Covid19-twitter: A Twitter-based Dataset for Discourse Analysis in Sentence-level Sentiment Classification

Shashank Gupta
Deakin University, Geelong, Australia
guptashas@deakin.edu.au

Mohamed Reda Bouadjenek
Deakin University, Geelong, Australia
reda.bouadjenek@deakin.edu.au

Antonio Robles-Kelly
Deakin University, Geelong, Australia
Defence Science and Technology Group, Edinburg, Australia
antonio.robles-kelly@deakin.edu.au

Tsz-Kwan Lee
Deakin University, Geelong, Australia
glory.lee@deakin.edu.au

Thanh Thi Nguyen
Deakin University, Geelong, Australia
Monash University, Melbourne, Australia
thanh.nguyen9@monash.edu

Asef Nazari
Deakin University, Geelong, Australia
asef.nazari@deakin.edu.au

Dhananjay Thiruvady
Deakin University, Geelong, Australia
dhananjay.thiruvady@deakin.edu.au

Abstract

For the sentence-level sentiment classification task, learning Contrastive Discourse Relations (CDRs) like *a-but-b* is difficult for Deep Neural Networks (DNNs) via purely data-driven training. Several methods exist in the literature for dissemination of CDR information with DNNs, but there is no dedicated dataset available to effectively test their dissemination performance. In this paper, we propose a new large-scale dataset for this purpose called *Covid19-twitter*, which contains around 100k tweets symmetrically divided into various categories. Instead of manual annotation, we used a combination of an Emoji analysis and a lexicon-based tool called Valence Aware Dictionary and sEntiment Reasoner (VADER) to perform automatic labelling of the tweets, while also ensuring high accuracy of the annotation process through some quality checks. We also provide benchmark performances of several baselines on our dataset for both the sentiment classification and CDR dissemination tasks. We believe that this dataset will be valuable for discourse analysis research in sentiment classification.

CCS Concepts

• **Computing methodologies → Neural networks**; *Artificial intelligence*; **Natural language processing**; **Discourse, dialogue and pragmatics**;

Keywords

Sentiment Classification, Discourse Analysis, Deep Learning.

ACM Reference Format:
Shashank Gupta, Mohamed Reda Bouadjenek, Antonio Robles-Kelly, Tsz-Kwan Lee, Thanh Thi Nguyen, Asef Nazari, and Dhananjay Thiruvady. 2024. Covid19-twitter: A Twitter-based Dataset for Discourse Analysis in Sentence-level Sentiment Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679120

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679120

1 Introduction

Deep Neural Networks (DNNs) are trained in a purely data-driven manner (i.e., no other external supervision is given) based upon a gradient-based optimisation algorithm [3, 19]. This type of training is often insufficient to learn some complex and desirable patterns in input data [27, 38]. For example, for the sentence-level sentiment classification task, it is challenging for DNNs to capture complex-linguistic patterns called *Contrastive Discourse Relations* (CDRs) [26] like *a-but-b* (refer Sub-section 2.1) via purely data-driven training [12, 17, 20].

To counter this drawback, a group of methods called Informed Machine Learning (IML) [6, 9, 42] have been proposed, which provides external supervision to the DNN as some prior knowledge [21] about the task during training. For the task of learning CDRs like *a-but-b*, several IML methods have been proposed [12, 17, 20], which model such relations as logic rules and include the information about dominant conjunct ("b" conjunct) with the DNN (also called CDR dissemination).

These IML methods are often evaluated on some existing general sentiment analysis datasets, like Stanford Sentiment Treebank (SST) [37], Movie Reviews (MR) [29] and Customer Reviews (CR) [16], which i) do not contain sentences with CDRs in large quantities, ii) the distribution of such sentences is skewed compared to regular sentences, and iii) identification of sentences containing a CDR is often confused with sentences just containing a CDR-syntactic structure (conjuncts do not contain contrastive sentiment-polarities). Moreover, the evaluation is usually limited to testing their sentiment classification accuracy, which is not correlated with their CDR dissemination performance [13]. Hence, these factors tend to make their analysis of CDR dissemination biased, and thus inaccurate.

In this paper, we propose *Covid19-twitter*, a large-scale dataset specifically designed to test the CDR dissemination performance of IML methods. It contains around 100k tweets symmetrically divided into various categories, which are specifically designed to test the CDR dissemination performance of IML methods effectively. Instead of manual annotation, we used a combination of Emoji analysis [36, 44] and a lexicon-based sentiment analysis tool called Valence Aware Dictionary and sEntiment Reasoner (VADER) [18] to perform automatic labelling of the tweets, while also ensuring high accuracy of the annotation process through some quality checks (for example, a label consistency check between VADER and Emoji analysis). Specific contributions of this paper are:

(1) A large-scale dedicated dataset specifically designed to test the CDR dissemination performance of IML methods for sentence-level sentiment classification.
(2) A rule-based approach to provide automatic labels for the data points instead of manual annotation to construct such a dataset.
(3) A comprehensive benchmark evaluation of several sentiment classifiers on the dataset for both sentiment classification and CDR dissemination tasks.

2 Dataset

In this section, we detail the *Covid19-twitter*[1] dataset starting from the inception of CDRs, construction, final distribution, and comparison against the existing datasets in the literature.

2.1 Contrastive Discourse Relations

Sentence-level sentiment classification is the task of determining the sentiment polarity of a sentence by classifying it, often as Positive, Negative, or Neutral. One important challenge in this regard is to model discourse relations between different segments (phrases or clauses) in a sentence and identify which segment will determine the sentence sentiment [26, 39].

Previous studies [17, 20] have shown that **Contrastive Discourse Relations** (CDRs) are hard to capture by DNNs like CNNs or RNNs for sentence-level sentiment classification. As per Mukherjee and Bhattacharyya [26], they introduce a sense of opposition between two ideas, which can confuse the DNN model to predict the correct sentiment, if it does not have a mechanism to learn and incorporate these structures. Thus, they need to be learned by the model while determining the overall sentence sentiment.

A sentence with a CDR has a syntactic structure like *a-keyword-b* where the conjuncts - *a* and *b* - are connected through a discourse marker (*keyword*) and have *contrastive sentiment polarities* [31]. In such case, the sentence-sentiment is determined as per the *dominant conjunct* [26], since using the opposing sentiment information in both conjuncts will confuse the DNN to provide the correct sentence-sentiment prediction [22, 26, 39, 45]. These relations can be further classified into (i) CDR_{Fol}, where the dominant clause is *following* (*b* conjunct), or (ii) CDR_{Prev}, where the dominant clause is *preceding* (*a* conjunct). Table 1 provides a list of CDRs used in our dataset.

Table 1: List of CDRs used in our dataset.

CDR	Keyword	Dominant conjunct	Sentence
a – **but** – *b*	*but*	*b* (CDR_{Fol}) [26]	The movie is good **but** *the casting is terrible*
a – **yet** – *b*	*yet*	*b* (CDR_{Fol}) [26]	Even though we can't travel **yet** *we can enjoy each other and what we have*
a – **though** – *b*	*though*	*a* (CDR_{Prev}) [26]	*You are having an amazing time* **though** we are having this awful pandemic
a – **while** – *b*	*while*	*a* (CDR_{Prev}) [1]	*Stupid people are not social distancing* **while** there's a global pandemic

2.2 Dataset Construction

2.2.1 ***Corpus Creation.*** To construct a high-quality dataset containing numerous distributions of various CDRs, we created a corpus of about *eight-hundred million* tweets on the COVID-19 topic. This was done by taking the tweet-IDs of such tweets from another large-scale Twitter-based dataset [23], and crawling the corresponding tweets from Twitter platform using API services[2].

2.2.2 ***Preprocessing.*** After creating the corpus, raw tweets were processed using a tweet pre-processor for sentiment analysis[3], which removed unwanted contents like #hashtags, URLs, @mentions and reserved keywords (like RT) to extract text part of tweets. Sentiment-sensitive information like "Emojis" and "Smileys" is extracted and preserved instead of being discarded, which is used for assigning sentiment labels. For example, a raw-tweet like "RT @tinuade01: This is highly disgusting and disturbing, why won't there be coronavirus😒😒 https://t.co/YJ3WKP1brB #hope #corona" is processed as "This is highly disgusting and disturbing, why won't there be coronavirus". We only processed *English* tweets containing 28 or more characters to get the average-length tweets.

2.2.3 ***Assigning Sentiment Label.*** Instead of manual annotation of the pre-processed tweets, whose numbers were in millions, we decided to explore alternative weakly supervised methods to perform automatic annotation. Digital pictograms like Emoticons associate a strong correlation with sentiment polarity of the sentence [14, 44] and hence, we designed an *Emoji analysis* method to assign sentiment labels to the pre-processed tweets. For each tweet, we check: i) whether it contains an emoji using an automatic emoji identification tool[4], ii) whether all emojis are present at the end of the tweet to make sure the tweet contains complete text[5], and iii) whether at least one emoji is present in the EmoTag1200 table [36] which associates 8 types of positive and negative emotions scores with an emoji - anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The emotion scores in the EmoTag1200 table [36] were assigned by human annotators, which convey to what extent they feel that an emoji is associated with a particular emotion. If the tweet passes all the checks, we calculate the sum of all emotion scores for each emoji present and get an *Aggregate Emotion Score*. This score is compared against emotion score thresholds for positive and negative polarities, which we found dynamically based on the dataset. These thresholds, 2.83 and -2.83, are such that they correspond to one standard deviation of aggregate emotion scores for a random sample of one million tweets. As a further consistency check, we used a lexicon-based sentiment analysis tool called

[1] The dataset is available at: https://github.com/shashgpt/Covid19-twitter

[2] Twitter developer tools: https://developer.twitter.com/en

[3] Tweet pre-processing tool: https://pypi.org/project/tweet-preprocessor/

[4] The emoji extraction tool is available at: https://advertools.readthedocs.io/en/master/

[5] This is to exclude tweets such as "I ♡NYC" as they are semantically incorrect.

Valence Aware Dictionary and sEntiment Reasoner(VADER) [18] to calculate the sentiment of the tweet and compared it with the sentiment label assigned by the emoji method. VADER is specifically attuned to sentiments expressed in social media [18], and has been shown to provide better performance than other lexicon-based sentiment analysers [2] and human raters as well [18]. We only kept those tweets in our dataset for which both VADER and emoji analysis assigns the same sentiment class.

2.2.4 *Assigning Rule Label.*

For each tweet that has been successfully assigned a sentiment label, we perform a *conjunction analysis* and identify if it contains a CDR-syntactic structure (*a-keyword-b* structure), listed in Table 1. The conjunction analysis involves i) identifying whether the tweet contains the *keyword*, ii) whether the *keyword* is not present at the end or beginning of the tweet (to make sure *a* and *b* conjuncts contain at least one word), and iii) the count of *keyword* in the sentence is just one (to avoid nested structures like *a-keyword-b-keyword-c*). If the tweet passes these checks, it is assigned the *a-keyword-b* rule label (*a-but-b* label for if the keyword is *but*) otherwise it is labelled as *No-rule*. Note that we only consider tweets that contain at most one structure (i.e., no multiple nested structures like *a-but-b-yet-c*).

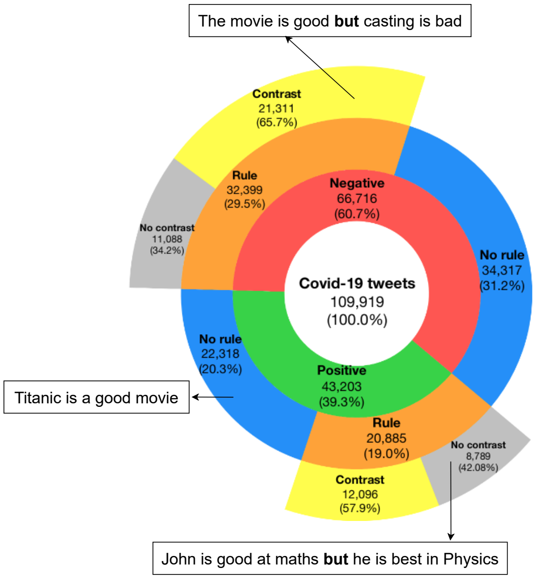
2.2.5 *Assigning Contrast Label.*

Rule labels for tweets denote whether they contain a CDR-syntactic structure, as listed in Table 1. To assign the CDR label to such tweets, a further check is necessary which determines whether the conjuncts contain contrastive sentiment polarities. Hence, we provide another binary label called *Contrast*, which takes the value as the result of this check. We again use the VADER tool and determine the sentiment polarity of each conjunct to compare whether they are similar or opposite. The former indicates no CDR but just the syntactic structure and is labelled as "No-Contrast" while the latter indicates a CDR and is labelled as "Contrast". Thus, collectively with the *Rule* label, the *Contrast* label determines whether the tweet contains a CDR or not.

2.3 Dataset Distribution

After processing the corpus and assigning all the labels, we obtain the final distribution as shown in Figure 1a. The dataset contains a total of 109,919 tweets, which are divided into multiple categories, each corresponding to a layer as depicted in Figure 1a. The first (most inner) layer denotes the tweets containing negative and positive sentiment polarities, accounting for about 60% and 40% of the dataset respectively. In the second layer, *Rule* denotes tweets with at most one syntactic structure corresponding to a CDR as outlined in Table 1, additionally *No-rule* refers to tweets with no syntactic structures. In the last layer, Rule subsets are further divided into *Contrast* and *No-contrast* categories. The *Contrast* contains tweets with CDRs and the *No-contrast* subset comprises tweets with no CDRs but just the corresponding syntactic structures where conjuncts do not exhibit contrastive sentiment polarities.

As can be seen in the Figure 1a, our dataset contains all tweets almost symmetrically divided into various classes, which means all classes are balanced, and the classifier can be trained without incurring any potential bias towards one class. In Table 1b, we provide distributions of individual CDRs listed in Table 1 in our dataset.



(a) Distribution of all categories in *Covid19-twitter* dataset.

Rules	Positive Contrast	Positive No-contrast	Negative Contrast	Negative No-contrast
$a - but - b$	9135	7091	17665	9002
$a - yet - b$	490	441	1072	761
$a - though - b$	962	443	625	268
$a - while - b$	1509	814	1949	1057

(b) Individual CDR distributions in the dataset.

Figure 1: Overall distribution of *Covid19-twitter* dataset.

2.4 Comparison with Existing Datasets

The most prominent sentence-level sentiment analysis datasets are Stanford Sentiment Treebank (SST) [37], Movie Reviews (MR) [29] and Customer Reviews (CR) [16] which contain 11,855, 10,662 and 3,775 sentences respectively. Compared to these numbers, our dataset contains around 100k data points. Other popular datasets like IMDb [25] and Amazon Reviews [5] contain big paragraphs instead of short texts and hence, cannot be used for evaluation on CDR dissemination task. A Twitter-based dataset called Sentiment140 [10] is closest to ours which was also automatically annotated based on emoji analysis, but we perform a sentiment consistency check with VADER which ensures the high accuracy of labels assigned. *TweetsCOV19* [8] is another Twitter-based dataset consisting of tweets on the COVID-19 topic. Their work also utilised a lexicon-based sentiment analysis tool called SentiStrength [40] for annotation and consists of about eight million tweets. Instead, in our work, we use the VADER sentiment analysis tool, which is more accurate and recent than SentiStrength [2, 33]. Moreover, our dataset contains a meticulously designed symmetric distribution of tweets between various discourse relations to effectively test the knowledge transfer performance of IML methods. Overall, to the best of our knowledge, *Covid19-twitter* is the first large-scale dataset specifically designed for testing the CDR dissemination performance of sentiment classifiers on the sentence-level sentiment classification task.

Table 2: Sentiment accuracy results.

Rule subset							
Sentiment Classifiers	**MLP**	**CNN**	**GRU**	**BiGRU**	**LSTM**	**BiLSTM**	**Transformer**
Flat Classifiers	0.836	0.892	0.845	0.871	0.867	0.872	0.889
Baseline Classifiers							
BERTweet [28]	0.96	0.959	0.963	0.959	0.959	0.962	0.957
GPT-2 [32]	0.962	0.967	0.965	0.965	0.969	0.966	0.961
XLNet [43]	0.969	0.973	0.968	0.967	0.967	**0.971**	0.969
RoBERTa [24]	**0.972**	**0.974**	**0.97**	**0.972**	**0.974**	**0.971**	**0.975**
DistilBERT [35]	0.968	0.972	0.969	0.969	0.967	0.967	0.962

Table 3: PERCY scores results.

Rule subset							
Sentiment Classifiers	**MLP**	**CNN**	**GRU**	**BiGRU**	**LSTM**	**BiLSTM**	**Transformer**
Flat Classifiers	0.025	0.07	0.039	0.07	0.065	0.07	0.058
Baseline Classifiers							
BERTweet [28]	0.11	0.111	0.108	0.109	0.11	0.109	0.109
GPT-2 [32]	0.104	0.101	0.098	0.094	0.096	0.097	0.099
XLNet [43]	0.109	0.101	0.111	0.113	**0.115**	0.113	0.114
RoBERTa [24]	0.118	0.111	**0.112**	0.11	0.111	0.112	0.108
DistilBERT [35]	**0.122**	**0.12**	0.111	**0.119**	0.113	**0.123**	**0.138**

Table 4: An example of $a - but - b$ CDR sentence, where the predicted sentiment was correct, but the decision was based on the *a* conjunct.

Sentences	Ground truth sentiment
absolutely right sad sad loss **but** the gentleman died of pneumonia another statistic for the covid regime its a joke	Negative

3 Benchmark Results

3.1 Setup

3.1.1 *Sentiment Classifiers.*

Firstly, we construct seven DNN models - MLP, CNN, GRU, BiGRU, LSTM, BiLSTM, and Transformer - as *flat classifiers* to get a measure of performances of simple DNN models on our dataset. This is to test how well such models can identify CDRs and base the sentiment decision as per the dominant conjunct. Each of these DNN models is implemented in their simplest possible architecture, that is, each model contains only one hidden layer. For the Transformer model [41], we modify it to perform text classification instead of neural machine translation by removing the decoder block and using the output of the encoder block to perform text classification via a sigmoid layer.

Krishna et al. [20] discovered that constructing contextual word embeddings [30] from input sentences and fine-tuning them on the downstream task, can inherently capture complex linguistic structures like CDRs in sentences. To test if using such word embeddings can help the DNN models in better capturing CDRs, we utilise several Pre-trained Language Models (PLMs) like BERT [7], and GPT-2 [32] to create contextual word embeddings, and create *baseline classifiers* by coupling each of them with a flat classifier. In this coupling, the word embeddings are constructed from the PLM (taken as the output from their last hidden layer), which are then fed to the flat classifier for sentiment classification. In our experiments, we utilise the most prominent PLMs like BERTweet [7] (a variant of BERT [7] proposed for sentiment classification), GPT-2 [32], XLNet [43], RoBERTa [24], and DistilBERT [35]. In total, we construct **42 sentiment classifiers** to provide a comprehensive benchmarking over our dataset.

3.1.2 *Metrics.*

We use the **Sentiment Accuracy** to quantify the sentiment classification performance of classifiers and a recently developed metric called **Post-hoc Explanation based Rule ConsistencY (PERCY)** score [13] to assess their CDR dissemination performance i.e, testing how effectively a classifier can recognise a CDR on input sentence and can predict the sentiment as per the dominant conjunct. PERCY uses feature-attribution-based AI Explanation frameworks like LIME [34] to calculate the contribution of each conjunct to the classifier prediction. These conjunct contributions are then compared against each other to determine whether the classifier based its decision on the dominant conjunct.

3.2 Results

3.2.1 *Sentiment Classification Performance.*

In Table 2, we show the sentiment accuracy results and find that the RoBERTa classifiers provide the best performance, while the DistilBERT and XLNet classifiers provide statistically comparable performances. This means that the difference between the p-values [4] of the scores is greater than 0.05 (p-value(RoBERTa scores, DistilBERT scores) > 0.05 and p-value(RoBERTa scores, XLNet scores) > 0.05) [11]. In particular, the performance improvement of baselines over the flat classifiers confirms that CDRs like *a-but-b* need to be learned by a DNN in order to provide better sentiment classification performance.

3.2.2 *CDR Dissemination Performance.*

In Table 3, we show the PERCY score [13] results and find that the DistilBERT classifiers outperform all the flat and baseline classifiers. This implies that, perhaps, using a Knowledge Distillation [15] training procedure can better enable the BERT classifier to capture CDRs. We also make an important note here that classifiers providing a high sentiment accuracy may not provide high PERCY score values as both metrics assess different tasks as described in Section 3.1.2. We show an anecdotal example in Table 4 where the classifier can provide a **correct** sentiment decision but based on the **wrong conjunct**. We observe that it is using some individual negative words in the *a* conjunct to base its decision.

4 Conclusion and Future Work

We presented a Twitter-based dataset to effectively test the CDR dissemination performance of sentiment classifiers on the sentence-level sentiment classification task. We designed a novel approach to automatically label the tweets using an Emoji analysis method and VADER [18] sentiment analysis tool. We benchmark our dataset for both sentiment classification and CDR dissemination tasks by conducting an exhaustive empirical evaluation of various general-purpose DNN models, and baselines constructed from pre-trained language models. Results show that the DistilBERT [35] classifiers provide the best CDR dissemination performance and comparable sentiment classification performance to the RoBERTa [24] classifiers. Future work involves exploring other types of discourse relations and enriching our dataset.

References

[1] Ritesh Agarwal, T. V. Prabhakar, and Sugato Chakrabarty. 2008. "I Know What You Feel": Analyzing the Role of Conjunctions in Automatic Sentiment Analysis. In *Proceedings of the 6th International Conference on Advances in Natural Language Processing*. Springer-Verlag, Berlin, Heidelberg, 28–39.

[2] Mohammed Al-Shabi. 2020. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *International Journal of Computer Science and Society.*

[3] Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Chris Ré, and Rob Malkin. 2019. Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale. In *Proceedings of the 2019 International Conference on Management of Data.*

[4] David Jean Biau, Brigitte M Jolles, and Raphaël Porcher. 2010. P value and the theory of hypothesis testing: an explanation for new researchers. *Clin. Orthop. Relat. Res.* (2010).

[5] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.*

[6] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. 2022. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports* (2022).

[7] 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).*

[8] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.*

[9] Artur S. d'Avila Garcez, Dov M. Gabbay, and Krysia B. Broda. 2002. *Neural-Symbolic Learning System: Foundations and Applications.* Springer-Verlag.

[10] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* (2009).

[11] Beatrice Grabowski. 2016. "P< 0.05" might not mean what you think: American statistical association ClarifiesPValues. *J. Natl. Cancer Inst.* (2016).

[12] Shashank Gupta, Mohamed Reda Bouadjenek, and Antonio Robles-Kelly. 2023. A Mask-Based Logic Rules Dissemination Method for Sentiment Classifiers. In *Advances in Information Retrieval.*

[13] Shashank Gupta, Mohamed Reda Bouadjenek, and Antonio Robles-Kelly. 2023. PERCY: A post-hoc explanation-based score for logic rule dissemination consistency assessment in sentiment classification. *Knowledge-Based Systems* (2023).

[14] Shatha Ali A. Hakami, Robert Hendley, and Phillip Smith. 2022. Emoji Sentiment Roles for Sentiment Analysis: A Case Study in Arabic Texts. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP).*

[15] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop.* http://arxiv.org/abs/1503.02531

[16] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[17] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 2410–2420.

[18] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* (2014).

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

[20] Kalpesh Krishna, Preethi Jyothi, and Mohit Iyyer. 2018. Revisiting the Importance of Encoding Logic Rules in Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

[21] Eyal Krupka and Naftali Tishby. 2007. Incorporating Prior Knowledge on Features into Learning. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research).* PMLR.

[22] Robin Lakoff. 1971. If's, And's and But's About Conjunction. In *Studies in Linguistic Semantics*, Charles J. Fillmore and D. Terence Langndoen (Eds.). Irvington, 3–114.

[23] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset.

[24] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.*

[25] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

[26] Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *Proceedings of COLING 2012.* The COLING 2012 Organizing Committee, Mumbai, India, 1847–1864.

[27] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *CoRR* (2014).

[28] 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.*

[29] Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05).*

[30] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).*

[31] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0.. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[33] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* (2016).

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery, New York, NY, USA, 1135–1144.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019.*

[36] Abu Awal Md Shoeb and Gerard de Melo. 2020. EmoTag1200: Understanding the Association between Emojis and Emotions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 8957–8967.

[37] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://www.aclweb.org/anthology/D13-1170

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.*

[39] Duyu Tang. 2015. Sentiment-Specific Representation Learning for Document-Level Sentiment Analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.*

[40] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* (2012).

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[42] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. 2023. Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[43] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: generalized autoregressive pretraining for language understanding.*

[44] Byungkyu Yoo and Julia Taylor Rayz. 2021. Understanding Emojis for Sentiment Analysis. *The International FLAIRS Conference Proceedings* (2021).

[45] Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-Grained Sentiment Analysis with Structural Features. In *Proceedings of 5th International Joint Conference on Natural Language Processing.*