
INFOMMDI Assignment

VISUALIZING LANGUAGE BIASES OVER TIME USING REDDIT COMMENTS DATA.

September 2020

Authors:

Otto Mättas,

Filippo M. Libardi

Supervisor:

Alexandru Telea

I. INTRODUCTION

In this paper we present a visualisation system that displays linguistic bias trends over time.

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

Hence a meaning of etymological conduct that outcomes 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’.

The work heavily relies on Ferrer et al. (2020). We intend to take the model there presented and display its results over a time-span. The authors *present a data-driven approach using word embeddings to discover and categorise language biases on the discussion platform Reddit.*

Reddit is a social media that is organised in clusters of members (i.e. subreddits). These can be seen as communities of people interested in the same topic. One main characteristic of Reddit is that is mostly a discussion based platform.

The aim of our research is to train the model over a few and specific subreddit’s comments, try to quantify the bias as a function of time and visualise its evolution. We assume that given such a visualisation, one could more easily spot and understand the rises and falls of bias as a consequence of real world events. More specifically, an online community’s moderator could better understand why linguistic bias raised in certain point in times by correlate it with events that took place within the same community.

From this it stems our research question:

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

In [section II](#) we present the mentioned paper as well as many papers entailing the field of language bias’ identification and discovery. We

will draw a line connecting each work to this project’s very scope.

In [section III](#) we aim to provide an overview of our approach and describe both the model utilised in this project and our own addition to the latter.

In [section IV](#) we propose and classify different visualisation methodologies that we could use for the final data visualisation. Following this, we will carry a discussion on the proposed visualisation methods and their pros and cons.

Furthermore in [section V](#) we describe the data’s type and how we process it in order to extrapolate the needed features. This section will also look at the usage of possible APIs to gather data on-the-go.

In [section VI](#) we will describe the validation method adopted and how we intend to apply it. The validation will entail collecting data about the usage of this tool and essentially prove its usability.

In [section VII](#), [section VIII](#) and [section IX](#) we will analyse the results obtained and discuss possible improvements of the tool.

Lastly in [section X](#) we will present a work plan.

II. RELATED WORK

This research project will be heavily based on a recent research Ferrer et al. (2020). Here authors make use of a word embedding model to estimate words embeddings and their related bias.

III. METHODOLOGY

In this section we aim to extensively describe our approach on 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 in [subsection III-A](#) we will describe how the previous method formalises the notion of language bias and we will try to deliver the intuition behind the authors choices. In [subsection III-B](#) we detail

out our addition and explain how we thought to approach its development.

A. Quantifying language bias

The model proposed by the authors tries to capture the language bias by training a NLP model on a given corpus of reddit comments. The model learns the corpus' embeddings that transform text into high-dimensional dense vectors and capture 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. Once the model has estimated all the embeddings of a corpus, it can be given an additional list of words and it will output the most biased words out of the corpus and towards the list provided. More details about this procedure will be seen in the following sections.

1) *Formalising language bias* : Given all the calculated embeddings of a corpus (Reddit comments in our case) and two lists of target words, the model identifies the most biased words towards these lists in the corpus.

Let $S_1 = \{w_i, w_i + 1, \dots, w_i + n\}$ and $S_2 = \{w_j, w_j + 1, \dots, w_j + n\}$ be sets of target words that try to convey a concept. For example S_1 could try to identify the concept of *male* (e.g. $\{he, son, his, him, father, male\}$) whereas S_2 could try 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 w is biased towards S_1 with respect to S_2 when the cosine similarity between the embedding of w is higher for \vec{c}_1 than for \vec{c}_2 .

$$\text{Bias}(w, c_1, c_2) = \cos(\vec{w}, \vec{c}_1) - \cos(\vec{w}, \vec{c}_2) \quad (1)$$

Where $\cos(u, v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2}$. 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 end visualisation. More details about this are given in [subsubsection III-A2](#).

2) *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 [subsection III-B](#) is to use the k most biased words in order to quantify bias. The latter translates into counting occurrences of these k over time.

3) *Sentiment analysis*: When talking about bias, there is often a deception over negative bias and positive bias. As bias is essentially a human tendency to associate two concepts with one another, it doesn't 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. In more details the model implements VADER, *A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, introduced in Hutto & Gilbert (2015). The model estimates the average of the sentiment of a set of words W as such:

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

Where SA returns a value $\in [-1, 1]$ corresponding to the polarity determined by the sentiment analysis system, -1 being strongly negative and $+1$ strongly positive. As such, [Equation 3](#) always returns a value $\in [-1, 1]$. We will make use of this sentiment analysis to estimate the connotation of the bias found.

B. Adding a time dimension

The first obstacle we encountered with the model proposed is that it is hard to attach

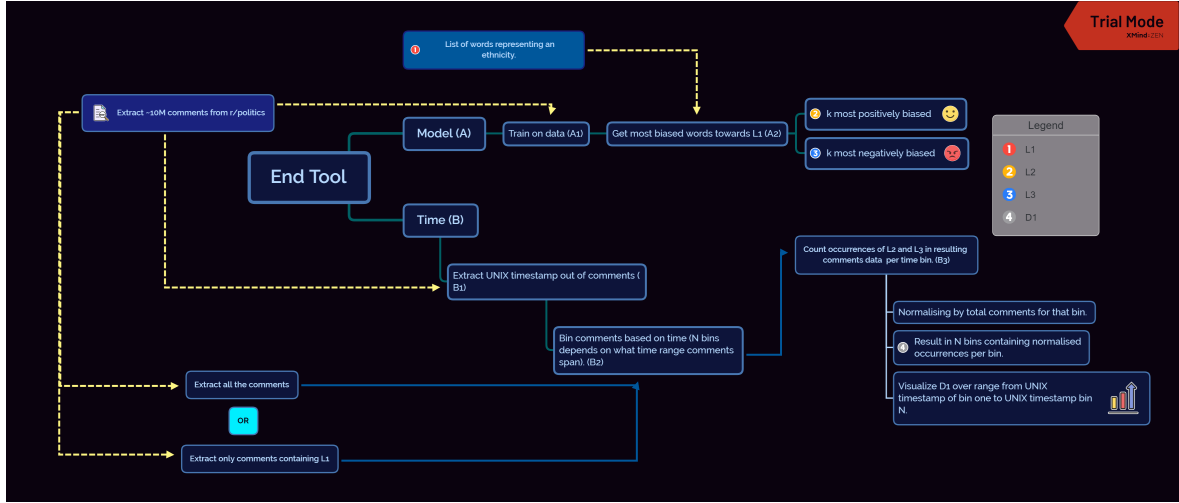


Fig. 1: An overview of the whole pipeline.

timestamps 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 line. Secondly we show what this temporal line entails and what it formally translates to. An overview of the procedure is shown in Figure 1.

Once the model has returned a list of k most biased words ($A1$ in Figure 1), we take only the ones that have a positive or negative connotation (through sentiment analysis) and that are tagged as adjectives (these are represented as $L1$ and $L2$ respectively in Figure 1), 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 polarity defined in Equation 3. Formally we define our *bias score* as follows.

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

Where C is a set of all the words present in all the comments of one day. The aggregated sentiment is normalised by the cardinality of the set (i.e. number of comments) in order not to favour very active days.

Because we are iterating through the comments in their original form, we can access the timestamp attached to them ($B1$ in Figure 1). As mentioned we bin the time on a daily basis (one bin is 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 1) 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 2. The plot shows a list of words that are the most negatively biased towards the Mexican ethnicity. There is only one word that was positively biased towards Mexicans, which is "Legal". This can be seen expanding on the positive side of the plane. Conversely, because the bias considered in this case mostly has a negative connotation (i.e. frequency of insults correlated/biased towards Mexicans), we see the line plot majorly extending on the negative side of the plane. The end tool will comprehend a visualisation of both the negative and positive bias for the words lists given to the model. Following this, we will discuss how to

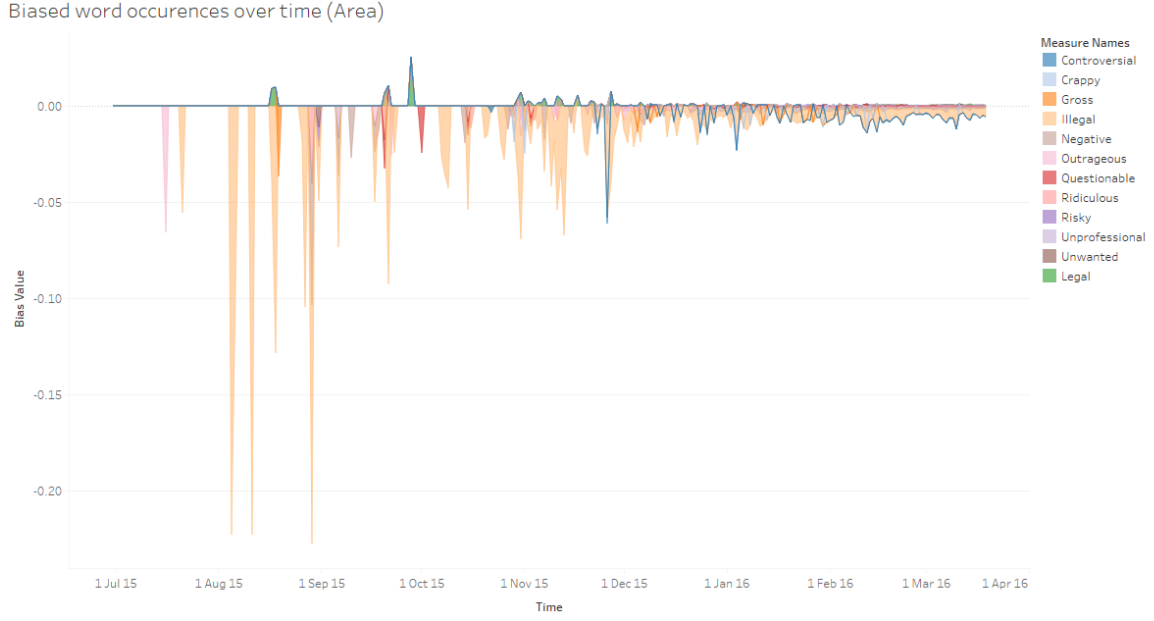


Fig. 2: *Bias Score* towards Mexican ethnicity evaluated by training the model on one million comment of the subreddit r TheDonald. The *Bias Score* is evaluated as shown in Equation 4.

visualise this in section IV.

IV. VISUALIZATION

In the following sections we would like to discuss some visualization taken into account and discussed.

A. Design

As we are trying to represent data that expands over time and that has a meaningful relationship with zero. The two designs considered for our visualisation are an Area Chart and a Normalised Area Chart (mock-ups of these are shown in Figure 3, (ref)). The idea is to have positive and negative bias expanding to both sides of the Cartesian Plane on the y axis. This goes along the intuition of negative bias being negative and positive bias being positive. Colors will more likely represent words and will be stacked one onto the other so to give an idea of which word is the "major contributor". Before stacking them up, words bias value will be sorted by area size (according to the design mocked in Figure 3, (ref)). All the visualization considered are displayed on [this Tableau Public Page](#).

1) *Run Chart*: This kind of visualisation is often use to represent data that evolves over time. Each data point is plotted and then connected with a line (this makes it a form of line plot). This visualisation would suit our needs, but on a mere visual perspective we thought it would be more appealing to highlight the area under the curve. This also emphasizes the magnitude of the bias contribution of each word.

2) *Stacked Area Chart*: An area chart is basically the same as the previously mentioned Run Chart. In both area charts and line graphs, data points are plotted and afterwards connected by a line to show the estimation of an amount at each point in time. Area charts are not quite the same as line charts as the zone between the x-axis and the line is filled in with shading or color. Additionally we aim to stack each variable on to the other so to make it more clear what word has had the most contribution at each time step with respect to all the other words.

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

efficiently.

B. Implementation

We will be using Tableau Desktop ([ref](#)) to visualise the data and export it in a dashboard on the provided cloud storage. As mentioned an implementation of each of the visualisation methods considered has already been implemented and is publicly available at [this Tableau Public Page](#). In this section, we will proceed onto describing the way we aggregate the data per word. Additionally we will also describe our conclusion and reasons to decide for one visualisation over others.

V. THE DATA

As researchers before us, we also chose for social news aggregation, web content rating, and discussion 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 of having ecological validity. In principle, we will be using open data APIs as we will investigate the comments shared publicly on internet.

A. Limitations from previous research

As we have investigated data collection methods from previous research and found it lacking, this section will see an expansion with sharing the weak backdrop.

B. Reddit API

At first, we investigated Pushshift API which was also mentioned in previous researcher Ferrer et al. (2020). As they have also referenced it from related research Baumgartner et al. (n.d.), 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 APIs.

Developed by third-party developers, Pushshift API has abstracted away from the official Reddit APIs and become more user-friendly to source linked data from with little 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 again, here 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 flow. 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 scalable infrastructure that is able to support taxing requests.

We moved on to investigating the PRAW Python framework for accessing the Reddit API. 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 give our client an appropriate user agent and we are set. Right now, we are 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.

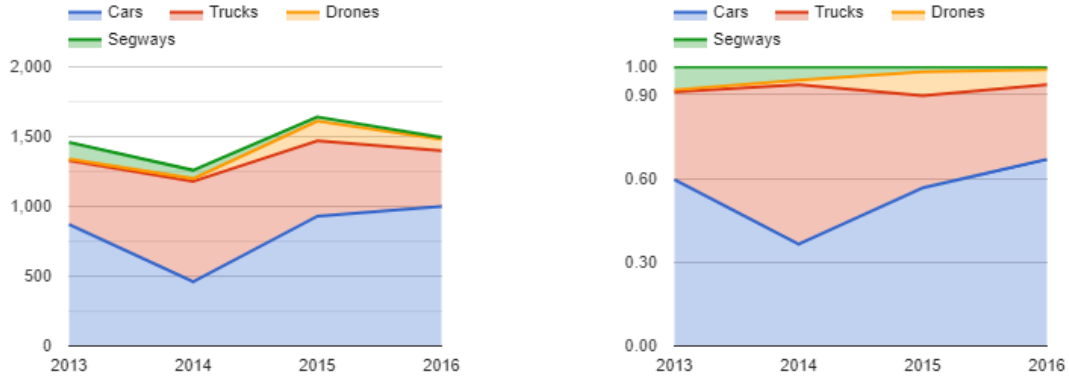


Fig. 3: The visualisation of the two considered designs. On the left the Area Chart and on the right the Normalised version. (ref).

C. Format

We are ingesting raw JSON 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 fed to the machine learning process explained in section III. The statistical model previously built plus our addition of the time dimension will output yet another CSV-formatted file, now only including the bias score (per word per bin) and time. This final file, including the comment body, timestamp and bias score will be forwarded to the visualisation process. In general, we aim to hold to web standards and follow best practices in our formatting.

D. Subreddits

Following, week 52 is planned for selecting subreddits. We are aiming to make an informed selection by working with the moderators. Overall, we are deciding for socially relevant and active communities. We are looking for ones that will be representative or with an user base, which is ideologically evenly distributed over the topics discussed.

E. API endpoints and queries

Last, week 53 is planned for harvesting data from selected subreddits. We start by using Pushshift API, additionally we might take to

the official Reddit APIs. We already found some constraints with the former, namely on real-time data processing. The need for such capacity will be found as we start with the implementation.

VI. VALIDATION

The visualisation proposed will be most likely validated by a user study. Users in this case would be moderators of selected subreddits, the new visualisation will be proposed to them and an evaluation form will try to quantify their satisfaction.

VII. RESULTS

We will develop more details about this section in week 2, according to our plan explained in section X. Overall, we aim for statistical significance to help bias research along, as well as provide clear ecological validity for our research.

VIII. DISCUSSION

After finalising analysis on results, we move on to discussing the implications. This section will be clarified on week 3 as explained in section X. Overall, we aim for sparking ideas also new to our team.

A. Limitations

B. Future research

There are many such sites with APIs available so our implementation might be duplicated to analyse other websites after Reddit too.

IX. CONCLUSION

This section will be clarified on week 3. We are planned for making summaries and finishing the report as explained in [section X](#).

X. TIMING

In this section we will provide a schedule of the project's development - of course we intend to respect it as much as possible. A more detailed overview of the work plan listed below is shown in [Figure 4](#)

- 7-13 December: Visualising the data over a simple line-plot. Familiarising with the API.
- 14-20 December: Familiarising with Tableau. Testing a bigger data-set. Researching possible visualisation design (consider feedback).
- 21-27 December: Selecting subreddits and contact moderators. Testing the selected visualisation design.
- 28-3 December/January: Harvesting data from selected subreddits and importing it into Tableau. Writing a form to send to moderators.
- 4-10 January: Working on Tableau visualisation. Sending moderators the form.
- 11-17 January: Preparing the presentation. Collecting and analysing the results.
- 18-24 January: Polishing report.
- 24-27 January: Hand in week.

Week	Date	Time	Activity	Comments	Comments	Deadline
49	04-Dec		Work on timeline	Gather data about course deadlines etc	Otto	10:30
	04-Dec		Agree on timeline	Timeline set up	Filippo/Otto	
	04-Dec	10:30	Weekly Friday Meeting	Read through previous research	Filippo/Otto	
	04-Dec		Work on baseline implementation	Read through previous research	Filippo	
50			Work on API data collection	Read through previous research	Otto	
	07-Dec	10:30	Weekly Monday Meeting	Implement previous research technical solution / API baseline	Filippo/Otto	
		11:00	Supervisor meeting	Baseline implementation / API data collection overview	Filippo/Otto	
	08-Dec	18:00	Send project draft for revision (Alex)	Gather feedback on the problem-solution combo during meeting	Filippo/Otto	
	09-Dec		Edit draft based on feedback	Carry over core improvements / Improve overall look	Filippo/Otto	
	09-Dec		Finish project draft	Confirm all the changes with final read-through	Filippo/Otto	
	09-Dec		Hand in final project draft	Finish project draft		18:00
				GOALS this week:		
				- Visualising the data over a simple line-plot.		
	11-Dec	10:30	Weekly Friday Meeting	- Familiarising with the API.		
				GOALS this week:		
				- Familiarising with Tableau.		
				- Testing a bigger data-set.		
51	14-Dec	10:30	Weekly Monday Meeting	- Researching possible visualisation design(consider feedback).		
	16-Dec	15:00	Class			
	18-Dec	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Selecting subreddits and contacting moderators.		
				- Testing the selected visualisation design.		
52	21-Dec	10:30	Weekly Monday Meeting			
	23-Dec	15:00	Class			
	25-Dec	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Harvesting data from selected subreddits and modelling.		
				- Importing results into Tableau.		
53	28-Dec	10:30	Weekly Monday Meeting	- Writing a form to send to moderators.		
	30-Dec	15:00	Class			
	01-Jan	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Working on Tableau visualisation.		
				- Sending moderators the form.		
1	04-Jan	10:30	Weekly Monday Meeting			
	06-Jan	15:00	Class			
	08-Jan	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Preparing the presentation.		
				- Collecting and analysing the results.		
2	11-Jan	10:30	Weekly Monday Meeting			
	13-Jan	15:00	Class			
	15-Jan	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Finalising the report.		
3	18-Jan	10:30	Weekly Monday Meeting			
	20-Jan	15:00	Class			
	22-Jan	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Hand-in week starts.		
				- Creating and rehearsing the presentation.		
4	25-Jan	10:30	Weekly Monday Meeting			
	27-Jan	14:00	Presentations in class	Project Presentation		
	29-Jan	10:30	Weekly Friday Meeting			
				GOALS this week:		
				- Hand-in week continues.		
5	01-Feb	10:30	Weekly Monday Meeting			
	03-Feb		Hand in assignment	Project Report Submission		?
	12-Feb		CELEBRATE			

Fig. 4: Detailed work plan.

REFERENCES

- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. (n.d.), 'The pushshift reddit dataset', p. 10.
- Beukeboom, C. (2014), *Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies.*, pp. 313–330.
- Ferrer, X., van Nuenen, T., Such, J. M. & Criado, N. (2020), Discovering and categorising language biases in reddit, *in* 'International AAAI Conference on Web and Social Media (ICWSM 2021) (forthcoming)'.
- Hutto, C. & Gilbert, E. (2015), Vader: A parsimonious rule-based model for sentiment analysis of social media text.