The Meanings of Class in Reddit Comments – An Explorative Study of Word Embeddings

Course: Computational Social Science, Lecturer: Prof. Karsten Donnay

Author: Jonas Schwenke (904634) – Universität Konstanz

Abstract: This paper explores word embeddings as a tool for culturomics. It is largely based on the paper '*The Geometry of Culture - Analyzing the Meanings of Class through Word Embeddings*' by Kozlowski et al. (2019) where vector representations of words are used to study the interplay of several class dimensions in digitalized written books (*Google Ngrams*) throughout the 20th century. Because of indications of bias towards scientific publications in the *Ngrams* corpus, I consult comments on the news aggregation platform Reddit for comparison. Analysis of the individual subreddits *science*, *AskReddit* and *worldnews* reveals only few differences to the *Ngrams* corpus when angles between vector representations of class dimensions are compared. Additionally, these differences do not support the bias hypothesis. Apart from this comparison, the paper aims at carefully documenting difficulties in preprocessing, evaluation and interpretation[1].

1. Introduction

Recently, Kozlowski et al. (2019) published a paper on the meanings of class in word embeddings. In this method, a corpus of text is transformed into a dictionary of distinct words, each represented as a vector in n-dimensional space. The goal of embeddings is to mathematically express the proximity of words as perceived by humans. This proximity is measured as the cosine similarity between vectors. Employing the *word2vec* embedding on the Google Ngrams corpus ranging back to 1900, Kozlowski and colleagues analyze different dimensions of the concept of class throughout the decades.

Reddit, which was founded in 2005, is a forum-like website with many different topic-specific subforums called *subreddits*. The page resides in the Top 21 of the most visited websites[2] and generated a total of 1.7B comments in 2019[3]. Roughly four years ago, user *u/Stuck_In_The_Matrix*

[1] Code available here: https://github.com/JonasSchwenke/RedditEmbeddingAnalysis

[2] https://www.alexa.com/topsites (Accessed on 28.04.2020)

[3] https://redditblog.com/2019/12/04/reddits-2019-year-in-review/

released the corpus of all publicly available Reddit comments[4] for research. After its release, scientists from a diverse set of fields have utilized this dataset (Singer, et al., 2014; Singer, et al., 2016; Soliman, et al., 2019). To offer more insight into the background of this work, I will start with summarizing Kozlowski and colleagues' research in more detail, pointing out which parts can and will be recreated in this paper.

1.1 Kozlowski et al. 2019

The study of semantic meaning among different class dimensions has been based on almost non-reproducible qualitative methods or limited topic modeling. With word embeddings – numerical representations of words in a vector space – Kozlowski et al. (2019) hope to gain fine-grained, reproducible results. The dimensions they seek to observe are derived from a range of classical and contemporary sociological publications.

In previous work the authors find contrasting ideas about the trajectory of class in the 20th century: some argue that non-economic dimensions have gained dominance over wealth and occupation in forming social structures, some deny this and some give most credit to *Cultivation* and *Status*. This ambiguity leads them to expect a mostly stable perception of class thoughout the 20th century.

The class dimensions are constructed with antonym word pairs. The intuition behind this is that a word vector orthogonally projected onto a vector reaching from e.g. *poor* to *rich* will lean towards the end that it mostly appears with in the corpora. For more robust results, the dimension vectors consist of not only one antonym pair but the average of several pairs. The antonyms are collected with help of contemporary and historical thesauri.

The validation of the word embedding is done in three different ways: First, a list of 59 terms from seven different categories are rated among the dimensions of *Affluence*, *Gender* and *Race* by online survey respondents.[5] The Pearson correlation coefficients of the resulting means with values from projections on three different word embeddings (*Google Ngrams word2vec*, *Google News word2vec and Common Crawl GloVe*)[6] show best values for gender (0.76, 0.88, 0.90), average for class (0.53, 0.58, 0.57) and worst for race (0.27, 0.75, 0.44). Second, historical validation is conducted by recreating a semantic differential survey from the 1950s with a Google Ngrams embedding from the same decade with all correlations being positive and statistically significant. Third, the decade-wise binning of Ngrams embeddings is repeated on sociological texts from *JSTOR* – according to rough qualitative analysis the results correspond with disciplinary trends. In my work, I will only consider the first validation method, since the other two are based on data from the 20th century.

Kozlowski and colleagues find mostly stable relations between the class dimensions. At the beginning of the century, *Status* and *Cultivation* yield the highest (positive) cosine similarity with

[4] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

[5] E.g. for *class* the words are rated from 0 = very upper-class to 100 = very working-class

[6] This Google Ngrams embedding constructed by Kozlowski et al. ranges from 2000 – 2012 (2012 is the latest date in the Google Ngram corpus). The other two embeddings are pre-trained and publicly available.

Affluence, followed by *Morality* and *Education* (positive), *Gender* (negative) and *Employment* (positive). They attribute the slightly negative relation between *Gender* and *Affluence* – which indicates a stronger association with femininity – to the instrumentalisation of women as displays of wealth. This theory is backed up by the words that score closest to the *Affluence* dimension: *fragrance*, *perfume* and *jewels*.

The most notable change during the 20th century is the rise of *Education*. Starting out on fourth position, it dominates the other dimensions in similarity to *Affluence* at the end of the century. This relationship holds true even when controlled for *Cultivation*, a suspected mediator of the positive link between the two. Considering all other dimensions and their respective projections, Kozlowski et al. note higher similarity among *Morality*, *Education* and *Cultivation* and lower values among *Status*, *Gender* and *Employment*. Notably, *Employment* yields a negative similarity with *Morality*, indicating a rather bad moral image of bosses and managers. In the next section I will describe characteristics of the Google Ngrams corpus.

1.2 The Google Ngrams Corpus

Kozlowski and colleagues admit to the caveats of their study: the version 2 of the Google Ngrams corpus is a sample of all books published (ca. 6%) and thus contains the output of a 'literary elite'. More precisely, the books were selected from over 40 university libraries (Michel, et al., 2011). Presumably, the sample is thus biased towards academic literature, underrepresenting prose. Unfortunately, the sampling procedure behind the Google Books Ngram Viewer (GBNV) is not made very transparent by the authors. Nevertheless, evidence for academic oversampling is provided by comparing the frequencies of the 1-grams *Figure* and *figures*, the first assumingly being more frequent in scientific publications (as in 'Figure 1: ...'). Looking at version 2 of both the English and the English Fiction corpora, the sharp rise of *Figure* and vast dominance at the end of the 20th century indicate academic oversampling (Pechenick, et al., 2015).

Another notable critique of the Ngram corpus is the neglect of popularity. Phrases from some obscure scientific paper are weighted the same as phrases from 'Harry Potter'. Arguably, the

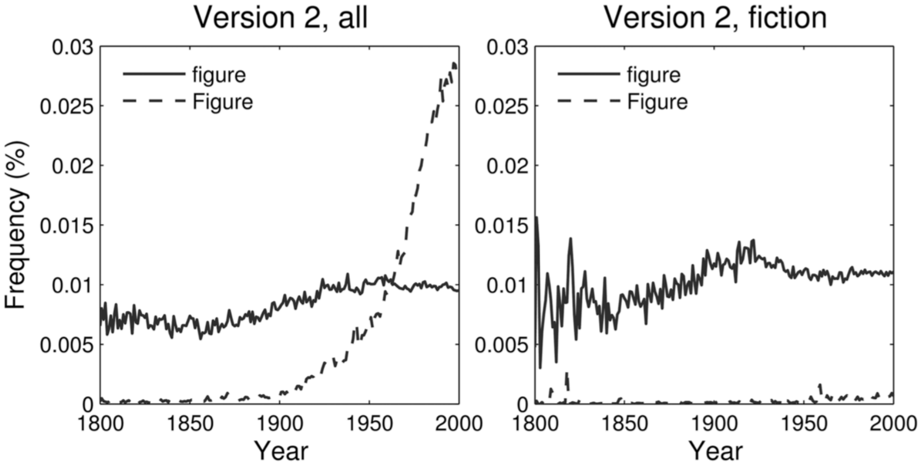

Figure 1. GBNV Frequencies of *Figure* and *figure* in Version 2 of the English and English Fiction Corpora (Pechenick, et al., 2015)

relationship of individual words of a bestseller has a bigger impact on culture and its collective understanding than the ones read only by a niche population. While a popularity weighting would be desirable when taking the Ngrams as a proxy for the whole population, in practice this seems unfeasible. After all, Kozlowski and colleagues understand this corpus as the output from a literary elite. However, even though Kozlowski et al. (2019) validated their word embeddings with a contemporary survey, including post-stratification weighting, further validation of and broad comparison with the Google Ngram corpus can be helpful to determine its reliability as a tool in culturomics.

1.3 The Reddit Corpus

If the Ngrams corpus reflects the perception of class from a 'literary elite' point of view, to which degree can the Reddit user base be considered as a proxy for a more casual authorship that reflects the point of view of the general public? To answer this question, three aspects must be considered: How does the internet as a medium compare to books? How might the characteristics of Reddit as a medium shape their users' behaviour? Which demography is represented by redditors?

1.3.1 The Medium Internet

One major difference between the internet and books as sources of authorship is anonymity. The equalization hypothesis of computer-mediated communication (CMC) expects anonymity to erase the power inequality that builds on top of social cues such as gender, race, age and others (Christopherson, 2007). Consequently, individuals might rather state their opinion in an anonymous environment, especially when this opinion is marginalized or sanctioned in non-anonymous environments. The cathartic effect of anonymity on the internet can optimally lead to people from minority groups openly discussing their struggles. This holds true as long as no censorships mechanisms are in place. Anonymity can also help to focus online contributions on the goal of the community. Additionally, there are gender differences, as men rather tend to dissolving anonymity to gain back the power differential that is common in face-to-face communication (Christopherson, 2007).

Apart from participation itself, the language used on the internet will vastly vary from language in books. Internet language is more efficient, uses creative spelling, abbreviations and is usually less complex with an increased readability compared to traditional forms of writing (Herring, 2002). Of course, these characteristics depend on the mode of communication, which can be synchronous (chat, texting) or asynchronous (e-mail, forums). Users of more synchronous modes of CMC use fewer words and shorter, simpler sentences (Herring, 2002).

Any comparison to communication via books is not very straight forward. An author of prose can let a character talk without having to fear his audience connecting it to them personally. That way, technical anonymity can be achieved in books as well. Regarding language, complexity and vocabulary richness must be expected to be higher in traditional writing, highly dependent on the part of the internet that is being studied.

1.3.2 Characteristics and Demography of Reddit

Because of their mostly casual authorship, comments on the social news aggregation platform Reddit might serve as an interesting corpus to validate the nature of Kozlowski and colleagues' class dimensions. The site was founded in 2005 and counts over 430M active users[7], mostly from the US (ca. 40%), followed by the UK and Canada (both below 10%)[8]. Users (usually called *redditors*) can post and comment on posts and comments in over 130K active *subreddits* – subversions of Reddit related to one specific topic. The subreddits with the most subscribers now are about quick entertainment (*funny*, *aww*, *pics*), interest (*gaming*, *science*, *worldnews*) and conversation/Q&A, sometimes with a philosophical touch (*AskReddit*, *todayilearned*, *ShowerThoughts*)[9]. While initially Reddit was designed as 'the front page of the internet' where users post links to external sites, it has experienced a sharp rise of self-referential content and a vast diversification of subreddits (Singer, et al., 2014). Unlike on Twitter, comments and posts on Reddit can highly vary in length with the current character limit set to 40.000[10]. Users can gain *karma* – points received or lost when other users upvote or downvote your contribution. These come either as a submission to a subreddit or comments on this submission and other comments. Consequently, each comment has parent and child comments and threads can go virtually infinitely deep. The most popular submissions of each subreddit are fed to each user's customized front page (or news feed) and appear first on each subreddit's page. More popular posts are thus more likely to receive comments while a large fraction of submissions are ignored. Shorter, easy to read titles are preferred. Additionally, the majority of redditors prefers passive browsing and only 16% of users generate 50% of interactions (*clicks*, *votes*, *pageloads*) (Medvedev, et al., 2018)[11]. Reddit's moderation system keeps a tight grip on what can be submitted and commented. The exact rules usually vary between the subreddits but the overall progressive ideology is enforced globally, resulting in bans of whole subreddits[12].

Two thirds of redditors are male and 64% are between 18 and 29 years old. Only 18% have no college experience – compared to 41% of US population. Accordingly, Reddit can be regarded as left-leaning, with 43% having a liberal and only 19% having a conservative political ideology[13]. Considering these demographics, the Reddit corpus might not differ from the Ngrams too much. Of course, the written comments will assumingly use a simple, everyday language but the average ideology behind it will be that of a well-educated, politically progressive male. Similar characteristics must be expected from the Ngram authorship: A large share consists of scientific

[7] https://www.redditinc.com (Accessed on 17.03.20)

[8] https://www.alexa.com/siteinfo/reddit.com (Accessed on 17.03.20)

[9] http://redditlist.com/

[10] https://www.reddit.com/r/changelog/comments/39hf9x/reddit_change_selfpost_character_limit_is_now/

[11] Because of time and space constraints, I will refrain from discussing what factors influence online participation and how these could be connected to demography and outcomes

[12] https://en.wikipedia.org/wiki/Controversial_Reddit_communities

[13] https://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

publications which are also dominated by men (Allagnat, et al., 2017). In order to find support for the hypothesis that the Ngrams corpus has a scientific bias, a comparison between subreddits can be helpful. The perception of class dimensions in the *science* subreddit must thus be expected to show more similarity to Ngrams results than other, less scientific subreddits. Even though there is no demographic information about individual subreddits, the premise is that more casual, everyday life-oriented subreddits have a user base closer to the general public. One possible difference in these groups might be less similarity of the *Affluence* dimension with *Education*, since more affluent, educated persons were found to be more likely to acknowledge this relationship (Manstead, 2018). More hypotheses about the difference of all dimensions' complex interplay would be highly speculative – at this point it will be better to yield results and then continue with interpretations.

2. Data

Since their API only allows for 1000 items per view, Jason Baumgartner's (aka *u/Stuck_In_The_Matrix)* data dump of all comments in 2015 and subsequent updating (up to 09.2019) allows for relatively easy access of historical data. Even though large-scale missing data has been reported on these datasets, the risk for any machine learning tasks is reported to be rather small (Gaffney & Matias, 2018).

2.1 Preprocessing

Reddit comments demand several different steps of preprocessing before being used for word embeddings. An inspection of a random sample from all complete years (2006-2018) revealed no notable occurrence of non-english comments. Consequently, non-english comments were not removed to save time and ressources. In a more drastic approach than other preprocessing on this corpus (Kim, et al., 2018; Singer, et al., 2016; Soliman, et al., 2019) I decided to remove all duplicate comments. Doing so I hope to erase comments that are (i) written by moderators who will not be removed in the steps described below (ii) are spam (iii) or *memes* or *copypastas* – repeated inside jokes. Additionally, all comments shorter than five tokens are removed. With this step, all deleted comments are removed[14] and it puts the Reddit corpus on-par with the 5-grams used by (Kozlowski, et al., 2019). Next, comments from deleted accounts are removed. This step is debatable, since the data of interest - the comment itself - is not missing. These authors might have been regular users but also bots, trolls or spam accounts and individual comments can not be traced back to either category. A future closer analysis of deleted accounts might shed more light on this issue.

One major source of noise stems from *bots* which are created both by Reddit moderation teams and by regular users. Bots comment automatically based on different events. Possible triggers are users breaking Reddit/subreddit rules or specific user interaction (e.g. *calling* a bot to state statistics). To

[14] which are indicated by the string `[deleted]`

avoid content from bots and moderators I remove all comments whose authors appear in an unofficial Reddit bot list[15,16] (Singer, et al., 2016; Soliman, et al., 2019). Since this list has not been updated in the last five years, I supplement it by removing all authors who match the strings bot, b0t and moderator.

After all preprocessing steps ca. 70% of comments remain. The word output of individual accounts roughly follows a power law with an extremely long tail - largely because of undetected bots. An inspection of the busiest accounts shows: only two out of the top ten authors with the most tokens in the sample are not bots. These eight bots (out of ca. 2.2M users) account for 0.28% of all tokens in the preprocessed sample and the top 0.01% of authors provide 1.28% of word output. Considering this I decide to handpick the top 50 authors for bots, moderators and semi-automated accounts and remove their output from the sample. 25 bots and moderators with a total of 1.4M words are removed, lowering the standard deviation from 502 to 419.

This method is far from perfect and while there exist several bot detection frameworks for Twitter (Varol, et al., 2017; Chu, et al., 2010), I could not find equivalent (academic) work for Reddit. There are approaches and code out there but they are either limited to specific subreddits or their application would go beyond the scope of this work.[17] In terms of efficiency, removal of bots and moderators might not be worth the effort after all. One reason is that moderators can comment off-duty and are often important contributors to their respective subreddit. Another reason is that they change daily and the harm to the reliability of results is hard to measure and highly depends on the task at hand. The risk here is that moderation and bot comments might dilute the content of interest, especially when only sampling is possible. Knowing about the percentage of irrelevant comments can help to choose appropriate sample sizes.

To prepare the comments for the embedding function, all links and line breaks are removed. They are then treated with the preprocessing function provided by the `gensim`[18] library, which includes removal of punctuation and numbers, lowercasing and tokenizing of words with an upper limit of 15 characters.

2.2 Discriptive Statistics

The full available corpus (12.2005 – 09.2019) contains over 5 billion comments. Starting out with just above 400K comments in the first full year 2006, the volume started to grow significantly in 2010, reaching around 100M each month in 2018. To study the corpus in respect of time and memory constraints, I draw a simple random sample of 15M comments across all complete years (2006-2018). After preprocessing, about 10.4M comments with a total of almost 342M tokens are left. There are circa 2.2M different authors who left an average of four comments. The average

[15] case-insensitive

[16] https://www.reddit.com/r/autowikibot/wiki/redditbots

[17] One possible idea is to remove comments that appeared below a certain time threshold after their parent comment – indicating an automated response. A future analysis of this method would be desirable.

[18] https://radimrehurek.com/gensim/index.html

comment has 34 tokens and the total output of individual authors follows a power law (as described above). The average number of words from one author is 158 but 78% stay below this number. Almost a third of authors produced less than 25 words in the sample. Comments reach lengths up to 3000 tokens but ~ 50% have 17 or fewer tokens – again a power law distribution. The same is true for the distribution of words across the 51K subreddits that were included in the sample: The top 10 subreddits account for 13% of all tokens. Table 1 lists the top 10 subreddits by percentage of words in the sample and shows their subscriber rank (data from 04.2020). Of course, there were changes in subscriber rank and not all subreddits have been around for the whole time, but there is indication that some topics draw deeper discussions than others. Especially sports related subreddits (*nfl*, *nba*, *CFB*) and politics (*politics*, *The_Donald*) yield much more tokens than their (current) subscriber rank would suppose.

Sample Rank	**Subreddit**	**% in Sample**	**Subscriber Rank**[19]
1	AskReddit	5.17	2
2	politics	1.98	55
3	worldnews	0.98	8
4	nfl	0.92	126
5	nba	0.76	70
6	news	0.75	13
7	CFB	0.65	403
8	unpopularopinion	0.65	276
9	The_Donald	0.64	376
10	funny	0.60	1

Table 1. Rank, Percentages and Subscriber Rank (excluding *r/announcements*) of the Top 10 Subreddits in Sample; Measured in Number of Tokens After Preprocessing (N=342M)

3. Method

In this section I will describe the method of word embeddings and go over hyperparameters of the *word2vec* implementation. After that, the results of different evaluation approaches are reported.

3.1 Word embeddings

Word embeddings are distributed, low-dimensional, real-valued representations of words of a given corpus (Turian, et al., 2010). Even though common word embeddings reach up to 300 dimensions, they are often considered low-dimensional since previous models based on word co-occurrence were significantly larger. Word embeddings can be used for a variety of natural language processing tasks; they are usually the first step before task-specific methods are applied

[19] http://redditlist.com (Accessed 23.04.20)

(Collobert, et al., 2011). Generally, word vector representations are constructed in a shallow neural network by updating the weights of the hidden layer. The weights, which are randomly initialized (= number of dimensions) are adjusted to minimize the error of predictions for each word. Words are fed by a stream of tokenized sentences. Kozlowski et al. (2019) apply the *word2vec* skip-gram architecture as presented by Mikolov et al. (2013). This model does not classify the current word but its surrounding words in a window of a given size. The classification task is conducted by training the weights to distinguish true word pairs from corrupted ones (negative sampling). The more distant a word, the less it is sampled during training. Increasing the window size results in higher quality embeddings. Trained on 783M words, this model was shown to perform best on a semantic-syntactic relationship test set. Because of its lower computational complexity *word2vec* trains considerably faster than previous models and is thus suitable for tasks were many embeddings are compared (Mikolov, et al., 2013; Kozlowski, et al., 2019). After *word2vec,* faster and better models have been published but for reasons of comparison with Kozlowski et al. (2019) and its relatively simple use I will focus on this approach.

In culturomics, word embeddings can be utilized to measure similarity between words of a given corpus, measured in cosine distance. However, because words are mapped based on their context words, two semantically distant words (e.g. `good` and `bad`) may appear close together if they have a similar syntactic function and thus often share similar context (Tang, et al., 2014; Kozlowski, et al., 2019). In their approach, Kozlowski and colleagues extended the idea of similarity between words to that of word dimensions. To that end, they made use of the geometric characteristics of word vectors that allow for simple algebraic operations: A dimension is built by averaging over the distance between normalized vectors representing antonym word pairs. A word dimension thus has the same number of vector dimensions as the individual words, which depend on the hyperparameter choice. The word pairs for all seven dimensions as well as the code for the construction of the dimensions are provided by the authors[20].

3.2 Hyperparameters

Various hyperparameters of the word2vec function can be tweaked for optimal results. For comparability, I decided to mainly go with the settings used by Kozlowski et al. (2019): They use the Skip-gram architecture with 300 dimensions, window size at 5 and negative sampling at 8. The Skip-gram mode as opposed to the Continous Bag-of-Words (CBOW) method was found to be overall better performing by *word2vec*'s inventors. While the number of dimensions can be increased to gain even better results, 300 was found to yield the best balance between cost and result (Mikolov, et al., 2013). All of my embeddings are trained with 5 iterations[21]. For a comparison between input sizes I set the minimum count parameter on 2 to gain a broad overview of vocabulary size while filtering out completely random utterances. In general practice however,

[20] https://github.com/KnowledgeLab/GeometryofCulture - includes word pairs for the three survey-relevant dimensions. Other word pairs are in the appendix of Kozlowski et al. (2019).

[21] It is unclear how many iterations were used by Kozlowski et al. (2019) since they didn't specify this parameter in their code and default values have changed between versions of `gensim`.

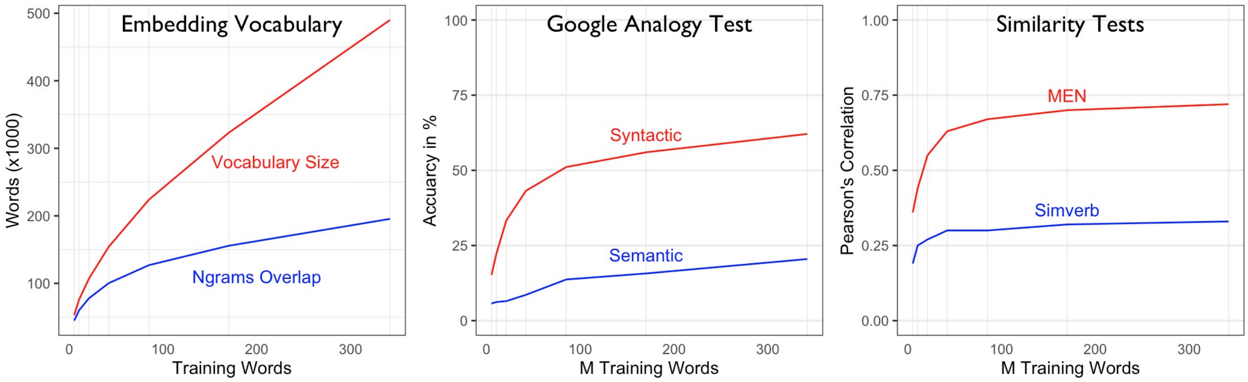

Figure 2. Several Evaluation Results of Embeddings with Increasing Number Of Input Words (5.25M - 10.5M - 21M - 42M - 85M - 171M - 342M). Random Samples from all of Reddit (complete years: 2006-2018).

this has several downsides: The model gets very big and words with very few appearances will most likely not yield meaningful vectors. Of course, the model can be trimmed after its completion. For the comparison between subreddits I set the minimum count on 10. In the next section I describe the evaluation of embeddings.

3.3 Evaluation

While Kozlowski and colleagues produced embeddings from the whole available corpus, this was not possible in my work. Therefore, an appropriate sample size must be found. There are many ways to evaluate the quality of an embedding. The evaluation can be either extrinsic or intrinsic. Extrinsic methods base the goodness on performance on downstream NLP tasks such as Named Entity Recognition or Semantic Role Labeling. Intrinsic methods compare the embedding with human-rated test sets on word relations (Bakarov, 2018).

To evaluate the Reddit embeddings, I first choose three different, easily available tests: The Google semantic-syntactic analogies provided by the *word2vec* creators (Mikolov, et al., 2013), the SimVerb-3005 verb similarity test (Gerz, et al., 2016) and the MEN similarity test, which includes human-annotated similarities of 3000 nouns (Bruni, et al., 2014).

To determine the overall Reddit performance and find indications for good sample sizes, I train word vectors on the biggest all-of-Reddit sample (342M words) and six more sizes, each roughly half the size of its bigger neighbor (171M, 85M, 42M, 21M, 10.5M, 5.25M). The first plot of figure 2 shows how the vocabulary grows almost linearly to close to 500K words in the biggest sample, while the growth of overlap with the Ngrams vocabulary noticeably declines. The same is true for the Google syntactic analogy test, where the samples of 85M, 117M and 342M score above 50%[22]. The semantic section, which includes currencies and capitals of rather unknown countries,

[22] Tests were run with exclusion of out-of-vocabulary (oov) words. Scores including oov words would assumingly be much lower.

Dimension	science	AskReddit	worldnews
Affluence (42)	85	91	88
Gender (10)	100	100	100
Race (7)	100	100	100
Education (9)	83	89	89
Employment (20)	98	100	100
Cultivation (23)	70	76	72
Morality (27)	96	96	98
Status (15)	87	93	87

Table 2. Percentages of Missing Words Used for Construction of Dimensions Across Subreddits. Parantheses Display Total Number of Antonym Pairs.

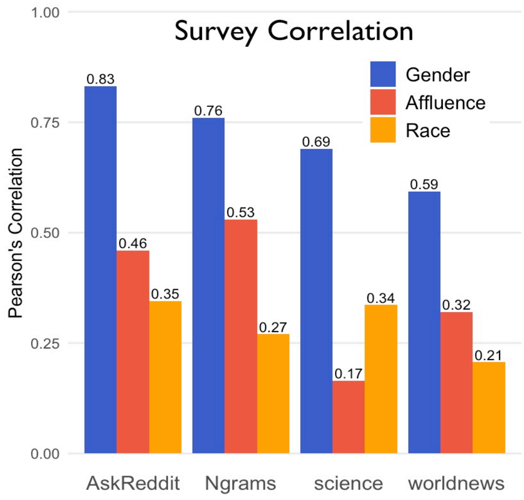

Figure 3. Pearson Correlation Coefficients of Ngrams and Three Subreddits with Survey Similarity Scores for 59 Terms Rated Across Three dimensions.

experiences a more steady growth but stays below 25%. Both similarity tests experience a sharp rise in Pearson's correlation coefficient and don't gain much beyond 42M input words. While these tests give some insight into the general performance of different input sizes, they are known to have issues (Drozd, et al., 2016) and can only be considered as exploratory. Best evaluation practice is to validate embeddings according to the task at hand.

First thing to consider is the number of missing words. As Kozlowski and colleagues point out: the more antonym word pairs are used to construct the class dimensions, the higher the correlation with the survey scores. This relationship follows Gossen's first law – adding more word pairs results in decreasing improvement of correlation coefficients. Consequently, dimensions which are built from fewer word pairs will suffer more from missing words. While Kozlowski reported one word missing in the Ngrams (2000-2012) vocabulary, the numbers are higher among Reddit data. Apart from *Gender* (10 word pairs) and *Race* (7), all dimensions have missings in the 342M-word embedding.

To support the Ngrams scientific-bias hypothesis, I produce word vectors for three different subreddits: *science*, *AskReddit* and *worldnews*. The *science* embedding is created from all its available comments, maxing out at ~250M words after preprocessing; the other two are sampled accordingly to control for word input size. Table 2 shows the percentages of present dimension words for the three subreddits. Interestingly, despite its larger vocabulary (116K words vs. *AskReddit*: 95K and *worldnews*: 84K) *science* has the most missing antonym words. One possible reason is that this subreddit is focused on the natural sciences. Only 4 out of 22 submission topic tags (*Psychology*, *Social Sciences*, *Economics* and *Anthropology*)[23] are somewhat related to *social class*. In the next step, the embeddings are compared to the survey scores from Kozlowski et al. (2019), described in section 1.1. The embedding scores are generated by multiplying the word vectors of the max. 59 survey terms with the three individual dimension vectors, generating the cosine similarity because vectors are normalized[24]. Figure 3 shows the Pearson correlation coefficients for the three subreddits and Ngrams for the dimensions rated by the annotators. Here the dimensions were constructed using only words that were present in all three subreddits. Missing survey words were also excluded from all three subreddits (*shanice*, *aaliyah*). Evidently, *AskReddit* has an overall higher similarity with the Amazon Mechanical Turk sample[25] than the other subreddits, with even higher scores than Kozlowski and colleagues' Ngrams on *Gender* and *Race*. A reason for the overall low scores of *worldnews* could be that here more international redditors take part in the discussions while the survey was only listed for people in the US. Additionally, both *AskReddit* and *worldnews* could be subject to sampling error; a future study should include bootstrapping confidence intervals.

4. Results

For further analysis, Kozlowski et al. (2019) dropped the *Race* dimension because of its low correlation with the survey similarity scores (.27). However, using this threshold value I would also have to drop the *Affluence* dimension, since here the subreddit *science* scored only .17. Also, there were no survey results for the other five dimensions which might have scored even lower. Considering this, I will use all original eight dimensions for the angle comparison. Figure 4 shows the heatmaps of cosine similarities between all eight dimensions for *Ngrams* (2000-2012) and the three subreddits. Positive values (red) indicate that e.g. higher *Affluence* is associated with higher *Cultivation*. Negative values (blue) indicate that e.g. higher *Morality* is associated with lower *Employment* or lower morality with higher *Employment*. For the nominal dimensions *Gender* and *Race* applies: male = positive; black=positive. More white tiles indicate a weaker association between two dimensions. Values are mirrored across the diagonal for better readability.

The *Ngrams* exhibit a rather high similarity between *Cultivation* and *Morality* as well as *Cultivation* and *Education*. Furthermore, a positive association is found between *Affluence* on the one side and *Education*, *Cultivation* and *Status* on the other. Astoundingly, there is no association between *Affluence* and *Employment*. Kozlowski et al. (2019) draw a line to theories that describe '*how relations of production undergirding systems of economic stratification are obscured while the outward trappings of class, displayed through consumption patterns, remain visible and culturally salient*' (p. 923). Compared to the last reported decade (1990-1999)[26], these values largely

[23] https://www.reddit.com/r/science

[24] For the list of terms and categories, see Appendix Part B. from Kozlowski et al. (2019)

[25] Post-stratified weighting applied

[26] Kozlowski et al. (2019) did not report the angles between dimensions for the validation *Ngrams* (2000-2012)

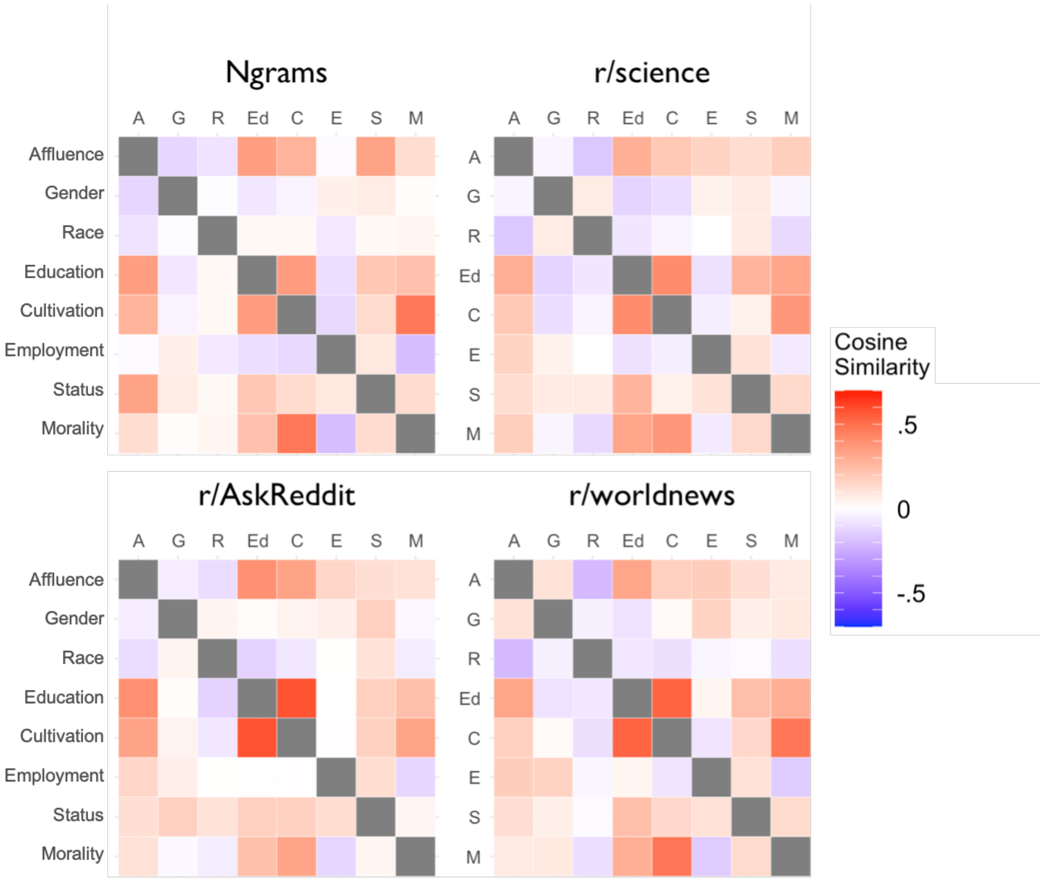

Figure 4. Cosine Similarity Between Eight Dimension Vectors for Ngrams and Three Subreddits. Values are Mirrored Across Grey Diagonal for Better Horizontal Readability.

remained stable; only *Cultivatio*n and *Morality* experienced a sharp rise from ~0.3 to 0.47 cos(θ). This indicates that dimensions of social distinction have become even more related in the new millenium. As before, *Gender* and *Employment* show no strong relationship to other dimensions. The same can be said about *Race*, which assumingly suffers from the ambiguity of some of its constituent words: *black* - *white*.

Looking at the angles produced by the subreddits, no remarkable changes can be found. With a few exceptions, all differences stay below 0.2 cos(θ). One larger gap can be observed for *Education* and *Cultivation*: While the *Ngrams* give 0.36 cos(θ) *AskReddit* and *worldnews* have their overall highest similarity with 0.58 and 0.54, respectively. For these subreddits, a cultivated taste has a stronger relationship with better education and vise versa. However, for *science* this association is only marginally bigger (+0.5) than in the *Ngrams*. Another notable difference between *Ngrams* and the subreddits can be found for *Affluence* and *Status*: In contrast to the Ngrams (0.33), all three Reddit groups yield lower similarities (~0.12). Apparently, the relationship between practical monetary wealth (*rich*, *lavish*, *luxury*…) and symbolic prestige (*honorable*, *influential*, *prominent*) is less pronounced in the online communities of Reddit. It is hard to determine where this variation has its roots but one possible reason could be language itself. As authors of prose try to artistically build their worlds, users in a fast-paced online environment are less inclined to embellish their output and thus use fewer symbolic adjectives, resulting in a bigger distance between the two dimensions. For a direct comparison between *Ngrams* and *AskReddit* (which scored best in the survey validation) see Appendix.

At this point, the comparison still lacks a proper measurement of full similarity between the corpora. One way to achieve this is the Euclidean distance between the sets of all 28 angles:

$$\sqrt{\sum_{i=1}^{28}(\cos(\theta)_{C1_i} - \cos(\theta)_{C2_i})^2}$$

where $C1$ and $C2$ are the two corpora that are being compared. A smaller distance between *science* and *Ngrams* than between the other two subreddits and the latter would bring support for the scientific-bias hypothesis of *Ngrams*. However, the results fom this measure are rather mixed: Even though the distance between *Ngrams* and *science* (0.54) is slightly smaller than for *Ngrams* and *worldnews* (0.55), it is even smaller for *Ngrams* and *AskReddit* (0.51). The distance in between subreddits is 0.44 for *AskReddit* and *worldnews*, 0.43 for *AskReddi*t and *science* and 0.4 for *science* and *worldnews*. Despite their different topics, these corpora share large portions of their user base and thus their perceptions of class. All in all, *Ngrams* and all subreddits are remarkably similar in how the class dimensions are associated with each other.

5. Discussion

As the demographics of the Reddit userbase (section 1.3.2) revealed, an overrepresentation of young, well-educated males had to be taken into account. The results of dimension angles showed remarkable similarity between the *Ngrams* corpus and three subreddits. Any differences can be due to individual demographics of subreddits, sampling (*AskReddit*, *worldnews*), preprocessing or the number of input words, which was almost exactly 1000 times higher for the Ngrams (250B). There are many parameters that influence the outcome of word vectors, producing errors in downstream analysis if not controlled for. This starts with preprocessing. While the objective for the Reddit corpus is to discard any content that does not reflect the opinions of its users, it remains unclear to what extent my choices influenced the outcome of word vectors and dimension angles. Future work should study the eventual influence of bot filter mechanisms and comment length thresholds. Later, the choice of hyperparameters can have a high impact on embedding performance (Levy, et al., 2015). Furthermore, when constructing dimensions, the choice of antonym words can be crucial for outcome. As oov words are skipped, the final number of available word pairs can highly vary between corpora. Additionally, more rare words will have overall weaker vectors, since their weights are trained less often. Even though it is possible to set a minimal count, the different outcomes dependent on word frequency are unclear. Word embeddings as a tool for culturomics remain fragile and complex, yet they promise great potential, especially when studying large corpora. The dimension approach by Kozlowski et al. (2019) and possible variations can offer insight into many more topics. To further validate this method, the robustness of dimensions when using different words must be evaluated. Especially when studying online environments, a choice of more modern terms and possibly abbreviations might be necessary to yield strong word vectors.

Furthermore, this paper did not find support for the scientific-bias hypothesis of *Ngrams* in regard of class dimensions. However, future work should focus on identifying characteristics of this corpus if it is continued to be used for sociological and linguistic analysis. While a popularity weighting will not be possible for *Ngrams* it could be achieved with Reddit comments, since meta information includes scores (upvotes – downvotes). Such an approach could reduce the influence of noise and additionally give a voice to users who vote but don’t write comments themselves.

Appendix

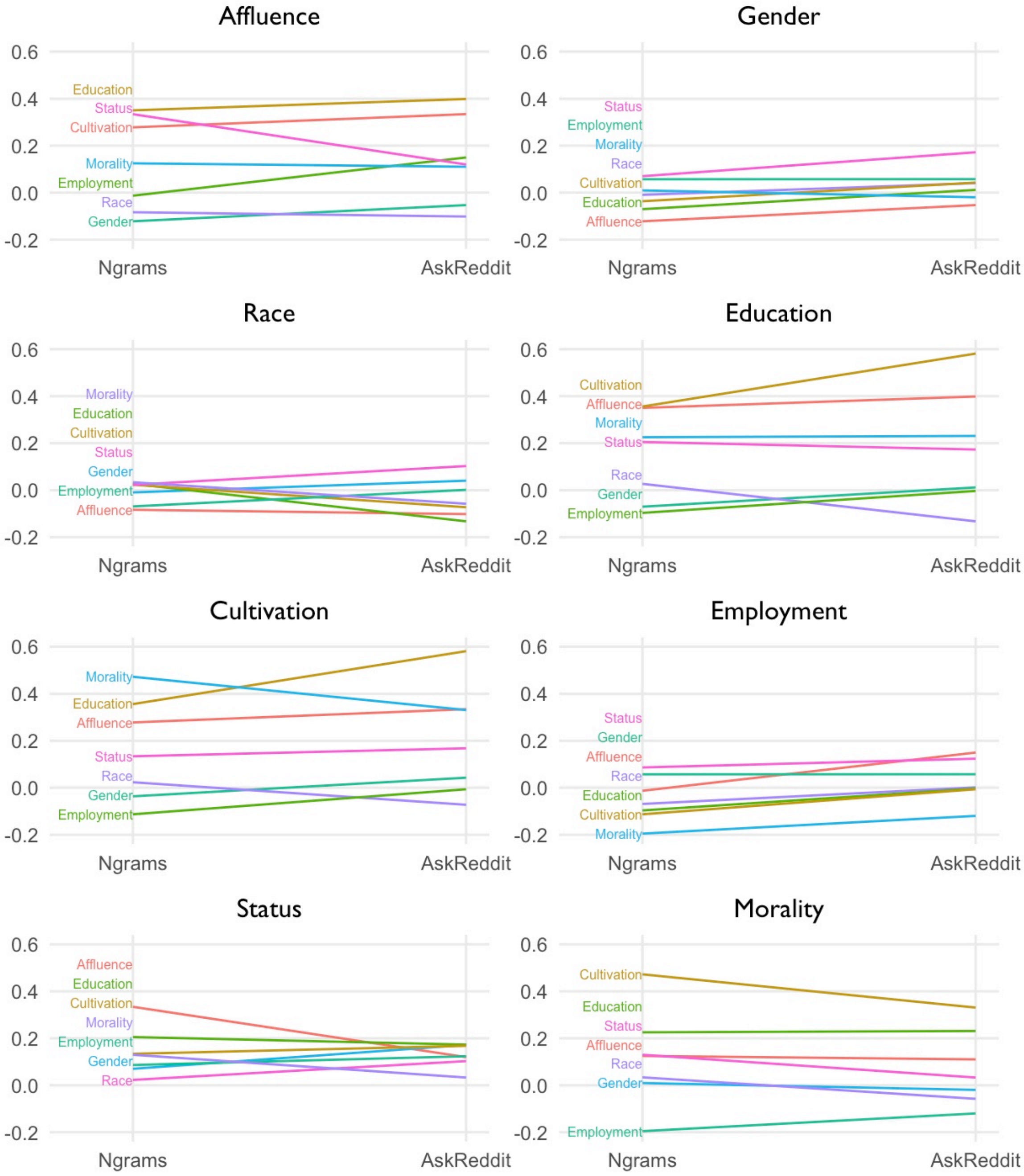

Figure 5. Comparison of Angles Between *Ngrams* and *AskReddit* for Eight Dimensions Measured in Cosine Similarity.

References

Allagnat, L. et al., 2017. *Gender in the Global Research Landscape,* s.l.: Elsevier.

Bakarov, A., 2018. A Survey of Word Embeddings Evaluation Methods.

Bruni, E., Tran, N.-K. & Baroni, M., 2014. Multimodal Distributional Semantics. *J. Artif. Intell. Res.,* 49: 1-47.

Christopherson, K. M., 2007. The positive and negative implications of anonymity in Internet social interactions: "On the Internet, Nobody Knows You're a Dog". *Comput. Hum. Behav.,* 23: 3038-3056.

Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S., 2010. Who is Tweeting on Twitter: Human, Bot, or Cyborg?. *ACSAC '10.*

Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 2.

Drozd, A., Rogers, A. & Matsuoka, S., 2016. Word Embeddings, Analogies, and Machine Learning: Beyond king - man + woman = queen. *COLING:* 3519-3530.

Gaffney, D. & Matias, J., 2018. Caveat emptor, computational social science: Large- scale missing data in a widely-published Reddit corpus. *PloS ONE.*

Gerz, D. et al., 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity.. *ArXiv, abs/1608.00869.*

Herring, S. C., 2002. Computer-mediated communication on the internet. *ARIST,* 36: 109-168.

Kim, B., Kim, H. & Kim, G., 2018. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks. *NAACL-HLT.*

Kozlowski, A. C., Taddy, M. & Evans, J. A., 2019. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *ArXiv, abs/1803.09288.*

Levy, O., Goldberg, Y. & Dagan, I., 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics,* 3: 211-225.

Manstead, A. S., 2018. The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour. *The British Journal of Social Psychology,* 57: 267 - 291.

Medvedev, A. N., Renaud, L. & Delvenne, J.-C., 2018. The anatomy of Reddit: An overview of academic research. *ArXiv, abs/1810.10881.*

Michel, J.-B.et al., 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331: 176-182.

Mikolov, T., Corrado, G., Chen, K. & Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR, abs/1301.3781.*

Pechenick, E. A., Danforth, C. M. & Dodds, P. S., 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE*, 10(10).

Singer, P. et al., 2016. Evidence of Online Performance Deterioration in User Sessions on Reddit. *PLoS ONE*, 11(8).

Singer, P. et al., 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?. *WWW '14 Companion*.

Soliman, A., Hafer, J. & Lemmerich, F., 2019. A Characterization of Political Communities on Reddit. *HT '19*: 259-263.

Tang, D. et al., 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. *ACL*, p. 1555–1565.

Turian, J., Ratinov, L. & Bengio, Y., 2010. Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 384–394.

Varol, O. et al., 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. *ArXiv, abs/1703.03107*.