VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

UNIVERSITY OF OXFORD

TOWARDS SYNTHETIC SOCIAL MEDIA DATA

AUTHORS:

Justina Mandravickaitė
justina.mandravickaite@vdu.lt

Milita Songailaitė
milita.songailaite@stud.vdu.lt

Veronika Gvozdovaitė
gvozdovaite.veronika@gmail.com

Danguolė Kalinauskaitė
danguole.kalinauskaite@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

INTRODUCTION

- Social network data (texts and network structure) are in high demand.
- Data protection regulations hamper the collection of the above-mentioned kind of data.
- As a result, interest in synthetic data is on the rise.
- Language models can be used for synthetic data generation.
- We propose a method for synthetic social media data generation.

GOAL

- The aim of our project was to create a prototype for a synthetic social media data generator.
- Our prototype, **Fabulator**, combines the use of graph structures and text generation to produce synthetic data to overcome the shortage of necessary data.

SYSTEM DESCRIPTION

1 Text generation

For text generation, DialoGPT-medium dialogue response generation model was used.

Two dialogue response generation models were created using relevant Reddit dialogues data: “Political” and “Conspiratorial” models.

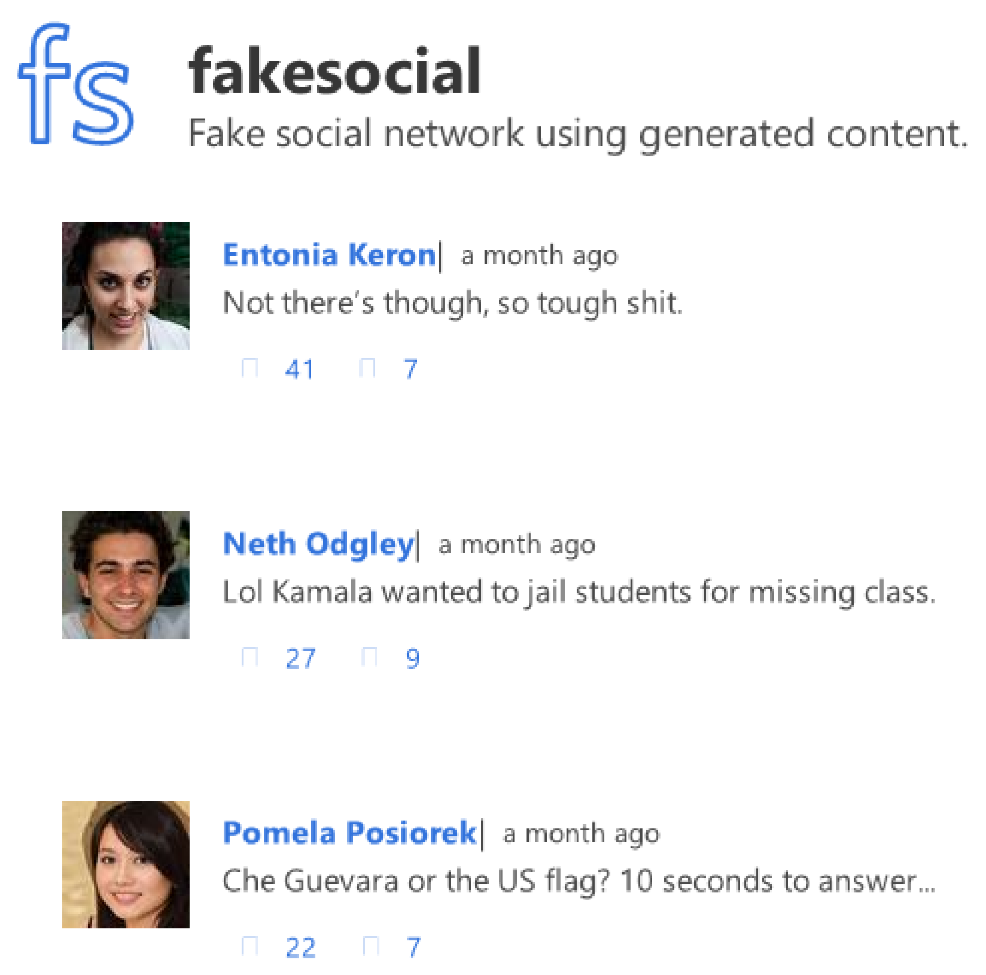
Each model consisted of two separate models pretrained to generate responses from the opposite point of conversation view.

2 Network structure

A simple social network of fake users who have connections, can make posts, comment and like was generated using ***fakesocial*** social network generator:

- *StyleGAN* neural network was used to generate people’s profile pictures.
- *Markov chain* generator was used to generate user profile names, locations and job titles.
- The generated profile data was randomly sampled to create the fake user profiles.

3 Social media site



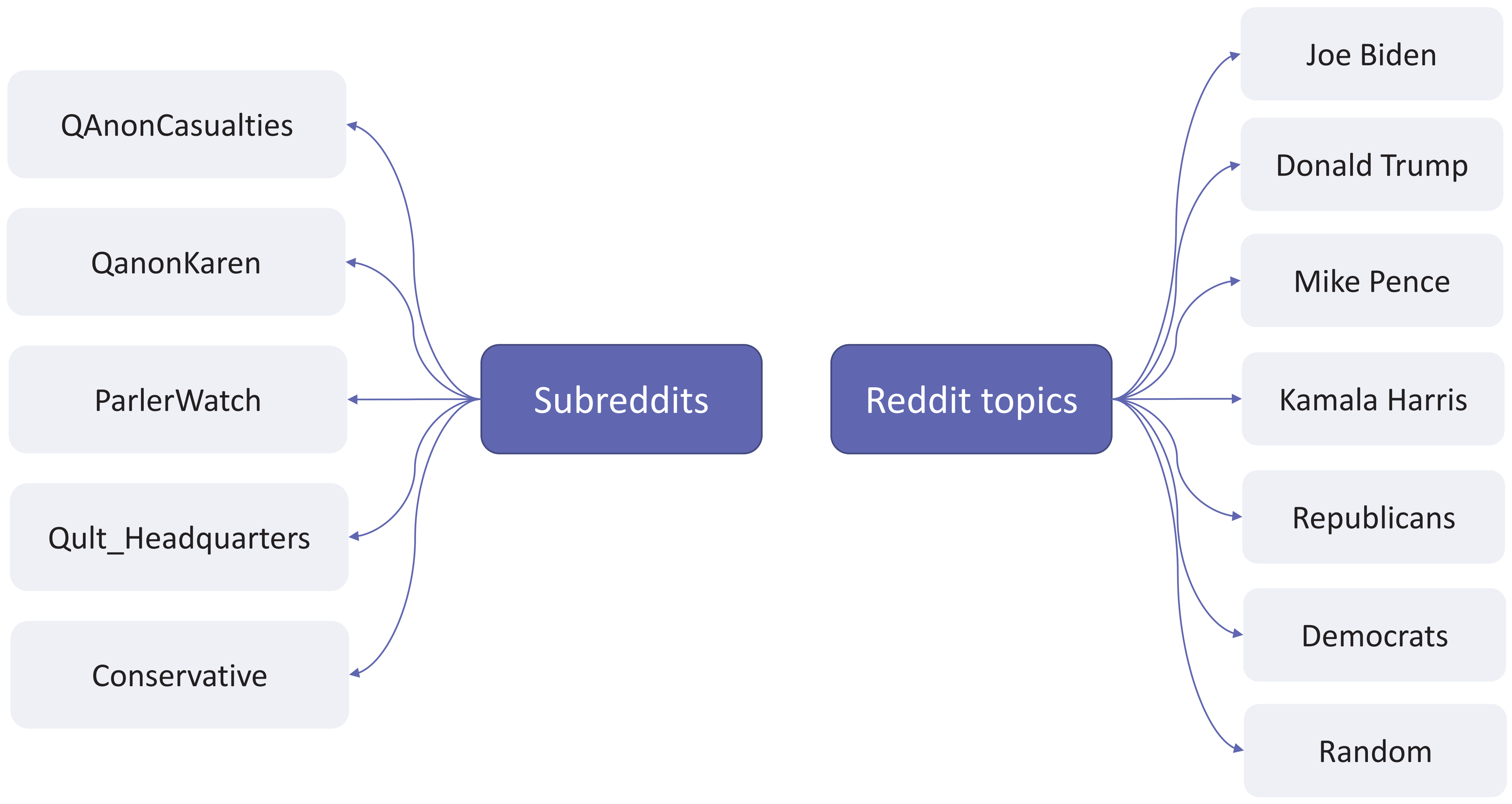
fakesocial
Fake social network using generated content.

Entonia Keron| a month ago
Not there's though, so tough shit.
41 7

Neth Odgley| a month ago
Lol Kamala wanted to jail students for missing class.
27 9

Pomela Posiorek| a month ago
Che Guevara or the US flag? 10 seconds to answer...
22 7

TOPICS



Subreddits: QAnonCasualties, QanonKaren, ParlerWatch, Qult_Headquarters, Conservative

Reddit topics: Joe Biden, Donald Trump, Mike Pence, Kamala Harris, Republicans, Democrats, Random

EVALUATION

Three metrics were chosen for the evaluation of the generated texts – **BLEU**, **ROUGE** and **perplexity**. The latter was used to evaluate the language model.

Classes of text generation models	BLEU	ROUGE
Republican	0.02	0.87
Qanon	0.01	0.65
Democrat	0.03	0.48
Conservative	0.02	0.32
Mean	**0.02**	**0.58**

Classes of text generation models	Perplexity
Republican	11.37
Qanon	11.95
Democrat	11.20
Conservative	12.01
Mean	**11.63**

CONCLUSIONS

- Our models reached 0.58 mean ROUGE value and 11.63 mean perplexity score, indicating good quality.
- We got a low mean BLEU score which indicates that generated texts differ in their structure, though they are grammatically correct and meaningful.

FUTURE PLANS

- Training more themed dialogue response generation models.
- Using different models for better text generation results.
- Focusing on other social media platforms text generation.
- Simulating various scenarios of cyber or propaganda attacks through social media texts.

CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT