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], Naïve 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.

2375-9259/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00028

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

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, Mortin, 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

TABLE I
DATA DISTRIBUTION

Strongly labeled		Weakly labeled	
Class	**#instances**	**Class**	**#instances**
Negative	22288	Negative	39986
Neutral	15761	Neutral	2130
Positive	50398	Positive	46331
Total	88447	Total	88447

TABLE II
DATA SPLIT

Class	Strongly labeled data			Weakly labeled data		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**
Negative	18053	2006	2229	32388	3599	3999
Positive	40822	4536	5040	37528	4170	4633
Total	58875	6542	7269	69916	7769	8632

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.

[2]https://sentic.net/api

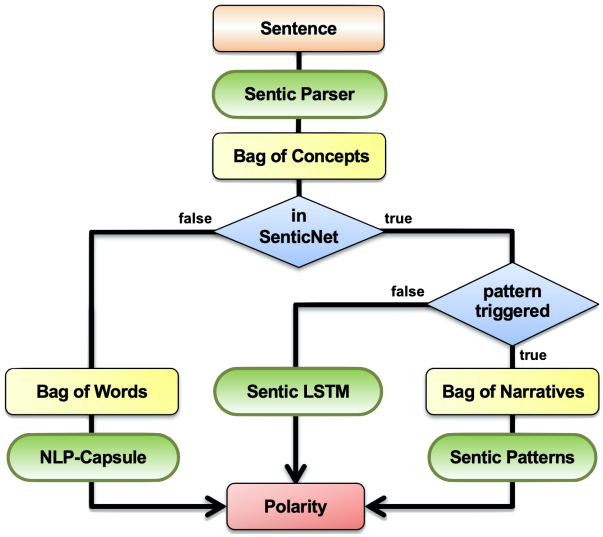

Fig. 2. Sentic API framework [25]

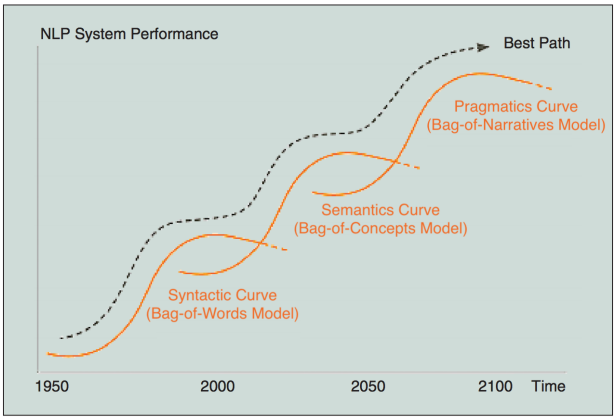

Fig. 3. Jumping NLP curves paradigm [26]

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).

TABLE III
CONFUSION MATRIX

Methods	Class	Strongly labeled Dataset						Weakly labeled Dataset					
		Training		Validation		Testing		Training		Validation		Testing	
		NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS
TC	NEG	17655	398	1599	407	1789	440	30621	1767	2868	731	3186	813
	POS	325	40497	339	4197	414	4626	946	36582	604	3566	675	3958
TC+All	NEG	17502	551	1589	417	1778	451	30804	1584	2934	665	3240	759
	POS	394	40428	317	4219	342	4698	1271	36257	657	3513	713	3920

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

TABLE IV
THE BERT BASE MODEL PERFORMANCE FOR THE STRONGLY LABELED DATASET

Class	Text comments (TC)						Text comments + All					
	Validation			Testing			Validation			Testing		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Negative	0.8251	0.7971	0.8109	0.8121	0.8026	0.8073	0.8337	0.7921	0.8124	0.8387	0.7977	0.8177
Positive	0.9116	0.9253	0.9184	0.9131	0.9179	0.9155	0.9101	0.9301	0.9200	0.9124	0.9321	0.9222
Macro	0.8683	0.8612	0.8646	0.8626	0.8602	0.8614	0.8719	0.8611	0.8662	0.8755	0.8649	0.8699
Micro	0.8860	0.8860	**0.8860**	0.8825	0.8825	**0.8825**	0.8878	0.8878	**0.8878**	0.8909	0.8909	**0.8909**
Weighted	0.8851	0.8860	0.8854	0.8822	0.8825	0.8823	0.8866	0.8878	0.8870	0.8898	0.8909	0.8901

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

TABLE V
THE BERT BASE MODEL PERFORMANCE FOR THE WEAKLY LABELED DATASET

Class	Text comments (TC)						Text comments + All					
	Validation			Testing			Validation			Testing		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Negative	0.8260	0.7969	0.8112	0.8252	0.7967	0.8107	0.8170	0.8152	0.8161	0.8196	0.8102	0.8149
Positive	0.8299	0.8552	0.8423	0.8296	0.8543	0.8418	0.8408	0.8424	0.8416	0.8378	0.8461	0.8419
Macro	0.8280	0.8260	0.8268	0.8274	0.8255	0.8262	0.8289	0.8288	0.8289	0.8287	0.8282	0.8284
Micro	0.8282	0.8282	**0.8282**	0.8276	0.8276	**0.8276**	0.8298	0.8298	**0.8298**	0.8295	0.8295	**0.8295**
Weighted	0.8281	0.8282	0.8279	0.8275	0.8276	0.8274	0.8298	0.8298	**0.8298**	0.8294	0.8295	0.8294

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

TABLE VI
RESULT COMPARISON

Models	Strongly Labeled Dataset				Weakly Labeled Dataset			
	TC		TC + All		TC		TC + All	
	Valid	Test	Valid	Test	Valid	Test	Valid	Test
LR	82.05	82.75	83.60	83.90	64.20	64.23	64.28	63.51
NB-SVM	82.12	82.94	83.80	83.77	63.15	63.72	63.50	63.42
BiGRU	84.61	84.81	85.72	86.67	75.36	75.46	75.78	75.90
BERT	**88.60**	**88.25**	**88.78**	**89.09**	**82.82**	**82.76**	**82.98**	**82.95**

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 punctuations 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

[1] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," *Knowledge-Based Systems*, vol. 69, pp. 1–2, 2014.

[2] M. Grassi, E. Cambria, A. Hussain, and F. Piazza, "Sentic web: A new paradigm for managing social media affective information," *Cognitive Computation*, vol. 3, no. 3, pp. 480–489, 2011.

[3] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional cnn-rnn deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294, 2021.

[4] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Systems with Applications*, vol. 135, pp. 60–70, 2019.

[5] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro, "Sentic computing for patient centered applications," in *IEEE ICSP*, 2010, pp. 1279–1282.

[6] A. Kathua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks," *Information Processing and Management*, vol. 56, no. 1, pp. 247–257, 2019.

[7] E. Cambria, T. Benson, C. Eckl, and A. Hussain, "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 533–10 543, 2012.

[8] Y. Zhang, S. Cui, and H. Gao, "Adverse drug reaction detection on social media with deep linguistic features," *Journal of biomedical informatics*, vol. 106, p. 103437, 2020.

[9] B. Guarita, V. Belackova, D. Van Der Gouwe, M. Blankers, M. Pazitny, and P. Griffiths, "Monitoring drug trends in the digital environment–new methods, challenges and the opportunities provided by automated approaches," *International Journal of Drug Policy*, p. 103210, 2021.

[10] M. E. Basiri, M. Abdar, M. A. Cifci, S. Nemati, and U. R. Acharya, "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques," *Knowledge-Based Systems*, vol. 198, p. 105949, 2020.

[11] WedMD, https://www.webmd.com/drugs/2/index.

[12] A. Valdivia, V. Luzón, E. Cambria, and F. Herrera, "Consensus vote models for detecting and filtering neutrality in sentiment analysis," *Information Fusion*, vol. 44, pp. 126–135, 2018.

[13] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 105–114.

[14] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[15] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

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

[18] 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.

[19] D. Demner-Fushman, S. E. Shooshan, L. Rodriguez, A. R. Aronson, F. Lang, W. Rogers, K. Roberts, and J. Tonning, "A dataset of 200 structured product labels annotated for adverse drug reactions," *Scientific data*, vol. 5, no. 1, pp. 1–8, 2018.

[20] D. Kuroshima and T. Tian, "Detecting public sentiment of medicine by mining twitter data," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 10, pp. 1–5, 2019.

[21] L. A. Ribeiro, D. Cinalli, and A. C. B. Garcia, "Discovering adverse drug reactions from twitter: A sentiment analysis perspective," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2021, pp. 1172–1177.

[22] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using bert," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, vol. 1. IEEE, 2019, pp. 1–5.

[25] E. Cambria, Y. Song, H. Wang, and N. Howard, "Semantic multi-dimensional scaling for open-domain sentiment analysis," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 44–51, 2014.

[26] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[27] R. S. Holt, P. A. Mastromarino, E. K. Kao, and M. B. Hurley, "Information theoretic approach for performance evaluation of multi-class assignment systems," in *Signal Processing, Sensor Fusion, and Target Recognition XIX*, vol. 7697. International Society for Optics and Photonics, 2010, p. 76970R.