Affective Event Classification with Discourse-enhanced Self-training

Yuan Zhuang, Tianyu Jiang, Ellen Riloff
School of Computing
University of Utah
{yyzhuang, tianyu, riloff}@cs.utah.edu

Abstract

Prior research has recognized the need to associate affective polarities with events and has produced several techniques and lexical resources for identifying affective events. Our research introduces new classification models to assign affective polarity to event phrases. First, we present a BERT-based model for affective event classification and show that the classifier achieves substantially better performance than a large affective event knowledge base. Second, we present a *discourse-enhanced self-training* method that iteratively improves the classifier with unlabeled data. The key idea is to exploit event phrases that occur with a coreferent sentiment expression. The discourse-enhanced self-training algorithm iteratively labels new event phrases based on both the classifier's predictions and the polarities of the event's coreferent sentiment expressions. Our results show that discourse-enhanced self-training further improves both recall and precision for affective event classification.

1 Introduction

In recent years, researchers have been tackling the problem of identifying *affective events*, which are events that have a positive or negative effect on people who experience the event. For example, events that are typically positive include being hired for a new job, breaking a sports record, or buying a home. Conversely, events that are typically negative include being fired from a job, breaking an arm, or having your house burn down. People's world knowledge about events and how they impact people is sufficient for humans to infer the affective state of someone who experiences such an event, even if that person does not explicitly express an emotion. Consequently, we will refer to these events as having positive or negative polarity with respect to an implicit affective state. Research has shown that recognizing affective events is important for a variety of natural language processing tasks, including narrative text comprehension and summarization (Lehnert, 1981; Goyal et al., 2013), dialogue systems (André et al., 2004), response generation (Ritter et al., 2011), and sarcasm detection (Riloff et al., 2013).

Much of the prior work on recognizing affective events has focused on producing lexical resources of verbs or event phrases with corresponding affective polarity values (Goyal et al., 2010, 2013; Rashkin et al., 2016; Ding and Riloff, 2016, 2018). These resources reflect substantial progress toward recognizing affective events in text, but their coverage is limited by their fixed content. We hypothesized that deep learning architectures that encode rich meaning representations could lead to a more effective approach for identifying affective events. Specifically, neural classification models have the capacity to generalize across lexically and syntactically different phrases that are semantically similar, and similar events are usually associated with the same affective polarity. To explore this approach, we created a BERT-based model for affective event classification and show that it recognizes affective events more effectively than a large affective event knowledge base.

Our research also introduces a *discourse-enhanced self-training* method that further improves affective event classification with unlabeled data. Self-training is a well-known method for using a classifier's own predictions on unlabeled instances to generate more training data. However, self-training has limitations. Using the most confident labels may not improve recall much because the new training instances are familiar, while using less confident labels often decreases precision because the training data becomes more noisy. To overcome these issues, we designed a *discourse-enhanced self-training* method that combines the

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5608–5617,
November 16–20, 2020.

classifier's predictions with information from local discourse contexts to robustly assign labels to new training instances.

The key to this approach is to exploit unlabeled event phrases that occur near coreferent sentiment expressions. Specifically, we extract event phrases that are followed by a sentiment expression in a syntactic structure that suggests it likely refers to the event. For example, consider the statement *"I got engaged today. It is exciting."*. "It" refers to the act of getting engaged, so the positive sentiment applies to that event. Our algorithm then predicts the affective polarity for unlabeled events using both the classifier's prediction for the event phrase as well as the associated sentiment expressions. We show that our discourse-enhanced self-training method improves both recall and precision for affective event classification.

2 Related Work

Several lines of research have focused on the problem of recognizing events that have implicit affective states. Research on narrative understanding used bootstrapped learning to identify *patient polarity verbs*, which impart affective polarity to their patients (Goyal et al., 2010, 2013). Vu et al. (2014) extracted "emotion-provoking events" using the seed pattern *"I am* < EMOTION > *that* < EVENT >, pattern expansion, and clustering. Reed et al. (2017) learned lexico-syntactic patterns associated with first-person affect to improve affective sentence classification alongside supervised learners. Li et al. (2014) extracted "major life events" from Twitter by clustering tweets that occurred with speech act words, such as "congratulations" or "condolences". But their work did not assign affective polarity to events, and focused only on major life events that prompt expressive speech acts. Our work has a broader scope, aiming to recognize everyday events as well (e.g., *being hungry* is negative, and *seeing a rainbow* is positive).

Work in opinion analysis created a +/- EffectWordNet (Choi and Wiebe, 2014) to recognize the effects of events on entities, although the effects are not necessarily "affective" because the entities need not be animate (e.g., *baking a cake* has a positive effect on the cake because it is created). Subsequent work developed implicature rules to use +/- effects for opinion analysis (Deng and Wiebe, 2014, 2015). There has also been work on recognizing the connotation of words and senses (Kang et al., 2014) and connotation frames (Rashkin et al., 2016), which infer connotative polarities for a verb's arguments from the writer's and entity's perspective. These efforts associated polarity with individual verbs, not event phrases.

Saito et al. (2019) used discourse relations to propagate affective polarity from seeds using a Japanese web corpus. They extracted events that co-occur with seeds in a large corpus, then used discourse relations as constraints in the learning process. Another line of related work is Emotion Cause Extraction, which links emotion expressions to the events that caused the emotion (Gui et al., 2016, 2017; Chen et al., 2018; Li et al., 2018; Xia and Ding, 2019). This research uses datasets created from Chinese news and microblogs that contain an explicitly mentioned emotion. This work assigns polarity to events in the context of a specific text passage. In contrast, our work aims to identify the *prior* affective polarity of an event, irrespective of context. Consequently, our classifier can be used to predict the affective polarity of events in contexts that do not contain any explicit emotion or sentiment indicators.

Our research is most closely related to the work by (Ding and Riloff, 2016, 2018), which identifies stereotypically affective events and their *prior polarity*, irrespective of context. The Affective Event Knowledge Base (AEKB) produced by (Ding and Riloff, 2018) contains over half a million event phrases coupled with polarity labels. These events were extracted from nearly 1.4 million personal blog posts in the ICWSM 2009 and 2011 Spinn3r datasets[1]. The polarity labels were generated automatically using a weakly supervised method. Their approach optimizes for semantic consistency over a graph of event nodes that are linked by edges capturing three types of semantic relations.

Our discourse-enhanced self-training algorithm adds a new twist to traditional self-training methods (Mihalcea, 2004; Kehler et al., 2004; McClosky et al., 2006). The approach is also reminiscent of co-training (Blum and Mitchell, 1998), which trains two classifiers based on independent views of the data. However in co-training, each classifier must be able to make reliable predictions on its own. We do not expect the coreferent sentiment expressions used by our approach to be sufficient by themselves because they are quite noisy (e.g., due to imperfect coreference, imperfect sentiment

[1] http://www.icwsm.org/data/

Method	F1	POS		NEG		NEU	
		Pre	Rec	Pre	Rec	Pre	Rec
Blogs	71.4	75.7	55.1	70.4	63.3	79.3	88.5
Twitter-found	65.2	72.2	40.6	78.7	60.8	65.6	87.9
Twitter-all	50.8	72.2	26.2	78.7	37.1	65.6	61.8

Table 1: Performance of AEKB across data sets.

labels, and issues like sarcasm). The strength of our method is that this signal can serve alongside the main classifier to produce a diverse new set of high-quality labels.

3 Creating Affective Event Classifiers

3.1 Motivation

Ding and Riloff (2018) created an Affective Event Knowledge Base (AEKB) that contains over 571,000 English event phrases labeled with affective polarity (positive, negative, or neutral). The AEKB was automatically generated from a corpus of personal blogs and is currently the largest resource of event phrases with polarity labels for the English language. We were curious to understand how effective the AEKB is at recognizing affective events in new texts. Twitter is another form of social media where we expect to find many affective expressions, so we created a new data set for affective event recognition in tweets to evaluate the generality of the AEKB and our new classifiers.

We produced a new dataset for affective events that contains 1,500 event phrases extracted from Twitter paired with manually assigned polarity labels. Section 4 describes the data creation process and gold standard annotation effort in detail. We represented events using a 4-tuple similar to the event representation in the AEKB: ⟨*Agent*, *Predicate*, *Theme*, *Prepositional Phrase (PP)*⟩.[2] We then evaluated the coverage and accuracy of the AEKB on our Twitter data. Every Twitter event was matched against the AEKB and, if a match was found, the polarity found in the AEKB was assigned to the event. Table 1 shows the results as a macro-averaged F1 score, alongside recall and precision for each of the three polarities: positive (POS), negative (NEG), and neutral (NEU).

The first row (Blogs) shows the results originally reported in (Ding and Riloff, 2018) for events extracted from blog posts, for comparison. Of the 1,500 Twitter events, only 997 events (66%) were found in the AEKB. The second row of Table 1 (Twitter-found) shows results for these 997 events. The overall performance is fairly similar across Twitter and blogs, except that recall for positive polarity is substantially lower. The lower precision for neutral polarity suggests that many positive Twitter events are labeled as neutral in the AEKB.

Another issue is that one third (34%) of the Twitter events were not found in the AEKB at all. The third row of Table 1 (Twitter-all) shows the results across *all* 1,500 Twitter events, where the missing events are left unlabeled. Overall, only 37% of the negative events and 26% of the positive events could be recognized by the AEKB.

These results show that despite its large size, the AEKB cannot recognize many affective events for two reasons: (1) the AEKB's precision is not perfect, so some positive and negative events are labeled as neutral, and (2) many affective events are not present in the knowledge base. Our research addresses these limitations by exploring whether classification models can achieve better coverage and accuracy by generalizing across events.

3.2 A BERT-based Affective Event Classifier

Our goal is to design a classifier that can label an event tuple with affective polarity. Representations produced by the transformer-based BERT model (Devlin et al., 2019) have achieved state-of-the-art performance across a variety of NLP tasks, so we used the pre-trained $\text{BERT}_{\text{BASE}}$ as the basis for our classifier and performed fine-tuning during the training.

The input is the sequence of tokens that comprise an event tuple. For example, ⟨*I, ride, bike, -*⟩ is converted into the sequence *"I ride bike"*. We use the uncased version of the *BERT base* model as our encoder. We use the 768-dimension output embedding of the special token [CLS], and pass the output vector of the special token [CLS] to a fully connected layer with softmax to produce a probability distribution over the three polarity classes. Each input event is then assigned the polarity with the highest probability value. We will refer to this model as **Aff-BERT**.

3.3 Experimental Results for Blogs Data

Baselines We developed two baselines to compare with Aff-BERT. The first model is a 1-layer LSTM. We first use ELMo (Peters et al., 2018) to encode an event sequence and feed the last layer of ELMo's outputs into the LSTM. The LSTM outputs a polarity distribution for the event. The

[2]The main difference is that we also allowed adjectival modifiers in noun phrases.

Method	F1	POS		NEG		NEU	
		Pre	Rec	Pre	Rec	Pre	Rec
AEKB	71.4	**75.7**	55.1	70.4	63.3	79.3	**88.5**
Aff-BERT(AEKB)	73.6	73.2	56.6	75.6	69.5	80.9	**88.5**
ELMo+Linear(Gold)	62.3	56.0	53.7	56.2	51.3	78.2	81.4
ELMo+LSTM(Gold)	70.5	71.4	60.8	70.8	57.3	81.3	**88.5**
Aff-BERT(Gold)	**77.4**	71.7	**66.2**	**78.2**	**77.2**	**85.0**	87.4

Table 2: Performance on the blogs test set.

second baseline is a linear classifier, which takes as input the average of the last layer of ELMo's outputs and produces a polarity distribution.

The LSTM has a hidden size of 512 and a dropout rate of 0.2. The learning rate is 0.01 for the LSTM, 0.1 for the linear classifier, and 1e-5 for Aff-BERT. We train all models for 5 epochs with a batch size of 50 and a linear warmup rate of 10% using AdamW optimizer.

Experiments Our first set of experiments evaluates Aff-BERT on the same blogs data that Ding and Riloff (2018) used to evaluate their AEKB. The validation and test data sets contain 490 and 1,000 manually annotated events, respectively.

The first row of Table 2 shows the results originally reported by (Ding and Riloff, 2018) for comparison. The second row shows the results when training Aff-BERT with the events that have polarity labels with predicted scores ≥ 0.6 in the AEKB.[3] It shows that Aff-BERT trained with the AEKB data performs better than the AEKB itself. The substantial recall gain for negative events is likely due to the generalization power of BERT's representations.

Next, we experimented with learning from gold labeled data by performing 10-fold cross-validation on the blogs test data. The third, fourth, and fifth rows of Table 2 show the results for the linear classifier, LSTM and Aff-BERT, respectively, trained with gold data. While the linear classifier and the LSTM do not perform as well as the AEKB, Aff-BERT trained on gold labeled data performs substantially better than both the AEKB and Aff-BERT trained on the AEKB. This shows that fine-tuning BERT on a relatively small amount of gold labeled data produces a strong affective event classifier, with respect to both recall and precision.

The strength of this model led us to wonder whether classification performance could be further improved by self-training with unlabeled data. As we will describe in Section 5, standard self-training produced only a small improvement, but we developed a new discourse-enhanced self-training algorithm that achieved bigger performance gains. In the next section, we describe how we collected events with coreferent sentiment expressions for the discourse-enhanced self-training algorithm.

[3]We tried score thresholds from 0 to 1 with the increment of 0.1, and 0.6 gave the best result.

4 Harvesting Events with Coreferent Sentiment Expressions

The key idea behind our approach is to create a self-training method that uses not only the classifier's own predictions but also a secondary source of information derived from local discourse contexts. Intuitively, the secondary signal confirms the classifier's prediction when they agree, or creates doubt about the classifier's prediction when they disagree. By taking both signals into account, we can assign high-quality labels to a diverse set of new examples in each cycle, which creates a robust self-training process.

From this point on, we turn our attention to Twitter because it is a vast resource that we can query to acquire a large set of event phrases in specific contexts, and where people share their everyday experiences. We acquire our unlabeled data by searching for event phrases on Twitter that occur with coreferent sentiment expressions. We use a heuristic to identify sentiment expressions that are likely to refer to an event in the preceding sentence. Specifically, we look for sentiment expressions that begin a sentence and match one of the following forms:

(a) {**this/that/it/I**}, {**be/feel/seem**}, {**ADJ+**}
(b) {**this/that/it**}, {**be/feel/seem**}, {**ADJ* N+**}

where the head adjective (ADJ) or head noun (N) is a sentiment term with positive or negative polarity. The sentiment expression cannot be followed by any events in its sentence and must follow a sentence that contains at least one event. Given these restrictions, the pronouns "this', "that", and "it" are likely referring to an event in the previous sentence, although this is not guaranteed. Similarly, the pronoun "I" is referring to the speaker who is likely expressing their sentiment toward something that was just mentioned, which is often (though not always) the prior event. We will call the phrases that match these patterns **coreferent sentiment expressions** because they express a sentiment that refers back to something mentioned earlier.

We found that the syntactic constructions above typically convey a sentiment about an event in the

Tweet1:	I rode a horse today! *That was fun.* ⟨**I, ride, horse, -**⟩
Tweet2:	Someone was abducted on the street right next to mine. *It's terrifying.* ⟨**-, abduct, someone, on street**⟩
Tweet3:	Disrupting my daily routine and alienating many people. *I am angry* ! ⟨**-, disrupt, my daily routine, -**⟩ ⟨**-, alienate, people, -, -**⟩

Table 3: Examples of harvested tweets and extracted events.

prior sentence, but this heuristic is not perfect. For example, the sentiment sometimes applies to an object in the prior sentence and not an action (e.g., *"I bought a book. It is excellent"* describes an excellent book and not an excellent buying experience). Nevertheless, the self-training algorithm will use this data in the aggregate, so some noise can be tolerated. In the following sections, we describe each step of the Twitter data harvesting process.

4.1 Creating Sentiment Queries

We create an initial set of sentiment queries for Twitter by instantiating the syntactic patterns shown earlier with 3,010 subjective adjectives and 2,023 nouns from the MPQA lexicon (Wilson et al., 2005). We also use the 1,147 words labeled with "anypos" in MPQA as an adjective and a noun to instantiate the patterns. For example, given the adjective *"good"*, we exhaustively generate all phrases that match the regular expression: "{that/this/it/I} {be/feel/seem} *good*", such as *"That is good"* and *"I feel good"*.

We then download tweets that contain these phrases. If the context around the sentiment expression satisfies the constraints mentioned earlier, then we extract the events in the previous sentence as affective event candidates. Table 3 shows three tweets that were retrieved with queries for the sentiment expressions in *italics* along with the events extracted from each tweet in **boldface**.

4.2 Creating Event Queries

Next we can use the extracted events to harvest more tweets with coreferent sentiment expressions.

Searching for phrases that match an event is not trivial. The Twitter API only supports exact phrase matching but an event is represented as a tuple (<Agent, Predicate, Theme, PP>). Furthermore, the components in an event tuple contain lemmatized head words. We want to construct queries that will retrieve phrases containing morphological variations (e.g., "drove" for the lemma "drive") as well as modifiers preceding heads (e.g., "a fancy car" instead of just "car"). To circumvent this problem, we generate text spans for each event tuple from the original tweets that it was extracted from. The text span contains all words between the leftmost word and the rightmost word of the tuple. Then we apply the PrefixSpan algorithm (Saraf et al., 2015) to compute the frequency of all subsequences of words. For each event tuple, we create queries from the 20 most frequent subsequences that contain all words in the event tuple. For example, <he, drive, car> might yield queries such as *"he drove a fancy car"*, *"he has driven my car"*, etc.

After we retrieve tweets that match an event query, we apply the same constraints as before but in reverse: the sentence that mentions the event must be followed by a coreferent sentiment expression matching our patterns. In this step, we assume that unknown terms in the ADJ or N position of the patterns are sentiment-bearing, allowing us to identify new sentiment expressions. We found this heuristic to be quite good and produced some interesting affective terms that are not in the MPQA lexicon. For example, the new negative terms include *"cyberbullying"*, *"yucky"* and *"gutless"*, and the new positive terms include *"record-breaking"*, *"reassuring"* and *"heart-warming"*.

4.3 Iteratively Harvesting Events

The first step of data harvesting creates sentiment queries from the MPQA lexicon and extracts new event phrases. The second step of data harvesting creates event queries and extracts new sentiment phrases. Given these building blocks, we create a cycle that alternates these steps, iteratively harvesting new events with associated sentiment expressions. In each iteration, we form queries for sentiment or event phrases that have frequency ≥ 5 and have not been used as queries previously. We download 5,000 tweets for each event query and 1,000 tweets for sentiment expression query.[4] Finally, we discard retweets and duplicated tweets[5]. To be consistent with the criteria used for affective

[4]Many tweets collected by event queries contain no coreferent sentiment expression, so we downloaded more tweets for event queries to increase the number of matched instances.

[5]A tweet is duplicated if it shares 6 or more consecutive words with another tweet.

events in the AEKB by (Ding and Riloff, 2018), we also discarded events that did not contain a first-person reference or a family member term.[6]

We ran the harvesting process over Twitter for 4 iterations, after which few new events were found. The final dataset contains 2,068,600 unique event tuples and 15,494 unique sentiment expressions.

4.4 Gold Dataset Creation

We created a gold standard dataset for affective events from Twitter (**Twitter Dataset**) by having two human annotators label 1,500 randomly selected events of frequency ≥ 5. Each event was labeled as positive, negative, or neutral using the same criteria defined by (Ding and Riloff, 2018) for the AEKB. The pairwise inter-annotator agreement using Cohen's kappa was .75. The two annotators then adjudicated their disagreements to produce the final set of gold labels. The final dataset contains 435 (29%) positive, 348 (23%) negative and 717 (48%) neutral events. This new evaluation dataset and the collection of the unlabeled harvested events are publicly available at https://www.cs.utah.edu/~yyzhuang/.

5 Discourse-enhanced Self-training

We designed an enhanced self-training algorithm that learns from unlabeled data by iteratively labeling new instances using both the affective event classifier's prediction as well as polarities associated with the event's discourse contexts. We will refer to this method as Discourse-enhanced Self-training. The intuition is that (1) new instances are labeled only if both sources of information agree, which yields high-quality labels, and (2) a more diverse set of instances will be labeled than if only the classifier's most confident predictions were used.

Figure 1 illustrates how an unlabeled event is scored during Discourse-enhanced Self-training. Each event is paired with the set of coreferent sentiment expressions that occurred with it in our Twitter dataset. The affective event classifier is applied to the event and generates a probability distribution over the three polarity values. In parallel, an external sentiment classifier produces a probability distribution over the polarity classes for each of the coreferent sentiment expressions. The probability distributions are then averaged to produce

[6](Ding and Riloff, 2018) also discarded events that only mentioned other people, but we did not apply this restriction due to the difficulty of recognizing people terms in tweets.

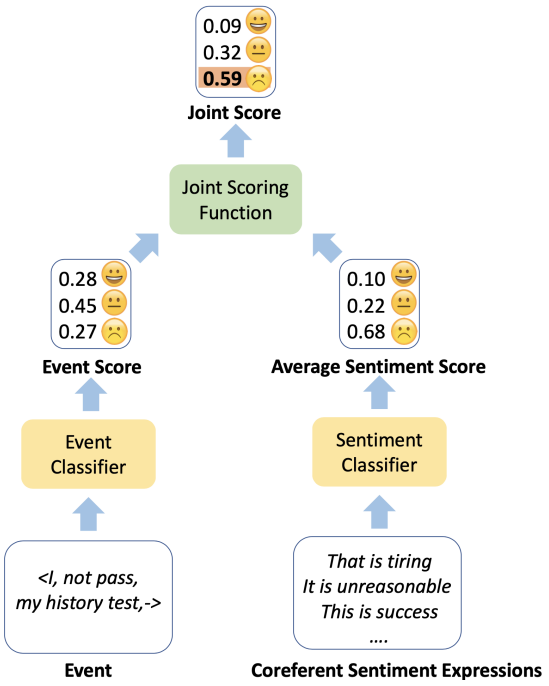

Figure 1: Illustration of Discourse-enhanced Scoring

an average probability distribution for the set of sentiment expressions as a whole. Finally, a joint scoring function takes the two probability distributions and produces a joint probability distribution for the event. The polarity with the highest probability is used as the event's label.

Algorithm 1 outlines our Discourse-enhanced Self-training procedure in detail. The process begins with a gold labeled set of events E_L, a set of unlabeled events E_U where each event e_i in E_U is paired with a set of coreferent sentiment expressions CSE_i, an external sentiment classifier, and two confidence thresholds θ_{jnt} and θ_{neu}. Each iteration starts by training the event classifier on E_L. The event classifier is then applied to every unlabeled event e_i in E_U to produce an event score vector s_{e_i}. Next, the sentiment classifier is applied to every coreferent sentiment expression cse in CSE_i to produce a polarity distribution. Then the polarity distributions of all cse in CSE_i are averaged to produce an average polarity distribution $\bar{s}_{\text{CSE}_i}$ for the whole set CSE_i. The joint scoring function then produces a joint score vector s_{jnt_i} for the event e_i by the equation below:

$$s_{jnt_i} = \frac{s_{e_i} \odot \bar{s}_{\text{CSE}_i}}{s_{e_i} \cdot \bar{s}_{\text{CSE}_i}}, \quad (1)$$

where $\odot$ denotes element-wise multiplication and $\cdot$ denotes dot product. Conceptually the joint scoring function gives equal weight to the event classifier

Algorithm 1: Discourse-enhanced Self-training

Input: Labeled events E_L, Unlabeled events E_U where each event e_i has an associated set of coreferent sentiment expressions CSE_i, an external Sentiment Classifier, and thresholds θ_{jnt} and θ_{neu}

while *E_U is not empty and not maximum iteration* **do**
- Train the Event Classifier over E_L
- For each $e_i \in E_U$, apply the Event Classifier to get an event score
- For each $e_i \in E_U$, apply the Sentiment Classifier to each $cse \in CSE_i$ and compute the average cse sentiment score
- Compute the joint score for each $e_i \in E_U$ by Eqn. 1
- Label new events (E_{jnt}) based on the joint scores and θ_{jnt}
- Label additional neutral events (E_{neu}) based on the event scores and θ_{neu}
- Update E_L and E_U:
  $E_L = E_L \cup E_{jnt} \cup E_{neu}$
  $E_U = E_U - E_{jnt} - E_{neu}$

end

and the sentiment classifier in the final decision of the label. Finally, each event e_i is assigned the polarity with the highest value in s_{jnt_i}.

We generate a set of new labeled events E_{jnt} by assigning labels to unlabeled events that have a polarity probability $\geq \theta_{jnt}$ based on the joint scores. All other events remain unlabeled. However, we found that this process labels relatively few events as neutral. Since neutral events can also co-occur with positive and negative sentiment expressions, they may have relatively low neutral scores. To better maintain the distribution of events over all three polarities, we also add a new set of events E_{neu}, which the event classifier predicts as neutral with confidence $\geq \theta_{neu}$.

Discourse-enhanced Self-training needs an external sentiment classifier, so we fined-tuned a BERT-based model with the gold standard Twitter dataset from SemEval-2017 (Rosenthal et al., 2017) following the experiment setups in Section 3.2 and Section 3.3. In our experiments, we set θ_{neu} to 0.9 and θ_{jnt} to 0.95 based on the model's performance over the validation set.

6 Experimental Results

We performed 10-fold cross validation over the gold Twitter Dataset, where each of the 10 runs used 80% of the data (8 folds) for training, 10% of the data (1 fold) for validation/tuning, and 10% of the data (1 fold) for testing. We compare Discourse-enhanced Self-training (**DEST**) with strictly supervised learning and traditional self-training. During each iteration of the traditional self-training model, the affective event classifier Aff-BERT is applied to each unlabeled event. Events with polarity score $\geq$ 0.9 are selected as new labeled data. We chose 0.9 as the threshold based on the model's performance on the validation set.

For the DEST model, to ensure a rich set of discourse contexts, we only used unlabeled events that (a) had at least 10 distinct coreferent sentiment expressions and (b) did not include "this", "that" or "it" as a subject or object of the event phrase because an event is often vague without knowing what the pronoun refers to. This resulted in 8,532 events in the unlabeled event set.

Method	Precision	Recall	F1
Supervised	76.5	75.2	75.7
Self-training	77.6	77.2	77.0
DEST	**79.6**	**78.7**	**79.0**

Table 4: Results for learning from unlabeled data.

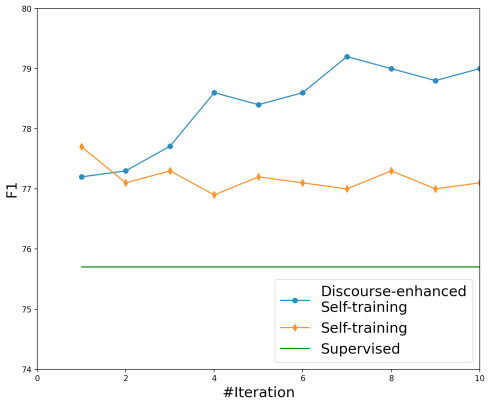

Figure 2: Learning curves through 10 iterations.

6.1 Results

Table 4 reports the performance of the models after 10 iterations of learning with unlabeled data, where the first row shows the results for Aff-BERT trained only with gold labeled data for comparison. For both self-training models, no new examples were labeled after 10 iterations. Table 4 shows that ordinary self-training produced small gains in both precision and recall. Our Discourse-enhanced Self-training algorithm achieved larger gains, improving precision over the supervised model from 76.5% $\rightarrow$ to 79.6% and improving recall from 75.2% $\rightarrow$ 78.7%.

Figure 2 shows the learning curves for each method over the 10 iterations based on their F1

Method	POS		NEG		NEU	
	Pre	Rec	Pre	Rec	Pre	Rec
Supervised	74.4	71.5	**79.0**	74.0	76.1	80.1
DEST	**81.8**	**74.8**	78.4	**80.0**	**79.4**	**82.4**

Table 5: Recall and precision across polarities.

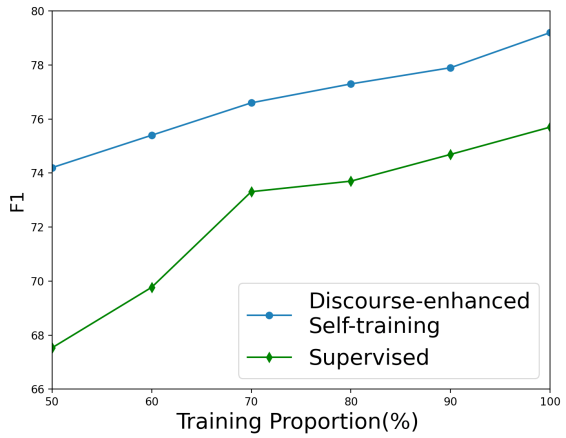

Figure 3: Learning curves of models with training sets of different sizes.

score. The flat line is the F1 score for Aff-BERT trained with only gold labeled data. Ordinary self-training produced its highest F1 score after the first iteration, then declined and stayed stable without further improvement. In contrast, the learning curve of Discourse-enhanced Self-training gradually ascends, reaching its peak in iteration 7 and showing signs that it could potentially exceed that peak with more unlabeled data.

Table 5 shows the performance breakdown across the three polarities. Discourse-enhanced Self-training improved both precision and recall for all polarities, except that precision was slightly lower for negative polarity. Most notably, DEST achieved a 6.0% absolute gain in recall for negative polarity, and a 3.3% absolute gain in recall for positive polarity, alongside a 7.4% absolute gain in precision.

We also generated learning curves for the supervised learner and Discourse-enhanced Self-training when trained with different amounts of labeled data . Figure 3 shows results when using 50% to 100% of the gold training data in increments of 10%. Discourse-enhanced Self-training showed even greater relative improvement over the supervised learner when only 50% of the gold data was used for training. In addition, when using about 60% of the gold data, DEST achieved performance comparable to the supervised learner trained with 100% of the data.

Incorrect → Correct	
Neutral → Positive:	
⟨I, see, exhibit, -⟩	⟨I, sleep, -, through whole night⟩
⟨I, get, tip, -⟩	⟨I, start, my new job, -⟩
Neutral → Negative:	
⟨I, need, air, -⟩	⟨-, separate, child,from parent⟩
⟨I, not get, reply, -⟩	⟨someone, unfollow, me, -⟩
Correct → Incorrect	
Neutral → Positive:	
⟨I, have, your book, -⟩	⟨I, watch, guy, -⟩
Neutral → Negative:	
⟨I, have, brace, -⟩	⟨I, have, comment, -⟩

Table 6: Examples of labels that are changed by the joint scoring function.

Overall, the Discourse-enhanced Self-training approach produced substantial gains over fully supervised learning, and achieved more robust learning from unlabeled data than ordinary self-training. This approach could be applied to many other types of problems as well, when a secondary source of information relevant to the task is available.

7 Analysis

To better understand the behavior of the resulting classifier, we did a manual analysis of events whose polarity was impacted by the coreferent sentiment expressions. The top portion of Table 6 shows examples of events for which the affective event classifier assigned an incorrect polarity but the joint scoring function produced the correct polarity. We saw many cases like these where the event phrase contained neutral words but the coreferent sentiment expressions revealed consistently positive or negative discourse contexts.

The bottom portion of Table 6 shows examples of events for which the affective event classifier assigned a correct polarity but the joint scoring function assigned an incorrect polarity. We observed two types of issues that caused this behavior. One common problem was incorrect coreference. Sometimes the sentiment was coreferent with the subject or object of the event, but not the event itself. For example, ⟨*I, have, your book,* -⟩ was followed by sentiments about the book itself (e.g., *"It is well-written"* and *"That is inspiring"*). In other cases the sentiment was coreferent with an event earlier in the discourse. These errors suggest that incorporating a better event coreference resolution

algorithm would likely improve results.

We also found some events that were correctly labeled as positive by the affective event classifier but labeled as negative by the sentiment classifier with high confidence, and consequently the event classifier's correct predictions were overridden. Most of these cases were expressions of love or empathy in response to negative events, such as ⟨*God, help, us, -*⟩, ⟨*my heart, go, -, to family*⟩, ⟨*you, have, my sympathy, -*⟩. This is an interesting phenomenon that may require better discourse modeling, including the recognition of expressive speech acts.

8 Conclusion

In this work, we proposed a BERT-based supervised classifier for affective event recognition and showed that it substantially outperforms a large affective event knowledge base. We also designed a novel discourse-enhanced self-training algorithm to leverage unlabeled data iteratively. By combining both the affective event classifier's prediction and the polarities of coreferent sentiment expressions, our algorithm substantially improved upon the supervised learning results. The resulting classification model is substantially more effective for affective event recognition than previous methods. We also believe that the general idea behind our discourse-enhanced self-training approach could be useful for many other types of problems where additional information can be extracted from larger contexts to serve as a secondary signal to help confirm or disconfirm a classifier's predictions.

References

Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp. 2004. Affective Dialogue Systems: Tutorial and Research Workshop. In *Lecture Notes in Computer Science*, volume 3068. Springer.

A. Blum and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint Learning for Emotion Classification and Emotion Cause Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014)*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment Propagation via Implicature Constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*.

Lingjia Deng and Janyce Wiebe. 2015. Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

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 (HLT/NAACL 2019)*.

Haibo Ding and Ellen Riloff. 2016. Acquiring Knowledge of Affective Events from Blogs using Label Propagation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*.

Haibo Ding and Ellen Riloff. 2018. Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

A. Goyal, E. Riloff, and H. Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2013. A Computational Model for Plot Units. *Computational Intelligence*, 29(3):466–488.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A Question Answering Approach for Emotion Cause Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-Driven Emotion Cause Extraction with Corpus Construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning Connotation over the Word+Sense Network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

A. Kehler, D. Appelt, L. Taylor, and A. Simma. 2004. Competitive Self-Trained Pronoun Interpretation. In *Proceedings of the Annual Meeting of the North American Chapter of the Associ ation for Computational Linguistics: Human Language Technologies (HLT/NAACL 2004)*.

Wendy G Lehnert. 1981. Plot Units and Narrative Summarization. *Cognitive Science*, 5(4):293–331.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A Co-Attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

D. McClosky, E. Charniak, and M Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Annual Meeting of the North American Chapter of the Associatio n for Computational Linguistics: Human Language Technologies (HLT/NAACL 2006)*.

R. Mihalcea. 2004. Co-training and Self-training for Word Sense Disambiguation. In *Proceedings of the Eighth Conference on Natural Language Learning (CoNLL 2004)*.

Matthew 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 (HLT/NAACL 2018)*.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Lena Reed, JiaQi Wu, Shereen Oraby, Pranav Anand, and Marilyn A. Walker. 2017. Learning lexico-functional patterns for first-person affect. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*.

Jun Saito, Yugo Murawaki, and Sadao Kurohashi. 2019. Minimally Supervised Learning of Affective Events Using Discourse Relations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP/IJCNLP 2019)*.

Pratik Saraf, R. Sedamkar, and Sheetal Rathi. 2015. PrefixSpan Algorithm for Finding Sequential Pattern with Various Constraints. *International Journal of Applied Information Systems*, 9:37–41.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a Dictionary of Emotion-Provoking Events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Rui Xia and Zixiang Ding. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.