CSD 346 SEMINAR

RIMJHIM SINGH
1910110317
Computer Science Engineering, 2023

Published: February 25, 2021

A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

Furqan Rustam , Madiha Khalid,
Waqar Aslam, Vaibhav Rupapara,
Arif Mehmood, Gyu Sang Choi

01

INTRODUCTION

What arose the need for this research?

02

OBJECTIVE

What is being researched?

03

DATASET

What data has been used?

04

METHODOLOGY

How is this study being conducted?

05

CONCLUSIONS

What conclusions have been derived?

INTRODUCTION

During the lock-down, traffic on social networking has increased tremendously.

Twitter has outclassed other competitors in the timely spreading of Covid-19 news.

Overwhelming part of these news is subjective due to the involvement of personal opinions and bias, hence giving rise to (un)intentional fake information, uncertainty, and negativity in human social circles.

01

02

OBJECTIVE

In this study, **Covid 19 tweets sentiment analysis** has been performed using a **supervised machine learning** approach. Identification of Covid-19 sentiments from tweets would allow informed decisions for better handling the current pandemic situation

RESEARCH QUESTIONS

QUESTION 1

How is the performance comparison of machine learning models for Covid-19 sentiment analysis on tweets?

QUESTION 2

Can we improve the performance of machine learning models by feature engineering?

DATASET

1. The dataset used in this study is obtained from the **IEEE data port** on May 31, 2020
2. It contains the tweet IDs and sentiment scores of **7528 tweets**.
3. Filters used: language “**en**” and keywords “**corona**”, “**coronavirus**”, “**covid**”, “**pandemic**” and variants such as “**sarscov2**”, “**nCov**”, “**covid-19**”, “**ncov2019**”, “**2019ncov**” and their hashtags.

03

Table 2. A sample of the dataset available at IEEE data port [30].

Tweet ID	Sentiment Score
1266588391854030000	-0.0571428571428571
1266588391858200000	-0.4
1266588392109760000	0.25

https://doi.org/10.1371/journal.pone.0245909.t002

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

7528 TWEETS

80 : 20

6022 Tweets : TRAINING | 1506 Tweets : TESTING

CONTRIBUTIONS OF THIS RESEARCH

Performance analysis of five supervised machine learning models, for Covid-19 sentiment analysis of tweets.

Propose a feature extraction technique based on BoW and TF-IDF and evaluate its performance

METHODOLOGY

- Extract tweets using an in-house built tweets crawler.

- The tweets are cleaned using preprocessing techniques, which include removal of non-supported information and extraction of meaningful text.

- Next sentiment scores are found using the **TextBlob toolkit**. These scores are classified as **positive, negative and neutral**.

TEXTBLOB

TextBlob

For NLP Preprocessing

1. **Polarity Score [-1,1]:** Represent negative, neutral and positive statements

2. **Subjectivity Score [0,1]:** Expression of opinions, evaluations, feelings and speculations

Table 11. The TextBlob performance on original data.

Tweet ID	Tweet Text	Score	Label
1266588391858208770	@whozak @bonifacemwangi She can't walk but can dance oops @AtwoliDza is cursed but anyway biwott was also there and . . . https://t.co/7knyPDDh0	-0.4	Negative
1266588393162575872	Ppl not scared of COVID-19 no more? Our numbers definitely going back up next week.	-0.0625	Negative

Table 12. TextBlob performance after preprocessed data.

Tweet ID	Tweet Text	Score	Label
1266588391858208770	walk dance oops cursed anyway biwott also	0.0000	Neutral
1266588393162575872	ppl scare covid number definitely go back next week	0.0000	Neutral

SS1

Sentiment scores that is provided by IEEE data port

SS2

Sentiment scores found after preprocessing the data

A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

Table 3

A comparison of sentiment scores between those provided by IEEE data port (SS1) and those determined by our approach after preprocessing (SS2).

ID	Tweets	SS1	SS2
1266588393594650000	RT @ziwe: caná€˜t believe corona blew a 28-3 lead to racism	0	0
1266588391858200000	@whozak @bonifacemwangi She can't walk but can dance oops @AtwoliDza is cursed but anyway biwott was also there andá€¦ https://t.co/7knyPDDh06	-0.4	0
1266588392109760000	RT @FLOTUS: Our country allows for peaceful protests, but there is no reason for violence. Iá€˜ve seen our citizens unify & take care of oneá€¦	0.25	0.25

https://doi.org/10.1371/journal.pone.0245909.t003

doi: https://doi.org/10.1371/journal.pone.0245909.t003

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

METHODOLOGY

- For feature extraction, techniques such as term frequency-inverse document frequency (**TF-IDF**), **bag-of-words** (BoW) are used.
- A new feature technique has been proposed by **concatenating TF-IDF and BoW** features.

MACHINE LEARNING MODELS TRAINED TO TEST

Random forest (RF)

XGBoost classifier

Support vector classifier (SVC)

Extra trees classifier (ETC)

Decision tree (DT)

LSTM

FEATURE EXTRACTION TECHNIQUES

TF-IDF

TF-IDF finds the weight of each feature in a document using the product of term frequency (TF) and inverse document frequency (IDF)

BoW

Generates the vocabulary all the uniques words and their occurrence frequencies for the training of learning models

Concatenation

Concatenation of BoW and TF-IDF features to boost the accuracy.

TF-IDF Features

Doc	Feature1	Feature1	Feature n	Label
1	0.645	0	0.454	True
2	0.648	0.333	0	False

BoW Features

Doc	Feature1	Feature1	Feature n	Label
1	1	0	1	True
2	1	1	0	False

Concatenation of BoW and TF-IDF

Doc	TF-IDF Feature1	TF-IDF Feature1	TF-IDF Feature n	BoW Feature1	BoW Feature1	BoW Feature n	Label
1	0.645	0	0.454	1	0	1	True
2	0.648	0.333	0	1	1	0	False

Fig 3. Our proposed approach based on concatenation of BoW and TF-IDF.

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

EVALUATION PARAMETERS

ACCURACY SCORE

Fraction of correct predictions

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

PRECISION SCORE

Exactness of the classifiers

$$precision = \frac{TP}{(TP + FP)}.$$

RECALL SCORE

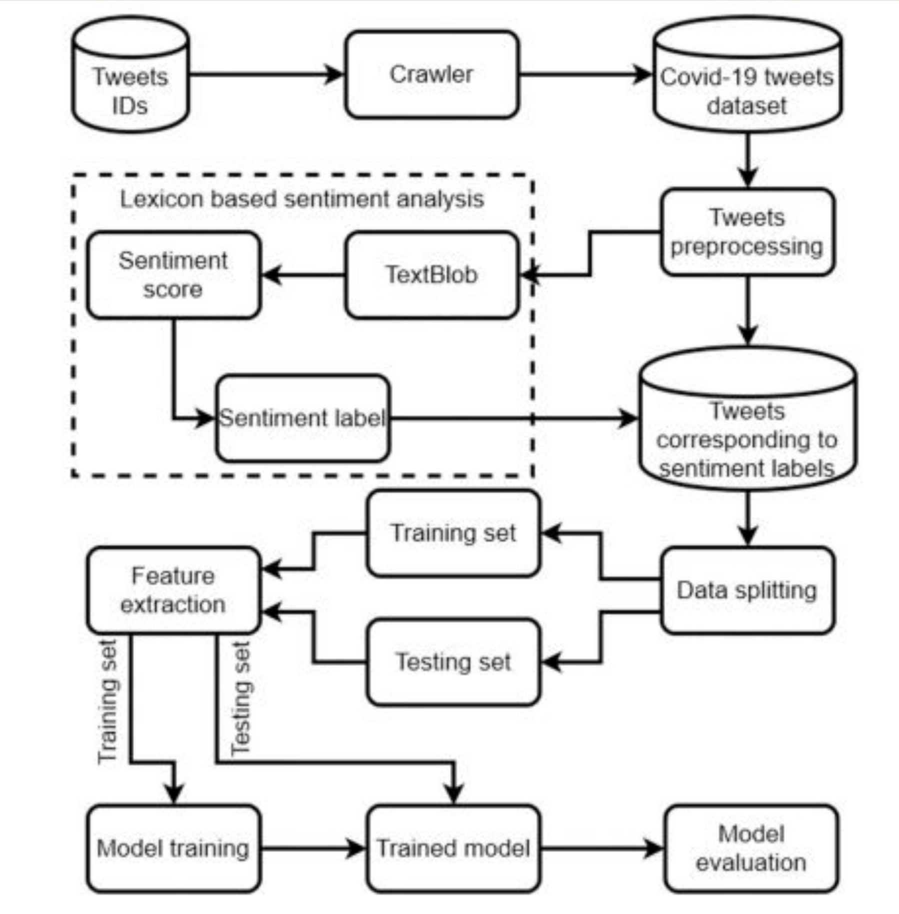
Completeness of a classifier

$$recall = \frac{TP}{(TP + FN)}.$$

F1 SCORE

Harmonic mean of precision and recall scores

$$F_1 = 2 * \frac{(precision \ * \ recall)}{(precision \ + \ recall)}.$$



Proposed Methodology

Figure: Thematic of the the proposed methodology

Removal of usernames and links

Removal of punctuation marks and conversion to lowercase

Removal of stopwords and numeric values

Stemming

REMOVAL OF USERNAMES AND LINKS

Table 7. Sample tweets after removing usernames and links.

Tweets Before Removal	Tweets After Removal of Usernames and Links
RT @skaijackson: Donald Trump called the coronavirus the Chinese virus	Donald Trump called the coronavirus the Chinese virus
RT @latimes: owners go to jail for breaking coronavirus rules? https://t.co/tltJJvIT8R	owners go to jail for breaking coronavirus rules?
RT bryanbehar: 40 million out of work #covid	40 million out of work #covid

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

REMOVAL OF PUNCTUATION MARKS AND CONVERSION TO LOWERCASE

Table 8. Sample tweets after removing punctuation marks and conversion to lower case.

Tweets Before Removal	Tweets After Removal of Punctuation Marks and Conversion to Lower Case
Donald Trump called the coronavirus the Chinese virus	donald trump called the coronavirus the chinese virus
owners go to jail for breaking coronavirus rules?	owners go to jail for breaking coronavirus rules
40 million out of work covid	40 million out of work covid

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

REMOVAL OF STOPWORDS AND NUMERIC VALUES

Table 9. Sample tweets after removing stopwords and numeric values.

Tweets Before Removal	Tweets After Removal of Stopwords and Numeric values
donald trump called the coronavirus the chinese virus	donald trump called coronavirus chinese virus
owners go to jail for breaking coronavirus rules	owners go jail breaking coronavirus rules
40 million out of work covid19	million out work covid19

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

STEMMING

Table 10. Sample tweets after stemming.

Tweets Before Stemming	After Stemming
donald trump called coronavirus chinese virus	donald trump call coronaviru chines viru
owners go jail breaking coronavirus rules	owner go jail break coronaviru rule
million out work covid19	million out work covid19

PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis

LET'S TALK RESULTS

What does the performance comparison look like?

Results with TF-IDF

1. **ETC model** performs better for both SS1 and SS2. It has a higher accuracy score 0.92 for SS2. It is due to the fact that TextBlob performs better when given a cleaned data.
2. Tree-based models give better performance for both SS1 and SS2. ETC correctly predicts 1390 tweets for SS2, while it makes 116 incorrect predictions. For SS1, ETC makes 250 incorrect predictions

Table 18. Models performance for SS1 using TF-IDF.

Model	Accuracy	Class	Precision	Recall	F_1 score
RF	0.90	0	0.84	0.99	0.91
		1	0.94	0.83	0.88
		2	0.97	0.86	0.91
		macro avg	0.92	0.89	0.90
		weighted avg	0.91	0.90	0.90
XGboost	0.90	0	0.84	0.97	0.90
		1	0.93	0.84	0.89
		2	0.95	0.86	0.90
		macro avg	0.91	0.89	0.90
		weighted avg	0.90	0.90	0.90
SVC	0.89	0	0.86	0.93	0.90
		1	0.89	0.89	0.89
		2	0.96	0.86	0.91
		macro avg	0.89	0.89	0.89
		weighted avg	0.89	0.89	0.89
ETC	0.92	0	0.88	0.99	0.93
		1	0.95	0.87	0.91
		3	0.97	0.89	0.93
		macro avg	0.93	0.92	0.92
		weighted avg	0.93	0.92	0.92
DT	0.89	0	0.88	0.92	0.90
		1	0.89	0.86	0.87
		2	0.90	0.88	0.89
		macro avg	0.89	0.89	0.89
		macro avg	0.89	0.89	0.89

Results with bag-of-words

1. **Tree-based models** RF, XGBoost, and DT perform slightly better for BoW than TF-IDF, while SVC perform better for TF-IDF.

2. ETC achieved high accuracy levels of 0.92 and 0.88 for SS2 and SS1, respectively.

1390 / 1506

FOR SS2

1326 FOR SS1

Table 20. Models performance for SS2 using BoW.

Model	Accuracy	Class	Precision	Recall	F_1 score
RF	0.91	0	0.86	0.99	0.92
		1	0.95	0.85	0.90
		2	0.97	0.87	0.92
		macro avg	0.93	0.90	0.91
		weighted avg	0.92	0.91	0.91
XGboost	0.91	0	0.87	0.95	0.91
		1	0.94	0.87	0.90
		2	0.94	0.89	0.92
		macro avg	0.92	0.91	0.91
		weighted avg	0.91	0.91	0.91
SVC	0.87	0	0.82	0.94	0.88
		1	0.88	0.80	0.84
		2	0.93	0.84	0.88
		macro avg	0.88	0.86	0.87
		weighted avg	0.87	0.87	0.87
ETC	0.92	0	0.88	0.99	0.93
		1	0.95	0.88	0.91
		2	0.98	0.88	0.93
		macro avg	0.93	0.92	0.92
		weighted avg	0.93	0.92	0.92
DT	0.91	0	0.92	0.97	0.94
		1	0.93	0.91	0.92
		2	0.93	0.90	0.92
		macro avg	0.93	0.93	0.93
		weighted avg	0.93	0.93	0.93

TF-IDF & BoW Concatenation

1. **Overall the performance is boosted with the combination.**

2. ETC accuracy improves to 0.93 when we train it using clean data. XGBoost and RF also outperform their previous results.

3. The feature set size increases so the model has more features to learn and to improve its accuracy. The improvement in results is due to the concatenated features

1412 / 1506

94 INCORRECT

Table 22. Models performance for SS2 using TF-IDF and BoW concatenation.

Model	Accuracy	Class	Precision	Recall	F_1 score
RF	0.92	0	0.87	0.99	0.93
		1	0.95	0.85	0.89
		2	0.96	0.87	0.92
		macro avg	0.92	0.91	0.92
		weighted avg	0.92	0.92	0.92
XGboost	0.92	0	0.88	0.95	0.91
		1	0.95	0.88	0.91
		2	0.95	0.90	0.92
		macro avg	0.93	0.92	0.92
		weighted avg	0.92	0.92	0.92
SVC	0.89	0	0.85	0.94	0.88
		1	0.91	0.84	0.86
		2	0.92	0.85	0.88
		macro avg	0.89	0.88	0.89
		weighted avg	0.89	0.89	0.89
ETC	0.93	0	0.89	0.93	0.91
		1	0.90	0.86	0.88
		2	0.91	0.88	0.90
		macro avg	0.90	0.89	0.89
		weighted avg	0.90	0.90	0.90
DT	0.91	0	0.92	0.97	0.94
		1	0.93	0.91	0.92
		2	0.93	0.90	0.92
		macro avg	0.93	0.93	0.93
		weighted avg	0.93	0.93	0.93

Results using Long Short-Term Memory

1: Shows very poor performance, with accuracy of 0.577.

2: The dataset is too small for a deep learning model.

WHAT WAS CONCLUDED?

The study concludes that **ETC is the best performer** with the features concatenation engineering approach.

CONCLUSIONS

1

ETC outperforms others.

The study concludes that ETC is the best performer with the features concatenation engineering approach.

2

Concatenation outperforms others

Of all three feature extraction techniques used, TF-IDF, BoW, and concatenation, the concatenation approach gave better results.

Models accuracy performance for SS1, SS2 and Vader using concatenated feature engineering technique

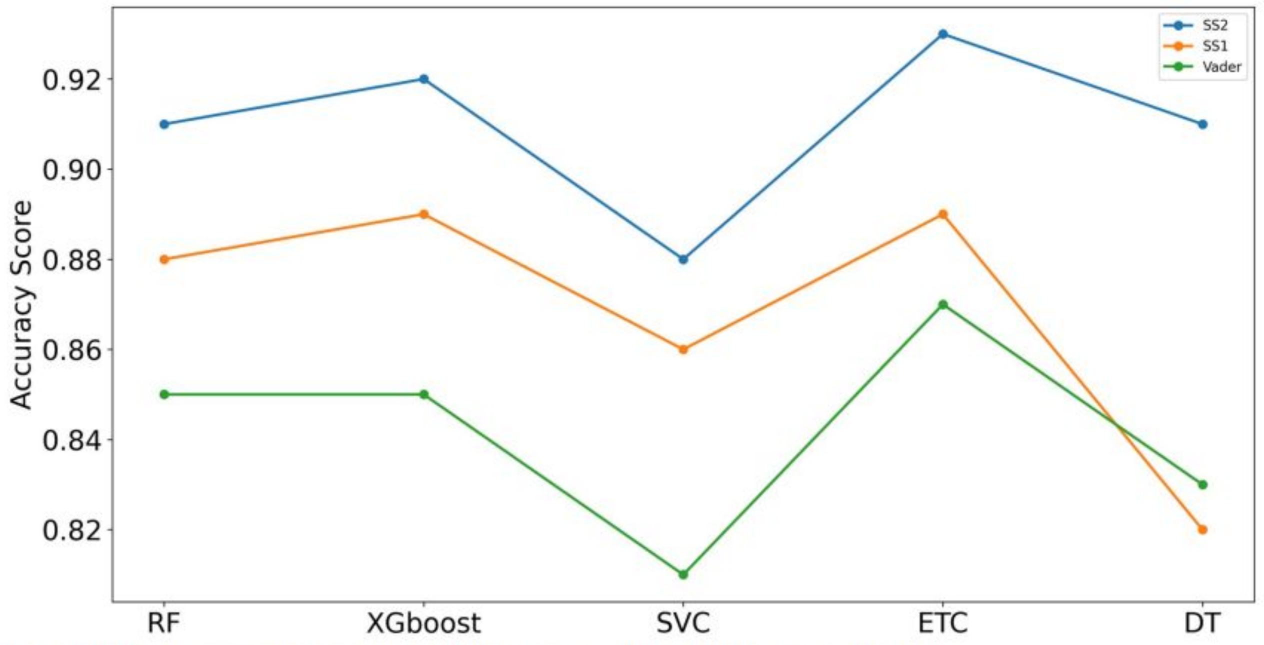


Fig 8. Models accuracy performance for SS1, SS2 and Vader using concatenated feature engineering technique.

Research questions undertook in this study

1: How is the performance comparison of machine learning models for Covid-19 sentiment analysis on tweets?

2: Can we improve the performance of machine learning models by feature engineering?

Answer 1

1

The question was addressed by acquiring a tweets dataset, cleaning it, and finding its sentiment scores using TextBlob.

TF-IDF and BoW are used for feature extraction. **ETC outperforms others.**

For the sake of completeness, we also trained and tested one deep learning model, **LSTM which exhibits the lowest performance.**

Answer 2

2

This question is addressed by proposing a feature set by concatenation of TF-IDF and BoW.

Again our proposed feature set outperforms the two standard techniques, TF-IDF and BoW.

This is due to the fact that an extended feature set allows more training points, thus increasing the chances of the test point to lie within closer proximity of one of those training points.

CITATION

- Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS (2021) A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE 16(2): e0245909. https://doi.org/ 10.1371/journal.pone.0245909

THANK YOU!

All data has been taken from here:
PLOS ONE: A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis