# Oasis: Online Analytic System for Incivility Detection and Sentiment Classification

Leyu Liu[*], Xin Huang[*], Jianliang Xu[*], Yunya Song[#]

[*]*Department of Computer Science, Hong Kong Baptist University*
[#]*Department of Journalism, Hong Kong Baptist University*
621leyu@gmail.com, {xinhuang, xujl}@comp.hkbu.edu.hk, yunyasong@hkbu.edu.hk

***Abstract*—Incivility detection is an important task to identify offensive language on online social media platforms. Sentiment analysis is an essential task of natural language processing to identify the emotions of given sentences. In this demonstration, we propose an online processing system, called Oasis, to perform incivility detection and sentiment prediction for short texts (e.g., tweets). Oasis offers several useful features including the CNN-LSTM classification models, incivility detection and sentiment analysis for real-time tweets and user-interested queries, and personalized search of trending topics in Twitter.**

## I. Introduction

Online social media platforms (e.g., Twitter, Facebook, Instagram, and Yelp) are popular for users to express opinions and share their daily life [1], [2]. Lots of such online posts are presented in the form of short texts, which have obvious indications of emotions and attitudes. Online short texts affect on individuals, business, culture, and even politics of society. For instance, individuals can make their own judgments towards emerging events; companies can gather product reviews to formulate the right marketing strategies and improve customer satisfaction. Meanwhile, online incivility frequently happens on social media in various kinds of ways, e.g., offensive language, profanity, cyberbullying, flaming, trolling, and even the more subtle sarcasm [3], [4], [5]. Curbing incivility while still allowing for free flow information in online communities, not only concerns the interest of individual citizens, but also matters to corporations and governments. To achieve this goal, an essential task is incivility detection, which may contribute to the development of cost-effective long-term solutions to online uncivil behavior. In addition, sentiment analysis [1], [6] is also useful to extract features for detecting offensive content, e.g., specific emotions of anger and sadness were used to identify victims.

In this demonstration, we present an **O**nline **A**nalytic **S**ystem for **I**ncivility detection and **S**entiment classification (Oasis). We modify deep learning models for incivility detection and sentiment classification in Oasis. Many existing profanity processing systems utilize a pre-defined list of hostile words or patterns, and apply a simple match-and-replace method [7]. However, a simple match-and-replace method would likely rule out inoffensive words. Thus, one significant challenge is how to develop an effective prediction model to classify incivility. Moreover, there exists a huge amount of online user-generated contents to be processed efficiently. To address these challenges, Oasis develops efficient classification models for real-time incivility and sentiment analytics, which are based on deep learning techniques of Convolutional Neural Network (CNN) [8], [9] and Long Short-Term Memory (LSTM) [10], [11]. Our CNN-LSTM classification models are trained by three datasets of tweets with ground-truth offensive and sentiment labels. Our classification models achieve appreciable results via cross-validation testing.

Oasis exhibits three useful functions as follows. (1) Real-time incivility detection and sentiment analysis for emerging tweets are presented with intuitive emotional results; (2) Oasis provides a user-friendly interface for users to input a query sentence for online sentiment and incivility analysis. (3) It offers users with the personalized search, exploration, and visualization of tweets on trending topics.

## II. System Overview

Fig. 1 depicts the system architecture of Oasis. Oasis is an online analytic system for *incivility detection* and *sentiment classification*, which offers user-friendly visual interfaces to analyze, monitor, and interact with real-time tweets and input queries. It is formed by one key component of deep learning models for incivility detection and sentiment classification, and two modules of input and output.

The *classification models* are constructed on deep learning techniques. It consists of three classifiers: one incivility classifier and two sentiment classifiers. All classifiers are built upon CNN and LSTM. We train our models on two kinds of training datasets, which are tweets with ground-truth sentiments and tweets with offensive/unoffensive labels respectively. We improve the models through optimizations.

The *input module* accepts two kinds of input data: real-time tweets and online queries raised by users. Real-time tweets are crawled from our following accounts on Twitter via APIs. Moreover, it allows users to issue their own interested queries and interact with our system. Oasis can take these data as the input of incivility and sentiment classification models.

The *output module* mainly consists of two display components: (1) real-time monitoring wall, and (2) user-interactive wall. Real-time monitoring wall can display the online analysis results of real-time collected tweets. It consists of several interesting functions. First, the word cloud reflects the important keywords of the up-to-date popular Twitter topics. Second, sentiment classifier can predict each tweet into one of five

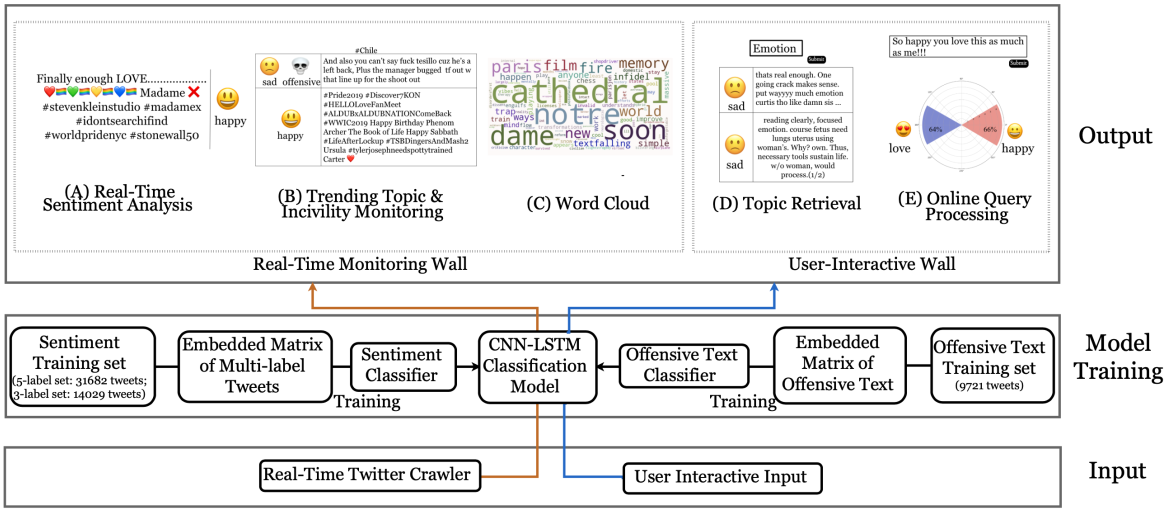

Finally enough LOVE.................. ❤️🏳️‍🌈💚🏳️‍🌈💛🏳️‍🌈💙🏳️‍🌈 Madame ❌ #stevenkleinstudio #madamex #idontsearchifind #worldpridenyc #stonewall50
happy
(A) Real-Time Sentiment Analysis
#Chile
sad
offensive
And also you can't say fuck tesillo cuz he's a left back, Plus the manager bugged tf out w that line up for the shoot out
happy
#Pride2019 #Discover7KON #HELLOLoveFanMeet #ALDUBxALDUBNATIONComeBack #WWIC2019 Happy Birthday Phenom Archer The Book of Life Happy Sabbath #LifeAfterLockup #TSBDingersAndMash2 Ursula #tylerjosephneedspottytrained Carter ❤️
(B) Trending Topic & Incivility Monitoring
paris film fire memory happen anyone infidel cathedral notre world dame new soon trap train ways mind textfalling simple
(C) Word Cloud
Real-Time Monitoring Wall
Emotion
Submit
sad
thats real enough. One going crack makes sense. put wayyyy much emotion curtis tho like damn sis ...
sad
reading clearly, focused emotion. course fetus need lungs uterus using woman's. Why? own. Thus, necessary tools sustain life. w/o woman, would process.(1/2)
(D) Topic Retrieval
So happy you love this as much as me!!!
Submit
64%
66%
love
happy
(E) Online Query Processing
User-Interactive Wall
Output
Sentiment Training set (5-label set: 31682 tweets; 3-label set: 14029 tweets)
Embedded Matrix of Multi-label Tweets
Sentiment Classifier
Training
CNN-LSTM Classification Model
Offensive Text Classifier
Training
Embedded Matrix of Offensive Text
Offensive Text Training set (9721 tweets)
Model Training
Real-Time Twitter Crawler
User Interactive Input
Input

Fig. 1. System Architecture of Oasis

classes {"happy", "neutral", "sad", "hate", "anger"}. Third, our online incivility model detects whether a tweet contains offensive language or not. It also allows users to overview many recent tweets related to trending topics in Twitter. Furthermore, the user-interactive wall shows the sentiment results of input queries and user interested tweets.

## III. Models

In this section, we introduce the deep learning methods for incivility detection and sentiment classification in Oasis.[1]

### A. CNN-LSTM based Classification Models

Oasis adopts deep neural network models for incivility detection and sentiment classification. It makes use of bidirectional LSTM [11] and CNN [9] techniques. Oasis puts an emphasis on the order of words in the analyzed texts. As a useful technique of NLP, bidirectional LSTM preserves feature information of texts from past (backward) and future (forward) states simultaneously [11]. Meanwhile, CNN takes the embedded texts as input and detects patterns of the expressions using the convolution layers and pooling layers to match the features with the corresponding classes [9].

The architecture of our training model is presented as follows. It applies *data preprocessing* on training texts for word-to-vector transformation. Then, the embedded sequences are processed to be the same input for four different modules simultaneously. Specifically, the first module uses two submodules of LSTM-CNN structure, each of which is consisted of a LSTM layer, a convolutional layer, and a concatenate layer; the second module has two submodules of CNN-LSTM structures; the third module adopts the structure of LSTM with two LSTM layers; the fourth module adopts the structure of CNN with two convolutional layers. The products of these four modules are concatenated via the max pooling layer, dropout, flatten layer, and dense layers. Finally, our CNN-LSTM classification model produces the output.

### B. Data preprocessing

Given an input tweet, the first step of data processing removes all unnecessary stop words and symbols, such as hyperlinks, @, and hashtags. The labels are transformed into binary vectors by one-hot encoding. Next, Keras [12] framework is adopted for text preprocessing and model construction. A Keras embedding layer provides the word embedding where words are represented with dense vectors with weights for deep learning model to learn. We use Keras tokenizer to perform text tokenization, and transform texts into word indexes. The size of word indexes is the input dimension of a Keras embedding layer. As Keras embedding layer requires each text with the same length, each tweet is reshaped into padding sequences with the length of 40 through post-sequence padding. To learn word embedding with suitable weights, we use pretrained Stanford GloVe vectors of Twitter with 200 dimensions to build a corresponding embedding matrix [13]. Lastly, the embedding matrix is used to initialize the weights of the embedding layer in our model.

### C. Optimization

During the construction process of training models, we adjust some parameters of deep learning networks to achieve good results for multi-label classification. The loss function of categorical loss entropy [14] is used for predicting more than 2 class labels. After several rounds of training and testing, our model achieves good performance by adopting 24 output filters, the convolution window with the size of 2, and a dropout rate of 0.2 to prevent overfitting in the dropout layer. Meanwhile, Oasis applies Adadelta optimizer using a default learning rate of 1.0, which produces encouraging and robust results of gradient descent.

### D. Sentiment Classification

We apply the CNN-LSTM model framework to build two multi-label sentiment classifiers. The first one is a 3-label sentiment classifier. We use a training dataset of 20,939 tweets, which contains 3 sentiment labels of "love", "worry", and "neutral" [6]. We apply 3-fold cross-validation to evaluate our

[1] https://github.com/621Alice/Oasis

| label | precision | recall | $F_1$-score | support |
|---|---|---|---|---|
| neutral | 0.484 | 0.439 | 0.460 | 3157 |
| happy | 0.669 | 0.716 | 0.692 | 5394 |
| sad | 0.634 | 0.679 | 0.656 | 5225 |
| hate | 0.904 | 0.677 | 0.775 | 1453 |
| anger | 0.94 | 0.714 | 0.813 | 377 |
| macro avg / total | 0.727 | 0.645 | 0.679 | 15606 |

TABLE I
Accuracy Performance of CNN-LSTM model on the 5-Label Sentiment Classification

| label | precision | recall | $F_1$-score | support |
|---|---|---|---|---|
| worry | 0.655 | 0.718 | 0.685 | 2765 |
| love | 0.679 | 0.627 | 0.652 | 1256 |
| neutral | 0.667 | 0.627 | 0.646 | 2889 |
| macro avg / total | 0.667 | 0.658 | 0.661 | 6910 |

TABLE II
Accuracy Performance of CNN-LSTM model on the 3-Label Sentiment Classification

| label | precision | recall | $F_1$-score | support |
|---|---|---|---|---|
| non-offensive text | 0.996 | 0.978 | 0.987 | 2421 |
| offensive text | 0.978 | 0.996 | 0.987 | 2367 |
| macro avg / total | 0.987 | 0.987 | 0.987 | 4788 |

TABLE III
Accuracy Performance of CNN-LSTM model on Offensive Text Classification

training model and report $F_1$-scores of accuracy performance by scikit-learn [15]. Our model achieves good classification results with a macro average $F_1$-score of 0.66 shown in Table II. In addition, we build another multi-label sentiment classifier using a 5-label sentiment dataset. The training dataset contains 47,288 tweets with 5 sentiment labels. These labels include “happy”, “neutral”, “sad”, “hate”, and “anger”. Table I shows the performance of our training model with a macro average $F_1$-score of 0.64.

### E. Incivility Detection

To identify offensive texts as incivility, we apply the CNN-LSTM model framework to construct an offensive text classifier. We use a training dataset of 14,709 tweets [4] to build our prediction model. We partition all tweets into two classes: offensive and non-offensive. Offensive tweets are texts containing offensive language and hate speech, and the others are non-offensive. Table III reports that our model gets a macro average $F_1$-score of 0.98.

## IV. Related Systems

Oasis is related to *sentiment analysis* and *incivility detection*. Sentiment analysis classifies target texts into different sentiment classes. Many state-of-the-art models have been proposed to use machine learning algorithms and deep learning approaches for different sentiment analysis tasks [1]. Fan et al. [16] proposed a convolutional memory network with an attention mechanism to analyze complicated multi-words expressions in sentences. With useful sentiment signals in textual terms on social media, unsupervised sentiment analysis model [17] was proposed to analyze signed social networks. Severyn et al. [18] proposed a unsupervised neural language model that tuned the parameter weights of CNN to improve sentiment analysis. Wang et al. [19] proposed a demonstration system to analyze sentiment of real-time tweets about the 2012 US president election cycle. In addition, online incivility has emerged to be a big issue on online social networks. Various studies have been proposed to study online incivility. A bidirectional LSTM network was proposed for offensive language detection, categorization and target identification [20]. Moreover, to solve the problem of unrestricted use of offensive language, Santos et al. [5] developed a encoder-decoder to transform offensive language into non-offensive ones. In contrast to above studies, Oasis develops deep learning models based on CNN and LSTM to perform incivility detection and sentiment analysis for real-time tweets and user-interested queries.

## V. Demonstration Overview

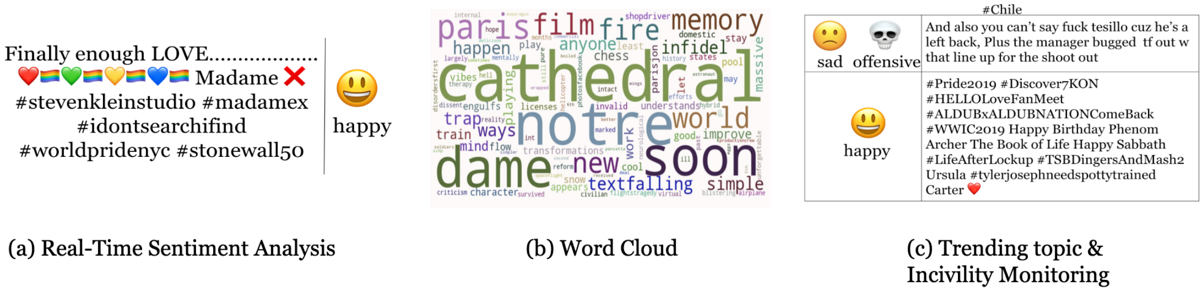
We implement Oasis system using Django, Word Cloud [21], and Tweepy [22]. We train the CNN-LSTM based classification models using Python, Keras, and scikit-learn. We develop websites to display the results of online incivility and sentiment analysis. The output module mainly consists of two parts: real-time monitoring wall and user-interactive wall, as shown in Fig. 2 and Fig. 3 respectively. The monitoring wall presents the results of sentiment and incivility classification for real-time tweets. In order to obtain real-time tweets from emerging events in Twitter, we create a Twitter account to follow multiple famous accounts of news, travel, and culture, such as “@abcWNN”, “@TravelLeisure”, and “@Complex”. Since a lot of news is likely to be negative or neutral, our account also follows popular accounts for funny tweets (e.g., “@JokesMemesFacts”), which can better demonstrate our sentiment model that is not limited to negative and neutral emotions. In addition, the user-interactive wall provides an interactive interface for audiences to input queries based on their own interest and explore results in a user-friendly manner.

### A. Real-time Tweets for Incivility and Sentiment Analysis

Fig. 2(a) displays the sentiment result of a real-time tweet as “happy”. Oasis collects ten up-to-date tweets every minute from our following twitter accounts. These obtained tweets are taken as inputs for sentiment classification and incivility detection. For each tweet, we apply the 5-label sentiment classifier to predict the sentiment label from “happy”, “neutral”, “sad”, “hate”, and “anger”. The results are kept in the local file and then displayed one tweet at a time on the real-time monitoring wall.

### B. Trending Topic Analysis

The module of trending topic analysis uses the Tweepy trends_place function and selects one recent hot topic. Oasis displays a list of English tweets up to 10 entries, and analyzes them using sentiment and incivility models. Fig. 2(c) shows the sentiment and incivility classification results of tweets w.r.t. a trending topic. Tweets are associated with different sentiment categories of happy and sad. If there is any offensive language involved in the tweet, one warning emoji will be highlighted to notify users. Oasis provides an intuitive interface for better visualization using emojis.



Finally enough LOVE.................. ❤️🏳️‍🌈💚🏳️‍🌈💛🏳️‍🌈💙🏳️‍🌈 Madame ❌ #stevenkleinstudio #madamex #idontsearchifind #worldpridenyc #stonewall50

#Chile

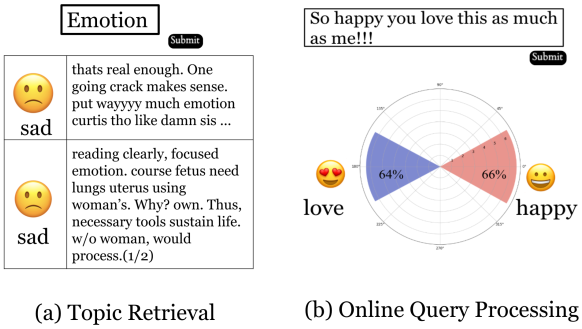
| | |
|---|---|
| 🙁 sad 💀 offensive | And also you can't say fuck tesillo cuz he's a left back, Plus the manager bugged tf out w that line up for the shoot out |
| 😃 happy | #Pride2019 #Discover7KON #HELLOLoveFanMeet #ALDUBxALDUBNATIONComeBack #WWIC2019 Happy Birthday Phenom Archer The Book of Life Happy Sabbath #LifeAfterLockup #TSBDingersAndMash2 Ursula #tylerjosephneedspottytrained Carter ❤️ |

(a) Real-Time Sentiment Analysis

(b) Word Cloud

(c) Trending topic & Incivility Monitoring

Fig. 2. Real-Time Tweet Monitoring



Emotion

Submit

| | |
|---|---|
| 🙁 sad | thats real enough. One going crack makes sense. put wayyyy much emotion curtis tho like damn sis ... |
| 🙁 sad | reading clearly, focused emotion. course fetus need lungs uterus using woman's. Why? own. Thus, necessary tools sustain life. w/o woman, would process.(1/2) |

So happy you love this as much as me!!!

Submit

(a) Topic Retrieval

(b) Online Query Processing

Fig. 3. User-Interactive Querying for Incivility and Sentiment

### C. Word Cloud

The module of word cloud shows the most frequent keywords of our collected real-time tweets using different sized fonts in Fig. 2(b). To acquire high-quality keywords, we apply data clean to remove the punctuation, hyperlinks and English stop words. Fig. 2(b) shows an emerging event of "Notre-Dame de Paris fire" happened on July 15, 2019. The larger font of one keyword is, the more occurrence of this keyword in real-time tweets is. This display component gives an effective way to learn about up-to-date popular topics on Twitter.

### D. Personalized Topic Search and Exploration

Oasis offers the feature of topic filtering in the user-interactive wall. Audiences can search for a topic that they are interested in. Then, Oasis uses the Tweepy search function to find a list of tweets related to the input topic. The discovered tweets will be presented with their sentiment labels as shown in Fig. 3(a). This function provides an interactive interface for users to explore the results of their interested tweets.

### E. Online Querying Sentiment and Incivility

Oasis provides an interesting interactive feature for online querying sentiment and incivility. Our playground web page demonstrates the interfaces of query input and sentiment classification report in Fig. 3(b). With the user input field, audiences can type in texts for sentiment and incivility classifications. For each input sentence, two models of 3-label and 5-label sentiment classifiers are applied to the tested tweet simultaneously. Oasis aims at providing comprehensive sentiment labels for the testing text. It also gives users a complete report of sentiment classifications with the corresponding precision scores of predictions. Fig. 3(b) shows two prediction results of "love" and "happy" for the input text.

## ACKNOWLEDGMENT

This work is supported by the NSFC No. 61702435, RGC Nos. 12200917, 12244916, 12632816, C6030-18GF, and HKBU No. RC-IG-FNRA/17-18/04.

## REFERENCES

[1] Z. Lei, W. Shuai, and L. Bing, "Deep learning for sentiment analysis: A survey," *ICIICT*, 2018.

[2] X. Huang, L. V. Lakshmanan, and J. Xu, *Community Search over Big Graphs*. Morgan & Claypool Publishers, 2019.

[3] P. B. Osullivan and A. J. Flanagin, "Reconceptualizing flamingand other problematic messages," *New Media & Society*, vol. 5, no. 1, pp. 69–94, 2003.

[4] D. Thomas, W. Dana, M. Michael, and W. Ingmar, "Automated hate speech detection and the problem of offensive language," *ICWSM*, 2017.

[5] N. d. S. Cicero, M. Igor, and P. Inkit, "Fighting offensive language on social media with unsupervised text style transfer," *ACL*, 2018.

[6] L. Timothy, H. K. Tong, and Y. K. Chia. Multi-class emotion classification for short text. [Online]. Available: https://github.com/tlkh/text-emotion-classification

[7] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," in *Canadian Conference on Artificial Intelligence*, 2010, pp. 16–27.

[8] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP*, 2014.

[9] B. Denny. Understanding convolutional neural networks for nlp. [Online]. Available: http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for

[10] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[11] O. Christopher. Understanding lstm networks. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs

[12] Keras documentation. [Online]. Available: https://keras.io/

[13] P. Jeffrey, S. Richard, and D. M. Christopher. Glove: Global vectors for word representation. [Online]. Available: https://nlp.stanford.edu/projects/glove/

[14] Loss functions. [Online]. Available: https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html

[15] Documentation of scikit-learn 0.21.2. [Online]. Available: https://scikit-learn.org/stable/documentation.html

[16] F. Chuang, G. Qinghong, D. Jiachen, G. Lin, X. Ruifeng, and W. Kam-Fai, "Convolution-based memory network for aspect-based sentiment analysis," *SIGIR*, 2018.

[17] C. Kewei, L. Jundong, T. Jiliang, and L. Huan, "Unsupervised sentiment analysis with signed social networks," *AAAI*, 2017.

[18] S. Aliaksei and M. Alessandro, "Twitter sentiment analysis with deep convolutional neural networks," *SIGIR*, 2015.

[19] W. Hao, C. Dogan, K. Abe, B. Franois, and N. Shrikanth, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," *ACL*, 2012.

[20] S. M. A. Lutfiye, B. S. lex, and S. Horacio, "Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bi-lstm model," *ACL*, 2019.

[21] Word cloud api. [Online]. Available: https://wordcloudapi.com/

[22] Tweepy documentation. [Online]. Available: http://docs.tweepy.org/en/latest/