DUSE: A New Benchmark Dataset for Drug User Sentiment Extraction

Ashok Kumar J
Information Science and Technology
Anna University, Chennai, India
jashokkumar83@auist.net

Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University, Singapore
cambria@ntu.edu.sg

Tina Esther Trueman
Information Science and Technology
Anna University, Chennai, India
tina_trueman@auist.net

Abstract—Social media continuously produce a huge volume of data in different formats and different domains. In particular, patients' and caregivers' written medical texts play an important role among individuals, medical doctors, and drug developers for understanding drug users' sentiment. However, automatic sentiment detection is a challenging problem in medical settings due to a lack of data with age group, gender, treatment duration, and so on. Therefore, we present a drug review dataset for the most reviewed 100 drugs. Especially, we collected 88K instances from WebMD which is one of the largest online health service providers. Empirically, we explore strongly labeled data and weakly labeled data for automatic sentiment detection using BERT, which learns context-dependent features. We show that the BERT model yields better accuracy than the baseline models.

Index Terms—Sentiment classification, drug user sentiment, transformers, BERT

I. INTRODUCTION

Nowadays, social media allow Internet users to interact, create, or share their interests, feelings, and ideas about an individual, an organization, or a product [1]–[4]. The individual (or organization) uses the Internet users information to build a business or monitor a product. In particular, social media plays a vital role in online health forums [5]–[7]. These forums allow drug users (patients or caregivers) to express their reactions or experiences on drugs and medications in the form of reviews or texts [8]. Drug users information can be used to improve the condition of the patients by looking at their drug usage level, side effects, causes, and effectiveness. Also, it is useful to medical doctors, drug developers, and individuals to understand the drug users condition and their experiences with a particular drug [9].

However, analyzing the large volume of data becomes a more challenging task to the medical doctors, drug developers, and individuals due to the age group, gender, treatment duration, and patient condition. Therefore, sentiment analysis is used to analyze the drug users’ experience on drugs and medication reviews. Sentiment analysis determines a personal feeling of an individual on a particular drug [10]. The personal feelings of an individual can either be positive, e.g., “Citalopram oral has helped me feel more like myself. Easy to take” or negative, e.g., “Unable to manage anxiety on a day to day basis”.

However, there is a lack of larger drug user sentiment detection datasets with age-group, gender, treatment duration, opinion giver, and prescribed condition of the drug for advanced computational models. To address these problems, we introduce the strongly labeled drug user sentiment extraction (DUSE) dataset. This dataset contains 88447 instances with a comment, age-group, gender, treatment duration, and opinion giver, condition, satisfaction, effectiveness, ease of use, and overall rating scores. The instances in DUSE are collected from WebMD [11], which is one of the most important online health information providers. For each instance, a sentiment label is assigned based on the overall rating scores such as positive and negative sentiments but also neutral [12].

In addition, the overall rating score of a text may not describe the accurate sentiment of the text. The manual annotation of these larger texts is impossible due to cost and time. Therefore, we introduce a strongly labeled dataset using rating-score (DUSE) using SenticNet [13], a neurosymbolic artificial intelligence (AI) framework for sentiment analysis. Empirically, we evaluate these datasets with the baseline models such as logistic regression (LR) [14], Na ¨ıve Bayes-support vector machine (NB-SVM) [15], and gated recurrent unit (GRU) [16]. The LR and NB-SVM represent BoW (bag of words) features, and the GRU represents the context-independent features. Furthermore, we introduce a Bidirectional Encoder Representation from Transformers (BERT) [17] to detect drug user sentiment. This model uses context-dependent features in a long-range input sequence. Our experiment indicates that the BERT model improves the performance of the baseline models.

The rest of this paper is organized as follows: Section II describes the drug user sentiment detection dataset with a labeled data and weakly-labeled data; in Section III, the automatic sentiment detection task is presented for drug user reviews; Section IV presents results and discussion; finally, this paper concludes with future works in Section V.

II. DRUG REVIEW DATASETS

Online social media is one the best source for identifying drug users’ experience and their opinion. Specifically, Grer et al. [18] constructed drug users experiment dataset from two web pages, namely, Drugs.com and DrugLib.com. They obtained 215063 reviews from the first webpage and 3551 reviews from the second webpage.
These datasets consist of user reviews, related conditions, and a user rating (10 stars). The authors derived the overall patient satisfaction with three sentiment polarity labels such as negative, neutral, and positive. Similarly, the side effects and effectiveness are derived with three sentiment polarity labels. Demner-Fushman et al. [19] selected the 200 approved drugs for generating the distinct labeled adverse drug reactions (ADRs) database and the annotated dataset of the structured product labels. The authors also verified the quality of ADRs to avoid bias. Kuroshima et al. [20] collected 10000 tweets of the self-reported patients from the Twitter feed for four pain killers, namely, Aleve, Motrin, Advil, and Tylenol. The authors collected this data for three months and labeled the sentiment polarity of these tweets. Then, they presented a computational method to detect the sentiment of the drug. Their results show a 70.7% precision score for the validated data. Moreover, Ribeiro et al. [21] created a database of 30000 labeled tweets for four distinct drugs namely, Fluoxetine, Quetiapine, Tamoxifen, and Venlafaxine. The authors also constructed ontology for improving the extraction of ADRs. Their study indicated that Twitter is one of the main sources for identifying ADRs. Even though many researchers introduced drug-related datasets for ADRs and opinion mining, there is no dataset available based on gender, age group, treatment duration, and drug opinion giver. These factors are more important to drug users. In particular, each drugs reaction varies from person to person based on their age group, gender, and treatment duration. Therefore, we introduce a labeled data based on drug users’ rating scores and weakly labeled data based on neurosymbolic AI (SenticNet) for identifying drug user sentiment.

First, the overall rating score based DUSE dataset\(^1\) contains 88447 comments from WebMD for 100 drugs [11]. Each of these comments is associated with drug name, condition, date and time, age group, gender, opinion giver, treatment duration, effectiveness rating, ease of use rating, satisfaction rating, overall rating, comments, and sentiment polarity label. The sentiment polarity label is assigned based on the overall rating score. A negative sentiment label is assigned for the rating score of 1 to 2, a neutral sentiment label is assigned for the rating score of 3, and a positive sentiment label is assigned for the rating score of 4 to 5. Fig. 1 shows some random examples for this dataset. The number of instances for the negative, neutral, and positive categories is shown in Table I. Second, the overall rating score of a text may not describe a correct polarity of a text always. For instance, the text “I am hungry all the time and I cannot sleep more than two or three hours a night” in Abilify oral indicates the overall rating score of three. This score shows a neutral sentiment polarity of the text. However, the text represents a negative sentiment polarity. Therefore, we use neurosymbolic AI to detect sentiment polarity for drug users comments. Neurosymbolic AI uses rules or formulas to represent real-world problems or applications in terms of properties and relations.

\(^1\)https://sentic.net/downloads

---

**Drug Name:** Abilify oral  
**Condition:** Additional Medications to Treat Depression  
**Date and Time:** 3/9/2017 1:48:00 AM  
**Age group:** 19-24  
**Gender:** Unknown  
**Opinion giver:** Patient  
**Treatment duration:** 6 months to less than 1 year  
**Effectiveness:** 5 star  
**Ease of use:** 5 star  
**Satisfaction:** 5 star  
**Overall rating:** 5 star  
**Comment:** adding this to my treatment helped me greatly.  
**Sentiment:** Positive

---

**Drug Name:** Acetaminophen oral  
**Condition:** Pain  
**Date and Time:** 10/1/2009 1:13:00 PM  
**Age group:** Unknown  
**Gender:** Female  
**Opinion giver:** Patient  
**Treatment duration:** Unknown  
**Effectiveness:** 1 star  
**Ease of use:** 1 star  
**Satisfaction:** 1 star  
**Overall rating:** 1 star  
**Comment:** I gain no relief from this treatment, and it also causes me to have stomach pain.  
**Sentiment:** Negative

---

**Drug Name:** Abilify oral  
**Condition:** Additional Medications to Treat Depression  
**Date and Time:** 6/22/2014 12:33:00 AM  
**Age group:** 45-54  
**Gender:** Female  
**Opinion giver:** Patient  
**Treatment duration:** 1 to 6 months  
**Effectiveness:** 4 star  
**Ease of use:** 3 star  
**Satisfaction:** 2 star  
**Overall rating:** 3 star  
**Comment:** Excellent help to my depression. But gained weight, even on 2 1/2 mg, and worse, I became borderline diabetic.  
**Sentiment:** Neutral

---

Fig. 1. Some random examples from DUSE
In particular, we use Sentic APIs\(^2\) to detect the sentiment polarity of a text or comment into a negative, neutral, or positive (Fig. 2). These APIs leverage neurosymbolic AI to detect concepts in a text and it assigns their contextual sentiment polarity based on the Jumping NLP curves paradigm (Fig. 3). The polarity label assignment is done through the dependency relations via sentic patterns. Here, there is no involvement of human expertise in the dataset for polarity assignment. Therefore, it is called a weakly labeled dataset.

### III. Automatic Sentiment Detection

We use machine learning, deep learning [22], and transformers-based [23] models to develop automatic sentiment detection for our datasets. In this task, we solve a binary sentiment classification problem for both strongly labeled and weakly labeled datasets. The following research questions are addressed in this task.

- How well can conventional machine learning, deep learning, and transformers-based models classify a text or comment into a fine-grained sentiment category?
- Can we design a transformer-based architecture to integrate patient-related meta-data with the comment to improve the performance of sentiment detection?

Recently, BERT architecture has achieved a greater performance on various classification datasets [17], [24]. It is built with the encoder representations of the transformer model. The model learns bidirectional context information from both directions i.e, from left to right and right to left. It is mainly designed to create pre-training language representations for fine-tuning specific tasks such as entity recognition, question answering, and classification. The pre-training language representation is constructed with two variants, namely, BERT Base pre-trained model and BERT Large pre-trained model using Wikipedia and BookCorpus datasets. Firstly, the BERT Base pre-trained model is built with 12 encoder or transformer layers, 12 self-attention heads, and 768 hidden units for representing 110M parameters. Secondly, the BERT Large pre-trained model is built with 24 encoder or transformer layers, 16 self-attention heads, and 1024 hidden units for representing 340M parameters. Each transformer layer contains two components, namely, a self-attention and feed-forward neural network. Firstly, self-attention relates each token position with other tokens in terms of queries (Q), keys (K), and values (V) [23]. Secondly, the feed-forward neural network normalizes the output and learns backpropagation. In this work, we classify the drug user sentiment using the BERT Base model. Especially, a sigmoid activation is employed on the top of the BERT transformer.

### IV. Results and Discussion

In this section, we present the experimental settings, results, and comparison of various models. We define the sentiment polarity label for the obtained DUSE dataset in two ways, rating-based sentiment polarity (strongly labeled) and Sentic API-based sentiment polarity (weakly labeled).

---

\(^2\)https://sentic.net/api

---

**TABLE I**

**DATA DISTRIBUTION**

<table>
<thead>
<tr>
<th>Class</th>
<th>Strongly labeled #instances</th>
<th>Weakly labeled #instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>22288</td>
<td>39986</td>
</tr>
<tr>
<td>Neutral</td>
<td>15761</td>
<td>2130</td>
</tr>
<tr>
<td>Positive</td>
<td>50398</td>
<td>46331</td>
</tr>
<tr>
<td>Total</td>
<td>88447</td>
<td>88447</td>
</tr>
</tbody>
</table>

**TABLE II**

**DATA SPLIT**

<table>
<thead>
<tr>
<th>Class</th>
<th>Strongly labeled data</th>
<th>Weakly labeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Valid</td>
</tr>
<tr>
<td>Negative</td>
<td>18053</td>
<td>206</td>
</tr>
<tr>
<td>Positive</td>
<td>40822</td>
<td>4536</td>
</tr>
<tr>
<td>Total</td>
<td>58875</td>
<td>6542</td>
</tr>
</tbody>
</table>

Fig. 2. Sentic API framework [25]

Fig. 3. Jumping NLP curves paradigm [26]
Table III

<table>
<thead>
<tr>
<th>Methods</th>
<th>Class</th>
<th>Strongly labeled Dataset</th>
<th>Weakly labeled Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Validation</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>NEG</td>
<td>POS</td>
<td>NEG</td>
</tr>
<tr>
<td>TC</td>
<td>17655</td>
<td>398</td>
<td>1599</td>
</tr>
<tr>
<td>POS</td>
<td>325</td>
<td>40997</td>
<td>339</td>
</tr>
<tr>
<td>TC+All</td>
<td>394</td>
<td>40428</td>
<td>317</td>
</tr>
</tbody>
</table>

NEG-Negative, POS-Positive, TC-Text comments

Table IV

The BERT BASE MODEL PERFORMANCE FOR THE STRONGLY LABELED DATASET

<table>
<thead>
<tr>
<th>Class</th>
<th>Text comments (TC)</th>
<th>Text comments + All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Negative</td>
<td>0.8251</td>
<td>0.7971</td>
</tr>
<tr>
<td>Positive</td>
<td>0.9116</td>
<td>0.9253</td>
</tr>
<tr>
<td>Macro</td>
<td>0.8683</td>
<td>0.8612</td>
</tr>
<tr>
<td>Micro</td>
<td>0.8860</td>
<td>0.8860</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.8851</td>
<td>0.8860</td>
</tr>
</tbody>
</table>

P-Precision, R-Recall, F1-F1 Score

Table V

The BERT BASE MODEL PERFORMANCE FOR THE WEAKLY LABELED DATASET

<table>
<thead>
<tr>
<th>Class</th>
<th>Text comments (TC)</th>
<th>Text comments + All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Negative</td>
<td>0.8260</td>
<td>0.7969</td>
</tr>
<tr>
<td>Positive</td>
<td>0.8299</td>
<td>0.8552</td>
</tr>
<tr>
<td>Macro</td>
<td>0.8280</td>
<td>0.8260</td>
</tr>
<tr>
<td>Micro</td>
<td>0.8282</td>
<td>0.8282</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.8281</td>
<td>0.8282</td>
</tr>
</tbody>
</table>

P-Precision, R-Recall, F1-F1 Score

Table VI

RESULT COMPARISON

<table>
<thead>
<tr>
<th>Models</th>
<th>Strongly Labeled Dataset</th>
<th>Weakly Labeled Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TC</td>
<td>TC + All</td>
</tr>
<tr>
<td></td>
<td>82.05</td>
<td>82.75</td>
</tr>
<tr>
<td>NB-SVM</td>
<td>82.12</td>
<td>82.94</td>
</tr>
<tr>
<td>BiGRU</td>
<td>84.61</td>
<td>84.81</td>
</tr>
<tr>
<td>BERT</td>
<td>88.60</td>
<td>88.25</td>
</tr>
</tbody>
</table>

TC-Text comments

The strongly labeled dataset contains 88447 instances with negative (22288), neutral (15761), and positive (50398) polarities. Similarly, the weakly labeled dataset contains 88447 instances with negative (39986), neutral (2130), and positive (46331) polarities. In this paper, we performed a binary class classification for both datasets. Therefore, we randomly split both datasets for training, validation, and testing based on positive and negative sentiment polarity in the ratio of 80:10:10 as shown in Table II. For both datasets, we applied a word contraction map for expanding short texts, and punctuation removal except periods, single and double-quotes. We then used five models as baselines, namely, logistic regression (LR), Naïve Bayes and Support Vector Machine (NB-SVM), Bidirectional Gated Recurrent Units (BiGRU), and BERT base model.

We used the ktrain python library for implementing these baseline models. In particular, we handled the missing information in each instance with the word unknown. Firstly, we used a bag of words features for LR and NB-SVM with various hyperparameters such as 150 maximum input sequence length, 0.001 triangular learning rate, and 20000 maximum word features. Secondly, we used fast text word embedding features for BiGRU with 300 dimension input vectors, 150 maximum input sequence length, 0.001 triangular learning rate, and 20000 maximum word features. Finally, we used context-dependent embedding features for the BERT base language model with 768 dimension input vectors, 320 maximum input sequence length, 2e-5 one-cycle learning rate, and 20000 maximum word features. The batch size of 14 is chosen based on trial and error for all baseline models.
In particular, we performed the experiment with only text (TC) and text and text-related meta-data (TC+All) in both datasets. The performance of the baseline models is evaluated with a confusion matrix, precision, recall, and F1 score and their corresponding macro, micro, and weighted scores [27]. Table III shows the confusion matrix of the BERT base model for both strongly labeled and weakly labeled datasets with TC and TC+All features. Moreover, it describes the summary of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for training, validation, and testing. The evaluation result of the strongly labeled dataset is shown in Table IV and the weakly labeled dataset is shown in Table V. These tables indicate that the BERT base model achieves 88.25% for text comments and 89.09% for text and text-related meta-data in the strongly labeled dataset, and 82.76% for text comments and 82.95% for text and text-related meta-data in the weakly labeled dataset. Table VI shows the result comparison of all baseline models. This table indicates that the BERT base model outperforms the LR, NB-SVM, and BiGRU models. We also found that there is a higher performance for text and text-related meta-data.

V. CONCLUSION

We introduced DUSE, a new dataset for automatic sentiment detection. This dataset enables the development of computational approaches with drug users meta-data such as age group, gender, treatment duration, condition, and opinion giver. Specifically, we obtained the sentiment polarity of texts with drug users’ overall rating score and neurosymbolic AI for a strongly labeled dataset and weakly labeled dataset, respectively. In this paper, we show that the BERT model significantly improves the performance with text and meta-data for both datasets. In particular, our empirical results indicate a higher accuracy for the strongly labeled dataset. In future work, our dataset can be used for gender-based and age group-based drug user sentiment detection tasks.

ACKNOWLEDGMENT

This work was supported by the University Grants Commission, Government of India under the National Doctoral Fellowship.

REFERENCES