2023

Medical Concept Mention Identification in Social Media Posts using a Small Number of Sample References

Vasudevan Nedumpozhimana
vasudevan.nedumpozhimana@tudublin.ie

Sneha Rautmare
Trinity College Dublin, Ireland, sneha.rautmare@adaptcentre.ie

Meegan Gower
Trinity College Dublin, Ireland, meegan.gower@adaptcentre.ie

See next page for additional authors

Funder: ADAPT Centre

Medical Concept Mention Identification in Social Media Posts using a Small Number of Sample References

Vasudevan Nedumpozhimana[1], **Sneha Rautmare**[2], **Meegan Gower**[2], **Maja Popović**[3], **Nishtha Jain**[2,5], **Patricia Buffini**[3], **John Kelleher**[1,4]
ADAPT Centre
[1]Technological University Dublin, [2]Trinity College Dublin, [3]Dublin City University, [4]Maynooth University,
([5]now at Spoke.ai, Berlin, Germany)
name.surname@adaptcentre.ie

Abstract

Identification of mentions of medical concepts in social media text can provide useful information for caseload prediction of diseases like Covid-19 and Measles. We propose a simple model for the automatic identification of the medical concept mentions in the social media text. We validate the effectiveness of the proposed model on Twitter, Reddit, and News/Media datasets.

1 Introduction

Caseload information of diseases like Covid-19 and Measles are likely reflected in social media posts in the form of mentions of relevant medical concepts. For example, increase in the mentions of medical concepts like *fever*, *headache*, *cough*, *loss of smell etc.* in social media text is a potential indication of increasing covid caseloads. Therefore models which identify the mentions of medical concepts from social media text can provide useful features for the caseload prediction of such diseases.

State-of-the-art natural language processing techniques mostly rely on huge pre-trained language models and such models can be utilized for identifying the mentions of medical concepts in social media texts. In this work, we propose a simple and effective model to automatically identify the presence of 24 selected medical concepts in social media text by using a small number of reference texts and a pre-trained language model.

2 Related Works

The basis for the medical concept mention identification method carried out in this work is the research on medical concept normalization. In the literature, the medical concept normalization problem is addressed by using different approaches. Traditionally lexicon-based string-matching approaches and rule-based approaches are used for medical concept normalization. For example, Aronson and Lang (2010) used a knowledge-intensive approach for concept normalization which is based on symbolic language processing.

Leaman et al. (2013) approached the medical concept normalization problem by learning the similarity between mentions and concept names. Limsopatham and Collier (2015) approached this medical concept normalization as a phrase-based machine translation problem and they translated social media phrases into formal medical concepts.

In another approach, Limsopatham and Collier (2016) used simple deep-learning-based models like CNN, and RNN with pre-trained LM and improved the performance of the medical concept normalization. Lee et al. (2017) further improved this performance by refining the dataset and leveraging the neural embeddings of health-related text.

Bornet et al. (2023) showed that language models can learn the semantics of medical concepts. They found that subword information is crucial for learning medical concept representation and global word co-occurance information is more useful for downstream tasks using these representations. This suggests the suitability of language models that have both subword information and global co-occurrence information for medical concept normalization.

More recently Kalyan and Sangeetha (2020) used the transformer-based BERT pre-trained language model for the medical concept normalization. In this method, they generated the embeddings of concepts and mentions by using the pre-trained RoBERTa language model (Liu et al., 2019). They further enriched concept embeddings using synonym information by using a retrofitting method proposed by Faruqui et al. (2015). Then the relations between concepts and mentions are calculated by using cosine similarity between their embeddings. Xu and Miller (2022) also proposed a similar and simple model for medical concept normalization by using pre-trained language model

Proceedings of Recent Advances in Natural Language Processing, pages 777–784
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_084

SAPBERT and cosine similarity. Based on these approaches we formulated our new model to extract medical concept features from social media text. However, in our approach, instead of retrofitting the concept embedding by using synonyms, we use information from manually selected positive and negative samples along with synonyms information and propose a novel closed-form optimization formulation for generating concept representations.

3 The Proposed Model

Our proposed model to automatically identify the mentions of a set of medical concepts from the social media text is inspired by the medical concept normalization model proposed by Kalyan and Sangeetha (2020). The proposed model utilizes a small number of preselected positive and negative samples along with the name and synonyms of the medical concepts to learn an anchor vector representation (distributed representation) of each of these concepts. In order to learn the anchor vector of concepts we first generate distributed representations of all the selected positive and negative samples, name of the concept, and its synonyms.

Then we will learn an anchor vector for each concept (V_c) by solving the optimization with the following objectives:

1. Cosine similarity between V_c and the distributed vector representations of positive samples of concept should be maximum. That is, maximize $cos(V_c, V_{ps})$, where, V_{ps} is the distributed vector of any positive sample of the concept

2. Cosine similarity between V_c and the distributed vector representations of negative samples of concept should be minimum. That is, minimize $cos(V_c, V_{ns})$, where, V_{ns} is the distributed vector of any negative sample of the concept

3. Cosine similarity between V_c and the distributed vector representations of its synonyms should be maximum. That is, maximize $cos(V_c, V_{ss})$, where, V_{ss} is the distributed vector of any synonyms of the concept

4. Cosine similarity between V_c and the distributed vector representations of its name should be maximum. That is, maximize $cos(V_c, V_n)$, where, V_n is the distributed vector of the name of the concept

We formulated this multiobjective optimization problem as a single objective optimization by defining a single aggregate objective function by taking the weighted sum of these objectives. The final objective will be:

Maximize $\{cos(V_c, V_n) + \lambda_p \sum_{ps} cos(V_c, V_{ps}) - \lambda_n \sum_{ns} cos(V_c, V_{ns}) + \lambda_s \sum_{ss} cos(V_c, V_{ss})\}$

Where $\lambda_p, \lambda_n, \lambda_s$ are positive weights corresponding to each of three objectives and without loss of generality we can set the weight corresponding to the fourth objective (first term in the single aggregate objective) as 1.

If we add a constraint that the V_c is a unit vector we will get a nice closed-form solution for this optimization problem, which is:

$$V_c = \frac{V_n + \lambda_p \sum_{ps} V_{ps} - \lambda_n \sum_{ns} V_{ns} + \lambda_s \sum_{ss} V_{ss}}{1 + \lambda_p + \lambda_n + \lambda_s}$$

This closed-form solution enables us to learn the anchor vector representation of each of the concepts by calculating the exact solution for the proposed optimization problem with linear time complexity. The samples used to learn the proposed model are very small and therefore we can easily learn the anchor vector representations of concepts without much computational resources.

Once we learn the anchor vector representations of each of the concepts, we can easily calculate the components of these concepts in any of the social media texts by taking the cosine similarity between the distributed representation of the social media text and the anchor vector representation of the corresponding concept.

4 Experimentations

Medical concepts considered for this work are based on the Covid-19[1] and Measles[2] symptoms mentioned by the World Health Organisation. We selected 24 key medical concepts related to symptoms of Covid-19 or Measles.

For each of the selected 24 medical concepts, we manually identified a set of synonyms. We used two sources of information for this process, first we consulted the SNOMED CT Browser[3], and then we also manually reviewed the top webpages returned

[1] https://www.who.int/health-topics/coronavirus

[2] https://www.who.int/health-topics/measles

[3] https://browser.ihtsdotools.org/

in response to a general web search using the medical concept as the keyword to identify potential synonyms. The number of synonyms identified ranged from a minimum of 6 synonyms for the concept *cough* to 41 synonyms for the concept *loss of mobility*[4]. Then we manually collected 10 positive and 10 negative sample tweets for each of the 24 medical concepts. A tweet which contains the mention of a medical concept is selected as the positive sample for that concept. In order to select the negative samples we considered tweets which can be misinterpreted as a positive sample. For example, a tweet which contains the term 'pink-eye shadow' may be misinterpreted as being relevant to the medical concept *conjunctivitis* due to the relation to 'pink-eye'. But the term 'pink-eye shadow' is not related to the medical concept *conjunctivitis* and therefore explicitly providing the information that this social media text is not related to *conjunctivitis* will be helpful for the model. Therefore we selected such misinterpretable samples as negative samples for all 24 medical concepts. If for a given medical concept we can't find 10 tweets that contain mentions to concepts that can be confused with the target concept then we select arbitrary samples which are not related to the corresponding medical concept as the remaining negative samples.

We generated 768-dimensional distributed representations of concept name, synonyms, and it's manually selected 10 positive sample tweets and 10 negative sample tweets of each of the 24 medical concepts by using a pre-trained sentence BERT language model (*all-mpnet-base-v2*) (Song et al., 2020). Then we learned the 768-dimensional anchor vector representations corresponding to each of these 24 concepts. We used this generated anchor vector representation for all our experiments on Twitter, Reddit, and News/Media datasets.

4.1 Cosine similarity between concept representations

As part of the evaluation of the proposed model, first, we analysed how the learned anchor vector representations of concepts are located in the embedding space by measuring cosine similarities between all pairs of anchor vectors. If all anchor vectors are located together in the embedding space then the cosine similarity between all concept pairs will be high (close to 1). We are also interested in whether the medical concepts are well separated from non-medical concepts and to investigate this we introduced a separate non-medical concept for this evaluation. Anchor vectors of the non-medical concepts are learned in a similar way the medical concepts are learned and for learning this anchor vector we considered 10 positive samples which are not related to any of the selected medical concepts and 10 negative samples which are related to at least one of the selected medical concepts. The heat map of cosine similarities between all pairs of these 25 concepts (24 medical concepts and one non-medical concept) is shown in Table 1. The cosine similarities between many of the concept pairs are small which indicates that the concepts are well distributed in the embedding space. The cosine similarities between the non-medical concept and all medical concepts are very small, less than 0 for many cases, which shows that there is a clear separation between medical and non-medical concepts in the embedding space.

4.2 AUC-ROC Evaluation

To evaluate our model we manually annotated a sample of 1017 tweets, where each tweet contained mentions for at least one of the 24 medical concepts. We have not considered the non-medical concept because our primary focus here is to evaluate how the proposed model performs on 24 medical concepts. We annotated each of these 1017 sample tweets with binary labels for each concept, that is, if a sample is mentioned to a particular concept then it is annotated as 1 for that concept and otherwise 0. So, at the end of this annotation process, each sample had 24 binary labels associated with it. We then adapted our proposed model to act as a multi-label classifier by considering the 24 medical concepts as labels. For each sample tweet, we calculated cosine similarities between the distributed representation of the sample tweet and the anchor vector of each of the 24 medical concepts and treated the cosine similarity score between the representation of the sample and a concept's anchor vector as the prediction probability corresponding to that concept. From these prediction probabilities, we calculated the AUC-ROC score for each of the 24 medical concepts and these are also shown in Table 2. We also generated a box plot from these AUC-ROC scores and showed it in Fig. 1. For

[4]Note, for the cumulative gain experiments we report later we did an analysis of whether the number of synonyms identified for a concept affected the performance of our model and we found weak negative correlations between the number of synonyms and the cumulative gain.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
aches and pains	1	1.00	0.63	0.47	0.46	0.52	0.70	0.54	0.40	0.53	0.63	0.68	0.73	0.36	0.59	0.45	0.28	0.53	0.54	0.48	0.37	0.47	0.39	0.62	0.65	0.00
chest pain	2	0.63	1.00	0.40	0.37	0.66	0.47	0.79	0.26	0.45	0.44	0.54	0.63	0.34	0.40	0.44	0.37	0.48	0.58	0.57	0.59	0.37	0.47	0.65	0.61	-0.10
confusion	3	0.47	0.40	1.00	0.46	0.47	0.44	0.53	0.34	0.40	0.65	0.50	0.61	0.32	0.51	0.50	0.48	0.52	0.48	0.47	0.36	0.32	0.40	0.40	0.39	0.02
conjunctivitis	4	0.46	0.37	0.46	1.00	0.43	0.43	0.38	0.47	0.52	0.42	0.47	0.52	0.49	0.28	0.37	0.22	0.43	0.54	0.59	0.37	0.60	0.49	0.47	0.57	-0.04
cough	5	0.52	0.66	0.47	0.43	1.00	0.50	0.70	0.28	0.54	0.45	0.67	0.57	0.42	0.39	0.51	0.41	0.58	0.75	0.79	0.70	0.38	0.70	0.76	0.60	-0.09
diarrhoea	6	0.70	0.47	0.44	0.43	0.50	1.00	0.53	0.28	0.44	0.55	0.60	0.56	0.34	0.45	0.53	0.20	0.61	0.54	0.52	0.31	0.40	0.38	0.57	0.49	-0.01
difficulty breathing	7	0.54	0.79	0.53	0.38	0.70	0.53	1.00	0.34	0.41	0.62	0.62	0.55	0.33	0.54	0.57	0.46	0.59	0.64	0.63	0.64	0.33	0.55	0.60	0.55	-0.07
discolouration of skin	8	0.40	0.26	0.34	0.47	0.28	0.28	0.34	1.00	0.31	0.35	0.44	0.32	0.45	0.43	0.37	0.31	0.36	0.28	0.34	0.24	0.61	0.31	0.23	0.40	-0.07
ear infections	9	0.53	0.45	0.40	0.52	0.54	0.44	0.41	0.31	1.00	0.39	0.49	0.59	0.40	0.30	0.42	0.30	0.48	0.61	0.60	0.42	0.40	0.41	0.59	0.60	-0.03
fatigue	10	0.63	0.44	0.65	0.42	0.45	0.55	0.62	0.35	0.39	1.00	0.61	0.54	0.19	0.61	0.47	0.30	0.47	0.48	0.42	0.34	0.31	0.32	0.43	0.44	0.10
fever	11	0.68	0.54	0.50	0.47	0.67	0.60	0.62	0.44	0.49	0.61	1.00	0.63	0.38	0.45	0.47	0.29	0.56	0.61	0.63	0.54	0.47	0.49	0.62	0.64	0.09
headache	12	0.73	0.63	0.61	0.52	0.57	0.56	0.55	0.32	0.59	0.54	0.63	1.00	0.35	0.47	0.47	0.35	0.54	0.65	0.57	0.37	0.41	0.46	0.63	0.61	0.01
Koplik spots in mouth	13	0.36	0.34	0.32	0.49	0.42	0.34	0.33	0.45	0.40	0.19	0.38	0.35	1.00	0.18	0.31	0.24	0.43	0.38	0.51	0.39	0.61	0.38	0.44	0.54	-0.17
loss of mobility	14	0.59	0.40	0.51	0.28	0.39	0.45	0.54	0.43	0.30	0.61	0.45	0.47	0.18	1.00	0.43	0.53	0.42	0.32	0.30	0.36	0.33	0.31	0.33	0.39	0.01
loss of smell	15	0.45	0.44	0.50	0.37	0.51	0.53	0.57	0.37	0.42	0.47	0.47	0.47	0.31	0.43	1.00	0.39	0.87	0.66	0.62	0.40	0.32	0.49	0.47	0.42	-0.12
loss of speech	16	0.28	0.37	0.48	0.22	0.41	0.20	0.46	0.31	0.30	0.30	0.29	0.35	0.24	0.53	0.39	1.00	0.43	0.28	0.30	0.39	0.24	0.32	0.33	0.36	-0.15
loss of taste	17	0.53	0.48	0.52	0.43	0.58	0.61	0.59	0.36	0.48	0.47	0.56	0.54	0.43	0.42	0.87	0.43	1.00	0.67	0.62	0.48	0.38	0.45	0.59	0.53	-0.16
nasal congestion	18	0.54	0.58	0.48	0.54	0.75	0.54	0.64	0.28	0.61	0.48	0.61	0.65	0.38	0.32	0.66	0.28	0.67	1.00	0.88	0.55	0.35	0.69	0.74	0.59	-0.11
nasal discharge	19	0.48	0.57	0.47	0.59	0.79	0.52	0.63	0.34	0.60	0.42	0.63	0.57	0.51	0.30	0.62	0.30	0.62	0.88	1.00	0.62	0.49	0.79	0.72	0.61	-0.07
pneumonia	20	0.37	0.59	0.36	0.37	0.70	0.31	0.64	0.24	0.42	0.34	0.54	0.37	0.39	0.36	0.40	0.39	0.48	0.55	0.62	1.00	0.31	0.49	0.52	0.50	-0.18
rash	21	0.47	0.37	0.32	0.60	0.38	0.40	0.33	0.61	0.40	0.31	0.47	0.41	0.61	0.33	0.32	0.24	0.38	0.35	0.49	0.31	1.00	0.44	0.40	0.60	-0.06
sneezing	22	0.39	0.47	0.40	0.49	0.70	0.38	0.55	0.31	0.41	0.32	0.49	0.46	0.38	0.31	0.49	0.32	0.45	0.69	0.79	0.49	0.44	1.00	0.51	0.43	-0.12
sore throat	23	0.62	0.65	0.40	0.47	0.76	0.57	0.60	0.23	0.59	0.43	0.62	0.63	0.44	0.33	0.47	0.33	0.59	0.74	0.72	0.52	0.40	0.51	1.00	0.72	0.01
swollen glands	24	0.65	0.61	0.39	0.57	0.60	0.49	0.55	0.40	0.60	0.44	0.64	0.61	0.54	0.39	0.42	0.36	0.53	0.59	0.61	0.50	0.60	0.43	0.72	1.00	-0.06
Non-Medical Concept	25	0.00	-0.10	0.02	-0.04	-0.09	-0.01	-0.07	-0.07	-0.03	0.10	0.09	0.01	-0.17	0.01	-0.12	-0.15	-0.16	-0.11	-0.07	-0.18	-0.06	-0.12	0.01	-0.06	1.00

Table 1: Cosine similarities between concept representations

most of the medical concepts we found above 90% AUC-ROC scores and for many concepts, we got more than 95%. All medical concepts achieved more than 85% AUC-ROC and the average score is 93.91%. This validates the effectiveness of the proposed model on the Twitter dataset.

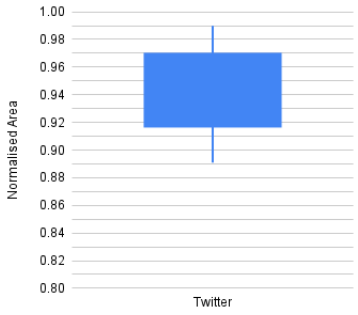

Figure 1: The boxplot of the area under ROC curve for 24 medical concepts on the Twitter dataset

4.3 Cumulative gain evaluation

The AUC-ROC evaluation required a sufficient number of manually annotated samples and therefore that evaluation method is not easily extendable to other social media. In order to validate the effectiveness of the proposed model across different social media sources, we adopted a cumulative gain evaluation method by using a very small set of manually selected samples. First, we selected a small set of positive samples (10 samples) corresponding to every 24 medical concepts. To evaluate each medical concept, we inserted the selected positive samples corresponding to that concept into a large set of random samples (around 100,000 samples). Then we calculated the cosine similarity between the anchor vector representation of that concept and each sample in the dataset. Then we sorted the samples based on cosine similarity so that samples which contain the mentions of the medical concept will come earlier.

We then check the position of the selected samples in the sorted order. If the model generates a high cosine similarity value for the selected positive samples then they should come earlier in the sorted list. We select the first k samples from the sorted samples and check the cumulative gain, that is how many of the inserted samples are in the first k. We increase the k and see how quickly the model achieves 100% cumulative gain. We then plot this cumulative gain chart where the x-axis is the k (number of samples from sorted sample set) and the y-axis is the percentage of cumulative gain. If the model is performing well then the cumulative gain chart will reach 100% quickly. The faster a cumulative gain chart rises to 100% (i.e., the lower the number of samples k that a model needs to retrieve all the positive examples) the better the model. Consequently, the larger the area under the cumulative curve for a model the better the model. Therefore we use the area under the cumulative gain curve as the performance metric to evaluate the proposed model.

4.3.1 Evaluation on Twitter dataset

To evaluate the performance of the proposed model on the Twitter dataset by using the cumulative gain

Medical Concepts	AUC-ROC	AUC-CG
aches and pains	0.8645	98.30
chest pain	0.9272	99.94
confusion	0.9311	99.52
conjunctivitis	0.9706	99.95
cough	0.9319	99.96
diarrhoea	0.9260	99.90
difficulty breathing	0.9164	99.71
discolouration of skin	0.9695	99.94
ear infections	0.9898	99.93
fatigue	0.8909	98.58
fever	0.9148	99.41
headache	0.8983	99.70
Koplik spots in mouth	0.9690	100
loss of mobility	0.9177	99.74
loss of smell	0.9730	99.57
loss of speech	0.9817	99.46
loss of taste	0.9815	100
nasal congestion	0.9117	99.78
nasal discharge	0.9121	99.91
pneumonia	0.9422	99.97
rash	0.9591	99.88
sneezing	0.9697	99.99
sore throat	0.9442	99.93
swollen glands	0.9456	99.96
Average	0.9391	99.71

Table 2: Area Under the cumulative gain chart and ROC of the proposed model on the Twitter dataset

evaluation method, first, we scraped English tweets from Texas state in the United States of America from the time period 24/05/2020 to 13/09/2020. We selected this time period because the first peak of Covid-19 cases in Texas happened in this period. We scraped 2,574,783 tweets from this period by using the Academic track of the Twitter API and *twarc* library[5] implementation.

Then from this set of tweets, we selected 10 positive sample tweets corresponding to each medical concept and added them to randomly selected 100,000 tweets. Then for concept, we performed a cumulative gain assessment on the dataset of the sample of 100,010 tweets and recorded the increase in cumulative gain across as k increase. Fig. 2 plots the resulting 24 cumulative gain charts obtained. For all 24 medical concepts, we got 100% cumulative gain within the 15 percentile of entire sorted tweets. In other words, for all medical concepts, all 10 positive samples appeared within the first 15 percentile of more than 100,000 tweets sorted according to the scores generated by using the model. We then calculated the area under this cumulative gain chart (AUC-CG) for each medical concept, these are listed in Table 2. The areas under the cumulative gain curve for all 24 medical concepts are above 98% and the average area under the curve is 99.71%. Such high values indicate that the proposed model is performing well on the Twitter dataset.

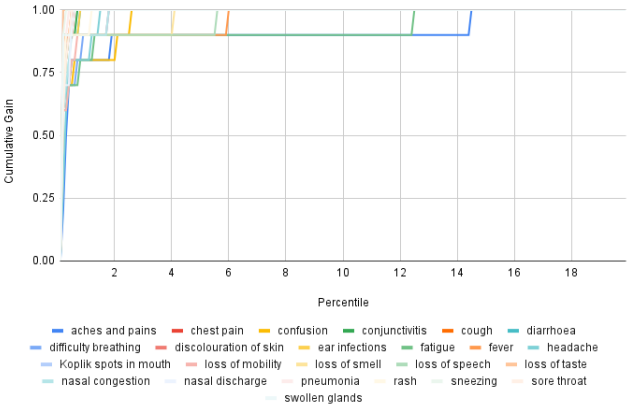

Figure 2: Cumulative gain chart of the proposed model on the Twitter dataset

4.3.2 Evaluation on Reddit dataset

The cumulative gain evaluation of the proposed model is further performed on the Reddit dataset by using already trained anchor vectors using Twitter samples. Similar to our previous evaluation, first we scraped English Reddit social media data from Texas state from the time period of the first peak of Covid-19 cases in Texas (24/05/2020 to 13/09/2020). To collect Reddit social media data we used Pushshift API [6] and a python module called *pmaw* and this doesn't require any credential information from our end. We scraped total 15,845 reddit submissions and 809,997 reddit comments.

Then we manually selected 10 positive Reddit samples corresponding to each medical concept from these scraped samples. We couldn't find 10 positive samples for some medical concepts and in such cases, we excluded such concepts from this evaluation. Then we added positive samples of each concept separately into a random set of 100,000 samples of Reddit comments and conducted the cumulative gain evaluation. The cu-

[5] https://twarc-project.readthedocs.io/en/latest/api/client2/

[6] https://github.com/pushshift/api

mulative gain chart obtained from this evaluation is plotted in Fig. 3. Out of 14 medical concepts used for this evaluation, for 13 medical concepts, we got 100% cumulative gain within the 20 percentile of entire sorted samples. In other words, for all these 13 medical concepts, all 10 positive samples appeared within the first 20 percentile of more than 100,000 samples sorted according to the cosine similarity scores calculated using the corresponding anchor vector.

We then calculated the area under this cumulative gain chart for each medical concept, see Table 3. The areas under the cumulative gain curve for all of these 14 medical concepts are above 90% and the average area under the curve is 98.71%. We can see that, except for the medical concept *swollen glands*, all other medical concepts have more than 98% area under the cumulative gain curve. Such high values show that the proposed model is performing well on the Reddit dataset also.

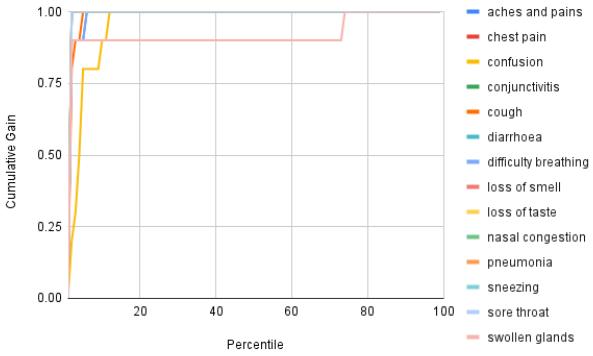

Figure 3: Cumulative gain chart of the proposed model on the Reddit dataset.

Medical Concepts	AUC CG
aches and pains	98.6
chest pain	99.0
confusion	95.8
conjunctivitis	99.0
cough	98.6
diarrhoea	99.0
difficulty breathing	98.6
swollen glands	91.8
sneezing	99.0
loss of smell	99.0
pneumonia	99.0
loss of taste	99.0
nasal congestion	99.0
sore throat	99.0
Average	98.17

Table 3: Area under the Cumulative Gain chart of the proposed model on the Reddit dataset

4.3.3 Evaluation on News/Media dataset

After evaluating the performance of the proposed model on the Twitter and Reddit datasets, we evaluated the performance of the model on News/Media data by using the cumulative gain evaluation method. Unlike Twitter and Reddit, there are no specific APIs for News/Media data scraping. To collect News/Media data from Texas, we manually scraped text from a list of 15 available online media from Texas.

Similar to Twitter and Reddit data scraping, we selected News/Media articles (940 articles) from Texas state which is published between 24/05/2020 and 13/09/2020 (the first peak of Covid-19 in Texas). We used the python library *BeautifulSoup* (Richardson, 2007) to parse the data in HTML format. Then we tokenized the News/Media data at the sentence level and treated each sentence as a separate sample. However, even after considering each sentence as a separate sample, the total number of New/Media samples (27,336) is small compared to Twitter and Reddit. Therefore we considered News/Media samples from two more time periods, 24/10/2020 to 22/2/2021 and 1/08/2021 to 31/12/2021, in which the number of Covid-19 caseloads peaked in Texas. After including these two more time periods we were able to collect 79,729 News/Media samples.

For the evaluation, we selected 10 positive News/Media samples for each medical concept. Some of the medical concepts do not have 10 positive samples and we excluded such concepts from our evaluation. Then for each of the medical concepts, we inserted these selected positive samples into a set of all available News/Media samples and conducted the cumulative gain evaluation. The cumulative gain chart from this evaluation is shown in Fig. 4. All 10 positive samples of 7 medical concepts except *fatigue*, *difficulty breathing*, and *headache* are gained from the first 15 percentiles and all 10 samples of *difficulty breathing* are gained from the first 25 percentiles of more than 79,729 sorted News/Media samples.

We then calculated the area under this cumulative gain chart for each medical concept, see Table 4. The areas under the cumulative gain curve for all medical concepts except *fatigue* are above 95% and

the average area under the curve is 96.33%. This indicates that the proposed model is also performing well on the News/Media dataset.

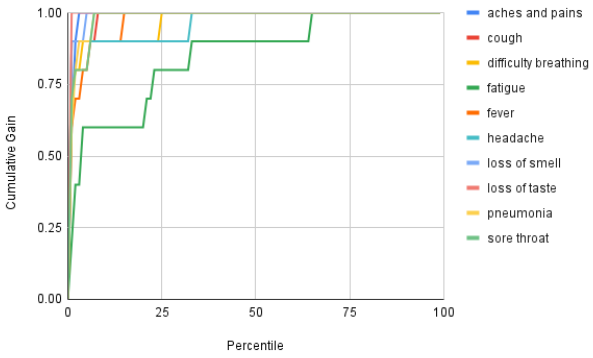

Figure 4: Cumulative gain chart of the proposed model on the News/Media dataset

Medical Concepts	AUC CG
aches and pains	98.5
cough	98.3
difficulty breathing	96.2
fatigue	84.4
pneumonia	98.0
fever	96.7
headache	95.8
loss of smell	98.6
loss of taste	99.0
sore throat	97.8
Average	96.33

Table 4: Area under the Cumulative Gain chart of the proposed model on the News/Media dataset

5 Discussion

The basis of the proposed model is for each medical concept that we wish to identify mentions of we learn an anchor vector (embedding). In our experiments, we used selected Twitter samples to learn this anchor vector. One interesting question is how effective this model which is trained by using data from one social media on another social media data. We already evaluated the model on two other social media, Reddit and News/Media. In order to compare the performance of the model on Twitter with the performance on Reddit and News/Media we generated box plots of these three datasets, see in Fig. 5. For each of these three datasets, the corresponding box plot shows the minimum, first quartile, median, third quartile, and maximum AUC-CG scores across all medical concepts considered for the evaluation.

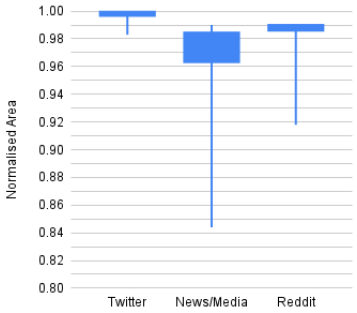

Figure 5: The boxplot of the area under the cumulative gain curve for medical concepts on the Twitter, Reddit, and News/Media datasets

From the box plots, we can see that the performance of the model on Reddit is comparable with Twitter for most of the concepts, but on News/Media data we can see a performance drop. We note that compared to Twitter samples the social media texts in News/Media dataset are longer and therefore the model's anchor vectors trained using Twitter samples may be less effective for samples from News/Media. Topical divergence between samples from Twitter and News/Media may also affect the performance of the model. In order to improve the performance we may need to include samples from News/Media for training the model.

6 Conclusion

We proposed a simple model to automatically identify the mentions of medical concepts in social media text by using a pre-trained language model and a small set of carefully selected samples. We validated the effectiveness of the proposed model on three social media sources Twitter, Reddit, and News/Media particularly focusing on medical concepts related to Covid-19 and Measles.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at DCU, Trinity College Dublin and TU Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme. Special thanks to Matthew Erskine, Dominik Dahlem and Danita Kiser from *Optum*.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Alban Bornet, Dimitrios Proios, Anthony Yazdani, Fernando Jaume-Santero, Guy Haller, Edward Choi, and Douglas Teodoro. 2023. Comparing neural language models for medical concept representation and patient trajectory prediction. *medRxiv*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2020. Target concept guided medical concept normalization in noisy user-generated texts. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 64–73, Online. Association for Computational Linguistics.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Kathy Lee, Sadid A Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. 2017. Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 462–469.

Nut Limsopatham and Nigel Collier. 2015. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1675–1680, Lisbon, Portugal. Association for Computational Linguistics.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Dongfang Xu and Timothy Miller. 2022. A simple neural vector space model for medical concept normalization using concept embeddings. *Journal of Biomedical Informatics*, 130:104080.