Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5428–5435
Marseille, 20-25 June 2022

Classifying Implant-Bearing Patients via their Medical Histories: a Pre-Study on Swedish EMRs with Semi-Supervised GAN-BERT

Benjamin Danielsson[*], Marina Santini[§], Peter Lundberg[†], Yosef Al-Abasse[**], Arne Jönsson[*], Emma Eneling[‡], Magnus Stridsman[‡]
[*]Department of Computer and Information Science,
Linköping University, Sweden
benda425@student.liu.se, arne.jonsson@liu.se

[§]RISE Research Institutes of Sweden, Sweden
marina.santini@ri.se

[†]Radiation Physics, Center for Medical Imaging and Visualization (CMIV), Linköping University, and
Department of Health, Medicine and Care, Linköping University, Linköping, Sweden
peter.lundberg@liu.se

[**]Radiation Physics, and Department of Health, Medicine and Care,
Linköping University, Linköping, Sweden
yosef.al-abasse@regionostergotland.se

[‡]Unit for Technology Assessment, Testing and Innovation, and Department of Biomedical Engineering,
Linköping University, Linköping, Sweden
emma.eneling@regionostergotland.se, magnus.stridsman@regionostergotland.se

Abstract

In this paper, we compare the performance of two BERT-based text classifiers whose task is to classify patients (more precisely, their medical histories) as having or not having implant(s) in their body. One classifier is a fully-supervised BERT classifier. The other one is a semi-supervised GAN-BERT classifier. Both models are compared against a fully-supervised SVM classifier. Since fully-supervised classification is expensive in terms of data annotation, with the experiments presented in this paper, we investigate whether we can achieve a competitive performance with a semi-supervised classifier based only on a small amount of annotated data. Results are promising and show that the semi-supervised classifier has a competitive performance when compared with the fully-supervised classifier.

Keywords: text classification, BERT, GAN-BERT, electronic medical records, EMR, clinical text mining

1. Introduction

The paradigm shift in NLP (Natural Language Processing) (Sun et al., 2021) based on large pre-trained language models that can be fine-tuned to downstream tasks has boosted the performance of many NLP tasks when tested on traditional benchmarks (Min et al., 2021).

In the experiments presented in this paper, we investigate a downstream task that is related to a real-world need, where standard benchmarks are not available, namely the automatic identification of patients who bear implant(s) in their body.

Identifying the presence of implants in certain patients is important for radiologists and other clinical professionals because some implants are not compatible with MRI scanning. The current process to ascertain the presence of implants in a patient is manual (Kihlberg and Lundberg, 2019), slow and error prone, especially if patients are elderly and might have forgotten about implants they had when they were younger. Additionally, even if implants are removed, parts (like leads) might remain in the body and cause damage (burns or scorches) to the patient during the MRI scanning.

The core idea of these experiments is then to classify implant-bearing patients on the basis of their medical histories. We rely on the following assumption: if patients underwent an implant operation, certainly this information is stored in patients' medical records. By sieving through medical records, we circumvent the risk of overlooking crucial clinical information that are useful for professional staff.

Therefore, rather than relying on patients' memory, we propose addressing this problem as a text classification task that leverages on state-of-the-art NLP. One of the most successful models of this generation is BERT (Devlin et al., 2018). BERT has proved to achieve the state of the art in full-supervised text classification in different languages and in different genres (González-Carvajal and Garrido-Merchán, 2020). Being fine-tuned as a fully-supervised classifier, BERT classification models normally need large amounts of labelled data, a possibility that is often prohibitive in real-world scenarios.

This issue has already been pointed out and addressed by Croce et al. (2020) who proposed GAN-BERT, a semi-supervised model based on Semi-Supervised General Adversarial Networks (SS-GANs) used to cut down the need of annotated data without negatively affecting the performance of the classifier. In their experiments, GAN-BERT achieved promising results in downstream tasks (namely, topic classification, question classification, sentiment analysis and natural language inference) evaluated on English benchmarks.

Inspired by these findings, we decided to start out our own investigation by re-using the code provided with semi-supervised GAN-BERT for the real-world task of classifying implant-bearing patients via their medical histories. Code reuse and experimental replication are pillars of modern research. They support both research and development at low cost, and are a must before initiating any ad-hoc modelling. Essentially, GAN-BERT will be re-used for a different downstream task (i.e. the classification of implant-bearing patients), on a different language (i.e. Swedish) and on a different genre (i.e. electronic medical records, EMRs).

The main research question we would like to answer with the experiments presented in is paper is whether and to what extent it is possible to achieve a competitive performance with a semi-supervised GAN-BERT model based only on a relatively small amount of annotated EMRs written in Swedish for the classification of patients bearing implant(s). The results are also compared to an SVM baseline which puts the BERT results into a greater context in regards to the NLP field at large. The answer to the research question will provide the following contributions to the community: (1) the re-use and evaluation of an existing semi-supervised model (GAN-BERT) on text classification, i.e a downstream task where the model was not tested upon; (2) the applicability of an existing model tailored on English to the Swedish language, i.e. a different language; (3) the applicability of an existing model on a difficult genre, namely EMRs. Results are indeed informative and lead the way to future experiments that capitalize on unlabelled data.[1]

2. Previous Work

We are not aware of any previous study on the classification of implant-bearing patients via their medical histories. Computationally speaking, this task is treated as a text classification problem and in this pre-study it is a binary classification problem because we are going to use only two classes, one positive class and one negative class, with unbalanced distribution (as explained in Section 3).

BERT has proved to achieve the state of the art in text classification in fully supervised settings (Sun et al., 2019; González-Carvajal and Garrido-Merchán, 2020). It has also been shown, however, that although BERT outperforms baselines in standard datasets with large training sets, when the training sets are small, simpler methods, like fastText (Joulin et al., 2017) combined with domain-specific word embeddings perform equally well or better than BERT (Edwards et al., 2020).

In real-world scenarios, obtaining reliable annotated data is expensive and time-consuming, even if the training set is supposed to be small. Conversely, the unlabeled examples that represent a specific target task are often abundant, but remain unutilized in fully-supervised approaches. This is why in practical empirical settings the semi-supervised approach would be ideal.

A recent approach proposed by Croce et al. (2020) implements a BERT-based semi-supervised approach that capitalizes on unlabelled data. This approach has been inspired by promising research in image processing and has been successfully transferred to NLP. Such a method is based on Semi-Supervised Generative Adversarial Networks (SS-GANs) (Salimans et al., 2016a) and has proved to be effective in several NLP tasks when evaluated on standard benchmarks (Croce et al., 2020; Owen, 2020; Breazzano et al., 2021; Zaharia et al., 2021). Croce et al. (2020) showed that GAN-BERT cuts off the need of labelled examples and their experiments show that with fewer than 200 labelled examples results competitive with fully supervised settings can be achieved.

As pointed out earlier, given GAN-BERT's promising performance on standard benchmarks, on a range of tasks and on range of different languages, in the experiments presented here we test GAN-BERT on a real-world dataset (not an academic one) representing a difficult genre (EMRs), on a specific practical downstream task (classification of patients with implants) and on an untested language (Swedish).

3. Data, Datasets and Annotation

The data that we use in the following experiments is a small random sample of Swedish EMRs from Region Östergötland (Sweden). This random sample has been extracted from a much larger corpus of EMRs coming from four clinics, i.e. neurology, cardiology and two different orthopedic departments. The corpus includes EMRs covering a period of 5 years.

The random sample was built using EMRs belonging to the cardiology and neurology clinics. In this sample, a patient is represented by a varying number of EMRs. Some patients are represented only by a single EMR, most of the patients are represented by several EMRs. Essentially, the medical history of a patient is made of one or multiple EMRs. Consequently, the length of medical histories varies a lot across the patients represented in the sample. 184 medical histories have a length between 4 and 64 words; 1309 medical

[1] The study was approved by the Swedish Ethical Review Authority (Etikprövningsmyndigheten), authorization nr.: 2021-00890.

histories have a length between 65 and 512 words; 328 medical histories have a word length between 513 and 1000; 1166 medical histories have a length greater than 1000 words.

The sample contains the medical histories of 2987 patients, out of which 1203 were labelled and 1784 are used as unlabelled. The labelled random sample is divided into two classes, namely the Y(es) class, representing medical histories of patients that have or have had an implant, and the N(o) class, i.e. patients who have no implants. The decision of having or not having an implant is based on the mentions of terms indicative of implants in the medical histories. If the medical history of a patient has no mentions of terms indicative of implants, then we assume that the patient has no implant. In previous studies on the same corpus of EMRs (Jerdhaf et al., 2021), we identified a list of terms that are indicative of implants. This list was evaluated by two domain experts (MRI physicists), and in these experiments we use only the terms where the two experts have a 100% agreement in assessing that a term is indicative of implants. These terms were used as keywords to automatically tag patients' medical histories.

The 1784 unlabelled medical histories can belong to classes other than Y or N. For example, they can represent medical histories where the domain experts felt "unsure" or disagreed on the final labels.

The labelled dataset is unbalanced, as unbalanced is the number of patients wearing implants in real life, since the majority of patients has NO implants. Therefore, the class distribution of our dataset well represent a real-life population. The distribution of the classes in the dataset is the following: out of 1203 labelled medical histories, 250 represent patients with implants (the Y class), and 953 represent patients with NO implants (N class).

Since the dataset contains information that can be traced back to patients, staff and locations, it cannot be released at present.

4. Methods: Swedish BERT

In this section we briefly describe the pre-trained (Malmsten et al., 2020) and the fine-tuned Swedish BERT (Jerdhaf et al., 2021), as well as the fully supervised BERT classifier, the semi-supervised GAN-BERT and an SVM classifier.

4.1. SVM: Traditional ML as Baseline

The baseline used for comparisons is a Support Vector Machine (SVM), a method that attempts to find the maximum margin hyperplane between two categories (Joachims, 1998). The SVM classifier was implemented using the scikit-learn toolkit (Buitinck et al., 2013), using TF-IDF as a feature space for linear classification. The classifier's hyperparameters were optimized using a gridsearch algorithm.

4.2. Pre-Trained Model

The pre-trained BERT model used in these experiments is the *bert-base-swedish-cased* released by the National Library of Sweden (Malmsten et al., 2020)[2]. To provide a representative BERT model for the Swedish language, the model was trained on approximately 15-20 gigabyte of text (200M sentences, 3000M tokens) from a range of genres and text types including books, news, and internet forums. The model was trained with the same hyperparameters as first published by Google and corresponded to the size of Google's base version of BERT with 12 so-called transformer blocks (number of encoder layers), 768 hidden units, 12 attention heads and 110 million parameters.

A BERT model has a predefined vocabulary. This vocabulary is a set of words known to the model and it is used to tokenize words. A token can in this case be a common word, a common subpart of a word or a single letter. Each object in the vocabulary of the model has a known embedding. To use the model for finding the embedding of a new word the model was used to tokenize the word, which means that it would try to rebuild the word using as few tokens from the vocabulary as possible. The pre-trained BERT model used in this study had a vocabulary of 50325 words. Pre-trained model hyperparameters are listed in Table 1.

Hyperparemeter	Dimensions/Value
Dropout	0.1
Hidden Activation	GELU
Hidden Size	768
Embedding Size	512
Attentional Heads	12
Hidden Layers	12
Forward Size	3072
Vocabulary Size	50325
Trainable Parameters	$11 \cdot 10^7$

Table 1: Pre-training parameters

4.2.1. Fine-Tuning the Pre-Trained Model

In the first fine-tuning step, the decisions about how to set parameters were made partly based on the original BERT paper (Devlin et al., 2019), partly on previous findings based on electronic health records notes (Li et al., 2019), partly on the observation of our current data. Hyperparameters used for fine-tuning in this study are shown in Table 2. We relied on the Adam algorithm with default values for its hyperparameters as indicated by Kingma and Ba (2014). The pre-processed EMRs and the pre-trained model were fed into a Python script. For the first fine-tuning, the corpus was split into sentences. The model was fine-tuned with MLM (Masked Language Model), a technique which allows bidirectional training. MLM consists in replacing 15% of the words in each sequence with a [MASK] token before

[2] https://github.com/Kungbib/swedish-bert-models

Hyperparameter	Dimension/Value
Epochs	3
Batch Size	32
Block Size	64
Learning Rate	$5e-5$

Table 2: Parameters used for fine-tuning

feeding word sequences into BERT. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. The block size was set to 64, which means that sequences with fewer than 64 tokens are padded to meet this length, and sequences with more than 64 tokens are truncated. Actually, the value of 64 is generous since according to our current calculations the average sentence length in tokens is 12.

4.3. Further Fine-Tuning: Fully-Supervised BERT Classifier

The fully-supervised BERT classifier leverages on the fine-tuned model described above and it is further fine-tuned as a BERT classifier (Sun et al., 2019). Our implementations are based on *Hugging Face* libraries and public code.

The BERT classifier makes use of the prepended [CLS] token which acts as a representation of the sentence. The [CLS] token is fed into an activation function and based on the result, an optimizer further trains the BERT model which results in the [CLS] token being improved through repetition (training)[3]. The activation function used in the classification is arbitrary, but the default function used in Hugging Face's classification is a tanh (hyperbolic tangent) function layered between two linear layers[4], the first of which extracts the classification from the [CLS] token and the second layer outputs the representation for training.

The classifier was trained on the labelled data described in Section 3. The maximum allowed length (sequence length) for a BERT model is 512 tokens. In the experiments below we used both 512 tokens, but also 64 tokens to allow the comparison with GAN-BERT as implemented on our computer. Fully-supervised BERT was trained for 5 epochs with a batch size of 32 (see Table 3.

4.4. GAN-BERT

Croce et al. (2020) further extends BERT fine-tuning with a SS-GAN's perspective. In the SS-GAN's perspective (Salimans et al., 2016a), a discriminator is trained to distinguish fake examples artificially created by a generator from the real examples. In SS-GAN, the labelled examples are used to train the discriminator, while both the unlabelled data and the fake examples are used to improve data representation. Similarly, in GAN-BERT (Croce et al., 2020), a generator produces fake examples resembling the actual data distribution, while BERT is used as a discriminator. This enables semi-supervised learning since the generated samples are labelled by the SS-GAN model (Salimans et al., 2016b). Sagaciously, GAN-BERT exploits BERT's potential to produce accurate representations of input examples (in its role of discriminator) and leverages on the unleashed power of unlabeled data material (via the generator) to help in the generalization needed for the final task.

The implementation of the GAN-BERT model used in this paper is in *PyTorch*[5]. Our implementation was trained with 10 epochs, a batch size of 64 and a sequence length of 64 tokens (see Table 4), although BERT allows up to 512 tokens. The sequence length was set to 64 tokens because GAN-BERT is computationally expensive for our current computing resources (see Table 5).

Hyperparameter	Value
Epochs	5
Batch size	32
Sequence len	512 and 64

Table 3: Hyperparameters used for BERT classifier

Hyperparameter	Value
Epochs	10
Batch size	64
Sequence len	64

Table 4: Hyperparameters used for the GAN-BERT classifier

Label	Description
CPU	Intel Xeon - 12x(E5-2620 v3)
GPU	NVIDIA Quadro M4000 [8GB(VRAM)|20GB(Shared)]
Clock Speed	2.40GHz
Memory (RAM)	40GB

Table 5: Details of computing resources.

5. Experiments

Seven experiments were set up in order to understand the performance advantages and disadvantages of

[3] https://discuss.huggingface.co/t/what-is-the-purpose-of-the-additional-dense-layer-in-classification-heads/526

[4] https://github.com/huggingface/transformers/blob/09a2f40684f77e62d0fd8485fe9d2d610390453f/src/transformers/modeling_bert.py#L476

[5] https://awesomeopensource.com/project/crux82/ganbert-pytorch

using semi-supervised GAN-BERT rather than fully-supervised BERT. For fully-supervised BERT we used both 512 and 64 tokens, and for GAN-BERT only 64 tokens for the reasons explained in Section 4.4. All the experiments were gauged against the SVM baseline.

In **Experiment 1**, we compared the performance of BERT and GAN-BERT on a training set of 903 instances and a test set of 300 instances. In all models, training set and test set are the same. This experiment applies the traditional partition of the data of 70% for training and 30% for testing. GAN-BERT relies also on 1784 unlabelled instances.

In **Experiments 2, 3 and 4**, we compared the performance of BERT and GAN-BERT on a small but increasing size of the training set, namely 100, 200 and 300 instances, and a fixed test set of 300 instances. In the three models, GAN-BERT relies on 1784 unlabelled instances.

In **Experiments 5, 6 and 7**, we compared the performance of BERT and GAN-BERT on a small but increasing size of the training set, namely 100, 200 and 300 instances for training and a fixed test set of 300 instances for fully supervised BERT (the same as in previous experiments). Conversely for GAN-BERT we use large test sets of varying size. The challenge in this set of experiments is for GAN-BERT, since it is trained on small training sets but tested on large test sets. As in previous experiments, GAN-BERT relies on 1784 unlabelled instances. In this set of experiments we use four types of data: 1) labelled data for supervised training and testing; 2) de-labelled data for semi-supervision; 3) re-labelled data for testing; 4) unlabelled data for semi-supervision only.

By *de-labelled*, we refer to instances that have a label, but their label is ignored during the semi-supervised learning with GAN-BERT, so they get the status of unlabelled data. However, since the de-labelled data have a label, we use the re-labelled version (*re-lab* in Table 6) of the de-labelled data for testing. Essentially, the labels of the de-labelled data are not seen during the semi-supervised training. At testing time, we use the ignored labels to assess the performance of the model on the de-labelled instances. This validation technique is normally used for *external cluster validation*, and consists in comparing the results of a cluster analysis to externally provided class labels to measure the extent to which cluster labels match externally supplied class labels.

The rationale of this experimental setting is to understand to what extent unlabelled/de-labelled data contribute directly to the correct classification of classes of the test set. We know that unlabelled data is used by GAN-BERT to create data representation, but it is not clear so far if there exists a direct relation between the representation learned from unlabelled data and the final classes of the test set. This is why the use of de-labelled data could help us shed some light on this point.

6. Results

Table 6 presents the results of the 7 experiments. Best results are in bold and competitive results in italics.

We can observe that fully-supervised BERT achieves the best results when trained with 512 tokens in a traditional evaluation settings, i.e. 70% training vs 30% testing by reaching a weighted F1 of **0.97** (Exp 1). Fully-supervised BERT (512 tokens) performs well also when the training set is as large as the test set with a promising results of **0.93** (Exp. 4). Both BERT models outperforms the SVM baseline. However, fully-supervised BERT does not perform well when trained with 512 tokens on very small training sets, because it achieves *0.15* and *0.10* weighted F1 when trained on 100 and 200 instances respectively, and tested on 300 instances (Exp. 2 and Exp. 3). This is a poor performance in comparison with the SVM baseline of 0.76 and 0.75. For some reasons, the performance is slightly better when fully supervised BERT is trained and tested in the same conditions but with 64 tokens rather than 512. We speculate that it might be possible that tokens between 65 to 512 are not representative of the data, and therefore misguide the model into false global contextuality. This in turn could hinder its performance during the classification task. This speculation however must be tested in future experiments.

GAN-BERT largely outperforms fully-supervised BERT when trained on small training sets (100 and 200 instances) and tested on 300 instances using 64 tokens (experiments 2, 3 and 4). It achieves an exciting **0.67** (vs. 0.31) and **0.84** (vs. 0.33) in experiments 2 and 3. When trained on 300 instances (Exp.4), fully supervised BERT'(64 token) soars to 0.85, while the increase of GAN-BERT's performance (i.e. 0.87) is small when compared with that achieved in Exp. 3, but it is still competitive. Only in Exp. 2, all BERT-based models underperform the SVM baseline.

In Experiments 5, 6 and 7, no tangible benefits can be observed when semi-supervised GAN-BERT models are trained with de-labelled data and tested on re-labelled data. In all the three training configurations, i.e. 100, 200 and 300 training instances, performance is low, reaching respectively a weighted F1 of 0.24, 0.26 and 0.29, a score much lower than the SVM baseline. When semi-supervised GAN-BERT models are trained with unlabelled data (1784 instances) and tested on large test sets, the performance goes slightly up to 0.34, 0.37 and 0.42.

Interestingly, when semi-supervised GAN-BERT models are trained with both unlabelled and de-labelled data, results are definitely more encouraging, and reach a weighted F1 of 0.44 (with 100 training instances and 1103 testing instances), 0.55 (with 200 training instances and 1003 testing instances), and **0.80** (with 300 training instances and 903 testing instances). It is intriguing to observe that performance of 0.80 (higher than the SVM baseline) is achieved in very difficult conditions, where the proportions of training set (i.e. 30% training instances) and test set (70% test instances) are the inverse of the proportions used in Exp.1.

7. Discussion

The results presented in the previous section provide answers to our research question, i.e. whether and to what extent it is possible to achieve a competitive performance with a semi-supervised GAN-BERT model based only on a small amount of annotated EMRs written in Swedish for the classification of patients bearing implant(s). The answer is: yes, it is possible to achieve competitive performances using GAN-BERT in some of the scenarios that we have set up in our experiments. The most promising results of GAN-BERT's potential are returned by experiments 3 and 4 on a fixed test set of 300 instances. However, we find much more inspiring the results achieved in experiment 7 where a small training set of 300 instances can achieve an astonishing performance of 0.80 when tested on a test set that is 3 times larger than the training set. This means in our opinion that the semi-supervised approach is profitable and efficient in the difficult scenarios of real-world tasks, where it is just impossible to label all the data, especially if domain-specific.

GAN-BERT was able to achieve the highest score in Exp. 3, and competitive results in Exp. 4. The fully-supervised SVM baseline scores vary very little across the 7 experiments, while fully-supervised BERT models vary from as low as 0.10 (Exp. 3) and as high has 0.97 (Exp. 1), depending on the amount of labelled training instances.

It is also to be noticed that experiments based on de-labelled instances only (i.e without the 1784 unlabelled instances) and tested on re-labelled instances are extremely informative. They show that GAN-BERT is not biassed by having seen the de-labelled data during the semi-supervision phase: results on the re-labelled test sets were so bad that we cannot surmize that the de-labelled data have somehow unexpectedly positively affected the performance. This interpretation is confirmed by the results achieved with unlabelled data: it seems that it is the sheer size of the data that makes the trick with semi-supervised GAN-BERT, not the labels. This independence between the semi-supervision phase and the test phase is evident in experiment 7 where the performance with de-labelled data only (0.29) and with unlabelled only (0.42) is much lower than the performance of 0.80 with de-labelled and unlabelled data added together. This is, we reckon, the most revealing finding of our set of experiments because it indicates that we can use as much de-labelled and unlabelled data as possible during semi-supervision without fearing that the actual performance is biassed or tweaked, because it is the amount of instances that matters, nothing else.

We can then conclude: 1) that GAN-BERT is suitable for a real-world downstream task, such as the classification of implant-bearing implants, 2) that it can profitably be applied to the Swedish language, and finally 3) that it can make sense of a difficult genre, like EMRs.

8. Conclusion and Future Work

In this paper, we investigated whether and to what extent BERT-based semi-supervised text classification is viable in real-world settings, where no standard benchmarks are available for evaluation.

Results are promising and informative. It is indeed possible to create BERT-based semi-supervised classification models based on small training sets that capitalize on unlabelled and de-labelled data.

However, several challenges lie ahead for the improvement of the approach. The first challenge is to find the ideal size of the training set for a specific task: would it be possible to find an automatic way to determine the size of the best performing training set for a specific downstream task thus overriding tedious empirical tries with training sets of different sizes? Obviously, learning curves can help, but they do not tell the whole story since they are based on a single training set.

Another challenge is to overcome the token limitation (our experiments are based either on 512 tokens or 64 tokens). This restriction is unrealistic for the text types we want to classify, i.e. medical histories whose length varies from a few words up to thousands of words. Several solutions have been proposed to safely apply BERT to long texts – e.g. (Devlin, 2018; Fiok et al., 2021; Adhikari et al., 2019) – and we will investigate them all.

Many of the algorithms that have been proposed for state-of-the-art NLP are computationally and environmentally expensive. Thus, a further challenge is for us to find the trade-off between the CO2 impact and the hardware limitations with respect to the most advanced NLP solutions, that are invariably too computationally and environmentally demanding for real-world empirical settings, such a public hospitals.

9. Acknowledgements

This research was funded by **Vinnova** (Sweden's innovation agency), https://www.vinnova.se/
Project title: *Patient-Safe Magnetic Resonance Imaging Examination by AI-based Medical Screening.* **Grant number**: 2021-01699 to Peter Lundberg.

10. Bibliographical References

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Docbert: Bert for document classification.

Breazzano, C., Croce, D., and Basili, R. (2021). MT-GAN-BERT: multi-task and generative adversarial learning for sustainable language processing. In Elena Cabrio, et al., editors, *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event,*

		Labelled training	Labelled test set	Unlabelled	De-labelled	Weighted F1		
						512	64	SVM
1	SVM (baseline)	903	300	n/a	n/a	n/a	n/a	0.78
	BERT	903	300	n/a	n/a	**0.97**	*0.91*	n/a
	GAN-BERT	903	300	1784	n/a	n/a	*0.90*	n/a
Testing both BERT and GAN-BERT on 300 examples								
2	SVM (baseline)	100	300	n/a	n/a	n/a	n/a	0.76
	BERT	100	300	n/a	n/a	0.15	0.31	n/a
	GAN-BERT	100	300	1784	n/a	n/a	0.67	n/a
3	SVM (baseline)	200	300	n/a	n/a	n/a	n/a	0.75
	BERT	200	300	n/a	n/a	0.10	0.33	n/a
	GAN-BERT	200	300	1784	n/a	n/a	**0.84**	n/a
4	SVM (baseline)	300	300	n/a	n/a	n/a	n/a	0.77
	BERT	300	300	n/a	n/a	**0.93**	*0.85*	n/a
	GAN-BERT	300	300	1784	n/a	n/a	*0.87*	n/a
Testing BERT on 300 examples and GAN-BERT on re-labelled data								
5	*SVM (same as 2)*	*100*	*300*	*n/a*	*n/a*	*n/a*	*n/a*	0.76
	BERT (same as 2)	*100*	*300*	*n/a*	*n/a*	0.15	0.31	n/a
	GAN-BERT	100	1103 (re-lab)	n/a	1103	n/a	0.24	n/a
	GAN-BERT	100	1103 (re-lab)	1784	n/a	n/a	0.34	n/a
	GAN-BERT	100	1103 (re-lab)	1784	1103	n/a	0.44	n/a
6	*SVM (same as 3)*	*200*	*300*	*n/a*	*n/a*	*n/a*	*n/a*	0.75
	BERT (same as 3)	*200*	*300*	*n/a*	*n/a*	0.10	0.33	*n/a*
	GAN-BERT	200	1003 (re-lab)	n/a	1003	n/a	0.26	n/a
	GAN-BERT	200	1003 (re-lab)	1784	n/a	n/a	0.37	n/a
	GAN-BERT	200	1003 (re-lab)	1784	1003	n/a	0.55	n/a
7	*SVM (same as 4)*	*300*	*300*	*n/a*	*n/a*	*n/a*	*n/a*	0.77
	BERT (same as 4)	*300*	*300*	*n/a*	*n/a*	0.93	0.85	*n/a*
	GAN-BERT	300	903 (re-lab)	n/a	903	n/a	0.29	n/a
	GAN-BERT	300	903 (re-lab)	1784	n/a	n/a	0.42	n/a
	GAN-BERT	300	903 (re-lab)	1784	903	n/a	*0.80*	n/a

Table 6: Results: Comparing the performance

November 29, 2021, volume 3015 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Croce, D., Castellucci, G., and Basili, R. (2020). Ganbert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Devlin, J. (2018). Github issue comment. `https://github.com/google-research/bert/issues/27`.

Edwards, A., Camacho-Collados, J., De Ribaupierre, H., and Preece, A. (2020). Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529.

Fiok, K., Karwowski, W., Gutiérrez, E., Davahli, M. R., Wilamowski, M., Ahram, T. Z., Aljuaid, A. M., and Zurada, J. M. (2021). Text guide: Improving the quality of long text classification by a text selection method based on feature importance. *CoRR*, abs/2104.07225.

González-Carvajal, S. and Garrido-Merchán, E. C. (2020). Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Jerdhaf, O., Santini, M., Lundberg, P., Karlsson, A., and Jönsson, A. (2021). Implant term extraction from swedish medical records–phase 1: Lessons learned. In *Swedish Language Technology Conference and NLP4CALL*, pages 35–49.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Kihlberg, J. and Lundberg, P. (2019). Improved workflow with implants gave more satisfied staff. In *SMRT 28th Annual Meeting 10-13 May 2019*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*, pages arXiv–1412.

Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., and Yu, H. (2019). Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: An empirical study. *JMIR medical informatics*, 7(3):e14830.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden – making a swedish bert.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Owen, L. (2020). semi-supervised intent classification with gan-bert.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016a). Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016b). Improved techniques for training gans.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Sun, T., Liu, X., Qiu, X., and Huang, X. (2021). Paradigm shift in natural language processing. *arXiv preprint arXiv:2109.12575*.

Zaharia, G.-E., Avram, A.-M., Cercel, D.-C., and Rebedea, T. (2021). Dialect identification through adversarial learning and knowledge distillation on romanian bert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 113–119.