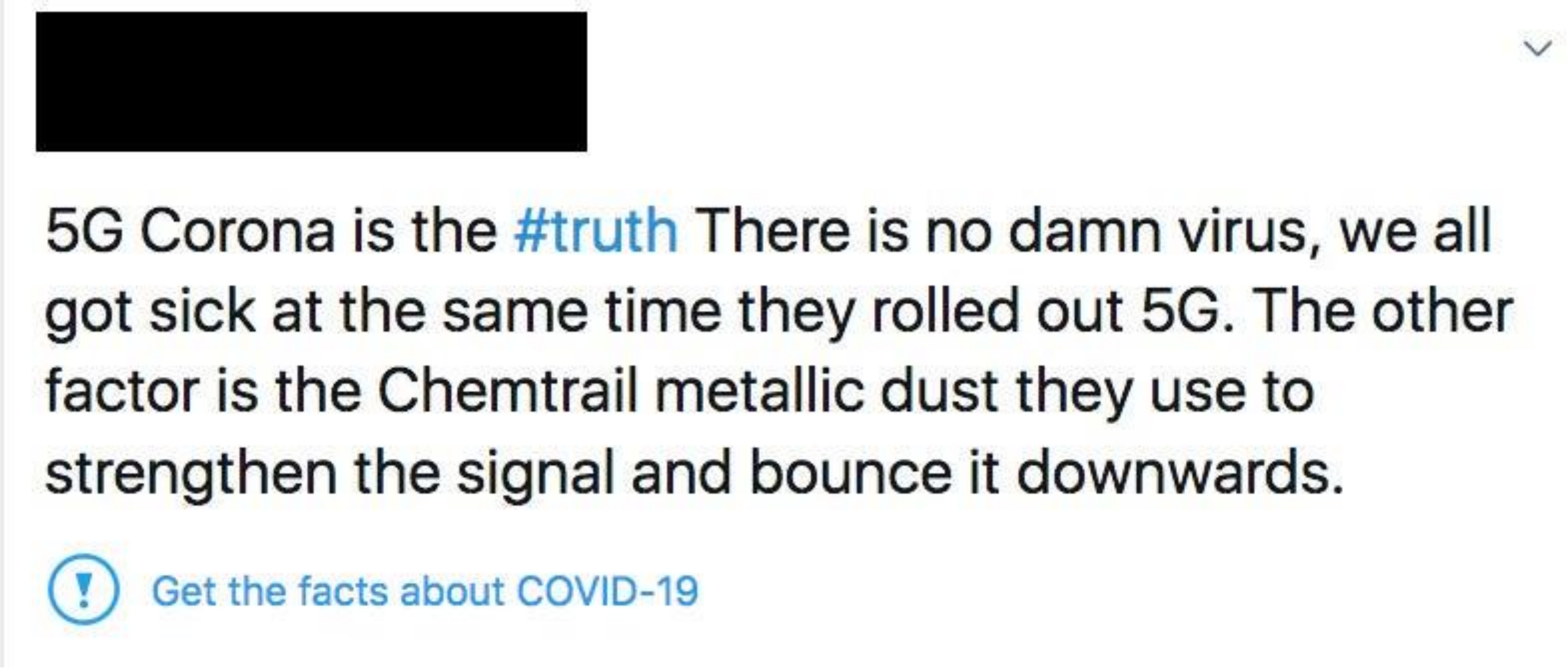


[redacted]

5G Corona is the #truth There is no damn virus, we all got sick at the same time they rolled out 5G. The other factor is the Chemtrail metallic dust they use to strengthen the signal and bounce it downwards.

# DL-TXST FakeNews: Enhancing Tweet Content Classification with Adapted Language Models

**Task:** FakeNews: Corona Virus and Conspiracies Multimedia Analysis (MediaEval 2021)

**Authors:** Muhieddine Shebaro, Jason Oliver, Tomiwa Olarewaju, Jelena Tešić

Computer Science, Texas State University, San Marcos TX, USA

{m.shebaro, jasonoliver, tro24, jtesic}@txstate.edu

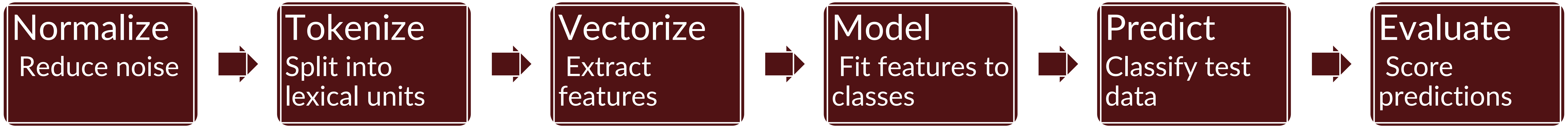
TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

# Our Message

Contributions incorporate:

- □ New Normalizations
- □ Encoding & Decoding Dependent Variables
- □ Data Fusion
- □ Logistic Regression, which performed best in previous participation, on new integrated dataset
- □ Ensemble Learning
- □ Use of BERTweet pretrained model instead of last year’s utilization of BERT-uncased model by Google



**Normalize** Reduce noise

**Tokenize** Split into lexical units

**Vectorize** Extract features

**Model** Fit features to classes

**Predict** Classify test data

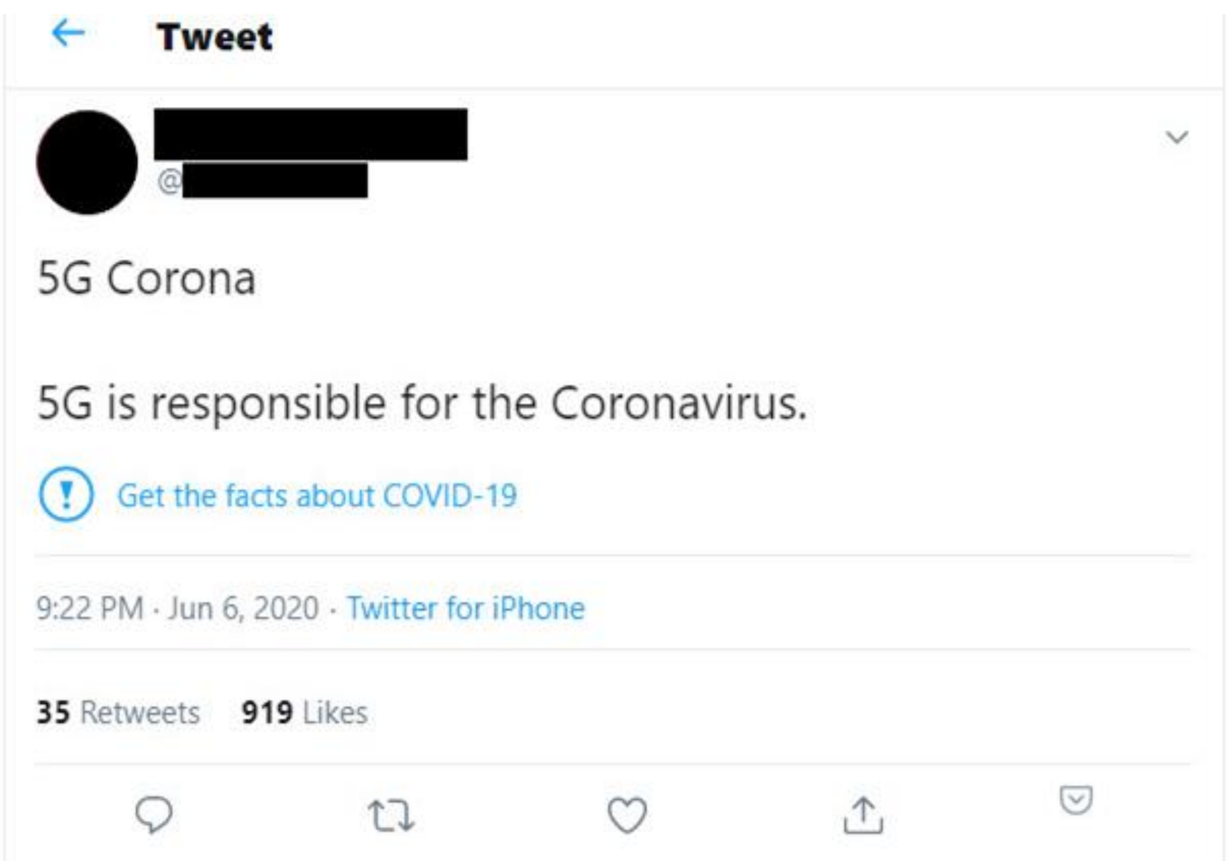
**Evaluate** Score predictions

# Our Challenge

Some of the challenging aspects are:

- ❑ We observed that there are some discrepancies between these two old and new development sets which would impede the flow of the integration process.
- ❑ When we looked at the development & testing set for Subtask 2 and 3, the tweet was divided into several parts and each part was present in a separate column.
- ❑ Since our data in subtask 2 & 3 contains multiple target variables, this is beyond the models' inference capabilities of more than one dependent variable at once.
- ❑ We wanted to experiment with Ensemble Learning on this task, but it was a taxing job to carefully select the models to combine.
- ❑ We observed that the more the number of class labels (encodings) we have , the worse the ability of our models to generalize.

MediaEval
Multimedia
Benchmark



Tweet

[redacted]
@[redacted]

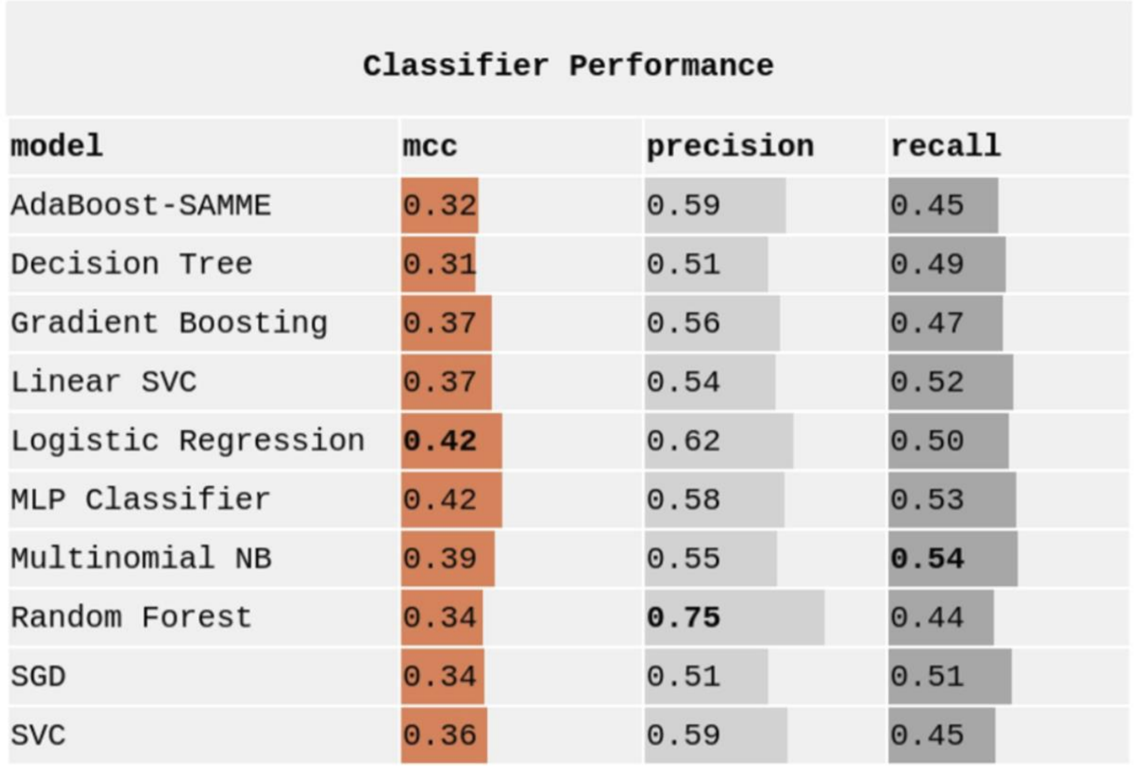
5G Corona

5G is responsible for the Coronavirus.

Get the facts about COVID-19

9:22 PM · Jun 6, 2020 · Twitter for iPhone

35 Retweets 919 Likes



Classifier Performance

| model | mcc | precision | recall |
|---|---|---|---|
| AdaBoost-SAMME | 0.32 | 0.59 | 0.45 |
| Decision Tree | 0.31 | 0.51 | 0.49 |
| Gradient Boosting | 0.37 | 0.56 | 0.47 |
| Linear SVC | 0.37 | 0.54 | 0.52 |
| Logistic Regression | **0.42** | 0.62 | 0.50 |
| MLP Classifier | 0.42 | 0.58 | 0.53 |
| Multinomial NB | 0.39 | 0.55 | **0.54** |
| Random Forest | 0.34 | **0.75** | 0.44 |
| SGD | 0.34 | 0.51 | 0.51 |
| SVC | 0.36 | 0.59 | 0.45 |

**Figure 5. Final results for fine-grained classification**

# Approach

## New Normalizations

Previous +

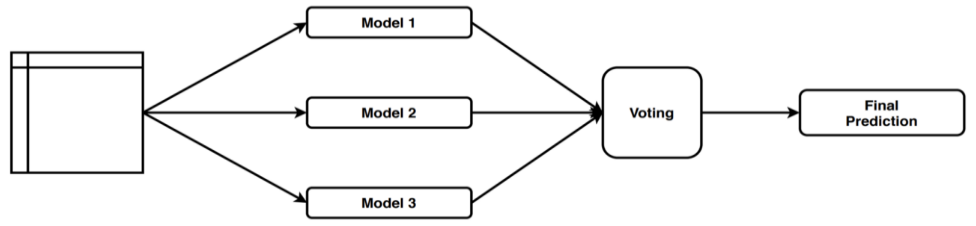
- ❑ Removing usernames
- ❑ Remove all special characters
- ❑ Removing hashtags
- ❑ Remove contractions (e.g., convert “won’t” to “will” and “not”)
- ❑ Removing non-English tweets if present
- ❑ Removing links which not only incorporates “https://t.co/”, but also “http” and “www”
- ❑ Removing Emoticons

## Logistic Regression

- ❑ We encode every occurrence of a combination of binary target variables into a single target variable and vice versa.
- ❑ Threshold applied is 20.
- ❑ Before fusion, the number of tuples of the old dataset was 5,946 rows. After integration, we got a total of 6,769 tuples.
- ❑ We chose it as a baseline model since it performed the best in the control experiment (new hyperparameter tuning).

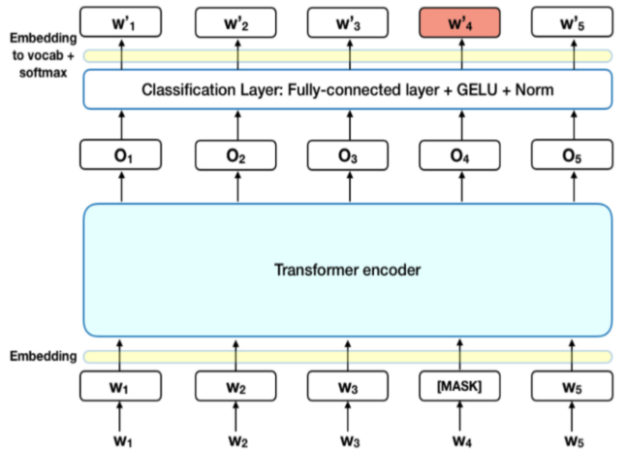
## Ensemble Learning

- ❑ We utilized the voting classifier to combine several selected models.
- ❑ They are SVC, Multinomial NB, Logistic Regression, and Random Forest Classifier.
- ❑ The voting type was set to “hard”.



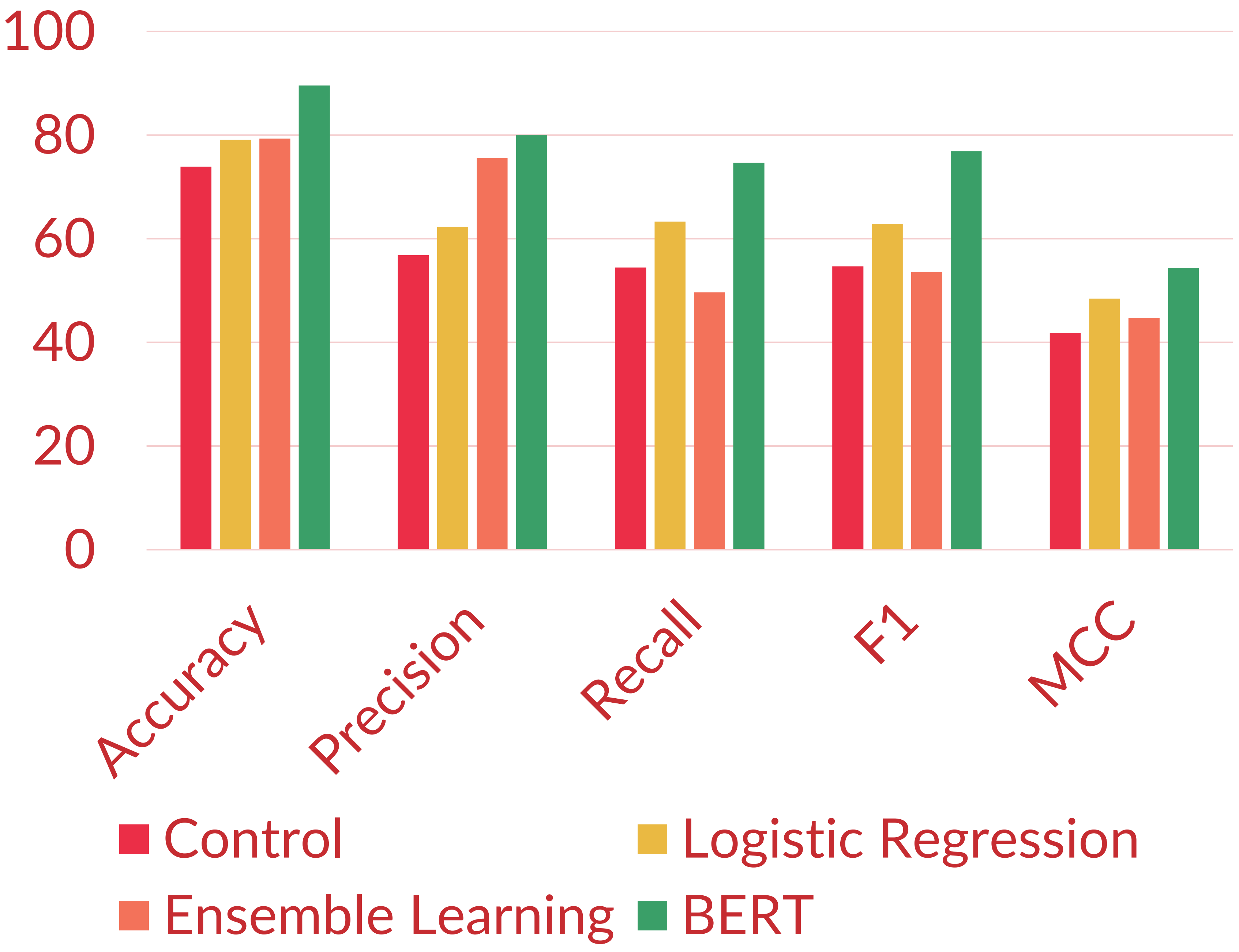
## BERTweet

- ❑ We stumbled upon BERTweet by “Vinai”.
- ❑ We based our code on a similar work that is already done on Kaggle but for disaster tweets.
- ❑ We kept the same hyperparameters (5 epochs and batch set size to 8) and changed the num_classes parameter.

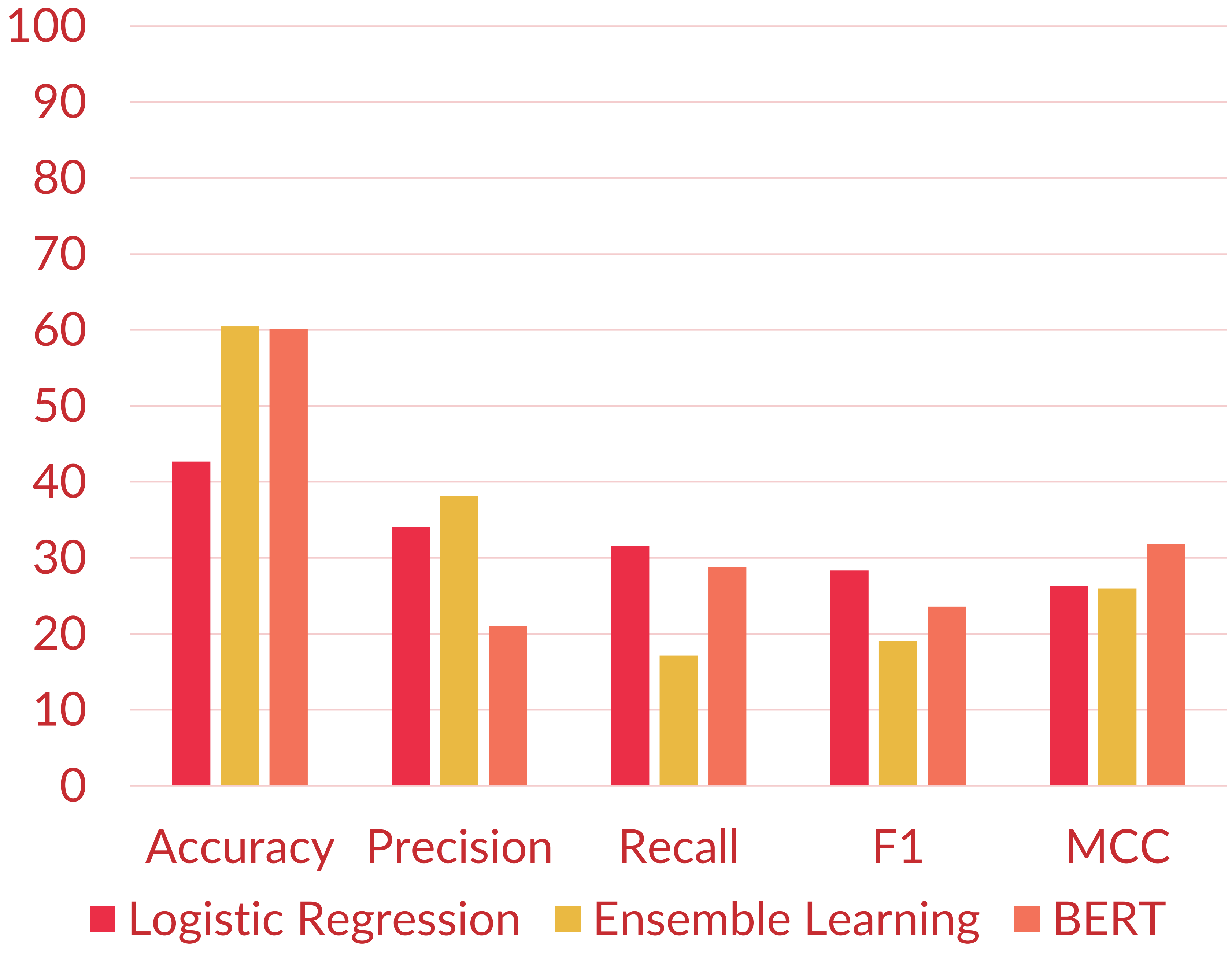

w’1
w’2
w’3
w’4
w’5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

# Results

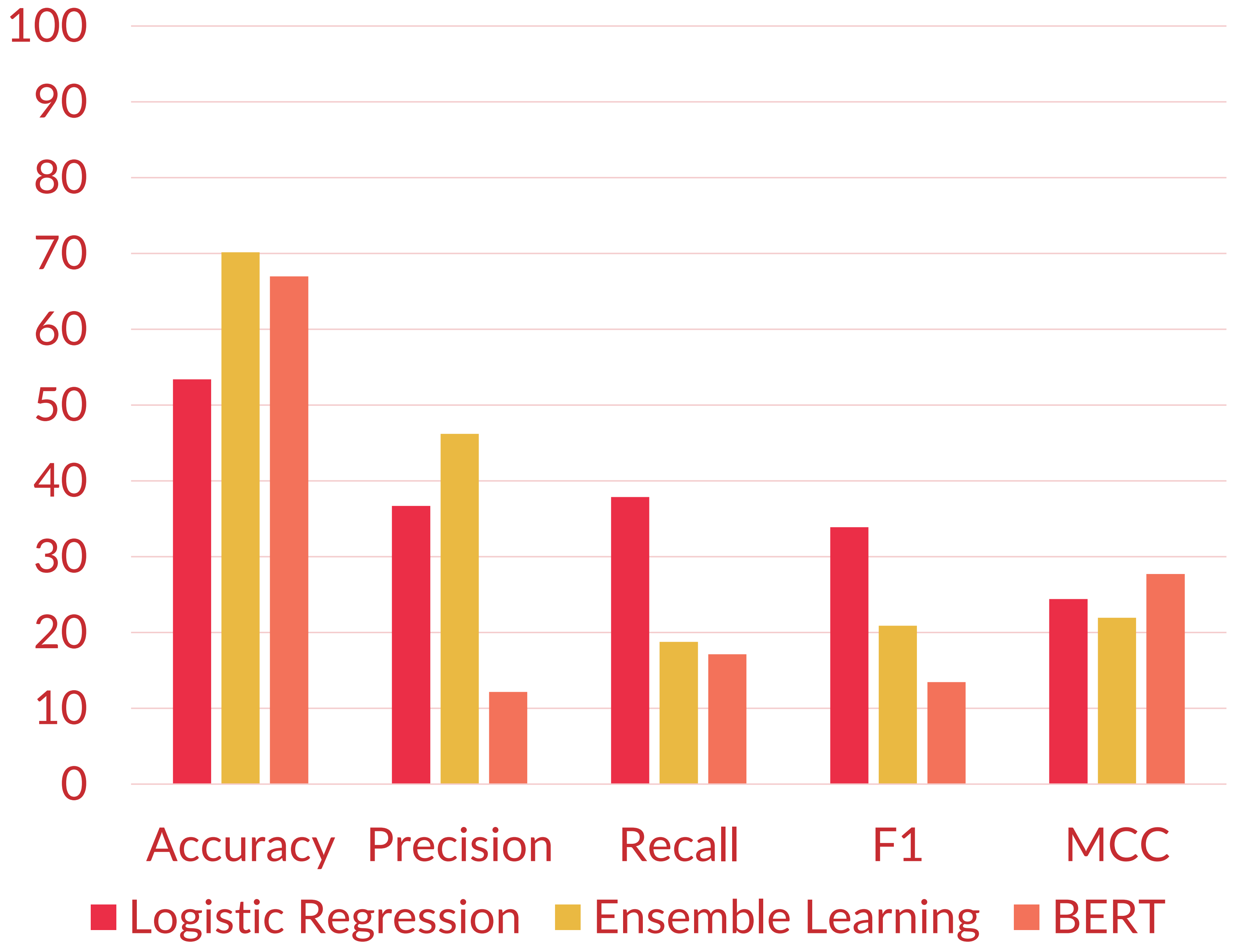
Validation Set Results for Subtask 1



Validation Set Results for Subtask 2



Validation Set Results for Subtask 3



❑ Tweet content normalization techniques improve the predictive power of the pipeline.

❑ BERT was significantly better at predicting the subtask 1 data.

❑ Logistic Regression performed the best in both subtasks 2 and 3.

❑ Logistic regression has the best recall and F1 for Subtasks 2 and 3, BERT has the comparable score and highest MCC.

# Lessons Learned (Future Research/Outlook)

Finding more optimal hyperparameters for BERTweet model

Experimenting more with many combinations of different models

Exploring other possible and effective ways to deal with more than dependent variables

Applying more state-of-the-art and net-positive normalizations

Attempting Facebook's fastText text classification and representation learning pretrained model

# Questions?