Casting the Same Sentiment Classification Problem

Erik Körner **Ahmad Dawar Hakimi** **Gerhard Heyer** **Martin Potthast**

Leipzig University
erik.koerner@uni-leipzig.de

Abstract

We introduce and study a problem variant of sentiment analysis, namely the "same sentiment classification problem", where, given a pair of texts, the task is to determine if they have the same sentiment, disregarding the actual sentiment polarity. Among other things, our goal is to enable a more topic-agnostic sentiment classification. We study the problem using the Yelp business review dataset, demonstrating how sentiment data needs to be prepared for this task, and then carry out sequence pair classification using the BERT language model. In a series of experiments, we achieve an accuracy above 83% for category subsets across topics, and 89% on average.

1 Introduction

At the sixth argument mining workshop ArgMining 2019 (Stein and Wachsmuth, 2019), the same side stance classification problem has been introduced by Stein et al. (2021) as a shared task to the argument mining community. Identifying the stance of an argument towards a topic is a fundamental problem in computational argumentation. The task presents a new problem variant, namely to classify whether two arguments share the same stance without the need to identify the stance itself. The underlying hypothesis is that this can be achieved in a topic-agnostic manner since only the similarity of two given arguments needs to be assessed. Similarly, in the authorship analysis community, the authorship verification problem (Koppel and Schler, 2004) is the task of determining for a given pair of texts whether they have been written by the same author. Here, too, instead of classifying a given text into predefined author classes, as is the case with authorship attribution, the verification problem casts the problem as a pairwise similarity-based classification task.

In this paper, we recast sentiment analysis in the same manner: Given two texts of unknown sentiment polarity, determine whether their sentiment is the same. Unlike for the same side and the same author classification problems, which suffer from a lack of large-scale training data, due to many resources available for sentiment analysis, scaling up does not prove to be a problem for the same sentiment problem. We see three major contributions in studying this task variant: (1) Focused research on topic-agnosticity, enabling direct observations of the effect of topic and that of agnostic modeling. (2) Potentially easing generalization across domains. (3) In time, a new paradigm of approaches may emerge (whereas the prevailing one still rules today). Our contributions are as follows: We demonstrate how to prepare standard sentiment data for meaningful training and evaluation, introduce an approach based on the transformer neural network architecture where we adapt the sequence pair classification task to the same sentiment problem, and evaluate our model in various experiments.[1] In what follows, Section 2 reviews related work, Section 3 introduces our approach and explains the dataset and its preparation, and Section 4 reports on our evaluation.

2 Related Work

Sentiment analysis has a wide range of applications in many languages and a variety of methods were developed to refine results and adapt to use-cases (Feldman, 2013; Terán and Mancera, 2019). Its main task is to determine the opinion or attitude of an author, either a single person or a group, about something, be it a product, brand, or service (Tedmori and Awajan, 2019). It has importance for businesses, in campaigns, and the financial sector, among others, and as a result, it has undergone

[1] Code and data: https://github.com/webis-de/EMNLP-21

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 584–590

much research to improve accuracy using different models and forms of data representation.

In recent years, sentiment analysis is increasingly being performed using deep learning approaches (Zhang et al., 2018). Johnson and Zhang (2017) designed a deep pyramid CNN which could efficiently represent long-range associations in text and thus more global information for better sentiment classification. Howard and Ruder (2018) have developed ULMFiT, a simple efficient transfer learning method that achieves improvements for various NLP Tasks such as sentiment classification. Another model that performs well on sentiment classification is BERT (Devlin et al., 2019), where pre-trained language models can be fine-tuned without substantial effort to suit different tasks. Sun et al. (2019) showed that decreasing the learning rate layer-wise and further pre-training enhance the performance of BERT. Another approach from Xie et al. (2019) improves the performance of BERT with the usage of data augmentation. It was shown that another current language model XLNet (Yang et al., 2019) achieves the best results for the sentiment classification task.

Based on the idea of the same side stance classification task by Stein et al. (2021) as well as the authorship verification problem (Koppel and Schler, 2004), our underlying hypothesis is that the more complex single sentiment problem may be able to be simplified to the semantic similarity of sentiment text pairs. This can then reduce the demand for topic-specific sentiment vocabulary usage (Hammer et al., 2015; Labille et al., 2017). As there is no prior work about *same sentiment* classification, our work uses well-known approaches from semantic text similarity (STS) about which several shared tasks have been organized (Agirre et al., 2013; Xu et al., 2015; Cer et al., 2017) and a variety of datasets (Dolan and Brockett, 2005; Ganitkevitch et al., 2013) have been compiled. While prior approaches have employed syntactic, structural, and semantic similarity, to evaluate sentence similarity, single models have gained more popularity in recent times. Mueller and Thyagarajan (2016) show the application of siamese recurrent networks for sentence similarity. With the introduction of contextualized word embeddings, Ranasinghe et al. (2019) evaluate their impact on STS methods compared to traditional word embeddings in different languages and domains.

3 The Same Sentiment Problem

In the following, we will introduce our model for same sentiment prediction and explain how to prepare training and test data.

3.1 Sequence Pair Classification Model

Our approach is based on the sequence pair classification task using the well-known transformer language models. The classification model employs the standard pre-trained BERT model architecture (Devlin et al., 2019) with an additional classification layer, consisting of a dropout of 0.1 and a dense layer with sigmoid activation. This layer accepts a pooled vector representation from the model based on the last hidden state of the `[CLS]` token, the first token for each input sequence intended to represent the whole sequence.

We fine-tuned the publicly available pre-trained model `BERT-base-uncased` using pairs of same or different sentiments reviews, generated as described in the following Section 3.2, with a training, validation, and test split of 80:10:10. 512 to include both input sequences with almost no truncation. Batch sizes were dependent on GPU memory and model sequence length, so we used 32 samples per batch for a sequence length of 128, but only 6 for a length of 512. Gradient accumulation was used to account for the small batch sizes. We kept the Adam optimizer with a learning rate of $5e{-}5$ and epsilon of $1e{-}8$. Typically, but depending on the number of training samples, between 3 and 5 epochs of fine-tuning seem to be enough to reach a plateau with further epochs only marginally improving prediction accuracy. The best model setup trained for 15 epochs only added 1% of accuracy but may very well have lost its ability to generalize for unknown topics. We used a single output for binary classification with a sigmoid binary cross-entropy loss function as it performed better than two outputs for classes same or not same.

3.2 Data Acquisition and Preparation

For our analysis, we required texts with clear stances or sentiments, with both positive and negative samples about the same topic. As we wanted to do cross-topic comparisons, multiple topics with enough samples for standalone training or fine-tuning of a model were necessary.

Those requirements were fulfilled by the sentiment datasets from the business reviews of the Yelp Dataset Challenge (Asghar, 2016) and Ama-

zon product reviews (Ni et al., 2019).[2] The IMDb dataset[3] commonly used in sentiment analysis was not useful as it only contained both a single positive and negative review per movie, and was, therefore, more suited for sentiment vocabulary analysis.

We chose to focus on the Yelp business review dataset as it contains a variety of categories for cross evaluations and qualitatively better review texts compared to Amazon. The dataset is a snapshot with reviews not older than 14 days at its time of creation and is officially being provided as several JSON files from which we only used general business information, such as category, and the customer reviews with text and ratings. It contains 6,685,900 user reviews about 192,127 businesses,[4] in 22 main categories.[5] Businesses are mostly assigned a single main category with related subcategories and seldom overlap. Previous general examinations by Asghar (2016) show extreme variance of the number of reviews and businesses between categories. The reviews required no further textual preprocessing as transformer models use a SentencePiece tokenizer (Kudo and Richardson, 2018) to handle arbitrary text input. It should be noted that those models can only handle some predefined sequence lengths, so text sequences after tokenization will be truncated to fit. With a sequence length of 512, we were able to sufficiently cover most review pairs, as the average number of tokens was about 150 for a single review.

Training Data Generation: For the sequence pair classification, we matched random pairs of reviews about the same business. The star rating of 1 to 5 was translated into binary labels, *good* or *bad*, with reviews being considered *good* if their ranking was above 3 stars. We filtered out businesses that had less than 5 positive and negative reviews each. The remaining reviews were randomly combined per pair type, i. e. 2 – 4 sentiment pairs each for *good-good*, *good-bad*, *bad-bad*, and *bad-good*.[6] This, we will show in Section 4, sufficed to fine-tune the model, even if we omitted in some cases more than 10,000 reviews for specific businesses. The pair generation resulted in a balance of positive / negative reviews and also samples of same sentiment pairs (*good-good*, *bad-bad*) or not (*good-bad*, *bad-good*). The number of businesses varied much between each major category, so cross-category training data also varied in quantity.

[2] https://nijianmo.github.io/amazon/index.html

[3] https://ai.stanford.edu/~amaas/data/sentiment/

[4] https://www.yelp.com/dataset

[5] https://www.yelp.com/developers/documentation/v3/all_category_list

[6] Our compiled datasets (review ids of pairs and splits) is available at https://webis.de/data.html#webis-samesentiment-21. The actual reviews have not been included as using the Yelp dataset requires agreementto their Dataset License.

4 Evaluation

To thoroughly inspect our approach we conducted a series of experiments to test which hyperparameters are necessary to fine-tune a model in general, how well the model is able to generalize by artificially separating topics in training and evaluation, and how it performs for each category specifically.

Baseline As baseline models, we started with linear models, SVM, and Logistic Regression classifiers, where we represented reviews as n-gram count vectors, TF-IDF word vectors, and as Doc2Vec (Le and Mikolov, 2014) embeddings. Using count and TF-IDF vectors, we were only able to achieve about 50% accuracy. With Doc2Vec embeddings, our accuracy improved to about 57%. Those results most likely meant that those approaches were not a good fit for sentiment pair similarity prediction.

We then used a Siamese Recurrent Network architecture (Neculoiu et al., 2016; Mueller and Thyagarajan, 2016) that has been successfully applied to semantic textual similarity problems. Words were represented by pre-trained 50-dimensional GloVe (Pennington et al., 2014) embeddings. We set a maximum input sequence length of 256, 50 LSTM cells in both bidirectional LSTM layers and 50 hidden units.[7] Training plateaued at 15 epochs with 83% accuracy. We will use the same configuration in all the following experiments.

Overall performance Using BERT, we started with an initial sequence length of 128, batch size of 32, and 5 epochs of fine-tuning but otherwise standard parameter choices to see how the model performs in general. The dataset consisted of 2 sentiment pairs for all 4 pair combinations for each business with a train/dev/test split of 90:10:10. This achieved 81.3% accuracy overall. Increasing the sentiment pairs per business to 4 per type only increased accuracy to 82%, so the randomly chosen samples were enough to generally cover the dataset,

[7] For more details about dropout, etc. refer to our code.

Pairing	TN	FP	FN	TP	Acc.	Examples
bad-bad	–	–	2,719	14,892	84.6%	17,611
bad-good	15,533	2,098	–	–	88.1%	17,631
good-bad	15,248	2,345	–	–	86.7%	17,593
good-good	–	–	1,537	16,004	**91.2%**	17,541
all*	30,781	4,443	4,256	30,896	87.6%	70,376

Table 1: Test results per sentiment pair type. Dataset filtered with at least 10 reviews per business, 4 sentiment pairs generated per pair-type. Model: BERT-base-uncased, 256 SeqLen, 3 Epochs.

and more samples only increased time per epoch but not significantly the result. We achieved a final 89.1% accuracy (cf. Table 2) only by increasing the sequence length to 512 but decreasing the batch size to 6, to reduce truncating of longer input texts.

Per Sentiment Pair Type We further compared how a model trained on all sentiment pair types evaluates each type separately in Table 1. Using the first model setup with a shorter sequence length, all pair types achieved between 84.6% and 88.1% accuracy except the *good-good* paring with 91.2%. The number of samples per pair type averages about 17,600 $\pm$50. Using the final model with the maximum sequence length displayed similar results, with pair types using *bad* sentiment texts performing worse but not as extreme as with the shorter sequence length. The siamese baseline in comparison achieved best results for *bad-bad* with 86.1%, with the other pair types being at 83%.

Per Major Category Of special interest is the evaluation per category which better shows where the model works well and where it has difficulties, assuming different categories employ varied and distinct vocabulary and even semantics. The analysis is made more difficult by the fact that the distribution of businesses per category is not uniform in the training data. The model had been trained on the whole train dataset but was evaluated with the test set split into the major categories. It is therefore no real unbiased prediction as examples for each topic were present in the training data. Accuracies, as reported in Table 2, span between 84% to 95% but show no clear correlations between the number of businesses or reviews and prediction accuracy.

Cross-Category A more real-world example has been done with training on a single category and evaluation on the remaining categories as well as category k-fold cross-validation. We chose to train models using a sequence length of 128 for *Food* and *Arts & Entertainment*. Results with and without overlapping businesses between train and validation categories did not amount to significant accuracy differences (less than 1%). However, we detected a difference of about 10% for results from *Arts & Entertainment* compared to *Food* which can be explained with the difference of about 4.5 times as many businesses in *Food*. The *Food* model had a test accuracy of 76% on the same category but ranged from 71% to 83% on the other categories, whereas *Arts & Entertainment* had 62% accuracy itself and between 63% to 72% on other categories.

For the cross-validation experiment, we randomized the main categories and split them into 4 non-overlapping sets of businesses to simulate a situation where the model had to predict on completely unknown categories. We increased the number of sentiment pairs per pair-type to 4, so that we had 16 sentiment pairs per business in total, since a not insignificant number of businesses with more than one main category had to be discarded. We then trained a BERT model with a sequence length of 128 for 3 Epochs on each fold, and evaluated (a) on the remaining folds together, (b) on each fold separately, and (c) on each main category not in the training fold (cf. Table 3). Results for (a) are expected and slightly worse due to the shorter sequence length compared to other tables. For (b) prediction accuracies span between 79.4% and 92.3%, with a difference of 6 pp. for each fold. This is possibly due to more diverse training data which make predictions on unknown categories more robust. Using the baseline siamese model, we achieve similar results that span from 80.7% to 90.5% accuracy. Experiment (c) displays the highest variability as small single categories may differ more extremely compared to larger ones or sets of categories. Our BERT model has 71.5% to 95.3% accuracy, while our baseline model again has a slightly tighter range from 73.6% to 93.5%. The BERT model consistently performed slightly better by 1–3 pp. in all cross-validation experiments, while only being able to use at most 64 tokens per review. It, however, required much longer training.

5 Conclusion

Our contribution in this paper is the introduction of a new perspective on sentiment analysis. We showed how sequence pair classification can be used to achieve relatively good accuracy on the same sentiment pair problem. Initial results are promising but applying same sentiment models on

Category	General Statistics			Evaluation			
	Businesses	Reviews	Tokens	Prec.	Rec.	F1	Acc.
All Categories	192,127	6,685,900	128.9	89.05%	89.05%	89.05%	89.05%
Active Life	9,521	222,098	147.1	87.12%	87.12%	87.12%	87.12%
Arts & Entertainment	6,304	417,708	154.9	83.83%	83.82%	83.82%	83.82%
Automotive	13,203	267,164	142.1	93.99%	93.99%	93.99%	93.99%
Beauty & Spas	19,370	432,557	127.8	94.00%	94.00%	94.00%	94.00%
Education	3,314	44,321	150.5	88.54%	88.54%	88.54%	88.54%
Event Planning & Services	10,371	549,982	151.3	87.30%	87.30%	87.30%	87.30%
Financial Services	3,082	28,982	130.3	95.05%	95.05%	95.05%	95.05%
Food	29,989	1,511,092	121.1	86.96%	86.96%	86.96%	86.96%
Health & Medical	17,171	252,519	137.5	94.31%	94.30%	94.30%	94.30%
Home Services	19,744	288,764	147.0	94.45%	94.45%	94.45%	94.45%
Hotels & Travel	6,033	343,194	170.5	87.16%	87.16%	87.16%	87.16%
Local Flavor	1,444	92,816	135.7	84.41%	84.40%	84.40%	84.40%
Local Services	13,932	209,375	126.1	93.58%	93.58%	93.58%	93.58%
Mass Media	319	4,188	141.8	89.79%	89.77%	89.77%	89.77%
Nightlife	13,095	1,202,166	133.4	85.85%	85.85%	85.85%	85.85%
Pets	4,111	79,399	146.2	94.06%	94.06%	94.06%	94.06%
Professional Services	6,276	89,661	134.7	93.60%	93.59%	93.59%	93.59%
Public Services & Government	1,343	24,651	136.3	85.26%	85.25%	85.25%	85.25%
Religious Organizations	547	5,930	139.3	86.73%	86.72%	86.72%	86.72%
Restaurants	59,371	4,201,684	125.4	86.97%	86.97%	86.97%	86.97%
Shopping	31,878	519,479	133.9	89.05%	89.05%	89.05%	89.05%

Table 2: (left) General statistics per main business category, (right) Results per category using our best model. Dataset filtered with at least 10 reviews per business, 4 sentiment pairs generated per pair-type. The total number of examples in train/test is 633,384, 10% used as test split. Model: BERT-base-uncased, 512 SeqLen, 3 Epochs.

Category Split	Evaluation Accuracy Per			
	Businesses	(a) Rest	(b) Category split	(c) Single category
Shopping, Local Flavor, Health & Medical, Event Planning & Services, Restaurants, Public Services & Government	279,408	82.4%	79.4% – 85.8%	71.5% – 90.3%
Religious Organizations, Active Life, Arts & Entertainment, Professional Services, Hotels & Travel, Local Services	22,176	84.5%	81.5% – 86.0%	73.6% – 93.0%
Education, Automotive, Bicycles, Mass Media, Home Services	36,624	83.0%	80.9% – 87.6%	72.5% – 95.3%
Pets, Nightlife, Financial Services, Beauty & Spas, Food	89,376	85.2%	84.2% – 92.3%	75.0% – 93.3%

Table 3: Cross-Evaluation results, (a) on remaining businesses, (b) on each other split, and (c) per category not in train split. Model: BERT-base-uncased, 128 SeqLen, 3 Epochs.

different domains like same stance argument pairs or for authorship verification requires further studies. Looking ahead, we plan to investigate other transformer variants like DistilBERT (Sanh et al., 2019) or ALBERT (Lan et al., 2020) that have shown improved results on other sequence classification tasks compared to BERT as well as more elaborate models. With the application on other domains, we hope to ultimately find some common features for sameness that can be exploited in various ways to support and improve existing models.

Acknowledgements

This work was funded by the Development Bank of Saxony (SAB) under project "MINDSET" (project no. 100341518).

Ethics Statement

We used the Yelp dataset without any modifications to the data contained within. The dataset is a collection of opinionated texts obtained from publicly available and appropriately acknowledged sources respecting their terms and conditions. By reusing pre-trained models using the *Huggingface.co transformers* library, our approach might have inherited some forms of bias. We did not perform any evaluation of this potential problem. It is worth noting that our experiments show that our approach is far from being ready to be used within a product. Our goal is to advance the research on this task. In terms of computational resources, we restricted ourselves to variants of pre-trained models that can be fine-tuned with (relatively) fewer resources and are accessible to the majority of researchers.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Nabiha Asghar. 2016. Yelp Dataset Challenge: Review Rating Prediction. *arXiv preprint arXiv:1605.05362*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. 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 *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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ronen Feldman. 2013. Techniques and Applications for Sentiment Analysis. *Commun. ACM*, 56(4):82–89.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Hugo Hammer, Anis Yazidi, Aleksander Bai, and Paal Engelstad. 2015. Building Domain Specific Sentiment Lexicons Combining Information from Many Sentiment Lexicons and a Domain Specific Corpus. In *Computer Science and Its Applications*, IFIP Advances in Information and Communication Technology, pages 205–216. Springer International Publishing.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Rie Johnson and Tong Zhang. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570, Vancouver, Canada. Association for Computational Linguistics.

Moshe Koppel and Jonathan Schler. 2004. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 62, New York, NY, USA. Association for Computing Machinery.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kevin Labille, Susan Gauch, and Sultan Alfarhood. 2017. Creating Domain-Specific Sentiment Lexicons via Text Mining. In *Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, pages 1–8.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*, JMLR Workshop and Conference Proceedings, pages 1188–1196. JMLR.org.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning Text Similarity with Siamese Recurrent Networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2019. Enhancing Unsupervised Sentence Similarity Methods with Deep Contextualised Word Representations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 994–1003, Varna, Bulgaria. INCOMA Ltd.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Benno Stein, Yamen Ajjour, Roxanne El Baff, Khalid Al-Khatib, Philipp Cimiano, and Henning Wachsmuth. 2021. Same Side Stance Classification. In *Same Side Stance Classification Shared Task 2019*, volume 2921.

Benno Stein and Henning Wachsmuth, editors. 2019. *6th Workshop on Argument Mining (ArgMining 2019) at ACL*. Association for Computational Linguistics, Berlin Heidelberg New York.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Sara Tedmori and Arafat Awajan. 2019. Sentiment Analysis Main Tasks and Applications: A Survey. *JIPS*, 15(3):500–519.

Luis Terán and José Mancera. 2019. Dynamic profiles using sentiment analysis and twitter data for voting advice applications. *Government Information Quarterly*, 36(3):520–535.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised Data Augmentation for Consistency Training. *arXiv preprint arXiv:1904.12848*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.