Natural language processing of WhatsApp messages as a tool to identify unrecognized HIV prevention needs among young women in Kenya

1079

Jessica E Haberer[1,2], Bernard Rono[3], Rakesh Ravi[4], Vinal Bagaria[4], Sinchana Shetty[4], Shanmukh Moparthi[4], Kevin Oware[3], Josephine B. Odoyo[3], Lindsey Garrison[1], William Rand[4], Elizabeth A. Bukusi[3,5]

[1]Massachusetts General Hospital, Boston, MA, USA, [2] Harvard Medical School, Boston, MA, USA, [3] Centre for Microbiology Research, Kenya Medical Research Institute, Kisumu, Kenya, [4] North Carolina State University, Raleigh, NC, USA, [5] Departments of Obstetrics and Gynecology and Global Health, University of Washington, WA, USA

BACKGROUND

- Identifying HIV prevention needs is challenging
- Common approaches (e.g., public health campaigns, HIV risk scores) have limited sensitivity
- Natural language processing (e.g., topic modeling and computational linguistics) is a type of artificial intelligence that can extract meaning and subjective information from text
- This study used natural language processing techniques with WhatsApp messages to identify unrecognized HIV prevention needs among young women in Kenya

METHODS

- Extensive prior work established key ethical protections (e.g., privacy, confidentiality) for participants
- Women (18-24 years) who used WhatsApp were enrolled from four clinical sites in Kisumu, Kenya
- After completing a socio-demographic questionnaire, participants securely transferred all WhatsApp messages from the prior 6 months and provided anonymized relationship types (e.g., friend, sexual partner) for their 20 most frequent contacts
- Data were de-identified, and all incoming messages and photos were deleted
- WhatsApp data were pre-processed using a downloaded large language model to translate Swahili and Dholuo to English and account for slang linguistic patterns
- We explored multiple approaches to topic modelling in the WhatsApp data (e.g., latent Dirichlet allocation)
- Non-negative matrix factorization and uniform manifold approximation and projection allowed for comparison of participant word topics with the VOICE HIV risk score (range 0-9, 9 indicating high risk; *Balkus, JAIDS 2016*)

Natural language processing of routine WhatsApp messages may identify unmet HIV prevention needs, thus potentially reaching individuals not otherwise receiving services

RESULTS

- Of 413 eligible young women, 13 (3%) declined to provide WhatsApp data and 22 had insufficient data for analysis
- Participants had a mean age of 22 (SD 1.6) years, 2.5 (SD 2.4) sexual partners in the past 3 months, and a VOICE risk score of 5.7 (SD 1.4); 84% were single and 71% lived in a semi-urban setting
- The remaining 378 women provided a median of 14 chats (IQR 11-20) per participant with a median of 478 words (IQR 198-894) for a total of 6,607 files; chats lasted a median of 20 days (IQR 5-100)
- Figure 1: Dominant topics (i.e., 10-word clusters) were selected based on word distributions per participant and/or seeds (e.g., words like “sex” that might be related to HIV risk)
- Figure 2: The median VOICE risk score was 7 for women whose most messaged relationship was a sexual partner versus 5 for those messaging a friend most often (p>0.05)
- Figure 3: Topic models had differential correlation with the VOICE risk score that was highest when indicating romance/affection (e.g., topics 2 and 4) and lowest when indicating faith/positivity (topic 6) and transactions/family (topic 8)
- Figure 4: Dimensional analysis of linguistic similarity identified discrete participant populations with distinct VOICE risk scores, although the presence of overlap indicates some participants had blended conversational cues

Figure 1. Sample word cloud of WhatsApp messages



Figure 2. Distribution of the VOICE risk score by most messaged relationship (i.e., friend, sexual partner, family member, or other)

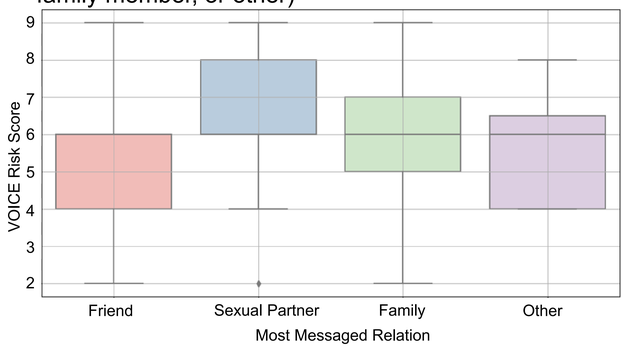
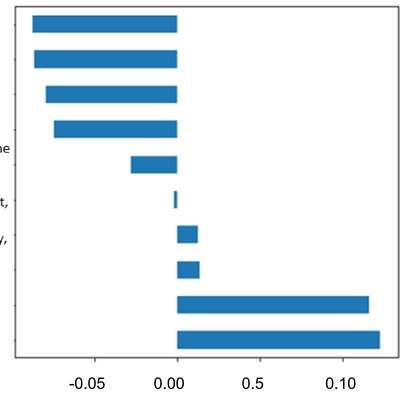


Figure 3. Non-negative matrix factorization showing correlation between sample topics and the VOICE HIV risk score

- Topic 6: much, well, around, also, name, great, low, amen, fish, someone, nice, auntie, pay, take, may
- Topic 8: send, money, number, give, pay, use, tomorrow, help, let, group, mother, ask, name, bring, phone
- Topic 3: special, fine, well, low, still, japanese, around, none, word, brother, safe_secure, maybe, alas, fuel, around
- Topic 5: fine, also, maybe, welcome, whoa, fuel, poor, dear, happy_birthday, mother, school, oooh, home, miss, long_time
- Topic 10: void, fine, fuel, happy_valentine, house, work, great, madam, leave, daddy, hbd, still, phone, why
- Topic 1: want, let, grow, even, give, leave, tomorrow, still, people, eat, sleep, time, day, think, back
- Topic 7: get, job, back, call, die, home, find, still, yet, tomorrow, today, pay, time, yup, miss
- Topic 9: tell, call, mmh, talk, ask, around, something, number, story, lie, busy, want, you, help, phone
- Topic 2: good, great, nice, good_night, look, welcome, also, day, friend, good_afternoon, miss, long_time, work, hope, dear
- Topic 4: love, babe, miss, baby, dear, sleep, good_night, want, beb, feel, eat, today, always, day, happy



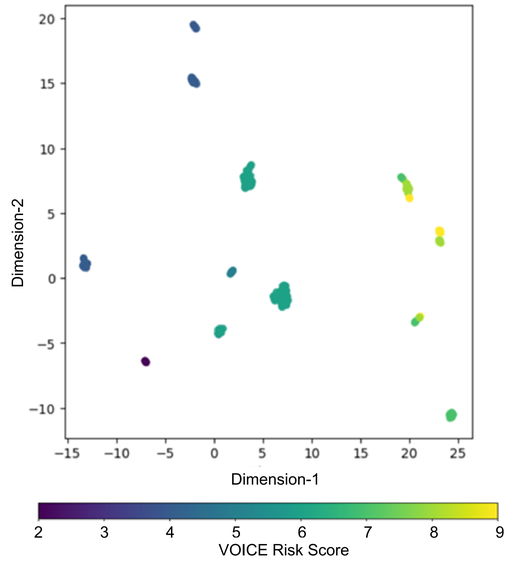
CONCLUSIONS

- This study provides proof-of-concept that natural language processing of routine WhatsApp messages can yield meaningful individual-level information about HIV prevention needs.
- Future work should compare such findings with HIV incidence to optimally determine its ability to identify unrecognized HIV prevention needs, thus potentially expanding HIV prevention services to individuals not currently engaged in care.

PLAIN LANGUAGE SUMMARY

Many people with HIV prevention needs are not aware of them and do not seek services. Automated analysis of word patterns in routine WhatsApp messages were found to correlate with HIV risk. Future app development with this type of artificial intelligence could alert people to their risk and guide them to seek HIV prevention services.

Figure 4. Uniform manifold approximation and projection visualization showing clustering of participants with linguistic similarity as related to the VOICE risk score



ADDITIONAL KEY INFORMATION

Funding: R21MH132436

Author Contact Information: jhaberer@mgh.harvard.edu

Acknowledgements: The authors would like to thank study participants and study staff for their valuable contributions