

Evaluating Crowdsourcing Behaviours in Identifying Online Misinformation

An Experimental Evaluation of the Healthy Internet Project

September 2021



Co-building the Accelerator Labs as a joint venture with:



UNDP
Core
Partners

About Us

UNDP Accelerator Labs

The United Nations Development Programme (UNDP) is the leading United Nations organization fighting to end poverty, inequality, and climate change. The UNDP Accelerator Labs is the world's largest and fastest learning network on wicked sustainable development challenges. The network of 91 labs covers 115 countries and taps into local innovations to create actionable insights and reimagine sustainable development for the 21st century. Learn more at acceleratorlabs.undp.org or follow us at @UNDPAccLabs

Busara Center for Behavioral Economics

Busara is a research and advisory firm dedicated to advancing and applying Behavioral Science in the pursuit of poverty alleviation in the Global South. Busara is spread across 5 offices in Africa and Asia; and works with clients to enable them to understand behaviors, and to design and test solutions to scale their interests. Busara has completed strategic partnerships with an array of private-sector organizations, as well as with governments, with NGOs, and with academic institutions, that are interested in leveraging behavioral science insights to improve upon outcomes. Learn more at: www.busaracenter.org



Acknowledgements

We would like to thank Renae Reints, Anand Upender and the entire technical team at the Healthy Internet Project for their support and input to the research, and for offering guidance where necessary.

We would like to thank Enock Nyariki, Christine Mutisya, Jessica Manim, Doreen Wainainah, and Collins Nabiswa for their time, efforts and contributions to the flag quality check exercise. We would like to thank Caroline Kiarie-Kimondo and Victor Awuor from UNDP Kenya for their Insights, guidance and feedback on this research.



How to cite this report:

Canagarajah, R., Ogutu, B., Mugi, N., Too, G., Njoro, L. (2021). *Evaluating Crowdsourcing Behaviors In Identifying Online Misinformation.*

List of Abbreviations

UNDP - United Nations Development Programme

HIP - Healthy Internet Project Incubated at TED

TED - Technology, Entertainment, Design

Table of contents

→ Executive Summary	5
→ Introduction	9
→ Key Findings	12
→ Evaluation Methods	16
→ Detailed Findings	29
→ Recommendations from Full Study	84

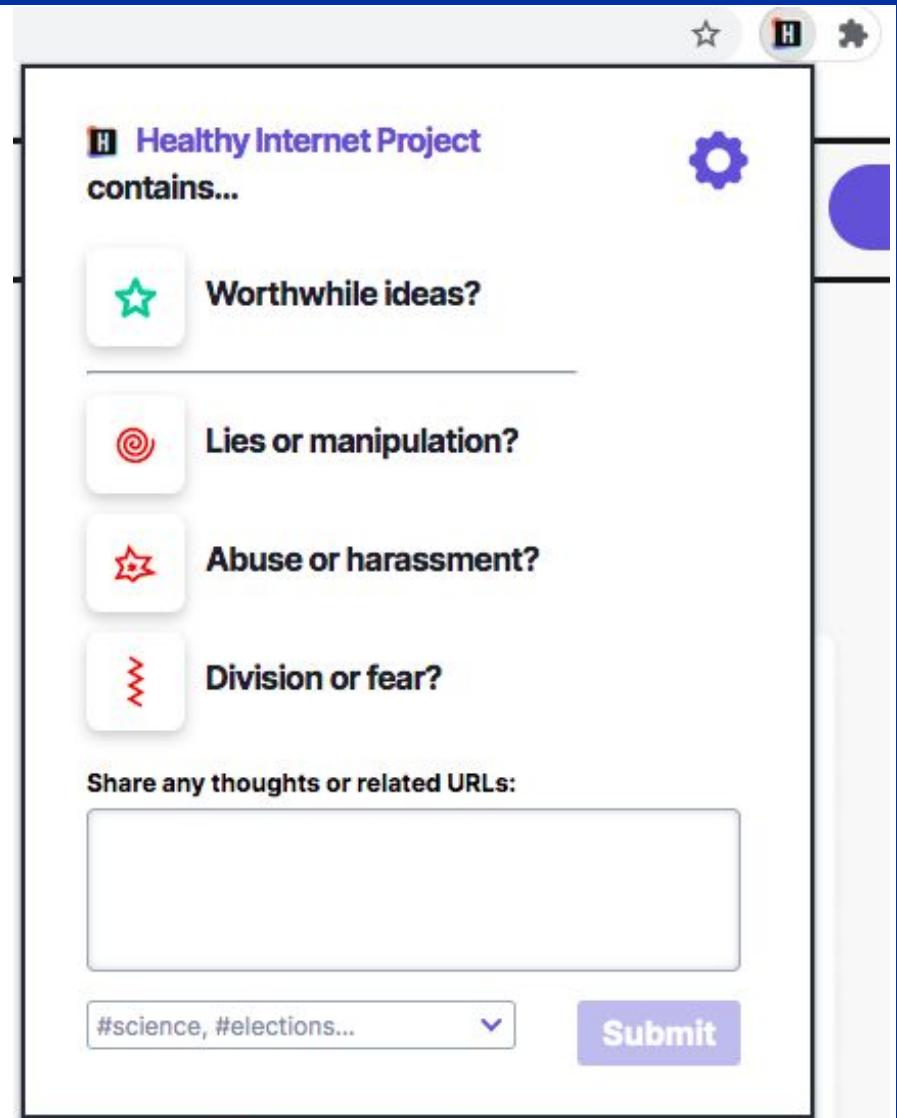
Executive Summary

Brief **overview** of **project activities**, a **summary** of **flagging trends** and our **recommendations**

Executive Summary: Part 1

Busara conducted a live experimental demonstration of the **Healthy Internet Project (HIP)** plugin, in collaboration with **UNDP Accelerator Lab Kenya**, and the **Healthy Internet Project (HIP)** incubated at **TED**. The **Healthy Internet Project** plug-in (displayed on the right) is an [open source web browser extension](#) that allows users to flag content online; it is intended to help curb the spread of lies, abuse, and fear mongering, as well as to uplift useful ideas on the internet. Users are able to mark flags as mild / minor, medium, or severe across the latter flagging categories.

Our experimental design sought to understand potential users' motivations, experiences, and practices in using the volunteer-driven, crowdsourcing platform to flag misinformation in a live experiment which encouraged natural behaviors.



Executive Summary: Part 2

We onboarded 205 users but only 128 used the platform. We followed up with 44 of these users in a qualitative exercise. In **user patterns** of flagging, the large majority of our respondents - also referred to as users- (N=109) fell into the Low Flagging Group (i.e., less than 5 flags). The majority of respondents (75%) flagged Worthwhile content. In **user perceptions**, there were concerns that flagging negative content was 1) more subjective; 2) might have led to harmful repercussions for those who are flagged; and 3) was personally risky, especially vis a vis political content. In terms of **user accuracy**, “misinformation” was viewed through negative user sentiments, such as a dislike for a topic, rather than as misinformation itself. Moreover, it was difficult for external fact checkers to know what constituted misinformation amongst flagged content, because the majority of flag-level comment sections went unused (81% of flags), or were ill used i.e., users hardly specified exactly what was misinforming about the websites they were on.

We conclude that the value-add of volunteer driven misinformation identification is limited without changes to user perception of safety, and of accuracy. There is need for a better understanding of misinformation to enhance objectivity in flagging activity. In order to improve platforms like HIP, we share the following recommendations:

Executive Summary: Part 3

- **User Experience**

- Ensure Anonymity: There should be more details to convince users of their anonymity to address the risks they feel on reporting misinformation.

- **User Accuracy**

- Provide a Misinformation Introduction: We believe a primer on misinformation should be present to increase accuracy of user reports.
- Remove Worthwhile Flag: We also believe there is value in reconsidering and potentially removing the Worthwhile flag to solidify the purpose of the plug-in.
- Add a Required “Misinformation Identification” Field: Finally, for the ease of fact checkers, we recommend adding a required field in which users specify the content they deem as misinformation. This should not be the website URL, but the actual content they believe to be misinformation.

Introduction

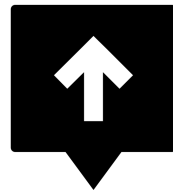
Project **timeline**, project **background**,
and project **activities**

Project Background

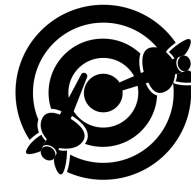
- **UNDP** and the Healthy Internet Project, incubated at TED engaged in a partnership to accelerate testing and to generate actionable data on TED's Healthy Internet Project (**HIP**) and its prototype tool. This crowdsourcing tool helps to identify and review worthwhile ideas as well as harmful content found on the internet. UNDP Kenya, through the Accelerator Lab, has partnered with **Busara** to conduct a pilot study in Kenya with a prototype of the HIP platform.
- The objectives of this study were to explore the following learning themes:



Usage **patterns**



Use **motivations and barriers**



User **accuracy**



Value proposition in addressing misinformation

- With a view to analyzing user accuracy, we engaged the support of [PesaCheck](#), Africa's largest indigenous fact-checking organisation, to validate a sample of the claims associated with flagging activity from the study.

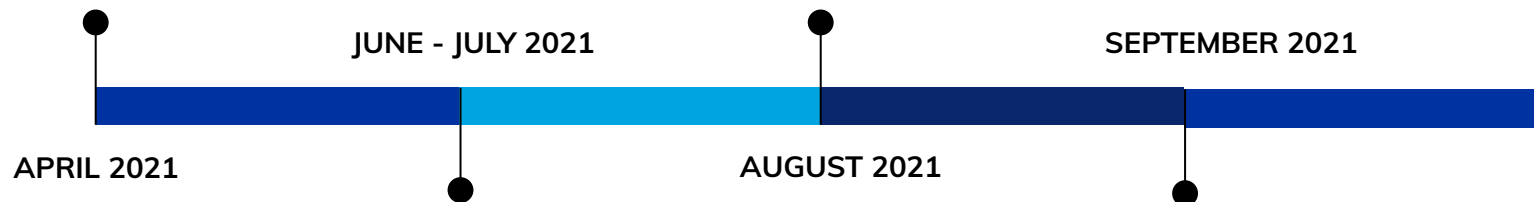
Project Timeline

Phase 0: Stakeholder Alignment

In collaboration with the UNDP Accelerator Lab Kenya, we proposed to conduct a study with the HIP platform to develop a deeper understanding of potential motivations and barriers for platform users.

Phase 2: Mixed Methods Behavioral Mapping

We used mixed methods in-depth interviews and a focus group discussion to analyze behavioral factors relevant to user engagement in crowdsourcing reporting approaches; with an aim to understanding context-specific motivations and barriers to HIP plug-in use.



Phase 1: Live Demonstration

We recruited 205 study participants from its database to pilot the HIP plug-in; we observed participant behavior across three key domains: user experience, user motivation, and user reporting accuracy.

Phase 3: Communications & Reporting

In this stage, we shared with UNDP insights from the study and recommendations on the way forward. We also collaborated with both organizations in creating blogs and publicly accessible documentation on our evaluation.

Key Findings

Overview of key **findings** based on quantitative and qualitative insights

On User Motivation and Perceptions



Majority Have Never Reported

Many respondents had never reported misinformation before HIP, despite knowing other methods. A smaller amount (<10) from qualitative study say that they flagged misinformation in other ways (i.e., Safaricom, report buttons).



Belief that HIP is a Good Tool

Participants consider HIP an appropriate tool for stopping the spread of misinformation. This tool may present a more formal approach to flagging misinformation given its explicit purpose in achieving this compared to social media avenues.



Risk in Reporting: Anonymity

Users fear that they will be identified through platform use, especially for political or relevantly sensitive information. When it comes to political actors, users are unconvinced of the platform's promise.



Risk in Reporting: Repercussions

Users do not know if their report on misinformation is accurate (i.e., objective vs. subjective). Moreover, they fear their actions may lead to negative consequences for those behind the websites or articles identified, or to themselves.

On User Engagement



Overwhelming Use of Worthwhile

Despite the intent to stop the spread of misinformation, most people use the tool to flag worthwhile content. The Worthwhile flags constituted the majority of flags (75%) used by our participants.



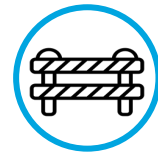
Most Use on Weekday Morning

User analysis suggests that our respondents used the platform most during weekday morning hours, and very infrequently over the weekend.



Lies+Manipulation Most “Severe”

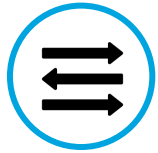
Despite Worthwhile being the most popular flag of use, it often received minor rankings. Lies and Manipulation however received the highest severity levels among all flags.



Barriers in Continued Use

Internet challenges and not frequently coming across harmful content partly explains the low usage of HIP. Other barriers include only being bound to PC for HIP use, and the lack of feedback mechanisms (i.e., receiving an update on the actions taken by the provider). Monetary incentives could increase usage of HIP.

On User Accuracy



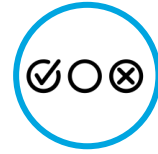
Mixed Self Perceptions on Accuracy

Some respondents believe that they flag accurately, despite the study finding otherwise in the monitoring phase. Yet others are not confident or comfortable in their interpretation of how to flag misinformation



Misinformation Trainings

Many users strayed away from flagging misinformation for two reasons: 1) they were not comfortable in the subjectivity; and 2) they did not want to harm anyone. Trainings and definition activities on misinformation and what happens post-flag validation could address this.



PesaCheck: Lack of Validation

Based on PesaCheck findings of 10 core flags, flagged content was shown to be sentiment-driven and often too broad to say what element of a website was misinformation or not.



Core Attention on Worthwhile Flag

Qualitative findings showed that more people than expected do more research on the validity of worthwhile content, than for harmful content.

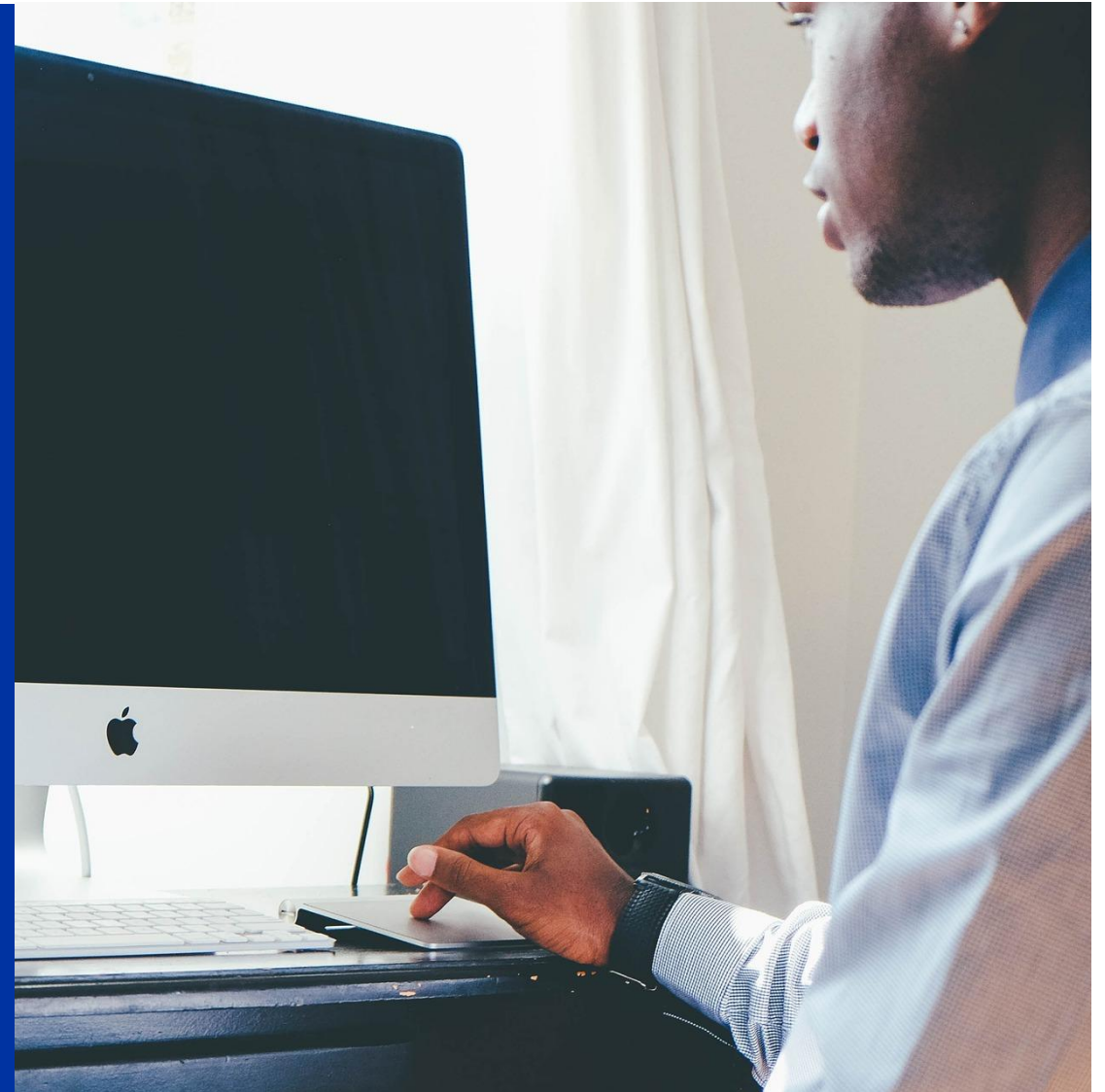
The background of the slide is a photograph of a coastal scene. In the foreground, there is a body of water with some ripples. In the middle ground, several boats are visible, including a prominent blue boat on the left and a smaller boat on the right. The sky is filled with soft, white clouds. A large, solid blue rectangular overlay covers the right two-thirds of the slide, containing the text.

Evaluation Methods

Insights into the live experiment, data analysis
and qualitative follow-ups

Quantitative Live Experiment

Overview of live experiment approach, data analysis limitations, and demographics



Purpose of the **live experiment study**:

To use a live experiment to observe “natural”* behaviors on the platform, leading to understanding user experience, user motivations, user accuracy, and segmented demographic trends of platform use based on volunteered engagement.**

*We exposed and trained users on the HIP platform, therefore their behaviors on the platform may not have been entirely natural. However, we did not have any other contact with them during the run of the live experiment.

**Respondents were not incentivised or paid to report. It wasn't until after the study that they knew they would be compensated.

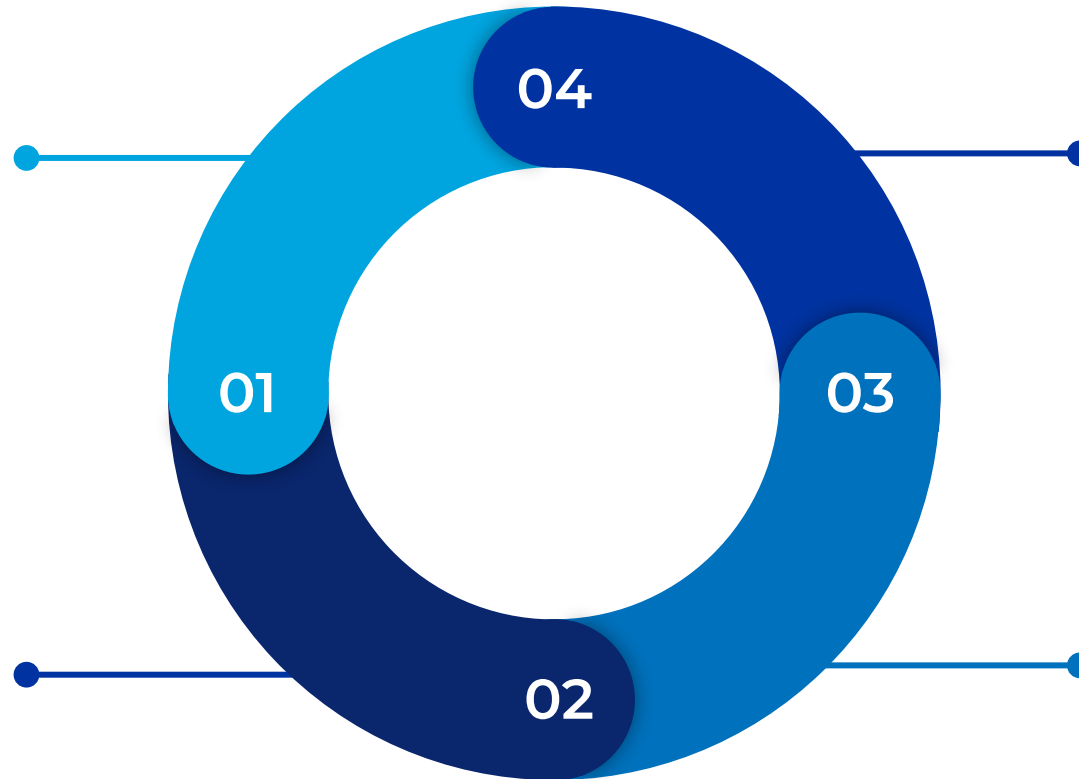
Live Experiment Approach

Live Experiment Set-Up

Prior to the evaluation, we determined demographics and locations of interest, particularly those on internet usage/access. We then designed a light-touch onboarding script to encourage natural behaviors and created a pre-analysis plan (PAP) based on metrics of interest.

Onboarding

Based on demographics segments, we used stratified random sampling, to recruit 205 final respondents (from Busara's existing database) in a staggered timeline. After each respondent was onboarded through the script, we collected user HIP IDs and merged them with Busara IDs.



Data Analysis and Findings

After respondents ended their 4 week trial of the HIP platform, we analysed usage patterns by demographics; user perceptions/ attitudes towards the platform; types of flags raised; and validity of flags, among other factors. we pre-coded flag categories and user behaviors according to trends in the data.

Live Demonstration

After light-touch onboarding, we allowed users to engage with the platform at their own will for 3+ weeks. Respondents were monitored on their HIP platform usage.

Summary for Live Experiment



371

Total Number
of Flags



57%

Male



43%

Female



128

Total Number of Users
(Post-attrition)

Demographic Summary for Live Experiment

Age Group

18 - 24	--	42%
25 - 39	--	52%
40 - 49	--	5%
50 - 60	--	≈ 1%

Education

Primary	--	< 1%
Secondary	--	44%
Post - Secondary	--	65%

Ethnic Group

Gĩkũyu	--	43%
Luo	--	17%
Kamba	--	13%
Kisii	--	9%
Luhya	--	8%
Meru	--	5%
Kalenjin	--	4%
Maasai	--	2%
Turkana	--	1%

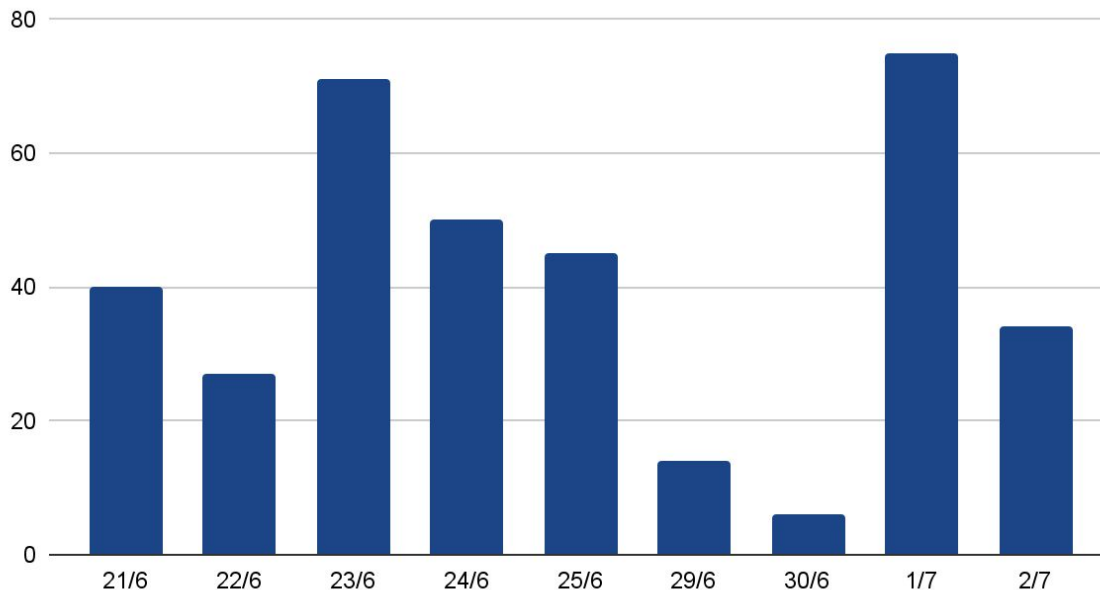
Location

Kiambu	--	34%
Nairobi	--	33%
Machakos	--	19%
Kajiado	--	10%
Muranga	--	5%

The study participants were onboarded on a staggered schedule

- User participation in the **onboarding sessions** was determined at **random**. All respondents were exposed to at minimum **2 weeks** on the platform. Onboarding entailed: explaining the study, guiding respondents in downloading the tool, and a practice session of flagging content.

Number of Participants Onboarded by Onboarding Date



healthy internet project plugin - ... contains...

- Worthwhile ideas?**
- Lies or manipulation?**
- Abuse or harassment?**
- Division or fear?**

Share any thoughts or related URLs:

#science, #elections...

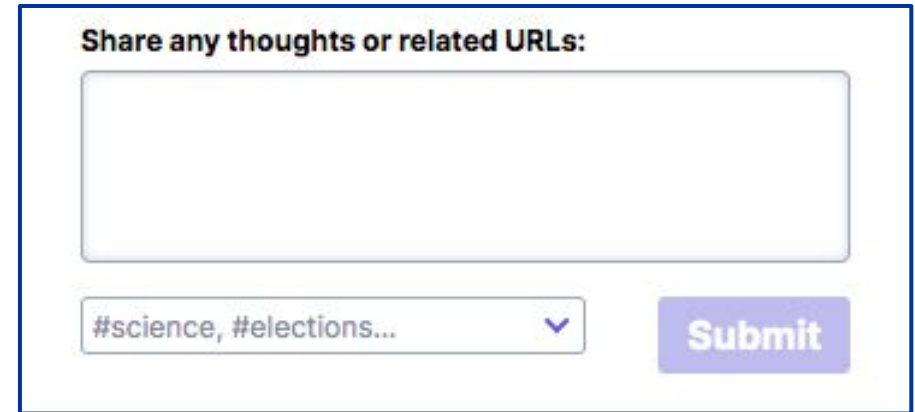
Submit

Limitations in Data and Analysis

1. **Non-Participation:** Of the **128 onboarded** study participants for whom data was included in this analysis, only **40% flagged more than one (1) item** using the HIP plug-in -- reducing data diversity thereby **impeding generalizability** of the conclusions from this analysis.
2. **Representation:** **100%** of study participants considered in this analysis, identified as **Christian**; **< 1%** had **completed primary school** education . This means that we cannot **generalize** the findings to a wider population since the whole population has varied religious denominations and education levels.

Limitations in Data and Analysis (CONT.)

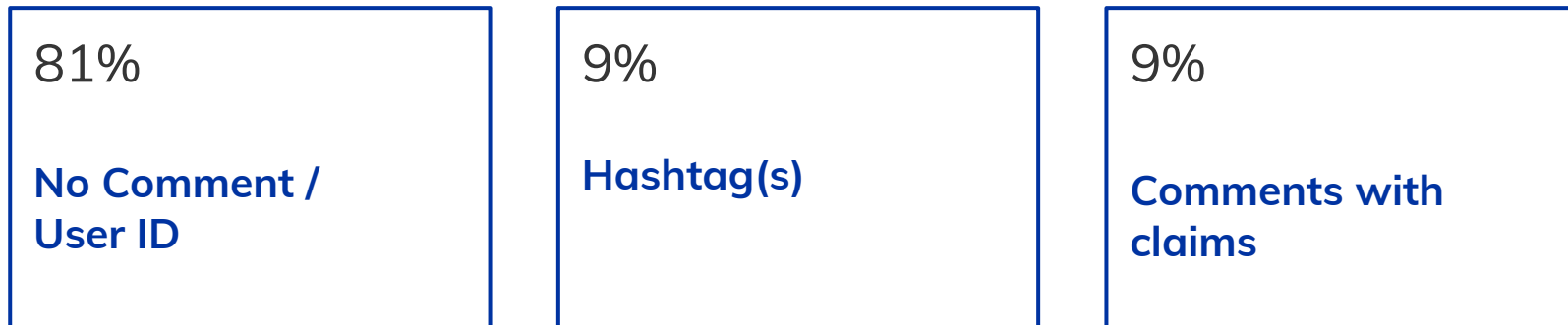
3. Broad categories: Comment categories are defined broadly (as illustrated on the right) -- **narrower** definitions would have required **claim categorization** at the individual comment level -- a task for which we were not able to bring **capacity** to bear upon for this project.



Share any thoughts or related URLs:

#science, #elections... ▼

Comment Categories:



Qualitative Interviews

Overview of our follow-up qualitative approach and demographics



Purpose of the **qualitative study**:

To understand context specific insights as it relates to users' demographics and motivations, and to get a more nuanced understanding of the findings captured in the live experimentation phase.

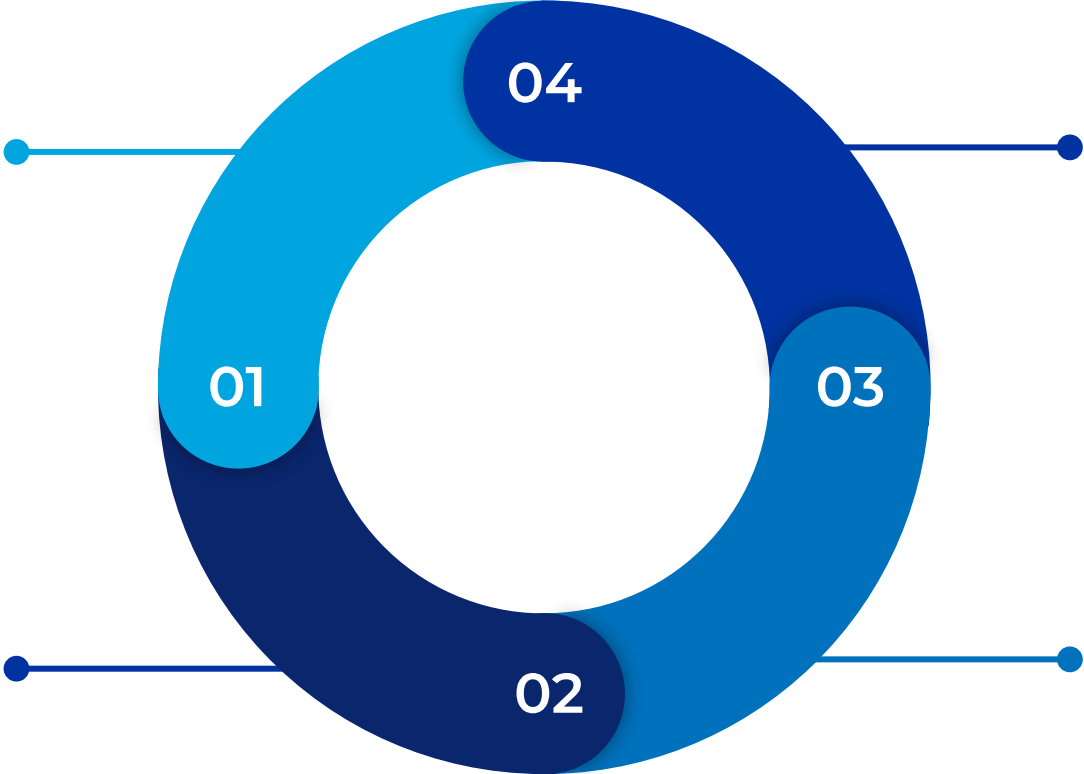
Qualitative Approach

Design of Qualitative Tools

We designed a qualitative instrument that was reviewed and approved by UNDP.

Conduct In-Depth Interviews (IDIs) and Focus Group Discussions (FGDs)

We conducted 39 phone in depth interviews and 1 in-person focus group discussion between August 24th 2021 and August 27th 2021. They were conducted with people from 9 counties.



Presentation of Findings

The analysis of the qualitative data brought out key findings that then informed the recommendations relevant for UNDP.

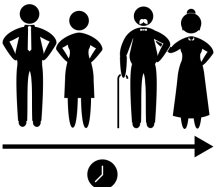
Analysis of Findings

The audio files from the in depth interviews and focus group discussions were transcribed and thereafter analyzed.

Demographic Summary for the Qualitative Study (N=44)

We conducted in depth interviews with 39 individuals, and held one focus group discussion with 5 people. Below is the demographic breakdown of this sample:

Age Group



20-25 - 65%
26-30 - 21%
31-35 - 14%

Gender



Male - 49%
Female - 51%

Education



Secondary School - 25%
Tertiary Education - 75%

Marital Status



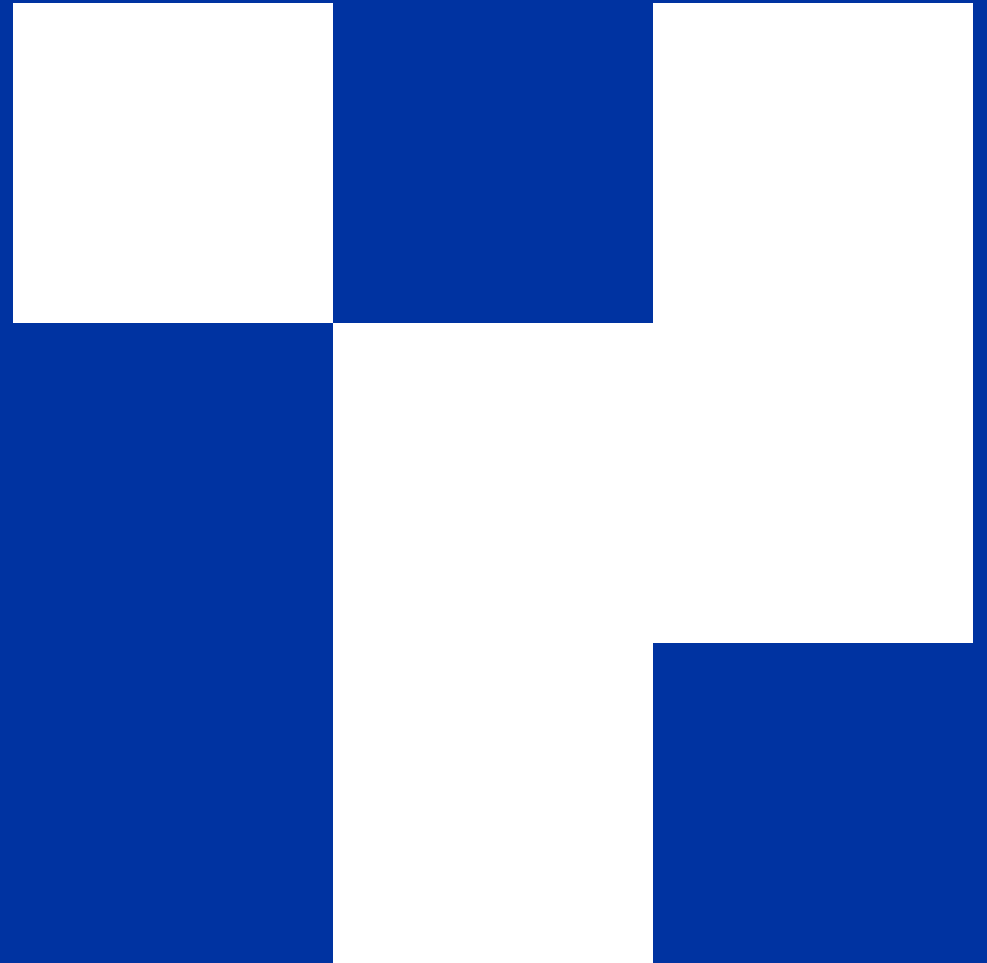
Single - 87%
Married - 13%

Detailed Findings

Insights into user backgrounds, user HIP practices, segmented user insights, user accuracy, and user motivations

HIP User Backgrounds

Overview of user practices on the internet and
exposure to misinformation



As expected, internet usage is prevalent amongst our respondent pool, which consists mostly of youth.

Personal Interests

- Our respondents were split equally in terms of paying attention to current affairs or not; while most respondents pay attention to social media. Additionally, most of the respondents have volunteering experience which they see as a prosocial responsibility.

Digital Services Usage

- All the respondents use both their phone and laptop or computer to access the internet. Phone usage is however predominant because of its portability and ease of information access. Laptops are mostly used for work.

Computer Proficiency

- Most of the respondents stated that they have advanced skills in computer use, with average years of computer usage being between 5 and 6 years.

Internet Usage

- Most participants use the internet to access information and share it with others. Many of them receive information from others through their social media platforms, and only a few access information from mass media channels, i.e., TV and radio.

Mass media channels, specifically televisions, are the most trusted source of information.

Data:

- Social media is the preferred method of accessing information. Some people attributed this preference to the convenience and accessibility of accessing it through the phone, compared to TVs and the radios.
- However, many participants cited that TVs are the most trusted source, because of the verifiable information they share, as opposed to rumors occasionally spread on social media platforms.

Analysis & Implication: Phones are convenient and easily accessible, which means that these are important tools to leverage and prioritize in the misinformation space. A phone-compatible version of the HIP tool will increase usage.

“Because I am mostly on my phone and I can easily access it.”
Female, 34yrs, Kajiado.

“Before something is posted it must be edited and so you trust it (TV). For social media, it might be rumors, there might be exaggeration and the likes.”
Female, 23yrs, Laikipia

“Mass media i.e., TV. because at least when they air something, it is something that you can see and they provide the evidence. But when you talk about things like twitter, it can be manipulated.”
Female, 23yrs, Nairobi

It is important for all the respondents that information is first verified before sharing.

Data:

- All the respondents emphasised the importance of verifying information first before sharing, in order to avoid misleading the public, spreading false information and rumors that might cause panic or fear.
- Most participants relied on their social network, like family, relatives, or friends to verify the information for them. On the other hand, some people indicated that they depend on their instincts to verify the content of the information themselves.

Analysis & Implication: Pair misinformation tools with education and awareness on effective ways of verifying information. This prevents users from flagging content based on their instincts or feelings.

"To avoid spreading rumors. Checking other Medias to see what they are saying about the something."
Male, 23yrs, Kiambu

"The degree of certainty is high when I do it on my own but when I am not certain on the subject I might consider doing so."
Male, 28yrs, Nairobi

"It is important because maybe you could be sharing information that is not true and you don't have any information about it and you can't even defend it. When you share something that is misleading, then it won't look good."
Male, 22yrs, Machakos

Besides HIP, some respondents were aware of other ways of reporting misinformation.

Data:

- Some respondents were aware of other ways of reporting misinformation including:
 - Calling the providers' contact number on the website to report the misinformation
 - Sending emails to the related sites, where the misinformation was found
 - Using the report button or icon of the social media platform
 - Reporting to the police
 - Reporting to Safaricom (though they could not explain this method)

Analysis & Implication: Reporting via social media was the most common because it was the only way people were aware they could report misinformation. This brings up, again, the need for creating more awareness on HIP.

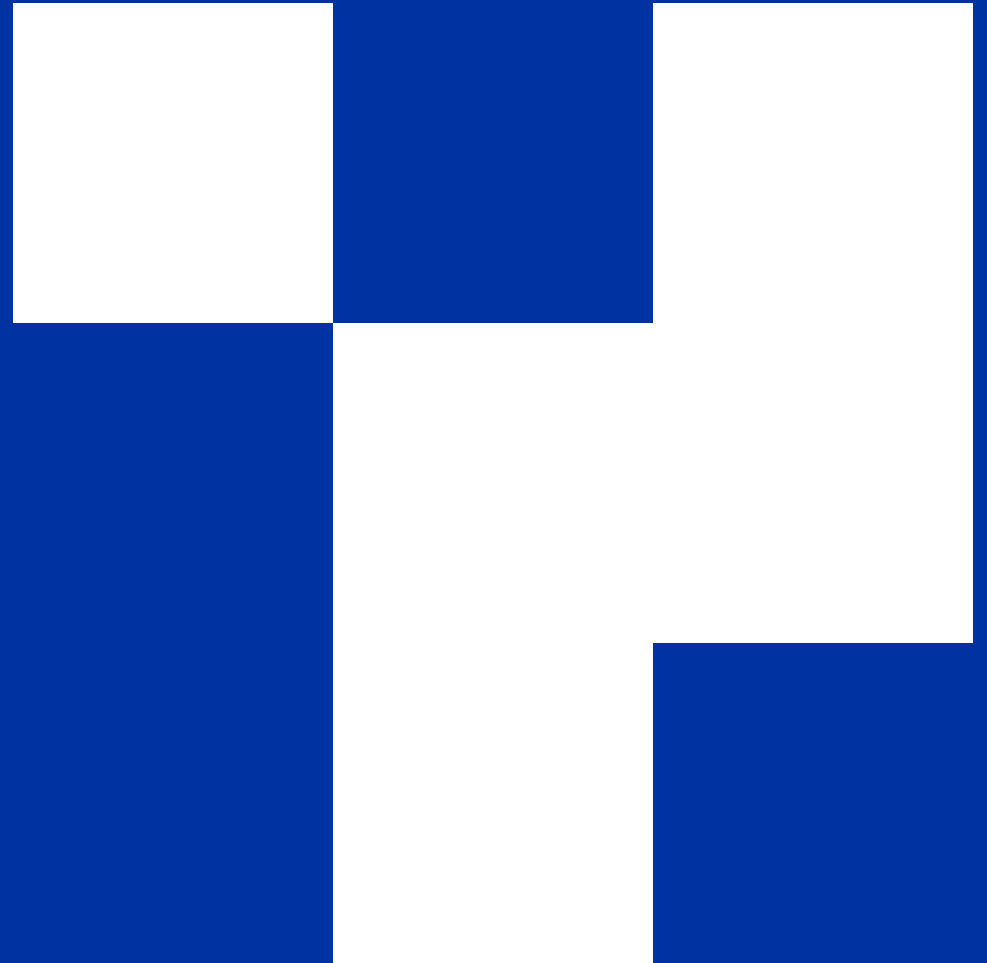
*“In the social media platforms there is a report button, that can be used to report misinformation.”
Male, 22yrs, Machakos*

*“With email you can forward the misinformation to the email account of the related sites.”
Female, 23yrs, Kiambu.*

*“For now maybe I use HIP but the problem is information passed by word of mouth cannot be reported in HIP.”
Male, 23yrs, Kirinyaga*

HIP User Experiences & Practices on HIP

Overview of general user behaviors and patterns on the HIP platform

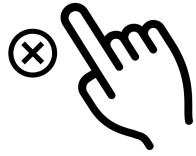


Summary of User Engagement and Experience with HIP

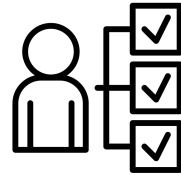
LEVERS



Improves the quality of information spread on the internet



Misleading content can be taken down, so less people see it.



In future posts, flagged sites will take caution not to share misinformation.

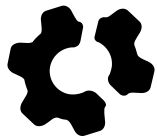


Easy to use: Simple interface with simple and clear language.



Users are completely anonymous.

BARRIERS



A few times the HIP tool fails to respond when a user clicks on it, therefore there are some delays.



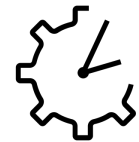
Lack of clarity on flag definitions: There is a bit of confusion on the flag definition is appropriate for certain types of content.



Lack of internet (bundles) and slow/ unstable internet.



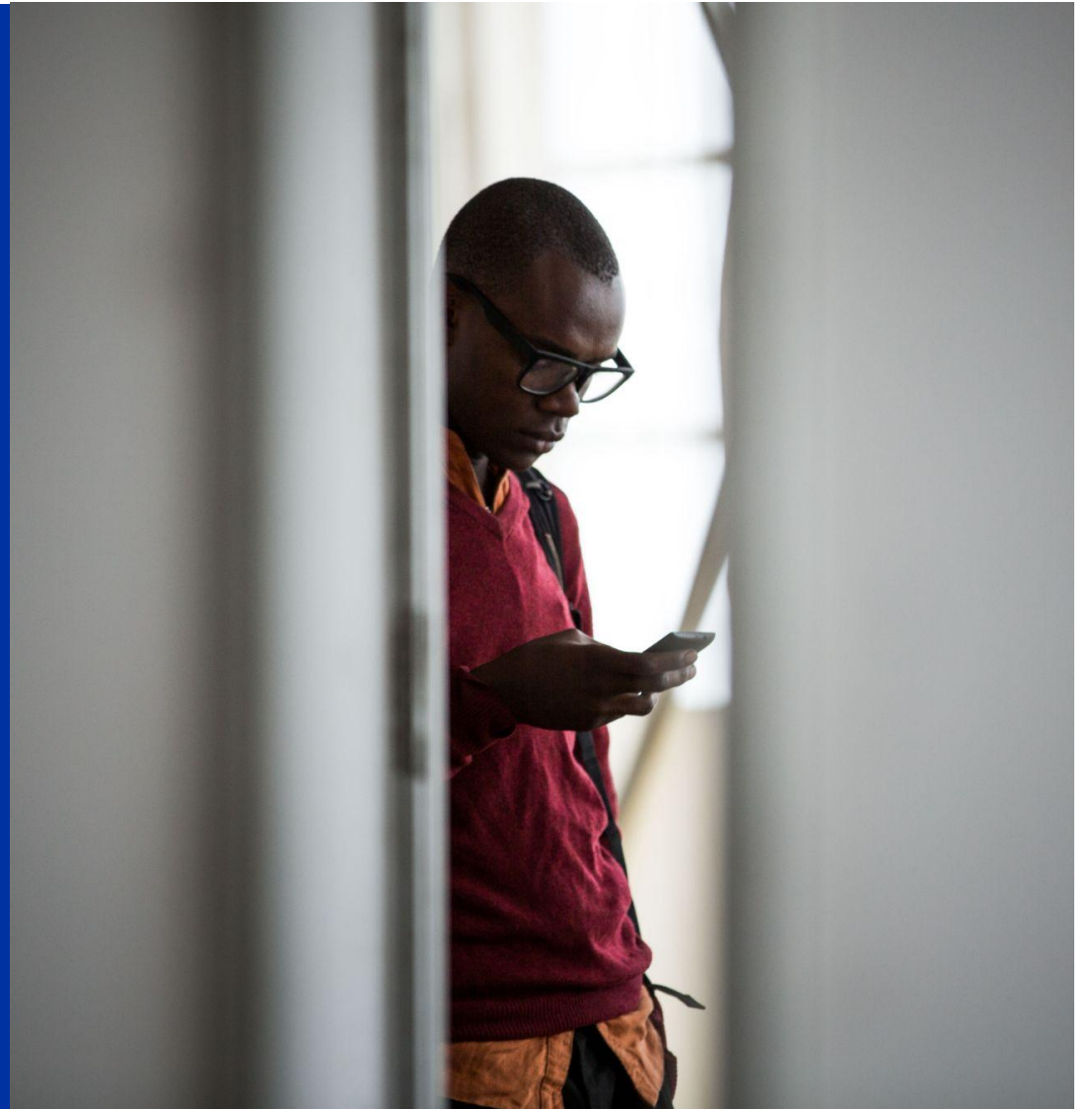
Limited browsers that can support the HIP tool.



What you flag is permanent, you cannot correct a mistake if you made one.

Overall Flagging Usage

Relationships between **flagging behavior** by demographic and non-demographic variables



People use HIP because they consider it an appropriate tool for stopping the spread of misinformation.

Data:

- Most of the respondents had used the HIP tool to flag misinformation. The main reason people use the HIP tool is because they believe it's a good and highly appropriate tool to: help indicate useful or important content; help people verify information and remove harmful content; and help stop the spread of misinformation.
- Other reasons that came from a few respondents is that it's very fast and easy to use, and they trust it.

Analysis & Implication: The study found that HIP is appropriate, easy to use and can change the quality of content online, and this confirms the need for such a tool, available to the general public.

"Because I thought that it was the most appropriate tool to use."

Male, 23yrs, Nairobi

"It is something I like to do so that I can help build an honest society and also to prevent misinformation and disinformation."

Male, 28yrs, Nairobi

"I trust it. So far have not had had any issues with it."

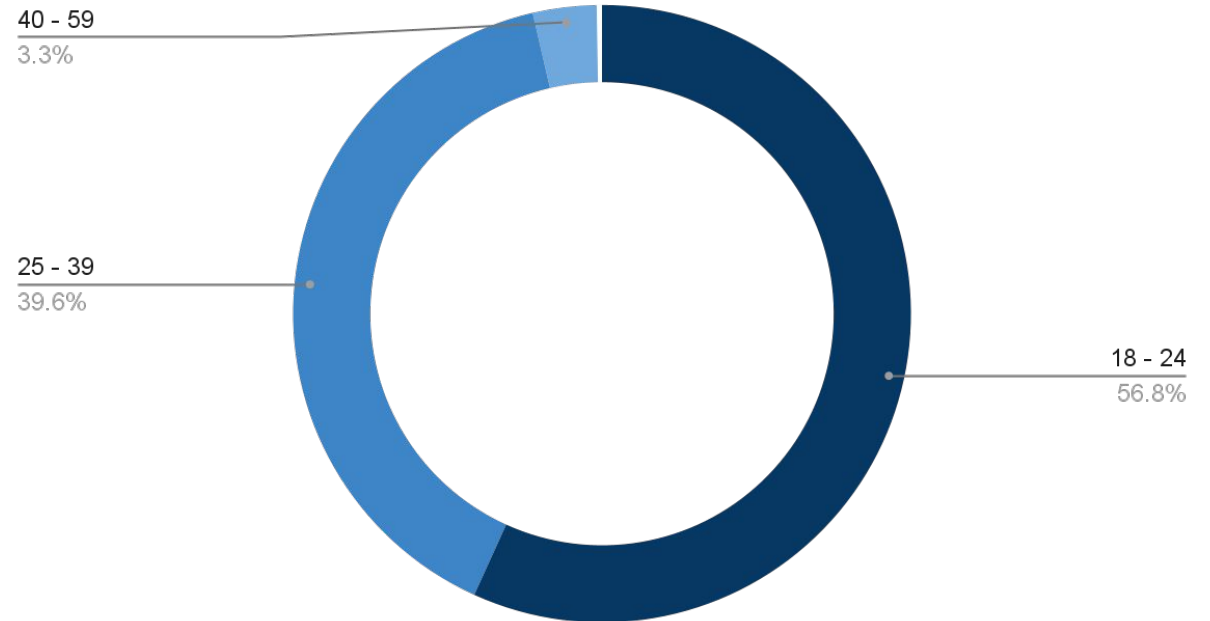
Male, 23yrs, Kajiado

Users between 18 to 24 years old were the most active on HIP; and accounted for $\approx 57\%$ of all flags.

The proportion of flags analyzed by age group:

- 18 - 24 years:
57% of flags (205)
42% of user population
- 25 - 39 years:
40% of flags (143)
52% of user population
- 40 - 49 years:
3% of flags (13)
5% of user population
- 50 - 60 years:
.3% of flags (1)
1% of user population

Total Number of Flags by Age Group



Female users registered a higher average number of flags per category than did male users

By **gender**, the **total** and the **average** numbers of flags per category were:

■ **Worthwhile Ideas**

Male: Total = 120 Average \approx 1.6
Female: Total = 166 Average \approx 3

■ **Lies or Manipulation**

Male: Total = 30 Average \approx .4
Female: Total = 9 Average \approx .2

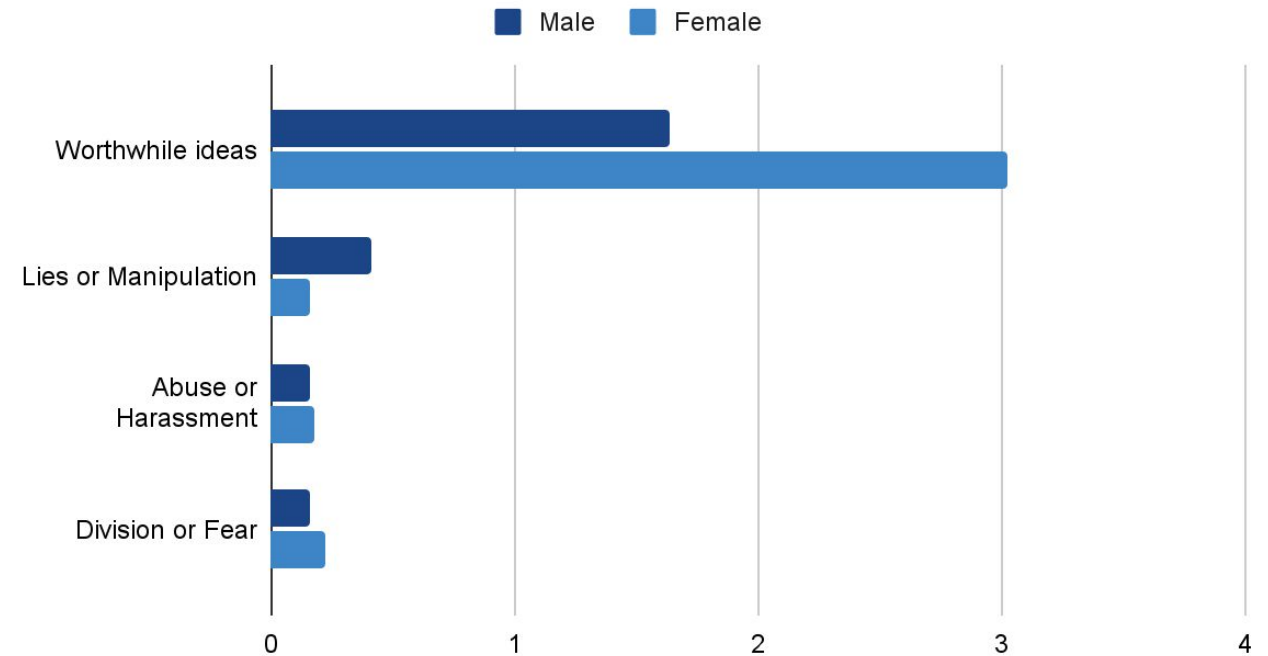
■ **Abuse or Harassment**

Male: Total = 12 Average \approx .2
Female: Total = 10 Average \approx .2

■ **Division or Fear**

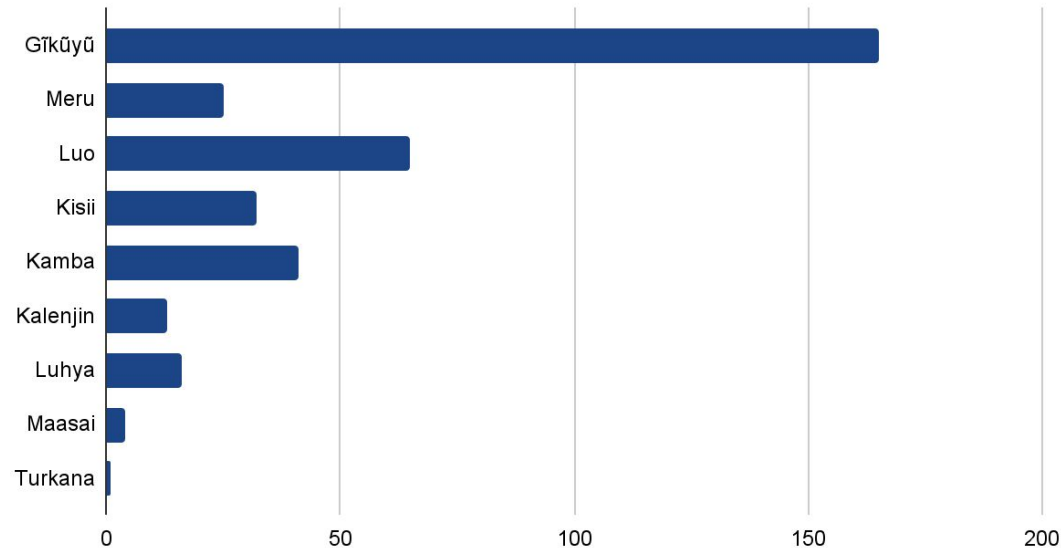
Male: Total = 12 Average \approx .2
Female: Total = 12 Average \approx .2

Average Number of Flags per Category by Gender



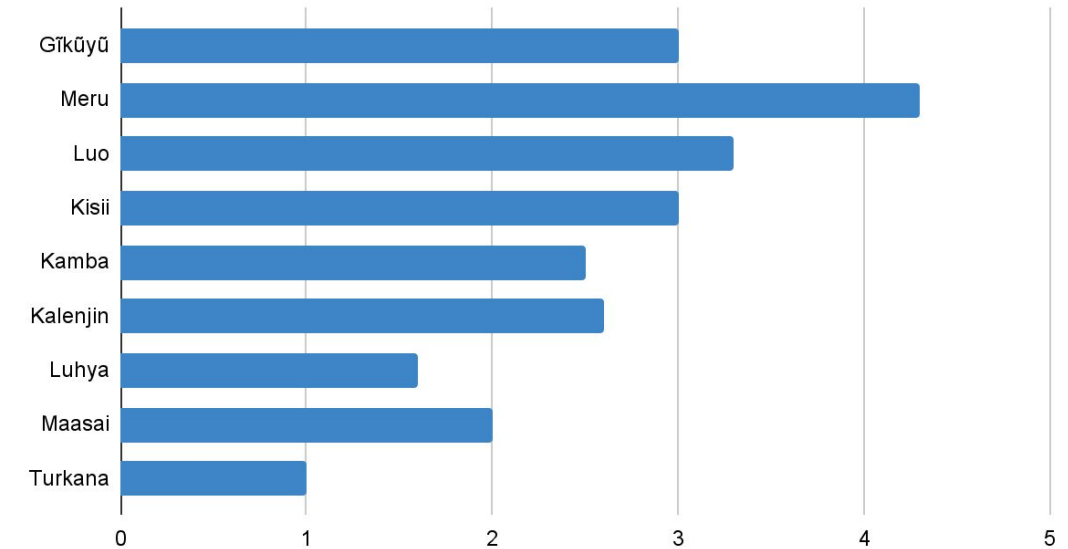
On average, more flags were registered by users from the Meru, Luo, Gĩkũyũ, and Kisii communities.

Total Number of Flags by Ethnic Community



- **Gĩkũyũ** (165 flags total per user), **Meru** (25 flags), **Luo** (65 flags), **Kisii** (32 flags), **Kamba** (41 flags), **Kalenjin** (13 flags), **Luhya** (16 flags), **Maasai** (4 flags), **Turkana** (1 flag)

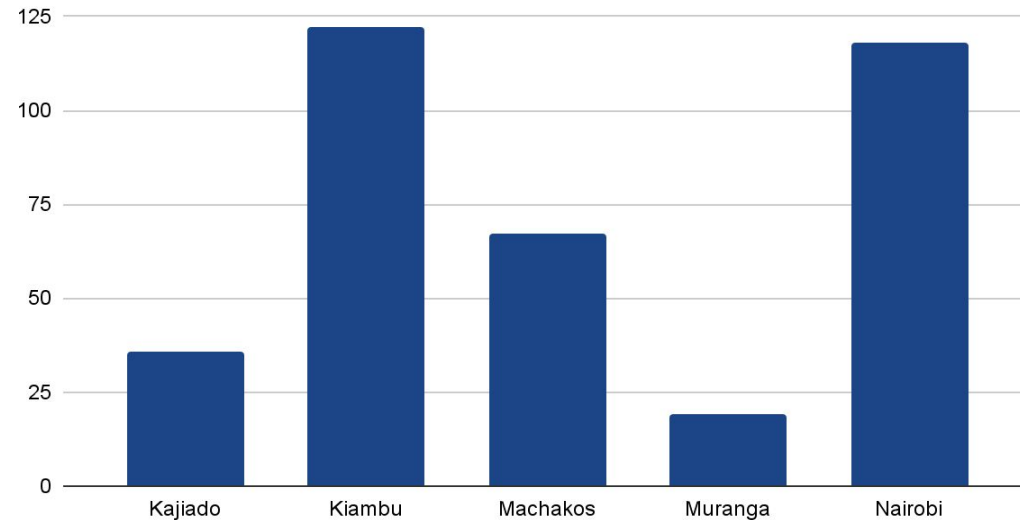
Average Number of flags by Ethnic Community



- **Gĩkũyũ** (3 flags average per user), **Meru** (4.3 flags), **Luo** (3.3 flags), **Kisii** (3 flags), **Kamba** (2.5), **Kalenjin** (2.6), **Luhya** (1.6), **Maasai** (2), **Turkana** (1)

On average, users from Kiambu were more active

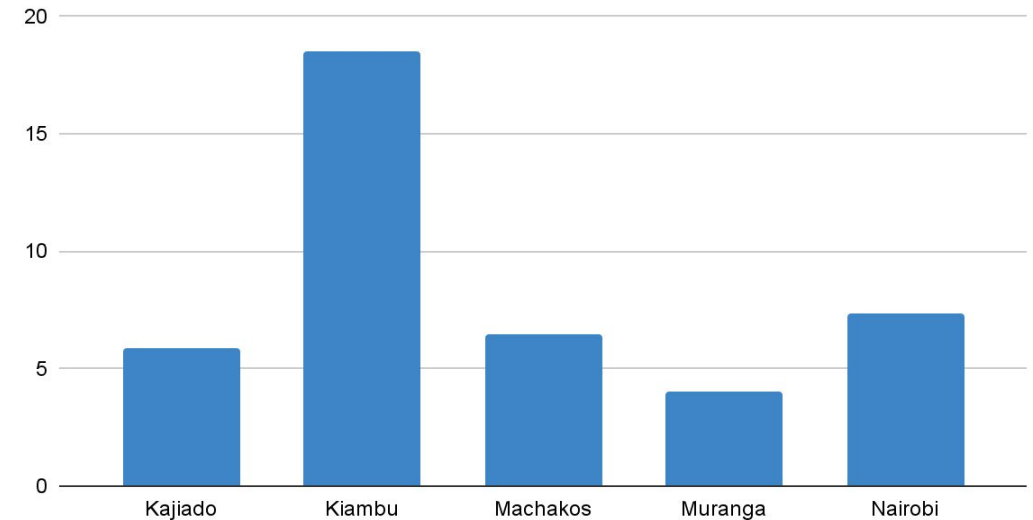
Total Number of Flags by Location



Total numbers of flags by location:

- Kiambu - **122** flags
- Nairobi - **118** flags
- Machakos - **67** flags
- Kajiado - **36** flags
- Muranga - **19** flags

Average Number of Flags by Location

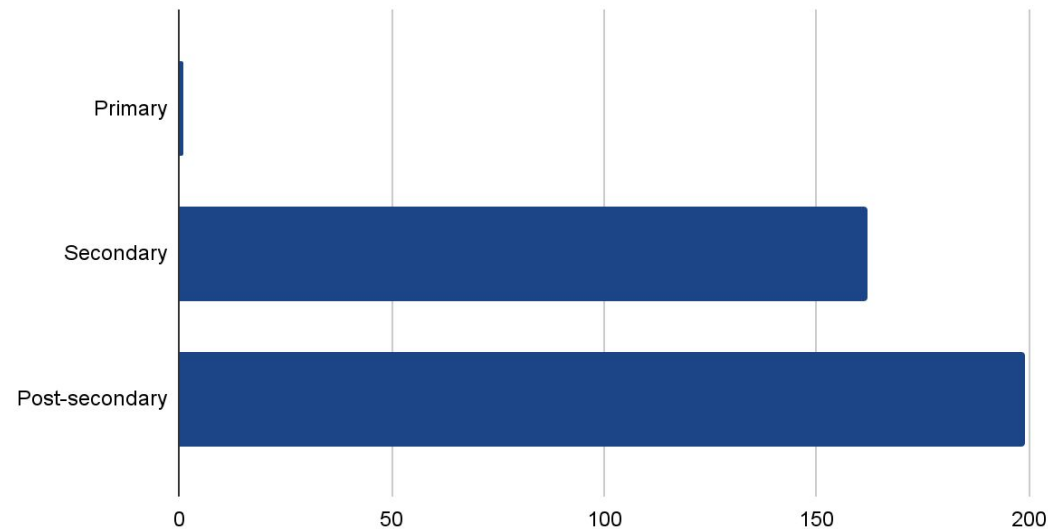


Average numbers of flags by location:

- Kiambu - **19** flags
- Nairobi - **7** flags
- Kajiado - **6** flags
- Machakos - **6** flags
- Muranga - **4** flags

On average, users with a secondary school level of education were more active.

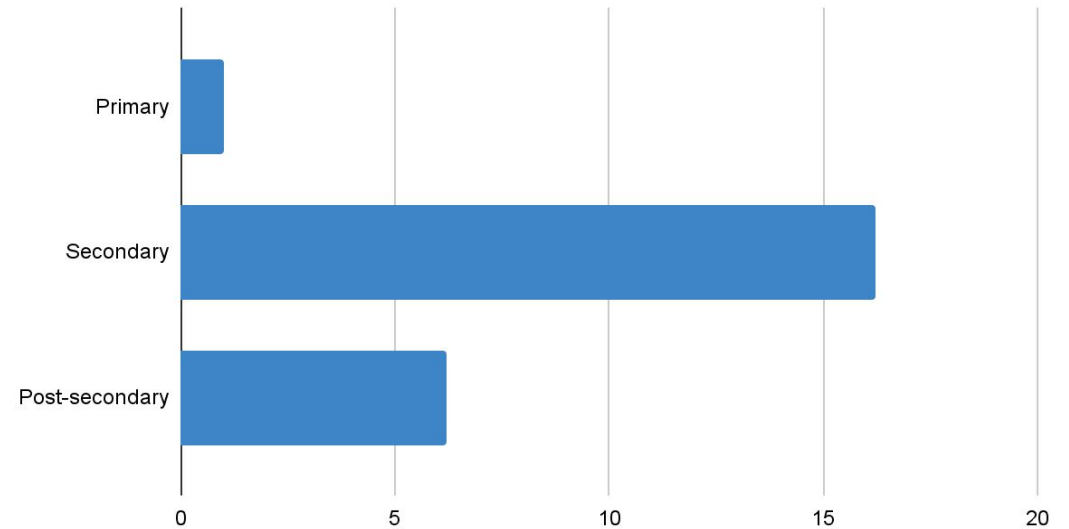
Total Number of Flags by Level of Education



Total numbers of flags by **level of education**:

- Post-secondary - **199** flags
- Secondary - **162** flags
- Primary - **1** flag

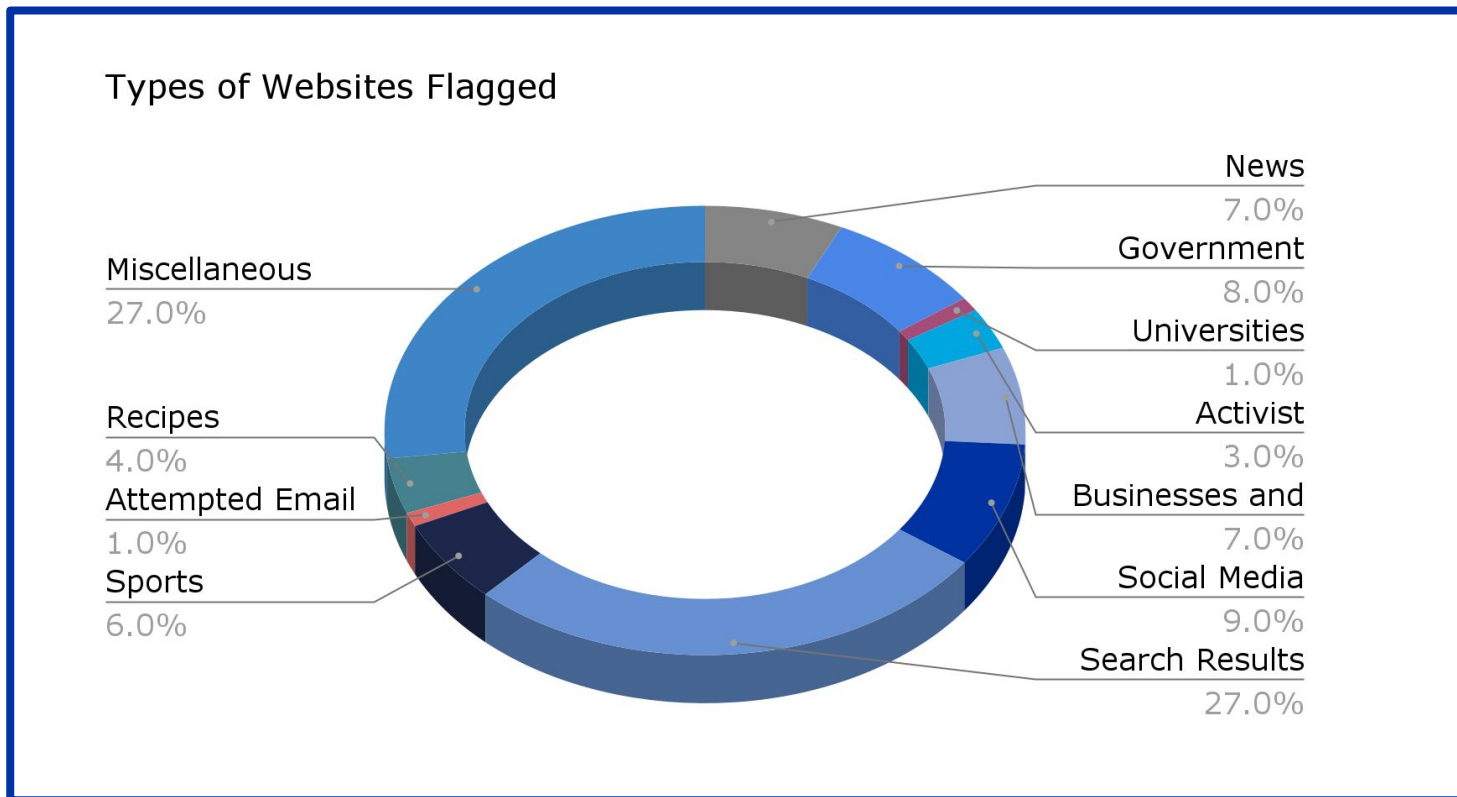
Average Number of Flags by Level of Education



Average numbers of flags by **level of education**:

- Secondary - **16** flags
- Post-secondary - **6** flags
- Primary - **1** flag

Besides search results, social media posts and articles were the most flagged type of content.



Legend:

- News
- Government Notifications
- Universities
- Activist Organizations
- Businesses and Services
- Social Media Posts
- Search Results
- Sports
- Attempted Email
- Recipes
- Miscellaneous

→ Themes included in **Miscellaneous**: health, art, self-help, natural science, events, and finance.

Academic and sports are the most flagged type of articles.

Data:

- Most people flag content in articles such as academic articles, sports articles and articles on politics.
- Many people flag content in social media (i.e., Facebook, Youtube, LinkedIn) and blogs, while a few people flag journals, online newspapers, and general websites on health, money, jobs, and scholarships.

Analysis & Implication: Despite the risk people attached to flagging political content, the study findings show that a good percentage of users still flag them. This shows that their social responsibility outweighs their personal safety.

“Articles and blogs because most of the time they are the writer’s opinions.”
Female, 23yrs, Kiambu

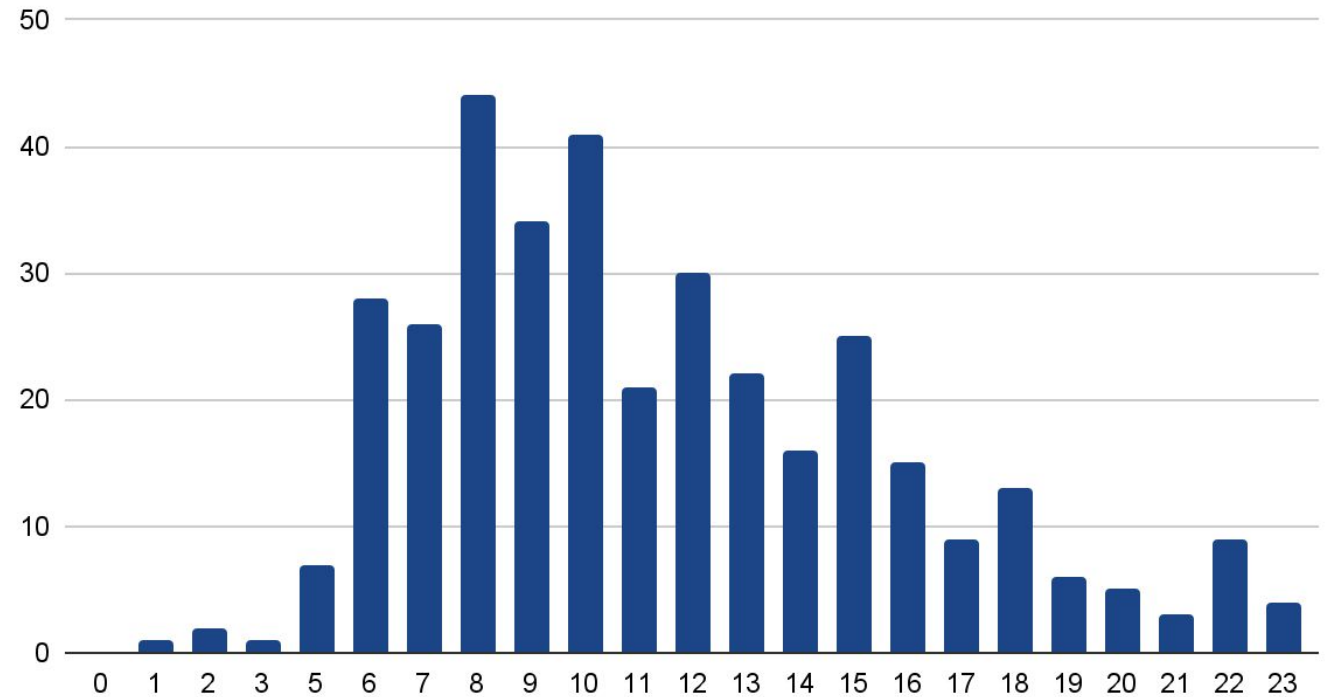
“Football articles and website, newsletters and some other apps like FlashScore* [which shares live football scores].”
Male, 25yrs, Nairobi

“Mostly on politics and from articles. This is because people mostly concentrate on politics.”
Male, 22yrs, Machakos

Mornings (7-11am) registered a relatively higher number of flag count

- Hour of the day with the **highest** number of flags is **8:00am**.
- Hour of the day with **lowest** number of flags is **3:00am**.

Count of Flags by Hour of the Day



The tool does not require much time so it is used by respondents at any time during the day.

Data:

- The tool is not too involving, so it doesn't take much time, therefore, many people use it for less than an hour in a day. One person who says it takes time is because they included the time involved in reading what they want to flag.
- Some respondents mentioned that they use the tool anytime, and so have no time preference since it does not take much time to use the HIP tool.

Analysis & Implication: A phone version of HIP can increase usage as people can make use of their other free times like during commutes, and meal times.

“Anytime, it’s something that you pop in and send, it doesn’t even take 20 seconds.”

Male, 22yrs, Machakos

“Mid morning and afternoon between 10am to 2pm, that’s the time when I’m more active in the internet.”

Female, 23yrs, Kiambu

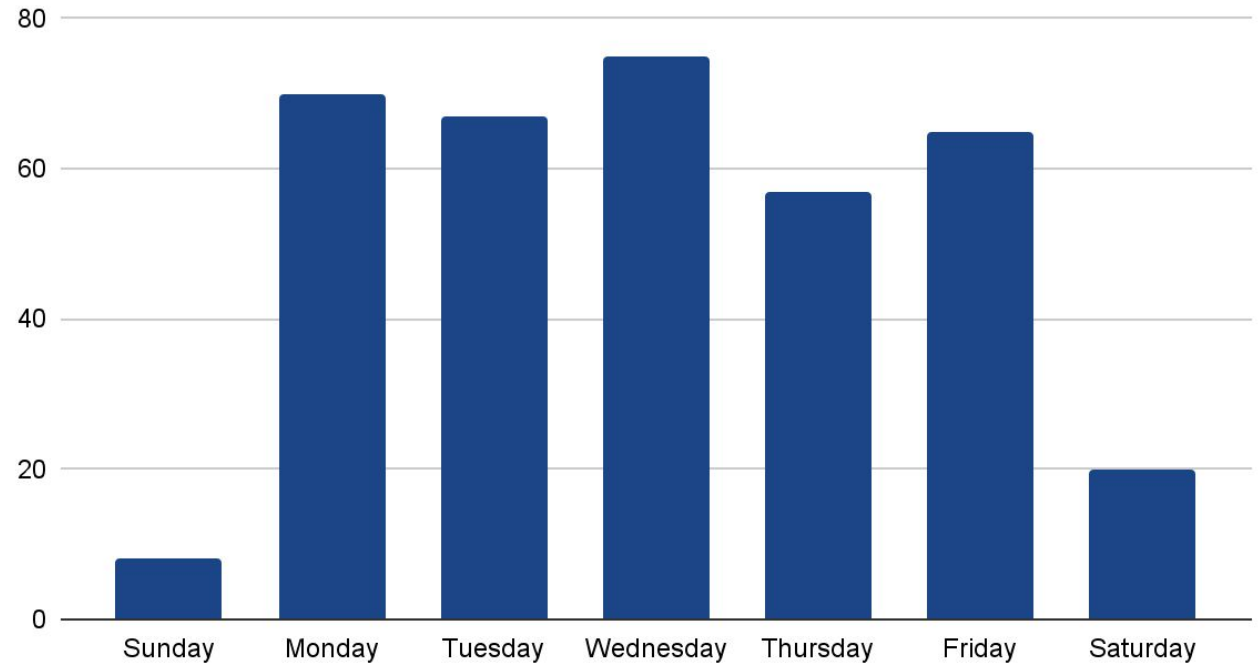
“I use it whenever I feel I need to use it.”

Male, 23yrs, Siaya

Weekends registered relatively lower numbers of flags.

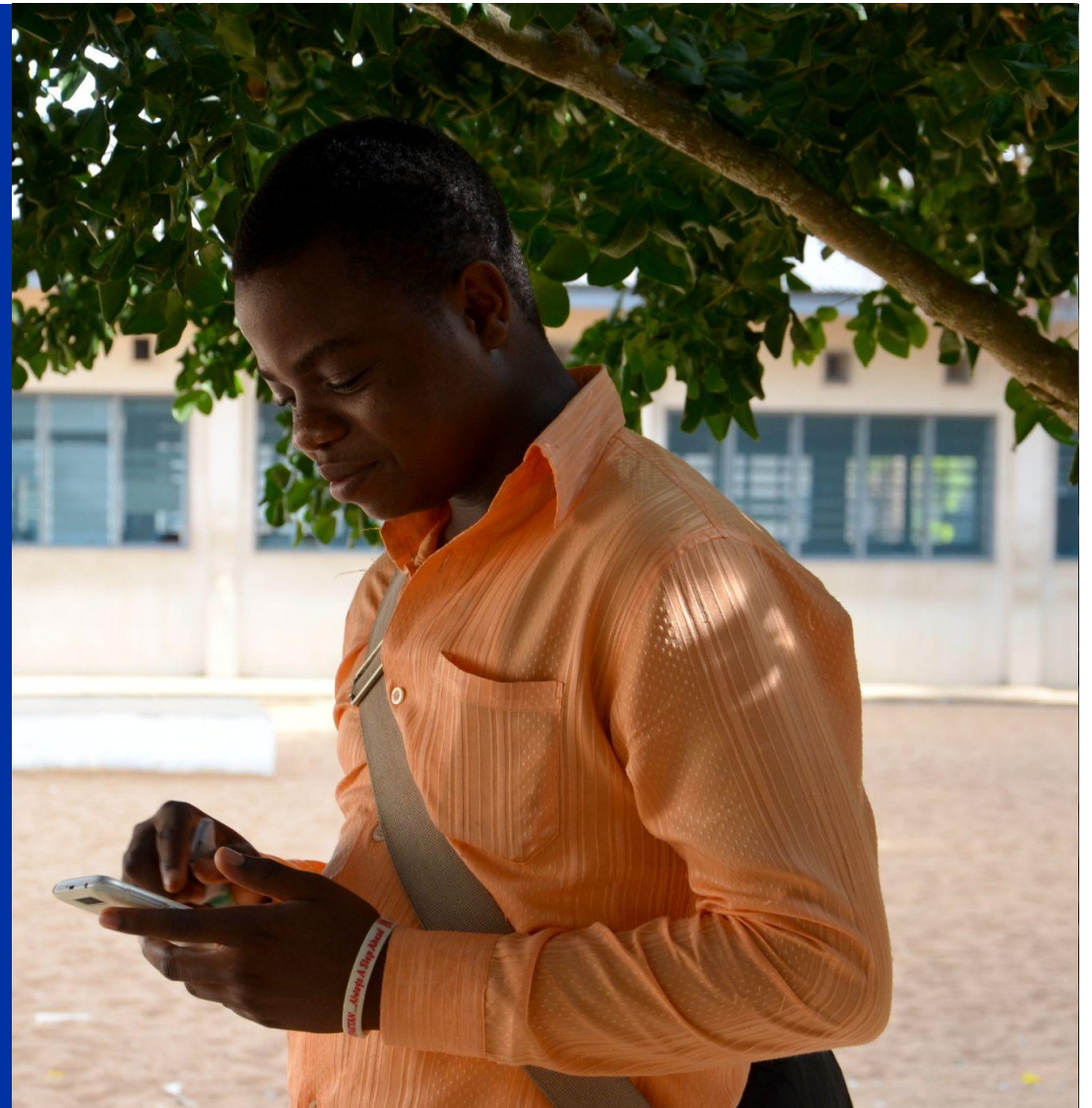
- Day of the week with **highest** number of flags is **Wednesday**.
- Day of the week with **lowest** number of flags is **Sunday**.

Total Number of Flags



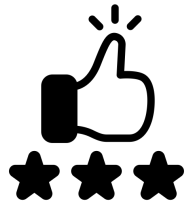
Usage by Flag Definitions

Assessing usage by flag definitions



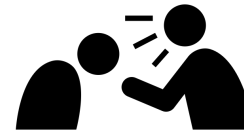
Overview of Usage by Flag Definition and the Demographic Breakdown

Worthwhile Ideas



- Total number of flags: **268**
- Median rank on the severity scale: Minor
- Female: **58%**
- Post secondary education: 54%
- 18 - 24 age group: **59%**

Abuse or Harassment



- Total number of flags: 22
- Median rank on the severity scale: Minor
- Male: **55%**
- Post secondary education: 41%
- 18 - 24 age group: **55%**

Lies or Manipulation



- Total number of flags: 39
- Median rank on the severity scale: **Severe**
- Male: **77%**
- Post secondary education: **74%**
- 18 - 24 age group: 49%
- 25 - 39 age group: 49%

Division or Fear



- Total number of flags: 24
- Median rank on the severity scale: Minor
- Female: 50%
- Post secondary education: 46%
- 25 - 39 age group: **63%**

Despite their intent to stop the spread of misinformation, most people use the tool to flag worthwhile content.

Data:

- Most people use the HIP tool to flag worthwhile content, while many people use it to flag both worthwhile and harmful content. Some people use it to flag misinformation or harmful content that they find online.
- One person only flagged worthwhile content because they described themselves as positive and thus they only spread information that can help someone. Another reason from another user is wanting to contribute to the society.

Analysis & Implication: As mentioned earlier, having the option to flag worthwhile content reduces the quantity of flags related to harmful content.

“It cannot lead to cyber bullying and the likes. I have mostly used it to spread worthwhile ideas.”
Female, 23yrs, Nairobi

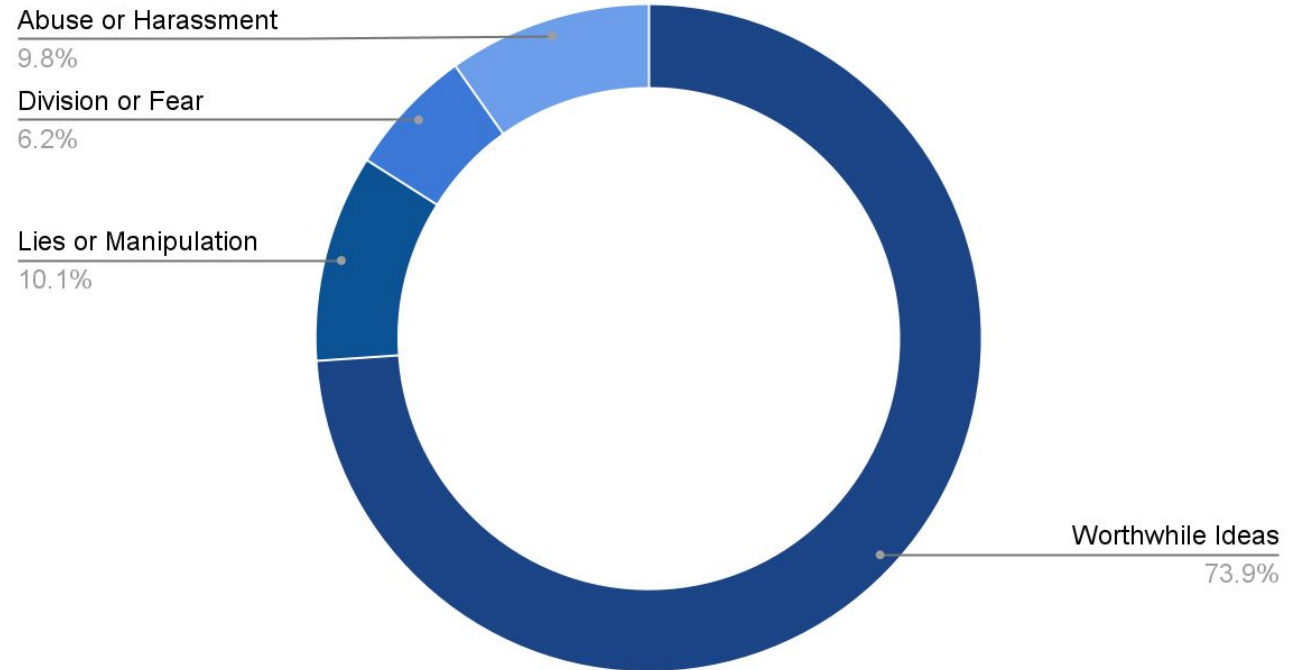
“I used it to share worthwhile ideas. The article was about how we can mitigate climate change, corals and I was just giving my idea and opinions. Adding to what they said.”
Female, 23yrs, Nairobi

“I am a positive person so I only flag the positive ones that can help someone.”
Female, 34yrs, Kiambu

A large majority of flags were in the “Worthwhile Ideas” category.

- Worthwhile Ideas: **73.9%**
- Lies or Manipulation: **10.1%**
- Abuse or Harassment: **9.8%**
- Division or Fear: **6.2%**

Proportion of Flags



More people than expected do more research on the validity of worthwhile content, than for harmful content.

Data:

- Most people use their personal judgement, often a hunch but sometimes through research, to decide whether the information they are flagging is worthwhile content. They do not receive any help in making the decision to flag because they believe in their own judgement or don't believe someone else will give them any new information that they do not already have.
- On the other hand, to verify the information, some people ask their social network like family, friends and peers; but mostly from friends. Many people prefer asking friends because they rely on them for guidance in areas or topics that they are not familiar with and they are also comfortable with them since they have known them for a long time.

“When I go through articles I look at the source, like if I get from New York Times or BBC. I know it's a reliable source because they are reputable media. One thing I look at is the publisher and the content of the article to see whether the writer has cited renowned people who are experts in a particular field.”
Female, 30yrs, Kiambu

“There is what you expect when researching on something so I used that judgement and no one influenced my decision I was just flagging the worthwhile ideas.”
Male, 22yrs,
Machakos

More people than expected do more research on the validity of worthwhile content, than for harmful content.

Data:

- Some people determine that the information is worthwhile if it is useful for them for what they are doing.
- While some people do their own research and check for reliable and credible sources, citations, credible publishers, known media houses (i.e., NTV and KTN) and trusted websites.
- Some check on feedback or ratings given on the site by other people. A few people check whether evidence has been attached (i.e., videos, photos).

Analysis & Implication: This explains why users provide more information when it comes to worthwhile content, making it easier to verify the quality of such flags.

"By checking on the content of the information. I also look at the source of the information are they credible. And also the different citations. Also checking on the feedback given on the same content."
Male, 28yrs, Nairobi

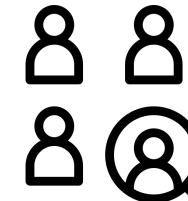
"If it helps me with whatever I wanted like if I was searching for something and I get the correct information I consider that as worthwhile content."
Female, 23yrs, Kiambu

Most people rely on their personal judgement to verify the quality of harmful content.

Data:

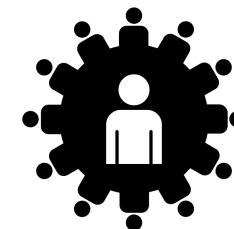
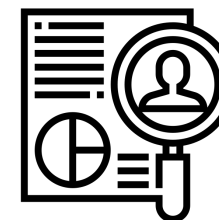
- Most people either use their personal judgement to determine whether the information is harmful content or check to see whether it would be harmful to them or others in the society.
- Some people check whether the content is targeting only certain groups of people and others use other sources or research to verify the information. Some people rely on their social network e.g community norms, peers, to help them verify whether the content is harmful. A few people review the feedback on the sites and check the sources.

Analysis & Implication: In terms of identifying harmful content, there is a huge lack of objectivity which reduces the quality of flags.



“The fact that I’m learned helps me know what is harmful and what is not so i just use my personal opinion.”

Female, 23yrs, Kiambu

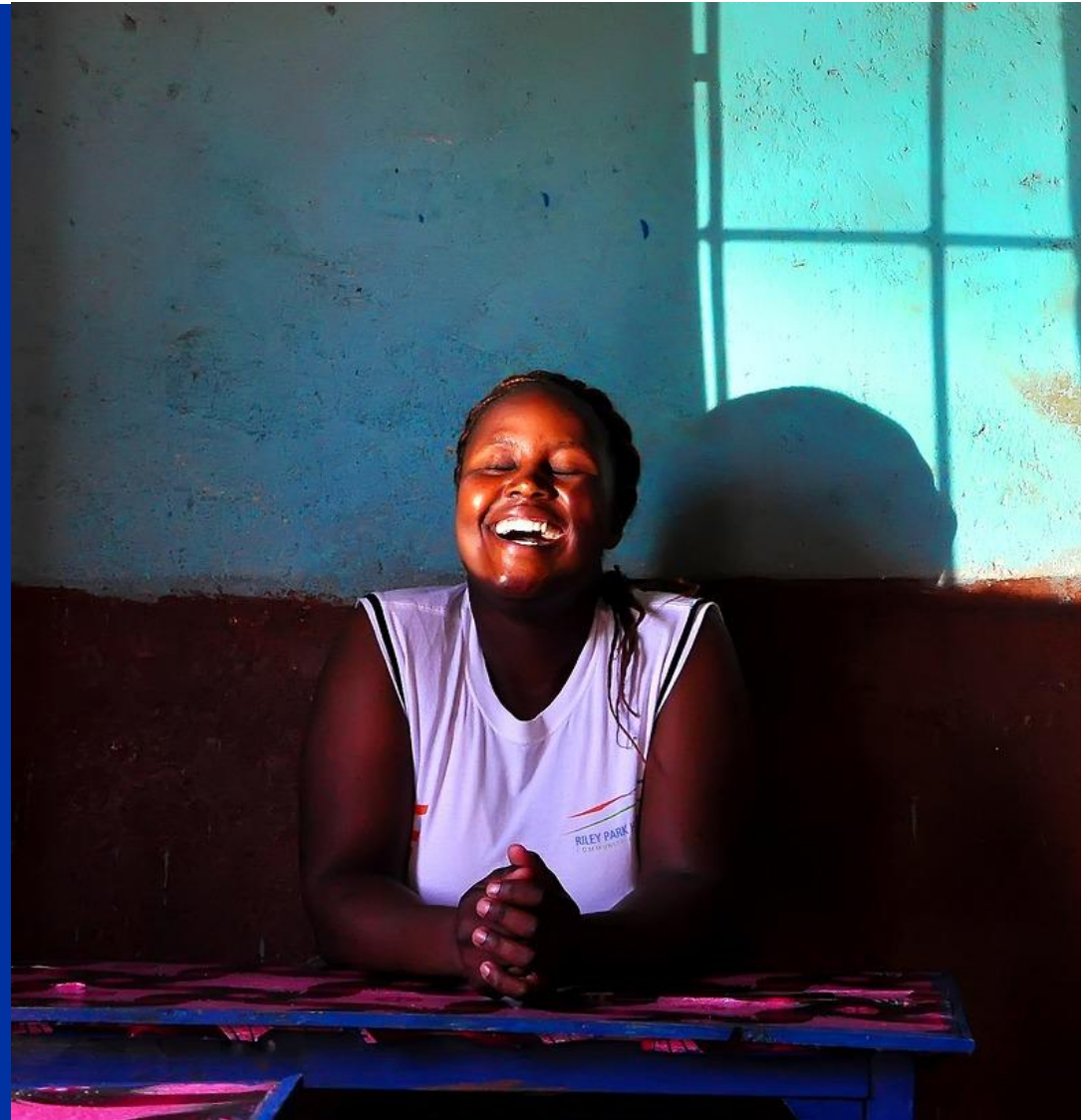


“Discuss with my friends and if most of them see the information as harmful then it is harmful.”

Male, 24yrs, Mombasa

Analysis by Flagging Severity

Trends in flagging analyzed by flag categories -- organized into severity ranks on a set severity scale running from “**Minor**” to “**Medium**” to “**Brilliant / Severe**”



People rely mostly on their personal judgement when it comes to attaching a severity level to a flag.

Data:

- Most people choose the severity level based on their own judgement or intuition; they read the content and check how it affects them or other people.
- For example, rape would severely affect them so it gets a severe rating while content on sports might not be severe.

Analysis & Implication: Users are not entirely sure how to use the severity rating, it could be based on how the content made them feel or whether it targets certain groups.

This feature needs a bit more clarity; more information can be provided so that users are on the same page about how to use it.

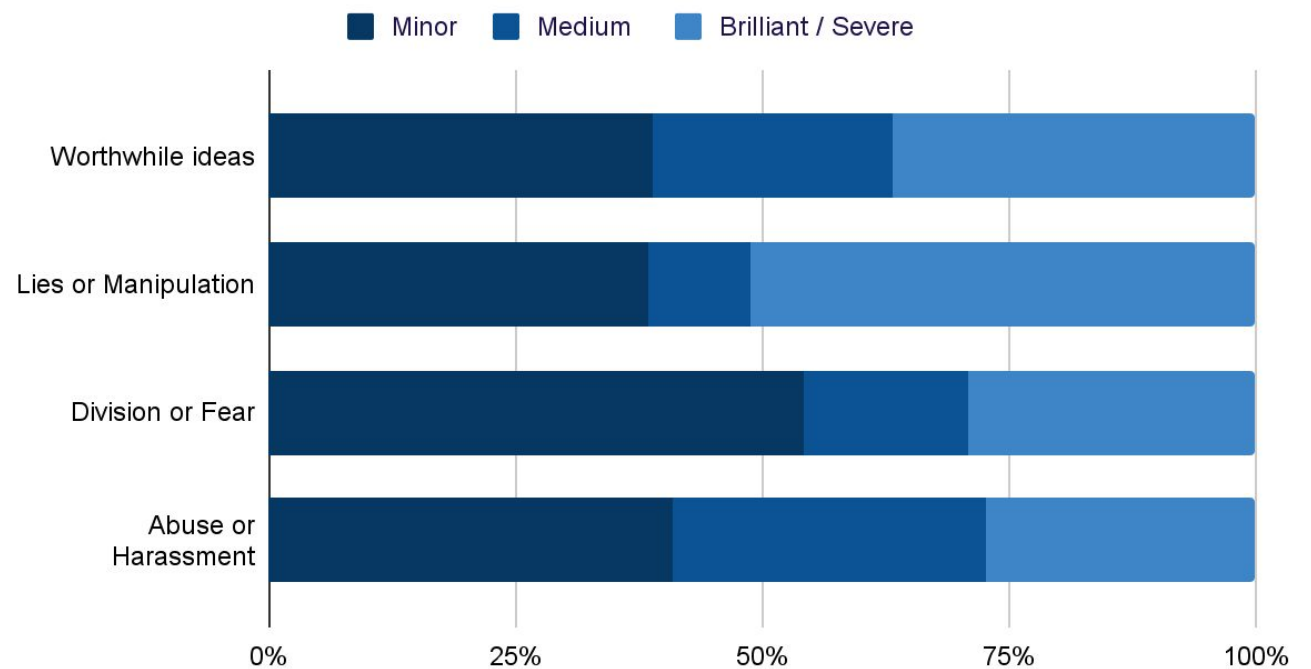
“You just make a judgement. For example if a girl was raped somewhere and someone is saying that it was their fault or they are making fun of that, it is severe. Then if it is just a joke, that one is not severe.”
Female, 23yrs, Laikipia

“Severity level varies from one person to another so what is misleading to me might not to another person. I just use my intuition.”
Female, 23yrs,
Kiambu

Severity levels differed by category, with Lies and Manipulation tending towards most severe rankings.

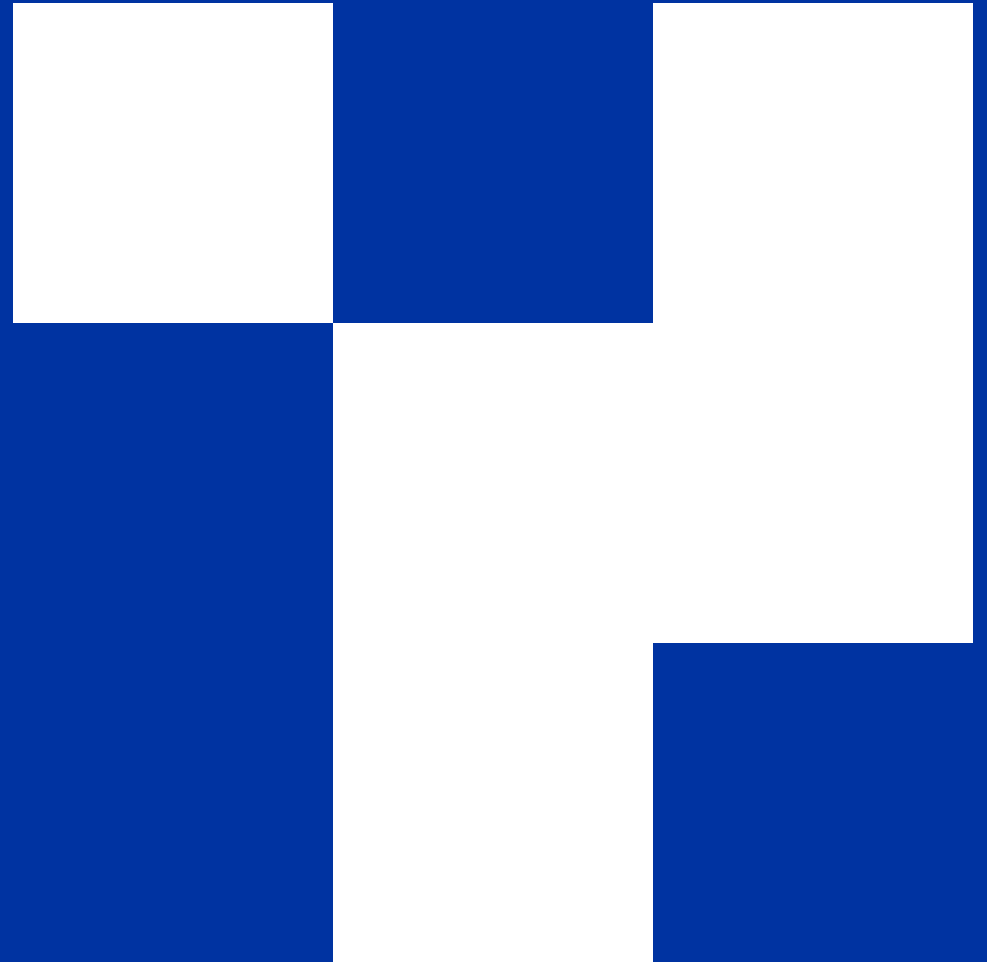
- **Division or Fear** received the **most 'minor'** rankings on the severity scale.
- **Abuse or Harassment** received the **most 'medium'** rankings on the severity scale.
- **Lies or Manipulation** received the **most 'severe'** rankings on the severity scale.
- **Abuse and Harassment** received the **least 'severe'** rankings on the severity scale.

Minor, Medium, and Brilliant / Severe

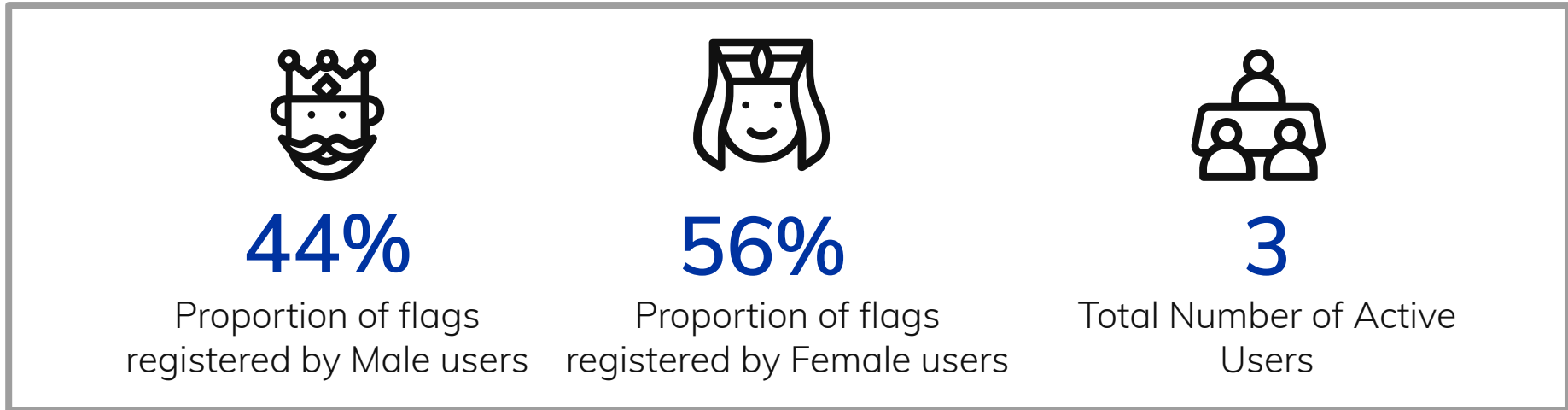
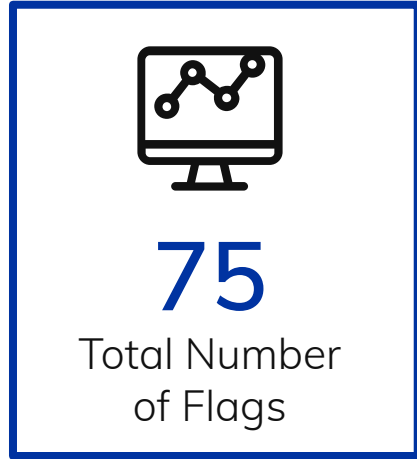


Segmented User Insights

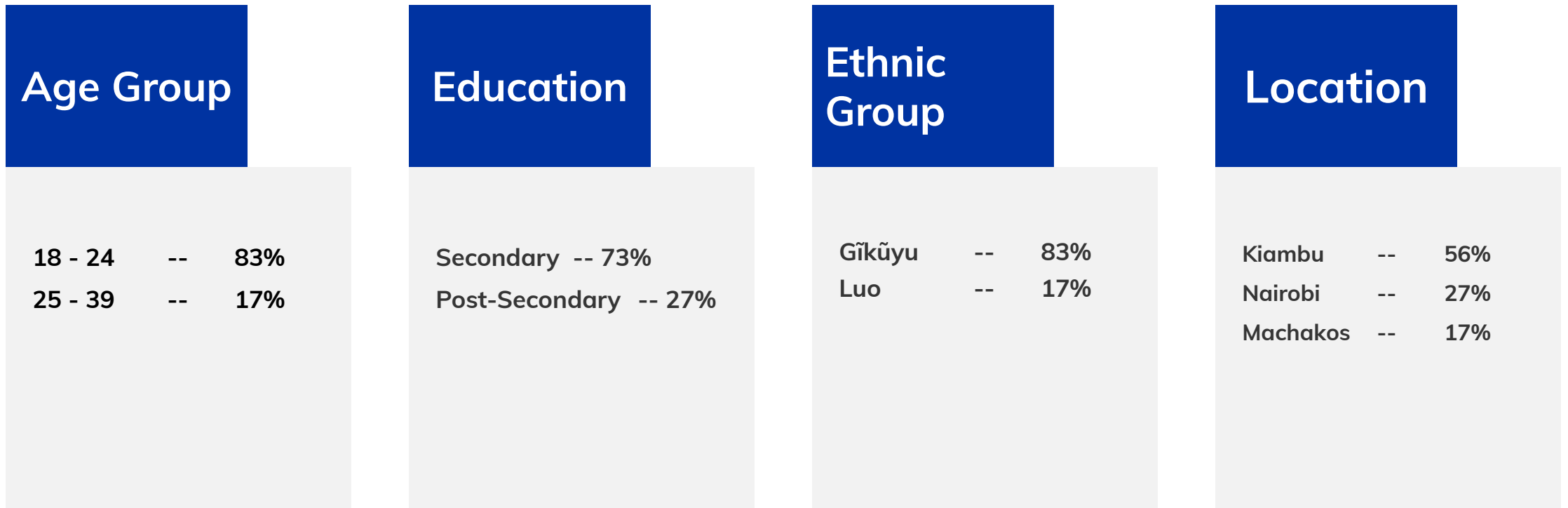
Analysis of user segmentation, based on high, mid and low flagging users



Only 3 people were considered Active Users (Flags > 15)



Demographic summary for the 75 flags registered by Active Users



16 users were considered Moderate Users ($5 < \text{Flags} < 15$)



130

Total Number
of Flags



29%

Proportion of flags
registered by Male users



71%

Proportion of flags
registered by Female users



16

Total Number of Active
Users

Demographic summary for the 130 flags registered by Middle Users

Age Group

18 - 24	--	60%
25 - 39	--	34%
40 - 49	--	6%

Education

Secondary	--	37%
Post - Secondary	--	63%

Ethnic Group

Gĩkũyu	--	29%
Kisii	--	18%
Kamba	--	16%
Luo	--	16%
Meru	--	16%
Kalenjin	--	5%

Location

Nairobi	--	39%
Kiambu	--	26%
Kajiado	--	10%
Machakos	--	19%
Muranga	--	6%

A majority of users were considered Low Users (Flags < 5)



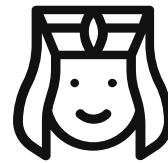
157

Total Number
of Flags



62%

Proportion of flags
registered by Male users



38%

Proportion of flags
registered by Female users



109

Total Number of Active
Users

Demographic Summary for the 109 flags registered by Low Users

Age Group

18 - 24	--	42%
25 - 39	--	55%
40 - 59	--	3%
50 - 60	--	< 1%

Education

Secondary	--	37%
Post-Secondary	--	62%
Primary	--	≈ 1%

Ethnic Group

Gĩkũyu	--	41%
Luo	--	20%
Kamba	--	13%
Luhya	--	10%
Kisii	--	6%
Kalenjin	--	5%
Maasai	--	3%
Meru	--	3%
Turkana	--	≈ 1%

Location

Kiambu	--	30%
Nairobi	--	30%
Machakos	--	19%
Kajiado	--	15%
Muranga	--	6%

There are varied reasons for people not using the HIP Platform.

Data:

- 75% of the people stated that they rarely use the HIP tool, that is, once a week to more than once in a month. The main reasons for not using the HIP tool include: rare usage of the internet (i.e., lack of bundles), not having a computer, not having time to flag, not coming across any harmful content, and the tool itself is not salient to them.
- The behavioural reasons include not wanting or liking to report other people and not being able to complete reporting after starting the process. One respondent thought that “people are reckless” so there is no point in reporting (i.e., misinformation will exist anyway).

“Yes, for worthwhile ideas, but for reporting no because nothing has come my way yet.”
Female, 34yrs, Kajiado

“Unless I become a blogger or someone that uses the internet all the time. If it can become like a job, then I can be using it severally.”
Male, 25yrs, Nairobi

“No, because I don’t like reporting things but when there is a lot of rumours being spread and I have the correct facts i would be motivated to report.”
Male, 22yrs, Nairobi

Internet challenges and not frequently coming across harmful content partly explains the low usage of HIP.

Data:

- Rare usage was attributed to internet issues, no option to report accounts, not being able to use it on the phone, forgetfulness, and they don't use the platform everytime they are online since it's not all the time that something needs to be flagged.
- For the available data on moderate users, there were a few varied reasons for why they keep using HIP, including a sense of responsibility, to ensure healthy internet usage, novelty of the tool, effectiveness of the tool, and the anonymity aspect.

Analysis & Implication: There is a likelihood that HIP will be used more by middle-upper class citizens since the low income users always think about the financial costs.

“Not every time that I’m online but only when it’s necessary that’s when I get to use it.”
Female, 28, Machakos

“Because we tend to flag things that are negative and they are not many.”
Male, 23yrs, Nairobi

“I spend most of my time on the phone and not on the laptop and when I am using the laptop most of the time it’s when I need to do something important.”
Female, 34yrs, Kajiado

User Accuracy

Assessment of user accuracy
in flagging misinformation from
flag sample validation exercise
by PesaCheck



PesaCheck found that respondents tend to use emotional reactions when flagging, as opposed to objectivity.

PesaCheck analyzed the validity of claims associated with flags from a sample of 15 flags (4%) that were registered by users for the study.

1. **Claims** made about flags that were in the **Worthwhile Ideas** category, were on average **more relevant and more fleshed out** than claims made in any of the other 3 flag categories: Lies or Manipulation, abuse or harassment, and Division or Fear.
2. Many of the claims from the sample that were associated with **Division or Fear** flags were made about websites that featured **political subject matter**.
3. PesaCheck found that claims associated with the relevant flags in the sample derive more users' **emotional reactions** to website content rather than **empirical disagreement** with the content itself.
4. For some claims, aggregation at the **website level** (as opposed to the individual flag level) would have made it difficult to identify **what aspects of each claim** to fact-check.

Respondents believe that they flag accurately, despite the finding from PesaCheck.

Data:

- Most people generally understand the flagging definitions and a large majority believe that the flagging definitions are easy to understand.
- Most people believe that they accurately flag based on the flag definitions (i.e., Worthwhile, Lies and Manipulation, Abuse and Harassment, Division or Fear). On a scale of 1 to 5, 45% rate themselves at a 4 while 39% rate themselves at a 5, when it comes to accuracy.

Analysis & Implication: It is expected that people believe that they flag accurately, since they mostly rely on their personal judgement. It is important to introduce more objectivity in flagging so as to improve the quality of flags.

"It is easy to understand if that person is literate. A primary person might not understand that but a highschool student might partially understand it and maybe a university student will fully understand it."

Female, 23yrs, Laikipia

"It's simple though at times I read the article and come to the options but I get stuck on where to place it. I have the opinion in my mind but I do not know what to call it. Like I would read something on Covid but I do not know if it's manipulation or spread of fear."

Female, 30yrs, Kiambu

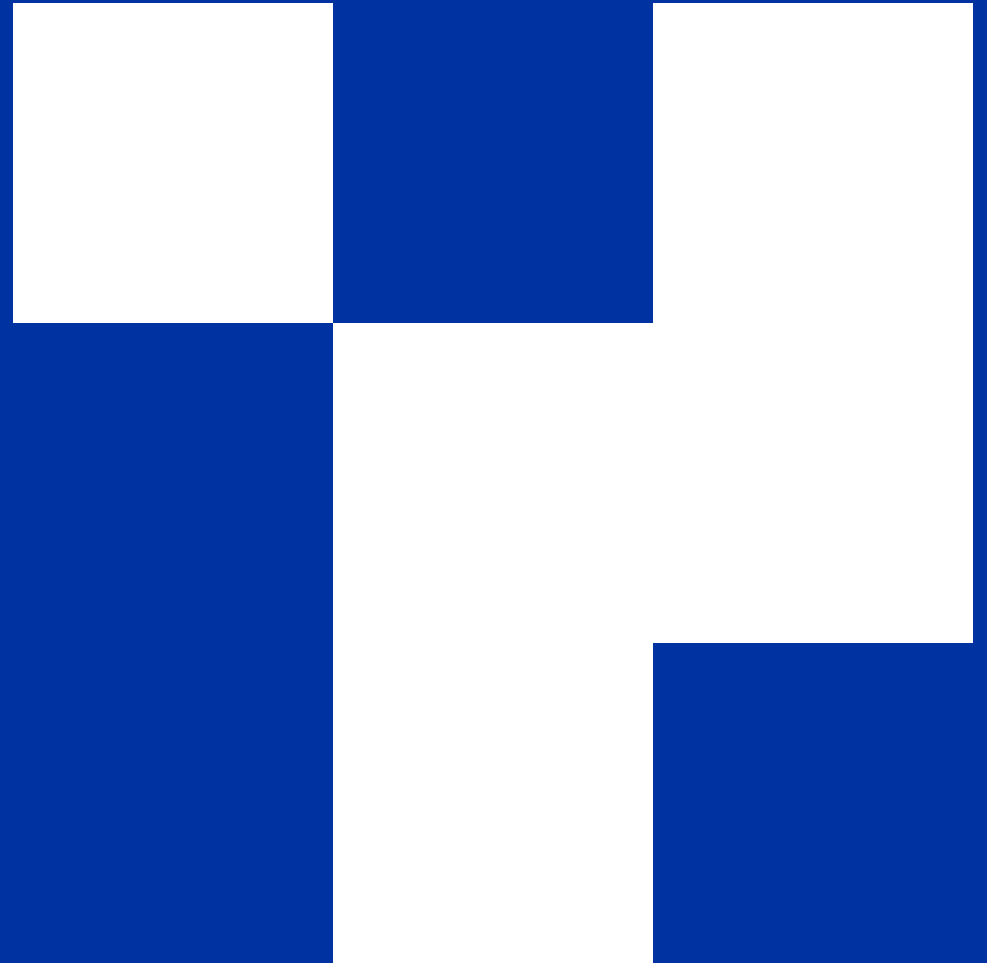
PesaCheck recommends an improvement on the platform design to improve the quality of flags.

Most of PesaCheck's recommendations were geared towards improving the platform design, with a view to making it possible for fact-checks to be conducted on claims associated with individual flags as registered by users.

1. Consider **adding the following fields** to the flagging platform, to **enable efficient fact-checking of claims** associated with **individual flags**:
 - a. A **text field per flag** that requires users to **specify which item** of website content that they would like to flag.
 - b. A **text field per flag** that requires users to **specify exactly why** they are flagging the specific item of website content that they would like to flag.

User Motivations and Perceptions of HIP

This section details what motivates a user to keep flagging content online, what they hope to gain from flagging content, and any shortcomings they believe crowdsourcing could have.



A majority of the participants have had misinformation shared with them.

Data:

- All of the respondents understood the term “misinformation” but only a few of them understood “disinformation”. Users mostly received misinformation by word of mouth from peers, as well as through their social media platforms (whatsapp and facebook) where misinformation is shared.
- Some respondents seek to verify the information before sharing, mostly by researching via the internet to verify the validity of the news or confirming from their peers.

Analysis & Implication: Users need to be educated about disinformation so that they can be able to identify and flag it as well.

“Through social media handles, I got it through Whatsapp and then I was later told it is fake, the sharing of promotions to five Whatsapp groups.”
Female, 34yrs, Kajiado

“Any information that is not true but has been shared and there is no reliable source for it.”
Female, 30yrs, Kiambu

“I read the article and it was not true since it was information from no reliable source. So I told the person who send it to me that those are people who are spreading fear.”
Female, 30yrs, Kiambu

Many respondents had never reported misinformation before HIP, despite being inclined to do so.

Data:

- Many respondents believe that flagging misinformation is important so as to stop the spread of false news and rumors that might be harmful or deceptive to the public. However, not all of them had flagged misinformation prior to our introduction.
- Some participants had never reported misinformation because they were unaware of the appropriate channels to report the false information they came across. Some also feared that there might be consequences to reporting.
- The study findings show that some respondents doubted if proper actions or measures would be taken to curb the spread of the news if they reported it.

"I was not sure anything would be done about it and it was not the first time seeing it so I just ignored it."
Female, 23yrs, Kiambu

"I used it because it was the only way available for reporting."
Male, 23yrs, Nairobi.

"Sometimes you do not know what you are supposed to do or which bodies to report to and also fear. You might report something and someone comes and bullies you."
Female, 23yrs, Laikipia

Many respondents had never reported misinformation before HIP, despite being inclined to do so.

Data:

- For participants who reported the misinformation, some of them used the report options on the social media platform to flag the users.
- Most participants indicated that they used these methods primarily because it was the only one they were familiar with to report.

Analysis & Implication:

Tools like HIP are not widely known. Therefore, while people have thought about reporting misinformation, they often are not aware of the avenues towards doing this. Stakeholders should not only create some awareness about the tool, but ensure user behaviors can sync naturally with tool use (i.e., mobile access). Stakeholders could also leverage the fact that people are likely to recommend the tool to their social networks, as a way of spreading information about available tools.

In addition to flagging misinformation, being able to share worthwhile content is considered one of HIP's value add.

Data:

- A majority of the respondents use the HIP tool because they want to stop the spread of misinformation and promote healthy internet usage. The major benefit of the tool is that they can share good content that others can benefit from.
- Other benefits include gaining knowledge, the tool is readily available (easy to find), it's easy to use and doesn't take a lot of time to use.

Analysis & Implication: Having the option to flag worthwhile content is definitely a value add but because it is highly preferred, it can reduce the amount of misinformation that is flagged, which is actually the main objective of the HIP tool.

"You get a chance to push not only bad but good content from the browser to that application."
Male, 22yrs, Machakos

"It creates awareness. It gives you knowledge through the feedbacks given. It can also give you editing skills with the articles they provide."
Male, 28yrs, Nairobi

"To reduce the rate of ignorance amongst us. Ensuring that there is the right information at the first instance is very important."
Female, 31, Kiambu

Receiving an update on the actions taken by the provider and a monetary incentive could increase usage of HIP.

Data:

- Some people believed that their flags, not through HIP but through social media, were effective because they saw actions taken, such as suspension or blocking of the users account. However, some other participants felt that their flags were not effective as they did not receive any feedback from the provider.
- Some participants cited how adding incentives will serve as a motivator to flag users spreading false news through HIP or any other platform. For a few others, while an incentive is an advantage, they see flagging as a prosocial responsibility and a personal interest to help.

“When a person reports, action should be taken and feedback given.”
Male, 23yrs, Siaya

“No. Availability of incentive will just be an added advantage but I will still report even without incentive.”
Male, 22yrs, Machakos

“Yes. Sometimes we don’t pay much attention to online misinformation but when an incentive is included, sometimes you will just go on social media to find information that you can report.”
Male, 23yrs, Kajiado

Receiving an update on the actions taken by the provider and a monetary incentive could increase usage of HIP.

Data:

- Most participants who would like incentives expressed their preference for financial incentives, with an average payment of 100 - 300 Kenya shillings per flag, that is, 5-10k per month.
- Only a few people expressed preference for non-financial incentives, such as recognition and a certificate of participation.

Analysis & Implication: Generally, most people will continue to flag, even without an incentive. However, providing non financial incentives would be a great way to encourage HIP usage. This can be through recognizing people who accurately flag information, or providing feedback on how their work has improved the internet.

“Monetary, once a week or based on articles you flag.”
Female, 22yrs, Kajiado

“Pay me with money and give me a number of articles that you would like me to flag. KES 1,000 for 5 flags.”
Female, 34yrs, Kiambu

“With comments like thank you and also giving a token of appreciation.”
Female, 26yrs, Kiambu

There is a risk to reporting misinformation, but receiving feedback can keep flaggers motivated to keep flagging.



Anonymity

Although most people trust that HIP is anonymous (due to the use of unique IDs, no public identifying information required and their trust in the provider), a few respondents are not completely sure because they cannot verify the truth of the matter.



Confidentiality

Some people are either not sure or are not convinced that the providers do not have access to their personal digital or mobile device. Though some people are convinced.

"I don't know because you can see the address of my device (i.e., IP address). I believe there is a way that one can see it." Female, 22yrs, Kajiado



Cyber Bullying

Some respondents expressed their fear of receiving threats from people, trolls online and being cyber bullied, which might lead to physical harm.

There is a risk to reporting misinformation, but receiving feedback can keep flaggers motivated to keep flagging.

Data:

- Whereas a few people did not appear to be worried about the risks involved, many participants stated that the risks make them worry, self-conscious and less likely to report because of fear of being attacked, bullied or having problems with other people.
- The respondents suggested that in order to keep people flagging despite the risks, platform providers should provide timely feedback to the flaggers to encourage them to report.

Analysis & Implication: Emphasise the anonymity aspect of the tool to boost the confidence of users and to enable them to keep flagging. Platforms like HIP should also provide timely feedback to users via a timeframe for all the actions that need to be taken should be strictly adhered to.

“You may become a target due to reporting through cyber bullying and people may end up causing harm to you physically if they know who you are.”
Male, 22yrs, Nairobi.

“I think they give me more courage to do so. There is still need to educate the population about having the right information.”
Female, 31yrs, Kiambu

“If nothing is being done every time I report, then that will not motivate me at all to report again. Increasing confidentially for the people doing the reporting.”
Female, 23yrs, Kiambu

Guaranteed anonymity and tagging worthwhile content also reduces the risk attached to flagging misinformation.

Data:

- A majority of people believe there is especially a risk in reporting misinformation for sensitive topics related to political or tribal issues, as well as reporting influential people or celebrities since they believe these people will be able to find the them.
- One person only tagged worthwhile content since they believed it would not bring them any harm, and one believed it's their opinion so there is no risk involved.
- Some people feel that it is a risk to report people if there is a possibility that the information could actually not be misinformation, since this puts their jobs or work or platforms at risk.

“Yes. The way I view a particular information that might not be the view of the one who was sending the information. So if I flag such it might affect their work so that's a risk.”

Female, 28yrs,
Machakos

“Yes, politics but as long as you are anonymous there will not be problems or risk but if you are not then there can be a problem.”

Male, 23yrs,
Kirinyaga

Guaranteed anonymity and tagging worthwhile content also reduces the risk attached to flagging misinformation.

Data:

- A few of those who didn't think there was a risk believe that the anonymity element protects them.

Analysis & Implication: To bring in some objectivity, users need to be educated on effective ways of identifying misinformation so that they do not feel that they are risking other people's careers by flagging subjectively.

Additionally, trust needs to be instilled in users so that they can feel safe about flagging controversial issues. Again, providing timely feedback can boost their confidence.

“Yes I believe there are risk if I report a manipulative message going round about political leaders and the information being spread is not true and I report that it would be risky for my safety compared to when I report manipulation about Covid 19 vaccine.”

Female, 30 yrs, Kiambu

Owing to the benefits of the HIP tool, users are likely to recommend it to their network.

Data:


- Most people absolutely plan on using the HIP tool again, especially if they can get compensation for it. On a scale of 1 to 5, 21% rate themselves at 4 for the likelihood of continuing to use the HIP tool, while 52% rate themselves at 5.
- Most people are likely to recommend the tool to people in their network.

Analysis & Implication: Misinformation platforms can leverage on people’s tendency to recommend products or services to their social network. During the education or awareness activities, this element can be incorporated into the content.

“Based on time and availability of internet, i will continue using it.”
Female, 22yrs, Kajiado

"I will use it but it will depend on whether action is taken on the flagged content. It I will be able to help someone else then I will use it."
Female, 23yrs, Kiambu

"So that we can have a large number of people out there fighting against the lies and security. Secondly, I also want my other friends to gain because it is a nice tool."
Female, 23yrs, Turkana

The background of the slide is a photograph of a beach scene. In the foreground, there is a sandy beach with some debris and a small boat. The water is calm and reflects the sky. In the distance, there are more boats and a small structure. The sky is filled with soft, white clouds. A large, solid blue rectangular overlay covers the right two-thirds of the slide, containing the main text.

Recommendations from the Full Study

A summary of **recommendations** for
improvements to the **HIP** platform.

In conclusion, volunteer-based crowdsourcing can be a useful addition to the current misinformation ecosystem.

Throughout the report, we highlight different ways in which volunteer-based crowdsourcing can be a value add to the existing misinformation ecosystem.

- There is an existing culture of volunteering amongst many populations as well as a sense of pro-social responsibility which organizations can leverage on when creating awareness for misinformation platforms.
 - We found that despite the risk people attached to flagging political content, users still flagged such content, owing to their sense of responsibility.
 - Even without an incentive, many users will continue to flag misinformation, which confirms that such crowdsourcing platforms have an opportunity to leverage on people's pro-social responsibility.
- Crowdsourcing platforms are preferable to people because of the highlighted HIP advantages.
 - Some people were aware of other ways of flagging misinformation, but preferred HIP once they were introduced to it.
 - The study found that HIP is appropriate, easy to use, and can change the quality of content online, which contributed to the reasons people used it.
 - Users also liked that the time needed to report misinformation using HIP was minimal.

In conclusion, volunteer-based crowdsourcing can be a useful addition to the current misinformation ecosystem.

Throughout the report, we highlight different ways in which volunteer-based crowdsourcing can be a value add to the existing misinformation ecosystem.

- The HIP tool and other crowdsourcing platforms can benefit from social networks because it can be simple to refer or introduce others in your network to the crowdsourcing tool.
 - All the respondents agreed that information needs to be verified first before sharing. Most of them do this by consulting with their social network (i.e., family, relatives and/or friends).
 - Users pointed out that the process of downloading and getting set up with HIP is easy and they did not need any help in flagging misinformation using it. Most of the users therefore plan on recommending the tool to other people.
- There are however some ways in which volunteer based crowdsourcing platforms can be improved so as to maximize their value add. The following slides have our recommendations on how this can be done.

User-Driven Recommendations on Improving User Engagement



Quality

To improve the quality of flags:

- Have more flagging options
- Simplify the flagging definitions
- Provide a section for putting images
- Worthwhile content should only have 1 rating



Usage

To increase usage:

- Have a version for the phone
- Improve the tool's responsiveness
- Provide an incentive for active users
- Enable people to flag social media
- Be able to see other worthwhile content people have flagged
- Translate to other languages



Salience

To increase salience of the tool:

- Have prompts on the icon or reminders
- Make the icon bigger, so that it can be salient to the users

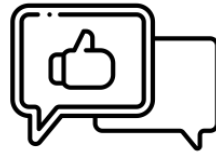
Our Recommendations on Improving User Engagement



Anonymity

Reinforce that the tool is anonymous:

Remind and reinforce the reality that using HIP can protect anonymity. Users said they trusted the plugin because of Busara, but not because of HIP itself. Show the steps HIP takes to preserve user anonymity (i.e., the technology used to how one manages the reports).



Feedback

Increase feedback mechanisms:

Provide a system in which users can understand, whether on an aggregate level or individual level, on how feedback is being actioned. Not only does this increase salience of the tool, but it proves that user behaviors make a difference. This can be an opt-in service as this recommendation conflicts with the next on anonymity.



Access

Enable mobile access:

Users in our qualitative follow-up believed that having phone access would enable them to more frequently access the HIP plugin because this is where they more frequently engage with online content, especially via social media. Moreover, stakeholders should create awareness of more tools like HIP so the public can utilize them.

Our Recommendations on Improving User Accuracy



Remove Worthwhile

Remove positive flag to orient preferred user behaviors:

Worthwhile flags were overwhelmingly used. This enables users to feel like they are actively using the plugin and deters the pressure of reporting misinformation. Alternatively, make Worthwhile one flag without the severity scale.



Add a Field

Create additional field to enable accurate identification of misinformation:

Make users identify which lines or tracts of text are considered to be “misinformation”. Failing to do this will result in fact-checkers dismissing most of the cases since users tend to flag entire websites, which make identifying misinformation impossible.



Create Training

Create a pre-onboarding training on misinformation:

Users will want to know how to identify misinformation. Our users felt uncomfortable reporting misinformation because they 1) weren't sure it was misinformation; and 2) didn't want to harm anyone in case they were wrong.

Areas for Further Analysis

Clarifying Misinformation

- It is clear from our experiment that users are uncomfortable with identifying misinformation, driven by self-doubt, uncertainty, risk and a proclivity towards flagging the positive elements of the internet.
- Given that misinformation is individualized towards one's echo chamber, what are ways to train users on identifying misinformation in an objective manner?

1

In the age of misinformation, how have other organizations approached how to objectively train others on identifying misinformation despite the influence of echo chambers?

2

How can HIP use the essentials of this finding to train volunteers on identifying misinformation?

Areas for Further Analysis

Feedback Mechanisms

- PesaCheck's current system involves a user flagging content using the Whatsapp platform, their team reviewing the content, and then providing feedback to the user in the Whatsapp group within a 24 hour window.
- However, the effectiveness of this feedback system hasn't been researched. This is also a gap in the literature. The mentioned research questions can be one way of filling this gap, and a qualitative approach can be used for the study.

1

What constitutes a timely, effective provision of feedback to users in order to sustain their HIP usage?

2

What are the best ways or methods to provide feedback to users?

Areas for Further Analysis

Risk and Willingness to Report

- The study findings show that despite users' intent to stop the spread of misinformation, many use the tool to flag Worthwhile content. They also provide more information on a Worthwhile flag as opposed to a harmful flag.
- We hypothesize that this is because of the risk people attach to flagging harmful content. It would be useful to further explore this through mixed methods research on risk preferences.

1

How do perceptions of risk emerge when it comes to reporting report harmful content, and what are patterns and trends according to website type (i.e., political vs. non-political)?

2

What tools and/or framings can be used to meaningfully address perceptions of risk on the internet?



Contact us for more information

acceleratorlab.ke@undp.org

| www.ke.undp.org

contact@busaracenter.org

| www.busaracenter.org

Co-building the Accelerator Labs as a joint venture with:



UNDP
Core
Partners



Action Partners

