Integrating Transformers and Knowledge Graphs for Twitter Stance Detection

Thomas Hikaru Clark[♠], **Costanza Conforti**[♠], **Fangyu Liu**[♠], **Zaiqiao Meng**[♠], **Ehsan Shareghi**[♣♠], **Nigel Collier**[♠]
[♠]Language Technology Lab, University of Cambridge
[♣]Department of Data Science and AI, Monash University
[♠]{thc44, cc918, fl399, zm324, nhc30}@cam.ac.uk
[♣]ehsan.shareghi@monash.edu

Abstract

Stance detection (SD) entails classifying the sentiment of a text towards a given target, and is a relevant sub-task for opinion mining and social media analysis. Recent works have explored knowledge infusion — supplementing the linguistic competence and latent knowledge of large pre-trained language models with structured knowledge graphs (KGs), yet few works have applied such methods to the SD task. In this work, we first perform stance-relevant knowledge probing on Transformers-based pre-trained models in a zero-shot setting, showing these models' latent real-world knowledge about SD targets and their sensitivity to context. We then propose novel knowledge-enriched stance detection models. We evaluate them on two Twitter stance datasets, achieving state-of-the-art performance on both.

1 Introduction

Stance detection (SD) involves identifying a text's stance towards a given target (for example, whether a tweet is supportive, against, or neutral towards *Joe Biden*). This is a challenging task with downstream use cases in opinion mining, fake news detection, and rumor verification (Küçük and Can, 2018; Fake News Challenge, 2017; Conforti et al., 2020). A major challenge for SD is the need for knowledge of current events and other fast-changing facts about the world.

Large pre-trained Transformer models trained on vast corpora (Vaswani et al., 2017; Devlin et al., 2019) have blurred the line between language models and knowledge bases, as shown by their performance on benchmarks like LAMA, which measure static factual knowledge (Petroni et al., 2020a; Radford et al., 2019). Recent works in SD have capitalized on the Transformer architecture; however, it remains uncertain how to adapt these models to the constantly shifting factual landscape found in SD tasks, for example in political tweets. At the same time, knowledge infusion (KI) approaches have had success in integrating KGs with Transformers for question-answering (QA) tasks, but there is a shortage of work on KI for SD.

Our contributions are as follows: 1) We perform stance-relevant knowledge probing on Transformers-based pre-trained models, showing these models' partial real-world knowledge and sensitivity to context, and 2) We train and evaluate knowledge-enriched stance detection models on two Twitter stance datasets, achieving state-of-the-art performance on both.

2 Previous Work

The original author baseline for the SemEval-16 SD task used an SVM classifier on hand-crafted features (Mohammad et al., 2017). More recent approaches for SD have achieved better performance using transfer learning with Transformer models, but without adding knowledge infusion (Ghosh et al., 2019; Schiller et al., 2021; Kaushal et al., 2021). This typically involves concatenating a tweet and target and feeding it into a Transformer model with a classification layer attached. To our knowledge, Kawintiranon and Singh (2021) is the only SD work using "knowledge enhancement", but this approach was based on identifying stance-signaling words rather than using KGs.

Most attempts to augment Transformer models with structured knowledge from KGs have focused on QA tasks, such as CommonsenseQA (Talmor et al., 2019), not SD. Bian et al. (2021) used ConceptNet (Speer and Havasi, 2013) to extract knowledge descriptions relating entities in each question to entities in each answer choice via multi-hop reasoning on a KG, with a BERT-based classifier to choose the best answer. Similarly, K-BERT (Liu et al., 2019) enriches entities in an input sentence based on lookup in a KG. The success of these methods suggests that downstream tasks can benefit from contextual priming, where the same input

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 304–312

supplemented with additional factual context leads to better predictions. This can be contrasted with approaches like K-Adapter and KnowBERT, which infuse knowledge by modifying the model architecture, rather than by adding context to the input (Wang et al., 2020; Peters et al., 2020). One advantage of contextual priming is the ability to leave a model's architecture largely unchanged, requiring only a method of collecting and generating useful factual context.

3 Probing Transformers for Stance-Relevant Knowledge

In this section, we seek to establish a lower bound for the stance-relevant knowledge already present in Transformer models before doing any knowledge infusion. Rather than testing recall of encyclopedic facts, we probe whether models make stance-related inferences regarding real-world entities in a human-like way. The three models we test are RoBERTa-Base (Liu et al., 2020), RoBERTa-Large, and Twitter-RoBERTa (Barbieri et al., 2020). RoBERTa is pre-trained on a large internet corpus, including news articles, while Twitter-RoBERTa is trained on $\sim 58M$ tweets, making these models good candidates for political SD on Twitter. Using each model in a masked language modeling setting, we feed it a sentence with a single word replaced by the special `[MASK]` token, returning a probability distribution over all vocabulary tokens. Rather than using an automatically-generated knowledge benchmark like LAMA (Petroni et al., 2020b), we draw on a human-in-the-loop paradigm (Nie et al., 2019) and manually design probes relevant to the SD task. Examples involve public figures and political issues that appear as targets in major SD datasets (Mohammad et al., 2017; Grimminger and Klinger, 2021), such as *Donald Trump*, *Hillary Clinton*, and *climate change*. We evaluate the models using **Accuracy@1**, which is the percentage of prompts for which the highest-probability token generated by a model is appropriate in context, e.g. factually correct or aligned with reasonable inferences. Some key example pairs are shown below:

(1) The Proud `[MASK]`, a far-right group, held a rally.

(2) The protests were sparked by the killing of George `[MASK]`, an unarmed black man in Minneapolis, Minnesota.

In example (1) above, all models were able to predict "Boys" for the masked token. In example (2), all models predicted "Zimmerman" rather than the desired "Floyd"; this illustrates how models can quickly become outdated when new words and names enter common usage after a model is pretrained.

(3) I think that climate change is such a `[MASK]`. Save the earth!

(4) I think that climate change is such a `[MASK]`. Drill, baby, drill!

For the above pair of probes, we test how the presence of pro- and anti-environmentalist slogans impact stance-relevant predictions. Both RoBERTa models have reasonable predictions, outputting "threat" and "problem" for (3) and "hoax" for (4). This shows how the models can sometimes leverage stance-relevant knowledge to make better predictions. This is not always the case, however:

(5) Joe has a bumper sticker that reads 'Drill Baby Drill'. He thinks climate change is a `[MASK]`.

(6) Joe has a bumper sticker that reads 'Drill Baby Drill'. He thinks climate change is `[MASK]`.

For (5), all models predict "hoax", which initially looks like a good stance-aware inference. For (6), however, all models predict "real". The deletion of a single article, "a", caused all models to make a stance-incongruent prediction.

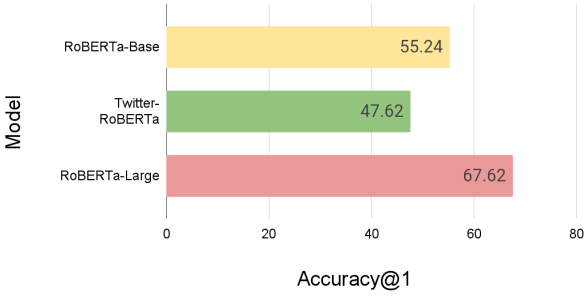

Figure 1: Accuracy@1 on a test set of 105 handpicked stance-relevant knowledge probes, showing the superior performance of RoBERTa-Large compared to the other models. The correctness of the answers was manually labeled by the first author.

We include additional examples of prompts in the *Appendix*, as well as a breakdown of probing performance by model in Figure 1. The key points of our probing analysis are that a) Transformer

models contain latent stance-relevant knowledge that can inform SD tasks, b) RoBERTa-Large beats smaller models, and c) stance-aware predictions can often be overruled by context. The results establish a promising lower bound on the stance-relevant knowledge of LMs, yet a significant gap remains. We therefore propose a way of using KGs to infuse knowledge specifically for SD tasks.

4 Knowledge Infusion for SD

4.1 Basic Knowledge Infusion

A KG is described by a list of triples of the form (e_1, r, e_2), where e_1 and e_2 are entities (nodes) linked by the relation (edge) r. To leverage such knowledge, we follow the intuition that short descriptions of unfamiliar entities may operate as a form of contextual priming. This is supported by the knowledge probing literature, as well as works like AUTOPROMPT (Shin et al., 2020) which learn to construct optimal contextual triggers for eliciting knowledge from a LM.

Given a tweet, we use the `spaCy` entity linker (Honnibal and Montani, 2020), which identifies spans in a text that refer to entities from the Wikidata KG (Vrandečić and Krötzsch, 2014). `spaCy` can identify different forms of an entity, and outputs short descriptions of any found entities. We then generate short knowledge descriptions of the form "[Entity], [Description]" for all entities found in a tweet. For example, a tweet containing the string 'Putin' would be paired with the following description: "Vladimir Putin, 2nd and 4th President of Russia". These descriptions are prepended to the tweet along with the stance detection target, separated from the tweet string by a special separator token. This enrichment process is done for both the training data and the testing data. We report results for this approach in Section 5.

4.2 Custom Knowledge Graph Construction and Pathfinding

The previous approach operates as a form of knowledge lookup, but does not exploit the informative relations between entities that may be contained in a KG. Prior works have exploited multi-hop knowledge paths within a KG to improve NLU performance (Bian et al., 2021), an approach we now apply to SD. To reduce the computational cost of finding knowledge pathways, we propose a customized collection approach of filtering for Wikidata triples within a small number of hops from the target entities of the SD datasets. Additionally, we include Wikidata triples relating to trending tropes, which we identify by finding words and collocations with a high temporal concentration in a large Twitter dataset spanning the timeframe of the SD task. Collocations are identified using Pointwise Mutual Information (PMI). Examples of trending tropes are provided in the *Appendix*. In short, we use several strategies to limit the size of the KG while keeping the entities and relations most likely to help with SD.

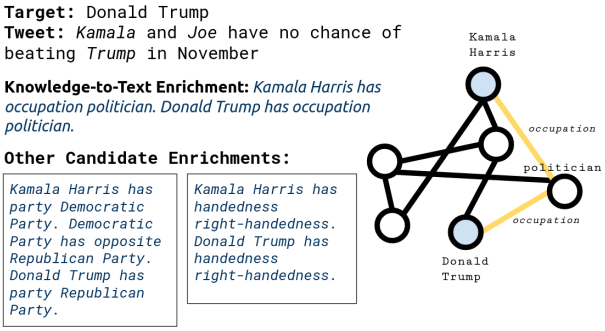

Figure 2: Illustration of our proposed path-based knowledge infusion method.

To infuse knowledge, we use our custom KG to find knowledge pathways connecting entities in a tweet to the SD targets. Since there are many possible pathways between two nodes in a KG, we limit paths to length 3 and choose the minimum cost path. We initially assign edge costs using a random walk strategy, which penalizes knowledge paths through less informative hub nodes. An example of low and high informativeness pathways found by the random walk strategy is shown below, reflecting the intuition that two people both holding the occupation President of the United States is a more informative relation than two people both working in Washington, D.C.

(7) *High Informativeness*: Donald Trump held the position of President of the United States. President of the United States has officeholder Joe Biden.

(8) *Low Informativeness*: Donald Trump has work location Washington, D.C. Presidential transition of Joe Biden has headquarters location Washington, D.C.

We turn any found knowledge pathways into natural language knowledge descriptions that are prepended to the tweet. The example shown in Figure 2 shows how this approach could plausibly

improve SD performance. Suppose a tweet mentions the entity "Kamala", but the model has not been exposed to many instances of this entity in its training data. The SD task is to determine the tweet's stance towards Donald Trump. Using a KG, the model establishes a knowledge pathway from Kamala Harris to Donald Trump, reflecting the knowledge that both are politicians.

4.3 Edge Cost Tuning

A major problem for our knowledge infusion approach is finding informative multi-hop knowledge paths. While the random-walk edge weighting method is a first step, it is highly dependent on the properties of the KG being traversed. Secondly, this method does not take advantage of the available training data to improve the estimates of edge cost. As a result, we propose a method called *Edge Cost Tuning* (ECT) for using the available training data to test KG edges for informativeness.

ECT builds on the previous path-based knowledge infusion model, using it to evaluate the helpfulness of various knowledge paths. For each tweet in the training set, our model finds the lowest-cost knowledge path from the target to an entity in that tweet. Both an enriched and unenriched version of the tweet are fed to the model. If the enriched version causes the model to assign a higher probability to the correct label than the unenriched version, the costs for all links along that knowledge path are reduced. Otherwise, the costs for all links along that knowledge path are increased. This causes unhelpful edges in the KG to accumulate high costs, while helpful edges are promoted. Importantly, this procedure has an interpretable result, as edges in the KG can be sorted by their change in cost to understand which pieces of context were most helpful or unhelpful in making stance predictions. The *Appendix* contains before-and-after examples of the ECT method.

5 Experiments

We consider two Twitter datasets for SD: SemEval-16 (Mohammad et al., 2017) and Grimminger & Klinger (G&K) (Grimminger and Klinger, 2021). The first involves a range of controversial political targets, such as abortion, atheism, and the 2016 U.S. presidential election. The task is to predict a class label from among {*favor, against, neither*}. The dataset contains 2914 training examples and 1249 test examples. The second centers exclusively

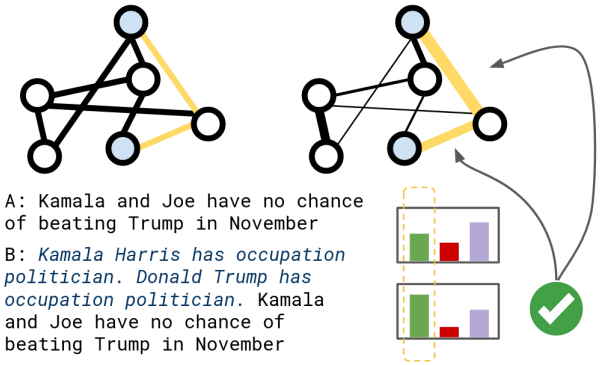

Figure 3: In this synthetic example, the tweet mentions Kamala Harris, who is connected to Donald Trump through a two-hop knowledge path. This knowledge path is tested against a no-enrichment baseline. Since the knowledge enrichment leads to a higher probability for the correct label, the corresponding edges in the KG are strengthened.

on the 2020 U.S. presidential election, and the prediction classes are *against, favor, neither, neutral*, and *mixed*. The dataset contains 2400 training examples and 600 test examples. Most SD models in the literature use SemEval-16 as a benchmark, and recent works have used BERT to achieve new state-of-the-art performance (Ghosh et al., 2019; Kaushal et al., 2021; Schiller et al., 2021; Kawintiranon and Singh, 2021).

Our general architecture for SD with Transformers involves the target (plus optional knowledge enrichments) and the tweet being concatenated with an intervening separator token before being fed into a RoBERTa model with a classification head. The weights of the entire network are updated during training. The architecture is very similar to that used in other Transformer-based SD models (Ghosh et al., 2019; Schiller et al., 2021; Kaushal et al., 2021). We compare our knowledge infusion models with base models fine-tuned on the same task data, as well as with K-Adapter, described in Section 2 (Wang et al., 2020).

As reported in Table 1, the best model for the SemEval-16 task was RoBERTa-Large with entity enrichment, while the best model for the G&K task was RoBERTa-Large with path enrichment and edge cost adjustment. One possible explanation for this is that ECT may work best when all examples for a task share a common topic (e.g. the 2020 election), as opposed to SemEval-16, which had 5 heterogeneous targets. Using a single KG with a single set of edge costs for such a task seems to underperform enrichment via direct entity lookup.

	SemEval-16		Grimminger & Klinger			
Model	**Acc.**	**F1**	**Ag.**	**Fav.**	**Nei.**	**Neu.**
Baseline*	-	69.1	79.0	89.0	95.0	53.0
RoBERTa-Base	71.8	71.4	81.8	91.4	93.9	60.0
Twitter-RoBERTa	71.4	71.7	82.7	90.0	94.5	56.4
K-Adapter	74.5	74.8	86.1	93.2	94.1	63.8
RoBERTa-Large	76.9	77.3	86.9	92.2	93.6	62.7
+ Entities	**77.2**	**78.5**	86.9	92.9	94.6	65.1
+ Paths	75.7	76.1	86.8	92.5	95.2	**68.3**
+ ECT	75.1	76.2	**87.0**	**93.7**	**96.0**	67.2

Table 1: Results for SemEval-16 Task: Mean Accuracy and F1 Scores (mean of Favor and Against labels). Results for Grimminger & Klinger Task: Mean F1 Scores by label: Against, Favor, Neither, and Neutral. The Mixed label was exceedingly rare in the dataset and no model ever predicted it, so all F1 scores for the Mixed label were 0. ECT = Edge Cost Tuning. *Baselines refer to author baselines from original SemEval and G&K papers.

The following are some examples from the G&K task that the knowledge-infused SD model with ECT correctly predicted while the unenriched model failed (knowledge descriptions are in italics):

(9) @realDonaldTrump It's today! The day I go to the polls and vote for Joe Biden and Kamala Harris. *donald trump held the position of president of the united states. president of the united states has officeholder joe biden.* (Correct Label: against Donald Trump)

(10) Trump must have stock in Regeneron. *stock, financial instrument. Donald Trump, 45th and current president of the United States. Regeneron Pharmaceuticals, pharmaceutical company.* (Correct Label: against Donald Trump)

(11) Biden or Kamala won't commit to their policy on packing the court, Joe's comment, "vote for me I'll let you know?" On fracking Joes flip flopping, Kamala is against fracking! VOTE RED! *donald trump has member of political party republican party. republican party has color red* (Correct Label: favor Donald Trump)

While there were also examples where the generated knowledge descriptions were irrelevant or noisy, these examples demonstrate how an appropriate knowledge description can improve downstream model performance. Examples (9) and (11) illustrate the utility of multi-hop reasoning, adding context relating Donald Trump to Joe Biden and the color red, respectively. Example (10) illustrates the usefulness of backing off to simple entity-enrichment in cases where no knowledge paths exist, providing useful additional context about the company Regeneron.

6 Conclusion

In this work, we highlighted three key points based on knowledge probing: Transformer models contain latent stance-relevant knowledge, RoBERTa-Large is better at this than the other models, and models can be misled by sentence context. We also established new state-of-the-art performance on two SD datasets using knowledge infusion. We introduce a novel method, *Edge Cost Tuning*, that uses training data to re-weight the connections in a knowledge graph, which produced best results on one of the two SD tasks. Our approach depends greatly on choice of KG and edge cost weighting method, so future work can explore additional ways of filtering for informative edges in a KG.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1.

Fake News Challenge. 2017. Fake News Challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance Detection in Web and Social Media: A Comparative

Study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11696 LNCS.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection.

Matthew Honnibal and Ines Montani. 2020. spacy. https://github.com/explosion/spaCy.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889, Online. Association for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An SVM-based approach.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019. K-BERT: Enabling language representation with knowledge graph.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in Tweets. *ACM Transactions on Internet Technology*, 17(3).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding.

Matthew E. Peters, Mark Neumann, Robert L. Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2020. Knowledge enhanced contextual word representations. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020a. How context affects language models' factual predictions.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020b. Language models as knowledge bases? In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners | Enhanced Reader. *OpenAI Blog*, 1(8).

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance Detection Benchmark: How Robust is Your Stance Detection? *KI - Kunstliche Intelligenz*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts.

Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A Large Semantic Network for Relational Knowledge.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020. K-ADAPTER: Infusing knowledge into pre-trained models with adapters.

A Trending Topics Identification

We hypothesize that knowledge of trending topics is important for political SD on Twitter for two main reasons: a) especially in the social media domain, trending topics can be very stance-signaling and b) pre-trained models will typically not have latent knowledge of these tropes because they are too temporally concentrated to be well-represented in the pre-training data. We implement a simple strategy for detecting tropes, based on three assumptions:

(1) Trending tropes will be relatively frequent n-grams.

(2) Trending tropes will be highly non-uniform in their distribution over time.

(3) Multi-word tropes will behave like collocations, with high pointwise mutual information between words.

For a given SD task, we sample a large selection of tweets from the same timeframe as the SD data (summer 2015 for SemEval-16, autumn 2020 for G&K). Within this sample, we choose uni-, bi-, and tri-grams that fit the above criteria. Figure 4 shows a sampling of discovered trending topics for the G&K SD task, each accompanied by a histogram of its occurrence over time.

B Results of Edge Cost Tuning

At the end of edge cost tuning, edges in the graph will have either lower, higher, or the same costs as before. Looking at the results, we see that the adjustments generally align with intuition. For example, in the G&K dataset, the triple `(politician, occupation_, Kamala Harris)` had one of the biggest decreases in cost after adjustment. This makes sense, because she may not have been a very prominent entity in the RoBERTa training data, but rose to much higher prominence in 2020 as Joe Biden's running mate. The decreased cost for that triple indicates that injecting this piece of knowledge generally helped predictions, while the lowered cost ensures that this piece of knowledge will be highly accessible to the model when evaluated on test data.

C Stance-Relevant Knowledge Probes

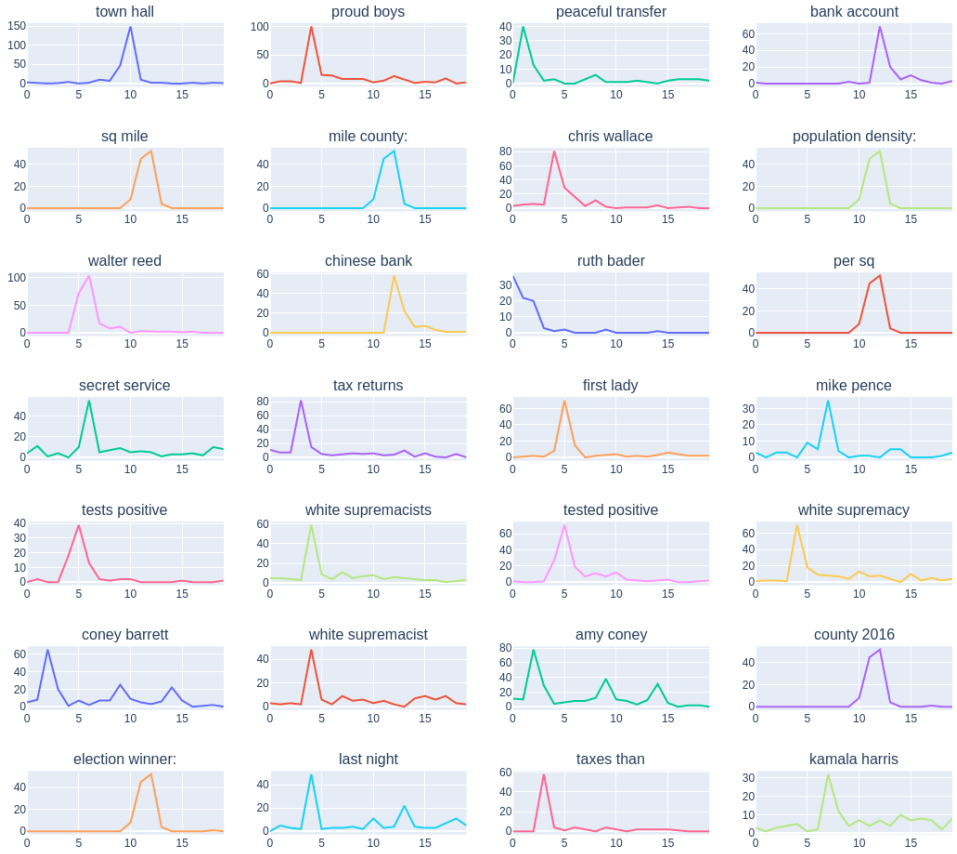

Figure 4: Examples of top bigram tropes for the G&K time frame on Twitter, restricted to tweets that mention Donald Trump.

Subject	**Relation**	**Object**
donald trump	occupation	politician
donald trump	position held	president of the ...
president of the ...	officeholder	joe biden
politician	occupation_	kamala harris
donald trump	member of political...	republican party
donald trump	award received	time person of the year
republican party	opposite of	democratic party
...	...	...
donald trump	award received	wwe hall of fame
trump tower	occupant	trump family
golf	country of origin	scotland
government	manifestation of	power
donald trump	significant person_	2021 storming of the...
2021 storming of the ...	instance of	demonstration

Table 2: Results of Edge Cost Tuning on G&K training data, arranged from biggest decrease in cost (top) to biggest increase in cost (bottom).

Prompt	RoBERTa-Large	RoBERTa-Base	Twitter-RoBERTa
Donald Trump spoke at a pro [MASK] rally on the anniversary of Roe v. Wade.	life	life	life
Hillary Clinton spoke at a pro [MASK] rally on the anniversary of Roe v. Wade.	abortion	life	abortion
Donald Trump is a member of the [MASK] Party.	Republican	Republican	Republican
Hillary Clinton is a member of the [MASK] Party.	Democratic	Democratic	Democratic
Ireland passed a referendum which repealed the constitutional ban on [MASK].	abortion	abortion	abortion
The Proud [MASK], a far-right group, held a rally.	Boys	Boys	Boys
Amy Coney [MASK] was nominated to the Supreme Court by President Trump.	Barrett	Barrett	worth
I believe in God, guns, and the Bible. Feminism is [MASK].	bullshit	bullshit	bullshit
I believe in LGBT rights and women's rights. Feminism is [MASK].	important	wrong	bullshit
I believe in God, guns, and the Bible. We should [MASK] be feminists.	all	not	all
I believe in LGBT rights and women's rights. We should [MASK] be feminists.	all	all	all
The remarks were delivered by [MASK], the 45th President of the United States.	Trump	Trump	Trump
The remarks were delivered by [MASK], the 44th President of the United States.	Obama	Trump	Obama
The remarks were delivered by [MASK], the 43rd President of the United States.	Bush	Obama	Obama
Kamala [MASK], the vice-presidential candidate, delivered a speech on Monday.	Harris	Harris	Harris
The protests were sparked by the killing of George [MASK], an unarmed black man in Minneapolis, Minnesota.	Zimmerman	Zimmerman	Zimmerman
The protests were sparked by the killing of Eric [MASK], an unarmed black man in New York City.	Garner	Garner	Garner
Please watch the documentary about melting ice caps. Climate change is [MASK].	real	real	real
Please watch the documentary about the global warming hoax. Climate change is [MASK].	real	real	real
I'm an atheist. The Bible is a book of [MASK].	lies	lies	God
I'm a Catholic. The Bible is a book of [MASK].	faith	faith	God

Table 4: A subset of the prompts used for stance-relevant knowledge probing. The prompts were manually chosen to relate to target entities from the SemEval and G&K SD tasks.