# Performant Text Classification with Naive Bayes for the Kaqchikel Mayan language

William J. Wakefield

March 2023

## Abstract

Kaqchikel is a language in the Mayan language family, spoken in Guatemala by about 410,000 people. As of this writing, there are no existing Natural Language Processing (NLP) application for this Mayan language, which leads to a lack of tools that could be implemented to help preserve the language in an advancing and globalized world. The text classification will use a Term Frequency- Inverse Document Frequency (TF-IDF) parallel model and a parallel Naive Bayes algorithm to be evaluated on the Kaqchikel Chronicles, a collection or rare pre-colonial texts. The NLP pipeline developed may make contributions to TextAnalysis.jl, MLJ.jl, and within the Julia NLP ecosystem in general. The code can be found here.

## NLP Pipeline

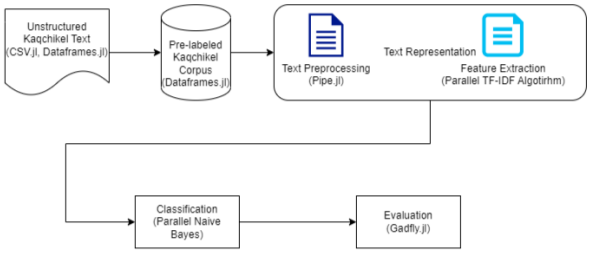


Figure 1: NLP Pipeline

## Corpus

The corpus used involves one novel corpus which "was constructed from existing religious texts, spoken transcripts, government documents, medical handbooks, and other educational books written in Kaqchikel" [11][12]. As from the metadate of the corpus, it contains approximately 0.7 million word tokens and 29,355 word types[11][12]. While the novel corpus is great as a huge potential for data mining purposes thank to the large extent of word tokens available, labeling the corpus for sentiment analysis and ensuring proper translation was not able to be done in time for the scope of the project. Instead, labeling sentiment and ensuring a one-to-one mapping for transla-

tion purposes, the other corpus used is the Kaqchikel Chronicles [7]. The issue with the Kaqchikel chronicles to note is that since it is mostly a religious and historical texts with a lot of it written in a poetic way in Kaqchikel and more formal, it would not be as accurate to how the modern Kaqchikel Maya speak as it would have done with examples of speeches from the novel corpus from Dr. Tang and Dr. Bennett. If you would like to view the novel corpus itself, please reach out to Dr. Tang and Dr. Bennett at [11] or [12].

## Procedure

1. Create a Kaqchikel Database from dictionaries or existing databases
2. Perform text preprocessing by removing cases, numbers, HTML tags and punctuation (except glottals?)
3. Create the TF-IDF matrix with TextAnalysis.jl
4. Create the Corpus object from the matrix
5. Pass Corpus object to default Naive Bayes Classifier from TextAnalysis.jl
6. Evaluate the Pipeline
7. Parallelize and optimize the TF-IDF matrix portion in terms of feature extraction [5]
8. Parallelize and optimize the TF-IDF matrix portion in terms of feature extraction [5]
9. Compare results with the new pipeline with previous pipeline

## Parallel Naive Bayes Classifier

$$P(c|x) = P(x|c) * P(c)/P(x)$$

The idea for the parallel Naive Bayes Classifier comes from source [5] and [10].

## Contributions to the Julia Text Ecosystem

While still in the initial stages, here is the pull request for adding Kaqchikel data to the Julia NLP libraries here. The pull request will allow TextAnalysis.jl to then perform the TF-IDF matrix and use the naive bayes classifier function on Kaqchikel.

## References


[1] TextAnalysis.jl documentation by Julia Hub — https://docs.juliahub.com/TextAnalysis/5Mwet/0.7.3/

[2] Vectorize everything with Julia by Bence Komarniczky — https://towardsdatascience.com/vectorize-everything-with-julia-ad04a1696944

[3] MLJ framework — a machine learning framework of Julia: https://alan-turing-institute.github.io/MLJ.jl/dev/

[4] MLJ Data Interpretation and Scitypes: https://juliaai.github.io/DataScienceTutorials.jl/data/scitype/

[5] Houda Amazal, Mohammed Ramdani, and Mohamed Kissi. 2018. A Text Classification Approach using Parallel Naive Bayes in Big Data Context. In Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications (SITA'18). Association for Computing Machinery, New York, NY, USA, Article 36, 1–6. https://doi.org/10.1145/3289402.3289536

[6] Annals of the Cakchiqueles https://www.gutenberg.org/files/20775/20775-h/20775-h.htm#THE_ANNALS

[7] Maxwell, Judith M. and Robert M., II Hill. Kaqchikel Chronicles: The Definitive Edition. University of Texas Press, 2006. Project MUSE muse.jhu.edu/book/45051.

[8] Naive Bayes Example https://www.geeksforgeeks.org/applying-multinomial-naive-bayes-to-nlp-problems/

[9] A corpus of K'iche' annotated for morphosyntactic structure (Tyers & Henderson, AmericasNLP 2021)

[10] Mapreduce function in Julia to parallelize the preprocessing and naive bayes classifier (https://docs.julialang.org/en/v1/base/collections/#Base.mapreduce-Tuple{Any,%20Any,%20Any})

[11] Tang, K., & Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (Kaqchikel Mayan). The Journal of the Acoustical Society of America, 144(2), 997-1017.

[12] Bennett, R., Tang, K., & Sian, J. A. (2018). Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan). Laboratory Phonology, 9(1), 9-9.