Bilingual Dialect Classification using NLP

Fazeeia Mohammed *, Patrick Hosein and Aqeel Mohammed

The University of the West Indies, St. Augustine, Trinidad
mindy.moh@gmail.com, patrick.hosein@sta.uwi.edu,
aqeel.mohammed.eng@gmail.com

Abstract. This study focuses on identifying preferential techniques for multi-classification of natural language when dealing with multiple dialects. This research investigates dialects in Trinidad and Tobago and addresses the challenges associated with analyzing and classifying such linguistically diverse data. Our research further explores effective strategies for data pre-processing, feature extraction and model selection.

Keywords: Natural Language Processing, Bilingual dialects, linguistics

1 Introduction

The recent development of text-based surveys, data collection/scraping techniques such as chatbots, social media and SMS has resulted in the development of the field of natural language processing or NLP. This is particularly relevant in the area of customer care which is associated with understanding the issues faced by consumers and providing timely analytics to drive company-wide decision-making actions. Great strides have been made in enhancing machine comprehension of intricate aspects such as syntax, sarcasm, and dialect.This has facilitated more accurate and nuanced analysis of customer feedback. In addition to the previous factors, globalization, and international migration has resulted in areas becoming "linguistically diverse melting pots". The Caribbean has a rich history of migration which has resulted in the emergence of numerous bilingual dialect pockets and various dialects including Creole, Patois and Spanglish.

With the assistance of a telecommunication company in Trinidad and Tobago a dataset was generated via surveys. The dataset reflects the bilingual nature of the region and through the analysis of this data set we can gain insights into the challenges faced when interpreting and classifying dialects as well as the potential applications for these models. This study utilizes the said dataset and aims to explore the performance of different multi-classification models on bilingual Creole dialect through traditional NLP techniques and then with a more modified approach.

The telecommunication market is highly competitive in the region and, as such, customer feedback plays a crucial role in shaping policies, business strategies and improving customer satisfaction. This makes processing and responsiveness a time critical step in mitigating any negative customer experience. This has the added benefit of being a proactive approach and demonstrates the

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18,
https://doi.org/10.2991/978-94-6463-314-6_5

company's commitment to customer satisfaction, trust, loyalty, and long-term relationships. The region has a complex history that has, in turn, affected the language, syntax rules, vernacular, and vocabulary.

Caribbean dialects often have their roots in European languages such as English, Dutch, Spanish, and French. In addition to these European languages, mixed in are some African, Indian, and indigenous languages. These languages exist in varying degrees depending on the particular Caribbean islands historical complexity. The linguistic mix is heavily influenced by history and there are varying dialects that exist within the same region with no consistency between grammar, vocabulary, or pronunciation [1].

Within the country of Trinidad and Tobago, "Trinidadian Creole" is frequently used. It is a unique blend of French, Spanish, African, and Amerindian languages developed during its colonial period. Trinidad Creole has its own grammar and vocabulary that changes and evolves over time with the addition of new migrants and globalization. Recently, the Caribbean has experienced an influx of Venezuelan nationals due to social, economic, and political issues in Venezuela [2]. This sudden mass movement has had a significant effect on the regions linguistics since they have brought their own dialect and language variations. The interaction between Venezuelan and Caribbean dialects has created a dynamic linguistic environment with the potential for the blending of language features. The impact of the recent migration events can be observed in various aspects of Caribbean society such as media, businesses and everyday interactions.

The dataset collected for this study was done through surveying users of a telecommunication network in Trinidad and Tobago. A shift is seen in the amount of Spanish-speaking comments collected over the period of time. Often these comments are a hybrid of Spanish and code-switching to English in hopes of being understood clearly and effectively to resolve issues they are facing. This form of language adaptation, code-switching, and multiple dialects must be preprocessed sufficiently to ensure the meaning behind the collected comments is retained.

2 Related Work

Multilingual natural language processing has made significant strides in exploring shared language structures especially within the context of dialect variations. Due to advancements in cloud computing and processing power, large-scale multilingual models have emerged. These models are being leveraged over multiple languages simultaneously rather than solely on pairwise language transfer. Studies done by [3], [4], [5], and [6] have demonstrated the effectiveness of these models in capturing cross-linguistic patterns. In addition, benchmark studies that eliminate language-specific feature engineering, extensive training and encompassing multiple languages provide invaluable insights into natural language modeling. Most notable is the research done by [7], [8],[9], and [10] which emphasizes the advantages of methodologies which are cognizant of biases and language differences.

Another approach to natural language processing involves the use of part of speech tagging. It serves as a crucial tool in chunking data and can enhance statistical language models. These approaches however require an understanding of the language and a large amount of annotated training data for supervised tagging. In cases where there is abundance of labeled data and standard lexicons, POS taggers are used due to their effectiveness. [11]

Arabic dialect requires manual transcribing of actual conversations. There are no universally accepted lexicon standards for Arabic which further complicates the development of NLP models for dialect Arabic. Some of the issues highlighted by the study included (a) Limited resources such as lexicons, tokenizers and morphological analyzers, (b) Difficulty in tagging the spoken language, (c) Rich morphology: Arabic has a complex morphology with an increasing out-of-vocabulary words. and (d) Lack of short vowel information [11].

In summary, there are numerous applications for multi-classification models on dialect data and they have shown promising results with various processing issues. Future research can delve into the effectiveness of other machine learning models and deep learning techniques on different datasets. Additionally, the development of more annotated datasets for dialect can greatly improve the performance of these models. It is important to note that these studies did not account for the temporal aspects of language. A study done by [12] used a supervised learning approach to predict the time period to which a word belongs based on its surrounding context. The idea of language changing as time passes was also delved into by [13] who employed a Latent Semantic Analysis to identify words from early to modern English by analyzing usage patterns to track how the meanings of words change over time. Finally, these ideas highlight that the meaning and interpretation of words can change over time which can impact the performance of NLP models.

3 Dataset Description and Processing

For this study the data set was collected between 2020 to 2023, consisting of 124207 comments from the Trinidad and Tobago population using a random selection.Of the 124207 comments 27% were Spanish and the remaining was English. The data set was sourced from a telecommunications company operating regionally. It is composed of comments from various touch points such as surveys and chat bots and the incoming data is used to address customer issues and queries. The data is collated and classified into; Plan, Network, Account, Billing, Customer Care, Live Chat and Other which is then used for KPI reporting. The Customer Care team carried out an involved multi-step process equating to twenty working hours split amongst the Customer Care team over a forty hour work week. To assist with this time consuming task a Multi-Classification model was proposed as a solution identify language and dialects. The intention is to reduce and streamline the processing time of both Spanish and English data. This solution would be a step towards the automation for KPI reporting, customer segmentation and targeted marketing .

Table 1. Table of Sample Data collected from various sources and classified.

Data Source	Subscriber Comment	Classification	Language
Chat-bot	Porque los datos caen y se recuperan cuando están listos.	Network	Spanish
Text Survey	Quality,attention,on time...clearing my doubts	Customer Care	English
Text Survey	No se puede obtener asistencia	Billing	Spanish

Pre-processing was used to address spelling errors, random characters, random capitalization and mixed language comments in the data.

Initially, we attempted to correct the spelling errors and clean the data set through standard natural language processing techniques. However, this proved challenging since there were both English and Spanish comments and their dialects and respective spell-correcting algorithms often interpreted some dialect words as misspelled Spanish or English words. Also, the Spanish dialect used in "Trinidad Creole" is different from the Spanish dialect used by Venezuelans entering Trinidad. Consequently, the original meaning and intent of the Spanish comments would have been altered leading to possible misinterpretations.

To address this issue, we detected the type of dialect and performed spelling corrections accordingly. This method was promising, however, it proved less effective in capturing the overall meaning of the comments, particularly due to the absence of colloquial terms and phrases in the dictionary. The omission of these colloquial terms introduced a potential loss in understanding the context of the statement. To improve the performance and preserve the intended meanings of the comments we implemented a colloquial dictionary into the pre-processing pipeline.To do this we relied on local colloquial dictionaries that were created to build a relationship between the colonial history and the common phrases used in today's language. This inclusion helped with some of the limitations associated with the colloquial language. The use of the colloquial dictionary helped preserve the meanings of the comments by converting the informal slang into standard language equivalents. By standardizing the language you can ultimately improve the accuracy. It is crucial to highlight that the colloquial dictionary employed had become outdated and may not have accurately represented the current slang being used internally. Additionally, it created a bias toward the Creole dialect which contains some Spanish in it which could negatively affect the Venezuelan dialect. In order to prevent biases and ensure fair treatment of both Venezuelan and Creole dialect comments, we made the decision to treat the data set as a distinct language during pre-processing. By considering the data set's unique characteristics and treatment of the comments as its own language, we aim to prevent any undue bias towards a particular dialect.

Ultimately, eliminating language-specific feature engineering is a necessary avenue of research since language changes so dynamically. The cleaned data was then tokenized in preparation for stemmatization and lemmatization. Stemma-

tization is a common natural language processing technique that involves the reduction of words to their base form or stem. The purpose is to simplify and normalize words to enable more efficient analysis of the classification of text data. The process of stemmatization often involves the removal of suffixes and prefixes and other morphological variations from words so that similar meanings and intents can be categorized. This reduction in dimensionality help improve the overall performance of the classification algorithms.

Stemmatization, can generally be applied to standardize words across different dialects and bring them to a common form. However, the stemmatization of colloquial words can alter the meaning of the word itself. Two examples within the corpus of text show there is a limitation of stemmatization results and a potential loss of information during processing.

- the word "dotish", a common word equating to stupid, had been stemmatized to 'dot' which then loses its meaning.
- the word "Mamaguy" is equated to being made fun of or to ridicule which then loses its meaning when stemmatized to Mama

The simplification of dialect words ,as prior, especially colloquial words that stem from various different languages, can lose the original nuance. In Trinidad and Tobago and the Caribbean region it is common that the current dialect is an amalgamation of different languages that over time has evolved to facilitate communication within a very diverse society. As such, there are words that don't exist in the formal language it originally stems from. This can impact the accuracy and effectiveness of the classification process since dialect-specific features may be lost or overlooked.

In addition, the stemmatization process is not always perfect and has the potential to introduce inaccuracies. The algorithm that is used for stemmatizing of a corpus of text relies on predefined rules which may not capture all the linguistical complexities or dialect trends accurately. Consequently, there are situations where stemmatization may produce incorrect stems which affect the reliability of the classification results. Despite these limitations, stemmatization is still a valuable pre-processing step for dialect data classification.

4 Performance Results

Standard metrics of Accuracy, Precision and F1 Score were used in evaluating and comparing the performance of these models and providing invaluable insights into the model's capabilities. Accuracy represents the models ability to correctly identify the class of the comment, precision represents the models resilience to false positives. Both accuracy and precision are important as we can effectively identify the comments and avoid false positives which depending on the business case may have implications (example: incorrectly targetting a client based on their comment resulting in advertising cost). F1 score strikes a balance of the two and is generally considered more applicable than Accuracy. The F1 score elaborates on its class-wise performance rather than an overall performance as

done by accuracy, representing the harmonic mean between precision and recall. It provides a more holistic view of the models performance taking into consideration the false negatives of the classes. For this study eXtreme Gradient Boosting (XGB), Multinomial Logistic Regression, Support Vector Machines (SVM), and Bidirectional Encoder Representations from Transformers (BERT) were used on the pre-processed data set as discussed above.

Table 2. Results of selected metrics on NLP models built for multi-classification.

Model	Precision	Accuracy	F1 Score
Multinomial Logistic Reg.	78	79	78
XGB	85	77	72
SVM	80	80	79
BERT	72	91	61

Support Vector Machine (SVM) had the highest F1 Score at 79% . Its precision and accuracy were also high at 80% each. The XGB model had an F1 score and accuracy were 72% and 77% respectively it had the highest precision of all models at 85% which was 5 % higher than the SVM. The BERT model had an accuracy of 91% which suggests it may have been over fitted since the F1 score was relatively low at 61% and may have required more data to be accurately compared to the other models. The Multinomial Logistic Regression model performed at a 78% precision and F1 score and at 79% accuracy. Overall, the SVM model outperformed the others achieving the highest F1 score but each model had its strengths and limitations. The results show the potential for models to effectively analyze and classify bilingual data with dialectal characteristics.

5 Conclusion and Future Research

The results have shown relatively good performance for multi-classification in spite of the linguistic challenges associated with multiple dialects in our dataset. We have shown with appropriate pre-processing techniques and data engineering we can achieve reasonable performance within the 80Th Percentile for our standard metrics. Future research will address the challenges when working with dialect data for classification, particularly when dealing with multiple linguistic influences. One area of development should be a Venezuelan dialect dictionary that captures the linguistic nuances and vocabulary used in daily communication allowing for efficient processing of Spanish comments. Lastly, future research should investigate the use of transfer learning techniques, such as pre-trained language models like GPT, for dialect data classification.

Acknowledgements

We acknowledge the support of the Digicel Group of Trinidad and Tobago.

References

1. in Cayman Islands, B.: Language variety and their history in the caribbean (2022). URL https://www.bdo.ky/en-gb/insights/featured-insights/language-variety-and-their-history-in-the-caribbean
2. International, R.: (2023). URL https://www.refugeesinternational.org/trinidad-and-tobago/
3. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502 (2019)
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
5. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
6. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)
7. Bender, E.M.: On achieving and evaluating language-independence in nlp. Linguistic Issues in Language Technology **6** (2011)
8. Tsarfaty, R., Bareket, D., Klein, S., Seker, A.: From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? arXiv preprint arXiv:2005.01330 (2020)
9. Ravfogel, S., Tyers, F.M., Goldberg, Y.: Can lstm learn to capture agreement? the case of basque. arXiv preprint arXiv:1809.04022 (2018)
10. Ahmad, W.U., Zhang, Z., Ma, X., Hovy, E., Chang, K.W., Peng, N.: On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. arXiv preprint arXiv:1811.00570 (2018)
11. Duh, K., Kirchhoff, K.: Pos tagging of dialectal arabic: a minimally supervised approach. In: Proceedings of the acl workshop on computational approaches to semitic languages, pp. 55–62 (2005)
12. Mihalcea, R., Nastase, V.: Word epoch disambiguation: Finding how words change over time. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 259–263 (2012)
13. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: Comparing word meaning across time and phonetic space. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, pp. 104–111 (2009)