Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Comparitive analysis of clickbait classifiers

NLP Project

March 22, 2023

Pradyumna Heddur Nagendra
Priya Chaudhary

Advisors
Prof. Dr. Paul Plöger
M.Sc. Tim Metzler

Clickbait detector

Introduction:

- Clickbait is a term that describes deceiving web content that uses ambiguity to provoke the user into clicking a link.
- It aims to increase the number of online readers in order to generate more advertising revenue.
- These baits may trick the readers into clicking, but in the long run, clickbait usually doesn't live up to the expectation of the readers and leave them disappointed.

Clickbait detector

Motivation: The automatic detection of clickbaits alerts readers of various media websites to the potential of falling for such headlines.

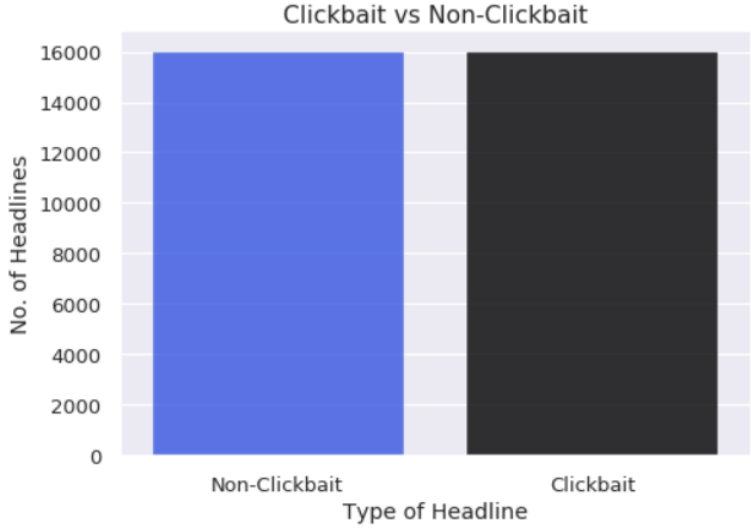
Dataset: Stop Clickbait Dataset: It contains 16,000 article headlines categorized as “clickbait” and “non-clickbait”. The clickbait articles have been pulled from websites including Buzzfeed and Upworthy, while the non-clickbait articles come from sites including Wikinews, The New York Times, and The Guardian.

Raw Data

	headline	clickbait
0	Should I Get Bings	1
1	Which TV Female Friend Group Do You Belong In	1
2	The New "Star Wars: The Force Awakens" Trailer...	1
3	This Vine Of New York On "Celebrity Big Brothe...	1
4	A Couple Did A Stunning Photo Shoot With Their...	1
...	...	...
31995	To Make Female Hearts Flutter in Iraq, Throw a...	0
31996	British Liberal Democrat Patsy Calton, 56, die...	0
31997	Drone smartphone app to help heart attack vict...	0
31998	Netanyahu Urges Pope Benedict, in Israel, to D...	0
31999	Computer Makers Prepare to Stake Bigger Claim ...	0

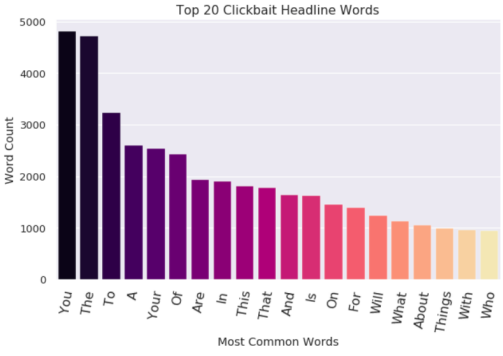
32000 rows × 2 columns

Data-Analysis

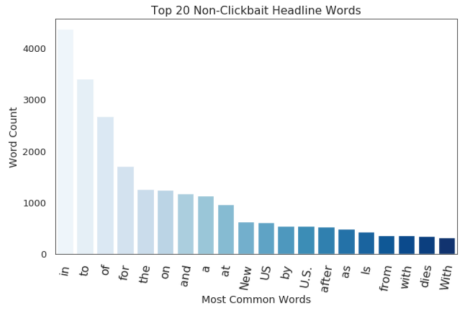


Data-Analysis

Top 20 clickbaits before processing

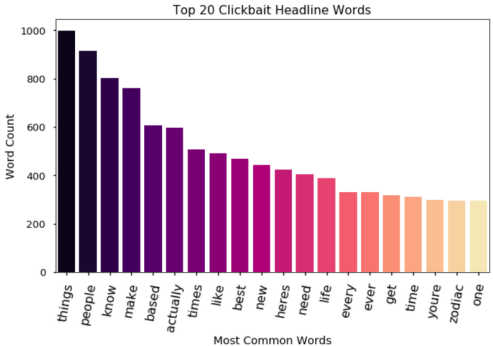


Top 20 non-clickbaits before processing

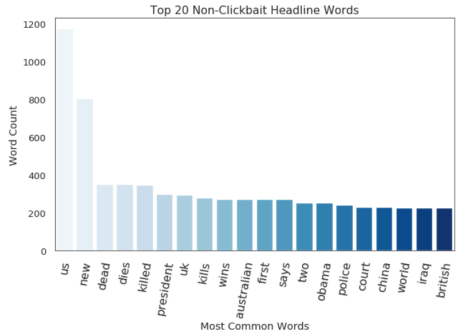


Data-Analysis

Top 20 clickbaits after processing



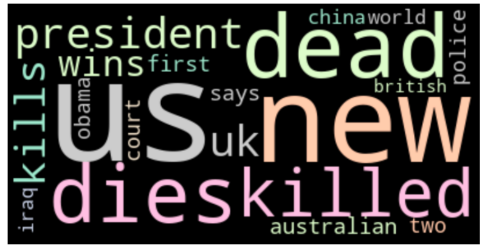
Top 20 non-clickbaits after processing



Data-Analysis

Clickbaits word cloud

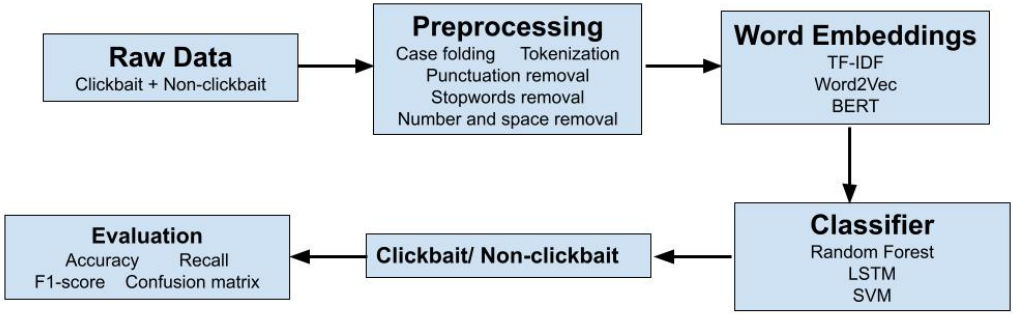
Non-clickbaits word cloud



Data-preprocessing

- Case folding
- Tokenization
- Punctuation removal
- Stopwords removal
- Number and space removal

Architecture



Vectorization

TF-IDF:

- Term frequency-inverse document frequency (TF-IDF) gives a measure that takes the importance of a word into consideration depending on how frequently it occurs in a document and a corpus.

Word2Vec:

- This model works using context. This implies that to learn the embedding, it looks at nearby words if a group of words is always found close to the same words, they will end up having similar embeddings.

BERT:

- BERT produces word representations that are dynamically informed by the words around them.

Classification networks

Machine learning:

- SVM
- Random Forest

Deep learning:

- LSTM

Result

TF-IDF:

Method	Accuracy	Recall	F1-Score
LSTM	0.961	0.953	0.961
SVM	0.962	0.962	0.962
Random Forest	0.912	0.969	0.917

Word2Vec:

Method	Accuracy	Recall	F1-Score
LSTM	0.959	0.952	0.958
SVM	0.812	0.720	0.797
Random Forest	0.825	0.770	0.816

BERT:

Accuracy	Recall	F1-Score
0.828	0.782	0.819

References

1. A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 9-16, doi: 10.1109/ASONAM.2016.7752207.
2. Suhaib R. Khater, Oraib H. Al-sahlee, Daoud M. Daoud, and M. Samir Abou El-Seoud. 2018. Clickbait Detection. In Proceedings of the 7th International Conference on Software and Information Engineering (ICSIE '18). Association for Computing Machinery, New York, NY, USA, 111–115. https://doi.org/10.1145/3220267.3220287
3. Pujahari, Abinash, and Dilip Singh Sisodia. "Clickbait detection using multiple categorization techniques." Journal of Information Science 47.1 (2021): 118-128.