Semi-Supervised Generative Adversarial Network for Sentiment Analysis of drug reviews

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

24-11-2021 / 08-12-2021

CITATION

Colón-Ruiz, Cristóbal (2021): Semi-Supervised Generative Adversarial Network for Sentiment Analysis of drug reviews. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17075054.v1>

DOI

[10.36227/techrxiv.17075054.v1](https://doi.org/10.36227/techrxiv.17075054.v1)

Semi-Supervised Generative Adversarial Network for Sentiment Analysis of drug reviews

Cristóbal Colón-Ruiz, Isabel Segura-Bedmar

Abstract—Sentiment analysis has become a very popular research topic and covers a wide range of domains such as economy, politics and health. In the pharmaceutical field, automated analysis of online user reviews provides information on the effectiveness and potential side effects of drugs, which could be used to improve pharmacovigilance systems. Deep learning approaches have revolutionized the field of Natural Language Processing (NLP), achieving state-of-the-art results in many tasks, such as sentiment analysis. These methods require large annotated datasets to train their models. However, in most real-world scenarios, obtaining high-quality labeled datasets is an expensive and time-consuming task. In contrast, unlabeled texts task can be, generally, easily obtained. In this work, we propose a semi-supervised approach based on a Semi-Supervised Generative Adversarial Network (SSGAN) to address the lack of labeled data for the sentiment analysis of drug reviews, and improve the results provided by supervised approaches in this task. To evaluate the real contribution of this approach, we present a benchmark comparison between our semi-supervised approach and a supervised approach, which uses a similar architecture but without the generative adversarial setting. Experimental results show better performance of the semi-supervised approach when annotated reviews are less than ten percent of the training set, obtaining a significant improvement for the classification of neutral reviews, the class with least examples. To the best of our knowledge, this is the first study that applies a SSGAN to the sentiment classification of drug reviews. Our semi-supervised approach provides promising results for dealing with the shortage of annotated dataset, but there is still much room to improvement.

Index Terms—Deep learning, drug reviews, generative adversarial network, machine learning, semi-supervised, sentiment analysis

I. INTRODUCTION

Since the beginning of the century, millions and millions of users' reviews are available on the Internet. In fact, online users are generating a huge amount of reviews on fresh topics every second, making it more difficult to massively label this amount of data in a short time [1].

As a result, sentiment analysis has become one of the most fruitful research fields in Natural Language Processing (NLP). This task covers a wide range of domains and could contribute to a variety of applications like financial forecasting [2], political strategy planning [3], public opinion analyzing [4], online decision making [5], and pharmacovigilance [6], among many others.

Pharmacovigilance, also known as drug safety, tries to prevent drug side effects. The early detection of side effects depends on clinical

This work was supported by the Research Program of the Ministry of Science and Innovation - Government of Spain (ACCESS2MEET project-PID2020-116527RB-I00), and by the Madrid Government (Comunidad de Madrid) under the Multiannual Agreement with UC3M in the line of "Fostering Young Doctors Research" (NLP4RARE-CM-UC3M) and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation; the Multiannual Agreement with UC3M in the line of "Excellence of University Professors (EPUC3M17)";

C. Colón-Ruiz is with the Computer Science Department, Universidad Carlos III de Madrid, Leganés, 28911, Spain (e-mail: ccolon@inf.uc3m.es).

I. Segura-Bedmar is with the Computer Science Department, Universidad Carlos III de Madrid, Leganés, 28911, Spain (e-mail: isegura@inf.uc3m.es).

trials and specific testing protocols, which are usually conducted under given conditions with a limited number of test subjects and time. As a consequence of the design of clinical trials, discrepancies in patient selection and treatment conditions can have a significant impact on drug efficacy and detection of potential adverse effects [7]. Additional information sources, such as user reviews, can offer great potential in this regard [8]. Automatic detection of negative user reviews could alert to unexpected drug behaviors and obtain relevant information to improve pharmacovigilance systems [9]. Moreover, the identification of positive user reviews could help to assess the drug efficacy and even discover new uses for a drug.

Approaches in sentiment analysis have evolved from deterministic rule-based systems [10] to machine-learning [11] and hybrid methods [12]. Over the last years, the emergence of deep learning techniques have brought remarkable breakthroughs in NLP, providing state-of-the-art results in many tasks [13]. Sentiment analysis is not oblivious to this success, and a wide range of deep learning-based systems have also achieved state-of-the-art results on standard sentiment analysis datasets [14], [15]. However, patient sentiment analysis, specifically of their experience with drugs, has received less research attention [16]–[18]. This is a challenging task because user opinions can have different degrees of subjectivity and could be focused on other different aspects (such as the quality of healthcare service or healthcare professionals), introducing noise in the classification task [19]. Moreover, the shortage of annotated data, which is essential for training machine learning algorithms, has become one of the greatest bottlenecks in NLP [20], [21], and, in particular also for the task of sentiment analysis [22].

To overcome the lack of labeled examples, the use of unsupervised and semi-supervised approaches have been applied to sentiment analysis [23]–[25]. However, these approaches do not outperform the traditional supervised machine learning methods.

More recently, transfer learning and cross-domain approaches have been applied to mitigate the lack of labeled data [17], [26]–[28]. Transfer learning is a machine learning approach where a first model, which is trained for a specific task, is then reused as input for training a new model for another task. A successful example of transfer learning in NLP is Bidirectional Encoder Representations from Transformers (BERT) [29], whose main advantage is that it can accurately represent the different meanings of a word. Then, a fine-tuned phase is performed for a specific task, such as text classification, text summarization or named entity recognition, among many others. These BERT-based models are achieving the current state-of-the-art results for many NLP tasks [29], [30]. However, the fine-tuning phase still requires thousands of annotated examples for the target task. Therefore, the quality of the results drops significantly when the number of labeled examples is insufficient [31].

To address the challenge caused by the shortage of labeled drug reviews, we propose to employ a semi-supervised method by implementing a Semi-Supervised Generative Adversarial Network (SSGAN) [32]. Generally, in Generative Adversarial Networks (GAN) [33], a generator is trained to produce synthetic examples from different data distributions. During training, the generator tries to mislead a discriminator, which is trained to distinguish between synthetic and real examples.

To the best of our knowledge, only a previous work [31] has used SSGAN for the task of sentiment analysis. Therefore, our study is the first work in applying SSGAN to the sentiment classification of drug reviews. To assess the real contribution of our semi-supervised approach, we also performed a comparison between the proposed semi-supervised approach and the supervised approach described in [6], which is based on the BERT-mini model [34] fine-tuned with a Bidirectional Long short-term memory (BiLSTM) layer. In this way, our approach exploits a more complex fine-tuning architecture for the task of sentiment classification than that proposed in [31], since they only used a softmax layer.

We use different splits of the training dataset, which are intended to show the performance of the semi-supervised compared to the supervised approach trained only with a small percentage of training dataset. Thus, these splits will allow us to show the amount of data should be annotated to obtain similar results than those obtained by the supervised approach.

The organization of this paper is as follows. After discussing prior work (Section II), we present the dataset used and describe the deep learning architectures studied in this work (Section III). Then, we evaluate the models and discuss their results (Section IV). Finally, we provide some conclusions extracted from the experimentation and present future research lines (Section V).

II. RELATED WORK

Although sentiment analysis has been widely applied to many application domains [2]–[6], [35], [36], the pharmaceutical domain is currently starting to receive research attention. In this section, we review previous research focused on sentiment analysis of drug reviews.

Early work on this task mainly used sentiment rules and lexicons (such as SentiWordNet [37]) [38], [39]. Later works such as [40]–[43] proposed approaches with Bag-of-Words (BoW) model to represent the input texts and different machine learning algorithms such as Decision Trees, K-Nearest-Neighbor (KNN), Support Vector Machine (SVM) or Naïve Bayes to estimate the polarity of patient posts in online health forums. One of the most popular studies is the work of [17] because the authors have created a labeled dataset of drug reviews, which has become a standard benchmarking corpus for the task. This dataset is described in more detail in Subsection III-A. The authors used a logistic regression to classify the drug reviews, achieving an accuracy of 92.2%.

More recently, deep learning approaches have shown remarkable results [18], [44], [45]. C. Colon-Ruiz and I. Segura-Bedmar [6] performed a benchmark comparison between different combinations of deep learning models such as Convolutional Neural Network (CNN), Bidirectional Long short-term memory (BiLSTM), and BERT (BERT-Mini [34]). Their experiments show that the BERT model obtained the best results with an accuracy of 90.4% on the dataset created by Grasser et al. Another study, using the same dataset, was described in [46], which explored both LSTM and BiLSTM layers followed by a self-attention mechanism, obtaining an accuracy of 92%.

Biseda and Mo [18] compared a variety of eight different BERT models for sentiment classification of drug reviews on the dataset proposed in [17]. Their experiments showed an accuracy of 90.6%. Recently, several neural network models with Embeddings from Language Model (ELMo) [47] and different transformers such as BERT, BERT variants, and XLNET have been explored in [45]. The best result, an accuracy of 94.3%, was obtained by a BERT architecture modified by replacing some multi-head attention layers with dynamic convolutions to reduce the temporal complexity.

Semi-supervised approaches could improve the efficiency of classifying text without the need for large labeled datasets. Wu et al. [48] implemented a semi-supervised approach with a variational autoencoder and recurrent LSTM networks to improve dimensional sentiment analysis (DSA) performance with considerably less labeled data. The authors evaluated their approach on several benchmark datasets, specifically, on a dataset consisting of status updates shared by Facebook users [49]. These texts were scored on a scale of 1 (most negative) to 9 (most positive). The semi-supervised approach was compared with a supervised LSTM model. The experiments showed that the semi-supervised approach overcame the supervised one with an improvement of approximately 15.9% of Pearson correlation, when only 10% of the training texts were used to train the models.

Croce et al. [31] proposed a supervised approach based on BERT and a semi-supervised Generative Adversarial Network to perform different NLP tasks using small training datasets. Between the different NLP tasks, a sentiment analysis task was performed on the Stanford Sentiment Treebank-5 (SST-5) dataset [50], which consists of movie reviews classified as negative, somewhat negative, neutral, somewhat positive, or positive. For this task, the BERT model was fine-tuned with a softmax layer to perform the classification. The semi-supervised approach achieved an improvement of approximately 10% accuracy compared with the supervised BERT-based approach, when only 1% of the reviews are labeled.

From the present review, we can say that semi-supervised methods based on deep learning architectures provide promising results, although they have been timidly applied to sentiment analysis. To the best of our knowledge, these semi-supervised approaches have not been used for sentiment classification of drug reviews.

III. METHODS

A. Dataset

For this work, we used the dataset of drug reviews created by [17]. In this dataset, the drug reviews, written by patients, were collected from Drugs.com and Druglib.com, two of the most visited pharmaceutical information websites offering information to both consumers and health care professionals. Each drug review includes a rating from 0 to 9, indicating the degree of patient satisfaction with the drug. For example, a patient with *epilepsy* posted the following comment: "I've had nothing but problems with the Keppera : constant shaking in my arms & legs, pins needles feeling in my arms & legs, severe light headedness, no appetite, etc.". The patient refers to drug *Keppera* and describes the adverse effects it has caused. The rating for the drug provided by the patient is negative with a value of 2. The reviews were grouped into three levels of polarity according to their rating: positive ($rating \geq 7$), negative ($rating \leq 4$) and neutral (rating [3,6]), such was proposed in [17]. The dataset, which contains a total of 215,063 drug reviews, is split into training and test sets with a ratio of 75:25 with stratified random sampling to maintain the same proportion of all classes. Additionally, as in [6], 15% of the total training set is used as the validation set, which is used to fit the hyperparameters of the models.

Figure 1 shows that the distribution of the different reviews according to their classes (degree of satisfaction) is strongly unbalanced in the training and test sets. Reviews with positive polarity (class 2) represent about 66% of the data, reviews with negative polarity (class 0) represent about 25% and, finally, reviews with neutral polarity (class 1) represent only 9%.

In order to prepare the training dataset for our comparison of the semi-supervised approach with the supervised one, we performed a series of splits (see Table I) to obtain several ratios between labeled and unlabeled reviews. By employing different ratios between labeled

Fig. 1: Class distribution in training and test datasets

TABLE I: Ratios between labeled and unlabeled reviews. For example, the ratio 1:99 indicates that only 1% of the dataset is labeled.

Split ratio	Number of labeled reviews			
	Total	Negative (0)	Neutral (1)	Positive (2)
1:99	1,371	341	121	909
2:98	2,742	682	243	1817
5:95	6,855	1,705	607	4,543
10:90	13,710	3,410	1,213	9,087
20:80	27,420	6,820	2,427	18,173

and unlabeled data, we will be able to observe how much annotated data we need to use with each approach to obtain similar results.

For example, in the split "20:80", only 20% of the training dataset is randomly selected to be used as labeled reviews, while the remaining 80% will be used as unlabeled reviews. In this particular split of 20%, there are 27,420 labeled reviews, where 6,820 belong to class 0 (negative), 2,427 belong to class 1 (neutral) and 18,173 belong to class 2 (positive). The splits are created using a stratified random sampling to keep the same class distribution in both partitions.

The semi-supervised approach will use labeled examples for training the classification model, but also the unlabeled examples to improve the distinction between automatically generated examples and real ones. On the other hand, the supervised approach will only use the labeled reviews to train its model.

B. Supervised approach: BERT-BiLSTM

To assess the real contribution of our semi-supervised approach, we also propose a supervised approach that employs a BERT model followed by a BiLSTM to perform the fine-tuning for the sentiment analysis task of drug reviews.

BERT provides a representation of their inputs as a result of a pre-training phase on large-scale corpora. These representations are able to provide a vector for each sense of a word taking into account the context where the word occurs. In this way, it can represent the different meanings of a word depending on the context.

To format the drug reviews in a BERT-based format, the texts were processed using the tokenizer provided by bert-for-tf2¹ and the vocabulary of the pre-trained BERT-Mini model. In addition, due to the difference in text lengths, it was necessary to pad and truncate the reviews based on the cumulative distribution function (Figure 2). This figure shows that almost 100% of the reviews have a length less than 250 tokens. Thus, the length of processed texts was set to 250 tokens.

¹<https://pypi.org/project/bert-for-tf2/>

As was mentioned above, one of the main advantages of BERT is its ability to provide contextualized word embeddings with a more accurate representation of the words. However, training BERT for a specific task is expensive from a parameter point of view [51]. Due to the computational cost, we used a pre-trained BERT model with only four encoder layers, concretely, the BERT-Mini model [34]. Moreover, we use adaptor modules [51] as a transfer learning mechanism that can adjust a pre-trained model for new tasks with no need to modify the weights of the original model. The adapter tuning strategy consists of injecting new layers into the original network forming a bottleneck architecture that projects the input features in lower dimensions, applies non-linearity and returns the input to the original dimension. The adapter module also incorporates an internal skip-connection to avoid problems with near-zero initializations. As a result, the original BERT model weights can remain unchanged to considerably reduce the number of trainable parameters during fine-tuning of the specified task.

BERT-Mini requires much less memory than the original model, enabling better training times, at the expense of worse results for the final task [34]. To address this problem, we add a BiLSTM layer, which has been shown to provide good results in sentiment analysis tasks [6]. Moreover, after this layer, we also added a fully connected perceptron layer with leaky rectified linear unit activation function (leaky ReLU) [52] to improve class prediction. As the last layer, we use a softmax layer, which outputs a vector containing probabilities for each of the classes (positive, neutral, negative).

Finally, for model training we used 200 epochs with the ADAM optimizer (learning rate of 0.001), a batch size of 200 and categorical cross-entropy as a loss function. The training parameters were selected empirically. The reader can find a detailed description of this approach in [6].

C. Semi-Supervised approach: GAN with BERT-BiLSTM

Our semi-supervised approach uses the same architecture as our supervised one, but also exploits an adversarial network to improve the internal representation of the examples for the classification task. In a GAN architecture [33], two models are trained: a generative model G and a discriminative model D . G learns the distribution of the data and generates synthetic examples. D predicts the probability that an example is synthetic (generated by G) or real (from the training dataset). The training of G consists of maximizing the probability that D makes a mistake. This scenario corresponds to a minimax game of two players [53], G and D , competing against each other. The generative model is confronted with an adversary, a discriminative model that learns to differentiate whether an example comes from G or belongs to the training data. This competition leads both models to improve their performances until the distributions generated by G are practically indistinguishable from those belonging to the real training data.

In this way, we use a semi-supervised GAN (SSGAN) [32] that adversarially improves the distributions generated by G and D , and thereby, provides a better representations for the classification task. To achieve this, the generator must produce a batch of new synthetic examples to try to mislead the discriminator. The generator uses as input (x_g) a uniform noise distribution and a multilayer perceptron to learn the synthetic examples. Then, the discriminator, which is also a multilayer perceptron, is trained on c classes (in our case, three classes: positive, negative and neutral) plus an additional binary class ($c+1$) to identify whether the input example comes from the generator or the real training data. The last layer of D is a softmax layer for the ($c+1$) classes. Figure 3 shows the whole architecture of our final approach (SSGAN-BL), which consists of the SSGAN model, whose

Fig. 2: Probability density function and Cumulative distribution function over the length of reviews

goal is to improve the internal representations of the examples, and the BERT-BiLSTM model, which learns the representation of the drug reviews according to the task of sentiment classification. Both models are trained concurrently.

In a first stage, while G and D try to adversarially improve the internal representations of both models, D take as input the output from G ($h_{fake} \in R^d$), but also the output from the BERT-BiLSTM model ($h_{real} \in R^d$). Ideally, D should distinguish synthetic examples from real examples. In other words, D should classify the synthetic examples coming from G with the binary class ($c+1$), but not the real ones. Moreover, as the training dataset was split into a percentage of labeled examples and the remaining ones to be considered as unlabeled, D should classify the labeled samples with one of the three classes (positive, negative or neutral).

During the training of the SSGAN-BL model, the weights of G and D are updated, but also the ones of the BERT-BiLSTM model. Once the SSGAN-BL model is trained, G is discarded, keeping the rest of the model for the sentiment classification task with the three classes (positive, negative and neutral) and ignoring the binary class ($c+1$) in the softmax layer of D .

We now describe in detail the loss functions used to train the models. From now on, x_b is a vector of a original review (input for BERT-LSTM) and x_g is a vector generated from a uniform random distribution (input for G). Once this is defined, we can express the probability P , provided by the discriminator D , as the probability of a given input, x_b or x_g , to be associated with one of the ($1, \dots, c+1$) classes. Using the probability distributions P , we can measure the error of each of the models. To train D , we must minimize the sum of errors produced by mis-classifying labeled reviews with wrong classes in ($1, \dots, c$), by misidentifying labeled and unlabeled reviews as "fake" reviews (that is, with the ($c+1$) class) and by misidentifying synthetic examples generated by G as "real" reviews. On other other hand, to train G , we must minimize the sum of errors produced by D in identifying an example generated by G as a synthetic one and by G in generating examples with distributions significantly different from the real reviews.

Therefore, the loss function of D is defined as $L_{D_{sup}} + L_{D_{uns}}$, where (1) provides the error in assigning a labeled review x_b in the original c -classes, (2) defines the error of recognizing a labeled or unlabeled review x_b as a synthetic example ($c+1$), and (3) defines the error of recognizing a synthetic example x_g as a real example ($c+1$). Finally, the error is back-propagated and the weights are updated for the D and BERT-BiLSTM models (adjusting the inner representations of the reviews).

In the meantime, the loss function of G is defined as $L_{G_{uns}} + L_{G_{match}}$. G must be able to generate synthatic examples similar

to the representation of the real ones. For this purpose, the average of the examples generated by G , in a batch, must be similar to the average of the real examples' representation. In (5), we define $f(x)$ as the activation on an intermediate layer of D , where x represents either the input data of BERT-BiLSTM or the input data of G . The loss function of G , in (4), also measures the error of recognizing a synthetic example x_g as a synthetic example, as G tries to mislead D with examples similar to the real ones.

$$L_{D_{sup}} = - \sum_{i=1}^c y_i \log[P(\hat{y} = y | x_b, y \in (1, \dots, c))_i] \quad (1)$$

$$L_{D_{uns}} = -\log[1 - P(\hat{y} = y | x_b, y = c + 1)] \quad (2)$$

$$- \log[P(\hat{y} = y | x_g, y = c + 1)] \quad (3)$$

$$L_{G_{uns}} = -\log[1 - P(\hat{y} = y | x_g, y = c + 1)] \quad (4)$$

$$L_{G_{match}} = \frac{1}{n} \sum_{i=1}^n |f(x_b)_i - f(x_g)_i|^2 \quad (5)$$

For training, we used 400 epochs with the ADAM optimizer (learning ratio of $4e-5$). The batch sizes can differ (64-200) depending on the number of labeled reviews used for training. This is due to the need to preserve batches with labeled reviews representation to avoid divergence produced by unsupervised components in adversarial training. On the other hand, for both G and D , we used Leaky Relu activation functions except in the last layer of the generator, where we used a hyperbolic tangent function ($\tanh$) as the activation function. The latter is to ensure that the h_{fake} representations produced by the generator are similar to the h_{real} representations provided by BERT-BiLSTM, where the last layer is composed of a BiLSTM recurrent network with hyperbolic tangent activation functions.

The pseudo-code of the training algorithm can be found in Algorithm 1.

Our source code is publicly available to enable the reproducibility of our experiments².

IV. EXPERIMENTAL RESULTS

In this section, we focus on the results obtained with our approaches on the test dataset described in Section III.

To evaluate both approaches, we used one of the standardized metrics for text classification tasks: F1. This metric can be extended for multi-classification problems with the micro-averaged and macro-averaged versions. With micro-averaged F1, we show the contributions of each of the classes, while with macro-averaged F1, we show the mean of the values obtained by each class independently.

²<https://github.com/ccolonruiz/SSGAN-DrugSentimentAnalysis/>

Fig. 3: Overview architecture of SSGAN-BL

Algorithm 1 Pseudo-code of SSGAN-BL

Inputs: epochs, batch size n , annotated ratio nr ($L|U$)

for $epoch$ in $epochs$ **do**

for $batch$ in $epoch$ **do**

$D \rightarrow trainable = True$

 Sample $l_1, \dots, l_{nr \cdot n}, u_1, \dots, u_{(1-nr) \cdot n}$ from $x_{b_1}, \dots, x_{b_n}$

 Sample $z_1, \dots, z_n$ from $N(\mu, \sigma^2)$

while $D \leftarrow train$ **do** $\triangleright B$ is also updated

$L_{D_{sup}} \leftarrow D(B(l))$

$L_{D_{uns}} \leftarrow D(B(l+u))$ and $D(G(z))$

$L_{D_{total}} = L_{D_{sup}} + L_{D_{uns}}$

end while

$D \rightarrow trainable = False$

while $G \leftarrow train$ **do**

$L_{G_{uns}} \leftarrow D(G(z))$

$L_{G_{match}} \leftarrow f(l+u)$ and $f(z)$

$L_{G_{total}} = L_{G_{uns}} + L_{G_{match}}$

end while

end for

end for

Figures 4 and 5 show the performance of the supervised and semi-supervised approaches using different ratios of labeled (Annotated %) and unlabeled reviews. In this way, we will see the difference in the results provided by the semi-supervised and supervised approaches when only limited subsets of labeled reviews are available.

Figures 4 compare both approaches when different percentages of labeled examples are used to train both models. We can observe that, with less than 10% of labeled examples, the semi-supervised approach provides better performance in terms of micro-F1 than the supervised approach (see Fig. 4b). Similarly, the semi-supervised approach overcomes the supervised one in terms of macro-F1, when we only use 20% of the trainin examples (see Fig. 4b). Therefore, when we use only 10% or 20% of the labeled examples to train the models, the semi-supervised approach starts to obtain worse results than the supervised one. This may be due to the fact that, from 10% onwards, the supervised approach has enough data to obtain good enough internal representations of the reviews on its own compared to the semi-supervised approach.

Moreover, we can observe in Figure 5b, the semi-supervised model obtains better results than the supervised model for the class with least instances (neutral class), regardless of the percentage of labeled data used. Thus, the model benefits more from the data representations adversarially enhanced by the SSGAN-BL model.

Fig. 4: Macro-F1 and Micro-F1 curves of BERT-LSTM and SSGAN-BL classifiers

Fig. 5: F1 curves of BERT-LSTM and SSGAN-BL classifiers for each class labels

The difference in the performance of the two models for the underrepresented class can be seen in Figure 4a. The Macro-F1 metric, described above, demonstrates a greater differentiation in the performance of both approaches regardless of the percentage of labeled data. Considering the performance of SSGAN-BL in class 1 (Neutral class), the mean of the values obtained for each class individually is higher than in the case of BERT-BiLSTM. However, Figure 4b shows that there is no statistically significant difference in terms of Micro-F1 (below 10% of labeled data). This may be due to a slight increase of failures in SSGAN-BL when classifying instances of the other classes (positive or negative) at the expense of improving performance with the under-represented class (neutral). As it can be seen in the confusion matrices (see Fig. 6), when we compare both approaches, there is a similar pattern of behavior. We can see how the number of reviews correctly classified as class 1 (neutral) is higher with SSGAN-BL (see Fig. 6f) than with BERT-LSTM (see Fig. 6c). However, the number of reviews correctly classified as class 0 (negative) is slightly lower with SSGAN-BL. This is repeated for each of the comparisons up to Figures 6g and 6i onwards. From this setup, there is also a decrease in the number of reviews correctly classified as class 2 (positive) by the SSGAN-BL model. The latter fact can be seen in Figures 4 and 5 from the 10% of labeled data.

Moreover, we can see a pattern of behavior in the confusion matrices that are not reflected by the F1 metrics. When we observe Figures 6b and 6e, we can notice that the SSGAN-BL model classifies reviews less often in the opposite polarity than BERT-BiLSTM. In contrast, BERT-BiLSTM is more inclined to classify a positive example as a negative one or a negative example as a positive one. It can be argued that misclassifying a polarized review as neutral is

preferable to misclassifying such a review in the opposite polarity (e.g., positive, as negative). For example, the following review is labeled in class 0 (negative): "Have been on Actos for almost a year, gained 24 pounds and have swelling in hands and feet and are retaining a lot of water in my thighs. My sugar levels are good. My doctor lowered my dosage from 30 mg to 15 mg but refused to take me off. Will get a second opinion because the side effects are too much.". In this case, the BERT-LSTM model classifies it as positive, while the SSGAN-BL model classifies it as neutral.

V. CONCLUSION

In recent years, the emergence of transformer-based approaches has revolutionized the field of NLP. Models such as BERT, have achieved state-of-the-art results for many NLP tasks. However, to fine-tune these models to specific tasks, it is still required a large amount of annotated corpora that are not always available.

In this paper, we propose a semi-supervised adversarial network (SSGAN) approach capable of dealing with the sentiment analysis of drug reviews when the size of the labeled dataset is limited. To evaluate the real contribution of our approach, we performed a comparison between the semi-supervised approach and a supervised one. Our experiment results show that the semi-supervised approach provides better results than the supervised approach in terms of Macro-F1. However, in terms of micro-F1, its performance is worse than the supervised approach when we employ more than 10% of labeled data. Moreover, the semi-supervised approach provides a noticeable improvement in the classification of underrepresented class reviews (neutral class). Conversely, when the availability of labeled reviews increases, its performance on overrepresented class

Fig. 6: BERT-LSTM and SSGAN-BL classifiers confusion matrices

reviews (positive and negative) decreases compared to the supervised approach.

The semi-supervised approach provides promising results, however, there is still much room for improvement. Therefore, we plan to propose new deep learning methods capable to automatically generate instances for training, reducing the dependence on annotated dataset. In addition, we will explore our semi-supervised approach for other NLP tasks such as named entity recognition, relation extraction or text summarization, among others. Moreover, we will study other transformer-based architectures such as GPT-2 [54] and XML [55] to deal with these tasks within the settings enabling semi-supervised approaches and few-shot transfer learning [56]. The basic idea of this technique is to reuse a model, which was trained to distinguish between different classes, to learn how to identify examples of unknown classes for the model, using only a few seed examples.

ACKNOWLEDGMENT

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R, ACCESS2MEET project-PID2020-116527RB-I00). It was also funded by the Community of Madrid under the Interdisciplinary Projects Program for Young Researchers (NLP4Rare-CM-UC3M project) and the line of Excellence of University Professors (EPUC3M17).

REFERENCES

- [1] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, 2018.
- [2] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods," *Cognitive computation*, pp. 1–16, 2021.
- [3] R. Sandoval-Almazán and D. Valle-Cruz, "Sentiment analysis of facebook users reacting to political campaign posts," *Digital Government: Research and Practice*, vol. 1, no. 2, pp. 1–13, 2020.
- [4] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covid-senti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, 2021.
- [5] C. Zuheros, E. Martínez-Cámara, E. Herrera-Viedma, and F. Herrera, "Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. case study of restaurant choice using tripadvisor reviews," *Information Fusion*, vol. 68, pp. 22–36, 2021.
- [6] C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews," *Journal of Biomedical Informatics*, vol. 110, p. 103539, 2020.
- [7] I. Segura-Bedmar and P. Martínez Fernández, "Pharmacovigilance through the development of text mining and natural language processing techniques," 2015.
- [8] S. A. Adams, "Using patient-reported experiences for pharmacovigilance?" in *Context Sensitive Health Informatics: Human and Sociotechnical Approaches*. IOS Press, 2013, pp. 63–68.
- [9] A. Wong, J. M. Plasek, S. P. Montecalvo, and L. Zhou, "Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 38, no. 8, pp. 822–841, 2018.
- [10] H. Liu, A. Gegov, and F. Stahl, "Categorization and construction of rule based systems," in *International conference on engineering applications of neural networks*. Springer, 2014, pp. 183–194.
- [11] P. Lin and X. Luo, "A survey of sentiment analysis based on machine learning," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2020, pp. 372–387.
- [12] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "Successes and challenges in developing a hybrid approach to sentiment analysis," *Applied Intelligence*, vol. 48, no. 5, pp. 1176–1188, 2018.
- [13] J. Chai and A. Li, "Deep learning in natural language processing: A state-of-the-art survey," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 2019, pp. 1–6.
- [14] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.
- [15] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.
- [16] S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López, "How do we talk about doctors and drugs? sentiment analysis in forums expressing opinions for medical domain," *Artificial intelligence in medicine*, vol. 93, pp. 50–57, 2019.
- [17] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," in *Proceedings of the 2018 International Conference on Digital Health*, 2018, pp. 121–125.
- [18] B. Biseda and K. Mo, "Enhancing pharmacovigilance with drug reviews and social media," *ArXiv*, vol. abs/2004.08731, 2020.
- [19] C. Sindhu, B. Sasmal, R. Gupta, and J. Prathipa, "Subjectivity detection for sentiment analysis on twitter data," in *Artificial Intelligence Techniques for Advanced Computing Applications*. Springer, 2021, pp. 467–476.
- [20] H. Li, "Deep learning for natural language processing: advantages and challenges," *National Science Review*, 2017.
- [21] I. Spasic and G. Nenadic, "Clinical text data in machine learning: systematic review," *JMIR medical informatics*, vol. 8, no. 3, p. e17984, 2020.
- [22] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalonde, S. Bhandari, R. Belfer, N. Kanagasabai *et al.*, "Sentiment analysis: It's complicated!" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1886–1895.
- [23] S. Lee and W. Kim, "Sentiment labeling for extending initial labeled data to improve semi-supervised sentiment classification," *Electronic Commerce Research and Applications*, vol. 26, pp. 35–49, 2017.
- [24] N. F. F. da Silva, L. F. Coletta, E. R. Hruschka, and E. R. Hruschka Jr, "Using unsupervised information to improve semi-supervised tweet sentiment classification," *Information Sciences*, vol. 355, pp. 348–365, 2016.
- [25] F. H. Khan, U. Qamar, and S. Bashir, "A semi-supervised approach to sentiment analysis using revised sentiment strength based on sentiwordnet," *Knowledge and Information Systems*, vol. 51, no. 3, pp. 851–872, 2017.
- [26] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware bert for cross-domain sentiment analysis," in *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 2020, pp. 4019–4028.
- [27] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He, "Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 568–579.
- [28] J. Liu, S. Zheng, G. Xu, and M. Lin, "Cross-domain sentiment aware word embeddings for review sentiment analysis," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 2, pp. 343–354, 2021.
- [29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.
- [30] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *China National Conference on Chinese Computational Linguistics*. Springer, 2019, pp. 194–206.
- [31] D. Croce, G. Castellucci, and R. Basili, "Gan-bert: 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*, 2020, pp. 2114–2119.
- [32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.
- [33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [34] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: The impact of student initialization on knowledge distillation," *ArXiv*, vol. abs/1908.08962, 2019.
- [35] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, p. 107134, 2021.

- [36] A. Mabrouk, R. P. D. Redondo, and M. Kaye, "Deep learning-based sentiment classification: A comparative survey," *IEEE Access*, vol. 8, pp. 85 616–85 638, 2020.
- [37] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *LREC*, vol. 6. Citeseer, 2006, pp. 417–422.
- [38] L. Goeuriot, J.-C. Na, W. Y. Min Kyaing, C. Khoo, Y.-K. Chang, Y.-L. Theng, and J.-J. Kim, "Sentiment lexicons for health-related opinion mining," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 2012, pp. 219–226.
- [39] M. T. Wiley, C. Jin, V. Hristidis, and K. M. Esterling, "Pharmaceutical drugs chatter on online social networks," *Journal of biomedical informatics*, vol. 49, pp. 245–254, 2014.
- [40] V. Bobicev, M. Sokolova, Y. Jafer, and D. Schramm, "Learning sentiments from tweets with personal health information," in *Canadian conference on artificial intelligence*. Springer, 2012, pp. 37–48.
- [41] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, "Can i hear you? sentiment analysis on medical forums," in *Proceedings of the sixth international joint conference on natural language processing*, 2013, pp. 667–673.
- [42] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. [Online]. Available: <https://www.aclweb.org/anthology/H05-1044>
- [43] A. Mishra, A. Malviya, and S. Aggarwal, "Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1402–1409.
- [44] Z. Min, "Drugs reviews sentiment analysis using weakly supervised model," in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2019, pp. 332–336.
- [45] N. Punith and K. Raketla, "Sentiment analysis of drug reviews using transfer learning," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2021, pp. 1794–1799.
- [46] H. Alhuzali and S. Ananiadou, "Improving classification of adverse drug reactions through using sentiment analysis and transfer learning," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 339–347.
- [47] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL*, 2018.
- [48] C. Wu, F. Wu, S. Wu, Z. Yuan, J. Liu, and Y. Huang, "Semi-supervised dimensional sentiment analysis with variational autoencoder," *Knowledge-Based Systems*, vol. 165, pp. 30–39, 2019.
- [49] D. Preotiu-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman, "Modelling valence and arousal in facebook posts," in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2016, pp. 9–15.
- [50] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [51] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.
- [52] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *ArXiv*, vol. abs/1505.00853, 2015.
- [53] K. Fan, "Minimax theorems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, no. 1, p. 42, 1953.
- [54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [55] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in Neural Information Processing Systems*, vol. 32, pp. 7059–7069, 2019.
- [56] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.