Clickbait Article Detection Using Deep Learning: These Results Will Shock You!

Kali Cornn, Department of Computer Science

Background

Social media utilizes clickbait headlines in order to lure users to read an article, since news outlets rely on users' clicks to generate revenue. Because such headlines are meant to be eye-catching, the article's content may not align up to readers' expectations.

Problem

Given an article headline from Reddit, classify it as "clickbait" or "non-clickbait".

Dataset

A total of 11,504 headlines were used for analysis.

- 9,522 headlines were from The Clickbait Challenge
- 1,982 headlines were scraped from Reddit (/r/savedyouaclick, /r/news, /r/worldnews).
- 5,751 (~50%) were clickbait
- 5,753 (~50%) were non-clickbait

Word cloud of words present in clickbait headlines

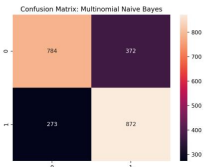
Word cloud of words present in non-clickbait headlines

Model Overview

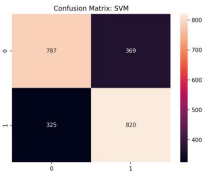
- Data split into train (60%) / dev (20%) / test (20%)
- Baseline models used **Term Frequency-Inverse Document Frequency (tf-idf)** as features
- Neural network models used **Global Vectors for Word Representation (GloVe)** as features
  - Adam optimization (learning rate = 0.0005)

Baseline Models

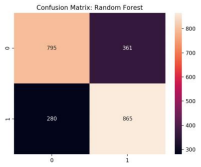
Multinomial Naïve Bayes



Support Vector Machines

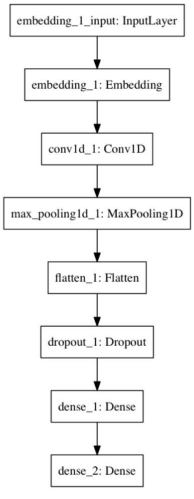


Random Forest Classifiers

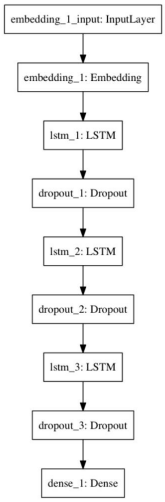


Neural Network Models

Convolutional Neural Network



Recurrent Neural Network



Results

Model	Accuracy	Precision	Recall	F1
MNB	0.720	0.721	0.720	0.719
SVM	0.698	0.700	0.698	0.698
RFC	0.721	0.723	0.721	0.721
CNN (w/o embed)	0.635	0.638	0.641	0.630
CNN (w/ embed)	0.729	**0.731**	0.728	0.723
CNN + LSTM (w/ embed)	0.710	0.705	0.723	0.707
RNN (w/ embed)	**0.738**	0.723	**0.770**	**0.742**

Discussion

- All models had > 50% accuracy (better than random guessing).
- Stemming headlines did not drastically improve model performance.
- CNN without embedding weights performed worse than did all baseline models.
- Train time for CNN models was much less than that of the RNN model
- RNN with embedding weights performed slightly better overall than all other models.

Future Work

- Utilize larger dataset for more diversity in headlines
- Use other features of articles (actual article text, presence of images, etc.).
- Further experiment with hyperparameter tunings and preprocessing methodologies.