4

Cultural and Linguistic Bias of Neural Machine Translation Technology

MENG JI

4.1 Introduction

Neural translations are not neutral. On the contrary, as a new dilemma for neural machine translation as neural machine translation systems have learned to recognize patterns in lexical and semantic units in human languages (Johnson et al., 2017; Pope et al., 2020; Grechishnikova, 2021) to translate more fluently, increasing cultural bias in the target language has emerged. Given that language use is heavily influenced by the culture of the host country and carries with it deeply ingrained perceptions, beliefs, and attitudes (Downes, 1998; Fishman, 2019; Thomas and Wareing, 1999; Montgomery, 1995), increasingly fluent translations can increasingly convey those cultural aspects, and sometimes bring cultural biases along with them. In this respect, machine biases induced by translation are inevitable consequences of algorithms designed to achieve near-native level linguistic naturalness and communicative fluency in automatic translation outputs (Weng et al., 2020; Feng et al., 2020; Martindale et al., 2019; Koehn, 2020;Wu et al., 2016). And in fact, University of Cambridge researchers have indeed discovered gender bias in machine translations of English into German, Spanish, and Hebrew, chosen for their distinct linguistic and cultural properties (Saunders and Byrne, 2020). Their studies revealed that, in MT output, gender bias in particular was an inevitable consequence of language use in training datasets that included genres such as news reports and speeches. Similarly, several studies of machine translation quality assessment revealed widespread racial, as well as gender, bias (Tomalin et al., 2021; Font and Costa-jussà, 2019; Salles et al., 2018; Best, 2017). In the machine translations of job titles in the U.S. Bureau of Labor Statistics, Prates et al. (2019) showed a strong tendency toward male defaults in as many as twelve languages

(Hungarian, Chinese, Japanese, Basque, Yoruba, Turkish, Malay, Armenian, Swahili, Estonian, Bengali, Finnish)[1].

While commendable progress has been made in developing scalable approaches to reducing such bias in machine translations, the problem persists. Significant human effort is still required to revise MT training data, and the resulting MT datasets still do not address all forms of social discrimination inherited from target-language datasets. In our current study, we will speak of this issue as inevitably arising from the social and cultural constraints of artificial intelligence. Since human thoughts and behaviors do have social and cultural contexts, sexist, racial, class, and other types of bias are inevitable in MT output; and artificial intelligence, as our brainchild, will inevitably amplify these tendencies. Nevertheless, they are predictable and preventable. We argue that MT quality assessment should incorporate social, ethical, and cultural sensitivity, rather than focusing solely on linguistic accuracy and fluency. And specifically, for materials generated by neural translation, it is necessary to develop mechanisms to support decision-making concerning the trade-offs involving linguistic fluency and cultural biases. Special attention is needed in specialized domains. One such, multicultural mental healthcare, provides the focus of our study.

Globally, anxiety disorders are the largest burden on mental health (3.76 percent in 2017). Countries with the highest prevalence of anxiety disorders (5 percent–6 percent) are some of the most advanced economies in their regions (Argentina, Brazil, Chile, Uruguay) and worldwide (U.S., Canada, UK, Germany, Australia, Sweden, Spain, France, Italy, Norway, New Zealand, Denmark, Ireland), as well as a few countries in the Middle East and Africa (Algeria, Iran). In Asia, only a few countries ranked within this range (5–6 percent), even though anxiety disorders are traditionally prevalent in countries like Japan, South Korea, and, more recently, India and China. Developing Latin American countries, too, have a tradition of anxiety. However, we can ask whether mental disorders are openly treated in different countries, and whether differences in openness might affect these statistics. The use of medical and mental healthcare can be subject to discrimination and stigma. And in fact, we do find that, in some rapidly developing countries, mental health issues are underrepresented, even though their populations are exposed to environmental, social, and economic stressors. This underrepresentation might be a result of traditional cultural beliefs stigmatizing people with mental illnesses, and perhaps from a related lack of access to mental healthcare support.

[1] Some of the language choices have been criticized, however. Chinese and Japanese are not gender neural languages, for example.

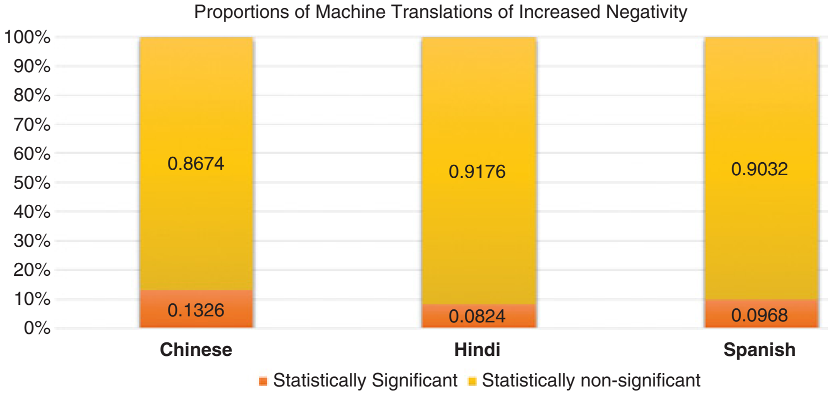


Figure 4.1 Ratios of machine translations of statistically increased negative emotions

We would expect such negative social sentiments or attitudes. Thus, to test our hypothesis, we gathered and generated translations into Chinese, Hindi, and Spanish of public health materials on anxiety disorders developed by health-promotion organizations in English-speaking countries. The back-translations from these three languages were paired with their original English health materials, and the distribution of negative emotion words in each pair was statistically analyzed (Figure 4.1).

4.2 Data Collection

We developed a set of quality control criteria based on five considerations for searching English public health materials on anxiety disorders. In gathering our training data, we screened online mental health information on this topic according to these criteria. Our intent has been to ensure the usefulness of the machine learning classifiers we have developed for health organizations and their wide applicability in research and clinical settings for effective, positive cross-lingual health communication concerning mental health disorders among multicultural populations.

- **Topic Relevance**: Our study focuses on anxiety disorders, due to their high prevalence
- **Information accessibility**: The materials we selected were written in an accessible, familiar style. Materials of this type are more likely to be translated by Google into language that the general public, as opposed to health

professionals, can understand. As the linguistic difficulty of English material increases, training data will be restricted to professional-oriented health resources. These will be less suitable for learning about the common language in a given country and the social attitudes toward mental disorders conveyed in its language.

- **Information credibility**: The English health materials selected were developed by national or charitable health-promotion organizations to ensure credibility (see Appendix 4).
- **Understandability**: Online health information materials may be intended for professionals or for patients. We chose English materials intended for the public, since their translations can be better understood by machine translation users from diverse cultural and linguistic backgrounds and thus can impact their opinions on mental health conditions. This criterion was significant because an important objective of our research was to develop machine learning classifiers that would improve translation quality – that is, that would help to produce less biased machine translations that could contribute to more positive understanding of anxiety disorders. With this goal in mind, we developed classifiers to process English health materials in an accessible and understandable manner.
- **User relevance:** While understudied, relevance to specific users is another key indicator of mental health resource quality. The causes, symptoms, and treatment of mental disorders vary considerably among people of varying demographic characteristics – people of different ages, genders, socioeconomic classes, and so on. We assume that health information can be significantly improved by tailoring it to specific user groups, and accordingly collected online English materials concerning anxiety disorders developed for children, teens, young adults, the elderly, men, women, and transgender people.

As part of the quality control process, we identified websites of national, charitable health-promotion organizations and selected original English health materials that met the above criteria. There were 557 original English health materials. There are three sets of natural language features annotated on the original English and back-translation health materials: multiple semantic categories using the university of Lancaster Semantic Annotation System (USAS); word frequency bands (WFB); and lexical dispersion rates (LDR), with the last two based on the British National Corpus. The total number of annotation classes was 153 including semantic classes (115), WFB (18), and LDR (20) (see Appendix 2).

4.3 Development of Machine Learning Classifiers

We collected 557 original English health materials regarding anxiety disorders that met all our search criteria. We generated their machine translations into Chinese, Hindi, and Spanish using the Google Translate API. We then compared the original English with their matching back-translations from the three languages and used the Linguistic Inquiry and Word Count System (LIWC) (University of Texas at Austin) to find the distribution of words expressing negative emotions in both sets of English materials, original and back-translation. A Wilkson signed-rank test found back-translations from the target languages showed a statistically significant increase (p*0.05) in expressions of negative emotions when compared to their original English texts.

The original English texts were chosen since they yielded back-translations showing statistically increased sentiment negativity concerning anxiety disorders. Since our goal was to develop neural programs that could distinguish texts relatively likely to produce biased translations, we then manually developed training corpora as follows. We classified the original English texts into risky (1) versus safe (0) classes: risky English texts were associated with back-translations of increased negativity in at least one language of Chinese, Hindi, or Spanish; and safe texts were associated with back-translations in which negativity increase was statistically insignificant (p>=0.05) in all three test languages. Of the 557 texts collected to train and test machine learning classifiers, 428 texts were classified as safe (class 0) and 129 texts as risky (class 1).

Again, our goal was to distinguish texts that were safe, or unlikely to contain biased language, from those that were risky, or like to contain such language. We faced some analytical problems, however, in that (1) the languages we studied differed in their respective degrees of negativity and (2) our corpus contained many more safe than risky texts. The most negative translations were found in Chinese (13.26 percent), followed by Hindi (8.24 percent) and Spanish (9.68 percent); and within the three target languages, the ratio of English materials associated with increased negativity in machine translations and those without any negative machine translations was 3:10.

In other words, in statistical terms, our data was imbalanced – as would be the case, for example, if we attempted to distinguish legitimate credit card transactions from fraudulent ones, since the former will greatly outnumber the latter in any corpus. Fortunately, various techniques have been developed for handling such data imbalance. Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) in Python was applied to improve the balance between the two classes of machine translation output in terms of negative emotion words. We divided the whole dataset, after oversampling, into training (70 percent) and testing datasets, and performed five-fold cross-validation on the training dataset (see Table 4.1).

Table 4.1 *Training and testing datasets*

Training/Testing Classifiers		Class 0	Class 1
Before Oversampling	Before total	428	129
After oversampling	Training (70%)	303	296
	Testing (30%)	125	132
	Total	428	428

4.4 Feature Optimization

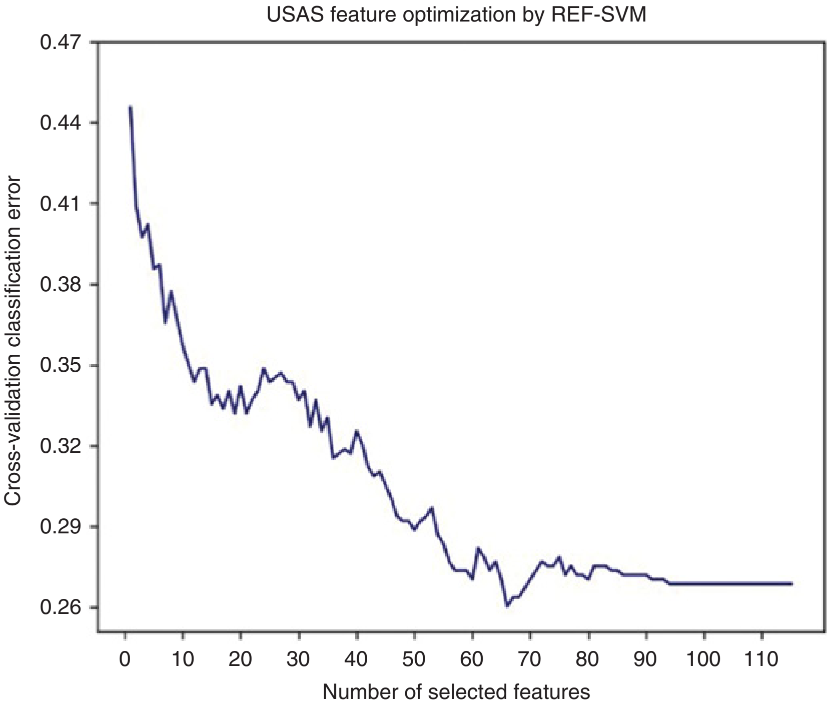
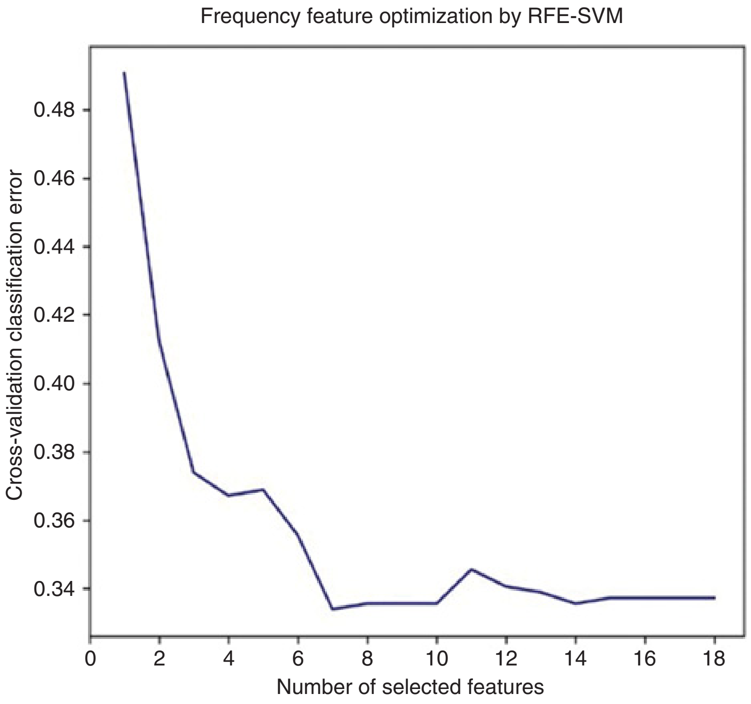
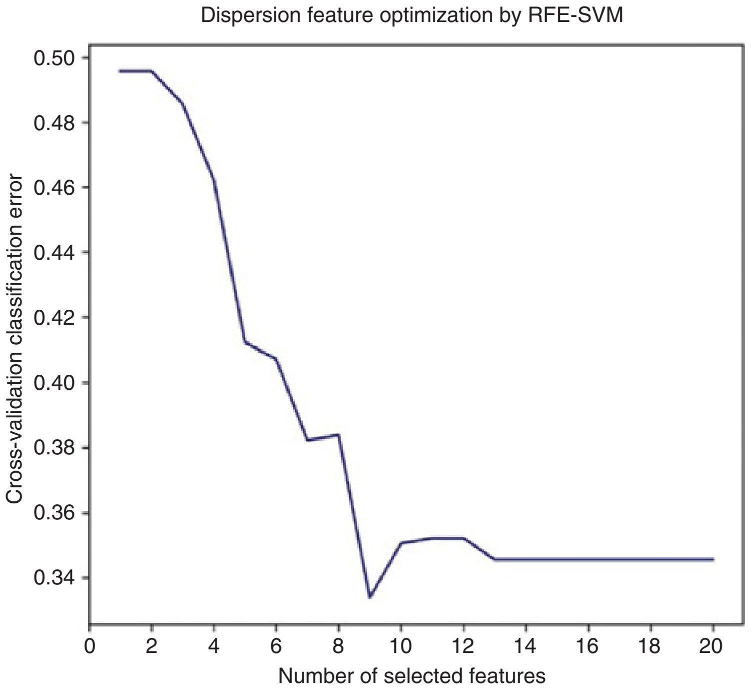


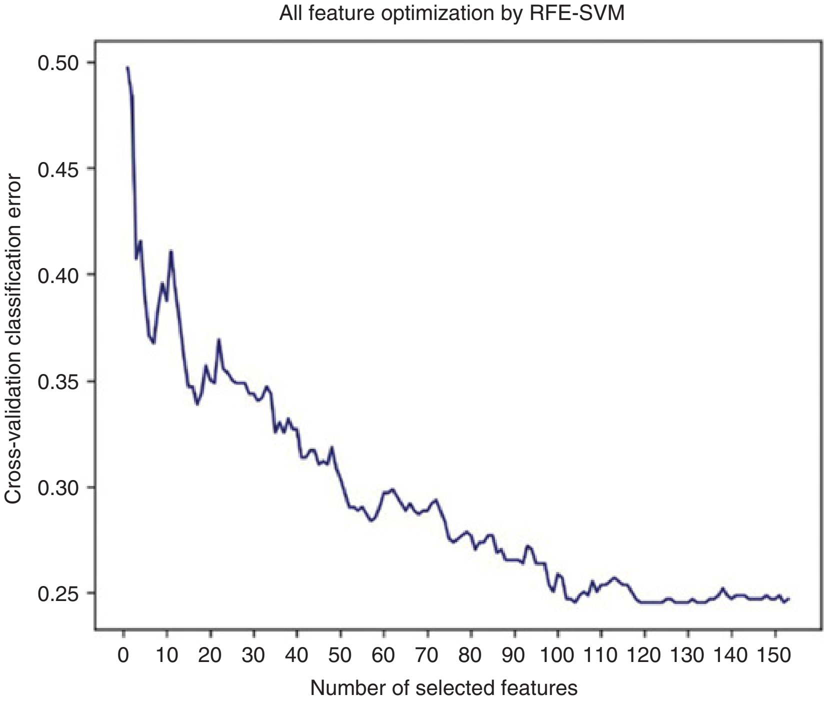
Figure 4.2 Recursive Feature Elimination with Automatic Feature Selection as the Base Estimator
Cross-validation classification error (CVCE)
(a) automatic optimization of English lexical dispersion features (from 20 to 9 feature, CVCE= 0.333)



(b) automatic optimization of English lexical frequency range features (from 18 to 7 features, CVCE= 0.333)



(c) automatic optimization of English semantic features (from 115 to 66 features, CVCE= 0.260)



(d) automatic optimization of all features (a, b, c) (from 153 to 119 features, CVCE=0.245)

4.5 Separate and Combined Feature Optimization

The dataset has 153 features, including 115 semantic classes, 18 WFB, and 20 LDR. By reducing high-dimensional feature sets, machine learning classifiers can be made more efficient and interpretable. Accordingly, support vector machine (SVM) methodology was used as the base estimator (RVM_SVM) in recursive feature elimination.

Figure 4.2 (a) shows the automatic optimization of English lexical dispersion rate features. After reduction of the LDR from 20 to 9, the cross-validation classification error reached its minimum (0.333). English lexis dispersion rates range from 0 to 1, with higher dispersion rates indicating wider distribution of the words across different textual genres, and thus indicating whether the relevant language is general or specialized. Both spoken and written dispersion rates were optimized: for very specialized words in spoken English (DiSp1:0.0–0.1, DiSp3:0.2–0.3); for general words in spoken English (DiSp6:0.5–0.6, DiSp9:0.8–0.9, DiSp10:0.9–1.0); and for medium-to-very-general words in written English (DiWr4:0.3–0.4, DiWr6:0.5–0.6, DiWr8:0.7–0.9, DiWr10:0.9).

In a comparable way, Figure 4.4 (b) shows the automatic optimization of English lexical frequency band features. When the number of features was reduced from 18 to 7, the minimal cross-validation classification error was obtained (0.333). In the British National Corpus (BNC), frequency bands refer to the ordinal ranges of word occurrence frequencies. For written materials, we listed nine frequency bands: FrWr1:0–500, FrWr2:500–1000, FrWr3:1000–1500, FrWr4:1500–2000, FrWr5:2000–2500, FrWr6:2500–3000, FrWr7:3000–3500, FrWr8:3500–4000, and FrWr9:4000–64420. The words that appear most frequently in the BNC corpus are those in the higher bands. For example, only 30 English words in the FrWr9 band occur more than 4,000 times in the entire database. Generally, the smaller the frequency bands, the less frequent or familiar the words are to the public. We also provided nine frequency bands for spoken materials: FrSp1:0–500, FrSp2:500–1000, FrSp3:1000–1500, FrSp4:1500–2000, FrSp5:2000–2500, FrSp6:2500–3000, FrSp7:3000–3500, FrSp8:3500–4000, and FrSp9:4000–57010. Again, higher frequencies indicate greater familiarity with words. We note that optimization of frequency band features reduced the original number of bands from 18 to 7: FrSp1:0–500, FrSp5:2000–2500, FrSp9:4000–57010, FrWr4:1500–2000, FrWr7:3000–3500, FrWr8:3500–4000, and FrWr9:4000–64420.

Finally, the automatic optimization of English semantic features is shown in Figure 4.4 (c). There were in total 115 semantic features covering as many as 21 semantic categories: 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); measurements (N1-N6, 6 features); substances, materials, objects and equipment (O1-O4, 4 features); education (P1), language communication (Q1-Q4, 4 features); social actions, states, processes (S1-S9, 9 features); time (T1-T4, 4 features); environment (W1-W5, 5 features); psychological actions, states and processes (X1-X9, 9 features); science and technology (Y1-Y2, 2 features); and names and grammar (Z0-Z9, Z99, 11 features).

The minimal classification error (0.260) was reached when the original semantic feature sets was reduced by almost half from 115 to 66: A12 (easy/difficult); A13 (degree, extent); A15 (safety/danger); A7 (probability); B1 (anatomy, physiology); B2 (health and disease); B3 (medicines, medical treatment); E2 (liking); E3 (calm/violent/angry); E4 (happiness, contentment); E5 (bravery, fear); E6 (worry, confidence); G2 (crime, law); I1 (money); I3 (employment); O4 (physical attributes); Q1 (linguistic actions, states, processes); S1 (social actions, states, processes); S2 (people); S8 (helping/hindering); S9 (religion); W1 (environment); W3 (geographical terms); X1 (psychological actions, states, processes); X3 (sensory); X4 (mental object); X5 (attention); X6 (Deciding); X7 (wanting, planning); X8 (trying); X9 (Ability); Z6 (negative); Z8 (pronouns); and so on. Figure 4.4 (d) shows the automatic optimization of the three sets of natural language features combined. The minimal classification error (0.245) was reached when the full feature set (153 features) was reduced to 119.

4.6 Classifier Training and Development

Relevance vector machine (RVM) methodology was used to develop Bayesian machine learning classifiers in Table 4.2. Different RVM models were compared using paired optimized and unoptimized feature sets, as well as their normalized versions, using three different techniques for feature normalization: min-maximal normalization (MMN), L2 normalization (L2 N), and Z-score normalization (ZSN). On the testing data, optimized feature sets of English LDR (Disp_9) achieved a higher area under the receiver operator characteristic (area under curve (AUC)=0.7023) than its matching unoptimized feature set (Disp_20) (AUC=0.7013). Feature normalization increased AUC of optimized and non-optimized feature sets to varying degrees. Optimization did not improve the performance of the feature set of WFB, since the AUC of Freq_7 on the testing data set (0.6626) was lower than Freq_18 (0.6784). By contrast, optimization did enhance the performance of RVMs using semantic features, as USAS_66 (0.7894) had a higher AUC than USAS_115 (0.773). With min-max normalization as the best technique, the AUC of the optimized model USAS_66 also increased.

The results of the separate optimizations of the English feature sets are shown in Table 4.2. Table 4.3 shows the results of combining the three feature sets. Although the optimized full feature set (F119) (AUC=0.778) did not achieve a higher AUC than the unoptimized full feature set (F153) (AUC=0.830), feature normalization significantly increased the AUC of classifier F119. The most effective normalization technique was min-max normalization, which increased the AUC of classifier F119 from 0.778 to 0.896, very similar to that of classifier F153 after the same normalization process (0.897).

Table 4.2 *Comparison of RVMs with full vs. separately optimized features sets*

	Training data	Testing data				
RVM	Mean AUC (STD)	AUC	Accuracy	Sensitivity	Specificity	Macro-F1
Full Feature Set (English LDR: Disp)						
Disp_20	0.6738 (0.0389)	0.7013	0.6381	0.6818	0.592	0.6367
Disp_20 (Min-Max normalization: MMN)	0.7943 (0.0329)	0.8147	0.7315	0.7121	0.752	0.7315
Disp_20 (L_2 normalization: L_2 N)	0.6632 (0.038)	0.7049	0.6304	0.7652	0.488	0.6212
Disp_20 (Z-score normalization: ZSN)	0.7899 (0.0275)	0.859	0.7899	0.7803	0.8	0.7899
Automatically Optimized Feature Set (English LDR: Disp)						
Disp_9	0.6709 (0.0409)	0.7024	0.6148	0.6742	0.552	0.6124
Disp_9 (MMN)	0.792 (0.0294)	0.8014	0.7588	0.7576	0.76	0.7587
Disp_9 (L_2 N)	0.6666 (0.0417)	0.7062	0.6459	0.7424	0.544	0.641
Disp_9 (ZSN)	0.8134 (0.0082)	0.8254	0.7626	0.7348	0.792	0.7626
Full Feature Set (English Lexical Frequency Bands: Freq)						
Freq _18	0.6906 (0.0429)	0.6784	0.6615	0.7652	0.552	0.6561
Freq_18 (MMN)	0.7911 (0.046)	0.8334	0.7626	0.8106	0.712	0.7615
Freq_ 18 (L_2 N)	0.6673 (0.0503)	0.652	0.6381	0.75	0.52	0.6317
Freq_18 (ZSN)	0.8052 (0.0326)	0.8343	0.786	0.8788	0.688	0.783
Automatically Optimized Feature Set (English Lexical Frequency Bands: Freq)						
Freq_7	0.6808 (0.0185)	0.6626	0.6381	0.7424	0.528	0.6324
Freq_7 (MMN)	0.6905 (0.0334)	0.6908	0.6148	0.6288	0.6	0.6144
Freq_7 (L_2 N)	0.6668 (0.0246)	0.6378	0.6226	0.7197	0.52	0.6174
Freq_7 (ZSN)	0.7163 (0.0277)	0.7905	0.7198	0.7348	0.704	0.7195

Table 4.2 *(cont.)*

	Training data	Testing data				
RVM	Mean AUC (STD)	AUC	Accuracy	Sensitivity	Specificity	Macro-F1
Full Feature Set (English Semantic Classes: USAS)						
USAS_115	0.7893 (0.0202)	0.773	0.7043	0.7045	0.704	0.7042
USAS_115 (MMN)	0.8623 (0.0302)	0.9219	0.856	0.8485	0.864	0.856
USAS_115 (L_2 N)	0.7767 (0.0413)	0.8092	0.751	0.7803	0.72	0.7503
USAS_115 (ZSN)	0.8652 (0.0302)	0.9092	0.856	0.8258	0.888	0.856
Automatically Optimized Feature Set (English Semantic Classes: USAS)						
USAS_66	0.8464 (0.0221)	0.7894	0.751	0.8106	0.688	0.7493
USAS_66 (MMN)	0.8814 (0.0347)	0.9053	0.8366	0.8182	0.856	0.8366
USAS_66 (L_2 N)	0.842 (0.0286)	0.8539	0.786	0.8485	0.72	0.7844
USAS_66 (ZSN)	0.8728 (0.0325)	0.8885	0.8249	0.8182	0.832	0.8249

Table 4.3 *Comparison of RVMs with full vs. jointly optimized features sets*

	Training data	Testing data			
Relevance Vector Machine (RVM)	Mean AUC (STD)	AUC	Accuracy	Sensitivity	Specificity
Full Feature Set (including Disp, Freq and USAS)					
Disp_20 + Freq _18 + USAS_115 = F153	0.780 (0.021)	0.830	0.755	0.765	0.744
F153 (MMN)	0.833 (0.045)	0.897	0.825	0.849	0.800
F153 (L_2 N)	0.774 (0.033)	0.780	0.697	0.735	0.656
F153 (ZSN)	0.863 (0.053)	0.878	0.809	0.788	0.832
Automatically Optimized Full Feature Set (including Disp, Freq and USAS)					
F119	0.776 (0.033)	0.778	0.689	0.674	0.704
F119 (MMN)	0.844 (0.058)	0.896	0.844	0.864	0.824
F119 (L_2 N)	0.788 (0.018)	0.803	0.735	0.765	0.704
F119 (ZSN)	0.846 (0.045)	0.893	0.817	0.803	0.832
Combinations of separately optimized feature Sets					
Freq_7 + Disp_9 + USAS_66 = F82	0.792 (0.013)	0.794	0.724	0.750	0.696
F82 (MMN)	0.853 (0.025)	0.906	0.837	0.841	0.832
F82 (L_2 N)	0.790 (0.025)	0.774	0.700	0.742	0.656
F82 (ZSN)	0.879 (0.023)	0.891	0.813	0.788	0.840

Table 4.3 *(cont.)*

	Training data	Testing data			
Relevance Vector Machine (RVM)	Mean AUC (STD)	AUC	Accuracy	Sensitivity	Specificity
Freq_ 7 + USAS_ 66 = F73	0.815 (0.022)	0.803	0.763	0.826	0.696
F73 (MMN)	0.856 (0.031)	0.903	0.825	0.856	0.792
F73 (L_2 N)	0.832 (0.032)	0.834	0.770	0.796	0.744
F73 (ZSN)	0.874 (0.014)	0.880	0.809	0.841	0.776
Disp_ 9 + USAS_ 66 = F75	0.811 (0.016)	0.805	0.739	0.765	0.712
F75 (MMN)	0.877 (0.039)	0.919	0.848	0.803	0.896
F75 (L_2 N)	0.803 (0.034)	0.783	0.732	0.780	0.680
F75 (ZSN)	0.878 (0.031)	0.890	0.837	0.833	0.840
Freq_ 7 + Disp_ 9 = F16	0.676 (0.03)	0.707	0.646	0.712	0.576
F16 (MMN)	0.768 (0.035)	0.793	0.728	0.735	0.720
F16 (L_2 N)	0.677 (0.029)	0.689	0.634	0.720	0.544
F16 (ZSN)	0.765 (0.030)	0.822	0.732	0.735	0.728

Pairwise comparisons were conducted of any two optimized feature sets to determine the best combination of features. The results show that the combination of optimized dispersion rates (Disp_9) and optimized semantic features (USAS_66) yielded the highest AUC on the testing data: F75 (AUC=0.919, sensitivity=0.803, specificity=0.896, accuracy=0.848), followed by the combination of all three optimized feature sets: F82 (AUC=0.906, sensitivity=0.841, specificity=0.832, accuracy=0.837). F75 thus emerged as the best model.

4.7 Statistical Refinement of the Optimized Classifier

In order to further improve the performance of the optimized classifier F75, we performed statistical analyses of the dispersion rate features and semantic features in the two sets of English mental health materials: labeled as 0, indicating no back-translation associated with statistically increased negative emotions, and labeled as 1, indicating back-translations with strong negative connotations in one or more of the three languages – Chinese, Hindi, and Spanish (Table 4.4). Appendix 3 shows the results of the Mann Whitney U test between the two sets of original English texts. As compared to "safe" original English materials, five features yielded statistically different distributions with respect to their respective probabilities of being translated into Chinese, Hindi, or Spanish with strong negativity: DiSp9:0.8–0.9 ($p<0.001$), DiSp10:0.9–1.0 ($p<0.001$), DiWr6:0.5–0.6

Table 4.4 *Comparison of RVMs with full vs. combined, separately optimized features sets*

	Training data	Testing data			
Relevance Vector Machine (RVM)	Mean AUC (STD)	AUC	Accuracy	Sensitivity	Specificity
Statistically Refined Feature Set based on the Automatic Optimization					
Disp_ 5 + USAS_66 = F71	0.817 (0.012)	0.807	0.755	0.796	0.712
F71 (L_2 N)	0.798 (0.029)	0.783	0.732	0.78	0.68
F71 (MMN)	0.884 (0.036)	0.865	0.829	0.849	0.808
F71 (ZSN)	0.867 (0.028)	0.886	0.856	0.856	0.856
Disp_ 9 + USAS_ 59 = F68	0.818 (0.018)	0.805	0.774	0.841	0.704
F68 (L_2 N)	0.805 (0.030)	0.766	0.728	0.803	0.648
F68 (MMN)	0.863 (0.034)	0.911	0.833	0.841	0.824
F68 (ZSN)	0.863 (0.035)	0.883	0.825	0.856	0.792
Disp_ 6 + USAS_ 59 = F65	0.815 (0.013)	0.806	0.751	0.788	0.712
F65 (L_2 N)	0.801 (0.028)	0.77	0.712	0.788	0.632
F65 (MMN)	0.866 (0.030)	0.881	0.848	0.856	0.84
F65 (ZSN)	0.867 (0.023)	0.885	0.833	0.841	0.824
Disp_5 + USAS_59 = F64	0.816 (0.017)	0.806	0.759	0.818	0.696
F64 (L_2 N)	0.804 (0.031)	0.767	0.716	0.796	0.632
F64 (MMN)	0.867 (0.039)	0.883	0.841	0.841	0.84
F64 (ZSM)	0.865 (0.021)	0.885	0.829	0.841	0.816
Disp 4 + USAS_59 = F63	0.822 (0.023)	0.799	0.743	0.773	0.712
F63 (L_2 N)	0.802 (0.036)	0.77	0.712	0.788	0.632
F63 (MMN)	0.866 (0.038)	0.885	0.825	0.826	0.824
F63 (ZSN)	0.868 (0.021)	0.887	0.829	0.841	0.816
Disp best 5 + USAS_59 = F best 64	0.817 (0.015)	0.805	0.77	0.826	0.712
F_ best 64 (L_2 N)	0.792 (0.045)	0.77	0.712	0.773	0.648
F_ best 64 (MMN)	0.870 (0.025)	0.886	0.848	0.849	0.848
F_ best 64 (ZSN)	0.865 (0.023)	0.893	0.852	0.864	0.84

(p=0.045), DiWr8:0.7–0.9 (p<0.001), and DiWr10:0.9–1.0 (p<0.001). The automatically selected features of LDR were reduced from 9 to 5. Similarly, the number of semantic features was reduced from 66 in the automatic feature selection (RVM_SVM) to 59. The classifier was subsequently fine-tuned by comparing four combinations of optimized LDR with optimized semantic features (USAS_59). (For details on the different dispersion rates used, see Appendix 2.) With ZSN, F (best 64) emerged as the best-performing classifier (AUC=0.893, accuracy=0.852, sensitivity=0.864, specificity=0.84).

RVM classifiers with different feature sets are compared in Tables 4.5 and 4.6. The comparison was to determine whether the sensitivity and specificity of the best-performing model were significantly higher than those of other classifiers.

Table 4.5 *Paired sample t test of the difference in sensitivity between the best model with other models*

No.	Pairs of RVMs	Mean Difference	S.D.	95% Confidence Interval of Difference		P value	Rank	(i/m)Q	Sig.
				Lower	Upper				
1	F best 64 (ZSN) vs. F75 (MMN)	0.0606	0.0093	0.0375	0.0837	0.0078	1	0.0056	**
2	F best 64 (ZSN) vs. F63 (ZSN)	0.0227	0.0039	0.0131	0.0323	0.0095	2	0.0111	**
3	F best 64 (ZSN) vs. F64 (MMN)	0.0227	0.0039	0.0131	0.0323	0.0095	3	0.0167	**
4	F best 64 (ZSN) vs. F68 (MMN)	0.0227	0.0039	0.0131	0.0323	0.0095	4	0.0222	**
5	F best 64 (ZSN) vs. F82 (MMN)	0.0227	0.0039	0.0131	0.0323	0.0095	5	0.0278	**
6	F best 64 (ZSN) vs. F153 (MMN)	0.0151	0.0026	0.0086	0.0216	0.0098	6	0.0333	**
7	F best 64 (ZSN) vs. F65 (MMN)	0.0075	0.0013	0.0042	0.0109	0.0103	7	0.0389	**
8	F best 64 (ZSN) vs. F71 (ZSM)	0.0075	0.0013	0.0042	0.0109	0.0103	8	0.0444	**
9	F best 64 (ZSN) vs. F119 (MMN)	0	0	0	0	1	9	0.0500	

To control for any false discovery rate, we applied the Benjamini–Hochberg correction procedure.

With respect to sensitivity, the results show that F (best 64) yielded statistically higher sensitivity than the other seven high-performing classifiers selected from the 72 classifiers we developed. There was no statistically significant difference between F (best 64) and F119 (jointly optimized features and normalized using min- max optimization). However, F (best 64) was much less complex with only 64 features.

With respect to specificity, F (best 64) gave statistically greater specificity than five high-performing classifiers (F63, F153, F68, F119, and F82), while F (best 64) gave statistically similar specificity to classifiers F64 and F65. The specificity of F (best 64) was statistically lower than that of F75 (MMN) and F71 (ZSM), but the sensitivity of F (best 64) was statistically higher than F75 (MMN) and F71 (ZSM).

As the primary aim of our study is to detect English texts that are more likely to be translated with strong negative connotations in the target languages, model sensitivity is more important than specificity. Therefore, F (best 64) was chosen as the best-performing classifier.

Table 4.6 *Paired sample t test of the difference in specificity between the best model with other models*

No	Pairs of RVMs	Mean Difference	S.D.	95% Confidence Interval of Difference		P value	Rank	(i/m)Q	Sig.
				Lower	Upper				
1	F best 64 (ZSN) vs. F63 (ZSN)	0.0240	0.0037	0.0149	0.0331	0.0077	1	0.006	**
2	F best 64 (ZSN) vs. F153 (MMN)	0.0400	0.0059	0.0255	0.0545	0.0071	2	0.011	**
3	F best 64 (ZSN) vs. F68 (MMN)	0.0160	0.0025	0.0098	0.0222	0.0080	3	0.017	**
4	F best 64 (ZSN) vs. F119 (MMN)	0.0160	0.0025	0.0098	0.0222	0.0080	4	0.022	**
5	F best 64 (ZSN) vs. F82 (MMN)	0.0080	0.0013	0.0048	0.0112	0.0083	5	0.028	**
6	F best 64 (ZSN) vs. F71 (ZSM)	−0.0160	0.0027	−0.0228	−0.0092	0.0095	6	0.033	**
7	F best 64 (ZSN) vs. F75 (MMN)	−0.0560	0.0108	−0.0827	−0.0293	0.0121	7	0.039	**
8	F best 64 (ZSN) vs. F64 (MMN)	0.0000	0.0000	0.0000	0.0000	1.0000	8	0.044	
9	F best 64 (ZSN) vs. F65 (MMN)	0.0000	0.0000	0.0000	0.0000	1.0000	9	0.050	

4.8 Model Stability

On the testing data set, Figure 4.3 shows how AUC varies when the size of the training data was adjusted from 150 to 550 on 100 intervals. The RVMs show themselves unlikely to have overfitting issues, unlike other classifiers such as extreme gradient boosting trees, random forests, and neural networks that require hyperparameter tuning. The RVM classifiers all demonstrated stability and scalability, as their performance (AUC) increased gradually as we increased the training dataset size. F (best 64) outperformed other classifiers when the size of training data exceeded that of testing data. Figure 4.4 shows the mean AUC of RVM classifiers on test data and Table 4.7 shows the paired sample t test of the AUC of these classifiers. Even though F (best 64) employed the smallest number of features, its mean AUC was comparable to that of other high-dimensional classifiers.

To review, then, we have succeeded in developing a high-performing relevance vector machine (RVM) classifier to predict the likelihood of a certain English health text being translated by Google as having statistically increased negative connotations when compared to the original English text.

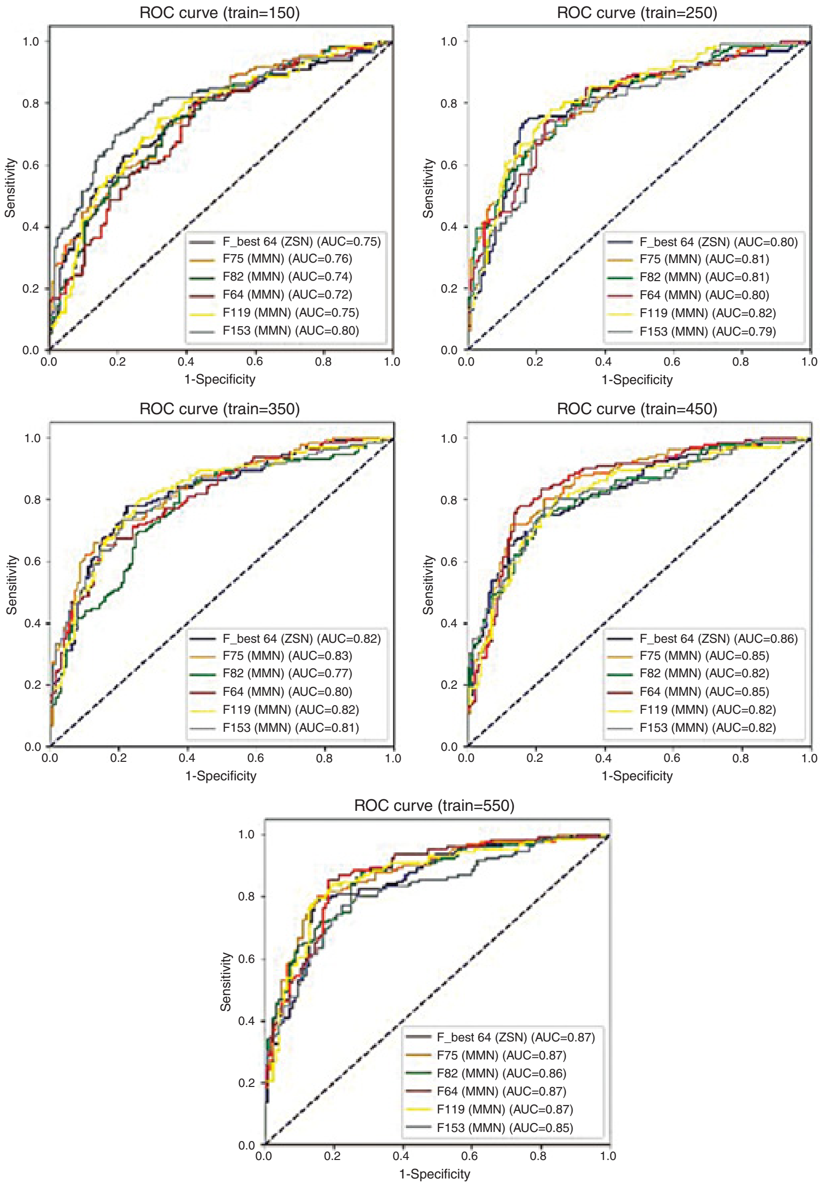


Figure 4.3 AUCs of RVMs on testing data using different training dataset sizes (150, 250, 350, 450, 550).

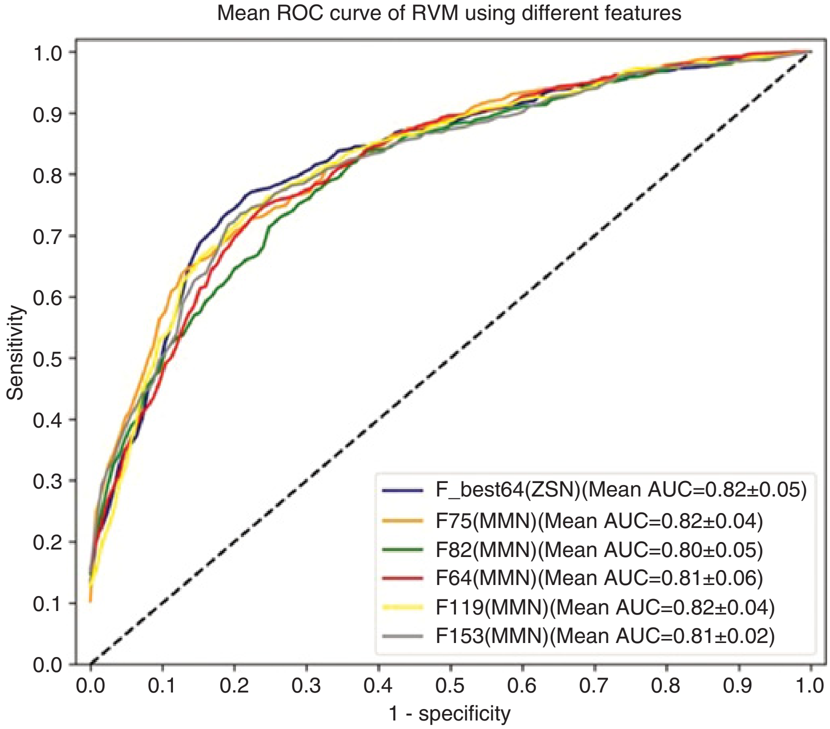


Figure 4.4 Mean AUC of RVMs on testing data using different training dataset size (150, 250, 350, 450, 550).

Table 4.7 *Paired sample t test of AUC of the best-performing classifier with other high-performing classifiers*

		Paired Mean Differences	Std. Error Mean	95% Confidence Interval of the Difference		Sig. (2-tailed)
				Lower	Upper	
Pair 1	Fbest 64 – F75	−0.0021	0.0053	−0.0167	0.0124	0.7039
Pair 2	Fbest 64 – F82	0.0201	0.0110	−0.0103	0.0506	0.1404
Pair 3	Fbest 64 – F64	0.0093	0.0047	−0.0037	0.0223	0.1176
Pair 4	Fbest 64 – F119	0.0043	0.0099	−0.0231	0.0318	0.6848
Pair 5	Fbest 64 – F153	0.0076	0.0156	−0.0357	0.0508	0.6526

Among the three languages we studied, the negative emotions and attitudes toward mental health – specifically, toward anxiety disorders – introduced by automatic translation were widespread. To understand the reasons, we carefully read the original English health texts in our database and their corresponding back-translations from the target languages. We found some illuminating examples, collected in Appendix 1. In some cases, mental health disorders have been translated by Google into Chinese as mental health diseases; people with anxiety disorders are describe in translation as mental illness patients; in Spanish, mental health conditions are translated as mental illnesses; in Hindi, shyness is translated as general shame, and panic is translated as nervousness. In Chinese, there is a subtle difference in word connotation: neural verbs such as “have a mental disorder” are replaced by “suffer from a mental disease.” As discussed above, neural machine translation emphasizes naturalness and fluency of translations, as compared with the more literal translations characteristic of statistical machine translation. The overall translation does improve significantly in terms of readability, fluency, and grammar; however, the accuracy of local translations may be compromised at the lexical and lexico-grammatical levels.

As Way noted in the following text:

> [Neural] MT output can be deceptively fluent; sometimes perfect target-language sentences are output, and less thorough translators and proofreaders may be seduced into accepting such translations, despite the fact that such translations may not be an actual translation of the source sentence at hand at all!

As we have seen, a direct result of the target-language-oriented approach to neural MT in mental health translation has been unintentional increased negativity and discrimination in the translation output, even though such connotations were absent in the original English mental health materials. And as discussed, we have developed machine learning classifiers mitigate this undesirable effect by detecting English mental health information that might lead to biased translation in the three languages.

Our RVM classifier reached its statistically highest sensitivity (mean=0.864, 95 percent C.I.: 0.805, 0.922) and specificity (mean=0.832, 95 percent C.I.: 0.766, 0.898) when the probability threshold was set at 0.5. However, the default threshold of 0.5 can be adjusted further to fine-tune the classifier. When the threshold of the classifier was increased, sensitivity decreased, while specificity increased; conversely, when the threshold was decreased, sensitivity increased, and specificity decreased. Thus, one can select the best sensitivity and specificity pairing according to the practical circumstances.

For example, high-sensitivity classifiers can be useful for screening purposes – to identify mental health materials in English which cannot be adequately translated by full, unverified automatic machine translation, due to their heightened likelihood of biased or discriminatory translations. By contrast, a low-sensitivity classifier is relatively unlikely to identify potentially problematic original English mental health information – that is, to screen out any materials that are not safe and suitable for neural machine translation. Thus, any social biases or discrimination against mental disorders still present in the target language would unfortunately be reinforced in machine-translated mental health resources, even if indirectly or unintentionally – clearly not the intent of global mental health promotion.

As for classifiers with high specificity, they are more suitable for identifying original English mental health texts which are suitable for neural machine translation, at least for the three languages we studied, Chinese, Hindi, and Spanish. On the other hand, when a classifier with a low specificity is used, there is a raised likelihood of false positive predictions: that is, even safe and suitable original English mental health information may be erroneously considered unsuitable for neural MT. While subsequent human post-evaluation could correct such inaccurate predictions, logistical burdens and staff costs would increase. This extra effort might well be impractical or prohibitive in low-resource healthcare service scenarios, often subject to tight budget constraints or lacking bilingual workers with sufficient knowledge of the relevant languages.

4.9 Conclusion

In comparison with earlier statistical machine learning models, current neural machine translation technologies exhibit greater linguistic fluency. However, as noted, while improving linguistic fluency, neural machine translation also learns, inevitably if unconsciously, to reflect the sentiments, attitudes, and biases of the target cultures, societies, and communities. Since the design favors the most natural sequence of translated words and phrases in the target language – its natural lexical and syntactic patterns – its results inevitably convey the social and cultural connotations of the cultures from which the languages sprang.

In many countries and cultures, mental disorders are still stigmatized and subject to discrimination. In associated languages, this deeply rooted sentiment comes to be reflected in conventional lexical and semantic units. Consequently, the neural machine translation of mental health information too often entails the transmission of negative social sentiments in the output, even when social prejudice against mental disorders is actually absent from the original

English materials. Our study has examined this understudied tendency in neural machine translation. We argue that this examination is appropriate and necessary as human communication technologies move rapidly toward more human-centric AI. Accordingly, we have developed Bayesian machine learning classifiers to assist with the probabilistic detection and prediction of socially biased neural machine translation outputs, using computational modeling and pairwise comparisons of original English and back-translations of neural machine translation outputs on anxiety disorders in Chinese, Hindi, and Spanish.

Via Google's Translate API, we collected and compared original English documents on anxiety disorders from U.S., UK, Canadian, and Australian health authorities and their back-translations from Chinese, Hindi, and Spanish. Through automatic, statistically informed feature optimization, RVM classifiers were developed. These models provided informative probabilistic predictions of the likelihood that an English text on anxiety disorders would be translated by Google into one of the three languages with subtle but strong negative connotations – again, because neural machine translation favors natural language translations.

The best-performing RVM (RVM_ best 64) contained 64 English linguistic features: 59 (semantic features) and 5 (LDR: DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr8:0.7–0.9, DiWr9:0.8–0.9, DiWr10:0.9–1.0). This result suggests that, in spoken and written English words belonging to certain semantic classes, words with high dispersion of meaning are relatively likely to produce negative neural machine translation results for anxiety disorders. The best-performing RVM (both optimized and normalized) achieved a mean AUC of receiver operator characteristic (0.893), accuracy (0.892), sensitivity (0.864), and specificity (0.84). Its sensitivity (SE) and specificity (SP) were statistically higher than those of unoptimized, normalized classifiers RVM_153 (min-max normalized MMN) (SE: p=0.0098, SP: p=0.007); of automatically optimized and normalized classifiers: RVM_82 (MMN) (SE: p=0.0095, SP: p=0.0083), RVM_75 (SE: p=0.0078, SP: p=0.0121); and of an automatically optimized and statistically refined classifier: RVM_71 (MMN) (SE: p=0.0103, SP; p=0.0095); RVM_68 (MMN) (SE: p=0.0095, SP: p=0.0080); RMV_63 (ZSM) (SE: p=0.0095, SP: p=0.0077). The stability of RVM_ best 64 appeared in its mean AUC (0.82, SD=0.05) when the training data sizes were reduced from 600 to 150.

Content analysis indicates that negative neural machine translations of anxiety disorders were primarily associated with increased linguistic fluency and communicative naturalness in the target Chinese, Hindi, and Spanish texts. However, once again, these stylistically enhanced phrasal patterns reflect persistent social attitudes toward mental health disorders in the relevant languages and cultures. And again, this bias is to be expected: as in any form of artificial intelligence, neural machine translation is designed to accommodate and understand human wants, needs, and thinking patterns.

This study demonstrates that phrasal patterning in target cultures does indeed yield increased negativity toward mental disorders as a consequence of greater translation naturalness. In compensation, however, we demonstrate that machine learning tools for the promotion of mental health translation can indeed detect instances of the automatic generation and dissemination of negative, discriminatory translation. In this way, neural tools can also promote positive and supportive social understanding and acceptance of mental disorders. This study confirms that, while neural machine translation technology is inevitably and increasingly culturally skewed, it can nevertheless be harnessed to foster more tolerant global health cultures.

References

Best, S. 2017. “Is Google Translate Sexist? Users Report Biased Results When Translating Gender-Neutral Languages into English,” The Daily Mail. Accessed January 28, 2020. www.dailymail.co.uk/sciencetech/article-5136607/Is-Google-Translate-SEXIST.html.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over- Sampling Technique,” *Journal of Artificial Intelligence Research*, 16, pages 321–357.

Downes, W. 1998. *Language and Society*, Volume 10, New York: Cambridge University Press.

Feng, Y., W. Xie, S. Gu, et al. 2020. “Modeling Fluency and Faithfulness for Diverse Neural Machine Translation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (01), pages 59–66. https://doi.org/10.1609/aaai.v34i01.5334.

Fishman, J. A. 2019. *The Sociology of Language: An Interdisciplinary Social Science Approach to Language in Society*. Rowley, MA: Newbury House.

Font, J. E. and M. R. Costa-jussà, 2019. “Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques,” *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Grechishnikova, D. 2021. “Transformer neural network for protein-specific de novo drug generation as a machine translation problem,” *Sci Rep* 11, 321. https://doi.org/10.1038/s41598-020-79682-4.

Johnson, M., M. Schuster, Q. V. Le, et al. 2017. “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation,” *Transactions of the Association for Computational Linguistics*, 5, pages 339–351.

Koehn, Philipp. 2020. *Neural Machine Translation*, Cambridge: Cambridge University Press.

Martindale, M. J., M. Carpuat, K. Duh, and P. McNamee. 2019. “Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation,” *Proceedings of MT Summit XVII*, 1, Dublin, August 19–23.

Montgomery, M. 1995. *An Introduction to Language and Society*. London: Routledge.

Popel, M., M. Tomkova, J. Tomek, et al. 2020. “Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals,” *Nat Commun*, 11, 4381. https://doi.org/10.1038/s41467-020-18073-9.

Prates, M. O. R., P. H. C. Avelar, and L. Lamb. 2019. Assessing Gender Bias in Machine Translation: A Case Study with Google Translate, https://arxiv.org/abs/1809.02208.

Ritchie, H. and M. Roser. 2018. Mental Health. Our World in Data. https://ourworldindata.org/mental-health.

Salles, A., M. Awad, L. Goldin, et al. 2019. “Estimating Implicit and Explicit Gender Bias Among Health Care Professionals and Surgeons,” *JAMA Netw Open*, 2 (7): e196545. DOI: 10.1001/jamanetworkopen.2019.6545.

Saunders, D. and B. Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. ACL. arXiv:2004.04498

Thomas, L. and S. Wareing. 1999. Language, Society and Power. London: Routledge.

Tomalin, M., B. Byrne, S. Concannon, et al. 2021. “The Practical Ethics of Bias Reduction in Machine Translation: Why Domain Adaptation Is Better Than Data Debiasing,” *Ethics Inf Technol*. https://doi.org/10.1007/s10676-021–09583–1.

Weng, R., H. Yu, X. Wei, and W. Luo. 2020. “Towards Enhancing Faithfulness for Neural Machine Translation,” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2675–2684, November 16–20, Association for Computational Linguistics.

Yonghui W., M. Schuster, Z. Chen, et al. 2016. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, https://arxiv.org/abs/1609.08144

Appendix 1 Examples of Back-Translations with Negative Connotations

Original	Back-Translation(from Chinese, Hindi, Spanish)
EN Those younger than 25 should be carefully watched for increased depression, agitation, irritability, *suicidality*, and *unusual* changes in behavior, especially at the beginning of treatment or when doses are changed.	CH People under 25 years of age should be carefully observed for depression, agitation, irritability, *suicide,* and *abnormal* behavior changes, especially at the beginning of treatment or when the dose is changed.

(cont.)

Original	Back-Translation(from Chinese, Hindi, Spanish)
EN	CH
People with Social Anxiety Disorder may feel *very uneasy* when talking with others, asking questions, going into a store, or ordering food in a restaurant.	People with social anxiety disorder may feel *very upset* when talking to others, asking questions, entering a store, or ordering food in a restaurant.
EN	CH
People with this *disorder* are afraid that others will judge them in a negative way and will lead to extreme embarrassment or rejection.	People with this *disease* are afraid that others will judge them in a negative way leading to extreme embarrassment or rejection.
EN	CH
Both males and females can *have* Social Anxiety Disorder.	Both men and women may *suffer* from social anxiety disorder.
EN	CH
Panic attacks are frequently mistaken for a medical event such as a *heart attack*.	Panic attacks are frequently mistaken for medical events such as a *heart disease*.
EN	CH
Family and other sources of social support can have a significant impact on the recovery process for *people* with panic disorder.	Family and other sources of social support can have a significant impact on the recovery process of *patients* with panic disorder.
EN	CH
Rather than denying their (people with social anxiety disorder) feelings, take the following steps to allow the person to feel seen and heard: *Remain supportive*	Rather than deny their (people with social anxiety disorder) feelings, take the following steps to make them feel seen and heard: *Stay supported*
EN	CH
The National Alliance on Mental Illness (NAMI) can help ***people with panic disorder*** and family members normalize the experience and help the individual know and realize, that they are not alone.	The National League for Mental Illness (NAMI) can help ***panic sufferers*** and family members normalize their experiences and help individuals understand and realize that they are not alone.

(cont.)

Original	Back-Translation(from Chinese, Hindi, Spanish)
EN For a person with panic disorder, social relationships can be an important way to cope with the symptoms of ***the condition***.	**SP** For a person with panic disorder, social relationships can be an important way of coping with the symptoms of ***the illness.***
EN A panic attack can be upsetting. It can sometimes be a ***challenging*** situation to deal with, but it is important to avoid seeming ***judgmental*** or upset.	**SP** A panic attack can be upsetting. It can sometimes be a ***difficult*** situation to deal with at times, but it is important to avoid coming across ***critical*** or upset.
EN The fear of Social Anxiety Disorder is extreme and is not the same ***ordinary shyness*** that many people sometimes feel.	**Hindi** The fear of Social Anxiety Disorder is extreme and is not the same ***general shame*** that many people sometimes feel.
EN Here are some possible symptoms of Social Anxiety Disorder: Anxiety or ***panic*** when interacting with others in social situation	**Hindi** Here are some possible symptoms of Social Anxiety Disorder: Anxiety or ***nervousness*** when interacting with others in social situation

Appendix 2 Description of Different Feature Sets

Feature Set Description	Abbrev.	Description of Items	Removed	Added
BNC Frequency Lists	Freq (20)	FrSp1:0–500, FrSp2:500–1000, FrSp3:1000–1500, FrSp4:1500–2000, FrSp5:2000–2500, FrSp6:2500–3000, FrSp7:3000–3500, FrSp8:3500–4000, FrSp9:4000–57010,		

(cont.)

Feature Set Description	Abbrev.	Description of Items	Removed	Added
		FrWr1:0–500, FrWr2:500–1000, FrWr3:1000–1500, FrWr4:1500–2000, FrWr5:2000–2500, FrWr6:2500–3000, FrWr7:3000–3500, FrWr8:3500–4000, FrWr9:4000–64420		
BNC Dispersion Lists	Disp (18)	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, DiWr10:0.9–1.0		
Original USAS List	USAS (115)	A1, A10, A11, A12, A13, A14, A15, A2, A3, A4, A5, A6, A7, A8, A9, B1, B2, B3, B4, B5, C1, E1, E2, E3, E4, E5, E6, F1, F2, F3, F4, G1, G2, G3, H1, H2, H3, H4, H5, I1, I2, I3, I4, K1, K2, K3, K4, K5, K6, L1, L2, L3, M1, M2, M3, M4, M5, M6, M7, M8, N1, N2, N3, N4, N5, N6, O1, O2, O3, O4, P1, Q1, Q2, Q3, Q4, S1, S2, S3, S4, S5, S6, S7, S8, S9, T1, T2, T3, T4, W1, W2, W3, W4, W5, X1, X2, X3, X4, X5, X6, X7, X8, X9, Y1, Y2, Z0, Z1, Z2, Z3, Z4, Z5, Z6, Z7, Z8, Z9, Z99		
Automatic selection RFE_SVM	Disp (9)	DiSp1:0.0–0.1, DiSp3:0.2–0.3, DiSp6:0.5–0.6, DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr4:0.3–0.4, DiWr6:0.5–0.6, DiWr8:0.7–0.9, DiWr10:0.9–1.0		
Automatic selection RFE_SVM	Freq (7)	FrSp1:0–500, FrSp5:2000–2500, FrSp9:4000–57010, FrWr4:1500–2000, FrWr7:3000–3500, FrWr8:3500–4000, FrWr9:4000–64420		
Automatic selection RFE_SVM	USAS (66)	A12, A13, A14, A15, A3, A7, B1, B2, B3, B4, B5, E2, E3, E4, E5, E6, G2, H1, H2, H3, H4, H5, I1, I2, I3, K4, K5, K6, L1, L2, M2, M3, M5, M6, M8, N2, O1, O3, O4, Q1, Q2, Q3, Q4, S1, S2, S8, S9, T1, T3, T4, W1, W3, X1, X3, X4, X5, X6, X7, X8, X9, Z3, Z4, Z6, Z7, Z8, Z99		

(cont.)

Feature Set Description	Abbrev.	Description of Items	Removed	Added
Statistical & Automatic selection RFE_SVM	USAS (59)	A12, A13, A14, A15, A3, A7, B1, B2, B3, B4, B5, E2, E3, E4, E5, E6, G2, H1, H2, H4, H5, I1, I2, I3, K4, K6, L1, M2, M3, M5, M6, M8, N2, O1, O4, Q1, Q2, Q3, Q4, S1, S2, S8, S9, T1, T3, T4, X3, X4, X5, X6, X7, X8, X9, Z3, Z4, Z6, Z7, Z8, Z99	H3,K5,L2, O3, W1,W3,X1	
Statistical & Automatic selection RFE_SVM	Disp (5)	DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr6:0.5–0.6, DiWr8:0.7–0.9, DiWr10:0.9–1.0	DiSp1:0.0–0.1, DiSp3:0.2–0.3, DiSp6:0.5–0.6, DiWr4:0.3–0.4	
Statistical & Automatic selection RFE_SVM	Disp (4)	DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr8:0.7–0.9, DiWr10:0.9–1.0	DiWr6:0.5–0.6	
Statistical & Automatic selection RFE_SVM	Disp (6)	DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr6:0.5–0.6, DiWr8:0.7–0.9, DiWr9:0.8–0.9, DiWr10:0.9–1.0		DiWr9:0.8–0.9
Statistical & Automatic selection RFE_SVM	Disp (best 5)	DiSp9:0.8–0.9, DiSp10:0.9–1.0, DiWr8:0.7–0.9, DiWr9:0.8–0.9, DiWr10:0.9–1.0	DiWr6:0.5–0.6	DiWr9:0.8–0.9

Appendix 3 Mann Whitney U Test of English Original and English Back-Translations of Chinese, Hindi, and Spanish Health Texts

	Mann Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
DiSp1:0.0–0.1	89648	181454	−0.78	0.435
DiSp2:0.1–0.2	91592	183398	0	1
DiSp3:0.2–0.3	90508	182314	−0.394	0.694
DiSp4:0.3–0.4	91592	183398	0	1
DiSp5:0.4–0.5	87469	179275	−2.233	0.026 **
DiSp6:0.5–0.6	88593	180399	−1.136	0.256 **
DiSp7:0.6–0.7	70673	162479	−5.817	0 **
DiSp8:0.7–0.8	58353.5	150159.5	−9.191	0 **
DiSp9:0.8–0.9	69068	160874	−6.228	0 **
DiSp10:0.9–1.0	76650	168456	−4.131	0 **
DiWr1:0.0–0.1	91592	183398	0	1
DiWr2:0.1–0.2	91592	183398	0	1
DiWr3:0.2–0.3	91592	183398	0	1

(cont.)

	Mann Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
DiWr4:0.3–0.4	90524.5	182330.5	−1.892	0.058
DiWr5:0.4–0.5	91592	183398	0	1
DiWr6:0.5–0.6	90736	182542	−2.004	0.045 *
DiWr7:0.6–0.7	91378	183184	−1	0.317
DiWr8:0.7–0.9	83319	175125	−2.829	0.005 **
DiWr9:0.8–0.9	77904.5	169710.5	−3.792	0 **
DiWr10:0.9–1.0	66993	158799	−6.801	0 **
FrSp1:0–500	63116.5	154922.5	−7.873	0 **
FrSp2:500–1000	76380.5	168186.5	−4.207	0 **
FrSp3:1000–1500	72000.5	163806.5	−5.424	0 **
FrSp4:1500–2000	73940.5	165746.5	−4.892	0 **
FrSp5:2000–2500	82849	174655	−2.425	0.015 **
FrSp6:2500–3000	84883	176689	−1.87	0.061
FrSp7:3000–3500	82505	174311	−2.54	0.011 **
FrSp8:3500–4000	87808	179614	−1.07	0.285
FrSp9:4000–57010	80821.5	172627.5	−2.982	0.003 **
FrWr1:0–500	66228	158034	−7.013	0 **
FrWr2:500–1000	69789.5	161595.5	−6.029	0 **
FrWr3:1000–1500	71272.5	163078.5	−5.621	0 **
FrWr4:1500–2000	76411.5	168217.5	−4.207	0 **
FrWr5:2000–2500	76381	168187	−4.304	0 **
FrWr6:2500–3000	77464.5	169270.5	−3.997	0 **
FrWr7:3000–3500	85839.5	177645.5	−1.628	0.104
FrWr8:3500–4000	86366.5	178172.5	−1.59	0.112
FrWr9:4000–64420	80821.5	172627.5	−2.982	0.003 **

Appendix 4 List of English Health Information Websites

https://au.reachout.com
https://familydoctor.org
https://foundrybc.ca
https://headspace.org.au
https://healthyfamilies.beyondblue.org.au
https://kidshealth.org
https://kidshelpphone.ca
https://medlineplus.gov
https://mindyourmind.ca
https://my.clevelandclinic.org
https://patient.info
https://psychcentral.com

https://riseabove.org.uk
www.anxietycanada.com
www.apa.org
www.betterhealth.vic.gov.au
www.beyondblue.org.au
www.blackdoginstitute.org.au
www.camh.ca
www.childline.org.uk
www.emedicinehealth.com
www.healthline.com
www.healthychildren.org
www.independentage.org
www.mayoclinic.org
www.medicinenet.com
www.menshealthforum.org.uk
www.mentalhealth.org.uk
www.msdmanuals.com/home
www.nami.org
www.papyrus-uk.org
www.postpartum.net
www.verywellmind.com
www.webmd.com
www.womenshealth.gov
https://youngmenshealthsite.org
https://youngminds.org.uk