WEB APIs & CLASSIFICATION

•••

PROJECT 3 | GROUP 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)
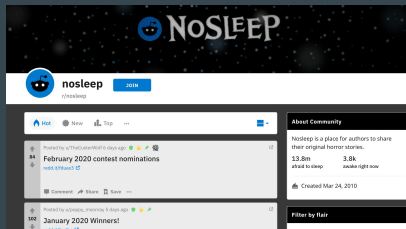
Collect Data
(Python)

Python's Requests
Library

Save Data
(JSON)

Parse Data
(Python)

URL:

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

* 25 posts per request

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

response.json()
JavaScript Object Notation

{"kind": "Listing", "data": {"modhash": "", "dist": 27, "children": [{"kind": "t3", "data": {"approved_at_utc": null, "subreddit": "nosleep", "selftext": "", "author_fullname": "t2_c446v4f", "saved": false, "mod_reason_title": null, "gilded": 0, "clicked": false, "title": "February 2020 contest nominations", "link_flair_richtext": [], "subreddit_name_prefixed": "r /nosleep", "hidden": false, "pwls": 6, "link_flair_css_class": null, "downs": 0, "hide_score": false, "name": "t3_fduh8s", "quarantine": false, "link_flair_text_color": "dark", "author_flair_background_id": null, "subreddit_type": "public", "ups": 83, "total_awards_received": 1, "media_embed": {}, "author_flair_template_id": null, "is_original_content": false, "user_reports": [], "secure_media": null, "is_reddit_media_domain": false, "is_meta": false, "category": null, "secure_media_embed": {}, "link_flair_text": null, "can_mod_post": false, "score": 83, "approved_by": null, "author_premium": false, "thumbnail": "", "edited": false, "author_flair_css_class": null, "author_flair_richtext": [], "gildings": {}, "content_categories": ["writing"], "is_self": false, "mod_note": null, "created": 1583458912.0, "link_flair_type": "text", "wls": 6, "removed_by_category": null, "banned_by": null, "author_flair_type": "text", "domain": "redd.it", "allow_live_comments": false, "selftext_html": null, "likes": null, "suggested_sort": null, "banned_at_utc": null, "view_count": null, "archived": false, "no_follow": false, "is_crosspostable": false, "pinned": false, "over_18": false, "all_awardings": [{"giver_coin_reward": null, "subreddit_id": null, "is_new": false, "days_of_drip_extension": 0, "coin_price": 200, "id": "award_f7562045-905d-413e-9ed2-0a16d40bfe349", "penny_donate": null, "coin_reward": 0, "icon_url": "https://i.redd.it/award_images/t5_22cerq/6vgr8y2li9741_PlusOne.png", "days_of_premium": 0, "icon_height": 2048, "resized_icons": [{"url": "https://preview.redd.it/award_images/t5_22cerq/6vgr8y2li9741_PlusOne.png?width=16&height =16&auto=webp&s=4c8f077e9e98a517968&cfd1e93d999c1f19e51e", "width": 16, "height": 16}, {"url": "https://preview.redd .it/award_images/t5_22cerq/6vgr8y2li9741_PlusOne.png?width=32&height=32&auto=webp&s =4d42bed81cdf5d406e3795ac1575819a2333336a", "width": 32, "height": 32}, {"url": "https://preview.redd.it/award_images

subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	link_flair_richtext	subreddit_nam
Thetruthishere	I am sharing this whenever I think would make...	t2_4qqdrtqg	False		NaN	0	False	Demonic sounds		r/Th
Thetruthishere	Does anyone else feel like their is something	t2_4w0h8res	False		NaN	0	False	Anyone else?		r/Th
Thetruthishere	Hi there! I just recently got this subreddit r...	t2_z0h2z1z	False		NaN	0	False	Demonic type of screams and other strange noie...		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

	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...

drop_duplicates(subset = 'title')

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 doubt. I had no doubt it was getting closer to me.\n\nThe thought of seeing it again forced me to my side where I vomited again.\n\n"Jane," Olivia said, scrambling to my side. "Jane, what happened?"\n\n"You're not next, Olivia," I said, my voice little more than a whisper. "I'm next. \n\n\\--------\n\n[YT](https://www.youtube.com/watch?v=ps0IBOCut5o)\n\n[More](https://jamesworchester.com/)'

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

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

EDA

PREPROCESSING AND MODELLING

Comparing Unique Words

Nosleep: face, eye, amp, right, hand, man, head, going, looked, look

Intersection: door, day, said, know, time, thing, room, saw, got, like

Truth: night, dream, think, went, really, friend, happened, ha, year, house
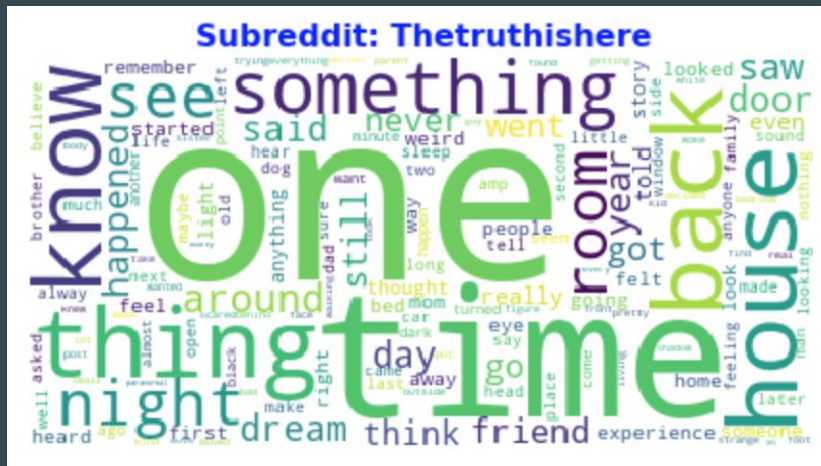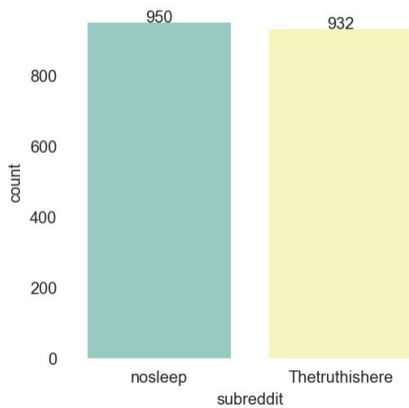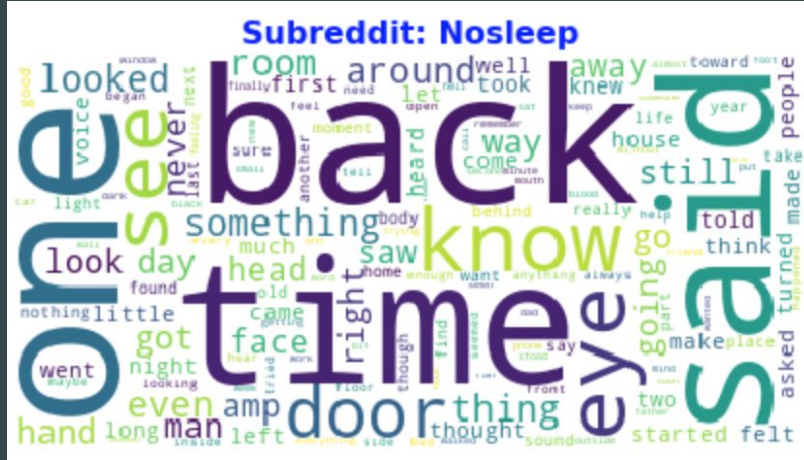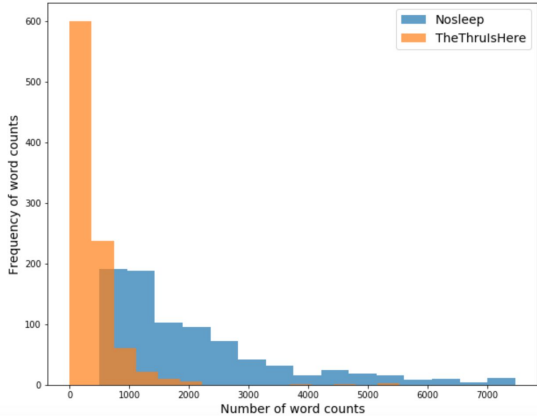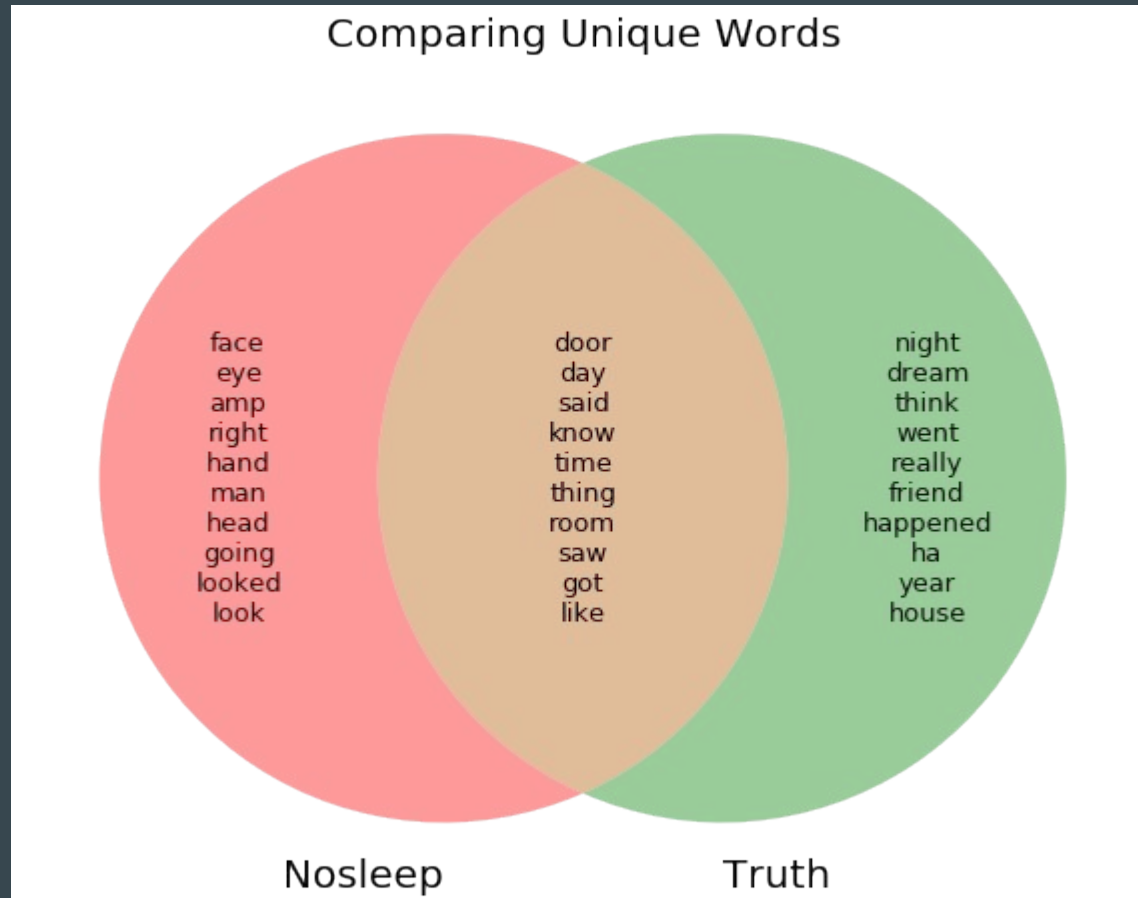
PREPROCESSING AND MODELLING

- Split our data into X and y

```python
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	woman
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

	Nosleep	Thetruthishere
target	0.504782	0.495218

ACCURACY

Models	Train Score	Test Score	Variance
CVEC + LR	0.999	0.953	-0.046
TDIF + LR	0.998	0.958	-0.040
CVEC + MultiNB	0.928	0.902	-0.026
TDIF + MultiNB	0.950	0.913	-0.037

CONFUSION MATRIX

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

LR + TD IDF	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) These predictive words can be potentially employed in targeted marketing campaigns to increase the "click- rate" of a particular ad to the target subreddit member.
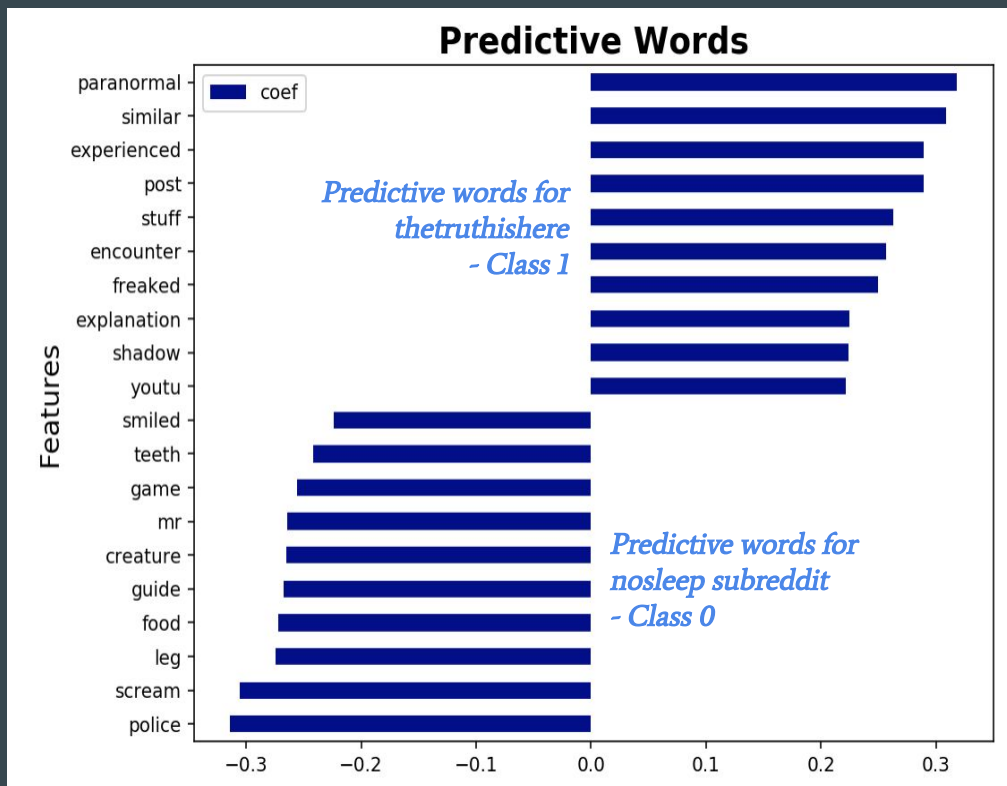
LIMITATIONS OF THE MODEL

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 synonyms, post agnostic words, common words like paranormal**
   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 13.8m users to customers; more bang for the buck
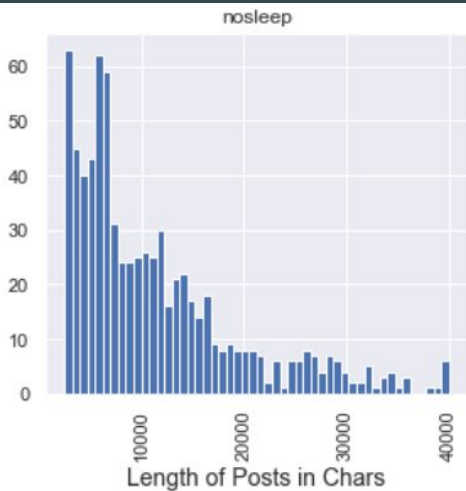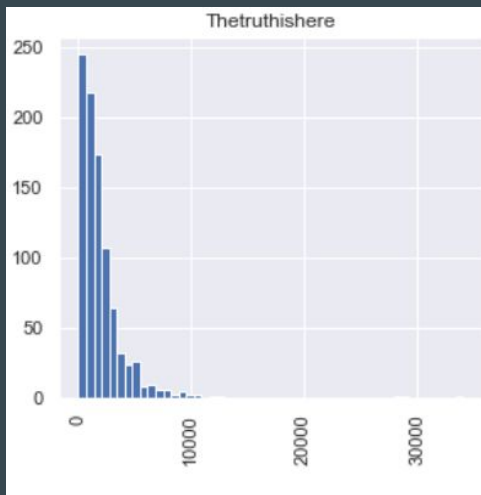
RECOMMENDATIONS

- 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'

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   •••