ENHANCING THE DETECTION OF POTENTIAL THREATS USING MACHINE LEARNING

MDATAGOV SYMPOSIUM 2019

FATIMA CHIROMA

CONTENTS

Background

Research

Placement

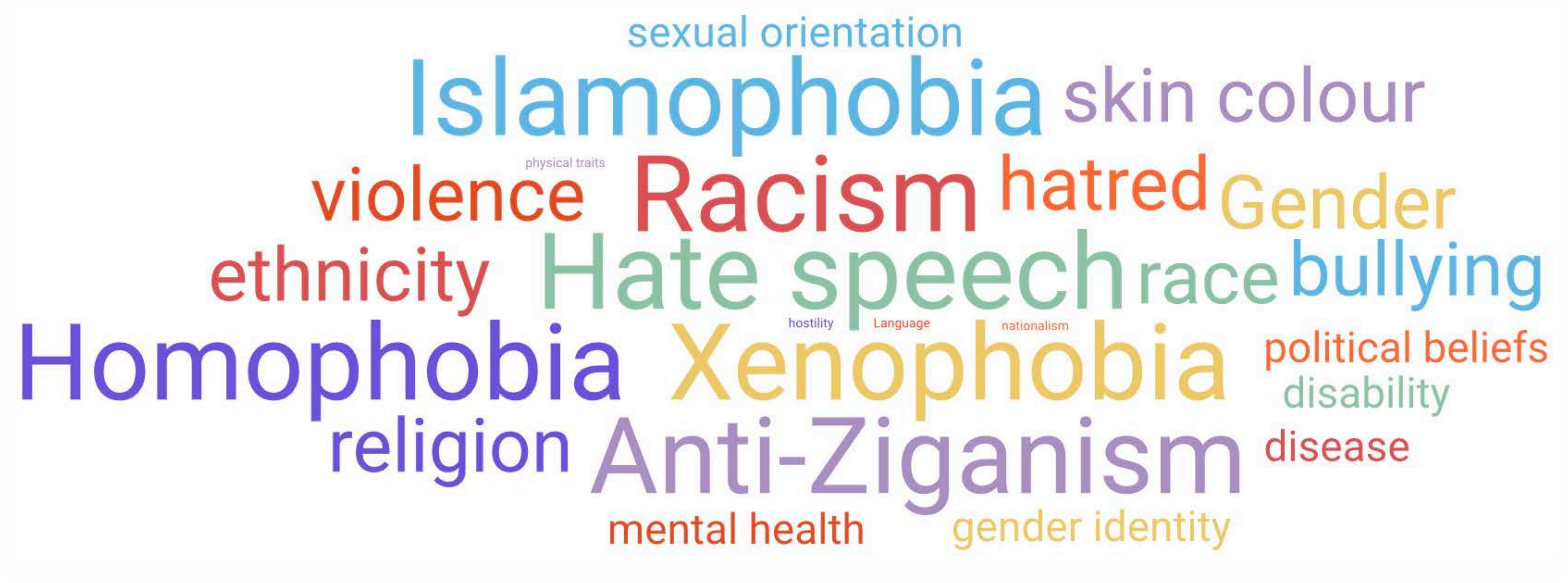
Conclusion

Acknowledgement

BACKGROUND

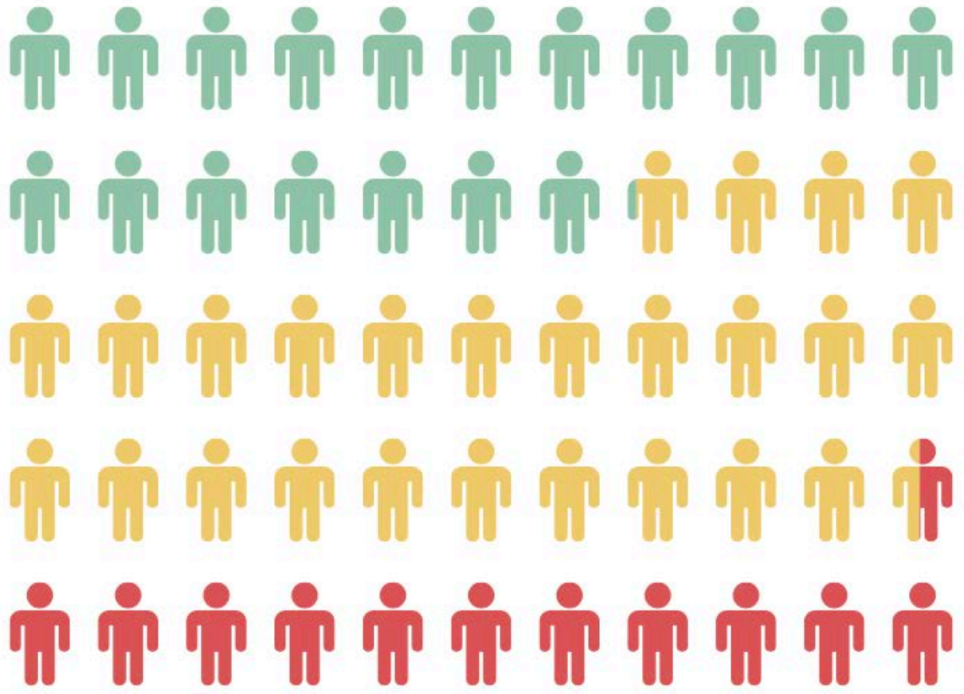
- Rapid increase in social media usage
- 2.47 Billion active social media users [1]
- Abundant user generated data
- Association between social media and suicidal behavior [2]

CYBERHATE


sexual orientation
Islamophobia
skin colour
physical traits
violence
Racism
hatred
Gender
ethnicity
Hate speech
race
bullying
hostility
Language
nationalism
Homophobia
Xenophobia
political beliefs
disability
religion
Anti-Ziganism
disease
mental health
gender identity

HATE MATERIALS

15 – 18 Years



67%

Exposed to hate materials[3]

21%

Victims

MISOGYNY

- Hate speech that is targeted towards women
- Online abuse has been linked to offline domestic violence against women

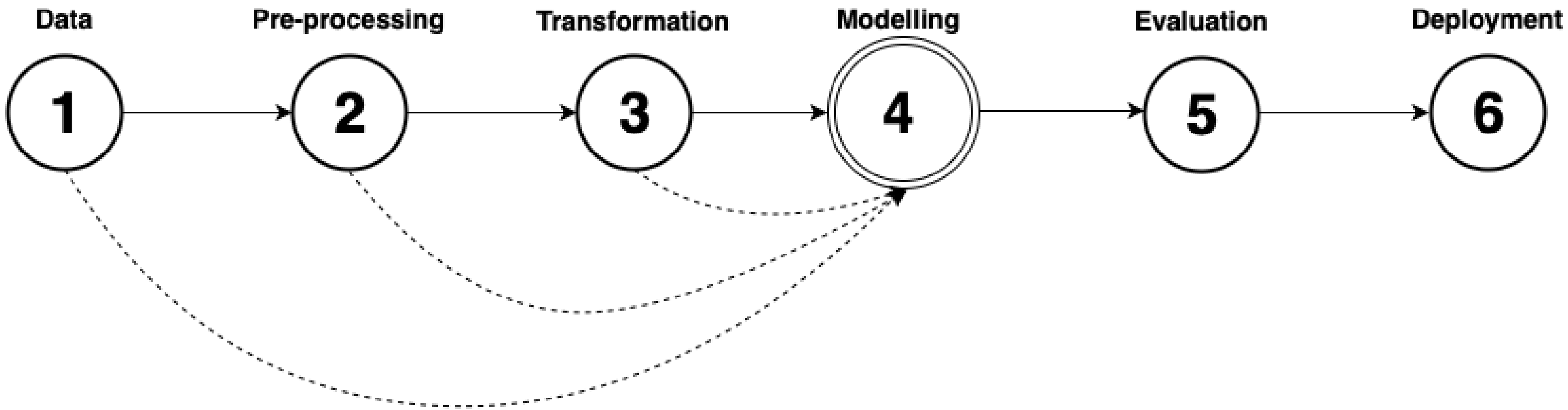
United Kingdom

48%

Experienced online abuse or harassment after leaving an abusive relationship[4].

RESEARCH

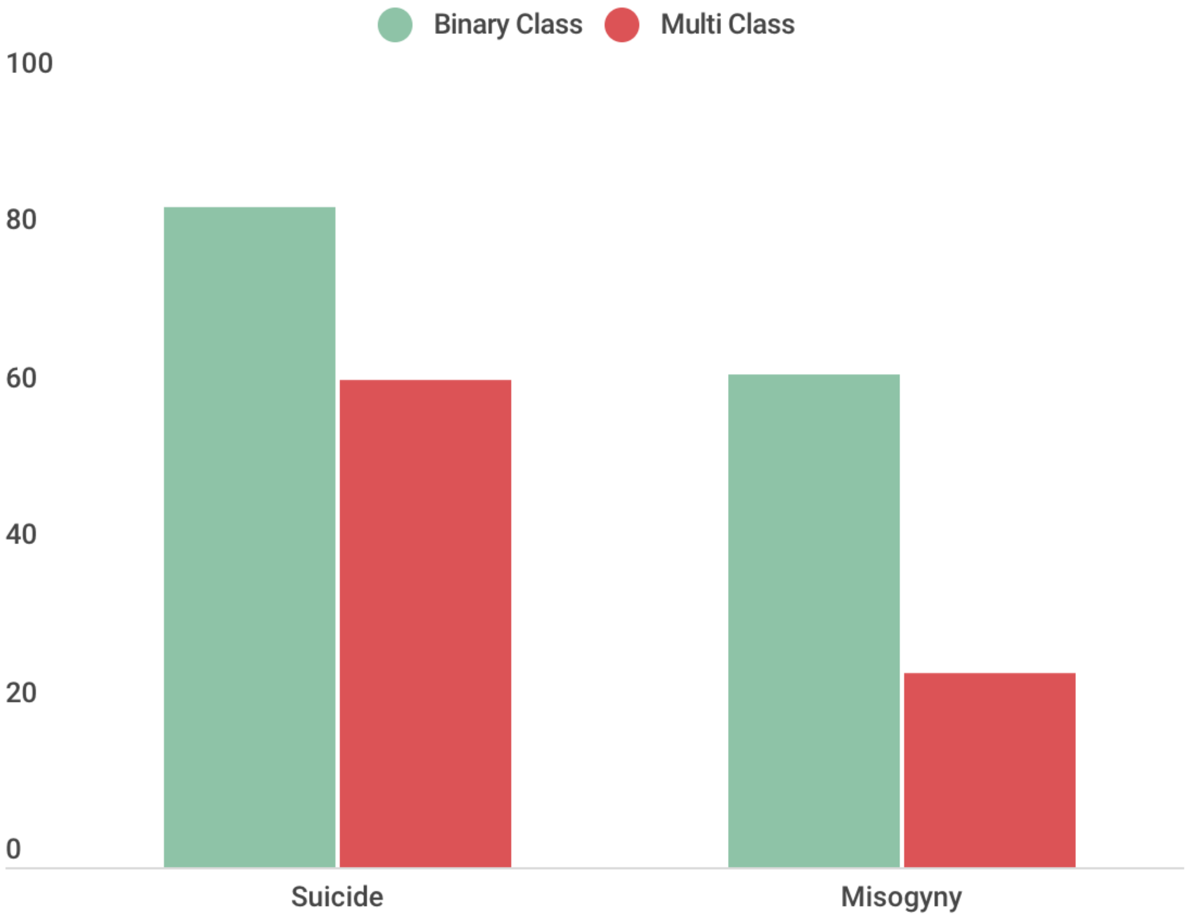
- How can machine learning be used for possible intervention?
- Improve the identification and classification of text containing worrying languages.
- Case studies: suicide related and misogynistic tweets (independently)



APPLICATION

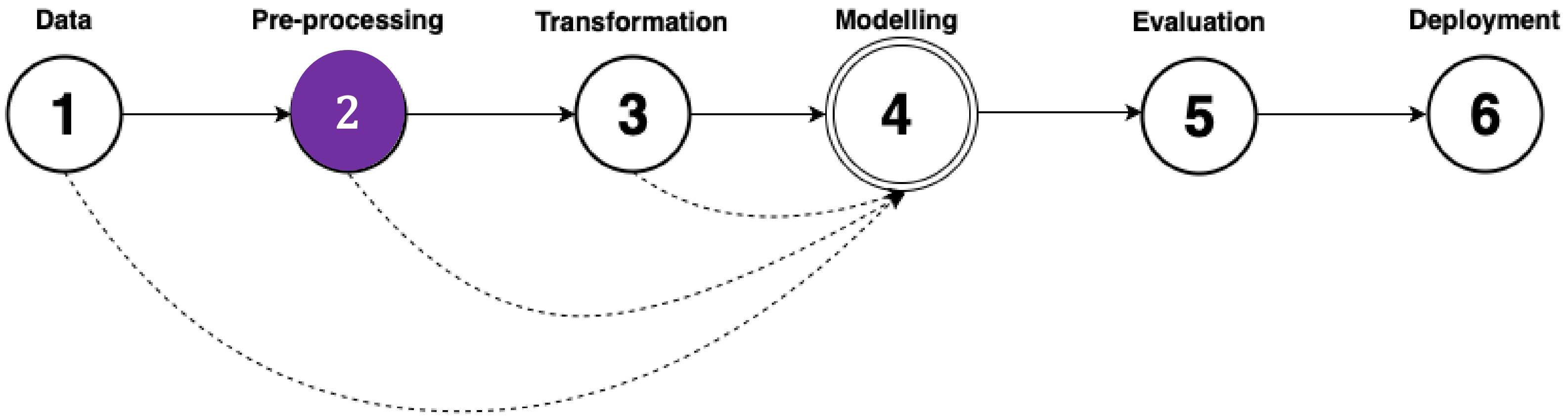
- Labelled tweets for suicide related communications and misogyny
- Measured the performance of existing machine learning approaches and techniques
- Impact of data manipulation
- Ensemble learning, rule-based learning
- Dimensionality reduction techniques

RESULT



PLACEMENT

- PyGrams: Python based app for extracting popular terms from large documents.
- Implement automated stop words
- 176 in-built English stop words e.g. I, any, haven't

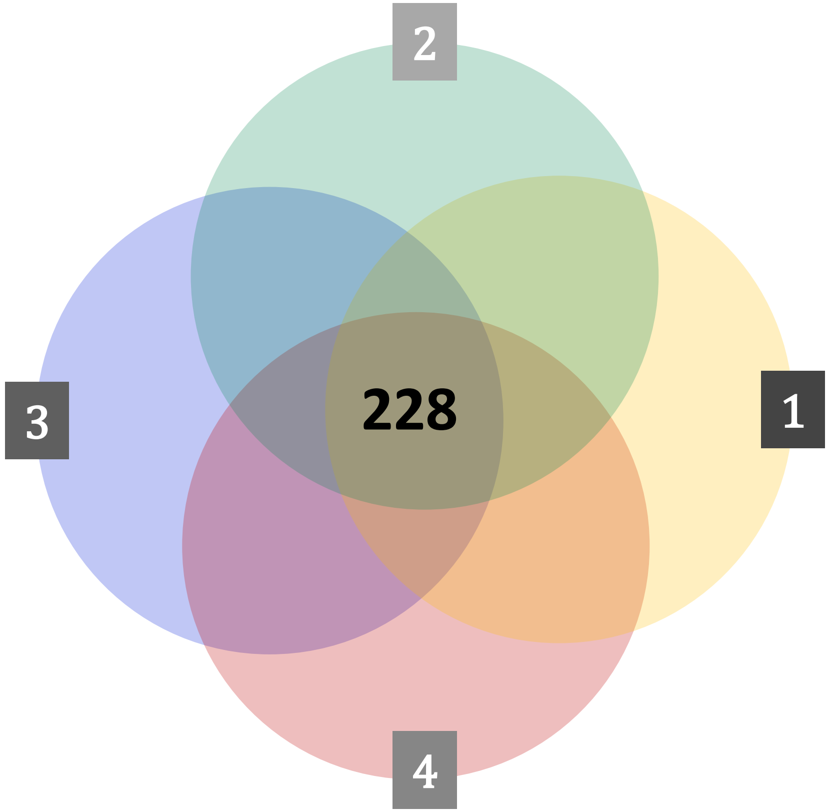


APPROACH

Sample Data: 3,526 words

Existing Methods	1	2	3	4
Entropy	🟨		🟦	
Term Frequency (TF)		🟩	🟦	🟥
TF1 (Singleton words)		🟩		
Inverse Document Frequency (IDF)		🟩		
TF/IDF	🟨			
Variance			🟦	🟥
Total Stop words	260	2613	319	500

Approach



RESULT

In-built		Auto-generated		Stop Words
176	+	228	=	404

NEXT

Placement

- In-built and auto-generated stop words
- Deploy auto-generated stop words

Research

- Auto-generated stop-words
- Rule-based learning
- Target and/or minority classes
- Domain specific

THANK YOU

ACKNOWLEDGEMENT

Office for National Statistics

Data Science Campus

University of Portsmouth

Petroleum Technology Development Fund

PTDF

Department of Health

Cardiff University

Prifysgol Caerdydd

REFERENCES

1. eMarketer. (n.d.). Number of social media users worldwide from 2010 to 2021 (in billions). In *Statista - The Statistics Portal*. Retrieved March 19, 2019, from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
2. Sueki, H. (2015). The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of affective disorders*, *170*, 155-160.Target and/or minority classes
3. Perry, B., & Olsson, P. (2009). Cyberhate: the globalization of hate. *Information & Communications Technology Law*, *18*(2), 185-199.
4. Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284-297.
5. https://www.freepik.com/free-vector/social-media-icons-globe_887251.htm
6. https://datasciencecampus.ons.gov.uk
7. https://ptdf.gov.ng
8. https://www.ons.gov.uk
9. https://www.port.ac.uk
10. https://www.cardiff.ac.uk
11. https://data.london.gov.uk/publisher/department-of-health