EdinburghNLP at WNUT-2020 Task 2: Leveraging Transformers with Generalized Augmentation for Identifying Informativeness in COVID-19 Tweets

Nickil Maveli

ILCC, School of Informatics, University of Edinburgh

Correspondence: n.maveli@sms.ed.ac.uk

Highlights

- **Identifying relevant information in tweets is challenging due to the low signal-to-noise ratio.**
- **We formulated this task as a binary text classification problem with "INFORMATIVE" and "UNINFORMATIVE" as the class names.**
- **We build an ensemble of Transformer models to leverage its strength in capturing contextual information.**
- **The data used to train these models is an augmented version carefully prepared to alleviate confirmation bias and thereby improve generalization.**

Methods

- Transformer models (RoBERTa, XLNet) and BERTweet ensemble learning.
- Generalized augmentation via pseudo-labeling - Data is carefully augmented with the help of pseudo labeling which is the process of adding confident predicted test data to the training data.
- Optimal thresholding via post-processing to adjust distribution of class labels in target - The idea here is to make the distribution of labels in dev/test set to match corresponding distribution of labels in train set to maintain the class ratio.

Abstract

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programmatically monitoring Twitter (disaster relief organizations and news agencies) and therefore recognizing the informativeness of a tweet can help filter noise from large volumes of data. In this paper, we present our submission for *WNUT-2020 Task 2: Identification of informative COVID-19 English Tweets*. Our most successful model is an ensemble of transformers including RoBERTa, XLNet, and BERTweet trained in a Semi-Supervised Learning (SSL) setting. The proposed system achieves a F1 score of 0.9011 on the test set (ranking 7th on the leaderboard) and shows significant gains in performance compared to a baseline system using fasttext embeddings.

Background

- The basic goal of WNUT-2020 Task 2 [1] is to automatically identify whether a COVID-19 English Tweet is Informative or not.
- Such Informative Tweet provide information about recovered, suspected, confirmed and death cases as well as location or travel history of the cases. About 4M COVID-19 English Tweets are daily being posted on twitter, the majority of which being not informative.

A hard to classify tweet - This example hints that in order to reach meaningful results, we may have to examine contextual linguistic features, model the annotator's bias, introduce adversarial examples.

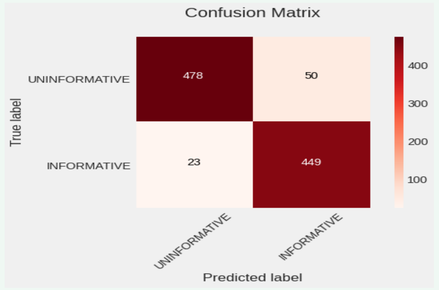
Model

- Texts are lowercased. Non-ascii letters, urls, @RT:[NAME], @[NAME] are removed. Break apart common single tokens; Eg: RoBERTa makes a single token for "...", so convert all single [...] tokens into three [.][.][.] tokens. Similarly, split "!!!". All Transformer models use this preprocessing strategy.
- Texts are normalized using TweetTokenizer. Some of the normalization steps are - Expand text contractions("can't" to "cannot", "M" to "million", etc.), text normalization ("p . m ." to "p.m.", etc.). All BERTweet models use this preprocessing strategy.
- We use only the dataset provided by the organizers to perform our experiments. Overall, there are a total of 10K Tweets split in the ratio of 70/10/20 into train/dev/test set respectively. However, for the final evaluation, 12K unlabeled noisy Tweets were provided, out of which 2K test Tweets were the actual ones the models were evaluated upon.
- To speed up training, sequence bucketing by removing unnecessary padding was employed [2]. To improve the robustness of neural networks, and improving resistance to adversarial attacks, Fast Gradient Method (FGM) was used [3] at the end of Transformer models.
- Multi-Sample Dropout [4] was used when using dropout before the last layer with p = 0.5, seemed to converge loss faster. Output of each dropout layer was then passed to a shared weight fc layer. Next, we took the average of the outputs from fc layer as the final output.

Results

Model	Without Augmentation			With Augmentation		
	Precision	Recall	F1	Precision	Recall	F1
RoBERTa$_{BASE_1}$	0.8652	0.9386	0.9004	0.9619	0.8833	0.9209
RoBERTa$_{BASE_2}$	0.8760	0.9280	0.9012	**0.9640**	0.8818	0.9211
XLNet$_{BASE_1}$	0.8583	0.9364	0.8956	0.9619	0.8798	0.9190
XLNet$_{BASE_2}$	0.8580	0.9343	0.8945	0.9619	0.8731	0.9153
BERTweet$_{BASE_1}$	0.8630	0.9343	0.8973	0.9534	0.8858	0.9184
BERTweet$_{BASE_2}$	0.8483	**0.9597**	0.9006	0.9449	0.8974	0.9206
Ensemble	0.8790	0.9386	0.9078	0.9513	0.8998	**0.9248**

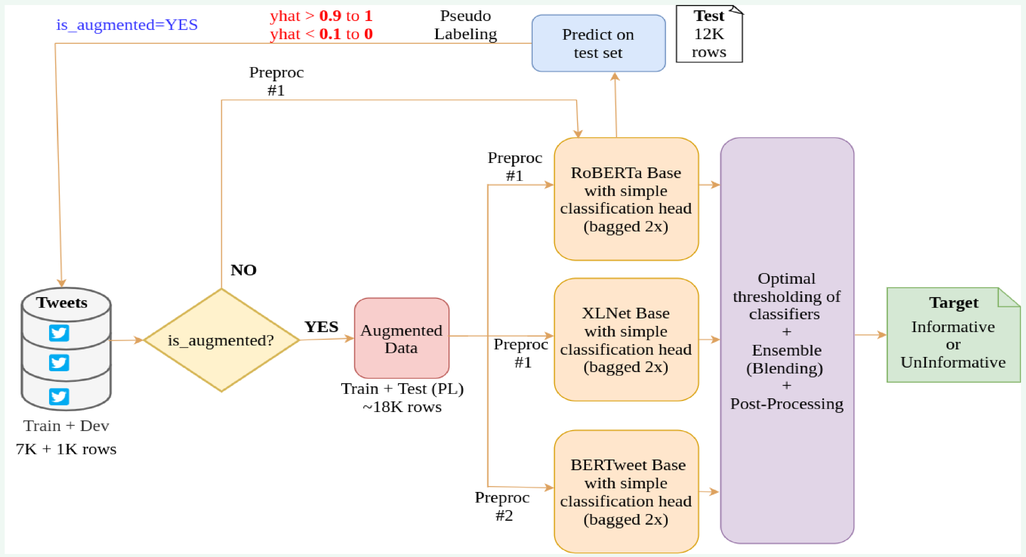
Results on Dev set



Model	P	R	F1
Baseline FASTTEXT	0.7730	0.7288	0.7503
RoBERTa-XLNet-BERTweet-Ensemble	0.8768	0.9269	0.9011

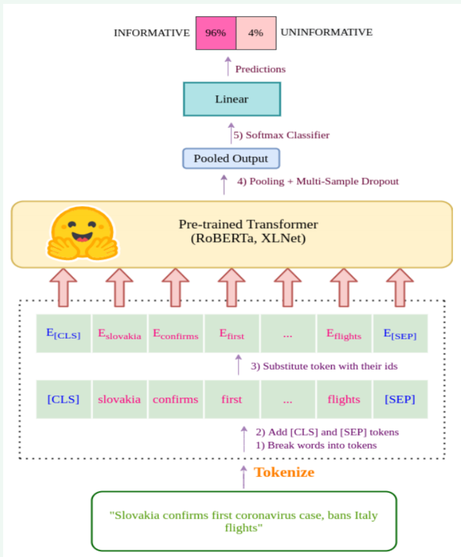
Results on Test set - Our model improves the organizer's baseline by 20%.

Model Architecture



A RoBERTa model does the classification on the 12K test-set, while being trained using 7K train-set. Later, 11K most confident predictions are appended to the train-set. The new concatenated data is fed to the ensemble models leading to better generalization and improved model performance.

Pre-trained Transformer model architecture for informativeness classification



Pretrained RoBERTa-base and XLNet-base-cased models with a single linear layer which is simply a feed-forward network that acts as a classification head were used.

Error Analysis

Tweet - I just remember this news recently China keeping two sets of coronavirus pandemic numbers? "Leaked" infection numbers over 154,000; deaths approach 25,000
Justification - Misinformation due to ambiguity and subjectivity. It could be well evident that some events may not really happen as the source of the news lacked credibility. This could have prompted inter-annotator disagreement.

Tweet - Writing 101: don't put 2 numbers side by side. The punctuation is easy to miss. I first read this as being 51,385 people have died in Ontario from Covid.
Justification - Inaccurate interpretation of contexts. Much of the attention weights are focused on the latter part. Our model may not capture this shift correctly given the long-distance dependency, which results in a false positive prediction.

Conclusion

- Combine user-related tweet features (followers, friends, favorite counts, etc.) and tweet-related meta features (retweets, creation date, sentiment, etc.) along with contextual representation.
- Extending to multilingual tweets is a potential future direction to pursue.

References

[1] Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In Proceedings of the 6th Workshop on Noisy User-generated Text.
[2] Viacheslav Khomenko, Oleg Shyshkov, Olga Radyvonenko, and Kostiantyn Bokhan. 2017. Accelerating recurrent neural network training using sequence bucketing and multi-gpu data parallelization. CoRR, abs/1708.05604.
[3] Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semisupervised text classification. In ICLR (Poster). OpenReview.net.
[4] Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. CoRR, abs/1905.09788.