On the Impact of Dataset Size: A Twitter Classification Case Study

Thi Huyen Nguyen
L3S Research Center, Leibniz Universität Hannover, Hannover, Germany

Hoang H. Nguyen
L3S Research Center, Leibniz Universität Hannover, Hannover, Germany

Zahra Ahmadi
L3S Research Center, Leibniz Universität Hannover, Hannover, Germany

Tuan-Anh Hoang
VNU University of Science, Hanoi, Vietnam

Thanh-Nam Doan
University of Tennessee at Chattanooga, Tennessee, USA

ABSTRACT

The recent advent and evolution of deep learning models and pre-trained embedding techniques have created a breakthrough in supervised learning. Typically, we expect that adding more labeled data improves the predictive performance of supervised models. On the other hand, collecting more labeled data is not an easy task due to several difficulties, such as manual labor costs, data privacy, and computational constraint. Hence, a comprehensive study on the relation between training set size and the classification performance of different methods could be essentially useful in the selection of a learning model for a specific task. However, the literature lacks such a thorough and systematic study. In this paper, we concentrate on this relationship in the context of short, noisy texts from Twitter. We design a systematic mechanism to comprehensively observe the performance improvement of supervised learning models with the increase of data sizes on three well-known Twitter tasks: *sentiment analysis, informativeness detection, and information relevance*. Besides, we study how significantly better the recent deep learning models are compared to traditional machine learning approaches in the case of various data sizes. Our extensive experiments show (a) recent pre-trained models have overcome big data requirements, (b) a good choice of text representation has more impact than adding more data, and (c) adding more data is not always beneficial in supervised learning.

CCS CONCEPTS

• **Information systems → Clustering and classification**.

KEYWORDS

Twitter classification, dataset size, extrapolation methods, empirical study, machine learning, neural network

ACM Reference Format:
Thi Huyen Nguyen, Hoang H. Nguyen, Zahra Ahmadi, Tuan-Anh Hoang, and Thanh-Nam Doan. 2021. On the Impact of Dataset Size: A Twitter Classification Case Study. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21), December 14–17, 2021, ESSENDON, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3486622.3493960

WI-IAT '21, December 14–17, 2021, ESSENDON, VIC, Australia

ACM ISBN 978-1-4503-9115-3/21/12…$15.00
https://doi.org/10.1145/3486622.3493960

1 INTRODUCTION

Social networks have become an essential part of the daily activities of people around the world. Among these networks, Twitter is emerging as one of the most popular platforms with over 330 million active users[1]. The platform allows users to express their own opinions in the form of short messages, known as "tweets", on various topics ranging from business and politics to casual conversation. Recently, its functionalities are extended to not only textual contents but also other types of media such as locations, videos, or even live streams. Such valuable and massive content has attracted the attention of research community. They leverage Twitter data to fulfill many tasks about real-world events, e.g., detecting and summarizing disasters/breaking news events [19, 30, 34], reporting incidents [5], or exploring user opinions [9]. Among these tasks, we focus on three well-known classification problems: *sentiment analysis*, *informativeness detection*, and *information relevance classification*. Respectively, *sentiment analysis* aims to predict whether a particular tweet contains a positive or negative opinion. *Informativeness detection* methods are employed to detect whether a tweet provides the informative content about a particular event. Besides, *information relevance* identifies whether a tweet is relevant to an occurring event or not.

Machine Learning methods have been successfully applied to these Twitter classification problems. For example, Naz *et al.* employ Support Vector Machine (SVM) to classify users' sentiment polarity on Twitter data [26]. Recently, after showing their success in various classification tasks, neural networks have gradually replaced traditional supervised techniques, such as SVM and logistic regression, as an effective method for Twitter classification. Moreover, pre-trained embedding methods gain more attention due to their ease of use and effectiveness. Despite such a rise in the use of neural networks, several crucial questions remain uninvestigated:

- First, the performance of the state-of-the-art methods on the Twitter classification problems, specifically the three aforementioned ones, with variable dataset size has not been explored. Such an investigation is especially interesting for the popular pre-trained embedding methods. Understanding

[1] https://about.twitter.com/

such connections helps build efficient machine learning systems in extreme cases, for instance, real-time applications or limited computational resources.
- Second, most pre-trained embedding methods are trained on well-written text, so their effectiveness on short-text and noisy Twitter datasets are still questionable. For instance, BERT [8] is trained on the articles of Wikipedia, and each article contains more than 1,500 words on average, while a Twitter post is limited to 140 characters. BERTweet [27] is the variant of the BERT model pre-trained on the Twitter data, but its performance on the classification tasks is still not thoroughly investigated.
- Third, existing works do not suggest concrete thresholds of data size to obtain good results or select suitable machine learning approaches for a specific Twitter classification task. Generally, supervised learning approaches require extensive training data to obtain good performances, yet it is usually challenging to secure large-scale labeled data for Twitter tasks. For instance, Twitter events evolve quickly; many incidents or disasters only occur in a short period, hence, accurately filtering event-related tweets is challenging [19]. The platform also imposes limits on the number of tweets downloaded per day. Besides, producing labeled data for supervised tasks is usually very expensive due to the high labor cost and time-consuming. Therefore, a threshold suggestion for the amount of data required to obtain the desired result is absolutely constructive.

In this paper, we study the performance of various machine learning algorithms, including both traditional methods and recent deep neural network models, across various dataset sizes and classification tasks. Theoretically, deep neural network models require big datasets to perform well. We examine whether the most recent deep learning techniques overcome that limitation and outperform traditional methods on limited data. Also, we observe how significantly different machine learning models improve their performance after adding more data. Thus, we conduct our experiments on multiple Twitter classification tasks with datasets of various sizes to compare the performance of those approaches and measure the impact of dataset size. Then, we propose functions to extrapolate the performance of machine learning models at a given dataset size. Our main contributions in this paper are as follows:

- We evaluate different machine learning models and text representation methods on three well-known Twitter classification tasks with various dataset sizes. Different from the previous known that deep learning requires big dataset to obtain good performance, the recent pre-trained models form the state-of-the-art even on small datasets.
- We examine the impact of dataset size on the performance of machine learning models and suggest the thresholds of data size for each model to obtain good classification performance. Our experiment suggests that thousands of data instances might be good enough to reach the optimal performance of a classification model.
- We illustrate the relationship between dataset size and models' performance by extrapolation functions. This experiment provides a guideline for the data needed to achieve the desired result.

2 RELATED WORKS

In this section, we briefly review several previous works closely related to ours. These works could be ordered into three principal groups: *classification problems on Twitter*, *the impact of dataset size on models' performance*, and *performance extrapolation*.

2.1 Classification Problems on Twitter

Following Twitter's success, various studies applied several natural language processing and machine learning methods on this social network's datasets in recent years. The most prominent topics include *sentiment analysis* [36, 37], *informativeness detection* [1, 33], and *information relevance classification* [2, 6]. These works implement and compare the performance of numerous machine learning approaches; however, the efficiency of the state-of-the-art models, such as embedding-based approaches [8, 25], have not been comprehensively observed. Moreover, the previous studies mainly focus on improving the classification models' results on the full datasets, but studies about the impact of the training data size on the models' performance have not been clearly elaborated.

2.2 Impact of Dataset Size on Models' Performance

The performance of supervised learning models highly depends on the dataset size. In general, the more data for training, the higher performance of classification we can achieve. Due to the strong theoretical correlation, several studies have emerged to investigate the impact of the dataset size on classification performance. Shawe-Taylor *et al.* [35] studied the constitution of small datasets. The authors proposed a Probably Approximately Correct (PAC) measurement to identify the sample size needed to guarantee accurate learning. Some recent studies investigated the extent to which the size of datasets impacts the classification performance [3, 24], yet focus on different domains such as object detection, information retrieval, or medial domain. Prusa *et al.* [32] proposed an empirical study about the effect of dataset size on training sentiment classifiers using Twitter data. Nevertheless, the authors only consider traditional machine learning models. Unlike the previous works, we investigate the impact of dataset size on the performance of various classifiers, ranging from traditional methods to the most recent deep learning techniques on the short-text datasets using Twitter. We observe on multiple classification tasks such as sentiment analysis, relevance detection, and informativeness detection.

2.3 Performance Extrapolation

Many prior works proposed learning curves to measure data size needed for a model family to reach a particular accuracy.

Frey and Fisher, in their early work [12], measured the expected data size by a decision tree to obtain the desired accuracy through some simple learning curve models such as linear, logarithmic, exponential, and power-law functions. Gu *et al.* [16] in a follow-up work improved the prediction performance of different machine learning models via six parametric extrapolating algorithms. In Machine Translation, Kolachina *et al.* [22] also presented six extrapolated learning curves to examine and predict how their statistical model performed when adding more data.

Cho *et al.* [7] presented a survey study on determining the optimum number of training medical images to obtain high classification accuracy with low variance. Accordingly, a learning curve [11] was studied to extrapolate the required size for training data by a systematic increase of sampling points to estimate the statistical mean accurately. Besides, Hestness *et al.* [18] on an empirical study validated that the power-law models improve the state-of-the-art ML and NLP models considerably, including complex deep learning methods. The authors also pointed out that the model size scales with the data size in a sublinear regime, and these scaling relationships have significant implications for deep learning progress. Also, Johnson *et al.* [20] proposed a feasible extrapolation methodology to estimate how well the system performs on a comprehensive dataset from a small pilot dataset with various extrapolations. This approach is applied to identify the biased power-law model with binomial weights for classification tasks, which makes it a stable baseline extrapolation model.

Inspired by those studies, we extrapolate the prediction performance of different models given dataset sizes in the sentiment classification problem. The extrapolation methods help observe the performance increase in a huge dataset and accurately model the relationship between data size and classifiers' performance.

3 METHODOLOGY

In this section, we introduce the details of our in-depth study. We aim to identify which model works better on a limited dataset and the level of classifier improvement after receiving more data. Besides, we observe the effectiveness of the recent deep neural networks compared to the traditional machine learning methods with respect to the changes in the data size. A previous study [32] found out an upper bound on the performance of a typical machine learning model for the sentiment classification task. We follow up the work and further identify the minimum data size required for machine learning models to obtain an acceptable result and whether adding more data is always beneficial for recent deep learning models. The upper bound performance is specified and verified for the case of huge data size by extrapolation curves. In the following sections, we first collect datasets for our classification experiments, and then we introduce base learners and experiment settings. Next, we illustrate the relationship between dataset size and models' performance by learning curves.

3.1 Datasets

We use three popular Twitter datasets in our experiments, each corresponding to one of the three classification problems of our focus. The details of these datasets are as follows:

- **COVID-19** [28]: It is relevant to the *informativeness detection* task and contains roughly 10,000 tweets about the pandemic Covid-19 with two labels: *informative* or *uninformative*.
- **CrisisLexT6** [31]: This dataset is dedicated to the *information relevance detection* task and contains about 60,000 English tweets of six crisis events in 2012 and 2013. Roughly 10,000 tweets are labeled by the experts for each event, according to relevance as *on-topic* or *off-topic*.
- **Sentiment140** [14]: The dataset is a sample of the *sentiment analysis* task. It consists of 1,600,000 tweets with two labels "*positive*" and "*negative*".

Datasets	#Tweets	
COVID-19	**#Informative**	**#Uninformative**
	3,775	4,225
CrisisLexT6	**#on-topic**	**#off-topic**
2012 Sandy Hurricane	6,138	3,870
2013 Alberta Floods	5,189	4,842
2013 Boston Bombings	5,648	4,364
2013 Oklahoma Tornado	4,827	5,165
2013 Queensland Floods	5,414	4,619
2013 West Texas Explosion	5,246	4,760
Sentiment140	**#positive**	**#negative**
	800,000	800,000

Table 1: Data Summary

The number of labeled tweets in each class of the datasets is shown in Table 1. The datasets are of various sizes with diverse degrees of class imbalance. To observe the changes in the models' performance with respect to dataset sizes, we construct sub-datasets of various sizes. Specifically, for the COVID-19 dataset and each of the CrisisLexT6 datasets, we construct sub-datasets of seven different sizes: 100, 500, 1000, 2000, 4000, 8000. At each step, we double the dataset size compared to the previous step, with an exception between the case of 100 and 500 examples. The observation on the extreme datasets of 100 or 500 examples helps us evaluate whether recent neural network models are able to overcome requirements of big data and outperform typical machine learning models. Similarly, we construct sub-datasets of 100, 500 examples for the Sentiment140 dataset. Besides, larger sub-datasets are also added, in which each subsequent sub-dataset triples the number of data instances, such as 1000, 3000, 9000, 27000, 81000 and 243000 examples. In total, we have sub-datasets of eight different data sizes for the Sentiment140 dataset.

3.2 Classification Models

We evaluate six machine learning models with diverse text representation methods on the classification tasks of various data sizes. We aim to cover both traditional machine learning models and the most recent deep learning models, both traditional text representation and state-of-the-art embedding representation methods. The list of these models is as follows:

- BERT Classifier (BERT-Cls) [8]: The state-of-the-art embedding technique, which pre-trains deep bidirectional representations from the unlabeled text. We implement a BERT model with a sequence classification head on top[2].
- BERTweet Classifier (BERTweet) [27]: BERTweet is a variant of BERT which is designed to generate pre-trained embeddings for Twitter texts. Similar to the BERT classifier, we apply BERTweet with a sequence classification head on top[3] to generate predictions on our datasets.
- Long-Short Term Memory [15] (B-LSTM): We employ a bidirectional LSTM layer followed by a fully connected layer with a Sigmoid function for classification. The input texts are represented by the pre-trained BERTweet embeddings.
- Convolutional Neural Network with a pre-trained word embeddings (W-CNN) [29]: A CNN-based classification model,

[2]https://huggingface.co/transformers/model_doc/bert.html#overview
[3]https://huggingface.co/transformers/model_doc/auto.html#

which was proposed for the relevance classification of crisis events on Twitter.
- Support Vector Machines (SVM) [17]: A well-known traditional machine learning model, yet a robust classification baseline method. We use the available implementation by Scikit-Learn[4] for our training process.
- Naïve Bayes (NB) [23]: A simple, robust probabilistic classifier, based on the Bayes' theorem. We consider this model due to its fast computation on big datasets. Besides, the model is also widely used as a baseline in many supervised learning tasks. We employ the provided version by Scikit-Learn[5].

Among the aforementioned models, SVM and NB are traditional machine learning methods, while W-CNN, B-LSTM, BERT-Cls and BERTweet are recent neural networks, which are widely adopted in many Twitter classification tasks.

3.3 Experiment Settings

We first pre-process the data: tweets are converted to lower case, and URLs and mentions are removed. Then, we remove stop-words, punctuation, and finally, extract TF-IDF of uni-grams as input feature vectors for traditional models such as SVM or NB. For the BERT-based models, each input tweet is tokenized and represented in the form of $< [CLS] \quad token_1 \quad token_2 \cdots token_N >$, where $token_i$ denotes the i^{th} token of the tweet, $i \in [1, N]$, and [CLS] is a unique token added at the beginning of each tweet, and it is used as an aggregate embedding representation of the input tweet.

We avoid any possible bias when selecting sub-datasets or evaluating models by performing several runs of 5-fold cross-validation on sub-datasets. Specifically, we sample four different sub-datasets for a given data size to avoid bias on how the data is chosen. Then, we perform 5-fold cross-validation on each sub-dataset, in which four folds are used as training data, and the remaining serves as the test data. The four runs of five folds cross-validations are repeated for all the data sizes. We use the default hyper-parameters of the Scikit-Learn library for SVM and NB, and the settings in the original paper [29] for W-CNN. The BERT-based models are trained for 10-epochs with the Adam optimizer, the warm-up strategy [21], an initial learning rate of 2e-5, and a batch size of 64.

3.4 Extrapolation methods

We observe the relationship between training data size and models' performance on huge data sizes by conducting the extrapolation task. The Sentiment140 dataset fully meets our goals since its size (1.6 M Tweets) is sufficient to produce reasonable results for evaluations. We investigate a large set of parametric curve models from the literature by Domhan *et al.* [10]. Particularly, we consider various parametric families such as linear, logarithmic, exponential, and power-law functions to extrapolate performance from smaller to larger datasets. Table 2 shows eleven different extrapolation models and their parametric formulas. These extrapolated learning curves tend to use maximum likelihood fits of each parametric model by itself, and the parameters are learned by utilizing the least square regression methods.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[5]http://scikit-learn.org/stable/modules/naive_bayes.html

Model name	Formula
vapor pressure	$\exp(a + \frac{b}{x} + c\log(x))$
pow_4	$c - (ax + b)^{-\alpha}$
pow_3	$c - ax^{-\alpha}$
exp_4	$c - e^{-ax^{\alpha}+b}$
exp_3	$c - e^{-ax+b}$
Janoschek	$\alpha - (\alpha - \beta)e^{-kx^{\delta}}$
logistic power	$\frac{a}{1+(\frac{x}{e^b})^c}$
$ilog_2$	$c - \frac{a}{\log x}$
logistic curve	$\frac{a}{1+e^{-k(x-b)}}$
$Hill_3$	$\frac{y_{max}x^{\eta}}{k^{\eta}+x^{\eta}}$
logx linear	$a\log(x) + b$

Table 2: Formulas of extrapolating learning curve models.

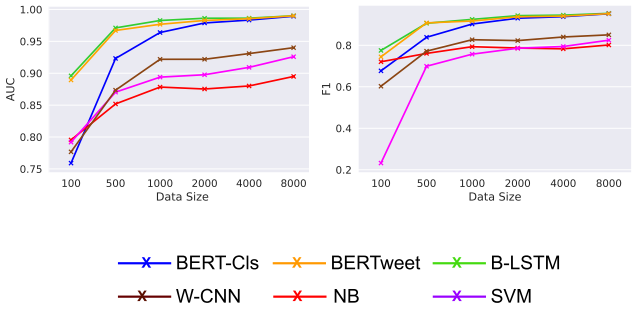

Figure 1: Performance of different models on COVID-19.

We fit the regression by models' performance at training data of sizes up to 243,000 examples. The actual classification results and the predicted value returned by the learning curve are then compared to evaluate the extrapolation task. For each classification model, we choose the best extrapolated learning curve presented in Table 2 to predict the results at a large dataset size. In this way, we reduce the computation time and resources of extensive experiments.

3.5 Evaluation Metrics

- **Classification evaluation**: Since our explored datasets are not always balanced (Table 1), we use both F1-score and AUC metrics to evaluate results. This ensures to generate a fair and comprehensive evaluation in all experiments.
- **Extrapolation evaluation**: We use *Root Mean Square Error* (RMSE) to evalutate the extrapolation functions for each classifier: $RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$, where $\hat{y}_i$ is the predicted value of data point i for a specific classifier at a given test data size n, and y_i is the corresponding label.

4 EXPERIMENT RESULTS

This section shows the performance of the classifiers with respect to the dataset sizes on the three classification problems, and then illustrates the relationship between dataset size and models' performance via extrapolation learning curves.

4.1 Classification Results

Informativeness Classification task acquires the following findings:

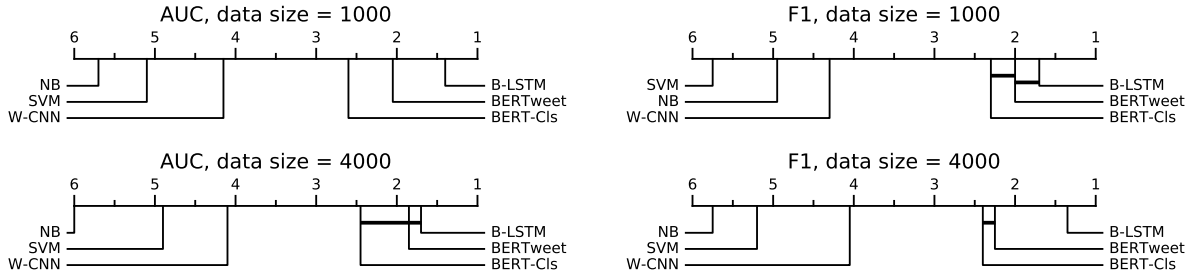

Figure 2: Critical difference diagram showing the pairwise statistical difference comparison of the classifiers on the COVID-19 dataset. A thick horizontal line groups a set of classifiers that are not significantly different (p-value>0.05).

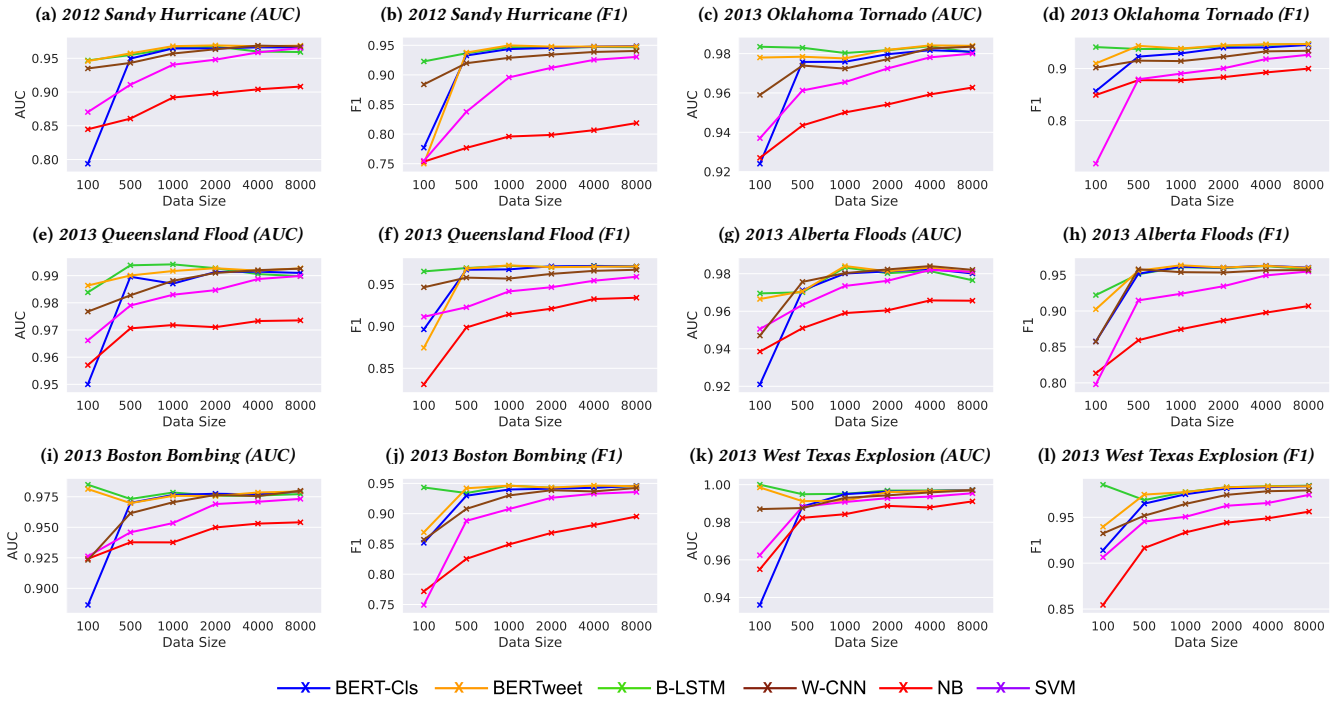

Figure 3: Performance of different models on the CrisisLexT6 datasets of various sizes.

Model	Dataset size			
	100	500	1000	2000
	F1-score st. dev			
BERT-Cls	0.09	0.03	0.02	0.01
BERTweet	0.14	0.03	0.02	0.01
B-LSTM	0.11	0.03	0.02	0.01
W-CNN	0.17	0.05	0.03	0.02
NB	0.10	0.04	0.04	0.02
SVM	0.19	0.04	0.03	0.02

Table 3: F1-score standard deviation on COVID-19.

- Figure 1 depicts the mean AUC and F1-score of each model over different cross-validation runs on the COVID-19 dataset. Generally, the performance of all classifiers improves with data size. However, the impact of adding more data diminishes significantly as data size increases. For example, the performance of SVM, W-CNN, and BERTweet models increases by 43.7%, 15.1%, and 19.2%, respectively, when the

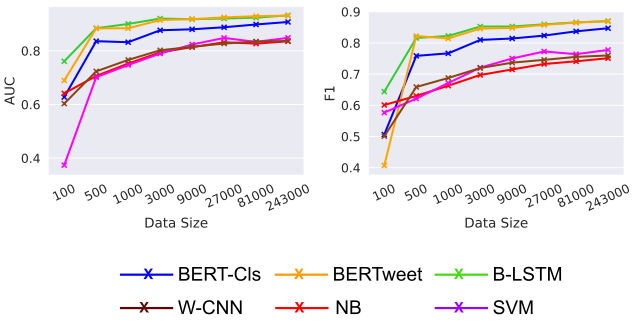

Figure 4: Performance of different models on the Sentiment140 dataset for various data sizes.

data size increases from 100 to 500 instances, while adding 500 more instances improves the performance of the three models by less than 5%. The pattern holds for the other models and the case of larger data size.

- Apart from SVM, all the classifiers achieve decent results (i.e., F1-score is higher than 60%) on an extremely small dataset of 100 examples. However, as shown in Table 3, all the classifiers require a data size of at least 1000 instances (or 2000 for NB) to obtain a stable prediction performance (i.e., the standard deviation across cross-validation runs is less than 3%).
- All the models incline to have an upper bound performance, and the recent pre-trained models require less data than traditional methods to achieve the upper bound. The three BERT embedding-based models indicate only a slight improvement after 1000 examples and tend to reach their relative upper bound performance when the data size is 4000. After that, adding more data is not clearly beneficial to these models (i.e., the prediction improves by less than 1% with doubling the data size). Meanwhile, W-CNN and other traditional models keep improving the performance after 4000 examples. However, the improvement is also less than 1.5%.
- The recent pre-trained BERT models generally perform better than the other methods for all data sizes. Interestingly, BERTweet and B-LSTM, which were trained on a small dataset of 100 examples, are comparative to the traditional models such as NB or SVM trained on the 8,000 data. Besides, an B-LSTM with pre-trained BERTweet embeddings performs notably better than a W-CNN with pre-trained word embeddings. This suggests that the input embedding representation has a significant impact on the model performance. The use of pre-trained embedding on the Twitter dataset helps BERTweet and B-LSTM achieve better performance than BERT-Cls in small datasets.
- Figure 2 illustrates the pairwise post-hoc analysis [4] with the Friedman test [13] to compare the six classifiers on two dataset sizes of 1000 (when the classifiers achieve stable results) and 4000 (when some models reach their relative upper bound performance). The purpose of the test is to check if the performance gap between two methods is statistically significant or not. The best ranks stand on the right side of the axis; hence, the three BERT-based models significantly outperform the others. Moreover, we clearly see that the two models, B-LSTM and BERTweet, which benefit from the Twitter-based pre-trained embeddings, are the best models. *Note* that similar findings are also observed in the other two tasks i.e. *Information Relevance Classification* and *Sentiment Classification* so we omit these experiments in the following parts due to space constraint.

Information Relevance Classification task results on the six events of the CrisisLexT6 dataset are illustrated in Figure 3:

- The behavior of the AUC and F1-score plots on the CrisisTexT6 dataset is relatively on par with the COVID19 dataset, except for the extremely small dataset of 100 examples, where in some events like *2013 Oklahoma Tornado* or *2013 West Texas Explosion*, BERTweet and B-LSTM achieve a better result compared to their performance on larger data sizes. That is due to the pre-trained embeddings on the Twitter data, though the results between different runs are not stable.
- All classifiers start to achieve stable performance (i.e., standard deviation across different runs is smaller than 3%) with 1000 examples. This is consistent with our observation in the COVID19 dataset.
- Generally, the information relevance detection problem on the CrisisLexT6 datasets can be solved with high accuracy by all the classifiers. Like the informativeness classification problem, B-LSTM and BERTweet are the best models in almost all cases of dataset sizes.
- In most of the cases, adding more data after 4000 examples is not helpful for the BERT-based models (i.e., BERTweet, BERT-Cls or B-LSTM).

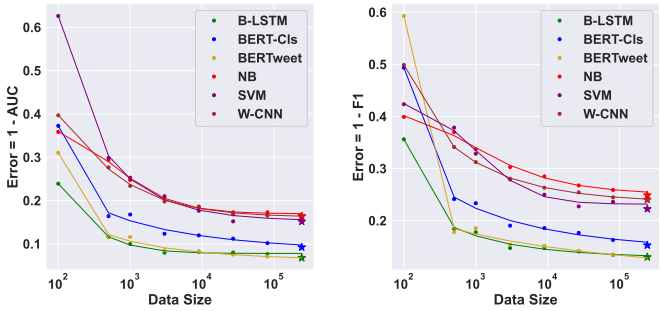

Figure 5: Extrapolation models on the AUC and F1-Score results. Lines are least-squares fits of the best learning curves for each classification model based on Table 4. The dots illustrate the actual errors. The stars are actual errors when training models on 243,000 examples.

Sentiment Classification task was carried out on the Sentiment140 dataset. Figure 4 illustrates the performance of our classifiers, and the findings are as follows:

- The performance of all models generally improves with the increase of dataset size. All the classifiers start to achieve more stable results with small deviation (less than 3%) over different runs at 1000 examples.
- Again, B-LSTM and BERTweet have the best results across different dataset sizes. The performance of BERT-based models improves quickly and tends to converge to an upper bound result faster than W-CNN or other traditional models. For example, at 3000 examples, BERT-based models achieve the relative upper bound performance, and the improvement with tripling the data size from 3000 to 9000 is less than 1%. In Figure 4, the three BERT-based models only show minor or unnoticeable changes in the improvement after 3,000 examples. Meanwhile, NB and SVM and W-CNN notably improve the performance after 3,000 examples, yet also get minor or almost no AUC improvement after 27,000 examples.

4.2 Extrapolation Results

We further enhance our observations on the upper bound performance by different extrapolation methods and identify whether classifiers change significantly on massive datasets. We consider extrapolating tasks to compute and evaluate how classification models' performance changes when varying data size. We train

Classification Model	Extrapolation Model	RMSE based-on AVG. AUC	RMSE based-on AVG. F1-Score
BERT-Cls	hill3	0.01119	0.01108
	exp3	0.01841	0.01886
	exp4	0.01031	0.00948
	ilog2	0.03428	0.04248
	Janoschek	0.01087	0.01034
	logistic power	0.01119	0.01108
	logistic curve	0.01919	0.02002
	logx linear	0.05551	0.06845
	pow3	0.01027	0.00947
	pow4	**0.00731**	**0.00566**
	vap	0.00937	0.00896
BERTweet	hill3	0.00784	0.0131
	exp3	0.01308	0.01554
	exp4	0.04104	0.02958
	ilog2	0.03451	0.08197
	Janoschek	0.00778	0.01267
	logistic power	0.00784	0.01310
	logistic curve	0.01342	0.01571
	logx linear	0.05310	0.11432
	pow3	0.04494	0.01194
	pow4	**0.00465**	**0.00570**
	vap	0.00721	0.02431
B-LSTM	hill3	0.00234	0.00622
	exp3	0.00538	0.01165
	exp4	**0.00230**	0.02976
	ilog2	0.02380	0.03013
	Janoschek	0.00275	0.00611
	logistic power	0.00234	0.00622
	logistic curve	0.00569	0.01215
	logx linear	0.03702	0.04767
	pow3	0.00241	0.00568
	pow4	0.00232	**0.00410**
	vap	0.00249	0.00528
W-CNN	hill3	0.00312	0.00388
	exp3	0.01122	0.01471
	exp4	0.00329	0.00223
	ilog2	0.01483	0.02115
	Janoschek	0.00329	0.00298
	logistic power	0.00312	0.00388
	logistic curve	0.01209	0.01609
	logx linear	0.03672	0.04391
	pow3	0.00399	0.00219
	pow4	**0.00262**	**0.00092**
	vap	0.00875	0.00305
NB	hill3	0.00970	0.00752
	exp3	0.00945	0.00921
	exp4	0.00441	0.00472
	ilog2	0.01110	0.01132
	Janoschek	0.00441	0.00472
	logistic power	0.00970	0.00752
	logistic curve	0.01039	0.00957
	logx linear	0.02562	0.01152
	pow3	0.01055	0.00783
	pow4	**0.00348**	**0.00335**
	vap	0.01585	0.01047
SVM	hill3	0.01229	0.01454
	exp3	0.02493	0.00944
	exp4	0.00806	0.00576
	ilog2	0.05380	0.01636
	Janoschek	0.00889	0.00576
	logistic power	0.01229	0.01454
	logistic curve	0.02895	0.01056
	logx linear	0.09424	0.02371
	pow3	0.00796	0.01543
	pow4	**0.00701**	**0.00493**
	vap	0.00860	0.02046

Table 4: Extrapolation models and RMSE evaluations based on average AUC and F1-Score. The better models have smaller RMSE. We highlight the best result of each method.

Classification Model	Data Size						
	3K	*9K*	*27K*	*81K*	*243K*	*729K*	*2187K*
	Actual AUC				Extrapolated AUC		
BERT-Cls	**0.877**	0.880	0.888	0.898	0.903	0.906	0.909
BERTweet	**0.915**	0.918	0.925	0.929	0.932	0.934	0.935
B-LSTM	**0.920**	0.918	0.920	0.924	0.922	0.922	0.922
W-CNN	0.802	0.815	0.827	**0.835**	0.836	0.837	0.838
NB	0.794	0.814	**0.832**	0.827	0.830	0.830	0.831
SVM	0.790	0.826	**0.848**	0.833	0.844	0.846	0.847

Classification Model	Data Size						
	3K	*9K*	*27K*	*81K*	*243K*	*729K*	*2187K*
	Actual F1-Score				Extrapolated F1-Score		
BERT-Cls	**0.810**	0.814	0.824	0.837	0.842	0.846	0.849
BERTweet	**0.846**	0.848	0.858	0.865	0.871	0.877	0.881
B-LSTM	**0.853**	0.853	0.860	0.866	0.867	0.869	0.870
W-CNN	0.720	0.737	0.745	**0.755**	0.759	0.761	0.763
NB	0.697	0.715	**0.732**	0.741	0.745	0.748	0.749
SVM	0.721	0.750	**0.773**	0.764	0.768	0.768	0.768

Table 5: The performance increment of the classifiers based on the actual and extrapolated AUC and F1-Score.

eleven extrapolating learning curve models as shown in Table 2 and examine root mean square errors (RMSE) with seven points of data size from 100 to 81000. Table 4 shows RMSE of the extrapolating learning curve functions for different models. Interestingly, the power-law function with four parameters (i.e. *pow4*) produces the best outcomes on almost all the classification models, except the AUC results for the B-LSTM model.

Based on the results in Table 4, we select the best extrapolation method for each classifier in our later evaluations. We illustrate the errors in Figure 5 and the extrapolated values of the average AUC and F1-Score on the larger data size in Table 5. These selected learning curves could efficiently model the relationship between performance and data size of the presented machine learning models with relatively low RMSE, especially W-CNN, NB, and B-LSTM. Accordingly, Figure 5 illustrates how the extrapolation functions generalize classification errors on both AUC and F1-Score. Table 5 confirms minor changes of BERT-based models when the data size is more than 3000. Specifically, by tripling the data from 3000 to 9000 examples, B-LSTM, BERTweet and BERT-Cls improve F1-score by 0%, 0.2%, and 0.4%, respectively. When running on the data size of 3000 and more than 2 million examples, the performance difference of these classifiers is also not significant. Meanwhile, other methods only start to observe minor or no improvement in the performance at the data size of 27000 with NB and SVM models or 81000 with W-CNN model.

4.3 Discussion

To sum up our extensive experiments, we observe:

- Deep learning models with pre-trained BERTweet have become the state-of-the-art techniques in solving various Twitter classification tasks. The methods overcome the requirement of a large dataset and perform well even on very small data of 100 or 500 examples.
- In general, to obtain stable results, we suggest collecting a dataset of more than 1000 examples.
- All the classifiers have a relative upper bound performance. The BERT-based models converge to the relative upper bound results generally faster than other models. For example, with 4000 examples for the *informativeness detection* or the the

relevance classification tasks, or 3000 examples for *sentiment classification* task. Hence, we do not need millions of labeled data to train a good classifier.

- The pre-trained embedding has more significant impact on model performance than adding more data. A good pre-trained model can help models obtain good results at a tiny dataset of 100 examples. Besides, With less than 1000 instances, a pre-trained model can obtain a competitive result to a traditional machine learning model on a dataset of thousands examples.
- Our extrapolation methods can effectively model the relationship between dataset size and models' performance. The methods can well predict the data needed for a model to reach its relative upper bound performance. Our experiments again ensure that thousands of examples might be adequate for a model to reach its relatively optimal performance rather than millions.

5 CONCLUSION

We studied the impact of dataset size on the performance of different machine learning models concerning three Twitter classification problems. Our experiments present a concrete view of how various machine learning approaches behave with the increment of data. We show that the recent BERT embedding-based models form the best classifiers and work well even on small datasets. Moreover, depending on the tasks, an upper bound performance exists for all classification methods, which can be achieved on datasets of thousands or tens of thousands of instances. We illustrate the dependency curve of the classification models' performance on dataset size. The extrapolation curves estimate the data size needed for machine learning models to achieve the desired score. Our findings are also in agreement with the previous studies [3, 7, 32].

ACKNOWLEDGMENTS

This work is supported by the DFG Grant (NI-1760/1-1) Managed Forgetting, the European Union's Horizon 2020 research and innovation program under grant agreement No. 832921 (project MIROR), and No. 833635 (project ROXANNE).

REFERENCES

[1] Piush Aggarwal. 2019. Classification approaches to identify informative tweets. In *Proceedings of the Student Research Workshop Associated with RANLP*.
[2] Ahmed Sulaiman M Alharbi and Elise de Doncker. 2019. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Cognitive Systems Research* (2019).
[3] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences* (2021).
[4] Alessio Benavoli, Giorgio Corani, and Francesca Mangili. 2016. Should we really use post-hoc tests based on mean-ranks? *Machine Learning Research* (2016).
[5] Cody Buntain, Jennifer Golbeck, Brooke Liu, and Gary LaFree. 2016. Evaluating Public Response to the Boston Marathon Bombing and Other Acts of Terrorism through Twitter.. In *ICWSM*.
[6] Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. 2017. Multi-class machine classification of suicide-related communication on Twitter. *Online social networks and media* 2 (2017), 32–44.
[7] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?. In *arXiv preprint arXiv:1511.06348*.
[8] Jacob Delvin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.
[9] Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *SIGCHI*.
[10] Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. 2015. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *IJCAI*.
[11] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. Predicting sample size required for classification performance. *BMC medical informatics and decision making* (2012).
[12] Lewis J Frey and Douglas H Fisher. 1999. Modeling decision tree performance with the power law. In *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR.
[13] Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics* (1940).
[14] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* (2009).
[15] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* (2016).
[16] Baohua Gu, Feifang Hu, and Huan Liu. 2001. Modelling classification performance for large data sets. In *International Conference on Web-Age Information Management*.
[17] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* (1998).
[18] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. In *arXiv preprint arXiv:1712.00409*.
[19] Tuan-Anh Hoang, Thi Huyen Nguyen, and Wolfgang Nejdl. 2019. Efficient Tracking of Breaking News in Twitter. In *WebSci*.
[20] Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman. 2018. Predicting accuracy on large datasets from smaller pilot data. In *ACL*.
[21] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
[22] Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *ACL*.
[23] David D. Lewis. 1998. The independence assumption in information retrieval. In *ECML*.
[24] Trond Linjordet and Krisztian Balog. 2019. Impact of Training Dataset Size on Neural Answer Selection Models. In *ECIR*.
[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*.
[26] Sheeba Naz, Aditi Sharan, and Nidhi Malik. 2018. Sentiment classification on twitter data using support vector machine. In *WI*.
[27] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *EMNLP: System Demonstrations*.
[28] 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 Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*.
[29] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks. In *ICWSM*.
[30] Thi Huyen Nguyen, Tuan-Anh Hoang, and Wolfgang Nejdl. 2019. Efficient Summarizing of Evolving Events from Twitter Streams. In *SDM*.
[31] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.
[32] Joseph Prusa, Taghi M. Khoshgoftaar, and Naeem Seliya. 2015. The Effect of Dataset Size on Training Tweet Sentiment Classifiers. *ICMLA* (2015).
[33] C Rossi, FS Acerbo, K Ylinen, I Juga, P Nurmi, A Bosca, F Tarasconi, M Cristoforetti, and A Alikadic. 2018. Early detection and information extraction for weather-induced floods using social media streams. *International journal of disaster risk reduction* (2018).
[34] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *TheWebConf*.
[35] John Shawe-Taylor, Martin Anthony, and N.L.Biggs. 1993. Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics* (1993).
[36] Bing Xiang and Liang Zhou. 2014. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *ACL*.
[37] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *TMIS* (2018).