A Dual-Channel Framework for Sarcasm Recognition by Detecting Sentiment Conflict

Anonymous ACL submission

Abstract

Sarcasm employs ambivalence, where one says something positive but actually means negative, vice versa. The essence of sarcasm, which is also a sufficient and necessary condition, is conflict between the literal and implied sentiments expressed in one sentence. However, it is difficult to recognize such sentiment conflict because of the sentiments are mixed or even implicit. As a result, the recognition of sophisticated and obscure sentiment brings in a great challenge to sarcasm detection. In this paper, we propose a Dual-Channel Framework by modeling both literal and implied sentiments separately. Based on this dual-channel framework, we design the Dual-Channel Net (DC-Net) to recognize sentiment conflict. Experiments on political debates (i.e., IAC-V1 and IAC-V2) and Twitter datasets show that our proposed DC-Net achieves state-of-the-art performance on sarcasm recognition.

1 Introduction

Sarcasm is a complicated linguistic phenomenon. Intuitively, it means that one says something positive on surface form, while he/she actually expresses negative, vice versa (Liu, 2012; Morrow, 2008). Take the sentence “Final exam is the best gift on my birthday” as an example, the literal sentiment on surface is positive, which is reflected the explicit sentiment words, i.e., “best gift”. However, the factual part of the text (i.e., “final exam happens on birthday”) implies that the sentiment to express is negative. This example suggests that it is the sentiment conflict that causes sarcasm linguistically.

However, modeling this linguistic nature of sarcasm is a great challenge due to the difficulty of digging sentiment conflict between the literal and implied meanings. We know that non-sarcastic texts do not contain implied meaning, so the literal sentiment is consistent with the actual sentiment. But for sarcastic texts, there is more than one meaning that coexist in one sentence. The literal meaning and the implied meaning are reflected in different sub-sentences. What’s more difficult, sentiments behind the two meanings are mixed or even implicit.

Many existing studies adopt generic classification models for sarcasm recognition (Lou et al., 2021; Ghosh and Veale, 2016). However, these methods directly model the entire sentence without considering the contradictory meanings behind sarcastic texts. Besides, there are studies using contrast patterns (e.g., phrase pair and word pair) as an indicator to detect sarcasm, which is approaching the linguistic essence of sarcasm. Riloff et al. (2013); Joshi et al. (2015) detect contrast or incongruity patterns, i.e., the co-occurrence of positive sentiment phrases and negative situational phrases. Tay et al. (2018); Xiong et al. (2019) use an attention mechanism to measure the sentiment conflict between word pairs in sarcastic texts. However, these methods emphasize too much on the explicit sentiment conflict on surface form (i.e., word/phrase level), which mainly reflect the literal meaning. As a result, the factual text is underestimated, which expresses the implied sentiment.

Dual-Channel Framework. In this paper, we propose a dual-channel framework to model the literal sentiment and the implied sentiment simultaneously. So we can exploit the conflict between the two channels in a comprehensive way. Figure 1 depicts the proposed dual-channel framework. Literal channel and implied channel are used to detect the surface and the hidden meanings separately. Once sentiment conflict is detected, we then determine the existence of sarcasm. The design of the dual-channel framework balances the effect of literal and implied inputs and avoids focusing too much on either one channel while ignoring the other. Our framework covers existing sarcasm patterns, and could be further enhanced to detect more sentiment conflict patterns.
Based on this framework, we develop the Dual-Channel Net (DC-Net) to detect sarcasm. DC-Net contains four modules: decomposer, literal channel, implied channel, and analyzer. In general, sentiment words directly reflect the surface sentiment, while the text without sentiment words reflects the implied sentiment. Hence, we split the sentiment words of input text into literal channel, and put the remaining words to implied channel by decomposer. Then we use the literal channel to model surface meaning, and the implied channel to model hidden meaning. Lastly, we use analyzer to recognize the conflict. Experiments on three benchmark datasets (i.e., IAC-V1, IAC-V2 and Tweets) show that our proposed DC-Net model achieves state-of-the-art performance.

The main contributions of this paper are two fold. First, to the best of our knowledge, the dual-channel framework is the first attempt to explicitly separate literal meaning and implied meaning, for detecting sentiment conflict between them. Second, experiments conducted on benchmark datasets (i.e., IAC-V1/V2 and Tweets) show that the proposed model achieves state-of-the-art performance.

2 Related Work

Prior works on sarcasm recognition can be divided into traditional models and neural network based models. There are also methods considering context information, e.g., history posts (Hazarika et al., 2018; Zhang et al., 2016), user profile (Poria et al., 2016; Kolchinski and Potts, 2018). However, such context and user profile information may not be always available.

2.1 Traditional Models

Most traditional approaches adopt the machine learning methods such as SVM with manually crafted rules or features. The features includes sentiment lexicons (González-Ibáñez et al., 2011), pragmatic features (i.e., emoticons (González-Ibáñez et al., 2011), capitalization, and punctuation (Joshi et al., 2015)), and pattern-based features (Riloff et al., 2013) et al.. Hee et al. (2018b) combine common sense to assist sarcasm detection on Twitter. Accordingly, the accuracy for sarcasm recognition highly depends on the quality of features.

Rewriting key part of sentence(s) manually is an expensive but effective method. Ghosh et al. (2015) believe that sarcasm involves a figurative meaning which is usually the opposite of literal meaning. So they re-frame sarcasm recognition as literal/sarcastic word sense disambiguation problem. Last, they paraphrase sarcastic texts manually to obtain target words that cause sarcastic disambiguation. This work is novel but heavily relies on manual paraphrasing and labeling of datasets to find target words. Moreover, target words are mostly limited to sentiment words. As a result, the model is dominated by these explicit sentiment words and ignores the implied channel.

2.2 Neural Networks Based Models

As attention mechanism has led to improvements in various NLP tasks, Tay et al. (2018); Xiong et al. (2019) use attention to capture the relationship of word pairs along with an LSTM to model the entire sentence. Ghosh and Veale (2016) propose a model composed of CNN, LSTM, and DNN for sarcasm detection. Lou et al. (2021) design a GCN-based model combining SenticNet (Cambria et al., 2020), dependency tree and LSTM with GCN (Kipf and Welling, 2017) together, which achieves state-of-the-art performance. Similar to previous studies, to better understand sarcasm, many approaches are able to utilize external information such as emoji expressions (Felbo et al., 2017), affective knowledge (Babanejad et al., 2020) and commonsense (Li et al., 2021). (Joshi et al., 2017) provide a more comprehensive survey. Moreover, there have been many systems developed for a shared task (Ghosh et al., 2020). We observe that these models are rarely designed to reflect the essential features of sarcasm phenomenon so that they can be applied to other classification tasks.
3 The Dual-Channel Network (DC-Net)

The architecture of the proposed DC-Net is shown in Figure 2. It consists of four modules: the decomposer, the literal channel, the implied channel, and the analyzer. Given an input text, we use decomposer to split it into two sub-sentences corresponding to the two channels. Then we use literal and implied channels to derive literal and implied representations independently. Lastly, analyzer predicts whether the text is sarcastic or not through detecting the sentiment conflict.

3.1 Decomposer

The decomposer module is designed to split input text to the literal and implied channels. In multiple sarcastic corpus, we observe that sarcastic texts often contain evident sentiment words. More specifically, the sentiment words directly reflect the literal sentiment, while the remaining text expresses the implied sentiment. For example, sentiment words of input text (e.g., “best gift”) represent positive, while the remained part (e.g., “Final exam is on my birthday”) implies the negative sentiment. Shown in Table 1, proportion of texts that contain sentiment words ranges from 88% to 96% in three datasets. Hence, using sentiment words to split input is well-suited for this scene.

Considering a text \( W_T = \{w_1, w_2, \ldots, w_N\} \) with \( N \) words, we decompose it into two pieces: the sentiment words \( W_L \), and the remaining text \( W_D \) (see Figure 2). \( W_L \) is fed to the literal channel, and \( W_D \) to the implied channel. In this process, we use sentiment lexicon (Wilson et al., 2005) to pick up sentiment words. If no sentiment words are matched from the given text, the original text is used as the literal channel’s input, which is the same as the implied channel. Note that sentiment words are mostly adjectives or adverbs, deleting them from sentences has no much influence on semantics. Although the text is not normative as expected after decomposing, we do not fill in the full text with placeholders like <MASK>.

Sarcasm as a complex linguistic phenomenon has various patterns. For example, text with word/phrase pair sentiment conflict is another pattern. Fortunately, sentiment conflicts are common in sarcasm texts. In this sense, the dual-channel framework could be further developed to detect more sentiment conflict patterns.

3.2 Literal Channel

The literal Channel includes an encoder, two linear layers, and a softmax classifier. \( \text{Encoder}_L \) is used to encode the literal text \( W_L \). Then we can get the literal representation \( v_L \) through

\[
v_L = \text{Encoder}_L(W_L).
\]

Next, we use a softmax layer to compute the literal sentiment distribution based on the literal representation \( v_L \).

\[
\mathcal{P}_l = \text{softmax}(W_r v_L + b_r),
\]

where \( W_r \) and \( b_r \) are parameters of the linear layer.

Considering the semantic complexity of sarcastic texts, a single representation of sentiment words may lose context information. So we adopt another \( \text{Encoder}_T \) to encode the original text \( W_T \) and obtain the representation of the entire text \( v_T \) through

\[
v_T = \text{Encoder}_T(W_T).
\]

Last, we concatenate the literal state \( v_L \) and the entire text’s state \( v_T \), followed by a linear layer and ReLU activation function to reduce dimension. Briefly, the final representation \( v'_L \) of the literal channel is formulated as:

\[
v'_L = \text{ReLU}(W_l[v_L; v_T] + b_l),
\]

where \( W_l \) and \( b_l \) are parameters of the second linear layer.

3.3 Implied Channel

In the implied channel, we also adopt an Encoder with the same structure but different parameters to encode the implied input text \( W_D \), and the representation of the implied channel is formulated as:

\[
v_D = \text{Encoder}_D(W_D).
\]
Similarly, we use $\text{softmax}$ to calculate the implied sentiment distribution based on the implied hidden state $v_D$:

$$P_d = \text{softmax}(W_z v_D + b_z),$$

where $W_z$ and $b_z$ are parameters.

Again, we concatenate the implied hidden state $v_D$ with the entire text’s hidden state $v_T$, followed by a linear layer and activation layer ReLU. The final representation $v'_D$ of the implied channel is formulated as:

$$v'_D = \text{ReLU}(W_d[v_D; v_T] + b_d),$$

where $W_d$ and $b_d$ are parameters.

Note that structures of the two channels are symmetrical, as both channels use Encoder models. However, the two encoders in the two channels do not share parameters, and their inputs are different. Since both channels are not specific to particular encoders, the Dual-Channel Framework is able to adapt to mainstream encoders, e.g., LSTM (Hochreiter and Schmidhuber, 1997), CNN (Kim, 2014), Recursive Neural Network (Socher et al., 2011), BERT (Devlin et al., 2019) et al. In DC-Net, we adopt Bi-LSTM as encoders for both channels.

### 3.4 Analyzer

The analyzer is designed to measure the conflict between the literal and the implied channels. We concatenate the literal representation $v'_L$ and the implied representation $v'_D$ and feed the result to a $\text{softmax}$ layer. Other analyzers such as subtraction or cosine similarity also fit our design.

$$P_s = \text{softmax}(W_p([v'_L; v'_D]) + b_p),$$

where $W_p$ and $b_p$ are parameters.

Although sarcasm has a strong correlation to literal sentiment and implied sentiment, we do not have gold labels for both sentiments. Hence, requesting the model to directly output sentiments on both channels may confuse the model. For this reason, we develop the objective function of sarcasm classification by adding objectives of the literal and implied channels.

### 3.5 Training Objective

The training objective of the proposed DC-Net model considers three aspects. One is to minimize the cross-entropy loss of the sarcasm probability distribution. The other two are to minimize the cross-entropy loss of the literal and the implied sentiment probability distribution respectively.

**Sarcasm Objective.** The sarcasm objective is to ensure the basic ability of detection. Hence, we use the cross-entropy loss of sarcasm classification. The objective $J_s$ is formulated as:

$$J_s(\theta) = \sum \text{cross-entropy}(y_s, P_s),$$

where $P_s$ denotes the sarcasm probability distribution of the text. The groundtruth of the sarcasm label is $y_s$.

**Literal Sentiment Objective.** In our implementation, we determine the literal sentiment label based on the number of words with positive sentiment and words with negative sentiment in the text. For sarcastic texts, if the number of positive words is greater than negative words, the literal sentiment label is positive and the implied sentiment label is negative, vice versa. For non-sarcastic texts, both the literal sentiment label and the implied sentiment label are the same, determined by the number of positive/negative sentiment words.

The literal sentiment classification objective is then formulated as:

$$J_l(\theta) = \sum \text{cross-entropy}(y_l, P_l),$$

where $P_l$ is the literal sentiment probability distribution. The label generated by the labeling processing of the literal sentiment is $y_l$.

**Implied Sentiment Objective.** We observe that literal sentiment and implied sentiment of sarcastic texts are often opposite. Using the implied labels based on the automatic labeling processing, we calculate the implied sentiment classification objective by

$$J_d(\theta) = \sum \text{cross-entropy}(y_d, P_d),$$

where $P_d$ denotes the implied sentiment probability distribution. The label generated by the labeling processing of the implied sentiment is $y_d$.

Considering the three objectives, we obtain the final objective function $L$ by adding them together:

$$L(\theta) = \lambda_1 J_s(\theta) + \lambda_2 J_l(\theta) + \lambda_3 J_d(\theta),$$

where $\theta$ is the parameter set of the model. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are used to leverage the contributions of the three objectives.
4 Experiment

4.1 Datasets and Implementation Details

We conduct experiments on three benchmark datasets: IAC-V1, IAC-V2, and Tweets. These datasets do not contain context information such as tweet posts and user profiles. All of them have been widely used in evaluating sarcasm detection. IAC-V1 is collected from the online political debates forum1. It is the subset of the Internet Argument Corpus (Lukin and Walker, 2017). Compared to tweets, texts of IAC are much longer and more normative. IAC-V2 (Oraby et al., 2016) contains more data than IAV-V1 (the two versions have a few overlaps). IAC-V2 divides sarcasm into three sub-types. We use the largest subset (general sarcasm) in our experiments. Tweets dataset is proposed in SemEval 2018 Task 3 (Hee et al., 2018a). There are three variations of the text in this dataset: (i) original texts, (ii) texts with hashtags removed, and (iii) texts with hashtags and emoji expressions removed. Hashtags like "#irony", are originally obtained from users. The hashtags are also used as prior knowledge for collecting sarcastic posts. In our experiments, we used the version without hashtags.

Table 1 reports the statistics. We observe that more than 88% of the texts contain sentiment word(s). Hence, it is reasonable to decompose the original text into sentiment words and non-sentiment words, as inputs to literal channel and implied channel, respectively. The number of instances in the three datasets is between 1k and 6k. Due to the small size, the split of train/valid/test is important to avoid over-fitting. For Tweets dataset, we follow the official train/test split. Then we randomly select 5% from training as valid sub-dataset. There is no official train/valid/test split for the two IAC datasets, so we split IAC datasets following the same ratio of Tweets. The baselines papers do not provide the split (or not conduct experiments on IAC datasets). As a result, we cannot directly adopt the results of baselines reported in their original papers. Hence, we re-implement all baseline models on IAC-V1 and IAC-V2 datasets.

There are another three datasets for sarcasm detection. Riloff et al. (2013) and Ptáček et al. (2014) propose another two datasets based on Tweets, but they only provide tweet IDs. Due to modified authorization status, lots of tweets are unavailable or deleted. For this reason, we could not experiment on the two Tweet datasets. Khodak et al. (2018) build a large self-annotated dataset from the Reddit forum platform. This dataset contains rich context information including posts, comments, responses, and authors. Since our work focuses on text-based sarcasm recognition, we do not use this dataset.

Implementation Details. We use Glove (Pennington et al., 2014) embeddings to initialize word vectors. The dimension of the vector is 300. There is a checkpoint every 16 mini-batch, and the batch size is 32. For Tweets dataset, the dropout on embeddings is set to 0, while for IAC datasets it is set to 0.5. Adam (Kingma and Ba, 2015) is used to optimize our model. The parameters $\beta_1$ and $\beta_2$ of Adam are set to 0.9 and 0.999. The learning rates for model parameters except word vectors are 1e-3, and 1e-4 for word vectors. Our model is implemented with Pytorch2 (version 1.7.0).

On IAC datasets, all of the loss contributions $\lambda_1, \lambda_2, \lambda_3$ of our DC-Net model are set to 1. On Tweets, they are set to 1, 1e-4, and 0.3, respectively. The hyperparameters are searched over the validation sub-dataset.

4.2 Compared Methods

We evaluate our model against the following baselines:

**UCDCC** (Ghosh and Veale, 2018) is a Siamese LSTM model exploiting Glove word embedding features. The method has designed a lot of rules in preprocessing Twitter data. It achieves the best performance on SemEval 2018 Task 3 Subtask A.

**THU-NGN** (Wu et al., 2018) consists of densely connected LSTMs based on word embeddings, sentiment features, and syntactic features. It ranks second on SemEval 2018 Task 3 Subtask A.

---

1http://www.4forums.com/political/

2https://pytorch.org

3https://nlds.soe.ucsc.edu/sarcasm1

4https://nlds.soe.ucsc.edu/sarcasm2

5https://github.com/Cyvhee/SemEval2018-Task3

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
<th>Avg $\ell$</th>
<th>$s$ ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAC-V1</td>
<td>1,596</td>
<td>80</td>
<td>320</td>
<td>68</td>
<td>91%</td>
</tr>
<tr>
<td>IAC-V2</td>
<td>5,216</td>
<td>262</td>
<td>1,042</td>
<td>43</td>
<td>96%</td>
</tr>
<tr>
<td>Tweets</td>
<td>3,634</td>
<td>200</td>
<td>784</td>
<td>14</td>
<td>88%</td>
</tr>
</tbody>
</table>

Table 1: Statistics of datasets. Avg $\ell$ denotes the average length of texts in the number of tokens. $s$ ratio is the proportion of texts that contain sentiment words.
Table 2: The precision, recall, and macro F1 of sarcasm recognition. The results marked with * are from Hee et al. (2018a). The best results are in boldface and second-best underlined.

<table>
<thead>
<tr>
<th>Model</th>
<th>IAC-V1</th>
<th>IAC-V2</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCDCC</td>
<td>58.6</td>
<td>58.6</td>
<td>58.5</td>
</tr>
<tr>
<td>THU-LSTM</td>
<td>64.4</td>
<td>64.3</td>
<td>64.2</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>64.6</td>
<td>64.6</td>
<td>64.6</td>
</tr>
<tr>
<td>AT-LSTM</td>
<td>65.9</td>
<td>65.5</td>
<td>65.3</td>
</tr>
<tr>
<td>CNN-LSTM-DNN</td>
<td>61.5</td>
<td>61.2</td>
<td>60.9</td>
</tr>
<tr>
<td>MIARN</td>
<td>65.6</td>
<td>65.2</td>
<td>64.9</td>
</tr>
<tr>
<td>ADGCN</td>
<td>64.3</td>
<td>64.3</td>
<td>64.3</td>
</tr>
<tr>
<td>DC-Net</td>
<td>66.6</td>
<td>66.5</td>
<td>66.4</td>
</tr>
</tbody>
</table>

**Bi-LSTM** (Hochreiter and Schmidhuber, 1997) is a variant of RNN, which could learn long-term dependencies and bidirectional information.

**AT-LSTM** (Wang et al., 2016) is an LSTM model followed by a neural attention mechanism. It could attend the important part of the input.

**CNN-LSTM-DNN** (Ghosh and Veale, 2016) is a combination of CNN, LSTM, and DNN. It stacks two layers of convolution and two LSTM layers, then passes the output to a DNN for prediction.

**MIARN** (Tay et al., 2018) learns the intra-sentence relationships of word pairs and the sequential relationships of the given text.

**ADGCN** (Lou et al., 2021) is a GCN-based method with a sentence graph and dependency graph. The initial input of GCN is the hidden state of Bi-LSTM.

### 4.3 Main Experiment Results

Table 2 shows that our DC-Net achieves the best macro F1 results across all datasets. On Tweets dataset, DC-Net achieves about 3.5% improvement in F1 score than the best baseline. On IAC-V2 dataset, our model outperforms the second-best baseline by 1.2% in F1. Surprisingly, compared with basic encoder model Bi-LSTM, our DC-Net boosts the performance up to 5% and 3% respectively on Tweets and IAC-V2, demonstrating the effectiveness of our dual-channel design. For Tweets dataset, the average length of texts is 14 words, which leads to a lack of information for sarcasm recognition. Nevertheless, our DC-Net improves 3.5% on F1 compared with the previous state-of-the-art ADGCN. It proves the effectiveness of our dual-channel framework.

Interestingly, UCDCC achieves the best precision 78.8% and accuracy 79.7% on Tweets dataset. Besides, THU-NGN gets the best recall 80.1% on Tweets. UCDCC designs targeted rules to preprocess the input text and it achieves the best performance on SemEval 2018 Task 3 Subtask A. Rules could improve precision effectively, but they cannot take recall into account at the same time. So the F1 is not good enough. The last place performance of UCDCC on IAC-V1/V2 also proves this point. These are because the designed rules are hard to fit missing instances and other domains. Similarly, THU-NGN uses linguistic knowledge such as semantic and syntactic, so it achieves the highest recall on Tweets but it cannot perform equally well on other datasets. That is, rules have limitations in handling this task.

The previous state-of-the-art ADGCN achieves second-best on IAC-V2 and Tweets. However, on IAC-V1 dataset, ADGCN performs not as well as the result reported in their paper. IAC-V1 dataset is relatively small so that the train/valid/test split has a significant impact. Our experiments also show that MIARN’s performance is not as good as expected. This indicates that the basic utilization of word pair correlation is not enough to improve the performance of sarcasm detection. Bi-LSTM, AT-LSTM, and CNN-LSTM-DNN methods are all based on LSTM. Thus the performances of these models on Tweets and IAC-V1 are close.

### 4.4 Comparison with BERT

BERT has contributed to significant improvements on various NLP tasks. To do a comprehensive comparison, we apply the Dual-Channel Framework to BERT (Devlin et al., 2019) model by using BERT as the encoder. The new model with BERT is named DC-Net (w/ BERT). Table 3 reports the experimental results.
Table 3: The precision, recall, and macro F1 of models including BERT, DC-Net with BERT as Encoder, and DC-Net with Bi-LSTM as Encoder.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre.</td>
</tr>
<tr>
<td>BERT</td>
<td>69.1</td>
</tr>
<tr>
<td>DC-Net (w/ BERT)</td>
<td>70.2</td>
</tr>
<tr>
<td>DC-Net (w/ Bi-LSTM)</td>
<td>76.4</td>
</tr>
</tbody>
</table>

As expected, the DC-Net (w/ BERT) model achieves significant improvement compared with the basic BERT. This result shows that our dual-channel framework is adaptable and effective. Interestingly, we observe that BERT-based methods perform not well enough compared with its huge improvement on other NLP tasks. This can be attributed to the fact that the corpus of pre-trained BERT contains more deterministic data (e.g., only one meaning without sentiment conflict). However, sarcasm is a niche linguistic phenomenon. The poor performance of BERT further reinforces that sarcasm recognition is a difficult task. It tells us that applying well-performed text classification methods directly does not lead to good performance.

4.5 Ablation Study

Recall that the model training objective (see Section 3) contains three objectives: sarcasm recognition, literal sentiment classification, and implied sentiment classification. To study the effect of the three objectives, we conduct ablation study on Tweets dataset.


<table>
<thead>
<tr>
<th>Objective</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre.</td>
</tr>
<tr>
<td>$J_s$</td>
<td>74.6</td>
</tr>
<tr>
<td>$J_s+J_d$</td>
<td>74.2</td>
</tr>
<tr>
<td>$J_s+J_l$</td>
<td>73.0</td>
</tr>
<tr>
<td>$J_s+J_l+J_d$</td>
<td>76.4</td>
</tr>
</tbody>
</table>

Table 4 lists results of ablation study. The model using only sarcasm recognition loss performs the worst. By adding literal and implied sentiment classification losses, the model’s performance improves 1.5 points on macro F1 score. As expected, the model with both literal and implied losses performs the best. It reveals that the dual channels complement each other. Conflict detection could recognize sarcasm when both of them are considered.

4.6 Effectiveness of DC-Net by Visualization

To verify the rationality and effectiveness of our proposed DC-Net, we adopt t-SNE (Van der Maaten and Hinton, 2008) to visualize high-dimensional vector representations based on the test sub-dataset of IAC-V2 (with largest data).

To figure out the effect of each channel, we visualize the representations of the literal channel and the implied channel. Figure 3(a) shows the visualization of literal representation $v_l$ and implied representation $v_d$. Recall that the decomposer module splits the original text into sentiment words and the remaining. We observe that there is a clear separation between literal and implied representations from Figure 3(a). This strongly indicates that our dual-channel framework is capable of effectively separating the representations of the two channels.

To get into the essence of sarcastic and non-sarcastic texts, we visualize the sarcastic and non-sarcastic representations from the overall and separate perspectives. Figure 3(b) shows the sarcastic and non-sarcastic representations in analyzer from the overall perspective. We observe that non-sarcastic texts focus on the upper right corner,
while sarcastic texts scatter on the lower left corner. It reveals that the sarcasm patterns are complex and changeable. Luckily, the dividing line between the two is relatively clear. Figure 3(c) and Figure 3(d) show the sarcastic representation and non-sarcastic representation of each channel, respectively. Figure 3(c) shows that for literal channel, the dividing line is not clear. This verifies that building literal representation is not enough to distinguish sarcasm. As a comparison, Figure 3(d) shows that for the implied channel, the sarcastic and non-sarcastic representations are well divided. It reveals that the implied channel is more distinguishing when detecting sarcasm. To this end, explicitly separating the literal and implied channels is necessary and effective.

### 4.7 Flexibility of Dual-Channel Framework

**Flexibility of encoder.** The dual-channel framework is flexible and generic, and can be realized by plugging in existing sarcasm recognition models, e.g., MIARN, or classification models, e.g., AT-LSTM, Bi-LSTM, and BERT. Therefore, we use these methods as the encoder to examine the flexibility of our proposed framework. The changing range on macro F1 from original baseline models to Dual-Channel models is shown in Table 5.

As expected, the performance of baseline models has different degrees of improvement on all datasets after applying dual-channel Framework. For relatively simple models such as MIARN and Bi-LSTM, the improvement could be up to 4.8%. Interestingly, for complex models like BERT, the improvement is up to 2.3%. As we mentioned earlier, the basic BERT performs not good enough because sarcasm is a niche linguistic scene and the training dataset of BERT contains few sarcasm texts. After applying the dual-channel framework to BERT, the performance for sarcasm recognition improves a lot. These prove that our designed framework is able to fit various encoders with a significant improvement.

**Flexibility of analyzer.** The analyzer module is used to measure the difference between the literal channel and the implied channel. As we described in Section 3.4, other analyzer methods such as concatenation and subtraction could be applicable. To this end, we compare different analyzer methods. Table 6 shows the results. We observe that concatenation performs better than subtraction on all datasets. It is because concatenation holds more useful information and DC-Net could compare the difference between the two input representations. However, subtraction only outputs the margin between the two representations. It loses the original values which also contain useful information.

### 5 Conclusion and Future Work

In this study, we argue that the essential characteristic of sarcastic text is the conflict between literal and implied sentiments in the same sentence. To this end, we propose a dual-channel framework to recognize sarcasm by decomposing the input text into the literal channel and the implied channel. Based on the framework, we develop DC-Net. DC-Net is capable of exploiting the literal sentiment by encoding the sentiment words of input text, and exploiting the implied sentiment by encoding the remaining text. Experiments show that the proposed DC-Net achieves state-of-the-art performance.

Our future work is planned in two main directions. For now, we use sentiment words as a static decomposer. To model more sarcasm patterns, dynamic decomposing will be explored. Accordingly, encoders for both channels should be re-designed based on the new dynamic decomposer. Another direction is to redesign the analyzer. In the current design, we adopt a soft weighting mechanism to detect sentiment conflict between the two channels. A redesigned analyzer may be able to output opposite sentiments directly, for more effective detection of sentiment conflict.
References


Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA, pages 161–169.


Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018b. We usually don’t like going to the dentist: Using common sense to detect irony on twitter. Comput. Linguistics, 44(4).


ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.


Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In COLING 2016, 26th International Conference on