Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.30

Are Emotions Conveyed Across Machine Translations? Establishing an Analytical Process for the Effectiveness of Multilingual Sentiment Analysis with Italian Text

Richard Anderson
Rutgers University
rick.anderson@rutgers.edu

Carmela Scala
Rutgers University
carmela.scala@rutgers.edu

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Vivek Kumar
University of Cagliari
vivek.kumar@unica.it

Parth Jain
Rutgers University
pj269@rutgers.edu

Abstract

Natural language processing (NLP) is being widely used globally for a variety of value-creation tasks ranging from chat-bots and machine translations to sentiment and topic analysis and multilingual large language models (LLMs). However, most of the advances are initially implemented within the English language framework, and it takes time and resources to develop comparable resources in other languages. The advances in machine translations have enabled the rapid and effective conversion of content in global languages into English and vice versa. This creates potential opportunities to apply English language NLP methods and tools to other languages via machine translations. However, although this idea is powerful, it needs to be validated and processes and best practices need to be developed and kept updated. The present research is an effort to contribute to the development of best practices and an evaluation framework. We present a systematic and repeatable state-of-the-art process to evaluate the viability of applying English language sentiment analysis tools to Italian text by using multiple English language machine translation mechanisms such that it can be easily extended to other languages.

Keywords *natural language processing, natural language understanding, sentiment analysis, machine translation, italian, emotion.*

1. Introduction

Natural language processing (NLP) as a domain has been experiencing unprecedented breakthroughs and an exponential adoption growth rate by businesses, institutions, governments, and individual users, driven by an increasing interest in textual data and the analytical and generative potential it presents (Samuel et al. 2022b). The global NLP market is expected to grow to $49.4 billion (United States dollars) by 2027, and there have been many notable developments in NLP since late 2021 (Markets and Markets 2022): Apple will provide an open-source reference PyTorch implementation of the Transformer architecture for its products, enabling global developers to effortlessly run Transformer models. At the end of 2021, Baidu introduced PCL-BAIDU Wenxin (ERNIE 3.0 Titan), a state-of-the-art knowledge-enhanced 260 billion parameters-based large language model (LLM) for the Chinese language. This model outperformed its predecessors easily, and more recently, in March of 2023, the controlled launch of OpenAI's multimodal (accepts images and text as input) Generative Pre-trained Transformer 4 (GPT-4), speculated to have around two trillion parameters, via ChatGPT Plus has further demonstrated the power and expanding capabilities of LLMs (Liu et al. 2023).

Large language models have been developed with a primary focus on English, and a few other LLMs such as ERNIE 3.0 Titan in the Chinese language have also been developed (Wang et al. 2021; Nguyen et al. 2023). Google's 2021 MUM language model was trained across 75 languages and is an exception to mainly English-focused language models. Google's VP of Search declared that MUM as "1,000 times more powerful than BERT" and that it has "…the potential to transform how Google helps [users] with complex tasks" (Nayak 2021). From a resource availability and allocation perspective, it would be expensive and probably unfeasible to expect such models to be built and kept updated for every human language in the short term. It is clear that LLM performance "among under-represented languages fall behind due to pre-training data imbalance" (Nguyen et al. 2023).

It is even more challenging for a large array of NLP tools, models, and methods available in Python, R and other languages to be readily extended to alternative and vernacular languages with the same level of effectiveness (Ranathunga and de Silva, 2022). While there have been recent localized efforts to develop NLP tools in other languages such as Welsh, Marathi, and Malayalam, it is evident that much work remains to be done (Cunliffe et al. 2022; Lahoti et al. 2023; Sebastian 2023). Given the growing importance of textual data analytics and NLP applications in a wide array of research, policy, socioeconomic, healthcare, business, and other domains, and in addressing global events such as the COVID-19 pandemic, it is important to address the challenge of multilingual data (Samuel et al. 2020a, 2020b; Rahman et al. 2021; Ali et al. 2021). This could be done using multiple approaches including the grassroots level development of local language NLP tools which would be time consuming and lag well behind English language tools, and also through the use of machine translations which could create opportunities for timely applications.

The key question therefore is: Given the NLP advances in one language such as English, can we extend the applications and benefits to other languages by machine-translating such languages into the language with advanced NLP models and tools and then draw implications back to the original languages effectively? Past research has shown that such an approach is feasible, and it is possible to use machine translations in conjunction with other NLP tools, including sentiment analysis with increasing effectiveness (Balahur and Turchi, 2012, 2014). However, there is a need to articulate a clear, updated, and repeatable process for applying NLP tools from one language to another with an evaluation mechanism to compare and gauge the effectiveness of such a process. To address this, we ran an experiment with a lab-developed Italian text corpus, using multiple machine translations and multiple sentiment analysis tools. In the next section, we conduct a literature review of relevant state-of-the-art NLP methods and tools, followed by a description of our dataset, process, evaluation methods, and analysis. We conclude with a discussion of our process and analysis, notes on limitations, future research, and concluding thoughts.

2. Literature Review

Extant research has emphasized the paucity of NLP tools for many languages, and past studies have experimented with the use of machine translations-based sentiment analysis for languages such as Arabic, researchers affirmed the usefulness of machine translations in spite of the lack of high levels of accuracy (Mohammad et al. 2016; Oueslati et al. 2020). Steering away from translating the text corpus, past multilingual sentiment analysis research has also obtained fair results using "automated translation of the dictionary" for legislative bills (Proksch et al. 2019). A recent study using French, Spanish, and Japanese machine translations analyzed the impact of indirect (pivot, using a mediating language) machine translations on automated sentiment analysis and highlighted weaknesses of sentiment classifiers when working with translated texts while also affirming the usefulness of machine translations-based analysis (Poncelas et al. 2020). Going further, recent research has also posited that with certain languages, machine translations based on sentiment analysis using English language tools yielded better results than the language-specific tools used for sentiment analysis (Araújo et al. 2020).

More recently, Kumar et al. (2023) used "a zero-shot learning-based cross-lingual sentiment analysis (CLSA)" to demonstrate the viability of using machine translations-based sentiment analysis for the Sanskrit language. So also machine translation has been shown to work well with classifier performance for the Bengali language (Sazzed and Jayarathna, 2019; Sazzed 2020). Berard et al. (2019) applied sentiment analysis and focused on the benefits of improving the quality of machine translation using French language user-generated content. This is useful because extant research has highlighted numerous challenges with machine translation-based approaches including "sparseness and noise in the data" and the failure of translation mechanisms to "translate essential parts of a text, which can cause serious problems, possibly reducing well-formed sentences to fragments" (Dashtipour et al. 2016). A number of NLP-based studies in the Italian language have used sentiment analysis, such as performing sentiment analysis on Italian Twitter data, the use of cross-lingual transfer learning for analyzing the sentiment of Italian TripAdvisor review data, application of sentiment analysis and text mining for generating insights from YouTube Italian videos on vaccination and a comparison of lexicon-based and Bert-based methods (Basile and Nissim, 2013; Porreca et al. 2020; Catelli et al. 2022a, 2022b). Similarly, there has been a fair amount of research on NLP and machine translations of the Italian language, including basic translation automation effort and more advanced applied research (Russo et al. 2012; Wiesmann 2019; Bawden et al. 2020; Modzelewski et al. 2023). However, in spite of numerous multilingual studies in Italian, we did not find any comparable combination of NLP tools and machine translations-based studies for the Italian language.

3. Data and Method

In this section, we describe the development of the Italian dataset and the 'gold standard' human expert-assigned sentiment classifications and visualize a few key features as shown in Figures 1a and 1b. We then explain the machine translation process and report on the two translation models we applied (Figures 2a and 2b). We present our analysis of the accuracy and nuances of the machine translations from Italian to English, then report our findings from applying sentiment analysis to the English translations. We compare the sentiment assigned to the English translations to the original Italian language gold standard sentiment classes and present our findings.

3.1 Data

The unique dataset of sentences from Author 2 were human-generated Italian Sentences and not from public sources. These sentences were created to have a clear positive, neutral, or negative sentiment. This information was recorded in the dataset. For this experiment, we used two translation methods, one was the web tool for Google Translate (Han 2022). We used that as a common and popular source of translations. To get the same results as the web tool for Google Translate method, we used the googletrans Python library. This library provides a convenient interface to Google Translate, allowing for consistent translation operations within Python scripts. Next, we chose the Marian Machine Translation Transformer-based technique for translations (Junczys-Dowmunt et al. 2018). It had a documented method of translation that could be reproduced.

We used the Marian Neural Machine Translation (Marian MT) and model to translate Italian to English using the Marian MT method:

opus-mt-tc-big-it-en Neural machine translation model for translating from Italian (it) to English (en). This model is part of the **OPUS-MT project**, an effort to make neural machine translation models widely available and accessible for many world languages (Tiedemann 2020; Tiedemann and Thottingal, 2020). All models are originally trained using the framework of **Marian NMT**, an efficient MT implementation written in pure C++ (Junczys-Dowmunt et al. 2018). The models have been converted to pyTorch using the transformers library by Huggingface. Training data is taken from **OPUS**, and training pipelines use the procedures of **OPUS-MT-train**.

(a) Complete Italian no stop words.

(b) Complete Italian stop words removed.

Figure 1: Word clouds for complete Italian text with and without stop words.

(a) Word Cloud for Google Translate Text where both translations are true.

(b) Word Cloud for Marian MT Italian to English Text where both translations are true.

Figure 2: Word clouds for Google Translate text and Marian MT Italian to English text where both translations are true.

Then we ran VADER sentiment evaluation on each of the translated sentences. Starting from a total of 167 sentences. Sixty-five sentences are correct in translation. Thirty-nine percent of the sentences were accurate from both datasets. The official source data frame will include the data where sentences are accurate in both translations. These "both true" sentences will be used in the remaining analysis. That way, we are measuring the good translations from here on out.

Google Translate had 91 good translations, while Opus Translate had 119. This indicates that Opus Translate performed slightly better in terms of translation quality in this specific dataset. There were 65 instances where both Google Translate and Opus Translate had good translations. This suggests some overlap in the quality of translations between the two engines.

BLEU and chrF scores are commonly used to measure the quality of a corpus and how well it adheres to accepted translations for the corpus. We used both techniques on our dataset to determine whether either would give us an automated method of evaluating the translated sentences. Our dataset includes the true sentiment and the "correct" translation. Either method might have biases based on their method of calculation. When we compared, the data BLEU and chrF scores on our dataset matched human the approved gold standard sentences. The BLEU and chrF metrics vary for how far off a translation is from what is expected, but both agree that the same sentences the expert has said are good translations. Either metric is good at confirming the human-chosen good translations for our dataset.

3.2 Machine Translations EVALUATION (- RICK to DESCRIBE MT and EVAL METRICS)

The process of analysis was automated in the following Colab Notebook: https://github.com/rianders/mtnlpxlmsentiment/blob/main/SentimentAnalysisAll.ipynb.

We used this dataset:

https://github.com/rianders/mtnlpxlmsentiment/blob/main/data/SentinmentALL-20230508.csv.

Evaluation tools: Marian MT OPUS Italian Dataset Google Translate VADER BLEU chrF

The notebook fetches the source data created by Dr. Scala and Dr. Samuel. These data include the true and gold standard sentiment, source sentence, and official translation sentence information. The next step cleans the data, runs the BLEU and chrF comparisons, and adds that information to the dataset. Then the translation and machine translation quality checks and graphs are created to confirm quality and accuracy. These quality checks and graphs include word frequency, sentence length, BLUE, and chrF scores. Translation comparisons are performed between Google Translated and Marian MT using OPUS Italian Data.

Now that the quality of translation has been determined, a review of sentiment distribution is shown. We use the VADER method where -1 is negative, 0 is neutral, and 1 is positive.

Then a new data frame is created that only contains the "correct" translations. Then word clouds are generated from that data frame. We calculate the confusion matrix, word frequency, and sentence length for each translation method. Then identify and show the outliers.

This process can be repeated with the same or updated dataset. The evaluation process will be the same and can give accurate feedback.

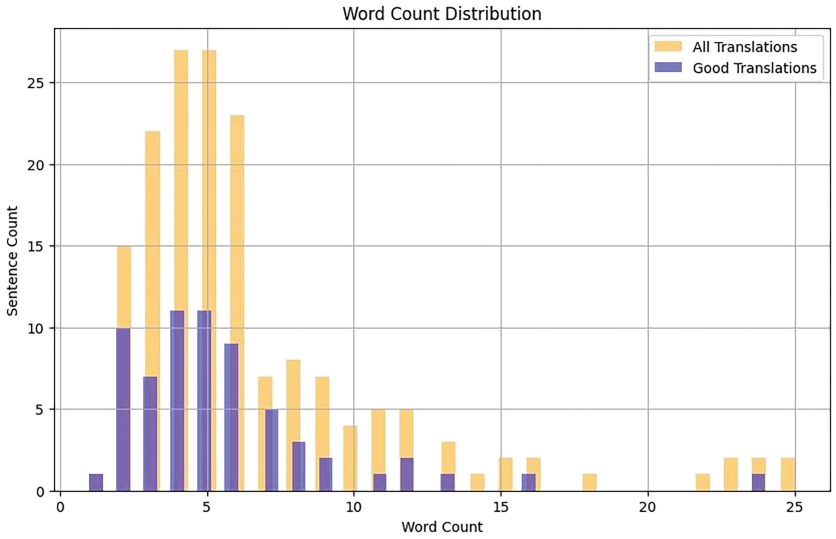

Figure 3: Histogram of the count of sentences by number of words.

3.3 Machine Translations—Observations

These word clouds show us something exciting about languages and language translations. Generally speaking, both translation software performed much better with short sentences and had some problems with longer ones (Figure 3). If we look closely at them, the Italian one does not have many words in common with the various English images, which seem very similar. We have a predominance of nouns, adjectives, and adverbs in the English clouds. In the Italian one, we have primarily articles, conjunctions (che in the specific), prepositions, and verbs. This discrepancy between the Italian word clouds and the English ones is easy to explain whether we consider the typical sentence structure of the Italian language. Italian uses articles, prepositions, and conjunctions (especially che) much more than English, and the word clouds captured this difference perfectly.

Generally speaking, both translation software performed much better with short sentences and had some problems with longer ones. In the sentence below, for example, in the second part, Google assumed the 'subject' was "Tom Cruise," thus translating "mi ha emozionato" with "he excited me" when it should have been "it excited me." In the original sentence, the subject was the movie, not the actor. Google translations were also more "literal"; hence they did not always produce sensible sentences in English. On the contrary, Opus's second set of translations was more accurate from an "idiomatic" point of view. In fact, Opus could better identify the idiomatic peculiarities of the sentences. In the sentence below, for example, in the second part, as noted above, Google Translate assumed the "subject" was "Tom Cruise," thus translating "mi ha emozionato" with "he excited me" when it should have been "it excited me." In fact, the subject was the movie and not the actor. Opus, instead, provided the perfect translation, identifying the right subject, "it.":

Italian: *Ho visto il nuovo film di Tom Cruise, "Maverick", e devo dire che mi ha emozionato perché mi ha riportato alla mia gioventù.*

English: *I saw Tom Cruise's new film, "Maverick," and I must say that he excited me because he brought me back to my youth.*

Also, it is essential to point out that some translations would have been correct in British English but are considered incorrect or inaccurate in USA English. Here are some examples:

1. Giovanna ha scelto di giocare a calcio: Giovanna chose to play football.
2. Ieri siamo andati allo stadio a vedere una partita di calcio: Yesterday we went to the stadium to watch a football match.

   Football is accurate in British English but in the United States, "calcio" is referred to as soccer.
3. Abbiamo deciso di cambiare casa: We decided to change homes.
4. E stata una vacanza da sogno: It was a dream vacation! OR It was a dream holiday! "We decided to change homes" and "It was a dream holiday" could be accepted in British English, but they sound wrong in United States English.

3.4 Data Analysis

The original dataset containing all translations had a range of sentiments that did not show coherence between translation methods. When we reviewed the sentiment for sentences considered accurate, VADER generated linear agreement across the negative, neutral, and positive categories. This is observed in Figures 4 and 5. Figure 4 shows the range of sentiments that include inaccurate translations. Figure 5 shows the VADER sentiment of accepted translations and that they stay along a linear path across the sentiment values.

When we categorize the remaining sentiment after removing the inaccurate translations, Figure 6 shows that 49.2% of the remaining sentences are positive. The Negative is 23.1%, and the Neutral is 27.7%.

The values for VADER sentiment are shown for Google Translate in Figure 7 and Marion MT in Figure 8. The VADER sentiment was the same for both translation techniques in terms of percentages. The way to tell the difference was to observe the outliers.

Figure 9 shows the VADER sentiment Confusion Matrix for Google Translate and Figure 10 shows the Confusion Matrix for the Marion MT translations. These two graphs show that the misclassifications were similar and that to determine the type of outlier, is to look at those misclassification cases. Those outliers are listed in listings 1 through 3. Figures 11 and 12 show the outliers and that neutral values did not have a clear cluster towards neutral. That one value in the negative was in the positive range.

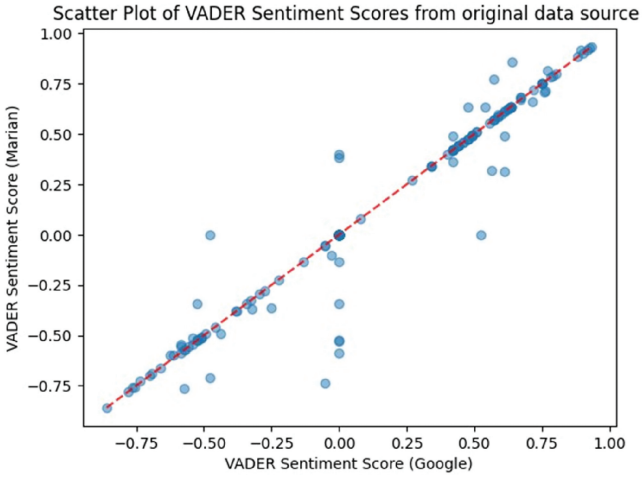

Figure 4: VADER sentiment for Marion MT and Google.

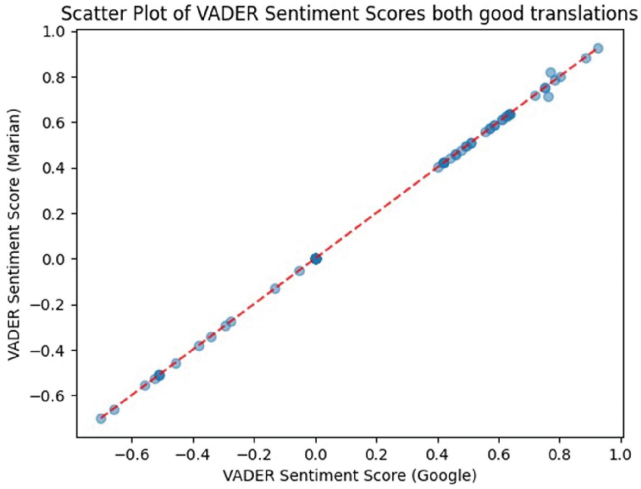

Figure 5: VADER sentiment for Marion MT and Google Translate including original bad translations translate accepted translations.

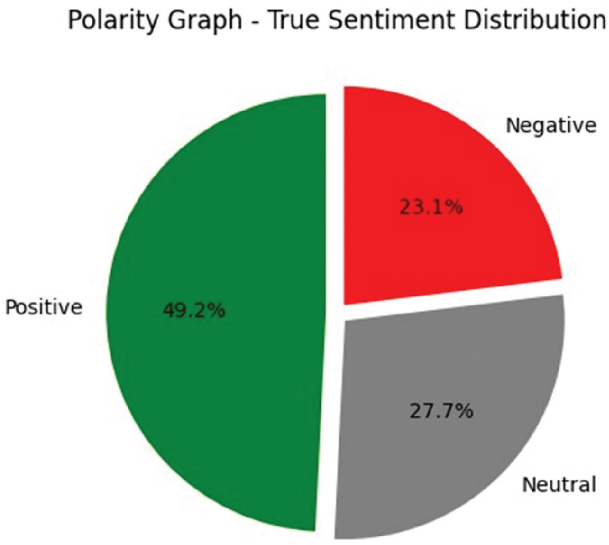

Figure 6: The true sentiment distribution for accurate translations.

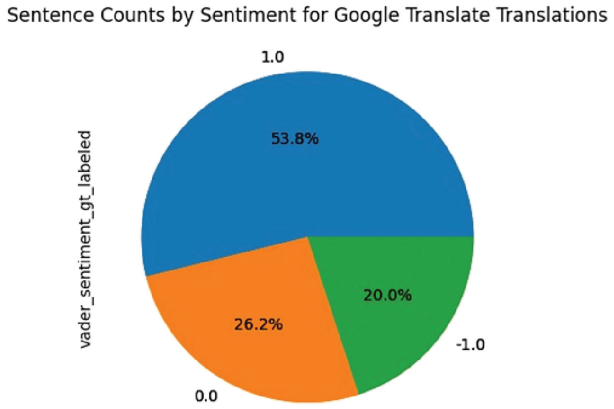

Figure 7: VADER sentiment for Google Translate by percentage number of sentences in category.

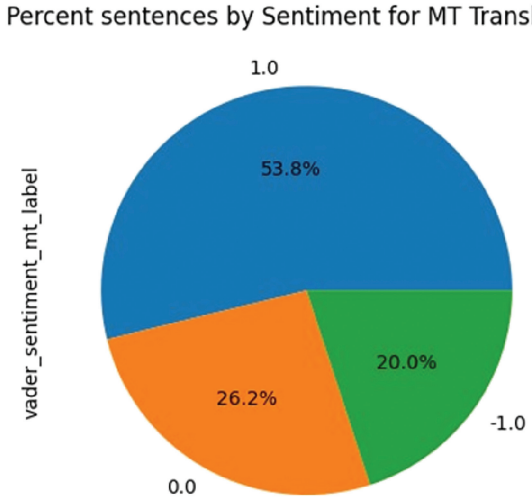

Figure 8: VADER sentiment for Marian Machine Translation by percentage number of sentences in category.

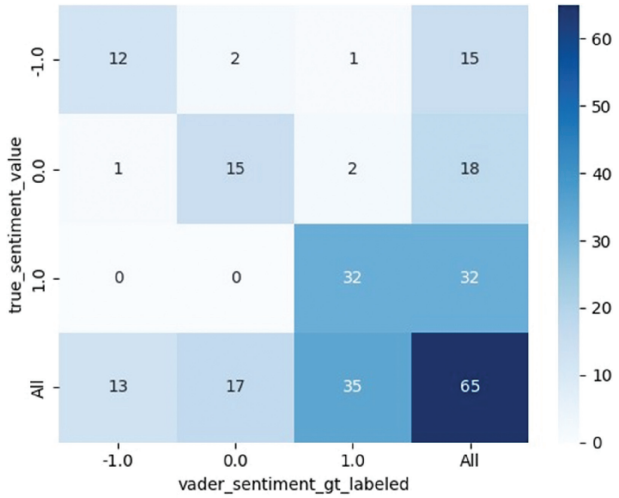

Figure 9: VADER sentiment confusion matrix for Google Translate.

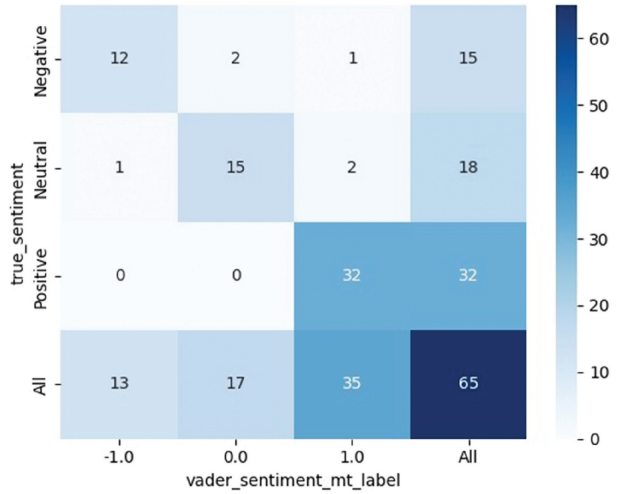

Figure 10: VADER sentiment confusion matrix for Marian Machine Translate.

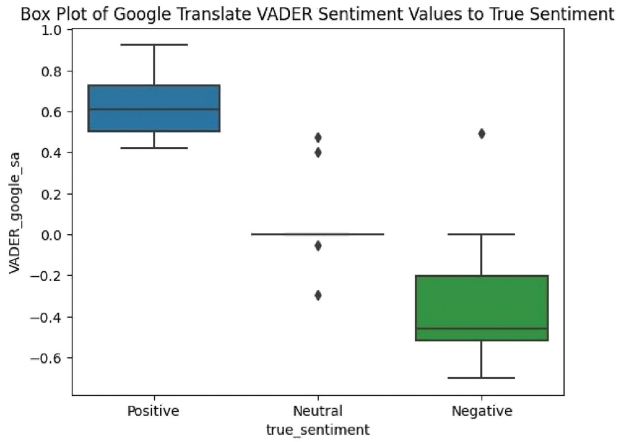

Figure 11: VADER sentiment box plot for Google Translate.

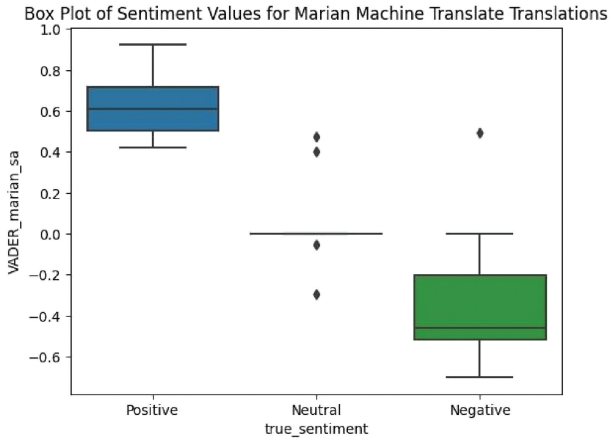

Figure 12: VADER sentiment box plot for Marian Machine Translate.

When examined, the outliers are the same; both calculated VADER values for either Google Translate or Marian MT are the same so all outliers are the same. This would be an area where more overall data would be useful.

4. Discussion

For both Google Translate and Marian MT we compared the true values with the rated values and looked for patterns in the outliers. What were there any after we removed the bad translations. We will review the outliers by Google Translate Positive, Neutral, and Negative. Then do the same for Marian MT. In this subsection, we review outliers for Google Translate output. There were no Google Translate positive outliers, implying no true positive sentiment statements falsely classified as neutral or negative after being translated into English.

The Google Translate sentence had no positive outliers.

Listing 1: Outliers for Neutral

```
Outliers for Neutral:
Sentence: La musica americana attrae sempre molti giovani italiani
Google Translation: American music always attracts many young Italians
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment: 0.4019
_____________

Sentence: I miei amici sono venuti a farmi visita
Google Translation: My friends came to visit me
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment: 0.4767
_____________

Sentence: Non ho preferenze su cosa fare stasera
Google Translation: I have no preferences on what to do tonight
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment: −0.296
_____________
```

Sentence: Domani partiamo per andare in Italia

Google Translation: Tomorrow we leave to go to Italy

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment: −0.0516

Outliers Marian Machine Translation

Using Marian MT, we had no positive outliers. Listing 3 and 4 show the neutral and negative outliers.

Listing 2: Marian MT Outliers for Neutral

Outliers for Neutral:

Sentence: La musica american attrae sempre molti giovani italiani

Marian MT Translation: American music always attracts many young Italians

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment (MT): 0.4019

Sentence: I miei amici sono venuti a farmi visita

Marian MT Translation: My friends came to visit me

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment (MT): 0.4767

Sentence: Non ho preferenze su cosa fare stasera

Marian MT Translation: I have no preference on what to do tonight

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment (MT): −0.296

Listing 3: Marian MT Outliers for Negative

Outliers for Negative:

Sentence: Sei un essere abominevole.

Marian MT Translation: You are an abominable being.

True Sentiment: Negative

True Sentiment Value: −1.0

VADER Sentiment (MT): 0.0

Sentence: Disapprovo la tua scelta!

Marian MT Translation: I disapprove of your choice!

True Sentiment: Negative

```
True Sentiment Value: −1.0
VADER Sentiment (MT): 0.0
_____

Sentence: Buono a nulla!
Marian MT Translation: Good for nothing!
True Sentiment: Negative
True Sentiment Value: −1.0
VADER Sentiment (MT): 0.4926
_____
```

All the sentences in the original Italian have no connotation of positive or negative sentiment; hence, they are considered neutral. Indeed, they do not express an opinion but simply state facts. So, why are they perceived differently in translation? In the case of the second and third sentences, we can assume that the sentiment was perceived as negative due to the presence of some negative signifiers, such as "no" (in "I have no preference…") and the verb "leave" (in "Tomorrow we leave…"). "Leave" implies a separation, and that is why it was probably perceived as negative. As for the first sentence, which was neutral but perceived as positive, we can infer that the presence of verbs indicating "company," such as "came" and "visit," triggered the positive interpretation. As mentioned above, these misinterpretations are also due to the absence of context and "voice inflection."

4.1 Why Was Polarized Italian Made Neutral?

The neutral score assigned to these sentences was a bit of a surprise. Let's analyze the original sentiments. Here are possible explanations for each of them: "Sei un essere abominevole" (translated correctly as "You are an abominable being") clearly has a negative sentiment. Even in the absence of overtly negative signifiers (such as "no," "not," "never," etc.), the word "abominable" sets the tone for a negative interpretation. However, since NLP models are primarily trained to identify the presence of positive and negative words to determine the sentiment of a sentence, in really short sentences where the rest of the signifiers are neither positive nor negative, the NLP model might make a "decision" to assign a neutral sentiment. In this case, that might be why it was perceived as neutral.

As for the second sentence, "Disapprovo la tua scelta" (translated as "I disapprove of your choice"), "disapprove" is the only overtly negative word, making the sentiment clearly negative.

Regarding the last sentence, "Buono a nulla" (translated as "Good for nothing"), it is possible that the NLP model was confused by the equal presence of positive and negative signifiers: "buono" (good) is positive, while "nulla" (nothing) is negative. Consequently, it might have perceived the conflicting sentiments as canceling each other out and opted for a neutral score.

In summary, the different interpretations of the original sentiments in translation could be attributed to the NLP model's reliance on identifying positive and negative words to determine sentiment and the specific words present in each sentence that contribute to the overall sentiment.

Sentiment Analysis General observations:

The sentiment analysis presented some interesting "challenges," more so in dealing with neutral sentences.

Let us look at some examples.

1. "L' esperienza studio in Italia è stata unica. Mi ha cambiato letteralmente la vita e mi ha aperto gli occhi su una nuova realtà." (Original positive) [The study experience in Italy was unique. It literally changed my life and opened my eyes on a new reality.]

   All three engines, Google, Opus, and NLTK assigned a score of 0, neutral, to this sentence. The original sentence bears a clear positive message: the study abroad experience was mindblowing, and it changed the student's life forever (it is implicit that it changed it in a positive way.) Yet this positivity did not translate into the English version even though the sentence's translation was correct for both Google and Opus.

   I believe the problem here was the word "unique," which can have positive, negative, and neutral connotations according to the context. If unique is intended as being "peculiar," it has a negative meaning; if it is used to point out that something or someone is just "different," the word carries a neutral connotation. If it is used to indicate that something or someone has "no equal," then it is positive. It is possible that the

neutral scores are justified by the fact that the engines perceived the study abroad experience as being simply "different." Furthermore, reflecting on the second part of the original sentence, "mi ha cambiato letteralmente la vita" (it changes my life completely), one could argue that a change in life is not always a positive event. It depends on the context and the person's perception of the events. From a linguistic point of view, the expression "L' esperienza studio in Italia è stata unica" in Italian is undoubtedly positive. In Italian, something that is unique to you is "positive." You would never use this expression to talk about something that was indifferent to you or negative. If an event is perceived as neutral, one would say, for example, "è stata un'esperienza normale," or "è stata un'esperienza come un'altra" ("it was a normal/uneventful experience," or "it was an experience like any other." If it is negative, then one would say: "E' stata una brutta esperienza" or "Un'esperienza negativa." ("It was a bad experience" or "It was a negative experience.")

2. "Sei raggiante!" (original positive) [You are glowing! (Opus); You are radiant! (Google)] Google sentiment score: 0.5255 (positive) Opus Sentiment score: 0 (neutral) NLTK Sentiment score:1 (neutral)

   This is another interesting case. The original sentence is clearly positive. To tell someone they are glowing in Italian is to compliment them. Yet Opus and also NLTK assigned it a neutral score. Opus's score is even more interesting because the translation it provided is more accurate than the one provided by Google. The most sensible explanation for this mistake in sentiment analysis would be that the word "glowing" in English is used in a variety of expressions that also carry negative feelings.

   "glow with something. 1. Lit. [for something] to put out light, usually because of high heat. The embers glowed with the remains of the fire. The last of the coals still glowed with fire. 2. Fig. [for someone's face, eyes, etc.] to display some quality, such as pride, pleasure, rage, health. Her healthy face glowed with pride. Her eyes glowed with a towering rage." "[https://idioms.thefreedictionary.com/glowing:. . . . .: text=

   I believe that this different use of the word "glowing" contributed to the final calculation of the sentiment score and justified the neutral rating assigned by Opus. The NLTK score averages out the sentiment scores provided by Google and Opus and leans towards the neutral sentiment. However, it also provides a 0.629 score for positive sentiment, thus recognizing the intent of the original.

3. È stata una vacanza da sogno! (Original positive) [It was a dream vacation! (Opus); It was a dream holiday! (Google)]

   Google Sentiment score: 0.6114 (positive) Opus Sentiment score: 0.3164 NLTK: positive score 0.433; neutral score: 0.567

   This one is worth discussing for the disparity among the different scores. Although all of the "datasets" assigned a positive score, there was a significant difference between Google and the score provided by Opus and NLTK. Google's assignment of the score appeared to be much more confident; Opus and NLTK provided a positive score with a lower level of confidence (in fact, NLTK also assigned a higher neutral score to this sentence) How do we explain this? The translations are both good (even though the one provided by Google seems more proper in British English). A plausible explanation could be the fact that "a dream vacation" is something different for everyone, thus subjectivity plays a role in determining the positivity or neutrality of the sentiment.

4. Mi hai delusa! (Original negative) [You let me down! (Opus); You disappointed me! (Google)] Google sentiment score: −0.4767 Opus sentiment score: 0 (neutral) NLTK Neutral score: 1 NLTK Negative score: 0 TextblobSentimentpolarity: −0.1555556

   The translations provided are both good, with a slight preference for Opus, which is more exact from an idiomatic point of view. Google's score is perfect as it identifies the original sentiment. This is interesting because, as mentioned above, Google does not provide a better translation but still is on point with the sentiment score. However, despite providing a better translation, Opus read the sentence as neutral, and NLTK also assigned the sentence a neutral score. The sentiment polarity was a low negative.

   There is no doubt that the original sentence has a negative connotation, "mi hai delusa" is a sentence that expresses sadness, anger, and disillusionment. I would imagine that "You let me down!" works the same way. That is why the neutral score was a surprise and needs further investigation. In fact, as of now, there is no plausible explanation for the mistake.

5. Sei un inetto! (Original negative) [You're inept! (Opus); You're an inept! (Google)] Opus Google score: 0 (neutral) NLTK Neutral score: 1 NLTK Negative score: 0

Again, as in the case above, the original leaves no room for misunderstanding. In Italian culture, calling someone "inetto" is certainly an offense; hence the expression carries a negative sentiment. It is possible, however, that the sentence was read as a "personal opinion," which is obviously not universal and open to personal interpretation. This would justify the neutral score assigned.

6. Buono a nulla! (Original negative) [Good for nothing! (Opus & Google)] Google score: 0.4926 Opus score: 0.4926 NLTK Positive score: 0.615 NLTK Negative score: 0 NLTK Neutral score: 0.385

   In Italian, "Buono a nulla!" is another way to say "inept" and just like the sentence above expresses a negative sentiment. The error in sentiment analysis can be justified by the presence of the word "good," which is usually positive. The three datasets picked up the sentiment score carried by the word 'good' and consequently read the sentence as positive.

7. "In questo momento mi sento piuttosto calma non provo emozioni forti. (Original Neutral) [Right now, I feel rather calm; I don't feel strong emotions. (Opus); "At this moment, I feel quite calm I do not feel strong emotions. (Google)] Google score: −0.0281 Opus score: −0.1032 NLTK Positive score: 0.192 NLTK Negative score: 0.225 NLTK Neutral score: 0.583

   "Alla fine dei conti puoi fare quello che desideri, a me non interessa molto. (Neutral) [At the end of the accounts, you can do what you want. I don't care much. (Opus Google)] Google score: −0.3244 Opus score: 0.3705 NLTK Positive score: 0.076 NLTK Negative score: 0.166 NLTK Neutral score: 0.758

   "Non ho preferenze su cosa fare stasera." [I have no preference on what to do tonight.(Opus Google) Google score: −0.296 Opus score: −0.296 NLTK Positive score: 0 NLTK Negative score: 0.239 NLTK Neutral score: 0.787

   "Non ho mai favorito nessuno studente, per me sono tutti uguali." (neutral) [I have never favored any student, for me, they are all the same. (Opus Google)] Google score: −0.3252 Opus score: −0.3252 NLTK Positive score: 0 NLTK Negative score: 0.189 NLTK Neutral score: 0.811

   These sentences were presented as neutral in the original because they do not express positive or negative feelings or attitudes. Indeed, the "subjects" of the sentences are neither upset nor happy; neither in favor nor against a particular situation, they are simply "emotions/opinions free"; thus, the sentences are neutral. However, they were rated as negative by both Google and Opus, while the NLTK scores were more on point.

   Here are some possible explanations:

   The presence of the negative words "Non ho/I do not"; "Senza/Without" might have led the "analysis" in the wrong direction.

   Also, the absence of context might have had a role in leading to the wrong score.

   Indeed, some of these sentences could sound negative if pronounced with an upset tone. This is true, especially for these two sentences:

   "Alla fine dei conti puoi fare quello che desideri, a me non interessa molto." "A me non interessa" (I don't care much) can be negative if pronounced with an altered/upset tone. It can communicate a lack of "interest" and "feelings." However, if the same sentence (at least in Italian) is pronounced with a flat tone, then it just communicates "neutrality." "Non ho preferenze su cosa fare stasera." In this case, if the sentence is pronounced within the context of an argument, hence with an altered tone of voice, then it can have a negative feeling. But if a person says it just to express that "s/he would go with the flow," it is entirely neutral.

   Last but not least, two more sentences are worthy of attention:

8. "La musica americana attrae sempre molti giovani italiani." (Original neutral) ["American music always attracts many young Italians." (Opus Google)] Google score: 0.4019 Opus score: 0.4019 NLTK Positive score: 0.31 NLTK Negative score: 0 NLTK Neutral score: 0.69

   "Domani partiamo per andare in Italia." (Original neutral) ["Tomorrow we leave to go Italy." (Opus Google)] Google score: −0.0516 Opus score: −0.0516 NLTK Positive score: 0 NLTK Negative score: 0.167 NLTK Neutral score: 0.833

   These two sentences in Italian are plain statements. They express simple facts: American music is popular, and "tomorrow" we are going to Italy. Yet the first was rated with a positive score (only NLTK proposed

a neutral score). The presence of the word "attracts," which intrinsically has a positive meaning, possibly led the datasets to identify this sentence as positive.

As for the second one, it is plausible that the word 'leave' which indicates "separation," might have led to the negative score.

5. Limitations

Our research in machine translations-based NLP solutions is presented as a lead study to establish a robust process at the intersection of state-of-the-art machine translations, English language NLP tools and languages other than English. For the purposes of this study, we use apply sentiment analysis, and two machine translations of an original Italian language dataset. There are a few limitations and these serve as areas of future research. We initiated the pilot project with a set of expert-created Italian language sentences. First, the dataset is specifically created for this study and not "real world" data in the sense of it being secondary data from external sources such as social media posts or news articles or blogs. Secondly, it is a small dataset, especially in the context of the LLMs which uses large quantities of data. Hence the findings may have limited external validity. However, since this is a lead study aimed at establishing a research process, the use of a custom dataset with limited size is justified given the rigorous and thorough analytical framework which is critical for a sustainable process and this enhances internal validity.

The third limitation is that we have tested only two machine translation models and fourth, only three sentiment analysis methods were applied. However, this is sufficient because this approach meets the goals of process centricity and process validations for this study. The use of two translations and three sentiment analysis methods ensures a simplified but rigorous approach for process validation, ensuring external validity. Therefore, in spite of the aforementioned limitations, the research accomplishes the main objective of the lead study, which is to establish a transparent and repeatable process for further and extensive analysis of the reliability of machine translation-based NLP solutions.

6. Future Research

Our lead study using English translations of Italian text has conceptually illustrated the usefulness of such an approach for extending the use of English language NLP tools to Italian text. Our future research will include additional languages, expand the size of the data analyzed, increase the number of machine translations applied, and explore the use of additional sentiment analysis methods. Incorporating open data into future research will be useful to facilitate public benefit and greater application potential (Samuel et al. 2023). We will also include the validation of additional NLP solutions such as identifications of topics and named entity recognition (NER). There is a significant need to establish best practices for machine translation-driven application of NLP solutions and future research should aim to address this need. In spite of recent calls to slow down NLP research and development, there is sufficient reason to believe that we will see rapid developments in this domain over the next few years (Samuel 2023a). Furthermore, within the broader context of artificial intelligence (AI), defined as the ability of machines to "mimic the functions and expressions of human intelligence, specifically cognition, and logic," it will be valuable to explore combining machine translations and multimodal approaches, including the recognition of images and handwritten text (Samuel 2021; Jain et al. 2023; Liu et al. 2023).

7. Conclusion

Our research affirms past studies that have illustrated the viability of using English translations of native texts with machine translation mechanisms for applications of sense-making methods and tools such as sentiment analysis (Balahur and Turchi, 2012, 2014). Furthermore, our study has created a new Italian dataset and a simple, repeatable, and effective process for testing and validating the use of English translations for NLP applications—this will enable us and other researchers to quickly validate many global languages for machine translations-based NLP solutions. Despite recent concerns over risks and ethics, NLP, generative, and adaptive AI technologies are expected to grow exponentially over the next few decades and have a significant societal impact (Samuel et al. 2022a; Samuel 2023b). In this context, we anticipate it will become increasingly important to use machine translations in conjunction with other NLP tools and AI technologies to address complex individual, community, and societal problems effectively. We anticipate an increased emphasis on machine translation-based NLP solutions to address issues of public importance and expect our novel process contribution to help applied NLP researchers develop solutions with greater efficiency.

References

Ali, G. M. N., M. M. Rahman, M. A. Hossain, M. S. Rahman, K. C. Paul, J. C. Thill, and J. Samuel. 2021. "Public Perceptions of Covid-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics." *Healthcare* **9**, no. 9: 1110. doi: 10.3390/healthcare9091110

Araújo, M., A. Pereira, and F. Benevenuto. 2020. "A Comparative Study of Machine Translation for Multilingual Sentence-Level Sentiment Analysis." *Information Sciences* **512**: 1078–1102. doi: 10.1016/j.ins.2019.10.031

Balahur, A., and M. Turchi. 2012. "Multilingual Sentiment Analysis Using Machine Translation?" *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Jeju, Republic of Korea, The Association for Computer Linguistics, July 12.

Balahur, A., and M. Turchi. 2014. "Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis." *Computer Speech & Language* **28**, no. 1: 56–75. doi: 10.1016/j.csl.2013.03.004

Basile, V., and M. Nissim. 2013. "Sentiment Analysis on Italian Tweets," *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, Association for Computational Linguistics, June 14.

Bawden, R., G. M. Di Nunzio, C. Grozea, I. J. Unanue, A. J. Yepes, N. Mah, D. Martinez, A. Névéol, M. Neves, M. Oronoz, O. Perez-de-Viñaspre, M. Piccardi, R. Roller, A. Siu, P. Thomas, F. Vezzani, M. V. Navarro, D. Wiemann, and L. Yeganova. 2020. "Findings of the Wmt 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages." *Proceedings of the Fifth Conference on Machine Translation*, Online, Association for Computational Linguistics, November 19.

Berard, A., I. Calapodescu, M. Dymetman, C. Roux, J. L. Meunier, and V. Nikoulina. 2019. "Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness." Preprint, submitted October 31. https://doi.org/10.48550/arXiv.1910.14589

Catelli, R., L. Bevilacqua, N. Mariniello, V. S. di Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito. 2022a. "Cross Lingual Transfer Learning for Sentiment Analysis of Italian Tripadvisor Reviews." *Expert Systems with Applications* **209**: 118246. doi: 10.1016/j.eswa.2022.118246

Catelli, R., S. Pelosi, and M. Esposito. 2022b. "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian." *Electronics* **11**, no. 3: 374. doi: 10.3390/electronics11030374

Cunliffe, D., A. Vlachidis, D. Williams, and D. Tudhope. 2022. "Natural Language Processing for under-Resourced Languages: Developing a Welsh Natural Language Toolkit." *Computer Speech & Language* **72**: 101311. doi: 10.1016/j.csl.2021.101311

Dashtipour, K., S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou. 2016. "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques." *Cognitive Computation* **8**, no. 4: 757–771. doi: 10.1007/s12559-016-9415-7

Han, S. 2022. "googletrans: Free and Unlimited Google Translate API for Python." Accessed February 27, 2023. https://github.com/ssut/py-googletrans

Jain, P. H., V. Kumar, J. Samuel, S. Singh, A. Mannepalli, and R. Anderson. 2023. "Artificially Intelligent Readers: An Adaptive Framework for Original Handwritten Numerical Digits Recognition with OCR Methods." *Information* **14**, no. 6: 305. doi: 10.3390/info14060305

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, *et al.* 2018. "Marian: Fast Neural Machine Translation in C++." *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, Association for Computational Linguistics, April 14. http://www.aclweb.org/anthology/P18-4020

Kumar, P., K. Pathania, and B. Raman. 2023. "Zero-Shot Learning Based Cross-Lingual Sentiment Analysis for Sanskrit Text with Insufficient Labeled Data." *Applied Intelligence* **53**, no. 9: 10096–10113. doi: 10.1007/s10489-022-04046-6

Lahoti, P., N. Mittal, and G. Singh. 2023. "A Survey on NLP Resources, Tools, and Techniques for Marathi Language Processing." *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**, no. 2: 1–34. doi: 10.1145/3548457

Liu, H., R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang. 2023. "Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4." Preprint, submitted May 5. https://doi.org/10.48550/arXiv.2304.03439

Markets and Markets. 2022. "Natural Language Processing Market." Accessed September 5, 2022. https://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp

Modzelewski, A., W. Sosnowski, M. Wilczynska, and A. Wierzbicki. 2023. "Dshacker at Semeval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation." *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada, Association for Computational Linguistics, July 13.

Mohammad, S. M., M. Salameh, and S. Kiritchenko. 2016. "How Translation Alters Sentiment." *Journal of Artificial Intelligence Research* **55**: 95–130. doi: 10.1613/jair.4787

Nayak, P. 2021. "MUM: A New AI Milestone for Understanding Information." Google. Accessed March 2023. https://blog.google/products/search/introducing-mum/

Nguyen, X. P., S. M. Aljunied, S. Joty, and L. Bing. 2023. "Democratizing LLMs for Low-Resource Languages by Leveraging Their English Dominant Abilities with Linguistically-Diverse Prompts." Preprint, submitted June 20. https://doi.org/10.48550/arXiv.2306.11372

Oueslati, O.,. E. Cambria, M. B. HajHmida, and H. Ounelli. 2020. "A Review of Sentiment Analysis Research in Arabic Language." *Future Generation Computer Systems* **112**: 408–430. doi: 10.1016/j.future.2020.05.034

Poncelas, A., P. Lohar, A. Way, and J. Hadley. 2020. "The Impact of Indirect Machine Translation on Sentiment Classification." Preprint, submitted August 25. https://doi.org/10.48550/arXiv.2008.11257

Porreca, A., F. Scozzari, and M. Di Nicola. 2020. "Using Text Mining and Sentiment Analysis to Analyse Youtube Italian Videos Concerning Vaccination." *BMC Public Health* **20**, no. 1: 259. doi: 10.1186/s12889-020-8342-4

Proksch, S. O., W. Lowe, J. Wäckerle, and S. Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* **44**, no. 1: 97–131. doi: 10.1111/lsq.12218

Rahman, M. M., G. M. N. Ali, X. J. Li, J. Samuel, K. C. Paul, P. H. Chong, and M. Yakubov. 2021. "Socioeconomic Factors Analysis for COVID-19 US Reopening Sentiment with Twitter and Census Data." *Heliyon* **7**, no. 2: e06200. doi: 10.1016/j.heliyon.2021.e06200

Ranathunga, S., and N. de Silva. 2022. "Some Languages Are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World." Preprint, submitted October 20. https://doi.org/10.48550/arXiv.2210.08523

Russo, L., S. LoáIciga, and A. Gulati. 2012. "Improving Machine Translation of Null Subjects in Italian and Spanish." *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, Association for Computational Linguistics, April 16.

Samuel, J. 2021. *A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies*. Boston, MA: Scholars Strategy Network.

Samuel, J. 2023a. "Response to the March 2023 'Pause Giant Ai Experiments: An Open Letter' by Yoshua Bengio, Signed by Stuart Russell, Elon Musk, Steve Wozniak, Yuval Noah Harari and Others." Preprint, submitted March 29. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412516

Samuel, J. 2023b. "The Critical Need for Transparency and Regulation amidst the Rise of Powerful Artificial Intelligence Models," accessed August 2, 2023, https://scholars.org/contribution/critical-need-transparency-and-regulation

Samuel, J., G. Ali, M. Rahman, E. Esawi, Y. Samuel. 2020a. "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification." *Information* **11**, no. 6: 314. doi: 10.3390/info11060314

Samuel, J., M. Brennan, M. Pfeiffer, C. Andrews, and M. Hale. 2023. "Garden State Open Data Index for Public Informatics." NJSPL Report. https://policylab.rutgers.edu/report-release-garden-state-open-data-index/#_ftn1

Samuel, J., R. Kashyap, Y. Samuel, and A. Pelaez. 2022a. "Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations." *International Journal of Information Management* **65**: 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, J., M. M. Rahman, G. M. N. Ali, Y. Samuel, A. Pelaez, P. H. J. Chong, and M. Yakubov. 2020b. "Feeling Positive about Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics." *IEEE Access* **8**: 142173–142190. doi: 10.1109/ACCESS.2020.3013933

Samuel, J., R. Palle, and E. Soares. 2022b. "Textual Data Distributions: Kullback Leibler Textual Distributions Contrasts on GPT-2 Generated Texts with Supervised, Unsupervised Learning on Vaccine & Market Topics & Sentiment." *Journal of Big Data: Theory and Practice* **1**, no. 1. doi: 10.54116/jbdtp.v1i1.20

Sazzed, S. 2020. "Cross-Lingual Sentiment Classification in Low-Resource Bengali Language." *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*, Online, Association for Computational Linguistics, November 19. http://dx.doi.org/10.18653/v1/2020.wnut-1.8

Sazzed, S., and S. Jayarathna. 2019. "A Sentiment Classification in Bengali and Machine Translated English Corpus." *Proceedings of the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Los Angeles, CA, IEEE, July 30. https://doi.org/10.1109/IRI.2019.00029

Sebastian, M. P. 2023. "Malayalam Natural Language Processing: Challenges in Building a Phrase-Based Statistical Machine Translation System." *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**, no. 4: 1–51. doi: 10.1145/3579163

Tiedemann, J. 2020. "The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT." *Proceedings of the Fifth Conference on Machine Translation*, Online, Association for Computational Linguistics, November. https://aclanthology.org/2020.wmt-1.139

Tiedemann, J., and S. Thottingal. 2020. "OPUS-MT–Building Open Translation Services for the World." *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, European Association for Machine Translation, November 3–5. https://aclanthology.org/2020.eamt-1.61

Wang, S., Y. Sun, Y. Xiang, Z. Wu, S. Ding, W. Gong, S. Feng, *et al.* 2021. "Ernie 3.0 Titan: Exploring Larger-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation." Preprint, submitted December 23. https://doi.org/10.48550/arXiv.2112.12731

Wiesmann, E. 2019. "Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German." *Comparative Legilinguistics* **37**, no. 1: 117–153. doi: 10.14746/cl.2019.37.4