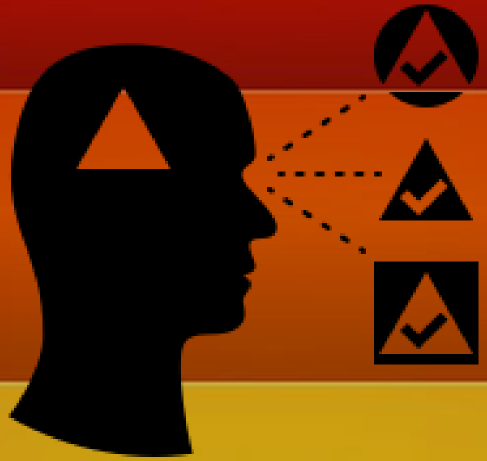


A Tool to Visualize Language Biases Over Time Using Reddit Comment Data

Filippo M. Libardi & Otto Mättas

Related Work

- Very recent 2020 paper;
- NLP model to identify biased terms towards a list of concepts given a text corpus;
- Method validated on Google News;
- The model only returns a list of biased words, the usage of these is not studied at all;
- Their method allows to formalize language bias;
- We aimed at improving on the method by displaying the usage of each biased word trough time.

Discovering and Categorising Language Biases in Reddit*

Xavier Ferrer⁺, Tom van Nuenen⁺, Jose M. Such⁺ and Natalia Criado⁺
⁺ Department of Informatics, King's College London
{xavier.ferrer_aran, tom.van_nuenen, jose.such, natalia.criado}@kcl.ac.uk

Abstract

We present a data-driven approach using word embeddings to discover and categorise language biases on the discussion platform Reddit. As spaces for isolated user communities, platforms such as Reddit are increasingly connected to issues of racism, sexism and other forms of discrimination. Hence, there is a need to monitor the language of these groups. One of the most promising AI approaches to trace linguistic biases in large textual datasets involves word embeddings, which transform text into high-dimensional dense vectors and capture semantic relations between words. Yet, previous studies require predefined sets of potential biases to study, e.g., whether gender is more or less associated with particular types of jobs. This makes these approaches unfit to deal with smaller and community-centric datasets such as those on Reddit, which contain smaller vocabularies and slang, as well as biases that may be particular to that community. This paper proposes a data-driven approach to automatically discover language biases encoded in the vocabulary of online discourse communities on Reddit. In our approach, protected attributes are connected to evaluative words found in the data, which are then categorised through a semantic analysis system. We verify the effectiveness of our method by comparing the biases we discover in the Google News dataset with those found in previous literature. We then successfully discover gender bias, religion bias, and ethnic bias in different Reddit communities. We conclude by discussing potential application scenarios and limitations of this data-driven bias discovery method.

1 Introduction

This paper proposes a general and data-driven approach to discovering linguistic biases towards protected attributes, such as gender, in online communities. Through the use of word embeddings and the ranking and clustering of biased words, we discover and categorise biases in several English-speaking communities on Reddit, using these communities' own forms of expression.

Reddit is a web platform for social news aggregation, web content rating, and discussion. It serves as a platform for multiple, linked topical discussion forums, as well as a network for shared identity-making (Papacharissi 2015). Members can submit content such as text posts, pictures, or direct links, which is organised in distinct message boards curated by interest communities. These 'subreddits' are distinct message boards curated around particular topics, such as /r/pics for sharing pictures or /r/funny for posting jokes[1]. Contributions are submitted to one specific subreddit, where they are aggregated with others.

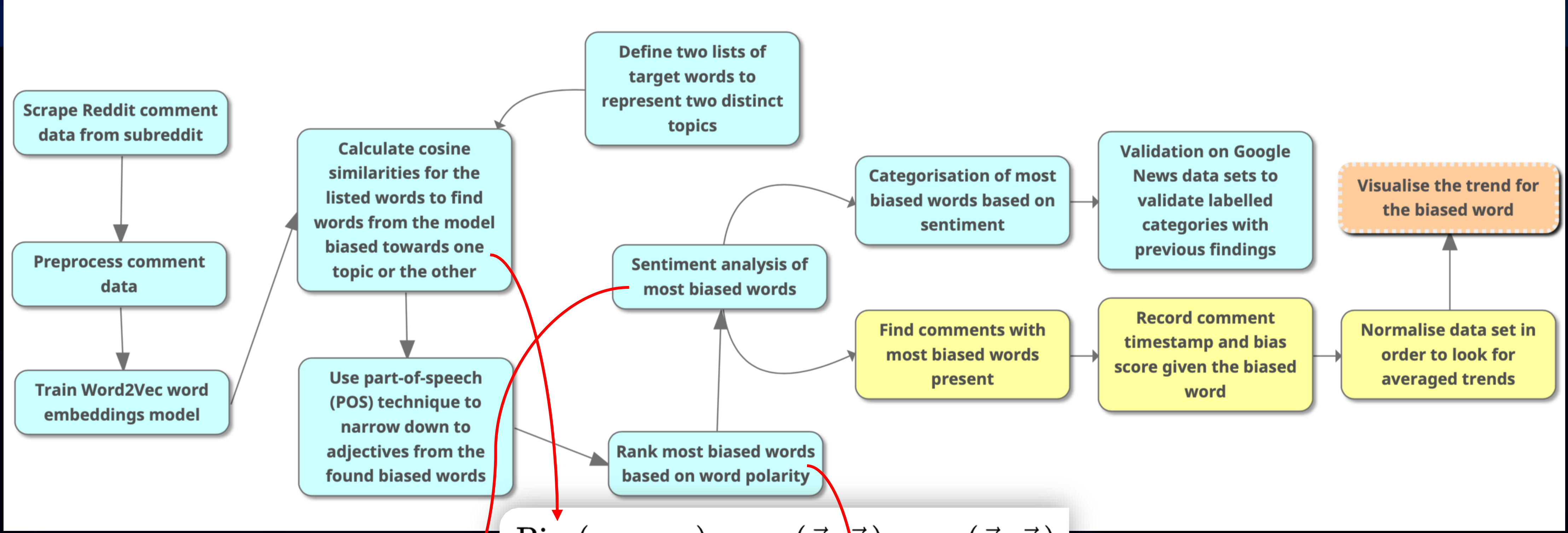
Not least because of its topical infrastructure, Reddit has been a popular site for Natural Language Processing studies – for instance, to successfully classify mental health discourses (Balani and De Choudhury 2015), and domestic abuse stories (Schrading et al. 2015). LaViolette and Hogan have recently augmented traditional NLP and machine learning techniques with platform metadata, allowing them to interpret misogynistic discourses in different subreddits (LaViolette and Hogan 2019). Their focus on discriminatory language is mirrored in other studies, which have pointed out the propagation of sexism, racism, and 'toxic technocultures' on Reddit using a combination of NLP and discourse analysis (Mountford 2018). What these studies show is that social media platforms such as Reddit not merely reflect a distinct offline world, but increasingly serve as constitutive spaces for contemporary ideological groups and processes.

Such ideologies and biases become especially pernicious when they concern vulnerable groups of people that share certain *protected attributes* – including ethnicity, gender, and religion (Grgić-Hlača et al. 2018). Identifying language biases towards these protected attributes can offer important cues to tracing harmful beliefs fostered in online spaces. Recently, NLP research using word embeddings has been able to do just that (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). However, due to the reliance on predefined concepts to formalise bias, these studies generally make use of larger textual corpora, such as the widely used Google News dataset (Mikolov et al. 2013). This makes these methods less applicable to social media platforms such as Reddit, as communities on the platform tend to use language that operates within conventions defined by the social group

*Author's copy of the paper accepted at the International AAAI Conference on Web and Social Media (ICWSM 2021).

[1] Subreddits are commonly spelled with the prefix '/r/'.

Ferrer et al. (2020)



Scrape Reddit comment data from subreddit

Preprocess comment data

Train Word2Vec word embeddings model

Define two lists of target words to represent two distinct topics

Calculate cosine similarities for the listed words to find words from the model biased towards one topic or the other

Use part-of-speech (POS) technique to narrow down to adjectives from the found biased words

Rank most biased words based on word polarity

Sentiment analysis of most biased words

Categorisation of most biased words based on sentiment

Validation on Google News data sets to validate labelled categories with previous findings

Find comments with most biased words present

Record comment timestamp and bias score given the biased word

Normalise data set in order to look for averaged trends

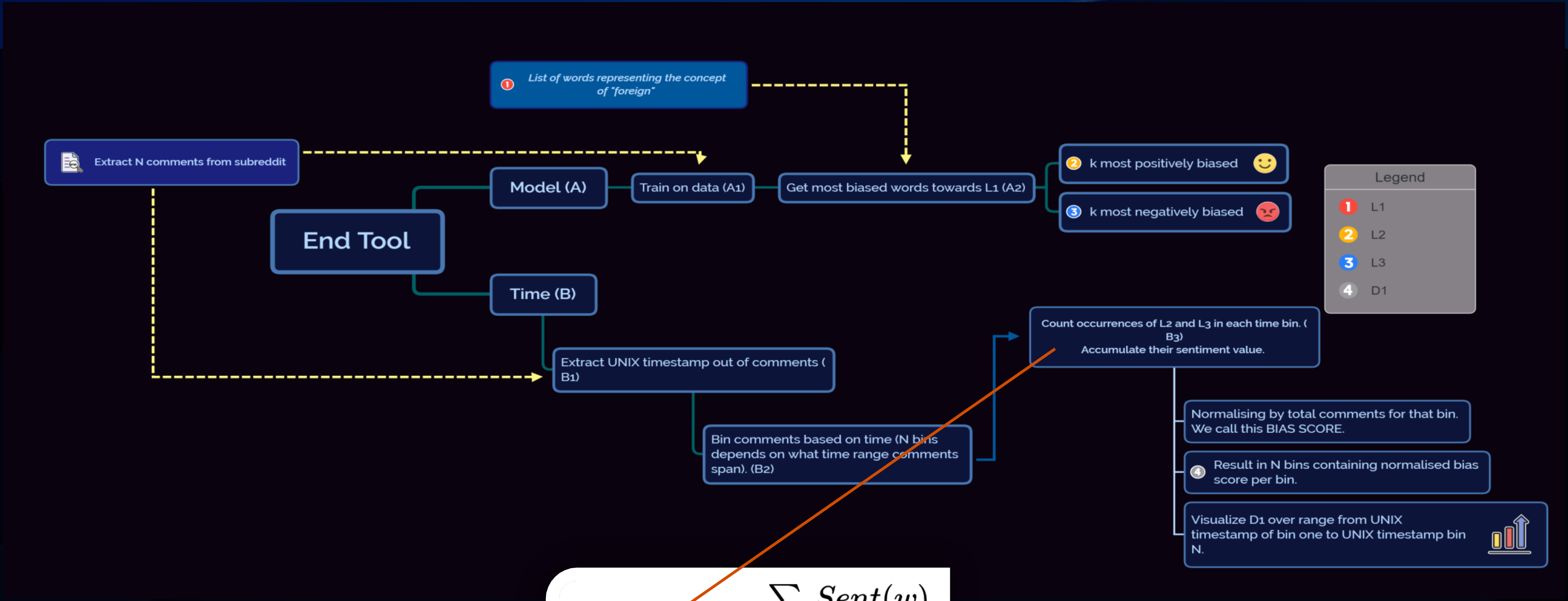
Visualise the trend for the biased word

$$\text{Bias}(w, c_1, c_2) = \cos(\vec{w}, \vec{c_1}) - \cos(\vec{w}, \vec{c_2})$$

$$\text{Sent}(W) = \frac{1}{|W|} \sum_{w \in W} SA(w)$$

$$\text{Most Biased}(V, c_1, c_2) = \arg\max_{w \in V} \text{Bias}(w, c_1, c_2)$$

Our Additions



1 List of words representing the concept of "foreign"

Extract N comments from subreddit

End Tool

Model (A)

Train on data (A1)

Get most biased words towards L1 (A2)

2 k most positively biased

3 k most negatively biased

Time (B)

Extract UNIX timestamp out of comments (B1)

Bin comments based on time (N bins depends on what time range comments span). (B2)

Count occurrences of L2 and L3 in each time bin. (B3)
Accumulate their sentiment value.

Normalising by total comments for that bin. We call this BIAS SCORE.

4 Result in N bins containing normalised bias score per bin.

Visualize D1 over range from UNIX timestamp of bin one to UNIX timestamp bin N.

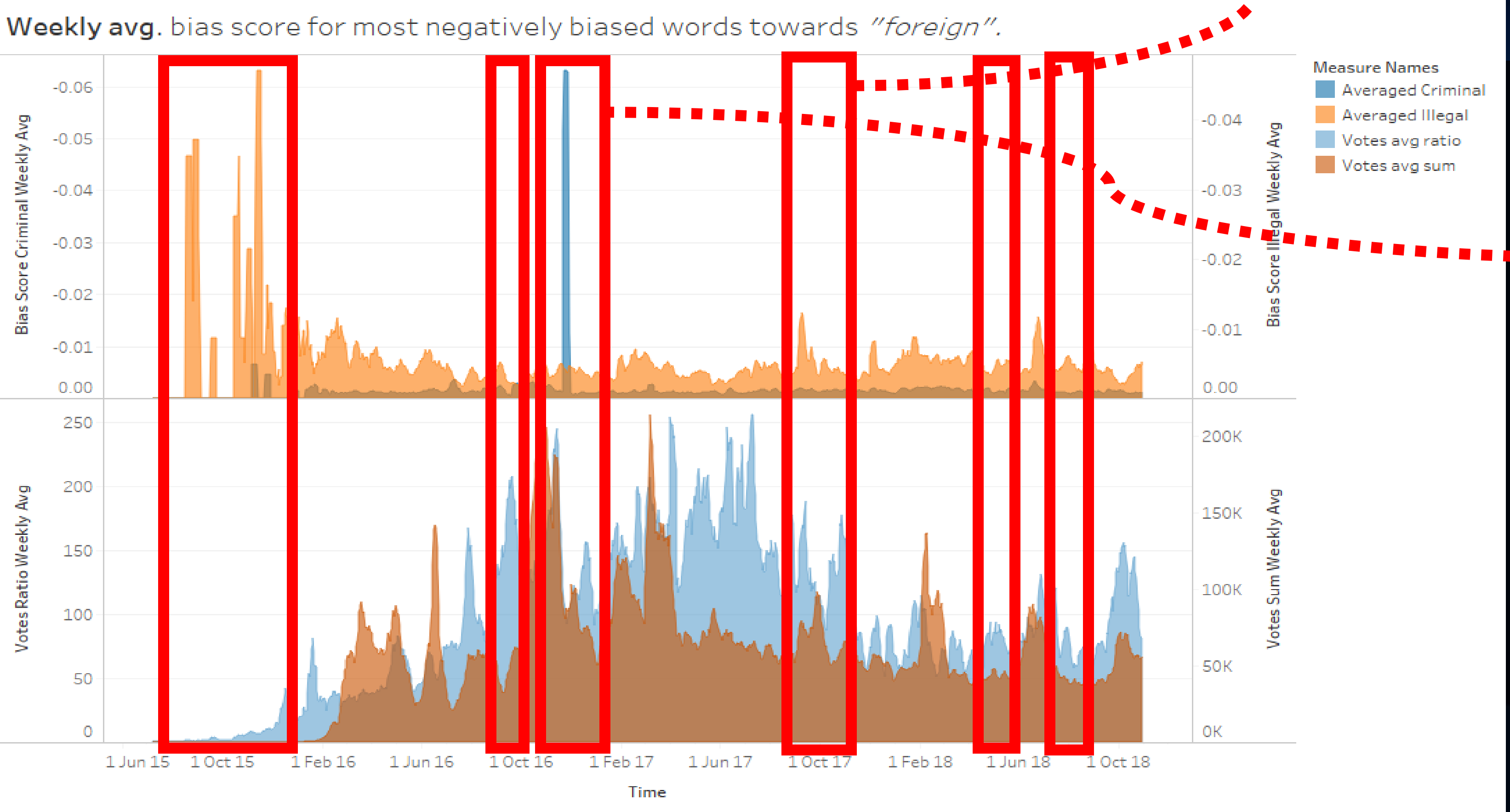
Legend
1 L1
2 L2
3 L3
4 D1

$$BiasScore = \frac{\sum_{w \in C} Sent(w)}{|C|}$$

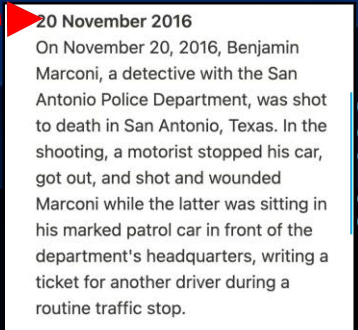
Issues

- Ferrer et al. (2020)
- APIs
- Formatting
- Subreddits
- Ingestion
- Processing

Results and Validation



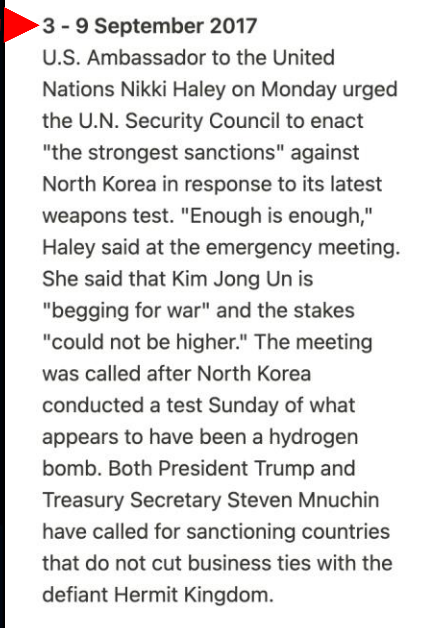
Weekly avg. bias score for most negatively biased words towards *"foreign"*.



20 November 2016

On November 20, 2016, Benjamin Marconi, a detective with the San Antonio Police Department, was shot to death in San Antonio, Texas. In the shooting, a motorist stopped his car, got out, and shot and wounded Marconi while the latter was sitting in his marked patrol car in front of the department's headquarters, writing a ticket for another driver during a routine traffic stop.

Source



3 - 9 September 2017

U.S. Ambassador to the United Nations Nikki Haley on Monday urged the U.N. Security Council to enact "the strongest sanctions" against North Korea in response to its latest weapons test. "Enough is enough," Haley said at the emergency meeting. She said that Kim Jong Un is "begging for war" and the stakes "could not be higher." The meeting was called after North Korea conducted a test Sunday of what appears to have been a hydrogen bomb. Both President Trump and Treasury Secretary Steven Mnuchin have called for sanctioning countries that do not cut business ties with the defiant Hermit Kingdom.

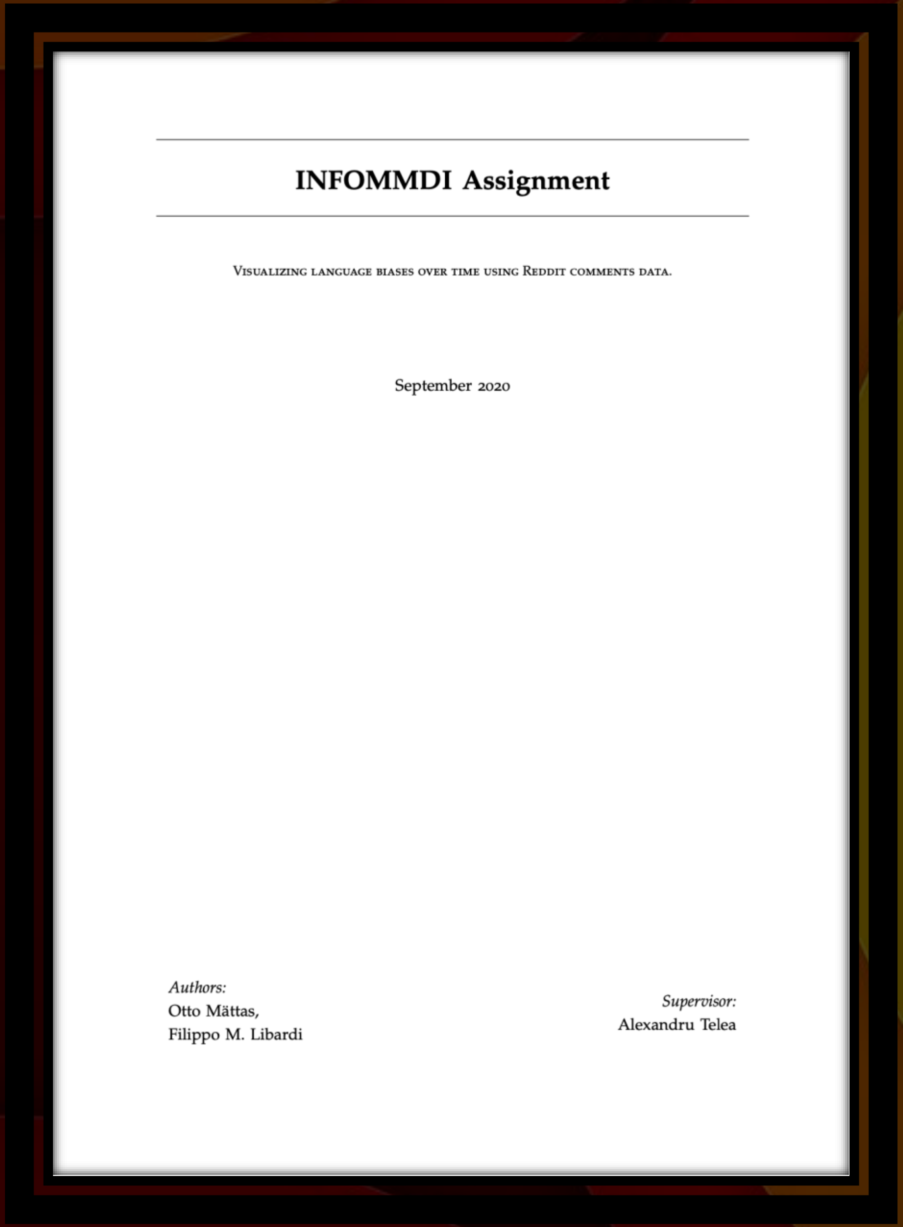
Source

Progress

- We have further specified and improved the methodology proposed by Ferrer et al. (2020);
- We have given a detailled description of the technical implementation;
- We have validated the results and the methodology for our purpose;

- We invite you to inform our proposal for future research!

And to read our report, of course…



INFOMMDI Assignment

Visualizing language biases over time using Reddit comments data.

September 2020

Authors:
Otto Mättas,
Filippo M. Libardi

Supervisor:
Alexandru Telea