



**NAVAL
POSTGRADUATE
SCHOOL**

MONTEREY, CALIFORNIA

THESIS

**PREDICTING THE AUTHENTICITY
OF CODE-SWITCHED TEXT GENERATED
BY A LARGE LANGUAGE MODEL**

by

Lucas J. Horan

September 2023

Thesis Advisor:

Co-Advisor:

Second Reader:

Ruriko Yoshida

Jefferson Huang

Samuel E. Buttrey

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.			
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE September 2023	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE PREDICTING THE AUTHENTICITY OF CODE-SWITCHED TEXT GENERATED BY A LARGE LANGUAGE MODEL		5. FUNDING NUMBERS	
6. AUTHOR(S) Lucas J. Horan			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.		12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) Japan is a crucial partner in the U.S. Navy's effort to remain the premier naval force in an increasingly contested Indo-Pacific region. However, in the current era of generative technologies, like the large language model (LLM) Chat Generative Pre-trained Transformer (ChatGPT), malevolent actors worldwide now possess an unprecedented capability to generate text-based synthetic media able to sow disarray among allies. Consequently, alliances between the United States and its non-English speaking allies, like Japan, can be tested by text-based deepfakes seeking to reinforce their credibility by using the native languages of both countries; fabricated bilingual diplomatic statements, military communiqués, or news articles all possess the potential to upend U.S. global partnerships. Employing the tools of natural language processing (NLP), our research seeks to examine whether we can detect if bilingual text—that which may be created to undermine the relationship between the U.S. and Japan—is “authentic” (that is, human-made) or “inauthentic” (that is, generated by an LLM, namely ChatGPT). We achieved 96% accuracy in our limited trials using logistic regression, with similar results for support vector machine (SVM), k-nearest neighbor (KNN), and naive Bayes models, with each model presenting slightly different misclassifications.			
14. SUBJECT TERMS large language models, natural language processing, ChatGPT, Indo-Pacific, Japan, Japanese, code-switching		15. NUMBER OF PAGES 57	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**PREDICTING THE AUTHENTICITY OF CODE-SWITCHED TEXT
GENERATED BY A LARGE LANGUAGE MODEL**

Lucas J. Horan
Lieutenant Commander, United States Navy
BS, University of Pittsburgh, 2011

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
September 2023**

Approved by: Ruriko Yoshida
Advisor

Jefferson Huang
Co-Advisor

Samuel E. Buttrey
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Japan is a crucial partner in the U.S. Navy's effort to remain the premier naval force in an increasingly contested Indo-Pacific region. However, in the current era of generative technologies, like the large language model (LLM) Chat Generative Pre-trained Transformer (ChatGPT), malevolent actors worldwide now possess an unprecedented capability to generate text-based synthetic media able to sow disarray among allies. Consequently, alliances between the United States and its non-English speaking allies, like Japan, can be tested by text-based deepfakes seeking to reinforce their credibility by using the native languages of both countries; fabricated bilingual diplomatic statements, military communiqués, or news articles all possess the potential to upend U.S. global partnerships. Employing the tools of natural language processing (NLP), our research seeks to examine whether we can detect if bilingual text—that which may be created to undermine the relationship between the U.S. and Japan—is “authentic” (that is, human-made) or “inauthentic” (that is, generated by an LLM, namely ChatGPT). We achieved 96% accuracy in our limited trials using logistic regression, with similar results for support vector machine (SVM), k-nearest neighbor (KNN), and naive Bayes models, with each model presenting slightly different misclassifications.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
2	Background and Literature Review	5
2.1	Natural Language Processing	5
2.2	Corpora	6
2.3	Text Classification Preprocessing	7
2.4	Language and International Allies.	8
2.5	Japan’s Distinct Writing System	8
2.6	Sentence Order	11
2.7	Code-switching	11
3	Data and Methodology	15
3.1	Data Collection	15
3.2	Text Preprocessing.	16
3.3	Model Selection.	20
4	Results	23
4.1	Results	23
4.2	Discussion	24
5	Future Work and Conclusion	27
5.1	Different Classification Techniques	27
5.2	Language Variation	27
5.3	Preprocessing Refinement	28
5.4	Data Modification	29
5.5	Conclusion.	29
	List of References	31

List of Figures

Figure 1.1	Bilingual Synthetic Text Generated by ChatGPT. Source: ChatGPT (2023a)	2
Figure 2.1	Japanese Alphabetic Representations of “Moon”	9
Figure 3.1	ChatGPT Response to Request for Code-Switched Sentences Source: ChatGPT (2023b)	16
Figure 3.2	Foreign Loanword Error	17
Figure 3.3	Foreign Loanword Error, Corrected	17
Figure 3.4	Pronoun Label Error	18
Figure 3.5	Pronoun Label Error, Corrected	18
Figure 3.6	Digit Error	18
Figure 3.7	Digit Error, Corrected	18

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 4.1	Comparison of Test Results Using Four Different Text Classifiers .	23
-----------	--	----

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AI	Artificial Intelligence
ChatGPT	Chat Generative Pre-Trained Transformer
DOD	Department of Defense
JSDF	Japanese Self-Defense Force
JWS	Japanese Word Segmentation
KNN	K-Nearest Neighbors
LLM	Large Language Models
ML	Machine Learning
NDS	National Defense Strategy
NLP	Natural Language Processing
NHK	Nippon Hōsō Kyōkai
PLA	People’s Liberation Army
PRC	People’s Republic of China
RNN	Recurrent Neural Network
SOV	Subject + Other Details + Verb
SVO	Subject + Verb + Other Details
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Japan is a pivotal United States ally and arguably the United States' most important ally in the Indo-Pacific region. However, the current era of advanced information technology brings with it an ability to sow unrest among allies that is unprecedented: the mass availability of Chat Generative Pre-trained Transformer (ChatGPT) and other generative artificial intelligence (AI) tools arms malicious actors worldwide with advanced multilingual text generation capabilities that they can use to imitate representatives of other countries. The purpose of this work is to examine whether we can detect if bilingual text—that which may be used by someone seeking to undermine the relationship between the U.S. and Japan—is authentic (that is human-made) or inauthentic (that is, made by ChatGPT). This type of identification mechanism can help maintain U.S. alliances by safeguarding against heinous actors.

Given the broad spectrum of possible synthetic textual content, our research will focus on text that is purportedly from a bilingual speaker posing as a DOD (Department of Defense) or JSDF (Japanese Self-Defense Force) official who seeks to convey language mastery by changing languages between English and Japanese mid-sentence—otherwise known as intrasentential “code-switching.” We employed the tools of natural language processing (NLP) to assist us in meeting our objective. Both the author and one faculty advisor have studied Japanese extensively, making it a fitting choice for our research.

The basic methodology of our research involved procuring data, preprocessing the data so that it was ready for the experiment, and selecting models appropriate for the task. Given the limited availability of textual data available for this scenario, a corpus of textual language data was built containing intrasententially code-switched Japanese-English text.

Authentic data was created by deriving a basic Japanese sentence from simplified news articles written by the Japanese Broadcasting Corporation, otherwise known as Nippon Hōsō Kyōkai (NHK). The textual data point was then manually code-switched by a bilingual speaker to emulate natural code-switching conversation ultimately resulting in over 200 authentic code-switched sentences. Conversely, the inauthentic data was generated by a large language model (LLM), namely ChatGPT, version 3.5. A query was submitted to

ChatGPT in May 2023 requesting for code-switched Japanese-English text. For example, “Generate 10 code-switched Japanese-English sentences” was submitted as a basic prompt, with a few additional queries submitted thereafter to refine the results. This effort produced 300 inauthentic code-switched sentences.

The data was preprocessed using libraries available in Python that supported tokenizing and lemmatizing in Japanese and English, respectively. Tokenizing the data involved breaking each sentence into individual words. With lemmatizing, we reduced the number of forms of the same word thereby reducing data size and increasing processing speed. Both steps were completed in advance of performing the experiment.

Finally, models were selected to use in the effort. We picked logistic regression as our baseline model because it is a simple, generalized linear classification model. From there, we sought to deploy various models that have had success when leveraged for text classification tasks: support vector machine (SVM), k-nearest neighbors (KNN), and naive Bayes. The train set was composed of 399 data points, and the test set was composed of 101 data points.

As a basic text classifier, the logistic regression model performed remarkably well on this task, accurately classifying at a rate of 96 percent. Our second model, SVM was accurate at a rate of 98 percent, only misclassifying two data points: one false positive and one false negative. Interestingly, both the false positive and false negative were also misclassified by the logistic regression model. Our naive Bayes model performed at the second highest accuracy rate at 97 percent and did not produce any false positive misclassifications. Our KNN classifier did not perform quite as well as the SVM classifier, but still performed almost on par with the baseline logistic regression model. Our KNN model had the same number of false positives as our SVM model, but the most false negatives of any model.

Across all models, including duplicate sentences, there were 14 total incorrect classifications out of a total of 404, resulting in a successful classification rate of almost 97 percent. Further subdividing these results, there were five false positive and 9 false negative misclassifications. Of note, three of the four models were more apt to misclassify as a false negative than as a false positive.

Overall, we are pleased with how the four classification models we chose performed on this specific task. While our SVM model led the way with a near-perfect classification rate, the

other three models all performed above 95 percent. Even though this is a limited data set and very specific task, it gives us hope that follow-on work in this area may prove to be just as successful. When we analyzed specific sentence misclassifications, there were possible trends in syntax or proper noun use that may have caused the models to misclassify, but these areas require further research to confirm.

The novelty of LLMs and the minimal research on code-switching allows for an abundance of future research to be performed, including harnessing different binary classification techniques, performing the task with various languages, leveraging different techniques for data preprocessing, and modifying the data.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

This work came together through months of brainstorming with Dr. Ruriko “Rudy” Yoshida about the topic of code-switched language and its implications for international partnerships. A sincere thank you to Rudy for her willingness and excitement to work on this topic. I’d also like to thank Dr. Jefferson Huang who guided me through numerous technical roadblocks and helped me conceptualize many esoteric elements of this work. Similarly, I am indebted to Dr. Sam Buttrey, who provided countless thoughtful suggestions for revision to the draft thesis.

My father, aunt, and brother have all served to influence and support this project in one way or another, and I’m grateful to each of you. My grandparents, whose final resting place lies not far from the grounds of NPS, are likewise a source of frequent inspiration.

Ready with alacrity to conquer any hill on the Monterey Peninsula, my greyhound Jet deserves abundant praise for his steadfastness during my innumerable hours of writing.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1: Introduction

The most recent *National Defense Strategy* (NDS) released in 2022 prioritizes U.S. security concerns in the Indo-Pacific region. The document specifically references the People’s Republic of China’s (PRC) “coercive and increasingly aggressive” efforts to destabilize the region (Department of Defense 2022, p. 4). The NDS goes on to say that: “The PRC seeks to undermine U.S. alliances and security partnerships in the Indo-Pacific region, and leverage its growing capabilities, including its economic influence and the People’s Liberation Army’s (PLA) growing strength and military footprint, to coerce its neighbors and threaten their interests” (Department of Defense 2022, p. 4). The United States perhaps has no more important Indo-Pacific ally, in the face of Chinese belligerence, than Japan.

Dating back to a partnership formed following Japan’s defeat by the Allies in World War II and the signing of the “Treaty of Mutual Cooperation and Security Between Japan and the United States of America” in the years after, Japan represents a vital part of the U.S. effort to counterbalance attempted Chinese hegemony in the Indo-Pacific region. In a “Joint Statement of the United States and Japan” released on January 13, 2023, these partner nations affirmed an “unbreakable bilateral relationship” and declared that the current “landscape demands the United States and Japan continue to strengthen our individual and collective capacity” (White House 2023). Together, the United States and Japan avowed that their security relationship represents “the cornerstone of peace, security, and prosperity of the Indo-Pacific” (White House 2023). With the U.S. Seventh Fleet based in Yokosuka, near the mouth of Tokyo Bay, and several other key U.S. military assets stationed throughout Japan, the United States is poised to maintain a robust relationship with the Japanese government and Japanese Self-Defense Force (JSDF) into perpetuity.

However, the current era of advanced information technology brings with it an ability to sow unrest among allies that is unprecedented: the mass availability of Chat Generative Pre-trained Transformer (ChatGPT) and other generative artificial intelligence (AI) tools arms malicious actors worldwide with advanced multilingual text generation capabilities that they can use to imitate representatives of other countries. In March of 2021, the

U.S. Federal Bureau of Investigation warned that “malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations” (Federal Bureau of Investigation 2021). The following year, Russia created an artificial video of Ukrainian President Volodymyr Zelenskyy ostensibly surrendering to Russia in the Russia-Ukraine War (Allyn 2022). Moreover, pro-Chinese operators are leveraging synthetic content to influence opinions toward the U.S. government (Graphika 2023, p. 2).

Thus far, the alliance between the U.S. and Japan has not yet been publicly tested by way of divisive synthetic content. However, with the development of the chatbot ChatGPT and other widely available generative AI tools, it is now easier than ever to do so. ChatGPT is a tool built on a collection of large language models (LLM) trained on massive amounts of text derived via the internet (Wilson 2023). It has the power to generate human-like multilingual text in response to prompts. For example, as depicted in Figure 1.1, it can create bilingual text posing as a public figure when given the prompt: “Please make a short statement from a Department of Defense (DOD) official about strained ties between the Japanese and U.S. Defense Secretaries. Make the statement bilingual in Japanese and English.”

"We deeply regret to inform that the U.S. and 日本の防衛大臣 (Japanese Defense Secretary) have reached a point where 我々は防衛関係を断絶 (we are severing defense relations). This is an unfortunate turn of events that affects both the 米国 (United States) and Japan. We have worked for years to 构建强大的联盟 (build a strong alliance), and it's disheartening to see those efforts come to an impasse. Our doors remain open for future 協力 (cooperation), but as of now, formal ties have been cut."

Figure 1.1. Bilingual Synthetic Text Generated by ChatGPT. Source: ChatGPT (2023a)

While ChatGPT was not built for this purpose, because it is now easily accessible, anyone in the world with access to this application can pose as a bilingual speaker or post bilingual content, including content intended for social engineering or other malicious purposes. Consequently, alliances between America and its non-English speaking allies like Japan can be tested by malevolent actors releasing false bilingual diplomatic statements, military

orders, or simply posing as a public figure to create mass confusion among the general public. Information warfare has always been present in war, but the reach of the current era’s misinformation capability is by far the greatest it has ever been.

However, there has already been a response to help counteract and unmask potential false information created by ChatGPT. For example, over concerns about AI plagiarism in academia, a student at Princeton University created an application to detect essays created by ChatGPT (Bowman 2023). Our research seeks to follow this lead by examining whether we can detect if bilingual text—that which may be used by someone seeking to undermine the relationship between the U.S. and Japan—is authentic (that is, human-made) or inauthentic (that is, made by ChatGPT). This type of identification mechanism can help maintain U.S. alliances by safeguarding against heinous actors.

Given the broad spectrum of possible synthetic textual content, our research will focus on text that is purportedly from a bilingual speaker posing as a DOD or JSDF official who seeks to convey language mastery by changing languages between English and Japanese mid-sentence—otherwise known as intrasentential “code-switching” (defined further in section 2.7). We will employ the tools of natural language processing (NLP) to assist us in meeting our objective. Both the author and one faculty advisor have studied Japanese extensively, making it a fitting choice for our research.

The structure of this thesis is as follows: in Chapter 2, we furnish background information and explore scholastic literature relevant to our work. In Chapter 3, we introduce our data set and review model selection. In Chapter 4, we examine and analyze our results. In Chapter 5, we conclude and survey possibilities for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2: Background and Literature Review

This section provides background information and explores scholastic literature relevant to our work.

2.1 Natural Language Processing

NLP is a subset of AI and machine learning (ML) that deals with leveraging the tools of modern computation to extract information from text or spoken words that would otherwise be laborious or impossible for a human to accomplish. Prior to the 1990s, NLP primarily involved “writing out symbolic rules of grammar. . . trying to replicate grammar and lexicons” (Manning 2017) with little application beyond this narrow scope. Then, the data revolution of the 1990s occurred during which, for the first time in history, massive amounts of text data became available via the internet. Manning (2017) underscores the importance of this pivotal revolution in access to text, noting that computers learn about the world through “the enormous amount that has already been written about the world. Every day, writers across the planet are writing about our world and how it all works. We create computational models that assign mathematical values to words and groups of words and use them to successfully read text and derive meaning.”

Due to the greater availability of text as well as the increase in processing power over the past several decades, NLP capabilities have increased dramatically. Computer scientists have constructed techniques that enable computers to process text and provide meaningful output in response to text or spoken inputs. Modern applications of NLP include spam detection, machine translation (e.g. Google Translate), virtual agents (e.g. Apple’s Siri), chatbots (e.g. ChatGPT), and text summarization, to name a few (IBM 2023). Yet, because of language complexity both within and between languages, NLP still remains an imperfect field. Computers are constrained or flummoxed by things like homonyms, idioms, metaphors, and sarcasm that humans often understand from context or experience (IBM 2023).

LLMs, such as ChatGPT, represent the vanguard of current NLP tools. Like humans, LLMs are a reflection of their teaching environment—LLMs trained on well-refined prose “speak”

much more eloquently than those trained on social media text data, like Twitter (Andersen 2023). There also is a finite quantity of high-quality human-manifested prose to train LLMs, and recent research suggests LLMs will exhaust this material by the year 2027 (Andersen 2023). These limitations do not currently inhibit the advancement of LLMs, however; introduced just this year, ChatGPT version 4 possesses a training set supposedly exceeding 10 trillion words and the capability to write entire novels (Andersen 2023). Our research employs the current publicly available versions of ChatGPT, version 3.5 and version 4.

2.2 Corpora

Corpora are massive compendiums of language data that provide the foundation necessary to perform many NLP tasks. Corpora can be monolingual or multilingual, and they can include textual data, spoken data, or both. Corpora are utilized in NLP text analytics tasks for things like topic detection, sentiment analysis, and machine translation (Kibble 2013, p. 33). Our work necessitates the use of a bilingual, code-switched Japanese-English corpus.

While the thirst for corpora has increased with the advent of LLMs, high quality corpora remain difficult to find. In the realm of spoken corpora used for speech recognition technologies, English and Chinese are the only two languages for which many hours of spoken corpora are available, and the most common Japanese spoken corpus is only slightly over half the size of its English counterpart (Takamichi et al. 2021). Takamichi et al. (2021) crafted a blueprint for creating an expanded Japanese corpus by compiling over 1300 hours of spoken data from YouTube, but the scope of this data is limited to monolingual speech recognition tasks.

Nakayama et al. (2018a) identified the need for a robust Japanese-English corpus given the rise in bilingual speakers and the need for training data to support bilingual speech recognition systems. More Japanese children are being born with at least one foreign parent and more Japanese students are studying abroad, setting the backdrop for increased code-switching phenomena in everyday life in Japan (Nakayama et al. 2018a). Since gathering bilingual code-switched dialogue could be costly or time-consuming, Nakayama et al. (2018a) identified a method utilizing data derived from a bilingual speaker and a text-to-speech system.

Nakayama et al. (2018a) finalized a code-switched corpus that included over 280,000 code-switched Japanese-English utterances. Although our efforts are focused on text, and we cannot heavily rely upon their spoken corpus, Nakayama et al.'s (2018a) research sketches an outline for building a bilingual, code-switched Japanese-English corpus. Inspired by their efforts, we utilized basic text phrases and applied translation to intrasententially code-switch the text. We discuss our corpus construction more in Chapter 3.

2.3 Text Classification Preprocessing

In advance of performing our classification task, we sought to bolster the tidiness and effectiveness of our corpus by refining our text data through morphological preprocessing. Preprocessing is a pivotal step in text classification because it mitigates confounding features of the text by, for example, correcting word misspellings and performing acronym disambiguation (HaCohen-Kerner et al. 2020). It is important to shape our preprocessing methods to our corpora since the value of a particular preprocessing technique depends on the type of text it is applied to (HaCohen-Kerner et al. 2020).

HaCohen-Kerner et al. (2020) reference numerous studies that vary preprocessing techniques and classification model types seeking to find the best task-dependent combination; several of the studies they reference are 15 years or older so it would be beneficial to redo the experiments using technology available to modern NLP practitioners. In their own experiment, HaCohen-Kerner et al. (2020) explored the accuracy of their classification task using blends of six basic preprocessing techniques and comparing the results to model accuracy without any preprocessing; their work exemplifies the value of attempting different preprocessing methods to test the impact on classification accuracy.

In our research, we opted to employ two advanced preprocessing techniques: tokenization and lemmatization. Tokenization involves breaking a string of text into individual units, either words or characters. For example, the sentence “The current is strong today” would be subdivided into: *the*, *current*, *is*, *strong*, and *today*. Lemmatization is a morphological technique that changes tokens into their basic “dictionary” forms, otherwise known as lemmas. By lemmatizing, it is possible to reduce the number of forms of the same word thereby reducing data size and increasing processing speed. For example, “is,” “was,” and “were” are all represented as “be” in a lemmatized corpus. Using lemmas shears flotsam

from the data. In our work, tokenization was essentially accomplished simultaneously with lemmatization, so moving forward we will refer to our preprocessing technique only as “lemmatization.”

2.4 Language and International Allies

It is crucial for DOD to be able to communicate with partner nations clearly and effectively at all levels of warfare: strategic, operational, and tactical. Language is a fundamental aspect of appropriate engagement with these international allies and regional partners. It also presents a challenge when the languages are vastly different. For example, during NATO medical exercises—most of the 31 member states speak different languages—body language or gestures are used to communicate (Schroeder 2018). While NATO has established English and French as its official languages (Schroeder 2018), it is helpful that many of the countries in the alliance share syntactically similar languages. For example, countries like Norway, Sweden, and the Netherlands have languages in the Germanic family like English. Moreover, the Romance languages have many similarities to English that make these language pairs relatively close in syntactic structure. Unlike NATO, however, Japan is both culturally and linguistically quite distant from the United States. Our work examines the capability of a malicious actor to exploit this distance.

2.5 Japan’s Distinct Writing System

Written language is generally subdivided into phonographic writing systems and ideographic writing systems. Phonographic writing systems are those whose symbols are linked to speech sounds, such as the Roman alphabet, while ideographic writing systems are those whose symbols are associated with ideas, such as Japanese kanji (Hasegawa 2015, p. 43). Modern written Japanese combines both ideographic symbols and phonographic syllabary and thus “involves unparalleled complexity” (Hasegawa 2015, p. 46).

Modern written Japanese is comprised of three distinct alphabets: hiragana, katakana, and kanji. Both hiragana and katakana are phonographic syllabaries that contain 46 characters denoting sounds or syllables; these two syllabaries are virtually identical in terms of the character sounds and differ only in their application. Hiragana is typically used for words

native to Japan or colloquial expressions, while katakana is often used for foreign loanwords or onomatopoeia.

Japanese kanji, on the other hand, are a collection of thousands of ideographic characters imported from the Asian mainland in the first few centuries of the common era along with rice agriculture, military technology, and religious ideas (Gordon 2020, p. 7). Each kanji possesses three fundamental attributes: shape, sound, and meaning (Halpern 2002, p. 17a). The shape of each kanji is distinct as is the stroke order used to make it; although, many kanji are homoglyphs of one another and appear nearly identical. With respect to sound, many kanji characters possess both an imported Chinese “on-yomi” reading and a native Japanese “kun-yomi” reading (Halpern 2002, p. 17a). Regarding meaning, every kanji character has at least one core meaning, but can boast several.

While each of these alphabets are distinct, there is significant interaction between them. Every kanji can be depicted with hiragana or katakana, and the readings of kanji are learned using hiragana. Moreover, the shapes of both the hiragana and katakana alphabets were originally derived from specific kanji shapes (Hasegawa 2015, p.47). Thus, there are echoes of kanji throughout the other two. Figure 2.1 seeks to underscore the complexity and inherent redundancy of the Japanese writing system by displaying the three Japanese alphabetic representations of the English word “moon.”

Hiragana: つき

Katakana: ツキ

Kanji: 月

Figure 2.1. Japanese Alphabetic Representations of “Moon”

Sampson (1985) comments on the origins of the complexities of written Japanese:

Japanese society, during much of the period in which the [hiragana and katakana were] developed, was characterized by the existence of an aristocratic class many members of which lacked political power or indeed any serious employ-

ment, so that their only role in life was as definers and producers of cultural norms, ways of civilized living. . . As a natural result many aspects of Japanese culture, including its writing, were greatly elaborated, made exquisite and intellectually rich rather than straightforwardly functional. (p. 172-73)

A unique facet of written Japanese that does not occur in English is the presence of particles. Particles include hiragana characters は, を, が, に, で, と, and others that mark a specific relationship with other Japanese parts of speech.

For example, this sentence uses three different particles:

昨日、僕は友達と和食を食べました。(Yesterday, I ate Japanese food with a friend.)

は – Marks the sentence topic “僕” or “I”

と – “With,” in this case “with a friend” or “友達”

を – “direct object” marker that indicates what the sentence topic or subject is acting upon

As seen above, some particles possess literal equivalencies in English, while others merely help provide context or semantic structure to the Japanese.

Furthermore, unlike English, written Japanese is unsegmented; that is, there is no spacing between words. Japanese word segmentation (JWS) is a process accomplished during preprocessing that inserts spacing between words and, when completed inaccurately, can create issues performing NLP tasks (Kitagawa and Komachi 2017). Thus, it is critical for us to choose preprocessing tools that will effectively perform JWS. Several approaches have been explored in JWS including constructing a word lattice and utilizing pointwise prediction (Kitagawa and Komachi 2017). JWS is similar to tokenization in that discrete units are separated prior to being used in the task. We discuss in Chapter 3 the specific Python library we employ to help complete our preprocessing, including effective JWS.

At its core the Japanese writing system is intricate, which adds an additional layer of complexity when assessing an ostensibly official statement’s credibility. This effort becomes even more daunting when that statement is made in the two languages of the partner nations, i.e. Japanese and English. Yet, this type of complexity is what excites us and motivates us when approaching this task.

2.6 Sentence Order

Apart from the writing system, sentence order represents the other marked difference between English and Japanese that has the potential to significantly confound NLP tasks that use both. Namely, Japanese is a verb-final language while English is a verb-medial language. For example, the general sentence pattern of each language reflects the following:

Japanese sentence structure - Subject + Other Details + Verb (SOV)

English sentence structure - Subject + Verb + Other Details (SVO)

For example:

Japanese: 昨日、僕は(S)友達と和食(O)を食べました(V)。

English: Yesterday, I (S) ate (V) Japanese food with a friend (O).

Distinct SOV and SVO languages each comprise about 40 percent of the languages throughout the world (Hahn and Xu 2022). This ratio hints that languages appearing in naturally occurring code-switched utterances may not exhibit the same inherent sentence order. Hahn and Xu (2022) suggest that languages evolved into one of the two groupings based largely upon whether users sought to keep together syntactically associated or contextually helpful language units. More research is needed to ascertain what these findings imply for code-switching using languages derived from different sentence order groupings. It is promising, however, that past research has unveiled phenomena unique to code-switching between SOV and SVO languages (Yoon 1992). For our task, this discovery implies that code-switched utterances may indeed embody an entirely unique language structure.

2.7 Code-switching

Our work seeks to analyze the validity of a sentence that will move between two languages, otherwise known as “code-switching.” Code-switching is a term that refers to a string of language, either typed or spoken, that alternates between languages within or between sentences (Nakayama et al. 2018a). There is no set rhythm or pattern for code-switching, but research indicates it is more than a random phenomenon (Nakayama et al. 2018a). This lack of a pattern introduces additional complexity when seeking to perform NLP.

Nakayama et al. (2018b) provide a broad overview of Japanese-English code-switching research and discuss some of the modern NLP challenges posed by code-switched Japanese-English utterances. Despite the presence of code-switching in daily life, Nakayama et al. (2018b) underscore the limited availability of spoken language technologies in support of code-switching and that more must be done to address this shortfall: “Since people may not always communicate in monolingual settings, spoken language technologies... must be developed that can handle the input in a multilingual fashion, not only Japanese or English but also Japanese-English [code-switching]” (p. 182). This need supports our thinking that DOD-level multinational interactions would benefit from this type of advance in code-switching technologies.

Taura (1996) explored Japanese-English code-switching in several different settings and concluded that “code-switching is deliberately and strategically used as one of the effective communication tools of bilinguals irrespective of their ages or linguistic preferences.” This insight suggests for our research that the “deliberate” or “strategic” use of code-switching might be a hallmark of human-generated code-switching, while machine-generated code-switching may be less strategic and more random.

Humans and computers “acquire” language in different ways. The human acquisition method of language is rooted in our biology: by leveraging the ability of the mouth to articulate the vibrations and resulting sounds from the vocal chords, humans engage in a “closed-loop speech chain” mechanism in which the ears receive input that enables learning to happen through repeated exposure and surrounding context (Nakayama et al. 2018b, p. 183). Matheson (2018) discusses how computers typically learn language through “sentences annotated by humans that describe the structure and meaning behind words.” However, researchers are seeking to develop a reliable “weakly supervised” approach that emulates how children acquire language (Matheson 2018). For our research, this development implies that the current state of LLMs may produce text in a more artificial way than they might in the future, and the task of classifying machine-generated text should be frequently revisited.

There are two general types of code-switching: intersentential and intrasentential (Nakayama et al. 2018a). The primary difference involves where the code-switching occurs. Intersentential code-switching occurs at the sentence level while intrasentential switching is done mid-sentence (Nakayama et al. 2018a). For example:

Intersentential:

“お腹が空いたよ. Can we stop for lunch?” = “I’m hungry. Can we stop for lunch?”

Intrasentential:

“その映画have you seen it?” OR “That movie 見たことがある？” = “Have you seen that movie?”

Conceptually, on the intersentential level, the shift occurs between language units that possess complete semantic meaning. Put another way, a speaker finishes a sentence in one language before shifting to a different language. Both “I’m hungry.” and “Can we stop for lunch?” could occur independently and retain their meaning and semantic relevance.

Conversely, on the intrasentential level, the switch happens between language units that often do not possess complete semantic meaning on their own. For example, both “その映画” and “That movie” fail to retain semantic value in the absence of the complete syntactic structure. Our research will focus on intrasentential code-switching because of its inherent complexity.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Data and Methodology

This chapter provides an overview of the methodology we utilized in seeking to construct a text classification model able to effectively identify code-switched Japanese-English text generated by an LLM. In order to successfully perform our classification task, we needed to collect data, perform data preprocessing, and select a suitable model. We address each of these areas in this chapter.

3.1 Data Collection

Given the limited availability of code-switched data, it was necessary for us to construct a unique code-switched Japanese-English corpus for this research. This data set was built with two distinct parts: “authentic” data and “inauthentic” data.

Authentic data is so called because it originated from human sources. A basic Japanese sentence was derived from simplified news articles written by the Japanese Broadcasting Corporation, otherwise known as Nippon Hōsō Kyōkai (NHK). The textual data point was then manually code-switched by the author to emulate natural code-switching conversation ultimately resulting in over 200 authentic code-switched sentences. The author is a native English speaker and a second language speaker of conversational Japanese. According to Taura (1996), “there is an interaction of situations and metaphorical meanings intended by interlocutors that render the [Japanese-English] code-switching meaningful” (p. 3). Our authentic data sought to preserve this essence of code-switching.

Conversely, the inauthentic data was generated by an LLM—namely ChatGPT, version 3.5. A query was submitted to ChatGPT in May 2023 requesting for code-switched Japanese-English text. For example, “Generate 10 code-switched Japanese-English sentences” was submitted as a basic prompt, with a few additional queries submitted thereafter to refine the results. ChatGPT used its own training data to generate the code-switched response, a sample of which is shown in Figure 3.1. Ultimately, this effort resulted in several hundred inauthentic code-switched Japanese-English sentences.



Certainly! Here are ten code-switched Japanese-English sentences without translations:

1. 昨日のパーティーで、友達とdancingを楽しんだ。
2. 明日の予定はbusyだけど、eveningにはfriendsとdinnerに行く予定だ。
3. 今日はコンビニでお弁当を買って、parkでpicnicしよう。
4. 新しいanimeを見つけた！今、English subtitlesでwatchしているところだ。
5. 今週のスケジュールはfullで、英語のmeetingと日本語のプレゼンが入っている。
6. このrestaurantのramenは本当にdeliciousだから、みんなにrecommendしたい！
7. 週末はshoppingに行く予定だ。新しいdressとshoesを探している。
8. イギリスへのvacationを計画している。LondonとEdinburghを訪れる予定だ。
9. 彼のパフォーマンスは本当にamazingだ！彼は日本語とEnglishの歌を披露する。
10. 暇なときにはNetflixでJapanese dramasをbinge-watchするのが好きだ。

Note: Code-switching refers to the practice of alternating between two or more languages or language varieties in a conversation or sentence. In these sentences, both Japanese and English words are used interchangeably to create a code-switched effect.

Figure 3.1. ChatGPT Response to Request for Code-Switched Sentences
Source: ChatGPT (2023b)

Each data point was assigned a value of “0” or “1” based on whether it was authentic or not. “0” was used for authentic data while “1” was used for inauthentic data. This is a helpful process when dealing with two discrete categories of textual data in an NLP classification task.

3.2 Text Preprocessing

To prepare the data for processing by a text classification model, we distilled the sentences into individual meaningful elements using the process of lemmatization. Since we used lemmatization tools that functioned on one language at a time, our first preprocessing step involved splitting the code-switched sentences into their respective Japanese and English sentence fragments. Once separated, we were able to begin the step of linguistically reshaping the sentence fragments.

3.2.1 Japanese Preprocessing

We performed Japanese lemmatization using “MeCab,” an open-source tool for performing Japanese morphological analysis (Hagiwara and McCann 2023), and “UniDic,” a detailed linguistic dictionary created by the National Institute for Japanese Language and Linguistics (McCann 2019). If UniDic is a set of directions for taking a word from “form A” to “form B,” MeCab is the engine that takes it there. “Fugashi” is the wrapper we utilized for MeCab.

Fugashi presented a few technical issues that were necessary for us to address in preprocessing. First, once lemmatized, Fugashi would mistakenly pair a katakana loanword from English with the English word itself, exhibited in Figure 3.2. We corrected this error by removing both the hyphen and English word, which would have otherwise confounded our data. The corrected katakana is depicted in Figure 3.3.

カレー-curry
カード-card
ピザ-pizza

Figure 3.2. Foreign Loanword Error

カレー
カード
ピザ

Figure 3.3. Foreign Loanword Error, Corrected

Next, we addressed an error in which, after lemmatization, any Japanese pronoun would return paired with the Japanese word for pronoun, “代名詞”, displayed in Figure 3.4. Retaining this extra information, ostensibly incorporated to identify the lemma as a pronoun, would confound our data since we wanted our dictionary to integrate only the lemma itself, not any linguistic tag. As with the loanword error, we removed both the hyphen and pronoun identifier to remedy the issue, as evidenced in Figure 3.5.

私-代名詞

僕-代名詞

君-代名詞

Figure 3.4. Pronoun Label Error

私
僕
君

Figure 3.5. Pronoun Label Error, Corrected

Finally, lemmatized digits were eliminated entirely and replaced with the word “None.” By eliminating digits, sentences can become incoherent. We corrected this error by ensuring that the digits were retained. As demonstrated in Figure 3.6 and Figure 3.7, “お菓子を5個食べた” (“I ate 5 sweets”) initially manifested without the digit “5” before being corrected.

御
菓子
を
None
個
食べる
た

Figure 3.6. Digit Error

御
菓子
を
5
個
食べる
た

Figure 3.7. Digit Error, Corrected

Below is an example of a Japanese sentence fragment before and after preprocessing. Of note, each character is separated into individual components, overcoming the challenge of JWS discussed in Chapter 2:

Before: どこまで屋根があるか見えなくなります。

After: 何処 まで 屋根 が 有る か 見える ない 成る ます。

3.2.2 English Preprocessing

The English preprocessing was more straightforward, requiring limited intervention to complete. A basic NLP library known as “TextBlob” was employed to perform English lemmatization. Since TextBlob is relatively simple, it lacks some of the nuance in capability that higher-level libraries possess. For example, TextBlob struggled with contractions like “I’m” and present participle verbs ending in “-ing.” To rectify this issue, we added a part-of-speech tagger to the preprocessing function. In the case of present participles, for example, it was important that they be labeled as a verb so that TextBlob would not confuse them for a gerund (i.e. noun) and leave them unlemmatized. Once the part-of-speech tagger was added, the English lemmatization was completed.

3.2.3 Final Preprocessed Data

Once both the Japanese and English sentence fragments were lemmatized, they were merged together to mirror their original order in the unprocessed sentence. By completing lemmatization, we effectively simplified our dictionary while simultaneously narrowing its scope to the most meaningful elements. Below is a single data point displayed before and after preprocessing was completed. Note the completion of JWS on the Japanese text and the lemmatization of the English verbs “exploring” and “are”, which became “explore” and “be,” respectively.

Before: I enjoy exploring 日本の庭園, they are so serene and peaceful.

After: I enjoy explore 日本の庭園 they be so serene and peaceful

3.3 Model Selection

We picked logistic regression as our baseline model because it is a simple, generalized linear classification model. From there, we sought to deploy various models that have had success when leveraged for text classification tasks. We opted to use a support vector machine (SVM) classifier (Cristianini and Ricci 2008) as our second model because they are known to be highly generalizable and adaptable (Liu et al. 2010). Moreover, they been shown to outperform other possible classifiers when performing a text classification task (Liu et al. 2010).

The third model we selected was a multinomial naive Bayes model (Raschka 2014). A multinomial naive Bayes model has been shown to succeed at a consistent rate whether the vocabulary size is low or high, beating out other classifiers in this regard (McCallum et al. 1998). This prior success irrespective of vocabulary size is beneficial for our scenario given our limited data. We opted to use a naive Bayes model because the results should be repeatable even in the absence of large amounts of code-switched textual data.

Lastly, we opted to use the k-nearest neighbor (KNN) algorithm (Guo et al. 2003) for our fourth and final model. KNN has been used in monolingual text settings to classify demanding languages like Arabic with a modest success rate (Al-Shalabi and Obeidat 2008). Extending KNN to classify in a code-switched setting of Nepalese-English and Spanish-English has shown marked success and achieved accuracy at nearly the same level as an SVM model (Barman et al. 2014). The work by Barman et al. (2014) is limited to word-level classification, but we believed these results could be duplicated at the sentence level.

The three models besides our baseline logistic regression model vectorized their data—that is, transformed their input data from text to numbers—utilizing the term frequency - inverse document frequency (TF-IDF) approach. TF-IDF increases the importance of a word depending on its rarity in the text and is especially helpful when common words like “the” or “and” are not removed in preprocessing (Raschka 2014), as is the case for our data.

Our data preprocessing code and classification model code were built using the Python coding language and run on a MacBook Pro using an Apple M1 chip and running macOS Monterey 12.6.8. For data preprocessing, we used the Fugashi and TextBlob Python libraries.

We leveraged ChatGPT, version 4.0, to assist in the final build of the data preprocessing code.

All models were built and run in a Jupyter Notebook. The initial test logistic regression model was built utilizing scikit-learn (Pedregosa et al. 2011), NumPy, Pandas, and PyTorch (Paszke et al. 2019) libraries. ChatGPT, version 4.0, was employed to help build and refine a second logistic regression model. All of the same Python libraries were used with the exception of PyTorch; we utilized PyTorch for our vectorization in the initial test model, but utilized scikit-learn for this step in our second model. Our final logistic regression model was sourced from Van Otten (2023) with the exception of the data loading and confusion matrix output which we integrated from our second model.

Our final SVM model utilized the basic classification template from Van Otten (2023) while leveraging the final logistic regression model's data loader and output code. ChatGPT, version 4.0, assisted in the vectorization of the data for our SVM model. Our naive Bayes model is nearly identical to our SVM model, but it incorporates a different classifier from scikit-learn for training the model. Our KNN model was built using ChatGPT, version 4.0.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4: Results

Our experiment consisted of training our models on about 80 percent of our data points, then using the remaining data to test the accuracy of our models.

4.1 Results

The results of our four experiments are shown in Table 4.1 with columns displaying final classification accuracy (in percent out of 100) as well as the breakdown of each classification type. False positive (“False +”) refers to an occurrence when the classifier predicted an authentic data point (“0”) but it was actually inauthentic (“1”). False negative (“False -”) refers to an occurrence when the classifier predicted an inauthentic data point when it was actually authentic. True positive (“True +”) and true negative (“True -”) represent occurrences when the classifier correctly classified the text as authentic, for True +, or inauthentic, for True -.

Table 4.1. Comparison of Test Results Using Four Different Text Classifiers

	Accuracy	False +	False -	True +	True -
Logistic Regression	96%	3	1	33	64
SVM	98%	1	1	33	66
Naive Bayes	97%	0	3	31	67
K-NN	95%	1	4	30	66

As a basic text classifier, the logistic regression model performed remarkably well on this task, accurately classifying at a rate of 96 percent. Our second model, SVM, was accurate at a rate of 98 percent, only misclassifying two data points: one false positive and one false negative. Interestingly, both the false positive and false negative were also misclassified by the logistic regression model. Our naive Bayes model performed at the second highest accuracy rate at 97 percent and did not produce any false positive misclassifications. Our

KNN classifier did not perform quite as well as the SVM classifier, but still performed almost on par with the baseline logistic regression model. Our KNN model had the same number of false positives as our SVM model, but the most false negatives of any model.

4.2 Discussion

Overall, we are pleased with how the four classification models we chose performed on this specific task. While our SVM model led the way with a near-perfect classification rate, the other three models all performed above 95 percent. Even though this is a limited data set and very specific task, it gives us hope that follow-on work in this area may prove to be just as successful.

Across all models, including duplicate sentences, there were 14 total incorrect classifications out of a total of 404, resulting in a successful classification rate of almost 97 percent. Further subdividing these results, there were five false positive and 9 false negative misclassifications. Of note, three of the four models were more apt to misclassify as a false negative than as a false positive.

Honing in on some of the specific misclassifications, we noticed that one sentence in particular was misclassified by all four models: “Hawaii and Korea だ行く 人が 多い 様 です” (“It seems there are many people going to Hawaii and Korea.”). Every model produced this sentence as a false negative; in other words, every model thought this was a data point generated by ChatGPT when it was actually one of our authentic data points.

One possible reason is that the models recognized a code-switching pattern in this authentic data point that aligned with the patterns it recognized in the ChatGPT data. However, many of the authentic data points mirror this same pattern in terms of having only a single code-switch and that code-switch occurring somewhere near the middle of the sentence. Moreover, the ChatGPT data was much more likely to possess a code-switching pattern that did not resemble the sentence in question; the ChatGPT data could contain three or more code-switches while an authentic data point contained at most two.

Next, it is possible that the sentence syntax suggested to the models that it had been generated by ChatGPT. Yet, nothing about the syntax as rendered bilingually suggests that it would be unusual as compared to the other authentic data points. The ending “が 多い 様 です”

appears in other authentic data points while not occurring in any of the inauthentic data.

It could also be the presence of nouns unique to this data point. “Hawaii and Korea,” as a pair of nouns, might as well be rendered in their katakana and kanji forms (ハワイ and 韓国), or replaced by any other location name, given that their impact to retaining the overall semantic meaning and syntactic structure is minimal. Future research would need to be conducted to more accurately determine the cause of this problem.

Both our KNN and naive Bayes models misclassified, as false negatives, two other sentences in common: “Dotonbori River は 逆も 賑やかだ 所に 有る ます” and “色々だ 料理 が 有る て、 and they be also sell in vend machine.” In the first sentence, we see another example of a proper noun unique to one data point beginning a misclassified sentence. While not necessarily a harbinger of misclassification, it is noteworthy that six of the total nine false negatives were in the pattern of [proper English noun] + [Japanese phrase].

The second sentence mentioned above is notable because the adjective “vending” that describes the type of “machine” doing the selling is actually treated as a present participle verb and the “-ing” is excised by our lemmatizer. We noted in Chapter 3 the rudimentary nature of the English lemmatizer we chose, and perhaps this sentence misclassification is a result of a failed lemmatization. It would be helpful in future work to explore if more advanced English lemmatizers resulted in a similar misclassification.

On the false positive side, one sentence was misclassified both by the logistic regression model and the SVM model: “私の family は 日本に 住む て 居る ます” (“My family is living in Japan”). A possible explanation for this misclassification is that this sentence shares many of the same features of other authentic data points, including two or fewer code-switches and use of polite Japanese.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5: Future Work and Conclusion

The novelty of LLMs and the minimal research on code-switching allows for an abundance of future research to be performed in this area. This section discusses some of the areas that can be addressed in future research, including harnessing different binary classification techniques, performing the task with various languages, leveraging different techniques for data preprocessing, and modifying the data.

5.1 Different Classification Techniques

One area that still needs to be addressed for the basic task is the use of additional binary classification techniques to compare technique performance on the assigned task. While we sought to incorporate several different classification models, there are other options available that can be explored. Random forest models would be an excellent option to test. Random forest models are commonly used in classification tasks and are effective even when the data is imbalanced (Jalal et al. 2022). Moreover, they have performed competitively in text classification tasks when pitted against other classifiers (Jalal et al. 2022). Recurrent neural networks (RNN) would also be an interesting avenue to explore as they also have shown tremendous success in text classification (Hu et al. 2020).

5.2 Language Variation

Another area of future work involves incorporating languages into this task other than English and Japanese. It is important to test the performance of different classifiers on a wide-range of languages. Possible code-switched combinations to explore include looking at two SOV or two SVO languages; looking at languages within the same etymological family, such as two Romance languages or two Germanic languages; and looking at two languages that use the same writing alphabet but are markedly different like English and Indonesian. Additionally, an analogous study to ours could be completed on English and another SOV language with a complex writing system, such as Hindi, to assess similarities and differences in the results. A spoken Hindi-English code-switched corpus was created

previously to address challenges of performing NLP in bilingual settings (Sreeram et al. 2018), so that corpus could be utilized in future work.

Fortunately, the digital age brings with it access to virtually every language on the planet. Precedent exists for performing NLP tasks by pruning web data in an array of varied languages; research has previously been completed by Grave et al. (2018) using Wikipedia and additional web sources to build word vectors for over 157 languages. Wikipedia is an especially appealing resource for NLP tasks on a diverse range of languages given the volume of tokens it possesses (Grave et al. 2018).

It is also worthwhile to explore this task using solely formal or informal textual Japanese. Our task utilized a mix of both formal and informal Japanese, which can possess significant differences in syntax, word use, and expression. Building data exclusively in one or the other would be helpful to look at different types of scenarios. Informal text will primarily emulate conversations while formal text will primarily emulate written Japanese.

5.3 Preprocessing Refinement

It is also important to look at different data preprocessing techniques. We prepared our data for the models by using Python NLP libraries available like Fugashi for Japanese and Textblob for English. However, there are numerous libraries available, especially for English, that may provide slightly better quality or accuracy when it comes to tokenizing or lemmatizing the data. HaCohen-Kerner et al. (2020) aver in their work that “it is always recommended to perform an extensive and systematic variety of preprocessing methods combined with [text classification] experiments because it contributes to improve [text classification] accuracy” (p. 20).

Research that analyzed various tokenization strategies for a Korean-English textual scenario found that the best tokenization technique depended on the specific type of NLP task to be performed (Park et al. 2020). We can extend this logic to our own work and reasonably assert that the best tokenization technique for Japanese, English, and code-switched Japanese-English depends on the specific task to be performed. Thus, there is significant value in exploring different preprocessing options to assess their contributions toward improved results.

5.4 Data Modification

It is important to retry this task with a larger data set. While procuring inauthentic data merely requires additional prompting of the source LLM, as has been previously noted, authentic data is more difficult to come by. One possible avenue for retrieving large amounts of authentic code-switched data is social media. Both Twitter and YouTube possess large amounts of untapped, mostly informal, text to select from. That being said, this raises the question if the comment section on these websites might resemble the text from an authentic code-switching conversation. Moreover, as Nakayama et al. (2018b) note, “Although it might still be possible to find a sufficient amount of only [code-switching] text or [code-switching] speech in social media, unfortunately, parallel speech and transcription of [code-switching] data are mostly unavailable that are suitable for training [automatic speech recognition] and [text-to-speech systems]” (p. 182). In other words, while the data may be available on social media, it might not be in a form that is suitable for performing various NLP tasks.

It would also be beneficial to generate the inauthentic data from different LLMs. The LLMs available for public use at the time of this research are trained on different data sets, so we would expect them to respond with slight variance when prompted for code-switched text. Moreover, it will be important to replicate this task over-time as the size of the LLM training data continues to grow with each new version release.

5.5 Conclusion

This work was intended as an initial attempt at understanding the immense capabilities that LLMs possess in terms of convincingly imitating complex human speech. LLMs are a disruptive new technology that provide tremendous benefit for society but also enable our adversaries to imitate us in ways that were never before possible. It is crucial that as an organization we continue to study this field and come up with techniques of detecting fake, LLM generated content quickly and reliably. It is up to the United States to ensure that we never allow malevolent actors to sow unrest or discord between the United States and her allies.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

- Al-Shalabi R, Obeidat R (2008) Improving KNN Arabic text classification with n-grams based document indexing. *Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt*, 108–112 , <https://api.semanticscholar.org/CorpusID:62110843>.
- Allyn B (2022) Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn. NPR. Accessed June 21, 2023, <https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia>.
- Andersen R (2023) What happens when AI has read everything? The Atlantic. Accessed June 21, 2023, <https://www.theatlantic.com/technology/archive/2023/01/artificial-intelligence-ai-chatgpt-dall-e-2-learning/672754/>.
- Barman U, Wagner J, Chrupała G, Foster J (2014) DCU-UVT: Word-level language classification with code-mixed data. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 127–132 (Doha, Qatar: Association for Computational Linguistics), <https://doi.org/10.3115/v1/W14-3915>.
- Bowman E (2023) A college student created an app that can tell whether AI wrote an essay. NPR. Accessed June 21, 2023, <https://www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism>.
- ChatGPT (2023a) ChatGPT prompt response dated 26 August 2023. Accessed August, 26, 2023, <https://www.openai.com/research/chatgpt>.
- ChatGPT (2023b) ChatGPT prompt response dated 28 August 2023. Accessed August, 28, 2023, <https://www.openai.com/research/chatgpt>.
- Cristianini N, Ricci E (2008) *Support Vector Machines*, 928–932 (Boston, MA: Springer US), https://doi.org/10.1007/978-0-387-30162-4_415.
- Department of Defense (2022) National Defense Strategy. Accessed May 5, 2023, <https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF>.
- Federal Bureau of Investigation (2021) Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations. Accessed June 21, 2023, <https://www.ic3.gov/Media/News/2021/210310-2.pdf>.
- Gordon A (2020) *A Modern History of Japan* (Oxford University Press, New York).

- Graphika (2023) Deepfake it till you make it: Pro-Chinese actors promote AI-generated video footage of fictitious people in online influence operation. Accessed June 21, 2023, <https://public-assets.graphika.com/reports/graphika-report-deepfake-it-till-you-make-it.pdf>.
- Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. *CoRR* abs/1802.06893, <https://doi.org/10.48550/arXiv.1802.06893>.
- Guo G, Wang H, Bell D, Bi Y, Greer K (2003) KNN model-based approach in classification. Meersman R, Tari Z, Schmidt DC, eds., *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 986–996 (Berlin, Heidelberg: Springer Berlin Heidelberg).
- HaCohen-Kerner Y, Miller D, Yigal Y (2020) The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE* 15:1–22, URL <http://dx.doi.org/10.1371/journal.pone.0232525>, <https://doi.org/10.1371/journal.pone.0232525>.
- Hagiwara M, McCann PO (2023) *Introduction to Japanese Natural Language Processing*. 60% complete as of March 15, 2023. <https://www.leanpub.com/japanesenlp>.
- Hahn M, Xu Y (2022) Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Psychological and Cognitive Sciences* <https://doi.org/10.1073/pnas.2122604119>.
- Halpern J (2002) *The Kodansha Kanji Learner's Dictionary* (Kodansha USA, New York).
- Hasegawa Y (2015) *Japanese: A Linguistic Introduction* (Cambridge University Press, Cambridge).
- Hu H, Liao M, Zhang C, Jing Y (2020) Text classification based recurrent neural network. *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 652–655, <https://doi.org/10.1109/ITOEC49072.2020.9141747>.
- IBM (2023) What is natural language processing (NLP)? Accessed June 25, 2023, <https://www.ibm.com/topics/natural-language-processing>.
- Jalal N, Mehmood A, Choi GS, Ashraf I (2022) A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University - Computer and Information Sciences* 34(6, Part A):2733–2742, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2022.03.012>.
- Kibble R (2013) Introduction to natural language processing. Accessed June 25, 2023, <https://www.london.ac.uk/sites/default/files/study-guides/introduction-to-natural-language-processing.pdf>.

- Kitagawa Y, Komachi M (2017) Long short-term memory for Japanese word segmentation. *CoRR* abs/1709.08011, <https://arxiv.org/abs/1709.08011>.
- Liu Z, Lv X, Liu K, Shi S (2010) Study on SVM compared with the other text classification methods. *2010 Second International Workshop on Education Technology and Computer Science*, volume 1, 219–222, <https://doi.org/10.1109/ETCS.2010.248>.
- Manning C (2017) How computers are learning to understand language. Stanford. Accessed on August 11, 2023, <https://biox.stanford.edu/highlight/how-computers-are-learning-understand-language>
- Matheson R (2018) Machines that learn language more like kids do. MIT. Accessed August 5, 2023, <https://news.mit.edu/2018/machines-learn-language-human-interaction-1031>.
- McCallum A, Nigam K, et al. (1998) A comparison of event models for naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*, volume 752, 41–48 (Madison, WI), <https://api.semanticscholar.org/CorpusID:7311285>.
- McCann PO (2019) Japanese tokenizer dictionaries. Dampkraft. Accessed on June 5, 2023, <https://www.dampkraft.com/nlp/japanese-tokenizer-dictionaries.html>.
- Nakayama S, Kano T, Do QT, Sakti S, Nakamura S (2018a) Japanese-English code-switching speech data construction. *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments*, 67–71, <https://doi.org/10.1109/ICSDA.2018.8693044>.
- Nakayama S, Tjandra A, Sakti S, Nakamura S (2018b) Speech chain for semi-supervised learning of Japanese-English code-switching ASR and TTS. *2018 IEEE Spoken Language Technology Workshop (SLT)*, 182–189, <https://doi.org/10.1109/SLT.2018.8639674>.
- Park K, Lee J, Jang S, Jung D (2020) An empirical study of tokenization strategies for various Korean NLP tasks. *CoRR* abs/2010.02534, <https://doi.org/10.48550/arXiv.2010.02534>.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

- Raschka S (2014) Naive bayes and text classification I - introduction and theory. *CoRR* abs/1410.5329.
- Sampson G (1985) *Writing Systems* (Stanford University Press, Stanford).
- Schroeder B (2018) Soldiers bridge language, cultural barriers at Combined Resolve Exercise. Accessed on June 25, 2023, <https://www.defense.gov/News/News-Stories/Article/Article/1514673/soldiers-bridge-language-cultural-barriers-at-combined-resolve-exercise/>.
- Sreeram G, Dhawan K, Sinha R (2018) Hindi-English code-switching speech corpus. *CoRR* abs/1810.00662, <https://doi.org/10.48550/arXiv.1810.00662>.
- Takamichi S, Kürzinger L, Saeki T, Shiota S, Watanabe S (2021) Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification. *CoRR* abs/2112.09323, <https://arxiv.org/abs/2112.09323>.
- Taura H (1996) A comparative study of Japanese/English bilingual code-switching in three different domains.
- Van Otten N (2023) How to implement logistic regression text classification in Python with scikit-learn and PyTorch. Spot Intelligence. Accessed September 2, 2023, <https://spotintelligence.com/2023/02/22/logistic-regression-text-classification-python/>.
- White House (2023) Joint statement of the united states and japan. Accessed May 5, 2023, <https://www.whitehouse.gov/briefing-room/statements-releases/2023/01/13/joint-statement-of-the-united-states-and-japan/>.
- Wilson M (2023) ChatGPT explained: Everything you need to know about the AI chatbot. TechRadar. Accessed June 21, 2023, <https://www.techradar.com/news/chatgpt-explained>.
- Yoon K (1992) New perspective on intrasentential code-switching: A study of Korean-English switching. *Applied Psycholinguistics* 13(4):433–449, <https://doi.org/10.1017/S0142716400005750>.

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California



DUDLEY KNOX LIBRARY

NAVAL POSTGRADUATE SCHOOL

WWW.NPS.EDU

WHERE SCIENCE MEETS THE ART OF WARFARE