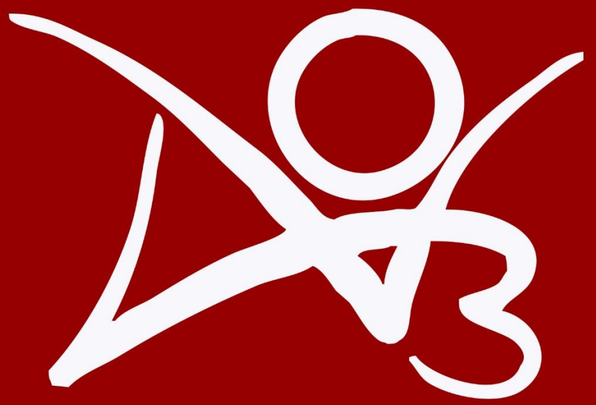


Abuse Detection

in Ao3 comments

1. Problem Statement
2. Data Collection & Cleaning
3. Exploratory Data Analysis
4. Modelling & Model Evaluation
5. Conclusions & Recommendations

Problem statement

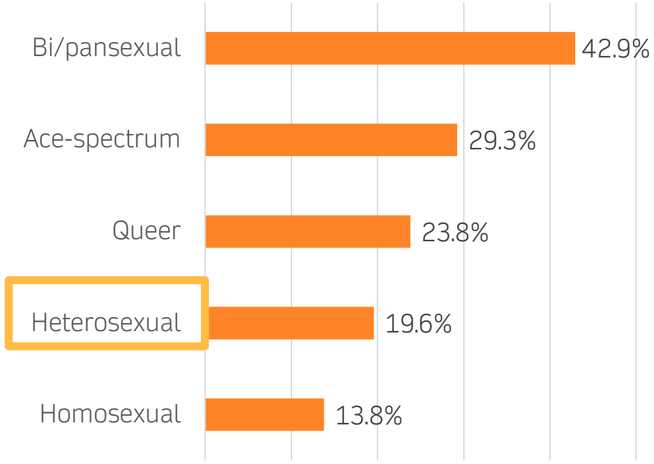
Search

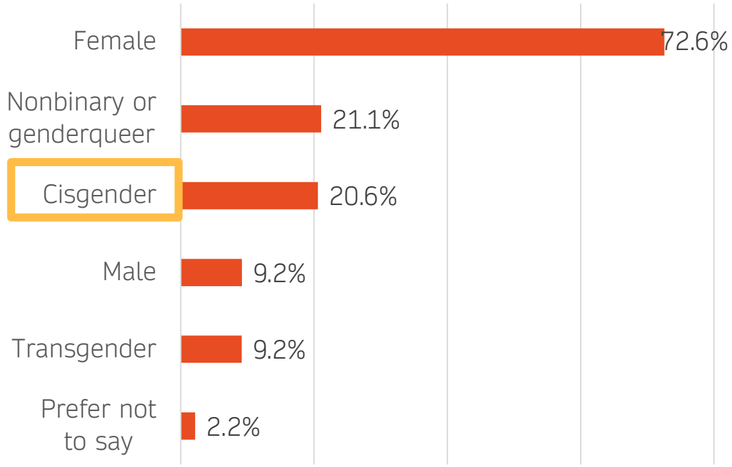
WHAT IS AO3?

- Ao3 is a nonprofit, open-source archive for fanfiction
  - 40 million+ daily visits
  - 7 million+ works
  - 4 million+ users
- Fanfiction is fiction written in an amateur capacity by fans, based on existing works of fiction

WHO CARES ABOUT FANFIC?

- Fanfiction is primarily written and consumed by queer women and genderqueer folks





Why Do Queer People Write Fan Fiction? To See Themselves in Mainstream Culture.

The history of queer representation is one long yellow brick road of **insulting punchlines**. Queer characters are traditionally relegated to **secondary status** and… if a queer person ever does get a shot at the spotlight, they often fall victim to a host of **unfortunate tropes** that are as insulting as they are tired and overdone.

For many queer viewers, favorite movies and TV shows offer a world in which complex, dynamic, actively queer people—people like them—**do not exist**.

Fanfiction is one of the few outlets that an increasingly frustrated queer audience has to engage with material that refuses to engage with them. It represents a challenge to the notion that being created in a straight world means one must live a straight life… It represents a refusal to be punchlines anymore, a refusal to express ourselves less because it's more convenient for everyone else.

In writing fanfiction, we do more than rewrite our favorite stories. We take those stories and make them strong enough to handle people like us.

Problem statement

Search

WHAT IS AO3?

- Ao3 is a nonprofit, open-source archive for fanfiction
  - 40 million+ daily visits
  - 7 million+ works
  - 4 million+ users
- Fanfiction is fiction written in an amateur capacity by fans, based on existing works of fiction

WHY CARE ABOUT FANFIC?

- Fanfiction is a crucial form of expression and representation for marginalized communities
- Online communities based around fanfiction are safe havens for marginalized individuals to meet, mingle, and bond over fan content

Problem statement

Search

CHALLENGES

Censorship by webhosts

- Livejournal Strikethrough
- Fanfiction.net's R ban
- Tumblr's NSFW ban

Legal threats

The Anne Rice crackdown on fanfic authors in 2000

BIRTH OF AO3

Established specially for hosting fanfiction

Anti-censorship

Legal advocacy

CRITICISMS

Abuse & harassment

Lack of features to block harassment & hate speech

- Turn off anon comments
- Comment moderation

Racism

Racist trolling

Racist harassment of Ao3 authors of colour

Problem statement

Search

My goal is to build a text classification model that can flag abusive comments. If successful, this model could serve as proof of concept for an “automated comment moderation” function, which could be used to:

(1) automatically filter abusive comments into a separate inbox

(2) flag trolls and hateful accounts to the Abuse Team

Hateful speech and online abuse is a serious issue

30% of polled users have been harassed by commenters on Ao3

Who is going to do the moderation? Ao3 is run by volunteers.

1. Problem Statement
2. Data Collection & Cleaning
3. Exploratory Data Analysis
4. Modelling & Model Evaluation
5. Conclusions & Recommendations

Data collection

Search

KEY CHALLENGE

Lack of a large and diverse enough dataset of hateful and abusive Ao3 comments

TRAINING DATASET

- 100,000 Wikipedia discussion comments rated for toxicity (1–5), aggression (1-7), and whether the comment contained a personal attack (0 or 1)
- Due to imbalanced classes, the non-abusive class was undersampled, leaving 15,000 comments with an even 50% split

FINAL TEST DATASET

- 320 Ao3 comments, 50% of which were flagged as abusive
- Comments were sourced from social media posts about trolls in a specific fan community
  - Two datasets of abusive comments
  - Several documents containing pictures of comments from known trolls
  - Pictures of comments posted directly to Twitter

Data cleaning

Search

WIKIPEDIA SET (TRAINING)

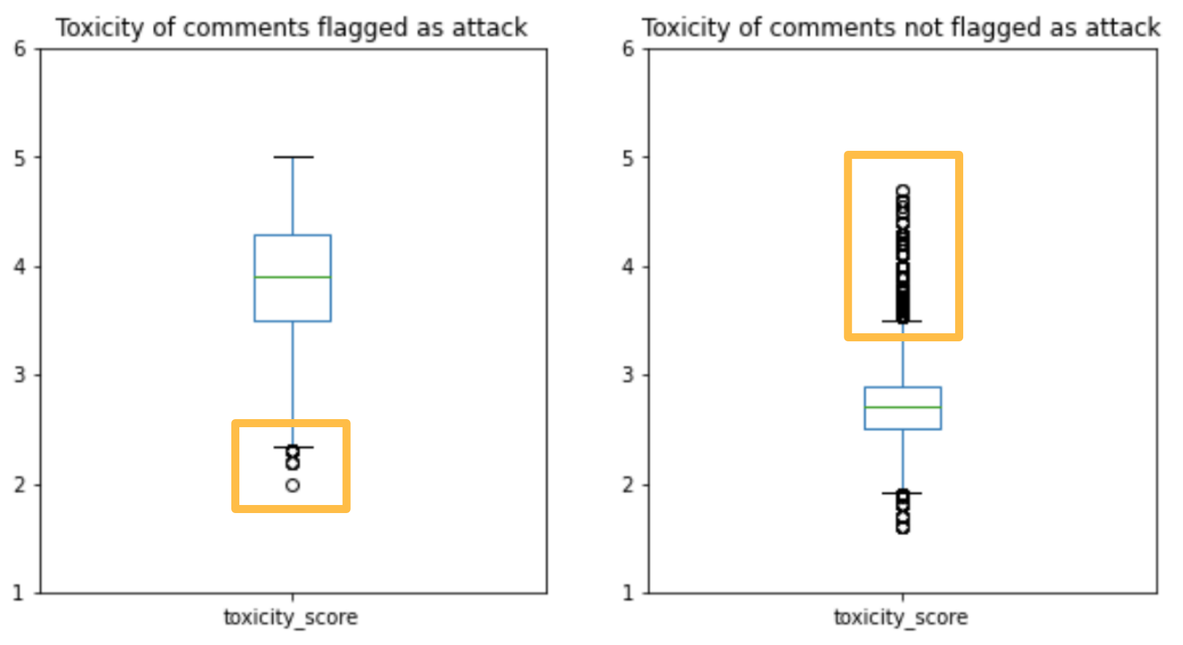
- Feature engineering
  - attack: 1 if majority scored 1, else 0
  - toxicity (1-5): mean score given
  - aggression (1-7): mean score given
- Removal of outliers (mean ± 3SD)
  - Comments that were abnormally long

Data cleaning

Search

WIKIPEDIA SET (TRAINING)

- Feature engineering
  - attack: 1 if majority scored 1, else 0
  - toxicity (1-5): mean score given
  - aggression (1-7): mean score given
- Removal of outliers (mean ± 3SD)
  - Comments that were abnormally long
  - Comments flagged “attack” but with low toxicity or aggression
  - Comments flagged “not attack” but with high toxicity or aggression
- Undersampling of majority class



HOPE YOUR HEAD GETS CUT OFF AND SOMEONE WIPES...
== you are dumb == Mexican Punk is the god of...
::Haha, I fucking pissed myself reading this ...
you people are cunts, bombing every ones mail ...

Data cleaning

Search

WIKIPEDIA SET (TRAINING)

AO3 SET (FINAL TEST)

Remove punctuation

Tokenize

Remove stopwords

Lemmatize

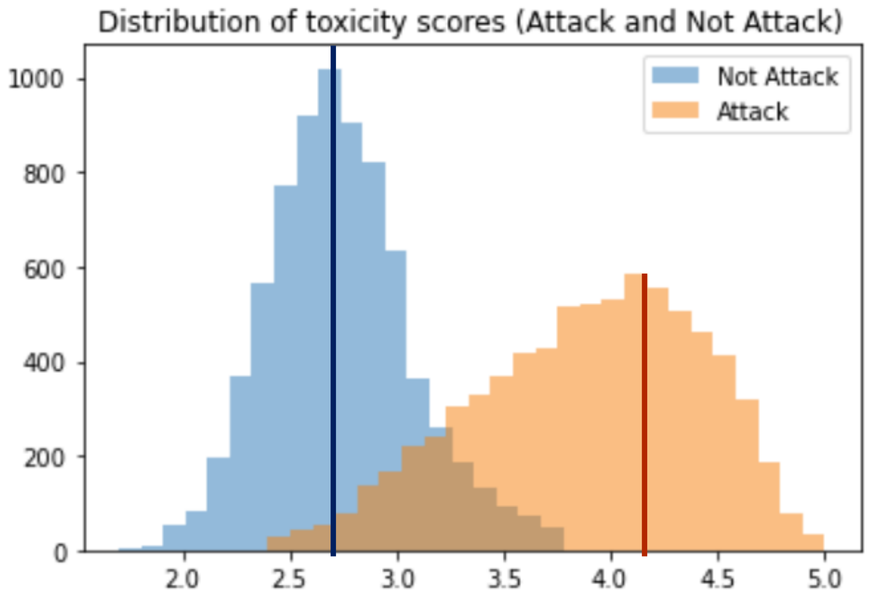
1. Problem Statement
2. Data Collection & Cleaning
3. Exploratory Data Analysis
4. Modelling & Model Evaluation
5. Conclusions & Recommendations

EDA (Wikipedia)

Search

TOXICITY SCORE

Average score of 3.3 with 3 representing
“Neither toxic nor healthy”

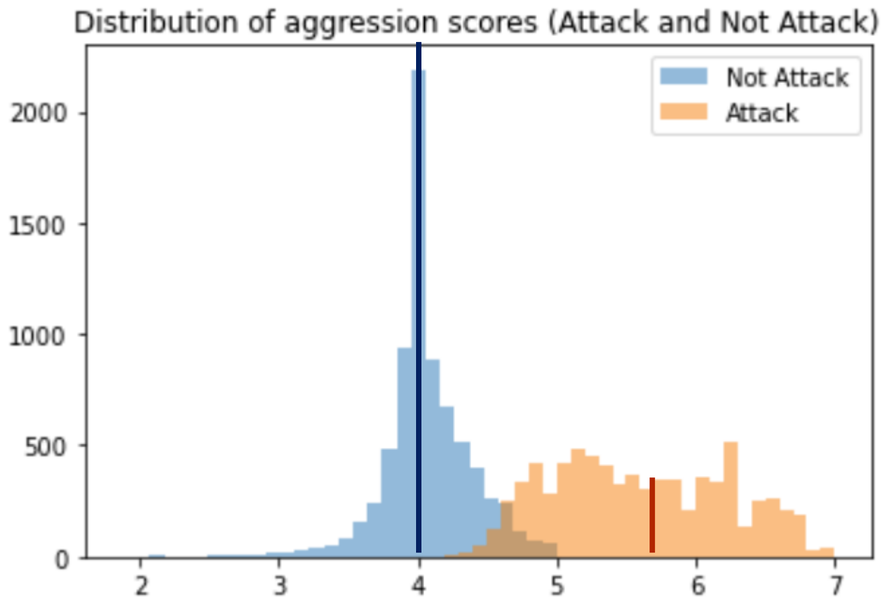


2.7
Not attack

3.9
Attack

AGGRESSION SCORE

Average score of 4.8 with 4 representing
“Neither aggressive nor friendly”



4.1
Not attack

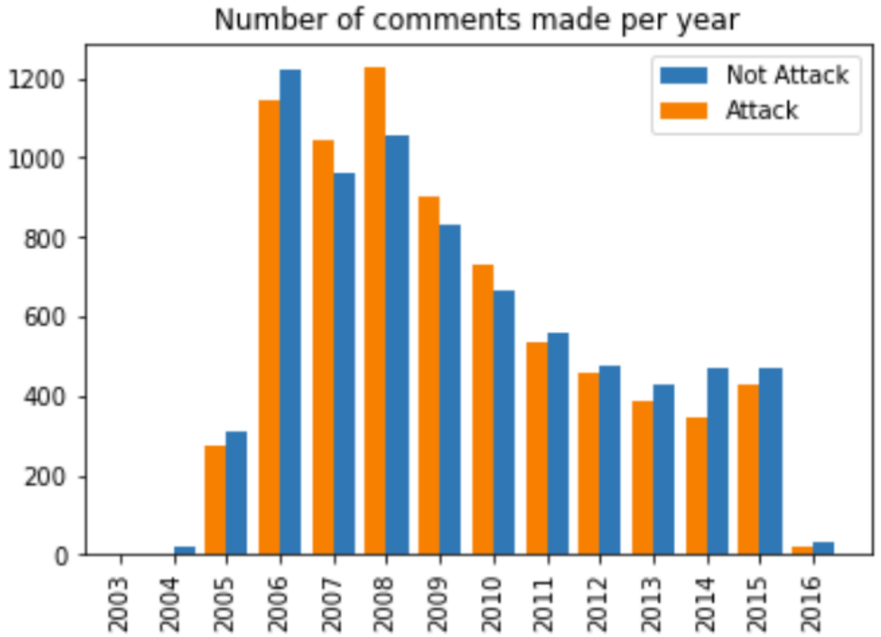
5.6
Attack

EDA (Wikipedia)

Search

YEAR OF COMMENT

- Posted between 2003 to 2016
- No difference between attack and not



LOGIN STATUS

- Majority of abusive comments were made by anonymous commenters

	Logged In	Attack	Not Attack	Total
0	Yes	46.3%	71.1%	58.7%
1	No	53.7%	28.9%	41.3%

- However, it's notable that there was still a significant number of logged-in users (46.3%) making abusive comments

EDA (Wikipedia)

Search

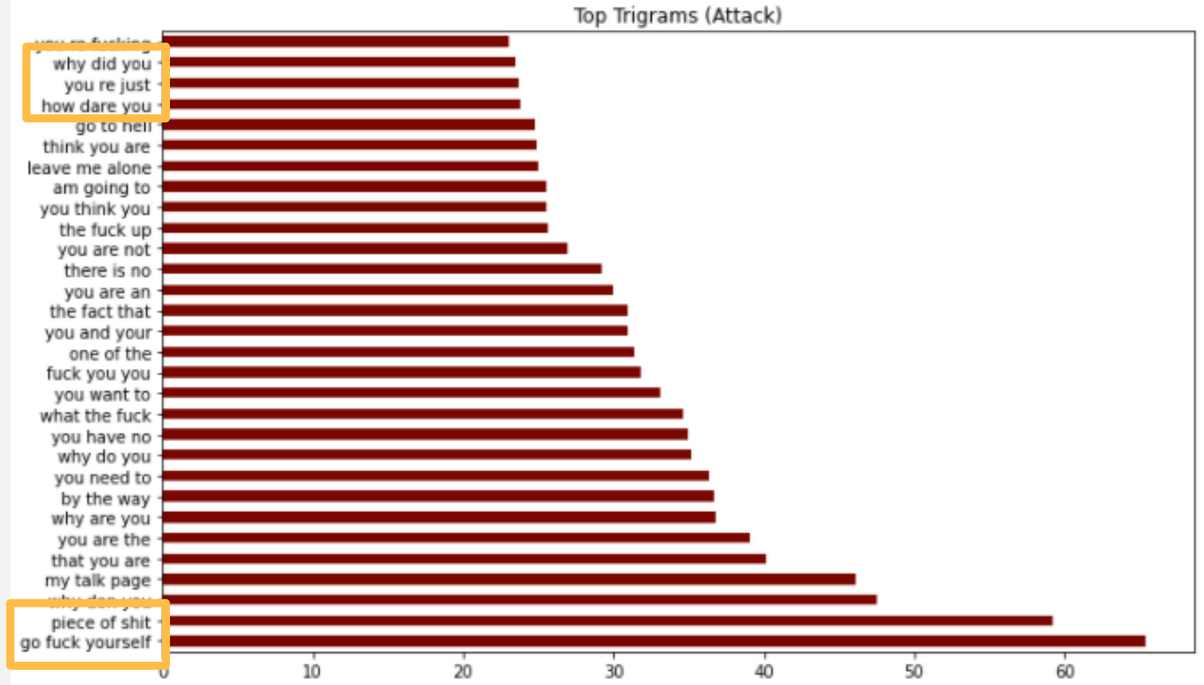
High in profanity

Polite and objective

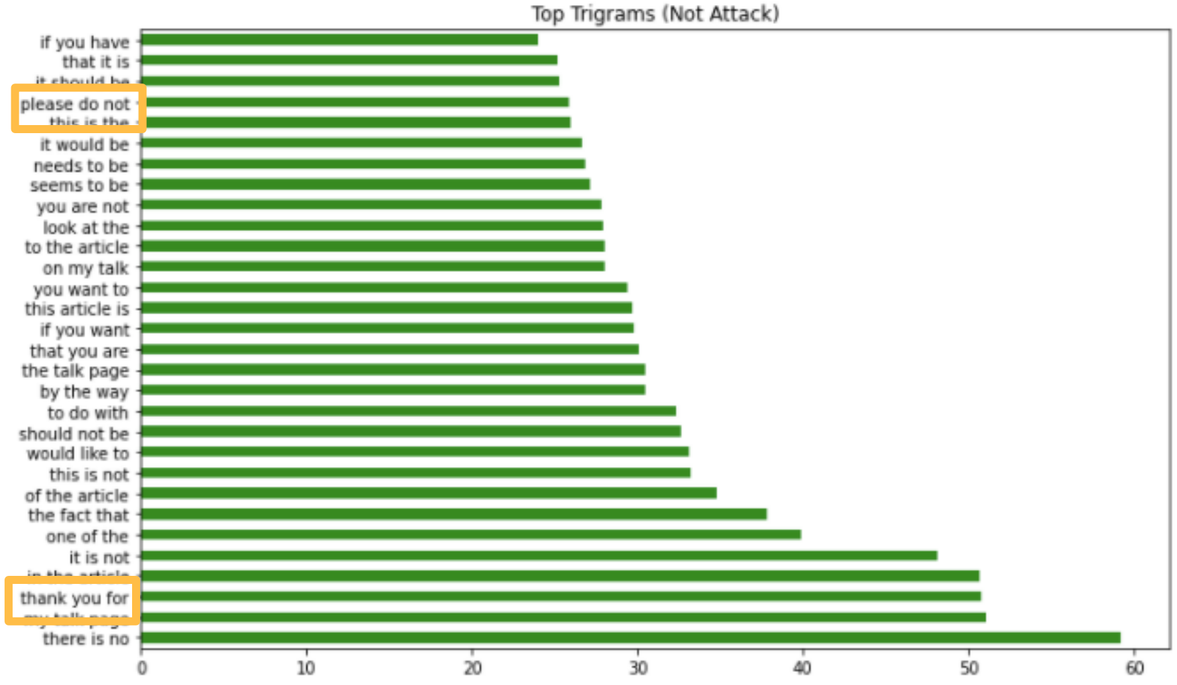
EDA (Wikipedia)

Search

High in profanity



Polite and objective



EDA (Ao3)

Search

AGGRESSORS (TROLLS)

	Comments	Logged in	Type of comments
Troll 1	45	69% of the time	Attacks on writing. Low to medium toxicity.
Troll 2	30	Always anon	Violent language . High toxicity.
Troll 3	32	62% of the time	Insults and hate speech. High toxicity.
Troll 4	42	Always anon	Attacks on writing. Medium toxicity. Sarcasm.
Unknown	11	18% of the time	Mixed

→ kill urself please

dear author, i hope someone precious to you dies- maybe ur parents?

→ still i just think you werent bullied enough for being half horse half rat freak bc you still seem to have the nerve to exist. how odd.

kinda amazing that anyone would want to display theyre a hideous middle aged tranny to everyone, but alas

EDA (Ao3)

Search

AGGRESSORS (TROLLS)

	Comments	Logged in	Type of comments
Troll 1	45	69% of the time	Attacks on writing. Low to medium toxicity.
Troll 2	30	Always anon	Violent language . High toxicity.
Troll 3	32	62% of the time	Insults and hate speech. High toxicity.
Troll 4	42	Always anon	Attacks on writing. Medium toxicity. Sarcasm.
Unknown	11	18% of the time	Mixed

→ Felt rushed. But other than that, a pretty okay story :)

Genuinely disgusted EW what the FUCK SHSMS

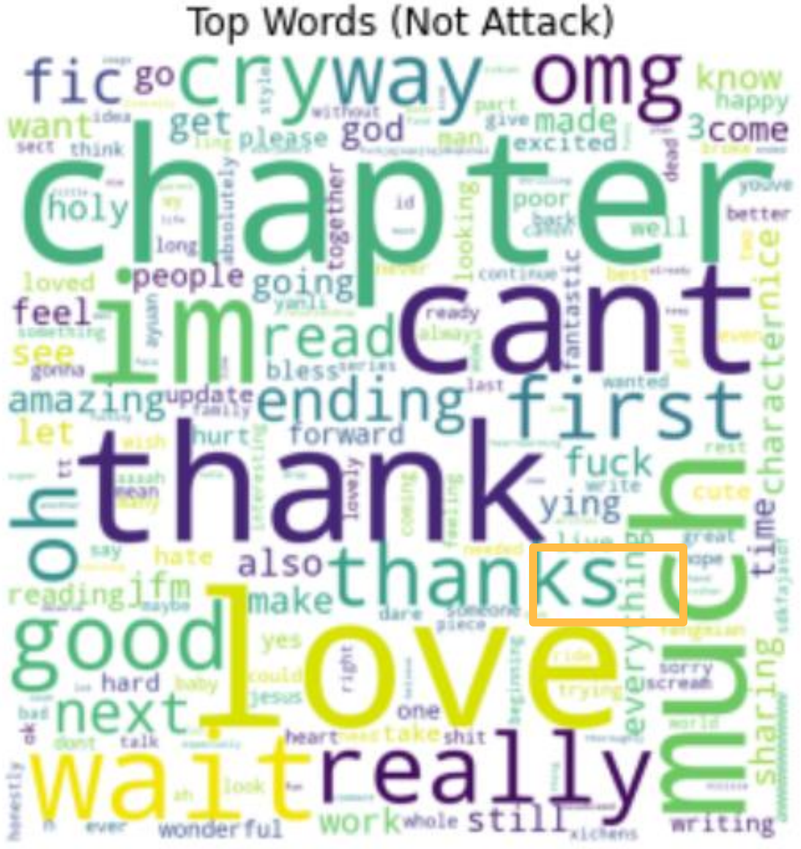
→ lol wow love how you completely gutted their personalities to write this! great job! :D :D :D

Fantastic self-insert where you could shove all your favorite pairings together no matter how illogical and change the plot to suit your own fantasies :D

EDA (Ao3)

Search

Fanfiction-specific vocab



Excited and grateful

EDA (Ao3)

Search

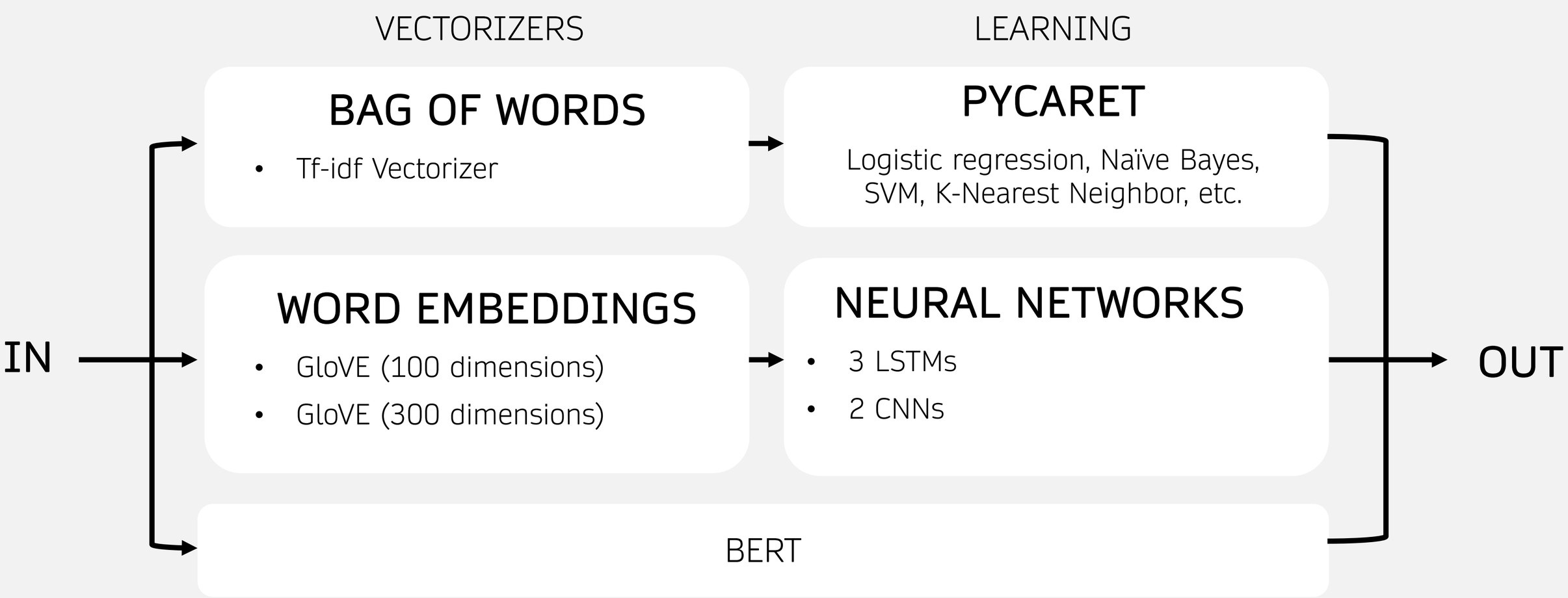
AGGRESSORS (TROLLS)

	Comments	Logged in	Type of comments
Troll 1	45	69% of the time	Attacks on writing. Low to medium toxicity.
Troll 2	30	Always anon	Violent language . High toxicity.
Troll 3	32	62% of the time	Insults and hate speech. High toxicity.
Troll 4	42	Always anon	Attacks on writing. Medium toxicity. Sarcasm.
Unknown	11	18% of the time	Mixed

1. Problem Statement
2. Data Collection & Cleaning
3. Exploratory Data Analysis
4. Modelling & Model Evaluation
5. Conclusions & Recommendations

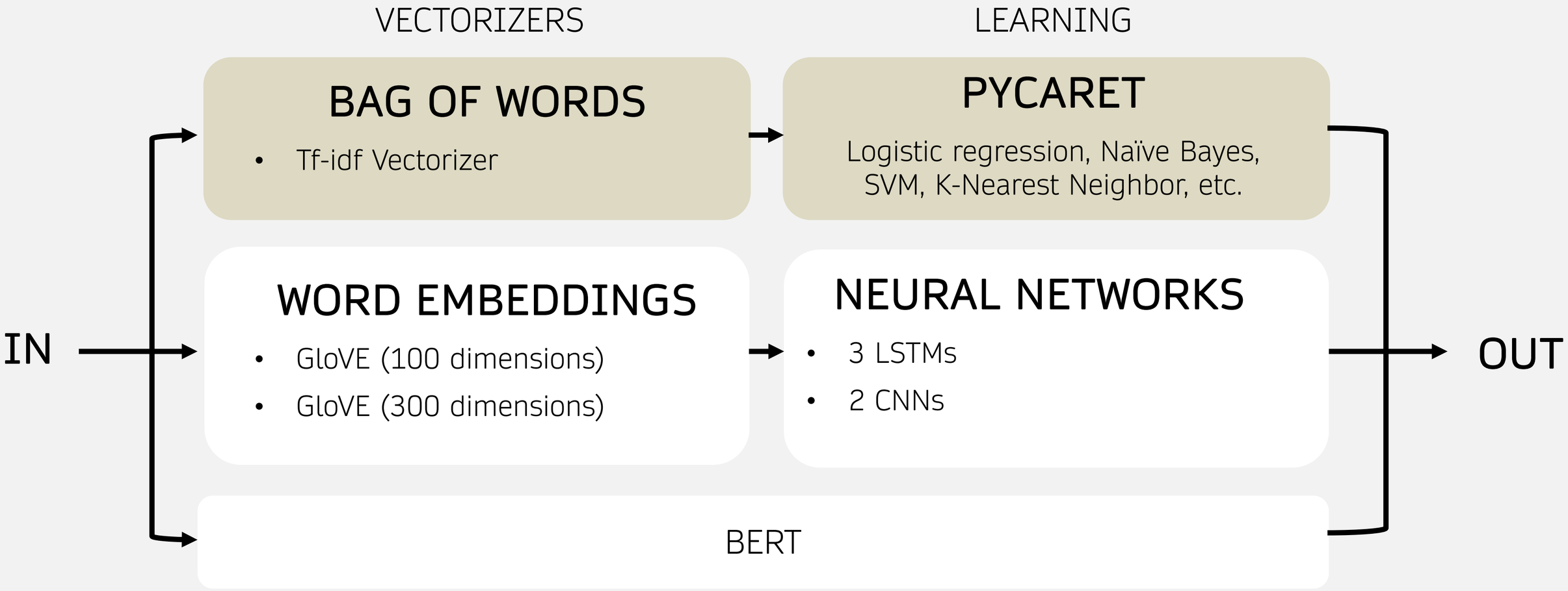
Summary of modelling

Search



Summary of modelling

Search



Model Evaluation (Bag-of-Words)

Search

INITIAL MODEL RESULTS

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
svm	SVM - Linear Kernel	0.8392	0.0000	0.7778	0.8872	0.8286	0.6785	0.6840	1.0400
lr	Logistic Regression	0.8387	0.9141	0.7864	0.8786	0.8297	0.6773	0.6814	2.5200
et	Extra Trees Classifier	0.8362	0.9111	0.7829	0.8766	0.8269	0.6724	0.6764	16.3910
rf	Random Forest Classifier	0.8257	0.9032	0.7607	0.8747	0.8135	0.6514	0.6572	8.9020
ridge	Ridge Classifier	0.8125	0.0000	0.7690	0.8426	0.8039	0.6249	0.6276	1.6290
ada	Ada Boost Classifier	0.7918	0.8521	0.6946	0.8627	0.7693	0.5836	0.5953	6.6680
gbc	Gradient Boosting Classifier	0.7916	0.8660	0.6417	0.9168	0.7547	0.5833	0.6116	24.8980
dt	Decision Tree Classifier	0.7808	0.7855	0.7894	0.7763	0.7827	0.5617	0.5619	14.6340
lda	Linear Discriminant Analysis	0.7421	0.8084	0.7428	0.7419	0.7421	0.4841	0.4844	45.5230
nb	Naive Bayes	0.7305	0.7330	0.8107	0.6989	0.7506	0.4609	0.4670	0.4740
qda	Quadratic Discriminant Analysis	0.5998	0.6000	0.2207	0.9135	0.3547	0.1998	0.3062	70.7290
knn	K Neighbors Classifier	0.5511	0.7281	0.9448	0.5402	0.6805	0.1019	0.1454	41.6640
dummy	Dummy Classifier	0.5002	0.5000	1.0000	0.5002	0.6669	0.0000	0.0000	0.1010

MODEL TUNING

- Alpha = 0.0002
- L1 ratio = 0.08
- Penalty = ElasticNet

	Train	Test	Ao3
Precision	88.4%	89.4%	59.3%
Recall	78.3%	80.5%	55.6%
Accuracy	84.0%	85.5%	58.8%

Error Analysis (Bag-of-Words)

Search

Misclassifies abusive comments that isn't profane but has criticisms that are specific to fanfiction

Misclassifies positive comments that have profanity, or that includes negative comments on an antagonist

Error Analysis (Bag-of-Words)

Search

"fuck, imma gonna crie"

"oh fuck oh fuck it's going down and im not ready. we're about to be massacred with feels"

"CAN SOMEONE FUCKING DECK HIS DAD IN THE FACE OMG HIS DAD IS SUCH A PIECE OF SHIT!!!!"

Misclassifies abusive comments that isn't profane but has criticisms that are specific to fanfiction

Misclassifies positive comments that have profanity, or that includes negative comments on an antagonist

Error Analysis (Bag-of-Words)

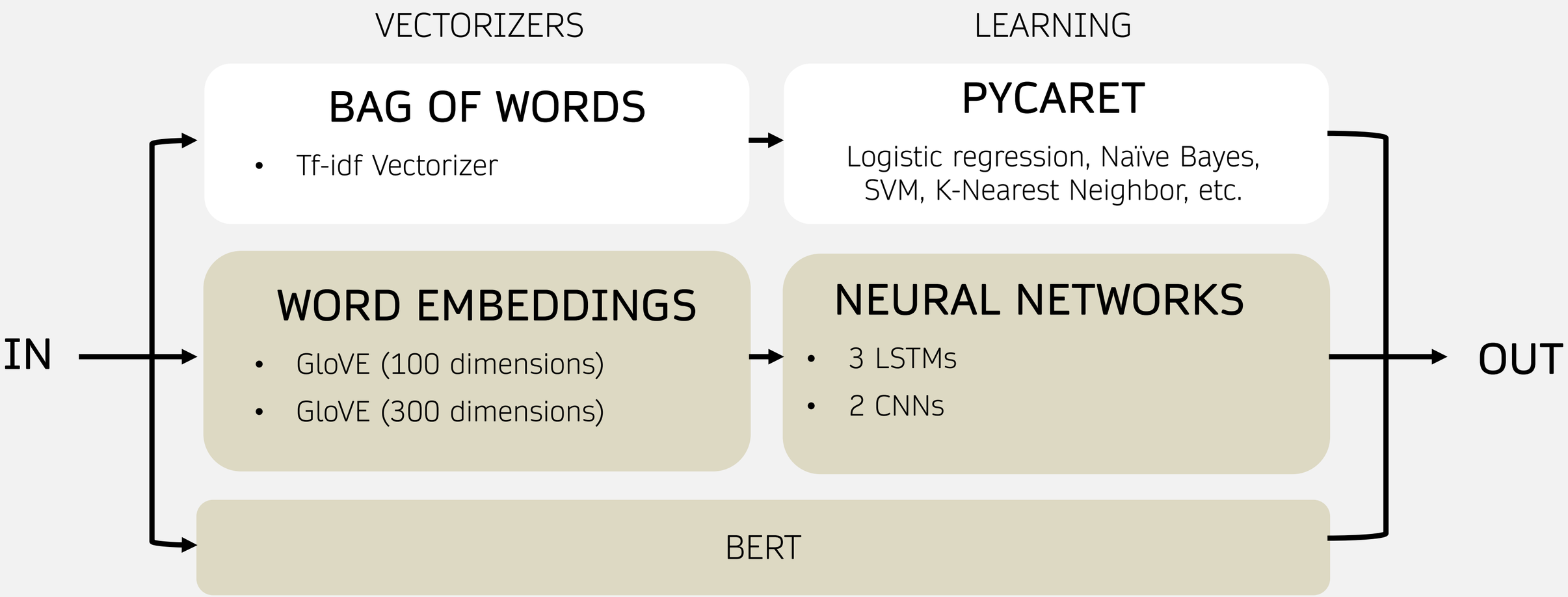
Search

AGGRESSORS (TROLLS)

	Comments	Flagged	Percent	Type of comments
Troll 1	45	16	35.6%	Attacks on writing. Low to medium toxicity.
Troll 2	30	21	70.0%	Violent language . High toxicity.
Troll 3	32	26	81.2%	Insults and hate speech. High toxicity.
Troll 4	42	21	50.0%	Attacks on writing. Medium toxicity. Sarcasm.
Unknown	11	5	45.5%	Mixed

Summary of modelling

Search

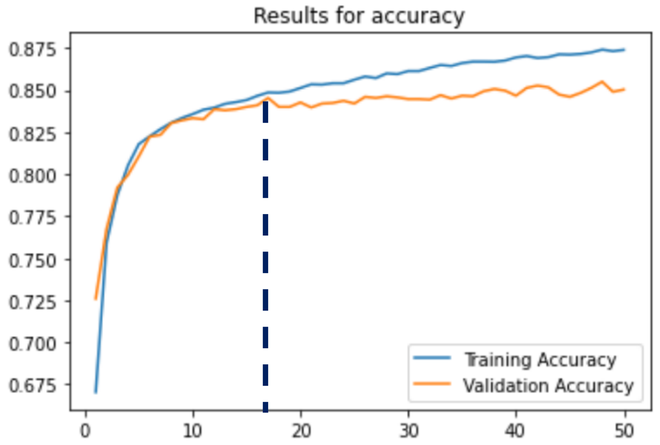


Model Evaluation (LSTMs)

Search

LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch



Model Evaluation (LSTMs)

Search

LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

Model Evaluation (LSTMs)

Search

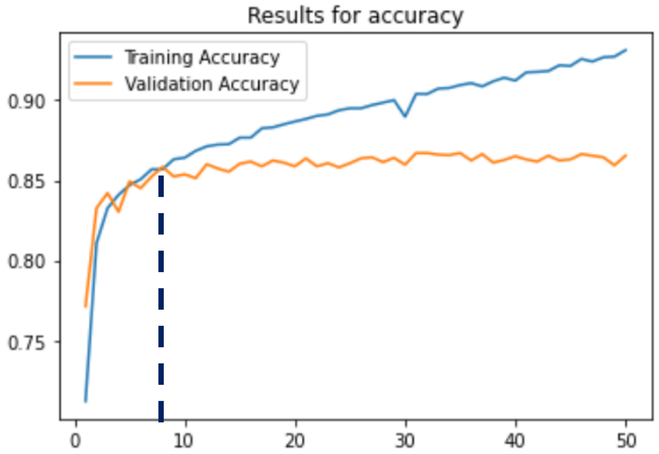
LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

LSTM with GloVE 300d

- 300-dimensional GloVE
- LSTM layer with 24 nodes
- Trained to 8th epoch



Model Evaluation (LSTMs)

Search

LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

LSTM with GloVE 300d

- 300-dimensional GloVE
- LSTM layer with 24 nodes
- Trained to 8th epoch

	Train	Test	Ao3
Precision	89.7%	87.1%	63.3%
Recall	84.3%	80.7%	50.6%
Accuracy	87.3%	85.1%	60.6%
Baseline accuracy	84.0%	85.5%	58.8%

Model Evaluation (LSTMs)

Search

LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

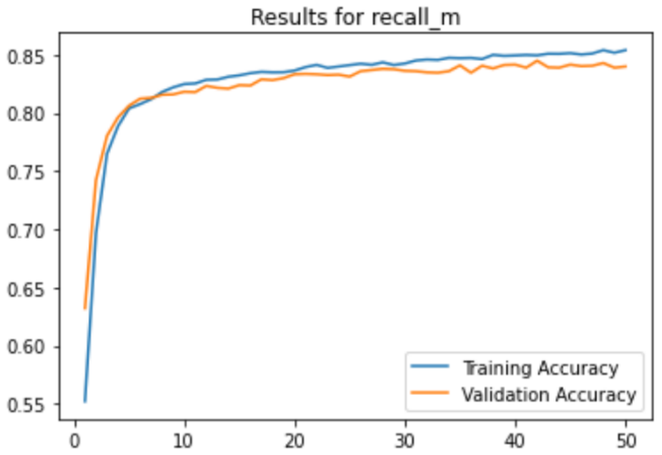
LSTM with GloVE 300d

- 300-dimensional GloVE
- LSTM layer with 24 nodes
- Trained to 8th epoch

	Train	Test	Ao3
Precision	89.7%	87.1%	63.3%
Recall	84.3%	80.7%	50.6%
Accuracy	87.3%	85.1%	60.6%
Baseline accuracy	84.0%	85.5%	58.8%

Bi-LSTM with GloVE 100d

- 100-dimensional GloVE
- Bi-LSTM layer with 10 nodes
- L2 weights regularization



Model Evaluation (LSTMs)

Search

LSTM with GloVE 100d

- 100-dimensional GloVE
- LSTM layer with 10 nodes
- Trained to 17th epoch

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

LSTM with GloVE 300d

- 300-dimensional GloVE
- LSTM layer with 24 nodes
- Trained to 8th epoch

	Train	Test	Ao3
Precision	89.7%	87.1%	63.3%
Recall	84.3%	80.7%	50.6%
Accuracy	87.3%	85.1%	60.6%
Baseline accuracy	84.0%	85.5%	58.8%

Bi-LSTM with GloVE 100d

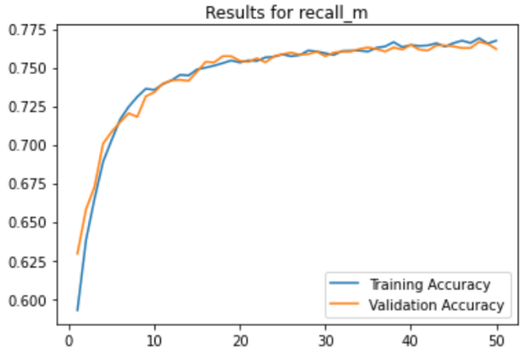
- 100-dimensional GloVE
- Bi-LSTM layer with 10 nodes
- L2 weights regularization

	Train	Test	Ao3
Precision	83.7%	83.5%	58.6%
Recall	88.3%	84.8%	59.4%
Accuracy	85.6%	84.0%	58.8%
Baseline accuracy	84.0%	85.5%	58.8%

Model Evaluation (CNNs)

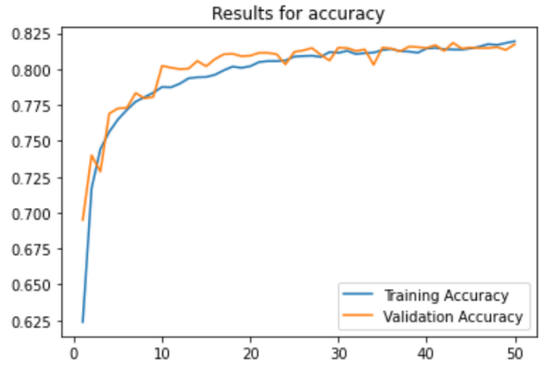
Search

1D CNN with GloVE 100d



	Train	Test	Ao3
Precision	79.2%	79.7%	51.9%
Recall	72.7%	70.3%	52.5%
Accuracy	76.8%	76.2%	51.9%
Baseline accuracy	84.0%	85.5%	58.8%

1D CNN with GloVE 300d

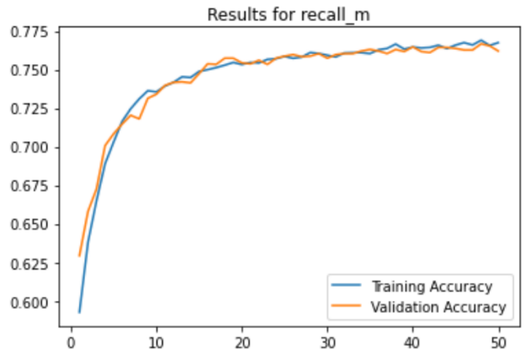


	Train	Test	Ao3
Precision	80.7%	81.9%	60.7%
Recall	83.2%	81.4%	63.7%
Accuracy	81.6%	81.7%	61.3%
Baseline accuracy	84.0%	85.5%	58.8%

Model Evaluation (CNNs & BERT)

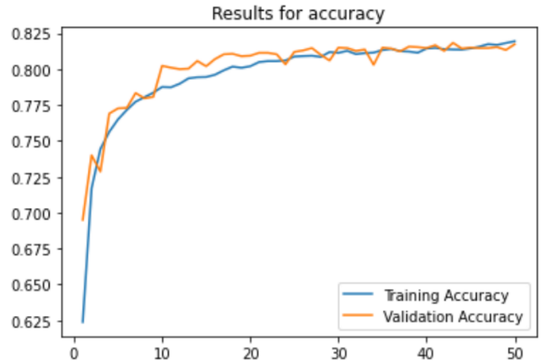
Search

1D CNN with GloVE 100d



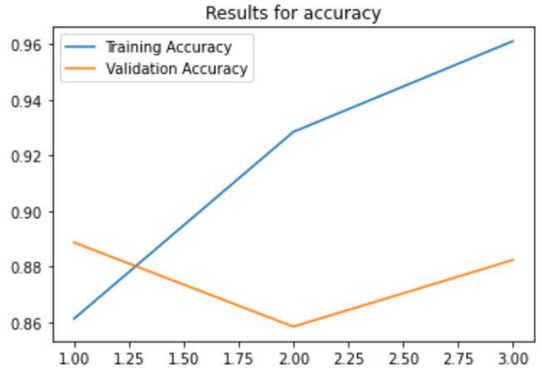
	Train	Test	Ao3
Precision	79.2%	79.7%	51.9%
Recall	72.7%	70.3%	52.5%
Accuracy	76.8%	76.2%	51.9%
Baseline accuracy	84.0%	85.5%	58.8%

1D CNN with GloVE 300d



	Train	Test	Ao3
Precision	80.7%	81.9%	60.7%
Recall	83.2%	81.4%	63.7%
Accuracy	81.6%	81.7%	61.3%
Baseline accuracy	84.0%	85.5%	58.8%

BERT



	Train	Test	Ao3
Precision	98.7%	87.4%	60.4%
Recall	99.1%	89.3%	54.4%
Accuracy	98.9%	88.2%	59.4%
Baseline accuracy	84.0%	85.5%	58.8%

Model Evaluation (NNs)

Search

LSTM with GloVE 100d

	Train	Test	Ao3
Precision	86.3%	87.1%	65.1%
Recall	83.8%	80.5%	60.6%
Accuracy	85.2%	84.3%	64.1%
Baseline accuracy	84.0%	85.5%	58.8%

LSTM with GloVE 300d

	Train	Test	Ao3
Precision	89.7%	87.1%	63.3%
Recall	84.3%	80.7%	50.6%
Accuracy	87.3%	85.1%	60.6%
Baseline accuracy	84.0%	85.5%	58.8%

Bi-LSTM with GloVE 100d

	Train	Test	Ao3
Precision	83.7%	83.5%	58.6%
Recall	88.3%	84.8%	59.4%
Accuracy	85.6%	84.0%	58.8%
Baseline accuracy	84.0%	85.5%	58.8%

1D CNN with GloVE 100d

	Train	Test	Ao3
Precision	79.2%	79.7%	51.9%
Recall	72.7%	70.3%	52.5%
Accuracy	76.8%	76.2%	51.9%
Baseline accuracy	84.0%	85.5%	58.8%

1D CNN with GloVE 300d

	Train	Test	Ao3
Precision	80.7%	81.9%	60.7%
Recall	83.2%	81.4%	63.7%
Accuracy	81.6%	81.7%	61.3%
Baseline accuracy	84.0%	85.5%	58.8%

BERT

	Train	Test	Ao3
Precision	98.7%	87.4%	60.4%
Recall	99.1%	89.3%	54.4%
Accuracy	98.9%	88.2%	59.4%
Baseline accuracy	84.0%	85.5%	58.8%

Error Analysis (LSTM)

Search

Top Words (False Positives)

Top Words (False Negatives)

Still misclassifies abusive comments that has criticisms that are specific to fanfiction

Still misclassifies positive comments that contain profanities

Error Analysis (LSTM)

Search

AGGRESSORS (TROLLS)

	Comments	Base	LSTM	Type of comments
Troll 1	45	35.6%	35.6%	Attacks on writing. Low to medium toxicity.
Troll 2	30	70.0%	70.0%	Violent language . High toxicity.
Troll 3	32	81.2%	81.2%	Insults and hate speech. High toxicity.
Troll 4	42	50.0%	64.4%	Attacks on writing. Medium toxicity. Sarcasm.
Unknown	11	45.5%	72.7%	Mixed

Error Analysis (LSTM)

Search

- **Model doesn't do well on positive comments that contain profanity**

  “that was fucking amazing! thank you!” – 0.8757

- **Model does decently on sarcastic comments**

  “amazingly illogical and senseless! good job!“ – 0.8310

- **But model is sensitive to typos - prediction changed quite drastically**

  “amazingly ilogical and senseless! good job!“ – 0.6574

- **Model doesn't do well with negative vocabulary that it wasn't trained on**

  “amazingly ooc! good job!” – 0.1783

1. Problem Statement
2. Data Collection & Cleaning
3. Exploratory Data Analysis
4. Modelling & Model Evaluation
5. Conclusions & Recommendations

Conclusions & Recommendations

Search

- Achieved baseline accuracy of 58.8% with SVM and tf-idf vectorization
- LSTM model outperformed baseline model with accuracy of 64.1%
  - Model doesn't do well on positive comments that contain profanity
  - Model doesn't do well with vocabulary that it wasn't trained on
  - Model is sensitive to typos - prediction changed quite drastically
  - Can be helped by using character-level rather than word-level embeddings.
  - Can’t be helped. Requires more tailored training content.
    - Choose better dataset to train on
    - Label Ao3 data

Conclusions & Recommendations

Search

1

More advanced models

- Deep learning models implemented are relatively shallow models
- Try deeper layers, hybrid models, better architecture
- Try tailored pre-trained models like HateBERT

2

Choose better train data

- Style of conversation and typing is more formal on Wikipedia than on Ao3
- Try to find train data that better matches the writing style and vocabulary used on Ao3
- This is potentially difficult

3

Collect and label Ao3 data

- Best way is to integrate a data collection system into the platform and let users flag content for you – but this requires action from the platform managers
- If manually labelled, it will be manpower intensive, but will probably lead to the best performance