Visualising language bias over time using Reddit comment data

February 2021

Authors:
Filippo M. Libardi (6715753),
Otto Mättas (6324363)

Supervisor:
Alexandru Telea

I. Abstract

We hereby propose a novel approach in language bias visualisation. We do not just present a mere visualisation tool but also explain the methodology for collecting what we call a bias score from a text corpus. We enhance a previously implemented method on a corpus of Reddit comments in order to visualise the trend of language bias over time. Some peculiar trends have been recognised in the usage of biased terms and this has been made possible by an added graphical visualisation of the trend overall.

Mentioned trends have subsequently been correlated with real world events, sparking interest in the study of when people use biased terms towards a given concept (which in this case translates to *foreign*). Given the results produced we claim our tool to be a valid aid in researching language bias.

II. Introduction

Generally, biases are identified as cognitive constructs that translate into a group of people having mutual ideas about how another group commonly behaves. Language works as both a reflection and a propagation of generalisations that individuals convey with them.

Accordingly, a meaning of etymological conduct deriving from these cognitive constructs is that of linguistic bias, defined in Beukeboom 2014 as *a systematic asymmetry in word choice as a function of the social category to which the target belongs*.

In this paper, we present a visualisation system that displays linguistic bias trends over time. The work hereby proposed relies heavily on Ferrer et al. 2020. We intend to take the modelling method presented there, train our own machine learning model and display the results over a time-span.

In Ferrer et al. 2020, the authors - quoting - *present a data-driven approach using word embeddings to discover and categorise language biases on the discussion platform Reddit.*

Reddit is a social media platform, organised in members' clusters (i.e. subreddits). These can be perceived as communities of people interested in the same topic. The range of topics is vast and varied, not only restricted to classical notions of similar interests like politics or funny pictures, but also specific world views on a very narrow topic. Reddit's main quality lays in its discussion-induced nature.

The aim of our research is to train a model over a specific subreddit's (i.e. r/TheDonald) comment corpus, to then quantify the bias as a function of time, and to finally visualise its evolution. We assume that given such visualisation, one could spot and understand the rises and falls of language bias as a consequence of real world events more easily. Specifically, an online community's user could better understand why linguistic bias grew at certain points in time by correlating it with events that took place within that same community. And this is where our research question stems from:

Do associations or correlations rise out of bias and sentiment visualisations after introducing the time dimension?

In section III, we present the above-mentioned paper by Ferrer et al. 2020 as well as other related papers entailing the field of language bias' identification and discovery. We will draw a line connecting each work to this project's very scope.

Furthermore, in section IV, we describe the data corpus and how we process it in order to extrapolate the needed features. This section will also look at the usage of possible methods to gather on-the-go data. Specific limitations with Ferrer et al. 2020 and related research is also covered.

In section V, we aim to provide an overview of our approach and describe both the model utilised in this project and our own addition to it.

In section VI, we propose and classify different visualisation methodologies usable for the final data visualisation. Following this, we will carry a discussion on the proposed visualisation methods and their respective strengths and weaknesses.

In section VII, we will present empirically collected results and give some further evidence to support them for validation. Also, we will focus on abstracting away from the technical details towards having a wider perspective on the topic of language bias.

In section VIII, we will discuss limitations with our work and propose possible threads to follow in the future.

In section IX, we will first collect our thoughts by explaining the premise of our research. We will then conclude by exposing our personal opinion on why this research is imperative so as to potentially build our way towards a more balanced society.

III. Related Work

There is great amount of literature explaining how language both reflects and propagates social generalisations. However, such investigations are mainly informed by human surveys, word references and subjective examinations - or knowledge on various languages as explained by Paugh 2005. The present project is heavily based on a recent research by Ferrer et al. 2020. Here, authors make use of a Word2Vec model to estimate word embeddings of a corpus and its related bias. We have created a visual representation of their technical pipeline, and our addition to it can be seen in yellow-orange colors on Figure 1.

A. Word2Vec

As the name suggests, Word2Vec translates words into a dense vector (i.e. a vector containing many non-zero values). The incorporated calculation process utilises a neural network model to understand word associations from an enormous corpus of text. When trained, a particular model can distinguish interchangeable words or predict extra words for an incomplete sentence. The vectors are evaluated carefully with the end goal being that a straightforward numerical capacity (the cosine similarity between the vectors) demonstrates the degree of semantic similitude between the words conveyed by those vectors. More about this Natural Language Processing method is explained by Mikolov et al. 2013.

As Ferrer et al. explain in their code repository [1], the research makes use of the gensim python package to build and train the model. More about this package can be read in Řehůřek 2020.

In addition, they add a layer of abstraction which makes use of the model to estimate bias and the most biased words. These concepts will be formally and further explained in section V-B1.

B. Word embeddings

Garg et al. 2018 present a computational approach to quantify 100 years of gender and ethnic stereotypes through word embeddings. The presented model is trained over 100 years of text data, gathered from Google News and other digital newspapers data sets. Researchers correlate the trends shown by the model with real life events and perform statistical tests to validate the model. Findings show concrete correlations, validating our ambition to use word embeddings for our purposes. We took inspiration from this latter methodology and adapted it to serve our purpose.

IV. Data processing and infrastructure

As researchers before us, we also chose social the news aggregation, web content rating, and discussion-based website Reddit. As the company's motto says, it is aiming to be "the front page of the internet" which also coincides with our goal to have ecological validity and real world correlations.

In principle, we are using open data APIs (abbreviation for Application Programming Interface) as we are investigating the comments shared publicly on the internet.

First, in section IV-A, we are giving an overview of limitations that we ran into, following Ferrer et al. 2020.

[1] https://github.com/xfold/LanguageBiasesInReddit

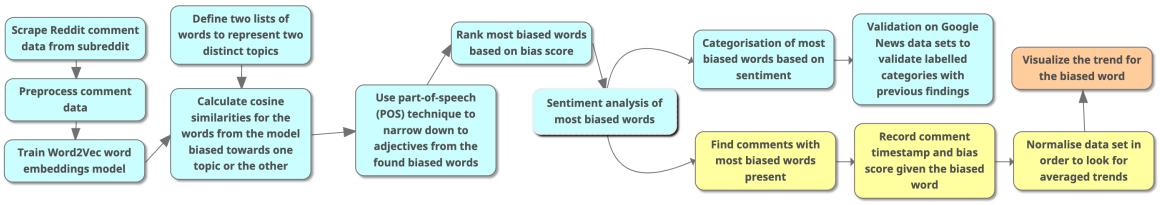

Figure 1: An overview of the whole pipeline as presented by Ferrer et al. 2020; shown in blue. Our addition is shown in yellow and orange color.

Next, in section IV-B together with section IV-E and section IV-F, we are sharing details on how we collected our data corpus and the issues we ran into.

Furthermore, we explain the decisions that informed our data corpus selection in section IV-D.

Finally, in section IV-C and section IV-G, we share steps taken to process the data to meet our needs.

A. Redoing work from previous research

We have investigated data collection methods as presented by Ferrer et al. 2020 and found it incomplete. This section will see an expansion with sharing the weak backdrop.

The main issue we found is that the experiment carried out by Ferrer et al. 2020 is not directly replicable.

Primarily, this is due to the given code incapability to handle data with size greater than their initial data set. Additionally, they do not provide complete details on technical specifications of the machine the experiment has been run on. After approaching the researchers, we finally received an indication of the technical specification as presented in Ferrer et al. 2021.

This confirmed our worries about replicating their work - the implementation presented by Ferrer et al. 2020 is only optimised for the data set they used originally.

Additionally, Ferrer et al. 2020 did not specify nor describe the exact data set they used. Furthermore, the sample data set presented in the repository does not directly reflect the actual data set used in the original experiment. It is a mere example comprised from a similar data

```
[ec2-user@ip-172-31-12-151 LanguageBiasesInReddit]$ python3 Run.py
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/ec2-user/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /home/ec2-user/nltk_data...

        ********************************************************************************
        Test run using a toy dataset of 1000 comments collected from r/TheRedPill
        ********************************************************************************

Training new model Datasets/toy_1000_trp.csv
->Starting with Datasets/toy_1000_trp.csv [body], output Models/toy_trp_model, window 4, minf 10, epochs 5, ndim 200
Traceback (most recent call last):
  File "pandas/_libs/parsers.pyx", line 1149, in pandas._libs.parsers.TextReader._convert_tokens
  File "pandas/_libs/parsers.pyx", line 1279, in pandas._libs.parsers.TextReader._convert_with_dtype
  File "pandas/_libs/parsers.pyx", line 1296, in pandas._libs.parsers.TextReader._string_convert
  File "pandas/_libs/parsers.pyx", line 1518, in pandas._libs.parsers._string_box_utf8
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xbf in position 1: invalid start byte
```

Figure 2: Formatting error in the original algorithm.

set but not the original. When trying to run the experiment without changing anything in the original code or the data, the implementation fails. Error seems to reflect on the sample data's formatting which can be worked around but should not.

In order to find clarity about the initial data set used, we went back another step.

As there was a clear reference to another research paper about collecting the data from Reddit, we investigated it - presented also in Baumgartner et al. 2020.

Unfortunately, it focuses on collecting the data overall. Also, no specific details about their technical implementation are shown. If presented, it would give a hint as to how the follow-up research on the language bias was conducted. We got many technical tips about the Pushshift platform and how it is built on a high level. Again, the details are missing.

We approached the researchers who built the Pushshift platform to ask for more details but they have not responded in time.

Though it was clear that the researchers had used Pushshift API for their data collection.

This originally made us think that we could train the model on two full data sets, respectively of 5 and 20 gigabytes in size. The action resulted in several different issues, mainly the code was not optimised to handle that size of data and

so the RAM of our computing clusters ran out quickly even just during the pre-processing steps.

Additionally, the training itself did not implement methodologies that the python package they used provides for training on large data sets (i.e. streaming from a file directly rather than storing in a variable).

Many changes in code have been implemented and the whole pipeline has been refactored to comply with a large data - or at least larger than the initial data set.

B. Reddit API

At first, we investigated Pushshift API which was also mentioned in the research paper by Ferrer et al. 2020. As they have also referenced it from related research Baumgartner et al. 2020, we investigated the underlying ingestion method. Unfortunately, it was not disclosed to any degree so we had to rely on our own skills and expertise in collecting information from publicly available APIs.

Pushshift API is developed by third-party developers and has abstracted away from the official Reddit APIs. Also, it has become more user-friendly to source linked data from with less effort. Unfortunately, the API is not very reliable as we receive many timeout errors while being connected for ingestion.

Additionally, they publish static data sets from archived content to be served and used freely. Unfortunately, the technical issues are even more visible as the data sets are stored in static files up to the size of 15GB which need to be downloaded in one continuous network session. Though the service is not reliable for accessing data sets larger than a few megabytes. From error responses received, the service does not seem to be running on a scalable infrastructure that is able to support taxing requests like ours.

We moved on to investigating the PRAW Python framework for accessing the Reddit API directly.

PRAW, an acronym for "Python Reddit API Wrapper", is a Python package that allows for simple access to Reddit's API. PRAW aims to be easy to use and internally follows all of Reddit's API rules. With PRAW, there's no need to introduce sleep calls in your code. We need to present our client with an appropriate user agent and start querying directly.

Right now, we are still waiting on PRAW developers to respond to specific questions about the framework in order to ingest our archive properly. To mitigate the need for data, we have started ingesting newly created comments since 20/12/2020 with a valid PRAW method.

We moved on to investigating the official Reddit API. Even though not meant for high-volume data aggregation and ingestion, the API is closest to the source of truth. Also, there are many specific questions that the Reddit developers are best suited to answer in the first place. We are also waiting on Reddit developers to respond to specific questions about the framework in order to ingest our archive properly.

C. Format

We are ingesting raw JSON (abbreviation for JavaScript Object Notation) comment data which will be transformed during ingestion. The ingestion process results in a CSV-formatted file, including the comment body and timestamp which then will be sent to the machine learning process explained in section V. The statistical model built by Ferrer et al. 2020 together with our addition of the time dimension will output yet another CSV-formatted file - now only including the bias score (per bin) and time. Final data, including timestamp and bias score, will be forwarded to the visualisation process.

In general, hold to web standards and follow best practices in our formatting.

D. Subreddits

We aimed at making an informed selection and chose socially relevant and active communities. We are looking for some that will be representative or with a user base, that might be ideologically and evenly distributed over the topics discussed. Also, activity is taken into consideration.

To start, we sought out preliminary statistics on the content of some political subreddit's comments. Namely, we are interested in race and bias towards it. This comes from the authors' intuition on what topic could be considered relevant in the current world. For example, we are using official racial and ethnic categories and definitions for National Institutes of Health (NIH) diversity programs. We count all the mentions of races (noun form) in the subreddits focused on US politics.

1) Race and ethnicity: The 2010 US Census included changes designed to more clearly distinguish Hispanic ethnicity as not being a race - as it is presented in *The U.S. Census Bureau Race Classification* 2020.

As we are looking at racial bias, we ought to exclude the ethnicity data but think keeping it allows to avoid some confusions with our audience. We are defining and using the following for our purposes:

- Hispanic or Latino. A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin," can be used in addition to "Hispanic or Latino."
- American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
- Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
- Black or African American. A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American."
- Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
- White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

Given the above we formed our target list (that tries to convey the concept of foreign) as follows: *["mexican", "foreign", "hispanic", "black", "brown", "immigrant", "african", "haitian"]*

2) Political subreddits: The authors made a preliminary selection from known regional subreddits for sampling:

- r\Politics
- r\US Politics
- r\American Politics
- r\Canada
- r\Europe
- r\American Government
- r\UK Politics
- r\Euro
- r\Palestine
- r\EU Politics
- r\Middle East News
- r\Israel
- r\India
- r\Pakistan
- r\Cascadia
- r\Iran
- r\TheDonald

3) Subreddit sampling: This is the list of words we are searching for in the comments for any subreddit:

- hispanic, latino, spanish origin
- american indian, alaska native
- asian
- black, african american, haitian, negro
- native hawaiian, other pacific islander
- white

E. API endpoints and queries

We planned to harvest data from selected subreddits via APIs. We started by using Pushshift API, then moved to Reddit APIs via different methods. As we found constraints with

them - namely on real-time data processing and platform reliability - we looked for static data instead.

F. Data ingestion

After many failed attempts to gather our own data set via the official Reddit API and via the Pushshift API, we went back to using a static data set found online. This did not include all comments from the decided subreddit r/The-Donald but held most comments. Namely, we found data from the inception of the subreddit until the end of 2018.

We decided to move forward with the incomplete data set and limit the scope of our research based on it. We could focus our efforts to the time before 2019 to work on proving our idea. Data from the start up to 2018 has been made available by Cornell University researchers for their own research purposes as presented by Chang et al. 2020.

G. Preprocessing

Trying to run the original algorithm on our data set failed. Mainly, the script was unoptimised for data other than the original authors used. Most often, the failure came from memory errors. The algorithm was not only training the model, but also pre-processing the data for training the model.

Unfortunately, the data set size exceeded what the original method was designed and developed for. We used Amazon AWS cloud resources, namely EC2 on-demand instances to run our computational workloads. We extended the technical specifications from 8 CPU cores / 16GB of RAM per instance up to 32 CPU cores / 72 GB of RAM per instance. Unfortunately, the original algorithm did not even get through the pre-processing part of the method without exhausting the memory available. We decided to break off the pre-processing step from the original method in order to keep the cost under control, hoping for a success. From certain thresholds, EC2 instances cost rises while it was not clear if we ever would be able to fit our data set into the memory with the original method. Cost can be estimated by looking at documents presented in *Amazon EC2 On-Demand Pricing* 2021.

1) Actions: As to give exact technical details of our implementation, we are going to outline all the actions taken in order to reach the final conclusive solution:

1) As the data set came in the form of one JSON file and was too large (20 gigabytes) in size to be worked on as a whole, we split it up in 100 pieces.
2) That allowed us to convert the JSON files into CSV files which is the de facto input format for the training part of the original algorithm.
3) As conversion also reduced the file size remarkably, we decided to merge the 100 CSV files back together in order to start training on the data. In case you wish to replicate this process, do not forget to keep only one CSV header instead of keeping all the headers from all the files. Keeping them could affect the training results as unnecessary noise is introduced.
4) We then uploaded the final CSV to cloud storage for safekeeping and as an accessible source location.

2) Additional findings: As to give more details on tried but failed approaches:

1) We tried exploring the AWS Sagemaker Studio which comes with more tools specific to training models on big source data. Also, it comes with built-in algorithms, one of which was Word2Vec, also employed within the original algorithm by Ferrer et al. 2020. Unfortunately, the original algorithm is still very much a developmental version in terms of clarity of the code. For example, no comments that would have helped us directly carry over the core of the method to AWS Sagemaker Studio were present. Even when we understood how the original method worked in principle, we could not replicate the whole algorithm as there were no clear

"building blocks" within the original method to be recreated outside of the method.

2) We soon abandoned the idea of continuing on AWS for the purpose of our project. This is still worth mentioning as the AWS environment is able to scale as the need for resources during training arises. It is also able to scale back the resources when not in active use. What's more - AWS Sagemaker might help cut costs by using spot instances which become available at random times during the day so the platform takes care of running the workloads for you.
3) We resulted to running the full script on a hefty 36 core CPU / 72 GB of RAM server. Unfortunately, this also failed as we already explained in section IV.

V. Method

In this section we aim to extensively describe our approach for building the visualisation tool. We will discuss the approach taken by Ferrer et al. on quantifying the language bias as well as describing what we want to introduce to the proposed method.

More closely, we will describe how the method proposed by Ferrer et al. 2020 formalises the notion of language bias and we will try to deliver the intuition behind the authors choices in section V-A.

In section V-C, we add detail to our contribution and explain how we thought to approach its development.

A. Quantifying language bias

The model proposed by Ferrer et al. 2020 tries to capture the language bias by training a Natural Language Processing model on a given data corpus of Reddit comments. The model learns the corpus' embeddings that transform text into high-dimensional dense vectors and captures semantic relations between words.

When the words are turned into vectors, it is possible to manipulate them as such. For instance, distances between vectors can be evaluated and semantic similarities can be drawn between the concepts that the vectors represent.

A classic example of this comparison is *Queen - Woman ; King - Man*. Intuition suggests that if the vector (i.e embedding) of the word *man* is added to the vector of the word *royal*, it results in the vector of the word *king*.

Similarly, the same process can be applied to the word *woman* which translates to the word *queen*, when the word *royal* is introduced. The visual representation of the latter example is given in Figure 3 as presented by Ethayarajh, Duvenaud, and Hirst 2019. In the figure vectors representing words in italic are depicted as arrows whether the others as point. This is due to the italic ones being displacement vectors whereas the others being position ones. For a summarised explanation, one might check Ethayarajh 2019 instead.

Once the model has estimated all the embeddings of the corpus, it can be given two other lists of words. These target lists contain the words towards which the bias wants to be evaluated.

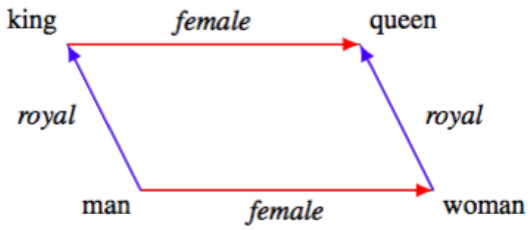

Figure 3: The visual representation of word embeddings in a vector space. All the concepts visualised here are vectors, for simplicity only the results of the operations are visualised in vector graphics (i.e. a line) as presented by Ethayarajh, Duvenaud, and Hirst 2019.

B. Formalising language bias

Given all the calculated embeddings of a corpus (Reddit comments in our case) and the lists of target words, the model identifies the most biased words towards these lists in the corpus.

Let $S_1 = \{w_i, w_{i+1}, ..., w_{i+n}\}$ and $S_2 = \{w_j, w_{j+1}, ..., w_{j+n}\}$ be sets of target words that try to convey two opposite concepts.

For example, S_1 tries to identify the concept of *male* (e.g $\{he, son, his, him, father, male\}$) whereas S_2 tries to identify an opposite concept, such as *female* (e.g $\{she, daughter, her, mother, female\}$).

The model computes the centroids of each target set ($\vec{c_1}$ for S_1 and $\vec{c_2}$ for S_2) obtained by averaging the embedding vectors of word $w \in S$. A word belonging to w is biased towards S_1 with respect to S_2 when the cosine similarity between the embedding of $\vec{w}$ is higher for $\vec{c_1}$ than for $\vec{c_2}$.

$$\text{Bias}\,(w, c_1, c_2) = \cos\left(\vec{w}, \overrightarrow{c_1}\right) - \cos\left(\vec{w}, \overrightarrow{c_2}\right) \quad (1)$$

where $\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$.

Positive values of *Bias* mean that a word w is more biased towards S_1, while negative values of *Bias* mean that w is more biased towards S_2.

Given this definition of bias, the model also estimates the k most biased words towards the given sets. This feature is fundamental towards the building of our final visualisation. More details about this are given in section V-B1.

1) Most biased words: Let V be the vocabulary of a word embeddings model.

The model identifies the k most biased words towards S_1 with respect to S_2 by ranking the words in the vocabulary V using *Bias* function from Equation 2:

$$\text{MostBiased}\,(V, c_1, c_2) = \arg\max_{w \in V} \text{Bias}\,(w, c_1, c_2) \quad (2)$$

The intuition behind the approach described in section V-C is to use the k most biased words in order to quantify bias. The latter translates into counting occurrences of these k over time.

2) Sentiment analysis: When talking about bias, there are often misconceptions and confusion over negative bias, positive bias and how they relate to each other. As bias is essentially a human tendency to associate two concepts with one another, it does not necessarily have positive or negative connotations attached to it.

In order to understand whether the bias is negatively or positively charged, the model makes use of sentiment analysis.

Specifically, the model implements VADER, *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, introduced by Hutto and Gilbert 2015. VADER estimates the average of the sentiment over a set of words W as follows:

$$\text{Sent}(W) = \frac{1}{|W|} \sum_{w \in W} SA(w) \quad (3)$$

where SA is a sentiment analysis system which inner working will not be explored in this report. Equation 3 returns a value $\in [-1, 1]$ corresponding to the averaged sentiment (or polarity) determined by the sentiment analysis system, -1 being strongly negative and $+1$ strongly positive.

We will make use of Equation 3 to estimate the connotation of the bias found.

C. Adding a time dimension

We found challenges in interpreting the meaning of visualisations presented by our predecessors in their research. Namely, it is hard to understand, what is signified by their graphs, how exactly does the bias evolve in time. Moreover, the practical challenge in our research scope came from attaching a time factor to the bias.

While training, the model does not take time as a feature into account. This means we had to find our own way of including time to the model output.

Hereby, we aim to describe our addition to the model proposed. The depiction of our approach is twofold: firstly, we will describe how we aggregate the bias values returned by the model so as to fit into a temporal timeline. Secondly, we show what this timeline entails and what it formally translates to.

An overview of the procedure is shown in Figure 4.

Once the model has returned a list of k most biased words ($A1$ in Figure 4), we take only the words that have a positive or negative connotation (found through sentiment analysis) and that

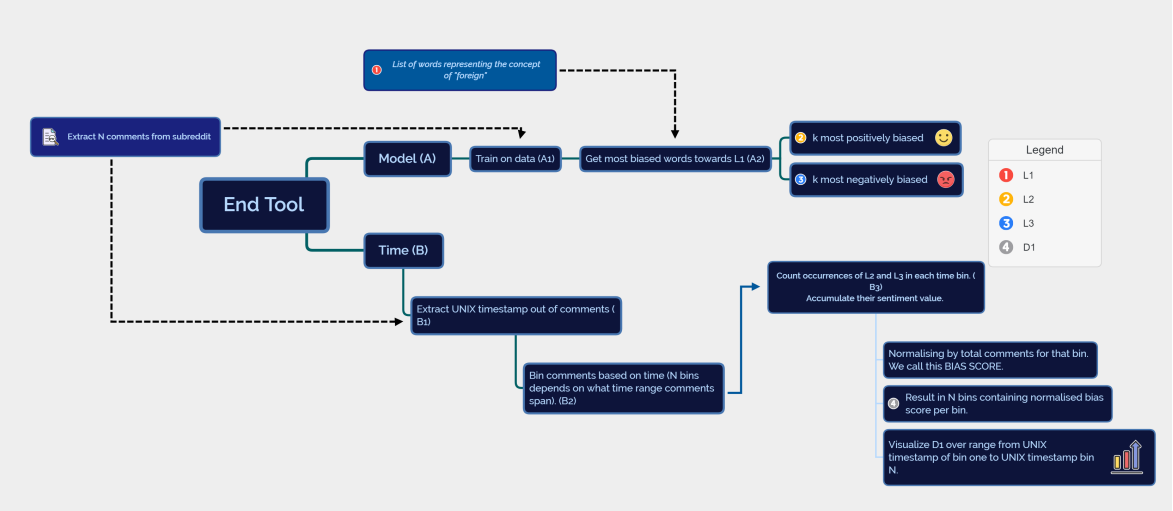

Figure 4: An overview of our whole pipeline. Yellow dotted line indicates where data corpus has been directly used. Green line indicates action taken within formalised technical solution. Blue line indicates manual action. White line indicates manual action specific for the visualisation creation.

are tagged as adjectives (these are represented as $L1$ and $L2$ respectively in Figure 4), leaving out the biased words that have no polarity.

Then, we iterate through the same comments the model has been trained on and count the occurrences of each word in the list.

A negative and positive *bias score* is accumulated while doing so.

The *bias score* is given by accumulating the sentiment defined in Equation 3. Formally, we define our *bias score* as follows.

$$BiasScore = \frac{\sum_{w \in C} Sent(w)}{|C|} \tag{4}$$

where C is a a set of all the words present in all the comments of one time bin.

The aggregated sentiment is normalised by the cardinality of the set (i.e. number of comments) in order not to favor days where users were commenting more actively than on average.

Because we are iterating trough the comments in their original form, we can access the timestamp attached to them ($B1$ in Figure 4).

As mentioned, we bin the time on a daily basis (one bin signifies one day), where for each bin we sum up all the bias scores and normalise by the total number of comments present in that specific bin ($B3$ in Figure 4), following Equation 4. Once k bins have been evaluated with their respective normalised *bias score*, we can display them over time.

An implementation of the thought approach is shown in Figure 5. The plot shows a list of words that are the most negatively biased towards the foreign people. Because the bias considered in this case only has a negative connotation (i.e. frequency of insults correlated/biased towards the concept of foreign), we see the line plot extending on the negative side of the plane.

An first approach saw the end tool being able to comprehend a visualisation for both the negative and positive bias for the word lists given to the model.

The problem with the latter approach is that the positive and negative bias often have a major discrepancy in the order of magnitude they are expressed in. This means that the negative *bias score* has much greater values than the positive one. Accordingly, we finally decided to only visualise the negative bias towards a related concept for the purposes of our research. Following this, we will discuss how to approach and visualise this in section VI.

D. Data processing decisions

We applied different data processing methods during the whole pipeline, following these will

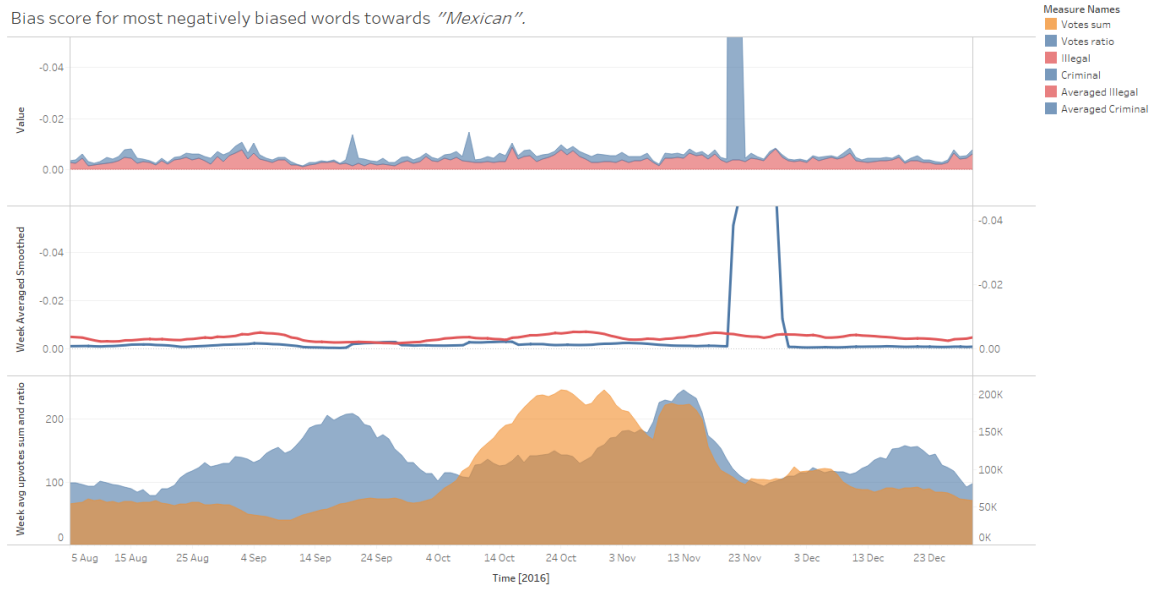

Figure 5: *Bias Score* towards *foreign* ethnicity's evaluated by training the model on ~20 million comments of the subreddit r/TheDonald. The *Bias Score* is evaluated as shown in Equation 4. The middle graph introduces a smoothed view by averaging over the value of a week for bias score. The bottom one shows the accumulated score of the day (upvotes + downvotes) as well as the ratio between negative and positive comments (# of positive / # of negative).

be explained.

Firstly, to reduce the size of the data at hand we decided to subsample the comments contained in the dataset.

Secondly, to aid us and future users find correlation of bias score trends and real life events we displayed what we call vote ratio and sum.

1) Subsampling: An additional note is to be made on the size of the data corpus. As described in section IV-A, the data we retrieved extends into gigabytes in size. This introduced a novel issue, which is not only related to the model implementation but also to training time and the robustness of the training method itself.

After realising that the time needed for training a model on such a corpus largely exceeded the time available to us, we decided to subsample from this set of comments, in order to reduce the size while still maintaining the meaningfulness of the data.

There are several possible ways of subsampling from a distribution, we chose simple random sampling and the reason is twofold.

Firstly, as mentioned, the data at hand did not make it easy to perform any sort of operation like dividing into subgroups or performing an analysis of the word distribution. Only running line by line creates significant computational load and increases processing time. Secondly, the random sampling process has shown to maintain the essence of the data, conveyed in the end, as presented by Taherdoost 2016 and West 2016.

A formalisation of the methodology applied is described below in Equation 5.

$$\begin{aligned} P &= 1 - \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \dots \cdot \frac{N-n}{N-(n-1)} \\ &\overset{\text{Canceling:}}{=} 1 - \frac{N-n}{N} \\ &= \frac{n}{N} \end{aligned} \tag{5}$$

where N is the total number of data points in the set (~482031200 in our case) and n is the sampled population (~4820312). This means we are sampling around 1% of the total population P.

After reducing the population size, we applied the method to extract the *bias score* explained in Equation 4 and exported everything into a final .csv file, only containing two columns (the date and the bias score of a particular comment).

2) Votes ratio and sum: Lastly, one additional action on the data has been performed.

Reddit is known for its upvoting and downvoting system, where users can not only show their support towards a comment or post (upvote), but also express a negative feeling about it (downvote). The final score of a comment or post is the sum of both down- and upvotes given to it.

We tried to understand if the bias score's rise and fall over time might correlate to people having a more active discussion on the platform, overall. The *activity* in this case translates to controversy - that is, posts with many down and up voted comments attached to them.

Logically, a post will be more controversial, if it has a great number of negative per positive comment. This is captured by the *votes ratio* (blue area in bottom graph of Figure 5), and is expressed by $|positivecomments|/|negativecomments|$.

Additionally, we display the overall sum of votes in a day, this is to see if users prefer days where a great amount of *bias score* is present (orange area in bottom graph of Figure 5). Again, a low number of overall votes could indicate a controversial day as it would mean that either people did not like comments at all that day or they have equally up and down voted them.

These metrics can be seen to validate our intuition. Both, the overall sum of votes and the ratio go down when there is a spike in usage of biased words (in the chart, this can be seen around November 2016).

Essentially, looking at the ratio of negative comments versus positive comments means, that there were comments that people from the subreddit did not agree on, sparking discussions. This is to say that if one comment has many upvotes but is also followed by other comments having a great number of downvotes, then there is probably a debate being discussed.

This suggests that when users engage in debates with other users, their language becomes more biased. Whereas when all users are agreeing on what is said the language can be described as "normally" biased or biased as it would regularly be for each user.

This factor could not be seen without our addition to the research paper by Ferrer et al. 2020. We retain it an interesting and significant finding, proving our method holds ecological validity.

VI. Visualisation

In the following sections we would like to discuss visualisation methods taken into account.

A. Design

We are trying to represent data that expands over time and has a meaningful relationship with zero. The two designs considered for our visualisations are an Area Chart and a Normalised Area Chart (mock-ups of these are shown in Figure 6).

The idea is to have negative bias expanding to the positive side of the y-axis in the plane (i.e. the visualisation is reversed), for a merely aesthetic purpose.

All the visualisation considered are displayed on a publicly available Tableau Public page, as presented by Libardi and Mättas 2021.

1) Run Chart: This kind of visualisation is often used to represent data that evolves over time. Each data point is plotted and then connected with a line: this makes it form a line plot. This type of visualisation would suit our needs in principle. From the more aesthetic perspective, we thought it would be more appealing to highlight the area under the curve. This also emphasises the magnitude of the bias contribution of each word.

2) Area Chart: An area chart is basically the same as the previously mentioned Run Chart. On both area charts and line graphs, data points are plotted and then connected by a line to show their value at each point in time. Area charts are not quite the same as line charts as the zone between zero (on the x-axis) and the line is filled in with shading or color. Area charts can be stacked or not stacked. In the stacked version variables value is stacked one upon the other, so

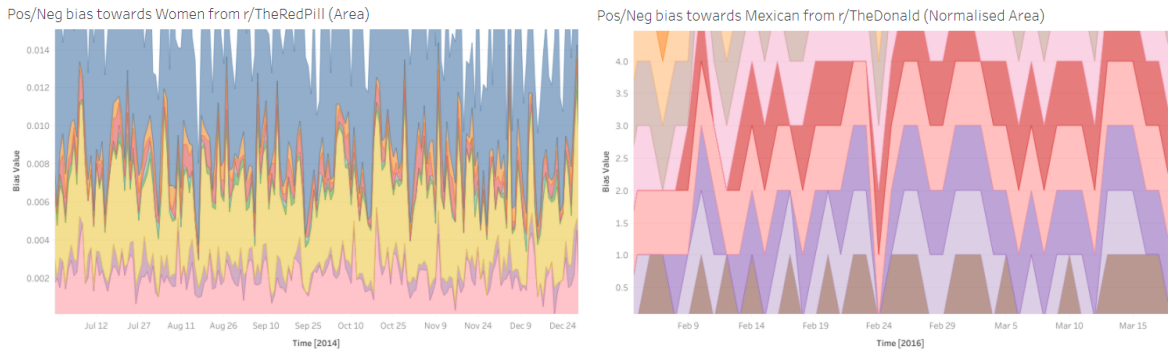

Figure 6: The visualisation of the two considered designs. On the left the Stacked Area Chart and on the right the Normalised version. These graph were made with preliminary data so they do not show valid results.

as to make it more clear what word has had the most contribution at each time step with respect to all the other words. In the not stacked version, the areas are visualised one over the other, like in Figure 7.

3) Normalised Area Chart: A variation of the area chart is the percent (or normalised) stacked area chart. It is essentially the same but data points are normalised at each time stamp. That allows the viewer to study the percentage of each variable in relation to the whole more efficiently.

B. Implementation

We will be using Tableau Desktop to visualise the data as described in *Tableau Desktop* 2021. Additionally, we will export it in a dashboard on the provided public cloud storage solution as explained in *Tableau Public* 2021.

The Tableau visualisation software eased the making of the final implementation. No preparation of the data was needed on the Tableau side. Before feeding it to the model, only preprocessing of the corpus was performed.

After seeing the result of each individual visualisation, an analysis of the more insightful (while still visually pleasing) methodology has been performed as presented by Libardi and Mättas 2021.

We implemented both the run chart (for the smoothed version) and the area chart. The only visualisation that has been excluded is the normalised area chart. The main issue with this visualisation is that our normalisation method does not have common minimum and maximum of all the normalised values.

More accurately, we normalise each day's *bias score* by the number of comments posted during that day. This results in each day having its own minimum and maximum values and if we were to display this in a normalised area chart, the percentage towards which each day contributed would not add up to 100. This could create confusion in the viewer. If we were to normalise by the greatest number of comments overall, then we could plot the area chart in a normalised fashion, being sure that each data point contributes equally to form the final area, leaving no blank space.

Finally we decided not to stack the areas in order to be able to look at each of them separately as the relationship of one biased word with the other is not meaningful in the scope of our research.

VII. Results

In this section we aim to give an exact description of the phenomena we observed through our tool and also propose a plausible explanation for these (i.e. an event that might have caused such interesting behaviours). The main events that we recognised to be concurrently happening with events described in the news are depicted in Figure 7.

We have validated the tool by systematically observing evidences that support our research. This kind of empirical validation does not entail presenting qualitative and quantitative measures

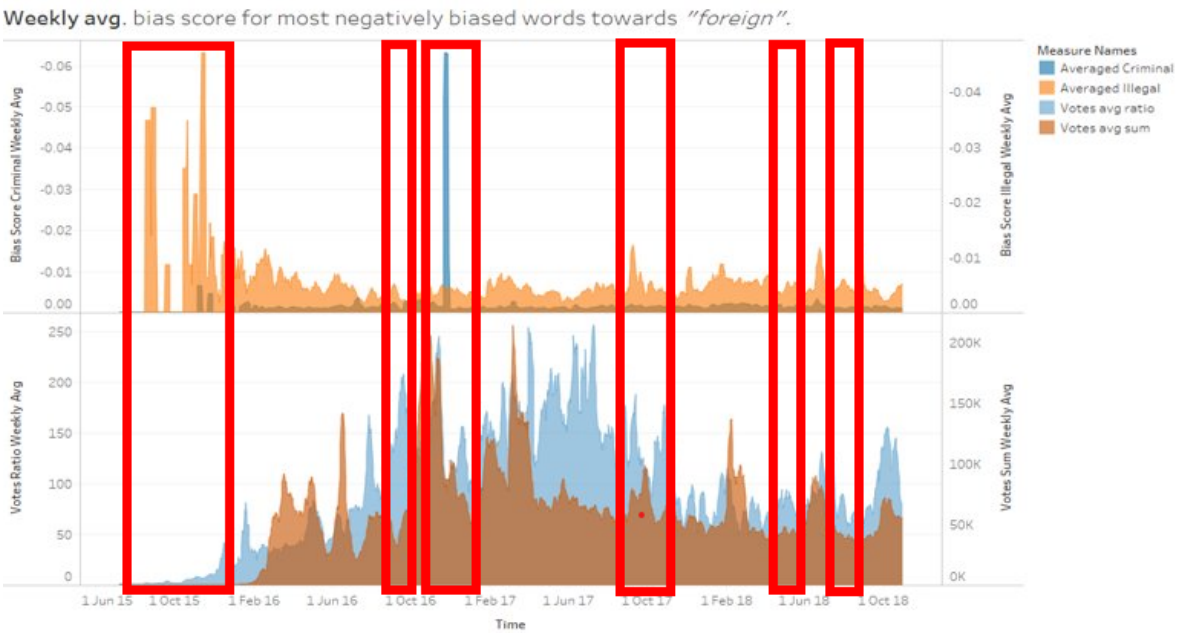

Figure 7: Main visualisation with highlighted the events we found correlation of.

with which we can evaluate our method, as it rather tries to assess the possible usage of the tool overall. Even though this might not be a scientifically exact nor strong validation method, we believe that performances can be analysed through experimentation and methodical perception (for example the aggregation of supporting examining proof) instead of hypothesis alone. We hereby assess the validity results we gathered ourselves using the tool.

As we will be discussing in section VII, we have found many relevant points in time that resulted to be interesting both on our visualisation tool and from a media point of view. The specific points in time will be discussed in more details in the next sections.

The procedure we constructed to validate the methodology proposed sees two main steps: first, isolate peculiar trend of bias score shown in our tool, and then try to match them with some real world data. In order to convey awareness on a social perspective, we have also added features like votes ratio and votes sum that have been extensively discussed in section V-D2. With the aid of these, we managed to gather more useful insights on days that spiked controversial discussions.

We have found several points that also correlated with some news and made sense both on a mere temporal perspective, but also when factoring in the controversy parameter.

Overall, we claim our tool aided us in the research of points in time where a more biased language was used and also helped us in finding the reason for the use of such language.

A. Validation

- June - March 2015: Interestingly enough we can see the trend of the bias score swinging between very high spikes and very steep downfalls. Given the way we decided to normalise the bias score (explained in section V) it shows that even in the very early days, there were very few comments with a very high frequency of biased words.
- 20 September 2016: On this day in New York, a bomb suspect has been arrested after shootout: *On Monday, police arrested Ahmad Khan Rahami — the suspect in bombings in New York City and New Jersey over the weekend — after a shootout that left Rahami and two police officers injured in Linden, New Jersey (New York Times).* This event is strictly correlated with a high spike in bias score on our visualisation -

to be more precise the rightmost screenshot in Figure 8. We can see that this did not have a peculiar behaviour for what concerns the controversy of the day, as conversely it was a very approved day as shown from the steady high value of both votes ratio and sum.

- 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. (Wikipedia)*. We can spot a peculiar trend in both bias and votes ratio/sum. It can be noted from the middle picture in Figure 8 that both the high usage of the term Criminal correlated with foreign as well as the controversy of the day (low vote ratio and score). The murder was in fact debated for a long time even after the judge sentenced the committed murderer.
- 3 - 9 September 2017: This is maybe the most interesting among all the findings. We can spot a period of time where the bias score trend has several spikes, not just a very high one, but repeated rises. When investigating that period it was evident to us what these spikes correlated with. This trend is clearly visible in the leftmost picture of Figure 8. During this period (very beginning of September 2017), the U.S. *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. (Wikipedia)*

Overall, the results proved our tool to be a valid aid when researching periods in time where language correlates with real world news and facts.

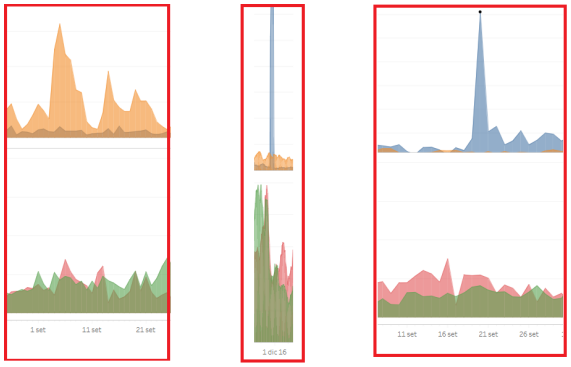

Figure 8: Three main events we presented and discussed as results.

VIII. Discussion

In this section we aim to give a clear overview of possible threats to validity that our method might entail. Additionally, we will explain some possible additions that can stem from our visualisation tool. Furthermore a conclusion will be presented as to close up the paper.

A. Effort dilution regarding data collection

Previous research failed to present a valid method to recreate the work which was needed as our starting point. We had to spend much of our effort in validating and duplicating previous research and decided to write it down ourselves. This would also leave future researchers with a reference point in how to replicate our work properly.

After many failed attempts to gather our own data set via the official Reddit API and via the Pushshift API, resulted to using an incomplete static data set found online. Additionally, there were many different issues with data processing, as already explained in section IV.

In essence, it took away quite some time from our initial focus but at the same time, gave us confidence in our work. After all, we were able to provide valuable technical insights for others coming after us.

B. Threats to validity

1) Negatively biased words used in positive context: First, it is worth talking about some words which polarity we assume to always be negative, but sometimes they might actually result to be used in a positive context.

For instance, when people are quoting other comments to criticise another one, the frequency is counted in the model but that specific comment does not really contribute towards the bias score of that single day.

2) Bias induced by the target word list: When selecting the target list, attention is to be focused towards the selection of the words. More specifically, one needs to be careful when choosing the words towards which bias needs to be identified. This is because those very same words might induce bias in turn.

For instance, if one wants to identify bias towards the concept of "American". When "gringo" is also annotated in the list of words that identify "American", then a bias is being induced. Simply, a negative word is being used to describe the target concept.

Generally, a good check could be to estimate the polarity of the words in the list and remove them if they result to be negatively charged.

3) Misinterpretation of the data: Correlation with news can often be misinterpreted. News selected for a given day needs to be correlated not only on the time of the event but also in its context.

In our specific case, a murder was correlated with the high usage of the word criminal (referred to "foreign"). This was not only the first piece of news for those days coming up on news websites, but it also correlated semantically (i.e. the word "criminal" falls within the same semantic sphere as the word "murder").

4) Weak scientific validation: The validation we applied is definitely not bulletproof. Scientifically, it actually has no real evidence of relevance. Even though they logically make sense, our results are not supported by a statistical test so we cannot claim their scientific validity. A statistical test could be assessed as a plan for future work.

C. Future research

There are many such sites with APIs available so our implementation might be duplicated to analyse other websites after Reddit too. Really, this model could be extended to any text corpus (as long as a timestamp is attached to each data point), so even looking at very large news data sets could spike interests in certain periods of time.

Additionally, extensions could be made by giving each user the possibility to see how much contribution from its own personal comments towards the bias score has been applied.

IX. Conclusion

We might claim that people use language as a tool to reflect and propagate generalisations in the form of bias. This can affect not only the individuals carrying the bias, but really anyone surrounding them. Consequently, there might be both negative and positive effects on the real life events and proceedings. Until now, it has been considered a very subjective and delicate topic. Each individual is entitled to stand for their own opinion and right for free speech, however biased or hateful this might be.

In the search for truth, it is imperative to understand how language can reflect the state of the world. And it is just as important to understand how language can be used to incite emotions that have real world consequences. First steps on this path could lead us to quantify - and if possible to visualise - language bias and connect them to actual events.

Having gone through a series of examples, we have highlighted specific trends in language bias. These are purposefully chosen to expose the emotional roller-coaster related to the US presidential election process. It is not only an evident place to look for language bias, but also a clear reflection of the American society thus far. This only adds more reason to investigate and bring clarity.

In the specific case of our research, radical supporters of a conservative candidate have made a perfect case to look into the language used. Namely, the research provides empirical evidence on the correlations of negative events sparking negative language use.

It is not yet clearly visualised how language can be used to directly call for action. At the same time, what we can see is that emotional topics spark extreme distancing between positive and negative sentiment and the people who carry them.

There are and always will be groups of people that are outnumbered by others. This makes us all effectively a minority in some way, however niche the topic might be. If we cannot pursue fairness towards all and directly aim at lowered language manipulation and bias, there is always opportunity for ill-will. People tend to gravitate towards negative topics as this is a evolutionary tactic that has helped us survive as a species. Unfortunately, there are those who seek to exploit this tactic.

In conclusion, we hope that this research has helped clarify what type of issue takes place when it comes to language bias. By creating such conversation, we hope to have shed a light on the nature of the problem, allowing us to openly think about the solutions this might require, keeping the pursuit of an ever-more-balanced society at heart.

References

[1] *Amazon EC2 On-Demand Pricing*. 2021. URL: https://aws.amazon.com/ec2/pricing/on-demand/.

[2] Jason Baumgartner et al. "The Pushshift Reddit Dataset". In: (2020), p. 10.

[3] Camiel Beukeboom. "Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies." In: Jan. 2014, pp. 313–330.

[4] Jonathan P. Chang et al. *ConvoKit: A Toolkit for the Analysis of Conversations*. 2020. arXiv: 2005.04246 [cs.CL].

[5] Kawin Ethayarajh. *Word Embedding Analogies: Understanding King - Man + Woman = Queen*. 2019. URL: https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html.

[6] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. "Towards Understanding Linear Word Analogies". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3253–3262. DOI: 10.18653/v1/P19-1315. URL: https://www.aclweb.org/anthology/P19-1315.

[7] Xavier Ferrer et al. "Discovering and Categorising Language Biases in Reddit". In: *International AAAI Conference on Web and Social Media (ICWSM 2021) (forthcoming)*. 2020. arXiv: 2008.02754 [cs.CL].

[8] Xavier Ferrer et al. *Issue: What runtime parameters are required?* Jan. 2021. URL: https://github.com/xfold/LanguageBiasesInReddit/issues/1.

[9] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (Apr. 17, 2018), E3635–E3644. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720347115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1720347115 (visited on 12/01/2020).

[10] C.J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: Jan. 2015.

[11] Filippo M. Libardi and Otto Mättas. *3 Years Bias Score From r/TheDonald*. 2021. URL: https://public.tableau.com/profile/filippol#!/vizhome/ThreeyearsrTheDonald/Dashboard1.

[12] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[13] Amy Paugh. "Acting Adult: Language Socialization, Shift, and Ideologies in Dominica, West Indies". In: Jan. 2005, pp. 1807–1820.

[14] Radim Řehůřek. *Gensim: topic modelling for humans*. 2020. URL: https://radimrehurek.com/gensim/auto_examples/index.html.

[15] *Tableau Desktop*. 2021. URL: https://www.tableau.com/products/desktop.

[16] *Tableau Public*. 2021. URL: https://www.tableau.com/products/public.

[17] Hamed Taherdoost. "Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research". In: *International Journal of Academic Research in Management* 5 (Jan. 2016), pp. 18–27. DOI: 10.2139/ssrn.3205035.

[18] *The U.S. Census Bureau Race Classification*. 2020. URL: https://www.census.gov/topics/population/race/about.html.

[19] Philip West. "Simple random sampling of individual items in the absence of a sampling frame that lists the individuals". In: *New Zealand Journal of Forestry Science* 46 (Dec. 2016). DOI: 10.1186/s40490-016-0071-1.