WNUT-2020 Task 2:
Identification of Informative COVID-19 English Tweets

Dat Quoc Nguyen[1,*], **Thanh Vu**[2,*], **Afshin Rahimi**[3], **Mai Hoang Dao**[1],
Linh The Nguyen[1] and **Long Doan**[1]

[1]VinAI Research, Vietnam; [2]Oracle Digital Assistant, Oracle, Australia;
[3]The University of Queensland, Australia

v.datnq9@vinai.io; thanh.v.vu@oracle.com; a.rahimi@uq.edu.au
{v.maidh3, v.linhnt140, v.longdct}@vinai.io

Abstract

In this paper, we provide an overview of the WNUT-2020 shared task on identification of informative COVID-19 English Tweets. We describe how we construct a corpus of 10K Tweets and organize development and evaluation stages for this task. In addition, we also present a brief summary of results obtained for final system evaluation submissions from 55 teams, finding that (i) many systems obtain very high performance, up to 0.91 F_1 score, (ii) the majority of the submissions achieve substantially higher results than the baseline fastText (Joulin et al., 2017), and (iii) fine-tuning pre-trained language models on relevant language data followed by supervised training performs very well in this task.

1 Introduction

As of late-September 2020, the COVID-19 Coronavirus pandemic has led to about 1M deaths and 33M infected patients from 213 countries and territories, creating fear and panic for people all around the world.[1] Recently, much attention has been paid to building monitoring systems (e.g. The Johns Hopkins Coronavirus Dashboard) to track the development of the pandemic and to provide users the information related to the virus,[2] e.g. any new suspicious/confirmed cases near/in the users' regions.

It is worth noting that most of the "official" sources used in the tracking tools are not kept up to date with the current pandemic situation, e.g. WHO updates the pandemic information only once a day. Those monitoring systems thus use social network data, e.g. from Twitter, as a real-time

alternative source for updating the pandemic information, generally by crowd-sourcing or searching for related information manually. However, the pandemic has been spreading rapidly; we observe a massive amount of data on social networks, e.g. about 3.5M of COVID-19 English Tweets daily on the Twitter platform (Lamsal, 2020) in which the majority are uninformative. Thus, it is important to be able to select the informative Tweets (e.g. COVID-19 Tweets related to new cases or suspicious cases) for downstream applications. However, manual approaches to identify the informative Tweets require significant human efforts, do not scale with rapid developments, and are costly.

To help handle the problem, *we propose a shared task which is to automatically identify whether a COVID-19 English Tweet is informative or not*. Our task is defined as a binary classification problem: Given an English Tweet related to COVID-19, decide whether it should be classified as INFORMATIVE or UNINFORMATIVE. Here, informative Tweets provide information about suspected, confirmed, recovered and death cases as well as location or travel history of the cases. The following example presents an informative Tweet:

> **INFORMATIVE**
>
> Update: Uganda Health Minister Jane Ruth Aceng has confirmed the first #coronavirus case in Uganda. The patient is a 36-year-old Ugandan male who arrived from Dubai today aboard Ethiopian Airlines. Patient travelled to Dubai 4 days ago. #CoronavirusPandemic

The goals of our shared task are: (i) To develop a language processing task that potentially impacts research and downstream applications, and (ii) To provide the research community with a new

---

[1]https://www.worldometers.info/coronavirus/
[2]https://coronavirus.jhu.edu/map.html

dataset for identifying informative COVID-19 English Tweets. To achieve the goals, we manually construct a dataset of 10K COVID-19 English Tweets with INFORMATIVE and UNINFORMATIVE labels. We believe that the dataset and systems developed for our task will be beneficial for the development of COVID-19 monitoring systems. All practical information, data download links and the final evaluation results can be found at the CodaLab website of our shared task: https://competitions.codalab.org/competitions/25845.

2 The WNUT-2020 Task 2 dataset

2.1 Annotation guideline

We define the guideline to annotate a COVID-19 related Tweet with the "INFORMATIVE" label if the Tweet mentions suspected cases, confirmed cases, recovered cases, deaths, number of tests performed as well as location or travel history associated with the confirmed/suspected cases.

In addition, we also set further requirements in which the "INFORMATIVE" Tweet has to satisfy. In particular, the "INFORMATIVE" Tweet should not present a rumor or prediction. Furthermore, quantities mentioned in the Tweet have to be specific (e.g. "*two new cases*" or "*about 125 tested positives*") or could be inferred directly (e.g. "*120 coronavirus tests done so far, 40% tested positive*"), but not purely in percentages or rates (e.g. "*20%*", "*1000 per million*", or "*a third*").

The COVID-19 related Tweets not satisfying the "INFORMATIVE" annotation guideline are annotated with the "UNINFORMATIVE" label. An example of uninformative Tweet is as follows:

> **UNINFORMATIVE**
>
> Indonesia frees 18,000 inmates, as it records highest #coronavirus death toll in Asia behind China HTTPURL

2.2 COVID-19 related Tweet collection

To be able to construct a dataset used for our shared task, we first have to crawl the COVID-19 related Tweets. We collect a general Tweet corpus related to the COVID-19 pandemic based on a predefined list of 11 keywords, including: "coronavirus", "covid-19", "covid_19", "covid_2019", "covid19", "covid2019", "covid-2019", "CoronaVirusUpdate", "Coronavid19", "covid" and

"SARS-CoV-2". In particular, we utilize the Twitter streaming API to download real-time English Tweets containing at least one keyword from the predefined list.[3]

We stream the Tweet data for four months using the API from 01st March 2020 to 30th June 2020. We then filter out Tweets containing less than 10 words (including hashtags and user mentions) as well as Tweets from users with less than five hundred followers. This is to help reduce the rate of Tweets with fake news (our manual annotation process does not involve verifying fake news) with a rather strong assumption that reliable information is more likely to be propagated by users with a large following.[4] To handle the duplication problem: (i) we remove Retweets starting with the "RT" token, and (ii) in cases where two Tweets are the same after lowercasing and removing hashtags and user mentions, the original earlier Tweet is kept and the subsequent Tweet will be filtered out. Applying these filtering steps results in a final corpus of about 23M COVID-19 English Tweets.

2.3 Annotation process

From the corpus of 23M Tweets, we select Tweets which are potentially informative, containing predefined strings relevant to the annotation guideline such as "confirm", "positive", "suspected", "death", "discharge", "test" and "travel history". We then remove similar Tweets with the token-based cosine similarity score (Wang et al., 2011) that is equal or greater than 0.7, resulting in a dataset of "INFORMATIVE" candidates. We then randomly sample 2K Tweets from this dataset for the first phase of annotation.

Three annotators are employed to independently assign each of the 2K Tweets with one of the two labels "INFORMATIVE" and "UNINFORMATIVE". We use the "docanno" toolkit for handling annotations (Nakayama et al., 2018). We measure the inter-annotator agreement to assess the quality of annotations and to see whether the guideline allows to carry out the task consistently. In particular, we use the Fleiss' Kappa coefficient for assessing the annotator agreement (Fleiss, 1971). For this first phase, the Kappa score

---

[3] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

[4] We acknowledge that there are accounts with a strong following who participate in publication and propagation of misinformation.

Item	Training	Validation	Test	Total
#INFOR	3,303	472	944	4,719
#UNINF	3,697	528	1,056	5,281
Total	7,000	1,000	2,000	10,000

Table 1: Basic statistics of our dataset. #INFOR and #UNINF denote the numbers of "INFORMATIVE" and "UNINFORMATIVE" Tweets, respectively.

is 0.797 which can be interpreted as substantial (Landis and Koch, 1977). We further run a discussion for Tweets where there is a disagreement in the assigned labels among the annotators. The discussion is to determine the final labels of the Tweets as well as to improve the quality of the annotation guideline.

For the second phase, we employ the 2K annotated Tweets from the first phase to train a binary fastText classifier (Joulin et al., 2017) to classify a COVID-19 related Tweet into either "INFORMATIVE" or "UNINFORMATIVE". We utilize the trained classifier to predict the probability of "INFORMATIVE" for each of all remaining Tweets in the dataset of "INFORMATIVE" candidates from the first phase. Then we randomly sample 8K Tweets from the candidate dataset, including 3K, 2K and 3K Tweets associated with the probability $\in [0.0, 0.3)$, $[0.3, 0.7)$ and $[0.7, 1.0]$, respectively (here, we do not sample from the existing 2K annotated Tweets). The goal here is to select Tweets with varying degree of detection difficulty (with respect to the baseline) in both labels.

The three annotators then independently assign the "INFORMATIVE" or "UNINFORMATIVE" label for each of the 8K Tweets. The Kappa score is obtained at 0.818 which can be interpreted as almost perfect (Landis and Koch, 1977). Similar to the first phase, for each Tweet with a disagreement among the annotators, we also run a further discussion to decide its final label annotation.

We merge the two datasets from the first and second phases to formulate a final gold standard corpus of 10K annotated Tweets, consisting of 4,719 "INFORMATIVE" Tweets and 5,281 "UN-INFORMATIVE" Tweets.

2.4 Data partitions

To split the gold standard corpus into training, validation and test sets, we first categorize its Tweets into two categories of "easy" and "not-easy", in which the "not-easy" category contains Tweets with a label disagreement among annotators before participating in the annotation discussions.

We then randomly select 7K Tweets for training, 1K Tweets for validation and 2K Tweets for test with a constraint that ensures the number of the "not-easy" Tweets in the training, validation and test sets are equal. Table 1 describes the basic statistics of our corpus.

3 Task organization

Development stage: Both the training and validation sets with gold labels are released publicly to all participants for system development. Although we provide a default training and validation split of the released data, participants are free to use this data in any way they find useful when training and tuning their systems, e.g. using a different split or performing cross-validation.

Evaluation stage: The raw test set is released when the final stage of system evaluation starts. To keep fairness among participants, the raw test set is a relatively large set of 12K Tweets, and the actual 2K test Tweets by which the participants' system outputs are evaluated are hidden in this large test set. Our CodaLab website allows each participant to upload at most 2 submissions during this final evaluation stage, in which the submission obtaining higher F_1 would appear in the leaderboard.

Metrics: Systems are evaluated using standard evaluation metrics, including Accuracy, Precision, Recall and F_1 score. Note that the latter three metrics of Precision, Recall and F_1 will be calculated for the "INFORMATIVE" label only. The system evaluation submissions are ranked by the F_1 score.

Baseline: fastText (Joulin et al., 2017) is used as our baseline, employing the default data split.

4 Results

In total, 121 teams spreading across 20 different countries registered to participate in our WNUT-2020 Task 2 during the system development stage. Of those 121 teams, 55 teams uploaded their submissions for the final evaluation stage.[5]

We report results obtained for each team in Table 2. The baseline fastText achieves 0.7503 in F_1 score. In particular, 48 teams outperform the baseline in terms of F_1. There are 39 teams with

---

[5] CXP949 is not shown on our CodaLab leaderboard because this team unfortunately makes an incorrectly-formatted submission file name, resulting in a fail for our CodaLab automatic evaluation program. We manually re-evaluate their submission and include its obtained results in Table 2.

Team	F_1	P	R	Acc.	Team	F_1	P	R	Acc.
NutCracker	**0.9096**	0.9135	0.9057	**0.9150**	CUBoulder-UBC	0.8841	0.8606	0.9089	0.8875
NLP_North	**0.9096**	0.9029	0.9163	0.9140	Sic Mundus	0.8823	0.8832	0.8814	0.8890
UIT-HSE	0.9094	0.9046	0.9142	0.9140	LynyrdSkynyrd	0.8805	0.8567	0.9057	0.8840
#GCDH	0.9091	0.8919	0.9269	0.9125	Dartmouth CS	0.8757	0.8818	0.8697	0.8835
Loner	0.9085	0.8918	0.9258	0.9120	L3STeam	0.8754	0.8654	0.8856	0.8810
Phonemer	0.9037	0.8934	0.9142	0.9080	XSellResearch	0.8739	0.8857	0.8623	0.8825
EdinburghNLP	0.9011	0.8768	0.9269	0.9040	Linguist Geeks	0.8715	0.9130	0.8337	0.8840
TATL	0.9008	0.8588	**0.9470**	0.9015	DSC-IITISM	0.8715	0.8343	0.9121	0.8730
SunBear	0.9005	0.8728	0.9301	0.9030	AmazingAI	0.8714	0.8637	0.8792	0.8775
InfoMiner	0.9004	0.9102	0.8909	0.9070	Siva	0.8527	0.8115	0.8983	0.8535
NEU	0.8992	0.8959	0.9025	0.9045	CSECU-DSG	0.8198	0.8155	0.8242	0.8290
Not-NUTs	0.8991	0.8787	0.9206	0.9025	IIITBH	0.7979	0.7991	0.7966	0.8095
UET	0.8989	0.8891	0.9089	0.9035	NLPRL	0.7854	0.8335	0.7426	0.8085
Emory	0.8974	0.8744	0.9216	0.9005	Kai	0.7772	0.7540	0.8019	0.7830
NJU ConvAI	0.8973	0.8751	0.9206	0.9005	IBS	0.7765	0.7692	0.7839	0.7870
IDSOU	0.8964	0.8988	0.8941	0.9025	MrRobot	0.7648	0.7515	0.7786	0.7740
ComplexDataLab	0.8945	**0.9195**	0.8708	0.9030	ISWARA	0.7631	0.8073	0.7235	0.7880
UPennHLP	0.8941	0.9028	0.8856	0.9010	TheWalkingBy	0.7614	0.7709	0.7521	0.7775
DATAMAFIA	0.8940	0.8857	0.9025	0.8990	KZhu	0.7580	0.7788	0.7383	0.7775
NIT_COVID-19	0.8914	0.8594	0.9258	0.8935	IRLab@IITBHU	0.7508	0.7904	0.7150	0.7760
CXP949	0.8910	0.8698	0.9131	0.8945	Baseline–fastText	0.7503	0.7730	0.7288	0.7710
NHK_STRL	0.8898	0.8985	0.8814	0.8970	Amrita_CEN_NLP	0.7496	0.8078	0.6992	0.7795
COVCOR20	0.8887	0.8655	0.9131	0.8920	intelligentCyborgs	0.7417	0.6507	0.8623	0.7165
CIA_NITT	0.8887	0.8772	0.9004	0.8935	BhagwanBharose	0.7269	0.7723	0.6864	0.7565
honeybee	0.8884	0.8956	0.8814	0.8955	IITKGPPHD	0.7132	0.7535	0.6769	0.7430
BANANA	0.8881	0.8853	0.8909	0.8940	NITK_NLP	0.6826	0.7581	0.6208	0.7275
SU-NLP	0.8881	0.8895	0.8867	0.8945	36H102	0.5800	0.5015	0.6875	0.5300
VT	0.8846	0.8723	0.8972	0.8895	TMU-COVID19	0.5789	0.5000	0.6875	0.5280

Table 2: Final results on the test set. **P**, **R** and **Acc.** denote the Precision, Recall and Accuracy, respectively. Teams are ranked by their F_1 score.

an F_1 greater than 0.80, in which 10 teams are with an F_1 greater than 0.90. Both NutCracker (Kumar and Singh, 2020) and NLP_North (Møller et al., 2020) obtain the highest F_1 score at 0.9096, in which NutCracker obtains the highest Accuracy at 91.50% that is 0.1% absolute higher than NLP_North's.

Of the 55 teams, 36 teams submitted their system paper, in which 34 teams' papers are finally included in the Proceedings. All of the 36 teams with paper submissions employ pre-trained language models to extract latent features for learning classifiers. The majority of pre-trained language models employed include BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), BERTweet (Nguyen et al., 2020) and

especially CT-BERT (Müller et al., 2020).

Not surprisingly, CT-BERT, resulted in by continuing pre-training from the pre-trained BERT-large model on a corpus of 22.5M COVID-19 related Tweets, is utilized in a large number of the highly-ranked systems. In particular, all of top 6 teams including NutCracker, NLP_North, UIT-HSE (Tran et al., 2020), #GCDH (Varachkina et al., 2020), Loner and Phonemer (Wadhawan, 2020) utilize CT-BERT. That is why we find slight differences in their obtained F_1 scores. In addition, ensemble techniques are also used in a large proportion (61%) of the participating teams. Specifically, to obtain the best performance, top 10 teams, except NLP_North, #GCDH and Loner, all employ the ensemble techniques.

5 Conclusion

In this paper, we have presented an overview of the WNUT-2020 Task 2 "Identification of Informative COVID-19 English Tweets": (i) Details of the task, data preparation process, and the task organization, and (ii) Report the results obtained by participating teams and outline their commonly adopted approaches.

We receive registrations from 121 teams and final system evaluation submissions from 55 teams, in which 34/55 teams contribute detailed system descriptions. The evaluation results show that many systems obtain a very high performance of up to 0.91 F_1 score on the task, using pre-trained language models which are fine-tuned on unlabelled COVID-19 related Tweets (CT-BERT) and are subsequently trained on this task.

References

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Priyanshu Kumar and Aadarsh Singh. 2020. NutCracker at WNUT-2020 Task 2: Robustly Identifying Informative COVID-19 Tweets using Ensembling and Adversarial Training . In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Rabindra Lamsal. 2020. CORONAVIRUS (COVID-19) TWEETS DATASET. *IEEE Dataport*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.

Anders Giovanni Møller, Rob van der Goot, and Barbara Plank. 2020. NLP North at WNUT-2020 Task 2: Pre-training versus Ensembling for Detection of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human. Software available from https://github.com/doccano/doccano.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Khiem Tran, Hao Phan, Kiet Nguyen, and Ngan Luu Thuy Nguyen. 2020. UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Hanna Varachkina, Stefan Ziehe, Tillmann Dońicke, and Franziska Pannach. 2020. #GCDH at WNUT-2020 Task 2: BERT-Based Models for the Detection of Informativeness in English COVID-19 Related Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Anshul Wadhawan. 2020. Phonemer at WNUT-2020 Task 2: Sequence Classification Using COVID Twitter BERT and Bagging Ensemble Technique based on Plurality Voting. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

J. Wang, G. Li, and J. Fe. 2011. Fast-join: An efficient method for fuzzy token matching based string similarity join. In *Proceedings of the 27th IEEE International Conference on Data Engineering*, pages 458–469.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763.