A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

David Trye[1*], Vithya Yogarajan[2*], Jemma König[1], Te Taka Keegan[1],
David Bainbridge[1] and Mark Apperley[1]

[1]School of Computing and Mathematical Sciences, University of Waikato, New Zealand
[2]Strong AI Lab, School of Computer Science, University of Auckland, New Zealand
[*]*dgt12@students.waikato.ac.nz, vithya.yogarajan@auckland.ac.nz*

Abstract

Most large-scale language detection tools perform poorly at identifying Māori text. Moreover, rule-based and machine learning-based techniques devised specifically for the Māori-English language pair struggle with interlingual homographs. We develop a hybrid architecture that couples Māori-language orthography with machine learning models in order to annotate mixed Māori-English text. This architecture is used to label a new bilingual Twitter corpus at both the token (word) and tweet (sentence) levels. We use the collected tweets to show that the hybrid approach outperforms existing systems with respect to language detection of interlingual homographs and overall accuracy. We also evaluate its performance on out-of-domain data. Two interactive visualisations are provided for exploring the Twitter corpus and comparing errors across the new and existing techniques. The architecture code and visualisations are available online, and the corpus is available on request.

1 Introduction

"Ko te reo te mauri o te mana Māori.
Ko te kupu te mauri o reo Māori."

Translated to English as *The language is the life force of the mana Māori. The word is the life force of the language* (Higgins and Keane, 2015), this famous saying by Tā Hēmi Hēnare (Sir James Hēnare) encapsulates the importance of the Māori language to Māori, the Indigenous people of Aotearoa[1] New Zealand.

Te reo Māori is both endangered and low-resourced, with limited corpora and Natural Language Processing (NLP) techniques available (James et al., 2020). Data annotation currently has to be done manually by language experts, making the process time-consuming and resource-intensive. These obstacles hinder technological advances that could assist in maintaining the language and, consequently, the culture of Māori.

The Māori language used today is frequently interspersed with English, either in the form of *code-switching* (Holmes and Wilson, 2017; Marras Tate and Rapatahana, 2022) or *borrowing*. Here, the borrowing process is bidirectional, resulting in both English loanwords in Māori (Harlow, 1993) and Māori loanwords in English (Calude et al., 2020). The latter are not only used by bilingual Māori speakers, but also by monolingual English-speaking New Zealanders. Linguists are interested in determining the frequency of these patterns, which are reflective of Aotearoa New Zealand's unique bicultural identity.

The interweaving of Māori and English is a key consideration for developing robust technologies that can accommodate practical, everyday usage of te reo Māori and New Zealand English. Leveraging the abundance of relevant data on Twitter, our research focuses on the following task:

> *Automatic language identification for bilingual Māori-English text at both the token (word) and tweet (sentence) level.*

Differentiating between Māori and English text is not straightforward. This is because both languages use the Roman script, and *interlingual homographs*—words that are spelt the same but differ in meaning across languages (Dijkstra, 2007)—are prolific. These words present a major challenge for classifying mixed-language text, especially if they are highly frequent in both target languages (Barman et al., 2014). Consider the following tweets in which interlingual homographs are emphasised:

(a) **Here** is **to a more** productive day tomorrow
(b) Ka **kite** koe **i a** koe!
(c) **He** is at **a** tangi in Ruatoki. Doubt **he** did

In terms of annotation, the desired tweet-level labels are (a) English, (b) Māori, and (c) Bilingual. These are determined with recourse to the individual token labels: all tokens in (a) are English, all

[1]Aotearoa is increasingly used as a Māori name for New Zealand. Te reo Māori means 'the Māori language'.

tokens in (b) are Māori, and (c) contains a mixture of tokens from both languages, with 'tangi' (funeral) and 'Ruatoki' (a place name) being labelled Māori. According to our approach, all words of Māori origin are tagged as Māori, even if they are used as borrowings in English.

In order to obtain accurate tweet and token-level labels, we utilise knowledge and understanding gained from Māori researchers, Māori technology developers and the Māori community. Our methodology involves combining machine learning techniques with Māori orthography, thereby instantiating the pipeline recommended by Hämäläinen (2021). We hypothesise that doing so will improve the overall accuracy of language identification for bilingual Māori-English text.

This paper makes the following contributions:

1. Development of a hybrid architecture[2] to detect Māori and English words for a given bilingual text input.
2. The *Māori-English Twitter (MET) Corpus*, a first-of-a-kind dataset comprising bilingual and monolingual tweets, annotated at the token- and tweet-level by deploying our architecture.
3. Evidence that the hybrid architecture improves both language detection of interlingual homographs and overall accuracy when compared with two existing techniques.
4. Two interactive visualisation tools for exploring the corpus and comparing label errors across the different systems.

2 Background and Related Work

2.1 Māori Data Sovereignty

The Māori language is the natural medium through which Māori express their cultural identity, construct the Māori worldview and convey their authenticity (Marras Tate and Rapatahana, 2022; Rapatahana, 2017; White, 2016). It is crucial to highlight that Māori data needs to be handled with care, because of the injustices caused by colonisation and its effect on the vitality of the language (Smith, 2021). We strongly believe that any NLP resources that are developed from this research, either directly or indirectly, should be created for the good of the Māori-language community and not for the capital gain of others; more generally, Indigenous data should not be commodified at the expense of Indigenous communities (Bird, 2020).

[2] *https://github.com/bilingual-MET/hybrid*

2.2 Challenges and Bias in Māori NLP

Key challenges in developing Māori speech and language technology arise from the lack and limitations of resources (James et al., 2020), phonological differences from English, and the lexical overlap between written Māori and English, including more than 100 interlingual homographs.[3] These obstacles hinder NLP advances that could facilitate the maintenance of Māori language and culture.

Existing large-scale technologies such as cloud-based language-detection tools and voice assistants are predominantly designed for English. These tools fail to recognise or correctly pronounce Māori words, even when used as borrowings in New Zealand English (James et al., 2022b). Our goal is to redress that inequity in NLP resources, and thus mitigate the bias that existing tools have towards the more dominant English language.

2.3 Code-Switching in NLP

Bilingual and multilingual code-switching, especially between resource-rich and low-resourced languages, has gained traction as a challenging but important NLP problem (Aguilar et al., 2020; Molina et al., 2016; Solorio et al., 2014). A myriad of studies investigating code-switching on social media has emerged, showcasing challenges and possibilities for many different language pairs (Jose et al., 2020; Maharjan et al., 2015; Barman et al., 2014).

While an overview of Māori-language corpora is given in Trye et al. (2022), we detail three that are particularly relevant here. The *Hansard Dataset* (James et al., 2022a) comprises two million Māori, English and bilingual sentences, annotated by hand at both the token and sentence levels. The *MLT Corpus* (Trye et al., 2019) is a publicly-available collection of English tweets with Māori borrowings, albeit lacking token-level labels. The *RMT Corpus* (Trye et al., 2022) contains predominantly-Māori tweets and is also publicly-available. We use the hand-crafted rules from the RMT Corpus to detect candidate Māori words based on Māori orthography (Section 3.2).

Research using machine learning techniques for te reo Māori is relatively young, and is restricted by the limited scope of available resources. Although cloud-based services offered by corporations such as Google and Microsoft support Māori-language detection, the accuracy of these services

[3] *https://github.com/TeHikuMedia/reo-toolkit*

is poor (Keegan, 2017; James et al., 2022b).

Recently-developed language identification and code-switching detection models for the Māori-English pair make use of Skipgram-based fastText models to pre-train embeddings (Dunn and Nijhof, 2022; James et al., 2022b). James et al. combine pre-trained embeddings with recurrent neural networks (RNNs) to identify Māori text and code-switching points between the Māori-English pair. Their embeddings were pre-trained on a large collection of bilingual and monolingual data, and shown to outperform open-sourced English-only equivalents. Our hybrid architecture uses the fastText pre-trained embeddings and Hansard training set from James et al. (2022b).

3 Methodology

This section details the process used to collect Twitter data (Section 3.1) and the techniques underpinning our hybrid architecture. We combine language rules (Section 3.2) with neural networks (Section 3.3), as suggested by Hämäläinen (2021).

3.1 Data Collection and Pre-processing

In order to create a bilingual Twitter corpus on which to deploy our architecture, we combined tweets that were originally gathered for the RMT Corpus with more recent tweets from the same users.[4] Tweets that included 30-80% Māori text under the RMT system were chosen, as it was deduced these would primarily contain instances of Māori-English code-switching. The collected tweets were pre-processed to mitigate noise in the dataset. A series of tweets was removed, including retweets, similar and identical tweets, tweets posted by bots, and tweets containing fewer than four words. Non-Roman characters were stripped from the remaining tweets and common English contractions were expanded. 20,000 foreign-language tweets were then removed via manual and automatic checks, which included searching for symbols denoting glottal stops in the middle of tokens (characteristic of several Polynesian languages related to, but distinct from, Māori). This yielded 178,192 tweets in total. Finally, when extracting the tokens in each tweet, links, user mentions, hashtags, punctuation, emoticons and Arabic numerals were all ignored. The rationale for excluding hashtags is that they often contain abbreviations and/or multiple words, sometimes even combining languages (Trye et al., 2020), making them difficult to annotate without additional pre-processing.

3.2 Hand-Crafted Rules

Trye et al. (2022) employ hand-crafted rules to identify Māori tokens in tweets, referred to as the *RMT system* throughout this paper. This technique adapts hand-crafted rules implemented by Te Hiku Media, an Indigenous Māori organisation.[5] The rules are as follows:

- Tokens must contain only characters from the Māori alphabet, which comprises five vowels (*i, e, a, o, u*) and ten consonants (*p, t, k, m, n, ng, wh, r, w, h*).
- Lengthened vowels may be indicated with a macron (*ā*), or using double-vowel orthography (*aa*).
- Tokens must adhere to Māori syllable structure: they must follow consonant/vowel alternation, end with a vowel, and be free of consonant clusters (excluding the digraphs *ng* and *wh*).
- For input to the algorithm, some further adjustments were made to identify as many candidate Māori words as possible.[6]

When applied to bilingual text, a major limitation of these rules is that tokens of the same type are always classified the same way (typically as Māori), which is problematic for interlingual homographs.

3.3 Machine Learning Component

The hybrid architecture uses Bidirectional Gated Recurrent Units (Cho et al., 2014) with an attention layer as the machine learning component. Text is represented using fastText (Bojanowski et al., 2017) Skipgram-model word embeddings (Mikolov et al., 2013) with 300 dimensions, pre-trained on a collection of Māori and bilingual corpora (James et al., 2022b). The attention layer used is based on the Bahdanau attention mechanism (Bahdanau et al., 2015). Our preliminary experiments favoured the use of Bi-GRU with an attention layer over other deep learning models such as CNNs and LSTMs.

To the best of our knowledge, there is no large bilingual Twitter dataset annotated accurately by experts at the token- or tweet-level. Hence, for training Bi-GRU, we use the Hansard Dataset containing transcribed formal Māori and

[4] Users were identified via *Indigenous Tweets* (*http://indigenoustweets.com/*).

[5] *https://github.com/TeHikuMedia/nga-kupu*

[6] Words like 'a', 'i', 'to' and 'no' were omitted from the original RMT system due to their high frequency in English.

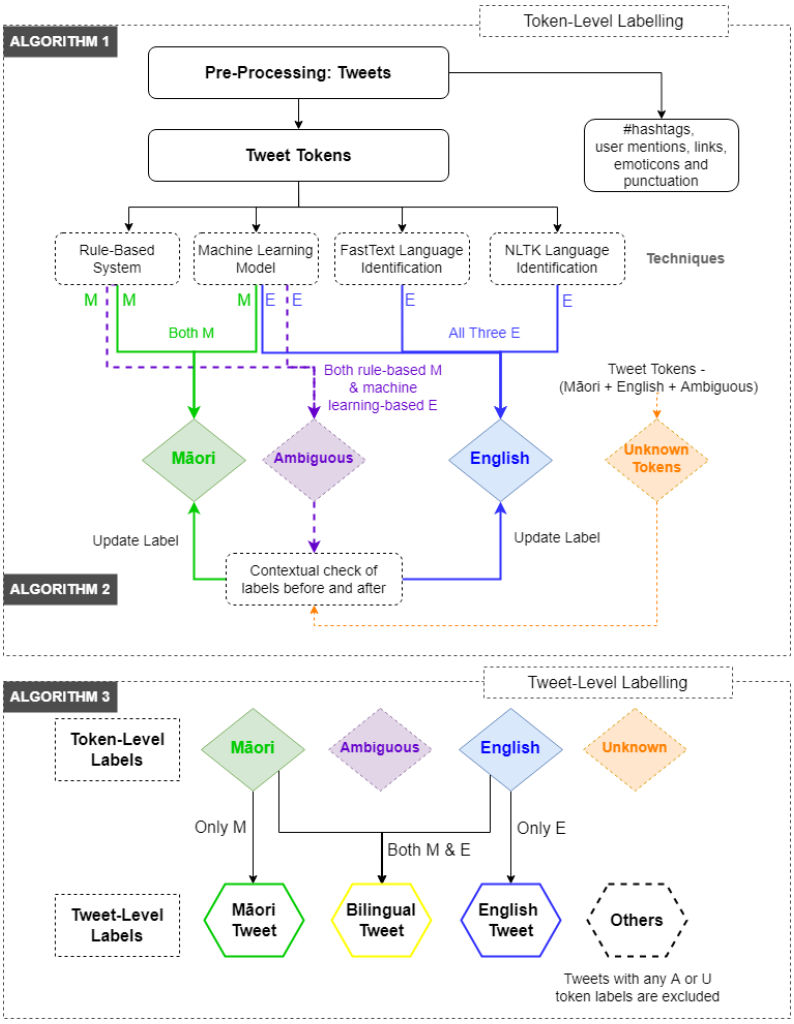


Figure 1: Flow chart detailing token- and tweet-level labelling.

English (James et al., 2022b). The Bi-GRU model is trained to predict Māori, English or bilingual sentences, using default settings in Keras/Tensorflow. Adam (Kingma and Ba, 2015), an adaptive learning rate optimisation algorithm, was employed as the optimiser for the networks. Softmax activation is leveraged in the output layer. To avoid over-fitting, we use a combination of dropout (Srivastava et al., 2014) with a rate of 0.5 and early stopping (Zhang et al., 2017).[7]

4 Hybrid Architecture

The hybrid architecture for labelling bilingual Māori-English datasets at both the token (word) and tweet (sentence) levels builds upon the RMT and ML techniques described in the previous section. Figure 1 outlines the process used to label the tweets in our cleaned dataset, and references the algorithms in Appendix A. The architecture can also be directly applied to Māori-English corpora with longer text sequences.[8]

4.1 Token-Level Labels

Multiple techniques are used to determine the appropriate label for each token (Algorithms 1 and 2). Initially, tokens are deemed to be Māori only if they are labelled 'M' by both the modified rules from the RMT Corpus and the pre-trained machine learning model. In a similar vein, English tokens are labelled by combining the outcome of using the machine learning model with fastText (Joulin et al., 2017, 2016) and NLTK (Bird and Loper, 2004) language identification models. These techniques have proven high accuracy in detecting English, providing confidence in the 'E' labels. Due to the informal nature of tweets, the language-specific tags include colloquial language and textspeak (e.g.

[7] Model trained on 12 core Intel(R) Xeon(R) W-2133 CPU @ 3.60GHz, GPU device GV100GL.

[8] *https://github.com/bilingual-MET/hybrid*

	Tweets	Bilingual (B)	English (E)	Māori (M)
Tweets	76,416	67,713	7847	856
Tokens	781,381	-	465,292	316,089
Users	2417	2347	1148	283
Avg tokens/tweet	10	11	6	6
Avg tweets/user	32	29	7	3

Table 1: Summary statistics for the MET Corpus.

'u' for 'you' in English).

Any tokens that are labelled 'M' by the modified RMT system and 'E' by the machine learning model are initially classified as ambiguous. The knowledge gained from neighbouring tokens is then used to re-classify these words as Māori or English (Algorithm 2). Crucially, the MET Corpus only includes tweets comprising 'M' and 'E' token-level labels; all remaining tokens that could not be re-classified with certainty led to the removal of the corresponding tweet, and are left for future research.

4.2 Tweet-Level Labels

The updated token labels are used to generate appropriate tweet-level labels (Figure 1, Algorithm 3). If a tweet consists solely of 'M' or 'E' tokens, then the tweet-level label is Māori or English, respectively. Tweets that contain at least one 'M' and 'E' token are considered bilingual; this includes single-word borrowings in otherwise monolingual contexts. For further confidence, the tweet-level labels were compared with the pre-trained machine learning model, and it was found that 90% of these labels matched the hybrid model.

5 The Māori-English Twitter Corpus

The steps detailed in the previous two sections resulted in the formation of a new bilingual dataset: the *Māori-English Twitter (MET) Corpus*. Key summary statistics for this collection of 76,000 tweets are presented in Table 1. Almost 90% of tweets in the corpus are labelled Bilingual, 10% are English and only 0.1% are Māori. This distribution is expected, given the chosen threshold and characteristics of the RMT system used to filter tweets in the data collection phase. In terms of individual words, 60% of tokens in the MET Corpus are labelled English and 40% are Māori. The 20 most frequent tokens are shown in Figure 2. Most of these tokens are function words rather than content words, apart from 'Māori' and 'reo' (language), whose presence would suggest that many tweets in the corpus pertain specifically to Māori language and culture.

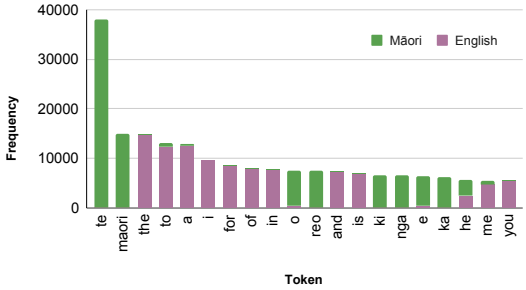


Figure 2: The 20 most frequent tokens in the MET Corpus: **Māori words**, **English words** and **homographs**.

5.1 Visualisation of the MET Corpus

We provide an interactive visualisation for exploring the MET Corpus;[9] see Figure 3. The visualisation includes a scrollable table of tweets and allows the user to select and filter data according to several dimensions. Key features include a treemap (and associated search bar) displaying token frequencies for the selection, a line chart of the distribution of selected tweets over time, and a bubble chart summarising the relative contribution of each user. In addition, selections can be made on both the tweet and token-level labels. The percentage of tweets that is currently visible (with respect to the entire corpus) is indicated at the top left of the display.

5.2 Gold Standard Labels

A manual annotation process was used to obtain gold standard labels for a random one percent sample of the data (N=850 tweets), including tweets that were ultimately filtered out of the corpus. This process consisted of two phases. In phase one, two of the authors manually tagged the true tweet-level label of each tweet in the sample, so that this could be compared against the predicted label for each system. Furthermore, the coders identified which tokens, if any, had been mislabelled by each system. Tokens were considered to be Māori if they were listed in the Māori dictionary,[10] constituted Māori slang (e.g. 'ktk' is the Māori equivalent of 'lol'), or were Māori named entities. It was decided that even Māori borrowings in otherwise English tweets should be tagged as Māori, because applications such as a New Zealand English text-to-speech tool would be required to correctly identify and pronounce words of Māori origin, regardless of how they are categorised from a theoretical point of view.

In the sample tweets, the coders encountered

[9] *https://bilingual-met.github.io/hybrid*
[10] *https://maoridictionary.co.nz/*

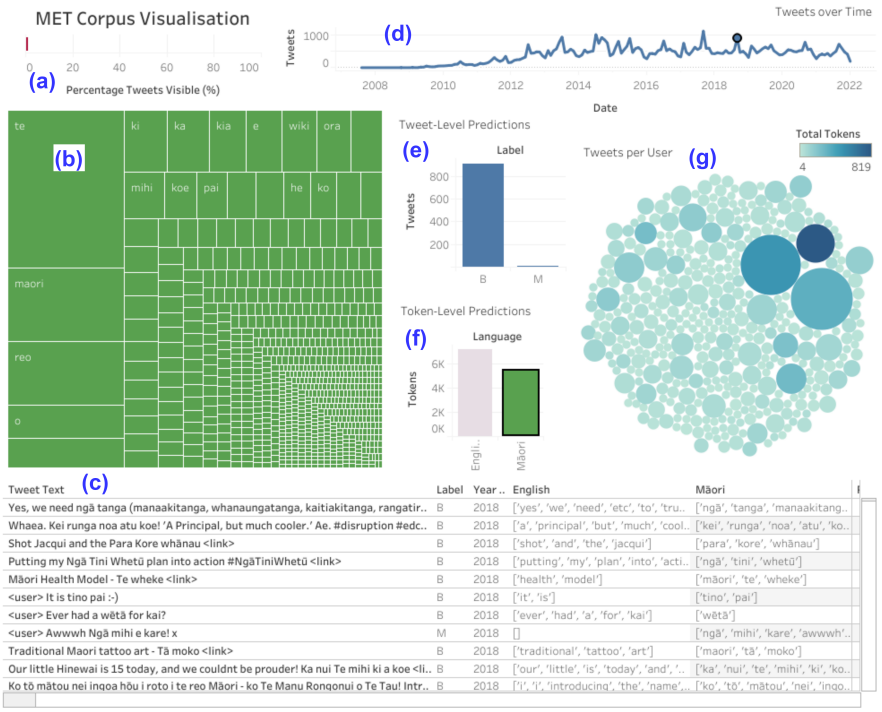


Figure 3: Interactive tool for exploring the *MET Corpus*: (a) percentage of corpus visible, (b) selected tokens by frequency, (c) tweet table, (d) tweets by year, (e) tweet predictions, (f) token predictions, (g) tweets by user.

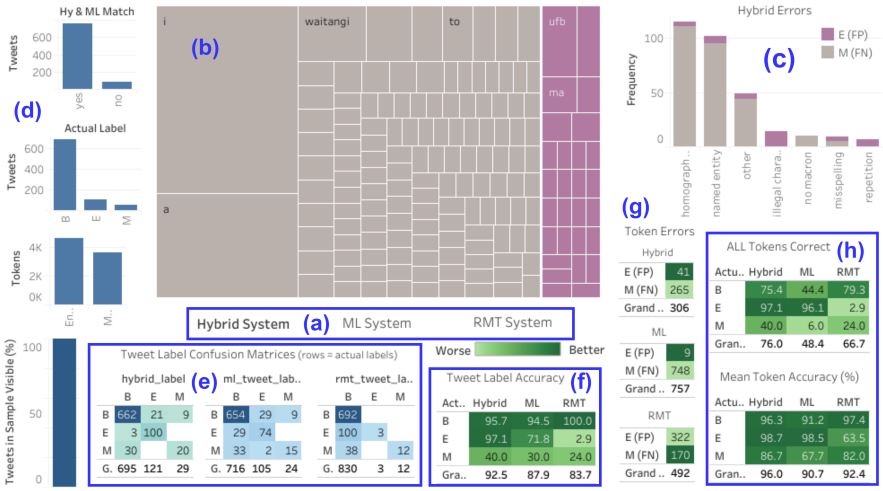


Figure 4: Interactive tool for comparing system errors: (a) navigation menu, (b) misclassified tokens, (c) error types, (d) filtering by labels, (e) tweet label confusion matrices, (f) tweet accuracy, (g) token mistakes, (h) token accuracy.

five foreign tweets (0.6%), which were discarded, since the individual tokens could not be accurately tagged as either English or Māori. In order to assess the efficacy of phase one of the annotation process, Cohen's kappa was computed for a subsample of 200 tweets. This yielded a score of 0.816, indicating a strong level of agreement.

For the second phase, one of the authors went through the data again, and, for each mistaken token, noted whether it was a Māori token that had been mislabelled as English (false negative), or an English token that had been mislabelled as Māori (false positive). Where possible, they recorded further information about the specific type of error. Common error types included short-length homographs, named entities (including names of people, places, tribes, organisations and events), the presence of one or more non-Māori characters, misspellings and missing macrons.

6 Experiment Results and Analysis

This section compares the performance of the newly-developed hybrid system with the standalone RMT (Trye et al., 2022) and ML (James

Tweets	Tweet Labels				Token-Level Errors (FP, FN)		
	Actual	RMT	ML	Hybrid	RMT	ML	Hybrid
1. Teaching ate me alive <link> via <user> #classroomreality	E	**B**	E	E	ate, me	-	-
2. <user> **ka pai**! Some **reo** and hugs! What more does one need:) #BFC630NZ	B	B	B	B	more, one	-	-
3. <user> <user> **Kia ora** Bronwyn. Hope to catch up while we are here!	B	B	B	B	hope, here	Kia	-
4. <user> **Ata marie** John, hope you're well mate.	B	B	B	B	marie, hope, mate	-	
5. **E hoa ma, nga mihi o te tau hou**! #Matariki #MaoriNewYear #BN-Zatm #respect <link>	M	M	**B**	M	-	E, o, tau	-
6. **Maori** Party welcomes **Waitangi** Tribunal report	B	B	B	B	-	Waitangi	Waitangi

Table 2: Example tweets indicating **actual Māori tokens**, **tweet-level errors** and unidentified Māori tokens.

System	Twitter Sample																		
	Tweet-Level													Token-Level					
	English				Māori				Bilingual				Overall	English			Māori		
	F1	P	R	S	F1	P	R	S	F1	P	R	S	Accuracy	F1	P	R	F1	P	R
RMT	0.06	**1.00**	0.03	**1.00**	0.39	1.00	0.24	**1.00**	0.91	0.83	**1.00**	0.10	0.84	0.90	0.93	0.87	0.87	0.88	0.85
ML	0.71	0.70	0.72	0.97	0.40	0.62	0.30	0.98	0.93	0.91	0.95	0.60	0.88	0.94	**0.94**	0.94	0.85	**0.96**	0.79
Hybrid	**0.89**	0.83	**0.97**	0.96	**0.51**	0.69	**0.40**	0.98	**0.95**	**0.95**	0.96	**0.78**	**0.93**	**0.95**	**0.94**	**0.95**	**0.94**	0.92	**0.97**
	Hansard Test Set																		
	Sentence-Level													Token-Level					
RMT	0.33	**0.71**	0.21	0.88	**0.96**	**1.00**	**0.91**	**1.00**	**0.95**	**0.91**	**0.95**	0.55	**0.92**	0.87	0.91	0.84	**0.86**	**0.86**	**0.86**
ML	**0.60**	0.43	0.97	**0.91**	0.32	**1.00**	0.19	0.99	0.79	0.90	0.70	0.55	0.68	0.92	0.91	0.91	0.66	0.70	0.64
Hybrid	0.52	0.35	**1.00**	0.89	0.38	**1.00**	0.24	0.99	0.85	**0.91**	0.79	**0.64**	0.77	**0.93**	**0.92**	**0.92**	0.71	0.73	0.70

Table 3: Tweet and token-level system evaluation for both the Twitter sample and Hansard test set. Recall (R), precision (P), F-score (F1), specificity (S) and overall accuracy are presented, with **best scores** emphasised.

et al., 2022b) systems. We also use a test set from the Hansard Dataset (James et al., 2022a) to evaluate our hybrid architecture with data from another domain. For brevity, we refer to interlingual homographs simply as *homographs*.

6.1 Visualisation of System Errors

To facilitate analysis of our manually-coded sample of tweets (hereafter, the *Twitter sample*), we have developed an interactive tool for comparing errors between the three systems of interest.[11] The visualisation helps users to explore the relationship between the tweet- and token-level labels for each system, and to better understand which kinds of tokens are responsible for the errors. Figure 4 provides a screenshot of this interactive tool, which guided the subsequent analysis.

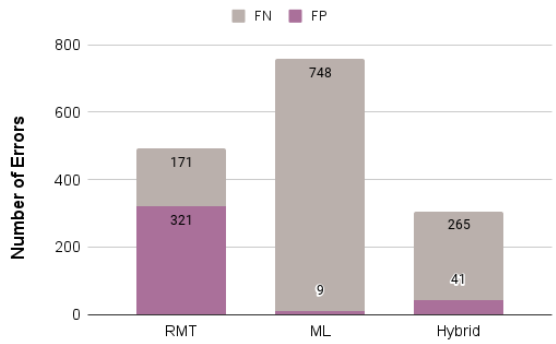
6.2 Overall Accuracy

Table 2 characterises the state of play for the hybrid system and the two existing systems, using six example tweets. All token-level errors are given, together with the resulting tweet labels. The token-level errors obtained using the RMT system's hand-crafted rules are mostly homographs, whereas those for the ML system are mostly Māori words. The hybrid architecture performs well by comparison, correctly identifying all but one Māori token.

Table 3 provides a synopsis of the system evaluations, broken down by tweet/sentence and token labels for both the Twitter sample and the Hansard test set. Looking at the Twitter sample, the Hybrid system has the highest overall accuracy. The Hybrid system's F1-scores are consistently better than the other two systems' at both the tweet and token level. The specificity of the Hybrid system is good across all tweet-level labels. Notably, the RMT system's specificity is extremely poor for bilingual tweets, indicating that the system is overly eager to find a positive result, even when it is not present. All systems do poorly at identifying Māori-only tweets; most are classified as Bilingual instead. This is likely because 'i' and 'a' are frequent in Māori but nearly always classified as English.

The Hansard test set included 10,000 bilingual, 1,000 Māori and 1,000 English sentences. The sentence-level accuracy for the RMT system is much better than the other systems; this is likewise true of the F1-scores for both Māori and bilingual sentences. One of the main reasons for this is that the test set contains predominantly bilingual sentences, and in most cases the RMT system identifies at least one Māori and English token. However,

[11] *https://bilingual-met.github.io/hybrid/sample*



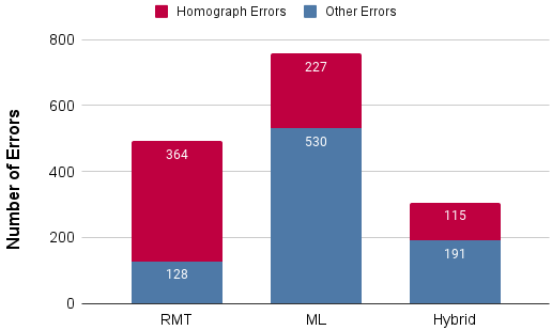


Figure 5: Token-level errors in the Twitter sample, showing **false positives**, **false negatives** and **homograph errors**.

System	False Positives	False Negatives
RMT	me, one, more, he, make, here, hope, take, o, nana, u	i, a, to, marie, no, ō, noho
ML	nana, ma	o, e, kia, i, he, a, tau, makaurau, waitangi, me, tūhoe, waatea, au, mo, kai, ō, to, kohanga, matatini, no, ā, morena, horipū, tuhoe
Hybrid	nana, ma, ufb	i, a, waitangi, waatea, to, no, tau, tuhoe

Table 4: Common token-level errors in the Twitter sample, including **homographs**.

the Hybrid system still has superior specificity for bilingual sentences. At the token-level, the Hybrid system does best for English tokens and the RMT system does best for Māori tokens.

6.3 Error Analysis

Figure 5 and Table 4 present a summary of token-level errors in the Twitter sample for all three systems, and highlight errors specifically caused by homographs. All systems struggle with short-length homographs (comprising fewer than five letters) like 'i' and 'a', which are pervasive in both languages. Nevertheless, the hybrid system fares considerably better than the other systems, with the ML and RMT systems having nearly double and over triple the number of homograph errors, respectively.

The vast majority of errors in the Hybrid system are Māori words that are mislabelled as English. Among these false negatives, short-length homographs constitute 42% of mistakes and named entities constitute 35%. While these are the two largest groups of errors, the Hybrid system still consistently classifies many of these kinds of words correctly (e.g. 'hope', 'Aotearoa').

System	Hansard Token-Level Errors
RMT	we, are, he, one, more, where, take, here, make, too, rate, none, rape, hope, reiterate, moe, mai, oki
ML	death, moe, mai, rā, hiamoe, kui, ki, te, pō, oti, atu, ai
Hybrid	moe, mai, rā, kui, ki, te, pō, oti, atu, ai

Table 5: Common token-level errors in the Hansard test set, including **homographs** mislabelled as 'M'.

These results indicate that the errors produced by the Hybrid system occur on a smaller scale than the ML system and are easier to fix than those for the RMT system. For instance, it is straightforward to update the labels for all tokens that contain non-Māori characters (like 'ufb'), and named entity accuracy (for tokens such as 'Waitangi') could be improved using an exhaustive list of non-ambiguous Māori place names.

A breakdown of the most prolific errors in the Hansard test set is given in Table 5. The most commonly misclassified homographs in both corpora are 'i', 'a', 'to' and 'no', which are all Māori particles that tend to be classified as English. Typically, such words are embedded inside larger segments of Māori text, so it is surprising that these instances are not correctly identified by our hybrid system's contextual check. One of the potential reasons is because the ML component of our hybrid architecture always classifies these tokens as English.

Like the Hybrid system, the ML system tends to mislabel Māori words as English rather than English words as Māori. Many of the same kinds of errors occur, though there are more false negatives and fewer false positives. The ML system frequently misclassified the particles 'e', 'o' and 'kia' in phrases such as "Miharo **e** hoa!", "Te Wiki **o** Te Reo Maori" and "**kia** ora". In contrast, the Hybrid system always labelled these correctly.

The RMT system differs from the others in that it has more false positives than false negatives. As a rule-based system, it always assigns the same label to each word type, even if it is valid in both languages. Words that are consistent with Māori orthography are generally tagged as Māori; as a result, the RMT system is considerably better at correctly classifying Māori named entities, including personal and place names. However, the RMT system performs considerably worse than the other two when classifying tweets with a large proportion of English text. Over 85% of false positives are short-length homographs, with 'me', 'one', 'more', 'he', 'make' and 'here' being the worst offenders. Like the other two systems, there are also some instances of Māori words that are misclassified as English (especially 'i', 'a', and 'to'), due to the stoplist that was used.

7 Limitations

The research presented in this paper has some limitations that need to be acknowledged. The hybrid architecture uses a single neural network-based model, but we have experimented with variations in the neural networks and parameter choices. Given the available data and resources, bidirectional RNNs performed the best.

We found that our hybrid architecture does not label Māori named entities consistently, and short-length homographs like 'i' and 'a' are problematic. This requires further investigation, perhaps involving a special look-up for Māori place names, and ensuring that a context check is always carried out for frequent homographs, especially function words.

In addition, our approach for identifying foreign-language tweets is not exhaustive, and in some cases, tokens that are neither Māori nor English will have been erroneously labelled as such. Our foreign-language processing currently focuses on manually identifying problematic tweets in a small subset of the data, then extrapolating this into the wider dataset. This approach could be further developed, or a more automated system could be implemented.

Our labels do not distinguish between borrowings and code-switches (Álvarez Mellado and Lignos, 2022). This means it is not possible to automatically extract tweets where Māori borrowings are used in otherwise English contexts, or vice versa, although the number of tokens identified in each language could serve as a useful proxy.

Finally, we discarded a proportion of the collected tweets as our algorithm was not optimised for dealing with undue levels of noise. The discarded tweets with unknown labels are not vital to the MET Corpus presented in this research; however, they require further investigation, and may constitute useful additions to the corpus.

8 Conclusions and Future Work

This paper presents an architecture for labelling bilingual Māori-English text, by bringing together machine learning and knowledge of Māori orthography, an approach that could also be fruitful for other endangered languages. We use this architecture to create the first large-scale corpus of bilingual Māori-English tweets annotated at both the token and tweet level. Both this corpus and the Hansard Dataset are used to illustrate the strengths of our approach, including superior token-level accuracy, especially with respect to interlingual homographs. In particular, the specificity scores for bilingual data favour the Hybrid system, while highlighting a major weakness of the RMT system. Additional insights can be gleaned from two exploratory visualisations for interrogating the corpus and comparing system errors.

Future work towards enhancing the bilingual corpus could involve extending this research to classify hashtags as these are currently ignored. Moreover, the architecture lends itself to annotating other bilingual datasets, such as the MLT Corpus (Trye et al., 2019), and could assist in the creation of new resources. A further avenue of exploration would be assigning part-of-speech tags to each token in the corpus, based on the language identified. This could be achieved using newly-developed tools for Māori (Finn et al., 2022) in conjunction with established part-of-speech taggers for English. Such developments are important for ensuring better representation of the Māori language in digital applications and environments.

Acknowledgements

We are indebted to researchers from the University of Auckland and Te Hiku Media for kindly sharing the Hansard Dataset. We thank Andreea Calude and three anonymous reviewers for their helpful comments and suggestions. DT acknowledges funding from the University of Waikato Doctoral Scholarship.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Andreea Calude, Louise Stevenson, Hēmi Whaanga, and Te Taka Keegan. 2020. The use of Māori words in national science challenge online discourse. *Journal of the Royal Society of New Zealand*, 50(4):491–508.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Ton Dijkstra. 2007. Task and context effects in bilingual lexical processing. In *Cognitive aspects of bilingualism*, pages 213–235. Springer.

Jonathan Dunn and Wikke Nijhof. 2022. Language identification for Austronesian languages.

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2022. Developing a part-of-speech tagger for te reo Māori. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 93–98.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*, pages 1–11. Rootroo Ltd.

Ray Harlow. 1993. Lexical expansion in Maori. *The Journal of the Polynesian Society*, 102(1):99–107.

Rawinia Higgins and Basil Keane. 2015. Te reo Māori – the Māori language - language decline, 1900 to 1970s', Te Ara - the encyclopedia of New Zealand.

Janet Holmes and Nick Wilson. 2017. *An introduction to sociolinguistics*. Routledge.

Jesin James, Isabella Shields, Rebekah Berriman, Peter J Keegan, and Catherine I Watson. 2020. Developing resources for te reo Māori text to speech synthesis system. In *International Conference on Text, Speech, and Dialogue*, pages 294–302. Springer.

Jesin James, Isabella Shields, Vithya Yogarajan, Peter J. Keegan, Catherine Watson, Peter-Lucas Jones, and Keoni Mahelona. 2022a. The development of a labelled te reo Māori-English bilingual database for language technology.

Jesin James, Vithya Yogarajan, Isabella Shields, Catherine Watson, Peter Keegan, Peter-Lucas Jones, and Keoni Mahelona. 2022b. Language models for code-switch detection of te reo Māori and English in a low-resource setting. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 136–141. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, pages 427–431.

Te Taka Keegan. 2017. Machine translation for te reo Māori. In Hemi Whaanga, Te Taka Keegan, and Mark Apperley, editors, *He Whare Hangarau Māori Language, Culture & Technology*, pages 23–28. Te Pua Wānanga ki te Ao/Faculty of Māori and Indigenous Studies.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84.

Joanne Marras Tate and Vaughan Rapatahana. 2022. Māori ways of speaking: Code-switching in parliamentary discourse, Māori and river identity, and the power of Kaitiakitanga for conservation. *Journal of International and Intercultural Communication*, pages 1–22.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Vaughan Rapatahana. 2017. English language as thief. In *Language and Globalization*, pages 64–76. Routledge.

Linda Tuhiwai Smith. 2021. *Decolonizing methodologies: Research and indigenous peoples*, third edition. Bloomsbury Publishing.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

David Trye, Andreea S Calude, Felipe Bravo-Marquez, and Te Taka Keegan. 2020. Hybrid hashtags: #YouKnowYoureAKiwiWhen your tweet contains Māori and English. *Frontiers in artificial intelligence*, 3:15.

David Trye, Andreea S Calude, Felipe Bravo-Marquez, and Te Taka Adrian Gregory Keegan. 2019. Māori loanwords: a corpus of New Zealand English tweets. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 136–142.

David Trye, Te Taka Keegan, Paora Mato, and Mark Apperley. 2022. Harnessing indigenous tweets: The Reo Māori Twitter corpus. *Language resources and evaluation*, pages 1–40.

Te Hau White. 2016. A difference of perspective? Māori members of parliament and te ao Māori in parliament. *Political Science*, 68(2):175–191.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires re-thinking generalization. In *Proc. International Conference on Learning Representations 2017*, pages 1–15.

Elena Álvarez Mellado and Constantine Lignos. 2022. Borrowing or codeswitching? Annotating for finer-grained distinctions in language mixing.

A Algorithms

Algorithm 1 Token-Level Labelling

Input: Pre-processed tweets, list of Māori labels obtained from RMT system, pre-trained ML model, and tokenizer
Output: Labels at token-level
class_label = [ML model output]
english_list = [tokens with class_label 'E']
maori_list = [tokens with class_label 'M']
rmt_list = [Māori tokens from RMT system]
ambiguous_list = [rmt_list ∩ english_list]
if len(ambiguous_list) != 0 **then**
    Remove ambiguous tokens from rmt_list & english_list
end if
for each tweet i **do**
    for each token j in i **do**
        if j in english_list **then**
            if j is detected as an English word using fastText and NLTK language detection tools **then**
                Assign label for j as **E (English)**
            end if
        else if j in rmt_list **then**
            if j in maori_list **then**
                Assign label for j as **M (Māori)**
            end if
        else if j in ambiguous_list **then**
            Assign label for j as **A (Ambiguous)**
        else if Token j not in 'E', 'M', 'A' **then**
            Assign label for j as **U (Unknown)**
        end if
    end for
end for

Algorithm 2 Context-Check for Ambiguous Items

Input: Pre-processed tweet tokens, list of Māori tokens, English tokens, and Ambiguous tokens obtained from token-level labelling
Output: Updated labels at token-level
for each tweet t **do**
    maori_list = [Māori words in t]
    english_list = [English words in t]
    ambiguous_list = [Ambiguous words in t]
    tokens = [all tokens in t]
    if len(ambiguous_list) != 0 **then**
        for amb_token in ambiguous_list **do**
            if amb_token contains {ā,ē,ī,ō,ū} **then**
                Assign label as **M (Māori)**
                Remove from ambiguous_list
            else
                before = tokens[index-1]
                after = tokens[index+1]
                before_before = tokens[index-2]
                after_after = tokens[index+2]
                if before & after in maori_list **then**
                    Assign label as **M (Māori)**
                    Remove from ambiguous_list
                else if before & after in english_list **then**
                    Assign label as **E (English)**
                    Remove from ambiguous_list
                else if before is null, i.e. amb_token is the first token in the tweet **then**
                    if after & after_after in maori_list **then**
                        Assign label as **M (Māori)**
                        Remove from ambiguous_list
                    else if after & after_after in english_list **then**
                        Assign label as **E (English)**
                        Remove from ambiguous_list
                    end if
                else if after is null, i.e. amb_token is the last token in the tweet **then**
                    if before_before & before in maori_list **then**
                        Assign label as **M (Māori)**
                        Remove from ambiguous_list
                    else if before_before & before in english_list **then**
                        Assign label as **E (English)**
                        Remove from ambiguous_list
                    end if
                end if
            end if
        end for
    end if
end for

Algorithm 3 Tweet-Level Labelling

Input: Bilingual tweets with token-level labels obtained using Algorithm 1 and Algorithm 2
Output: Labels at tweet-level
for each tweet t **do**
    maori_list = [Māori words in t]
    english_list = [English words in t]
    unknown_list = [Unknown words in t]
    ambiguous_list = [Ambiguous words in t]
    if len(maori_list) == 0 & len(unknown_list) == 0 & len(ambiguous_list) == 0 **then**
        tweet_label of t is **E (English)**
    else if len(english_list) == 0 & len(unknown_list) == 0 & len(ambiguous_list) == 0 **then**
        tweet_label of t is **M (Māori)**
    else if len(ambiguous_list) == 0 & len(unknown_list) == 0 **then**
        tweet_label of t is **B (Bilingual)**
    else
        tweet_label of t is **O (Other)**
    end if
end for
for each tweet t **do**
    label_ML = ML tweet-label for t
    if label_ML == tweet_label **then**
        Final tweet-level label for MET Corpus
    else
        Further investigation needed
    end if
end for