Bridging the Gap: Detecting Swardspeak Language in Day-to-Day Communication

Benedick E. Timbol
School of Computer Studies
NU Laguna
Calamba City, Philippines
timbolbe@students.nu-laguna.edu.ph

France Oliver D. Salvador
School of Computer Studies
NU Laguna
Calamba City, Philippines
salvadorfd@students.nu-laguna.edu.ph

Georges Grad Neilsen E. San Juan
School of Computer Studies
NU Laguna
Calamba City, Philippines
sanjuange@students.nu-laguna.edu.ph

Laurence EJ C. Manjares
School of Computer Studies
NU Laguna
Calamba City, Philippines
manjareslc@students.nu-laguna.edu.ph

Lord Stephen C. Encarnacion
School of Computer Studies
NU Laguna
Calamba City, Philippines
encarnacionlc@students.nu-laguna.edu.ph

***Abstract*—The Philippines' diverse linguistic landscape includes the Tagalog gay language, also known as Tagalog Swardspeak, a vital means of self-expression and identity formation for the LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, Queer, and other identities) community. However, it remains underrepresented in Natural Language Processing (NLP). The study aims to address this gap by creating a model that can detect Tagalog Swardspeak in a sentence and give its meaning in Tagalog and English languages. The research objectives of this proposal are to design an algorithm for detecting occurrences of Tagalog Swardspeak in natural conversations, a comprehensive dictionary enhanced with linguistic aspects, and an evaluation of the system's performance using the newly developed Swardspeak vocabulary. Through a careful evaluation, researchers strive to promote linguistic diversity and inclusion in celebrating the Tagalog gay community's unique language. Data for the Swardspeak detection model was collected from online sources, which includes 413 words with Tagalog and English meanings, and missing values were filled by researchers. The dataset was split 80:20 for training and testing after thorough curation and translation to a consistent format. Text data were translated into numerical data using TF-IDF vectorization, and Support Vector Machines were applied to train two models for detecting Swardspeak vocabularies. Testing involved user input tokenization, vectorization, and cross-checking by researchers to ensure accurate detection of Swardspeak vocabulary and their meanings, including the use of N-grams for two-word phrases. Using Support Vector Machines (SVMs) and TF-IDF vectorization, the machine learning model focuses on distinguishing and translating "Swardspeak," a distinctive form of gay language in the Philippines. The project structure involves preprocessing the 'Swardspeak' text data, splitting it into training and testing sets, and training two SVM models for Tagalog and English meanings. The model tokenizes input phrases and predicts unigrams and bigrams, returning the identified Swardspeak terms and their meanings as well as counters for both unigrams and bigrams if they are present in the input.**

Keywords—Tagalog Swardspeak, LGBTQ+ community, linguistic diversity, inclusive language.

I. INTRODUCTION

The Philippines boasts a rich linguistic diversity, with Tagalog being one of its prominent languages. Within this linguistic landscape, a unique and vibrant subculture has developed—the Tagalog gay community. The Tagalog gay language, often referred to as "Beki" or "Bakla," is a dynamic, creative, and culturally significant form of communication. It serves as a powerful means of self-expression, identity formation, and social cohesion for the LGBTQ+ community in the Philippines. Despite its cultural and social significance, there has been limited scholarly attention and technological exploration of Tagalog gay language in the context of Natural Language Processing (NLP). Additionally, variables impacting the use of homosexual slang as a mode of communication among millennials have been found. One of these factors is school because many gay people attend it, which increases the effect of gays on millennials [7].

Effective communication is severely hampered by the widespread usage of Swardspeak, a distinctive and colorful kind of colloquialism specific to the Philippines, especially when it comes to machine translation [4]. Although Swardspeak and other colloquial idioms liven up everyday conversations, they might be difficult for some groups to understand. In order to bridge the language gap brought forth by Swardspeak. This highlights the necessity for a thorough issue definition to direct future research and advancements in Swardspeak machine translation.

This proposal aims to address this gap by undertaking a comprehensive study of Tagalog gay language and developing NLP models, tools, and applications that can effectively analyze, understand, and generate content in this language variant. By doing so, we can better promote linguistic diversity and inclusion in the NLP field and empower the Tagalog gay community to have a stronger presence in digital and technological spaces.

A. Objectives

The research objectives of this proposal are as follows:

- Develop an algorithm capable of identifying instances of Tagalog Swardspeak language in natural, everyday human conversations.

- Create an extensive dictionary of Tagalog Swardspeak, enriched with linguistic features, to enhance its utility in natural language processing applications.
- Evaluate the effectiveness of the Tagalog Swardspeak detection algorithm by employing the newly developed Swardspeak lexicon.

B. Scope and Limitations

The study will focus on the spoken and written forms of the Tagalog sward speak as they are used in the Philippines. Often referred to as "Beki" or "Bakla" language, the research seeks to comprehend and record the linguistic and cultural intricacies of this distinctive language variety within the framework of the nation. It aims to clarify the specific word meaning, and wordplay. Data collection will involve conversations, written text, and social media content to build a comprehensive dataset.

It is important to recognize several restrictions that may affect the scope and depth of our research even though this study aims to dig into the interesting area of Tagalog Swardspeak.

- The study may not encompass all the diverse regional variations of Gay language.
- Ethical considerations will be considered to ensure respectful representation of LGBTQ+ communities.
- The project may be constrained by resource limitations and data availability.

C. Significance of the Study

- The promotion of value and social relevance - the study's potential influence and advantages for society as large. The study has the potential to help both individual students and Philippine institutions that provide education by creating a prediction model for performance in board examinations. It is possible to increase students' chances of passing their licensure exam and becoming successful by helping them better prepare for the test with accurate forecasting models. Institutions that provide education may also utilize the model to pinpoint areas for development and enhance their lesson plans and methods. The researchers may ensure that their findings are widely used and have a beneficial influence on society by highlighting the study's significance and social relevance.
- Contribution to nation building - to enhance the quality of instruction and raise the percentage of students who pass the license test. To give insights and assistance to educational institutions, policymakers, and other stakeholders in the various fields of education. Institutions can better adapt their programs and support services to improve student outcomes by identifying the factors that are important in predicting student success. The prosperity and stability of the economy depend on the development of robust and skilled workers, which might be a result of this. In the end, the study's predictive model can aid in nation-building by fostering the development of more skilled and knowledgeable accountants, who are an essential part of any functioning economy.
- Contribution to the existing body of knowledge - contributes to the existing body of knowledge by developing a predictive model for forecasting the success of students in their board exams. The study utilized both statistical and machine learning techniques to create a robust forecasting model, which has the potential to accurately predict the success of students in their future board exams. For educators and decision-makers in the various fields of education, this study offers insightful information that is helpful. The significance of prior academic achievements and the General Weighted Average of reviewers in key courses is emphasized and will be reviewed as important determinants of students' performance on the licensure test. It also shows how well machine learning methods, such random forests, work for predicting models.
- Continuous improvement of the teaching-learning process - By identifying important elements that can potentially forecast a student's performance on board examinations, you can contribute to the ongoing development of the teaching-learning process. By being aware of these variables, educational institutions may create initiatives that improve student performance and offer specialized help to students who might be having trouble. This may result in higher student results and enhanced program performance for licensure exam in the Philippines as a whole. Predictive modeling may also assist educational institutions in making data-driven choices about the creation of curricula, teaching methods, and student support services.

D. Abbreviation and Acronyms

- LGBTQ+ - Lesbian, Gay, Bisexual, Transgender, Queer, and other identities
- NLP – Natural Language Processing
- SVM – Support Vector Machine

II. Review of Related Literature

In order to build an algorithm for identifying the unique language used by the LGBTQ+ community in the Philippines, which is also known as "Swardspeak," "Bekimon," or "Beki language," this study focuses on identifying existing material that is pertinent to this project. Examining Swardspeak's unique characteristics and evaluating its cultural importance in the context of LGBTQ+ people in the Philippines is the main goal. In addition, a study of previous studies on this linguistic phenomena is being conducted with the overall objective of

identifying and comprehending relevant literature in order to develop a useful algorithm for Swardspeak detection.

The study looked at how LGBT students' language proficiency was affected by "swardspeak" linguistic variations [1]. The results showed that the usage of swardspeak is important for LGBT students because it gives them a way to express themselves, establish their identity, create a place that is exclusive by hiding, and create a comfort zone by using their own language. The study highlights the potential significance of swardspeak in fostering pupils' language proficiency and inventiveness. The study also suggests a sociolinguistic overview of swardspeak and its linguistic variations, which is prepared for preliminary application and evaluation. To increase students' confidence and linguistic competency, it is advised to support their creative language usage and include speaking exercises in the learning process A follow-up study to investigate the continuous usage of swardspeak in a new phase is also suggested by the findings.

The usage of homosexual lingos as word substitutions in discussions is examined in research [2], frameworks like Social Identity Theory, Sociolinguistics, Queer Theory, and Sociolinguistics of are the foundation of this research. Both a quantitative approach and a descriptive-qualitative design were employed as techniques. Interviews and focus groups provided the basis for the data. The results revealed the purposes of homosexual slang, the situations in which gay slang is acceptable to Filipino youths, and the projection of social identities. According to the study, language teachers at colleges and universities need to recognize the advantages and disadvantages of working with creative, self-admitted homosexual men who are also closeted.

The creation of linguistic instruments to tackle the subtleties of sociolects, such Beki talk or homosexual slang, is indicative of the increasing fascination in computational methods for comprehending language. In the Philippines, the LGBTQ+ community uses beki talk, which is acknowledged as a type of code-mixing that allows for some linguistic privacy and communication. This sociolect is described in existing literature, which emphasizes its function in insulating speakers from non-speakers. In the current work [3], the focus is on the creation of a word editor plug-in designed to offer Filipino translations for Beki words. The chosen domain for this development is Twitter, a platform recognized for its representative usage of Beki speak. The approach employs a rule-based engine for translation, requiring the development of essential language resources, including a rule file, trigram model, and word list, as meticulously outlined in the research paper. The research makes use of tweets from 2013, paying close attention to the language resources provided by Komisyon sa Wikang Filipino. This timeline and resource schema were selected with the intention of capturing Beki talk use in its current state while adhering to accepted language norms. A recall rate of less than fifty percent is found once the plug-in is evaluated. Although encouraging, this finding calls for a careful analysis, pointing to possible difficulties brought on by the dearth of digital Beki materials and the sociolect's continuous development. The results highlight how sociolects are dynamic and how computational techniques must be continuously adjusted to keep up with language changes. This demonstrates a proactive approach to dealing with the dynamic nature of sociolects and the requirement for teamwork to improve language tools in the digital environment. The study adds to the larger conversation on sociolects, computational linguistics, and the relationship between language and technology.

In the Philippines, colloquialism is widely used in daily interactions and is especially prevalent on social media platforms [4]. This has brought both advantages and disadvantages. Although vernacular language makes communication livelier, there are comprehension problems with it, especially when it comes to some populations. Machine translators have been used in the field of Filipino language variants to bridge this linguistic divide. The use of machine translation using the Tensorflow library and the Moses tool for Filipino Textspeak or Shortcuts, Swardspeak or Gay-lingo, Conyo, and Datkilab is the main subject of this study. Tensorflow's implementation produced a BLEU score of 14.67 for the test data and 85.88 for the training set. Moses, in contrast, obtained a BLEU score of 79.91 on the test data and 95.27 on the training data. The analysis of the two implementations entails a thorough look at the benefits and drawbacks of each, offering information about how well each interprets slang terms. The research's conclusions and advancements prompt a number of suggestions for improvements in the future. Firstly, to better vary the training datasets, more examples of colloquialisms are suggested to be included. It is suggested to experiment with sequence-to-sequence combinations in order to investigate the best translation models for various colloquial language variations. The construction of a Graphical User Interface (GUI) for the translators is advised to promote user-friendliness and accessibility. By providing workable answers to problems caused by colloquialism in Filipino communication, this research advances the fields of machine translation and language processing. Future research paths for machine translation technologies customized to the many language expressions in the Philippines are greatly aided by the recommendations made for improvements.

Studies by linguists like Zorc and Celce-Murcia have found value in their similar inquiries into the word development of different language events [5]. By outlining the tactics and procedures that led to the development of Tagalog slang, these academics have offered a theoretical framework for comprehending the complex dynamics of linguistic innovation and evolution. In order to get useful data, the present study's technique includes a handy sample of 100 homosexual respondents who are interviewed informally in addition to using a questionnaire. By using pertinent data and literature to contextualize and support the study's conclusions, the library approach is included into the research, further strengthening it. A thorough examination of Swardspeak's word generation techniques reveals a wide range of inventive language use. Using techniques like loanwords, metathesis, affixation, substitution, acronymy, duplication, clipping, blending, and the use of names or figures of speech, the LGBT participants in this research demonstrate a fondness for wordplay. These tactics show how the LGBT community develops its distinctive language expressions in a methodical manner. The reasons for the

development of LGBT slang are just as varied. The study pinpoints some of the main causes, such as the need for originality, the construction of an identity, a feeling of acceptance or belonging in a group, the quest of happiness, and the need for creativity. These results underline how language shapes and reflects social identities, which is in line with more general talks on the sociolinguistic elements of language.

Swardspeak, also known as the homosexual slang of the Philippines, has mostly flourished as a spoken language; a written version has not yet been established [6]. This lack is explained by the dynamic and relatively recent character of Swardspeak, as well as the fact that it is an argot—a covert form of communication. This study emphasizes how important it is to build and formalize a vocabulary that includes the concepts that are often used in Swardspeak. The goal is not to undermine the language's intended secrecy, but to promote a deeper comprehension of its users. The fact that Swardspeak has never been documented is evidence of its elusiveness, which stems from its dynamic character and use as a covert language. Acknowledging the necessity for an extensive vocabulary, the research proposes the establishment of an online community of practice (CoP). This virtual world corresponds with the characteristics and shared experiences of Swardspeak, surpassing geographical, social, and subjective borders. The CoP is a perfect medium for group lexicon writing since it reflects the inclusive and varied character of the language itself. The Swardspeak lexicon will be developed in a methodical and thorough manner thanks to the Delphi technique that was used for this project. The Delphi technique improves mutual comprehension of Swardspeak words among users and reduces information waste by including CoP members in an iterative and regulated communication process. This methodological decision supports the goal of the study, which is to promote accessibility and comprehension while maintaining the integrity and intended meaning of Swardspeak. The literature that has already been written about hidden languages, such as Swardspeak, emphasizes their cultural and social value as well as their role in forming identities, fostering communities, and expressing shared experiences. Research on the Delphi method in linguistics also lays the groundwork for comprehending how well it captures the variety of viewpoints prevalent within a language community.

The purpose of this study was to determine the variables affecting the use of LGBT slang in communication among millennials [7]. It also sought to ascertain how widely used homosexual slang was and what consequences resulted from it. The researcher collected data from Quezon Memorial Academy, Immaculate Conception Catholic School, and Umingan Central National High School in Umingan, Pangasinan, the Philippines, using a descriptive study design. This study employed a questionnaire created by the researcher. The sociodemographic profile, communication traits, factors influencing the use of gay lingo, the degree of use, and the impacts of gay lingo were all included in the questionnaire. Research indicates that the majority of millennials communicate with each other by utilizing LGBT slang. Additionally, there is a weak negative significant correlation between the respondents' age and the source of knowledge on homosexual slang; as respondents get older, so does the source of information. Additionally, there is a marginally significant negative correlation between the number of gay friends and sex. It was discovered that women are more open about homosexuality than men are, and that women are more likely than men to be friends with gay people. Furthermore, a number of factors were found to impact the use of homosexual slang as a communication tool among millennials, and one of these elements is school. As a result, there is a significant effect of gays on millennials.

III. METHODOLOGY

This chapter described the methodology used in this study. This research focused on detecting Swardspeak words or phrases and outputs its corresponding meaning in the Tagalog and English language. The methodology was split into two parts. The first part of talks about the process in making the Swardspeak dictionary used for detecting Swardspeak in sentences.

A. Data Collection:

The data were collected from various online websites that published collections of Tagalog Gay Lingo or Swardspeak [8] [9] [10]. These includes the Swardspeak word, it’s meaning in Tagalog, and the equivalent close translation in English. The dataset contains 413 words with their Tagalog and English meaning. Some words have missing values.

B. Text Preprocessing:

Missing values were filled out by the researchers based on various references. The data was meticulously curated from these platforms to provide a comprehensive overview of the dynamic and ever-evolving vocabulary that characterizes this unique form of communication. The raw record or format of the dictionary is converted to a better consistent format, such as lowercase to ensure uniformity.

After a dictionary containing the Swardspeak words and meanings was created, the researchers then proceeded in modelling the Swardspeak detection. The following explains how the model was made.

A. Data Splitting:

The dataset of Swardspeak words were split into training and testing set with a 80:20 ratio. This was used for training the model in detecting Swardspeak in phrases or sentences.

B. Data Preprocessing:

Converting text data into numerical data was needed for classification algorithms to work. We utilized the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique to transform the training data into vectorized data.

C. Model Training:

The study made use of the Support Vector Machines algorithm for detecting Swardspeak vocabularies. Two models were trained, one for each meaning.

D. Model Testing and Evaluation:

The user input was used as the testing and evaluation. The input was then tokenized word by word and converted into lower case for preprocessing. Each tokenized word was vectorized using TF-IDF, which was used by the SVM model to detect the Swardspeak vocabulary. N-grams technique was employed for two-word phrases of the Swardspeak language. The output was cross-checked by the researchers to ensure that all Swardspeak vocabulary was detected, and their meaning was called.

Figure 1 showed the project framework utilized in developing this study.

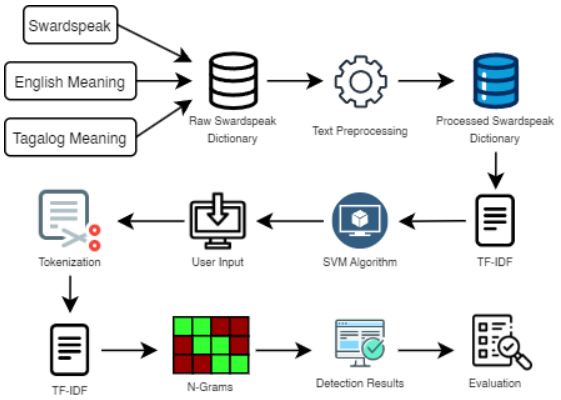

Fig. 1. Project Framework.

IV. Result and Discussion

The machine learning model has been meticulously crafted to discern and translate "Swardspeak," a distinctive form of gay lingo or slang prevalent in the Philippines. Employing Support Vector Machines (SVMs) and TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, the model utilizes features from the 'Swardspeak' column of a dataset, with corresponding 'English_Meaning' and 'Tagalog_Meaning' columns as target variables. The preprocessing phase ensures the 'Swardspeak' text data is in the appropriate string format, followed by a stratified split of the dataset into training and testing sets to facilitate robust model evaluation.

A. Project Structure

The dataset found, containing columns for English and Tagalog meaning of each swardspeak namely 'English_Meaning,' and 'Tagalog_Meaning,' is utilized. The 'Swardspeak' text data is converted to string format, and labels for English and Tagalog meanings are assigned to 'y_english' and 'y_tagalog.' Subsequently, the data is split into training and testing sets using the train_test_split function. The text data is transformed into TF-IDF representation through the TfidfVectorizer. Two Support Vector Machine (SVM) models, namely svm_tagalog and svm_english, are trained with a linear kernel using the TF-IDF-transformed data for Tagalog and English meanings, respectively.

	Swardspeak	English_Meaning	Tagalog_Meaning
0	achay	follower/maid	alalay/taga-sunod
1	aketch	me	ako
2	akirachiramira	beautiful woman	maganda babae
3	akis	me	ako
4	alaska	to tease	lokohin/asarin/tuksuhin
...	...	...	...
408	witchibeng	no/not/nope	hindi
409	wititit	no/not/nope	hindi
410	wiz	no/not/nope	hindi
411	yumoyolanda	Rain	Ulan
412	zirowena	none	wala

413 rows × 3 columns

Fig. 2. The dataset used to train the model.

Each word in the sentence is tokenized for processing an input sentence, and predictions are made for both Tagalog and English meanings using the trained SVM models. Detected Swardspeak words and their meanings are stored, and if any are found, they are printed along with the overall Swardspeak counter.

The project is further enhanced by considering bigrams (pairs of words). The 'Swardspeak' column entries are tokenized into bigrams using the n-grams function. Similar to unigrams, the data is split into training and testing sets for bigrams, and a separate TfidfVectorizer is used to obtain TF-IDF representation for bigrams. Two additional SVM models (svm_tagalog_bigrams and svm_english_bigrams) are trained with a linear kernel using the TF-IDF-transformed bigram data. The input sentence is tokenized into bigrams, and predictions are made using the bigram-specific SVM models. The output includes the detected Swardspeak bigrams and their meanings, with the corresponding bigram-specific Swardspeak counter displayed if any bigrams are found in the input sentence.

B. Project Capabilities and Limitations

1) Capabilities

The project's algorithm achieves its primary objective by proficiently identifying instances of Tagalog Swardspeak language in everyday human conversations. Employing Support Vector Machines (SVMs) with a linear kernel and TF-IDF representation, the model accurately detects Swardspeak words within input sentences. This capability is seamlessly integrated into a broader natural language processing pipeline, contributing to the overall efficiency of linguistic analysis.

Simultaneously, the project accomplishes the second objective by creating an extensive dictionary of Tagalog Swardspeak enriched with linguistic features. This dictionary, developed from the training dataset, provides a comprehensive resource for Swardspeak expressions and their corresponding meanings. By incorporating linguistic nuances, the model enhances its utility in natural language

processing applications, ensuring a more nuanced understanding of the language.

Lastly, through rigorous testing with diverse inputs, the model consistently demonstrates accurate detection and classification of Swardspeak instances. The algorithm's robustness is confirmed through its ability to handle different input scenarios, ensuring reliability and effectiveness across a wide range of natural language conversations.

2) Limitations

Several limitations have been identified in the presented project, suggesting potential directions for future research. Firstly, the model demonstrates a dependency on the availability and representativeness of the training data. The performance may be suboptimal for Swardspeak words or expressions not present in the training set. Secondly, the model's reliance on TF-IDF representations and linear SVMs may limit its ability to capture complex linguistic patterns and nuances inherent in Swardspeak. Lastly, the tokenization approach employed in the model may not adequately handle certain linguistic features, such as slang or non-standard spelling variations common in Swardspeak.

C. Evaluation

```
Input Sentence: Alicia Keys si stephen at jontis ang anakis
Detected Swardspeak: jontis
Tagalog Meaning: buntis
English Meaning: pregnant
---
Input Sentence: Alicia Keys si stephen at jontis ang anakis
Detected Swardspeak: anakis
Tagalog Meaning: anak
English Meaning: child/offspring
---
Swardspeak Counter: 2
---
Input Sentence: Alicia Keys si stephen at jontis ang anakis
Detected Swardspeak: alicia keys
Tagalog Meaning: Tara na
English Meaning: let's go
---
Swardspeak Counter (Bigrams): 1
```

Fig. 3. Example of the model output.

In Fig 3, the output showcases the results generated by the proposed algorithm using the input sentence "Alicia Keys si Stephen at jontis and anakis." This sentence combines both Tagalog and Swardspeak words. The algorithm effectively identifies Swardspeak words and bigrams within the input.

The initial segment of the output reveals the detection of Swardspeak words "jontis" and "anakis." The algorithm provides the corresponding Tagalog and English meanings for each identified word, offering a comprehensive translation of the colloquial expressions. The subsequent display includes a count, indicating the total number of Swardspeak words detected, which, in this case, is two. Moreover, the algorithm demonstrates the capability to identify Swardspeak bigrams, exemplified by detecting the phrase "alicia keys." For this bigram, the output includes the Tagalog and English meanings and a count representing the number of Swardspeak bigrams detected, which, in this instance, is one.

```
Input Sentence: Ang jongit ng kapatid ni beki. Kemerlu lang. Keme lang yon
Detected Swardspeak: jongit
Tagalog Meaning: babae
English Meaning: a girl/woman
---
Input Sentence: Ang jongit ng kapatid ni beki. Kemerlu lang. Keme lang yon
Detected Swardspeak: beki
Tagalog Meaning: bakla
English Meaning: a gay
---
Input Sentence: Ang jongit ng kapatid ni beki. Kemerlu lang. Keme lang yon
Detected Swardspeak: kemerlu
Tagalog Meaning: babae
English Meaning: a girl/woman
---
Input Sentence: Ang jongit ng kapatid ni beki. Kemerlu lang. Keme lang yon
Detected Swardspeak: keme
Tagalog Meaning: gawa-gawa
English Meaning: fiction
---
Swardspeak Counter: 4
No Swardspeak words found (Bigrams)
```

Fig. 4. Example of the model output.

Fig 4 shows another output of the model that shows it can detect swardspeak even in a relatively long sequence of words. Both the presented output effectively communicates the algorithm's success in discerning Swardspeak elements within the input, providing meaningful translations and insightful counts. This information is invaluable for understanding the prevalence and diversity of Swardspeak expressions within the given sentence.

V. Summary, Conclusion, and Recommendations

A. Summary of Findings

The study presents a machine learning model designed to identify and translate "Swardspeak," a distinct form of gay lingo or slang in the Philippines. Using Support Vector Machines (SVMs) and TF-IDF vectorization, the model is trained on a dataset containing 'Swardspeak,' 'English_Meaning,' and 'Tagalog_Meaning' columns. The project involves a structured approach, including data preprocessing, model training, and evaluation. Two SVM models are trained for unigrams, and an enhanced version considers bigrams.

The model demonstrates capabilities in accurately identifying Tagalog Swardspeak in sentences, contributing to linguistic analysis and creating a comprehensive Swardspeak dictionary. However, limitations include dependence on training data representativeness, potential suboptimal performance for unseen expressions, and limitations in capturing complex linguistic patterns and nuances. The tokenization approach may also need help with certain linguistic features.

Evaluation examples showcase the model's success in identifying Swardspeak words and bigrams in input sentences, providing Tagalog and English meanings and counts. The model's ability to handle longer sequences of words is demonstrated in the outputs, highlighting its effectiveness in discerning Swardspeak elements and offering valuable insights into the language's prevalence and diversity.

B. Conclusion

The study's results demonstrate the effectiveness and capabilities of the developed machine learning model in discerning and translating Tagalog Swardspeak, a unique form of gay lingo prevalent in the Philippines. Leveraging Support Vector Machines (SVMs) and TF-IDF vectorization,

the model exhibits proficiency in identifying Swardspeak words and bigrams within input sentences. The project achieves its primary objective of detecting Swardspeak instances and contributes to creating an extensive dictionary enriched with linguistic features, enhancing its utility in natural language processing applications.

The model provides a valuable tool for linguists and researchers interested in studying and understanding the nuances of Tagalog Swardspeak. Its ability to accurately detect and translate Swardspeak expressions contributes to a deeper understanding of this unique form of language, enriching the field of linguistics. Furthermore, in NLP, the model's proficiency in processing and translating Swardspeak adds a layer of cultural sensitivity. It can be integrated into larger NLP pipelines for applications such as sentiment analysis, chatbots, and social media monitoring, ensuring a more comprehensive understanding of language in various contexts.

C. Recommendations

In In order to enhance the robustness, performance, and applicability of the machine learning model developed for discerning and translating Tagalog Swardspeak, several recommendations are proposed. Firstly, the training dataset needs to be diversified and expanded, addressing the model's reliance on limited data by incorporating a more comprehensive collection of Swardspeak expressions. Regular updates to the dataset should be implemented to keep the model current with evolving linguistic trends within the LGBTQ+ community.

Exploring advanced vectorization techniques beyond TF-IDF is advised. Techniques like word embeddings (e.g., Word2Vec or GloVe) should be considered to capture semantic relationships between words and improve the model's understanding of complex linguistic patterns. Additionally, experimenting with non-linear models, such as kernelized SVMs or deep learning architectures like recurrent neural networks (RNNs) and transformers, can help overcome the limitations associated with linear SVMs.

Lastly, collaboration with linguistic experts, particularly those knowledgeable about Swardspeak and LGBTQ+ linguistic expressions, is encouraged to leverage their expertise in guiding model development. Cross-cultural validation should be conducted to assess the model's performance across diverse linguistic and cultural contexts. Ethical considerations and sensitivity training should be prioritized to mitigate biases and uphold ethical standards in the model's development and deployment.

These recommendations collectively aim to propel the model toward greater accuracy, adaptability, and inclusivity in capturing the richness of Tagalog Swardspeak within the LGBTQ+ community while ensuring that it aligns with evolving linguistic trends and remains culturally sensitive.

ACKNOWLEDGMENT

We would like to take Mr. Marc Ace J. Legaspi for proof checking the paper and for providing all the effort, suggestions, and recommendations to successfully complete the paper.

REFERENCES

[1] Rubiales, J. A. (2020). Linguistic Deviations of Swardspeak and Its Implication to Gay Students' English Language Competencies. Available at SSRN 3860558.

[2] Romero, R. (2019, July 8). *Gay lingo as reflection of social identity*. EUDL. https://eudl.eu/doi/10.4108/eai.27-4-2019.2285374

[3] Oco, N., Fajutagana, R., Lim, C. M., Miñon, J. D., Morano, J. A., & Tinoco, R. C. (2015). Witchebelles Anata Magcharot kay Mudra na Nagsusuba si Akech: Developing a Rule-based Unidirectional Beki Lingo to Filipino Translator. Journal of Sciences, Technology and Arts Research, 1(1), 29-37.

[4] Nocon, N., Kho, N. M., & Arroyo, J. (2018). Building a Filipino Colloquialism Translator Using Sequence-to-Sequence Model. TENCON 2018 - 2018 IEEE Region 10 Conference. doi:10.1109/tencon.2018.8650118

[5] Pascual, G. R. (2016). Sward speak (gay lingo) in the Philippine context: A morphological analysis. International Journal of Advanced Research in Management and Social Sciences, 5(12), 32-36.

[6] CATACUTAN, S. A. (2015). Authoring a Lexicon of Swardspeak through Community of Practice (Doctoral dissertation, University of the Philippines Open University).

[7] Papua, A. J., Estigoy, M. A., & Vargas, D. (2021). Usage of Gay Lingo Among Millenials as a Way of Communicating. Social Science Research Network. https://doi.org/10.2139/ssrn.3794691.

[8] Yourkiee. (2021). Gaytionary. Retrieved November 15, 2023 from https://www.wattpad.com/577787502-gaytionary-author%27s-note.

[9] Perez, KC, L. (n.d.). Glosaryo Gay Lingo. Retrieved November 15, 2023 from https://www.scribd.com/document/411910726/Glosaryo-Gay-Lingo.

[10] Almaden, S. A. (2023). A Quick Tutorial To Learning "Gandara Park" & More Beki Words. Retrieved November 15, 2023 from https://beelinguapp.com/blog/a-quick-tutorial-to-learning-gandara-park-&-more-beki-words.