Using MT for multilingual covid-19 case load prediction from social media texts

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

Abstract

In the context of an epidemiological study involving multilingual social media, this paper reports on the ability of machine translation systems to preserve content relevant for a document classification task designed to determine whether the social media text is related to covid-19. The results indicate that machine translation does provide a feasible basis for scaling epidemiological social media surveillance to multiple languages. Moreover, a qualitative error analysis revealed that the majority of classification errors are not caused by MT errors.

1 Introduction

The work reported in this paper was carried out as part of a covid-19 case load forecasting project. Similar to other work on covid-19 forecasting, e.g. (Rahimi et al., 2021; Wang et al., 2022; Namasudra et al., 2023), our baseline system used an auto-regressive approach to case-load prediction. Several studies, however, have pointed to social media as a useful information source for this task (Yousefinaghani et al., 2021; Drinkall et al., 2022). Consequently, we wished to supplement our auto-regressive forecasting with information from social media. Specifically, we used the prevalence of mentions of covid-19 and related concepts in the social media emanating from a location to inform the case load predictions for that location.

Given the global nature of covid-19, we wished to make the solution scalable to multiple languages. One approach would be to use multilingual classifiers. However, having reviewed the literature (see Section 1.1) a decision was made to use machine translation (MT) to translate data sources in other languages into English and then to focus only on developing the text classification and prediction for English. This approach has several technical advantages, such as: (i) many existing NLP resources are designed to work in English (ii) adding new languages involves building a new MT system rather than developing out a new classification and prediction pipeline for a new language.

Interestingly, in this context, as the goal of MT was not translated text as such but enabling downstream text classification, we did not use any of the usual intrinsic MT evaluation strategies (automatic scores, human evaluation of translation quality criteria, error annotation and classification), but extrinsic evaluation, namely assessing and analysing the performance of the classifier on the translated English data. Our research questions were:

RQ1 How useful is MT for this classification task? In other words: how close is the classification accuracy on translated text to the accuracy on original English texts?

RQ2 What is the relation between classification and translation errors? In other words: how many of classification errors happened because important terms were not translated correctly?

In order to enable reproducibility and further research, all annotated data are publicly available.[1]

[1] https://github.com/m-popovic/corona-mexican_tweets

1.1 Related Work

Prior to the covid-19 pandemic, the majority of work on harnessing social media data for disease surveillance focused on the prediction of influenza outbreaks. However, given that influenza and covid-19 are both respiratory infections diseases, this prior work is relevant to our research.

(Schmidt, 2012) provides an overview of some of the early work on online search term and social media analysis for flu surveillance, including the *Google Flu Trends* and *HealthMap* systems. Some of this early work focused on the analysis of English twitter to identify key phrases whose prevalence tracked either with flu or the H1N1 outbreaks (Lampos and Cristianini, 2012). More recently, (Samaras et al., 2020) report on the relative benefit of using Google or Twitter as a flu surveillance platform. The study examined the predictive power of the frequency of two key terms (two Greek terms for the English term 'influenza') on each platform. The Google frequencies were obtained via Google Trends as weekly counts. The Twitter frequencies were obtained as daily and weekly counts of tweets containing these terms. Both frequencies had high Pearson correlations with the Twitter correlation being slightly stronger.

(Sooknanan and Mays, 2021) provides a review of current work on using social media data for disease modelling, both in terms of analysing Twitter data to understand public opinions (e.g., about mask wearing) and also in terms of using data from social media as inputs to into compartmental prediction models (e.g., by using social media to estimate the relative sizes of disease aware and unaware populations and tailoring risk factors for these populations).

None of the above work focused on multilingual approaches to social media analysis. Multilingual disease outbreak identification was explored in (Mutuvi et al., 2020). They compare different text classification methods for classifying news articles from six different languages into those about disease outbreaks and others. They report that a fine-tuned deep learning models based on a pre-trained multi-lingual BERT produced best results. However, in the data set used in these experiments, the topic (label) of the news articles was determined by their URL and titles, so there was no need for manual labelling. As a result, the data set contained hundreds to thousands of labelled samples per language. Our work is focused on analysing Twitter rather than news articles, and it was not possible to determine the topic of each tweet, nor was there an available large data set of tweets with appropriate labels for covid-19. Consequently, using deep learning based multilingual word representations would require manual labelling for each language, and this would make the scalability to other languages difficult.

The work reported in (Verma et al., 2022) demonstrated the feasibility of using MT to scale social media analysis to multiple languages. The authors used MT for cross-lingual cyberbullying detection. This work was based on an Italian data set of adolescent WhatsApp messages that was annotated for cyberbullying (Sprugnoli et al., 2018). The original Italian WhatsApp messages were translated into English both by professional translators and by MT systems. The reported F-scores on human translations were around 0.8, and on MT outputs around 0.7–0.75, and these results were on par with classifiers trained on the original Italian messages. Overall, their results indicate that MT can be useful for this task despite of relatively low automatic scores (25 BLEU, 48 chrF). Building on these results, we chose to use machine translation to translate tweets into English and to develop our covid-related text classification models for English.

It is noteworthy that (Verma et al., 2022) also report an analysis of both translation and labelling errors on the human translations of the Italian corpus. The labelling error analysis assessed whether the labels applied to the original Italian tweets were still valid for the English translations. However, no error analysis was applied to the MT outputs. Indeed, to the best of our knowledge, a detailed analysis of multilingual text classification errors potentially caused by MT has not yet been reported in the literature.

2 Method

The focus of our experiment was on Spanish and English social media data from online regional (North American) sources. The Spanish data were translated by five MT systems trained on different corpus sizes and domains, and then given to the classifier trained on English manually labelled data.

The task of case load prediction was based on the text topic, namely whether it is related to covid-

19 or not. for this purpose, English set of Tweets was manually labelled and used to develop a classifier. This classifier is then used both for originally written English texts as well as for Spanish texts after they were translated by five MT systems.

Building an appropriate MT system for the given task poses some challenges. For the English–Spanish language pair, being one of the "high-resourced" language pairs, there are generally a lot of parallel data available. Still, for social media texts such data is not available. The genre also poses several additional challenges such as informal language, spelling and grammar errors, emoticons, hashtags, and a large number of domains/topics. In addition, different Spanish dialects might represent a challenge, too: European Spanish is generally different than American Spanish, and there are differences between different American countries and regions as well.

Since there is no available data in the desired genre and domain (social media data related to corona virus), we built the initial systems on publicly available data which are partially similar to the data to be translated. Since we had a decent amount of monolingual English in-domain data, we used this data to augment the initial training data by synthetic in-domain parallel corpora by back-translation which is a widely used practice in NMT (Sennrich et al., 2016; Burlot and Yvon, 2018; Poncelas et al., 2018).

As for evaluation of the MT systems, not only *training* parallel in-domain data were unavailable, but also *development* and *test* data. In-domain data were written either in Spanish or in English, and they were not translated into the other language by human translators. Therefore, using automatic metrics such as BLEU (Post, 2018), chrF (Popović, 2015) or the newest neural-based ones such as COMET (Rei et al., 2020), was not possible.

A possible solution could be to find a translator to generate the corresponding reference English tweets thus enabling automatic intrinsic MT evaluation. However, automatic scores would give only an overall idea about the translation quality, which is not necessarily correlated to the performance on the final task, namely the classification accuracy. Moreover, the translation process requires effort and time. Another solution would be human evaluation of translated texts, however the usual quality criteria (adequacy, fluency, readability) or error annotation and/or classification do not necessarily correlate with the performance of the final task.

Therefore, the evaluation and comparison of MT systems was performed extrinsically, by measuring the classification performance on translated texts. Overall scores are presented in the form of classifier accuracy and compared to the accuracy on the English test data. Furthermore, all classifier errors are analysed in depth to examine to which extent they are related to MT errors, and what are the differences between MT systems in this aspect. A qualitative analysis was performed too, to examine the nature of relevant MT errors. For this kind of evaluation, only correct reference labels of the Spanish text were necessary, which requires significantly less time and effort than translation or manual MT evaluation.

3 Data

3.1 Classification data

Our data of interest was scrapped from different social media sources such as Twitter, Reddit and news/media for specific time periods from the last two years. The time periods were divided into three stages: early (beginning of 2022), middle (end of 2022/beginning of 2021) and recent (end of 2021), in order to capture at least three different covid-19 peak uptakes, peaks and peak downfalls.

The raw data extracted from these sources contains highly unstructured and redundant information. To overcome these issues, the preprocessing of text is performed by utilising different techniques. The social media text usually contains hashtags in order to highlight the topic. The removal of hashtags might lose some important information or the context from the text. Therefore, we decided to split these hashtags into separate words, and considered it inside the text instead of totally removing it. For example, #HappyLife get converted into <hashtag>Happy Life <end_hashtag>. Another typical information in social media text are the emoticons or emojis. The emojis give us ideas about the sentiments or expressiveness of people towards a particular topic. Instead of removing it, we mapped emojis to their text description and kept it in original text. For example, <emoji>Happy_face_smiley<end_emoji>. Some text samples also contain URLs to give more information on particular topics by redirecting to the url. Instead of extracting full URL content,

	annotated tweets	related to covid (%) *no*	*maybe*	*yes*
en	3024	54.5	5.2	40.2
es	898	63.5	7.2	29.3

Table 1: Statistic of data used for classification experiments: number of annotated English and Spanish tweets together with the distribution (percentages) of the three labels.

we decided to extract the URL title, which gives the overview of entire URL content, for example <url_title>title <end_url_title>.

Annotation

In order to be able to train and test the classifier, each of the selected tweets was manually assigned one of the following three class labels:

- 0 ('*no*') - The text is not related to covid-19
- 1 ('*maybe*') - Not sure whether the text is related to covid-19
- 2 ('*yes*') - The text is related to covid-19

The annotators were given the following guidelines: if you are at least 70% confident that the tweet content is relevant to covid-19 - irrespective of the tweeters intention – then '*yes*'. If symptoms are mentioned but not explicitly related to covid-19, then '*maybe*'. Things like depression or similar which could be but are not explicitly talking about covid-19 are '*no*'. Parties and similar are '*yes*' only if there is explicit reference to covid-19/pandemic social norms. Emojis like 'face_with_medical_mask' are taken as '*yes*'.

Due to time and resource constraints, each text was annotated by one annotator, therefore it was not possible to estimate inter-annotator agreement. The English annotator was a native speaker who also provided the guidelines. The Spanish annotator was fluent both in Spanish and English, and had experience in translation.

The statistics of annotated tweets is presented in Table 1. It can be seen that the distribution is similar in both languages, especially for the label '*maybe*' which is clearly the least frequent one. As for '*yes*' and '*no*' labels, both texts are skewed towards '*no*', especially the Spanish text.

3.2 MT data

3.2.1 Training data

Medical corpus This corpus consists of corona-related corpus[2] provided by TAUS together with the EMEA part of the OPUS[3] corpus (Tiedemann, 2012).

The Spanish-English part of the TAUS corpus consists of about 800,000 sentences of a conversational genre about different medical topics including corona virus. The domain and the genre of this corpus are similar to those of the analysed texts although it cannot be called 'in-domain'.

The EMEA corpus consists of various medical concepts written in a formal way, however they are not related to corona. The goal of this corpus is to provide general medical terminology necessary for the given task.

Subtitles The analysed data do not consist only of medical topics, therefore the training material should be enriched with non-medical texts. For this goal, we used OpenSubtitles part of the OPUS corpus consisting of conversational sentences from movie subtitles because they are partially similar to social media texts due to its informal language and conversational nature.

Synthetic in-domain corpora These corpora consist of monolingual scrapped in-domain English data from different sources and their machine translations into Spanish. A MT system in the opposite direction (English to Spanish) is trained on subtitles and medical texts and used to "back-translate" the English in-domain data. For this purpose, about 8 million English sentences from Twitter, 5.8 million English sentences from Reddit and 79 thousand English sentences from News were used.

3.2.2 Development data

As previously mentioned, human translations of the in-domain data were not available. Therefore, we used a part of the publicly available corona-related parallel TICO corpus (Anastasopoulos et al., 2020) as development set for all the systems.

3.2.3 Test data

The official in-domain test set for MT are the annotated tweets from Mexico described in Section 3.1. In total, there are 898 tweets consisting

[2] https://md.taus.net/corona
[3] https://opus.nlpl.eu/

of 1377 sentences/segments. As previously mentioned, human reference translations for the test set were not available, only manual labels about whether the tweets are corona-related or not.

(a) Training

	segments	words Spanish	words English
medical	1,999,966	39M	35M
subtitles	6,000,000	49M	55M
twitter	8,009,223	/	111M
reddit	5,848,187	/	7M
news	78,884	/	2M

(b) Development + Test

	segments	words Spanish	words English
dev (TICO)	500	8,787	7,800
test (Mexican tweets)	1,377	20,906	/

Table 2: Statistics of data used for MT experiments: number of segments and running words in each corpus.

4 Experimental set-up

4.1 Classification/prediction

We used *bert-base-uncased* pre-trained BERT model (Devlin et al., 2018) to fine-tune for automatic text classification according to relatedness to covid-19. We fine-tuned the model with 20 training epochs with 500 warm-up steps and we used 'huggingface' library implementation for training the model (Wolf et al., 2020).

To build our final classification model, we trained four separate models by using four different data sets: 801 tweets from a short initial early stage time frame, and 801 tweets for each of the three stages mentioned in Section 3.1, namely early, middle and recent. These four models are then combined using ensemble approach based on summation of logits of predictions of each model. In this ensemble model, we do not directly take the prediction of each model, but the logit of each model's prediction. These logits are then summed and the label with the highest logit sum is selected as the final label. The advantage of the logit sum strategy is that we can account for the confidence of each model: labels predicted by individual models with high confidence will get higher priority when selecting the final label than those predicted with low confidence.

The initial classifier was trained on 80% of annotated English data in order to be tested on the remaining 20%. Afterwards, another classifier was trained on the entire English corpus, which was then further used for classifying additional English data from different sources as well as translated Spanish tweets used for MT testing.

4.2 MT systems

All our systems are based on the Transformer architecture (Vaswani et al., 2017) and built using the first version of the Sockeye implementation (Hieber et al., 2018). The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016) with 32,000 BPE merge operations both for the source and for the target language texts.

All the systems have Transformer architecture with 6 layers for both the encoder and decoder, model size of 512, feed forward size of 2,048, and 8 attention heads. For training, we use Adam optimiser (Kingma and Ba, 2015), initial learning rate of 0.0002, and batch size of 4,096 (sub)words. Validation perplexity is calculated on the development set after every 4,000 batches (at so-called "checkpoints"), and if this perplexity does not improve after 20 checkpoints, the training stops.

The following five MT systems have been developed using different data for training:

M (medical) trained on the two medical texts (corona corpus and EMEA).

MS (medical+subtitles) trained on the two medical texts and subtitles.
+*reverse* MS, a system trained on the same corpus in the opposite direction (English to Spanish), in order to generate synthetic parallel in-domain data by "back-translating" English data.

MST (medical+subtitles+twitter) trained on the medical texts and subtitles together with the synthetic Twitter corpus.

MSTRN (medical+subtitles+twitter+reddit+news) trained on the medical texts and subtitles together with the synthetic Twitter, Reddit and News corpora.

MSTRN$^+$(medical+subtitles+twitter+reddit+news with domain labels) trained on the medical texts and subtitles together with the synthetic

Twitter, Reddit and News corpora; each sentence in the synthetic corpora has a label indicating the domain (analogously to the language labels in multilingual MT systems (Johnson et al., 2017)).

5 Evaluation

The first step Although reference translations were not available for the test set, the very first step was to check the sanity of the initial MT systems (M and MS) on the TICO development set. The BLEU and crhF scores for these systems (trained on medical data and subtitles) were very high (63.9/63.2% BLEU, 78.8/78.0% chrF), which indicated that the systems can be used and further developed. It has to be taken into account, though, that the development set is coming from the same domain and also contains the same Spanish variant as the training data, therefore those scores are too optimistic for the actual task at hand.

Dis/similarity of MT outputs Although it was not possible to use any automatic metrics for evaluation, it was possible to use some automatic methods to estimate the similarity between the five MT outputs. If some of the outputs were (almost) identical, detailed analysis of all of them would not be necessary. For this purpose, we calculated normalised edit distance (Levenshtein, 1966) (word error rate, WER) and chrF for all pairs of MT outputs in order to obtain an idea how different (if at all) they are. These scores showed that all the outputs are in general different, so that all of them were further analysed in details.

5.1 Classification accuracy

The extrinsic evaluation process for each of the five MT systems consisted of the following steps:

1. translate the Spanish test set into English
2. pass the translated English text to the classifier
3. calculate the accuracy by comparing the predicted labels with the labels manually assigned to the Spanish original text
4. higher accuracy score indicates better MT performance *for the given task*, not necessarily in terms of *translation quality*.

The classification accuracy of the outputs of the five MT systems together with the accuracy of the original English text used for classifier evaluation are shown in Table 3.

language	MT system	accuracy
en	/	85.4
es	M	81.7
	MS	86.2
	MST	83.9
	MSTRN	83.8
	MSTRN$^+$	84.4

Table 3: Classification accuracy (%): en = original English text, es = Spanish texts translated into English by five MT systems.

It can be seen that the accuracies achieved on MT outputs are comparable to the accuracy on the English text, indicating that the translation process preserved most of the information important for the classification process.

It should be taken into account, however, that, as mentioned in Section 3.1, the classifier evaluated on English data was trained on 80% of annotated English data, whereas the classifier used for MT outputs was trained on the entire labelled data set. Therefore, the comparison might seem too optimistic (for example, classifying MS output being more accurate than classifying original English data). Despite of that, the accuracies obtained on MT outputs can be considered as high enough, so that in general using MT is suitable for the given task.

As for different MT systems, the M system, trained only on medical texts, yielded the lowest accuracy, as it could be expected. Somewhat surprisingly, the best accuracy was not achieved by adding Twitter training data (MST) but by the system trained only on medical texts and subtitles (MS). The three systems which used additional synthetic training data (MST, MSTRN, MSTRN$^+$) have similar accuracies, ranged between the best and the worst one, the MSTRN$^+$ being slightly better than the other two.

5.2 Relations between classification and translation errors

While the accuracy scores are giving an idea about the usefulness of an MT system for the given task, it still remains unclear whether and how the classifier errors are related to MT errors, as well as whether there are any differences between the MT systems in this aspect.

In order to explore this, we calculated:

- percentage of all classification errors related to MT errors

- percentage of each type of classification error (confusion) related to MT errors

In order to enable this analysis, the test set was annotated in the following way:

1. for each incorrectly labelled tweet, check MT errors
2. if there are MT errors involving words important for assigning the label, it is considered that the classification error is related to MT errors
3. if the important words are correct, it is considered that the classification error is not related to MT errors (regardless whether there are other MT errors)
4. if there are MT errors which might have affected the classification process, the relation is considered unclear

This annotation was carried out by the same annotator who assigned the labels to the original Spanish tweets.

Table 4 shows eight examples of misclassified texts and different types of MT errors. In the first three examples, classification error is not related to MT errors: 1) because there are no MT errors, 2) because MT errors do not involve the important signals for the classifier (in this case "overwhelmed health care system"), 3) because MT error in the important part is only word order, not the meaning.

The relation between classification and MT errors in the next two examples is unclear: 4) "work remote" instead of "I work remotely" could have had influence 5) "stand healthy" instead of "stay healthy" could be the reason; furthermore, the entire source text is in English.

MT errors in the last two examples triggered the classification error: 6) "downpour" (heavy rain) is translated as "lockdown" thus creating a false signal about non-existing relation to corona 7) "cover" instead of "face mask" or "medical mask" removes the signal for the relatedness to covid 8) the important hashtag "Quedate en casa" ("Stay at Home") was not translated correctly.

Table 5 shows the percentage of classification errors which are related to MT errors, together with the percentage of those not related to MT errors and those potentially related ('unclear'). While the percentage of potentially MT-related errors is relatively low and similar for all MT systems, there are notable differences in MT-related errors between the systems.

Overall, the MS system results in the lowest number of MT-related classification errors (less than one third), and the M system results in the highest number (more than a half). The differences between the MST and the MSTRN systems are small while MSTRN$^+$ has a lower percentage than those two, but notably higher than the MS system.

5.2.1 Analysis of confusions

Table 6 presents separated percentages for each of the classification confusions.

It can be seen that the majority of incorrect classifications of the label '*maybe*', either as '*yes*' or as '*no*', are not related to MT errors, except of the M system for '*maybe→no*' confusion. Qualitative analysis revealed that for all MT systems, predicting '*maybe*' as '*no*' is often related to problems with hashtags such as 'quedateencasa', 'stayathome', and similar. As for MT-unrelated confusions, one possible reason is the low frequency of the label '*maybe*' in the English training data, and another is the uncertainty of the meaning of the text which made it difficult even for human annotators to decide whether it is related to covid or not.

As for '*no→yes*' confusions, they are highly MT-related only for the M system. Qualitative analysis of those errors showed that this system overly generates medical terms such as 'hospital', 'symptoms', 'lockdown', 'disease', 'outbreak', 'tests', etc.) (by mistranslating non-medical words in the source or adding medical-related hallucinations) thus creating many false signals for relatedness to covid. For other four systems, the situation is opposite, namely the vast majority of this type of confusions is not related to MT errors.

Finally, the '*yes→no*' confusion is generally the most MT-related classification error, but notably less for the M and MS systems than for the other three. Qualitative analysis of this problem showed that the majority of MT-related errors come from incorrect translation of 'cubrebocas/tapabocas' meaning 'medical mask' or 'face mask': while M and MS usually translate this term correctly, the other three systems usually fail – the translations are sometimes completely incorrect, and sometimes 'face covering' or only 'cover, covering', so that the important information is lost in the translation process. Also, the emoticon description 'face-with-medical-mask' which represents an important signal is often changed. This

		– all MT errors are underlined – those related to the classifying error are in bold	relationship between errors
1)	source	La nueva normalidad no se va a lograr...	no (no MT errors)
	MT	The new normality isn't going to be achieved...	
	correct	The new normality isn't going to be achieved...	
2)	source	El sistema de salud de Torreón está colapsado, cuídense mucho amigos de la comarca.	no (the crucial part "overwhelmed health care system" is correct)
	MT	The Torreon health care system is overwhelmed, so be plenty of friends in the area.	
	correct	The Torreon health care system is overwhelmed, so take much care, friends in the area.	
3)	source	Ya hay variante lambda, jajaja	no (the crucial part "lambda variant" is correct (even though not in the right order))
	MT	Ya *variant lambda*, jajaja	
	correct	there is already lambda variant, hahaha	
4)	source	Trabajo a distancia y horarios escalonados, las opciones para la nueva normalida	unclear ("work remote" instead "I work remotely" could be the reason)
	MT	**x** Work **remote** and chill schedules, the options for the new normalcy	
	correct	I work remotely and staggered schedule, the options for the new normalcy	
5)	source	remember feelgood goodvibes goodnight stayhealthy	unclear (source in English; "stand" could be the reason)
	MT	beender feelgod gods goodnight **stand** healthy	
	correct	remember feelgood goodvibes goodnight stayhealthy	
6)	source	Desde cuando mi mamá cree que estoy apta para atravesar la ciudad en pleno aguacero	yes ("lockdown" instead of "downpour")
	MT	From the time my mom thinks I am fit to go through the city in the middle of a **lockdown**.	
	correct	Since when my Mum thinks I am fit to go through the city in the middle of a downpour	
7)	source	Cubrebocas 3 capas para mayor protección	yes ("cover" instead of "face/medical mask)
	MT	**Cover** 3 layers for greater protection	
	correct	Face masks 3 layers for greater protection	
8)	source	<hashtag> Quedate En Casa <end_hashtag> en Los Tulipanes, Cuernavaca	yes ("Quedate" remained untranslated, "House" not perfect, "at" missing)
	MT	<hashtag> **Quedate x House** <end_hashtag> in Las Tulipes, Cuernavaca	
	correct	<hashtag> Stay at Home <end_hashtag> in Los Tulipanes, Cuernavaca	

Table 4: Examples of relations between classification and MT errors; all MT errors are underlined, and those related to the classification error are in bold.

problem could be diminished by special focus on such terms.

6 Summary

This work explored the ability of MT systems to preserve relevant content for a document classification task designed for covid-19 case load prediction.

The results of extrinsic evaluation (classification performance) show that classification performance on the MT tweets is comparable with the performance on original English tweets, indicating that MT does provide a feasible basis for scaling epidemiological social media surveillance to multiple languages. Furthermore, a detailed analysis of classification errors revealed that the majority of them are not caused by MT errors. Moreover, most of those MT errors which triggered a classification error are related to specific terminology and can be improved in future work. Other directions for future work include specific data selection for MT training, other methods for domain-adaptation and terminology translation, as well as using multilingual word representations from intermediate network layers instead of full translations.

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 Dublin City University, Trinity College Dublin and Technological University 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

MT system	% of classification errors which are MT-related	MT-unrelated	unclear
M	56.0	30.7	12.6
MS	**30.1**	**56.9**	12.2
MST	39.9	47.6	11.9
MSTRN	39.6	49.3	10.4
MSTRN$^+$	36.7	51.8	10.8

Table 5: Percentage of MT-related, MT-unrelated and potentially MT-related ('unclear') classification errors for each MT system.

	maybe → *yes*			*maybe* → *no*		
	MT-related	MT-unrelated	unclear	MT-related	MT-unrelated	unclear
M	11.1	55.6	33.3	40.7	37.0	22.2
MS	14.3	57.1	28.6	25.0	60.7	14.3
MST	0	50.0	50.0	30.5	54.2	15.2
MSTRN	0	87.5	12.5	29.6	53.7	16.7
MSTRN$^+$	0	80.0	20.0	28.1	56.1	15.8

	no → *yes*			*yes* → *no*		
	MT-related	MT-unrelated	unclear	MT-related	MT-unrelated	unclear
M	85.3	13.1	1.6	45.0	42.5	12.5
MS	8.3	83.3	8.3	44.7	46.8	8.5
MST	7.1	92.9	0	57.6	33.3	9.1
MSTRN	6.7	93.3	0	60.6	31.8	7.6
MSTRN$^+$	12.5	87.5	0	55.0	36.7	8.3

Table 6: Analysis of confusions: percentages of MT-related, MT-unrelated and potentially MT-related ('unclear') confusions for each MT system.

Matthew Erskine, Dominik Dahlem and Danita Kiser from *Optum* and to Patricia Buffini from *ADAPT Centre @ Dublin City University*.

References

Anastasopoulos, Antonios, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Burlot, Franck and François Yvon. 2018. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, pages 144–155, Belgium, Brussels, November.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Drinkall, Felix, Stefan Zohren, and Janet Pierrehumbert. 2022. Forecasting covid-19 caseloads using unsupervised embedding clusters of social media posts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1471–1484.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018)*, pages 200–207, Boston, MA, March.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, May.

Lampos, Vasileios and Nello Cristianini. 2012. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–22.

Levenshtein, Vladimir Iosifovich. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, February.

Mutuvi, Stephen, Emanuela Boros, Antoine Doucet, Adam Jatowt, Gaël Lejeune, and Moses Odeo. 2020. Multilingual epidemiological text classification: A comparative study. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, Barcelona, Spain (Online), December.

Namasudra, Suyel, S. Dhamodharavadhani, and R. Rathipriya. 2023. Nonlinear Neural Network Based Forecasting Model for Predicting COVID-19 Cases. *Neural Processing Letters*, 55(1):171–191, February.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Back translation in Neural Machine Translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pages 249–258, Alicante, Spain, May.

Popović, Maja. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 186–191, Brussels, Belgium, October.

Rahimi, Iman, Fang Chen, and Amir H. Gandomi. 2021. A review on COVID-19 forecasting models. *Neural Computing and Applications*, February.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November.

Samaras, Loukas, Elena García-Barriocanal, and Miguel-Angel Sicilia. 2020. Comparing social media and google to detect and predict severe epidemics. *Scientific Reports*.

Schmidt, Charles. 2012. Trending now: Using social media to predict and track disease outbreaks. *Environmental Health Perspectives*, 120, January.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 86–96, Berlin, Germany, August.

Sooknanan, Joanna and Nicholas Mays. 2021. Harnessing social media in the modelling of pandemics – challenges and opportunities. *Bulletin of Mathematical Biology*, 80(5).

Sprugnoli, Rachele, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey, May.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA, December.

Verma, Kanishk, Maja Popović, Alexandros Poulis, Yelena Cherkasova, Cathal Ó hÓbáin, Angela Mazzone, Tijana Milosevic, and Brian Davis. 2022. Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents. *Natural Language Engineering*, page 1–23.

Wang, Yanding, Zehui Yan, Ding Wang, Meitao Yang, Zhiqiang Li, Xinran Gong, Di Wu, Lingling Zhai, Wenyi Zhang, and Yong Wang. 2022. Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. *BMC Infectious Diseases*, 22(1):495, May.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yousefinaghani, Samira, Rozita Dara, Samira Mubareka, and Shayan Sharif. 2021. Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada. *Frontiers in public health*, 9:656635–656635, April. Place: Switzerland.