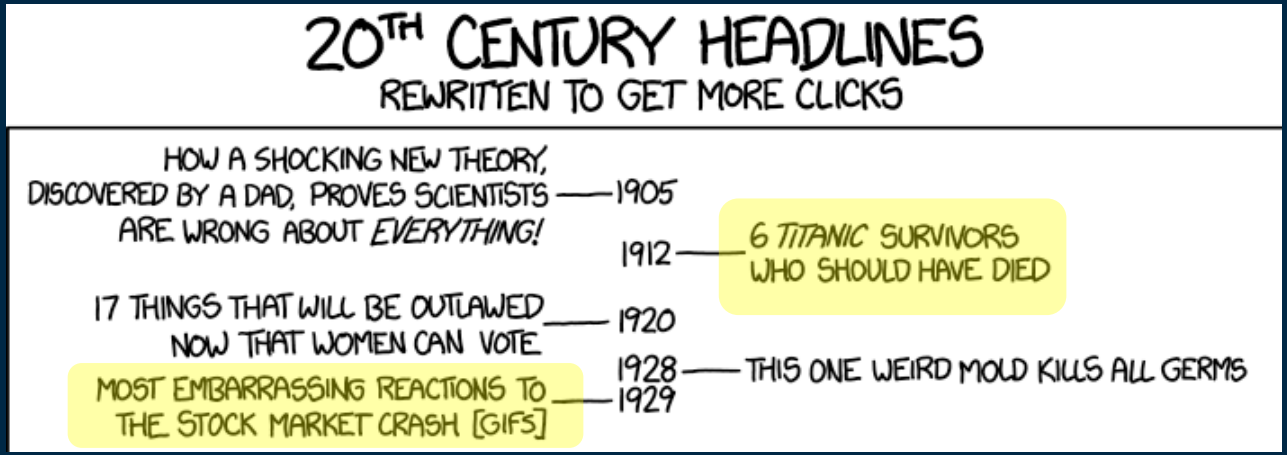
Project 3: Subreddit NLP

Uplifting or Absurd: The Art of High Click-Through Rate Titles

Alyssia Oh


20TH CENTURY HEADLINES
REWRITTEN TO GET MORE CLICKS
HOW A SHOCKING NEW THEORY, DISCOVERED BY A DAD, PROVES SCIENTISTS ARE WRONG ABOUT EVERYTHING!
1905
1912
6 TITANIC SURVIVORS WHO SHOULD HAVE DIED
17 THINGS THAT WILL BE OUTLAWED NOW THAT WOMEN CAN VOTE
1920
1928
THIS ONE WEIRD MOLD KILLS ALL GERMS
MOST EMBARRASSING REACTIONS TO THE STOCK MARKET CRASH [GIFS]
1929

Background

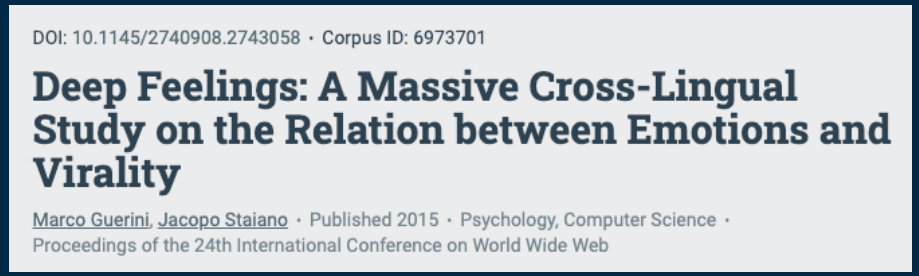
- Clickthrough rate (CTR):
  - A ratio showing **how often people** who see your ad or product listing **end up clicking** it.
  - Clickthrough rate (CTR) can be used to **gauge how well your keywords** and ads are **performing**.

Problem Statement

What are some characteristics of high CTR news titles?

1. how long should it be
2. what words to include
3. what is the sentiment

r/Uplifting News vs r/Not the Onion



DOI: 10.1145/2740908.2743058 · Corpus ID: 6973701

Deep Feelings: A Massive Cross-Lingual Study on the Relation between Emotions and Virality

Marco Guerini, Jacopo Staiano · Published 2015 · Psychology, Computer Science · Proceedings of the 24th International Conference on World Wide Web

A CONCEPTUAL MODEL FOR UNDERSTANDING THE IMPACT OF TEXTUAL EMOTION MINING.

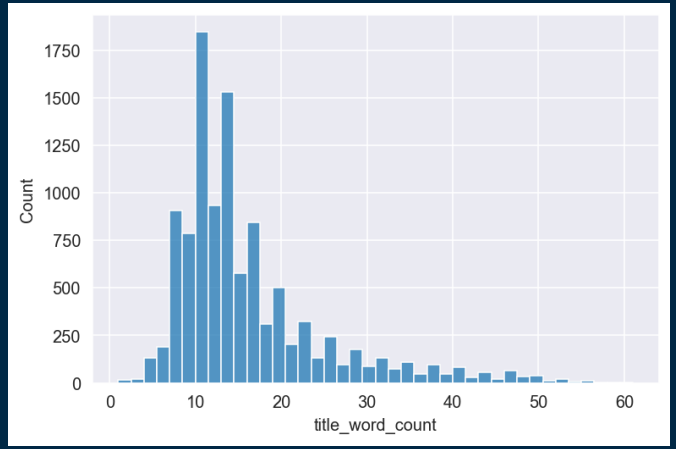
- **Source:** International Journal of Advanced Research in Computer Science . Jan/Feb2018, Vol. 9 Issue 1, p860-863. 4p.
- **Author(s):** Chawla, Shivangi; Mehrotra, Monica

r/UpliftingNews
- 16.6m subscribers
- trust, joy, anticipation

r/NottheOnion
- 18.5m subscribers
- surprise, anger, disgust

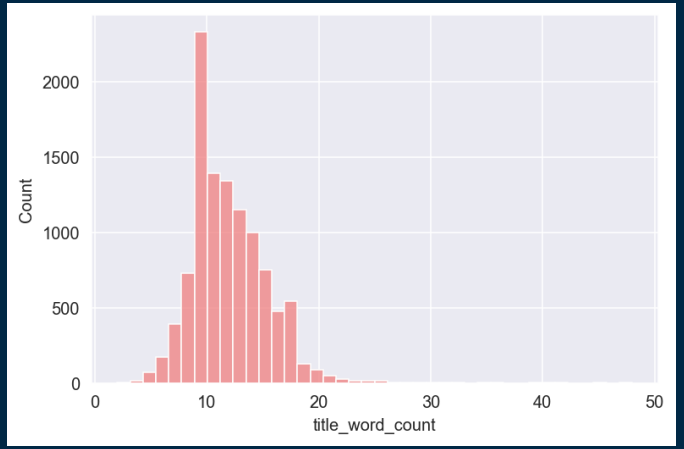
Is there an ideal length for title?

Based on 10,000 posts with comments & upvotes >= 10



r/Uplifting News

15.9 words +/- 9 words



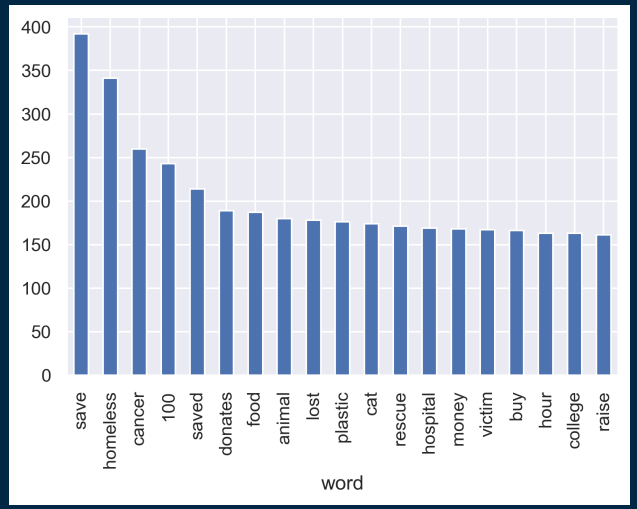
r/Not the Onion

12 words +/- 3.4 words

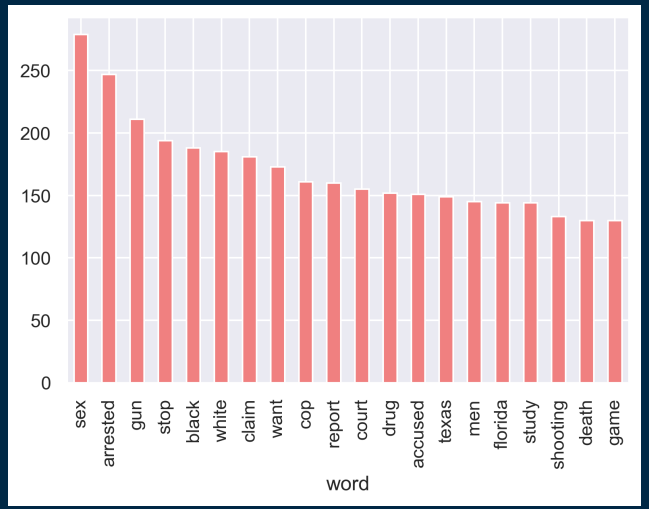
What are the most frequently appearing words?

Save
Homeless
Cancer
Donate
Food
Animal
Lost
Plastic
Cat
Rescue
Hospital
College
Money
Victim
Raise
Buy

r/Uplifting News



r/Not the Onion



Sex
Arrested
Gun
Stop
Black
White
Claim
Want
Cop
Court
Drug
Accused
Texas
Florida
Study
Shooting
Death
Game

Highest scoring posts

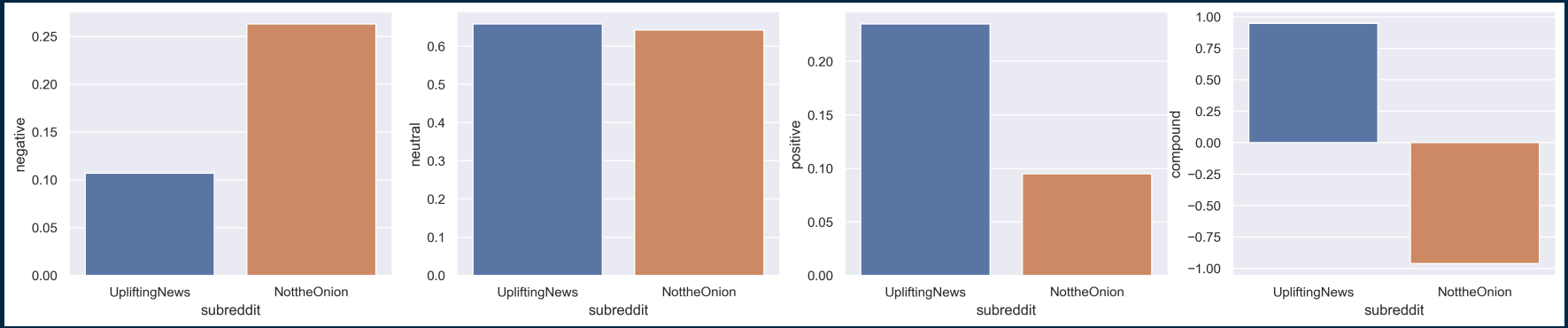
r/Uplifting News

title
Over a Million People Sign Petition Calling For KKK to Be Declared a **Terrorist** Group
when China **demands** names of airline's employees who **protested**, CEO lists only himself
Chattanooga's Police Chief has updated his department's Code of Conduct, saying they have a duty to stop others in the department from committing **illegal** activities including acts of **brutality** and **abuse** of authority.
Saudi Arabian heir to the crown has **declared war** on radical clerics, he also said "We are returning to what we were before, a country of moderate Islam that is open to all religions and to the world."

r/Not the Onion

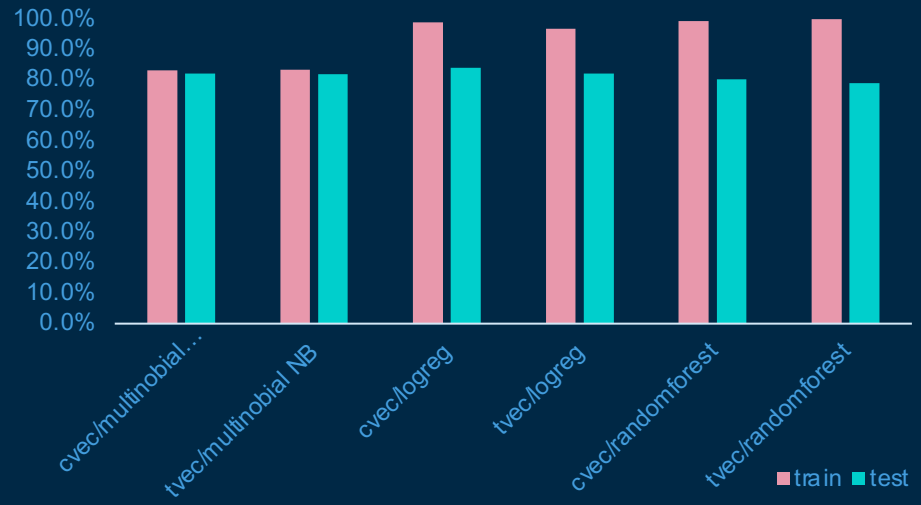
title
Nat Geo hires Jeff Goldblum to walk around, being professionally **fascinated** by things
Man **rescued** from Taliban didn't believe Donald Trump was President
Trump **dedicates** golf trophy to hurricane **victims**
'They’re going to get over it': Missouri Gov. insists kids must go back to **school** even though 'they will' get COVID-19

How are the sentiments of the titles?



Negative **Neutral** **Positive** **Compound**

Model testing



Preprocessing

tokenizer = RegexpTokenizer
lemmatizer = WordNetLemmatizer

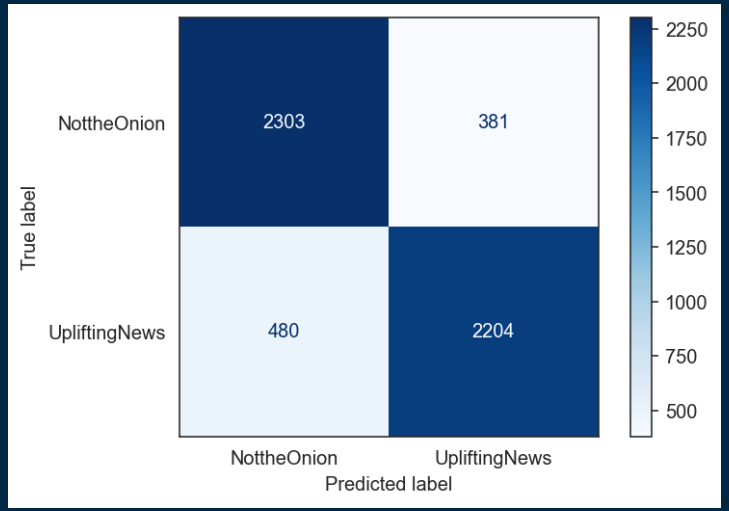
Transformers

- CountVectorizer
- TfidfVectorizer

Tested models

- Multinomial Naïve Bayes
- Logistic regression
- Random forest

Best prediction – Logistic Regression (overfit)



Baseline: 50%

CountVectorizer/LogReg

Accuracy : 84.0%

Precision : 85.4%

Highest Coefficient

Features	Coefficient
donates	8.752
rescued	8.540
muslim	8.528
cancer	7.048
victim	6.059
bullied	5.900
save	5.582
puppy	5.525
plastic	5.417
same sex	5.070

Lowest Coefficient

Features	Coefficient
say	0.148
tell	0.175
porn	0.186
poop	0.197
naked	0.208
admits	0.213
claim	0.217
jesus	0.218
warns	0.226
russia	0.236

Conclusion

What are some characteristics of high CTR news articles?

1. length – between 10-20 words
2. words
   - help, rescue, donate, animals
   - sex, crime, racial tension, police violence
3. sentiment
   - Either extreme positivity or extreme negativity (including sarcasm/absurdity) to evoke strong emotional response

However, due to the complex nature of language, no single superior model to generate the high CTR