Sarcasm Detection

Saba Suhail
July 20, 2022

Overview

- Sarcasm means opposite of what one wants to say
- To insult/irritate/being funny
- Limits of sarcasm not so well defined
- Sarcasm is subjective-> Wordplay
- Non-native speakers or readers may not get it
- **Goal:** To predict whether a headline is sarcastic or not

Dataset

- Taken from Kaggle available at(26709 entries):

https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

- Each record consists of three attributes:

  `**is_sarcastic**`: 1 if the headline is sarcastic 0 otherwise

  `**headline**`: the headline of the news article

  `**article_link**`: link to the original news article

- **Success Metrics:** Accuracy, precision, Recall, Confusion Matrix

Understanding the project

Step 1

EDA-Visualizations, Feature Engineering

Preprocessing-Stemming, Lemmatizing, Stopwords

Vectorization: Countvectorizer

Step 2

Classical ML Models

SelectKBest(100 features):

↓

Random Forest, Logistic Regression

PCA on all columns:

↓

Random Forest

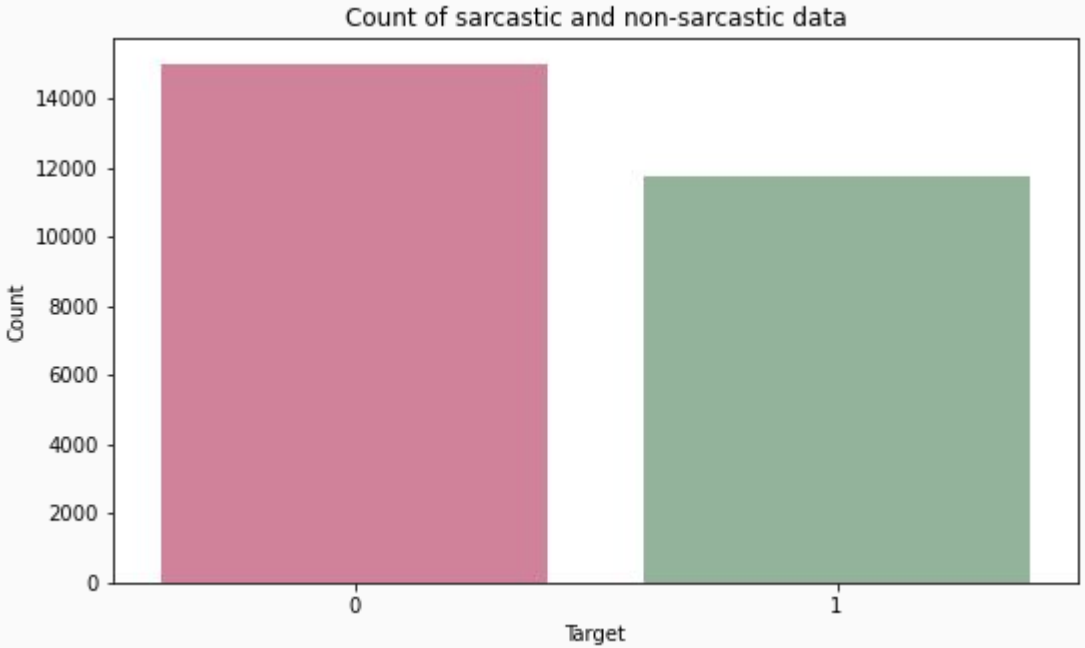
Step 3

Deep Learning

↓

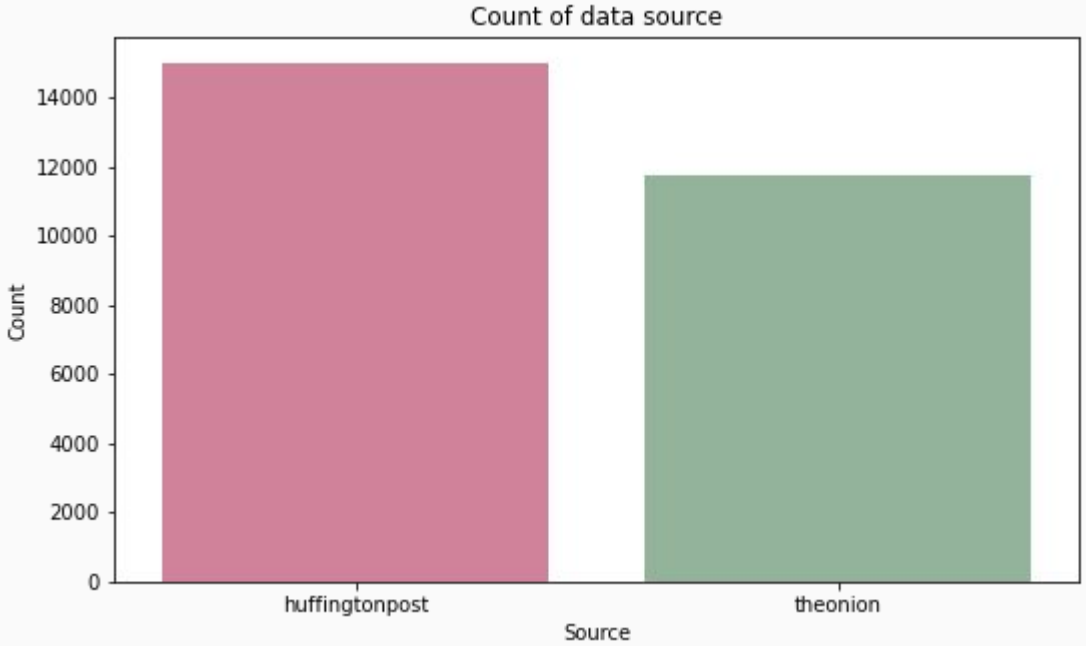
BERT

Distribution of target variable



Baseline:56%

Source of the Data



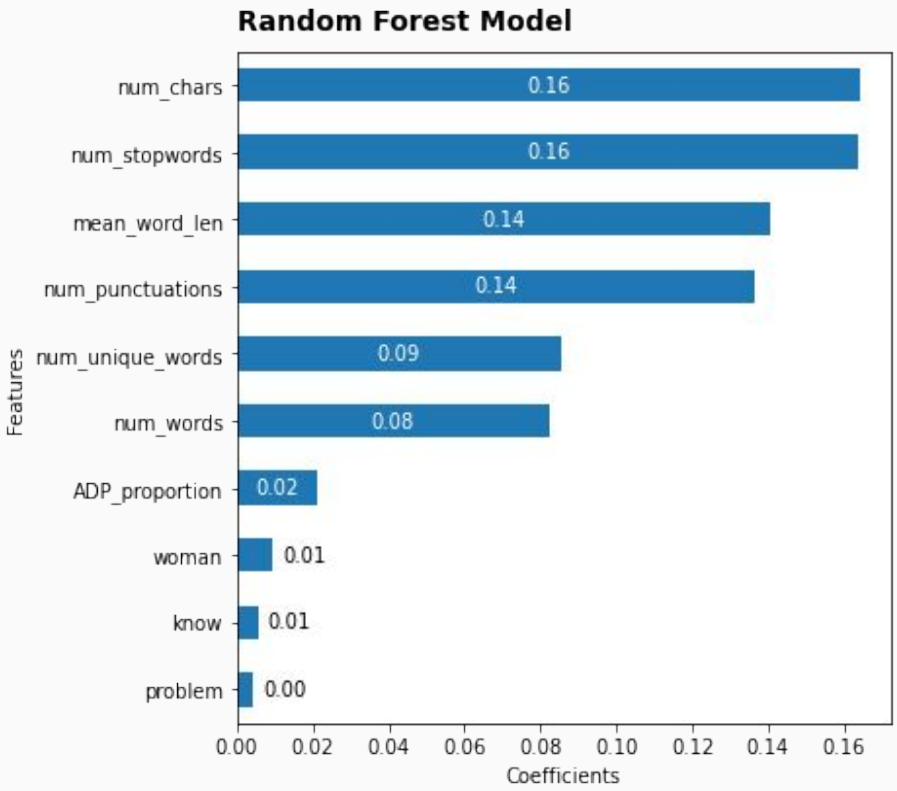
Sarcastic headlines come entirely from theonion

Non-sarcastic headlines come entirely from Huffington Post

Meta-features

- `num_words`
- `num_unique_words`
- `num_chars`
- `num_stopwords`
- `num_punctuations`
- `mean_word_len`

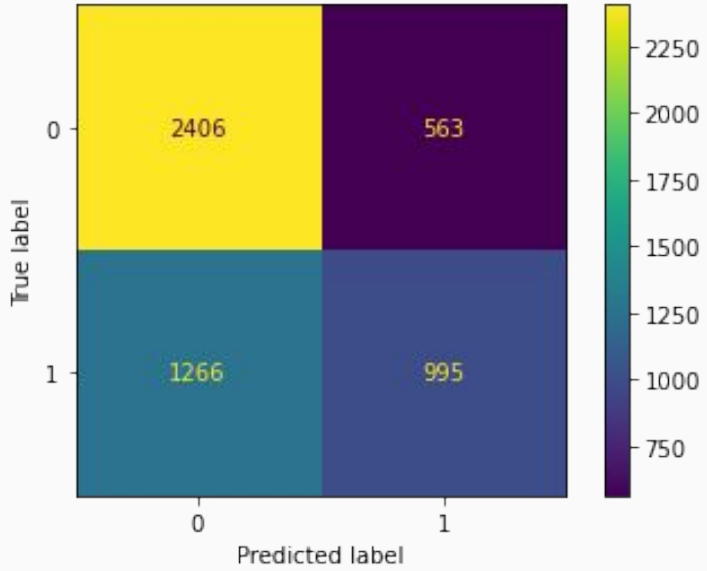
- Number of Punctuations in non-sarcastic headlines is noticeably more than that in sarcastic headlines which means it is the wordplay which creates sarcasm.
- Contrary to informal written language probably which uses more punctuation to create sarcasm.

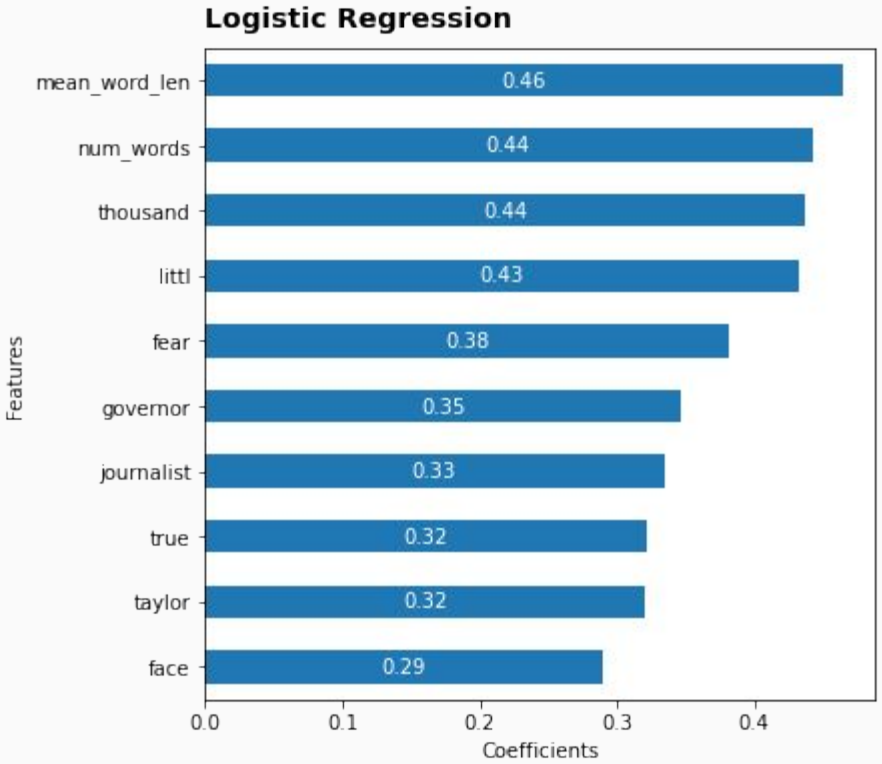


Random Forest on top 100 features

Random Forest Model Details

	precision	recall	f1-score	support
0.0	0.66	0.81	0.72	2969
1.0	0.64	0.44	0.52	2261
accuracy			0.65	5230
macro avg	0.65	0.63	0.62	5230
weighted avg	0.65	0.65	0.64	5230

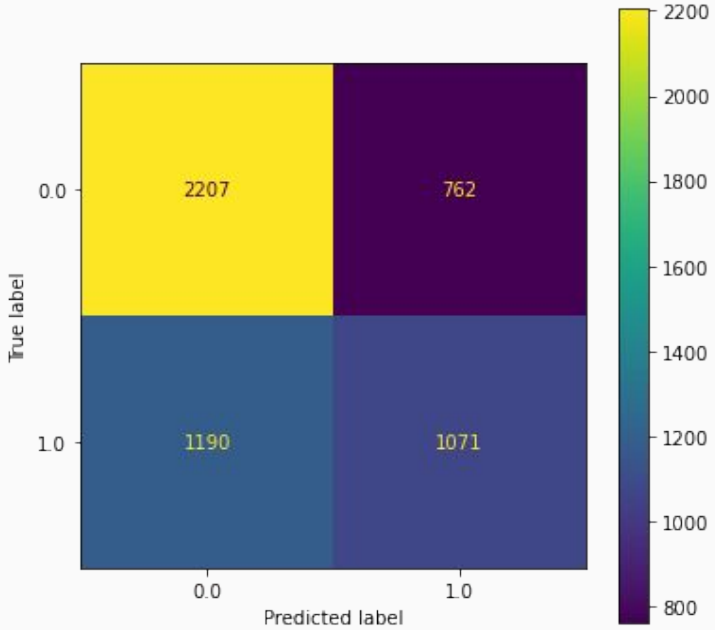




Logistic Regression on top 100 features

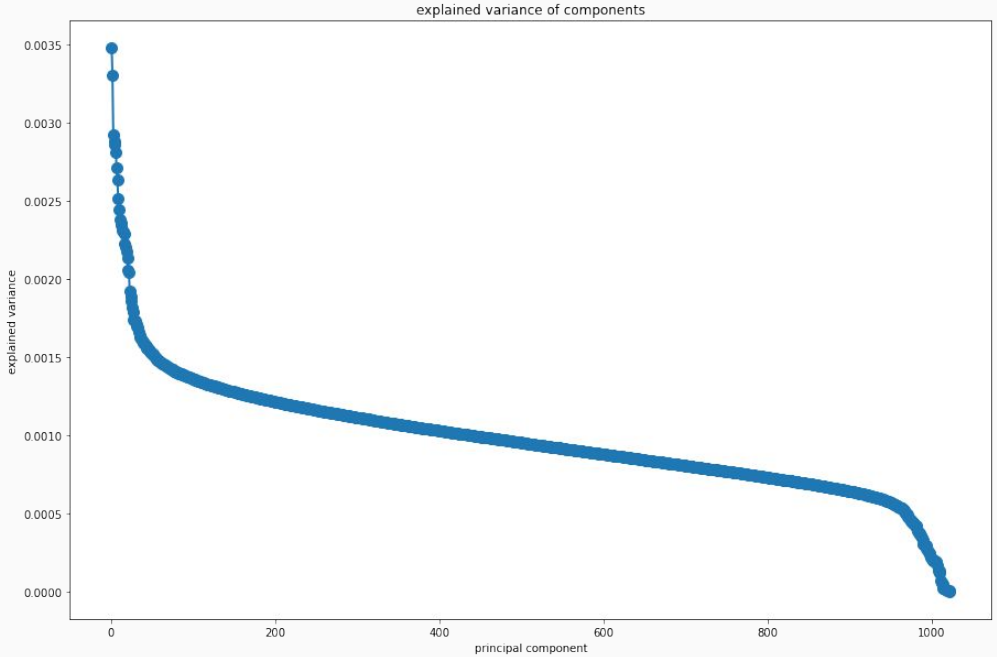
Logistic Regression Model Details

	precision	recall	f1-score	support
not_sarcastic	0.65	0.74	0.69	2969
is_sarcastic	0.58	0.47	0.52	2261
accuracy			0.63	5230
macro avg	0.62	0.61	0.61	5230
weighted avg	0.62	0.63	0.62	5230

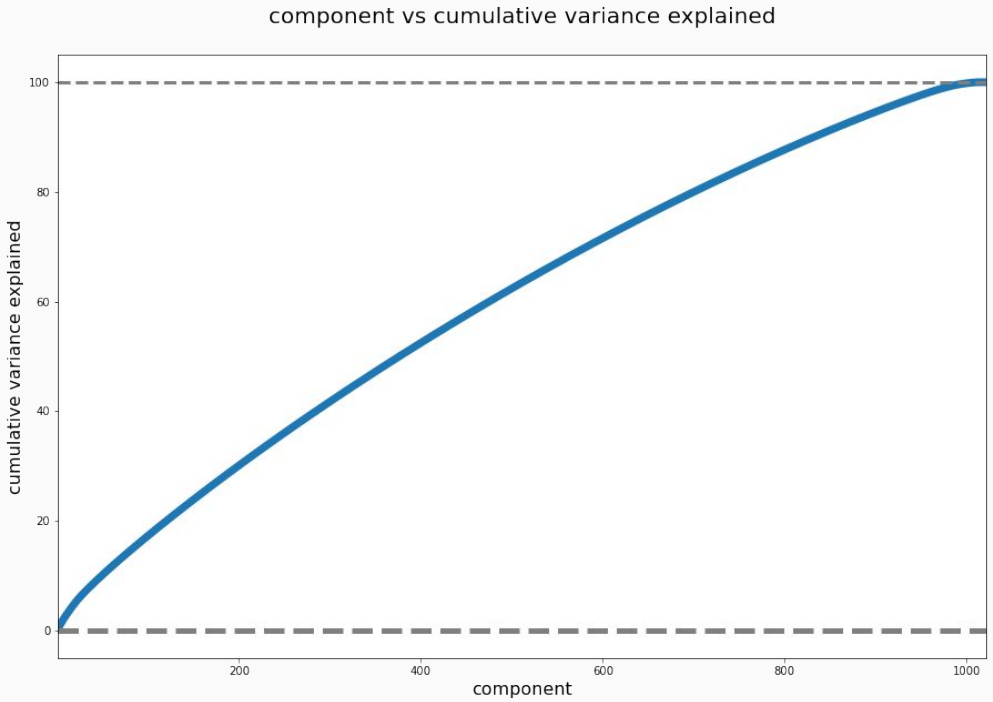


PCA

- Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.
- Tried PCA for dimensionality reduction and having those features which have high explained variance.



PCA



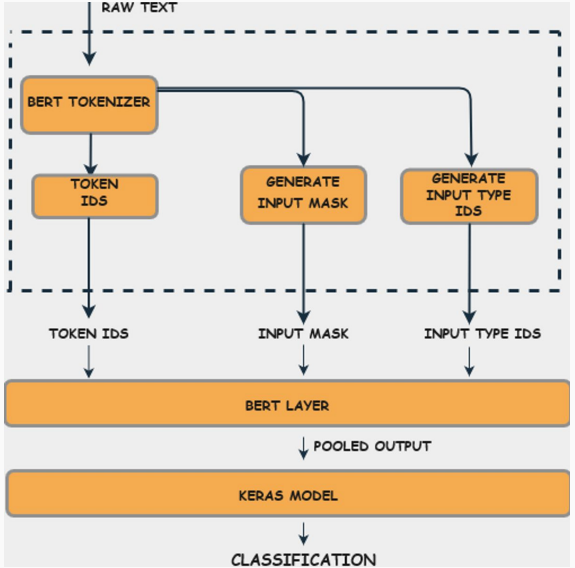
Accuracy for Random Forest Model is 58% just 2% above baseline.

ML Models for NLP

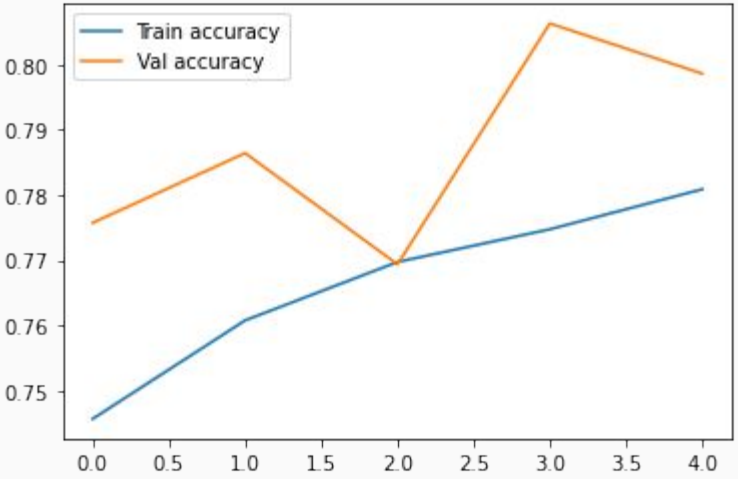
- “Machine Learning” really means “machine teaching.” We know what the machine needs to learn, so our task is to create a learning framework and provide properly-formatted, relevant, clean data for the machine to learn from.
- A machine learning model is the sum of the learning that has been acquired from its training data. The model changes as more learning is acquired.
- There are three aspects to any given chunk of text:
  - **Semantic Information:** Semantic information is the specific meaning of an individual word.
  - **Syntax Information:** The second key component of text is sentence or phrase structure, known as syntax information
  - **Context Information:** one must understand the context that a word, phrase, or sentence appears in
- Purely rules-based text analytics is a dead-end: encountered in classical ML methods
- Certain aspects of machine learning are very subjective. You need to tune or train your system to match your perspective. Therefore, BERT.

BERT

- *"BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training."*
- Bidirectional model
- This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).



BERT Model



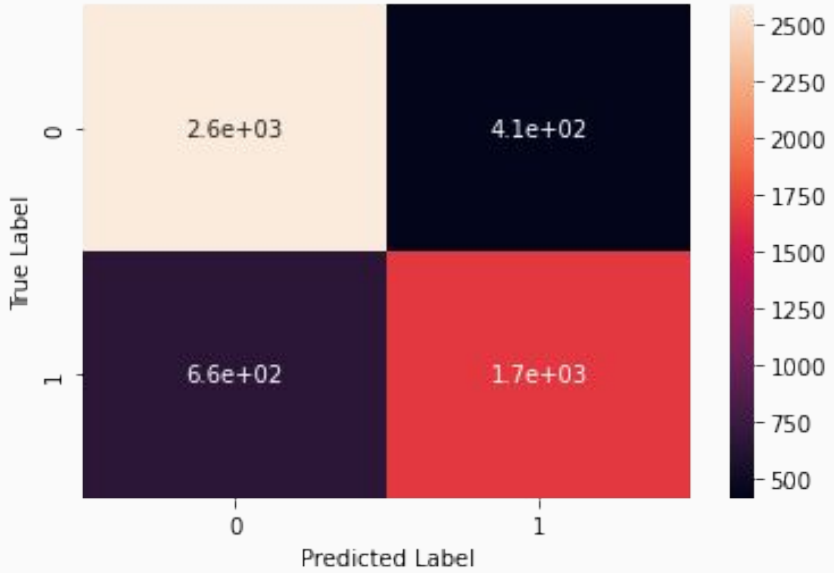
-The best test accuracy is achieved at epoch=3.

- Regularization mechanisms, such as Dropout and L1/L2 weight regularization, are turned off at testing time. They are reflected in the training time loss but not in the test time loss. Also, the testing loss for an epoch is computed using the model as it is at the end of the epoch, resulting in a lower loss.

BERT Details

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.86	0.83	2997
1	0.80	0.72	0.76	2345
accuracy			0.80	5342
macro avg	0.80	0.79	0.79	5342
weighted avg	0.80	0.80	0.80	5342



Conclusion

Conclusion : Context matters for sarcasm and hence BERT seems apt for text classification and performed best among the models in this project.

Recommendation: BERT should be the production model.

Limitations: Only applicable for English language

Next Steps: Explore more contextual neural nets and techniques

Applications

- This analysis can be used to design intelligent chatbots
- Can help in adding emotions into voice detection systems
- E-commerce websites can use such it to have a better understanding of product/service reviews posted by users