Playing Offense Against Hate: Detecting Hate Speech

CloverWorks: Daryl, Joel, Lionel, Wenzhe

Disclaimer: Presentation includes offensive language.

Hate Speech in the News

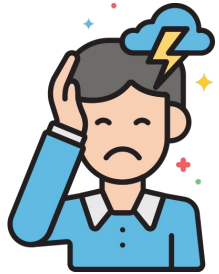
TikTok removes Singapore-born comedian's controversial video, citing hate speech

Govt must intervene early before hate speech disrupts racial, religious harmony in S'pore: Shanmugam

Singapore takes zero tolerance approach on hate speech: Teo Chee Hean

Background

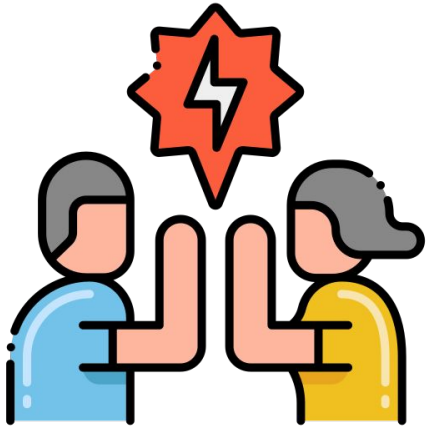
- **Mental Health**: State of mental well-being that enables people to cope with the stressors of life.
- Poor mental health in Singapore increased from **13.4% in 2020** to **17.0% in 2022**.
- **Anxiety disorders** are a type of mental health condition.
- **Hate Speech** can causes a **rise** in **clinical anxiety levels** among viewers.
- Productivity lost due to depression and **anxiety** may cost Singapore up to **$16b** a year.



Background

Beyond anxiety, **Hate Speech**:

- Causes **economic negative externalities** on non-consenting viewers.
- **Inhibits** people's ability to **understand one another**.
- **Prevents / discourages** participation in **social/political discourse**
  - e.g. political participation
- Can lead to **dangerous divisions** in society.

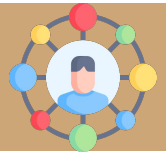


Problem Statement

- We are a **Cyber Wellness Non-Profit** in Singapore.
- Specifically, we want to create a safer online browsing experience for users by:
  - Blocking **hateful / offensive posts**
  - Blocking **users** that make such posts

Statement: We want to empower internet **users** and **community leaders (moderators)** alike to **safeguard their mental well being** against online hate speech.

Who it Matters to?

	 Doe	 Ray	 Far
who	Tired **student** who **unwinds** by browsing **social media**	**Parent** who is **introducing** internet and social media to **children**	**Community leader / moderator** on a social media platform
pains	**Disturbed** by hateful / offensive posts	**Concerned** about exposure to unsafe and unsavory content	Too many posts to review
goal	Wants to **avoid** hateful / offensive posts		Wants to **identify** potentially hateful / offensive posts and their users

Data - Source

- Random sampling of Tweets that contains words in a lexicon from *Hatebase.org*
- Twitter data set used for research
- Classified and labelled by a panel, which was provided the definitions and details for labelling

Categories of Speech

- Hate
- Offensive
- Neither

Categories of Speech

Hate

- Expresses hatred towards a **targeted (disadvantaged social) group / members of the group** or
- Intends to be **derogatory**, to **humiliate**, or to **insult** the members of the group

Offensive

- Includes **profanity**, **obscene** language
- Can cause people to **feel offense, upset**

Sources:

Automated Hate Speech Detection and the Problem of Offensive Language (https://arxiv.org/pdf/1703.04009v1.pdf)

Categories of Speech - Examples

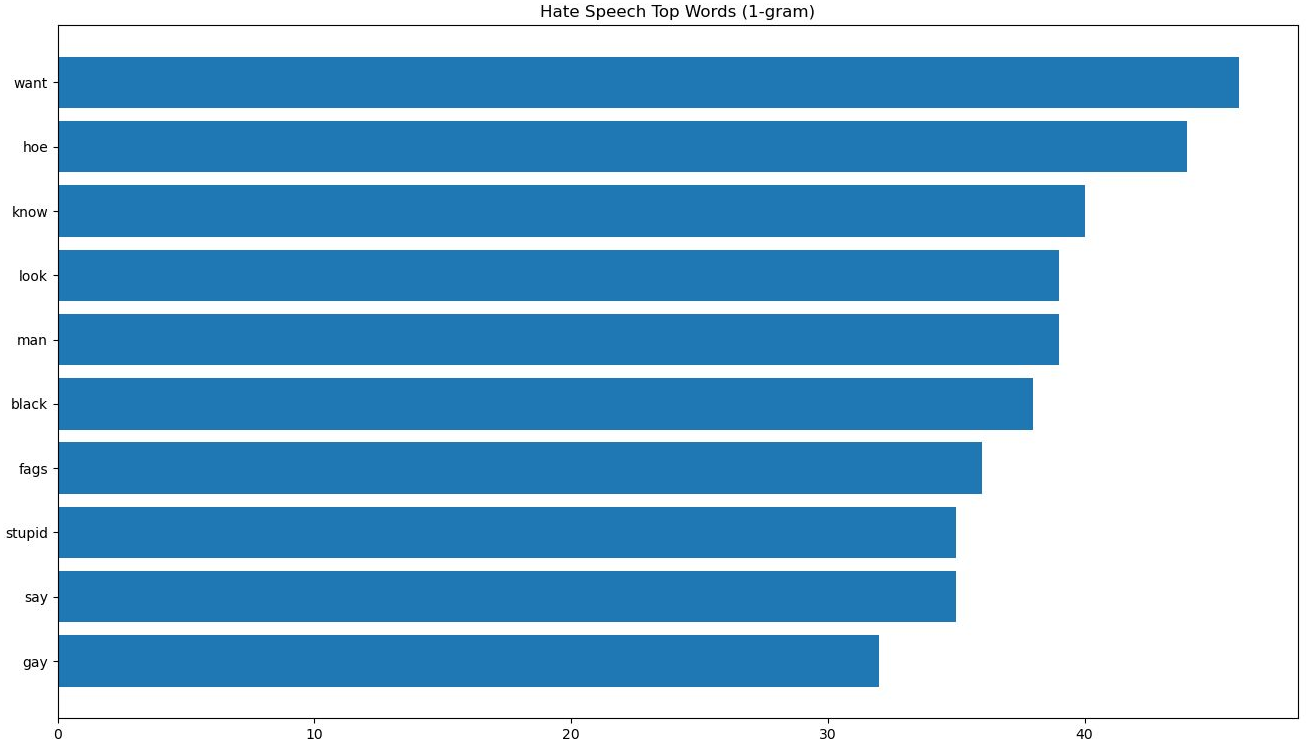
Hate	Offensive
• At least I'm not a n*g**r	• i spend my money how i want b**ch its my business

Sources:
Automated Hate Speech Detection and the Problem of Offensive Language (https://arxiv.org/pdf/1703.04009v1.pdf)

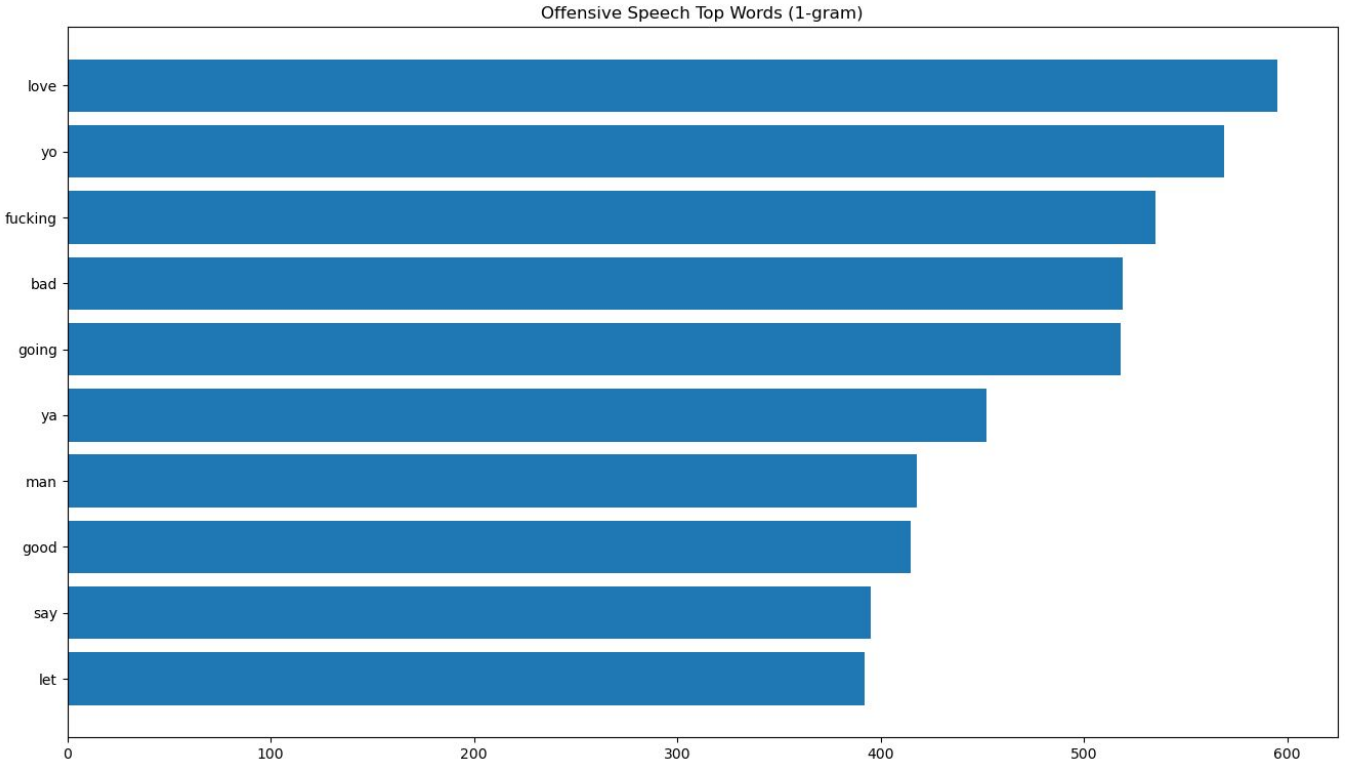
Data - Exploratory Data Analysis (EDA) and Cleaning

- Data Cleaning
  - Line or tab characters
  - Website urls
  - Hashtags and emojis
  - Tweet mentions
  - Retweet Substrings
  - Numbers
  - Punctuations

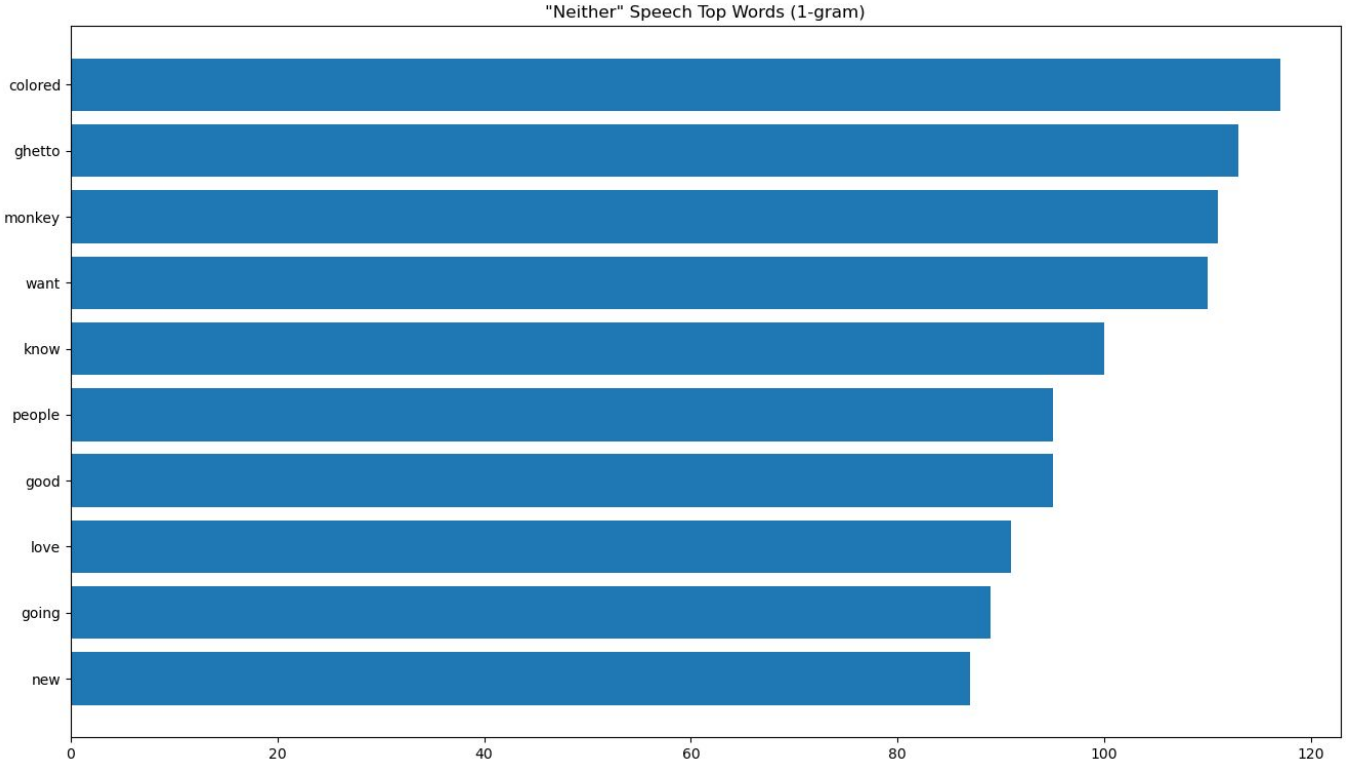
Data - Hate Speech Top Words (1-gram)



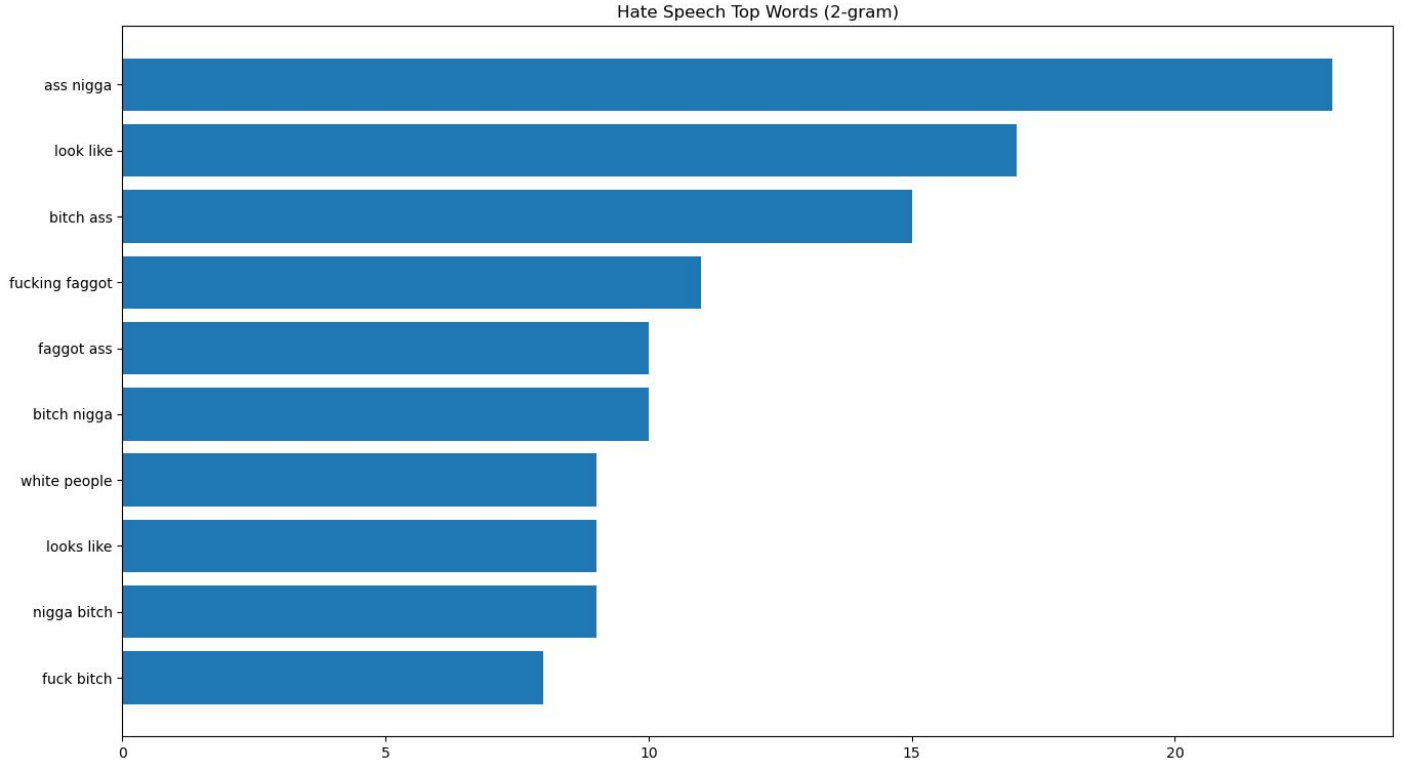
Data - Offensive Speech Top Words (1-gram)



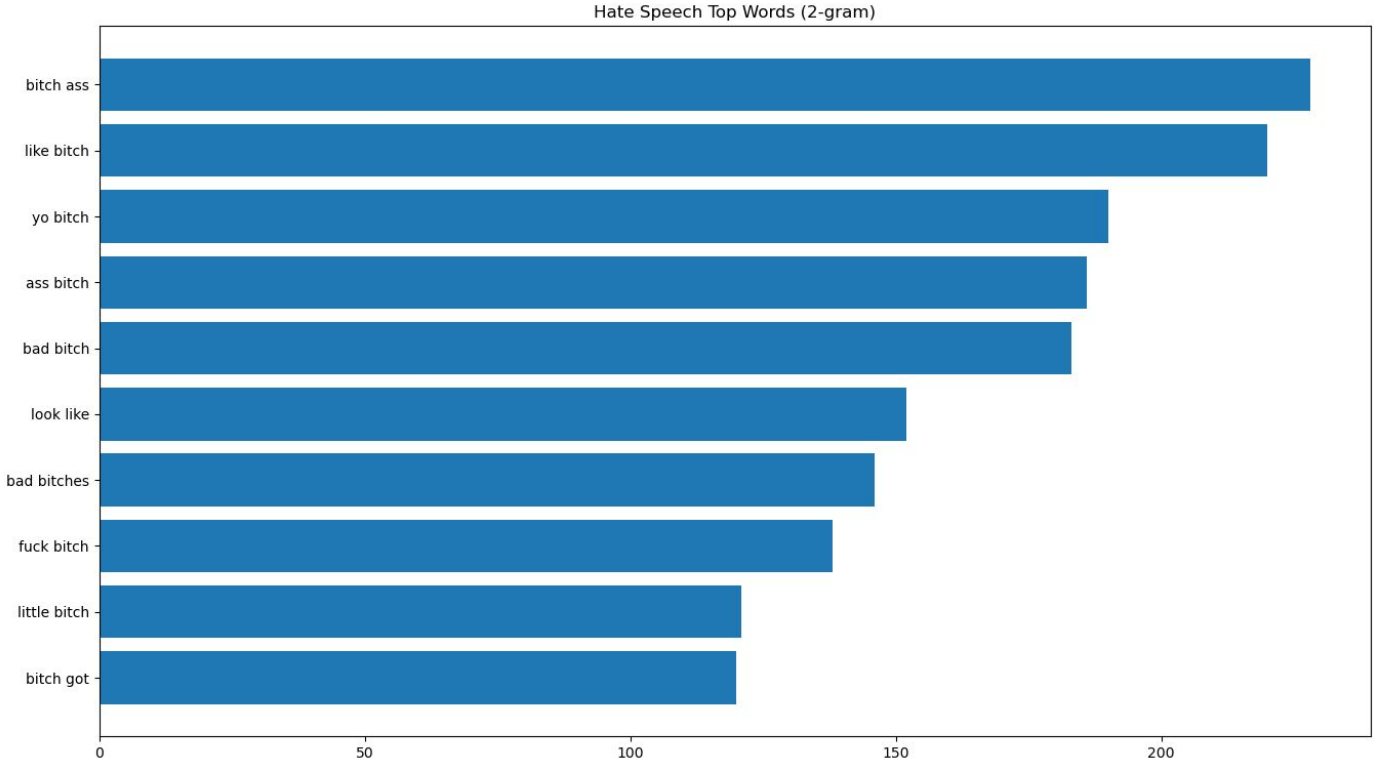
Data - “Neither” Speech Top Words (1-gram)



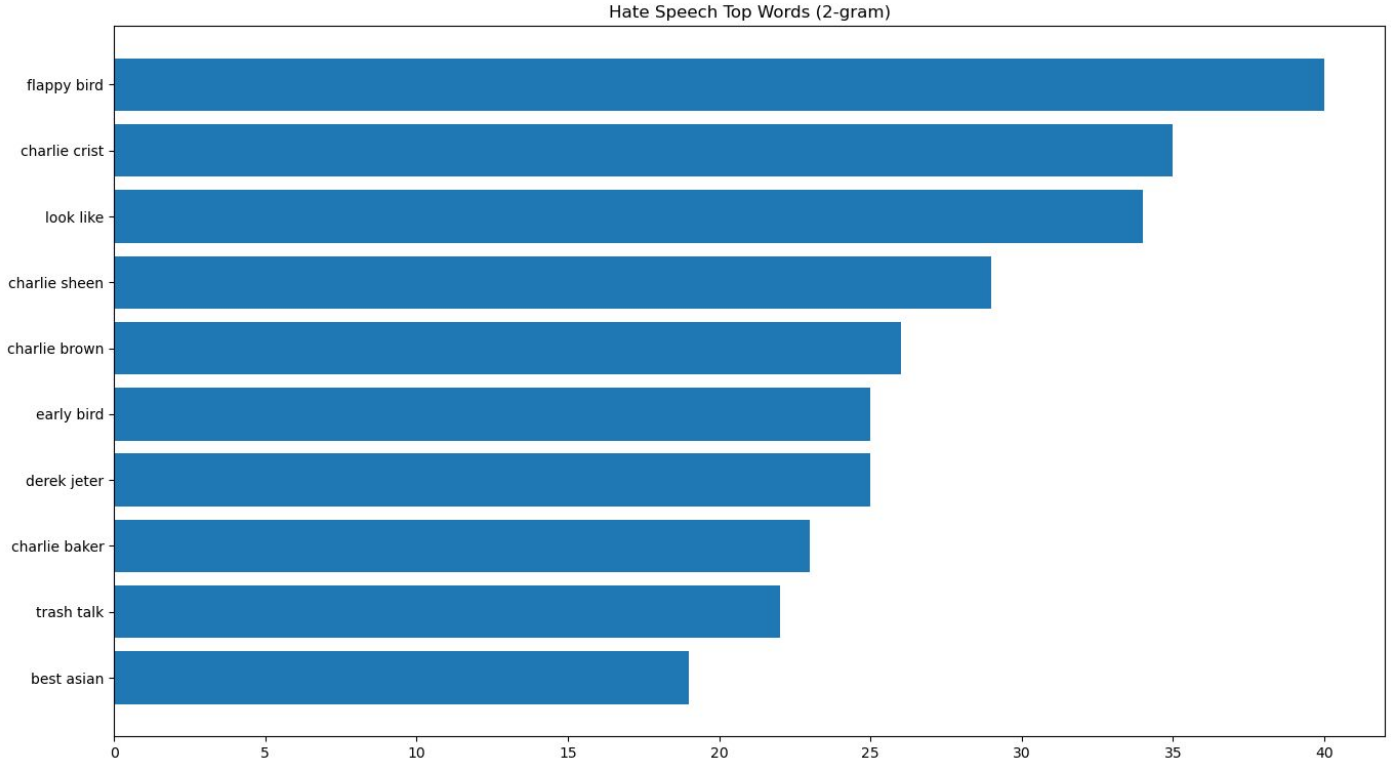
Data - Hate Speech Top Words (2-gram)



Data - Offensive Speech Top Words (2-gram)



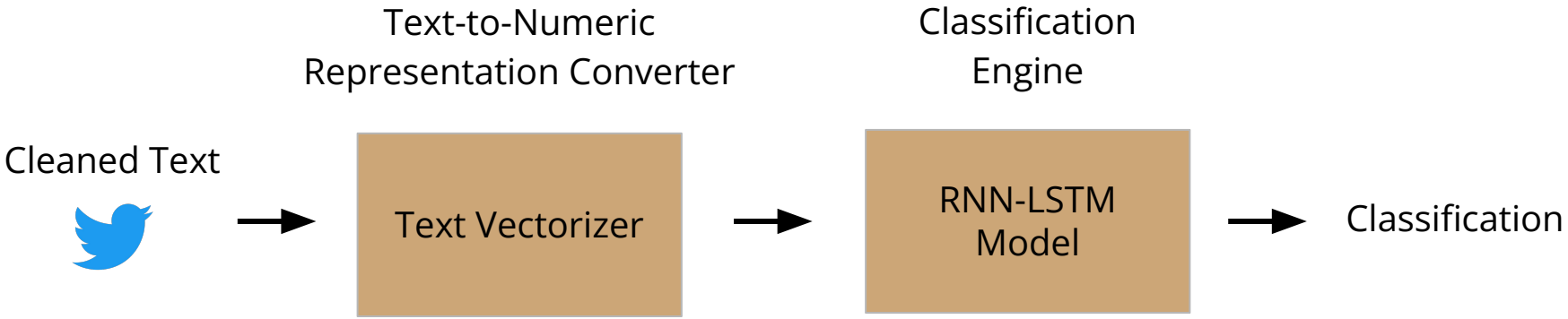
Data - “Neither” Speech Top Words (2-gram)



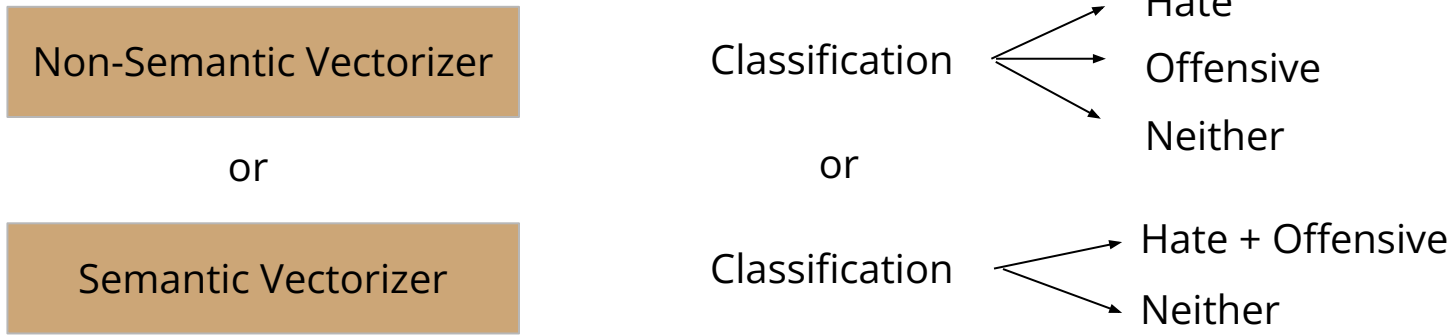
Top Words Trends

- Hate (Class 0): Racial Slurs, Derogatory Terms, Cuss Words
- Offensive (Class 1): Profanities
- Neither (Class 2): Non-Offensive/Hate

Overarching Approach Employed to Identify Hate Speech



Options Explored



Evaluation Metrics of Interest

Metric	**Description / Rationale**
Accuracy	% of predictions that are correct
F1-Score	Composite measure that considers both (a) “precision” and (b) “recall” (a) “precision”: % of predicted “positives” that are true *Relevant as a low precision score on detecting hate speech indicates that we are wrongly censoring non-hate speech that could be of interest to readers* (b) “recall”: % of actual “positives” that are predicted correctly *Relevant as we wish to identify as high a proportion of hate speech that exists*

Insights from Exploring Approaches to Identifying Hate Speech

- Semantic vectorizer offers better model performance
- Identifying hate speech separately is challenging due to class imbalance

Multiclass Classification with (a) Non-Semantic and (b) Semantic (GloVe) Text Vectorizers

"Train"	**Rebalancing**	
% Obs	Before	After
-	-	-
-	-	-
Hate	**6%**	~33% (Upsampled)
Offensive	77%	~33% (Downsampled)
Neither	17%	~33% (Retained)

"Test"	**Text Vectorizer**	
Evaluation	(a) Non-Semantic	**(b) Semantic**
Accuracy	0.86	**0.88**
F1-Scores	-	-
Hate	0.40	**0.40**
Offensive	0.91	**0.93**
Neither	0.82	**0.86**

Finalised Approach: A Broad-based Identifier of Hurtful Text

- Vectorizer: Semantic
- Classification: Binary [Hate + Offensive , Neither]

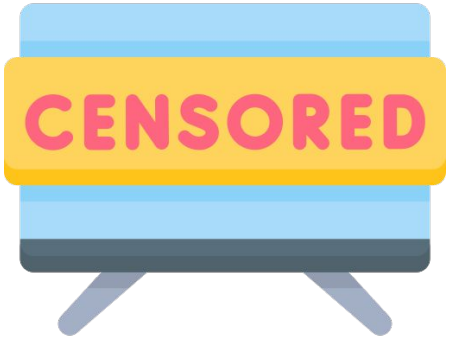
Multiclass Classification with Semantic (GloVe) Text Vectorizer & Binary Classification

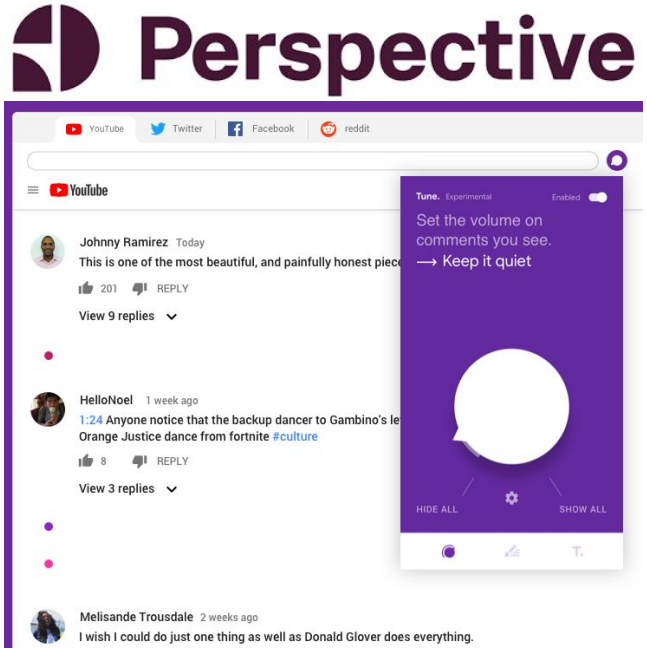
"Train"	**Rebalancing**	
% Obs	Before	After
-	-	-
-	-	-
Hate	**6%**	~17% (Retained)
Offensive	77%	~33% (Downsampled)
Neither	17%	~50% (Retained)

"Test"	**Text Vectorizer**
Evaluation	**Semantic**
Accuracy	**0.95**
F1-Scores	-
Hate + Offensive	**0.97**
Neither	**0.86**

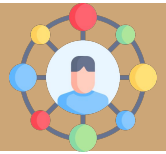
Hate Speech Detection Landscape

- Social medias have their own filters
  a. Site-wide moderation may be limited
- Browser Extension: Tune by Perspective API
  a. Sends posts to an external Machine Learning API
- Browser Extension: Shut Up
  a. Completely removes comment sections on platforms





Demo

	 Doe	 Ray	 Far
who	Tired **student**, **unwinds** by browsing **social media**	**Parent** who is **introducing** internet and social media to **children**	**Community leader / moderator** in a social media platform
pains	**Disturbed** by hateful / offensive posts	**Concerned** about exposure to unsafe and unsavory content	Too many posts to review
goal›	Highlight potentially harmful users to block		Flag out potentially harmful posts and users for moderation

We want to provide a middle ground solution to empower users and community leaders

Cost Benefit Analysis

1. Internet penetration of Singapore is at **96%**
2. An approximation of savvy users would be **37%** who use adblock globally
3. IPS 2021 indicated that **2 out of 10** SG residents are concerned about online hatred pertaining to race

Estimated 7.1% of 5.8M residents: **400k** people may adopt our solution

With **26%** prevalence of anxiety, **104k** people may benefit more from the mitigation against online hatred-induced anxiety

Cost Benefit Analysis

Benefit

- If **104k** people who benefitted may reduce cost of lost productivity **($16b)** and save ~5%, resulting in **~$1b** saved from productivity
- Other benefits not resulting from productivity

Cost

- Average cost of building a browser extension ~$20k
- Average cost of building machine learning model ~$80k
- Estimated total: **$100k**

Future Work

1. Processing of offensive images will require **contextualising images** with more sophisticated means
2. Generalise the model better by **covering other** social media **platforms**
3. Explore **localised dataset** to provide better coverage of SG media

Conclusion

1. Hate speech disrupts harmony and **negatively impacts mental health**
2. The **benefits** of mitigating online hate speech far **outweighs** the costs
3. Therefore, we offer to provide solution to **empower users** to **mitigate** the impact of online hate speech
4. Our **model predicts** hate / offensive texts **accurately** with good **f1-score**
5. Future work to **enhance** the solution for better **coverage** locally and across platforms

References

Hate Speech in the News

1. https://www.todayonline.com/world/tiktok-removes-singapore-born-comedians-controversial-video-citing-hate-speech-2189341
2. https://www.todayonline.com/singapore/govt-must-intervene-early-hate-speech-disrupts-racial-religious-harmony-spore-shanmugam
3. https://www.channelnewsasia.com/singapore/singapore-zero-tolerance-hate-speech-teo-chee-hean-2720241

References

Background

1. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
2. https://www.channelnewsasia.com/singapore/poor-mental-health-young-adults-seek-help-moh-survey-3802531
3. https://my.clevelandclinic.org/health/diseases/9536-anxiety-disorders
4. https://safeparty.ucdavis.edu/hate-public-health-issue
5. https://www.todayonline.com/singapore/depression-anxiety-lost-productivity-cost-singapore-billions-2159496

References

Background

1. https://repositories.lib.utexas.edu/server/api/core/bitstreams/5c14762d-f787-4b0c-8893-186128ce8af4/content
2. https://www.nature.com/articles/s41598-023-31146-1
3. https://www.coe.int/en/web/combating-hate-speech/what-is-hate-speech-and-why-is-it-a-problem-

References

Cost Benefit Analysis

1. https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2022&locations=SG&start=1990&view=chart
2. https://www.coe.int/en/web/combating-hate-speech/what-is-hate-speech-and-why-is-it-a-problem-
3. https://lkyspp.nus.edu.sg/docs/default-source/ips/ips-exchange-series-22.pdf
4.