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

Cynthia Amol[1] Lilian Wanzare [2] James Obuhuma [3]

[1,2,3] Maseno University, Kenya

[1]cynthia@maseno.ac.ke

Abstract

In the age of freedom of speech, users of the social media platform Twitter post millions 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 dataset introduced by this study, which a novel Swahili-English misinformation classification dataset, contains 2,472 Swahili-English texts, 8,104 English texts and 33 Swahili texts. The texts are labelled as fact, fake or neutral as compared to a fact-checked dataset also created for this study. The results of experiments conducted using pre-trained language models prove the dataset's usefulness in training Swahili-English code-switched misinformation classification models.

Misinformation on Social Media

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 which offer real-time, more interactive and uncensored avenues to share news and comment on current topics [1].Twitter users often take advantage of the anonymity that comes with an online persona to spread misinformation by posting false texts, links or reposting other users' misleading news. 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 [2]. Misinformation on political issues often polarizes 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 contested elections [3].

In a multicultural country like Kenya with 42(+) tribes [4] 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 [5] 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 as shown in Table 1.

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

Table 1: Variations in code-switched text.

The semantic complexities of code-mixing in social media texts [1] 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.

Code-switching: Use of Swahili-English on Social Media

The Swahili language (Kiswahili) is an African language of the Bantu family with a constant changing vocabulary [6] 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 [7] because of the inadequate data for Natural Language Processing [8]. Low-resource languages are usually neglected due to few resources and research efforts [7].

The youth, who are the numerically dominant population and the most users of Twitter mostly use Swahili-English in informal communication. In the Kenyan post-independence era, Swahili is a national language next to English [6]. So far, there is hardly any work involving East African languages in detecting misinformation from social media [9]. If we want machines to partake in human-human conversations, they need to also be able to understand what is being said in varied registers [5]. This creates the need to build models trained using data in languages spoken by humans.

Misinformation Classification

Misinformation classification involves identifying text as misinformation or not by comparing the text against fact checked text. Twitter uses a combination of human review and technology to detect and label or remove misleading content. In the case of the 2022 General Elections in Kenya, the application of these labels was not done extensively [2].

Development of automated tools for misinformation detection involves either building of a novel dataset to train and test a model or use of pre-existing datasets to train and test a model. A code-switched Swahili-English misinformation classification dataset does not exist. Semantic complexities of code-mixing in social media texts [1] makes models that are trained with single language datasets less accurate when classifying a code-mixed dataset. This necessitates research into curation of datasets for low-resource languages such as Swahili. These datasets can be used to train models with better accuracy in classifying misinformation in code-switched texts.

Code-switched Datasets

Creation of code-switched datasets is done by collection of data, processing the data to remove any single-language only texts then labeling the text. Some of the publicly available code-switched datasets are Indian Political Memes (IPM) by the study [10] in Hinglish (Hindi-English) for hate speech classification and the Luganda-English code-mixed dataset for COVID-19 misinformation classification [9].

The PolitiKweli Dataset

The PolitiKweli dataset is a Swahili-English misinformation classification dataset based on Twitter posts relating to events during the electioneering period for the 2022 General Elections. It has a total of 10,609 labelled texts in three languages: Swahili-English (swa-eng), English (eng) and Swahili (swa) distributed as shown in Table 2. Data curation process of the PolitiKweli dataset involved data collection, data processing and data annotation.

Language	Number of texts
swa-eng	2472
eng	8104
swa	33

Table 2: Number of texts per language

Data Collection

Data on Twitter was collected using Twitter Academic API which offers access to both historical and real-time data as in [11]. The study collected 20,000 tweets posted from 4th October, 2021 which was the first day of voter registration in preparation for 2022 General Elections in Kenya to 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, #KenyaDecides2022 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.

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 labelled the data in four categories: 'Swahili-English', 'Swahili', 'English' or 'Other'. This resulted in four sets data grouped data according to language. All unusable text was removed from the first three categories of text during data cleaning. After cleaning, all usernames in the tweets were converted to @user to protect identities of users mentioned.

Data Annotation

A team of ten annotators labelled the processed data. Annotation guidelines were set and a pilot annotation session conducted to measure inter-annotator agreement. The texts were labelled as: fake (news that could be proven as untrue compared to the fact-checked dataset), fact (news that could be proven as true compared to the fact-checked dataset) or neutral (news that is neither fact nor false, merely a Twitter users' opinion that could not be fact-checked).

Table 3 shows sample tweets and their labels.

Text	Label
spoilt ballots haziingi kwa ballot box, hio ni rejected	Fact
if a voter is given six ballot papers and he only vote one, he rest of the ballot papers are return to the clerks because he hasnt marked the papers	Fake
if everything aligns sisi tutakua kwa ballot paper pia 2027	Neutral

Table 3: Sample labeling

The annotation process resulted in three categories of data as shown in Table 4.

Data	swa-eng	eng	swa
Fact	479	2277	2
Fake	414	1130	10
Neutral	1571	4679	21

Table 4: Annotated data per language

The dataset was tested using the pre-trained language model, BERT.

References

[1] 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).*

[2] Mozilla. https://foundation.mozilla.org/en/blog/new-research-in-kenya-disinformation-campaigns-seek-to-discredit-pandora-papers/ last accessed 2022/12/15

[3] Mukhongo, L. L. (2020). Participatory Media Cultures: Virality, Humour, and Online Political Contestations in Kenya. *Africa Spectrum, 55(2), 148-169.*

[4] 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.*

[5] 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.*

[6] Kresse, K., & Vierke, C. (2022). Swahili language and literature as resources for Indian Ocean studies. *History Compass, e12725.*

[7] 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.*

[8] Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. *Data in brief, 31, 105951.*

[9] 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.*

[10] Rajput, K., Kapoor, R., Rai, K., & Kaur, P. (2022). Hate Me Not: Detecting Hate Inducing Memes in Code Switched Languages. *arXiv preprint arXiv:2204.11356.*

[11] 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.*