Clickbait Predictor and Headline Analysis

Amanda White

What is “Clickbait”?

- Clickbait is content designed to get more views or clicks to increase traffic on a website

- Clickbait headlines are rarely representative of the actual news story and perpetuate the sensationalization of news

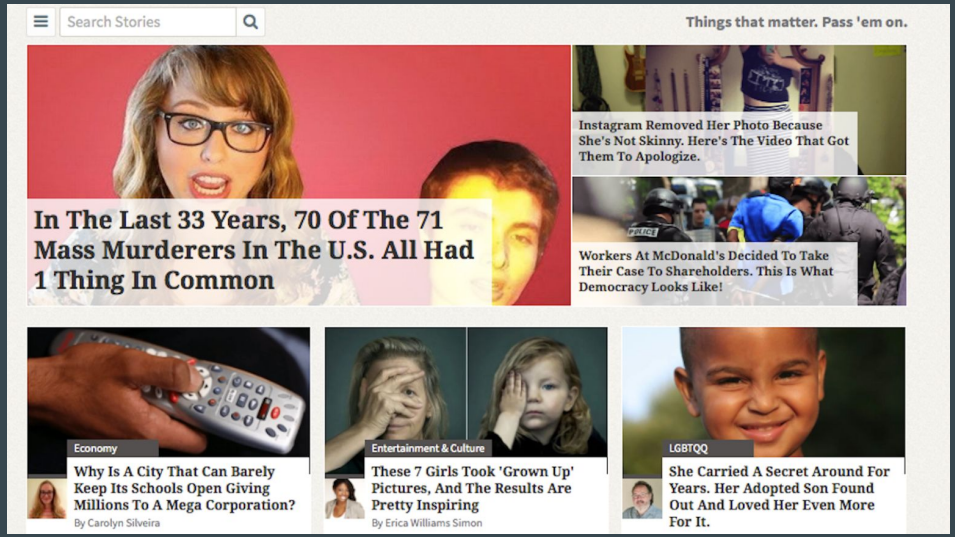


Image source

The Idea

Create a ML algorithm that can identify whether or not a headline is “clickbait”

This serves as another step in the editing process to ensure a news site is not producing misleading content

The Data

Five years of headline data from two news sites:

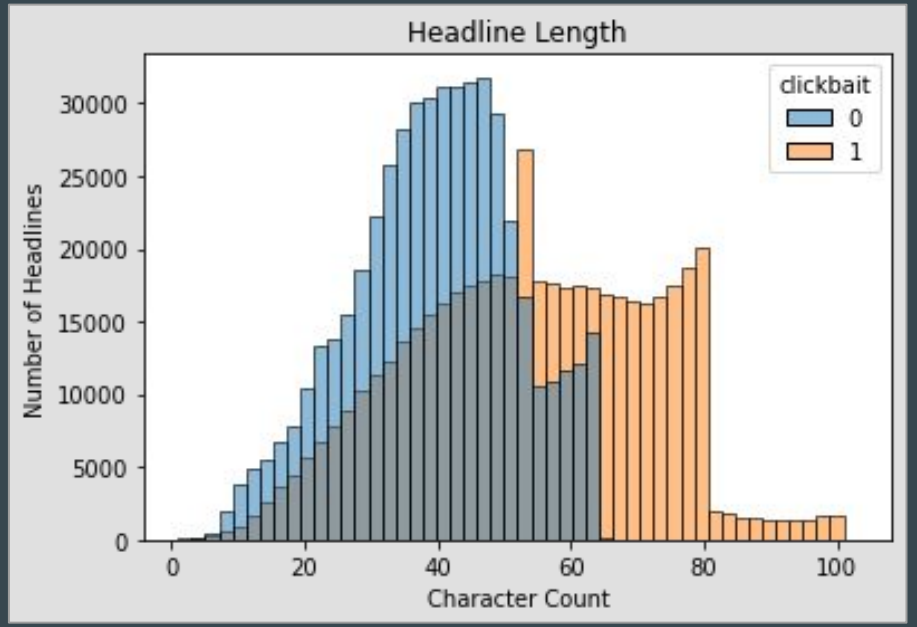
- Clickbait data from The Examiner (a site known for producing less than accurate and clickbait content)
- Non-Clickbait data from ABC (Australian Broadcasting Company)

The Process

1. Cleaning the data: removing stop words, punctuation, numbers, etc.
    a. Only keep the key features of a headline
2. Perform initial analysis on headlines
3. Preprocess the data: apply embedding and tokenize the data
    a. Turning a headline from words into numbers so that an algorithm can make sense of it all
4. Train a Recurrent Neural Network to distinguish between clickbait and not clickbait

Initial Analysis--Headline Length

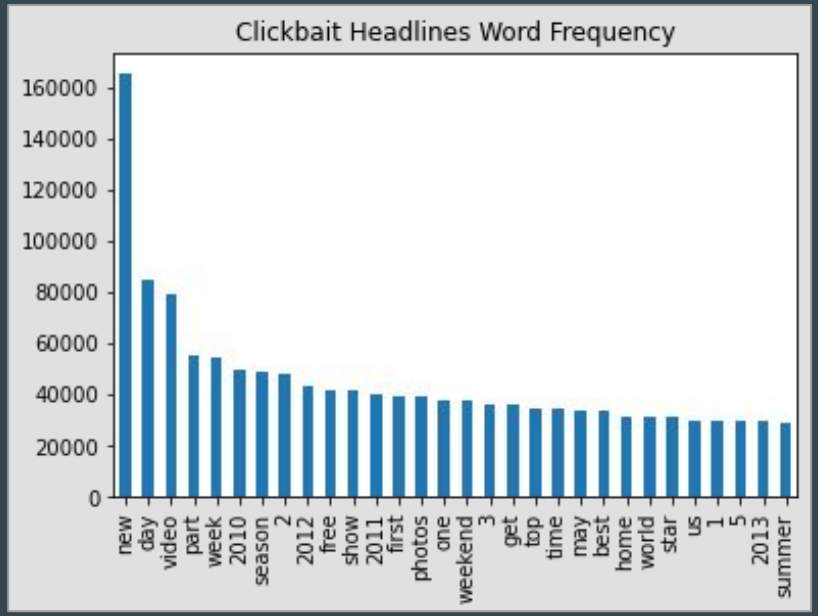
- When comparing headline length non clickbait headlines have a more narrow spread
- Overall, clickbait headlines can be longer, but there is too much overlap to be a reliable predictor

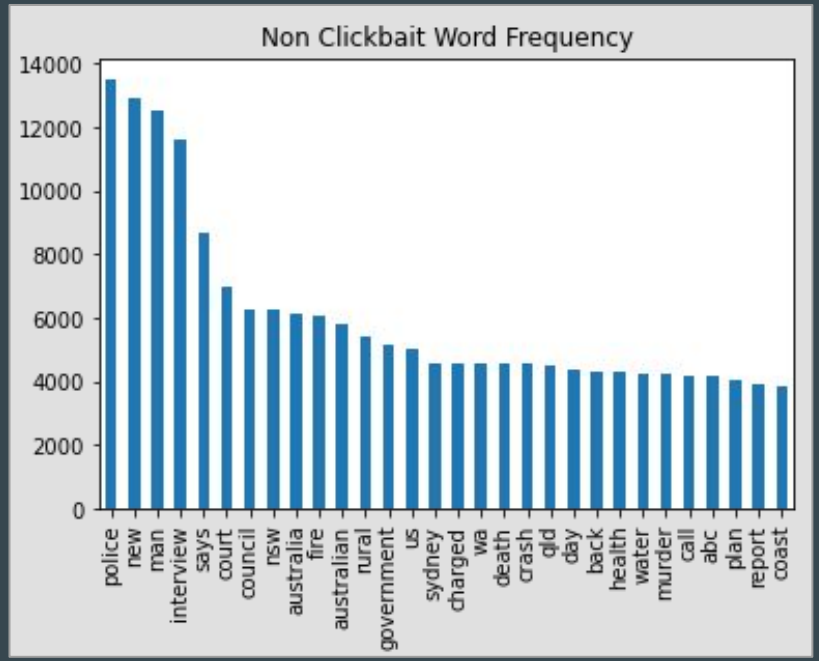


Initial Analysis--Word Frequency

Below are the 30 most commonly used words for each dataset

Numbers and years are used frequently in clickbait headlines





Modeling

The final model: Recurrent Neural Network (RNN)

Predicts clickbait with 91% accuracy

How to use: a function that takes the headline of interest as the only input will process and run through the model

How to use the model

This model is to be used in conjunction with the judgement of the writer and editor.

It should be used as a tool to flag possible clickbait headlines that need to be reviewed and possibly revised

Further Research

Conduct cosine similarity study to determine if headlines have become more like clickbait over time

Update model with training data that are validated clickbait or not clickbait headlines (rather than making assumptions)