Natural Language Processing

By
Dr Monday Oshoikpor

Presentation Outline

- Problem Statement
- Data Collection
- Data Pre-Processing
- Model Analysis
- Conclusions & Recommendations

Problem Statement

- The management at Reddit want to increase their ability to track posts that include hate speech, posts that incite violence, and posts that might include potential real life threats.
- I will present a solution that demonstrates the value of machine learning applied to natural language processing to solve this problem.
- Given a post from one of two subreddits, my solution will correctly classify whether or not a post came from the r/LanguageTechnology thread.

Data Collection

I selected and collected submissions from two subreddits r/LanguageTechnology and r/videos

	subreddit	title
0	LanguageTechnology	Looking for a table to text codebase
1	LanguageTechnology	Allennlp: What in the frig is a Predictor?
2	LanguageTechnology	Just finished my first proper NLP project
3	LanguageTechnology	T-V Distinction Classifier
4	LanguageTechnology	Styleformer performance. Or anything that turn...
...	...	...
95	videos	Found a anti-abortion video whose comment sect...
96	videos	Battlefield 2042
97	videos	Mtn Dew Frostbite Review (Ft. Mighty Meat Scep...
98	videos	Can you solve these riddles in less than 7 sec...
99	videos	World Record Progression: The Deadlift

200 rows × 2 columns

Data Cleaning

Cleaning:

- Removed duplicates
- Removed non-standard characters like #\ #\
- After cleaning: 200 => 196 records

Preprocessing:

- Lemmatization
- Add non-meaningful words from the dataframe to the stop words dictionary to eliminate words that are so commonly used that they carry very little useful information.
- Train/test split (used .33 test, stratify)
- Classes are balanced, 50%

Data Pre-Processing

Most Frequent Words Lists

LanguageTechnology

nlp	25
model	20
text	15
new	9
data	9
language	9
bert	7
looking	6
transformer	6
processing	6
training	6
using	6

videos

video	6
home	4
official	3
internet	3
baby	3
movie	3
talk	3
best	3
dog	3
like	3
mod	3
trailer	3
tour	3
reddit	3
new	3

Modeling Analysis

I ran three models to see which one worked best:

1) **Logistic Regression with CountVectorize**
   - Train Score GS1B: 0.8682170542635659
   - Test Score GS1B: 0.7692307692307693
2) **KNN with Tfidf**
   - Train Score GS3A: 0.7209302325581395
   - Test Score GS3A: 0.6923076923076923
3) **Random Forest Classifier**
   - Train Score GS4A: 0.9224806201550387
   - Test Score GS4A: 0.7384615384615385

Conclusions & Recommendations

- The best scoring model is the Logistic Regression with CountVectorizer at 77%.
- The algorithms beat the baseline accuracy, but Logistic Regression with CountVectorizer scored higher than KNN with Tfidf and Random forest Class.
- The objective to train a model to predict with 77% accuracy those posts belonging to r/LanguageTechnology was met.

Conclusions & Recommendations

To improve the model, I recommend getting more submissions and using different models to score the algorithms.

References

- https://hbr.org/2021/01/social-media-companies-should-self-regulate-now