Spirituality vs. Religion

A Classification Solution

by Joseph Tay

RELIGION OR SPIRITUALITY?

DUALITY	ONENESS
DIVINE POWER IS OUTSIDE	DIVINE POWER IS INSIDE
A BELIEF	A FEELING
KNOWLEDGE	WISDOM
SOMEONE ELSE'S EXPERIENCE	YOUR OWN INDIVIDUAL EXPERIENCE
DOGMATIC	OPEN TO INTERPRETATION
CALLS FOR SIMPATHY	CALLS FOR EMPATHY
DRIVES SEPARATENESS	DRIVES CLOSENESS
GENERATES BIGOTRY	GENERATES TOLERANCE
DEMANDS SOBRIETY, CONTROL & ASCETIC BEHAVIOR	ALLOWS LAUGHTER, FREEDOM & CONSCIOUS ENJOYMENT
SELF IMPROVEMENT THROUGH REPRESSION AND DENIAL	SELF GROWTH THROUGH LOVING ACCEPTANCE OF WHAT IS
FOCUSES ON OTHER'S EXPERIENCES AND GROWTH	FOCUSES ON ONE'S OWN EXPERIENCE AND GROWTH
LOVING TOWARDS OTHERS/ABNEGATION/PUSHOVER	LOVING TOWARDS SELF/SELFISHNESS/ASSERTIVE
FOR PEOPLE WHO ARE AFRAID OF THE DARK	FOR PEOPLE WHO'VE ALREADY BEEN THERE
FEAR-BASED	LOVE-BASED

HEALING MAGICK

> Religion is belief in someone else's experience. Spirituality is having your own experience. – Deepak Chopra

Problem Statement

- There have been frequent postings from religious groups that try to disrupt the discussion and senior forum members wants to stop such postings in the forum
- To develop a quick way to allow the moderators to identify such posts **accurately** and stop them from being posted

Approach:

- Using data scrapped from subreddits **/r/spirituality** and **/r/religion**, a **classification model** together with NLP will be trained to predict **Spirituality as the positive class** and **top prediction features** to be identified

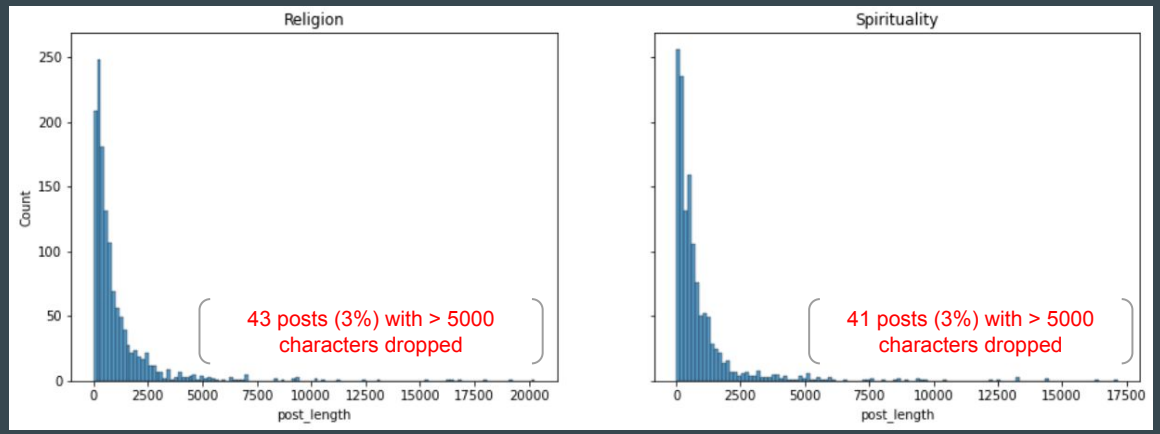
Measure of Success:

- The classification model would be assessed on its **test accuracy** and **specificity score**.

EDA

- For a balanced dataset, 1342* posts from each subreddit were included
- Postings dated from 2009 to 2020, with >85% from 2020, earlier posts were probably from top or hot posts
- For a more generalised model, postings with >5000 characters were dropped:
  - For religion, 43 posts were dropped
  - For spirituality, 41 posts were dropped

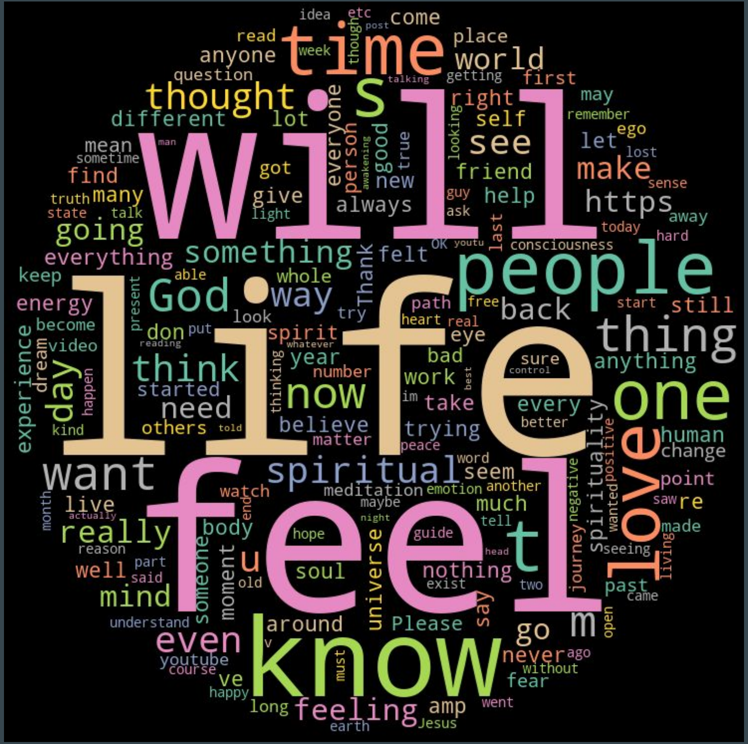
subreddit	created_yr	
spirituality	2020	1168
	2019	99
	2018	43
	2017	6
	2016	14
	2015	8
	2014	1
	2013	1
	2012	2
religion	2020	1163
	2019	92
	2018	15
	2017	5
	2016	10
	2015	9
	2014	7
	2013	12
	2012	12
	2011	7
	2010	6
	2009	4



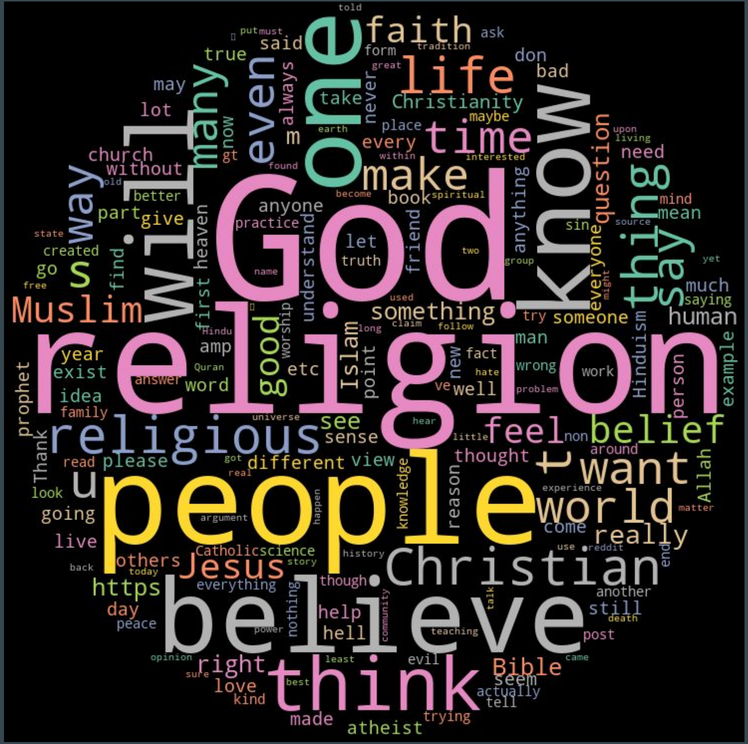
* Based on the subreddit with lesser posts, after removing duplicates and nan

Word Clouds

Spirituality



Religion



Pre-Processing

- Includes:
  - Using re.sub to remove all non-letters
  - Converting to lower-case
  - Removing ‘english’ stop words imported from NLTK.corpus, also removing words like ‘religion’ and ‘spirituality’ which is the subreddit itself
  - Using WordNetLemmatizer to simplify words to its base form

EDA with CountVectorizer

- After running CountVectorizer, a total of 12990 terms were created
- More than half of top-10 terms is common to both subreddits
  - like, life, know, people, one, god

Top-10 Terms for Spirituality

like	630
life	580
feel	510
know	477
people	466
time	440
one	416
thing	403
love	401
god	399

Top-10 Terms for Religion

god	1349
people	822
like	641
one	553
believe	536
would	531
know	487
christian	423
think	420
life	403

Logistic Regression

- Optimal model for LR found:
  - Using TfidfVectorizer
  - Accuracy score: 0.913
  - Specificity score: 0.872
  - Some overfitting

Pipeline

```
pipe = Pipeline([
    ('tvec',TfidfVectorizer()),
    ('lr',LogisticRegression(random_state=42))
])
```

Hyperparameters Tested

```
pipe_params = {
    'tvec__binary': [True,False],
    'tvec__max_features': [8000,10000,12000],
    'tvec__min_df': [2,3,4],
    'tvec__max_df': [0.8,0.85,0.9],
    'tvec__ngram_range': [(1,1),(1,2)]
}
```

Optimal Parameters

```
{'tvec__binary': False,
 'tvec__max_df': 0.8,
 'tvec__max_features': 12000,
 'tvec__min_df': 2,
 'tvec__ngram_range': (1, 2)}
```

Classification Metrics

	Model	Training Accuracy	Testing Accuracy	ROC AUC	Sensitivity (Recall)	Specificity	Precision	F1_score
0	Logistic Regression with CountVectorizer	0.998	0.888	0.955	0.921	0.857	0.860	0.890
1	Logistic Regression with TfidfVectorizer	0.958	0.913	0.964	0.957	0.872	0.877	0.915

Multinomial Naive-Bayes Classifier

- Optimal model for NB found:
  - Using TfidfVectorizer
  - Accuracy score: 0.904
  - Specificity score: 0.880
  - Some overfitting

Pipeline

```
pipe = Pipeline([
    ('tvec',TfidfVectorizer()),
    ('nb',MultinomialNB())
])
```

Hyperparameters Tested

```
pipe_params = {
    'tvec__binary': [True,False],
    'tvec__max_features': [8000,10000,12000],
    'tvec__min_df': [2,3,4],
    'tvec__max_df': [0.8,0.85,0.9],
    'tvec__ngram_range': [(1,1),(1,2)],
    'nb__alpha': [0.8,0.9,1.0]
}
```

Optimal Parameters

```
{'nb__alpha': 0.8,
 'tvec__binary': False,
 'tvec__max_df': 0.8,
 'tvec__max_features': 9000,
 'tvec__min_df': 2,
 'tvec__ngram_range': (1, 2)}
```

Classification Metrics

	Model	Training Accuracy	Testing Accuracy	ROC AUC	Sensitivity (Recall)	Specificity	Precision	F1_score
3	Naive Bayes with TfldfVectorizer	0.955	0.904	0.964	0.929	0.880	0.881	0.904
2	Naive Bayes with CountVectorizer	0.944	0.896	0.949	0.917	0.876	0.876	0.896

RandomForest Classifier

- Optimal model for RFC found:
  - Using CountVectorizer
  - Accuracy score: 0.917
  - Specificity score: 0.887
  - More overfitting

Pipeline

```
pipe = Pipeline([
    ('cvec', CountVectorizer()),
    ('rfc', RandomForestClassifier(random_state=42))
])
```

Hyperparameters Tested

```
pipe_params = {
    'cvec__binary': [True,False],
    'cvec__max_features': [8000,10000,12000],
    'cvec__ngram_range': [(1,1),(1,2)],
    'rfc__n_estimators': [360,370,380]
}
```

Optimal Parameters

```
{'cvec__binary': False,
 'cvec__max_features': 10000,
 'cvec__ngram_range': (1, 2),
 'rfc__n_estimators': 360}
```

Classification Metrics

Model	Training Accuracy	Testing Accuracy	ROC AUC	Sensitivity (Recall)	Specificity	Precision	F1_score
RandomForest with CountVectorizer	1.0	0.917	0.969	0.949	0.887	0.889	0.918
RandomForest with TfidfVectorizer	1.0	0.912	0.968	0.945	0.880	0.882	0.913

Production Model

- Multinomial NB was chosen as the final production model as it shows the least amount of overfitting while maintaining a high accuracy and specificity score.

Classification Metrics

	Model	Training Accuracy	Testing Accuracy	ROC AUC	Sensitivity (Recall)	Specificity	Precision	F1_score
0	Logistic Regression with CountVectorizer	0.998	0.888	0.955	0.921	0.857	0.860	0.890
1	Logistic Regression with TfidfVectorizer	0.958	0.913	0.964	0.957	0.872	0.877	0.915
2	Naive Bayes with CountVectorizer	0.944	0.896	0.949	0.917	0.876	0.876	0.896
3	Naive Bayes with TfIdfVectorizer	0.955	0.904	0.964	0.929	0.880	0.881	0.904
4	RandomForest with CountVectorizer	1.000	0.917	0.969	0.949	0.887	0.889	0.918
5	RandomForest with TfidfVectorizer	1.000	0.912	0.968	0.945	0.880	0.882	0.913

Top Important Features

- Top-10 important features associated with the positive class of Spirituality

Top-10 NB feature importance

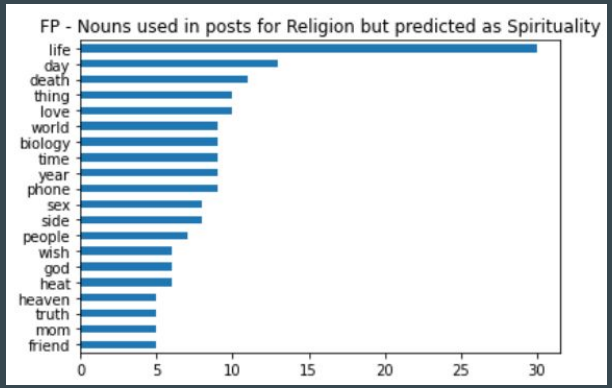
	Importance
life	0.001895
feel	0.001889
like	0.001829
love	0.001704
know	0.001594
time	0.001551
people	0.001484
thing	0.001380
want	0.001313
one	0.001303

False Positives

- 32 postings were misclassified as false positives
- These 6 terms from these postings were causing the misclassification

LR feature odds

soul	5.207924
day	4.644754
life	3.176231
time	2.780972
thing	1.194352
year	1.107591



Test Model with New Data

- 50 new posts each from both subreddits were downloaded and tested with the final model
  - Accuracy score: 0.917, higher than using test data
  - Specificity: 0.971, highest achieved so far

Model	Training Accuracy	Testing Accuracy	ROC AUC	Sensitivity (Recall)	Specificity	Precision	F1_score
Naive-Bayes with New posts	NaN	0.917	0.955	0.88	0.971	0.978	0.926

Conclusion

- The model is able to distinguish content of Spirituality and Religion quite well, with an ROC AUC score of 0.955 on test data.
- On new data, the model also performed well with a high accuracy score of 0.917 and a high specificity of 0.971.

Recommendation

- Deploy the model to start identify negative class postings quickly and remove them from the forum.
- For positive class posts, if they contain these 6 keywords: soul, day, life, time, thing and year, look at them separately before allowing the post as these are likely candidates for false positives.

The End