Data Augmentation Approach for Language Identification in Imbalanced Bilingual Code-Mixed Social Media Datasets

Mohd Suhairi Md Suhaimin [1,2], Mohd Hanafi Ahmad Hijazi [1,3,*], Ervin Gubin Moung [1], Mohd Azwan Mohamad Hamza [4]

[1] Data Technology and Applications Research Group
Faculty of Computing and Informatics, Universiti Malaysia Sabah
Sabah, Malaysia

[2] Polytechnic and Community College Education Department
Ministry of Higher Education Malaysia
Putrajaya, Malaysia

[3] Creative Advanced Machine Intelligence Research Centre
Faculty of Computing and Informatics, Universiti Malaysia Sabah
Sabah, Malaysia

[4] Knowledge Engineering & Computational Linguistic (KECL) Research Group
Universiti Malaysia Pahang
Pahang, Malaysia
Email: mohd_suhairi_di21@iluvums.edu.my, hanafi@ums.edu.my, ervin@ums.edu.my, azwan@ump.edu.my

2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) | 979-8-3503-0415-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/IICAIET59451.2023.10292108

***Abstract*—Addressing the problem of language identification in code-mixed datasets poses notable challenges due to data scarcity and high confusability in bilingual contexts. These challenges are further amplified by the associated imbalance and noise characteristic of social media data, complicating efforts to optimize performance. This paper introduces an augmentation approach designed to enhance language identification in bilingual code-mixed social media data. By incorporating reverse translation, semantic similarity, and sampling techniques alongside customized preprocessing strategies, our approach offers a comprehensive solution to these complex issues. To evaluate the effectiveness of the proposed approach, experiments were conducted on language identification at both the sentence and word levels. The results demonstrated the potential of the approach in optimizing language identification performance, offering a compelling combination of generation techniques for addressing the challenges of language identification in code-mixed data.**

Keywords—language identification, code-mixing, data augmentation, social media

I. INTRODUCTION

Code-mixing[1], a phenomenon where words or phrases from multiple languages are combined to convey a message, is a longstanding occurrence in bilingual and multilingual communities due to tradition and commonplace. This blending of languages often materializes in written text communication, public opinion expression, and information transmission on social media platforms. In this environment, users' posts naturally alternate between languages, increasing the complexity of language identification tasks. The role of language identification serves as the critical initial step in processing specific languages for natural language processing (NLP) tasks. It proves beneficial in preprocessing bilingual code-mixing data, enabling accurate spellchecking, and facilitating effective feature extraction. However, low accuracy in language identification can potentially weaken the performance of subsequent NLP tasks [1].

Building a language identification model for code-mixing language poses a considerable challenge due to the unpredictable nature of language mixing and switching in written text [1]. Augmenting existing datasets can address the data scarcity and expense associated with large-scale code-mixing data production. Nonetheless, this solution presents new hurdles such as dealing with imbalanced and noisy data from social media platforms. Imbalanced data refers to a significant discrepancy in the number of words available from each language class, with some languages heavily represented while others scarcely [2]. This issue is further complicated in the context of mixed-code data embedded in noisy social media content, characterized by typos, slang, abbreviations, and emoticons, which worsen the language identification task [3]. Data augmentation, however, can provide the model with a broader, more representative range of language data, thereby enhancing its overall predictive capabilities.

The aim of this paper is to introduce a data augmentation approach explicitly designed for imbalanced bilingual mixed-code data, and later to empirically measure its performance in improving language identification within the noisy context of social media.

The main contribution of this paper is an approach for handling imbalanced code-mixing social media data, with the potential to enhance language identification systems through improved performance. This proposed approach could serve as the foundational approach for subsequent NLP tasks involving code-mixing data. The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the proposed approach and Section 4 details the dataset and experiments conducted. Section 5 discusses the result and analysis, while Section 6 concludes the paper and suggests a potential direction for future work.

*Corresponding Author: Mohd Hanafi Ahmad Hijazi
Ministry of Higher Education Malaysia [FRGS/1/2022/ICT02/UMS/02/3]

[1] Code-mixing refers to intra-sentential mixing, whereas code-switching refers to inter-sentential mixing of languages. Given the former is more general, we use code-mixing in this paper to encompass both.