Augmenting Reddit Posts to Determine Wellness Dimensions impacting Mental Health

*Chandreen Liyanage, **Muskan Garg, *Vijay Mago, **Sunghwan Sohn
* Lakehead University, Thunder Bay, ON P7B 5E1, Canada
**Mayo Clinic, Rochester, MN 55901, USA

Abstract

Amid ongoing health crisis, there is a growing necessity to discern possible signs of Wellness Dimensions (WD)[1] manifested in self-narrated text. As the distribution of WD on social media data is intrinsically imbalanced, we experiment the generative NLP models for data augmentation to enable further improvement in the pre-screening task of classifying WD. To this end, we propose a simple yet effective data augmentation approach through prompt-based Generative NLP models, and evaluate the ROUGE scores and syntactic/semantic similarity among *existing interpretations* and *augmented data.* Our approach with ChatGPT model surpasses all the other methods and achieves improvement over baselines such as Easy-Data Augmentation and Backtranslation. Introducing data augmentation to generate more training samples and balanced dataset, results in the improved F-score and the Matthew's Correlation Coefficient for upto 13.11% and 15.95%, respectively.

1 Introduction

The social determinants of health (SDOH) refer to various factors present in the surroundings where individuals are born, reside, acquire knowledge, work, engage in leisure activities, practice religion, grow older, impacting a broad range of health-related outcomes, risks and quality-of-life indicators.[2][3] A rapid expansion of research in SDOH 2030 encourages the social NLP research community to design and develop
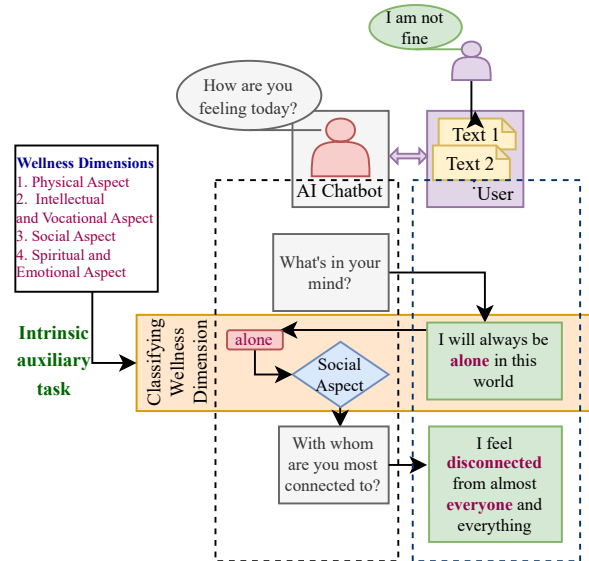

Figure 1: Overview of the task. Generating balanced dataset through data augmentation to facilitate the development of classifiers screening Reddit posts through a lens of Wellness Dimensions.

computational intelligence models for enhancement of an individual's well-being (Bompelli et al., 2021). In this work, we choose to pre-screen human-writings for biomedical therapy by investigating latent indicators of *wellness dimensions* in Reddit posts (see illustration in Figure 1). Wellness dimensions (WD) refer to different aspects of an individual's overall well-being that contribute to their physical, spiritual, social, emotional, intellectual, occupational, environmental, and financial well-being. The disturbed WD, if remains unaddressed, have adverse impact on mental health of an individual. As social media becomes integral part of our daily lives (Wang et al., 2020), studies in the past suggest that individuals tend to express their thoughts and emotions impacted by one or more wellness dimensions more easily on social media platforms as compared to during in-person sessions with clinical psychol-

---

[1]The concept of Wellness Dimensions is often used in holistic approaches to health, recognizing that well-being encompasses multiple areas of life.

[2]https://health.gov/healthypeople/priority-areas/social-determinants-health

[3]Social Determinants of Health (SDOH) are the social and economic factors that influence an individual's health outcomes.

ogists and mental healthcare (Zhang et al., 2023; Garg, 2023). We construct, annotate and observe the original (natural) composition of WD dataset as an imbalanced dataset. In this work, we augment a multi-class dataset on WD to facilitate design and development of NLP models for classifying WD impacting mental health in Reddit posts during mental health screening. Pre-screening filters are helpful in biomedical therapy by facilitating early detection of WD impacting mental health, which if left untreated may cause severe mental disorders. Dunn highlights holistic nature of wellness in 1961 as a *high-level wellness*, denoting a superior level of healthy living (Printz-Markó and Ivancsóné Horváth, 2017).

We reduce multiple WD to four key dimensions of well-being based on the frequency and recognition in human writings: Physical Aspect (PA), Intellectual and Vocational Aspect (IVA), Social Aspect (SA), Spiritual and Emotional Aspect (SEA) (Wickramarathne et al., 2020; Dillette et al., 2021). Our major contributions (as illustrated in Fig. 1) include (i) the applicability of generative NLP models for domain-specific data augmentation, (ii) examining the diversity among generated and original instances through semantic and syntactic similarity measure, (iii) test and validate the efficacy of data augmentation by investigating classifiers' performance.

2 Background

According to Weiss (1975), sociologists put forth a theory that outlines six social needs to prevent loneliness: attachment, social integration, nurturance, reassurance of worth, sense of reliable alliance, and guidance in stressful situations (Weiss, 1975). The Self-Determination Theory (SDT)[4] highlights the importance of balancing relatedness, competency, and autonomy for intrinsic motivation and genuine self-esteem, which contribute to overall well-being. Neglecting mental disturbance can escalate subclinical depression to clinical depression by activating interpersonal risks. This research seeks to examine the origins and outcomes of mental disturbance to mitigate these risks.

Corpus Construction We present a new dataset with 3,092 instances and 72,813 words to identify wellness dimensions impacting mental disturbance: PA, IVA, SA, and SEA. A senior clinical psychologist, a rehabilitation councilor, and a social NLP researcher framed annotation schemes and perplexity guidelines for text annotation through pre-defined wellness dimensions. Our experts trained three postgraduate students to annotate the data based on predefined dimensions. The annotations were validated using Fleiss' Kappa interobserver agreement, resulting in a kappa score of 74.39%. Final annotations were determined through majority voting and expert verification. The experts achieved a kappa score of 87.32% for the selection of explanatory text spans. Despite slight confusion between PA and SEA, there was a higher agreement for the selection of explanations. To facilitate future research and developments, we publicly release our dataset at Github.[5]

Problem Formulation: We collect and annotate Reddit data from subreddits `r/depression` and `r/suicidewatch` for the task of identifying WD and found imbalanced dataset in its natural composition, suggesting the need of data augmentation. To evaluate the effectiveness of generative NLP models for data augmentation, we frame the task of *augmenting Reddit posts* as a *text generation* problem. We compare and contrast the performance of model trained on data augmented with (i) Generative Pre-trained Transformers (GPT) (Yang et al., 2023), and (ii) conventional data augmentation approach for NLP such as Easy Data Augmentation (EDA) (Wei and Zou, 2019) and Back Translation (BT) (Sennrich et al., 2016).

3 Experiments and Evaluation

We first generate the data using two-fold measures: (i) traditional data augmentation methods for NLP - EDA and BT, and (ii) prompt-based Generative Pre-trained Transformer models (Ding et al., 2023). We further investigate the diversity of the generated samples in comparison to the original samples and

---

fine-tune BERT language model to observe improvements in WD classification, if any. Hinged on the classification results and similarity measures, we select the best model for augmenting WD dataset.

3.1 Methods: Data Augmentation

We use pre-trained generative models[6] (Meng et al., 2021) for this task: (i) ChatGPT models: *gpt-3.5-turbo and gpt-3.5-turbo-0301*, and (ii) other GPT-3 models: *text-curie-001 and text-davinci-003*. The original dataset consists of 3092 samples, with 740, 592, 1139 and 621 records from classes PA, IVA, SA, and SEA respectively. We first split the dataset such that we maximize the number of training samples required for each WD.[7] After augmentation, the training set comprises a total of 4376 records, with an equal distribution of 1094 records per class.

Prompt Design and Parameter Setup: As shown in Figure 2, we design following prompts to produce, a) text similar to the original text (Topic and text), and b) an explanation of newly generated text (text and explanation).

While designing prompts according to OpenAI prompt design instructions[8], we begin with explanation through instructions and examples or both. During the text creation, we only provide instructions to the model. As every text belongs to one of the four pre-defined WD, we provide class name as an input, for example, "Physical Aspect", hypothesizing its contribution towards contextual consciousness required for enhancing similar text generation. Furthermore, the explanation generation is developed as a few-shot learning approach (Brown et al., 2020), where we provide *five* text-explanation pairs as examples. The selective examples ensure the representation of all four classes and are made static for every call. We keep temperature as 0.7 to preserve the creativity/ randomness of generated text.

3.2 Method: Similarity Measures

First we calculate ROUGE scores <ROUGE-1, ROUGE-2 and ROUGE-L> to examine similar-

---

Prompt Designs for GPT models

Considering the given topic, generate similar text to the given text.
Topic: ≪class label as a string≫
Text: ≪original text sentence≫

Similar text:

- -

Consider the examples and generate a very short explanation of the given text.

text: ≪example1-text≫
explanation: ≪example1-explanation≫
...
text: ≪example5-text≫
explanation: ≪example5-explanation≫
text: ≪original text sentence≫
explanation: ≪original explanation≫

text: ≪augmented text sentence≫
explanation:

Figure 2: The prompt designs for generating Text and Explanation.

ity[9]. Next, for *semantic similarity,* we calculate the embedding for each sentence through eleven pre-trained language models (Ormerod et al., 2021; Reimers et al., 2019).[10] The resulting sentence embeddings[11] of each original and augmented data instance were then compared using *cosine similarity.* Lastly, for *syntactic similarity*, we first parsed given sentences into syntactic trees and then mapped them into vector representations using the "en_core_web_md" English pipeline in the spaCy library[12]. Next, these vector representations are used to compute the similarity score between sentences. Furthermore, we compute the set overlap between the Part-of-speech (POS) tag sequences of the original and augmented sentences to determine their similarity[13].

---

[6] https://platform.openai.com/docs/models
[7] see Appendix A for more details.
[8] https://platform.openai.com/docs/guides/completion/prompt-design

[9] https://pypi.org/project/rouge/
[10] see Appendix B for more details.
[11] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings
[12] https://spacy.io/models/en
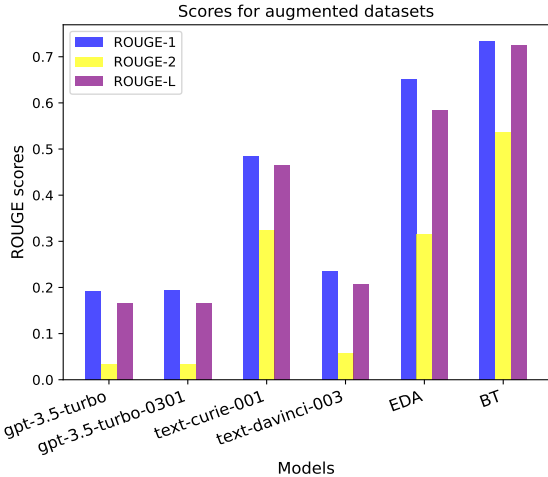[13] https://www.nltk.org/api/nltk.tag.pos_tag.html

Figure 3: We present ROUGE scores for six different augmentation mechanisms leveraging the augmented samples in comparison to the original text, averaged over all the texts generated.

3.3 Classification with BERT

As the final evaluation, we build BERT (Kenton and Toutanova, 2019), a baseline classifier, with 6 augmented datasets and compare its performance with the BERT classifier built over original data. We used the training data in WD dataset for finetuning for 10 epochs with a batch size of 32 and a learning rate of 3e-5. To preserve the lengths of texts, we set the max_length to 256 during tokenization. We use the validation set (20% of the training set) and testing set (180 samples) to examine the efficiency and effectiveness of a classifier through F-score and Matthew's Correlation Coefficient (MCC), respectively.

4 Results and Discussion

Similarity Analysis: We report three types of ROUGE scores: ROUGE-1, ROUGE-2 and ROUGE-L between the original and augmented text. The ChatGPT models show the lowest ROUGE scores, and *gpt-3.5-turbo* versions surpases all other augmentation methods (see Figure 3). We further examine semantic and syntactic similarities through average of all 13 models in Figure 4(a) and 4(b). We observe high diversity and low similarity with GPT based models where ChatGPT based models illustrate the least similarity. However, compared to the other GPT models, text-curie-001 shows a notably higher similarities through all the similarity models.

Table 1: Improvement in classifiers. M1: gpt-3.5-turbo, M2: gpt-3.5-turbo-0301, M3: text-curie-001, M4: text-davinci-003.

Type	Val-A	T-P	T-R	T-F	T-A	T-MCC
Original	0.427	0.65	0.63	0.61	0.63	0.514
M1	0.504	**0.70**	**0.69**	**0.69**	**0.69**	**0.596**
M2	0.499	**0.69**	**0.68**	**0.67**	**0.68**	**0.581**
M3	0.498	0.63	0.63	0.62	0.63	0.519
M4	0.502	0.66	0.67	0.66	0.67	0.559
EDA	0.498	0.63	0.63	0.62	0.63	0.518
BT	0.504	0.65	0.64	0.63	0.64	0.527

Classification Performances: We obtain the validation accuracy (Val-A), and testing results with precision (T-P), recall (T-R), F-score (T-F), accuracy (T-A) and MCC value (T-MCC) with experimental results for evaluation (see Table 1). Even though we keep the testing dataset to be a small chunk of 180 samples, we observe significant difference in the results in training on the *original imbalanced dataset* and *augmented dataset*. The *gpt-3.5-turbo* model over testing dataset outperforms all the baseline models, specifically the original dataset by 2.99% F-score and 1.47% Accuracy followed by the second best model: *gpt-3.5-turbo-0301*. Moreover, compared to the best traditional augmentation method (BT), the top ChatGPT model shows 7.81% improvement in testing accuracy. Notably, the datasets from text-curie-001 and EDA which gained higher similarity values have shown lowest performance on all classification measurements.

We further examine the MCC values to determine the effectiveness of the classifier (Boughorbel et al., 2017). MCC values vary between -1 and 1 such that values closer to 0 and 1 suggest increased randomness and perfect prediction towards decision making correspondingly. We found 15.95% improvement in MCC score when model is trained on augmented training samples with *gpt-3.5-turbo* model. Overall, the augmented text with lowest ROUGE scores, syntactic and semantic similarities showed the highest classification performance on BERT.

Moreover, the following Table 2 compares the class-vise classification performance between the original and M1 (best performed dataset) datasets. We notice a significant improvement in all the measurements of all the classes after augmenting data. Additionally, compared to other classes, the IVA- class with the least
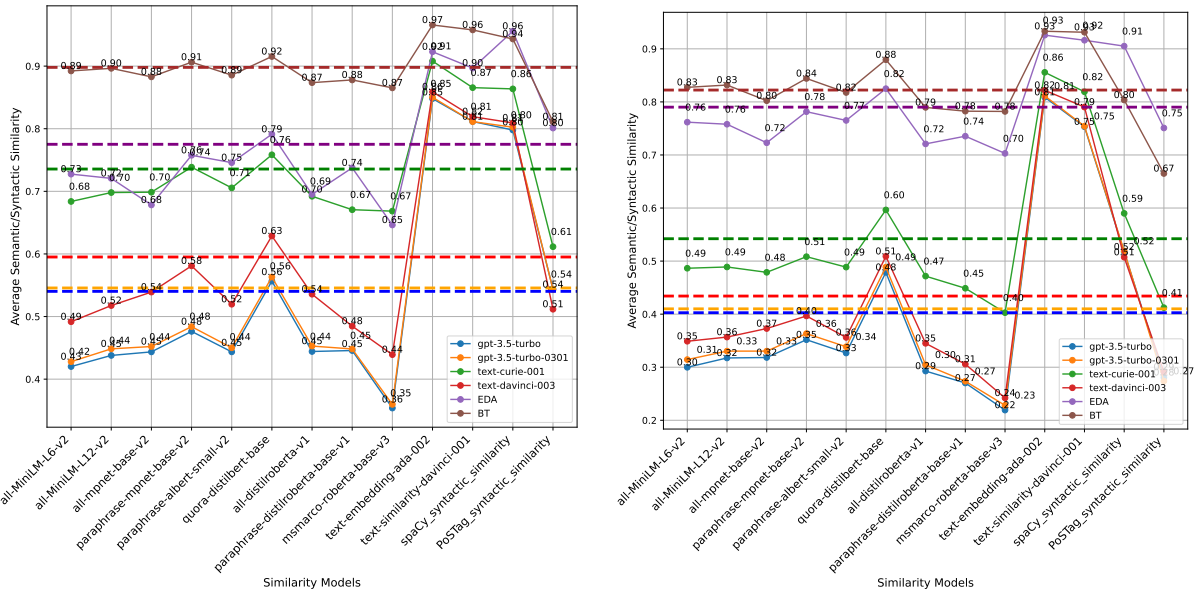
Figure 4: (a): Average Textual Similarity among Original and Augmented Text. (b): Average Textual Similarity among Original and Augmented Explanations.

Table 2: Class-vise classification performance. NS: Number of correctly classified samples, INS: Improvement in NS (in %), OD: Original dataset, AD: Dataset augmented by M1 method.

Class	Type	T-P	T-R	T-F	NS	INS
PA	OD	0.78	0.71	0.74	32	4.44
	AD	0.76	0.76	0.76	34	
IVA	OD	0.67	0.31	0.42	14	17.78
	AD	0.69	0.49	0.57	22	
SA	OD	0.61	0.76	0.67	34	2.22
	AD	0.69	0.78	0.73	35	
SEA	OD	0.53	0.73	0.62	33	2.22
	AD	0.65	0.76	0.70	34	

number of original samples shows a significantly higher improvement in the number of correctly classified samples.

5 Conclusion and Future Scope

In this work, we augment the Reddit posts for a four-class classification problem of determining Wellness Dimensions impacting mental health. The GPT models are outperforming in terms of generating diverse text by preserving the context of the corresponding original text. In future, we plan to experiment with different parameter settings and prompts for generating datasets and develop improved classifiers to determine WD in a well balanced dataset. Furthermore, we will evaluate the classification performance of short explanation text we generated in this dataset.

Ethics and Broader Impact

The data used in this study is obtained from Reddit, a platform designed for anonymous posting, and the user IDs have been anonymized. Furthermore, all sample posts displayed in this study have been obfuscated, paraphrased, and anonymized to protect user privacy and prevent any misuse. As annotation is subjective in nature, we acknowledge that there may be some biases present in our gold-labeled data and the distribution of labels in Wellness Dimensions dataset. We urge researchers to be mindful of the potential risks associated with WD dataset based on personal textual information. To prevent this, human intervention by a moderator is necessary. We acknowledge that we do not release user's metadata and the augmented samples further increase the privacy. The dataset and the source code required to replicate the baseline results can be accessed at Github.[14]

---

[14]https://github.com/drmuskangarg/WellnessDimensions

References

Anusha Bompelli, Yanshan Wang, Ruyuan Wan, Esha Singh, Yuqi Zhou, Lin Xu, David Oniani, Bhavani Singh Agnikula Kshatriya, Joyce (Joy) E Balls-Berry, and Rui Zhang. 2021. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: a scoping review. *Health Data Science*, 2021.

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alana K Dillette, Alecia C Douglas, and Carey Andrzejewski. 2021. Dimensions of holistic wellness as a result of international wellness tourism experiences. *Current Issues in Tourism*, 24(6):794–810.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pretrained language models. *Nature Machine Intelligence*, pages 1–16.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, pages 1–24.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2021. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Mark Ormerod, Jesús Martínez del Rincón, and Barry Devereux. 2021. Predicting semantic similarity between clinical sentence pairs using transformer models: Evaluation and representational analysis. *JMIR Medical Informatics*, 9(5):e23099.

Erzsébet Printz-Markó and Zsuzsanna Ivancsóné Horváth. 2017. Applicability of american wellness research methods in case of central-european countries. In *DIEM: Dubrovnik International Economic Meeting*, volume 3, pages 825–842. Sveučilište u Dubrovniku.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 671–688. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. 2020. Public sentiment on governmental covid-19 measures in dutch social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Robert Weiss. 1975. *Loneliness: The experience of emotional and social isolation.* MIT press.

PDV Charika Wickramarathne, Jeong Chun Phuoc, and Ahmad Rasmi Suleiman Albattat. 2020. A review of wellness dimension models: For the advancement of the society. *European Journal of Social Sciences Studies*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

A   Training and Testing Split

Consider the data containing D documents representing a collection of Reddit posts $\{D =$

Table 3: Language models used to evaluate generate sentence embedding

Base Model	Version
BERT	all-MiniLM-L6-v2
BERT	all-MiniLM-L12-v2
MPNet	all-mpnet-base-v2
MPNet	paraphrase-mpnet-base-v2
Albert	paraphrase-albert-small-v2
DistilBERT	quora-distilbert-base
DistilRoberta	all-distilroberta-v1
DistilRoberta	paraphrase-distilroberta-base-v1
Roberta	msmarco-roberta-base-v3
GPT-3	text-embedding-ada-002
GPT-3	text-similarity-davinci-001

$d_1, d_2, ..., d_n$} where $n = 3092$. For each document d_i, there exist a tuple representing $< E_i, C_i >$ where E_i is text-span/ explanation and C_i is the aspect class for i^{th} instance. Thus, the original WD dataset consists of three columns for 3092 samples: $< D_i, E_i, C_i >$. The aspect class $C_i \in \alpha$ where $\alpha =$[PA, IVA, SA, SEA] and the composition of original dataset contains imbalanced distribution of aspect classes (see Table 4). The number of samples for every class $\alpha[j]$ where $1 \leq j \leq 4$ suggests the need of data augmentation to facilitate development of NLP models over balanced dataset. To this end, we propose the algorithm - *Required annotation count* to decide the number of samples that needs to generated for each WD. Given an input α as a list of the number of text samples for different WD (PA, IVA, SA, SEA) where PA, IVA, SA, SEA defines the count of instances for each class.

As such, our goal is to achieve a balanced dataset by obtaining 1094 samples for each WD, resulting in $1094 * 4 = 4376$ data samples. We observe that all the samples for IVA class must be augmented while no augmentation is required for SA class.

B Semantic Similarity Models

We use 11 different semantic similarity models as shown in Table 3. Sentence Transformers are a set of state-of-the-art language models implemented in Python for generating text embeddings. The two different GPT-3 models used for this task accessed the embeddings API endpoint.

Algorithm 1: Required Augmentation Count

Result: Return $AS=[]$
// AS: augmented sample
Input: α : [PA, IVA, SA, SEA]
Set: $\beta=[]$, R, Red=[], RC=[]
Set: $min_{value} := min(\alpha)$ // Get the record count of minority class
for j in count(α) **do**
  $\beta[j] := max(\alpha) - \alpha[j]$
  /* For each class, get the record count difference from majority class   */
end
// Estimate the size of the test set
$R = min_{value} - max(\beta[j])$
for j in count(α) **do**
  /* For each class, calculate the Reduction Percentage (percentage of reduction after separating the testing set)   */
  $Red[j] = \frac{R}{\alpha[j]} * 100$
  /* For each class, calculate the Reduced Composition (number of training records before augmentation)   */
  $RC[j] = \alpha[j] - Red[j]$
end
/* Get the maximum number of records per class for augmentation   */
$max_{RC} = max(RC)$
for j in count(α) **do**
  /* Augment each class up to the maximum record count */
  $AS[j] = max_{RC} - RC[j]$
end
// return the augmented dataset
return AS

WD	α	Red	RC	AS	Tot.
PA	740	6.0	695	399	1094
IVA	592	7.6	547	547	1094
SA	1139	4.0	1094	0	1094
SEA	621	7.2	576	518	1094

Table 4: The statistics of original composition (α), the reduction percentage (Red), reduced composition (RC), the number of augmented samples (AS) and total number of samples (Tot.)