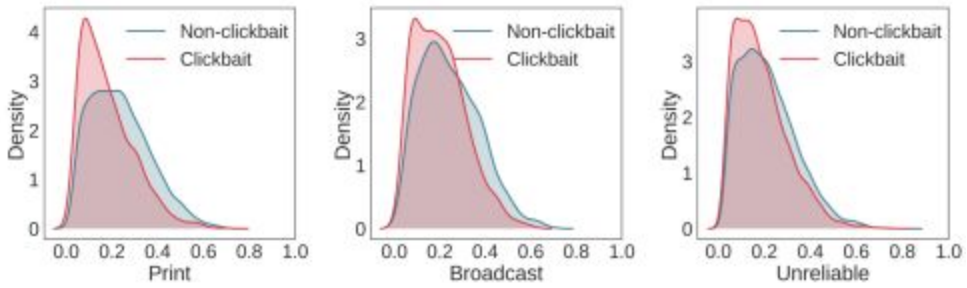
Qualitative Analysis: Topic Modeling

- Use BTM (Bi-term Topic Modeling) for topic detection
- BTM performs better on short text than the traditional topic modeling algorithm
- Take 5 topic for each type and each topic contains 10 words
- Clickbait headlines in print and broadcast media represent more personalized, sensationalized and entertaining topics
- Non-clickbait headlines highlight topics of collective problems such as public policies

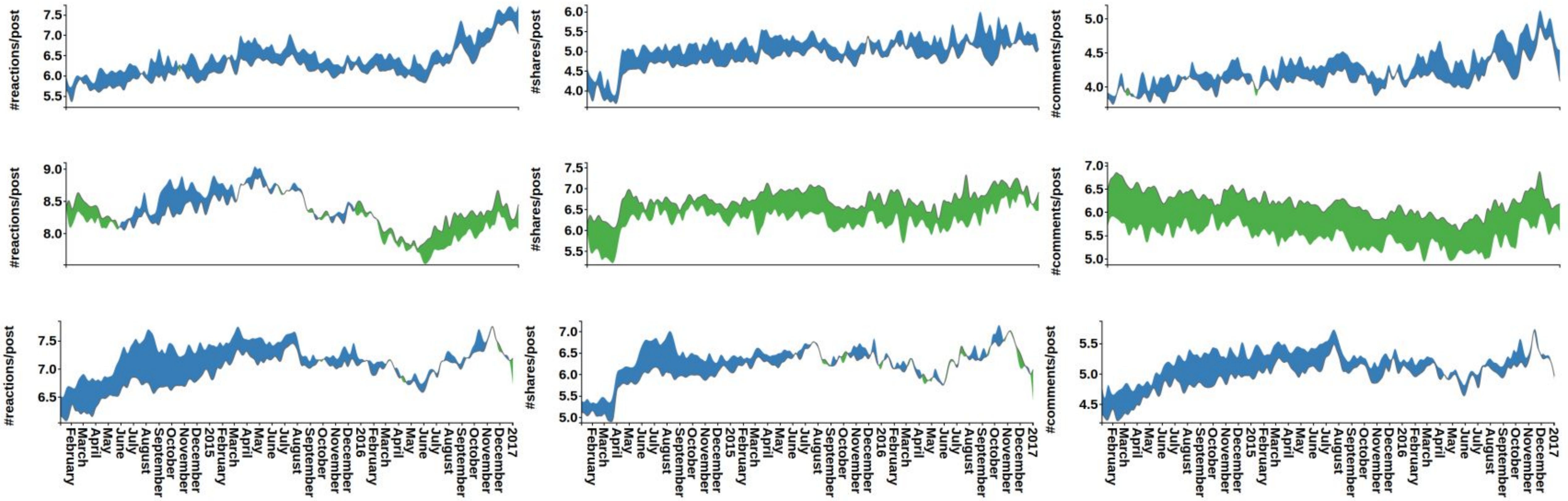


Qualitative Analysis: Headline-Body Relevance

- Hypothesis: Clickbait headlines are less
  relevant to the body content.
- Cosine similarity was used to measure the
  relevance between a headline and the body



Impact Analysis



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.

Future Work

- Headline – Body Similarity
- Deception Mining

Questions?

Thank You