Toward Sentiment Aware Semantic Change Analysis

Roksana Goworek[1] **Haim Dubossarsky**[1,2]
[1] Queen Mary University of London
[2] Language Technology Lab, University of Cambridge
{r.goworek, h.dubossarsky}@qmul.ac.uk

Abstract

This student paper explores the potential of augmenting computational models of semantic change with sentiment information. It tests the efficacy of this approach on the English SemEval of Lexical Semantic Change and its associated historical corpora. We first establish the feasibility of our approach by demonstrating that existing models extract reliable sentiment information from historical corpora, and then validate that words that underwent semantic change also show greater sentiment change in comparison to historically stable words. We then integrate sentiment information into standard models of semantic change for individual words, and test if this can improve the overall performance of the latter, showing mixed results. This research contributes to our understanding of language change by providing the first attempt to enrich standard models of semantic change with additional information. It taps into the multifaceted nature of language change, that should not be reduced only to binary or scalar report of change, but adds additional dimensions to this change, sentiment being only one of these. As such, this student paper suggests novel directions for future work in integrating additional, more nuanced information of change and interpretation for finer-grained semantic change analysis.

1 Introduction

Lexical Semantic Change is a crucial aspect in the study of linguistics, offering insights into how the meanings of words evolve over time. This phenomenon reflects the dynamic and ever-changing nature of language, revealing how cultural, societal, and historical contexts influence linguistic expression. Current research primarily detects semantic change either as a binary classification (whether a word's meaning has changed between two corpora) or as graded change scores (the extent of meaning change). Despite the nuanced analysis of different types of semantic change that has existed in historical linguistics research for many years, current approaches in NLP are still lagging behind (Hengchen et al., 2021). Analysis typically involves comparing cosine distances between word embeddings across corpora from different time periods. The two prevalent methods are APD (Average Pairwise Distance) (Kutuzov and Giulianelli, 2020; Schlechtweg et al., 2018), and PRT (Inverted cosine similarity over word prototypes) (Kutuzov et al., 2022).

Semantic change detection can be analysed using type-based approaches (Schlechtweg et al., 2020; Kutuzov and Giulianelli, 2020), analyzing shifts in semantic vector spaces, or using token-based methods. BERT-based (Devlin et al., 2018) and ELMo-based (Peters et al., 2018) models have been very commonly used due to their high semantic encoding abilities and possibility for further fine-tuning for specific tasks (Laicher et al., 2021). Using contextualised word embeddings for semantic change detection (Giulianelli et al., 2020) relies on the distributional hypothesis, the assumption that words with similar meanings share contexts. The current state-of-the-art (SOTA) model, XL-LEXEME by Cassotti et al. (2023), demonstrates exceptional accuracy in the SemEval-2020 Task 1 on multiple languages. Specifically it achieves a 0.757 Spearman semantic change rank correlation between predictions and graded scores on the English dataset, by creating comparable context-based word embeddings.

Over the years, the volume of research in this area has expanded significantly. This is largely due to the increasing availability of large digital text corpora, such as the SemEval dataset, and advances in natural language processing (NLP) techniques. More recently, the integration of machine learning, particularly deep learning models like BERT,

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 350–357

has revolutionized the field (Laicher et al., 2021; Beck, 2020). The primary objective of these advancements has been to improve the accuracy of detecting and quantifying semantic shifts. The secondary objective has been to expand the task to more languages, which SemEval has successfully addressed by providing datasets for English, German, Latin and Swedish (Schlechtweg et al., 2020). However, semantic change of language is multifaceted, extending beyond binary and graded one-dimensional classifications. It includes various forms like broadening, narrowing, complete shifts, and notably amelioration (shift to positive connotation) and pejoration (shift to negative connotation), both particularly noteworthy for their direct impact on sentiment. This complexity necessitates a more nuanced approach to semantic analysis which includes finer-grained semantic change classification, recognizing that words can shift along a spectrum of meanings and connotations influenced by diverse cultural and societal factors (Hengchen et al., 2021). Even if not explicitly undergoing amelioration or pejoration, by changing meaning and hence the context it appears in, a word may change in its associated sentiment. Sentiment analysis is a highly-researched task in NLP, providing a lot of publicly-available resources which can be used to enhance semantic change detection models beyond detecting a shift in the distribution of word embeddings, by additionally considering a shift in their associated sentiment.

This research explores the interplay between semantic and sentiment change, as it parallels amelioration and pejoration, a major aspect of semantic change (Closs Traugott, 1985). Our findings indicate that even subtle semantic shifts can affect sentiment. Hence, the goal is to refine semantic change detection by integrating sentiment change analysis, using sentiment model data to improve upon existing semantic change detection models.

2 Related Work

The few studies exploring finer-grained semantic change in recent literature have taken various paths. Cook and Stevenson (2010) focused on identifying the most polar words in different corpora and analyzing their changes without directly linking them to semantic change in general. Research on large-scale sentiment change, such as the studies by Xie et al. (2020) and Fernández-Cruz and Moreno-Ortiz (2023), did not specifically investigate the role of sentiment change information in semantic change detection. Some innovative methods for more interpretable semantic change detection have emerged, like the approach by Giulianelli et al. (2023) that clusters tokens into interpretable word senses using definition generation. Additionally, Giulianelli et al. (2021) proposed a method based on grammatical profiling, focusing on morphosyntactic behavior changes, offering an alternative perspective in detecting semantic shifts. More closer to our study is a recent work that tried to enrich models for semantic change by first fine-tuning them on a range of NLP tasks (Zhou et al., 2023). However, they did not directly take sentiment score into account in their models like this study does. And, to the best of our knowledge, no work has attempted to enhance semantic change analysis through sentiment change analysis.

3 Methodology

In order to determine the sentiment change of words we used five publicly-available BERT-based models, fine-tuned for sentiment analysis sourced from huggingface.co. The models, named for simplicity, are not the official "BERT" and "RoBERTa" but are based on these architectures. Note that other models might also be based on these architectures, but were fine-tuned differently: BERT[1], SST BERT[2], sbcBI[3], RoBERTa[4], Reviews[5]. The models differ in their training data and some in architecture, they were selected precisely to ensure the results are robust regardless of the sentiment model used, as long as it passes the later-described validation test we developed for sanity checks. We also use VADER (Hutto and Gilbert, 2014) as another sentiment analysis model. VADER is a rule-based sentiment analysis tool, which combines a dictionary of sentiment-laden words with a set of rules that consider grammatical and syntactical conventions for expressing sentiment. We added this to our analysis to demonstrate that

[1]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment, Accessed November 2023

[2]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english, Accessed November 2023

[3]https://huggingface.co/sbcBI/sentiment_analysis_model, Accessed November 2023

[4]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest, Accessed November 2023

[5]https://huggingface.co/juliensimon/reviews-sentiment-analysis, Accessed November 2023

the findings of this study hold even for sentiment analysis models not based on contextualised language models.

Sentiment scores of all models were standardized to be in the range of 0 (most negative) to 1 (most positive). We stress that sentiment scores are assigned to each sentence, and when word level analysis is done the averaged sentiment score is used. When models produce distribution of sentiment scores across binary or ordinal categories (e.g., assigned .25 weight for 0 and .75 for 1) the weighted average is computed. This method provided a uniform approach to quantify sentiment across various models. We analysed the sentiment shift of sentences from the SemEval-2020 Task 1 corpora (Schlechtweg et al., 2020), focusing on the 37 target words with binary and graded semantic change scores, derived from human-annotated semantic word in context similarity judgments. Corpus 1 consists of $\tilde{6}$ million tokens from 1810-1860, and corpus 2 is also made up of $\tilde{6}$ million tokens from 1960-2010, both based on the Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) dataset. Note that target words were tagged with their pos tags in this dataset.

To maintain fairness and accuracy in comparing sentiment differences for a word between Corpus 1 and Corpus 2, we analyzed an equal number of sentences from each corpus for every word. Specifically, we used the smaller number of sentences containing the target word found in either corpus. For the corpus with a larger number of sentences for the target word, we randomly sampled an equal number of sentences to match the other corpus. This sampling involved first extracting all sentences with the target word and then randomly selecting the same number of sentences (as in the smaller set) using a random seed of 42.

4 Validation of Sentiment Models

To ensure the validity of our approach, we first verify that our evaluation of sentence sentiment is robust to the inherent noise associated with different sentiment scores produced by different models. If a model captures sentiment reliably, then the agreement of a word sentiment should be greater within each model than between different models. We created random splits for each corpus (A-B splits). Then for each model, we computed the average sentiment for each word in each split (A or B), and then computed the correlation between the averaged sentiment scores of A and B, both within each model, and between models. We posit that if models provide reliable sentiment scores, than the correlation between two independent splits should be higher within a model relative to different models.

Table 1 shows that all models demonstrate much higher correlations between sentiment scores of the same model relative to other models. This indicates a high level of agreement of the models' sentiment scores. However, the medium range correlation scores across different models also indicate that different models show a fair amount of agreement as well. Overall, all six models were deemed suitable for subsequent sentiment analysis.

5 Results - Semantic and Sentiment Change

The Mean Sentiment Change of a word w is measured as the absolute difference of the mean sentiment of sentences containing the word in corpus 1 and corpus 2:

$$\Delta S_w = \left| \frac{1}{N} \sum_{i=1}^{N} \mathcal{S}_{C_1,w(i)} - \frac{1}{N} \sum_{j=1}^{N} \mathcal{S}_{C_2,w(j)} \right| \quad (1)$$

Where $\mathcal{S}_{C.,w}$ is the sentiment score of the word w as it appears in a single sentence in a particular corpus.

As can be seen in Table 2, this sentiment change is greater for changed than stable words, which confirms the hypothesis that semantic change is associated with sentiment change.

As can be seen in Table 3, this result also emerges on the German SemEval-2020 Task 1 dataset, which has the same format as the English dataset, and contains 48 target words with binary and graded scores of semantic change. The experiment set up was the same as for English, except with the following five huggingface models:

	C1 A-B correlations		C2 A-B correlations	
Model	Within	Avg Cross-Model	Within	Avg Cross-Model
BERT	0.860	0.481	0.792	0.341
SST BERT	0.832	0.412	0.877	0.302
sbcBI	0.849	0.394	0.835	0.309
RoBERTa	0.931	0.467	0.918	0.514
Reviews	0.901	0.260	0.925	0.119
VADER	0.829	0.419	0.801	0.418

Table 1: Within and average cross-model correlation in A-B splits in the two corpora.

	BERT	SST BERT	sbcBI	RoBERTa	Reviews	VADER
Changed	0.048	0.086	0.070	0.041	0.051	0.059
Stable	0.044	0.073	0.069	0.032	0.032	0.053

Table 2: Mean Sentiment Change of Semantically Changed and Stable Words in English

BERT[6], twitter[7], gereval[8], news[9], sentiment[10], fine-tuned for sentiment analysis in German. All five models passed the sanity checks described above.

We can inspect the degree and direction of the sentiment change of each word, as shown in Figure 1, and observe that the word "record", for example, has undergone the most amelioration (i.e., positive change), according to the Reviews model, which could be explained by the fact that it has evolved from being associated with documented information to musical records, which are more associated with entertainment, hence the more positive sentiment. Meanwhile the word "prop" has developed more negative connotations (i.e., pejoration), which can be due to its newly developed association with artifice and superficiality, particularly in entertainment, politics, and media, where it implies manipulation or a lack of authenticity, as it began to be used less for denoting a physical supporting object to more metaphorical usages. Such detailed analysis of semantic change holds the potential to categorize words that have evolved over time into distinct types of change, such as amelioration and pejoration, while also quantifying its extent within each semantic aspect.

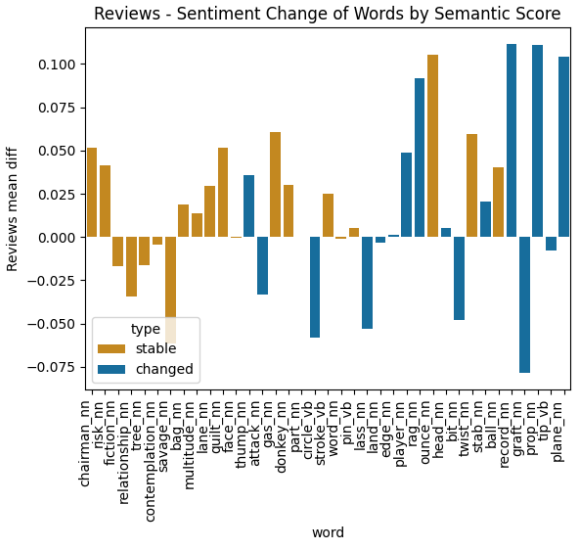

Figure 1: Mean Sentiment Change of Words from Corpus 1 to Corpus 2, Scored by the Reviews Model

6 Integrating sentiment to semantic change models

Now we attempt to use this sentiment change information to improve the performance of a baseline semantic change detection model, based on APD distances, described in §1. We experiment with two models; Logistic Regression to predict the binary semantic change labels, and Linear Regression to predict the graded semantic change scores. We use these models to test our hypothesis that integrating sentiment information with semantic change information can improve overall semantic change prediction accuracy. We use the predictions from

[6]https://huggingface.co/oliverguhr/german-sentiment-bert, Accessed November 2023

[7]https://huggingface.co/JP040/bert-german-sentiment-twitter, Accessed November 2023

[8]https://huggingface.co/deepset/bert-base-german-cased-sentiment-Germeval17, Accessed November 2023

[9]https://huggingface.co/mdraw/german-news-sentiment-bert, Accessed November 2023

[10]https://huggingface.co/aari1995/German_Sentiment, Accessed November 2023

	BERT	twitter	gereval	news	sentiment
Changed	0.019	0.021	0.002	0.037	0.046
Stable	0.015	0.020	0.001	0.031	0.046

Table 3: Mean Sentiment Change of Semantically Changed and Stable Words in German

an APD-based model as the performance baseline (Kutuzov and Giulianelli, 2020), which we try to improve upon by adding sentiment change information to the models. The sentiment change ratings are derived by means of various statistical analyses of the distributions of sentiment scores of sentences containing the target word derived as follows:

- Mean Diff: Absolute difference between the average sentiment scores of two corpora.
- T-test: Statistical test comparing sentiment scores between the two corpora to assess significance of their mean difference.
- Ratio: Ratio of average sentiment scores of the two corpora, calculated with the smaller average as the numerator.
- KL Divergence: Measures the divergence in sentiment score distributions between the two corpora.

For the logistic regression model, we balanced the two classes of words by removing 5 stable words, resulting in a dataset of 16 changed and 16 stable words. This approach was adopted because fitting the model with imbalanced class sizes led to a classifier bias, where it predominantly chose the larger class. Because we use the existing SemEval dataset to train our model, it cannot be used as a test set. Instead, for evaluation we report averages of 6-fold cross validation in order to make the results based on this small sample as reliable as possible.

The performance of the logistic regression model is the accuracy of its binary predictions, while the performance of the linear regression model is the Spearman's rank correlation of the model's predicted values with the true semantic change scores.

Using only the semantic change APD model's predictions as input to the models, logistic regression achieves 0.56 accuracy, while linear regression achieves 0.61 correlation.

Results show (Table 4) that adding sentiment information does not reliably improve the accuracy of logistic regression model. Out of 20 model combinations, only one exhibits improved accuracy. In contrast, linear regression shows mixed results for integrating sentiment information (Table 4), improving the results in 8 out of 20 conditions, and by a large margin.

Although sentiment change is evidently related to semantic change, as our results for both English and German demonstrate (see Table 2, and Table 3), it seems that integrating sentiment information is not straightforward. In its simplest form of a single metric of sentiment change between two corpora, sentiment does not systematically improve the performance of semantic change detection models. This lack of regularity could be related to statistical variations rather than reliable results which would generalise to another dataset.

The specific conditions in which sentiment does contribute to semantic change needs to be further explored. One of the potential reasons behind this could be the lack of numerous data points for meaningful analysis. To address this point we look to the DWUGs dataset of sentence pairs (Schlechtweg et al., 2021).

7 Considering Sentence Pairs

The DWUGs dataset (Schlechtweg et al., 2021) consists of sentence pairs containing the same target words as the SemEval dataset, annotated with word sense-similarity judgments by human annotators. A rating of 4 is assigned to sentence pairs where the target word is used in exactly the same sense, while a rating of 1 indicates that the word is used in very different senses in the two sentences. A rating of 0 is used for uncertain or ambiguous cases. This dataset offers a more granular insight into sentiment change, contrasting with the limitations of a single average sentiment change score per word offered by the SemEval dataset. The same sentiment models were used to get the sentiment scores of sentences in the pairs.

As seen in Table 5, there is a small correlation of sentence pair sentiment differences with the semantic sense judgments. A deeper examination of the sentence pair sentiment differences for words with the highest and lowest average sense

Statistic	BERT		SST BERT		sbcBI		RoBERTa		Reviews		VADER	
Regression:	Log	Lin	Log	Lin	Log	Lin	Log	Lin	Log	Lin	Log	Lin
Mean diff	0.56	0.51	0.56	0.59	0.56	0.53	0.56	**0.66**	0.56	**0.72**	0.56	0.59
T-test	0.46	0.60	0.51	**0.64**	0.43	**0.64**	0.56	0.56	0.54	0.52	0.46	**0.62**
Ratio	0.56	0.51	0.51	0.60	0.56	0.56	0.56	**0.65**	**0.59**	**0.75**	0.56	0.59
KL divergence	0.56	0.59	0.45	**0.68**	0.56	0.56	0.56	0.60	0.56	0.58	0.56	0.57

Table 4: Average k-fold accuracies for Logistic Regression (Log) and Spearman Correlations with Linear Regression Predictions (Lin). Bolded results outperform baseline (0.56 and 0.61, respectively).

Judgment	BERT	SST BERT	sbcBI	RoBERTa	Reviews	VADER
0	**0.228**	**0.456**	**0.294**	0.204	0.283	0.238
1	0.218	0.449	0.280	0.199	**0.266**	0.245
2	0.217	0.445	0.273	**0.212**	0.255	**0.259**
3	0.222	0.427	0.274	0.209	0.243	**0.259**
4	0.212	0.422	0.267	0.201	0.238	0.251

Table 5: Average Sentiment Difference for Each Judgment Group
The highest value in each column is boldened, while the lowest value is underlined.

similarity (see Figure 2) reveals distinct patterns. Notably, sentences with consistent semantic usage of a target word tend to have smaller sentiment differences compared to sentences where the word's usage is more semantically varied. However, this pattern does not uniformly apply across all words, as the distributions of semantically changed and stable words are not easily distinguishable for most cases.

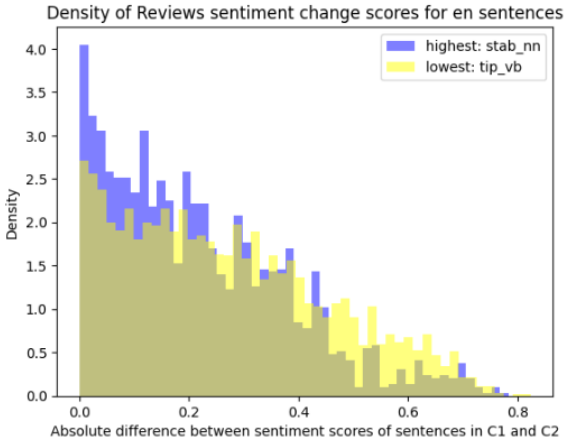

Figure 2: Sentence Pair Sentiment Differences for Least and Most Sense-Stable Words

We further explored the impact of relative sentiment change. We can see which words changed sentiment in a similar way to others and which words diverged in the sentiment of their usages by calculating the Jensen Shannon distances between the sentiment difference distributions of all words. We look at various statistics of these Jensen Shannon distances of a word to all the other words, specifically the min, max, mean and standard deviation, to examine whether these relative comparisons of sentiment difference distributions could improve semantic change detection. The same experiments using linear and logistic regression as described previously were carried out, using these statistics. The logistic regression model failed to learn anything, and to improve upon its baseline accuracy, regardless of the sentiment change statistic used as an additional input feature, hence it is not reported in a table. The results of the linear regression model experiments can be found in Table 6.

Similarly to the previous experiments, the performance improvements in the simple semantic change detection models are scarce and irregular for the linear regression model. As a result, the hypothesis that sentiment change information can improve performance of semantic change detection models is not supported. The differences between sentiment scores of two sentences are also related to the sense similarity of the usage of the target word in those sentences, however, this measure of sentiment change, based on many data points, also cannot be used to improve the performance of semantic change detection models.

We propose several reasons for this negative find-

Statistic	BERT	SST BERT	sbcBI	RoBERTa	Reviews	VADER
Means	**0.62**	0.61	0.54	0.60	0.59	0.58
Minima	0.61	0.58	0.51	**0.63**	0.59	0.51
Maxima	**0.66**	0.57	0.53	0.58	**0.62**	0.56
Standevs	0.58	**0.62**	0.53	**0.69**	0.59	0.60

Table 6: Logistic Regression Results Using DWUG Sentence Pairs, Baseline = 0.61

ing. To begin with, neither amelioration nor pejoration are the most common types of semantic change. It could be that most words in the SemEval do not explicitly undergo these changes, which in turn is reflected by no change to the words' sentiment. Second, and related to the above, this effect is also related to the small number of words, 37, that exist in the English SemEval, which is still too small for meaningful analysis. Third, the average sentiment of sentences in which a word appears may not be the optimal method to evaluate the sentiment of individual target words, because it may contain too much noise from the sentence to provide any valuable insights about the sentiment shift of that word.

8 Limitations & Future Research

This study acknowledges several limitations in its approach to measuring sentiment change of individual words. Currently, there is no established methodology for assessing such sentiment changes. Unlike in semantic change detection, where clustering usages into word senses across corpora is a common practice, sentiment change analysis lacks similarly sophisticated methods. The approach adopted in this study, focusing on the average sentiment of sentences in which a target word appears, may not fully capture the nuanced sentiment contributions of the word itself. Future research should aim to develop more intricate techniques that specifically evaluate the sentiment contribution of a word within its sentence context.

Another limitation is the potential narrow applicability of our method. It may be best suited as a supplementary tool to refine and elaborate on semantic changes already detected by SOTA models. Sentiment shift may not be relevant for words whose semantic shifts do not necessarily entail amelioration or pejoration.

Additionally, our method's approach to quantifying sentiment differences—by taking the absolute difference of sentiment scores between two sentence usages of a word—represents a rather simplistic estimate. This approach is somewhat analogous to measuring semantic differences by computing the cosine distance between sentence embeddings, which is a relatively basic and possibly insufficient method for assessing nuanced semantic shifts in word usage. As such, our findings must be interpreted within the context of this methodological simplicity, and future studies should explore more refined approaches for enhancing semantic change analysis with sentiment change information, as it may be a promising avenue of research, given more appropriate methods for evaluating sentiment change of a word between two corpora.

Further research in this area could enable trend analysis in digital humanities, provide insights into societal and cultural shifts by examining how word sentiments evolve. Additionally, it could aid in monitoring language changes, reflecting evolving societal attitudes and behaviors.

9 Conclusion

This paper provides evidence that sentiment change is associated with semantic change: Words that are deemed to change semantically (according to expert human annotators) also show greater change to their sentiment, on average. However, the hypothesis that sentiment information could be used to aid the task of semantic change detection ends with a null result on the 37 words from the SemEval English dataset.

The results confirm that words that change meaning are also more likely to change their associated sentiment, even if they didn't directly undergo amelioration or pejoration. However, this sentiment shift, in the simple ways we measured it, is not a reliable estimate of semantic change.

In summary, while our study provides valuable insights into the relationship between semantic and sentiment changes, it also highlights the need for more advanced methodologies in this emerging area of research.

10 Acknowledgments

This work was partially funded by the research program Change is Key!, supported by Riksbankens Jubileumsfond (reference number M21-0021).

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte Im Walde. 2020. Ccoha: Clean corpus of historical american english. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966.

Christin Beck. 2020. Diasense at semeval-2020 task 1: Modeling sense change via pre-trained bert embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585.

Elizabeth Closs Traugott. 1985. On regularity in semantic change.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Javier Fernández-Cruz and Antonio Moreno-Ortiz. 2023. Tracking diachronic sentiment change of economic terms in times of crisis: Connotative fluctuations of 'inflation'in the news discourse. *Plos one*, 18(11):e0287688.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernandez. 2020. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2021. Grammatical profiling for semantic change detection. *arXiv preprint arXiv:2109.10397*.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. *arXiv preprint arXiv:2305.11993*.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In *Computational approaches to semantic change*, pages 341–372. Berlin: Language Science Press.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Andrey Kutuzov and Mario Giulianelli. 2020. Uio-uva at semeval-2020 task 1: Contextualised embeddings for lexical semantic change detection. *arXiv preprint arXiv:2005.00050*.

Andrey Kutuzov, Erik Velldal, and Lilja Ovrelid. 2022. Contextualized language models for semantic change detection: lessons learned. *arXiv preprint arXiv:2209.00154*.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving bert performance on lexical semantic change detection. *arXiv preprint arXiv:2103.07259*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. Dwug: A large resource of diachronic word usage graphs in four languages. *arXiv preprint arXiv:2104.08540*.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. *arXiv preprint arXiv:1804.06517*.

Jing Yi Xie, Renato Ferreira Pinto Jr, Graeme Hirst, and Yang Xu. 2020. Text-based inference of moral sentiment change. *arXiv preprint arXiv:2001.07209*.

Wei Zhou, Nina Tahmasebi, and Haim Dubossarsky. 2023. The finer they get: Combining fine-tuned models for better semantic change detection. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 518–528.