Statistical Journal of the IAOS 39 (2023) 785–791
DOI 10.3233/SJI-230063
IOS Press

Classifying respondent comments from the 2021 Canadian Census of Population using machine learning methods[1]

Joanne Yoon
Data Science and Innovation Division, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6, Canada
Tel.: +1 343 542 5625; E-mail: joanne.yoon@statcan.gc.ca

Abstract. To improve the analysis of respondent comments from the Canadian Census of Population, data scientists at Statistics Canada compared and evaluated traditional machine learning, deep learning and transformer-based techniques. Cross-lingual Language Model-Robustly Optimized Bidirectional Encoder Representations from Transformers (XLM-R), a cross-lingual language model, fine-tuned on census respondent comments yield the best result of 89.91% F1 score overall despite language and class imbalances. Following the evaluation, the fine-tuned model was implemented successfully to objectively categorize comments from the 2021 Census of Population, with high accuracy. As a result, feedback from respondents was directed to the appropriate subject matter analysts, for them to analyze post-collection.

Keywords: Deep learning, multilingual text classification, machine learning, census respondent comments, natural language processing, young statistician prize 2023

1. Introduction

Once every five years, the Canadian Census of Population [1] provides a detailed and comprehensive statistical portrait of Canada and its population. The census is the only data source that provides consistent statistics for both small geographic areas [2] and small population groups including Indigenous populations [3] across Canada. Census analytical products [4] can be used at all levels to make informed decisions and research about Canadian society.

Comments that respondents fill in at the end of their census form help Statistics Canada understand respondents' experience with the census, comprehension of the concepts being measured and perceptions of the census. Two years in advance of every census, Statistics Canada conducts a Census Test on a population sample to help ensure that the questions are relevant and clear [5]. Collected feedback is factored into the final questionnaire content. Comments about specific questionnaire content, from both the Census Test and the Census of Population, are categorized by subject matter areas such as education or demography and shared with the corresponding experts.

Nevertheless, manually going through millions of comments is time-consuming and unfeasible. Traditionally, analysts have analyzed comments using keyword-based methodologies, which are reliable enough to inform decision-making, but machine learning can be a more comprehensive, timely and accurate alternative. Given a data set of comments labeled with their respective census topics, machines can learn relationships between the text and its topic labels to classify comments in the new census.

To select the best model to classify comments from the 2021 Census of Population, the following

[1]This paper was a joint first-place prize winning submission to the 2023 IAOS Young Statistician Prize. The author is grateful for the support from Statistics Canada, especially Data Science and Innovation Division and Census Subject Matter Secretariat.

four text classification methods have been trained and evaluated on census comments: Support Vector Machines (SVM), Convolutional Neural Networks (CNN), semi-supervised Bidirectional Long Short-Term Memory Network (BiLSTM), and Cross-lingual Language Model-Robustly Optimized Bidirectional Encoder Representations from Transformers (XLM-R). The dataset contains arbitrary French and English text written by various people in Canada and covers a wide range of census topics. Some comments are labeled with multiple classes if respondents address multiple topics in a single comment. This research will explore machine learning algorithms of varying levels of complexity and analyze their results on this multi-lingual and multi-label supervised learning problem.

Following the research, the best-performing fine-tuned XLM-R model was used to efficiently classify 1.9 million comments received throughout the 2021 Census of Population cycle, further modernizing methods in Statistics Canada. Categorized comments were shared with the appropriate expert analysts to help with relevant and timely statistics.

The rest of this paper is organized as follows: Section 2 explains each classification model. Section 3 describes the classes and labeled data used for model training. Resulting findings and observations presented in section 4 show that the fine-tuned XLM-R model was superior overall and in different languages and classes. Section 5 describes how classified 2021 Census of Population respondent comments were used. The last section concludes the work and discusses potential work for the next census given the success of this work for 2021 Census of Population.

2. Methods

2.1. Support vector machines

The first model was built using Term Frequency-Inverse Document Frequency (TF-IDF) [6] and SVM [7, 8]. Before TF-IDF, comments were lowercased, lemmatized, and spell-checked so that variations of the same root word were interpreted as one. TF-IDF calculated the importance of each word to a comment in the dataset by calculating the number of times the word appears in a comment but offset by the word's frequency in the dataset. Cleaned comments were converted into a matrix of TF-IDF features for SVM to learn to distinguish vectorized representations of one class from the rest via a vector presentation. A hyperplane in the embedding space was optimally chosen to separate relevant comments in the vector space and to maximize their distances from nearby data points. Classes were predicted based on the partition of space formed by the set of hyperplanes.

Even though SVM was one of the most efficient machine learning algorithms [9], feature vectors fed into SVM did not preserve word orders. For example, if a comment spoke positively about the census overall but negatively about a topic not related to the census, TF-IDF would not link the emotion to their corresponding topic and misclassify the comment. Deep learning methods that overcome this problem were thus evaluated.

2.2. Convolutional neural networks

CNN was initially developed to classify images but have been applied for text classification [10]. The sequence of words in a comment first passed through an embedding layer where a lookup table mapped each word to a dense vector. A convolution filter was applied to each sliding window of words in the sequence to output a one-dimensional convolution that was made up of word embedding n-grams. Thus, words that might not have been descriptive alone were grouped with neighbouring words. Max pooling selected the highest value from the convolutional feature map to output the most prominent features which described the semantic meaning of the current phrase. The final sigmoid activation function predicted a class.

CNN identified phrases relevant to each class, but its architecture did not learn to ignore unnecessary words or to identify important words based on the comment's context. CNN model might classify "*I am happy to live in Canada*" into the *General comments – positive experience with the census* class even though the happiness is not attributed to the census. This would be problematic because any positive comments about a topic unrelated to the census should have been classified as *General comments – unrelated to the census*. BiLSTM architecture responded to the problem by capturing long-term dependencies between word sequences.

2.3. Semi-supervised bidirectional long short-term memory networks

Single-layer BiLSTM models employing both supervised and semi-supervised training approaches have shown to be competitive with their complex approaches [11]. Since Statistics Canada had historical

respondent comments from previous census cycles, and labeling was expensive, leveraging unlabeled data to train the model was favourable. Like CNN, the architecture's embedding layer represented text as a vector. BiLSTM encoders [12] captured long-term dependencies between word sequences and learned to retain or forget the previous state's information. Outputs from forward LSTM (processing words left to right) and backward LSTM (processing words right to left) were concatenated to concatenate past and future context in a hidden state. The max value was pooled from each hidden state to represent the most prominent and useful features from the input sentence. A linear layer translated features into classes, and the most likely classes were predicted using a softmax activation function.

Maximum likelihood estimation (MLE) and adversarial training [13] were used to train the model on the labeled data. MLE determined the parameters of the neural network to minimize loss between labeled data's estimated class probability and true label. The model used adversarial training furthermore to be robust to small perturbations to the input word embedding. This regularization method improved the quality of the word embedding.

Entropy minimization and virtual adversarial training used unlabeled data to make the model more generalizable. Minimum entropy regularization minimized the conditional entropy of the estimated class probabilities for an unlabeled observation [14]. Virtual adversarial training (VAT) added noise to the input word embedding of an input text and calculated a loss based on how similar the label distribution of the modified input with noise was to that of the original data point [15]. With VAT, the model produced more consistent predictions on small perturbations to the input.

Semi-supervised BiLSTM outperformed SVMs and CNNs, but it was worth finetuning transformers that offered additional advantages such as self-attention and pre-trained language models.

2.4. *Fine-tuned cross-lingual language model-robustly optimized bidirectional encoder representations from transformers*

Bidirectional Encoder Representations from Transformers (BERT) were successful on text classification tasks by transfer learning weights pretrained on various large datasets [16]. A model pretrained on a wealth of unclassified text and fine-tuned on a small dataset performed better than models trained on the same dataset but starting with a random initialization of weights. Also, compared to other neural network architectures such as CNN and BiLSTM, transformers excelled in representing dependencies between words in long texts. Its encoder architecture used self-attention [17] to learn the importance of each word in a sentence relative to the other words.

Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) [18] further improved BERT model by dynamically generating a different masking pattern when a sequence was fed into the model and increasing the training time, batch size and training data size. These modifications made the model learn more generalizable and robust word representations. Multilingual RoBERTa models soon followed. XLM-R [19] extended RoBERTa by being pretrained on 2.5Tb of text in 100 languages and effectively transferring what it learned from other languages.

3. Data

3.1. *Census respondent comments*

At the end of the Census of Population questionnaire, respondents were presented with a text box to share feedback on their experience completing the questionnaire. The instructions were phrased as follows:

Please use the space provided below if you have concerns, suggestions or comments to make about:

- the steps to follow or the content of this questionnaire (for example, a question that was difficult to understand or to answer)
- the characteristics of the questionnaire (for example, the design, the format, the size of the text).

Responses in this space were collected and analyzed to support decision-making regarding content determination for the next census and to monitor factors such as response burden.

3.2. *Comment classes*

To categorize the collected responses, seventeen classes were defined and listed in Table 1. Comments could be classified into one or more of the classes, except for the *General comments – unrelated to the census* class, which was exclusively for comments not falling into any other census-related class.

3.3. *Labeled data*

Of the 44,539 labels, most labels came from respondent comments from the 2019 Census Test, while a

Table 1
List of classes covering diverse aspects of the questionnaire and their proportions in the labeled data

Class name	Proportion of labels (%)
Activities of daily living	2.85
Coverage/Methodology	3.81
Demography	3.34
Education	2.15
Electronic questionnaire	11.83
Ethnocultural	6.20
General comments – positive experience with the census	8.29
General comments – unrelated to the census	6.99
Geography	2.12
Indigenous	2.34
Labour market activities, journey to work and place of work	8.23
Language of instruction	2.99
Languages	7.48
Payments and housing	10.16
Response burden	28.31
Sex at birth and gender	3.82
Veterans	0.71

small percentage came from the 2016 Census of Population and the 2021 Census of Population. The 2019 Census Test was conducted before the 2021 Census of Population on a small sample to evaluate the modified questionnaire and collection procedures. Even though the 2019 Census Test and the 2021 Census of Population covered the same topics, respondent comments about certain topics could have differed because of events in 2021 that did not impact Canadians in 2019; such as the COVID-19 pandemic. Some 2021 Census of Population respondent comments received early in the survey period have thus been labeled to help train the model.

Labels composed of 86.77% English comments and 13.23% French comments had a notable language imbalance. Since Statistics Canada received more comments from census respondents in English than in French, this imbalance was expected and representative of the Canadian population.

Analysts from Census Subject Matter Secretariat (CSMS), who have analyzed respondent comments from each census cycle, identified the classes and labeled the comments appropriately. Consequently, the labels were consistent and of high quality. The class distribution in the labeled data, shown in Table 1, was naturally imbalanced as it reflected the actual expected distribution.

4. Results

4.1. Experiment setup

To compare the performance of various models, the 44,539 labeled data entries were split into stratified random sets of mutually exclusive training data (80%), validation data (10%), and test data (10%) with equal proportions of English and French comments.

SVM and CNN models were trained on the training set. Semi-supervised BiLSTM was also trained on the training set but had 600,476 unlabeled comments from previous census cycles for unsupervised learning. XLM-R-Large model was fine-tuned on the training set. The validation set was used for each model's hyperparameter tuning, and each model's F1 score was reported based on the test set.

4.2. Overall performance

SVM, CNN, semi-supervised BiLSTM, and fine-tuned XLM-R were evaluated with an F1 score. F1 score measured the model's performance on each class as a harmonic mean of precision and recall, which are two important model evaluation metrics. Precision measured relevance of retrieved samples by calculating the ratio between the number of correctly classified samples and the number of samples predicted to the class. Recall measured the proportion of correctly retrieved relevant samples by calculating the fraction of correctly classified sample over samples that were actually relevant. F1 score was selected over accuracy because it measured the model's performance better on an imbalanced dataset. Overall F1 score was a weighted average of the F1 scores for each class, where the weight depended on the number of true labels for each class.

As shown in Table 2, the fine-tuned XLM-R model performed the best with an overall F1 score of 89.91% on the bilingual test data followed by the semi-

Table 2
F1 scores of SVM, CNN, semi-supervised BiLSTM and fine-tuned XLM-R models on bilingual comments as well as comments just in English or French

Model	Bilingual F1 (%)	English F1 (%)	French F1 (%)
SVM	81.30	82.45	72.47
CNN	81.43	82.21	75.79
Semi-supervised BiLSTM	85.11	85.94	79.36
Fine-tuned XLM-R	89.91	90.26	87.19

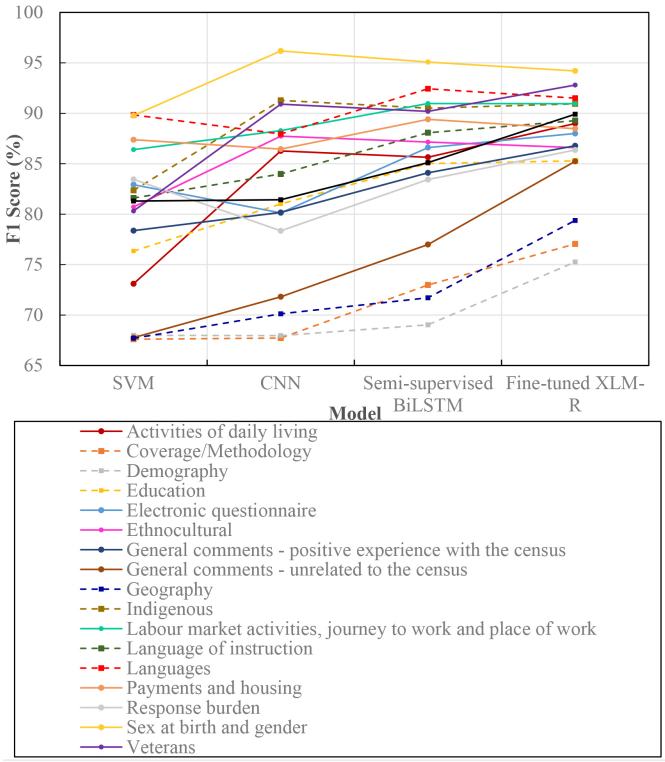

Fig. 1. F1 Score of SVM, CNN, semi-supervised BiLSTM and fine-tuned XLM-R models on all classes. Each line represents a class.

supervised BiLSTM model getting 85.11%. CNN and SVM models performed the worst, producing F1 score of 81.43% and 81.30% respectively.

Fine-tuned XLM-R, the best overall model, also outperformed other models on French test data, getting F1 score of 87.19%. It also yielded the smallest F1 score gap between the two languages (3.07%). Having a small difference in the F1 score between the two languages made the model dependable to classify 2021 Census of Population respondent comments well regardless of which official language they were written in.

4.3. *Performance of each class*

Fine-tuned XLM-R model performed well overall and the best in most classes. Figure 1 displayed each model's F1 score per class. Notably, the following two types of classes saw greater improvements using the fine-tuned XLM-R model: General census themed classes where the context in a sentence was important and *Demography*, *Coverage/Methodology*, and *Geography* classes that were closely related enough to be multi-classed in a single comment.

Fine-tuned XLM-R model excelled at correctly classifying classes that were not about a certain census content (i.e., *General comments – unrelated to the census, General comments – positive experience with the census* and *Response burden*). By capturing context and dependencies between words, fine-tuned XLM-R model classified comments to *Response burden* only when it was regarding the census; comments such as, *"Taxes should be made clearer"*, that expressed concerns outside of the census were correctly classified as *General comments – unrelated to the census*. As a

result, fine-tuned XLM-R model had the highest F1 score of 85.24% on *General comments – unrelated to the census* class compared to semi-supervised BiLSTM, CNN and SVM models that had F1 scores of 76.99%, 71.81% and 67.76% respectively on the same class. Fine-tuned XLM-R model also had the best performance on the other general classes, showing F1 score differences of 2.68%–8.41% on *General comments – positive experience with the census* class and 2.91%–8.01% on *Response burden* class. Great performance in these three classes influenced the overall performance because these three classes made up 43.59% of the labeled data.

Moreover, fine-tuned XLM-R model was the best at correctly differentiating *Demography, Coverage/ Methodology, and Geography* classes. Since the topics were similar and many comments covered more than one of these topics, these classes were often misclassified. Figure 1 showed that these three classes consistently performed the worst across different models compared to other classes, so improvements brought using fine-tuned XLM-R model were important. For example, the fine-tuned XLM-R model predicted "*My wife died on January 1 2018*." correctly as both *Demography* and *Coverage/Methodology,* but other models did not capture all the two classes in their predictions. Compared to other models trained from scratch, fine-tuned XLM-R model achieved better F1 scores on *Demography* (by 6.23–7.30%), *Coverage/Methodology* (by 4.07–9.45%), and *Geography* (by 7.67–11.69%).

5. Applying fine-tuned XLM-R model to classify 2021 Census of Population respondent comments during collection

Following the evaluation, the fine-tuned XLM-R model was used to classify respondent comments from the 2021 Census of Population. Comments about a specific questionnaire topic were sent to the corresponding subject matter experts to read and analyze. Categorizing these respondent comments helped triage the comments and see changes in the proportion of predicted classes over time.

Changes in respondents' experience with the census were first observed. Respondents who filled out the census early in May 2021 left more *General comments – positive experience with the census* compared to *Response burden* and *General comments – unrelated to the census.* As time passed by, the proportion of comments predicted to *General comments – positive experience with the census* decreased. Comments from people who did not respond to the census until August 2021 had more *General comments – unrelated to the census* than *Response burden,* and there were fewer *General comments – positive experience with the census* than *Response burden*. The comments about the *Electronic questionnaire* and other subject matter area topics maintained a stable distribution over time.

Class predictions also helped compare long-form census comments to that from the short-form. Comments from the long form had 18.0% more comments predicted to *Response burden* and 15.5% fewer comments related to *Positive experience with the census* compared to that from the short form. *Payment and housing* and *Labour market activities, journey to work and place of work* topics that were only asked in the long-form census questionnaire were naturally discussed more frequently in the comments from the long form than in the short form by 13.1% and 7.6% respectively. These topics were still asked about and voluntarily described in comments from the short form, but these topics made up less than 1% of the comments from the short form.

6. Conclusion

In this study, we selected, trained and evaluated four different models for the classification of bilingual respondent comments from the Canadian Census of Population. Fine-tuned XLM-R model performed best overall, in addition to achieving a strong performance across different classes and languages despite the imbalances in the training dataset. These successes led Statistics Canada to apply this model to automatically classify comments collected during the 2021 Census of Population. This new approach alleviated the manual workload of triaging comments and relayed comments to the appropriate subject matter experts in a timely manner.

Given the success of this work, Statistics Canada will continue to use machine learning to classify census respondent comments. In preparation for the next Census of Population in 2026, more cross-lingual transformers models could be explored. If unlabeled text from thousands of bilingual Statistics Canada products could be used on top of a pretrained model, the fine-tuned model may classify Canadian census comments better.

Acknowledgments

The author thanks Shirin Roshanafshar, Alexandre Istrate, and Najeeb Qazi for their contributions to the

Census of Population respondent comments classification project. She also thanks Nicholas Denis, Kimberley Flak, Geneviève Tilden, Tania Hinchcliff, Anne-Marie Rollin, and Claudiu Motoc for their helpful comments and suggestions.

References

[1] Statistics Canada. Census of Population. [Internet]. Ottawa (CA): Statistics Canada; 2022. Available from: HYPERLINK https://www12.statcan.gc.ca/census-recensement/index-eng.cfm. [Accessed 30 March 2022].

[2] Bollman RD. An overview of rural and small town Canada. Canadian Journal of Agricultural Economics/Revue Canadienne D'agroeconomie. 1991; 805-817.

[3] Steffler J. The indigenous data landscape in Canada: An overview. Aboriginal Policy Studies. 2016.

[4] Statistics Canada. 2021 Census of Population – Analytical products. [Internet]. Ottawa (CA): Statistics Canada; 2022. Available from: HYPERLINK https//www12.statcan.gc.ca/census-recensement/2021/as-sa/index-eng.cfm. [Accessed 30 March 2022].

[5] Statistics Canada. 2021 Census of Population Consultation Results: What we heard from Canadians. [Internet]. Ottawa (CA) 2019. Available from: HYPERLINK https://www12.statcan.gc.ca/census-recensement/2021/consultation/92-137-x/92-137-x2019001-eng.cfm. [Accessed 30 March 2022].

[6] Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 1972.

[7] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In European Conference on Machine Learning. Springer. 1998; 137-142.

[8] Manevitz LM, Yousef M. One-class SVMs for document classification. Journal of Machine Learning Research. 2001; 139-154.

[9] Karamizadeh S, Abdullah SM, Halimi M, Shayan J, Javad Rajabi M. Advantage and drawback of support vector machine functionality. In 2014 International Conference on Computer, Communications, and Control Technology (I4CT). IEEE. 2014; 63-65.

[10] Kim Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014; Conference on Empirical Methods in Natural Language Processing (EMNLP); Doha, Qatar: Association for Computational Linguistics. 1746-1751.

[11] Sachan DS, Zaheer M, Salakhutdinov R. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In Proceedings of the AAAI Conference on Artificial Intelligence. 2019; 6940-6948.

[12] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997; 1735-1780.

[13] Miyato T, Dai AM, Goodfellow I. Adversarial Training Methods for Semi-Supervised Text Classification. In Proceedings of the 5th; International Conference on Learning Representations. 2017.

[14] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. Advances in Neural Information Processing Systems. 2004.

[15] Miyato T, Maeda Si, Koyama M, Ishii S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018; 1979-1993.

[16] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Minneapolis, Minnesota: Association for Computational Linguistics. 4171-4186.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017.

[18] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019.

[19] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th; Annual Meeting of the Association for Computational Linguistics. 2020; 8440-8451.