A Mask-based Logic Rules Dissemination Method for Sentiment Classifiers

Shashank Gupta$^{1[0000-0001-8283-6597]}$, Mohamed Reda Bouadjenek$^{1[0000-0003-1807-430X]}$, and Antonio Robles-Kelly$^{2[0000-0002-2465-5971]}$

[1] School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, VIC 3216, Australia
{guptashas,reda.bouadjenek}@deakin.edu.au
[2] Defence Science and Technology Group, Edinburg, SA 5111, Australia
antonio.robleskelly@defence.gov.au

Abstract. Disseminating and incorporating logic rules inspired by domain knowledge in Deep Neural Networks (DNNs) is desirable to make their output causally interpretable, reduce data dependence, and provide some human supervision during training to prevent undesirable outputs. Several methods have been proposed for that purpose but performing end-to-end training while keeping the DNNs informed about logical constraints remains a challenging task. In this paper, we propose a novel method to disseminate logic rules in DNNs for Sentence-level Binary Sentiment Classification. In particular, we couple a Rule-Mask Mechanism with a DNN model which given an input sequence predicts a vector containing binary values corresponding to each token that captures if applicable a linguistically motivated logic rule on the input sequence. We compare our method with a number of state-of-the-art baselines and demonstrate its effectiveness. We also release a new Twitter-based dataset specifically constructed to test logic rule dissemination methods and propose a new heuristic approach to provide automatic high-quality labels for the dataset.

Keywords: Logic Rules · Sentiment Classification · Explainable AI

1 Introduction

Deep Neural Networks (DNNs) provide a remarkable performance across a broad spectrum of Natural Language Processing (NLP) tasks thanks to mainly their Hierarchical Feature Representation ability [5], However, the complexity and non-interpretability of the features extracted hinder their application in high-stakes domains, where automated decision-making systems need to have a human understanding of their internal process, and thus, require user trust in their outputs [23]. Moreover, a huge amount of labeled training data is required to construct these models, which is both expensive and time-consuming [2].

To fight against the above-mentioned drawbacks, it is desirable to make DNNs inherently interpretable by augmenting them with domain-specific or task-specific Expert Prior Knowledge [4]. This would complement the labeled training data [26], make their output causally interpretable [23] to answer the *why?* question and help the model learn real-world constraints to abstain from providing strange outputs, in particular for high-stakes domains. For example, for the binary sentiment classification task, given a sentence containing an *A-but-B* syntactic structure where A and B conjuncts have contrastive senses of sentiment (*A-but-B* contrastive discourse relation), we would like the model to base its decision on the B conjunct – following the *A-but-B* linguistically motivated logic rule [14]. However, in practice, such rules are difficult to learn directly from the data [10,13].

In this paper, we propose to model Expert Prior Knowledge as First Order Logic rules and disseminate them in a DNN model through our Rule-Mask mechanism. Specifically, we couple a many-to-many sequence layer with DNN to recognize contrastive discourse relations like *A-but-B* on input sequence and transfer that information to the DNN model via Feature Manipulation on input sequence features. The task of recognizing these relations is treated as binary token classification, where each token in the input sequence is classified as either 0 or 1 creating a rule-mask of either syntactic structure (e.g., $0-0-1$ or $1-0-0$), where only tokens corresponding to the rule-conjunct are classified as 1. This mask is then applied to the input sequence features via a dot product and the output is fed to the DNN model for the downstream task. Compared to existing methods, our method is jointly optimized with the DNN model and so it maintains the flexibility of end-to-end training, being straightforward and intuitive. Thus, the key contributions of this paper are summarized as follows:

1. We introduce a model agnostic Rule-Mask Mechanism that can be coupled with any DNN model to ensure that it will provide prediction following some logical constraints on the downstream task. We test this mechanism on the task of Sentence-level Binary Sentiment Classification where the DNN model is constrained to predict sentence sentiment as per linguistically motivated logic rules.
2. We release a dataset for the Sentence-level Binary Sentiment Classification task which contains an equal proportion of the sentences having various applicable logic rules as contrastive discourse relations. This dataset was constructed to test our method's ability to recognize the applicable logic rule in the input sentence and disseminate the information in the DNN model (i.e. help the DNN model to constrain its prediction as per the logic rules).
3. Instead of manual labeling of the dataset, we propose a new heuristic approach to automatically assign the labels based on Emoji Analysis and using a lexicon-based sentiment analysis tool called VADER [12]. We validate this approach by labeling a sample of tweets where we find high consistency between automatic labels and human labels.
4. We present a thorough experimental evaluation to demonstrate the empirically superior performance of our method on a metric specifically constructed to test logic rule dissemination performance and compare our results against a number of baselines.

2 Related Work

Even before the advent of modern Neural Networks, attempts to combine logic rules representing domain-specific or task-specific knowledge with hierarchical feature representation models have been studied in different contexts. For example, Towell and Shavlik [26] developed Knowledge-Based Artificial Neural Networks (KBANN) to combine symbolic domain knowledge abstracted as propositional logic rules with neural networks via a three-step pipelined framework. Garcez et al. [4] defined such systems as Neural-Symbolic Systems, which can be viewed as a hybrid model containing the representational capacity of a connectionist model like Neural Network and inherent interpretability of symbolic methods like Logical Reasoning. Our work is related to the broader field of Neural-Symbolic Systems, where we construct an end-to-end model, which embeds the representational capacity of a Neural Network and is aware of the logical rules when making inference decisions on the input. Thus, we review below both implicit and explicit methods to construct Neural-Symbolic Systems.

2.1 Implicit Methods to Construct Neural-Symbolic Systems

While not originally proposed to construct a Neural-Symbolic System, these works show that certain existing models can implicitly capture logical structures without any explicit modifications to their training procedure or architecture. For example, Krishna et al. [13] claimed that creating Contextualized Word Embeddings (CWE) from input sequence can inherently capture the syntactic logical rules when fine-tuned with the DNN model on downstream sentiment analysis task. They proposed to create these embeddings using a pre-trained language model called ELMo [19]. More recent state-of-the-art models like BERT [3] and GPT-2 [21] can also be used to create contextual representations of words in the input sequence.

However, as we show in our experimental results, such contextual representation of words alone is not sufficient to capture logical rules in the input sequence and pass the information to the DNN model. We instead show that implicit learning can be used to learn a rule-mask by a sequence model which then can be used to explicitly represent logic rule information on the input features to the downstream DNN model via Feature Manipulation.

2.2 Explicit Methods to Construct Neural-Symbolic Systems

These methods construct Neural-Symbolic systems by explicitly encoding logic rules information into the trainable weights of the neural network by modifying either its input training data, architecture, or its objective function.

Focusing on sentence-level sentiment classification, perhaps the most famous method is the Iterative Knowledge Distillation (IKD) [10], where first-order logic rules are incorporated with general off-the-shelf DNNs via soft-constrained optimization. An upgraded version of this method is proposed in [11] called Mutual Distillation, where some learnable parameters ϕ are introduced with logic rules when constructing the constrained posterior, which are learned from the input data. Instead of formulating constraints as regularization terms, Li and Srikumar [16] build Constrained Neural Layers, where logical constraints govern the forward computation operations in each neuron. Another work by Gu et al. [6] uses a task-guided pre-training step before fine-tuning the downstream task in which domain knowledge is injected into the pre-trained model via a selectively masked language modeling.

In contrast to these methods, our approach does not encode the rule information into the trainable parameters of the model but instead uses Feature Manipulation on the input through rule masking so as to disseminate the rule information into the downstream model. Thus, our method can incorporate logic rules without any such complicated ad-hoc changes to either input training data, architectures, or training procedures. Overall, the current literature lacks any method to construct a Neural-Symbolic model for sentiment classification which is straightforward, intuitive, end-to-end trainable jointly with the base neural network on training data and that can provide empirically superior performance.

3 Methodology

This section provides a detailed description of our method starting with the inception of Logic rules from domain knowledge to disseminating them with a DNN model.

3.1 Sources of Logic Rules

Previous work has shown that Contrastive Discourse Relations (CDRs) are hard to capture by general DNN models like CNNs or RNNs for sentence-level binary sentiment classification through purely data-driven learning [10,13,27]. Thus, Prasad et al. [20] define such relations as sentences containing *A-keyword-B* syntactic structure where two clauses A and B are connected through a discourse marker (*keyword*) and have contrastive polarities of sentiment. Sentences containing such relations can be further classified into (i) CDR_{Fol}, where the dominant clause is *following* and the rule conjunct is B (sentence sentiment is determined by B conjunct), or (ii) CDR_{Prev}, where the dominant clause is *preceding* and the rule conjunct is A. Mukherjee and Bhattacharyya [17] argue that these relations need to be learned by the model while determining the overall sentence sentiment. Hence, for our experiments, we identify these relations as expert prior knowledge, construct First Order Logic rules from them and incorporate these rules with the DNN model through our mask method. Table 1 lists all the logic rules we study in this paper.

3.2 Rule-Mask Mechanism to Disseminate Logical Information

Our task is to build an end-to-end system, which provides sentence-level sentiment predictions and bases its predictions on linguistically motivated logic rules.

Table 1: List of logic rules used in this analysis. Rule conjunct denotes the dominant clause during the sentiment determination and is italicized in examples.

Logic rule	**Keyword**	**Rule conjunct**	**Example**
A – **but** – B	*but*	B [17]	Yes there is an emergency called covid-19 **but** *victory is worth celebration*
A – **yet** – B	*yet*	B [17]	Even though we can't travel **yet** *we can enjoy each other and what we have*
A – **though** – B	*though*	A [17]	*You are having an amazing time* **though** we are having this awful pandemic
A – **while** – B	*while*	A [1]	*Stupid people are not social distancing* **while** there's a global pandemic

Specifically, given an input sentence S containing a rule-syntactic structure like $A - keyword - B$ where $keyword$ indicates an applicable logic rule in Table 1 and A & B conjuncts have contrastive senses of sentiment, we would like the classifier to predict the sentiment of S as per the B conjunct if the rule conjunct is B, otherwise, to predict the sentence sentiment as per A if the rule conjunct is A.

A straightforward method to create such a system is to use Feature Extraction [8] on the input data, where features corresponding to the rule conjunct are extracted and fed as input to the classifier. Specifically, given the input sentence, S containing *A-keyword-B* syntactic structure, Gupta et al. [8] proposed to manually compute a rule mask M of the structure $0 - 0 - 1$ if the rule conjunct is B, otherwise, $1 - 0 - 0$ if the rule conjunct is A. Then, they propose to compute a post-processed instance $X_{conjunct} = X * M$ as the dot product between S and M, where $X_{conjunct}$ can be regarded as an explicit representation of the applicable logic rule. $X_{conjunct}$ is then passed as input to the sentiment classifier and hence, the classifier predicts the sentiment as per the rule conjunct. The mask M is applied during both the training and testing phases of the classifier.

Although the Feature Extraction method proposed in [8] is quite simple, intuitive, and can determine whether the sentence contains *A-keyword-B* structure, it lacks the adaptability to the more nuanced nature of language since it cannot determine whether the conjuncts have *contrastive* polarities of sentiment and hence, cannot determine whether the sentence has a CDR or not. Moreover, simply removing a part of the input sequence entirely often leads to a loss of sentiment-sensitive information which can affect the sentiment classification performance on sentences that contains rule-syntactic structure but no CDR. Besides, as pointed out in [11], human knowledge about a phenomenon is usually abstract, fuzzy, and built on high-level concepts (e.g., discourse relations, visual attributes) as opposed to low-level observations (e.g., word sequences, image pixels). Thus, logic rules constructed from human knowledge should have these traits in the context of the dataset under consideration.

This necessitates a mechanism based on predictive modeling for the rule mask, which can: (i) determine whether the input sentence has a CDR instead of just rule syntactic structure, (ii) be learned from the training data, (iii) coupled

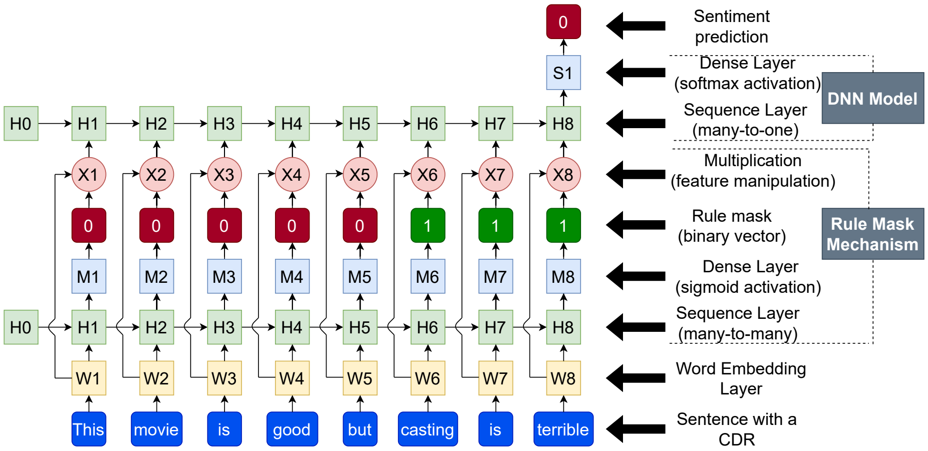

Fig. 1: Architecture of our rule-mask mechanism coupled with a DNN model. In mask block, the sequence layer predicts a rule mask M containing binary values corresponding to each token in input sentence S. Rule mask M is then multiplied with word embeddings W of S and the result is fed to the downstream DNN model.

with the classifier instead of being applied in a pipelined fashion, and (iv) jointly learned with the classifier on the training data to create a truly end-to-end system. Thus, we present a mechanism, in which given an input sentence S, it identifies whether it contains a logic rule structure like $A - keyword - B$ with A & B conjuncts having contrastive polarities of sentiment. If both conditions are met, it predicts a rule mask of a syntactic structure $0 - 0 - 1$ if the rule conjunct is B (mask values corresponding to tokens in A and $keyword$ parts are zero) or, otherwise, of structure $1 - 0 - 0$ if the rule conjunct is A (mask values corresponding to tokens in B and $keyword$ parts are zero). If there is no sentiment contrast between conjuncts or there is no rule-syntactic structure, it predicts a rule mask of a structure $1 - 1 - 1$. We optimize both the rule-mask mechanism and the DNN model jointly as:

$$\min_{\theta_1, \theta_2 \in \Theta} \quad L(y, p_{\theta_1}(y|x)) + \Sigma_{t=1}^{n} L(y_t, p_{\theta_2}(y_t|x_t)) \tag{1}$$

where $p_{\theta_1}(y|x)$ is the sentiment prediction of the DNN model and $p_{\theta_2}(y_t|x_t)$ is the mask value for t^{th} token in the input sequence $x = [x_1 \cdots x_n]$ and tackle the task of rule mask prediction by casting it as a token-level binary classification problem, where we predict either 0 or 1 tags corresponding to every token in the input sentence. We choose L as the Binary Cross-Entropy loss function.

Note that the proposed rule-mask mechanism can also be used with popular transformer-based DNN models BERT [3] where token embeddings can be first used to calculate the rule mask and then used to calculate the Masked Language Modeling (MLM) output.

4 Covid-19 Twitter Dataset

To conduct effective experimentation for testing the logic rule dissemination capability of our method, we constructed a dataset that contains an equally proportional amount of sentences containing logic rules (shown in Table 1) and no rules as shown in Figure 2. Further, the rule subset contains an equal proportion of sentences containing CDRs (contrast labels) and no CDRs (no contrast labels). The reason behind constructing our own dataset is that we wanted to get the specific distribution of sentences as shown in Figure 2 to test the logic rule dissemination performance of our method in an unbiased manner. Such distribution in sufficient quantities is very difficult to find in existing popular sentiment classification datasets like SST2 [25], MR [18], or CR [9].

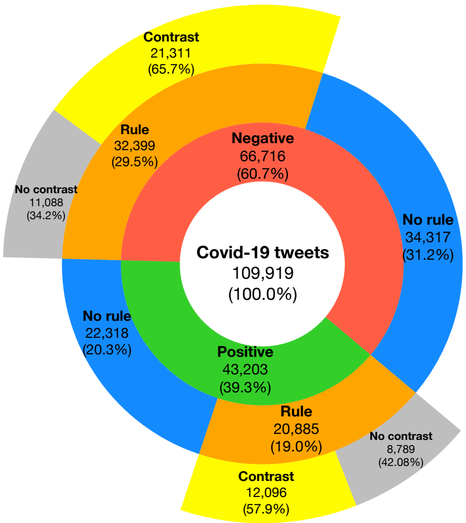

Fig. 2: Sector map of the constructed dataset denoting the overall distribution of tweets. The 1st layer denotes the proportion of tweets containing negative and positive sentiment polarities. In the 2nd layer, the Rule sector denotes tweets having at-most one of the logic rules applicable in Table 1 and the No-Rule sector denotes tweets with no applicable logic rule. In the last layer, the Contrast sector denotes tweets containing a CDR as defined in Section 3.1 and the No-Contrast sector denotes tweets without a CDR but contains a logic rule.

To get this distribution, we created a corpus of tweets from Twitter on the Covid-19 topic where the tweet IDs were taken from the Covid-19 Twitter dataset [15]. Raw tweets were then pre-processed using a tweet pre-processor[3], which removes unwanted contents like hashtags, URLs, @mentions, reserved keywords, and spaces. Each pre-processed tweet was then passed through a series of steps as listed in Figure 3 so as to obtain the following: (1) Sentiment Label, which indicates the polarity of the sentence, (2) Logic-Rule Label corresponding to either of the applicable rules listed in Table 1, and (3) Contrast Label which determines if the sentence containing a logic rule has a CDR or not (conjuncts A and B have a contrastive sense of sentiments). In the following sub-sections, we provide more details on the definition of these labels, why they need to be assigned, and how they were assigned to each tweet.

[3] Tweet pre-processing tool used here is accessible at `https://pypi.org/project/tweet-pre-processor/`

4.1 Sentiment Labels

Previous works [28] have shown that emojis indicate a strong correlation with associated sentence sentiment polarity and hence, we designed an Emoji Analysis method to assign sentiment labels to pre-processed tweets. Specifically, for each pre-processed tweet, we check whether it contains an emoji using an automatic emoji identification tool in texts[4], whether all emojis are present at the end of the tweet to make sure the tweet contains complete text[5] and whether at least one emoji is present in the EmoTag1200 table [24] which associates 8 types of positive and negative emotions scores with an emoji - anger, anticipation, disgust, fear, joy, sadness, surprise, and trust - and contains a score for each emotion assigned by human annotators. If the tweet passes the above checks, we then calculate the sum of all emotion scores for each emoji present and get an Aggregate Emotion Score for the tweet. 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 1 million tweets. As a further consistency check, we used a lexicon-based sentiment analysis tool called VADER [12] and only kept those tweets in our dataset for which both VADER and emoji analysis assigns the same sentiment class.

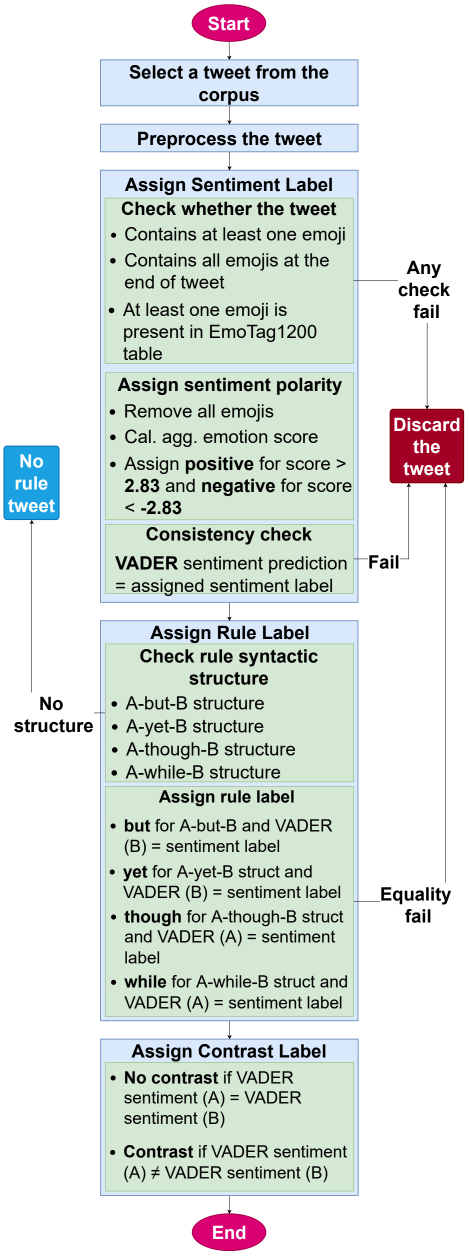
Start
Select a tweet from the corpus
Preprocess the tweet
Assign Sentiment Label
Check whether the tweet
• Contains at least one emoji
• Contains all emojis at the end of tweet
• At least one emoji is present in EmoTag1200 table
Any check fail
Discard the tweet
Assign sentiment polarity
• Remove all emojis
• Cal. agg. emotion score
• Assign positive for score > 2.83 and negative for score < -2.83
Consistency check
VADER sentiment prediction = assigned sentiment label
Fail
No rule tweet
Assign Rule Label
Check rule syntactic structure
• A-but-B structure
• A-yet-B structure
• A-though-B structure
• A-while-B structure
No structure
Assign rule label
• but for A-but-B and VADER (B) = sentiment label
• yet for A-yet-B struct and VADER (B) = sentiment label
• though for A-though-B struct and VADER (A) = sentiment label
• while for A-while-B struct and VADER (A) = sentiment label
Equality fail
Assign Contrast Label
• No contrast if VADER sentiment (A) = VADER sentiment (B)
• Contrast if VADER sentiment (A) ≠ VADER sentiment (B)
End

Fig. 3: Covid-19 tweets dataset construction flowchart.

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

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

4.2 Rule Labels

For each tweet that has been successfully assigned a sentiment label, we perform a conjunction analysis and identify if it contains any rule-syntactic structure listed in Table 1. Note that we only consider tweets that contain only one structure (i.e. no multiple nested structures like *A-but-B-yet-C*). The absence of any structure in the tweet is labeled as No-Rule otherwise, we check the corresponding rule applicability condition on the tweet which for example, for $A - but - B$ structure checks whether the sentiment polarity of the tweet is consistent with the sentiment polarity of B conjunct. We again use VADER to determine the sentiment polarity of the rule conjunct. If the rule applicability condition holds, we assign the corresponding rule label to the tweet, otherwise discard it to avoid noise in our dataset.

4.3 Contrast Labels

Contrast labels are important as performance on this subset (Rule-Contrast) is expected to indicate how effectively a method disseminates contrastive discourse relations (CDRs) in the DNN model. As mentioned in Section 3.1, general DNNs cannot capture CDRs in sentences and hence, cannot determine their sentiment correctly. Therefore, we need to provide another label to tweets containing a rule-syntactic structure, which determines whether they contain a CDR or not. For such tweets, we provide another binary label called Contrast, which determines whether their conjuncts contain contrastive senses of sentiments or not. To determine this label, we again use VADER and determine the sentiment polarity of each conjunct to compare whether they are similar indicating "No-Contrast" or opposite indicating a CDR and labeled as "Contrast". We again maintain an equal proportion of sentences labeled with "Contrast" and "No-Contrast" so to train classifiers that can effectively determine the CDR, not just the rule-syntactic structure.

4.4 Constructed Dataset

After processing the corpus (flowchart shown in Figure 3) and assigning all the labels, we obtain the final distribution as shown in Figure 2. The dataset contains a total of 109,919 tweets assigned either positive or negative sentiment labels accounting for about 60% and 40% of the dataset respectively. Further, each sentiment subset is divided into 2 subsets - Rule, which contains tweets having one of the logic rule labels listed in 1, and No-Rule tweets, which do not contain any logic rules. The Rule subsets are further divided into Contrast and No-Contrast subsets, where the former contains tweets containing logic rules and CDRs (A and B conjuncts have contrastive senses of sentiment), and the latter contains tweets having applicable logic rule but do not contain a CDR (A and B conjuncts do not have contrastive senses of sentiment). In Table 2, we show a small sample of tweets annotated manually for all the labels in our dataset as shown in Figure 2 to validate our heuristic approach of dataset labeling.

Table 2: Sample of tweets labeled manually to validate the heuristic approach.

(a) No rule tweets labeled with Positive Sentiments.

finally have decent ppe in the care home.
love this idea we are living through history and this is a great way to capture it.
we went to crawley, it was well organised and we felt looked after so thanks indeed.
ederson still my best performing city player since lockdown.
u are well i hope you are staying safe much love from montreal canada.
ms dionne warwick you are giving me so much lockdown joy.

(b) No rule tweets labeled with Negative Sentiments.

the provincial governments are drastically failing its people.
this quarantine makes you to attend a funeral just to cry out.
duterte threatens to jail those who refuse covid vaccines.
my professor just sent us an email saying he got covid there will be no class.
got covid yesterday and today pumas lost what a shit weekend.
i told my mam i filled out my application for my vaccine and she called me a bitch.

(c) Rule tweets labeled with positive sentiment and contrast.

A lot has been said against our president **but** I think he is doing his best.
it's a covid 19 pandemic ravaged tennis season **yet** carlos alcaraz is still won 28 lost 3.
first game after lockdown started with a birdie **though** good scoring didnt last.
friends in brazil posting festivals **while** ive been in lockdown since march.
He's in quarantine **but** still looking good and handsome as always.
feku wrote the book on how to lie non stop **but** his supporters still believe him.

(d) Rule tweets labeled with positive sentiment and no contrast.

michael keaton is my favorite batman **but** lori lightfoot is my favorite beetlejuice.
best boy band and **yet** so down to earth and always down for fun bts best boy.
awww it's such a cute corona **though** i want to hug it.
happy birthday have all the fun **while** staying covid safe.
well said we always try to improve as human nature **but** corona teach us very well.
this research is funny **but** also might encourage some mask use.

(e) Rule tweets labeled with negative sentiment and contrast.

I want to get a massage **but** of course, that's not such a good idea during a pandemic.
kaaan it has been one freakin year **yet** people still dont take this pandemic seriously.
absolutely disgusting that fans would gather even **though** corona virus is a thing.
niggas having social events **while** its a pandemic out.
thats looks fun **but** covid 19 destroyed our habitat shame on that virus.
i got a plan for a trip **but** chuck it i know it's gonna get cancel.

(f) Rule tweets labeled with negative sentiment and no contrast.

this is so sad i want churches to reopen too **but** i also dont want to see this happening.
stage 4 cancer **yet** its corona that killed him.
people are getting sick on the vaccine **though** i know people who have it very bad.
there is nothing safe about this **while** theres a pandemic still going on i mean wtf.
i may come off as rude **but** during the pandemic ive forgotten how to socialize sorry.
hes never stayed away from me **but** i know he misses them and i have to work.

5 Experimental Results

In this section, we discuss the performance results of our method and baselines under study for the task of sentence-level binary sentiment classification on our dataset [6].

5.1 Dataset Preparation

We divide the Covid-19 tweets dataset into the Train, Val, and Test splits containing 80%, 20%, and 20% proportion of sentences respectively. Each split contains similar distributions for various subsets - No-Rule Positive, No-Rule Negative, Rule Positive Contrast, Rule Negative Contrast, Rule Positive No-Contrast, and Rule Negative No-Contrast - as presented in the complete dataset Figure 2. This ensures the classifiers are trained, tuned, and tested on splits containing proper distributions of every category of sentences.

5.2 Sentiment Classifiers

To conduct an exhaustive analysis, we train a range of DNN models as Base Classifiers - RNN, BiRNN, GRU, BiGRU, LSTM, and BiLSTM - to get the baseline measures of performances. Each model contains 1 hidden layer with 512 hidden units and does not have any mechanism to incorporate logic rules. We then train these models again coupled with a rule dissemination method proposed in Iterative Knowledge Distillation (IKD) [10], Contextualized Word Embeddings (CWE) [13] and our Rule-Mask Mechanism to construct Logic Rule Dissemination (LRD) Classifiers. For our method, we train a wide range of possible configurations to provide an exhaustive empirical analysis. These configurations are {RNN base classifier, BiRNN base classifier, GRU base classifier, BiGRU base classifier, LSTM base classifier, and BiLSTM base classifier} × {RNN mask layer, BiRNN mask layer, GRU mask layer, BiGRU mask layer, LSTM mask layer, and BiLSTM mask layer}, which totals up to 36 LRD classifiers to exhaustively test the empirical performance of our method. We want to compare the performance of our method with other dissemination methods and propose the best method for a particular base classifier.

5.3 Metrics

While Sentiment Accuracy is the obvious choice given the task is sentiment classification, it fails to assess whether the classifier based its decision on the applicable logic rule or not. For example, a classifier may correctly predict the sentiment of the sentence *"the casting was not bad but the movie was awful"* as negative but may base its decision as per the individual negative words like *not* in the *A* conjunct instead of using *B* conjunct. Hence, we decided to use an alternative metric called PERCY proposed in [7] which stands for *Post-hoc*

[6] Code and dataset are available at `https://github.com/shashgpt/LRD-mask.git`

Explanation-based Rule ConsistencY. It assesses both the accuracy and logic rule consistency of a classifier for the sentiment classification task. Briefly, we compute this score as follows:

1. Given a sentence s which is an ordered sequence of terms $[t_1 t_2 \cdots t_n]$ and contains a logic rule structure like *A-keyword-B*, we use *LIME* explanation framework [22], which maps it to a vector $[\tilde{w}_1 \tilde{w}_2 \cdots \tilde{w}_n]$ with $\tilde{w}_n$ indicating how much the word t_n contributed to the final decision of the classifier.
2. Next we define the contexts $C(A) = [\tilde{w}_1 \cdots \tilde{w}_{i-1}]$ and $C(B) = [\tilde{w}_{i+1} \cdots \tilde{w}_n]$ as respectively the left and a right sub-sequences w.r.t the word *keyword* indexed by i.
3. Finally, we select top $k = 5$ tokens by their values from $C(A)$ as $C_k(A)$ and $C(B)$ as $C_k(B)$ and, propose that a classifier has based its decision on B conjunct if $\mathbb{E}_w[C_k(B)] > \mathbb{E}_w[C_k(A)]$ otherwise on A conjunct if $\mathbb{E}_w[C_k(A)] < \mathbb{E}_w[C_k(B)]$, where $\mathbb{E}$ is the expectation over conjunct weights. Hence, we define the PERCY score as the following:

$$PERCY(s) = (P(y|s) = y_{gt}) \wedge (\mathbb{E}_w[C_k(A)] \lesseqgtr \mathbb{E}_w[C_k(B)]) \quad (2)$$

   where the first condition $(P(y|s) = y_{gt})$ tests the classification accuracy ($P(y|s)$ denotes classifier prediction on sentence s and y_{gt} is the ground-truth sentiment) and the second condition ($\mathbb{E}_w[C_k(A)] < \mathbb{E}_w[C_k(B)]$ or $\mathbb{E}_w[C_k(A)] > \mathbb{E}_w[C_k(B)]$) checks whether the prediction was based as per the rule-conjunct (if the logic rule present is *A-but-B* or *A-yet-B*, the rule-conjunct is B whereas if the logic rule is *A-though-B* or *A-while-B*, the rule-conjunct is A).

5.4 Results

In this section, we analyze the PERCY scores for the classifiers as discussed in Section 5.2 obtained on **rule-contrast subset** of Covid-19 tweets test dataset (yellow color portion of the distribution as shown in Figure 2), which contains sentences with Contrastive Discourse Relations as discussed in Section 3.1. Remember that the task of our method is to identify applicable CDRs in the sentences and disseminate the information in the downstream DNN model. Therefore, we show the results only on the rule-contrast subset.

Here, we find that our method outperforms all the base classifiers as well as the other logic rule dissemination methods proposed in [10] and [13]. This implies that the base classifiers cannot learn CDRs in sentences while determining their sentiments, and hence, they perform poorly. Further, we observe that the bidirectional mask models perform the best which implies that bidirectional models can identify the applicable CDRs and learn the rule mask better than unidirectional ones. It could be argued that the mask method uses the explicit representation of logic rules on input features instead of probabilistic modeling like other methods and, hence, is expected to provide the best empirical performance.

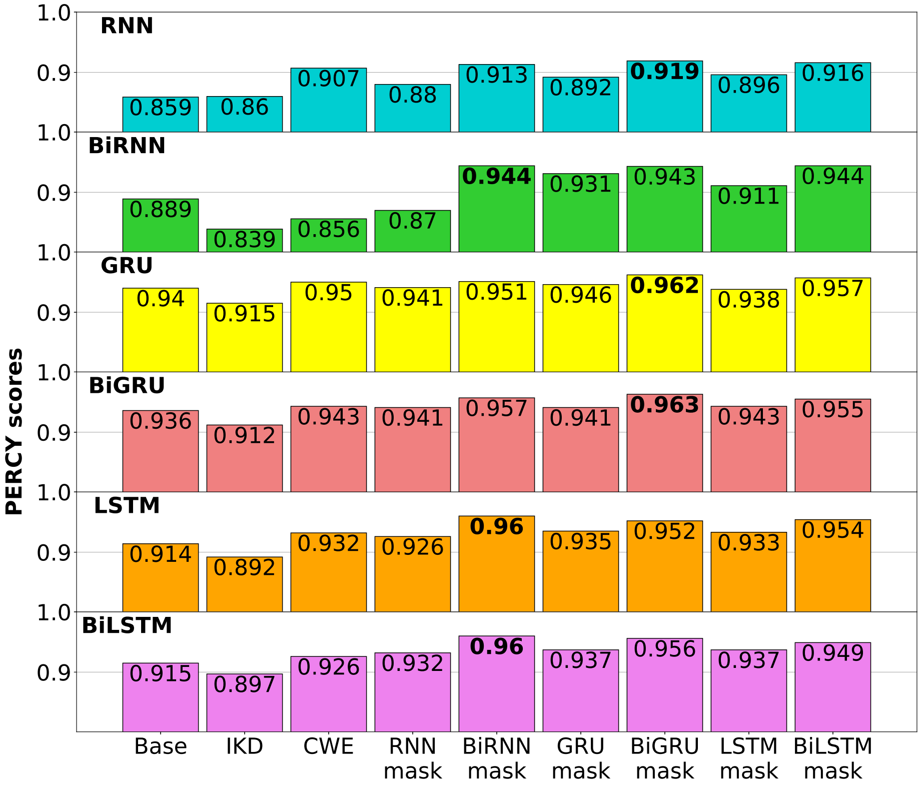

Fig. 4: PERCY scores for the classifiers obtained on **rule-contrast** subset of Covid-19 tweets test dataset. We show a total of 6 bar plots each corresponding to a base classifier (RNN, BiRNN, etc.) and each plot contains results for 9 classifiers as discussed in Section 5.2 with the best value highlighted in bold.

6 Conclusion

In this paper, we presented a novel method to disseminate contrastive discourse relations as logical information in a DNN for the sentiment classification task. This is done by coupling a rule mask mechanism with the DNN model, which identifies applicable CDR on the input sequence and transfers the information to the model via feature manipulation on the input sequence. Compared to existing methods, ours is end-to-end trainable jointly with the DNN model, does not require any ad-hoc changes to either training or, architecture, and is quite straightforward. We constructed our own dataset of tweets using a heuristic approach to conduct an unbiased analysis. We have shown results for various configurations of our method on different DNN models and compared it with existing dissemination methods. Our experimental results demonstrate that our method consistently outperforms all baselines on a both sentiment and rule consistency assessment metric (PERCY score) when applied to sentences with CDRs.

References

1. Agarwal, R., Prabhakar, T.V., Chakrabarty, S.: "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. p. 28–39. GoTAL '08, Springer-Verlag, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85287-2_4, https://doi.org/10.1007/978-3-540-85287-2_4
2. Bach, S.H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., Kuchhal, R., Ré, C., Malkin, R.: Snorkel drybell: A case study in deploying weak supervision at industrial scale. In: Proceedings of the 2019 International Conference on Management of Data. p. 362–375 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
4. Garcez, A.S.d., Broda, K., Gabbay, D.M., et al.: Neural-symbolic learning systems: foundations and applications. Springer Science & Business Media (2002)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
6. Gu, Y., Zhang, Z., Wang, X., Liu, Z., Sun, M.: Train no evil: Selective masking for task-guided pre-training. In: EMNLP (2020)
7. Gupta, S., Bouadjenek, M.R., Robles-Kelly, A.: An analysis of logic rule dissemination in sentiment classifiers. In: Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 118–124. Springer International Publishing, Cham (2022)
8. Gupta, S., Robles-Kelly, A., Bouadjenek, M.R.: Feature extraction functions for neural logic rule learning. In: Structural, Syntactic, and Statistical Pattern Recognition. pp. 98–107. Springer International Publishing (2021)
9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177 (2004)
10. Hu, Z., Ma, X., Liu, Z., Hovy, E., Xing, E.: Harnessing deep neural networks with logic rules. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2410–2420. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1228, https://www.aclweb.org/anthology/P16-1228
11. Hu, Z., Yang, Z., Salakhutdinov, R., Xing, E.: Deep neural networks with massive learned knowledge. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1670–1679. Association for Computational Linguistics (2016)
12. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media **8**(1), 216–225 (May 2014), https://ojs.aaai.org/index.php/ICWSM/article/view/14550
13. Krishna, K., Jyothi, P., Iyyer, M.: Revisiting the importance of encoding logic rules in sentiment classification. In: Proceedings of the 2018 Conference on Empirical

Methods in Natural Language Processing. pp. 4743–4751. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1505, https://www.aclweb.org/anthology/D18-1505

14. Lakoff, R.: If's, and's and but's about conjunction. In: Fillmore, C.J., Langndoen, D.T. (eds.) Studies in Linguistic Semantics, pp. 3–114. Irvington (1971)
15. Lamsal, R.: Coronavirus (covid-19) tweets dataset (2020). https://doi.org/10.21227/781w-ef42, https://dx.doi.org/10.21227/781w-ef42
16. Li, T., Srikumar, V.: Augmenting neural networks with first-order logic. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 292–302. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1028, https://aclanthology.org/P19-1028
17. Mukherjee, S., Bhattacharyya, P.: Sentiment analysis in twitter with lightweight discourse analysis. In: COLING (2012)
18. Pang, B., Lee, L.: 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) (2005)
19. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of NAACL (2018)
20. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)
21. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. Tech. rep., OpenAI (2018)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
23. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019)
24. Shoeb, A.A.M., de Melo, G.: EmoTag1200: Understanding the association between emojis and emotions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8957–8967. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.720, https://aclanthology.org/2020.emnlp-main.720
25. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://www.aclweb.org/anthology/D13-1170
26. Towell, G.G., Shavlik, J.W.: Knowledge-based artificial neural networks. Artificial Intelligence **70**(1), 119–165 (1994). https://doi.org/https://doi.org/10.1016/0004-3702(94)90105-8, https://www.sciencedirect.com/science/article/pii/0004370294901058
27. Yang, B., Cardie, C.: Context-aware learning for sentence-level sentiment analysis with posterior regularization. In: Proceedings of the 52nd Annual Meeting of

the Association for Computational Linguistics (Volume 1: Long Papers). pp. 325–335. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-1031, https://aclanthology.org/P14-1031

28. Yoo, B., Rayz, J.T.: Understanding emojis for sentiment analysis. The International FLAIRS Conference Proceedings **34** (Apr 2021). https://doi.org/10.32473/flairs.v34i1.128562, https://journals.flvc.org/FLAIRS/article/view/128562