RedditEM: Unveiling Diachronic Semantic Shifts in Social Network Discourse [*]

Jiajun Zou zjj21@mails.tsinghua.edu.cn
Tsinghua University

Sixing Wu wusx@ncepu.edu.cn
North China Electric Power University

Jinshuai Yang yjs20@mails.tsinghua.edu.cn
Tsinghua University

Minghu Jiang jiang.mh@mail.tsinghua.edu.cn
Tsinghua University

Yongfeng Huang yfhuang@mail.tsinghua.edu.cn
Tsinghua University, Zhongguancun Laboratory

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Humans employ words to convey abstract concepts. The evolution of lexical semantics holds significance not only in Natural Language Processing applications but also in the realm of social computing research. However, the scarcity of diachronic word representations persists due to the substantial computational demands, particularly evident in the absence of large-scale and enduring diachronic word embeddings for social network texts. Herein, we introduce RedditEM, a comprehensive collection of diachronic word representations derived from Reddit English comment texts, featuring one word embedding per month spanning from January 2010 to December 2021. To assess the diachronic semantic shifts of words, we employ cosine distance metrics and juxtapose the embeddings' neighborhoods. Our experimental findings underscore the utility of RedditEM in detecting alterations in word meanings within social networks and advancing social computing endeavors. Researchers interested in accessing this resource are cordially invited to contact us without hesitation.

Keywords: Social computing; Social network texts; Diachronic word embedding; Semantic change.

1. Introduction

Language, characterized by its ability to make infinite use of finite media, stands as a cornerstone of communication, intricately intertwined with human behavior and cultural identity. The evolution of society and culture finds its genesis in the transformative power of languages, which undergo multifaceted changes over time, encompassing phonetic alterations, spelling variations, lexical modifications, semantic nuances, and syntactic adaptations Braun (2022). This dynamic process involves the addition of new meanings to existing words, shifts in prevalence, and even the replacement of former meanings Shoemark et al. (2020). The exploration of lexical semantic change assumes pivotal importance not only

[*] This work was supported by by the National Key R&D Program of China (Grant No. 2022YFC3302104) and the National Key Natural Science Foundation Project of China under Grant 62036001.

in the realm of Natural Language Processing (NLP) applications but also in the broader domains of text data mining, social sciences, politics, and culture. For instance, a comprehensive analysis of historical shifts and stabilities in social group representations spanning centuries underscores the significance of studying lexical semantics as a window into societal transformations Charlesworth et al. (2022).

With the advancement of NLP technology, word embeddings have emerged as indispensable tools across diverse application domains, ranging from sentiment analysis and named entity recognition to information retrieval and beyond, encompassing fields such as social sciences, marketing, and finance. These embeddings leverage dense, low-dimensional real-number vectors to represent words in geometric spaces, enabling computational operations to calculate semantic similarities between words or sentences. Rooted in the distributional hypothesis Lenci (2008), which posits that words with similar meanings tend to occur in similar contexts Tsakalidis et al. (2021), word embeddings offer a pathway to uncovering subtle shifts in word meanings that may elude manual inspection Shoemark et al. (2020). In recent years, diachronic distributional models, particularly diachronic word embeddings, have proven instrumental in tracking semantic changes and shifts in language use across various contexts Pedrazzini and McGillivray (2022); Braun (2022); Shoemark et al. (2020), underscoring the urgent need to study lexical semantic changes and extract valuable insights through diachronic word embeddings. Over the past years, there has been an increase in the amount of studies on diachronic word embeddings in NLP research Kutuzov et al. (2018); Tahmasebi et al. (2021). Based on neural networks, these diachronic word embeddings turn word input into vectors with generally 50–300 dimensions, such as word2vec Mikolov et al. (2013) and Glove word embedding Pennington et al. (2014). And a large number of resources are used to train these embeddings, which include Twitter, Common Crawl, Gigaword and Wikipedia Tsakalidis et al. (2021). Although there have been many researchers proposed diachronic word embedding, for example in Hamilton et al. (2016b); Grayson et al. (2016); Kim et al. (2014); Shoemark et al. (2019); Tsakalidis et al. (2021), as far as we know, comprehensive word embeddings derived from social network texts remain scarce due to computational constraints and limitations in text sources.

While Tsakalidis Tsakalidis et al. (2021) presented DUKweb (1996-2013), this diachronic word embedding is based on web texts rather than the very anonymous social network texts. Social media texts, characterized by their irregularity and variability, offer rich ecological signals reflecting individuals' emotions, thoughts, behavioral patterns, and traits Liu et al. (2022), making them invaluable sources for studying lexical semantic changes. Shoemark's work Shoemark et al. (2019) based on Twitter contains the sole example of trained diachronic word embeddings encompassing a short and recent time period, spanning 5.5 years, from January 1, 2012 to June 30, 2017, but it has become outdated and cannot reflect the semantic changes of social network language in recent years, especially in the impact of COVID-19 epidemic Huang et al. (2020). Consequently, there exists an urgent need for large-scale, long-term word embeddings based on social network texts to advance the study of semantic changes in social network vocabularies.

In this study, we provide a solution to the above issues. And our main contributions are as follows:

- Oriented social network texts, for the first time, we present a large-scale, long-term diachronic word embeddings derived from Reddit English comments spanning from January 2010 to December 2021, covering 144 months of linguistic evolution.
- Through rigorous experimentation employing cosine distance metrics and neighborhood comparisons, we demonstrate the efficacy of RedditEM in capturing and quantifying lexical semantic changes over time.
- Our findings underscore the potential of RedditEM to advance research in understanding semantic shifts and unraveling social group traits embedded within the evolving lexicon of social network discourse.

2. Related Works

In this section, we will briefly introduce related works from diachronic word embedding and semantic change measurement.

2.1. Diachronic Word Embedding

In recent years, there has been a surge in studies focusing on diachronic word embeddings. Hamilton et al. Hamilton et al. (2016b) introduced a method for identifying semantic change utilizing word embeddings trained on the extensive Google Ngram corpus Lin et al. (2012), comprising 8.5 billion words from historical writings in English, French, German, and Chinese. This pioneering work stands as a significant exemplar in this emerging field. Charlesworth et al. Charlesworth et al. (2022) leveraged these historical word embeddings to conduct an extensive quantitative and qualitative analysis of social group representations. Kim et al. Kim et al. (2014) presented word2vec embeddings at five-year intervals, trained on a substantial portion of the Google Ngram corpus' English fiction section. Additionally, Grayson et al. Grayson et al. (2016) released diverse word2vec embeddings trained on the Eighteenth-Century Collections Online dataset spanning 1700-1799, derived from 150 million randomly selected words from its "Literature and Language" component across five twenty-year intervals.

Another notable contribution is DUKweb Tsakalidis et al. (2021), a diachronic word embedding derived from the JISC UK Web Domain Dataset (1996–2013), encompassing a vast archive of Internet resources from domains ending in '.uk'. This resource comprises series of word co-occurrence matrices and two types of word embeddings for each year within the dataset. However, it is worth noting that most existing diachronic word embeddings are primarily based on written or web texts, thus potentially overlooking lexical semantic changes in social media discourse.

In contrast, social media texts offer a distinct advantage due to their anonymized nature, which allows for a more authentic reflection of individuals' thoughts and behaviors, characterized by spontaneity and variability. Shoemark et al. Shoemark et al. (2019) collected tweets from Twitter's 'statuses/sample' streaming API endpoint from January 1, 2012, to June 30, 2017, training word embeddings using gensim's Řehřek and Sojka (2010) implementation of the continuous bag of words model Mikolov et al. (2013). However, this dataset only extends to 2017, limiting its applicability in capturing the latest semantic changes in social network texts. Consequently, there is a pressing need to develop a large-scale and long-term diachronic word embedding specifically tailored to social network texts to facilitate comprehensive research in this domain.

In the realm of training diachronic word embeddings, two predominant approaches are typically adopted. The first method involves continuous training, where embeddings for a specific time-step t are initialized with those trained at the preceding time-step t-1 Kim et al. (2014). Conversely, the second approach entails training embeddings independently for each time-step and subsequently aligning them posthoc Hamilton et al. (2016b); Kulkarni et al. (2015). Shoemark Shoemark et al. (2019) observed that independently trained and aligned embeddings outperform continuously trained embeddings, particularly over extended time periods.

Given these findings, we opt to pursue the latter approach for training diachronic embeddings.

2.2. Measuring Semantic Change

Various methods exist for quantifying semantic change, among which cosine distance stands as a prominent example. The formula for cosine distance is as follows:

$$cos_{distance}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - cos_{similarity}(\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{1}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is a vector, respectively. The cosine distance, closely linked to cosine similarity, denotes similarity between two vectors. When the cosine distance approaches zero, it signifies a higher degree of similarity between the vectors. Previous studies Dubossarsky et al. (2017); Kim et al. (2014); Stewart et al. (2017) have employed cosine distance as a metric for measuring semantic change, utilizing vector representations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the same vocabulary at different time points, respectively.

An alternative approach involves comparing the neighborhoods of word embeddings Hamilton et al. (2016a), which has garnered widespread adoption Shoemark et al. (2019); Braun (2022); Pedrazzini and McGillivray (2022); Robertson et al. (2021). Initially, it entails establishing the ordered collection, F_t^w, comprising the k closest neighbors of word embedding $\boldsymbol{w}_t$ in lexical representation space using cosine similarity at each time-step t. Subsequently, the second-order vector $\boldsymbol{S}_t$ is constructed for any two time-steps by amalgamating F^w of the two nearest neighbor sets. Each element S_t^i of $\boldsymbol{S}_t$ contains the cosine similarity of $\boldsymbol{w}_t$ to the vector $\boldsymbol{w}_t^i$ from the neighboring word embedding collection F_t^w at time t. Therefore, the following formula is obtained:

$$S_t^i = cos_{similarity}(\boldsymbol{w}_t, \boldsymbol{w}_t^i),\ \boldsymbol{w}_t^i \in F_t^w. \tag{2}$$

Then the local neighborhood change is as follows:

$$D(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}) = cos_{distance}(\boldsymbol{S}_t, \boldsymbol{S}_{t+1}). \tag{3}$$

This method, akin to the cosine distance approach, underscores the importance of neighboring vocabulary surrounding the target word, offering a wealth of contextual information.

In this study, we leverage both aforementioned methods to gauge semantic changes in words.

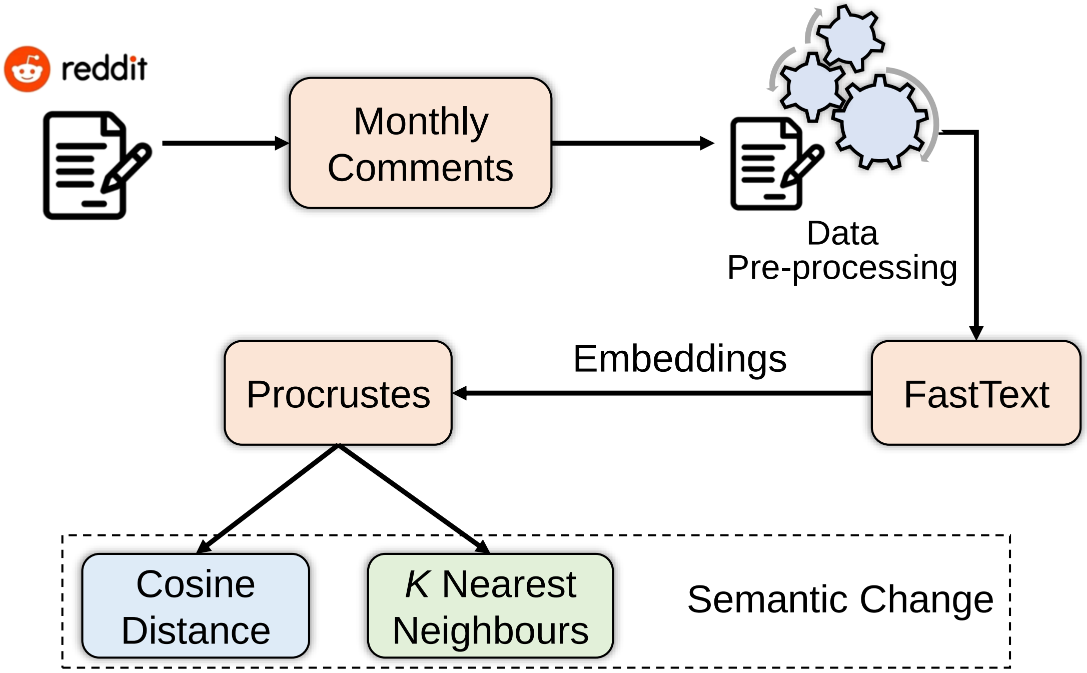

Figure 1: Flowchart of the creation and use of RedditEM.

3. Method and Experiment

The process leading from social network texts to the development of RedditEM is illustrated in Figure 1. Subsequent sections provide a comprehensive account of the RedditEM creation process and elucidate how RedditEM facilitates the extraction of lexical semantic changes.

3.1. Data and Pre-processing

Reddit, boasting a user base exceeding 430 million monthly active users Guo et al. (2021), stands as one of the largest social media platforms globally, rendering it a fitting subject for our investigation. To procure textual data, we leverage pushshift Baumgartner et al. (2020), a platform offering comprehensive access to Reddit's public posts and comments.

To ensure a robust dataset, we have filtered out comments predating 2010 and compiled monthly comment data spanning from January 2010 to December 2021, amounting to a total of 10.2 billion comments. Subsequently, we proceed to extract monthly comment texts from these files and subject them to a thorough cleansing process. This involves the removal of "[deleted]" comments, elimination of excess whitespace, and filtering out of URLs, HTML codes, hashtags, handles, and any residual whitespace in each comment text. Notably, we employ "langid"[1] to discern English comments accurately.

[1] https://github.com/saffsd/langid.py

Table 1: Words with semantic change measured by cosine distance

experimental group	blackberry	tweet	nfc	tablet	AI	btc
	raspberry	metaverse	python	science	fabs	cloud
	XR	ML	DL	VR	transformer	bert
control group	white	black	asian	irish	hsipanic	men

Table 2: The Selected Social Groups

Social Groups (All in lowercase form)					
white	black	asian	irish	hispanic	men
women	old	young	fat	thin	rich
poor	america	democrats	conservative	vagabond	backpacking

3.2. Training Word Embeddings

We employ fastText[2] Bojanowski et al. (2017); Joulin et al. (2017), an unsupervised predictive learning algorithm, to craft our diachronic word embeddings. Key parameters of the model include its dimensionality and the range of subword sizes. To ensure comprehensive information capture, we set the dimensionality to 300. Subword sizes range from 2 to 5, allowing for nuanced representation. The learning rate is established at 0.01, while the epoch parameter retains the default value of 5. Due to the benefit seen in Shoemark et al. (2019), Building upon the insights gleaned from Shoemark et al. (2019), we opt to individually train word embeddings for each month, subsequently aligning these embeddings for enhanced coherence.

3.3. Aligning Diachronic Embeddings

In order to facilitate comparison across different time periods, we employ Procrustes analysis, as outlined in Hamilton et al. Hamilton et al. (2016b), to align the semantic spaces, thereby ensuring that all embeddings are referenced within a unified coordinate system. Let $\boldsymbol{W}^{(t)} \in \mathbb{R}^{n \times m}$ denote the matrix of word embeddings in month t, where n represents the vocabulary size and m signifies the dimensionality. The Orthogonal Procrustes problem, detailed in Schönemann (1966) entails identifying the rotation matrix $\boldsymbol{Q} \in \mathbb{R}^{m \times m}$ that best aligns the matrices $\boldsymbol{W}^{(t)}$ and $\boldsymbol{W}^{(t+1)}$. The solution is as follows:

$$\underset{\boldsymbol{Q}}{min} \left|\left| \boldsymbol{W}^{(t)}\boldsymbol{Q} - \boldsymbol{W}^{(t+1)} \right|\right|_F, \text{ s.t. } \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}, \tag{4}$$

where $\boldsymbol{I}$ is the $m \times m$ identify matrix, and $||...||_F$ is the Frobenius norm Pedrazzini and McGillivray (2022). The optimisation problem in equation (4) can be solved by using the singular value decomposition of $\boldsymbol{W}^{(t)}(\boldsymbol{W}^{(t+1)})^T$ Tsakalidis et al. (2019, 2021); Pedrazzini and McGillivray (2022). Once the embedding spaces have been aligned, we employ the cosine distance between vectors of a word across different months to ascertain the semantic changes of that word.

[2] https://fasttext.cc/

3.4. Cosine Distance to Measure Semantic Change

Over the past decade, the rapid advancement of science and technology has led to the emergence of new meanings for certain words. For instance, "ML" traditionally referred to mean motor launch[3], but its usage has evolved to increasingly denote Machine Learning in recent years. This lexical evolution is also evident in social media platforms. Thus, in this subsection, we employ cosine distance to assess the semantic changes of select technology-related words using the trained and aligned diachronic word embeddings.

The chosen words, sourced from the Oxford English Dictionary (https://www.oed.com/), a highly authoritative and comprehensive English lexicon, are presented in Table 1. We establish an experimental group comprising 18 technological terms and a control group consisting of words with relatively stable semantics, such as "white", "black", "men" and so forth. we compute the cosine distance between its embedding in December 2010 and its aligned vectors in the semantic space for every month spanning from 2011 to 2021.

3.5. Neighborhood Embeddings to Measure Semantic Change

In contrast to the cosine distance method, the neighborhood embedding approach also emphasizes the relevance of neighboring words, providing insights into the distribution of words most pertinent to semantic changes of a given word. In this subsection, we leverage RedditEM to explore the historical depictions of social groups on social media from January 2011 to December 2021 using neighborhood embeddings, thus discerning changes in trait words most associated with these social groups. This experiment comprises two phases.

Initially, drawing from previous studies Charlesworth et al. (2022); Waller and Anderson (2021), we identify 18 social groups, as delineated in Table 2. Given the minimal semantic fluctuation in these groups, we employ characteristic words to indirectly reflect their historical changes and stability. Utilizing a pool of over 600 trait adjectives Peabody (1987), downloadable from the provided link[4]. We refine the word embedding of a social group for a specific month by incorporating 25 neighborhood embeddings of the social group. Subsequently, we compute the cosine distance between all social group representations and trait adjectives, identifying the top 10 trait associates most strongly associated with each social group (Table 3). This allows for an intuitive grasp of trait changes within social groups.

In the subsequent phase of the experiment, building upon the initial findings, we determine the sentiment valence of these top 10 trait words. Following methodologies akin to prior studies Waller and Anderson (2021); Garg et al. (2018); Bolukbasi et al. (2016); Kozlowski et al. (2019), we devise a sentiment valence dimension to gauge the valence of these trait adjectives. Employing very positive words (e.g., "good", "wonderful", "great") and negative words (e.g., "bad", "awful", "sad") as seed words, we calculate vector differences between word embeddings of these seed words. The average of these vector differences serves as the sentiment valence dimension vector. Subsequently, we project the identified trait adjectives onto the valence dimension vector, yielding their corresponding valence scores. This

[3] https://www.oed.com/view/Entry/111722?rskey=THn5lb&result=1&isAdvanced =false#eid38438193

[4] https://osf.io/kbuhn

process can be described by the following formula:

$$\boldsymbol{V}_{np} = \frac{\sum_{i}^{m}(\boldsymbol{v}_{ni} - \boldsymbol{v}_{pi})}{m \cdot m}, \tag{5}$$

$$SC = \frac{\boldsymbol{s} \cdot \boldsymbol{V}_{np}}{||\boldsymbol{s}|| \cdot ||\boldsymbol{V}_{np}||}, \tag{6}$$

where $\boldsymbol{v}_{ni}$ and $\boldsymbol{v}_{pi}$ are embeddings of seed words, respectively. $\boldsymbol{s}$ is the social group's representation, and $\boldsymbol{V}_{np}$ is the valence dimension. SC is the valence score, m is the number of seed words. Then we can learn about the changes in the valence of social groups from 2011 to 2021.

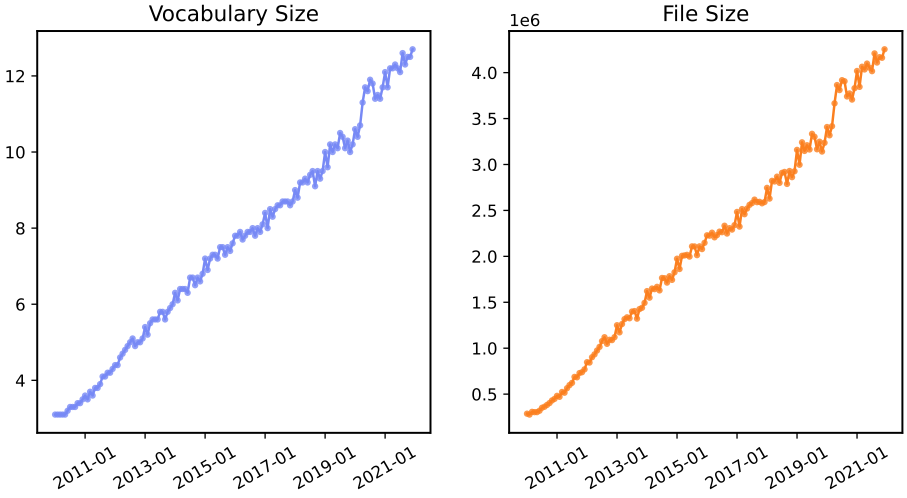

Figure 2: The vocabulary and file size of changes over every month.

4. Results and Analysis

4.1. The Diachronic Word Embeddings

Figure 2 illustrates the evolution of vocabulary size and file size across the 144-month word embeddings (RedditEM). It is evident that the vocabulary embedding scale increases progressively over time. From January 2010 to January 2021, the scale has surged by approximately 15-fold. Despite the exclusion of comment texts predating 2010, the volume of comment texts in recent years is substantial. For instance, the size of the comment text file for each month in 2021 is anticipated to exceed 40GB. Consequently, ensuring uniformity in the size of word embeddings across different time periods becomes challenging. This phenomenon is to be expected, given the expanding user base of Reddit over time, leading to the introduction of numerous new words and a proliferation of vocabulary. Hence, these word embeddings outcome offer a more authentic portrayal of the evolving landscape of social network vocabulary.

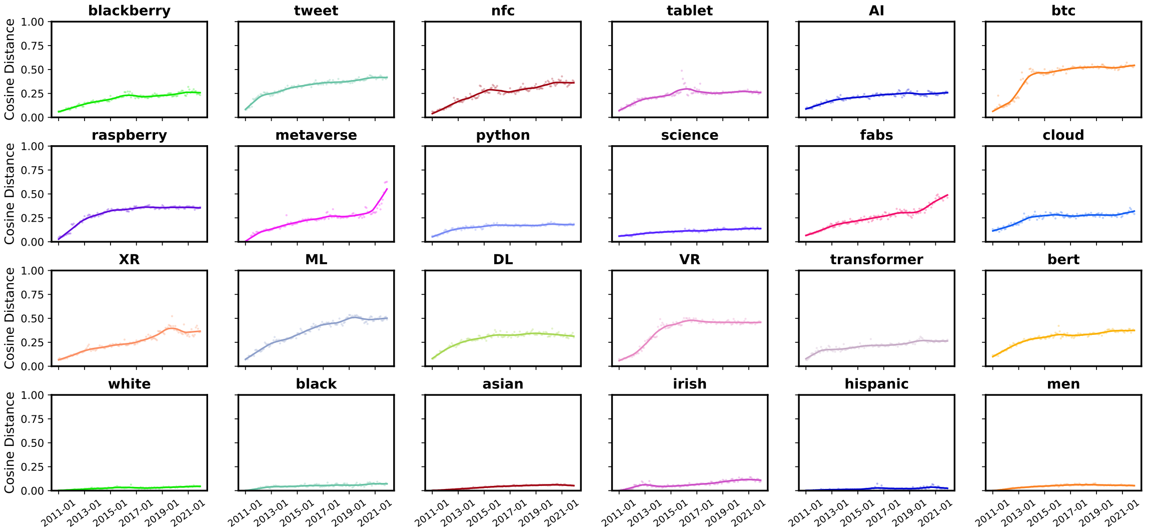

Figure 3: Temporal evolution of cosine distance for the selected words. The time series is smoothed using LOWESS, a non-parametric regression technique that fits distinct linear regressions for each data point by incorporating neighboring data points to estimate slope and intercept.

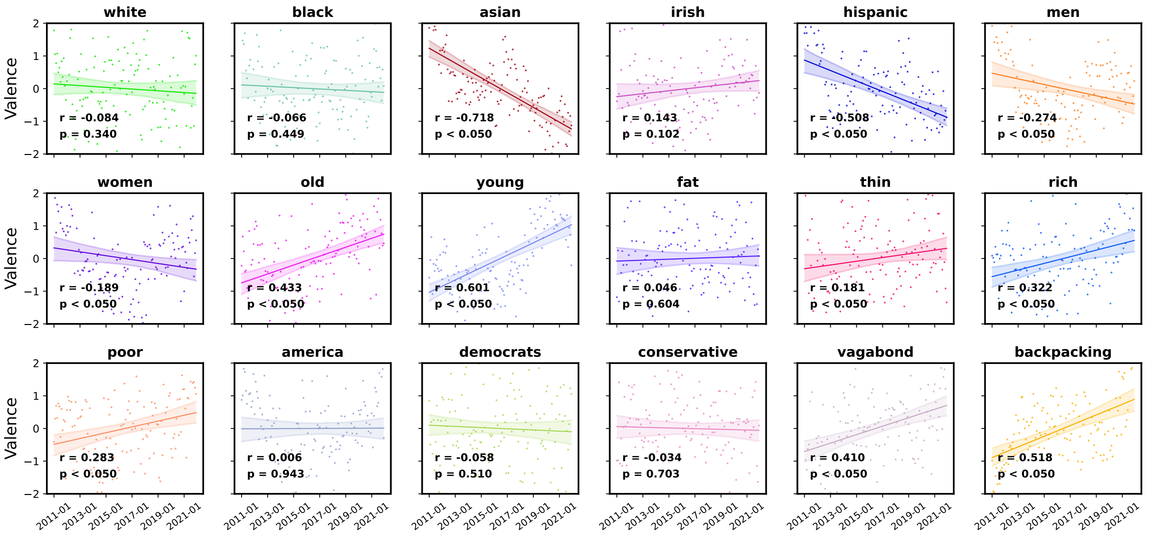

Figure 4: Temporal dynamics of trait valence associated with social groups. The y-axis represents the projection scores of words on the valence dimension. Statistical analysis was performed on these sentiment scores across time intervals

4.2. Words Cosine Distance Change Across Time

Figure 3, utilizing LOWESS[5] for time series smoothing, illustrates the evolution of cosine distances for selected words from January 2011 onwards, with semantic anchors set in December 2010. This visualization reveals notable semantic shifts over time for most technology-related terms. Conversely, words like “white”, “black”, “asian”, “irish”, “hiapanic” and “men” exhibit minimal variation, maintaining a cosine distance close to 0 throughout the observed period. Detailed meanings of these words can be referenced at https://www.oed.com/.

In terms of technological words, we observe a trend towards semantic stabilization for certain words following initial fluctuations, as exemplified by “btc”, “raspberry”, “cloud”, “DL” and “VR”. Semantic shifts often align with the emergence or surge in popularity of corresponding technologies. Take “btc” for instance, originally denoting a peer-to-peer electronic cash system (Bitcoin) Nakamoto (2008) in 2008. Its semantic evolution unfolded during the period from 2008 to 2014, coinciding with its introduction and adoption by online communities, eventually stabilizing around 2015.

Conversely, some terms exhibit ongoing semantic drift in recent years rather than attaining relative stability. Examples include “metaverse” and “fabs”. Initially, “metaverse” referred to a hypothetical virtual reality environment or a conceptual space where users interact via networked computers[6]. However, in contemporary usage, it encompasses a novel Internet application and social construct integrating various new technologies Ning et al. (2023). The renaming of Facebook to Meta in 2021 further propelled “metaverse” into the spotlight, expanding its semantic horizons over the past three years, as depicted in Figure 3.

4.3. Words Neighborhoods Change Across Time

The cosine distance method provides insight into the semantic drift of words but does not elucidate the related words in their vicinity. The ranked top 10 trait adjectives most closely associated with individual social groups over time are presented in Table 3 and Table 4.

Examining specific social groups, similarities emerge in the traits attributed to White and Black, albeit with the latter characterized by descriptors such as “discourteous”, “shrewd” and “loyal”. Notably, biases and stereotypes towards Asian are evident, including terms like “sly”, “resigned”, “evasive”, while Hispanic is linked with traits like “listless”, “forward” and “insonent”. America is predominantly associated with “helpful” and “ecasive”.

Traits attributed to Men and Women exhibit similarities, featuring words such as “original”, “reliable” and “sluggish”. Conversely, Young is described as more “direct”, while Old is associated with traits like “meditative”. Interestingly, Fat is characterized by terms like “communicative”, “diplomatic” and “solemn”, while Thin lacks such associations. Rich is linked with “alert” and “prompt”, whereas Poor is more aligned with “serene”, “sensual” and “sexy”. Additionally, Vagabond is described as “sly”, “bold” and “courteous”, whereas Backpacking is associated with “moral”, “submissive” and “direct”.

[5] https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

[6] https://www.oed.com/view/Entry/271922?redirected From=metaverse#eid

Table 3: Top 10 traits adjectives most strongly and relatively associated with social groups

	2011-01	2014-01	2017-01	2020-01
white	**forward**,reliable,direct, emotional,hopeful,calm, withdrawn,happy, patient,nervous	withdrawn,**forward**, **meditative**,reliable, trust,worthy,orderly, humble,loyal,patient, optimistic	**meditative**,resigned, withdrawn,friendly, prompt,spiritual,fickle, direct,indirect,active	**meditative**,withdrawn, patient,cooperative, resigned,charitable, friendly,reliable,prompt, disorderly
black	forward,emotional, reliable,moral,happy, hopeful,natural,spiritual, earnest,calm	meditative,withdrawn, forward,earnest, trustworthy,reliable, **loyal**,humble,analytical, charitable	meditative,resigned, spiritual, direct, emotional, indirect, friendly,**loyal**,withdrawn, **shrewd**	meditative,withdrawn, patient,**shrewd**,gracious, resigned,courteous, reliable,wholesome,concise
america	**helpful**,bossy,thorough, disorderly,flexible, forward,orderly,active, playful,verbal	**evasive**,listless,sensual, prompt,solemn,resigned, bossy,submissive, disorderly,orderly	**evasive**,**helpful**,solemn, listless,methodical, meditative,guarded, defensive,disorderly, versatile	**evasive**,meditative, sensual,**helpful**,disorderly, gentle,methodical, unreliable,orderly, neat
asian	**sly**,forward,defensive, **resigned**,sober,warm, calm,happy,hopeful, prompt	**resigned**,prompt, obstructive,forward, defensive,reliable,**sly**, alert,insolent,cold	**resigned**,obstructive, prompt,orderly,alert, solemn,reliable, concise, defensive, listless	obstructive,**evasive**, **resigned**,prompt, insolent,orderly,withdrawn, unselfish,solemn,reliable
irish	controlled,logical, reliable,capable, defensive,direct, forward,social,disorderly, responsible	reliable,prompt, disorderly, capable,orderly,critical, cooperative,direct, resigned,evasive	obstructive,reliable, disorderly,prompt, direct,orderly,helpful, competitive,defensive, evasive	evasive,obstructive, orderly,disorderly, prompt,reliable, courteous,discourteous, direct,dedicated
hispanic	**forward**,sly,hopeful, warm,direct,prompt, reliable,capable, glum,flexible	prompt,**forward**,cold, **listless**,reliable,sly, careful,brave, efficient,inefficient	**listless**,prompt,erratic, brave,cold,**forward**, careful,alert, solemn,upright	evasive,**insolent**,direct, prompt,jolly,serene, meditative,solemn, reliable,brave
men	**original**,alert,forward, direct,refined,controlled, hopeful,**reliable**,calm, prompt	**reliable**,abrupt,**original**, alert,versatile,stable, sluggish,efficient, unreliable,finn	sluggish,**reliable**, unreliable,**original**, alert,versatile, jolly,listless, friendly,wary	sluggish,**reliable**,**original**, versatile,evasive,polished, neat,alert,jolly,unreliable
women	**original**,alert,forward, direct,refined,**reliable**, controlled,hopeful,calm, composed	**reliable**,abrupt,**original**, versatile,alert,stable, sluggish,finn,unreliable, efficient	sluggish,**reliable**, unreliable,versatile, **original**,alert,jolly, listless,efficient,wary	sluggish,**reliable**,versatile, polished,**original**,evasive, neat,jolly,alert,inefficient
old	direct,moral,social, natural,indirect, controlled,ethical, logical,economical, peaceful	**meditative**,prompt, moral,evasive, ethical,indirect, direct,helpful,objective, withdrawn	moral,**meditative**, objective,peaceful, ethical,indirect, defensive, obstructive, negativistic, bold	obstructive,helpful, evasive,indirect, **meditative**, constructive, objective,direct, concise, tense
young	moral,controlled, refined,**direct**, alert,objective, artificial,ethical, logical,natural	evasive,controlled, prompt,autocratic, efficient,alert,reliable, jolly,informal,abrupt	**direct**,prompt, obstructive, submissive,indirect, evasive,objective, controlled,solemn, alert	**direct**,obstructive, evasive,thorough, dedicated, helpful, prompt, controlled, indirect,reliable
fat	direct,**communicative**, organized,**diplomatic**, alert,moral,peaceful, prompt,innovative, cooperative	**diplomatic**,alert, direct,cooperative, organized,loyal, guarded,thorough, innovative,resigned	direct,**diplomatic**, **solemn**,cooperative, prompt,thorough, brilliant,organized, alert,verbal	**solemn**,inquisitive, dedicated, prompt,**diplomatic**, guarded,direct, alert,cooperative, meddlesome
thin	alert,responsible,direct, social,peaceful, organized, controlled,forward, informal,original	verbal,withdrawn, prompt,alert, meditative,peaceful, retiring,trustworthy, reliable,cooperative	meditative,prompt, alert,direct, submissive,orderly, spiritual,retiring, verbal, responsible	alert,prompt, wholesome, courteous,meditative, direct,responsible, meddlesome,verbal, evasive
rich	**alert**,forward,verbal, punctual,prompt,active, sensual,cordial,casual, formal	prompt,verbal,**alert**, forward,calm,active, defensive,submissive, reliable,adaptable	prompt,submissive, solemn,**alert**,verbal, active,defensive, inactive,evasive, forward	verbal,prompt,active, inactive,helpful, changeable,forward, evasive,submissive, courteous
poor	alert,active,forward, calm,casual, deep,natural, sensual,**sexy**,cordial	**serene**,jolly, sensual,bold, calm,timid,alert, meditative, cordial,solemn	sensual,solemn, meditative, prompt,active, submissive, alert, jolly, **serene**,**sexy**	sensual,meditative, active,evasive, **sexy**,solemn, helpful,prompt, inactive, inhibited
democrats	**cordial**,intense, sly,refined, sensual,alert, soft,casual, flexible,forward	serene,adaptable, intense,sensual,tranquil, **versatile**,moody,soft, gentle,flexible	**versatile**,sensual, **listless**,serene, moody,tranquil, soft,flexible,sexy, adaptable	**versatile**,serene, playful,sensual, evasive,dedicated, refined,soft, discreet,moody
conservative	**sly**,forward,alert, **sexy**,warm, cold,capable, original,bossy,deep	**sly**,tranquil, serene,**finn**, **jolly**,**sexy**,silent, sensual,capable,prompt	**jolly**,serene,tranquil, sensual,versatile, prompt,**finn**, grim,direct,alert	serene,sensual, **sexy**, **jolly**, versatile,tranquil, **sly**,**finn**, restless, direct
vagabond	**sly**,forward,lazy,alert, smug,proud,liberal,cold, calm,lively	**sly**,moderate, humble,prompt, biased,versatile, reliable,**bold**, alert,forward	**bold**,prompt,moderate, innovative,reliable, forward,alert,obstructive, biased,sluggish	**bold**,evasive,alert, **courteous**,innovative, dedicated,unoriginal, sluggish,reliable,bright
backpacking	**moral**,disorderly,ethical, immoral,natural, expressive,intellectual, religious,principled,social	**submissive**,timid, benevolent,disobedient, abrupt,argumentative, loyal,**direct**, indirect,biased	disobedient,observant, obedient,indirect, credulous,inquisitive, **submissive**,**moral**, loyal,negativistic	**submissive**,deceitful, sensual,credulous, dominant,indirect, **direct**,disorderly, truthful, constructive

Democrats are notably connected with traits like “cordial”, “serene”, “versatile” and “listless”, whereas traits like “sly”, “jolly”, “finn” and “sexy” are strongly associated with Conservative.

Table 4: Top 10 traits adjectives most strongly and relatively associated with social groups in January 2021

	Top 10 traits adjectives
white	**meditative**,withdrawn,patient,courteous,emotional,prompt,active, wholesome,disorderly,sexy
black	meditative,patient,withdrawn,courteous,concise,emotional,helpful, active,**discourteous**,disorderly
america	**evasive**,listless,solemn,inhibited,methodical,playful,self-disciplined, prompt,sensual,**helpful**
asian	**evasive**,obstructive,orderly,prompt,insolent,disorderly,**resigned**, withdrawn,concise,helpful
irish	obstructive,orderly,courteous,disorderly,evasive,helpful,discourteous, reliable,dedicated,competitive
hispanic	evasive,solemn,**insolent**,prompt,dedicated,crafty,meditative,wasteful, sluggish,thorough
men	sluggish,alert,friendly,**reliable**,unreliable,erratic,versatile,noisy,jolly, **original**
women	sluggish,alert,friendly,**reliable**,versatile,erratic,unreliable,jolly,noisy, polished
old	prompt,sensual,helpful,constructive,indirect,**meditative**,concise, evasive,courteous,direct
young	**direct**,courteous,obstructive,thorough,helpful,concise,prompt,indirect, discourteous,controlled
fat	**solemn**,courteous,discourteous,organized,direct,inquisitive,informal, prompt,alert,cooperative
thin	courteous,alert,prompt,wholesome,discourteous,abusive,disorderly, meditative,organized,peaceful
rich	prompt,evasive,verbal,helpful,active,courteous,inactive,inhibited, meditative,**alert**
poor	**sexy**,sensual,meditative,active,**serene**,solemn,prompt,evasive, courteous,helpful
democrats	**versatile**,**listless**,evasive,serene,sexy,dedicated,inhibited,talented, sensual,moody
conservative	finn,**sexy**,versatile,sensual,serene,tranquil,restless,**jolly**,talented, courteous
vagabond	**courteous**,discourteous,prompt,alert,dedicated,reliable,innovative, friendly,sluggish,**bold**
backpacking	**direct**,indirect,inquisitive,**submissive**,obedient,disobedient,disorderly, courteous,vulgar,**moral**

Furthermore, the valence changes of social group traits are depicted in Figure 4, along with statistical correlation and significance values. Notably, 11 out of 18 social groups exhibit significant slopes of positivity or negativity over time. Seven groups, including Old, Young, Thin, Rich, Poor, Vagabond, and Backpacking, demonstrate increasing valence over the last decade (all P-values<0.05). Conversely, Asian, Hispanic, Men, and Women show negative changes over time, with Asian displaying a particularly strong correlation (-0.718). Seven social groups exhibit stable valences over time, including White, Black, Irish, Fat, America, Democrats, and Conservative, suggesting minimal changes in stereotypes among these groups on Reddit over the last decade. These results underscore the utility of RedditEM for computational social science.

5. Conclusion

Currently, there is a notable absence of large-scale and enduring diachronic word embeddings derived from social network texts for studying semantic changes and social computation. Addressing this gap, we introduce RedditEM in this study, offering a comprehensive diachronic word representation sourced from Reddit comment texts spanning a period of 12 years (2010-2021) across 144 months. Leveraging cosine distance and neighborhood words, we gauge the shifts in lexical semantics post-alignment of vectors facilitated by RedditEM. The experimental outcomes underscore the efficacy of RedditEM. Moving forward, we envisage utilizing RedditEM to delve deeper into research domains such as social group stereotype analysis and conceptual knowledge mining.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Daniel Braun. Tracking semantic shifts in german court decisions with diachronic word embeddings. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 218–227, 2022.

Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119, 2022.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145, 2017.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. Novel2vec: Characterising 19th century fiction via word embeddings. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016*, 2016.

Xiaobo Guo, Yaojia Sun, and Soroush Vosoughi. Emotion-based modeling of mental disorders on social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 8–16, 2021.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access, 2016a.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, 2016b.

Chaolin Huang, Ye ming Wang, Xing wang Li, Lili Ren, Jianping Zhao, Y. Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jia'an Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rong meng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang, and Bin Cao. Asymptomatic infection of covid-19 and its challenge to epidemic prevention and control. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi*, 41 12: 1985–1988, 2020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5): 905–949, 2019.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635, 2015.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174, 2012.

Tingting Liu, Lyle H Ungar, Brenda Curtis, Garrick Sherman, Kenna Yadeta, Louis Tay, Johannes C Eichstaedt, and Sharath Chandra Guntuku. Head versus heart: social media reveals differential language of loneliness from depression. *npj Mental Health Research*, 1 (1):16, 2022.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, page 21260, 2008.

Huansheng Ning, Hang Wang, Yujia Lin, Wenxi Wang, Sahraoui Dhelim, Fadi Farha, Jianguo Ding, and Mahmoud Daneshmand. A survey on the metaverse: The state-of-the-art, technologies, applications, and challenges. *IEEE Internet of Things Journal*, 2023.

Dean Peabody. Selecting representative trait adjectives. *Journal of personality and social psychology*, 52(1):59, 1987.

Nilo Pedrazzini and Barbara McGillivray. Machines in the media: semantic change in the lexicon of mechanization in 19th-century british newspapers. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 85–95, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Radim Řehřek and Petr Sojka. Software framework for topic modelling with large corpora. 2010.

Alexander Robertson, Farhana Ferdousi Liza, Dong Nguyen, Barbara McGillivray, and Sc ott A. Hale. *Semantic Journeys: Quantifying Change in Emoji Meaning from 2012-2018.* 2021.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

P Shoemark, LF Ferdousi, D Nguyen, H Scott, and B McGillivray. Monthly word embeddings for twitter random sample (english, 2012–2018). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Zenodo*, 2019.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), pages 66–76. Association for Computational Linguistics, 2020.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. Measuring, predicting and visualizing short-term change in word representation and usage in vkontakte social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 672–675, 2017.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational approaches to semantic change*. Language Science Press, June 2021.

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. Mining the uk web archive for semantic change detection. Recent Advances in Natural Language Processing, 2019.

Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu, and Barbara McGillivray. Dukweb, diachronic word representations from the uk web archive corpus. *Scientific Data*, 8(1):269, 2021.

Isaac Waller and Ashton Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268, 2021.