[redacted] | 2,195 words, excl. references | [redacted]

Classifying Filipino Full Names using Conditional Random Fields

I. Introduction

Names are an integral part of our identity, which connects us to our history, family, and culture. In Filipino context, full names follow the structure of having first name, middle name, and last name. This is heavily influenced by Spanish and Western practices with some having two or more given names, the mother's maiden family name as the middle name, and the last name being the father's paternal family name [1].

Accurate parsing of names play an important role in banking and financial transactions, specifically in sending or receiving money. Anti-money laundering (AML) regulations [2] mandate the complete and accurate information exchange about the sender and recipients of payment instructions between financial institutions to facilitate transaction monitoring processes.

In the United Kingdom, the retail payments authority, Pay.UK, provides a service called Confirmation of Payee (CoP) [3] that ensures accuracy of the recipient's account details. However, this system relies on the full account name rather than segmented name components.

Conversely, in the Philippines, according to Bangko Sentral ng Pilipinas (BSP) Circular 980 [4], validation is only performed on the account number by the receiving bank. On the other hand, the Anti-Money Laundering Council, Philippines (AMLC) reporting guidelines [5] suggest that for individual senders, at least the last name should be present. However, due to the free-format nature of the name fields to accommodate various naming conventions, the process of separating names into their respective fields remain a manual task during report generation.

This study proposes that this process can be automated using a machine learning algorithm. Such approach can help with parsing names to improve efficiency and accuracy in AML compliance.

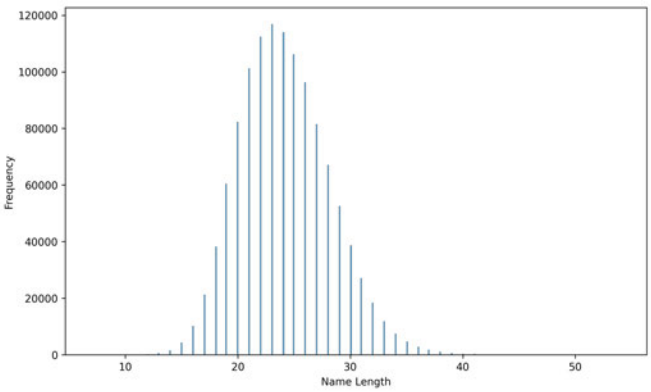

Fig. 1. Name Length Distribution

II. Data Background

A. Data Source and Collection

The dataset used for this study comprises names collected from Philippine professional licensure examinations (board exams) and bar exams. The primary sources of data include the Professional Regulation Commission (PRC), the Supreme Court of the Philippines, and reputable news agencies such as Rappler and GMA News Network. The names were collected by downloading pages containing the lists of names and then filtering the text to match the regular expression for the name format: LAST NAME, COMBINED FIRST AND MIDDLE NAMES.

B. Dataset Size and Composition

For this data collection, a total of 1,183,452 names were collected from 2,285 web pages and initially separated into last names and combined first and middle names. Through a series of rules, the combined first and middle names were further subdivided into their respective parts. The dataset attributes are shown in Table 1.

C. Exploratory Data Analysis

The names collected contains 447,941 unique first names, 87,521 unique middle names, and 116,120 unique last names. The average full name length is 24 characters, with a minimum of 7 and maximum of 54 characters. The total number of words in the full name averages at 3.58, with maximum at 9 words. Additional analyses with the names were also conducted to identify features that can be extracted from the dataset.

1) Name Length Analysis

Fig. 1 shows the distribution of the full name lengths and indicates that most names have a length between 22 and 26 characters. There also exists some names falling beyond this range, but it is relatively uncommon. Having a bell-shaped curve also indicates that the names are normally distributed around the average length.

Additionally, Fig. 2 shows the distribution of the lengths of names per component (first, middle, or last). As the first name is

TABLE I. Dataset Structure

Column	*Type*	*Description*	*Sample Value*
0	Integer	Non-unique identifier	230
1	String	Last name	DELA CRUZ
2	String	Combined first and middle name	JUAN CRUZ
A	String	Assumed first name	JUAN
B	String	Assumed middle name	CRUZ
is_surname_in_list	Boolean	Column B check if it is in last name column	True

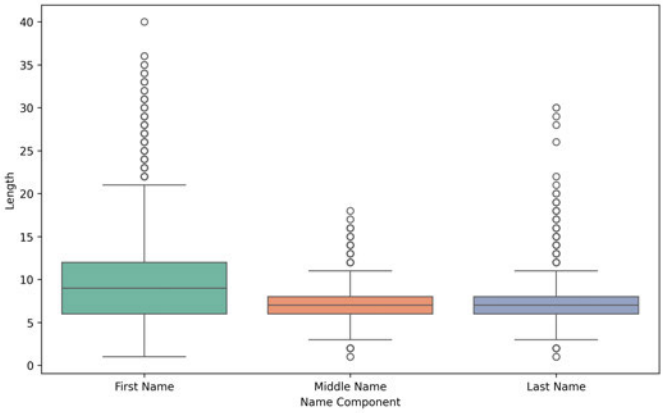

Fig. 2. Name Component Length Distribution

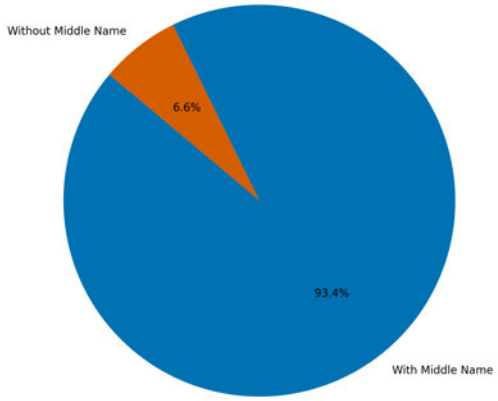

Fig. 3. Proportion of Names With and Without Middle Names

given by the parents or guardian, it tends to have more variability in length compared to the middle and last names. Consequently, last name lengths have a notable number of outliers on the upper end, possibly indicating the presence of compound surnames.

2) Name Component Analysis

Fig. 3 shows that 6.6% of the names in the dataset does not have a value for the middle name. On the other hand, Table 2 shows the distribution of compounded names. The name is considered compounded if they are hyphenated or consists of multiple words.

3) Frequency Analysis

To visualise the common names associated with each of the segments, a word cloud visualisation of first names, middle names, and last names was generated as shown in Fig. 4. The figure remains consistent with acknowledging the presence of Western and Spanish influence in the naming system such as presence of MA and JR in first names, as well as DE LA/DELA in middle and last names.

III. Methodology

The machine learning problem in this study is to be able to parse names with Filipino origins into their constituent parts: first, middle, and last names. It also requires to be able to identify certain nuances in Filipino names such as having generational designations (e.g. JR., II, III, IV) or having Spanish-influenced prepositional particles (e.g. DE, DE LA, DELA) in the middle or last names.

TABLE II. Compounded Names Distribution

	Hyphenated	***Multiple Words***	***Total Compounded***
first_name	546 (0.05%)	621362 (52.50%)	621908 (52.55%)
middle_name	6691 (0.57%)	29890 (2.53%)	36581 (3.09%)
last_name	6824 (0.58%)	33725 (2.85%)	40549 (3.43%)

A. Feature Selection

To determine the role of each word in a name, indicators such as suffixes (e.g., JR, II, IV) and prefixes (e.g., DE, STA, DELA) are extracted. Additional features like word length, relative position, and the presence of special characters are also considered to aid in classification.

B. Univariate vs. Multivariate Analysis

This study utilised multivariate analysis as this allows for the following cases:

1. Multiple Features: Various features of each word in the name, such as position, presence of nuances, etc., can be used to determine the classification of each word.
2. Complex Patterns: As Filipino names can have complex structure, considering multiple factors is required to accurately classify them.
3. Feature Interaction: The relationship between features, such as considering word position and suffix presence, can help significantly for accurate classification.

C. Algorithm Consideration

In considering the algorithm to be used for this ML problem, the following factors were considered:

1. Balance of Complexity and Performance: As this only relies on an average of 4 words per name, the model may not require a larger complexity as there will be less features than performing a full NER.
2. Task Suitability: Name parsing is a relatively structured task where relevant features can be easily identified. Having a model that can utilise these features can improve accuracy.

Based on these factors, the following algorithms were evaluated:

1) Conditional Random Fields

Conditional random fields was presented by Lafferty, McCallum, and Pereira as a framework to segment and label sequence data in [6]. CRF considers the neighbours when doing

Fig. 4. (L-R) Word Cloud of First Names, Middle Names, and Last Names.

the classification, whereas a classifier does not. This algorithm is a supervised machine learning technique given that it requires a labelled training data.

In relation to this, the training data is tagged in Beginning-Inside-Outside format presented by Ramshaw and Marcus in [7]. This scheme allows to identify the structure and positioning of the names in relation to the words beside them.

This combination of method has been frequently used for Named Entity Recognition (NER) machine learning problems. This framework has been used in NER in different languages such as Marathi [8], Portuguese [9], and Filipino [10] and have an F1-Measure of 75.51, 82.68, and 83.31, which shows its effectiveness in different languages.

2) Bidirectional LSTM with CRF layer (BiLSTM-CRF)

Long Short-Term Memory (LSTM) networks [11] are a type of recurrent neural network that manage the flow of information, enabling the model to retain long-term dependencies in sequences. Bidirectional LSTM enhances this by analysing sequences in both directions, forward and backward, producing a comprehensive representation for each word [12].

While using this approach, alongside bidirectional CRF, can lead to more accurate classification by capturing dependencies in both directions, it demands more computational resources during training and testing. Given the small size of input data, this may be excessive for this study. However, it could be explored for developing a language-independent model.

3) Naïve Bayes

The Naïve Bayes classifier is a probabilistic model that assumes all features are independent of one another. Although this assumption is frequently violated in real-world scenarios, the classifier still performs reasonably well under it [13].

However, because of this assumption, important features like word context and sequence are disregarded, resulting in suboptimal performance for Named Entity Recognition (NER) tasks. Moreover, the model struggles with ambiguity, which is a common challenge when dealing with named entities, particularly in this dataset.

D. Algorithm Selection

Due to the balance in performance and cost of training and maintaining the model, it is ideal to use **Conditional Random Fields**. This is a good compromise with the possible increase in accuracy provided by BiLSTM-CRF with the small number of words per full names as well as classifying only on three fields. On the other hand, Naïve Bayes classifier might be able to achieve the same tasks but having an algorithm that is more optimised can be beneficial in the long run.

E. Model Performance Evaluation

As the problem is approached using a classifier, key classification metrics are used to assess the model's performance. These include precision, recall, and F1-score which are computed through the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

1. **Accuracy** (1) measures correctness of the model.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. **Precision** (2) measures the proportion of correct predictions out of all instances classified as positive by the model.

$$P = \frac{TP}{TP + FP} \quad (2)$$

3. **Recall** (3) indicates the model's ability to correctly identify positive instances from all actual positive cases.

$$R = \frac{TP}{TP + FN} \quad (3)$$

4. **F1-score** (4) balances both precision and recall, providing a single metric that evaluates the trade-off between them, especially useful when one metric may affect the other.

$$F1 = 2 * \frac{P \cdot R}{P + R} \quad (4)$$

These metrics are evaluated for each label (B-FIRST, B-LAST, B-MIDDLE, I-FIRST, I-LAST, I-MIDDLE) to provide understanding on the model's performance in classifying different parts of the name.

Additionally, a 10-fold cross-validation was used to minimise the effect of any bias or variance caused by a specific set of training data. This is done by splitting the dataset into 10 subsets, with the model trained on 9 and tested on the remaining set. This process is repeated 10 times, with each subset acting as

a test set once, to ensure that the model is evaluated on all parts of the dataset and providing a more reliable estimate of its performance.

IV. Results and Discussion

A. Model Performance

The Conditional Random Fields (CRF) model showed excellent performance in classifying Filipino full names in all metrics. The result of the 10-fold cross-validation are as follows:

1. Accuracy: 98%
2. Macro-average F1-score: 0.97
3. Weighted-average F1-score: 0.98

The model consistently performed well across all name components, particularly for the beginning of first names (B-FIRST) and last names (B-LAST), as shown in Table 3. However, slightly lower performance was observed for middle names (B-MIDDLE), indicating challenges in differentiating middle names from other components.

B. Feature Performance

Analysis of the 10 most important features revealed which factors contributed to the model's decision-making process:

1. **Presence of comma:** This was a strong indicator for the beginning of the first names, reflecting the common "LAST, FIRST MIDDLE" format.
2. **Beginning of Sequence (BOS):** This feature was a key factor to identify the start of the first name, aligning with the "FIRST MIDDLE LAST" format.
3. **Specific Name Patterns:** Words such as "MAE", "TAMAYO", "JOY" were considered as important features, suggesting that it learned some Filipino name patterns.
4. **Previous Words:** Words like "LOS" were strong indicators of the middle names, suggesting that the model identified patterns in compound middle names (e.g. "DE LOS SANTOS") or those with Spanish influences.

C. Discussion

The high overall accuracy demonstrates the effectiveness of CRF model in classifying Filipino full names. This model performs well in identifying the beginning of each name component which contributes to the accurate classification. However, the slightly lower prediction for middle names suggests that the variability of the inner structure of Filipino names, such as having multiple given names and inconsistent use of middle names, makes it harder to find the beginning of the middle name.

The importance of features like comma presence and position-based indicators helps the model to be effective in learning the patterns in name formats which suggest this model can be generalised to other structured name formats beyond Filipino names.

Overall, the high performance of the model across different components demonstrates the model's ability to handle complex full name structures, such as having multiple words for the first, middle, and last names. This adaptability is particularly important given the diversity of Filipino naming conventions.

TABLE III. Compounded Names Distribution

Component	*Precision*	*Recall*	*F1-score*	*Support*
B-FIRST	1.00	1.00	1.00	1183452
B-LAST	0.98	0.98	0.98	1183452
B-MIDDLE	0.93	0.98	0.96	1105139
I-FIRST	0.99	0.92	0.95	665613
I-LAST	0.98	0.97	0.98	36458
I-MIDDLE	0.96	0.98	0.97	32366

V. Conclusion

This study demonstrates the effective application of Conditional Random Fields (CRF) in the domain of entity recognition, specifically for classifying Filipino full names. The model's high overall accuracy of 98% and consistently strong performance across different components highlights its potential in automating tasks.

Future research can be conducted where the following can be explored such as introducing controlled noise, expanding the dataset to other structure name formats, and exploring transfer learning approaches to adapt the model to other cultural naming conventions.

In conclusion, this study provides a promising framework for automated Filipino name classification using CRF. The high accuracy and ability to capture nuanced features of Filipino names makes it a valuable tool for various applications.

References

[1] Cultural Atlas (2023) Filipino culture: Naming. Available at: https://culturalatlas.sbs.com.au/filipino-culture/filipino-culture-naming (Accessed: 19 August 2024).

[2] FATF (2023) FATF Recommendations. Available at: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html (Accessed: 19 August 2024).

[3] Pay.UK (2023) Confirmation of Payee. Available at: https://www.wearepay.uk/what-we-do/overlay-services/confirmation-of-payee/ (Accessed: 19 August 2024).

[4] Bangko Sentral ng Pilipinas (2017) Circular No. 980 - Anti-Money Laundering Regulations. Available at: https://www.bsp.gov.ph/Regulations/Issuances/2017/c980.pdf (Accessed: 19 August 2024).

[5] Anti-Money Laundering Council (2021) AMLC Registration and Reporting Guidelines. Available at: http://www.amlc.gov.ph/images/PDFs/2021-AMLC%20REGISTRATION%20AND%20REPORTING%20GUIDELINES.pdf (Accessed: 19 August 2024).

[6] Lafferty, J.D., McCallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282-289.

[7] Ramshaw, L., & Marcus, M. (1995). Text Chunking using Transformation-Based Learning. https://doi.org/10.48550/arxiv.cmp-lg/9505040

[8] Patil, N., Patil, A., & Pawar, B. V. (2020). Named Entity Recognition using Conditional Random Fields. *Procedia Computer Science*, *167*, 1181–1188. https://doi.org/10.1016/j.procs.2020.03.431

[9] Nogueira, R., Lotufo, R., & Souza, F. (2019). Portuguese Named Entity Recognition using BERT-CRF. https://doi.org/10.48550/arxiv.1909.10649

[10] Alfonso, A. P. T., Villegas, J. T., Domingo, I. V. R., Villar, R. B., Sagum, R. A., & Galope, M. J. F. (2013). Named Entity Recognizer for Filipino Text Using Conditional Random Field. *International Journal of Future Computer and Communication*, 376–379. https://doi.org/10.7763/ijfcc.2013.v2.189

[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[12] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2023). Chapter 2 - Natural language processing and sentiment analysis: perspectives from computational intelligence. In *Computational Intelligence Applications for Text and Sentiment Data Analysis* (pp. 17–47). elsevier. https://doi.org/10.1016/b978-0-32-390535-0.00007-0

[13] Rish, I. (2001) An empirical study of the Naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, Seattle, 4 August 2001, pp. 41-46.