Entity Based Sentiment Analysis for Textual Health Advice

A Thesis Submitted to The Department of Computer Science of Dartmouth College In Partial Fulfillment Of the Requirements for the Degree Bachelor of Arts in Computer Science

Dae Lim Chung

Spring 2022

Supervisor: Sarah Masud Preum

Entity Based Sentiment Analysis for Textual Health Advice

Dae Lim Chung[1]
Dartmouth College
May 2022

Abstract

This work explores entity based sentiment analysis for textual health advice through deep learning. We fine tuned a pretrained BERT model to analyze sentiments across five different predetermined categories which consist of food, medicine, disease, exercise, and vitality for three different sentiments: positive, negative, and neutral. Original set of annotated medical dataset from Dartmouth College's Persist Lab was used to conduct the experiments. For the aim of tailoring the data for the purpose of entity based sentiment analysis, we explored data transformation techniques to generate optimum training examples. During the experiments, we were able to discover that the wide variety and complexity of terms for the medicine and the disease category posed difficulty on the BERT model which we utilized masking techniques to mitigate. We also demonstrated that our model struggles to learn neutral sentiments in cases where instances of all labels are balanced(in both training and testing). Our system was able to achieve an overall F1 score of 0.739 when conducting the experiment with masked text on a balanced dataset.

[1] I want to express my gratitude to my supervisor Sarah Masud Preum as well as Joseph Gatto and Parker Seegmiller for their guidance and support.

1 Introduction

1.1 Motivation

Although by 2010 digitization of healthcare was on a radical rise, the COVID-19 pandemic has significantly accelerated and necessitated the demand for virtual health care [6]. For instance, a survey done by the UK government's survey on patients indicated that 42% of the patients seek out information online for health advice in their everyday life [7]. Another study done by the NHS on user's motivation and individual habits estimated that nearly half of the user's rely on health advice from the web before consulting a doctor due to the fact that online health information is quick and simple [7]. As the pandemic clearly demonstrated the weakness of current health care infrastructure, consumers are now prioritizing convenient access to care and most importantly increasingly taking charge of their health decisions. Hence, it has become imperative that health systems prioritize improvements in customer engagement for their services in a way that allows customers to extract the information they need as easily as possible.

1.2 Problem Formulation

One of the primary hurdles for achieving maximum user experience with textual health advice comes from the way medical texts are structured. It is often the case that medical advice is convoluted with jargon that can confuse the readers. However, conducting a standard document level sentiment analysis(SA) on medical texts won't resolve this issue as in the context of medical text document level SA won't always yield results that are granular enough to provide true understanding of the text as it is often the case that several entities appear.

Text	Food	Medicine	Disease	Exercise	Vitality
the inflammation associated with rheumatoid arthritis can raise your chances of heart disease, so adding healthy fats to your diet is good for more than your joints, Sandon says, who also has RA. What's more, virgin olive oil contains a compound that has anti-inflammatory properties similar to nonsteroidal medications (your doctor may refer to them as NSAIDs) like ibuprofen. Extra-virgin olive oil comes from the first pressing of olives and can fight nflammation more than refined light versions.	healthy fats virgin olive oil	nonsteroidal meds NSAID	arthritis	N/A	Inflammation

Figure 1:Sample Medical Advice Text

As it is displayed in figure1, which is a row that has been extracted from the initial annotated data set, a short medical text consisting of 3 sentences can harbor as many as 6 entities. From this sample alone we can observe that utilizing document level on the text would be ineffective as in this case, the text describes the condition(disease) "arthritis" as an adverse factor that can raise your chances of heart disease whereas entities such as "healthy fats" and NSAID" that belong to food and medicine category are described positively. In such a manner, we can already pinpoint that when a text is trying to provide medical advice in response to a condition it is expected that it is coupled with opinion regarding both the disease and the treatment where the sentiment of the text towards the disease is expected to negative and positive for entities belonging to the treatment category. Therefore, such frequent occurrences of various entities with mixed sentiments can lead us to question the integrity of document level SA.

1.3 Method

In order to overcome this challenge inherent to medical texts the work looks into entity based sentiment analysis.

Figure 2:Entity Based Sentiment Analysis [8]

Entity based sentiment analysis is a type of sentiment analysis that aims to predict the sentiment expressed in respect to each individual entity within a text as explained by figure2. In this work, we categorize entities into five categories (food, medicine, disease, exercise and vitality) in correspondence with the annotated dataset and analyze the sentiments for each of the categories within a provided medical text with pre-train BERT model. For all categories we assigned three labels to indicate neutral, positive, and negative sentiment. Our experiments have demonstrated that in the case of a balanced dataset the approach is capable of accurately predicting the sentiment across all categories for all three sentiments.

Input Text
It's worth a try if you prefer a natural remedy. Some studies suggest it may ease IBS symptoms. Look for enteric-coated capsules, which are less likely to cause heartburn -- and check with your doctor first if you take other medications.

Entity	Category	Output
natural Remedy	MEDICINE	Positive
enteric-coated capsules	MEDICINE	Positive
IBS symptoms	VITALITY	Neutral
heartburn	VITALITY	Neutral
None	FOOD	Neutral
None	DISEASE	Neutral
None	EXERCISE	Neutral

Figure 3:Sample Input and Output

2. Related Works

Preum et al devise a novel semantic rule based solution which detects conflicting health advice through heterogeneous sources that uses linguistic rules and external knowledge bases [1]. The dominant problem of conflict within textual health advice that the work strives to confront allowed me to see the limitations of document level sentiment analysis on medical advice which consequently encouraged me to explore the effectiveness of aspect level sentiment analysis on capturing the granular sentiment of medical texts.

The aim of sentiment analysis is to summarize an opinion of a text to a single sentiment score. Walaa Medhat, Ahmed Hassan, and Hoda Korashy's survey paper on sentiment analysis algorithms and applications provides a comprehensive overview of SA implementations and applications [2], thereby allowing one to gain a full image of the field. Aspect based SA seeks to improve the accuracy of standard SA by constructing a lexicon consisting of opinion pairs where each opinion word belongs to an aspect(category) of the target entity and is assigned a sentiment label [2]. This technique was introduced in Wu et al in which they compared the performance of entity based SA to general SA and proved its superiority [3].

The invention of transformer architecture by Vaswani et al [4] was able to overcome the limitations of Long-Short-Term-Memory(LSTM) that was prone to overfitting and long training period which consequently gave rise to the development of BERT model by Delvin et al [5]. The BERT model was able to clear benchmarks in numerous NLP tasks and its state of art performance made it the model of choice for our implementation.

3. Models

Pre-trained BERT based model was used to conduct the classification task but before delving into BERT it is important to understand the Transformer architecture which lays the foundation for the BERT model.

3.1 Transformer

The Transformer architecture was introduced in the paper “Attention Is All You Need”, Vaswani et al. [4] and is used as the encoder in BERT. Although in the paper Transformer is composed of two parts, the Encoder and the Decoder, only the architecture of the Encoder is relevant to this study as the BERT model only utilizes the Encoder.

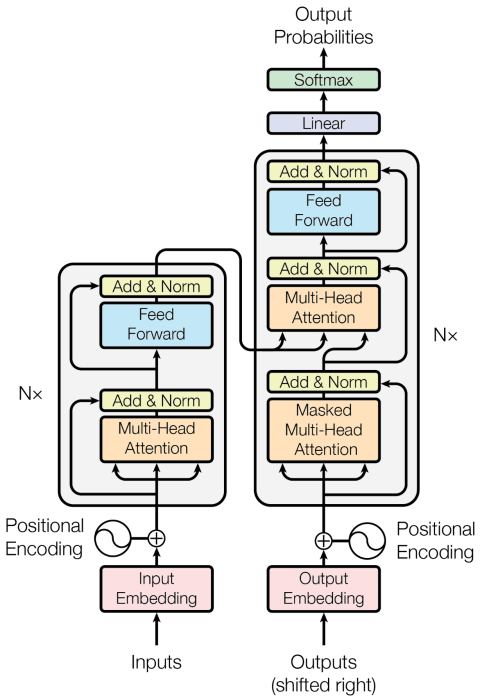

Figure 4:Transformer Architecture [4]

The transformer as displayed in figure 4 consists of an encoder-decoder architecture, with encoder being on the left side and decoder on the right.

The encoder maps a sequence of inputs to a vector that is the same in dimensions as the token embeddings(which acts as vector look-up for each token). For each token a corresponding vector is looked up, stacking each of the vector to obtain a matrix of dimensions: $input\ length\ \times\ token\ embeddings$. Once there is a matrix representation of the sequence, positional encodings are applied to modify the meaning of a word depending on its position. A total of N encoder blocks are stacked together to generate the Encoder's output. This iterative approach helps the neural network extract complex relationships between words in the input sentence as it is iteratively building the meaning of the input sequence as a whole. It is important to note that the powerfulness of the transformer is heavily dependent on the self attention mechanism which mirrors the human behavior of paying more attention to certain words in a sentence than others.

3.2 BERT

The Bidirectional Encoder Representation from Transformers(BERT) was first proposed by Devlin et al [5] as a general language representation model and is essentially a chain of multiple Transformer encoders. Although adding bidirectionality while using the Transformer Encoder to train a language model is unfeasible, BERT was able to achieve it by pretraing the model into the masked language model(MLM) and next sentence prediction(NSP). By overcoming the limitation of unidirectionality BERT is able to extract richer contextual feature representations compared to a unidirectional approach and thus has been the model of choice for various downstream tasks since its introduction.

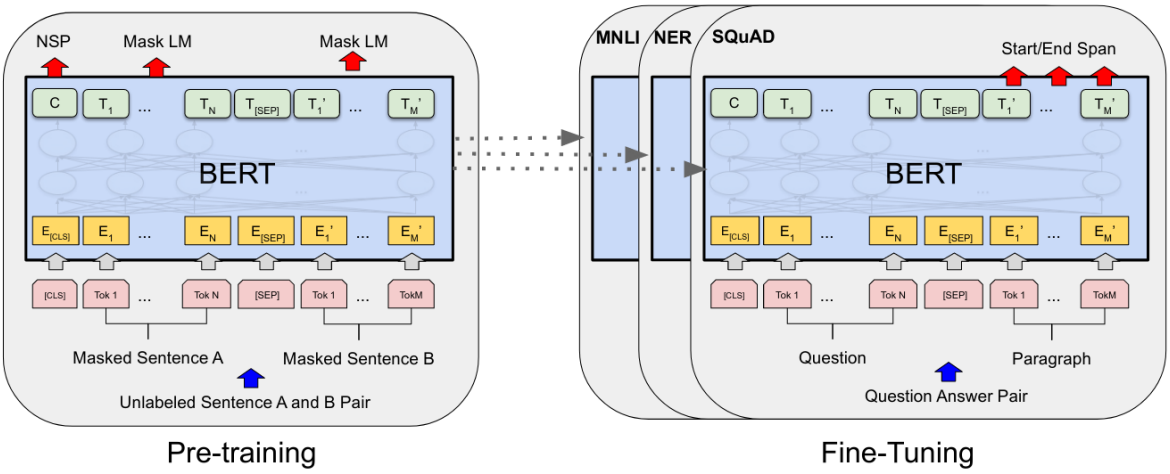

Figure 5:Pre-training of BERT [5]

The MLM demands the model to predict random words from within the sequence. For the case of BERT, 15% of the words that were fed in as input were masked [5]. However, not all tokens were masked in the same manner: 80% were placed by the masked token, 10% were replaced by a random token and the remaining 10% were left intact [5]. This approach allowed for the model to learn a contextual representation of all the words in an input sentence instead of having the model only focus on the contextual representation of the mask token.

The additional pre-training step with NLP is due to the importance of the model knowing how to relate two different sentences to perform downstream tasks e.g natural language inference. Because pretraining with MLM alone does not capture this knowledge, pre-training with NLP notably increased performance for downstream tasks. The pre-training corpus was constructed from BookCorpus and English Wikipedia and for pre-training sequences, the author randomly sampled batches of two, where 50% of sentences were pairs with actual adjacent sentences [5].

4. Dataset and Preprocessing

In total four experiments were conducted:masked text with unbalanced dataset, unmasked text with unbalanced dataset, masked text with balanced dataset, and unmasked text with balanced dataset. For all experiments, the transformer used was ,bert-base-uncased, and BertForSequenceClassification (a hugging face model) was employed to carry out the classification task. All models were trained on Google Colab Pro's High-RAM environment using a single NVIDIA P100 GPU. The following hyper parameters were used during training: batch size of 4, weight decay of 0.001, and learning rate of 2e-5. All other hyper-parameters were set to their default values.

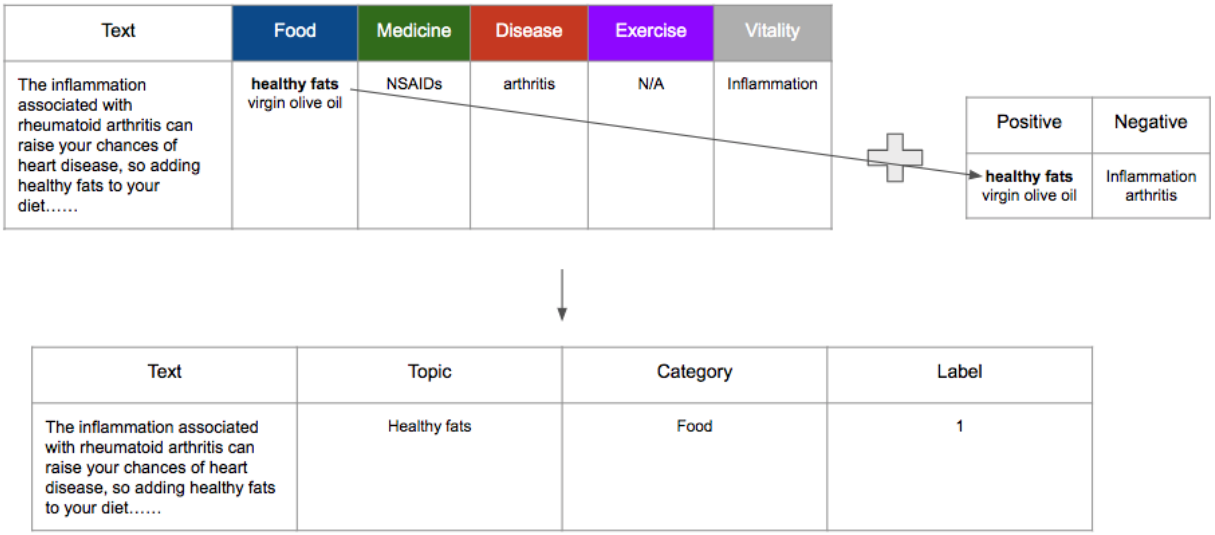

Figure 6:Data Transformation

5. Experiments

5.1 Masked Text with Unbalanced Dataset

The model was trained with the training dataset ($N = 4593$) and was evaluated using the test dataset ($N = 1956$). The labels for the training dataset were largely imbalanced as most instances were concentrated on label1(neutral sentiment). Overall, there were 3844 instances for

label1, 379 instances for label 0(negative sentiment), and 340 instances for label2(positive sentiment). Masked text(text with the entity masked with **[category]**) and the category of the entity were tokenized as input for the model. Figure 7 demonstrates the masking process for raw text where the entity "antinuclear antibodies" which was labeled as medicine is masked with **[medicine].** In instances where the entity appeared multiple times all instances were masked.

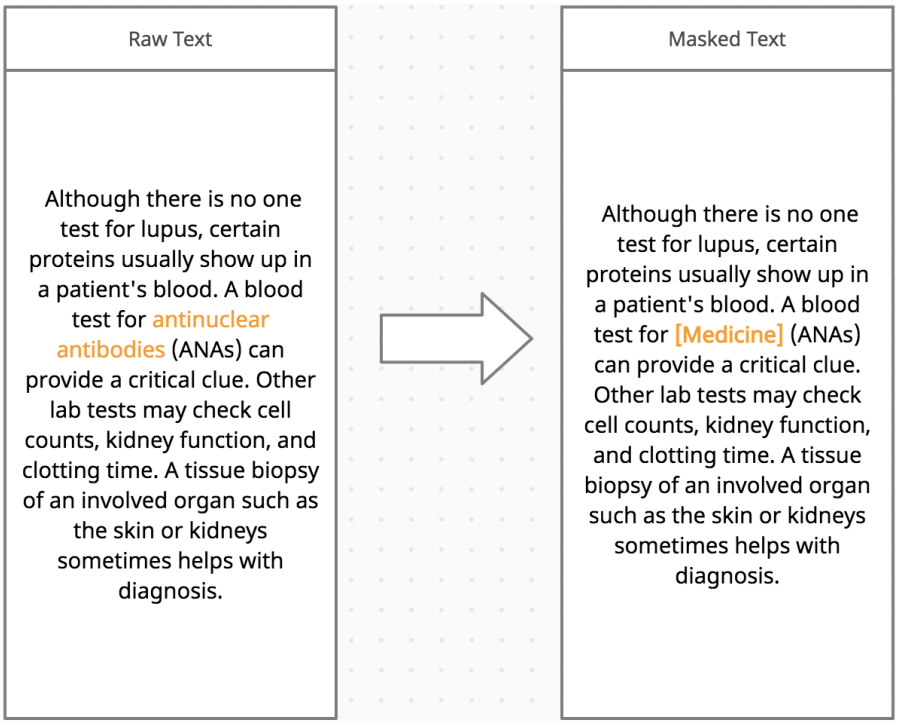

Figure 7:Masking Raw Text

5.2 Unmasked Text with Unbalanced Dataset

The sizes of training and test data were identical to the previous experiment and the imbalance in labels were similar, 1:10 ratio for label 0 and 2 against label1. There were 5491 instances for label 1, 543 instances for label 0, and 485 instances for label1. The key difference is that raw text was tokenized as input for the model in place of masked text.

5.3 Masked Text with Balanced Dataset

The model was trained with the training dataset($N = 987$) and was evaluated using the test dataset($N = 423$). This time the data was processed so that the distribution of labels in the training and test set were uniform. The balancing was done by first splitting the whole dataset into training and test data then for each, matching instances of two labels with the size of the label with the smallest number of instances. For example, if label 1 had the smallest number of instances within the dataset with 130 instances the remaining two labels, label 0 and label 2, would also be reduced to 130 instances to make the distribution of labels uniform. For the training data, each label was balanced to 329 instances and for the testing it was balanced to 146 instances. Masked text and the category of the entity were tokenized as input for the model.

5.4 Unmasked Text with Balanced Dataset

The model was trained with the training dataset($N = 906$) and was evaluated using the test dataset($N = 513$). The data was processed so that the distribution of labels in the training and test set were uniform which was done in an identical approach as the previous experiment. For the training data, each label was balanced to 302 instances and for testing it was balanced to 171 instances. Raw text and the category of the entity were tokenized as input for the model.

6. Results

This section presents the performance for the four experiments conducted. Table1 and 2 reports the precision, recall, and F1 score for each of the categories. Table 3 and 4 summarizes the performance of the model for each of the three labels(negative, neutral and positive). At a glance, we can see that our system is displaying good performance with F1 scores for

Masked+Unbalanced, Unmasked+Unbalanced, Masked+Balanced, and Unmasked+balanced being 0.897, 0.885, 0.739, and 0.732 respectively.

	Masked+ Unbalanced			**Unmasked+ UnBalanced**		
	F1	Precision	Recall	F1	Precision	Recall
Food	0.943	0.942	0.945	0.922	0.922	0.924
Medicine	0.835	0.833	0.843	0.825	0.823	0.833
Disease	0.821	0.827	0.833	0.854	0.858	0.864
Exercise	0.987	0.987	0.987	0.974	0.980	0.972
Vitality	0.961	0.960	0.963	0.943	0.937	0.952

Table 1

	Masked+ Balanced			**Unmasked+ Balanced**		
	F1	Precision	Recall	F1	Precision	Recall
Food	0.844	0.849	0.842	0.744	0.752	0.746
Medicine	0.785	0.787	0.794	0.703	0.702	0.711
Disease	0.717	0.715	0.735	0.644	0.655	0.693
Exercise	0.977	0.980	0.977	0.918	0.931	0.923
Vitality	0.755	0.763	0.767	0.803	0.853	0.836

Table 2

	Masked+ Unbalanced			Unmasked+ UnBalanced		
	F1	Precision	Recall	F1	Precision	Recall
Overall	0.897	0.895	0.9	0.885	0.883	0.892
Negative	0.691	0.755	0.638	0.522	0.500	0.522
Neutral	0.946	0.933	0.959	0.964	0.962	0.965
Positive	0.567	0.614	0.526	0.498	0.520	0.478

Table 3

	Masked+ Balanced			Unmasked+ Balanced		
	F1	Precision	Recall	F1	Precision	Recall
Overall	0.739	0.743	0.740	0.732	0.734	0.734
Negative	0.748	0.700	0.804	0.827	0.827	0.827
Neutral	0.677	0.738	0.628	0.672	0.632	0.719
Positive	0.797	0.797	0.797	0.766	0.744	0.796

Table 4

7. Error Analysis and Discussion

One of the first observations that we can make from table 3 is the fact that intuitively there is a noticeable difference between the F1 score of the neutral label versus the positive and negative label. The primary cause behind this behavior seems to be overfitting of the model on neutral labels which is not unexpected when we consider the fact that the imbalance between the positive and negative labels and the neutral label was approximately 1:10 for both experiments conducted with unbalanced dataset. This explanation is further affirmed by the result in table 4 presenting a relatively uniform F1 score when experiments were conducted with balanced labels.

What's also interesting is the fact that the F1 score for the neutral label is the lowest out of the three labels in the F1 score when the dataset is balanced which confirms the result in table3 was indeed due to overfitting. Hence, it seems like as long as the datasets are balanced, the model does not have particular difficulty in learning the relationship of a specific sentiment.

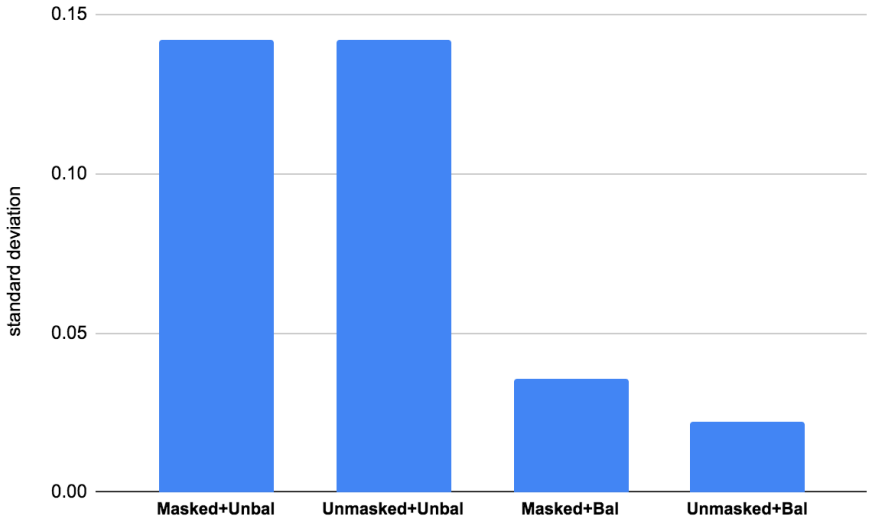

Figure 8:Standard Deviation of F1 Scores across Labels

In terms of category, it is clear to see that medicine and disease consistently have the lowest performance across all experiments. Our speculation is that the sheer variety and obscurity of the terms for these two categories makes it more difficult for BERT to learn the relationship between the sentiment and text compared to others. To support this hypothesis, we can observe that for the case of experiments conducted with balanced dataset, the F1 score of medicine and disease show much less disparity in comparison to other categories when the target entities were masked instead of being fed raw. Moreover, when looking at the class distribution of wrongly predicted labels we see that from unmasked to masked, the percentage for both disease and medicine decrease with the change in medicine being arguably significant (46.7% to 43.4%).

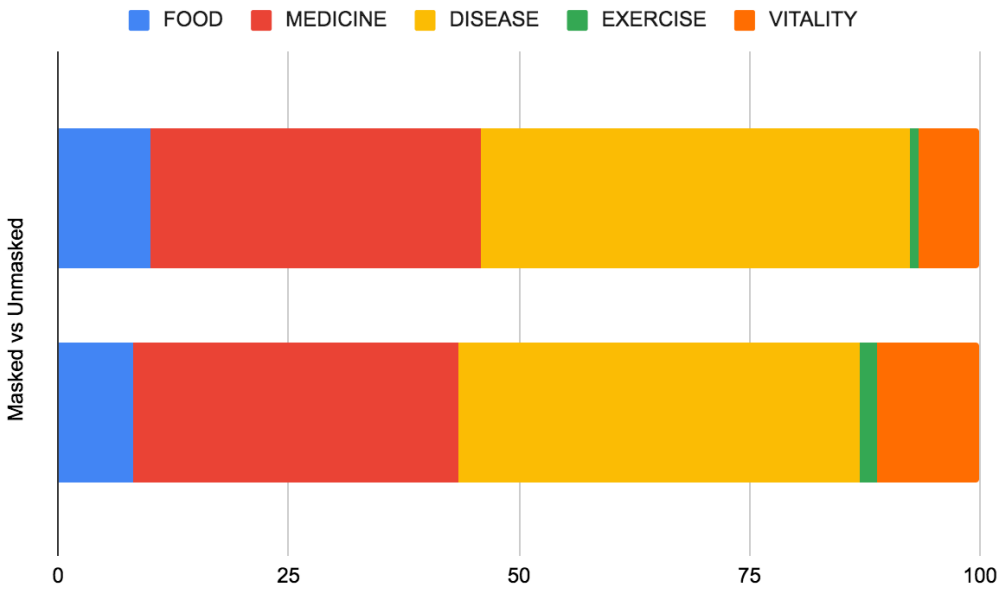

Figure 9: Graph showing the class distribution of wrongly predicted labels Above(Unmasked) / Below(Masked)

7. Conclusion

7.1 Limitations and Future Work

There are limitations to this work that are worth highlighting. When observing the F1 score of the three labels for the two experiments with the balanced dataset, we see that the F1 score lies in the low 0.7 range which is good but not ideal. As we are solely exploring our method based on the BERT model, we believe that our current performance indicates that we could have applied another powerful transformer model to this problem (e.g XLNET).

Aside from the choice of models, our system's performance on neutral labels under a balanced dataset is also something that needs further questioning. Although the disparity between the

performances of neutral labels and the remaining two is not as dramatic as it was for unbalanced dataset we can still observe that the model is struggling to learn neural sentiments in particular. We speculated that it may be due to the fact that we annotate a category as neutral when a text does not have any entity that belongs to the category but when conducting trial experiments with such annotations eliminated we were unable to see any meaningful improvements. Therefore, we believe that further exploration in regards to this matter is imperative for meaningful model improvements.

Finally, another problem that must be addressed is pinpointing the reason behind the low performance of medicine and disease category. Although we speculate that obscurity and unfamilitary of the terms of these two categories for the BERT model is part of the cause, we believe that an even more rigorous analysis must be conducted for the two categories to discover the principle cause which would allow implementations that would significantly improve our method.

7.2 Summary

In this work, we proposed entity based sentiment analysis for textual health advice with the aim of capturing granular sentiments for multiple entities which corresponds well with the nature of medical texts. We first introduced how we transformed our raw data to allow the model to learn the textual sentiment across our predetermined categories. We also demonstrated how our method based on pre-trained BERT model was able to predict sentiment of each category and label with satisfying accuracy without too much discrimination under the assumption that the dataset is balanced. Most importantly, we were able to observe how obscurity of terms for certain

categories hinder the learning for BERT model which could be overcomed to a certain extent through masking as demonstrated by our experiments.

References

[1] Preum, S., 2017. Preclude: Conflict Detection in Textual Health Advice. *2017 IEEE International Conference on Pervasive Computing and Communications*,. (2017)

[2] Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp.1093-1113. (2014)

[3] S.Wu, F.Wu, Y.Chang, C.Wu, Y.Huang Automatics construction of target-specific sentiment lexicon. *Expert Systems with Applications,* 116, pp.285-298, (2019)

[4] Vaswani, Ashish et al. "Attention Is All You Need". *31St Conference On Neural Information Processing Systems (NIPS 2017)*, (2017)

[5] Devlin, Jacob et al. "BERT: Pre-Training Of Deep Bidirectional Transformers For Language Understanding". *Association For Computational Linguistics*, pp. 4171-4186, (2019)

[6]Appleby, Chuck et al. "Digital Transformation: From A Buzzword To An Imperative For Health Systems". *Deloitte Insights*, (2021)

[7]Battineni, Gopi et al. "Factors Affecting The Quality And Reliability Of Online Health Information". *Digital Health*, vol 6, pp. 1-11, (2020)

[8] Noble, Dan et al."Label Text For Aspect-Based Sentiment Analysis Using Sagemaker Ground Truth". *AWS Machine Learning Blog*, (2022)