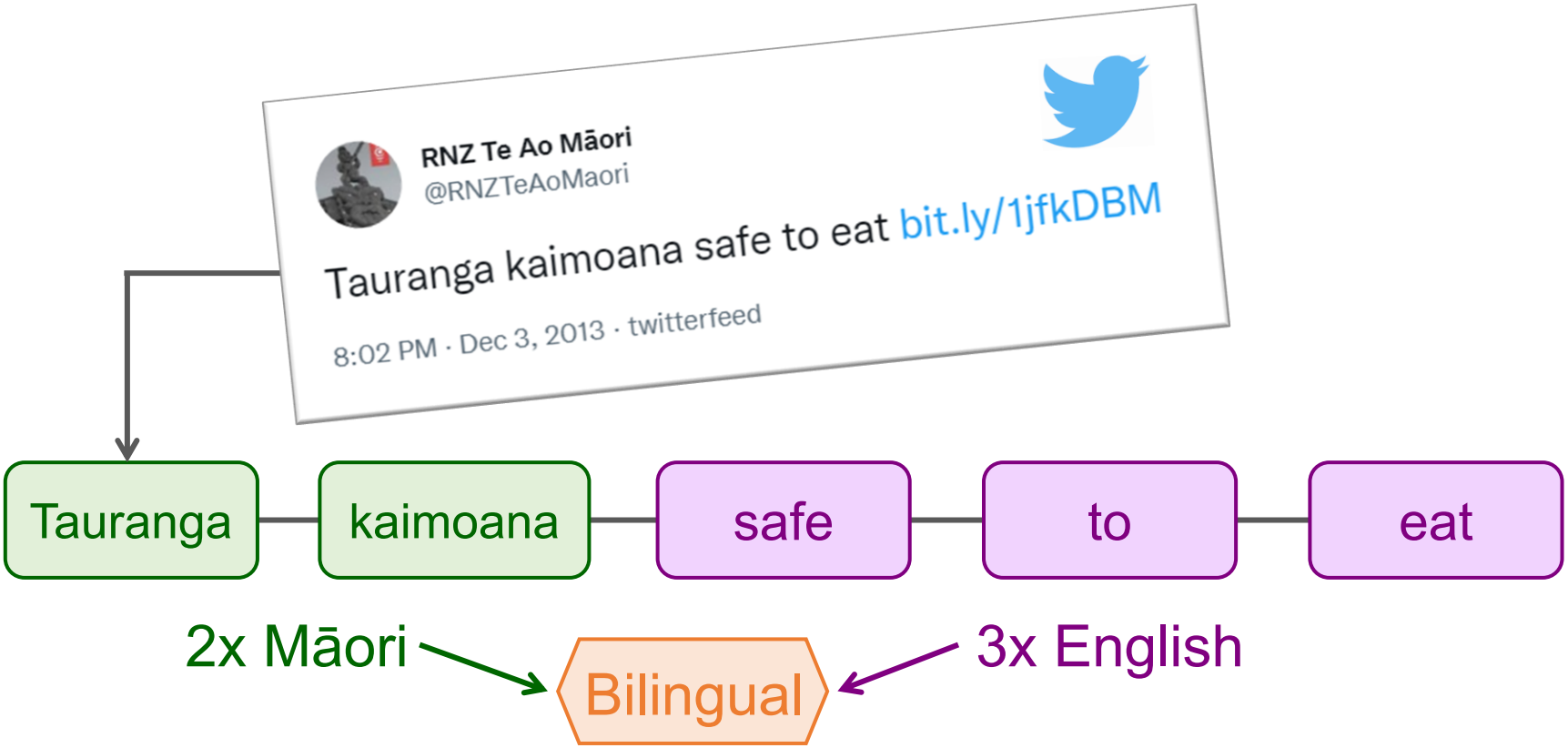
A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

David Trye, Vithya Yogarajan, Jemma König, Te Taka Keegan,
David Bainbridge & Mark Apperley

Research Aim

- To improve automatic **language identification** for Māori-English text
  - Focusing on (**noisy**) Twitter data



Contributions

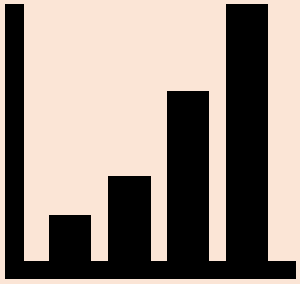
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

System

MET Corpus

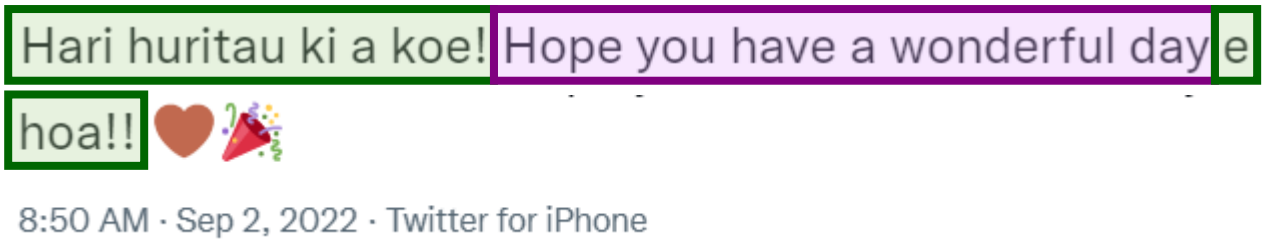
Evaluation

Vis Tools

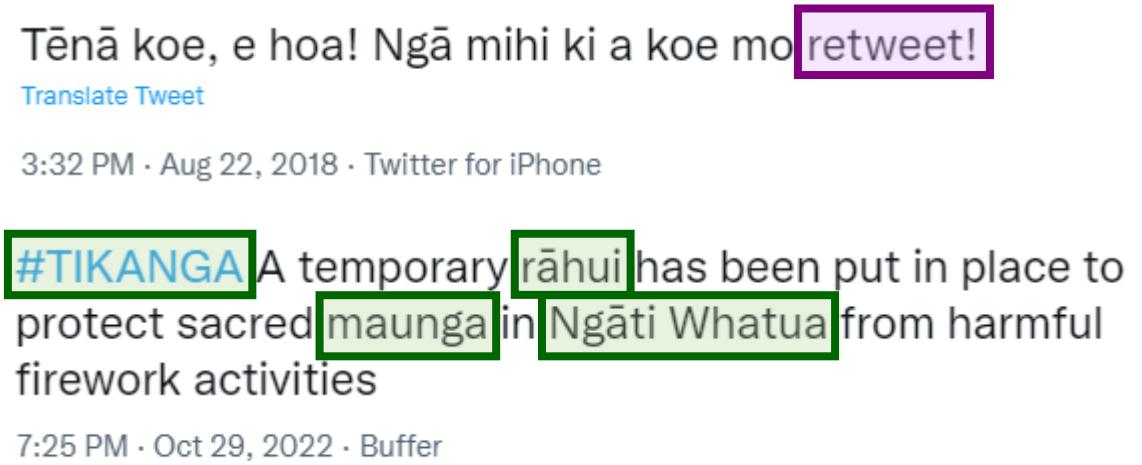


Background

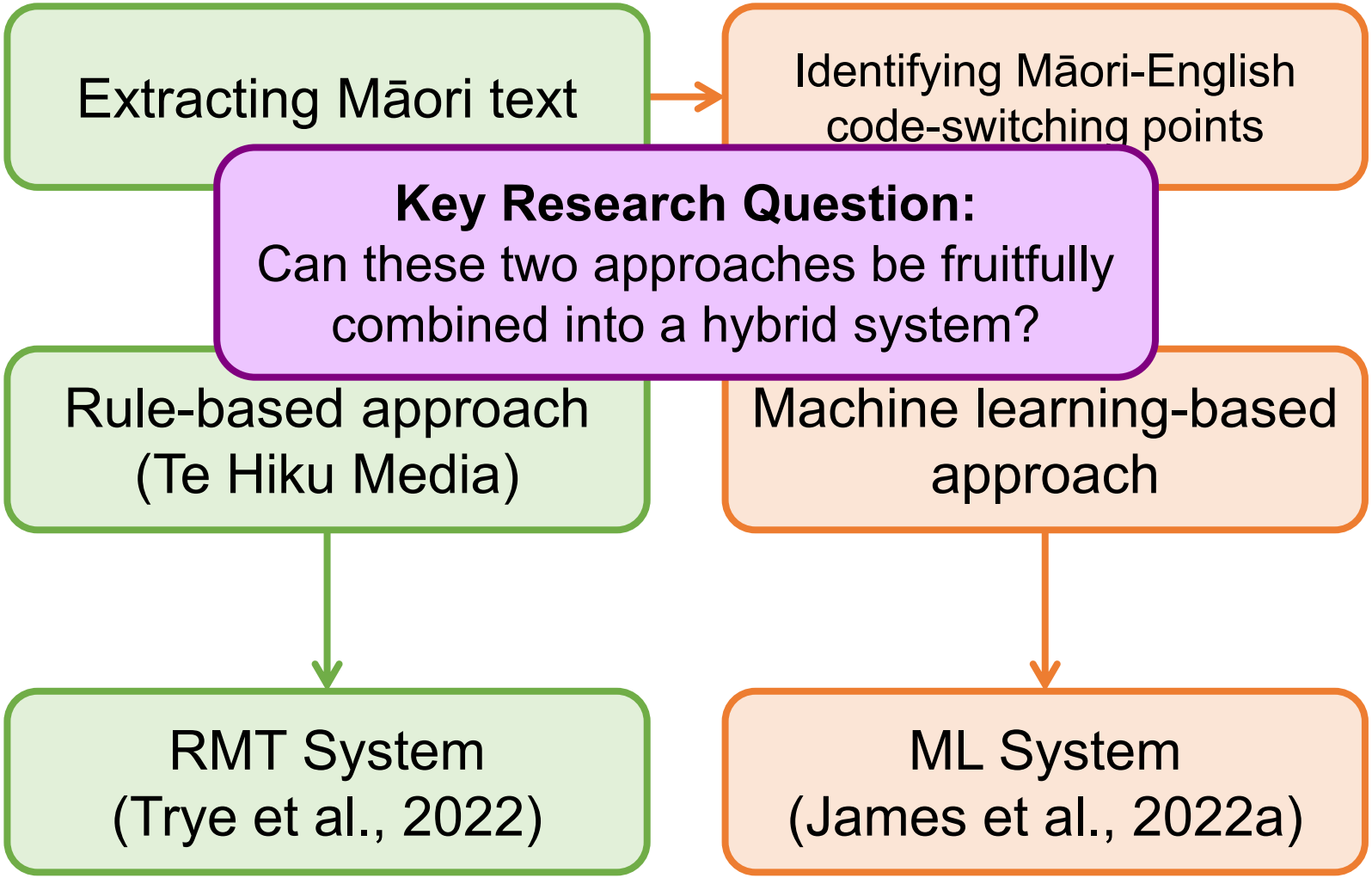
- Te reo Māori is often interspersed with English
  - **Code-switching:** ‘multi-word stretches’ (Poplack, 2018)



  - **Loanwords / borrowings:** (mostly) individual words



Related Work



The Bigger Picture

- Te reo Māori is fundamental to **Māori culture**
- Both Māori and New Zealand English are **under-represented** in speech and language technology
  - There is a critical need for new systems and resources to address this (James et al., 2020, 2022a)
- Existing NLP tools are **biased** towards (certain varieties of) English (Hovy & Prabhumoye, 2021)
  - These tools often fail to recognise or correctly spell/pronounce Māori words ("Kaitaia" → "Car Tyre")
- Our goal is to **reduce this inequity** in NLP resources

Key Challenges

- Lexical overlap
  - Both Māori and English use the Roman script
  - 100+ **interlingual homographs**
    - Words that are spelt the same but have different meanings across languages (Dijkstra, 2007)
    - *i, a, hope, here, more, kite,* etc.
- Social media language
  - Internet slang, abbreviations, acronyms
    - *haha, ktk (Māori equivalent of lol), amirite, cuzzie*
  - Misspellings, typos
  - Neologisms
  - Emojis, hashtags, GIFs, etc.

RMT System

- Based on rules by Te Hiku Media
- Tokens must contain valid Māori **characters**
  - 5 vowels (*i, e, a, o, u*)
  - 10 consonants (*p, t, k, m, n, ng, wh, r, w, h*)
- Tokens must follow Māori **syllable structure**
  - Consonant/vowel alternation: *(C)V(V)*, *(C)*$V_1V_1V_2$
  - No consonant clusters
  - End with a vowel
- Lengthened vowels may be indicated with a macron (*ā*) or double vowels (*aa)*

ML System

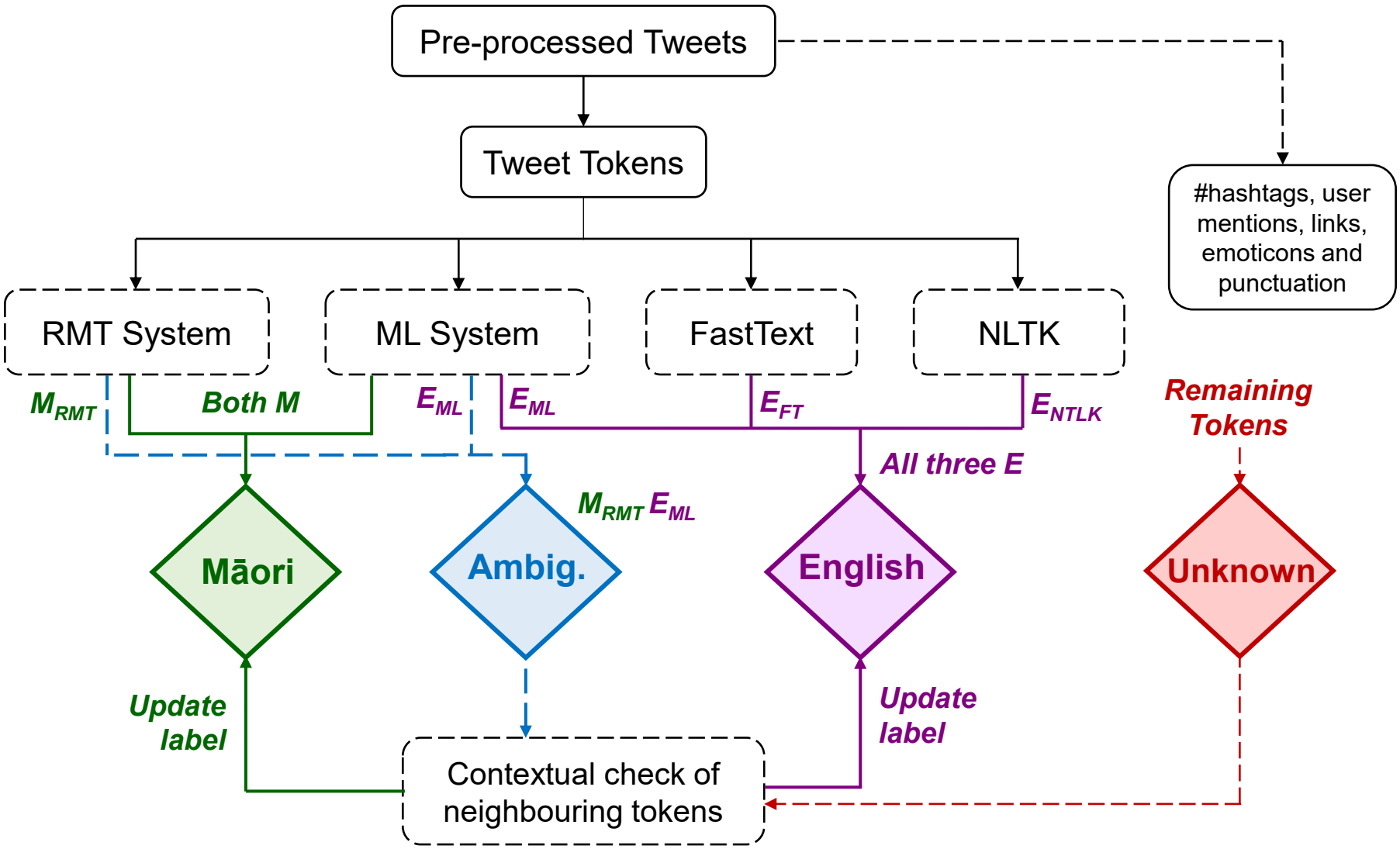
- Bidirectional Gated Recurrent Units (**Bi-GRU**)
  - Attention layer based on Bahdanau mechanism
  - Trained on Hansard dataset (James et al., 2022b)
- Text represented using **fastText** word embeddings
  - Skip-gram model with 300 dimensions
  - Pre-trained on Māori & Māori-English corpora (James et al., 2022a)
- Model trained to predict **M/E/B tweets**
  - Networks optimised with *Adam* (Kingma and Ba, 2015)
  - Softmax activation in output layer
  - Dropout rate of 0.5 and early stopping used

Pre-processing

- Collected tweets comprising roughly **30-80% Māori text** from known Māori-language users
  - Users identified via *Indigenous Tweets* (Scannell, 2022)
- Tweets were subsequently **cleaned**
  - Stripped non-Roman characters (漢字)
  - Standardised user mentions (@user) & links (<link>)
  - Expanded English contractions (isn't → is not)
- Discarded ~40,000 **irrelevant tweets**
  - Retweets, bots, duplicates, short tweets (<4 tokens)
  - Tweets containing other languages (not exhaustive)

īhi! Heivā i Tahiti ! Te ineine mai ra ! Pā'oti i ni'a, pā'oti i raro, tīfene, tīfene, 'ami, 'ami

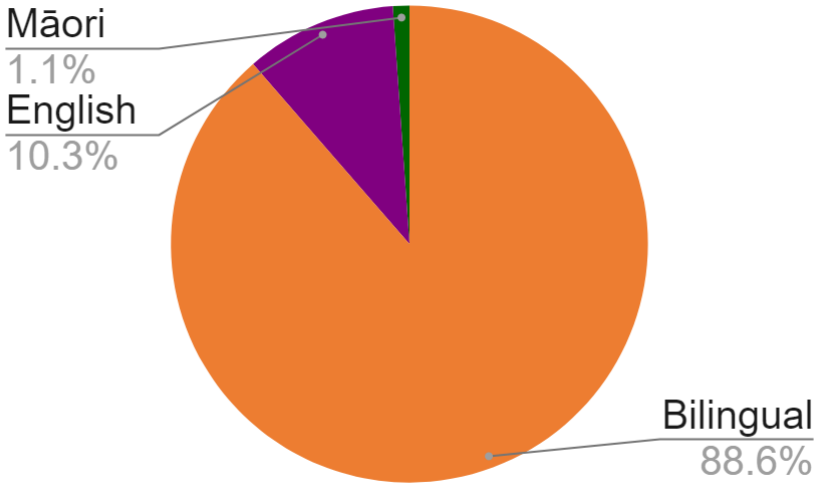
Token-Level Labels



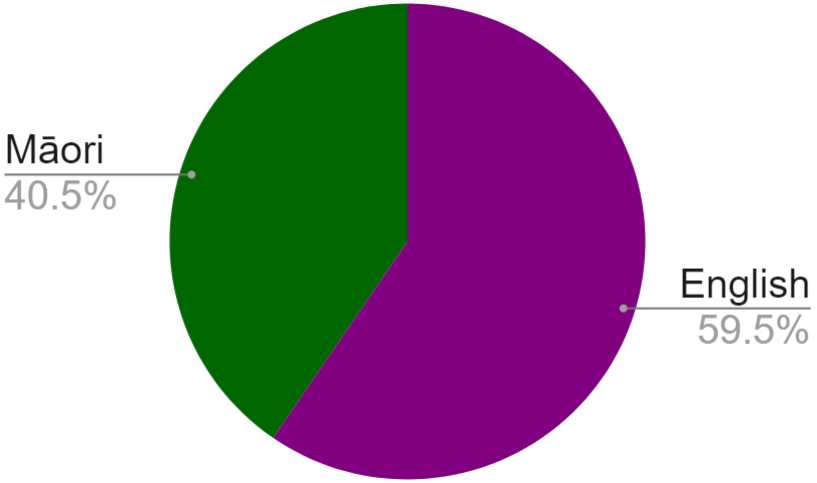
MET Corpus Summary

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

76,416 tweets

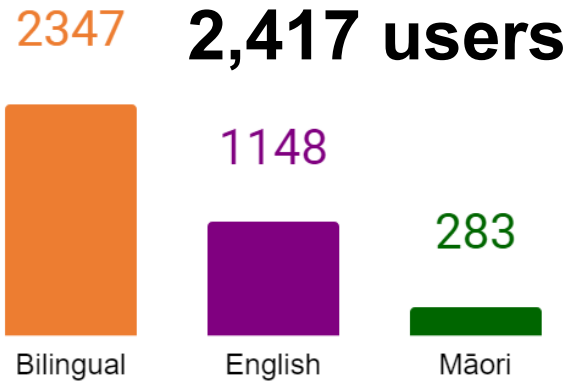


781,381 tokens



Limitation: Many tweets were filtered out of the corpus to improve accuracy, such as tweets with one or more ‘Unknown’ or ‘Ambiguous’ labels

2,417 users

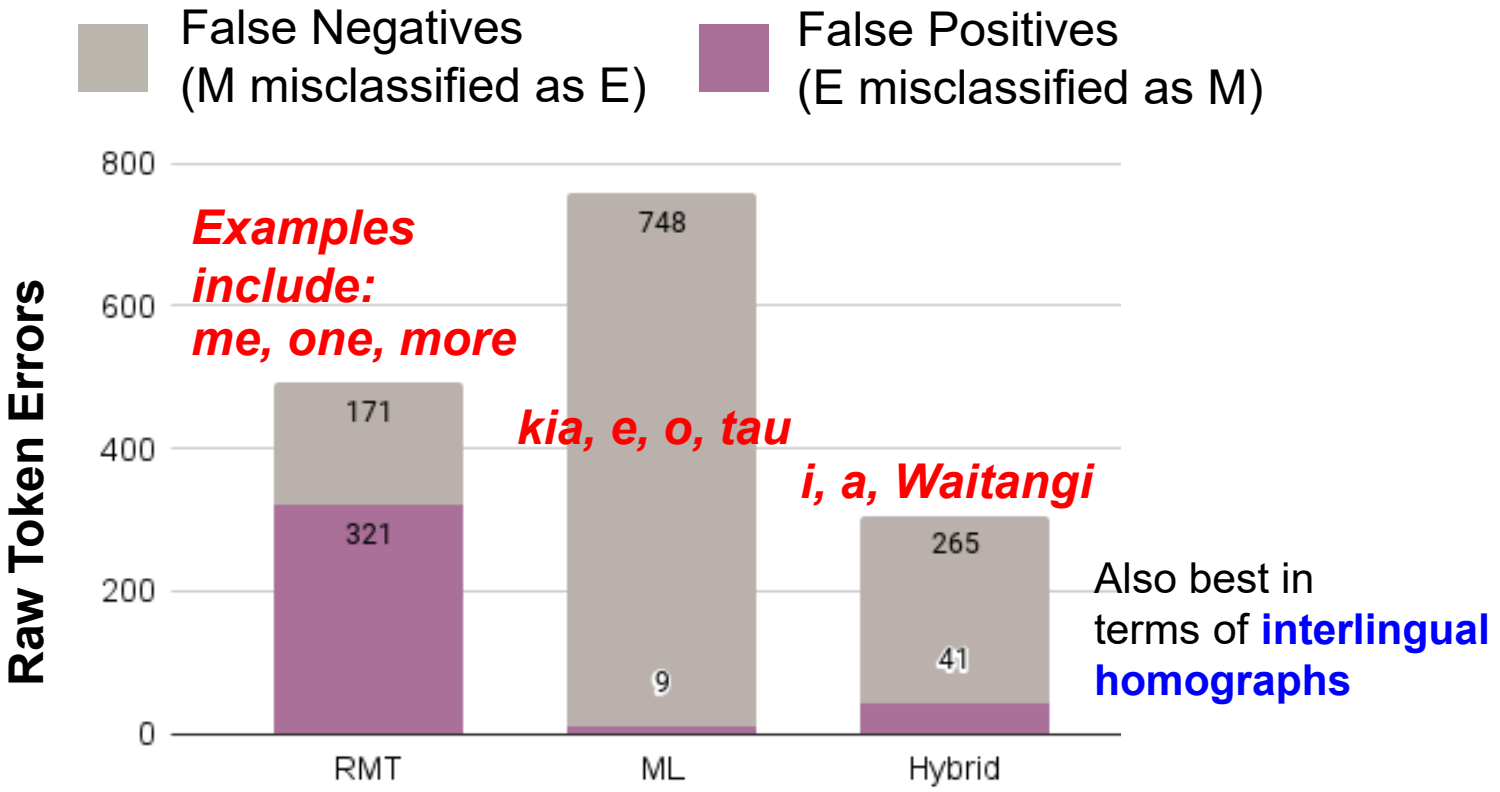


Manual Annotation

- We manually labelled **850 tweets** for evaluation purposes
  - All three systems (RMT, ML & Hybrid) at both the token and tweet level
  - Strong agreement between annotators
    - Cohen's κ = 0.816 for a subsample of 200 tweets
- Recorded information about each mistake
  - False negative (FN) or false positive (FP)?
  - Specific **error type**
    - Interlingual homograph
    - Named entity (person, place, iwi, organisation, event, etc.)
    - Illegal character(s)
    - Misspelling or missing macron(s)

Evaluating our system

- Hybrid system had the fewest token-level errors, followed by RMT system



Evaluation Metrics

Token-Level

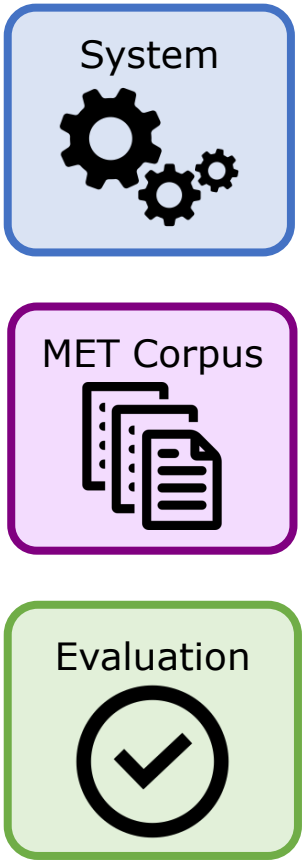
	F1-Score		Precision		Recall	
	E	M	E	M	E	M
RMT	0.90	0.87	0.93	0.88	0.87	0.85
ML	0.94	0.85	**0.94**	**0.96**	0.94	0.79
Hybrid	**0.95**	**0.94**	**0.94**	0.92	**0.95**	**0.97**

Tweet-Level

	F1-Score			Specificity			Overall Accuracy
	E	M	B	E	M	B	
RMT	0.06	0.39	0.91	**1.00**	**1.00**	0.10	0.84
ML	0.71	0.40	0.93	0.97	0.98	0.60	0.88
Hybrid	**0.89**	**0.51**	**0.95**	0.96	0.98	**0.78**	**0.93**

Wrapping Up

- We devised a **novel system** for labelling Māori/English text
- We used this system to create an **annotated corpus** of 76,000 tweets
- These developments can facilitate further NLP research for Māori and New Zealand English
- This work could also be impactful for research in other low-resourced languages



Thanks for listening!

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

Check out our interactive visualisation tools:

- https://bilingual-met.github.io/hybrid/
- https://bilingual-met.github.io/hybrid/sample

Contact me
David Trye
dgt12@students.waikato.ac.nz

References

- Dijkstra, T. (2007). Task and context effects in bilingual lexical processing. In *Cognitive aspects of bilingualism* (pp. 213-235). Springer, Dordrecht.
- Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8).
- James, J., Shields, I., Berriman, R., Keegan, P. J., & Watson, C. I. (2020, September). Developing resources for te reo Māori text to speech synthesis system. In *International Conference on Text, Speech, and Dialogue* (pp. 294-302). Springer, Cham.
- James, J., Yogarajan, V., Shields, I., Watson, C. I., Keegan, P., Mahelona, K., & Jones, P. L. (2022a). Language Models for Code-switch Detection of te reo Māori and English in a Low-resource Setting. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 650-660).
- James, J., Shields, I., Yogarajan, V., Keegan, P. J., Watson, C., Jones, P. L., & Mahelona, K. (2022b). The Development of a Labelled te reo M\= aori-English Bilingual Database for Language Technology. *arXiv preprint arXiv:2208.09778*.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Poplack, S. (2018). *Borrowing: loanwords in the speech community and in the grammar.* Oxford: Oxford University Press.
- Scannell, K. P. (2022). 41 Managing Data from Social Media: The Indigenous Tweets Project. *The Open Handbook of Linguistic Data Management*, 481.
- Trye, D., Keegan, T. T., Mato, P., & Apperley, M. (2022). Harnessing Indigenous Tweets: The Reo Māori Twitter corpus. *Lang Resources & Evaluation, 56*, 1229-1268. doi:10.1007/s10579-022-09580-w