Subreddit Classification & NLP

Distinguishing between r/TIFU & r/AITA

u/youngluck u/acidtwist

The Question

How well can we distinguish between post belonging to one of the two popular subreddits;

/r/TIFU	14.2m subs
/r/AmITheA**hole	632k subs

/r/TIFU

Short for 'Today, I F*ed Up', TIFU is always there for your daily dose of cringe, and a good laugh after it all.**

➔ **Honest**
Users share recent embarrassing experiences in painful detail

➔ **Humble**
Posters understand their causal role in the catastrophe

➔ **Humorous**
Playful banter and schadenfreude are the main attraction

/r/amithea**hole

Often abbreviated AITA, this subreddit serves as a court of public opinion to those desperate enough to seek the approval of random strangers on the internet.

- ➔ **Desperate**
  If you're comin' here for help, this isn't the only kind you need.
- ➔ **Judgemental**
  Readers get to play judge Judy
- ➔ **Scandalous**
  I'm just here for the trainwreck

Collecting the Data

1. Iteratively acquire .JSONs for each subreddit using the Reddit API
2. Extract the text portion from each post and assemble a corpus
3. Remove irrelevant HTML or markdown substrings (e.g. \n, ‘www’)
4. Tokenize strings and vectorize documents to obtain a bag of words
5. Bob has now become your uncle!

CountVectorizer vs TF-IDF

Using performance of a Logistic Regression Classifier for comparison:

```
cvec_params={
    'cvec__max_features':[3000,4000,5000],
    'cvec__max_df':[.9,.94,.98],
    'cvec__min_df':[2,4,6],
    'cvec__ngram_range':[(1,1),(1,2),(1,3)],
    'logreg__penalty':['l1','l2'],
    'logreg__C':[.4,.6,.8,1],
}
```

```
tfidf_params={
    'tfidf__max_features':[3000,4000,5000],
    'tfidf__max_df':[.9,.94,.98],
    'tfidf__min_df':[2,4,6],
    'tfidf__ngram_range':[(1,1),(1,2),(1,3)],
    'logreg__penalty':['l1','l2'],
    'logreg__C':[.4,.6,.8,1],
}
```

95.88% Accuracy

90.44% Accuracy

CountVectorizer vs TF-IDF

Using performance of a Logistic Regression Classifier for comparison:

```
cvec_params={
    'cvec__max_features':[3000,4000,5000],
    'cvec__max_df':[.9,.94,.98],
    'cvec__min_df':[2,4,6],
    'cvec__ngram_range':[(1,1),(1,2),(1,3)],
    'logreg__penalty':['l1','l2'],
    'logreg__C':[.4,.6,.8,1],
}
```

95.88% Accuracy

```
tfidf_params={
    'tfidf__max_features':[3000,4000,5000],
    'tfidf__max_df':[.9,.94,.98],
    'tfidf__min_df':[2,4,6],
    'tfidf__ngram_range':[(1,1),(1,2),(1,3)],
    'logreg__penalty':['l1','l2'],
    'logreg__C':[.4,.6,.8,1],
}
```

90.44% Accuracy

What is /r/TIFU saying?

	Word	AITA Count	TIFU Count	AITA Tf-Idf Position	TIFU Tf-Idf Position	difference
2018	forget	5	41	2672	561	2111
1920	grabbed	6	65	2437	361	2076
1918	finger	6	64	2434	365	2069
1930	tongue	6	47	2463	502	1961
2011	wet	5	26	2650	847	1803
1891	corner	6	37	2369	624	1745

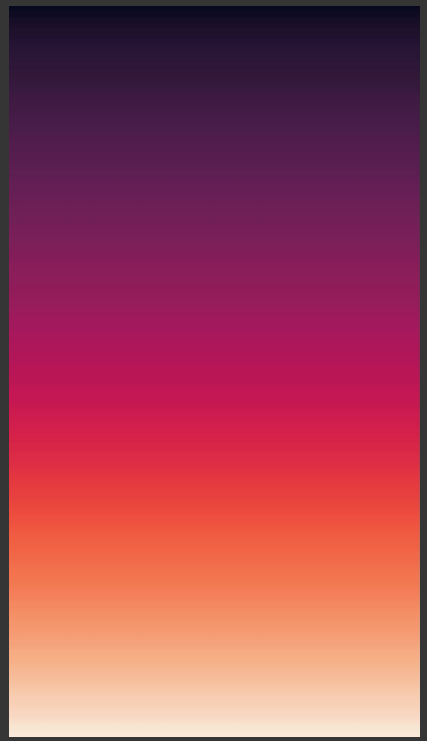
What is /r/AITA saying?

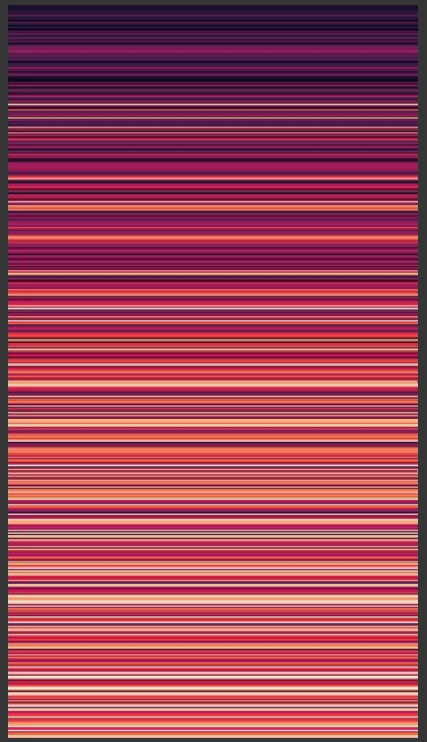
	Word	AITA Count	TIFU Count	AITA Tf-Idf Position	TIFU Tf-Idf Position	difference
470	comfortable	46	6	473	2673	-2200
471	cook	46	6	474	2653	-2179
774	advantage	26	5	791	2950	-2159
2018	forget	5	41	2672	561	2111
790	responsibility	25	5	808	2903	-2095
1920	grabbed	6	65	2437	361	2076

Relative Term Frequency

Words in:
/r/AITA

sorted by
TF-IDF in
/r/AITA





Words in:
/r/TIFU

sorted by
TF-IDF in
/r/AITA

Modelling: Accuracy

Model:	Train:	Test:
Multinomial Naive Bayes	95.2%	91.3%
Logistic Regression	98.1%	95.2%
K Nearest Neighbours	100%	59.6%
AdaBoost (Decision Tree)	100%	96.1%
Random Forest	94.1%	94.4%
Support Vector Classifier	99.5%	95.0%
Bagging Classifier (Decision Tree)	99.5%	94.1%

Modelling: Accuracy

Model:	Train:	Test:
Multinomial Naive Bayes	95.2%	91.3%
Logistic Regression	98.1%	95.2%
K Nearest Neighbours	100%	59.6%
AdaBoost (Decision Tree)	**100%**	**96.1%**
Random Forest	94.1%	94.4%
Support Vector Classifier	99.5%	95.0%
Bagging Classifier (Decision Tree)	99.5%	94.1%

Winner Winner Chicken Dinner

An adaptively boosted decision tree classifier can differentiate between posts made to the subreddits /r/TIFU and /r/AmITheAsshole with ~96% accuracy. Great success!

AdaBoost (Decision Tree) Parameters:

Learning rate: 0.8

Number of Estimators: 60

Confusion Matrix:

	/r/TIFU post	/r/AITA post
Predicted /r/TIFU	**190**	**4**
Predicted /r/AITA	**10**	**153**

Thank You!