Feature Selection for Sentiment Classification of COVID-19 Tweets: H-TFIDF Featuring BERT

Mehtab Alam Syed[1,4], Elena Arsevska[2,5], Mathieu Roche [1,4], Maguelonne Teisseire[3,4]
[1]CIRAD, UMR TETIS, F-34398 Montpellier, France
[2]CIRAD, UMR ASTRE, F-34398 Montpellier, France
[3]INRAE, UMR TETIS, Montpellier
[4]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[5]ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France
{mehtab-alam.syed,elena.arsevska,mathieu.roche}@cirad.fr,
maguelonne.teisseire@inrae.fr

Abstract

In the first quarter of 2020, the World Health Organization (WHO) declared COVID-19 a public health emergency around the globe. Different users from all over the world shared their opinions about COVID-19 on social media platforms such as Twitter and Facebook. At the beginning of the pandemic, it became relevant to assess public opinions regarding COVID-19 using data available on social media. We used a recently proposed hierarchy-based measure for tweet analysis (H-TFIDF) for feature extraction over sentiment classification of tweets. We assessed how H-TFIDF and concatenation of H-TFIDF with bidirectional encoder representations from transformers (BH-TFIDF) perform over state-of-the-art bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF) features for sentiment classification of COVID-19 tweets. A uniform experimental setup of the training-test (90% and 10%) split scheme was used to train the classifier. Moreover, evaluation was performed with the gold standard expert labelled dataset to measure precision for each binary classified class.

Keywords: Text Mining, Sentiment Analysis, Feature Selection, Twitter

1 Introduction

In the beginning of March 2020, the World Health Organization announced the COVID-19 outbreak as a global pandemic (Dubey, 2020). The lockdown at the beginning of the pandemic affected the social activities of millions of people around the world. During this lockdown, people used social networks, especially Twitter, to express their feelings and thoughts about COVID-19. These tweets resulted in different trends of global coronavirus (Fernandes et al., 2020). These trends were helpful for health officials and other stakeholders by realizing the health crisis and its impact over different regions (WHO, 2020), (Organization and others, 2020). Due to the massive number of tweets regarding the COVID-19 pandemic, it is difficult to analyze the information. Decoupes et al. (Decoupes et al., 2021) proposed a hierarchy-based measure for tweet analysis (H-TFIDF) features from COVID-19 tweets by considering spatial and temporal dimensions. H-TFIDF captures important features that reflect local concerns by taking into account spatiotemporal aspects (Decoupes et al., 2021). These features illustrate various ways of exploring tweets in the health context of the coronavirus COVID-19 pandemic. By using an adaptive interest of these features, global insight of the evolution of features over space and time is obtained. Furthermore, H-TFIDF features greater semantic information richness, which can be helpful for sentiment classification of COVID-19 tweets. Moreover, bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) have pretrained language models that can be helpful for extracting contextual features in the context of COVID-19 tweets (Hoang et al., 2019). The main objective of our work is to perform sentiment classification of COVID-19 tweets by taking into account both spatial and semantic aspects with H-TFIDF and concatenation of BERT and H-TFIDF (BH-TFIDF) features. The objective is achieved by using a supervised learning approach. Moreover, machine learning models, i.e., linear and nonlinear, are chosen to perform the sentiment classification task. These machine learning models are trained using a publicly labeled dataset (KazAnova, 2016). Moreover, the best model is chosen among them for sentiment classification. The model predicts results using different sets of features, i.e., Bag-

of-words (BOW), TF-IDF, H-TFIDF, BH-TFIDF, and BOW+BERT. Finally, the purpose of the proposed work is to evaluate how H-TFIDF features and BH-TFIDF perform over BOW features and TF-IDF features for sentiment classification of COVID-19 tweet data. This paper is structured as follows: Section 2 describes the state-of-the-art literature related to sentiment analysis of COVID-19 tweet data of January 2020. Section 3 presents the proposed methodology. Section 4 presents the results of the experiments and a discussion of the results. In Section 5, we discuss the advantages and limitations of the proposed work and propose some future perspectives.

2 State of the art

Social media, especially Twitter, provides trends on different topics by different users around the world (Ferrara, 2020), (Shen et al., 2019). These trends of topics on the recent COVID-19 pandemic are helpful to see the impact of different stakeholders on the health crisis, current situation, and economic influences (Allain-Dupré et al., 2020). (Schouten et al., 2017) proposed both supervised learning techniques and unsupervised learning techniques for performing sentiment analysis on different aspects of Twitter data. (Gulati, 2021) presented a comparative analysis of common machine learning-based classifiers, i.e., Linear Support Vector Classifier SVC, Perceptron, Passive Aggressive Classifier and Logistic Regression found Logistic Regression, and Linear SVC (the best for all sentiment classes). Another study (Sharma and Ghose, 2021) proposed a lexicon-based approach for sentiment classification of tweet data. However, it has severe accuracy issues over machine learning techniques. Further research (Mansoor et al., 2020) proposed long short-term memory (LSTM) and artificial neural networks (ANNs) for sentiment classification of COVID-19 tweets to see the impact of coronavirus on people's lives, especially work from home (WFH) and online learning. Another study (Wisesty et al., 2021) performed a comparative analysis of sentiment classification that was performed with word embedding (word2vec and GloVe) with LSTM and BERT (bidirectional encoder representations from transformers). In these experiments, BERT performed better than other word embedding techniques for sentiment classification. Feature selection is the most important perspective apart from selecting the best models or techniques to solve the sentiment classification (Kou et al., 2020). In sentiment classification, feature selection is a crucial process in both supervised learning and unsupervised learning. Improper large feature selection may degrade classifier performance and increase the computational cost (Kumar, 2014). Feature selection techniques can be used to select an optimal subset of features, reducing the computational cost of training a classifier and potentially improving classification performance (Prusa et al., 2015). (Madasu and Elango, 2020) proposed the term frequency inverse document frequency (TF-IDF) as a feature extraction technique to obtain results with different subsets of features. (Wang and Lin, 2020) proposed a new method when selecting a suitable number of features by using the chi-square feature selection algorithm to employ feature selection using a preset score threshold. Another study (Ansari et al., 2019) proposed recursive feature elimination to select the optimal feature set and an evolutionary method based on binary particle swarm optimization of the final feature subset. These approaches were validated for sentiment analysis in five different domain balanced datasets including movie reviews and Amazon product reviews. Further work (Rustam et al., 2021) proposed a comparison of sentiment classification using different features, i.e., Bag-of-words (BOW), TF-IDF, and concatenation of BOW and TF-IDF to boost the performance. In this paper, the concatenation of BOW and TF-IDF outperformed other features in sentiment classification of COVID-19 tweets. However, the issues with features were the computational cost of model learning and overfitting of the model. To address this research gap, (Decoupes et al., 2021) proposed a set of features that are extracted from a COVID-19 tweet dataset by considering the spatial and temporal aspects of COVID-19 data. In this work, the main focus was on the hierarchical characteristics of spatial and temporal dimensions for extracting a more relevant set of features in the context. These important features, i.e., hierarchical term frequency inverse document frequency (H-TFIDF) in the tweets for different regions and time, help determine the local situation, crisis management, and opinions of inhabitants. Moreover, these reduced sets of features (H-TFIDF) may be important for sentiment classification of COVID-19 tweets. Therefore, it is impor-

tant to analyze how well these H-TFIDF features perform in the sentiment classification of COVID-19 tweets. In the proposed work, we compare H-TFIDF features and BH-TFIDF features, and we show how these features outperform state-of-the-art BOW and TF-IDF features for sentiment classification of COVID-19 tweets.

3 PROPOSED METHODOLOGY

In this work, we performed sentiment analysis of COVID-19 tweets for sentiment classification using different features, i.e., H-TFIDF, BH-TFIDF, BOW, and TF-IDF. The flow of our experiments (training and prediction steps) is shown in Figure 1. There are two major types of learning techniques: supervised learning and unsupervised learning. In supervised learning, the models are trained and tested with labeled data. However, unsupervised learning learns using features and predicts unlabeled data. The dataset for sentiment analysis of COVID-19 tweets is unlabeled and needs to be classified. For sentiment classification, prediction of sentiment of these tweets is performed using machine-learning-trained models. The process of our proposed work has two phases:

3.1 Training Phase

In the training phase, we considered three machine learning models (linear and nonlinear) for performing the task: LR, SVM with a linear kernel, and RF. These models are mainly used for classification tasks, as already explained in Section 2. The next step is to choose the dataset for training these models. This is discussed in Section 3.1.1.

3.1.1 Training Dataset

The training dataset is the well-known kaggle Sentiment140 dataset for sentiment analysis of tweets in English only. The dataset is available at `https://www.kaggle.com/kazanova/sentiment140` (KazAnova, 2016). It has labeled data for supervised learning for the classification of tweets. The dataset contains 1.6 million tweets. Tweets are annotated as (0 = negative) and (4 = positive). Later, the trained model will be used to detect sentiments for COVID-19 tweet data. The training dataset for learning models will be used for the binary classification of tweets.

	virus	causes	mental	stress	deaths
D1	1	1	1	1	0
D2	1	1	0	0	1

Table 1: Document-term matrix

3.1.2 Data Preprocessing

We next preprocessed and cleaned texts by removing unwanted words, removing stop words, special characters, etc., using the Python library *tweet-preprocessor* (Özcan, 2016), which was specifically used for cleaning the text by removing URLs, hashtags, reserved keywords, etc. Punctuation in the text was removed using regular expressions. Text standardization was applied by converting text into lowercase text, which was later used to train the model.

3.1.3 Feature Extraction

The third step in the training phase is feature selection. We used a state-of-the-art feature selection model, i.e., BOW, for the learning machine model. The BOW model is very simple and flexible for extracting features from the model. A bag of words represents the following:

1. Vocabulary of known words in the corpus.
2. Measures of the presence of each vocabulary word in each document of the corpus.

This is represented in document-term matrix form. The document-term matrix is explained with an example below.
A corpus having two documents is

```
D1: virus causes mental stress.
D2: virus causes deaths.
```

The document-term matrix of the above corpus is shown in Table 1.

3.1.4 Training Models

For the experiments, we used linear and nonlinear models for sentiment classification of COVID-19 tweets. These models are logistic regression (LR), support vector machine (SVM), and random forest (RF). These models were trained using BOW features. Moreover, we applied cross-validation to evaluate the performance of models.

3.1.5 Model Selection

It is better to evaluate the performance of each model by calculating train-test chunks of data with a cross-validation strategy (Raschka, 2018).

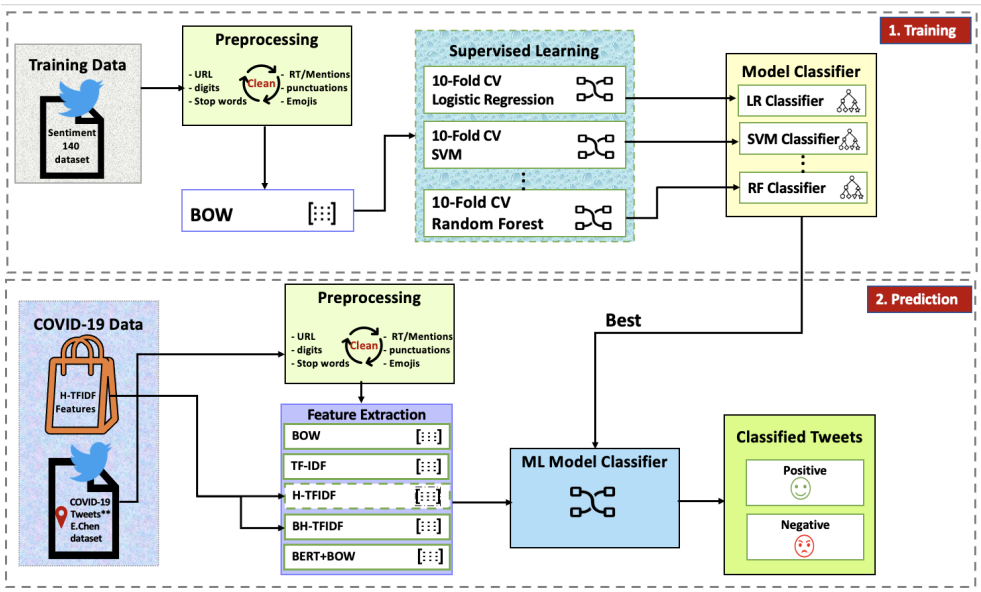


Figure 1: Process pipeline

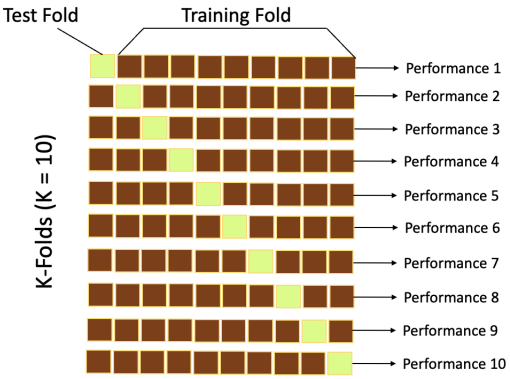


Figure 2: Cross-Validation

Cross-validation is a data resampling method to assess the generalization ability of predictive models and to prevent overfitting (Berrar, 2019). For the experimental setup, a train-test validation scheme of 90% and 10% is used with 10-fold cross validation. The dataset of 1.6 million is divided into 10 splits such that the first split has test data and the remaining nine splits are used for training in the first iteration. Similarly, in the second iteration, the first and last eight are used for training, the second iteration has test data, and a similar pattern is shown in Figure 2. The performance of each model is calculated after each iteration. The average performance of each model is shown in Table 2.

The average performance score of the three models with 10-fold cross validation (cv), i.e., 1) LR, 2) SVM, and 3) RF, are 79%, 70%, and 63%,

	BOW		
	Precision	Recall	F-Score
LR	80	79	79
SVM	71	70	70
RF	61	63	61

Table 2: Machine learning models performance with 10-fold cross validation

respectively, for the test dataset. It is clearly shown in Table 2 that LR is the best model with 10-cv for sentiment classification over other machine learning models.

3.2 Prediction Phase

In the second phase, which is the prediction phase, sentiment classification of the COVID-19 tweets is performed using the best model with different features, i.e., BOW, TF-IDF, H-TFIDF, BH-TFIDF, and BOW+BERT. As mentioned previously, we predict the sentiment classification on the tweets from January 2020, which are discussed in Section 3.2.1.

3.2.1 COVID-19 Dataset

In the second phase, we first selected the dataset of COVID-19 tweets that were extracted from *E. Chen dataset* (Chen et al., 2020). For the experiments, we extracted the COVID-19 tweets for the month of January 2020. The tweet IDs of COVID-19 were extracted using the Twitter Streaming API by using COVID-related keywords. The analysis

dataset contains 165,537 tweets. Each tweet contains the information ID, UserID, text, location, country, and its creation_date. Furthermore, data preprocessing was performed with the same strategy as discussed in section 3.1.2. Finally, sentiment analysis was performed using different sets of features, i.e., BOW, TF-IDF, H-TFIDF, and BH-TFIDF. These features are discussed in Section 3.3.1.

3.3 Data-Preprocessing

Similar to the training phase, the tweets were preprocessed through the Python library *tweet-preprocessor* (Özcan, 2016). Some examples of preprocessed tweets are as follows:

```
<Tweet1>:"@pearlylondon Don't
worry, if she does contract
a fatal dose of coronavirus
at least she will have a
dignified burial \n#Blackadder
https://t.co/8KdpMIItki".

<Tweet2>:"5 confirmed cases
of #coronavirus in Brighton.
In the meantime, local news...
#Brighton https://t.co/KTXkQCOApg"

<Preprocessed Tweet1>:"do not worry,
if she does contract a fatal
dose of coronavirus at least
she will have a dignified burial"

<Preprocessed Tweet2>:"confirmed
cases of in brighton in the
meantime, local news"
```

3.3.1 Feature Extraction

The results are calculated using BOW, TF-IDF, H-TFIDF, and BH-TFIDF. These features are discussed below.

1. **BOW**: In the first experiment, sentiment analysis is performed on the COVID-19 dataset using BOW features with the best model for classification (i.e., LR). These features were discussed in Section 3.1.3.

2. **TF-IDF**: The second experiment was performed using term frequency-inverse document frequency (TF-IDF) features using the LR model. TF-IDF is defined as in two parts. The term frequency (TF) indicates the frequency of each of the words present in the document or dataset. The second part is inverse document frequency (IDF), which actually tells us how important the word is to the document (Qaiser and Ali, 2018; Yahav et al., 2018). The basic purpose of this is to enable us to determine how each word is relevant in the document and the corpus (see equations below):

$$tf - idf(t,d) = tf(t,d) * idf(t) \quad (1)$$

$$tf(t) = f_t \Big/ f_{tot} \quad (2)$$

$$idf(t) = \log(N \Big/ df_t) \quad (3)$$

3. **H-TFIDF**: In the third experiment, a hierarchy-based measure for tweet analysis known as H-TDIDF features is used to perform sentiment analysis of the COVID-19 dataset. H-TFIDF features are the discriminative features extracted by considering spatial and temporal windows from the early beginning of the outbreak (Decoupes et al., 2021). H-TFIDF are defined in Equation (4) (Decoupes et al., 2021):

$$H - TFIDF(t, d_{(s_i,t_j)}, D_{(level_i,t_j)}) = TF(t, d_{(s_i,t_j)}) * IDF(t, D_{(level_i,t_j)}) \quad (4)$$

4. **BH-TFIDF**: In the fourth experiment, we used a combination of bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and H-TFIDF features to perform sentiment analysis of COVID-19 tweets. The main purpose of integrating BERT features is to enhance H-TFIDF features in terms of enhancing the contextual vocabulary. Moreover, due to semantic richness, it would also be helpful to improve the sentiment classification of COVID-19 tweets.

5. **BOW+BERT**: In the fifth experiment, we used a combination of bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and state-of-the-art BOW features to perform sentiment analysis of COVID-19 tweets. This combination is used to improve sentiment classification of COVID-19 tweets.

Predicted results using these features are represented by Equations (5) and (6).

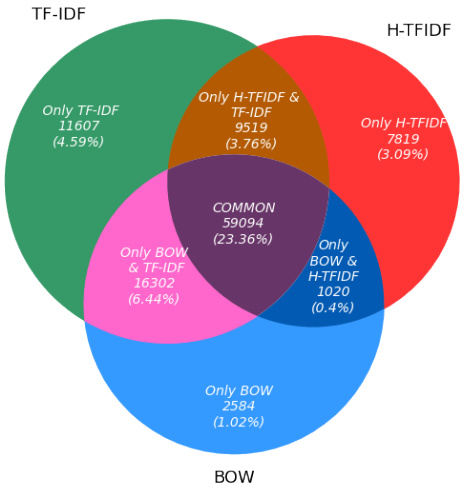
	Classification	
	Positive	**Negative**
BOW	79000	90538
TF-IDF	96522	73016
H-TFIDF	77452	92086
BH-TFIDF	97536	72002
BOW+BERT	80007	89531

Table 3: Overall sentiment classification count

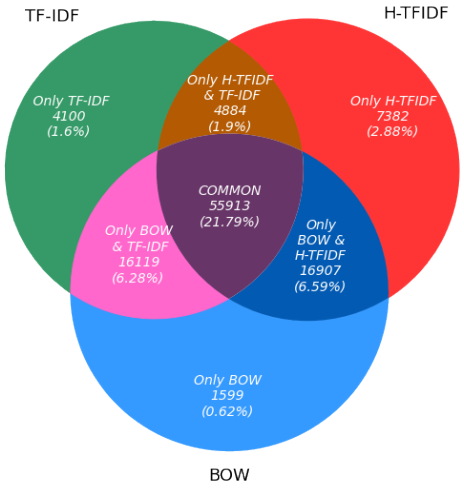
```
Let B be the BOW set
Let T be the TF-IDF set
Let H be the H-TFIDF set and
Let BH be the BH-TFIDF set
```

$$\begin{cases} Only(B) = B - (B \cap H) - (B \cap T) \\ Only(H) = H - (H \cap B) - (H \cap T) \\ Only(T) = T - (T \cap B) - (T \cap H) \\ Only(B \cap T) = (B \cap T) - (B \cap T \cap H) \\ Only(B \cap H) = (B \cap H) - (B \cap T \cap H) \\ Only(T \cap H) = (T \cap H) - (T \cap H \cap BH) \\ COMMON_{B,H,T} = B \cap H \cap T \end{cases} \tag{5}$$

$$\begin{cases} Only(B) = B - (B \cap H) - (B \cap BH) \\ Only(H) = H - (H \cap B) - (H \cap BH) \\ Only(BH) = BH - (BH \cap B) - (BH \cap H) \\ Only(BH \cap B) = (BH \cap B) - ((BH \cap B) \cap T) \\ Only(BH \cap H) = (BH \cap H) - (BH \cap H \cap T) \\ Only(H \cap B) = (H \cap B) - (H \cap B \cap BH) \\ COMMON_{B,H,BH} = B \cap H \cap BH \end{cases} \tag{6}$$



(a) Positive tweets using BOW, TF-IDF, and H-TFIDF

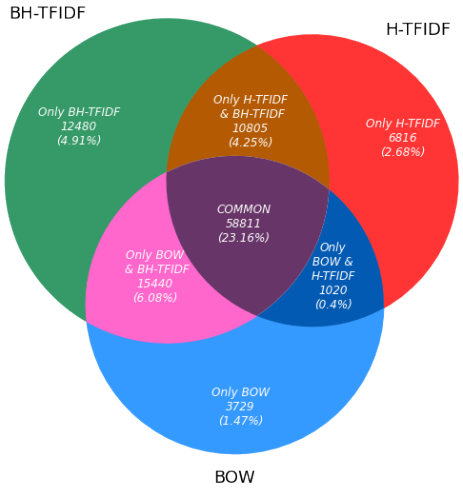


(b) Negative tweets using BOW, TF-IDF, and H-TFIDF

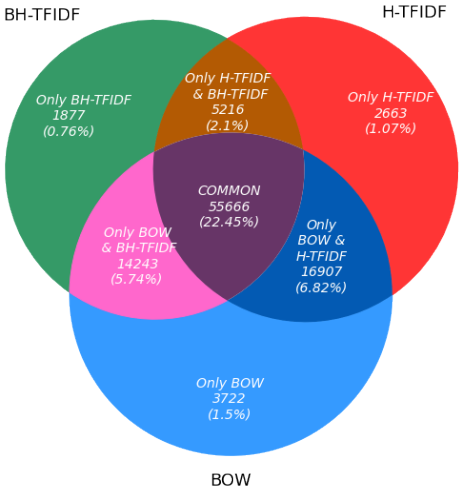
Figure 3: Positive tweet comparison by features

4 RESULTS & DISCUSSION

Binary classification of positive and negative was predicted with 4 different experiments. In each experiment, classification was performed using different sets of features, i.e., BOW, TF-IDF, H-TFIDF, and BH-TFIDF, using the LR machine learning model. The results have the final binary classification with positive and negative opinions. Overall classified positive tweets and negative tweets using different features are listed in Table 3. To compare different features, tweets with similar opinions for different features are further analyzed by the expert to find the correct classification label. The expert manually labeled 500 tweets as positive and negative, which was considered the gold standard . Furthermore, the state-of-the-art evaluation of the performance of a classification task was measured for each feature result, i.e., BOW, TF-IDF, H-TFIDF, BH-TFIDF, and BOW+BERT, with gold standards for classes "positive" and "negative," respectively. The classification matrix results in "true positives," "false positives," which results in precision for each binary class result. The precision for the "Positive" predicted class with different features is shown in Table 4. Similarly, precision for the "negative" predicted class with different features is shown in Table 5. The best feature for classifying the positive class for tweets was BH-TFIDF with a precision of 0.84. The best features for classifying negative tweets are BOW and BOW+BERT with precisions of 0.796 and 0.792, respectively. One perspective of the discussion is how discrete the extended features are over the state-of-the-art fea-



(a) Positive tweets using BOW, H-TFIDF, and BH-TFIDF



(b) Negative tweets using BOW, H-TFIDF, and BH-TFIDF

Figure 4: Negative tweet comparison by features

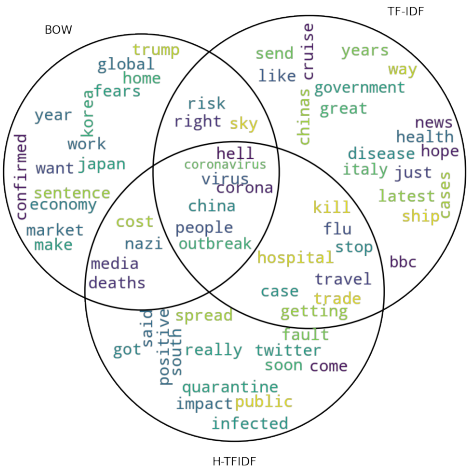
Features	Precision
BH-TFIDF	0.840
H-TFIDF	0.340
TF-IDF	0.808
BOW	0.414
BOW+BERT	0.436

Table 4: Positive Tweets: Precision, Recall, and F-Score

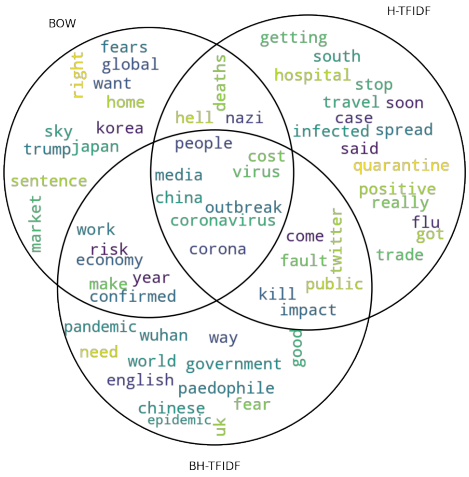
Features	Precision
BH-TFIDF	0.352
H-TFIDF	0.583
TF-IDF	0.354
BOW	0.796
BOW+BERT	0.792

Table 5: Negative Tweets: Precision, Recall, and F-Score

tures. Another perspective is how discrete the extended features performed sentiment classification over the state-of-the-art features. These perspectives were analyzed in two ways: 1) top ranked features and 2) sentiment level comparison. To compare specific and common features/tweets, we applied a visualization technique called a Venn diagram (Ho and Tan, 2021) (see Figures 5a, 5b, 3a, 3b, 4a, and 4b). Table 6 shows the top 10 feature terms in the corpus of COVID-19 tweets. In this table, features such as 'coronavirus' and 'China' are the most impacting features across different feature models. However, 'death' impacts H-TFIDF and BH-TFIDF features more than BOW features. Similarly, 'kill,' 'fault,' and 'impact' are less important features for state-of-the-art BOW and TF-IDF feature models. Similarly, in the table, these features overlap in each feature model but with differences in their rankings. Next, insight into the large set of features of all feature models was visualized using a Venn diagram. Figure 5a shows BOW, TF-IDF, and H-TFIDF features. It can be clearly visualized that the most influential features, e.g., 'coronavirus,' 'outbreak,' 'hell,' and 'China,' between them are visible in overlapping areas. However, in contrast, there are some discrete features, e.g., 'quarantine,' 'infected,' 'positive,' and 'fault,' in the H-TFIDF feature set that impact sentiment classification. Another comparison in Figure 5b shows BOW, H-TFIDF, and BH-TFIDF features. If we gain insight into the overlap between these features, then we clearly find some supreme features, e.g., 'coronavirus,' 'outbreak,' 'media,' and 'China.' However, there are some distinct influential features in H-TFIDF, e.g., 'quarantine,' 'infected,' 'stop,' 'trade,' and BH-TFIDF, e.g., 'pandemic,' 'epidemic,' 'paedophile,' and 'fear.' Conclusively, TF-IDF and H-TFIDF have more prevalent features than BOW. In addition, there are more similarities in the BOW and BH-TFIDF fea-



(a) Top BOW, TF-IDF, and H-TFIDF features



(b) Top BOW, H-TFIDF, and BH-TFIDF features

Figure 5: Top-ranked features

tures, as shown in Figure 5a. It is interesting that visualization shows a comparison of predicted results with different feature models. The first comparison provides a comparison of positively classified tweets. Figure 3a shows the results of positive tweets for the features of BOW, TF-IDF, and H-TFIDF. The exclusively predicted positive tweets using TF-IDF features are 4.59%, while those using H-TFIDF features are 3.09%, and those using BOW are 1.02%. There are 23.06% common positive tweets among them. This analysis concludes that TF-IDF results predict more positive tweets than H-TFIDF and BOW. Figure 3b shows the results of negative tweets for the features of BOW, TF-IDF, and H-TFIDF. The comparison by percentages of each solely negative tweet is H-TFIDF with 2.88%, BOW 0.62%, and TF-IDF 1.62%. Moreover, the common negative tweet percentage among all is 21.79%. In conclusion, the features that predicted more negative tweets are H-TFIDF over BOW and TF-IDF. Another interesting result is the classification of tweets of BOW and H-TFIDF with BH-TFIDF features. Figure 4a shows the results of positive tweets of features, i.e., BOW, H-TFIDF and BH-TFIDF. BH-TFIDF predicts more exclusive positive tweets, with a percentage of 4.91%, over BOW, with 1.47%, and H-TFIDF, with 2.68%. The common positive tweet percentage is 23.16% between them. Convincingly, BH-TFIDF results in more positive tweets than H-TFIDF and BOW. Figure 4b shows the results of negative tweets using features, i.e., . BOW, H-TFIDF and BH-TFIDF (BH-TFIDF). BOW predicts more exclusive negative tweets with a percentage of 1.5% than H-TFIDF with a percentage of 1.07% and BH-TFIDF with 0.76%. The prevalent negative tweet percentage between these features is 22.45%. The conclusion represented in Figure 4b clearly shows that BOW predicted more negative tweets than H-TFIDF and BH-TFIDF. The trends of the sentiment classification using different feature models are analyzed for both COVID-19 tweets and gold standard labeled tweets. These trends for the positive classification and negative classification are the same in both datasets. This clearly shows that BH-TFIDF features are more enriched toward positive classification of tweet data. On the other hand, BOW and BOW+BERT are more tilted toward negative classification of tweet data.

5 CONCLUSION

This paper proposed new feature selection measures for the sentiment classification of COVID-19 tweets. H-TFIDF features and BH-TFIDF features (both were enriched with contextual information) with other state-of-the-art features were used in the classification of tweets. These features carried out different COVID-19 aspects such as public opinions to provide insight into the local situation and government health concerns. In this work, we showed that BH-TFIDF features outperform H-TFIDF features and other state-of-the-art features, i.e., BOW and TF-IDF for classification of positive tweets. Moreover, state-of-the-art BOW features and BOW+BERT features performed better than TF-IDF, H-TFIDF, and BH-TFIDF for the negative classification of tweets.

In future work, we will focus on terminology

BOW	TF-IDF	H-TFIDF	BH-TFIDF
coronavirus	coronavirus	coronavirus	coronavirus
china	china	china	china
health	death	death	death
spread	health	health	health
cases	news	chinese	spread
deaths	pandemic	public	world
travel	want	kill	wuhan
disease	right	impact	fault
trade	travel	fault	kill
economy	hospital	travel	impact

Table 6: Top Features of BOW, TF-IDF, H-TFIDF, and BH-TFIDF

extraction approaches for the classification of COVID-19 tweets. The benefit of these approaches is that they are weakly supervised and unsupervised. The focus will be on term extraction of both single-word terms and multiword terms to further generate typed dictionaries of terminologies. The ultimate goal is to study the improvements in results in comparison with other classification methods. The proposed research focused on sentiment analysis of COVID-19 tweets during the beginning of the pandemic, as it may be useful to know about the public opinion during this period . We selected best machine learning model i.e., LR, among other models, i.e., SVM and RF, by applying cross-validation to evaluate model performance. Furthermore, experiments were performed with LR using different features (BOW, TF-IDF, H-TFIDF, and BH-TFIDF) to predict the sentiments of the tweets. Furthermore, the analysis of the results showed that BOW features performed better for predicting negative tweets. However, BH-TFIDF features were useful in predicting positive tweets in the COVID-19 dataset.

ACKNOWLEDGEMENTS

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD031. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

Dorothee Allain-Dupré, Isabelle Chatry, Varinia Michalun, and A Moisio. 2020. The territorial impact of covid-19: Managing the crisis across levels of government. *OECD*.

Gunjan Ansari, Tanvir Ahmad, and Mohammad Najmud Doja. 2019. Hybrid filter–wrapper feature selection method for sentiment classification. *Arabian Journal for Science and Engineering*, 44(11):9191–9208.

Daniel Berrar. 2019. Cross-validation. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology - Volume 1*, pages 542–545. Elsevier.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Rémy Decoupes, Rodrique Kafando, Mathieu Roche, and Maguelonne Teisseire. 2021. H-tfidf: What makes areas specific over time in the massive flow of tweets related to the covid pandemic? *AGILE: GIScience Series*, 2:1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Akash Dutt Dubey. 2020. Twitter sentiment analysis during covid-19 outbreak. *Available at SSRN 3572023*.

Blossom Fernandes, Urmi Nanda Biswas, Roseann Tan Mansukhani, Alma Vallejo Casarín, and Cecilia A Essau. 2020. The impact of covid-19 lockdown on internet use and escapism in adolescents. *Revista de psicología clínica con niños y adolescentes*, 7(3):59–65.

Emilio Ferrara. 2020. # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*.

Kamal Gulati. 2021. Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to covid-19 pandemic. *Materials Today: Proceedings*.

Sung Yang Ho and Tan. 2021. What can venn diagrams teach us about doing data science better? *International Journal of Data Science and Analytics*, 11(1):1–10.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland, September–October. Linköping University Electronic Press.

KazAnova. 2016. Sentiment140 dataset. https://www.kaggle.com/kazanova/sentiment140.

Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E Alsaadi. 2020. Evaluation of feature selection methods for text classification with small datasets using multiple criteria

decision-making methods. *Applied Soft Computing*, 86:105836.

S Vanaja K Ramesh Kumar. 2014. Analysis of feature selection algorithms on classification: a survey.

Avinash Madasu and Sivasankar Elango. 2020. Efficient feature selection techniques for sentiment analysis. *Multimedia Tools and Applications*, 79(9):6313–6335.

Muvazima Mansoor, Kirthika Gurumurthy, VR Prasad, et al. 2020. Global sentiment analysis of covid-19 tweets over time. *arXiv preprint arXiv:2010.14234*.

World Health Organization et al. 2020. Aparttogether survey: preliminary overview of refugees and migrants self-reported impact of covid-19.

Joseph D Prusa, Taghi M Khoshgoftaar, and David J Dittman. 2015. Impact of feature selection techniques for tweet sentiment classification. In *The Twenty-eighth international flairs conference*.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. 2021. A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. *Plos one*, 16(2):e0245909.

Kim Schouten, Onne Van Der Weijde, Flavius Frasincar, and Rommert Dekker. 2017. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275.

Ankita Sharma and Udayan Ghose. 2021. Lexicon a linguistic approach for sentiment classification. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 887–893. IEEE.

Chien-wen Shen, Min Chen, and Chiao-chen Wang. 2019. Analyzing the trend of o2o commerce by bilingual text mining on social media. *Computers in Human Behavior*, 101:474–483.

Zhaoxia Wang and Zhiping Lin. 2020. Optimal feature selection for learning-based algorithms for sentiment classification. *Cognitive Computation*, 12(1):238–248.

WHO. 2020. Who announces covid-19 outbreak a pandemic.

Untari N Wisesty, Rita Rismala, Wira Munggana, and Ayu Purwarianti. 2021. Comparative study of covid-19 tweets sentiment classification methods. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 588–593. IEEE.

Inbal Yahav, Onn Shehory, and David Schwartz. 2018. Comments mining with tf-idf: the inherent bias and its removal. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):437–450.

Said Özcan. 2016. tweet-preprocessor: Elegant tweet preprocessing. https://github.com/s/preprocessor.