Predicting Subreddits:

Classification of Reddit Posts using Natural Language Processing

Presented by Rebecca Wright

Telling truth from fan-fiction...

Can a Natural Language Process classifier determine if a reddit post is talking about an individual's personal supernatural experience, or if the speaker is talking about the paranormal television series "Supernatural".

Understanding the Subreddits:

r/Thetruthishere

r/Supernatural

r/Thetruthishere

A subreddit community of 375k members, chronicling non-fiction personal experiences with the unknown.

r/Supernatural

A 167k member community, dedicated to the CW television show Supernatural.

Web-Scraping Reddit and Building the Dataset

From 20,000 raw posts and over 80+ available features, data cleaning and feature selection reduced the dataset to 13,000+ posts featuring:

- Reddit post “selftext” : *the text content of the post*
- Reddit post “subreddit” : *the subreddit classifier*

Data Cleaning and Preprocessing

Data Cleaning

Removed rows of values:

- [removed]
- [deleted]
- *Nans*

This eliminates previous posts by no-longer active members.

Preprocessing

Cast text to lowercase.

Removed:

- punctuation
- digits
- generic stopwords
- 1-2 letter words

Lemmatized the text.

The Null Model

Establishing a baseline

Using the “most frequent” classifier, the null model has a baseline accuracy score of

66%

Building Models

Vectorizers

Count Vectorizer

Term-Frequency-and-Inverse-Document-Frequency Vectorizer

Classifiers

Multinomial Naive Bayes

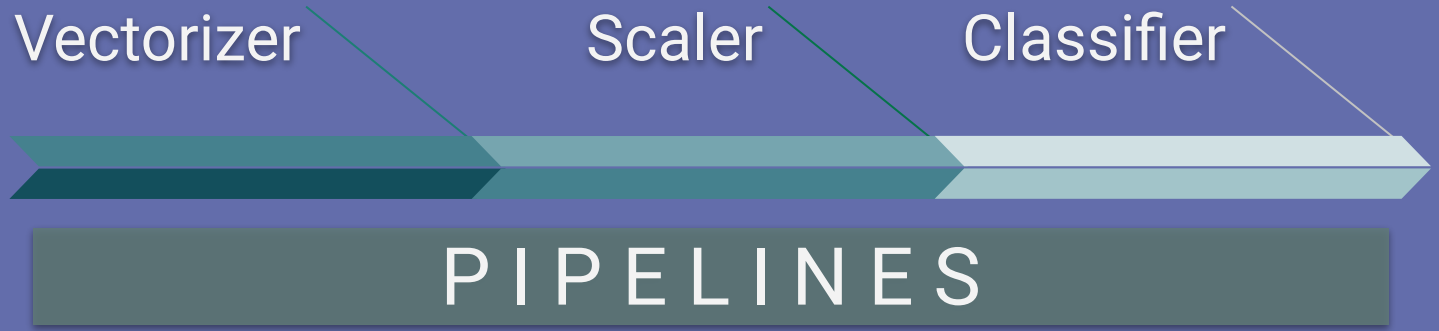
Logistic Regression

Decision Tree Classifier

Bagging Classifier

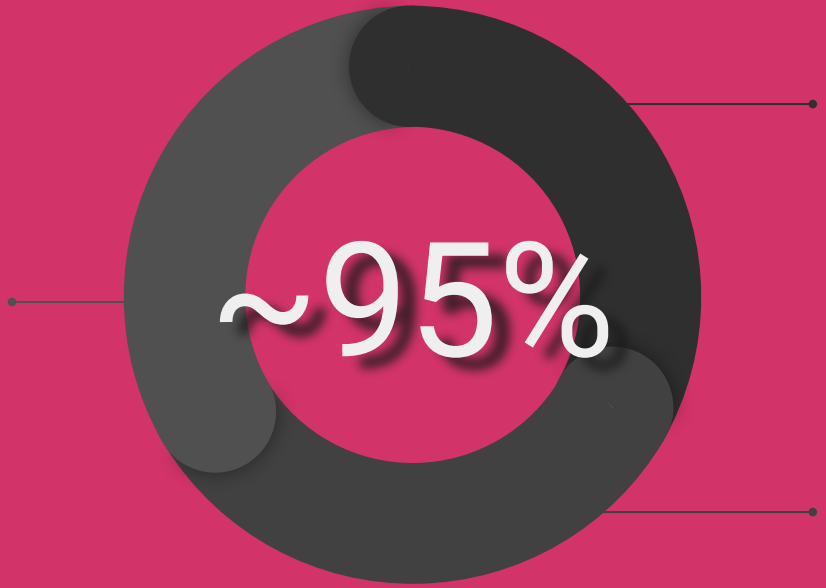
Random Forest Classifier

Building Models



Model Performance:

Accuracy Scores



Logistic Regression

Using either CountVectorizer or Tfidf Vectorizer

Bagging Classifier

Using either CountVectorizer or Tfidf Vectorizer

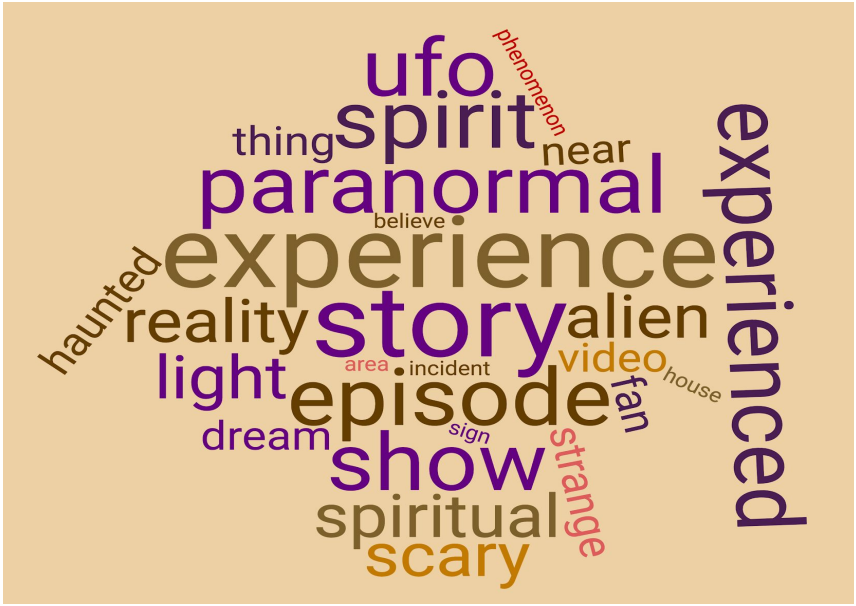
Multinomial Naive Bayes Classifier

Using either CountVectorizer or Tfidf Vectorizer

Why so many classification models performed so well ...

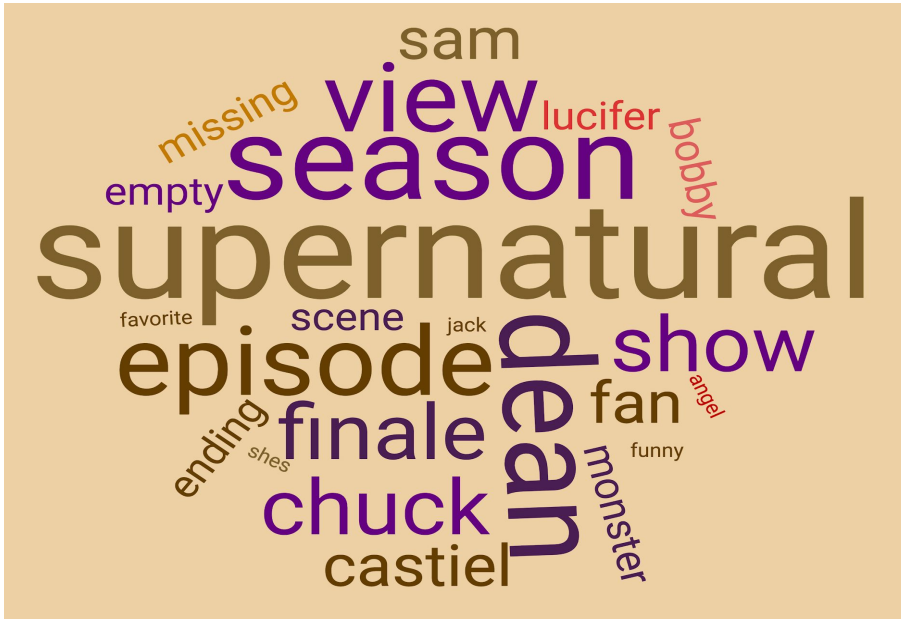
The profound impact of unique expressions on classification models.

Top Predictive Words r/Thetruthishere



- experience
- story
- paranormal
- experienced
- ufo
- video
- reality
- spiritual
- light
- scary
- alien
- dream
- thing
- strange
- haunted
- near
- incident
- believe
- area
- sign
- phenomenon
- house

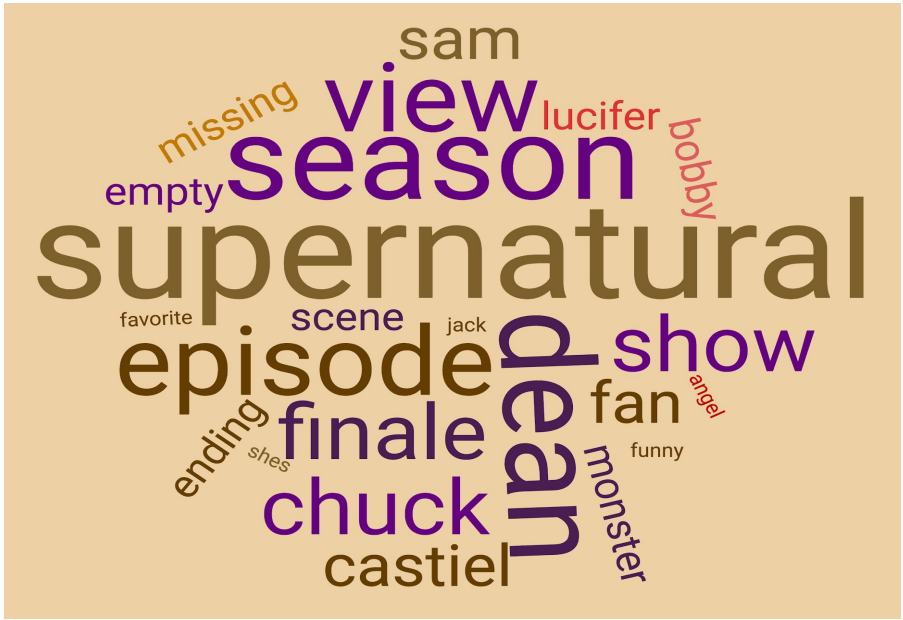
Top Predictive Words
r/Supernatural



- season
- dean
- supernatural
- episode
- show
- sam
- finale
- chuck
- ca
- winchester
- ending
- character
- scene

Supernatural Proper Nouns

Top Predictive Words
r/Supernatural



- season
- dean
- supernatural
- episode
- show
- sam
- finale
- chuck
- ca
- winchester
- ending
- character
- scene

Television Terms

Summary Findings

Even though the television show Supernatural and the non-fiction encounters as outlined in r/Thetruthishere both cover supernatural content and events, unanticipated expressions can exert surprising predictive influence on the classifier models.

Further study should be devoted to expanding the list of words to be stripped from a post to include proper nouns of the television show as well as television industry jargon to further evaluate the predictive capabilities of these classifier models.

Supplemental Observations:

The posting habits of different subreddit communities

An initial 10,000 posts were scraped from both subreddits.

r/Supernatural stats:

- Posts were made between 1/14/21 and 7/13/20.
- After cleaning to remove deleted posts or posts from discontinued users, 46% of posts remained.

r/Thetruthishere stats:

- Posts were made between 1/14/21 and 9/25/16.
- After cleaning, 88% of posts remained.

Members of r/Thetruthishere post less, but their posts last longer.