NLP Subreddit Classification

Sean Cleary
October 9, 2020

Overview

- Natural Language Processing (NLP) is a form of Machine Learning that uses tokenization, lemmatization, and stemming to classify and extract sentiment from text. There are many other applications of NLP that are outside the scope of this project.
- Can a classification model using posts from two subreddits accurately identify which category a post belongs to?

Understanding the problem

r/wallstreetbets

- A subreddit that focuses on the futures/options exchanges.

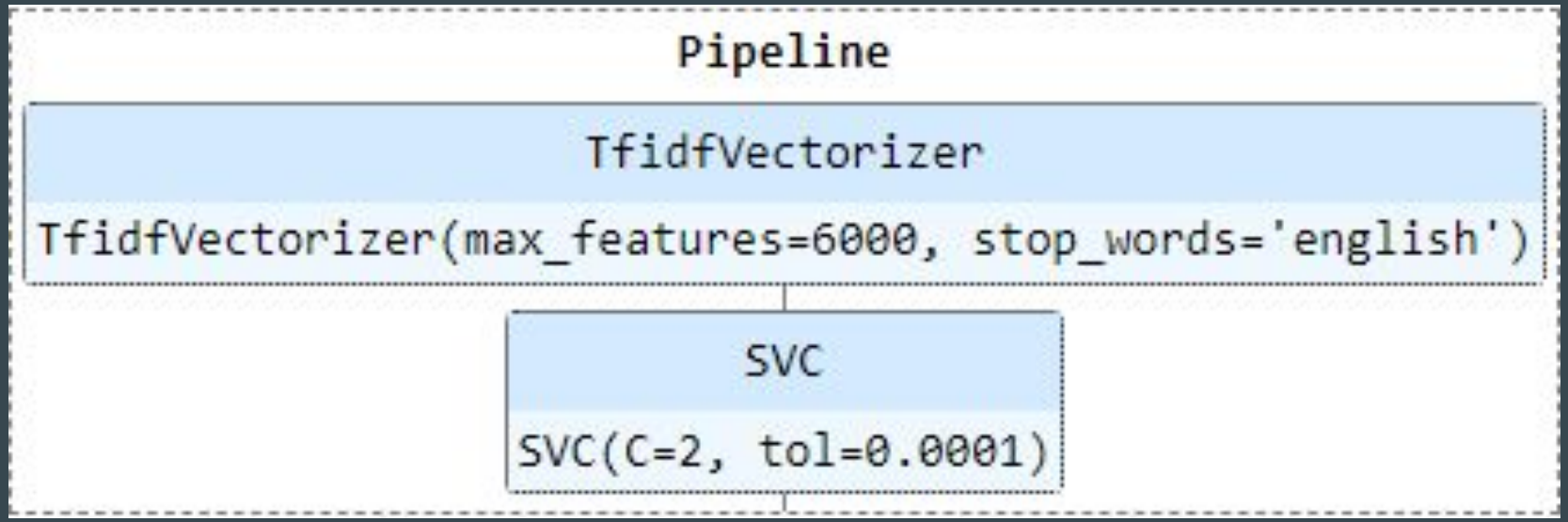
- Could be mistaken as a gambling subreddit.

r/MachineLearning

- A subreddit that focuses on Machine Learning and all of its glory.

-Great for finding new resources/techniques in the field of ML.

Model Selection

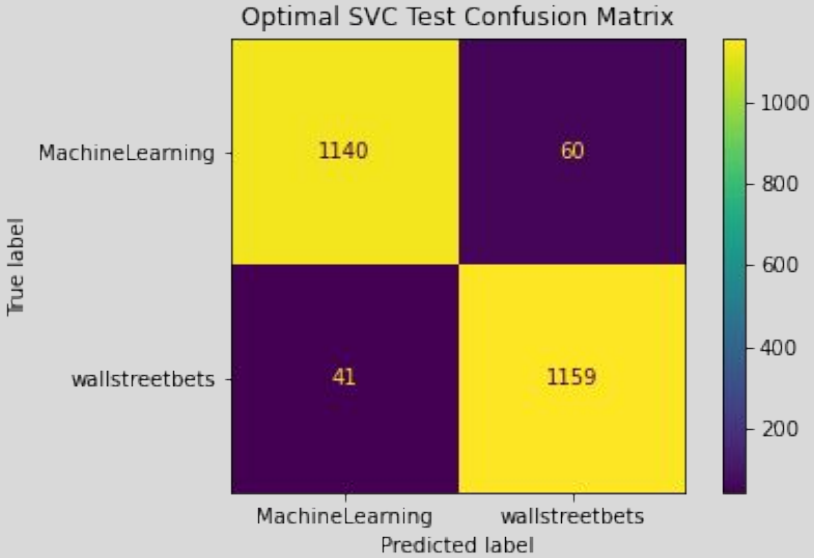


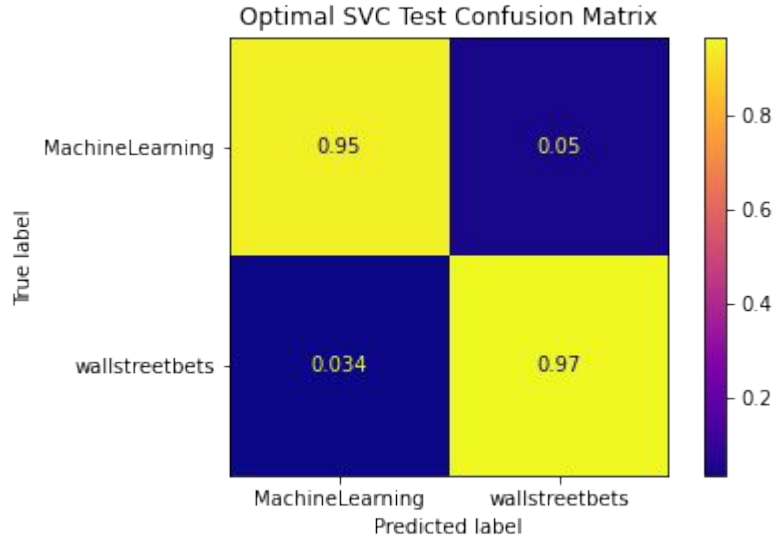
TfidfVectorizer

-Term Frequency Inverse Document

-Used to measure how important a term is in relation to document frequency.

Confusion Matrix





Classification Report

	precision	recall	f1-score	support
MachineLearning	0.97	0.95	0.96	1200
wallstreetbets	0.95	0.97	0.96	1200
accuracy			0.96	2400
macro avg	0.96	0.96	0.96	2400
weighted avg	0.96	0.96	0.96	2400

Demonstration:

Wall Street Bets:
“Why you may want to get into GME now. As in right now. After hours. Even at the top.”
“I know more pain than you, yet I told myself do not sell it will pop for sure”

Machine Learning:
“Forex machine learning strategy with Python: features processing”
“Deconstructing Nvidia Maxine”

Next Steps:

-Remove specific words from the dataset.

-Investigate my missing token in frequency weights.

-Apply this to a NN and see if deep learning can increase performance further.

-Let me know if you have any questions.