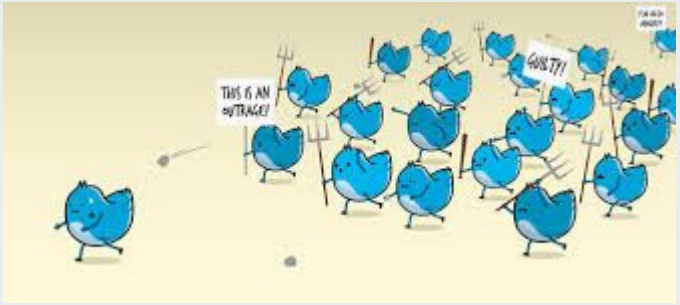


Hate Tweet Sentiment Analysis

Steven Yan
Ivan Zakharchuk

Agenda

- Background
- Business Proposal
- Data Sources
- Methodology
- Class Imbalance
- Modeling
- Analysis
- Next Steps
- Contact Information

Background

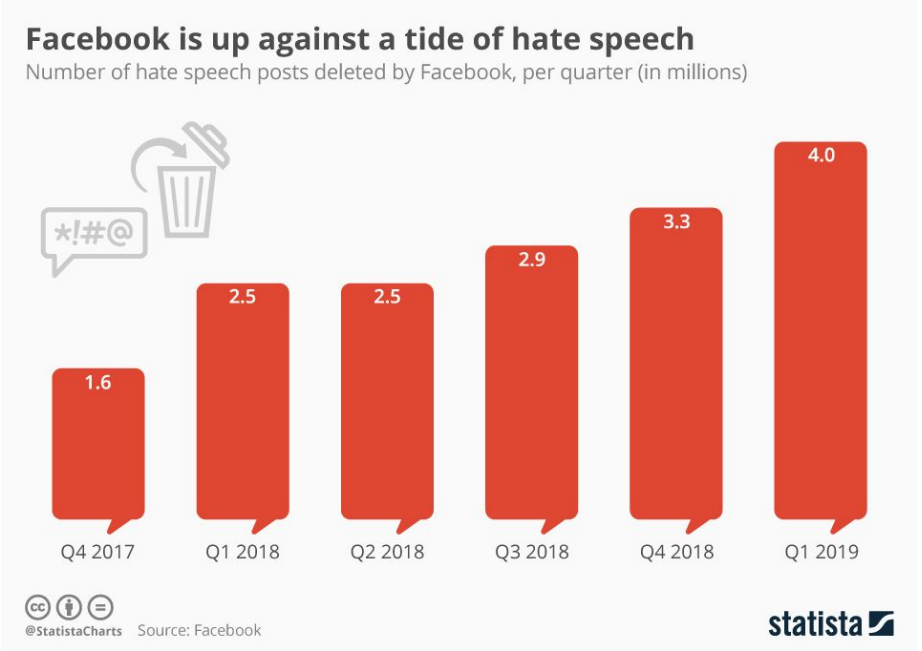
What is Hate versus Offensive Speech?

Hate Speech:

- any form of expression intending to vilify humiliate, or incite hatred against group orindividual on basis of race, religion, skin color, sexual or gender identity, ethnicity, disability, or national origin
- Key takeaway: incite violence and promote hatred

Forms of Content Moderation:

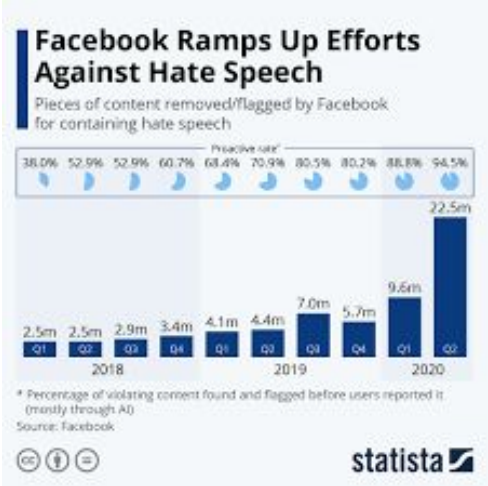
1) Human Moderation (Facebook)
2) Community Moderation (Reddit)
3) Algorithmic Moderation (Twitter, YouTube)
4) Multiple Tier (Algorithmic and Human or Community)

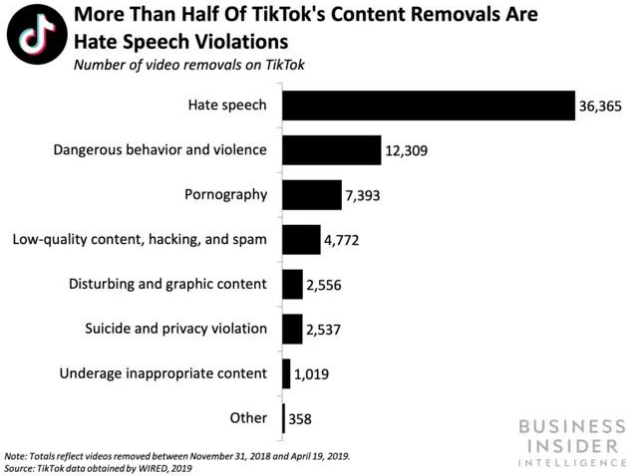


Business Proposal

Why should your company invest in an automated hate speech detection system?

- ***Community:*** Establish leadership presence in setting precedent in industry
- ***Business Partners:*** Negative press can affect willingness of businesses to engage or individuals from investing
- ***Employee:*** Promotes sense of stability and well-being among your employees
- ***Customer:*** Appeal to larger customer base in urban areas that lean more to the left





Data Sources

Analysis and Challenges

Cornell University Study:

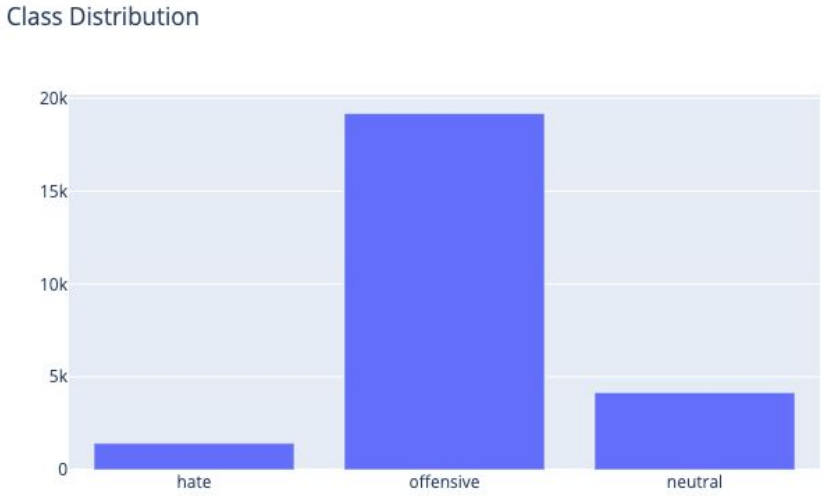
- Dataset biased towards racism and homophobic tweets
- Easier to label, misogynistic terms more ambiguous, human classifiers more biased against misogyny
- Huge class imbalance:

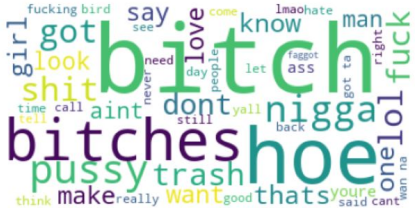
Aristotle University Study:

- Dataset included Sexism and Racism labeled tweets
- Acquire more hate-labelled tweets through API using Tweet ID lookup from datasets

Challenge:

- Differentiating between hate versus offensive tweets

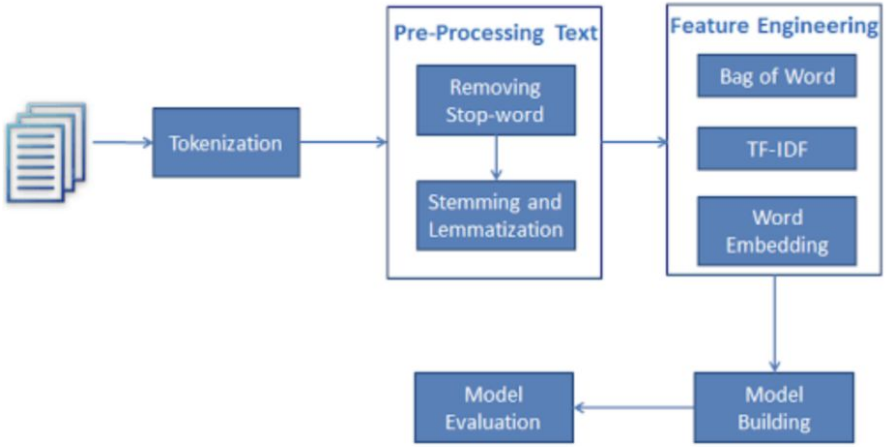




Methodology

With NLP, computers are taught to understand human language, its meaning and sentiments.

In order to translate complex natural human language into systematic constructed features, we need to follow some major steps shown on the next graph.



Class Imbalance

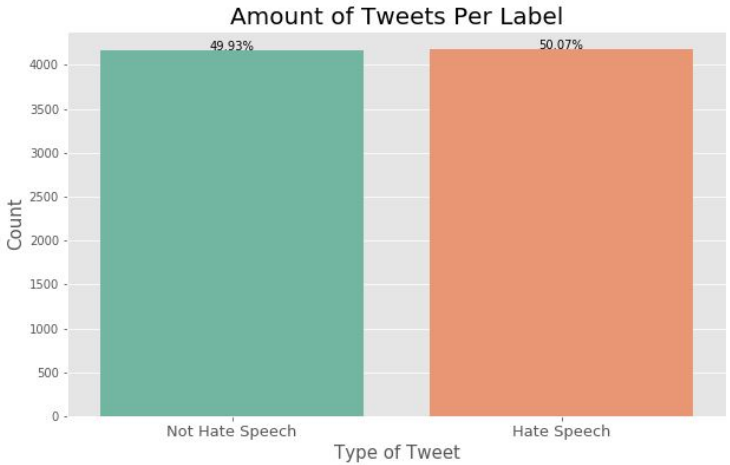
During EDA, we discovered that data from Cornell University appears to be unbalanced with minority class as hate speech and represented on the top graph.

Undersampling Methods:

- RandomUnderSampler
- ClosestNearestNeighbours
- SMOTE-ENN (Oversampling and Undersampling)

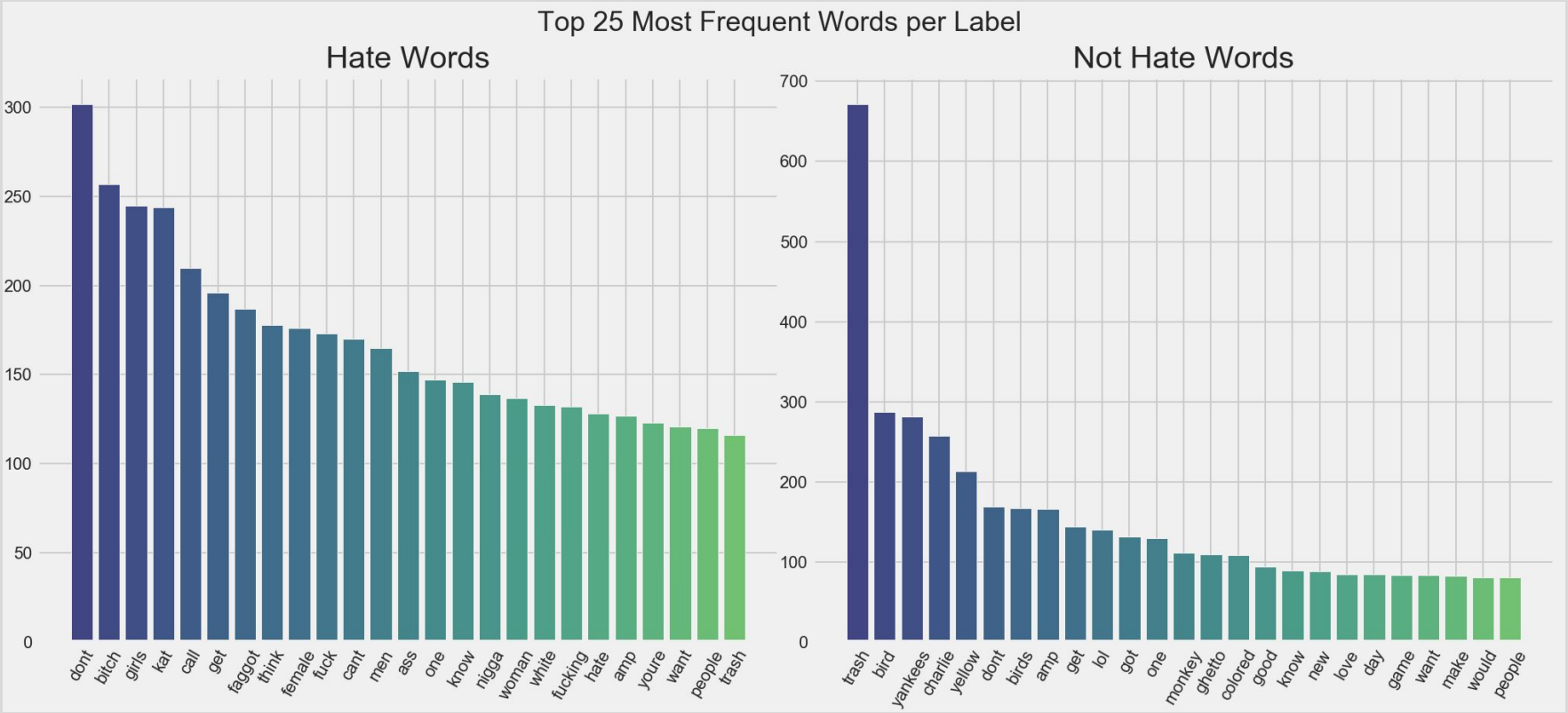
With API requests using labeled as hate speech tweets ids we were able to bring more data to our project and balance it. Bottom graph shows balanced data.

	F1 Score	Recall	Precision	PR AUC
Logistic Regression	0.1777	0.1080	0.5000	0.3512
Decision Tree	0.2795	0.2493	0.3180	0.1232
Random Forest	0.1609	0.0969	0.4729	0.3006
Random Forest RUS	0.344	0.7423	0.2238	0.3234
Random Forest CNN	0.1655	0.825	0.0920	0.2266
Random Forest SMOTE-ENN	0.2867	0.4044	0.3792	0.2963



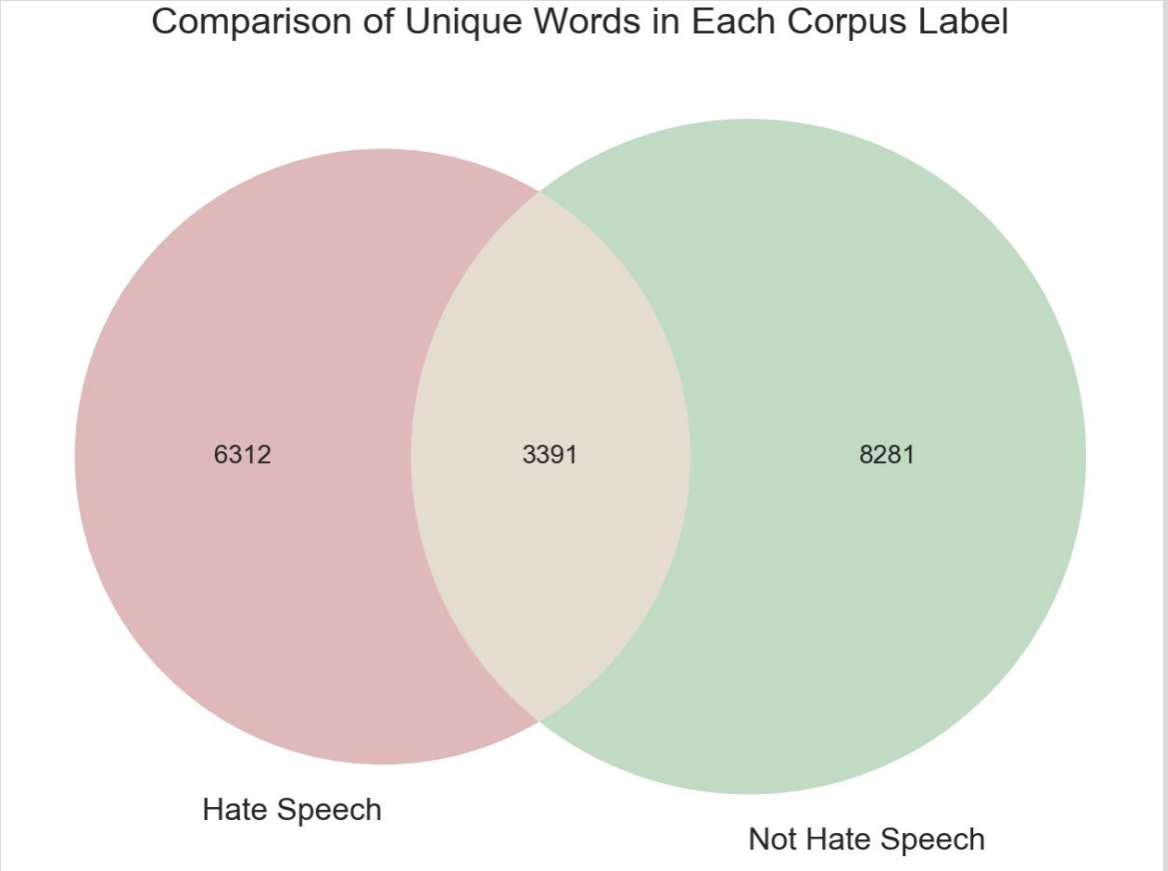
Analysis

Frequency distribution of words within the whole corpus.



Analysis

With further analysis we were able to find out and create vocabulary of only words that belong to tweets labeled as hate speech. We found 6312 words that exclusively belong to tweets labeled as hate speech. Majority of hate speech words are racist, sexist and homophobic slurs that exceed cultural slang.

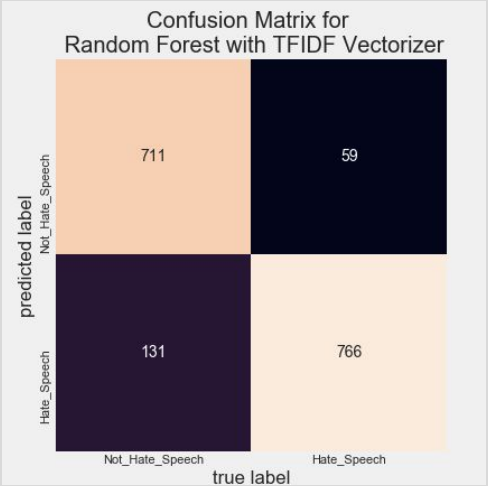


Modeling

Model	Precision	Recall	F1-Score
Random Forest	0.85	0.92	0.89
Logistic Regression	0.9	0.89	0.88
Naive Bayes	0.87	0.92	0.89

Random Forest with Hyper Parameters selected with GridSearch let us create final model with following results on testing data:

Precision: 0.7124
Recall: 0.937
Testing Accuracy: 0.7816
F1 Score: 0.8094



Next Steps

What are our future goals with the project?

- **Multi-class Classification**: Developing a model that can differentiate between the different nuances based on business needs
- **Neural Network Integration**: Allow for use of embeddings and better semantic understanding of the text analyzed
- **Preparing for Deployment**: Testing on unseen data to demonstrate generalizability of developed model

Contact Us

Any questions or concerns?
We can be reached via email, LinkedIn or Github.

Want more information?
Check out our Github repo or contact us directly.

Steven Yan

- Email: stevenyan@uchicago.edu
- Linkedin: www.linkedin.com/in/examsherpa
- Github: www.github.com/examsherpa

Github Repo:
https://www.github.com/examsherpa/Twitter-Sentiment-Analysis

Ivan Zakharchuk

- Email: ivan.zakharchuk@gmail.com
- Linkedin: https://www.linkedin.com/in/ivan-zakharchuk
- Github: https://github.com/vanitoz