Utilizing social media to investigate perspectives on the COVID-19 pandemic using natural language processing: a comparative study of four countries

Mengke Hu1 PhD, Mike Conway2 PhD
Department of Biomedical Informatics, University of Utah
mengke.hu@gmail.com, mike.conway@utah.edu

Introduction

- In Dec. 2019, several cases of pneumonia reported (Wuhan city)
- A new coronavirus—COVID-19 (defined by WHO)
- COVID-19 a pandemic (WHO) on March 11, 2020
- 85,000,000 confirmed cases and 1,860,000 deaths (Jan 05, 2021)

Reddit Data

- Rich autobiographical storytelling and advice seeking
- 40,000 characters per post
- 18% of the adults in US used Reddit (23% Men, 12% women), 2021
- 430M active users by May 2020
- 1.2M subreddits
- Over 70% users from English-speaking countries
- Text-based data well suited for analysis using Natural Language Processing (NLP)

Methods

- Reddit Data: 6 most popular COVID-related subreddits for US, UK, Canada and Australia (Feb. 2020–Oct. 2020)
- Topic modeling is a machine learning method for the automatic identification of topics (themes) in a corpus of text
- Train different Topic Models (Latent Dirichlet Allocation) for each country
- Manually annotate each topic from LDA output with a common topic

Results

- The volume of posting declined consistently across all four countries
- During lockdown events, the volume of posts peaked
- The UK and Australian subreddits contained much more policy discussion than US and Canadian subreddits

Reddit Data

- Rich autobiographical storytelling and advice seeking
- 40,000 characters per post
- 18% of the adults in US used Reddit (23% Men, 12% women), 2021
- 430M active users by May 2020
- 1.2M subreddits
- Over 70% users from English-speaking countries
- Text-based data well suited for analysis using Natural Language Processing (NLP)

Methods

- Reddit Data: 6 most popular COVID-related subreddits for US, UK, Canada and Australia (Feb. 2020–Oct. 2020)
- Topic modeling is a machine learning method for the automatic identification of topics (themes) in a corpus of text
- Train different Topic Models (Latent Dirichlet Allocation) for each country
- Manually annotate each topic from LDA output with a common topic

Results

- The volume of posting declined consistently across all four countries
- During lockdown events, the volume of posts peaked
- The UK and Australian subreddits contained much more policy discussion than US and Canadian subreddits

Lockdowns 2020 in US, UK, Canada and Australia

March 11 UK lockdown; US announced level 3 travel advisory
March 18 US and Canada suspend non-essential travel between 2 countries
March 23 UK lockdown
March 24 Australia bans all overseas travel
April 18 US protest the country’s lockdown
June 24 US 26 states case raise since easing lockdown restriction
July 3 UK announced an end to travel restriction except US
July 4 Melbourne, Australia tighten restriction on 12 suburbs
Sept. 5 Australia extends hard lockdown till end of Sept.
Oct. 12 UK announced new lockdown rules

Common Topics | Topic Summary | Example Posts
--- | --- | ---
Case Report | case report, report of interaction with hospital | "graph of daily local cases..."
Prevention | mask wearing, hand washing, transmission risk | "Masks have always been effective, don’t let anyone gaslight you..."
COVID Impact | work/finance, education, travel restriction, social distancing | "Shops such as Bunnings and Dan Murphy open after tomorrow?..."
Politics & News | policy announcement, news question and answer | "Daily Discussion Post...
"COVID-19 Survey..."
"ABC news..."