Multilingual Sentiment Analysis for Investigating Perceptions of Globalization

Anagha Ani and Erik Cambria
College of Computing & Data Science
Nanyang Technological University
Singapore
AN0003NI@e.ntu.edu.sg
cambria@ntu.edu.sg

***Abstract*—This paper aims to identify trends in perceptions on the topic of globalization by performing sentiment analysis on text data collected from social media posts. A novel methodology of extracting culture-informed sentiments is proposed and tested on a corpus containing text posts from the social media site Reddit in two languages: French and English. To do so, a graph convolutional network is used to train a polarity classification model by extracting commonsense culture-specific knowledge using the SenticNet knowledge base. A variety of sentiment analysis tasks including polarity classification, intensity ranking, toxicity spotting, wellbeing assessment, and personality classification are performed using Sentic APIs on two extracted subsets of the corpus and the resulting trends in the data are identified and analyzed.**

Keywords— Multilingual Sentiment Analysis, Polarity Classification, Graph Convolutional Networks, Culture-Specific Knowledge Extraction, Globalization

I. Introduction

A. Background

Globalization has long been a hotly debated topic, with far-reaching implications on all aspects of life, from culture to the economy to political positioning [1]. As a result, understanding the perception of globalization is of utmost importance in making informed decisions in various domains including electoral campaigning and policy making to resolve issues such as social inequality [2]. It is particularly interesting to understand differing perceptions and attempt to make causal connections to intelligently inform decision making processes. This leads to the motivation behind this study: understanding the impact of culture on the perception of globalization. The relationship between language and culture has been studied extensively, with compelling evidence backing up the strong relationship between the two and the impact of each on the other [3]. This relationship can be exploited to understand the cultural impact on perceptions by analyzing data in different languages.

Social media platforms such as Reddit and Quora are frequently used to express thoughts and beliefs on various topics, and enable users to act under anonymity if they so wish, increasing candor and revealing true opinions. It is extremely useful to analyze data collected from such platforms as they reflect the opinions, particularly on polarizing topics, that may be lurking beneath the surface. Natural Language Processing (NLP) is a field of Artificial Intelligence which allows systems to intelligently process human language to make useful inferences. As the data produced on these platforms expands, NLP for sentiment analysis stands out as a useful tool in improving our understanding of the sentiments and beliefs upheld by our societies.

B. Objectives and Scope

The objective of this study is to develop and analyze a corpus containing comments and opinions on the topic of globalization posted to social media sites in two languages: French and English. The aim was achieved in two broad phases. The first phase involves creation of a training corpus upon which the sentiment analysis was performed. Data representing public opinion on the topic of globalization was collected from Reddit. The training corpus can be split broadly into four sections: firstly, text data originally posted in the French language. Secondly, the same text data which is translated into English using machine translation tools. Thirdly, text data originally posted in the English language. The fourth consists of this data translated to French. The Sentic Graph Convolutional Network (GCN) model was trained separately on the two non-translated datasets to extract the affective commonsense knowledge that is unique to each of the two languages. The two trained models were then run on the two translated datasets.

In the second phase, the entries in each of the datasets that were misclassified by the two trained GCN models were extracted. Due to the syntactic similarities between French and English, it is asserted that the misclassification arises as a result of differences and ambiguities in the extracted commonsense knowledge between the two languages; hence, these extracted misclassified datasets represent the text that is heavily influenced by the cultural commonsense knowledge native to the origin language. Sentic Application Programming Interfaces (APIs) were then applied to these extracted datasets to analyze the sentiments on five axes: polarity classification, intensity ranking, toxicity spotting, wellbeing assessment, and personality detection. This allowed identification of trends within the misclassified data which were used to isolate the cultural impact on perception of globalization. Comparison of the trends in the five sentiment analysis tasks between the two culture-specific datasets allowed the identification of differences between the cultural impact of the two languages.

C. Contributions

This paper aims to investigate the application of NLP techniques, namely GCNs for commonsense knowledge extraction, to the problem of multilingual sentiment analysis. There exists a gap in current literature in the investigation of the isolated impact of culture on the commonsense knowledge extracted through this method, which this paper aims to address. The methodology proposed in this paper involves isolating culturally-influenced text data through extraction of misclassified data by applying a model trained to extract commonsense knowledge in one language to data translated from another language.

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00059

This methodology has not been explored in existing literature, this paper thus contributes in a novel way by investigating its utility. Theoretical insights are achieved through the application of this methodology to identify differences in analytic capability of GCN models in the native language based on differences in commonsense knowledge informed by culture. Finally, the practical contributions of this paper are realized through the analysis of trends in social media posts on the topic of globalization. Overall trends as well as culture-specific trends are discussed in this paper. The identified trends can have implications in domains including policy making, electoral campaigning, etc.

II. LITERATURE REVIEW

A. Review of NLP Techniques

1) Approaches to Sentiment Analysis

Sentiment analysis is a broad field of study that encompasses a range of smaller tasks including concept parsing, subjectivity detection, polarity classification, intensity ranking, emotion recognition, and many more. The sentiment analysis models that perform these tasks allow us to extract useful information from large amounts of data, including user opinions, interests, personalities, and attitudes. There are three major approaches to sentiment analysis model development in existing literature: learning-based methods, lexicon-based methods, and hybrid methods [4]. Deep learning has grown to be the most popular learning-based technique applied to sentiment analysis problems, achieving the highest prediction accuracies [5]. Hybrid machine learning and deep learning models combine the capabilities of dynamically-trained models with the predefined lexical and syntactic rules that make up languages.

2) Techniques for Multilingual Sentiment Analysis

Machine learning and deep learning-based techniques perform best with the availability of a large amount of data. The available data on the internet, which is the most abundant source of data today, is predominantly in the English language. Thus the majority of developments in the field of sentiment analysis have focused on English language texts due to its high resource availability. In contrast, low resource languages suffer in terms of the accuracy of sentiment analysis predictions. As a result, several methods have been developed to handle multiple languages in the sentiment analysis context.

a) Multilingual NLP Approaches

Language-specific approaches to non-English language sentiment analysis involve the development of models directly in the target languages [6]. However, the lack of resources for most languages yields unsatisfactory results in predictive or analytic capability. One solution to the resource problem is translating the English language Sentiment Analysis tools into the target language before performing the training with text in the target language. One study [7] investigated this solution and discovered the failure to retain subjectivity across languages, attributed to the ambiguity in the languages.

b) Cross Lingual NLP Approaches

The cross lingual approach involves the translation of data in low-resourced languages to high-resourced languages like English to directly make use of the available resources and tools. Following the translation, well-performing sentiment classification models trained on large amounts of data can be applied to the dataset. The most widely used approach to translation is Machine Translation (MT), which is the automated translation of text-to-text using popular tools such as Google Translate, Bing, Yandex, and DeepL. One study [8] established a significant difference between machine translated and manually translated text from high-to-low-resource languages, and the impact of machine translation on subsequent polarity classification skewed negatively, attributed to factors including slang, sarcasm, ambiguous words, idioms, and negation.

B. SenticNet

1) Motivation: Domain Specificity

The efficacy of Sentiment Analysis models, and Natural Language Processing in general, has been observed to be highly domain-dependent [9], with models performing well in one domain having limited performance quality in others. This feature of traditional NLP models calls into question their utility in real world scenarios, which often demand reusability and applicability to various differing contexts in which statistical models may be rendered useless. These techniques often rely on sentiments or polarity that are attached to affect keywords in sentiment lexicons. The larger the text input to such a model, the more accurate its analysis of sentiment, resulting from a higher number of affect keywords from which to derive insights on polarity. This results in poor performance on smaller units of text such as independent sentences. These problems with traditional sentiment models motivated the development of SenticNet, a commonsense NLP framework.

2) Sentic Computing

The Sentic Computing framework was developed with the goal of incorporating a number of domains that influence language evolution and interpretation, most importantly, common sense reasoning, into traditional statistical NLP and Sentiment Analysis methodology [10]. The proposed methodology is to incorporate into the Sentiment Analysis tasks, both symbolic models as well as sub-symbolic paradigms to effectively infer from statistical patterns while also making use of conceptual understanding derived from theoretical insights. The idea is to more closely replicate the human ability to express meaning through language in a way that incorporates contextual information without explicitly specifying it. This methodology has been effectively applied to a range of sentiment analysis tasks, including subjectivity detection, polarity classification, intensity ranking, emotion recognition, etc.

3) SenticNet in the Multilingual Context

SenticNet as a commonsense reasoning framework and sentiment analysis lexicon was originally created for opinion mining in English, and there have been efforts to extend the utility of the framework to other languages. One paper [11] proposed the methodology of using translation engines and dictionaries in combination with a concept disambiguation algorithm and sentiment conflict detection algorithm to overcome the challenge posed by lack of resources for non-English languages. Dependency parsing, a key aspect of natural language analysis that involves extracting the information contained in the patterning and positioning of tokens in multi-word expressions, is another area that has been proved to benefit from semantic understanding in multilingual contexts. A study [12] exploring the integration of graph embeddings into neural dependency parsing models found significant improvement in performance with the integration of learnings from dependency graphs. This finding emphasizes the importance of contextual clues in semantic and syntactic multilingual tasks.

4) Sentic GCN

A subset of sentiment analysis is aspect-level sentiment analysis, which is the fine-grained task of identifying various aspects within the text and their related individual sentiments. This task is particularly well addressed by GCNs. Traditional GCNs for aspect-based sentiment analysis considers the dependencies between text elements encoded as a graph [13-14] to extract information that informs sentiments. The syntactic dependencies can better extract contextual clues rather than relying emphatically on affective keywords. However, this approach fails to incorporate the key commonsense knowledge that informs meaning which is not explicitly contained within the text. This leads to the purpose of Sentic GCN [15]: integrating commonsense knowledge into the framework of GCNs to take advantage of both syntactic dependencies and the underlying knowledge that encodes meaningful information with regards to each aspect being analyzed. This is reflected by a current trend in AI that aims to use external knowledge to ground meaning into semiotic representations [16]. This is done through a hybrid approach to AI, also known as neurosymbolic AI, which Sentic GCN belongs to.

In particular, the proposed methodology of Sentic GCN to inform the constructed graph of the related commonsense knowledge is as follows: to capture syntactic dependencies, the graph is constructed as normal from the dependency tree; following this, the commonsense knowledge is integrated into the adjacency matrix using SenticNet to calculate an affective score to derive an enhanced aspect-specific dependency graph. On this enhanced graph, the GCN is applied to obtain the final aspect-based sentiment analysis results. The resulting model was found to outperform all investigated comparison models, including graph networks, BERT-based models, as well as deep neural networks, tested on four datasets. The results of the study effectively established the utility of commonsense knowledge integrated into syntactic inferences for aspect-based fine grained sentiment analysis.

III. METHODOLOGY

The overarching idea is to attempt to isolate the impact of culture on the perception of globalization to identify differences in trends through analysis of text data in English and French. This was achieved through the following three-stage methodology.

In the first stage, text data was collected through Reddit posts originally written in English and French. In this stage, the translated counterparts of each of these datasets were also be obtained and each of the datasets were labeled.

In the second stage, each dataset was used to separately train Sentic GCN to obtain two predictive models: one trained in English, the other trained in French. In this stage, the next step is to run the models on the translated datasets: the model trained in English was run on the dataset of posts originally in French which was translated to English, and vice versa. The next step is to identify and extract all the entries that are misclassified by the GCN. Because the GCN extracts affective commonsense knowledge from the training data which it uses to make predictions on the test data, the misclassified data formed the datasets that represent the statements that are heavily informed by those aspects of culture unique to the origin language.

The final stage involved using Sentic APIs to perform five key sentiment analysis tasks on these datasets: polarity classification, intensity ranking, toxicity spotting, wellbeing assessment, and personality detection. This allowed identification of the trends linked solely to the culturally unique text. The trends in polarity classification were compared to those in the original datasets containing the entries that do not involve the ambiguity introduced by cultural differences to isolate and identify the extent to which the analyzed perceptions are influenced by culture. Additionally, comparison between the five sentiment analysis task trends of the two culturally-influenced datasets revealed the differing perceptions as informed by culture.

A. Data Collection

In order to analyze the trends in perception on the topic of globalization, it was deemed necessary to access data from social media sites to get candid viewpoints undeterred by social pressures which are encouraged on these sites through the apparent anonymity that users can easily achieve.

The data was obtained from Reddit using a Python library named Python Reddit API Wrapper (PRAW) that allows for access to Reddit's API. The API allows access to subreddits, retrieval of posts and comments, and executing search operations.

1) Creating the French Dataset

In order to retrieve posts and comments relevant to the topic of globalization originally posted by users in French, the four keywords of 'mondialisation', 'globalisation', 'internationalisation', and 'universalisation', were used in the search query, all of which are synonyms in the French language for the English word globalization. The list of all relevant French subreddits was obtained from the subreddit r/annuaire that hosts the directory of all French-speaking subreddits. In order to ensure that the data is representative of the perceptions of French speakers as a whole and is not biased by the viewpoints of people more likely to speak on a particular topic, it was ensured that the subreddits comprised a range of diverse topics, from entertainment to politics. The final list of active subreddits consisted of 783 entries and the inactive subreddit list had 869 entries.

Subsequently, the list of subreddits were looped through, and PRAW was used to search for the keywords in each post, including the title and post text, as well as each comment. The resulting three text types: post titles, post texts, and comments were stored to a .csv file. This was treated as the dataset for text originally posted in French, and it contained a total of 626 entries.

2) Creating the English Dataset

To create a comprehensive dataset comprising diverse opinions on the topic of globalization, a comprehensive list of existing subreddits was retrieved from the subreddit r/listofsubreddits, and the 'general' list was selected, to ensure retrieval from those expressing opinions on a diverse range of topics, similar to the French dataset.

Similar to the creation of the French dataset, the list of subreddits were looped through to search for text which contains the word 'globalisation' or 'globalization'. The obtained dataset had a total of 2770 entries.

3) Translations

In order to carry out an investigation on the perceptions on a multilingual paradigm, it was required that the two final collated datasets (one in French, the other in English), each be translated to the other language. To do this, machine translation was employed. Machine translation is the procedure of automating translation activities rather than manually translating each entry. The Microsoft Azure AI Translator service was chosen for this purpose due to its speed, security, and reliability. The Azure Translator was called through the REST API, specifying the API endpoint, location, and key. Similarly, two further collated datasets were obtained: English data translated to French (2770 entries), and French data translated to English (470 entries).

B. Data Labeling

Data labeling was carried out using Sentic APIs. The SenticNet polarity classification algorithm can return two possible labels: Positive and Negative. The intensity ranking algorithm returns a value between 0 and 100 indicating the intensity of the emotion associated with the text being analyzed. In order to conform to the system of Sentic GCN polarity classification algorithm, both these algorithms of the Sentic APIs were used in conjunction to obtain labels, with the label 'Neutral' assigned to entries with intensity below the threshold of 25.

C. Training Sentic GCN

1) Data Preparation

Each of the datasets were split into a train and test set, containing 75% and 25% of the entries respectively. Subsequently, for each of the train and test sets, text files were created containing the data in the format: [sentence] \n [aspect] \n [true polarity label]. Due to the nature of this project, the task of aspect extraction was unnecessary, removing the uncertainty of inaccuracy in this dimension. Instead, the aspect of each sentence was automatically determined to be the word that was searched to obtain the data point (i.e. 'globalization' and its variants including synonyms and translations). In order to conform to the formatting accepted by the Sentic GCN model, the aspect word was extracted and replaced with the special sequence of characters used to identify the aspect position, around which the sentence is broken up for graph construction.

2) Graph Generation

Prior to beginning model training, three different graphs are generated for the dataset. The first is an ordinary dependency graph which captures the syntactical information of the given sentence. The second is a sentic graph, in which the SenticNet knowledge base containing a dictionary of words mapped to their corresponding affective score, is used in the creation of the sentence graph, effectively encoding the affective commonsense knowledge that each word lends to the sentence. The edges are weighted by the cumulative affective score corresponding to each token in the sentence. The third graph, which combines the SenticNet affective knowledge as well as the syntactic information contained in the sentence, encodes both the dependencies between the tokens as well as the relative importances of the tokens.

3) Hyperparameter Tuning

The Sentic GCN model takes in a number of arguments that can be specified while running the training command. The list of arguments and their corresponding details are consolidated in the table below.

TABLE I. HYPERPARAMETER DESCRIPTIONS

Argument	**Purpose**	**Data Type**	**Possible Values**	**Tested Values**
Optimizer	Function used to adjust the weights learned by the model during training	String	adadelta, adagrad, adam, adamax, asgd, sgd, rmsprop	adadelta, adagrad, adam, adamax, asgd, sgd, rmsprop
Initializer	Function used to initialize weights prior to training	String	xavier uniform, xavier normal, orthogonal	xavier uniform, xavier normal, orthogonal
Learning Rate	Specifies the size of steps taken during optimization. Increased learning rate increases speed and convergence of training	Float	(0, 1)	0.001, 0.01, 0.1, 0.002
L2Reg	Used to avoid overfitting by specifying the amount of regularization applied to the weights	Float	(0, 1]	0.00001
Num Epoch	Determines the number of times the dataset is passed through the neural net during training	Integer	(0, ∞)	50
Batch Size	Determines the number of data samples processed together in a training iteration	Integer	[1, {total no. of samples}]	16, 32, 64
Log Step	Defines the frequency at which the model is evaluated for test accuracy and f1. Lower log step = higher frequency	Integer	[1, {batch size * num epoch}]	5
Embed Dimensions	Specifies the size of the embedding vectors	Integer	[2, ∞)	300
Hidden Dimensions	Specifies size of the hidden layers in the neural net	Integer	[2, ∞)	100, 200, 300

Polarities Dimensions	Specifies the number of unique values the polarity classification output can possibly have	Integer	[2, 3]	3
Seed	Defines the random seed for initializing parameters and operations to ensure reproducibility	Integer	[0, ∞)	600, 776, 800, 900
Device	Specifies the hardware (CPU or GPU) that performs the computations	String	Cuda, cpu	Cuda, cpu

The number of epochs was fixed at 50 as it was observed that all the models converge to a maximum test accuracy and f1 average well before the completion of 50 epochs. In order to obtain the best possible model with the highest predictive capability, all the different combinations of hyperparameters were tested using a script that automated the hyperparameter specification in the train command. The list of combinations of values under the column Tested Values in Table 1 was generated and the models were trained using each combination. The corresponding maximum average test f1 and accuracy were recorded as evaluation metrics.

D. Extraction of Culture-Specific Datasets

The key hypothesis of this paper is that a highly accurate GCN trained and tested on the French language, when classifying a new dataset of English posts translated to French, will misclassify those data points that are heavily informed by the culture specific to the speakers of the English language. These misclassified data points, when extracted, can be used to decipher the sentiments informed by the culture that is exclusive to English language speakers.

Following this hypothesis, the extraction of the culture-specific sub-dataset in English is conducted by performing the polarity classification task on the English to French translated dataset using the model trained in French. The predictions of the model are compared against the true labels generated by the language-specific Sentic APIs. The misclassified data points are extracted and the original untranslated English text forms this culture-specific subdataset. In the same manner, the French culture-specific sub-dataset is extracted using the Sentic GCN trained and validated on the English dataset.

E. Trend Analysis on Culture-Specific Datasets

The extracted culture-specific datasets were analyzed to identify predominant trends in the perception of globalization of English and French speakers that are influenced by culture. To understand the statistics of the extracted datasets in comparison to the correctly classified datasets (unbiased by cultural influences), the true labels of the two subsets: Culturally Influenced and Culturally Common data was plotted to compare their relative class distributions for each of the English and French datasets.

In addition, four other sentiment analysis tasks were performed on the culture-specific datasets using the Sentic API. The overall trends in the results of these sentiment analysis tasks on the two culturally-biased subsets are compared against each other.

IV. Results

A. English Dataset

1) French Model Variation Results

A total of 3024 variants of the Sentic GCN model were trained and their corresponding Maximum Average Test F1 and Accuracy scores were recorded. The Sentic GCN model that achieved the highest maximum average test f1 score had the parameters specified in Table II.

TABLE II. Optimal Hyperparameter Values- French Model

Hyperparameter	Value	Hyperparameter	Value
optimizer	adam	batch_size	32
initializer	xavier_normal	log_step	5
learning_rate	0.002	embed_dim	300
l2reg	0.00001	hidden_dim	200
num_epoch	50	polarities_dim	3
		seed	600

The results of the evaluation metrics of the two best models are detailed in Table III.

TABLE III. Best Achieved Evaluation Metrics- French Model

Evaluation Metric	Value
max_avg_test_acc	0.66242
max_avg_test_f1	0.664393

2) Trends on Influence of Culture: English Speakers

After training the French model in such a way to maximize the test f1 score, the dataset of entries collected in English and translated to French was run through the French model to make predictions on the polarity of the text.

Subsequently, the predicted labels and true labels were compared to extract the misclassified subset of the original English dataset. This is the subset that represents the culturally-influenced data. The subset with matching true and predicted labels (i.e., correctly classified) represents the data that is not influenced by cultural differences. Fig. 1 depicts the frequencies of the true polarities of the culturally influenced and culturally non-influenced datasets.

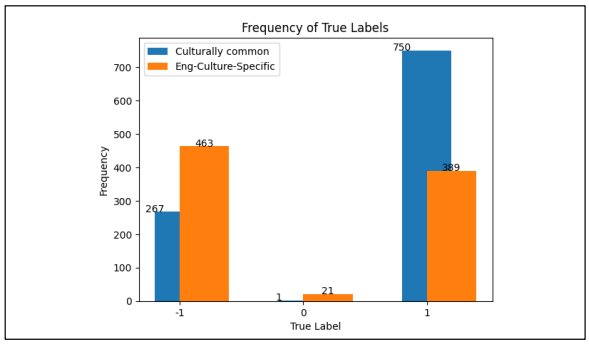

Fig. 1. Polarity comparison plot- English-Culture-Specific vs English Non-Culture-Specific

The subset of data points that are influenced heavily by those aspects of culture or commonsense knowledge that are not present in the French data and is specific to English speakers, are skewed more negatively in terms of polarity, whereas those data points that are culturally common to English and French are skewed more positively.

B. French Dataset

1) English Model Variation Results

The results of the hyperparameter variations for the English model were as follows. A total of 547 models were trained and the impact of the variations on the maximum test accuracy and f1 were recorded. Due to the large size of the train dataset as compared to the French model, the training time for the English model was significantly higher (~5 times slower); as a result, the full grid of hyperparameter combinations that were generated and trained for the French model could not be trained for the English model. Most notably, only two of the optimizers: adadelta and adagrad, were tested.

The model with the highest F1 score had the parameters detailed in Table IV.

TABLE IV. OPTIMAL HYPERPARAMETER VALUES- ENGLISH MODEL

Hyperparameter	Value	Hyperparameter	Value
optimizer	adagrad	batch_size	16
initializer	xavier_uniform	log_step	5
learning_rate	0.1	embed_dim	300
l2reg	0.00001	hidden_dim	300
num_epoch	50	polarities_dim	3
		seed	800

The two best performing models had the maximum average test F1 and accuracy scores detailed in Table V.

TABLE V. BEST ACHIEVED EVALUATION METRICS- ENGLISH MODEL

Evaluation Metric	Value
max_avg_test_acc	0.744589
max_avg_test_f1	0.663294

2) Trends on Influence of Culture: French Speakers

Following the training of the English model, it was applied on the dataset of French text posts translated to English. The subset of misclassified entries representing the French-culture-specific data points were extracted by comparing the true and predicted labels. The class distribution of the true polarity for the misclassified and correctly classified entries were as depicted in Fig. 2.

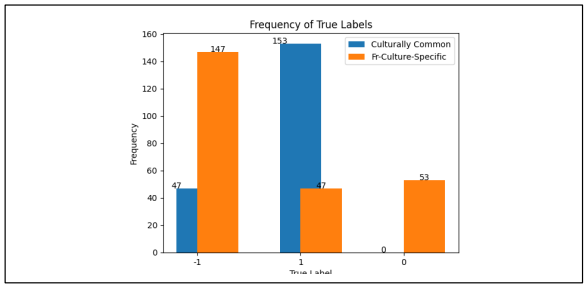

Fig. 2: Polarity comparison plot- French-Culture-Specific vs French Non-Culture-Specific

As can be observed from Fig. 2, the French culture-specific data is skewed far more negatively as compared to the subset of the data that is common to both cultures. The subset of data common to both cultures is skewed positively. This is in accordance with the findings from the analysis of the frequency of true labels in the English culture specific dataset which found that the subset of data that is correctly classified by the French model (i.e., that which is common to both cultures), is skewed positively.

C. Comparison of Culture-Specific Datasets

The two extracted datasets were then fed into Sentic APIs to generate the labels for the Sentiment analysis tasks. The results of the generated labels and the trends in these results will be described and analyzed in this section.

1) Intensity Ranking

Sentic API intensity ranking algorithm [17] returns an intensity value associated with the input text which will be within the range 0 to 100. The trends from Fig. 3 show that both the culturally-exclusive datasets contain text posts that express high intensity of emotions.

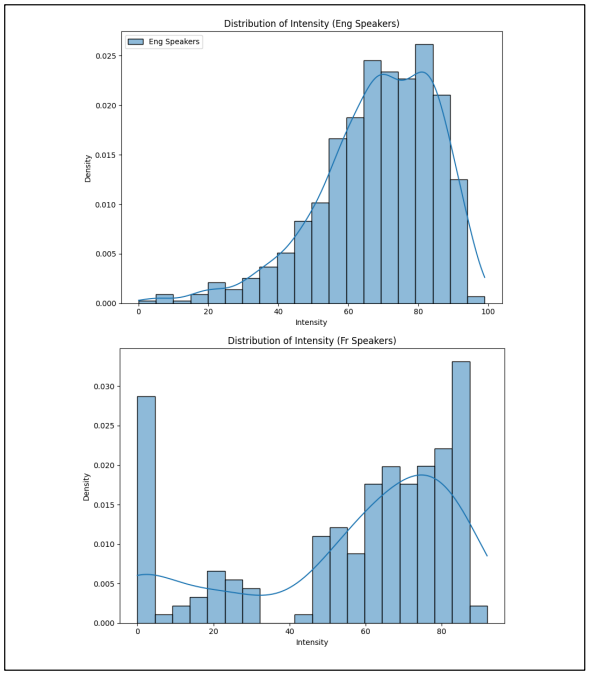

Fig. 3: Intensity Ranking Distributions of Culturally-Influenced Datasets

The distributions of the two datasets show that the English Culturally influenced dataset has higher intensity on average, with the majority of the data points concentrated at high intensity of around 80. In contrast, the French culturally influenced dataset has relatively higher spread, with a peak in density at intensity of 0 (representing neutral sentiment).

2) Polarity Classification

In order to compare the polarities across the two datasets, the counts of each of the labels were normalized by dividing by the size of the corresponding dataset, effectively representing the percentage of each of the datasets that were classified under each label.

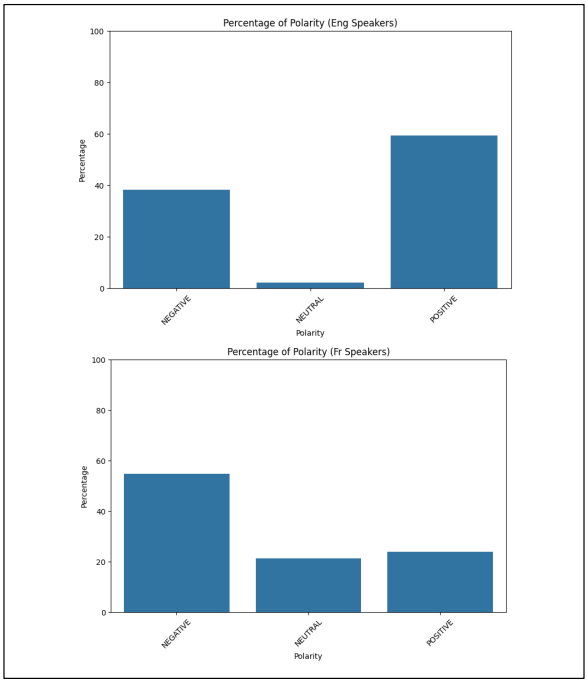

Fig. 4: Polarity Classification Distributions of Culturally-Influenced Datasets

Polarities of the English culture dataset skew towards the positive, whereas the polarities of the French culture dataset skew towards the negative, with equal distribution over neutral and positive sentiments. In contrast, there is low representation of the neutral sentiment in the English culture dataset.

3) Toxicity Spotting

The toxicity spotting algorithm analyzes the input text to determine how it reads in terms of harmful, hurtful, or unpleasant qualities [18]. It returns a value between 0 and 100, encoding no toxicity to maximum toxicity. The counts of each of the values for the two datasets were normalized by dividing by the size of the corresponding dataset to allow for better comparison between the two datasets. Comparing the trends in Fig. 5, it can be observed that the English-culture-specific dataset shows a relatively lower level of toxicity on average. The data that is exclusive to both cultures contains text expressing sentiments that are by and large not harmful.

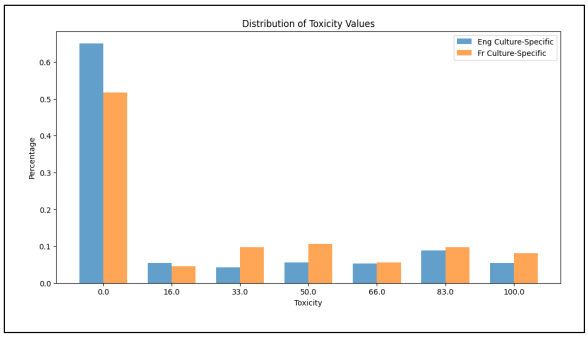

Fig. 5. Comparative Trends in Toxicity Spotting between Culturally-Influenced Datasets

4) Wellbeing Assessment

The wellbeing assessment API performs an evaluation of the author of the input text to determine their overall mental wellbeing, an evaluation that considers concepts including mental health, self-evaluation on a cognitive level, and aggregate emotions [19]. The output is a wellbeing score between -100 and 100, encoding the state of the author from high stress to high wellbeing. The results are shown in Fig. 6.

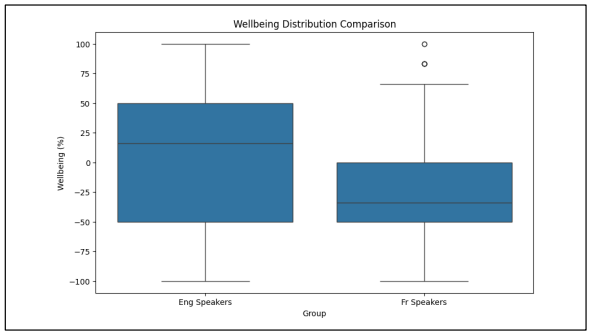

Fig. 6. Wellbeing Score Distribution Comparison between Culturally-Influenced Datasets

The dataset of French-culture-specific data points shows lower wellbeing, with a negative median value representing higher stress levels in these authors. In contrast, the dataset of English-culture-specific data shows comparatively higher wellbeing. A significant number of data points amongst the English culture-specific dataset have authors with positive wellbeing. Only a few outliers in the French dataset correspond to high wellbeing.

5) Personality Detection

The personality detection task involved generating the personality classes for each of the data points using the Sentic API [20—25]. The percentage of each of the datasets that fell into each category of personality was plotted for the two datasets to allow for comparison.

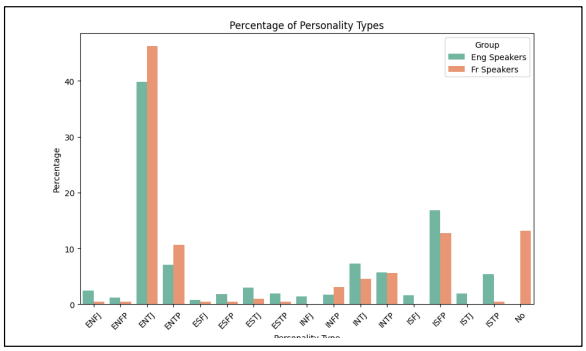

Fig. 7: Personality Type Classification of Culturally-Influenced Datasets

As can be observed from Fig. 7, the vast majority of entries in both datasets fall into the class ‘ENTJ’, which is the personality type characterized by assertiveness and confidence. This points to a disproportionately high representation of this personality type in the datasets, the reason for which may be interpreted as attributing to the composition of the datasets of text posts from the internet, particularly on a topic that is polarizing and may motivate political discussions.

V. Conclusion

To conclude, a novel methodology of extraction of culture-specific text data in English and French was proposed and implemented. This methodology involved making use of an affective commonsense knowledge extraction framework by training in a target language and testing the source text after conversion through machine translation into the target language. The misclassifications by the target language model on the translated text were hypothesized to be attributed to cultural commonsense knowledge differences due to the syntactic similarities between the two tested languages.

The analysis of the trends comparing the culturally-influenced and non-culturally-influenced datasets of text posts originally in English revealed that those data points that expressed sentiments in a manner that is highly specific to the culture of English speakers were mostly negative in true polarity. The same trend was observed in the subset of the original French dataset representing the points heavily influenced by French culture. In both culturally-common datasets, the skew was in the direction of positive polarity. The most notable of the sentiment analysis results on the two culturally-influenced datasets revealed lower average intensity in the French-culture dataset, along with lower wellbeing, i.e., higher stress.

Future work on the topic could include applying the methodology on a larger dataset of text posts which would ideally result in improvement of the accuracy of the two trained models. The size of the French dataset undoubtedly contributed towards the unsatisfactory performance of the trained French model. Sourcing a larger dataset in the French language would be beneficial to obtain a model with higher predictive capability. Additionally, efforts to manually annotate the datasets for polarity would serve to solidify the legitimacy of the findings.

Acknowledgment

This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005) and by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore.

References

[1] Debating globalization: Perceptions of the phenomenon based on political positioning and on ideological understandings of economy, culture, and the nation‐state−Griva−2015−European Journal of Social Psychology−Wiley Online Library. (n.d.). Retrieved from onlinelibrary.wiley.com.remotexs.ntu.edu.sg/doi/10.1002/ejsp.2163

[2] Globalization and perceptions of social—ProQuest. (n.d.). Retrieved from proquest.com.remotexs.ntu.edu.sg/docview/224005594

[3] Millrood, R. P. (2013). Language as a Symbol of Culture. YAZYK I KULTURA-LANGUAGE AND CULTURE, 22, 43–60.

[4] Amraouy, M., Himmi, M. M., Bellafkih, M., Talaghzi, J., & Bennane, A. (2023). Sentiment analysis in digital learning: Comparing Lexical, Traditional machine learning, and deep learning approaches. 14th International Conference on Intelligent Systems.

[5] Jain, K., & Kaushal, S. (2018). A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis. International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), 483–487.

[6] Araújo, M., Pereira, A., & Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences, 512, 1078–1102. https://doi.org/10.1016/j.ins.2019.10.031

[7] Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. In A. Zaenen & A. van den Bosch (Eds.), Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 976–983 https://aclanthology.org/P07-1123

[8] Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Abdullah, Batra, R., & Wani, M. A. (2021). The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing. IEEE Access, 9, 124478–124490. https://doi.org/10.1109/ACCESS.2021.3110285

[9] Turney, P. (2002). Thumbs Up or Thumbs Down? {S}emantic Orientation Applied to Unsupervised Classification of Reviews. Computing Research Repository - CORR, 417–424. https://doi.org/10.3115/1073083.1073153

[10] Cambria, E., & Hussain, A. (2015). Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer International Publishing. https://doi.org/10.1007/978-3-319-23654-4

[11] Vilares, D., Peng, H., Satapathy, R., Cambria, E. (2018). BabelSenticNet: A Commonsense Reasoning Framework for Multilingual Sentiment Analysis. Proceedings of IEEE SSCI, 1292-1298.

[12] Le-Hong, P., & Cambria, E. (2023b). Integrating graph embedding and neural models for improving transition-based dependency parsing. Neural Computing and Applications. https://doi.org/10.1007/s00521-023-09223-3

[13] Zhang, C., Li, Q., & Song, D. (2019). Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4568–4578). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1464

[14] Sun, K., Zhang, R., Mensah, S., Mao, Y., & Liu, X. (2019). Aspect-Level Sentiment Analysis Via Convolution over Dependency Tree. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5679–5688). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1569

[15] Liang, B., Su, H., Gui, L., Cambria, E., & Xu, R. (2022). Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowledge-Based Systems, 235, 107643. https://doi.org/10.1016/j.knosys.2021.107643

[16] Cambria, E., Mao, R., Chen, M., Wang, Z., & Ho, S.-B. (2023). Seven Pillars for the Future of Artificial Intelligence. IEEE Intelligent Systems 38(6), 62-69.

[17] Cambria, E., Zhang, X., Mao, R., Chen, M., & Kwok, K. (2024). SenticNet 8: Fusing Emotion AI and Commonsense AI for Interpretable, Trustworthy, and Explainable Affective Computing Proceedings of HCII, Washington DC.

[18] Kumar, AJ., Abirami, S., Trueman, TE., & Cambria. E. (2021). Comment Toxicity Detection via a Multichannel Convolutional Bidirectional Gated Recurrent Unit. Neurocomputing 441, 272-278.

[19] Rastogi, A., Liu, Q., & Cambria, E. (2022). Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study. Proceedings of IJCNN.

[20] Zhu, L., Mao, R., Cambria, E., & Jansen, BJ. (2024). Neurosymbolic AI for Personalized Sentiment Analysis. Proceedings of HCII.

[21] Cambria, E. (2024). *Understanding Natural Language Understanding*. Springer, ISBN 978-3-031-73973-6.

[22] Valdivia, A., Luzón, M., Cambria, E., Herrera, F. (2018). Consensus Vote Models for Detecting and Filtering Neutrality in Sentiment Analysis. Information Fusion 44, 126-135.

[23] Chaturvedi, I., Ong, Y., Tsang, I., Welsch, R., Cambria, E. (2016). Learning Word Dependencies in Text by Means of a Deep Recurrent Belief Network Knowledge-Based Systems 108, 144-154.

[24] Susanto, Y., Livingstone, A., Ng, B., Cambria, E. (2020). The Hourglass Model Revisited IEEE Intelligent Systems 35(5), 96-102

[25] Liu, Q., Han, S., Cambria, E., Li, Y., Kwok, K. (2024). PrimeNet: A Framework for Commonsense Knowledge Representation and Reasoning Based on Conceptual Primitives. Cognitive Computation 16