PolitiKweli: A Swahili-English Code-switched Twitter Political Misinformation Classification Dataset

Cynthia Amol[1] Lilian Wanzare[2] and James Obuhuma[3]
[1] Maseno University, Kenya
cynthia@maseno.ac.ke
[2] Maseno University, Kenya
[3] Maseno University, Kenya

1. Abstract

In the age of freedom of speech, users of the social media platform Twitter post mil-lions of messages per day. These messages are not always fact-checked resulting in misinformation which is false or misleading news. Misinformation classification involves identifying and classifying text as either false or fact by comparing the text against fact-checked news. On political matters, misinformation online can result in mistrust of political figures, polarization of communities and violence offline. Existing studies mostly address misinformation detection for messages written in a single language such as English. Among most bilingual or multilingual user groups in countries like Kenya, the use of Swahili-English code-switching and code-mixing is a common practice in informal text-based communication such as messaging on social media platforms like Twitter. There is therefore need for more research in low-resource languages such as Swahili. The PolitiKweli[1] dataset introduced by this study, which a novel Swahili-English misinformation classification dataset, contains 6,345 Swahili-English texts, 22,954 English texts and 211 Swahili texts. The texts are labelled as true, false or neutral as compared to a fact-checked dataset also created for this study. The dataset curation process including data collection, processing and annotation are explained. Challenges during annotation are also discussed. The results of experiments conducted using pretrained language models prove the dataset's usefulness in training Swahili-English code-switched misinformation classification models.

1 Introduction

The increasing popularity of social media has shifted the preference of news dissemination channels from mainstream newsrooms on cable television to digital platforms such as the social media platform Twitter (now X) which offer real-time, more interactive and uncensored avenues to share news and comment on current topics (Ombui, Muchemi & Wagacha, 2019). Twitter users often take advantage of the anonymity that comes with an online persona to spread

[1] dataset: https://github.com/jayneamol/kweli

misinformation. By posting false texts, links or re-posting other users' misleading news, Twitter users can easily amplify one false post causing it to trend and sometimes be picked up by mainstream news channels. Twitter defines misleading content ('misinformation') as claims that have been confirmed to be false by external, subject-matter experts or include information that is shared in a deceptive or confusing manner (Twitter, 2023).

In Africa, Kenya is among the leading countries in terms of blogging, and Twitter is an increasing platform for social commentary and political engagement (Mukhongo, 2020). Political events like the recently concluded 2022 General Elections in Kenya were widely tweeted about. Despite Twitter's set policies against misinformation amplification, multiple posts relating to misleading news like fake polls, unverified electoral results, and unsubstantiated statements from political parties and individuals were flagged (Mozilla, 2022). Misinformation on political issues often polarises the country and may result to violence in extreme cases. Twitter serves as a mobilization platform for Kenyan internet users to go to the streets and stage protests after con-tested elections (Mukhongo, 2020).

The Swahili language is an African language of the Bantu family with a constant changing vocabulary (Kresse & Vierke, 2022) and is spoken by over 100 million people in East and Central Africa. Despite this high number of speakers, Swahili is still classified as a low-resource language (Wanjawa et. al., 2022) because of the inadequate data for Natural Language Processing (Shikali & Mokhosi, 2020). Aside from the use of standard Swahili or English and other native languages, an ever-evolving slang (sheng') is often developed among Twitter users. So far, there is hardly any work involving East African languages in detecting misinformation from social media (Nabende et. al., 2021).

In a multicultural country like Kenya with 42(+) tribes (Balaton-Chrimes, 2021) each speaking an almost distinct language and two national languages, messages on Twitter are often code-switched. Code-switching, which denotes a shift from one language to another within a single utterance, is common in casual text found in social media (Sitaram et. al, 2019) among bilingual or multilingual communities. The shift in languages vary and can occur in the form of a whole phrase or word in a different language in one sentence. Table 1 shows some of the variations in code-switched Swahili-English texts.

Table 1: Variations in code-switched texts

Language	Text
|eng| swa| eng| swa| eng| swa	|propaganda| imeganda| proper| kama hauna| facts don't engage| watu haujui
|swa| eng	|kwa hivo kuibiwa pia ni makosa| why victimise him
|swa| eng	|mheshimiwa| voting is through secret ballot

The semantic complexities of code-mixing in social media texts (Ombui, Muchemi & Wagacha, 2019) makes models that are trained with single language datasets less accurate when classifying a code-mixed dataset. This necessitates building of code-switched datasets to train models that can perform classification tasks for code-switched data.

This study introduces a novel dataset: PolitiKweli, a Swahili-English misinformation classification dataset built from tweets relating to the highly contested 2022 General Elections in Kenya. The dataset containing tweets posted by Kenya's electoral commission, Independent Electoral and Boundaries Commission, IEBC was also created for fact-checking the text.

This dataset can be used to train and test text classification models which can be applied to Twitter and other social media platforms to flag cases of misinformation on political issues in Swahili-English code-switched texts. Misinformation classification helps preserve the integrity of news posted on social media and curb any adverse effects that may result from misinformation.

2 Related Work

Previous studies have built datasets that are used in training and testing classification models in machine learning. Some of the common classification tasks are misinformation classification, fake news detection and hate speech detection. Nabende et. al. (2021) curated a Luganda-English code-mixed COVID-19 misinformation classification dataset by comparing tweets and Facebook posts about COVID-19 against credible information posted by the Ugandan Ministry of Health. The posts were annotated as 'misinformation' or 'no-misinformation'. 'FACTOID' (Sakketou et. al., 2022) is an English language dataset designed for misinformation detection on Reddit built by classifying Reddit users as 'misinformation spreaders' versus 'real news spreaders' and assigning credibility scores to each user based on the factuality of the news sources.

The datasets 'FakeNewsNet' by Shu et. al. (2020) and 'ISOT' used by Amer, Kwak, & El-Sappagh (2022) are built for fake news detection tasks. It is in English and contains instances of fake news campaigns by Twitter users and who they follow which can be used to create a social network of fake news spreaders (Michail, Kanakaris & Varlamis, 2022). 'ISOT' used by the study Amer, Kwak, & El-Sappagh (2022) on the other hand, contains English language news stories sourced from Reuters, Wikipedia, Politifact. The real articles were gathered from the Reuters website, while the fake or false ones were from different sites identified as fictitious by Wikipedia and Politifact. In the study by Ombui, Muchemi and Wagacha (2019), a Swahili-English code-mixed dataset for hate speech detection was created from tweets from the 2017 and 2012 general election in Kenya.

Some of the publicly available datasets that focus on low-resource languages in-cluding Swahili are 'Afrisenti' (Muhammad, et. al., 2023), 'Kencorpus' (Wanjawa et. al., 2022), 'MasakhaNER' (Adelani et. al., 2021) and 'NER' (David, 2020). The pro-cess of creation of these datasets include data collection, data pre-processing and annotation. During the creation of PolitiKweli, we took inspiration from the studies by (Muhammad, et. al., 2023), Wanjawa et. al. (2022), Nabende et. al. (2021) and Ombui, Muchemi & Wagacha (2019) in dataset creation for code-switched texts and exploration of the low-resource language, Swahili.

3 Methodology

An exploratory research design was applied in data collection from Twitter, data pre-processing and processing to generate viable data to train the model and data annotation as either 'fake', 'false' or 'neutral'.

3.1 Data Collection

Data on Twitter was collected using Twitter Academic API which offers access to both historical and real-time data as in Muhammad et al. (2023). The study collected 50,000 tweets posted from 4th October, 2021 which was the first day of voter registration in preparation for 2022 elections in Kenya to the 5th September, 2022 when the Supreme Court of Kenya issued the ruling on the contested presidential election results.

The collection process involved selection of tweets with hashtags relating to elections that trended during the election period such as #KURA 2022, #Uamuzi2022 #KenyaElections2022, #Kenya and

#GE2022. In addition to hashtags, the most used key words such as elections, vote, voters, general elections, tallying, election results and key mentions such as @RailaOdinga, @WilliamsRuto and @IEBCKenya – the electoral body were used in the search. There were two sets of tweets: the general tweets posted by Twitter users about the 2022 General Elections and tweets posted by the electoral body (@IEBCKenya) which were regarded as fact-checked news.

3.2 Data Processing

The data processing stage included language identification, data cleaning, anonymization and lowercasing. Manual language detection was done due to semantic complexities of code-switched texts and the several other languages that may be present in one tweet. A team of ten annotators labeled the data in four categories: 'Swahili-English', 'Swahili', 'English' or 'Other'. This resulted in four sets data grouped data according to language. The data anonymization process involved changing of any mentions to @user to protect identities of Twitter users mentioned in tweets.

3.3 Data Annotation

A team of ten annotators labeled the clean data as fake, false or neutral. The grammatical complexity of code-switched texts necessitates that the human annotators be conversant with both languages used. Both inter-annotator agreement and measure of code-switching were computed. An average Free-Marginal Multirater Kappa score of 0.47 was achieved. The code-switched texts had an average M-index of 0.67 which indicated good code-switching.

Some of the challenges faced during annotation were the high cost of annotation given that the study did not have any external funding and decoding of satirical text.

4 Results

A summary of the respective labels per language is shown in Table 2. The texts are classified as Swahili-English (swa-eng), English (eng) and Swahili (swa).

Table 2: Summary of the respective labels per language

Data	swa-eng	eng	swa
Factual Data	1221	6065	8
Neutral Data	4094	13380	19
Fake Data	1030	3512	184

The Swahili-English texts which were the main focus of this study, were tested using BERT pre-training language model. An accuracy of 53.9% was achieved.

The English and Swahili texts are also this study's contribution and can be used for other machine learning tasks such as text classification and sentimental analysis in Swahili or English only tasks in the future.

5 Conclusion and Future Work

The contribution of this study is a dataset with three sets of texts: 6,345 Swahili-English texts, 22,954 English texts and 211 Swahili texts. The dataset can be used in training models that perform classification tasks for both Swahili-English code-mixed texts and Swahili only texts. With constant advancements in technology and increased use of code-switching in the new generations of language speakers, there is need for more research in low-resource languages such as Swahili. This study paves way for more research in code-mixing and how machine learning can be leveraged to provide solutions for constantly evolving problems such as misinformation detection on social media. We also hope to achieve better accuracy with pretrained multilingual models such a mBERT.

6 Acknowledgements

We acknowledge the contributions of Shamsuddeen Hassan, Martin Okech, Edwin Onkoba, Mary Gitaari, Elphas Otieno, Bowa Marita, Peter Gathuita and the Maseno University School of Computing and Informatics staff and students. We thank Eze-kiel Maina, Nelson Odhiambo, Stephen Otieno, Monicah Odipo, Harrison Kioko and Samwel Okonda for their contribution in data annotation.

References

1. Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., ... & Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116-1131
2. Amer, E., Kwak, K. S., & El-Sappagh, S. (2022). Context-Based Fake News Detection Model Relying on Deep Learning Models. Electronics, 11(8), 1255.
3. Balaton-Chrimes, S. (2021). Who are Kenya's 42 (+) tribes? The census and the political utility of magical uncertainty. Journal of Eastern African Studies, 15(1), 43-62.
4. Barnett, R., Codó, E., Eppler, E., Forcadell, M., Gardner-Chloros, P., Van Hout, R., ... & Wensing, S. (2000). The LIDES Coding Manual: A Document for Preparing and Analyz-ing Language Interaction Data Version 1.1--July 1999. International Journal of Bilingual-ism, 4(2), 131-271.
5. Davis David. 2020. Swahili: News classification dataset.
6. Githiora, 2018 — Githiora, C. J. (2018). Sheng: rise of a Kenyan Swahili vernacular. Boydell & Brewer.
7. Kresse, K., & Vierke, C. (2022). Swahili language and literature as resources for Indian Ocean studies. History Compass, e12725.
8. Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 363-374.
9. Martin, G., Mswahili, M. E., Jeong, Y. S., & Young-Seob, J. (2022, July). SwahBERT: Language Model of Swahili. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 303-313).
10. Michail, D., Kanakaris, N., & Varlamis, I. (2022). Detection of fake news campaigns using graph convolutional networks. International Journal of Information Management Data In-sights, 2(2), 100104.
11. Mozilla. https://foundation.mozilla.org/en/blog/new-research-in-kenya-disinformation-campaigns-seek-to-discredit-pandora-papers/ last accessed 2022/12/15
12. Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Ousidhoum, N., Adelani, D. I., Yimam, S. M., ... & Arthur, S. (2023). Afrisenti: A twitter sentiment analysis benchmark for african languages. arXiv preprint arXiv:2302.08956.
13. Mukhongo, L. L. (2020). Participatory Media Cultures: Virality, Humour, and Online Polit-ical Contestations in Kenya. Africa Spectrum, 55(2), 148-169.
14. Nabende, P., Kabiito, D., Babirye, C., Tusiime, H., & Nakatumba-Nabende, J. (2021). Mis-information detection in Luganda-English code-mixed social media text. arXiv preprint arXiv:2104.00124.
15. Okoth, G. B. W. (2020). How Kenyans on Twitter use visuals as a form of political protest. Journal Kommunikation. Medien, 1-27.
16. Ombui, E., Muchemi, L., & Wagacha, P. (2019, October). Hate speech detection in code-switched text messages. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-6).

17. Sakketou, F., Plepi, J., Cervero, R., Geiss, H. J., Rosso, P., & Flek, L. (2022). FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias. arXiv preprint arXiv:2205.06181.
18. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data, 8(3), 171-188., 8(3), 171–188.
19. Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. Data in brief, 31, 105951.
20. Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W. (2019). A survey of code-switched speech and language processing. arXiv preprint arXiv:1904.00784.
21. Twitter, https://help.twitter.com/en/resources/addressing-misleading-info, last accessed 2023/06/10.
22. Wanjawa, B., Wanzare, L., Indede, F., McOnyango, O., Ombui, E., & Muchemi, L. (2022). Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Lan-guage Processing Tasks. arXiv preprint arXiv:2208.12081.