Supporting Risk-Aware Use of Online Translation Tools in Delivering Mental Healthcare Services among Spanish-Speaking Populations

Xie, Wenxiu; Ji, Meng; Zhao, Mengdan; Qian, Xiaobo; Chow, Chi-Yin; Lam, Kam-Yiu; Hao, Tianyong

Published in:
Computational Intelligence and Neuroscience

Published: 01/01/2021

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1155/2021/1011197

Publication details:
Research Article

Supporting Risk-Aware Use of Online Translation Tools in Delivering Mental Healthcare Services among Spanish-Speaking Populations

Wenxiu Xie,1 Meng Ji,2 Mengdan Zhao,2 Xiaobo Qian,3 Chi-Yin Chow,1 Kam-Yiu Lam,1 and Tianyong Hao3

1Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
2School of Languages and Cultures, The University of Sydney, Darlington, Australia
3School of Computer Science, South China Normal University, Guangzhou, China

Correspondence should be addressed to Meng Ji; christine.ji@sydney.edu.au

Received 14 September 2021; Revised 1 October 2021; Accepted 15 October 2021; Published 28 October 2021

1.Introduction

Despite the increasing public awareness of the prevalence of mental health issues among populations from low and middle-income countries, accurate, scientific, non-discriminative communication of mental disorders remains a real challenge [1–3]. Within different societal, cultural systems, conventionalised linguistic constructs have been developed over years and decades to describe and convey the underlying social attitudes and understanding of different mental disorders like varieties of anxiety or depressive disorders. In English-speaking multicultural countries, the communication and interpretation of mental disorders and their treatment for non-English-speaking migrant populations pose important challenges to frontline health workers and clinicians [4–6]. The rapid development of machine translation technologies has offered necessary technical means to interact and engage with multicultural vulnerable communities and people who have limited access to mental healthcare services, despite the prevalence of mental health issues among such populations who are at higher risks of developing clinical mental disorders or other comorbidities such as chronic non-communicable diseases or physical health conditions that are likely to exacerbate...
their mental health issues. Currently, there is very limited research which examines the reliability, safety, or levels of risks in using state-of-the-art online translation tools such as Google Translate in clinical settings for communicating and talking with patients about mental disorders.

Much of current research shows that the use of automatic translation tools in primary healthcare settings is driven by a persistent lack of qualified bilingual health professionals [7–10]. The risk of an unchecked use of translation technologies in specialised health and medical settings which have been developed largely for general cross-lingual communication purposes is real and well documented [11–14]. However, the practical needs for cost-effective translation tools in disease diagnosis and medical treatments are increasing. Although the provision of proper training to a certain number of bilingual health professionals can help reduce healthcare inequality issues, real-life scenarios can be much more complex, uncertain, and dynamic for any health systems at various levels. For example, a recurring issue in clinical settings is the lack of adequately qualified health translators working with under-resourced languages. Even for resourced language pairs such as English-Spanish and English-Chinese, it can be challenging to find bilingual health translators with extensive in-depth knowledge of different medical specialties. That is, the quality of human translation can also be compromised by the complexity and speciality of medical communications. In fact, with simple, direct sentences, online translations tools can fulfil their function to support a meaningful exchange of information between doctors and patients. We argue that health communication technologies like neural translation tools are evolving rapidly and that they have important potential for scaled uptake in health systems, especially those serving vulnerable or disadvantaged communities in under-resourced local health districts. Health technologies can be leveraged to help close the gaps in current medical and healthcare structures and improve the quality and accessibility of healthcare resources to populations and people at risk. Research is needed to develop instruments and aids to improve the safety and reliability of available translation technologies to be adopted in health and clinical settings.

2. Materials and Methods

2.1. Data Collection. We collected authoritative health information on anxiety disorders from the websites of federal and state health agencies and not-for-profit health promotion organisations in the U. S., Australia, Canada, and the United Kingdom. The total database contains 557 full-length articles including 270 (48.47%) original English materials associated with automatic translations to Spanish which contained misleading errors. We labelled these materials as positive or “risk” samples. Around 51.53% of the total samples we collected were original English texts whose automatic translation into Spanish using Google Translate did not contain any misleading information. The evaluation of the Spanish translations by Google was through the comparison of the original English texts with their backtranslations from Spanish. This method known as forward and backward translation was endorsed by international health organisations such as the World Health Organisation [15]. We subsequently labelled such English texts as negative or “safe” cases for automatic translation. We divided the entire database into 67.65% training data (389) and 32.35% testing data. Within the training dataset, there were 187 positive “risk” English texts which were prone to automatic translation errors and 202 negative “safe” English texts which had been proven to be suitable and reliable for automatic translation to Spanish. Similarly, within the testing dataset, there were 83 positive samples and 85 negative samples for testing the performance of classification of the classifiers to be developed. We applied 5-fold cross-validation on the training dataset to develop the classifiers to help remove biases in the development of algorithms.

2.2. Annotation of Feature Sets. Traditionally, health translation mistakes are believed to be associated with or triggered by the linguistic difficulty or lack of readability of the original English materials including complex, sophisticated, structural, syntactic, and lexical features. However, with the rapid developments of machine translation technologies, more research shows that semantic polysemy, that is, the multiple meanings of a certain word across domains and other issues, could be more challenging for latest neural machine translation tools. As a result, we included four large sets of features to investigate possible reasons which have triggered significant mistakes in machine-translated health materials from English to Spanish.

We annotated both training and testing datasets with 4 large sets of linguistic features: structural features (24 in total using Readability Studio software), lexical dispersion rates based on the British National Corpus (20 features in total), English lexical semantic features (115 features in total), which we annotated using the USAS system developed by the University of Lancaster [16–18], UK, and lexical sentiment features (92 features in total) that we annotated using Linguistic Inquiry and Word Count software. Appendix A shows the details of these 4 linguistic features.

2.3. Bayesian Machine Learning Classifiers (MLCs). Bayesian MLC is a sparse classifier which can effectively counter model overfitting with relatively small datasets like ours. Bayesian classifiers are different from other supervised machine learning techniques in that they produce the posterior odds of a class dependent on the prior odds of an event and asymmetrical classification errors of the model, whereas frequentist ML classifiers only return a hard binary decision. In solving practical questions like ours, posterior odds are much more informative than a certain predicted binary outcome, as the Bayesian-style prediction using posterior odds helps practitioners and decision makers to appreciate the levels of risks of negative and positive cases over a continuous probability scale and assists in developing more effective intervention strategies to achieve optimal
outcomes. This advantage of Bayesian MLs suits the purpose of our study, as we aimed to identify original English mental health materials which are more likely to cause significant errors if translated using automatic tools without further human evaluation. This can help health agencies developing bilingual health materials to better invest their resource and minimise risks of using machine translation in healthcare settings.

3. Methods

To identify the best subset of features within each annotation category, as well as the best subset of features across annotation categories, we applied separate and combined feature optimisation techniques. The automatic feature selection technique we used was recursive feature elimination (RFE) with cross-validation in Python scikit-learn to increase the generalisability and accuracy of the Bayesian machine learning classifiers we developed. To identify and rank highly predictive features, recursive feature elimination used linear kernel support vector machine (SVM) as the base estimator. An optimal set of features was identified when the cross-validated classification error reached the minimal value. Figure 1 shows the results of the automatic optimisation of different feature sets: in Figure 1(a), the optimised features of lexical dispersion rates were 4, as the cross-validated classification error dropped from 0.425 with the full feature set (20 in total) to 0.393 when the features were reduced to 4. In Figure 1(b), the optimised feature set of English semantic features was 10, as the cross-validated classification error decreased from 0.40 with the full feature set (115) to 0.375 when the features were reduced to 10. Further elimination of features however led to a spike in the classification error. In Figure 1(c), the optimised feature set of English sentiment features (annotated using the LIWC software) was 10, as we observed that the minimal classification error of 0.416 was reached when the total number of sentiment features was scaled back from 92 to 10. In Figure 1(d), 5 optimal structural features (92 in total) were identified when the minimal classification error of 0.409 was reached. Lastly, in Figure 1(e), we conducted the combined feature selection by integrating the 4 feature sets (251 in total): lexical dispersion rates and semantic, sentiment, and structural features. The optimal number of features emerged from the combined optimisation was 33 which was associated with the minimal classification error of 0.383.

4. Results and Discussion

4.1. Results. Following automatic feature optimisation to enhance the classification accuracy of classifiers, we evaluated the performance of Bayesian models (relevance vector machine (RVM)) with different feature sets on both the training and testing datasets. As discussed earlier, we applied 5-fold cross-validation on the training dataset to minimise biases in the classifiers being developed. First, we compared the original feature sets with their respective optimised feature sets in Table 1–5. Next, we compared the performance of RVM classifiers with different pairs of optimised feature sets. Table 6 shows the comparison of RVM classifiers with double, triple, and quadruple optimised feature sets, respectively. Like feature optimisation, feature normalisation is another useful automatic technique to enhance the performance of machine learning classifiers. We applied three popular feature normalisation techniques: min-max, $L_2$-norm ($L_2$), and Z-score normalisation with each RVM classifier to see whether this could help balance asymmetrical classification errors within each model.

Table 1 shows the performance of RVM classifiers with lexical dispersion rates as features. It shows that after automatic feature selection, RVM with the reduced and optimised feature set (D4) reached a largely comparable performance to that of the classifier run on the full feature set: on the training dataset, the mean of area under the curve (AUC) of RVM (D4) was 0.617 (SD = 0.06), compared to 0.625 (SD = 0.06) of RVM (full 20 features), suggesting that feature reduction could also help encounter the issue of overfitting in training machine learning classifiers. On the testing dataset, the AUC of the RVM (D4) (0.648) was similar to that of RVM (All 20) (0.649). Sensitivity dropped slightly from 0.578 (RVM-All 20) to 0.566 (RVM-D4), and specificity remained unchanged at 0.753. Normalisation did not improve RVMs with the entire or optimised feature sets of lexical dispersion rates.

Table 2 compares the performance of RVM classifiers run on English semantic features. It shows that after automatic feature selection, the performance of the RVMs improved on both the training and the testing datasets: on the training data, the mean of AUC of RVM with the full semantic feature set (USAS115) observed a marginal improvement from 0.652 to 0.659 with a slightly reduced standard deviation (SD) from 0.052 to 0.045. On the testing dataset, the AUC of RVM (USAS115) saw an improvement from 0.692 to 0.714. Specificity improved from 0.729 of RVM (USAS115) to 0.777 of RVM (U10); sensitivity decreased from 0.590 of RVM (USAS115) to 0.578 of RVM (U10). Normalisation did not improve model performance.

Table 3 compares the performance of RVMs with English sentiment features annotated with the Linguistic Inquiry and Word Count (LIWC) software. It shows that after automatic feature optimisation, the performance of the RVM classifier (LIWC all 92) improved on the testing datasets. The AUC of RVM (L10) increased from 0.580 to 0.605. Model sensitivity increased from 0.518 to 0.651, but specificity decreased from 0.577 to 0.494. The impact of feature normalisation on RVMs with all and optimised feature sets was similar, while the classifier specificity improved, sensitivity decreased, and the overall model accuracy on the testing dataset however did not improve significantly.

Table 4 compares the performance of RVMs with various structural features which we annotated with the Readability Studio software. After automatic feature optimisation, the AUC of the classifier RVM (structural all 24) decreased from 0.636 to 0.621, which was due to decreased model sensitivity from 0.518 to 0.446, but the model specificity increased from 0.729 to 0.788. Feature normalisation helped to balance the asymmetric classification errors on the classifier RVM with both the entire feature set and the optimised feature set: the model specificity decreased and sensitivity increased.
Dispersion feature optimization by RFE-SVM

USAS feature optimization by RFE-SVM

LIWC feature optimization by RFE-SVM

Structural feature optimization by RFE-SVM

All feature jointly optimized by RFE-SVM

Figure 1: Automatic feature selection recursive feature elimination with SVM as base estimator.
However, the overall model accuracy or the AUC did not improve with different feature normalisation techniques. Table 5 shows the performance of the RVM with the combined feature sets of lexical dispersion rates and semantic, sentiment, and structural features, which represented 251 features in total. Automatic feature optimisation reduced the original feature set of 251 features to a parsimonious model containing 33 features only. With less predictive and noisy features involved in the model, the performance of the classifier also improved significantly on both the training and the testing datasets: on the training data, the model AUC was 0.642 (SD = 0.038). This increased to 0.658 (SD = 0.034) with the optimised RVM classifier. On the testing data, the AUC improved from 0.647 to 0.718.
With automatic feature optimisation, both sensitivity and specificity improved: sensitivity increased from 0.518 to 0.554 and specificity increased from 0.706 to 0.741. Importantly, feature normalisation played a critical role in balancing asymmetrical classification errors on RVMs with combined feature sets. Specifically, min-max normalisation increased sensitivity of the optimised classifier to 0.651, the highest so far, and retained the high specificity of the classifier at 0.741. This sensitivity and specificity pair was the best combination among the classifiers developed so far.

Table 6 compares the performance of classifiers with double and multiple optimised feature sets. We compared in total 10 different pairs of optimised feature sets and conducted feature normalisation with each combination of optimised features, and as a result, each RVM in Table 6 has four different versions: the unnormalised version, followed by normalised versions with min-max, L2, and Z-score normalisation. The 10 combinations of optimised features were as follows: optimised lexical dispersion rates (D4) and optimised semantic feature (U10) (F1–F4), optimised lexical dispersion rates (D4) and optimised sentiment features (L10) (F5–F8), optimised lexical dispersion rates (D4) and optimised structural features (S5) (F9–F12), optimised semantic features (U10) and optimised sentiment features (L10) (F13–F16), optimised semantic features (U10) and optimised structural features (S5) (F17–F20), optimised sentiment features (L10) and optimised structural features (S5) (F21–F24), and so on. We identified 5 high-performing models based on considerations of the overall model AUC, accuracy, sensitivity, and specificity: F13 was the unnormalised combination of optimised semantic (U10) and optimised sentiment features (L10). It had an overall AUC of 0.674 on the testing dataset, with sensitivity of 0.651 and specificity of 0.741. F35 was the normalised version (through L2 normalisation) of optimised semantic features (U10), sentiment features (L10), and optimised structural features (S5). It had an overall AUC of 0.683 on the testing dataset, with sensitivity of 0.627 and specificity of 0.694. Figure 2 shows the visualised comparison of the AUCs of these 5 high-performing classifiers, the RVM with the entire combined feature sets (251 features) with L2 normalisation, and the best-performing classifier we developed (RVM (All33)) with min-max normalisation.

Tables 7 and 8 show the paired sample t-tests assessing the significance levels of differences in sensitivity and specificity between the various competing high-performance classifiers and the best-performing RVM classifier we developed through the combined automatic optimisation of four different feature sets followed by automatic feature normalisation. To reduce false discovery rates in multiple comparison, we applied the Benjamini–Hochberg correction procedure [19–21]. The results show that sensitivity of our best-performing RVM classifier was significantly higher than that of most other high-performing models, except for F35 ($p = 0.0017$); specificity of our best-performing RVM classifier was statistically higher than that of all other competing models with $p$ values equal to or smaller than 0.004.

Table 9 shows the paired sample t-tests assessing the significance levels of differences in AUCs between various competing high-performance classifiers and the best-performing RVM classifier on testing data using different training dataset sizes (i.e., 100, 150, 200, 250, 300, and all 389 training samples). We applied Benjamini–Hochberg correction to reduce false discovery rates in multiple comparison. The results show that AUC under different training dataset sizes of our best-performing RVM classifier was significantly higher than that of most other high-performing models, except for F13 ($p = 0.0752$) and F27 ($p = 0.1698$). Figure 3 shows the visualised comparison of the mean AUCs of these 6 competitive classifiers and the developed best-performing classifier. As shown in Figure 3, our best-performing RVM classifier gained the highest mean AUC than all other competing models.

### Table 5: Performance of RVM classifiers with all (dispersion + USAS + LIWC + structural) features.

<table>
<thead>
<tr>
<th>RVM</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC mean</td>
<td>SD</td>
</tr>
<tr>
<td>ALL 251</td>
<td>0.642</td>
<td>0.038</td>
</tr>
<tr>
<td>ALL 251 with min-max</td>
<td>0.626</td>
<td>0.041</td>
</tr>
<tr>
<td>ALL 251 with L2</td>
<td>0.670</td>
<td>0.045</td>
</tr>
<tr>
<td>ALL 251 with Z-score</td>
<td>0.633</td>
<td>0.085</td>
</tr>
<tr>
<td>ALL33 (automatically optimised)</td>
<td>0.658</td>
<td>0.014</td>
</tr>
<tr>
<td>ALL33 with min-max</td>
<td>0.678</td>
<td>0.034</td>
</tr>
<tr>
<td>ALL33 with L2</td>
<td>0.710</td>
<td>0.015</td>
</tr>
<tr>
<td>ALL33 with Z-score</td>
<td>0.672</td>
<td>0.036</td>
</tr>
</tbody>
</table>

With automatic feature optimisation, both sensitivity and specificity improved: sensitivity increased from 0.518 to 0.554 and specificity increased from 0.706 to 0.741. Importantly, feature normalisation played a critical role in balancing asymmetrical classification errors on RVMs with combined feature sets. Specifically, min-max normalisation increased sensitivity of the optimised classifier to 0.651, the highest so far, and retained the high specificity of the classifier at 0.741. This sensitivity and specificity pair was the best combination among the classifiers developed so far.

Table 6 compares the performance of classifiers with double and multiple optimised feature sets. We compared in total 10 different pairs of optimised feature sets and conducted feature normalisation with each combination of optimised features, and as a result, each RVM in Table 6 has four different versions: the unnormalised version, followed by normalised versions with min-max, L2, and Z-score normalisation. The 10 combinations of optimised features were as follows: optimised lexical dispersion rates (D4) and optimised semantic feature (U10) (F1–F4), optimised lexical dispersion rates (D4) and optimised sentiment features (L10) (F5–F8), optimised lexical dispersion rates (D4) and optimised structural features (S5) (F9–F12), optimised semantic features (U10) and optimised sentiment features (L10) (F13–F16), optimised semantic features (U10) and optimised structural features (S5) (F17–F20), optimised sentiment features (L10) and optimised structural features (S5) (F21–F24), and so on. We identified 5 high-performing models based on considerations of the overall model AUC, accuracy, sensitivity, and specificity: F13 was the unnormalised combination of optimised semantic (U10) and optimised sentiment features (L10). It had an overall AUC of 0.674 on the testing dataset, with sensitivity of 0.651 and specificity of 0.741. F35 was the normalised version (through L2 normalisation) of optimised semantic features (U10), sentiment features (L10), and optimised structural features (S5). It had an overall AUC of 0.683 on the testing dataset, with sensitivity of 0.627 and specificity of 0.694. Figure 2 shows the visualised comparison of the AUCs of these 5 high-performing classifiers, the RVM with the entire combined feature sets (251 features) with L2 normalisation, and the best-performing classifier we developed (RVM (All33)) with min-max normalisation.

Tables 7 and 8 show the paired sample t-tests assessing the significance levels of differences in sensitivity and specificity between the various competing high-performance classifiers and the best-performing RVM classifier we developed through the combined automatic optimisation of four different feature sets followed by automatic feature normalisation. To reduce false discovery rates in multiple comparison, we applied the Benjamini–Hochberg correction procedure [19–21]. The results show that sensitivity of our best-performing RVM classifier was significantly higher than that of most other high-performing models, except for F35 ($p = 0.0017$); specificity of our best-performing RVM classifier was statistically higher than that of all other competing models with $p$ values equal to or smaller than 0.004.

Table 9 shows the paired sample t-tests assessing the significance levels of differences in AUCs between various competing high-performance classifiers and the best-performing RVM classifier on testing data using different training dataset sizes (i.e., 100, 150, 200, 250, 300, and all 389 training samples). We applied Benjamini–Hochberg correction to reduce false discovery rates in multiple comparison. The results show that AUC under different training dataset sizes of our best-performing RVM classifier was significantly higher than that of most other high-performing models, except for F13 ($p = 0.0752$) and F27 ($p = 0.1698$). Figure 3 shows the visualised comparison of the mean AUCs of these 6 competitive classifiers and the developed best-performing classifier. As shown in Figure 3, our best-performing RVM classifier gained the highest mean AUC than all other competing models.
5. Discussion

A few important findings emerged in our extensive computational analyses, especially the search for the best subset of features for developing Bayesian machine learning classifiers to address our core research question, which was to predict and assess the risk levels of original English mental healthcare materials in terms of their suitability for automatic neural machine translation targeting Spanish-speaking patients. Our study shows that separate feature optimisation on the four distinct featuresets did not achieve acceptable pairs of model sensitivity and specificity. Let us take a close look at features retained in each optimised feature set.

Table 10 summarises the list of separately and jointly optimised features. First, the optimised feature set of lexical dispersion (D4) contained DiSp8: 0.7–0.8, DiSp9: 0.8–0.9, DiSp10: 0.9–1.0, and DiWr10: 0.9–1.0. Lexical dispersion rate is a measurement of familiarity of language to the public. We used existing lexical dispersion rates of the British National Corpus which had 10 intervals between 0 and 1 for spoken and written materials, respectively. In both spoken and written materials, higher lexical dispersion rates like those in the optimised feature set (D4) indicate that automatic machine translation mistakes were strongly associated with lexical items of higher familiarity in spoken and written materials. We used non-parametric independent sample test and Mann–Whitney U test to compare samples labelled as “risky” and “safe” original English mental health materials for automatic machine translation. The result shows that all 4 optimised lexical dispersion rates had statistically higher means in “risky” than in “safe” English mental health materials: DiSp8: 0.7–0.8 (safe text class: mean \( M \) = 20.689, standard
deviation (SD) = 11.963, standard error (SE) = 0.943; risky
text class: M = 27.854, SD = 14.628, SE = 1.283,
p < 0.0001), DiSp9: 0.8–0.9 (safe texts: M = 44.553, SD = 15.679,
SE = 1.236; risky texts: M = 54.400, SD = 17.275, SE = 1.283,
p < 0.0001), DiSp10: 0.9–1.0 (safe texts: M = 78.217, SD = 21.115, SE = 1.664; risky texts: M = 88.423,
SD = 23.223, SE = 2.037, p < 0.0001), and DiWr10: 0.9–1.0
(safe texts: M = 147.453, SD = 47.024, SE = 3.706; risky texts:
M = 176.346, SD = 53.828, SE = 4.721, p < 0.0001).

With the optimised semantic feature set (U10), there were
10 semantic features identified as most relevant predictive
features for the classifier. Similar to the optimised feature set

---

**Figure 2:** AUCs of RVMs on testing data using different feature sets.

**Table 7:** Paired sample t-test of the difference in sensitivity between the best-performing model and other models.

<table>
<thead>
<tr>
<th>No.</th>
<th>Pairs of RVMs</th>
<th>Mean difference</th>
<th>SD</th>
<th>95% CI Lower</th>
<th>95% CI Upper</th>
<th>p value</th>
<th>Rank</th>
<th>(i/m) Q</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All33 (min-max) vs. F27</td>
<td>0.0361</td>
<td>0.0021</td>
<td>0.0319</td>
<td>0.0403</td>
<td>0.0012</td>
<td>1</td>
<td>0.0083</td>
<td>**</td>
</tr>
<tr>
<td>2</td>
<td>All33 (min-max) vs. F39</td>
<td>0.0241</td>
<td>0.0015</td>
<td>0.0212</td>
<td>0.0270</td>
<td>0.0013</td>
<td>2</td>
<td>0.0167</td>
<td>**</td>
</tr>
<tr>
<td>3</td>
<td>All33 (min-max) vs. All251 (L2)</td>
<td>0.0120</td>
<td>0.0008</td>
<td>0.0105</td>
<td>0.0136</td>
<td>0.0014</td>
<td>3</td>
<td>0.0250</td>
<td>**</td>
</tr>
<tr>
<td>4</td>
<td>All33 (min-max) vs. F13</td>
<td>0.0120</td>
<td>0.0008</td>
<td>0.0105</td>
<td>0.0136</td>
<td>0.0014</td>
<td>4</td>
<td>0.0333</td>
<td>**</td>
</tr>
<tr>
<td>5</td>
<td>All33 (min-max) vs. F31</td>
<td>0.0120</td>
<td>0.0008</td>
<td>0.0105</td>
<td>0.0136</td>
<td>0.0014</td>
<td>5</td>
<td>0.0417</td>
<td>**</td>
</tr>
<tr>
<td>6</td>
<td>All33 (min-max) vs. F35</td>
<td>-0.0121</td>
<td>0.0009</td>
<td>-0.0137</td>
<td>-0.0104</td>
<td>0.0017</td>
<td>6</td>
<td>0.0500</td>
<td>**</td>
</tr>
</tbody>
</table>

**Statistical significance at 0.05 level using Benjamini–Hochberg correction procedure.

---

**Table 8:** Paired sample t-test of the difference in specificity between the best-performing model and other models.

<table>
<thead>
<tr>
<th>No.</th>
<th>Pairs of RVMs</th>
<th>Mean difference</th>
<th>SD</th>
<th>95% CI Lower</th>
<th>95% CI Upper</th>
<th>p value</th>
<th>Rank</th>
<th>(i/m) Q</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All33 (min-max) vs. All 251 (L2)</td>
<td>0.1412</td>
<td>0.0110</td>
<td>0.1195</td>
<td>0.1628</td>
<td>0.0020</td>
<td>1</td>
<td>0.0083</td>
<td>**</td>
</tr>
<tr>
<td>2</td>
<td>All33 (min-max) vs. F35</td>
<td>0.1294</td>
<td>0.0105</td>
<td>0.1088</td>
<td>0.1500</td>
<td>0.0022</td>
<td>2</td>
<td>0.0167</td>
<td>**</td>
</tr>
<tr>
<td>3</td>
<td>All33 (min-max) vs. F31</td>
<td>0.0706</td>
<td>0.0068</td>
<td>0.0573</td>
<td>0.0839</td>
<td>0.0031</td>
<td>3</td>
<td>0.0250</td>
<td>**</td>
</tr>
<tr>
<td>4</td>
<td>All33 (min-max) vs. F39</td>
<td>0.0471</td>
<td>0.0048</td>
<td>0.0376</td>
<td>0.0566</td>
<td>0.0035</td>
<td>4</td>
<td>0.0333</td>
<td>**</td>
</tr>
<tr>
<td>5</td>
<td>All33 (min-max) vs. F13</td>
<td>0.0353</td>
<td>0.0038</td>
<td>0.0279</td>
<td>0.0427</td>
<td>0.0038</td>
<td>5</td>
<td>0.0417</td>
<td>**</td>
</tr>
<tr>
<td>6</td>
<td>All33 (min-max) vs. F27</td>
<td>0.0235</td>
<td>0.0026</td>
<td>0.0185</td>
<td>0.0286</td>
<td>0.0040</td>
<td>6</td>
<td>0.0500</td>
<td>**</td>
</tr>
</tbody>
</table>

**Statistical significance at 0.05 level using Benjamini–Hochberg correction procedure.
of lexical dispersion rates (D4), the 10 optimised semantic features also had statistically higher means in potentially “risky” than in “safe” English mental health materials with regard to their suitability for automatic machine translation: A2 (changes, modifications) (safe texts: \( M = 15.429, \ SD = 15.093, SE = 1.190 \); risky texts: \( M = 24.185, SD = 20.646, SE = 1.811, p < 0.001 \)); A3 (existing status of objects, things, people) (safe texts: \( M = 37.267, SD = 28.570, SE = 2.252 \); risky texts: \( M = 57.062, SD = 41.987, SE = 3.682, p < 0.001 \)); A4 (classification) (safe texts: \( M = 4.311, SD = 5.299, SE = 0.418 \); risky texts: \( M = 8.062, SD = 10.504, SE = 0.921, p < 0.001 \)); A6 (comparison) (safe texts: \( M = 12.689, SD = 11.913, SE = 0.939 \); risky texts: \( M = 20.846, SD = 22.697, SE = 1.991, p < 0.001 \)); A7 (probability) (safe texts: \( M = 24.373, SD = 19.357, SE = 1.526 \); risky texts: \( M = 37.262, SD = 25.795, SE = 2.262, p < 0.001 \)); A13 (degree adverbs) (safe texts: \( M = 9.857, SD = 9.073, SE = 0.715 \); risky texts: \( M = 15.654, SD = 12.487, SE = 1.095, p < 0.001 \)); E5 (trepidation, courage, surprise) (safe texts: \( M = 3.901, SD = 6.927, SE = 0.546 \); risky texts: \( M = 11.515, SD = 18.045, SE = 1.583, p < 0.001 \)); O4 (physical attributes) (safe texts: \( M = 5.509, SD = 5.671, SE = 0.447 \); risky texts: \( M = 8.115, SD = 6.668, SE = 0.585, p < 0.001 \)); Z5 (functional words) (safe texts: \( M = 217.242, SD = 151.680, SE = 11.954 \); risky texts: \( M = 326.238, SD = 222.992, SE = 19.558, p < 0.001 \)); and Z6 (negative particles) (safe texts: \( M = 7.944, SD = 7.246, SE = 0.571 \); risky texts: \( M = 12.062, SD = 9.090, SE = 0.797, p < 0.001 \)).

Next, we examined the optimised feature set of English sentiment features (L10). This includes clout expressions (negative particles) (safe texts: \( M = 86.714, SD = 13.275, SE = 1.046 \); risky texts: \( M = 81.142, SD = 16.857, SE = 1.478, p = 0.004 \)); emotional tones (safe texts: \( M = 29.106, SD = 32.179, SE = 2.536 \); risky texts: \( M = 18.367, SD = 27.907, SE = 2.448, p < 0.001 \)); words per sentences (safe texts: \( M = 18.714, SD = 5.075, SE = 0.400 \); risky texts: \( M = 19.728, SD = 4.813, SE = 0.422, p = 0.009 \)); and (third person pronouns) (safe texts: \( M = 0.746, SD = 0.660, SE = 0.052 \); risky texts: \( M = 1.017, SD = 0.953, SE = 0.084, p = 0.028 \)); affect words (safe texts: \( M = 8.581, SD = 2.517, SE = 0.198 \); risky texts: \( M = 9.382, SD = 2.604, SE = 0.228, p = 0.002 \)); negative emotions (safe texts: \( M = 4.830, SD = 2.609, SE = 0.206 \); risky texts: \( M = 6.127, SD = 3.165, SE = 0.278, p < 0.001 \)); anxiety words (safe texts: \( M = 3.049, SD = 2.128, SE = 0.168 \); risky texts: \( M = 4.088, SD = 2.659, SE = 0.233, p < 0.001 \)); tentativeness words (safe texts: \( M = 5.043, SD = 1.643, SE = 0.129 \); risky texts: \( M = 5.599, SD = 1.590, SE = 0.139, p = 0.005 \)); differentiation (safe texts: \( M = 4.176, SD = 1.515, SE = 0.119 \); risky texts: \( M = 4.717, SD = 1.333, SE = 0.117, p = 0.002 \)); and core drives and needs (reward focus) (safe texts: \( M = 1.731, SD = 1.083, SE = 0.085 \); risky texts: \( M = 1.339, SD = 0.858, SE = 0.075, p = 0.001 \)).

Within the optimised feature set of structural linguistic features (S5), there were 5 optimised features. Like the other three sets of optimised features, features retained in S5 had statistically higher means in “risky” texts than in “safe” English health texts: number of difficult sentences (more than 22 words) (safe texts: \( M = 10.491, SD = 8.821, SE = 0.695 \); risky texts: \( M = 16.200, SD = 14.048, SE = 1.23, p < 0.001 \)); number of monosyllabic words (safe texts: \( M = 560.186, SD = 358.796, SE = 28.277 \); risky texts: \( M = 811.446, SD = 515.400, SE = 45.204, p < 0.0001 \)); number of long (6+ characters) words (safe texts: \( M = 280.255, SD = 215.525, SE = 16.986 \); risky texts: \( M = 439.215, SD = 347.782, SE = 30.502, p < 0.0001 \)); number of sentences which use same words multiple times (safe texts: \( M = 10.814 \), risky texts: \( M = 5.427, SD = 4.123, SE = 0.339, p < 0.001 \)).

**Statistical significance at 0.05 level using Benjamini–Hochberg correction procedure.**
We found that the accuracy, sensitivity, and specificity of Bayesian classifiers based on these separately optimised features were suboptimal, in spite of the individual features retained in each optimised feature set being statistically significant features. Recent studies suggest that statistical significance and predictivity of features are often taken as exchangeable concepts, mistakenly [22–26]. Adding statistically significant features identified between case and control samples however do not necessarily improve the predictive performance of machine learning classifiers. This was verified in our study through joint optimisation of different feature sets combining lexical, semantic, sentiment, and structural features (251 features in total). The joint optimisation led to an optimised mixed feature set of 33 features, including 5 which did not have statistically different distribution in “safe” versus “risky” English mental health texts: K5 (leisure, activities) (safe texts: M = 2.596, SD = 5.473, SE = 0.431; risky texts: M = 2.292, SD = 2.900, SE = 0.254, p = 0.346); M3 (means of transport on land) (safe texts: M = 0.839, SD = 1.680, SE = 0.132; risky texts: M = 2.038, SD = 7.276, SE = 0.638, p = 0.076); Z3 (organisations names) (means of transport on land) (safe texts: M = 2.466, SD = 6.791, SE = 0.535; risky texts: M = 1.854, SD = 2.910, SE = 0.255, p = 0.893); average number of sentences per paragraph (safe texts: M = 2.160, SD = 3.394, SE = 0.267; risky texts: M = 2.434, SD = 6.268, SE = 0.550, p = 0.384); and number of questions (safe texts: M = 3.068, SD = 3.607, SE = 0.284; risky texts: M = 3.815, SD = 5.059, SE = 0.444, p = 0.312).

All other features in the jointly optimised feature set had statistically higher means in “risky” than in “safe” English mental health materials. Specifically, this included A3 (being/existing) (p < 0.001), A13 (degree adverbs) (p < 0.001), A15 (abstract terms of safety, danger) (p < 0.001), B2 (physical conditions) (p < 0.001), B3 (medical treatment) (p < 0.001), E5 (trepidation, courage, surprise) (p < 0.001), E6 (apprehension, confidence) (p < 0.001), K5 (leisure, activities) (M1 (movement), M3 (transport on land), M5 (transport by air), M6 (points of reference), N3 (measurement), O4 (physical attributes), S1 (social action, state, process), S5 (affiliation), T1 (time), X2 (reasoning, belief, scepticism), Z3 (organisations names), Z5 (functional words), Z8 (prouns))

<table>
<thead>
<tr>
<th>Table 10: Results of automatic feature optimisation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimised features (number)</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Lexical dispersion rates (4)</td>
</tr>
<tr>
<td>Semantic features (10)</td>
</tr>
<tr>
<td>Sentiment features (10)</td>
</tr>
<tr>
<td>Structural features (5)</td>
</tr>
<tr>
<td>All features (33)</td>
</tr>
</tbody>
</table>

Semantic features (21): A3 (being/existing), A13 (degree adverbs), A15 (abstract terms of safety, danger), B2 (physical conditions), B3 (medical treatment), E5 (trepidation, courage, surprise), E6 (apprehension, confidence), K5 (leisure, activities), M1 (movement), M3 (transport on land), M5 (transport by air), M6 (points of reference), N3 (measurement), O4 (physical attributes), S1 (social action, state, process), S5 (affiliation), T1 (time), X2 (reasoning, belief, scepticism), Z3 (organisations names), Z5 (functional words), Z8 (prouns)

Sentiment features (L4): clout expressions, affect words, negative emotions, anxiety

Structural features (S7): number of sentences using the same words multiple times (overused words), average number of sentences per paragraph, number of questions, number of proper nouns, number of unique multiple (3+) syllable words, number of unique long (more than 6 letters) words, out-of-dictionary words.

SD = 12.263, SE = 0.966; risky texts: M = 19.177, SD = 22.999, SE = 2.017, p < 0.0001; and passive voice (safe texts: M = 2.294, SD = 5.639, SE = 0.444; risky texts: M = 4.931, SD = 6.191, SE = 0.543, p < 0.0001).

The major advantage of the relevance vector machine classifier (RVM) based on the optimised and mixed feature set was the balanced sensitivity (0.651) and specificity (0.741), which made the instrument more applicable and useful in practical settings such as development and evaluation of mental health education and promotion resources for
Spanish-speaking patients. The list of optimised linguistic features included in the best-performing classifier also provides important opportunities for health professionals to make well-targeted, cost-effective interventions to English health materials to improve their suitability for automatic translation purposes. For example, health professionals could adjust the distribution patterns of relevant linguistic features, especially those associated with higher risks of causing automatic translation mistakes, and rerun the automatic assessment of the English input materials using our machine learning classifier, iteratively, until the predicted risk level reaches an acceptable level. Importantly, this process does not require any linguistic knowledge on the part of English-speaking medical professionals of patients’ language (in this case, Spanish).

6. Conclusions

Our paper developed probabilistic machine learning algorithms to assess and predict the levels of risks of using the Google Translate application in translating and delivering mental health information to Spanish-speaking populations. Our model can inform clinical decision making around the usability of the online translation tool when translating different original English texts on anxiety disorders into Spanish. This was achieved through the probabilistic prediction of Bayesian machine learning classifiers: if an input English text was assigned a high probability (over 50%) of causing erroneous and misleading automatic translation output, health professionals should become alert of the risk of using Google Translate; by contrast, if an input English text was assigned a low risk probability (below 50%), health professionals can feel reassured that the whole piece of English information can be translated safely to its intended user, using the online automatic translation tool. The smaller the risk probability of an English text is, the safer it is for the text to be translated automatically online. For original English materials which were labelled as non-suitable for automatic translation, our machine learning offers the opportunity to adjust, modify, and fine-tune the text to improve its suitability for automatic translation. This was achieved through the feature optimisation technique developed in our study. An important and useful feature of our model is that it does not require any linguistic knowledge on the part of English-speaking medical professionals of the patients’ language. The classifier can be applied as a practical decision aid to help increase the efficiency and cost-effectiveness of multicultural health communication, translation, and education.

Appendix

A. Annotation of Feature Sets

A.1. Structural Features (24). This includes average number of sentences per paragraph, number of difficult sentences (more than 22 words), average sentence length, number of interrogative sentence (question), number of exclamatory sentences, average number of characters, average number of syllables, number of numerals, number of proper nouns, number of monosyllabic words, number of unique monosyllabic words, number of complex (3+ syllable) words, number of unique 3+ syllable words, number of long (6+ characters) words, number of unique long words, out-of-dictionary words, repeated words, wording errors, redundant phrases, number of sentences which use the same word multiple times, wordy items, cliché, passive voice, and sentences that begin with conjunctions.

A.2. Lexical Dispersion Rates (20). This includes DiSp1: 0.0–0.1, DiSp2:0.1–0.2, DiSp3:0.2–0.3, DiSp4:0.3–0.4, DiSp5: 0.4–0.5, DiSp6:0.5–0.6, DiSp7:0.6–0.7, DiSp8:0.7–0.8, DiSp9: 0.8–0.9, DiSp10: 0.9–1.0, DiWr1:0.0–0.1, DiWr2:0.1–0.2, DiWr3:0.2–0.3, DiWr4:0.3–0.4, DiWr5:0.4–0.5, DiWr6: 0.5–0.6, DiWr7:0.6–0.7, DiWr8:0.7–0.9, DiWr9:0.8–0.9, and DiWr10: 0.9–1.0.

A.3. English Lexical Semantic Features (115) (USAS). This includes general and abstract terms (A1–A15, 15 features); the body and the individual (B1–B5, 5 features); arts and crafts (C1); emotion (E1–E6, 6 features); food and farming (F1–F4, 4 features); government and public (G1–G3, 3 features); architecture, housing and the home (H1–H5, 5 features); money and commerce in industry (I1–I4, 4 features); entertainment, sports and games (K1–K6, 6 features); life and living things (L1–L3, 3 features); movement, location, travel and transport (M1–M8, 8 features); numbers and measurements (N1–N6, 6 features); substances, materials, objects and equipment (O1–O4, 4 features); education (P1), language and communication (Q1–Q4, 4 features); social actions, states, processes (S1–S9, 9 features); time (T1–T4, 4 features); world and environment (W1–W5, 5 features); psychological actions, states and processes (X1–X9, 9 features); science and technology (Y1–Y2, 2 features); and grammar (Z0–Z9, Z99, 11 features).

A.4. Lexical Sentiment Features (92) (LWIC). This includes analytical thinking (describing the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns), clout (describing the relative social status, confidence, or leadership that people display through their writing or talking), authenticity (describing how people reveal themselves in an authentic or honest way), emotional tone (the higher the number, the more positive the tone; numbers below 50 suggest a more negative emotional tone), words per sentence, words longer than 6 letters, dictionary words, function words, pronouns, personal pronouns, personal pronouns (first per singular: I), personal pronouns (first person plural: we), personal pronouns (2nd person: you), personal pronouns (third person singular: she or he), personal pronouns (third person plural: they), impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, negations, verbs, adjectives, comparatives, interrogatives, numbers, quantifiers, affect words, positive emotions, negative emotions, anxiety, anger, sad, social words, family, friends, female referents, male
Acknowledgments

This study was supported by grants from the Australian Research Council (no. DP150102405).

References


