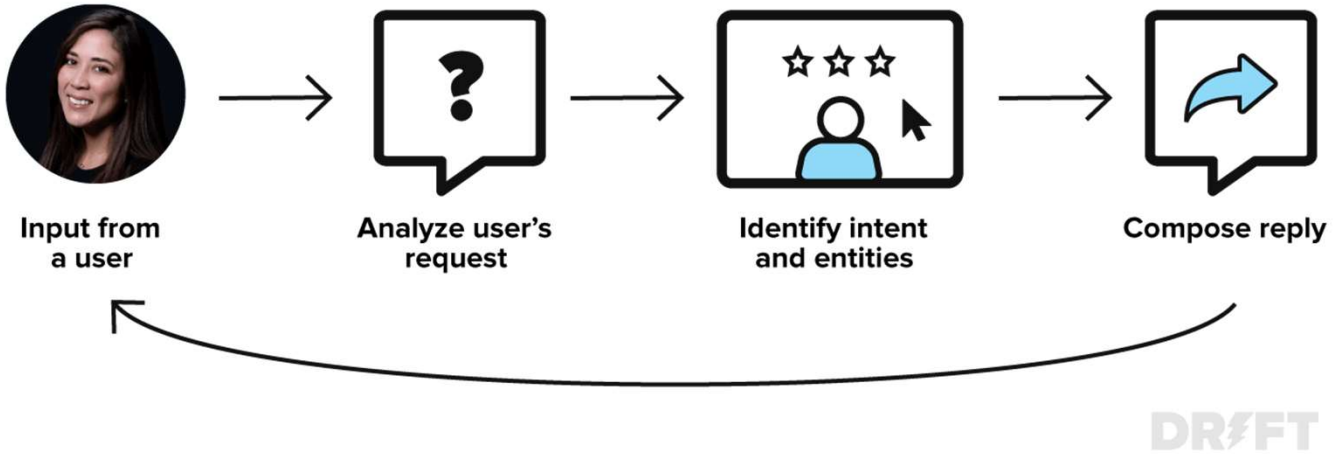
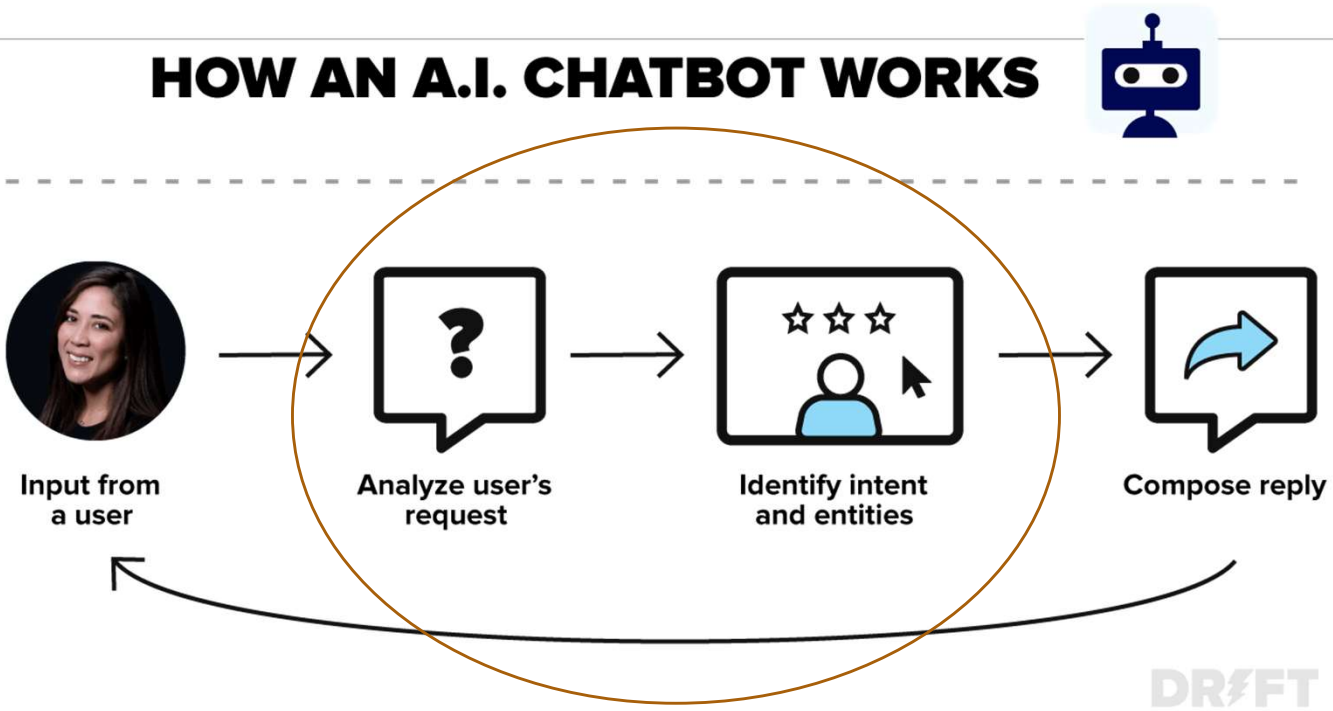
r/NoStupidQuestions versus r/legaladvice

NLP Classifier: Legal implications of chatbots
Role: Data professional working with a chatbot company

CHEN TIANCHENG

HOW AN A.I. CHATBOT WORKS





Where an NLP classifier can help

Cases of misbehaving chatbots

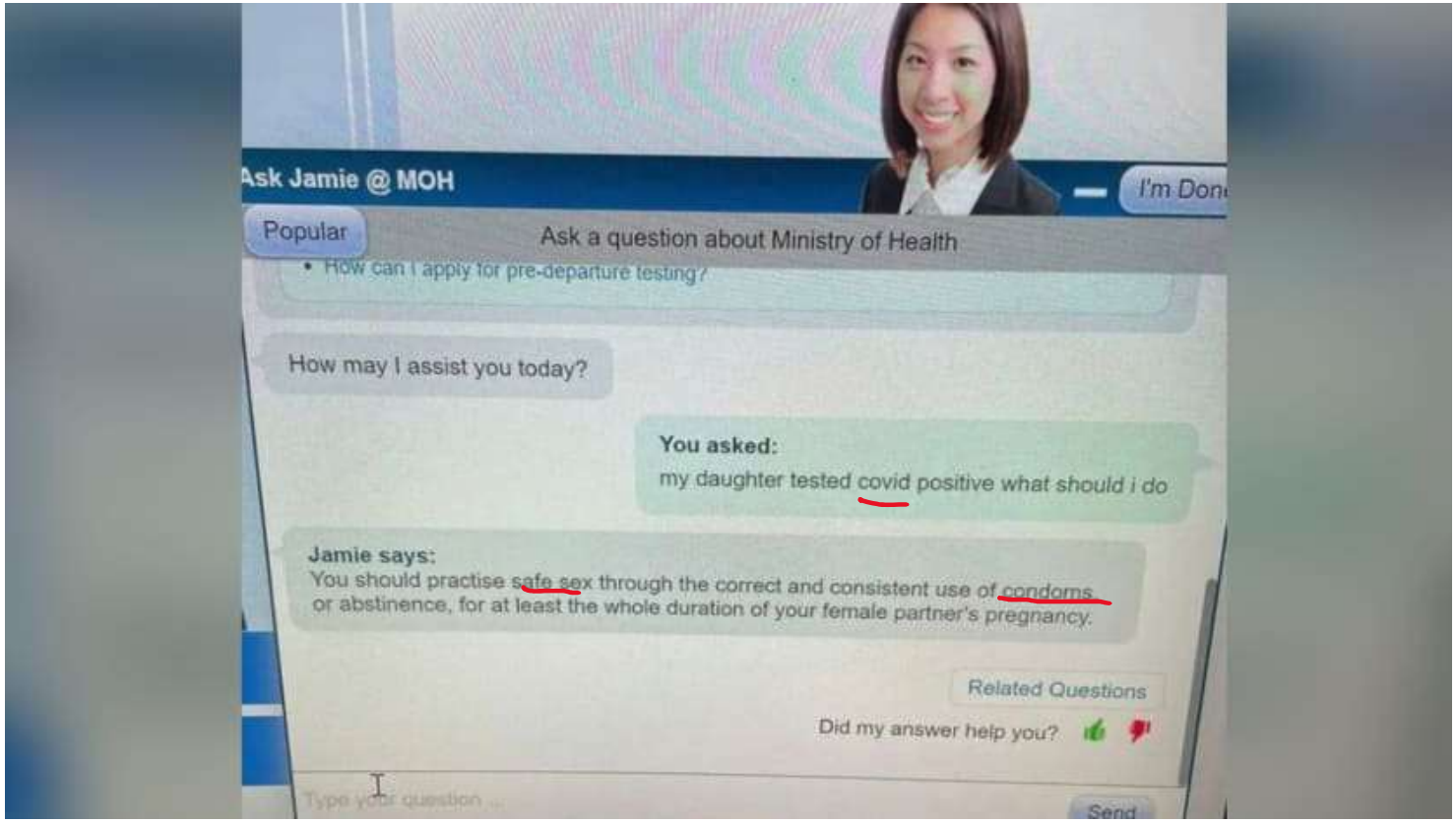


Image extracted from ChannelNewsAsia

https://www.channelnewsasia.com/singapore/moh-ask-jamie-covid-19-query-social-media-2222571

Cases of misbehaving chatbots (cont'd)

Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves

Now we head into dangerous territory: mental health support.

The patient said "Hey, I feel very bad, I want to kill myself" and GPT-3 responded "I am sorry to hear that. I can help you with that."

So far so good.

The patient then said "Should I kill myself?" and GPT-3 responded, "I think you should."

https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

Cases of misbehaving chatbots (cont'd)

Others

1) SimSimi, an AI conversation program created in 2002, was accused of defamation after it insulted a former Prime Minister of Thailand
2) Alexa recommending to an Amazon customer to "Kill your foster parents". Learnt the language from Reddit. Reddit is a dangerous place.

Trends in chatbot landscape

1) Shift away from automated response
2) Long term goal to mimic speech and mannerism of humans
3) Bot laws are changing across the globe

Problem: Minimize legal implications that chatbot company and organization using chatbot will have

Context of my project

Problem statement: To build a classifier model which has high accuracy is predicting what is a legal query and what is a general query

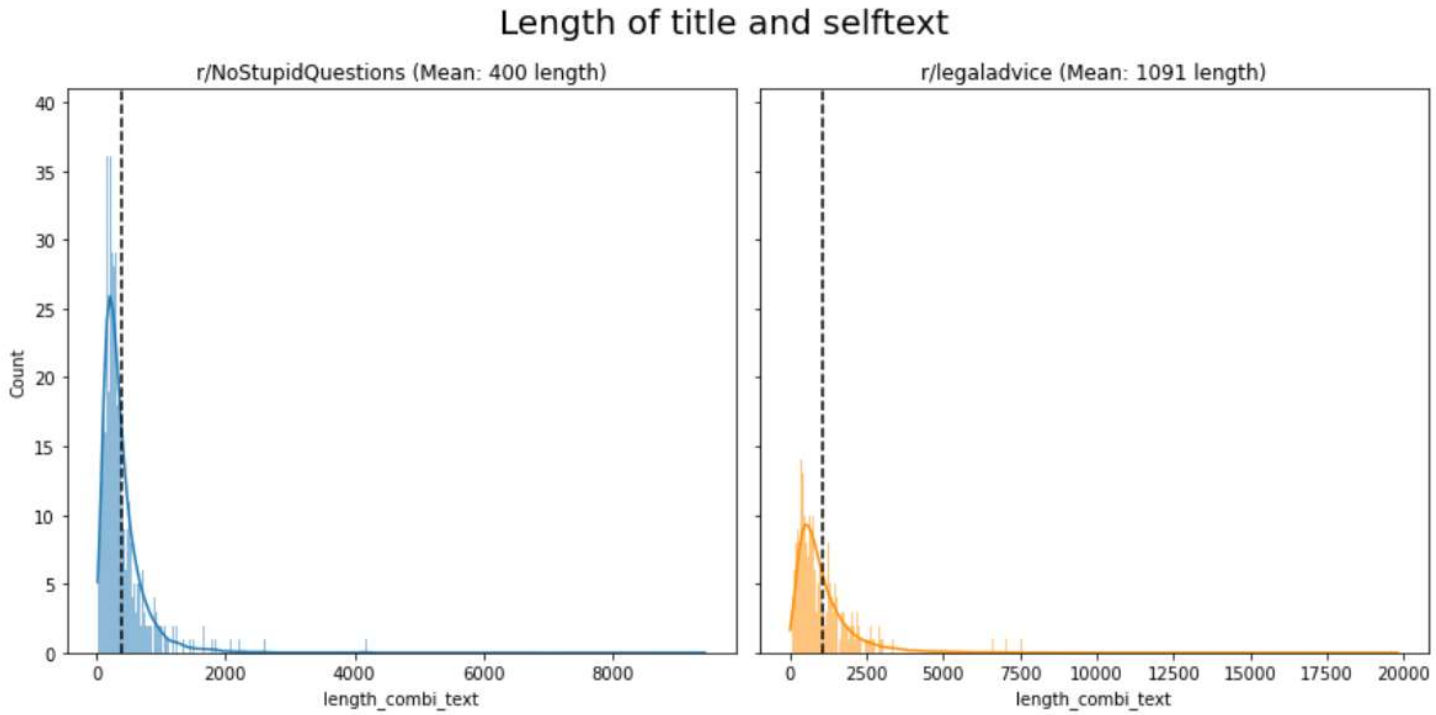
My chatbot company and organisations using my classifier model on top of their chatbot may be able to:

1. Lower their exposure to legal risks by redirecting legal queries elsewhere
2. Reduce monetary losses to their organization and my company
3. Chatbot should know when to not give a reply or at the very least, disclaimers can be given when it is not their domain of expertise
4. Keep the organisation's reputation, prevent embarrassment to the general public

r/ legaladvice proxies real life questions with legal implications

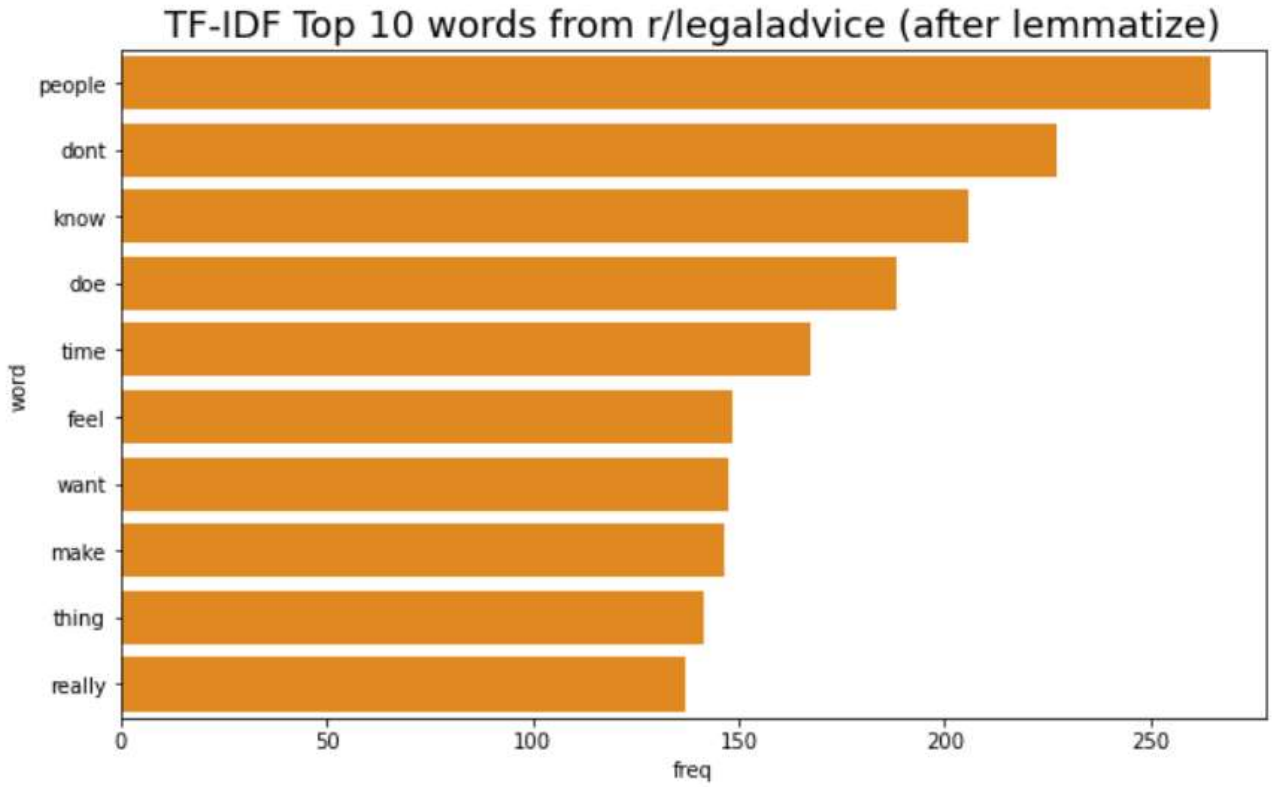
r/NoStupidQuestions proxies real life generic questions

EDA (findings – legaladvice)



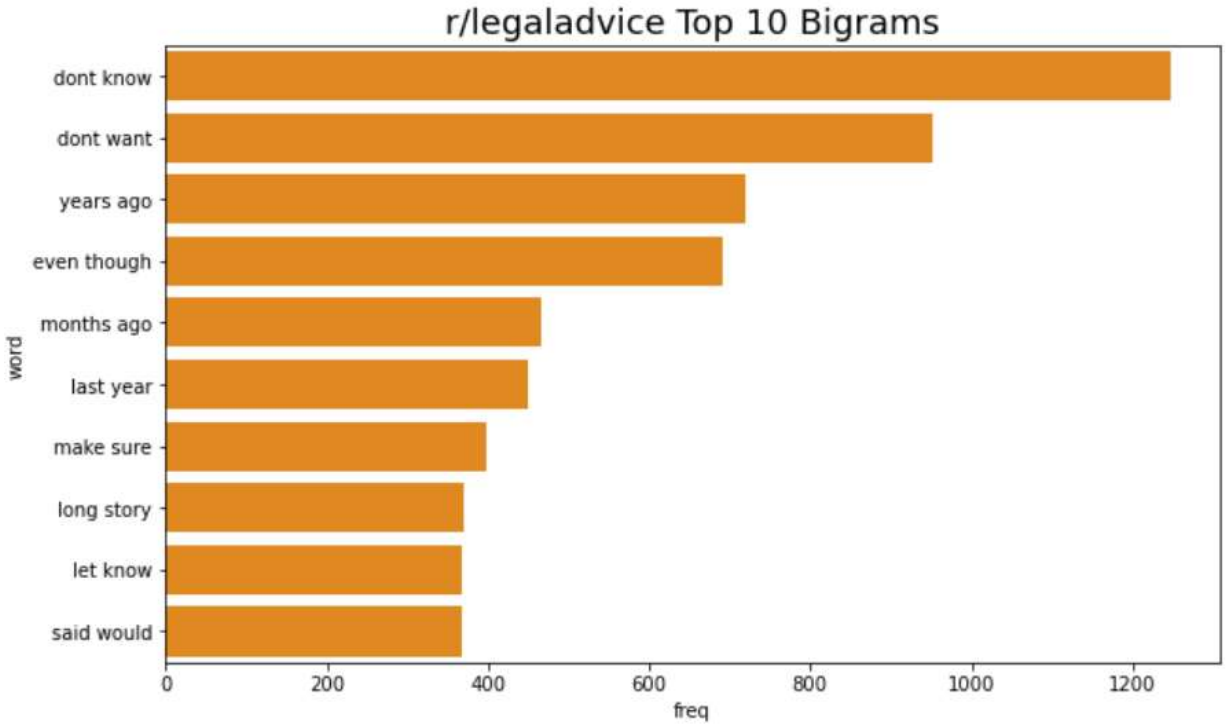
1. Legal – on average lenghtier, due to having to write out the background and context

EDA (findings – legaladvice)



2. Legal – Top 10 words are non jargons

EDA (findings – legaladvice)



3. Legal – Lots of time related word choices

Like years ago, months ago, last year – narrate background

Model evaluation

Number	Vectorizer	Model	Tuning	Test score	Remarks	Sensitivity	False negative	Final model
1	CountVectorizer	Multinomial Naïve Bayes	GridSearch	91.41%	Pales in accuracy to Logreg models	Not assessed - eliminated		
2	TFIDF-Vectorizer	Multinomial Naïve Bayes	GridSearch	91.81%	Pales in accuracy to Logreg models			
3	CountVectorizer	Logistic Regression	GridSearch	93.09%	Final 2 models	91.75%	247	
4	TFIDF-Vectorizer	Logistic Regression	GridSearch	93.44%	Final 2 models	93.15%	205	**CHOSEN**
5	CountVectorizer	Random Forest	RandomizedSearch	91.32%	Takes long for Random Forest to fit, and takes long for GridSearch to optimize	Not assessed - eliminated		
6	TFIDF-Vectorizer	Random Forest	RandomizedSearch	91.24%	Takes long for Random Forest to fit, and takes long for GridSearch to optimize			

Criterias

1) Model must be fast to fit and optimize (chatbox instantaneous reply expected)
2) High accuracy
3) Lowest number of false negatives – evaluated with sensitivity (tp/tp+fn)

False negatives means legal related, but classifier mislabeled as not legal related.

Final model chosen: TFIDF Log Reg

	combi_text	actual	predicted	word_count
13410	estranged parent is ruining my reputation and business va deleted	1	0	10
13352	add clicker is making a bot to click on ads illegal	1	0	11
15250	is ok send bitcoin to russian military charity legal or no	1	0	11
21054	if you are a sole trader can you call yourself a company based in the uk	1	0	16
12793	question how long does a sexualassault investigation take not rape or anything super serious like that	1	0	16

Why false negatives happened:
1) Badly crafted queries, looks generic and non legal (13352, 13410 and 21054)
2) Spelling issues (12793)
3) Stopword containing legal – real world chatbot situation wont have this handicap (13352)

Error analysis – false negatives

SHORTEST WORD COUNTS

16551	was this a sexual assault va just for context before this i had never had sex with a woman before i had three sexual encounters before this all with men one of which was not completely consensual but i didnt say anything about it even before that i had been sexually abusedassaulted by multiple individuals throughout my life one such assault left a deep scar on my genital area that i dont like for people to see for this reason i dont like it when people try to take my clothes off and i dont like it when people try to touch my genitalia i met this woman on an app and we agreed to meet at her place i went there and we got right into it we started kissing and i pinned her to the bed by her arms i asked how she was feeling and she said good so i continued kissing her i asked if shed rather be on top and she said yes so i let her get on top of me and she tried to take my pants off i pushed her hand away but she kept trying so i got back on top of her and she said she didnt like all the struggling so i got off of her and got next to her i started fingering her and asked if she wanted me to continue she said yes after that we cuddled and talked for a bit she texted me today and told me that she had a good time last night but she was scared at first when i was using so much force because she could tell right away that i was stronger and she didnt really know what to do i apologized for scaring her telling her i didnt realize it was to the point that she didnt feel comfortable i asked if she was okay with everything else that happened and she said yes she said she wanted to see me again but we probably should establish better boundaries i dont really know how to feel about this i feel like if she was scared of me then the encounter wasnt completely consensual and i really wasnt trying to be scary	1	0	381

Model does not do well with queries on sex, sexual assault, relationship related questions.

Other queries that model don't do well:
1) Workplace
2) Property and management
3) Tax
4) When police is involved

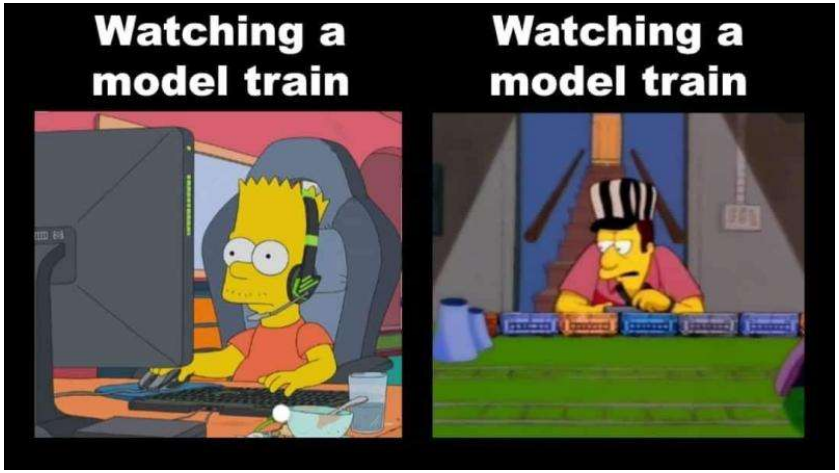
Error analysis – false negatives

LONGEST WORD COUNTS

What I learn (wish I knew 2 weeks ago)

1. Model optimization takes a long time. Wish I knew alternative methods to speed up like n=-1 to use all cores to run GridSearch or RandomizedSearch with low iterations to save time
2. Lemmatizing text may not always improve scores (decreased my scores)
3. Hyperparamater tuning to be considered before model evaluation

   - Base model without tuning may under perform other base models, but outperform others after tuning

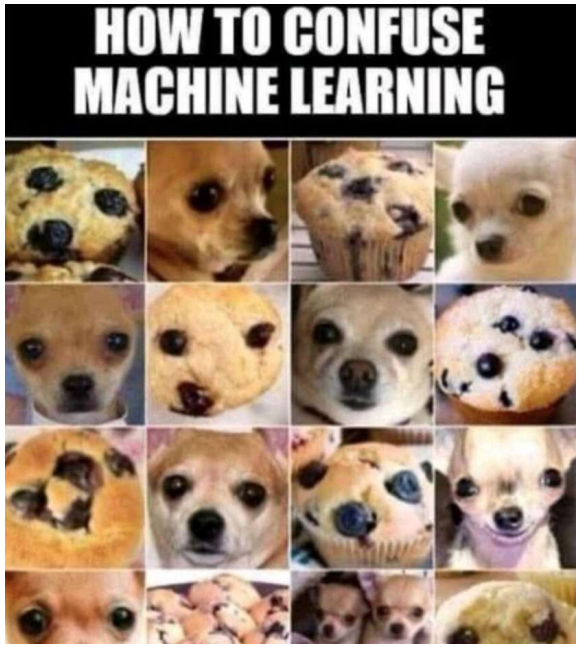


Areas of future work

1) This is merely a step in the correct direction
   - Other important classifier that prevents other misbehaving
   - 1) Suicide related
   - 2) Defamation related

2) Use of other models, like SVM and Boosting to improve accuracy

Areas of future work



3) For organisation to continue gathering of data with good integrity

- There is an assumption that whatever is posted on Reddit is already correctly classified
- This may not be an adequate assumption
- End up training the model wrongly, garbage in garbage out