TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media

Daniel Loureiro[◇]
Aminette D'Souza[*◇], **Areej Nasser Muhajab**[*◇], **Isabella A. White**[*◇], **Gabriel Wong**[*◇]
Luis Espinosa Anke[◇♡], **Leonardo Neves**[♣], **Francesco Barbieri**[♣], **Jose Camacho-Collados**[◇]
[◇] Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
[♣] Snap Inc., Santa Monica, California, USA; [♡]AMPLYFI, UK
[◇]`boucanovaloureirod@cardiff.ac.uk`, `cardiffnlp.contact@gmail.com`

Abstract

Language evolves over time, and word meaning changes accordingly. This is especially true in social media, since its dynamic nature leads to faster semantic shifts, making it challenging for NLP models to deal with new content and trends. However, the number of datasets and models that specifically address the dynamic nature of these social platforms is scarce. To bridge this gap, we present TempoWiC, a new benchmark especially aimed at accelerating research in social media-based meaning shift. Our results show that TempoWiC is a challenging benchmark, even for recently-released language models specialized in social media.

1 Introduction

One of the most studied challenges in NLP is lexical ambiguity. Solutions include word sense disambiguation (Navigli, 2009) or entity linking (Ling et al., 2015), where words are linked to sense inventories such as WordNet (Miller, 1995) or Wikipedia. Recently, notable progress has been made with the advent of Language Models (LMs) and contextualized embeddings, crucially well equipped for modeling meaning in context (Pilehvar and Camacho-Collados, 2020; Bevilacqua et al., 2020).

One notable limitation with current lexical semantics benchmarks, however, is that they are typically clean and time-invariant, where standard grammar is the norm, and have little to no account of language usage in real-world platforms like social media. However, there is ample agreement that modeling changes in language and topic distributions is crucial for modern NLP (Loureiro et al., 2022a). Thus, there is a rich body of literature concerned with, e.g., adapting existing word representations (mainly word embeddings) diachronically (Hamilton et al., 2016; Szymanski, 2017; Rosenfeld and Erk, 2018; Hofmann et al., 2021), exposing LMs to time-specific data (Lazaridou et al., 2021), or temporal adaptation in general (Luu et al., 2021; Agarwal and Nenkova, 2022; Jin et al., 2021; Loureiro et al., 2022a). The lack of real-world data to serve as ground truth has typically limited the evaluation of diachronic word-level NLP models. This limitation has been addressed in a myriad of ways, e.g., by comparing distributional similarities with human judgments (Gulordava and Baroni, 2011), contrasting change vs. frequency (Hamilton et al., 2016), comparing time-sensitive representations with stripped-down versions (Frermann and Lapata, 2016) or, more recently, determining whether a word has acquired new senses over time by looking at relatively large targeted subcorpora (Van Hee et al., 2018), or probing for diachronic awareness in settings reminiscent of knowledge base completion (Dhingra et al., 2021; Hofmann et al., 2021).

Despite the above, few works have attempted to model the connection between meaning shift and social media. Among them, Del Tredici et al. (2019) showed that trending words are a meaningful signal for predicting meaning shifts. This platform-specific insight is the main basis for the construction of our dataset. TempoWiC follows the simple formulation from the SuperGLUE Word-in-Context (WiC) challenge (Pilehvar and Camacho-Collados, 2019), which is particularly well suited for temporal meaning shift evaluation given that it is not reliant on a reference sense inventory. This change of paradigm has seen wide adoption, with multilingual extensions such as XL-WiC (Raganato et al., 2020), Am2ico (Liu et al., 2021) or MCL-WiC (Martelli et al., 2021), or reformulations such as WiC-TSV (Breit et al., 2021). In contrast to these, TempoWiC is crucially designed around meaning shift and instances of word usage tied to Twitter trending topics. As an additional contribution, along with the benchmark, we provide a set of robust baselines and analyses that highlight the challenging nature of the task.

[*] Annotation team.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3353–3359
October 12–17, 2022.

Tweet 1	Tweet 2	Label
2019-02 "I ain't gone let the ppl *frisk* me if I'm dirty homie"	*2020-02* "Set up a stop and *frisk* outside a white club and catch coke heads"	T
2019-04 "i wish i still had images of my old animal crossing new leaf *villager* he was good boy"	*2020-04* "How does *villager* trading in New horizons even work like tf"	T
2019-08 "This dude just said "Boys of the Backstreet" He made em sound like a whole *folklore*"	*2020-08* "my target app said they didn't have *folklore* cds but when i went inside they had some i'm so happy"	F
2019-08 "In case you were wondering facial devotion still worked with a face *mask* on"	*2020-08* "With these *mask* at work customers are forever confusing me and Reyna lmao"	F

Table 1: Examples from the training set of TempoWiC. Target words in *italic*. The label T (True) indicates that the word has the same meaning in the two tweets, the label F (False) indicates that the meaning is different.

2 TempoWiC: Temporal Word in Context

In this section, we describe our process to build our evaluation benchmark for detecting meaning shift in social media. The task is framed as a simple binary classification problem in which a target word is present in two texts (tweets) posted during different time periods. The goal is to decide whether the meaning corresponding to the first target word in context is the same as the second one or not. Table 1 lists a few examples.

2.1 Data collection

Word Selection. Since this work focuses on meaning shift, we do not consider neologisms and use lemmas from WordNet as an initial set of potential words of interest (82K lemmas, ignoring multi-word expressions, stopwords and numbers). From a corpus of 100M tweets collected from the Twitter API for the period between the start of 2019 and September of 2021, we compiled monthly frequency counts for this set of known words, and computed trending scores following Chen et al. (2021). Each trending word peak is estimated as the day with highest frequency during the year/month with most occurrences. As the prior date, we considered the same date exactly one year before. This is done in order to avoid seasonal confound factors, which are known to affect models in social media (Chae et al., 2012; Barbieri et al., 2018). Afterwards, we selected the top 10 words with highest trending scores from each month, resulting in 210 words which are candidates for annotation. For this selection, we ignored words with fewer than 100 occurrences in our corpus during their peak date.

Obtaining Paired Tweets. In this phase we collected, for each trending word, 100 tweets posted during the peak date, and 100 tweets posted during the prior date. For this phase we used the Twitter APIs, setting filters to request only English tweets, and ignoring replies and retweets.[1] We preprocessed each tweet using spaCy (Honnibal et al., 2020) and we randomly paired tweets from the prior and peak sets for specific words that match both in surface form and part-of-speech tag.

2.2 Annotation

Annotators. We recruited four annotators through our internal institution recruitment office. This ensured that the annotators were part of the process, trained and understood all the details of the task.[2] Annotators, who were all native or near-native English speakers, were all paid the equivalent of a research assistant per hour.

First stage. The annotation was split into two phases. First, we took a breadth-first approach in which a relatively short number of instances (i.e., 10) of a large number of the selected words (210 in total) were annotated. The motivation for this initial phase was to understand which words had some sort of meaning shift to start with. The selection of the words to be included in the dataset was then restricted to words which had more than 3 out of 10 instances with meaning shift.

[1] We retrieved additional tweets with the APIs as we included trending words that had a minimum occurrences of 100 tweets on peak day in our initial dataset, but some words included less than 100 tweets on the prior date.

[2] The full guidelines provided to the annotators are available in the task website.

	Word	# Instances	Trending	Agreement
		(% Diff. Meaning)	Date	(Krippendorff's α)
Train	frisk	99 (54%)	11/2/2020	0.718
	pogrom	99 (5%)	25/2/2020	0.482
	containment	100 (33%)	12/3/2020	0.274
	virus	96 (48%)	12/3/2020	0.254
	epicenter	100 (71%)	14/3/2020	0.124
	ventilator	99 (17%)	27/3/2020	0.541
	villager	100 (64%)	10/4/2020	0.546
	turnip	100 (95%)	10/5/2020	0.316
	bunker	98 (61%)	1/6/2020	0.408
	mask	99 (76%)	14/7/2020	0.255
	teargas	98 (3%)	18/7/2020	0.786
	paternity	100 (22%)	30/7/2020	0.289
	entanglement	99 (89%)	1/8/2020	0.623
	folklore	82 (92%)	3/8/2020	0.917
	parasol	100 (85%)	2/9/2020	0.446
Validation	impostor	99 (76%)	23/9/2020	0.544
	lotte	98 (43%)	27/9/2020	0.514
	recount	100 (28%)	6/11/2020	0.682
	primo	100 (77%)	9/11/2020	0.528
Test	milker	99 (50%)	4/3/2021	0.699
	moxie	97 (83%)	5/3/2021	0.755
	unlabeled	100 (90%)	10/3/2021	0.711
	pyre	100 (32%)	27/4/2021	0.243
	gaza	100 (60%)	15/5/2021	0.749
	ido	91 (83%)	27/5/2021	0.712
	airdrop	99 (40%)	6/6/2021	0.918
	bullpen	99 (9%)	16/6/2021	0.388
	crt	100 (68%)	26/6/2021	0.867
	monet	98 (94%)	8/7/2021	1.000
	burnham	100 (16%)	1/8/2021	0.964
	delta	100 (100%)	11/8/2021	1.000
	gala	100 (46%)	14/9/2021	0.498
	launchpad	99 (81%)	17/9/2021	0.558
	vanguard	99 (95%)	21/9/2021	0.421

Table 2: Details all the words included in TempoWiC. Maximum pairwise agreement is reported.

Second stage. The second phase was based on a depth-first approach in which 100 instances of all selected words from the first stage were annotated. We ensured that each instance was annotated by three annotators. The final label attributed to each instance was determined by majority vote.[3]

2.3 Statistics and Inter-annotator Agreement

The outcome of our annotation pipeline is a dataset of 3,297 instances divided in train/validation/test sets of size 1,428/396/1,473 instances, respectively. We measured inter-annotator agreement using Fleiss' Kappa at 0.446, and using Krippendorff's α at 0.439. Since each instance is assigned a majority vote label, we also computed the maximum pairwise Krippendorff's α at 0.627, which should be more revealing of the expected performance on this task. Words with Krippendorff's α below 0.1 were removed from the dataset. Table 2 provides a summary of the most relevant details of the dataset after annotation.

3 Evaluation

In this section, we report baseline results on TempoWiC using two different approaches which have proven successful on the WiC task that inspired this work. More concretely, we report results based on pretrained LMs using fine-tuning on the tweet pair as well as comparing the similarity of contextual embeddings.[4]

Evaluation metrics. The results are reported according to the standard Macro-F1 metric for multi-class classification problems. Accuracy is also reported for completeness but, given the unbalanced nature of the dataset, Macro-F1 should provide a more accurate representation of the performance.

3.1 Models

Our experiments include the following LMs: RoBERTa base and large pretrained on general domain corpora (Liu et al., 2019); RoBERTa base with continued training on tweets until the end of 2019, and a similar model trained with more tweets until the end of 2021 (Loureiro et al., 2022a, TimeLMs); LMs based on RoBERTa but trained from scratch on tweets, both base and large versions (Nguyen et al., 2020, BERTweet).

Each of these LMs is fine-tuned representing instances as "`<s>` Tweet 1 `</s>` Tweet 2 `</s>`", with each tweet represented by the encodings produced by each model's tokenizer.[5] Additionally, we trained a logistic regression classifier on the cosine similarity of the contextual embeddings corresponding to the target word on each tweet of the pair. This approach based on contextual embeddings is sensitive to the choice of layers from the LM used to represent embeddings. Following Loureiro et al. (2022b), we use SP-WSD layer pooling weights (model specific) that are suited for sense representation (see Appendix B for ablation results with alternative pooling strategies). Both

[3] Some words proved too difficult to reliably annotate according to feedback from the annotators as well as low agreement scores computed after annotation (more details in Section 2.3). Consequently, these words were removed from the dataset. Among the various challenges to be expected from annotating social media, mixed language (e.g., English and Hindi) was among the most frequent issues.

[4] Additional baselines are reported in Appendix A.

[5] We experimented concatenating the encodings for the target word at the end of the sequence as proposed by Wang et al. (2019) for WiC, but found no improvements.

	Model	Accuracy	Macro-F1
Fine-tuning	RoBERTa-base	66.89%	58.26%
	RoBERTa-large	66.49%	59.10%
	TimeLMs-2019-90M	66.46%	57.70%
	TimeLMs-2021-124M	65.04%	54.75%
	BERTweet-base	61.46%	51.27%
	BERTweet-large	67.93%	60.62%
Similarity	RoBERTa-base	67.96%	52.89%
	RoBERTa-large	72.98%	67.09%
	TimeLMs-2019-90M	**74.07%**	**70.33%**
	TimeLMs-2021-124M	71.01%	63.51%
	BERTweet-base	69.45%	65.16%
	BERTweet-large	69.18%	56.95%
Naive	Random	50.00%	50.00%
	All True	36.59%	26.79%
	All False	63.41%	38.80%

Table 3: Main results on the test set of TempoWiC. Fine-tuning results are the average of 3 runs.

Word	Accuracy		Macro-F1	
	Fine-tune	Similarity	Fine-tune	Similarity
airdrop	40.48%	**65.31%**	30.13%	**65.18%**
bullpen	**67.68%**	38.38%	**44.53%**	34.54%
burnham	27.67%	**83.00%**	27.62%	**69.15%**
crt	64.98%	**79.80%**	46.26%	**73.24%**
delta	85.37%	**98.98%**	46.05%	**49.74%**
gala	47.67%	**78.00%**	36.71%	**77.86%**
gaza	67.01%	**69.39%**	**66.74%**	64.61%
ido	83.88%	**90.11%**	63.04%	**75.78%**
launchpad	**81.82%**	80.81%	45.00%	**59.26%**
milker	46.13%	**64.65%**	37.28%	**63.44%**
monet	93.54%	**94.90%**	48.33%	**62.96%**
moxie	**89.35%**	68.04%	**74.75%**	56.48%
pyre	**64.29%**	51.02%	**62.78%**	50.84%
unlabeled	65.67%	**76.00%**	47.71%	**57.48%**
vanguard	**95.96%**	73.74%	**68.80%**	42.44%

Table 4: Performance by word on the TempoWiC test set. Using TimeLMs-2019-90M as the best similarity model, and BERTweet-large as the best fine-tuning model (average of 3 runs).

approaches are implemented with Wolf et al. (2020, Transformers).

3.2 Results

Our results on Table 3 show that TempoWiC is a challenging task with room for improvement. While the best results using both fine-tuning and similarity approaches are obtained by models adapted to the Twitter domain, this advantage isn't substantial over generic RoBERTa. Interestingly, we find that the straightforward similarity approach manages to substantially outperform fine-tuning, with a Twitter base model trained with data before any word's trending peak achieving the best performance. While this result may be surprising considering that fine-tuning performs better on WiC, this finding is in line with recent work in word sense disambiguation showing that approaches based on contextual embeddings can be more robust and generalizable than fine-tuning (Loureiro et al., 2021).

Analysis by word. Table 4 provides a detailed breakdown of the results of the best performing model (i.e., TimeLMs-2019-90M) by individual words. As can be observed, there are large differences between words, which are also due to the unbalanced natural distribution of certain words to start with (see Table 2). More interesting is perhaps the gaps between fine-tuning and similarity techniques. While similarity appears to be generally more robust, in words such as *bullpen* or *vanguard*, the tendency is reversed.

3.3 Future Work

This work only covers English, but future work should include additional languages and experiment with both multilingual and monolingual models. We leave an analysis explaining the difference between the 2019 and 2021 models for future work as well, alongside the development of methods that leverage the dates provided with each instance towards improved performance, similarly to Dhingra et al. (2021); Rosin et al. (2022).

4 Conclusion

This work introduced a new lexical semantics task and a dataset, TempoWiC, focused on meaning shift detection in Twitter. While meaning representation is at the core of the task, the challenges of this task go beyond simple word sense disambiguation with a focus on its temporal aspect. To make the task realistic, we extracted Twitter trending words for different periods and paired them with tweets from past periods. This makes the task more challenging and grounded in real-world applications for social media. We performed extensive experiments with standard meaning representation approaches based on language models. The results show that the task leaves ample room for improvement, with several avenues for future research on how to better integrate time-aware social media models with meaning representation techniques. The TempoWiC dataset and baseline scripts are available at github.com/cardiffnlp/tempowic.

Acknowledgements

Jose Camacho-Collados and Daniel Loureiro are supported by a UKRI Future Leaders Fellowship. We thank Snap Inc. for providing resources to support this project.

References

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Francesco Barbieri, Luis Marujo, Pradeep Karuturi, William Brendel, and Horacio Saggion. 2018. Exploring emoji usage and prediction through a temporal variation lens. *arXiv preprint arXiv:1805.00731*.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221.

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

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online. Association for Computational Linguistics.

Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE conference on visual analytics science and technology (VAST)*, pages 143–152. IEEE.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2021. Mitigating temporal-drift: A simple approach to keep NER models crisp. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 163–169, Online. Association for Computational Linguistics.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2021. Time-aware language models as temporal knowledge bases. *arXiv preprint arXiv:2106.15110*.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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, Berlin, Germany. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6970–6984, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2021. Lifelong pretraining: Continually adapting language models to emerging corpora. *arXiv preprint arXiv:2110.08534*.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, et al. 2021. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*.

Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022a. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022b. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305:103661.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, 47(2):387–443.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2021. Time waits for no one! analysis and challenges of temporal misalignment. *arXiv preprint arXiv:2111.07408*.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *NAACL*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2020. Embeddings in natural language processing: theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA. Association for Computing Machinery.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Additional Baselines

Besides the fine-tuning and similarity methods described in the main paper, we also experimented with additional approaches in order to better understand the difficulty of this dataset.

In this appendix we provide additional results based on an MLP trained with concatenated contextual embeddings (Table 5), and another MLP trained with the concatenation of the average of static embeddings from each tweet (Table 6). Embeddings are *L2* normalized after concatenation.

The hyper-parameters used with these MLPs were determined by grid search on 24 different configurations which were tested on the validation set. The parameters tested were *hidden layer sizes* ((*embedding size*, 100) or (100)), *solver* (adam or sgd), *batch size* (32 or 64), and *maximum number of iterations* (50, 100 or 200).

Static embeddings are based on fastText (Bojanowski et al., 2017) and learned from Twitter data on the same corpora used for Loureiro et al. (2022a). These embeddings are trained with skip-gram for 300-dimensions, min-ngram size 2 and max-ngram size 12.

Model	**Accuracy**	**Macro-F1**
RoBERTa-base	68.11% (0.77%)	55.00% (1.88%)
RoBERTa-large	**68.82%** (1.06%)	55.17% (2.60%)
TimeLMs-2019-90M	68.68% (0.68%)	**58.62%** (0.81%)
TimeLMs-2021-124M	68.23% (0.90%)	57.55% (3.57%)
BERTweet-base	65.51% (0.91%)	57.03% (3.56%)
BERTweet-large	66.55% (0.59%)	54.16% (0.09%)

Table 5: Performance of MLP trained with concatenation of the target word's contextual embeddings (SP-WSD pooling), tuned on the validation set. Reporting average of 3 runs, and standard deviation.

Model	**Accuracy**	**Macro-F1**
CommonCrawl	55.53% (0.78%)	48.98% (0.57%)
TimeLMs-2019-90M	**57.64%** (0.71%)	49.46% (0.28%)
TimeLMs-2021-124M	55.20% (0.31%)	**52.30%** (0.57%)

Table 6: Performance of MLP trained with concatenation of the average of static embeddings from each tweet. Reporting average of 3 runs, and standard deviation.

B Pooling Contextual Embeddings

Table 7 reports results using the similarity method described in the main paper with alternative choices for layer pooling. We considered the final layer and the sum of the last 4 layers as these are common choices in word sense disambiguation settings.

Model	**Final Layer**	**Sum Last 4**	**SP-WSD**
RoBERTa-base	40.89%	**60.33%**	52.89%
RoBERTa-large	38.80%	53.32%	**67.09%**
TimeLMs-2019-90M	59.93%	67.69%	**70.33%**
TimeLMs-2021-124M	53.14%	60.26%	**63.51%**
BERTweet-base	67.35%	**66.91%**	65.16%
BERTweet-large	38.80%	41.75%	**56.95%**

Table 7: Performance of Contextual Similarity method according to choice of layer pooling approach.