Reddit Classification -

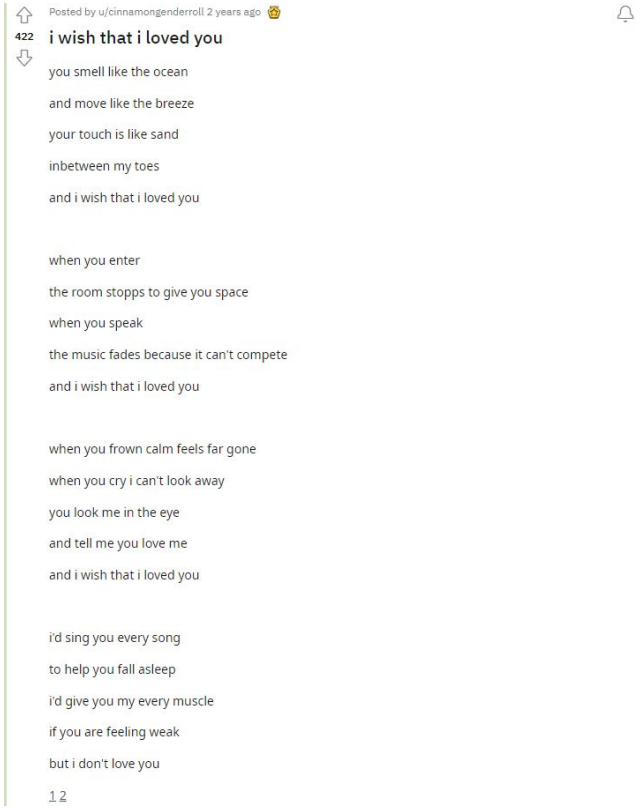
Scary Stories vs Poetry

Project overview

How well can an NLP model distinguish poetry from scary stories?

1. Background
2. Data Overview
3. Models
4. Conclusion

Sample Posts



Posted by u/cinnamongenderroll 2 years ago

422

i wish that i loved you

you smell like the ocean
and move like the breeze
your touch is like sand
inbetween my toes
and i wish that i loved you

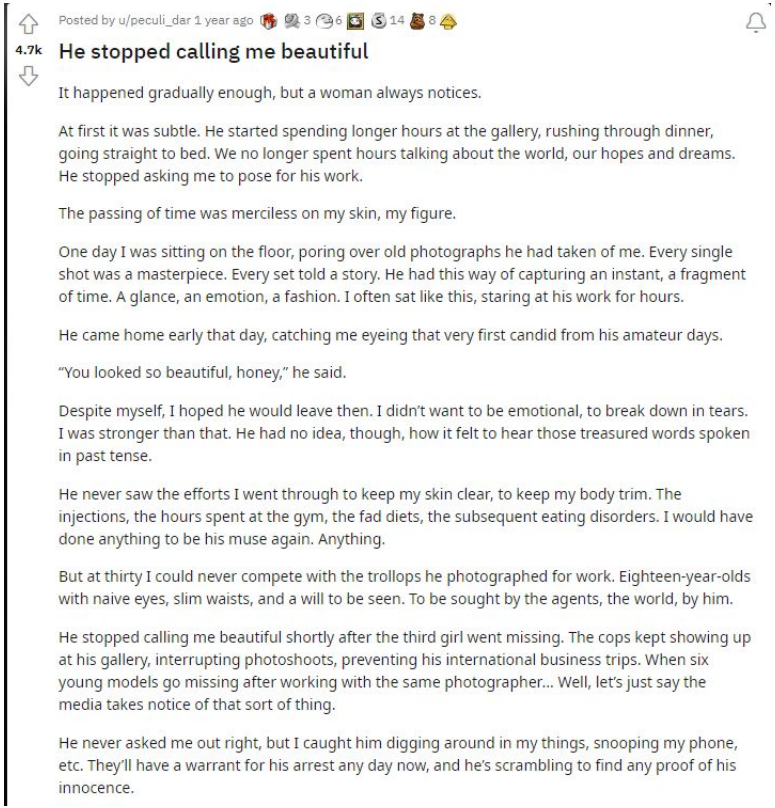
when you enter
the room stopps to give you space
when you speak
the music fades because it can't compete
and i wish that i loved you

when you frown calm feels far gone
when you cry i can't look away
you look me in the eye
and tell me you love me
and i wish that i loved you

i'd sing you every song
to help you fall asleep
i'd give you my every muscle
if you are feeling weak
but i don't love you

1 2

r/OCPoetry



Posted by u/peculi_dar 1 year ago 3 6 14 8

4.7k

He stopped calling me beautiful

It happened gradually enough, but a woman always notices.

At first it was subtle. He started spending longer hours at the gallery, rushing through dinner, going straight to bed. We no longer spent hours talking about the world, our hopes and dreams. He stopped asking me to pose for his work.

The passing of time was merciless on my skin, my figure.

One day I was sitting on the floor, poring over old photographs he had taken of me. Every single shot was a masterpiece. Every set told a story. He had this way of capturing an instant, a fragment of time. A glance, an emotion, a fashion. I often sat like this, staring at his work for hours.

He came home early that day, catching me eyeing that very first candid from his amateur days.

"You looked so beautiful, honey," he said.

Despite myself, I hoped he would leave then. I didn't want to be emotional, to break down in tears. I was stronger than that. He had no idea, though, how it felt to hear those treasured words spoken in past tense.

He never saw the efforts I went through to keep my skin clear, to keep my body trim. The injections, the hours spent at the gym, the fad diets, the subsequent eating disorders. I would have done anything to be his muse again. Anything.

But at thirty I could never compete with the trollops he photographed for work. Eighteen-year-olds with naive eyes, slim waists, and a will to be seen. To be sought by the agents, the world, by him.

He stopped calling me beautiful shortly after the third girl went missing. The cops kept showing up at his gallery, interrupting photoshoots, preventing his international business trips. When six young models go missing after working with the same photographer... Well, let's just say the media takes notice of that sort of thing.

He never asked me out right, but I caught him digging around in my things, snooping my phone, etc. They'll have a warrant for his arrest any day now, and he's scrambling to find any proof of his innocence.

r/shortscarystories

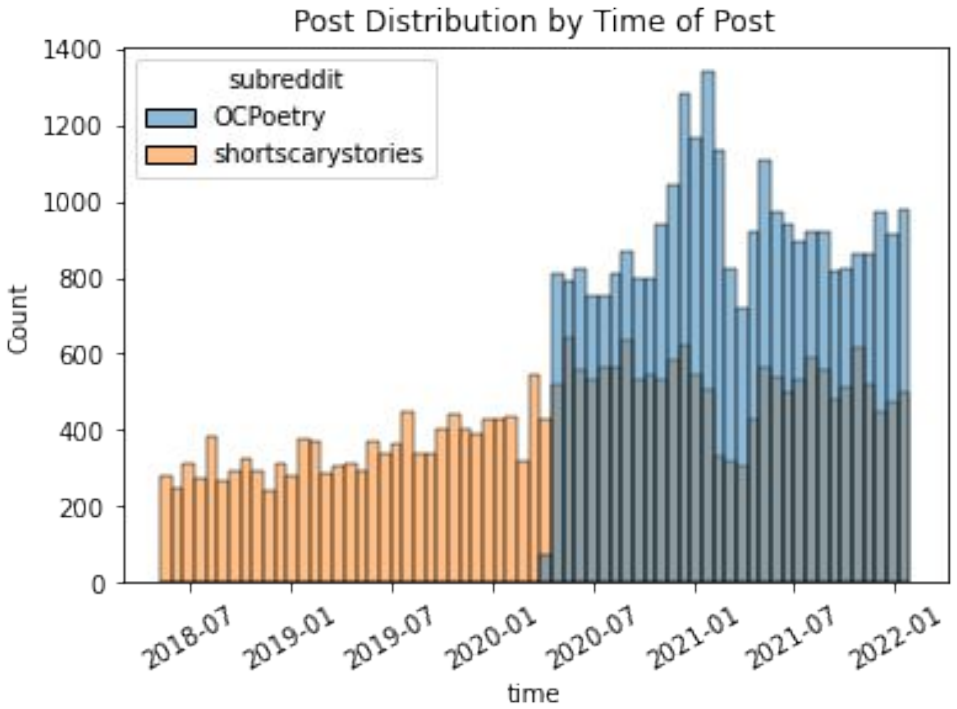
Data overview

- 56550 total posts from r/OCPoetry and r/shortscarystories
- Scraped using the push shift API
- Filtered for length, deleted posts, and duplicates

EDA Steps

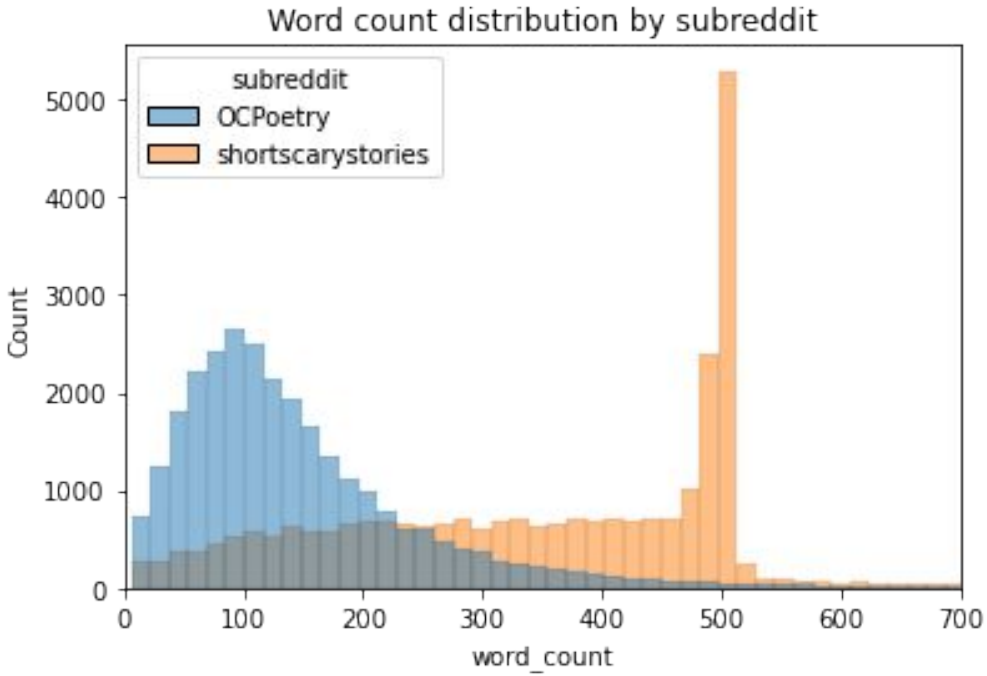
Date of Posts

- Data set is balanced but r/OCPoetry is more popular.
- r/shortscarystories have greater data range.



Word Count

- r/shortscarystory has rule requiring posts to be <= 500 words.
- r/OCPoetry posts are shorter, more normally distributed



Modelling Phase

Text Features only

- Variety of models trained on unigram features.
- Stop_words:
  - `['poem',`
  - `'poems',`
  - `'ocpoetry',`
  - `'poet',`
  - `'poets',`
  - `'poetry',`
  - `'link',`
  - `'links',`
  - `'feedback',`
  - `'story',`
  - `'stories',`
  - `'amp']`

Feature set

Model

	Tfidf 1k	Tfidf 5k	Count 1k
Logistic Regression	0.908	0.923	0.901
Multinomial Naive-Bayes	0.856	0.891	0.857
Random Forest	0.871		
Extremely Randomized Trees	0.857		

Feature Engineering

What is the one feature to rule them all?

Can you guess the single feature model with 90%+ accuracy!

Other Features

- For feature alone - unpenalized logistic regression.
- With text vectorization, used penalized logistic regression
- Text Tfidf logistic regression baseline: 0.908

Feature 2

Feature 1	None	Tfidf 1k
White Space	0.910	0.942
New lines	0.902	–
Parts of speech	0.776	0.913
Parts of speech + punctuation	0.830	0.919
Sentiment Scores	0.643	0.910
Word Length	0.678	0.909

All Feature Models

Logistic Regression, 1k unigrams	0.948
Logistic Regression, 5k unigrams	0.954
Logistic Regression, 1k unigrams, bigrams	0.945
Logistic Regression CV, 5k unigrams, bigrams	0.953
Multinomial NB, 5k unigrams, bigrams*	0.870
Gradient Boosting Classifier, 5k unigrams, bigrams	0.948
LightGBM Classifier, 5k unigrams, bigrams	0.960

Top Features

=== r/OCPoetry ===	
new_line_chars	-39.192570
wrote	-3.825403
heart	-3.410960
space_chars	-2.868872
too	-2.695912
yet	-2.691979
tears	-2.551375
in	-2.528592
lost	-2.366558
write	-2.311113

=== r/shortscarystories ===	
.	11.404029
horror	5.604725
word_count	3.839304
text_words	3.830713
immediately	3.354428
woods	2.887923
people	2.874973
police	2.837151
short	2.816935
human	2.679333

Conclusion

- Best classifiers achieved high accuracy distinguishing poetry from scary stories
- White space, word length, and part of speech tagging gave extra predictive power
- LightGBM was best model, followed by Logistic Regression, naive-Bayes.

Questions?