WEB APIs & CLASSIFICATION

...

PROJECT 3

PROBLEM STATEMENT

To **classify** Reddit posts from r/nosleep and r/thetruthishere using Natural Language Processing **(NLP)** and **Classification** modelling. The model evaluated with **Accuracy** scores is then expected to predict to which subreddit a given post belongs to.

The model should then help Reddit data science team to advise their advertisers on spending forecast for targeted marketing campaigns of products and services like online gaming targeting members of r/nosleep, calming supplements & therapy/ counselling services for members of r/thetruthishere depending on the predicted subreddit.

BACKGROUND

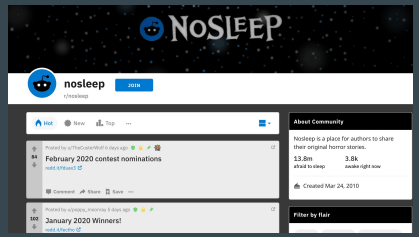
The objective is to find out the attributes of each group and their online behaviour.

With the information gathered, business can;

- Decide how much advertising dollars to spend and on which group to obtain better ROI
- Types of products and services to be advertised
- Frequency of advertisement
- Type of advertisement eg, static, pop-up, etc
- KPI expected from Reddit

DATA COLLECTION

HTML Data
(Reddit's API)



URL:

"https://www.reddit.com/r/nosleep.json"

** 25 posts per request*

Collect Data
(Python)

Python's Requests Library

```
headers = {'User-agent': 'Pony Inc 1.0'}
response = requests.get(url, headers = headers)
Add timer: time.sleep()
```

Save Data
(JSON)


```
response.json()
```
JavaScript Object Notation

Parse Data
(Python)

subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	link_flair_richtext	subreddit_nam
Thetruthisshere	I am sharing this whenever I think would make...	t2_4qq9tag	False	NaN	0	False	Demonic sounds	[]	r/Th
Thetruthisshere	Does anyone else feel like their is something ...	t2_4w0h8res	False	NaN	0	False	Anyone else?	[]	r/Th
Thetruthisshere	Hi there! I just recently got this subreddit c...	t2_r0h2z1z	False	NaN	0	False	Demonic type of screams and other strange nois...	[]	r/Th

DATA CLEANING AND EDA

DATA CLEANING:

	subreddit	selftext	title
0	nosleep	A half-dozen police cars crowded the gravel dr...	There's Been a String of Suicides in my Town. ...
1	nosleep	[Part V](https://www.reddit.com/r/nosleep/comm...	My wife and I bought a ranch in the mountains ...
2	nosleep	People called her Brittle.\n\nAlexa was differ...	My little sister had glass bones

drop_duplicates(subset = 'title')

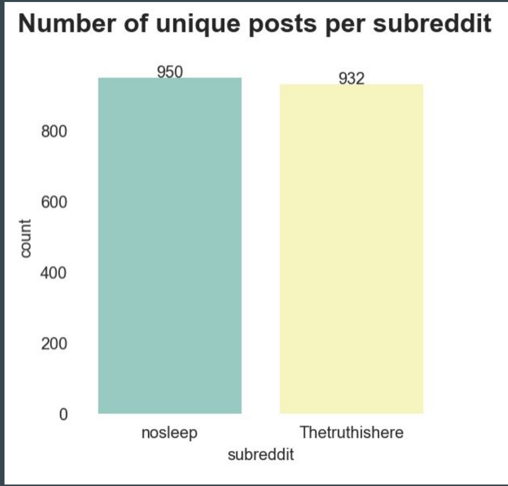
	subreddit	selftext	title
30	Thetruthishere	NaN	Have you ever met a reptilian shapeshifter?
35	Thetruthishere	NaN	Entity on babycam 12/02/20
38	Thetruthishere	NaN	Anybody any info on entities that feed usually...
43	Thetruthishere	NaN	Tell me your story
46	Thetruthishere	NaN	Hearing voices,simulation,demonic,paranormal,g...

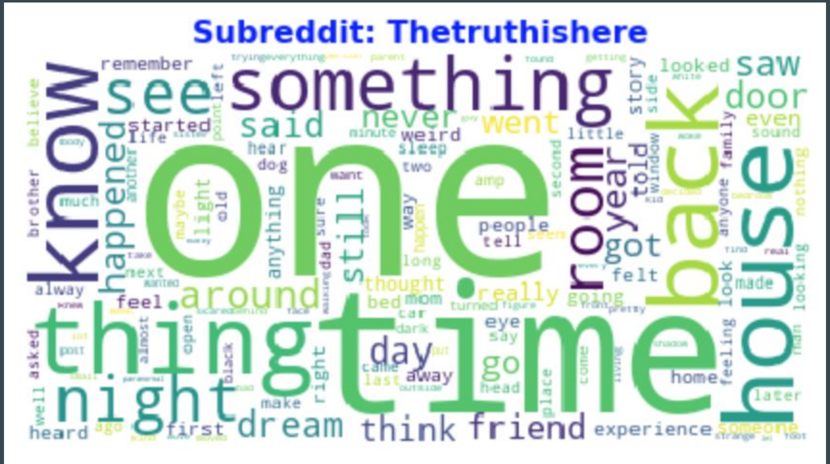
```
on my elbows, knees, and forehead, but I didn't remove my hands from my eyes.\n\nI lay in the grass, the warm sun on
my face, and I wished for nothing more than the will to claw my own eyes out. That thing was following me, I had no d
oubt. I had no doubt it was getting closer to me.\n\nThe thought of seeing it again forced me to my side where I vomi
ted again.\n\n"Jane," Olivia said, scrambling to my side. "Jane, what happened?"\n\n"You're not next, Olivia," I sai
d, my voice little more than a whisper. "I'm next.'\n\n\\--------\n\n[YT](https://www.youtube.com/watch?v=ps0IBOCut5
o)\n\n[More](https://jamesworchester.com/)'
```

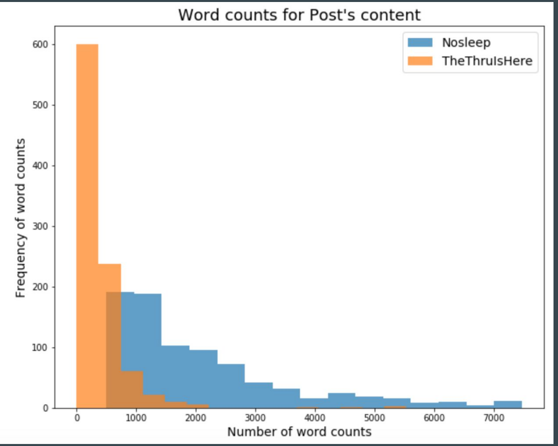
BeautifulSoup(raw_post).get_text()
re.sub()
.lower().split()

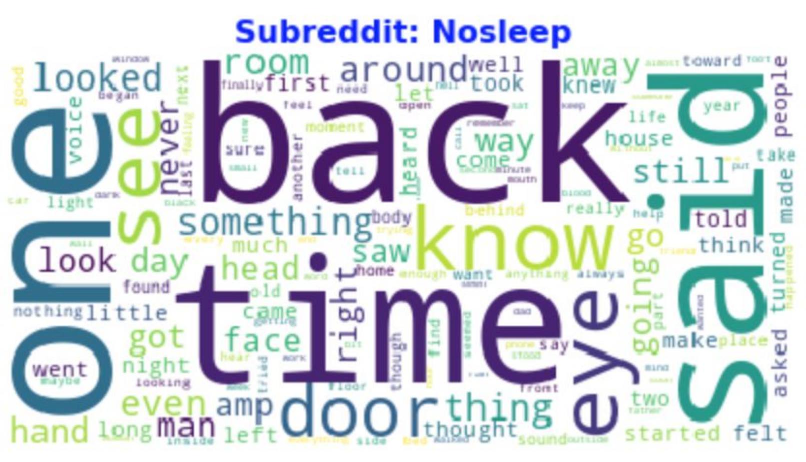
.lemmatize()
stopwords.word('english')

EDA

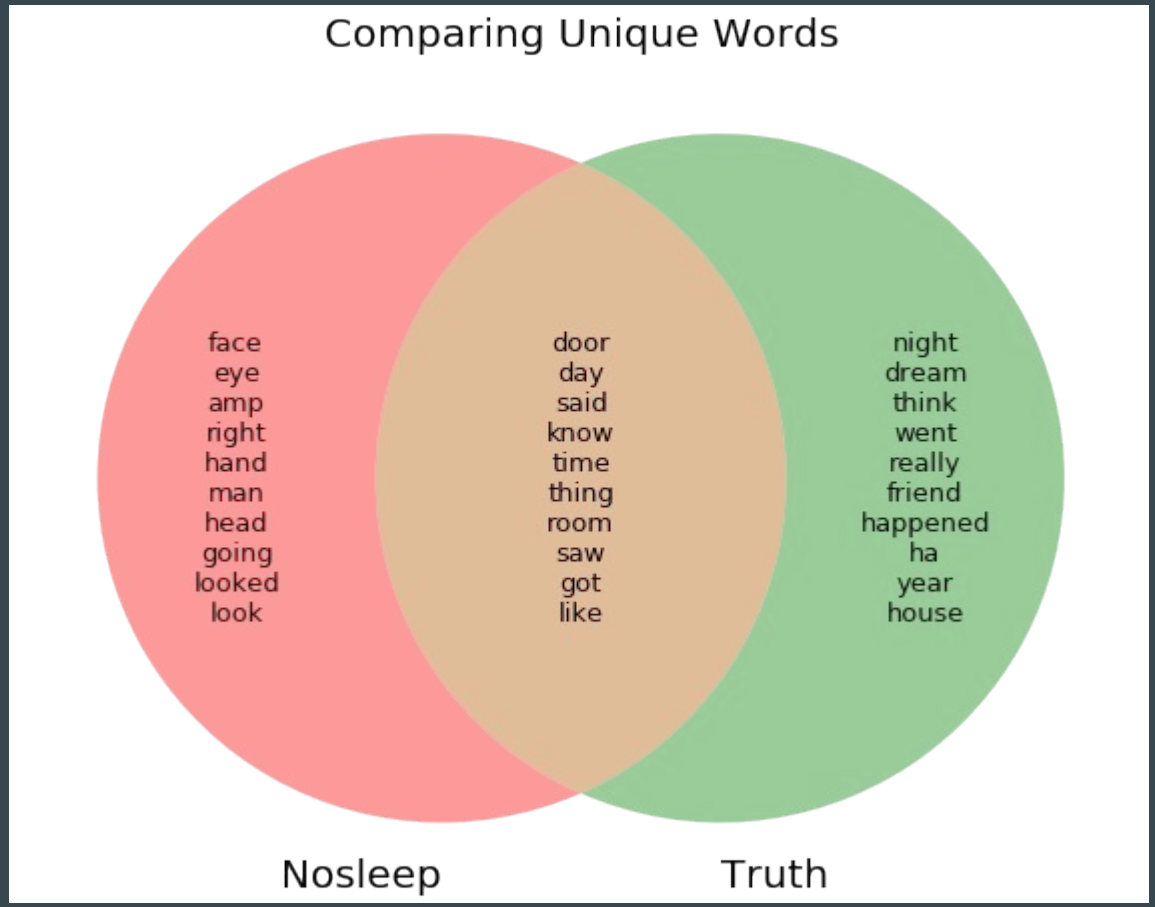

Number of unique posts per subreddit
950
932
800
600
400
200
0
count
nosleep
Thetruthishere
subreddit


Subreddit: Thetruthishere
one
time
thing
know
something
back
house
see
night
room
dream
think
friend
around
said
never
went
door
saw
happened
still
got
day
experience


Word counts for Post's content
Nosleep
TheThruIsHere
Frequency of word counts
Number of word counts
0
100
200
300
400
500
600
1000
2000
3000
4000
5000
6000
7000


Subreddit: Nosleep
back
time
one
said
know
see
eye
door
something
looked
room
around
thing
look
day
head
face
still
saw
went
hand
man

PREPROCESSING AND MODELLING



PREPROCESSING AND MODELLING

- Split our data into X and y

```
X = df_combined_clean[['combined']]
y = df_combined_clean['subreddit']
```

- Split our data into training and testing sets
- Turn our text into features
  - CountVectorizer
  - TF-IDFVectorizer

PREPROCESSING AND MODELLING

CountVectorizer

	able	actually	ago	alys	amp	arm	asked	ay	bad	bed	...	week	weird	went	white	window	woman	word	work	world	year
0	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
2	0	0	0	2	0	1	2	4	0	0	...	0	0	1	1	0	5	0	2	0	1
3	0	1	4	0	53	7	6	12	4	13	...	1	0	7	2	2	8	2	1	0	15
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

TF-IDFVectorizer

	able	actually	ago	alys	amp	arm	asked	ay	bad	bed	...	week	weird	went	white	window	womar
0	0.0	0.349599	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.00000	0.000000
1	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.00000	0.000000
2	0.0	0.000000	0.000000	0.080697	0.000000	0.049083	0.082776	0.141408	0.000000	0.000000	...	0.000000	0.0	0.033910	0.052101	0.00000	0.275428
3	0.0	0.007470	0.026131	0.000000	0.471109	0.051864	0.037485	0.064037	0.030006	0.084334	...	0.007056	0.0	0.035831	0.015729	0.01404	0.066522
4	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.00000	0.000000

PREPROCESSING AND MODELLING

Classifier Models	
Naive Bayes Model (MultinomialNB)	**Logistic Regression**
Columns are all integer counts	Response default falls into one of two categories.

EVALUATION

BASELINE SCORE

	0	1
target	0.504782	0.495218

ACCURACY

Scores	CVEC + LR	TDIF + LR
Training score	0.999	0.998
Testing score	0.953	0.958
Variance	-0.046	-0.040

CONFUSION MATRIX

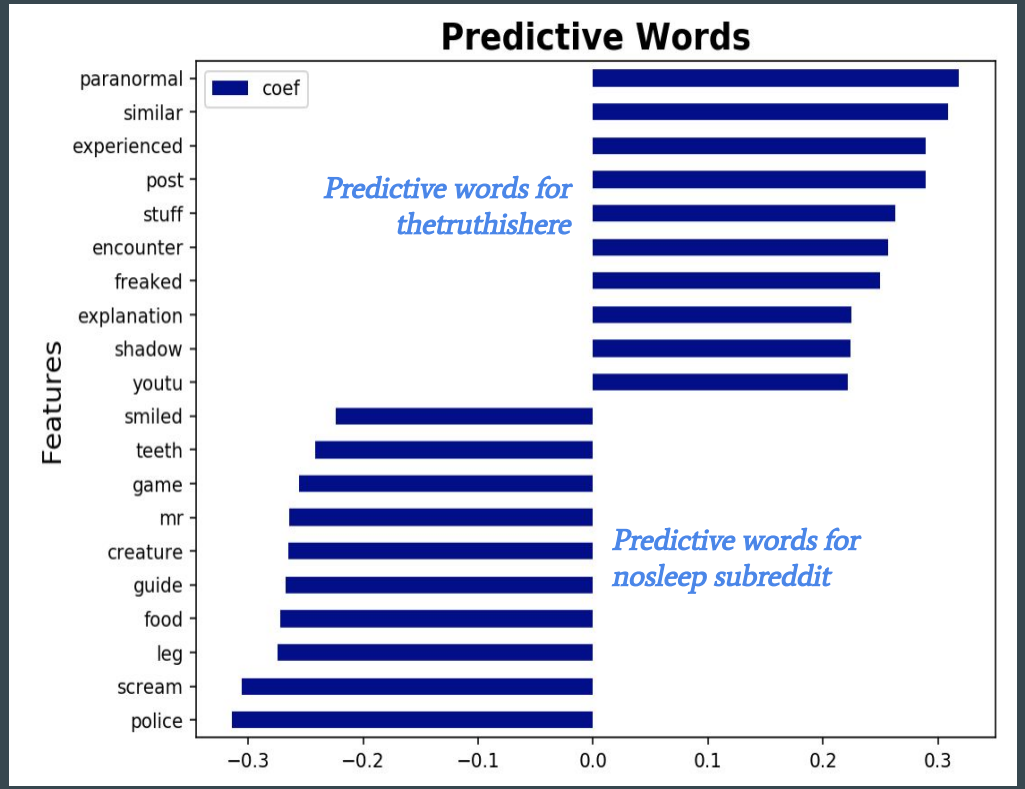
LR + CVEC	Predicted Nosleep	Predicted Thetruthishere
Actual Nosleep	228	10
Actual Thetruthishere	12	221

LR + TDIF	Predicted Nosleep	Predicted Thetruthishere
Actual Nosleep	234	4
Actual Thetruthishere	16	217

EVALUATION

1) With logistic regression, we can identify predictive words that belong to each subreddit.

2) This will help in forecasting spending for the different products/ services that are targeting members of each subreddit.

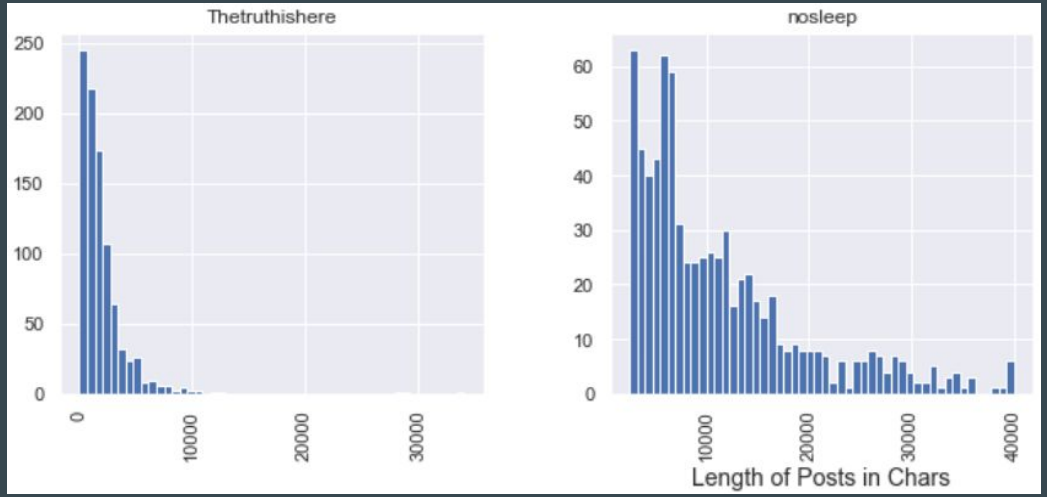


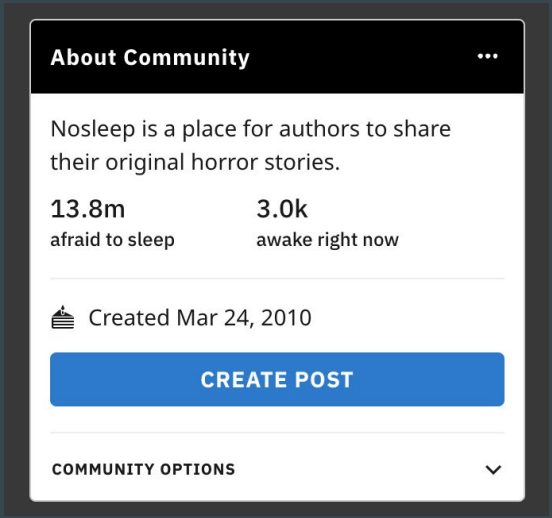
LIMITATIONS OF THE MODEL

1) The hyper-parameters of the model has been tuned to include features/ words as many as 1500 to reduce the variance between testing and training scores.
2) But in reality, it may make sense to reduce the number of words.
3) The model can be further enhanced :
    a) **To eliminate words of similar meanings**
    b) **To include phrases instead of just words for prediction.**

CONCLUSION

- Business should get data directly from Reddit using API or RESTful API which is significantly faster than scraping (limit on posts).
- Business may want to spend more advertising dollars on the Nosleep group as higher chance to convert users to customers; more bang for the buck





RECOMMENDATIONS

- Offer products and services such as games, online gaming, movie subscriptions, and food delivery services
- Advertise counselling services, calming supplements to promote good night sleep to the Truth group
- Business can consider using our Topic modeling service to segment users into groups by say, sci-fi, horror, fantasy, etc for;
    - pop-up advertisement according to genre, and
    - Implement adaptive or real-time advertisements at Reddit
- Business can work with Reddit on advertisement placements, get Reddit to provide analytics such as ‘click-rate’

5.0k

Posted by u/RichardSaxon 16 hours ago

42 years ago we sent Voyager 1 into space to look for extraterrestrial life, today we found it.

42 years, 6 months and 3 days ago, on the 5th of September 1977, a space probe was launched from Earth, and sent on an endless journey through space. The probe, which was affectionately named Voyager 1, contains a multitude of information regarding humanity, including our language, our art, and in a more metaphorical way; Our souls.

Today, the little machine is 22 billion kilometers away from Earth, the furthest reach of our species, though not a manned vessel, it's still a part of us. Despite its distance, we're still keeping contact with it, and during it's journey spanning almost half a century, it has given us an insight into the mysteries of the universe we thought we could only dream about.

My own father spent most of his life on the project, and I have since followed in his footsteps. I dreamed about taking over his work even as a twelve year old kid, and

203 Comments Share Save ...