A Linguistic Annotation Framework to Study Interactions in Multilingual Healthcare Conversational Forums

♡Ishani Mondal, ♡Kalika Bali, ♡Mohit Jain, ♡Monojit Choudhury, ♠Ashish Sharma,*
◇Evans Gitau, ◇Jacki O' Neill, ◇Kagonya Awori, ◇Sarah Gitau
♡Microsoft Research Labs, Bangalore, India
♠Paul G. Allen School of Computer Science and Engineering, University of Washington
◇Microsoft Africa Research Institute, Nairobi, Kenya
{t-imonda, kalikab, monojitc, jacki.oneill}@microsoft.com

Abstract

In recent years, remote digital healthcare using online chats has gained momentum, especially in the Global South. Though prior work has studied interaction patterns in online (health) forums, such as TalkLife, Reddit and Facebook, there has been limited work in understanding interactions in small, close-knit community of instant messengers. In this paper, we propose a linguistic annotation framework to facilitate the analysis of health-focused WhatsApp groups. The primary aim of the framework is to understand interpersonal relationships among peer supporters in order to help develop NLP solutions for remote patient care and reduce the burden of overworked healthcare providers. Our framework consists of fine-grained peer support categorization and message-level sentiment tagging. Additionally, due to the prevalence of multilinguality in such groups, we incorporate word-level language annotations. We use the proposed framework to study two WhatsApp groups in Kenya for youth living with HIV, facilitated by a healthcare provider.

1 Introduction

Good communication between patients and healthcare providers as a part of patient-centered care, where the patient is an equal partner in their healthcare decisions, has been shown to improve medical adherence (Schoenthaler et al., 2009; Ciechanowski et al., 2001) (i.e., the practice of consistently taking the medicines as prescribed) . However in the Global South, i.e. the countries, located primarily in the southern hemisphere, that have historically been identified as third world due to perceptions of their socio-economic status, technological advancements, and global dominance, high patient volumes and a need for cost-effective solutions can make patient-centered care difficult. In recent years, various online chat applications such as WhatsApp and WeChat are being increasingly used to ensure smooth access to both patient-provider communication and peer support beyond formal healthcare system (Karusala et al., 2021; Bhat et al., 2021; Karusala et al., 2020). Even though these peer support groups are effective in providing patient-centered care, they can be taxing for already overworked healthcare providers (Viswanathan et al., 2020). There is therefore a necessity for designing technology to support both the providers and patients using such groups.

Prior work on online health forums focuses on analyzing participation (Schultz et al., 2019; Sadeque et al., 2015; van Campen and Iedema, 2007), social connection and engagement (Sharma et al., 2020a; Kushner and Sharma, 2020; Smith-Merry et al., 2019; Halder et al., 2017), and modelling user behavior (Hosseini and Caragea, 2021; Sharma et al., 2020b; Buechel et al., 2018; Elliott et al., 2018; Pérez-Rosas et al., 2017; Choudhury et al., 2016; Gibson et al., 2016; Choudhury and De, 2014). Further, existing research focuses on the linguistic analysis of conversations in social media platforms like Facebook (Dehouche, 2020; Tian et al., 2017; Vyas et al., 2014; Das and Gambäck, 2013), Reddit and Twitter (Zomick et al., 2019; Rudra et al., 2019; Kiesling et al., 2018; Rijhwani et al., 2017; Tran and Ostendorf, 2016; Rudra et al., 2016; Celli and Rossi, 2012; Bak et al., 2012; Gouws et al., 2011), but has paid little attention to analyzing the conversations in small, close-knit WhatsApp communities. In this paper, we propose a hierarchical annotation framework to facilitate the analysis of health-focused WhatsApp groups.

Our proposed hierarchical framework consists of fine-grained peer support categorization and message-level sentiment tagging. Additionally, we address the prevalence of code-mixing in such

* This work was done when the author was a Research Fellow at Microsoft Research Lab India.

WhatsApp groups by incorporating word-level language annotation. We evaluate our framework on chat logs of two WhatsApp groups of Kenyan youth living with HIV, moderated by a healthcare provider. We also develop a user-friendly annotation interface to facilitate annotation. In the rest of the paper, we describe the annotation process using the framework, the challenges faced and our learnings. Our annotation study shows that the annotators achieve high agreement for all categories, thereby indicating the efficacy of our linguistic annotation framework. We believe that this framework can be used to analyze inter-personal relationships among group members and is the first step towards developing NLP solutions to support healthcare workers providing remote patient care.

While developing the framework, we emphasized on code-mixing. Multilingualism in Kenya contributes to a considerable linguistic diversity in a population that speaks Kiswahili, Kikiyu and English fluently (Dwivedi, 2015; Chumbow, 2010). This leads to a high prevalence of code-mixing in their linguistic interactions. This is further enriched by the use of Sheng, a Swahili and English-based cant (a mixed language or creole), by the urban youth in Kenya. It emerged as a result of urbanization and globalization (Githiora, 2002; ABD). Any annotation framework for such conversations, therefore, has to take into account this multilinguality.

Several annotation schemas have been proposed for multilingual code-mixed messages on social media platforms (Chakravarthi et al., 2021, 2020; Sirajzade et al., 2020; Vijay et al., 2018; Swami et al., 2018; Dhar et al., 2018; Jamatia et al., 2016; Begum et al., 2016; Maharjan et al., 2015; Barman et al., 2014; Bergsma et al., 2012). However, to the best of our knowledge, none of these look into instant-messaging based interactions among peer supporters using code-mixed African languages like Kiswahili, Sheng.

We developed our annotation framework to support an ethnomethodologically-informed ethnographic analysis of the chat: a qualitative approach used in Human-Computer Interaction (HCI) research to understand social interactions (Button and Sharrock, 1997; Hughes et al., 1994). Such analysis has proven useful in the design of computer systems as it reveals the practices and methods of those who will use these systems and which need to be designed for (Crabtree, 2003); being used to examine text interactions such as instant messages (O'Neill and Martin, 2003), internet forums (Martin et al., 2014) and chat messages (Wang et al., 2020; Jain et al., 2018a). The annotation framework was designed to be used to support and inform the ethnographic analysis, with the aim of identifying ways in which we might support both the healthcare provider and participants in these chat groups.

Additionally, the framework would be useful for sociolinguistic studies of multilingual conversations and the schema could be extended to understand other peer support settings such as mental health forums. Another area where the framework might prove useful is for research around improving human-machine conversational agents (Ashktorab et al., 2019; Jain et al., 2018b) by leveraging data annotated using these categories.

2 Dataset

We studied WhatsApp chat logs from two peer-support groups for Kenyan youth, living with HIV. There were a total of 1,655 messages in Group-1 (28 members, 14 female, 14 male, age=14-17 years) and 4,901 messages in Group-2 (27 members, 21 female, 6 male, age=18-24 years). Both the groups were facilitated by a healthcare provider with a background in public health, HIV testing, and counselling. The facilitator sent weekly messages on a range of topics, such as future goals, strategies for medical adherence, etc., responded to queries (within 12 hours), clarified any misinformation posted on the group, and referred medical-questions to a clinic. The chat records were in Kiswahili, English and Sheng, sometimes a mix of these languages in a single message. All the messages were anonymized and translated into English by a native speaker. Each chat message in the dataset had the following information: anonymized speaker ID, timestamp, original message and English-translated message. The data contains sensitive content dealing with people's deeply personal lives and tragedies, and therefore, even with anonymization there are ethical reasons for not making the dataset public. However, anyone wanting access to the data for research purposes can get in touch with the authors.

3 Annotation Framework

The primary objective of our proposed linguistic annotation framework is to facilitate studies

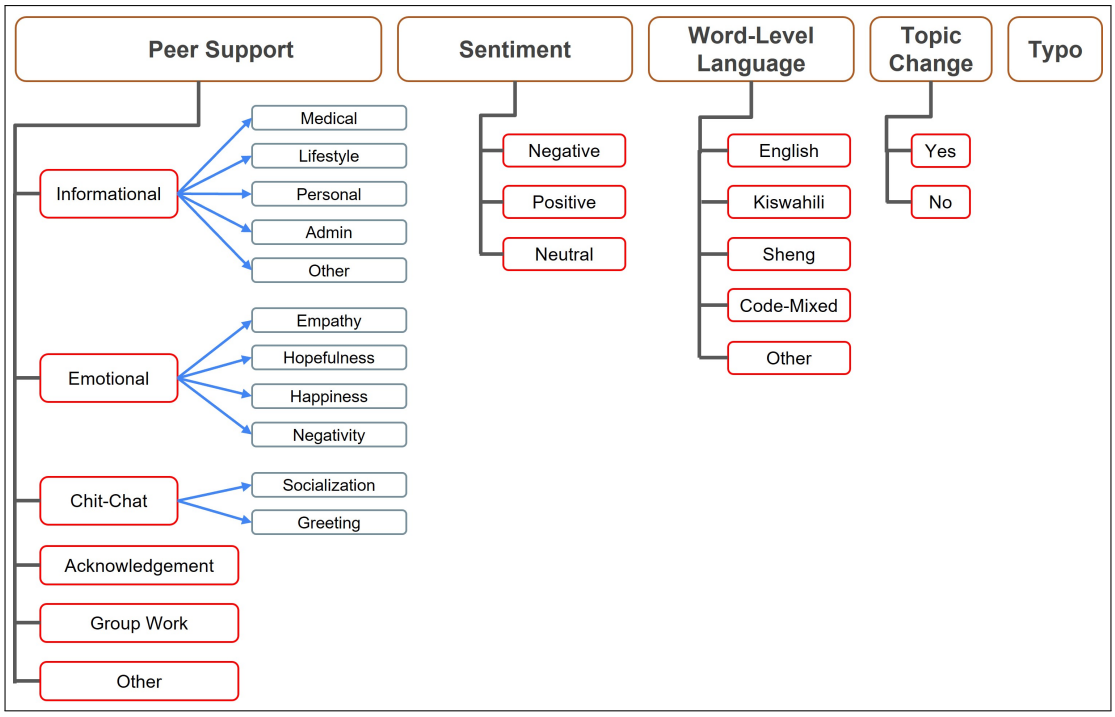

Figure 1: An overview of the linguistic annotation framework.

on inter-personal relationships among the peer-supporters. In order to satisfy the above objective, the framework is governed by the following guiding principles. 1) We want to model the conversational intent of the peer supporters behind the act of sending messages. 2) We need to assess the psychology of their behavior by understanding their sentiments and emotions. This is even more critical in a healthcare setup where it is important to maintain an overall positive outlook. 3) We also want to capture the morphosyntactic level of linguistic information which can be used to help us tease apart deeper interaction patterns and stylistic features of conversations. The framework (Figure 1) has a hierarchical schema with five top-level categories that deal with linguistic and pragmatic phenomena observed in the chat. The categories include *Peer Support* (principle 1, 3.1), *Sentiment* (principle 2, 3.2), *Word-Level Language*, *Topic Change*, and *Typo* (principle 3, 3.3, 3.4, 3.5). Note that this framework can be further augmented to include more linguistic features and can be extended to perform similar studies on other peer support groups. In this section, we define the broader and fine-grained categories.

3.1 Peer Support

Peer support refers to members using their own experiences to help others feel accepted and understood. Our framework defines the following sub-categories of peer support as:

3.1.1 Informational

The group members either seek or provide factual information, advice, or warning. These are further classified as:

- **Seek Information:** A group member is seeking information from fellow group members. E.g., "*explain further about circumcision*".
- **Provide Information:** A group member is providing information. E.g., "*Yeah, it might be a sign. Also it increases when you are stressed.*".

The informational sub-category is further divided into the following sub-types:

Medical: Seeking or providing factual information, advice, or warning about medicine, treatment, or disease.
E.g., [Original] "*Most people infected with the bacteria that cause tuberculosis don't have symptoms. When symptoms do occur, they usually include a cough (sometimes blood-tinged), weight loss, night sweats and fever*".

Lifestyle: Healthy lifestyle related information such as exercise and food.
E.g., [Original] "*You can take a lot of soup, njahe, omena. These works well for new moms. Don't*

drink coffe or majani. Tumia to cocoa".
[Translated] "You can take a lot of soup, turtle beans, fish. These work well for new moms. Don't drink coffee or tea. Use only cocoa".

Personal: Information regarding relationships with partner or other family members.
E.g., [Original] "*The problem is I have a partner na ajue my status am afraid day nitamwambia he can even kill himself*"
[Translated] "*The problem is that I have a partner who doesn't know my status and am afraid the day I decide to say he might kill himself.*"

Admin: Administrative information such as rules and regulations of the group and formal meetings.
E.g., [Original] "*We shall respect all members and not use words like HIV, ART, CD4 So that people feel comfortable*".

Other: This contains any information not included in the above categories.
E.g., [Original] "*Hi....This is a big warning...my mother in law has just lost over Kes. 70,000 to M-Shwari and KCB-MPESA con-women...*".

3.1.2 Emotional

This sub-category deals with seeking emotional support, or communicating empathy, love or concern towards fellow group members.
E.g., [Original] "*Haki usichoke mm pia nilipoteza hapitite BT ilirudi tu lakini jaribu uone daktari, saa zing one inakuwanga type ya drugdrug kukeep time yakutake drug pole dia*".
[Translated] "Please don't be am also loosing appetite but it come back just go see a doctor, sometimes it's usually the type of drug and your timing for your medication, sorry my dear".

The emotional sub-category of messages are further sub-categorized into the following:
Empathy: An explicit mention of the difficulty experienced by the seeker that validates his/her distress or an attempt to see things from the seekers' perspective.
E.g., [Original] "*pia Mimi na shida ingine Niko na mimba yake*".
[Translated] "Same here also and another problem is that he got me pregnant".

Hopefulness: A response encouraging others to stay hopeful and optimistic in difficult situations.
E.g., [Original] "*Sorry for this dear. Like 5002 has said some things need sacrifice and time. Never give up. Just try and overcome the hurdle, all will be fine.*"

Happiness: A response that indicates a happy mood or excitement.
E.g., [Original] "*Early this month, we were blessed with a baby girl and we thank God for His mercies on us*".

Negativity: A response that indicates a mournful mood or shows any negative emotion (such as anger/sorrow/frustration). E.g., [Original] "*Hello guys nfeel so bad today c ski poa nko n headache n pia cna strength pray 4 me*".
[Translated] "Hello guys am feeling so bad today i have a headache and I also dont have energy please pray for me".

3.1.3 Chit-Chat

The messages in this category include introductions, greetings and other responses that promote social interactions or demonstrate friendliness.
E.g., [Original] "*Lunch ilipita sasa ni supper*"
[Translated] "Lunch is over it's now supper time".

It can be further sub-categorized as: **Socialization** (E.g., [Original] "*Karibuni lunch guys*", [Translated] "Welcome for lunch guys") and **Greeting** (E.g. [Original] "*Happy Easter guys*").

3.1.4 Acknowledgement

Responses used to affirm, acknowledge and/or backchannel communication, are tagged under this category. .
E.g., [Original] "*Santy...pia ww doz poa*".
[Translated] "Thanks..you too sleep well".

3.1.5 Group Work

A response wherein members encourage each other to work as a group.
E.g., [Original] "*Let's bring topics even on the contemporary experiences so that we can continue keeping the group more active as it should be guys!*"

3.1.6 Other

Any response that does not belong to any of the above categories (like emoji, a system-generated message, etc).
E.g., [Original] "*Waiting for this message*". (a default system-generated message in WhatsApp)

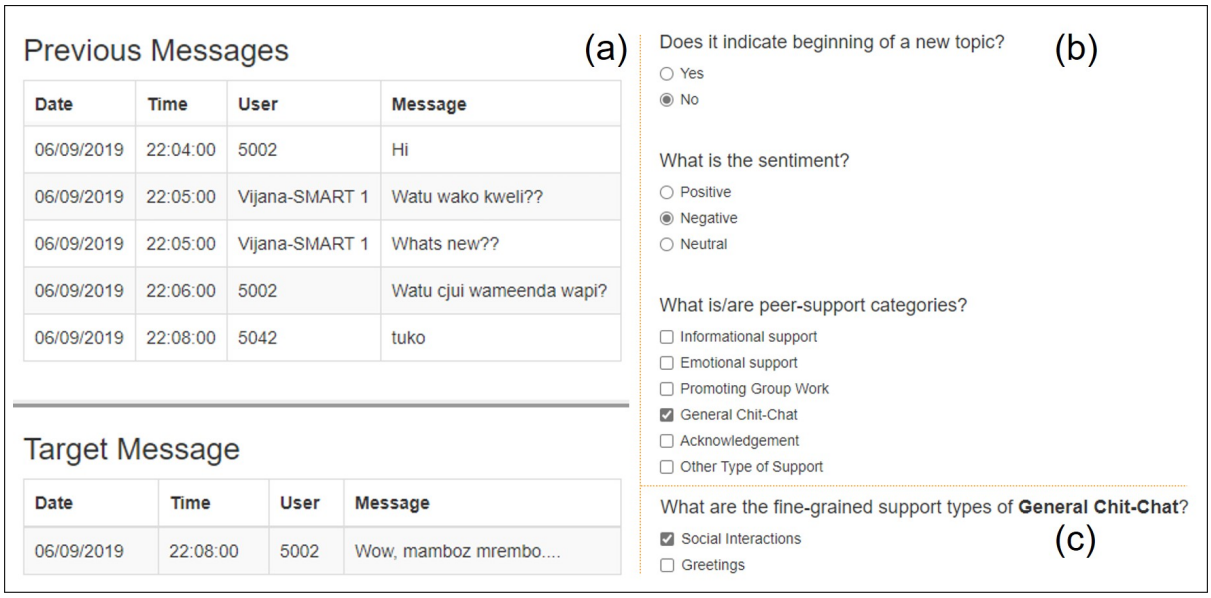

Figure 2: An overview of the annotation interface to annotate WhatsApp chat messages.

3.2 Sentiment

Sentiment is a view or opinion that is held or expressed by an individual. The sentiments are further categorized into the following:

Negative: It expresses a negative view or opinion.
E.g., [Original] "*Uko na shida ya kichwa c siri*".
[Translated] "You have a mental problem it's no secret".

Positive: A message is classified as positive if the speaker expresses a strongly positive opinion on something or someone. Note: Greetings are not included in this category.
E.g., [Original] "*I'm much happy to interact and share with you guys!*"

Neutral: When neither positive nor negative sentiment is expressed, the message is labelled as neutral. This could include a general comment, acknowledgement, chitchat or a simple greeting.
E.g., [Original] "*I suggest that we should try and at least do this for our health.*"

3.3 Word-Level Language

To understand the code-mixed interactions in the chat, word-level language annotation is needed. For our dataset, we used five language identification codes: *En* for English, *Sw* for Kiswahili, *Sh* for Sheng, *CM* for Code-Mixed words or phrases (E.g., '*nfeel*', '*tutawash*'), and *Oth* for miscellaneous (e.g., emojis, any other language).
E.g., [Original] "*Mimi Nilichukua my siz tukalost*".
[Language] (Sw) (Sw) (En) (En) (CM)
[Translated] "I took my sister and I got lost".
If a particular word is tagged as *Oth*, the annotators were asked to explain the reason behind choosing this particular tag. For instance, *Kyole 2* (name of a hospital) is assigned the language tag *Oth*, and the reason specified is *Named Entity*. Also, if the language is different from *En*, *Sh* and *Sw*, that needs to be specified in the reason correctly. Thus, our schema can be further extensible to annotate more languages.

3.4 Topic Change

Each message is labeled with a binary flag (Yes/No) to indicate a change in topic.

3.5 Typo

Shortened words, abbreviations and misspellings are pervasive in informal conversations, especially in instant messaging such as WhatsApp. It is important to identify and normalize such typos. The annotators were asked to correctly normalize the typos found during the annotation process. It is worthy to note that the typos and abbreviations need to be separately tagged and identified. E.g.,
1. [Original] "***Fuine*** *and how are you..*"
2. [Original] "*We shall respect all members and not use words like* ***HIV***, ***ART***, ***CD4***, ***VL***. *So that people feel comfortable.*"
The spelling mistake and abbreviations in the

above examples are written in bold.

4 Annotation Interface Overview

We developed an annotation interface (Figure 2) to facilitate user-friendly annotation. The annotators were provided with detailed guidelines on how to use the interface for annotating the chat messages based on our proposed annotation framework. Django framework (Python) was used on the server-side for implementing this interface. This interface can be also used to annotate the peer support conversations in the mental health chat forums such as Reddit, TalkLife. The annotators need to login into the interface with their credentials to start the annotation process as outlined in the guidelines provided to them. On logging in, a target message is displayed along with ten (or less) previous messages to provide context (Figure 2a). The right side of the screen shows question related to top-level categories: peer support, sentiment, and change in topic (Figure 2b). The annotators then have to choose appropriate labels. For the peer support category, the annotators have to further choose subcategory labels as well (Figure 2c).

As a message can be associated with multiple subcategories, the annotators are asked to highlight specific portions of the message for each subcategory. E.g., in the message "*Hi...he got me pregnant...*"—"*Hi*" is tagged as *Chit-Chat → Greeting* and "*he got me pregnant*" is tagged as *Informational → Personal*.

For *Typo*, the annotators are asked to highlight spelling mistakes and abbreviations, and correct them wherever possible. For instance, in the message "*Morning members from today, hence forth untill further notice am the S.T.O (sitting allowance officer) ask any querry you get the right answer*" the annotators needs to highlight the typos ("*untill*", "*querry*") and normalize those ("*until*", "*query*") using the interface. Moreover, they should also annotate the abbreviation ("*S.T.O*").

The annotation requires a language tag for each word in the message. Instead of specifying a language label for each word, the annotators are asked to highlight the word spans for each language category, facilitating quicker and less error-prone language identification. For example, in the message "*Wow, mamboz mrembo*" ([Translated]: "*Wow, Hello beautiful*")—"*Wow*" is highlighted as *English* and the word span "*mamboz mrembo*" is highlighted as *Kiswahili*.

The annotators are expected to completely annotate each message before moving on to the next message. The submitted annotations are saved in a cloud database. The annotators have an option to revise and modify their current or previously submitted annotations during the process. If the annotators are unsure of how to annotate a particular message, the interface allows them an option to note and skip that message. This annotation interface can be leveraged easily to annotate any complex hierarchical annotation schema comprising of multiple levels.

5 Annotation Experiments and Observations

The annotation was conducted in three phases. In Phase-1, a small sample of 395 English-translated messages (105 messages from Group-1, 290 messages from Group-2) were annotated, and the corresponding original multilingual messages of the same set were annotated in Phase-2. In Phase-3, the entire multilingual dataset was annotated. Below we discuss the challenges faced and the learnings from all the annotation phases.

5.1 Annotation Phase-1

The Phase-1 study was designed to ensure that the entire annotation process as well the guidelines were well-understood by the annotators. In this phase, the short-listed 395 chat messages were annotated by one of the co-authors, an expert English speaker, who led the development of the proposed annotation framework. The co-author, not being able to understand Swahili or Sheng, found it easier to initially develop the schema using the English-translated messages (this translation was manually carried out by a native speaker). We consider that as the gold standard annotation for Phase-1, and use it to assess the quality of the rest of the pilot annotations in Phase 1. Though the English-translations were not of great quality, but that did not affect our interpretations since we have carried out the final annotation on the multilingual messages (Phase-2 and Phase-3).

Three annotators (graduate students proficient in English) were asked to annotate the same 395 messages using the annotation guidelines. For quality estimation, Cohen's Kappa (κ) (Vieira et al., 2010) was used to assess inter-annotator agreement (IAA) between each annotator and the

expert annotator. We observed an average of 0.67 κ agreement for the fine-grained peer support subcategory tags and 0.89 κ for sentiment. While there was strong agreement for sentiment labelling, the peer support subcategories showed only moderate agreement.

We analyzed the data to understand the annotation differences. First, we found the highest annotation disagreement among *Informational* messages—46% of the total disagreements were between *Medical* and *Lifestyle*. Hence, we revised the annotation guidelines such that *Medical* was given priority over *Lifestyle*, i.e., the annotators were asked to select only *Medical*, or both *Medical* and *Lifestyle*, when unsure. For example, in this message, "*Hi? It's GXXX here I think you can try tell a person who has not accepted the status the importance of taking the drugs and how they work in our bodies. Yah remember to eat healthy*", the individual is stressing upon both the importance of medical adherence and consumption of healthy food. Hence, it should be tagged as both *Medical* and *Lifestyle* under *Informational*.

Second, the ambiguity between *Lifestyle* and *Personal* subcategories comprised 12% of the disagreements. For example: "*We should always protect ourselves using a condom during sexual intercourse*". Our guidelines were revised to clarify that the *Personal* tag should be applied only to messages dealing with personal relationships, while the *Lifestyle* tag should only be used to annotate lifestyle habits. Thus, in the above example, only *Lifestyle* tag is applicable.

Third, another source of confusion was between *Group Work* and *Admin* with 24% of total disagreements. To solve this confusion, we provided more examples in the annotation guidelines to clarify the differences between them. These examples were laid out to elucidate that encouraging more activity within the group members belong to *Group Work* whereas *Admin* dealt with the meetings arranged by the moderators or formal group regulations. For example, in "*With this group, we would like to support each other in staying healthy and supported. If you are having problems or questions about your health, please share them with the group, or inbox me individually. We are stronger together.*", the moderator is talking about the rules of communication within the group, thus it should be tagged as *Admin*. Whereas in "*How are we all doing this week? What are you most looking forward to about this group?*", the speaker is encouraging more participation in the group and should be annotated as *Group Work*.

Finally, other minor disagreements originated due to human error (provide versus seek information, *Greetings* versus *Acknowledgement*). This did not require any revision of the guideline.

5.2 Annotation Phase-2

The annotation guidelines were revised based on the learnings from Phase-1 annotations. We have calculated the pairwise Cohens kappa score using the Scikit-learn library of Python. In Phase-1, we wanted to understand how much each of the pilot annotators agree with the expert annotator using Kappa Agreement. Certain portions of annotation guidelines as well as schema were updated and revised based on their understanding.

Two professional annotators, based in Kenya and proficient in English, Kiswahili, Kikuyu and Sheng languages, were then asked to carry another pilot annotation in Phase-2 based on the revised guidelines. In Phase-2 IAA results (as illustrated in Figure-3), we calculate the pairwise agreement between two annotators of Africa for each of the peer support categories on the multilingual chat messages. No gold standard annotation was used to assess the quality of their annotation in Phase-2 unlike Phase-1. The key difference between Phase-1 and Phase-2 was the addition of language tagging as well as the annotation being undertaken on the original messages rather than the translations.

While analyzing IAA, we observed that tags of English-translations differed from those of the original chat logs, which was due to inaccurate language translation. For instance, "*Poa Sana*" ([Translated] "*Very good*") was tagged as *Acknowledgement* and *Socialization* by the annotators, whereas the corresponding English-translation "*Hey*" was tagged as *Greetings* by the expert annotator. Thus, with the help of the Kenyan annotators, a more accurate English translation was also obtained, and accordingly the gold standard annotation was corrected.

The annotators also found the language tags of certain named entities such as food items (e.g., 'soup', 'cocoa') confusing, mainly because some of the named entities can exist as borrowings in *Kiswahili* and *Sheng*. Hence, we modified the instructions to categorize such entities as *English*.

5.3 Annotation Phase-3

In this phase, the same annotators from Phase-2 annotated the entire dataset based on the revised guidelines. An iterative revision of the guidelines was conducted based on any observed confusion in this phase as well, and the annotators were expected to revise their annotations accordingly. Here we describe a few of them.

First, annotators were unsure about tagging specific entities, such as time (e.g., '1100hrs', '1pm'), symbols (e.g., '+'), laughter (e.g., 'hahaha') and numerals (e.g., '2','five'). This was resolved by asking the annotators to tag time, symbols, and numerals as *Other* language, along with specifying the reason as 'time', 'symbols' or 'numerals', respectively. Laughter expressions were tagged in the language they were written in, as different languages have different expressions for laughter, e.g., 'hahaha' was tagged as *English*.

Second, annotators also faced difficulties in tagging language category for typos (e.g., 'realtionships', 'apa', 'nni'). This was resolved by asking them to choose the language category of the normalized word (e.g., 'realtionships' → 'relationships' in *English*, 'apa' → 'hapa' in *Kiswahili*, 'nni' → 'nini' in *Kiswahili*). It is worthy to note that in order to tag the abbreviations and contracted forms of words, the annotators were also asked to tag it to the language category of the expanded form. We did not come across any ambiguities or unclarities regarding this during the annotation process.

Third, emojis posed another challenge for annotating the peer support categories of the chat messages. In the annotation guidelines, the annotators were instructed to tag non-textual content as *Other*. However, for certain messages, decoding the emoji was needed to obtain the correct tag. For example, "*Is* 🥂 *bad?*". The semantics of the message is dependent on decoding the meaning of the emoji ("*Is drinking alcohol bad?*"). Ideally, this message should be annotated as *Lifestyle* and not *Other* in the *Informational* category. Since the point of human annotation is to indeed identify deeper meaning which a naive algorithm might miss, the annotators were asked to tag the messages after decoding the meaning of the emoji.

5.4 Observations

The distribution of peer support category labels for Phase-3 in both the groups are summarized in Table 1. We observed that a majority of the messages in both the groups belong to *Chit-Chat* (65.9% in Group-1 and 66.2% in Group-2), followed by *Informational*. Within *Informational* category, *Medical* stands out as the most prominent for Group-2 (35.0% informational messages) while *Admin* is the most prominent in Group-1 (34.6% informational messages).

It should be noted that the IAA is relatively lower for *Hopefulness*, *Admin*, and *Group Work* indicating disagreement between the annotators. Overall, *Chit-Chat* and *Acknowledgement* are easier to annotate as is reflected in their IAA score. Further, we also observe that most messages are short in length with an average message length of 4.34 words in Group-1 and 5.6 words in Group-2.

Peer Support	Group-1	Group-2
Informational		
Medical	57	445
Lifestyle	24	81
Personal	50	392
Admin	86	195
Other	31	158
Emotional		
Empathy	21	42
Hopefulness	27	93
Happiness	5	31
Negativity	29	51
Chit-Chat		
Socialization	1168	2316
Greetings	232	695
Acknowledgement	242	615
Group Work	17	225
Other	138	304

Table 1: Overall distribution of peer support categories in both the peer support groups after Phase-3.

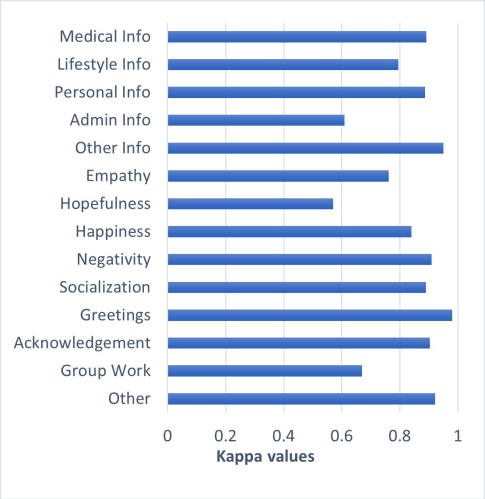

Figure 3: IAA for peer support categories in Phase-2

Word-Level Language	Group-1	Group-2
Monolingual		
En Only	805	2278
Sw Only	262	633
Sh Only	64	142
CM Only	3	9
Multilingual		
En Sw	171	706
En Sh	16	84
En Sw Sh	54	198
En CM	4	17
Sw Sh	70	182
En Sw CM	15	63
En Sh CM	2	5
Sw Sh CM	2	24
Sw CM	14	38
En Sw Sh CM	8	40

Table 2: Overall distribution of the messages containing different language category labels in both the groups after Phase-3. The nature of messages (monolingual or multilingual) is determined by the language tags for each of the words in the messages.

With respect to *Sentiment* labels, most of the messages were tagged as *Neutral* (93.4% in Group-1 and 87.1% in Group-2), followed by *Negative* (5.0% in Group-1 and 4.3% in Group-2) and then *Positive*. It was also found that the annotators agreed mostly on *Negative* sentiment (Phase-2 κ=0.92) and disagreed the most on *Positive* (Phase-2 κ=0.67).

Table 2 describes the results for language tagging after Phase-3. It shows that the IAA score for *English (En)* (κ=0.98) and *Kiswahili (Sw)* (κ=0.98) words is higher compared to those of *Sheng (Sh)* (κ=0.87) and *Code-Mixed phrases (CM)* (κ=0.85). It can also be observed that the messages contain a high amount of code-mixing (23.8% messages are code-mixed in Group-1 and 24.4% in Group-2). While English is the dominant language for the monolingual messages, code-mixed *English Kiswahili (En Sw)* is the dominant language-pair for the multilingual messages.

We found that a change of topic was indicated in 3% and 4% of the chat messages in Group-1 and Group-2, respectively. In Group-1, a topic change was initiated mostly by the admin (89% of topic changes), while in Group-2, other group members have also been found to initiate discussions on new topics (only 39% of total topic changes were triggered by the admin). Further, we observe that a topic change is mostly (95.2%) associated with the *Informational* or *Group Work* subcategories.

Due to the informal nature of WhatsApp conversations, a significant percentage of chat logs contain spelling mistakes (19.4% and 18.7% messages in Group-1 and Group-2, respectively) or abbreviations (15.7% and 10.9% messages in Group-1 and Group-2, respectively).

6 Conclusion and Future Work

In this paper, we presented a comprehensive annotation schema, combining several linguistic and pragmatic phenomena, to enable studies of WhatsApp based conversations of youth living with HIV. Unlike Twitter and Facebook, this requires understanding the content of previous discourse to annotate the sentiment and conversational intent of supporting the peers in specific examples. During the process of annotation, we have learnt that a good percentage of examples require understanding the context. To the best of our knowledge, we are the first to combine several standard annotation schemes to annotate the multilingual conversations in a small, close-knit community.

We believe our schema can be used to flag important messages (based on user concerns) to the healthcare providers. This will in turn help to alleviate the need to scroll through all the messages individually to identify the ones that require a response from the facilitators. Linguists can also use this framework to study the nature of multilingualism in such conversations from both sociolinguistic and structural perspectives. In future, we plan to use the framework to further study the nature of engagement and multilingual interactions in such peer-support groups in healthcare domain with the aim of enhancing remote patient care.

Acknowledgement

We thank Professor Keshet Ronen, Naveena Karusala, and other members from the University of Washington for providing us access to the datasets. We also thank Samuel Maina from Microsoft Africa Research Institute (MARI) for providing his feedback on the work, and the three annotators from Microsoft Research India Lab for putting a considerable amount of time in annotation, and giving us their feedback on the schema. Finally, we acknowledge the efforts of Dr. Seema Mehrotra from the National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore during the initial ideation around an annotation framework on healthcare conversations.

References

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. *Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns*, page 112. Association for Computing Machinery, New York, NY, USA.

JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64, Jeju Island, Korea. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *CodeSwitch@EMNLP*.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74, Montréal, Canada. Association for Computational Linguistics.

Karthik S. Bhat, Mohit Jain, and Neha Kumar. 2021. Infrastructuring telehealth in (in)formal patient-doctor contexts. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Sven Buechel, Anneke Buffone, Barry Slaff, L. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *EMNLP*.

Graham Button and Wes Sharrock. 1997. *The Production of Order and the Order of Production: Possibilities for Distributed Organisations, Work and Technology in the Print Industry*, pages 1–16. Springer Netherlands, Dordrecht.

Fabio Celli and Luca Rossi. 2012. The role of emotional stability in Twitter conversations. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 10–17, Avignon, France. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text.

M. D. Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

M. D. Choudhury, Emre Kcman, Mark Dredze, Glen A. Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

B. S. Chumbow. 2010. Linguistic diversity, pluralism and national development in africa. *Africa Development*, 34:21–45.

P. Ciechanowski, W. Katon, J. Russo, and E. Walker. 2001. The patient-provider relationship: attachment theory and adherence to treatment in diabetes. *The American journal of psychiatry*, 158 1:29–35.

Andy Crabtree. 2003. *Designing Collaborative Systems: A Practical Guide to Ethnography*. Springer, London.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. the last language identification frontier? *Trait. Autom. des Langues*, 54:41–64.

Nassim Dehouche. 2020. Dataset on usage and engagement patterns for facebook live sellers in thailand. *Data in Brief*, 30:105661.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and mt augmentation approach.

A. Dwivedi. 2015. Linguistic realities in kenya: A preliminary survey.

R. Elliott, A. Bohart, J. Watson, and D. Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55:399410.

James Gibson, Dogan Can, Bo Xiao, Zac E. Imel, David C. Atkins, P. Georgiou, and Shrikanth S. Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. In *INTERSPEECH*.

Chege Githiora. 2002. Sheng: Peer language, swahili dialect or emerging creole? *Journal of African Cultural Studies*, 15(2):159–181.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29, Portland, Oregon. Association for Computational Linguistics.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135, Copenhagen, Denmark. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13018–13026.

John Hughes, Val King, Tom Rodden, and Hans Andersen. 1994. Moving out from the control room: Ethnography in system design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, page 429439, New York, NY, USA. Association for Computing Machinery.

Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018a. Farmchat: A conversational agent to answer farmer queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4).

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018b. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 895906, New York, NY, USA. Association for Computing Machinery.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2016. Collecting and annotating indian social media code-mixed corpora. In *CICLing*.

Naveena Karusala, David Odhiambo Seeh, Cyrus Mugo, Brandon Guthrie, Megan A Moreno, Grace John-Stewart, Irene Inwani, Richard Anderson, and Keshet Ronen. 2021. *That Courage to Encourage: Participation and Aspirations in Chat-Based Peer Support for Youth Living with HIV*. Association for Computing Machinery, New York, NY, USA.

Naveena Karusala, Ding Wang, and Jacki O'Neill. 2020. *Making Chat at Home in the Hospital: Exploring Chat Use by Nurses*, page 115. Association for Computing Machinery, New York, NY, USA.

Scott F. Kiesling, Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han, and Jacob Eisenstein. 2018. Interactional stancetaking in online forums. *Computational Linguistics*, 44(4):683–718.

Taisa Kushner and Amit Sharma. 2020. Bursts of activity: Temporal patterns of help-seeking and support in online mental health forums. In *Proceedings of The Web Conference 2020*, WWW '20, page 29062912, New York, NY, USA. Association for Computing Machinery.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA. Association for Computational Linguistics.

David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*, CSCW '14, page 224235, New York, NY, USA. Association for Computing Machinery.

Jacki O'Neill and David Martin. 2003. Text chat in action. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '03, page 4049, New York, NY, USA. Association for Computing Machinery.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas. Association for Computational Linguistics.

Koustav Rudra, Ashish Sharma, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2019. Identifying and analyzing different aspects of english-hindi code-switching in twitter. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(3).

Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. 2015. Predicting continued participation in online health forums. In *Proceedings of the Sixth International Workshop on*

Health Text Mining and Information Analysis, pages 12–20, Lisbon, Portugal. Association for Computational Linguistics.

Antoinette Schoenthaler, William F. Chaplin, John P. Allegrante, Senaida Fernandez, Marleny Diaz-Gloster, Jonathan N. Tobin, and Gbenga Ogedegbe. 2009. Provider communication effects medication adherence in hypertensive african americans. *Patient Education and Counseling*, 75(2):185–191.

Rosalie Schultz, Stephen Quinn, Byron Wilson, Tammy Abbott, and Sheree Cairney. 2019. Structural modelling of wellbeing for indigenous australians: importance of mental health. *BMC Health Services Research*, 19(1):488.

Ashish Sharma, M. Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *ICWSM*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An annotation framework for Luxembourgish sentiment analysis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 172–176, Marseille, France. European Language Resources association.

Jennifer Smith-Merry, G. Goggin, A. Campbell, Kirsty McKenzie, Brad Ridout, and C. Baylosis. 2019. Social connection and online engagement: Insights from interviews with users of a mental health online forum. *JMIR Mental Health*, 6.

Sahil Swami, A. Khandelwal, Vinay Singh, S. Akhtar, and Manish Shrivastava. 2018. A corpus of english-hindi code-mixed tweets for sarcasm detection. *ArXiv*, abs/1805.11869.

Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. 2017. Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 11–16, Valencia, Spain. Association for Computational Linguistics.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Cretien van Campen and Jurjen Iedema. 2007. Are persons with physical disabilities who participate in society healthier and happier? structural equation modelling of objective participation and subjective well-being. *Quality of Life Research*, 16(4):635.

S. Vieira, U. Kaymak, and J. Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. *International Conference on Fuzzy Systems*, pages 1–8.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Ramaswamy Viswanathan, Michael F. Myers, and Ayman H. Fanous. 2020. Support groups and individual mental health care via video conferencing for frontline clinicians during the covid-19 pandemic. *Psychosomatics*, 61(5):538–543. 32660876[pmid].

Yogarshi Vyas, Spandana Gella, Jatin Sharma, K. Bali, and M. Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *EMNLP*.

Ding Wang, Santosh D. Kale, and Jacki O'Neill. 2020. *Please Call the Specialism: Using WeChat to Support Patient Care in China*, page 113. Association for Computing Machinery, New York, NY, USA.

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, Minnesota. Association for Computational Linguistics.