Predicting Ads in TV News Closed Captions

Final Report

Identifying and Removing Ads

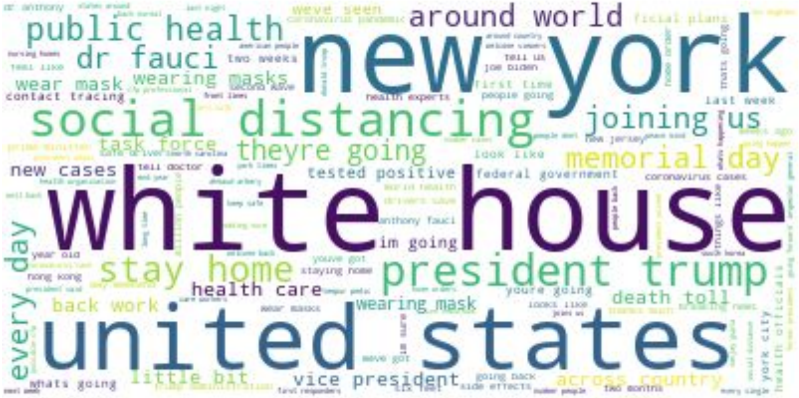
TV News closed captioning captions the ads as well as the news itself and when it’s contaminated with ads, it’s harder to extract information about the news itself.

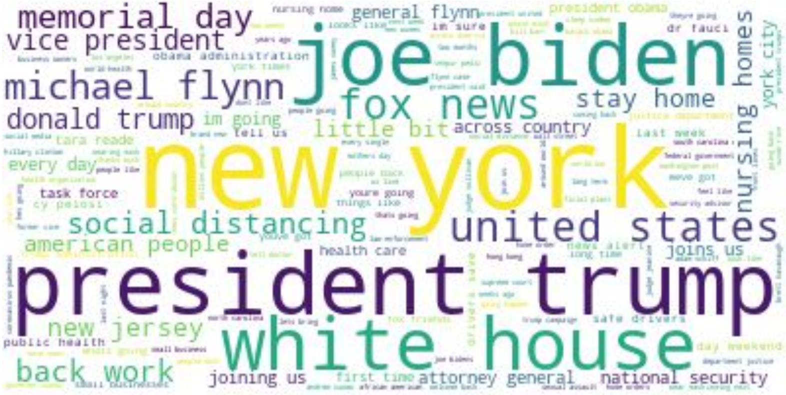
Knowing advertisers is also useful information for media watchdog groups, so identifying ads is an important processing step for NLP on TV News closed captions

Data Gathering and Wrangling

- Downloaded from archive.org’s collection of TV News closed captioning
- Focused on the last year of CNN and FOXNews shows
- Parsed XML metadata and HTML to create CSV of shows and snippets of text
- Parsed snippets into sentences

Word Clouds May 2020





CNN was still talking about Dr. Fauci, public health, and social distancing since the COVID-19 epidemic was still ongoing, while FOX was pushing stories about Michael Flynn

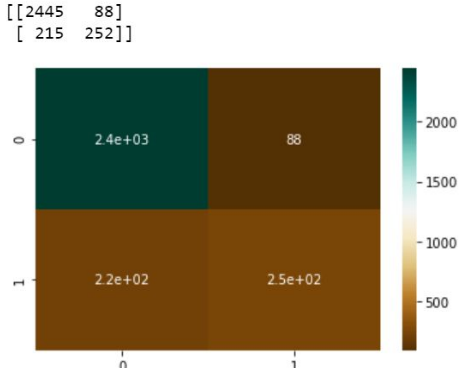
Hand-coding and Feature Creation

Steps to create features and aid in hand-coding:

- Lemmatize and use TFIDF vectorization to create words and bigrams
- KMeans clustering, then identifying ad and news clusters
- Create regression model based on KMeans clustering
- Apply model to snippets recorded output on sentences
- Apply topic modeling and record topic scores
- Create features to code for whether words like “applause” “welcome back” “up next” were nearby
- Hand-code 10,000 sentences each of FOXNews and CNN shows

Machine Learning

Classification models yielded better results for FOX than CNN, likely because FOX has fewer different advertisers.



```
[[2445   88]
 [ 215  252]]
```

Model	Accuracy CNN	Accuracy FOX
LogisticRegression	0.792	0.867
KNeighborsClassifier	0.750	0.863
SVC	0.715	0.844
LinearSVC	0.741	0.865
SGDClassifier	0.359	0.842
DecisionTreeClassifier	0.803	0.874
RandomForestClassifier	0.853	0.891
BaggingClassifier	0.822	0.899
GradientBoostingClassifier	0.786	0.880
AdaBoostClassifier	0.785	0.878

Feature Importance

Feature	Importance
start_snip	0.063977726
end_snip	0.060999728
snip_ad	0.025750646
president	0.012447353
easy	0.008710522
wa	0.005392661
doctor	0.004642851
think	0.004489554
topic_3	0.004409409
topic_17	0.004370745

CNN Top 10 Features

Feature	Importance
start_snip	0.044676518
snip_ad	0.044558297
end_snip	0.042547359
liberty	0.006252189
oh	0.005960919
car	0.005744653
easy	0.005208301
president	0.00491346
doctor	0.004716136
topic_73	0.004566933

FOX Top 10 Features

Further Testing the Models

- Testing the best models on 300 additional uncoded sentences that weren’t in the train or test set gave an accuracy of 95% for the CNN model and 86% for the FOX, which makes sense with such a small test size
- Testing the CNN model on FOX data gave an 88% accuracy, better than the CNN model on CNN data
- Testing the FOX model on CNN data gave a 75% accuracy

False Positives, False Negatives

CNN Falsely Identified Ads:

- question for you, is four years of college still worth it?
- us?
- bring it on o, bring it on.

CNN Falsely Identified News:

- with nine grams of protein and twenty-six vitamins and minerals.
- or current gm owners can get twenty seven fifty total cash allowance on this traverse.
- because, they really need their space.

FOX Falsely Identified Ads:

- of the party that have gained so much traction.
- for many nations, their sacrifice poured out in blood, courage and even death, to secure liberty for your enslaved children and to smashed here any, remains our moral touchstone.
- he's not here to defend himself

FOX Falsely Identified News:

- hey, that baker lady's on tv again.
- uncover the lost chapters of your family history with ancestry.
- i go on a trip.

Conclusions and Next Steps

Project envisioned as proof of concept which does have promise, but needs more and different feature engineering, as well as frequent model updates

Some potential next steps include:

- *Using entity extraction to identify brands and then creating a feature to show if sentences contain brand names or are near sentences that contain brand names*
- *Since ads are well predicted by the words in the sentence before them, try using a long short-term memory neural network*
- *Creating a web application to hand-checking and continually update the models, much the way that spam flagging works in an email system*