This is a Joke

An Exploration of Natural Language Processing

By Molly Baird

Problem Statement

Can we use natural language processing to distinguish between two joke-themed subreddits?

- /r/Jokes
- /r/DadJokes
- /r/MommaJokes
- /r/CleanJokes
- /r/DirtyJokes

Jokes vs DadJokes

I've run out of toilet paper and started using old newspapers instead...

I've run out of toilet paper and started using old newspapers instead...

...the times are rough

Posted on **/r/DadJokes** by u/HellsJuggernaut

Model Performance

-Size of dataset: 9,126

-Submission title and selftext only

-No stopword removal

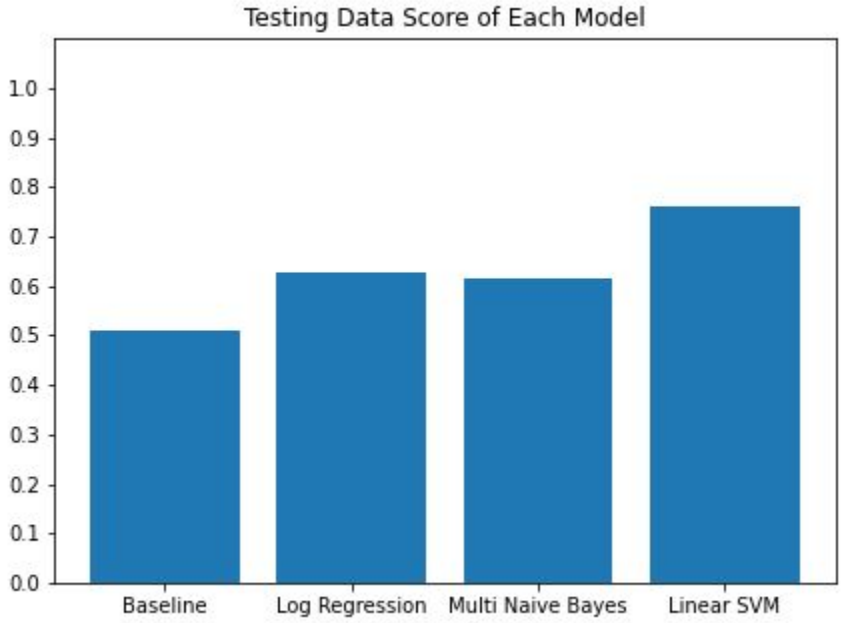
-N-gram range (1,1)

-Max features 500

-Lemmatizing/Stemming had little to no effect

-CountVectorize vs Tfidf had little to no effect

-Linear Support Vector Machine 76% accurate



Model Performance

-Size of dataset: 9,126

-Submission title and selftext only

-No stopword removal

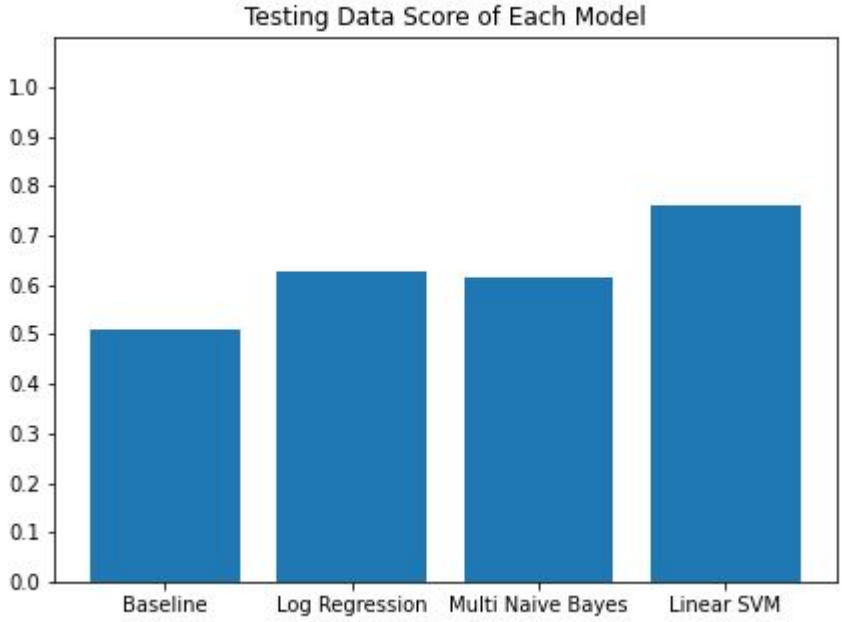
-N-gram range (1,1)

-Max features 500

-Lemmatizing/Stemming had little to no effect

-CountVectorize vs Tfidf had little to no effect

- Linear Support Vector Machine 76% accurate



Why isn't this model better?

My friend and I just started a business

where we weigh tiny objects...

...it's a small scale operation

What's the difference between juvenile humor

and dad jokes?

A dad joke is full groan

Posted on /r/Jokes by u/beenthereonce2

Jokes vs MommaJokes

Yo momma so fat...

Yo momma so fat...

...she got counted twice in the census

Posted on /r/MommaJokes by u/luxdesigns

Model Performance

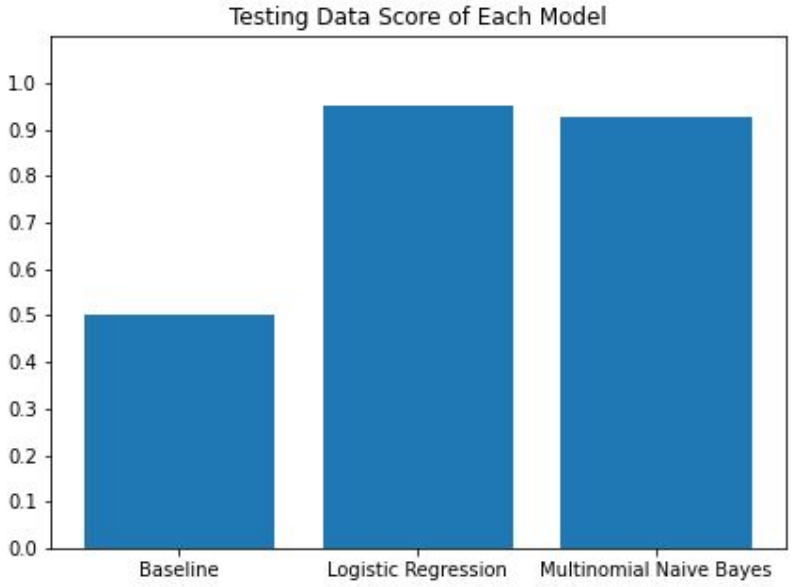
-Size of dataset: 2,744

-Submission title and selftext only

-No stopword removal

-N-gram range (1,2)

-Max features 500



CleanJokes vs DirtyJokes

Why do people in athens hate getting up so early?

Why do people in athens hate getting up so early?

Because dawn is tough on greece.

What kind of bees make milk?

What kind of bees make milk?

BooBees.

Posted on /r/DirtyJokes by u/TheDisappointment101

Model Performance

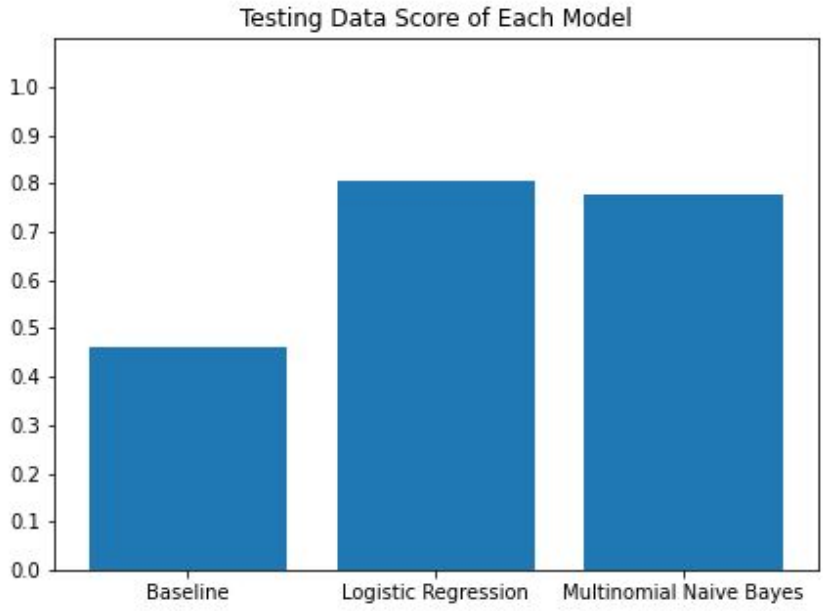
-Size of dataset: 9,276

-Submission title and selftext only

-English stopword removal

-N-gram range (1,1)

-Max features 500



Conclusions and continuations

- Jokes with standard phrasing are easier to classify
- Dad joke is not well defined
- Stemming/lemmatizing didn’t really help much
- Want to incorporate comments
- Bigger datasets?
- Want to know which jokes were misclassified
- Want to know which words got the most weight
- Multiclass model with all joke subreddits

Thank you!