Multiple Instance Learning for Offensive Language Detection

Jiexi Liu, Dehan Kong, Longtao Huang, Dinghui Mao, Hui Xue
Alibaba Group
{liujiexi.ljx,kongdehan.kdh,kaiyang.hlt,maodinghui.dh,hui.xueh}@alibaba-inc.com

Abstract

Automatic offensive language detection has become a crucial issue in recent years. Existing researches on this topic are usually based on a large amount of data annotated at sentence level to train a robust model. However, sentence-level annotations are expensive in practice as the scenario expands, while there exist a large amount of natural labels from historical information on online platforms such as reports and punishments. Notably, these natural labels are usually in bag-level corresponding to the whole documents (articles, user profiles, conversations, etc.). Therefore, we target at proposing an approach capable of utilizing the bag-level labeled data for offensive language detection in this study. For this purpose, we formalize this task into a multiple instance learning (MIL) problem. We break down the design of existing MIL methods and propose a hybrid fusion MIL model with **mutual-attention** mechanism. In order to verify the validity of the proposed method, we present two new bag-level labeled datasets for offensive language detection: **OLID-bags** and **MINOR**. Experimental results based on the proposed datasets demonstrate the effectiveness of the mutual-attention method at both sentence level and bag level.

1 Introduction

Detection of offensive content online has attracted widespread attention in recent years. Offensive content could not only be human-produced on social media platforms such as Twitter and Facebook, but also could be system-generated due to the pervasive usage of pre-trained language models (Gehman et al., 2020). To tackle the problem in practice, a general solution is to train models capable of identifying messages containing offensive language. Then the identified messages are checked and processed by human moderators. There have been a great many studies on dealing with offensive language based on deep learning approaches (Pitsilis et al., 2018; Pitenis et al., 2020).

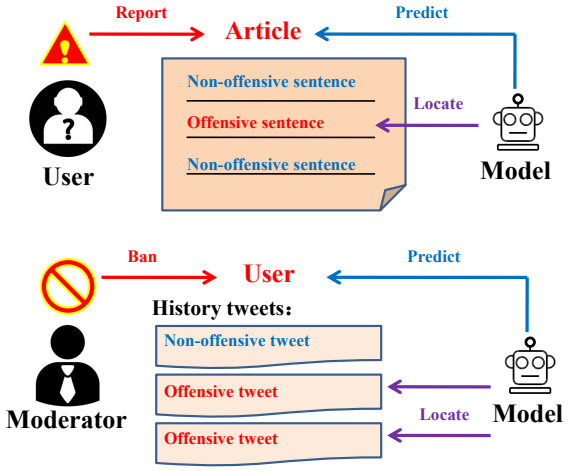


Figure 1: MIL scenarios in offensive language detection. Natural labels such as reports and bans are associated with the entire bag. The model is supposed to learn to predict the bag label and locate all offensive instances at the same time.

Previous researches mainly work on detection task based on sentence-level annotated corpus from a specific resource (Zampieri et al., 2019; Kumar et al., 2018), which requires massive manual effort. Such approaches are effective but resource-consuming, especially when transferring to a new platform or language. In social media, there are many existing information sources that could substitute for manual labeling. Historical records of online platforms like user feedback (report or dislike) and punishments made by moderators could act as supervision. Unfortunately, in many cases those "natural labels" are associated with a larger object (i.e. a user, an article or a dialogue) rather than one certain sentence, which are not suitable for fully supervised learning. However, by regarding each sentence as an instance and the article/dialogue as

WARNING: This paper contains tweet examples that could be offensive and biased.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7416 - 7425

a bag of instances, we can formalize this scenario into a multiple instance learning task. In this way, we are able to train neural models only using the natural bag-level labels from platforms. As illustrated in Figure 1, the model could not only predict bag-level tags, but also locate offensive sentences to provide more explainable results for moderators.

MIL is a typical weakly-supervised task where each label is associated with a bag of instances, following the rule that a bag is labeled positive if the bag contains at least one positive instance. In offensive language detection task, an offensive sentence will be regarded as a positive instance. Most works of MIL concentrate on the original main target of MIL, which is to predict test bag labels. In offensive language detection tasks, models are supposed to predict not only the bag label but also the sentence labels in order to locate the offensive contents.

As benchmarks are necessary for studying MIL in offensive language detection, we first reconstruct an existing supervised corpus—OLID (Zampieri et al., 2019) into bag-form. Regarding that bags pieced together randomly with independent instances may lack internal relevance, we first cluster the sentences and then sample bags inside each cluster. We also collect a new corpus named Multi-INstance Offensive Response (MINOR) dataset. Bags in MINOR are constructed with tweets and replies, which accord better with the practical application.

In order to study MIL methods for natural language processing tasks systemically, we break down the design of MIL methods into four categories according to where instance-level information is fused into bag level: text fusion, embedding fusion, score fusion and hybrid fusion. We notice that embedding fusion methods perform well in bag-level prediction, while score fusion ones have advantages in instance-level. So we propose a hybrid fusion method with mutual-attention mechanism, which enhances both instance and bag level representation at the same time. Experimental results demonstrate that our mutual-att method outperforms other models at both levels. Ablation studies further illustrate the effectiveness of each component in mutual-att.

To summarize, our contributions are as follows:

- We formalize offensive language detection into a MIL task to utilize coarse grained natural labels from online platforms.
- We present two datasets: **OLID-bags** and **MINOR** to study the multi-instance offensive language detection task.
- After categorizing the existing MIL methods and revisiting their relative merits on our datasets, we propose a new hybrid fusion MIL method—**mutual-att**, which outperforms existing methods at both bag-level and instance-level.

2 Related Work

Offensive Language Detection Offensive language detection has always been a concerned topic for researchers. Great effort has been made to collect corpus from social media (i.e. Twitter) (Waseem and Hovy, 2016) and establish benchmarks (i.e. OLID, TRAC) (Zampieri et al., 2019; Kumar et al., 2018). As offensive language online is a world-wide problem, researchers also constructed many non-English (Pitenis et al., 2020; Mubarak et al., 2021) and multi-language (Kumar et al., 2018) datasets. Semi-supervised dataset SOLID (Rosenthal et al., 2021) has also been proposed to provide large-scale training data without heavy annotation efforts.

Offensive language detection is a typical text classification task. Classic machine learning classifiers including naive Bayes and support vector machine have been widely employed to detect offensive language. Besides universal features such as bag-of-words (McEnery et al., 2000) and n-grams (Pendar, 2007), Chen et al. (2012) also developed task-specific feature extraction method. Neural models like LSTM and CNN have been applied in numerous recent studies, while pre-trained language models like BERT have achieved SOTA performances in numbers of challenges (Liu et al., 2019).

Multiple Instance Learning Multiple instance learning was originally proposed by Dietterich et al. (1997) for drug activity prediction. As the framework (Maron and Ratan, 1998) of MIL could be extended into various scenarios, it attracts attention from communities of many areas. Computer vision (CV) is the major application field of MIL. Numerous studies have applied MIL methods to CV tasks including image classification (Wu et al., 2015), object tracking (Babenko et al., 2009) and medical prediction (Yao et al., 2020; Li et al., 2021). Several natural language processing (NLP) tasks like document modeling (Pappas and Popescu-Belis,

2017) and sentiment analysis (Pappas and Popescu-Belis, 2014; Angelidis and Lapata, 2018; Ji et al., 2020) also meet the definition of MIL.

Various models have been adopted as the base model in MIL tasks. The base model of MIL varies from classic machine learning methods (Gärtner et al., 2002; Andrews et al., 2003) to deep models (Shi et al., 2020; Ilse et al., 2018) over time, and from convolutional networks (Wu et al., 2015; Li et al., 2021) to language models (Angelidis and Lapata, 2018; Ji et al., 2020) over application fields. Besides the choice of base model, we find in this study that MIL fusion methods are essential to MIL frameworks. In Section 3, we discuss when and how instance-level information is fused into bag level in detail.

3 MIL fusion

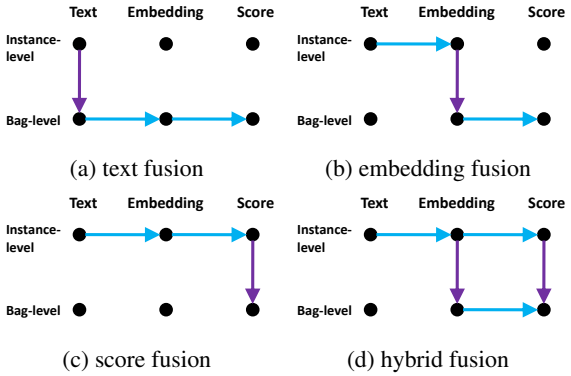
The input of multi-instance offensive language detection task is the text of several instances from a bag, while the supervision is the bag-level label. So the calculation process of a MIL method could be represented as a path from instance-level text to bag-level score, which is shown in Figure 2. According to when instance-level information is fused into bag level, we organize MIL methods for NLP tasks into four categories: text-level fusion, embedding-level fusion, score-level fusion and hybrid fusion. We also discuss different fusion operations in Section 3.2 including pooling methods and attention mechanism. Table 1 is the symbol description of the paper.

Symbol	Explanation
$\boldsymbol{X}/\boldsymbol{x}_i$	bag/instance text
$\boldsymbol{H}/\boldsymbol{h}_i$	bag/instance embedding
P/p_i	bag/instance score
$f()$	a neural network (e.g. BERT)
$g()$	a fusion operation (e.g. attention)
n	instance amount of a bag
σ	sigmoid activation
$\boldsymbol{W}/\boldsymbol{w}$	neural weight
$\boldsymbol{B}/\boldsymbol{b}$	neural bias

Table 1: Symbol description.

3.1 Fusion Level

Text-level Fusion Text fusion is an intuitive method for textual MIL tasks, as shown in Equation 1. First, n instance-level text inputs are fused into bag-level by concatenating them into long text.



(a) text fusion (b) embedding fusion

(c) score fusion (d) hybrid fusion

Figure 2: Categorization of MIL methods according to fusion level. Blue arrows stand for forward calculation of neural network, while purple arrows for fusion operations.

Then a neural model is applied for the long text classification. During the inference phase, instance-level predictions are made by taking a single sentence as the input text.

$$\begin{aligned} \boldsymbol{X} &= [\boldsymbol{x_1}; \boldsymbol{x_2}; ...\boldsymbol{x_n}] \\ \boldsymbol{H} &= f(\boldsymbol{X}) \\ P &= \sigma(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{B}) \end{aligned} \tag{1}$$

Embedding-level Fusion By combining n hidden sentence embeddings into bag-level as shown in Equation 2, embedding fusion methods retain informative representation for the final classification layer, which benefits for bag-level prediction. However, most embedding fusion methods do not have the ability to predict instance labels independently. Only attentional model (Ilse et al., 2018) is explainable in instance-level, but still can not make a direct prediction.

$$\begin{aligned} \boldsymbol{h_i} &= f(\boldsymbol{x_i}) \\ \boldsymbol{H} &= g(\boldsymbol{h_1}, \boldsymbol{h_2}, ...\boldsymbol{h_n}) \\ P &= \sigma(\boldsymbol{W}\boldsymbol{H} + \boldsymbol{B}) \end{aligned} \tag{2}$$

Score-level Fusion As shown in Equation 3, score-level fusion methods first predict n instance labels independently and then calculate the bag-score according to the instance-scores. They have advantages in instance-level prediction, but are weak in bag-level performance because representation information is lost before the bag-level deci-

sion.

$$\begin{aligned} \boldsymbol{h_i} &= f(\boldsymbol{x_i}) \\ p_i &= \sigma(\boldsymbol{w}\boldsymbol{h_i} + \boldsymbol{b}) \\ P &= g(p_1, p_2, ...p_n) \end{aligned} \tag{3}$$

Hybrid Fusion Hybrid fusion is a combination of embedding fusion and score fusion. By fusing embedding and score at the same time, the model could get a rich bag representation while having the ability to predict each instance label. Loss-base attention (Shi et al., 2020) is a typical hybrid attention method, which introduces a loss function over both bag-score and instance-score, as Figure 3 shows. Though the bag-level and instance-level scores are both trained via back propagation, they do not have interaction during the forward process. Thus, we develop our hybrid fusion method with mutual-attention, which allows both levels to enhance each other's prediction.

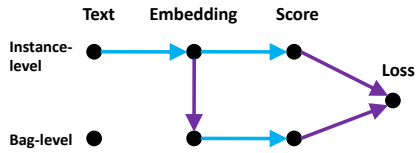


Figure 3: An example of hybrid fusion methods: loss-attention(Shi et al., 2020).

3.2 Fusion Operation

Pooling methods are adopted as fusion operations in many MIL works, of which the two most basic are max-pooling and mean-pooling. Max-pooling conforms well to the bag-labeling rule of MIL task. However, among each bag, only the instance/neurons with max output will be trained, which could cause low training efficiency and poor instance-level performance. Although mean-pooling could provide gradients for all instances, treating every instance equally apparently is not suitable for MIL tasks. Therefore, pooling methods including log-sum-exp pooling (Ramon and De Raedt, 2000) and noisy-or pooling (Zhang et al., 2005) are adopted to provide gradients for all instances while treating every instance differently. In order to develop a flexible and trainable fusion operation, attention mechanisms have been introduced as a MIL fusion operation (Ilse et al., 2018). Recent works (Shi et al., 2020; Li et al., 2021) with attention mechanism have obtained state-of-the-art performance. Besides, neural network such as CNN (Kotzias et al., 2015) and GRU (Karamanolakis et al., 2019) could also be used as a fusion operation in MIL model.

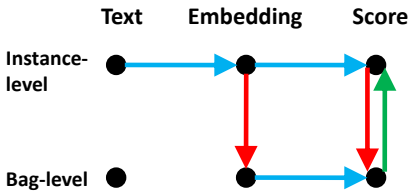


Figure 4: Mutual-attention mechanism. The red arrow stands for I2B-att, while the green one for B2I-att.

4 Method

We propose a mutual-attention mechanism composed of instance to bag attention (I2B-att) and bag to instance attention (B2I-att). As Figure 4 shows, representation and score of instances are both fused into bag-level via I2B-att, while bag score enhances instance scores by B2I-att.

Instance to Bag Attention Instance embedding and score are fused via the same attention—I2B-att. Following (Shi et al., 2020), we directly calculate the instance weight α_i according to the output of instance prediction layer z_i. In this way, instance weight is ensured to be consistent with the prediction probability, that is to say, the instance with a higher possibility to be positive has a larger attention weight. What's more, no extra parameters need to be introduced, so the model can remain high-efficient.

$$\begin{aligned} \boldsymbol{h_i} &= f(\boldsymbol{x_i}) \\ z_i &= \boldsymbol{w}\boldsymbol{h_i} + b \\ p_i &= \sigma(z_i) \\ \alpha_i &= \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}} \\ H &= \sum_{i=1}^{n} \alpha_i \boldsymbol{h_i} \end{aligned} \tag{4}$$

Bag to Instance Attention In order to avoid disagreement between instance and bag predictions, B2I-att is applied to constrain instance prediction with bag score. When the bag label is offensive, an instance label can be offensive or non-offensive, thus low constraint from bag level is supposed to be added to the instance score. By contraries, an instance label should be constrained to non-offensive

if the bag is not offensive. The trainable weight β in Equation 5 controls how much instance predictions are supposed to be influenced by the bag prediction.

$$
\begin{aligned}
Z &= \boldsymbol{W}\boldsymbol{H} + B \\
\beta &= \sigma(W_{B2I}Z + B_{B2I})
\end{aligned}
\tag{5}
$$

Fusions of Mutual-Attention Model Having I2B and B2I weight α and β, we then calculate the final prediction P_{final} following Equation 6, where λ is a hyper-parameter.

$$
\begin{aligned}
P_{bag} &= \sigma(Z) \\
p_i &\leftarrow (1-\beta)p_i + \beta P_{bag} \\
P_{ins} &= \sum_{i=1}^{n} \alpha_i p_i \\
P_{final} &= \lambda P_{bag} + (1-\lambda)P_{ins}
\end{aligned}
\tag{6}
$$

5 Experimental Setup

5.1 Dataset Construction

OLID-bags OLID (Zampieri et al., 2019) is an offensive language dataset containing annotated tweets. In order to study MIL task, we reconstruct it into bag-form. First, we cluster the sentences using Kmeans algorithm and TF-IDF feature. The number of clusters is set 595/67/43 for train/dev/test sets to make each cluster contains 20 sentences on average. Then bags are randomly sampled inside each cluster. Each bags contains 2 to 8 instances and its bag-label follows the definition of MIL. Table 3 shows the statistics of OLID-bags.

Since the bag-label is offensive when any instance in the bag is offensive, the proportion of offensive and non-offensive samples in instance-level and bag-level differs greatly. Offensive instances account for only about one-third, while most bags are offensive. This inconsistency of ratio between instance and bag level makes it more challenging for models to perform well at both levels.

MINOR dataset In order to construct more "real" bags, we collect and annotate the MINOR dataset. Each bag in MINOR is composed of a tweet and several corresponding responses. To obtain a sufficient proportion of offensive language, we get the IDs of tweets and responses from Stance in Replies and Quotes data (SRQ) (Villa-Cox et al., 2020), whose topics are very controversial.

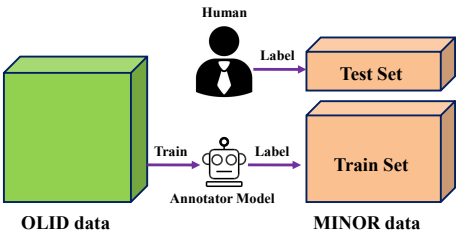


Figure 5: The construction process of MINOR dataset.

The test set of MINOR is manually annotated, which contains 389 bags and 1501 instances. Our definition of offensive and non-offensive text follows OLID(Zampieri et al., 2019). We use OLID sentences as examples in the annotation guideline. Each instance is labeled by 3 annotators with a Cohan's Kappa agreement (Cohen, 1960) of 0.868 and a Spearman correlation (Delgado and Tibau, 2019) of 0.881. As OLID and MINOR are both tweet data, we label the training set of MINOR using a BERT (Devlin et al., 2019) annotator model trained on OLID fully supervised. The accuracy of the annotator model is 84.8% on OLID and 80.8% on MINOR. Such semi-supervised labeling strategy could save manual efforts while providing a larger data scale.

Comparison Clearly, OLID-bags has higher label quality because it is manual-labeled while MINOR has larger data scale. As we can see in Table 2, due to the clustering step, instances in an OLID bag often share similar vocabularies or topics. Instances in an MINOR bag have more strong and direct connections as they are post and responses. Witness rate (WR) is a concept of MIL, which stands for the proportion of positive instances in positive bags. The WR of OLID-bags is 40.5% while MINOR's is 53.7% because MINOR has a higher inner-bag similarity. Previous studies(Carbonneau et al., 2018) have shown that tasks with lower WR are more challenging, so MINOR could be easier for a model even if it's automatically labeled.

5.2 Settings and Baselines

Experimental Settings In order to compare MIL methods fairly, we set the base model $f()$ to be BERT-base for all experiments. The learning rate of all methods is set to 1e-5 and the batch size is 8. The threshold for prediction is determined by searching on the dev set to get the best macro F1-score. λ value is set to 0.8 for mutual attention and

Text	Label
An example bag from OLID-bags	1
I just threw up in my mouth. #LockHerUp #MAGA URL	1
Where is my MAGA cap?! URL	0
So now someone has to photoshop a pic of Toad in a MAGA cap with a confederate flag on his kart.	0
An example bag from MINOR	1
Post: S**n H****ty and T*k** C***s*n Are getting the first interviews with Tr**p after P*t*n summit in other words No hard questions for Tr**p	0
Responses: In other words, Tr**p's lapdogs, H****ty, C***s*n, will sit obediently at P*t*n's feet, wagging their little tails	1
Another sit-down full of lies and self-aggrandizing. We'll hear LIE after LIE after LIE	1
He will and would lie anyway. He is meeting P*t*n privately for personal reasons and it is not about the United States.	1

Table 2: Example bags from OLID-bags and MINOR. We use * to mask names of public figures and some sensitive words.

Set	Level	Offensive	Non-offensive
Train	Instance	6805 (33.2)%	13691 (66.8)%
	Bag	3156 (76.0)%	996 (24.0)%
Dev	Instance	738 (32.2)%	1554 (67.8)%
	Bag	362 (78.4)%	100 (21.6)%
Test	Instance	380 (26.6)%	1051 (73.4)%
	Bag	202 (67.1)%	99 (32.9)%

Table 3: Data split of OLID-bags.

Set	Level	Offensive	Non-offensive
Train	Instance	14241(31.2%)	31442(68.8%)
	Bag	6516 (47.0%)	7339 (53.0%)
Dev	Instance	979 (30.1%)	2275 (69.9%)
	Bag	470 (47.0%)	530 (53.0%)
Test	Instance	680 (46.0%)	821 (54.0%)
	Bag	287 (73.8%)	102 (26.2%)

Table 4: Data split of MINOR.

loss-based attention(Shi et al., 2020). The detailed analysis of λ is carried out in Section 6.2.

Baselines In this paper, we will represent baselines using our categorization rather than the original name because the task and the base model are different, and it is more clear to show features of the method in this way. Specifically, each MIL method will be named by its fusion level + fusion operation. Corresponding relations between baselines and their names are shown in Table 5.

Baselines	Our Categorization
MI-Net(Wang et al., 2018)	emb+pooling
mi-net(Wang et al., 2018)	score+pooling
Gated-Attention(Ilse et al., 2018)	emb+att
Loss-Attention(Shi et al., 2020)	hybrid+loss-att

Table 5: Name from the original paper and our categorization of baselines.

We also show the result of instance-level supervised learning, but note that it is for reference only because label-levels are different.

5.3 Evaluation Metrics

Bag-level Prediction We evaluate bag-level performance by macro-F1 score and accuracy. For methods that are not able to make bag-level prediction (i.e. supervised), we inference the bag-level label from instance-level prediction according to the rule of MIL.

Instance-level Prediction Instance-level performance is also measured with macro-F1 score and accuracy. As mentioned in Section 3.1, some embedding fusion methods can not predict instance labels directly. So in this paper, we evaluate instance-level performance of embedding fusion methods by letting them predict the labels of single-instance bags. For emb+att model, instance-level predictions $p_i = \sigma(z_i)$ are made according to the logits value z_i before the soft-max calculation of attention mechanism.

6 Experiments

6.1 Main Results

Experimental results of our method and baselines on OLID-bags and MINOR are shown in Tabel. 6.

Supervised learning can be regarded as the ceiling of instance-level performance, as it is trained with full instance labels. Surprisingly, most MIL methods have comparable or even better bag-level performance on OLID-bags than fully supervised learning. Also, some MIL methods have close instance-level performance to supervised learning on both datasets, especially our mutual-att method. These results indicate that multiple instance learning is a feasible and promising solution to make use of natural bag labels in online offensive language detection.

Model	Bag-F1	Bag-Acc	Ins-F1	Ins-Acc
text fusion	0.746±0.013	0.778±0.021	0.776±0.008	0.829±0.005
emb+max	0.756±0.08	0.785±0.009	0.773±0.010	0.827±0.020
emb+mean	0.723±0.011	0.782±0.009	0.762±0.019	0.796±0.026
emb+att	0.768±0.011	0.809±0.009	0.772±0.010	0.830±0.009
score+max	0.755±0.09	0.784±0.013	0.775±0.013	0.808±0.009
score+mean	0.666±0.017	0.750±0.031	0.643±0.085	0.655±0.092
score+att	0.758±0.011	0.810±0.014	0.784±0.008	0.832±0.010
hybrid+loss-att	0.773±0.013	0.810±0.005	0.789±0.012	0.831±0.011
hybrid+mutual-att	**0.779***±0.009	**0.812**±0.008	**0.792**±0.009	**0.839***±0.011
supervised	0.770±0.006	0.803±0.004	0.801±0.012	0.852±0.006

(a) OLID-bags

Model	Bag-F1	Bag-Acc	Ins-F1	Ins-Acc
text fusion	0.912±0.013	0.921±0.011	0.817±0.012	0.826±0.008
emb+max	0.923±0.014	0.930±0.009	0.817±0.011	0.825±0.009
emb+mean	0.917±0.011	0.925±0.009	0.808±0.010	0.815±0.012
emb+att	0.928±0.010	0.938±0.006	0.817±0.008	0.825±0.007
score+max	0.917±0.008	0.927±0.009	0.820±0.012	0.825±0.012
score+mean	0.899±0.017	0.915±0.015	0.806±0.013	0.811±0.011
score+att	0.924±0.012	0.937±0.011	0.826±0.010	0.828±0.011
hybrid+loss-att	0.934±0.006	0.947±0.005	0.824±0.006	0.827±0.006
hybrid+mutual-att	**0.941***±0.005	**0.952***±0.005	**0.832***±0.008	**0.836***±0.007
supervised	0.944±0.006	0.954±0.006	0.845±0.004	0.849±0.004

(b) MINOR

Table 6: Main results on OLID-bags and MINOR. All experiments are conducted over 5 runs, and we report the averaged results along with standard deviations. The bold numbers stand for the best performance except supervised learning. The bold results with * are significantly better than the second highest ones ($\alpha = 0.05$). Underlined ones represent the 2nd and 3rd highest results.

Among the four fusion level categories, text fusion is the simplest method which has an average but reliable performance. As we mentioned in Section 3, embedding fusion methods are good at bag-level prediction, while score fusion methods have better instance-level performance. We can find in the table that emb+att and score+att get remarkable results at bag and instance level respectively which are second only to hybrid fusion methods. By combining their advantages, the two hybrid fusion methods achieve high performance at both levels. In particular, our mutual-att method outperforms loss-att at both levels and is more stable at instance level. As for fusion operations, we find that max-pooling and attention mechanism perform well, while mean-pooling has poor and unstable performance.

All models achieve much higher performance on MINOR than OLID-bags which implies that MINOR is less challenging even if it's annotated by model. There are two main reasons. One is that MINOR has more sufficient data, and the other is that MINOR has a higher witness rate as we mentioned in Section 5.1. We will make a detailed discussion about why could model get high results on such semi-supervised dataset in Section 6.3.

6.2 Ablation Study and Parameter Analysis

Ablation Study In order to investigate the effectiveness of each component of our mutual-attention method, we conduct an ablation study whose results are shown in Table 7. Note that in experiment "w/o I2B-att" we only remove score level I2B-att and only during the testing process.

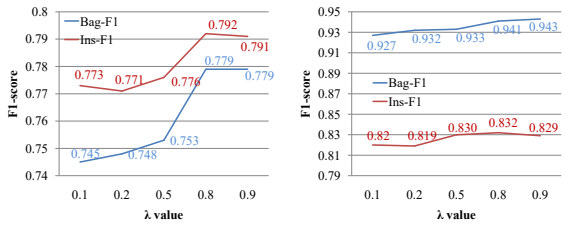
Model	Bag-F1	Ins-F1	Disagree
hybrid+loss-att	0.773	0.789	7.9%
hybrid+mutual-att	**0.779**	**0.792**	**3.0%**
w/o I2B-att	0.769	**0.792**	4.3%
w/o B2I-att	0.774	0.784	5.6%

Table 7: Results of ablation study on OLID-bags. Disagree stands for the proportion of bags whose instance predictions conflict with bag prediction.

Results imply that I2B-att mainly enhances bag predictions, while B2I-att mainly improves instance-level performance. We notice that such improvements may be made by eliminating disagreements between bag prediction and instance predictions. We find that existing MIL methods suffer from the disagreement problem. About 8% bags in loss-att model's prediction have conflicts between two prediction levels. For example, the model may predict the whole paragraph is non-offensive while predict one sentence in it is offen-

sive, which is illogical and will confuse the decision maker in practical use. Our B2I and I2B attention reduce those disagreements by letting the two predictions influence each other. When they are both applied, the disagreement rate is reduced to 3%.

Parameter Analysis We carry out experiments to investigate the influence of λ in Equation 6. Curves in Figure 6 show how the performance of our model changes when λ varies. Note that simply setting $\lambda = 0, 1$ will make the model unreasonable, so these results are not included in 6. The results of removing components of our model are discussed in the ablation study part.



(a) OLID-bags (b) MINOR

Figure 6: Parameter analysis for λ.

We find that F1 performance on OLID-bags dramatically drops when λ becomes lower than 0.5 especially at bag-level. The model is less sensitive to λ value on MINOR. From these results, 0.8 to 0.9 could be the proper range for λ. Empirically, we set $\lambda = 0.8$ in our main experiments.

6.3 Semi-Supervising

Since the training set of MINOR is labeled by a model trained on OLID, we doubt if a model directly trained on OLID-bags would perform better. We carry out an experiment to prove such semi-supervised strategy is effective and necessary. Results in Table 8 demonstrate that training on the auto-labeling data is far more effective than direct transferring from OLID data. Although the supervision is indirectly from OLID, the unique distribution of real tweet and response bags of MINOR is also necessary for the bag modeling. Besides, MINOR has a larger amount of data than OLID, which may provide the model with more diverse expressions.

What's more, since bag label is not necessarily associated with every instance label, MIL task may suffer less from label noise. For example, if one instance in a bag is labeled correctly as offensive,

Train set	Bag-F1	Bag-Acc	Ins-F1	Ins-Acc
OLID-bags	0.843	0.867	0.797	0.802
MINOR	0.941	0.952	0.832	0.836

Table 8: Comparing transferring and semi-supervised learning. The model is our hybrid+mutual-att method.

the bag label will be correct regardless of other instances. As Table 9 shows, the annotator model fails in 19.2% MINOR instances, while the bag-level error rate is only 12.1%. "Natural labels" such as moderator punishments and user reports also do not require manual efforts and may also contain noise. High performance on MINOR indicates that it is possible to utilize those free resources with MIL methods even if their label quality is limited.

Level	F1	Acc
Instance	0.799	0.808
Bag	0.860	0.879

Table 9: Performance of the annotator model on MINOR test set.

7 Conclusion and Discussion

In this paper, we propose that MIL can utilize the historical information from social media platforms as natural labels for offensive language detection. We present two multi-instance datasets for offensive language detection: OLID-bags and MINOR. OLID-bags is reconstructed from OLID while bags of MINOR are composed with real tweets and responses.

We systematically categorize MIL methods for textual tasks into four classes namely text fusion, embedding fusion, score fusion and hybrid fusion. We observe on OLID-bags and MINOR that embedding fusion has higher bag-level performance, while score fusion methods are good at instance-level prediction. Hybrid fusion methods could integrate the advantages of them, among which our proposed mutual-attention achieves state-of-the-art performance on both datasets at both levels. We also carry out a detailed ablation study to investigate the effectiveness of the proposed I2B-att and B2I-att and how they address the prediction disagreement problem. Discussion in Section 6.3 also verifies that our semi-supervised strategy is efficient and effective.

The proposed datasets, OLID-bags and MINOR, could support future studies on multi-instance offensive language detection. Also, the presented

MIL formalization could expand to other online risky content identification tasks. We hope our work could inspire researchers from social media platforms and be instantiated in real scenarios.

Ethical Statement

Data used in this study are all from public released datasets. We strictly follow the ethical implications in previous research related to the data source. The source of our data has a nature of anonymity in a certain extent. We further clean the private information by filtering out nicknames, phone numbers, and URL links in a rule-based setting. The annotators in this study are all co-authors and are only shown with anonymized tweets when annotating. Moreover, because of the natural bias in terms of political stance, race, gender, etc. when defining and annotating offensive languages, we urge the user to cautiously examine the ethical implications of offensive language detection models in real-world applications.

References

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, pages 577–584.

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. 2009. Visual tracking with online multiple instance learning. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 983–990. IEEE.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. 2002. Multi-instance kernels. In *ICML*, volume 2, page 7.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR.

Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. 2020. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7012–7023.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Weakly supervised attention networks for fine-grained opinion mining and public health. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 1–10.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 597–606.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Bin Li, Yin Li, and Kevin W Eliceiri. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328.

Ping Liu, Wen Li, and Liang Zou. 2019. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

Oded Maron and Aparna Lakshmi Ratan. 1998. Multiple-instance learning for natural scene classification. In *ICML*, volume 98, pages 341–349. Citeseer.

Tony McEnery, Paul Baker, and Andrew Hardie. 2000. Swearing and abuse in modern british english.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 455–466.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Nick Pendar. 2007. Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241. IEEE.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Jan Ramon and Luc De Raedt. 2000. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928.

Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. 2020. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5742–5749.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3460–3469.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Cha Zhang, John Platt, and Paul Viola. 2005. Multiple instance boosting for object detection. *Advances in neural information processing systems*, 18:1417–1424.

A Appendix