A Machine Learning Application for Raising WASH Awareness in the Times of *COVID-19* Pandemic

Rohan Pandey[1#], Vaibhav Gautam[1#], Chirag Jain[2$], Priyanka Syal[3$], Himanshu Sharma[4$], Kanav Bhagat[2&], Ridam Pal [2&], Lovedeep Singh Dhingra[5&], Arushi[5&], Lajjaben Patel[5&], Mudit Agarwal[5&], Samprati Agrawal[5&], Manan Arora[6], Bhavika Rana[2], Ponnurangam Kumaraguru[2], Tavpritesh Sethi[2*]

[1]Shiv Nadar University, Noida, Uttar Pradesh, India
[2]Indraprastha Institute of Information Technology, Delhi, India
[3] Microsoft India (R&D) Private Limited
[4]GL Bajaj Institute Of Tech and Management, Greater Noida, Uttar Pradesh, India
[5]All India Institute of Medical Sciences, New Delhi, India
[6]Plaksha University, Punjab, India
[*]tavpriteshsethi@iiitd.ac.in

Abstract

Proactive management of an Infodemic that grows faster than the underlying epidemic is a modern-day challenge. This requires raising awareness and sensitization with the correct information in order to prevent and contain outbreaks such as the ongoing *COVID-19* pandemic. Therefore, there is a fine balance between continuous awareness-raising by providing new information and the risk of misinformation. In this work, we address this gap by creating a life-long learning application that delivers authentic information to users in Hindi and English, the most widely used languages in India. It does this by matching sources of verified and authentic information such as the WHO reports against daily news by using machine learning and natural language processing. It delivers the narrated content in Hindi by using state-of-the-art text to speech engines. Finally, the approach allows user input for continuous improvement of news feed relevance daily. We demonstrate this approach for Water, Sanitation, Hygiene for containment of the COVID-19 pandemic. Thirteen combinations of pre-processing strategies, word-embeddings, and similarity metrics were evaluated by eight human users via calculation of agreement statistics. The best performing combination achieved a Cohen's Kappa of 0.54 and was deployed as *On AIr*, *WashKaro*'s AI-powered back-end. We introduced a novel way of contact tracing, deploying the Bluetooth sensors of an individual's smartphone and automatic recording of physical interactions with other users. Additionally, the application also features a symptom self-assessment tool based on WHO-approved guidelines, human-curated and vetted information to reach out to the community as audio-visual content in local languages.
WashKaro - http://tiny.cc/WashKaro

1 Introduction

Raising healthcare awareness for primary prevention of diseases is a challenge all across the globe. However, the modern world faces another challenge- the Infodemic that accompanies an epidemic and may spread faster than the latter itself. Yet, raising awareness with the correct information can save lives. Hygiene promotion is the most cost-effective health intervention if accurate content is delivered effectively. A majority of preventable diseases result from unhygienic practices. Water, Sanitation and Hygiene (WASH) measures such as hand-washing are also important in limiting the spread of pandemics such as the currently raging COVID-19. Further, the awareness-raising content is often not available to those who need it the most and in a format that they easily understand leading to profoundly wide socio-economic impacts of this lack. In 2017, around 55% of the global population did not make use of a safely managed sanitation service effected in part due to lack of awareness in addition to the lack of facilities at home [12]. Around 827,000 people in low and middle-income countries die as a result of inadequate water, sanitation, ,and hygiene each year. A significant proportion of these deaths can be averted through the dissemination of information about WASH practices and their critical role in preventing diseases by delivering authentic information content in local languages. This cuts across the Sustainable Development Goal 3 (Good Health and Well Being for All) and 6 (Adequate Sanitation and Hygiene for All).

India is the second-most populous country in the world, with more than 1 billion citizens where a staggering 344 million lack hygienic defecation facilities [11]. The World Health Organisation states that more than 500 children under the age of five die each day from diarrhoea in India alone [11] and estimates that 21 percent of communicable diseases in India are linked to unsafe water and the lack of hygiene practices [11].

Ironically, India is also one of the largest and fastest-growing markets for digital consumers, with 560 million Internet subscribers in 2018 [3] and about 60% of Indian users anticipate that the m-Health technologies will improve healthcare within the next three years [4]. This offers a unique opportunity to bridge the gap in information availability through m-Health technologies to reach out to those

[#]contributed equally [$]contributed equally [&]contributed equally

who need it the most, and in a medium that they understand the most, e.g. audios delivered in local languages, thus narrowing the divide between these resources and the masses.

The recent pandemic outbreak of Coronavirus (COVID-19) has demonstrated the need for proactive containment and prevention measures including repeated hand-washing. Every single day lost of proactive interventions has an exponential impact and countries that acted early were able to contain the disease effectively, thus saving thousands of lives and dollars [9]. Therefore, there exists a dire need for proactive information in addition to proactive testing while preventing the spread of misinformation.

In this work, we demonstrate an AI-Powered Infodemic Management solution *WashKaro* that uses NLP approaches, machine learning and m-Health to reach out to the community as audio-visual content in local languages. *WashKaro* provides authentic sources of information with daily news, complemented with a Bluetooth based Contact Tracer, WHO directive-based Symptom Self-Assessment tool and human-vetted information delivering these in Hindi, the most widely understood local language across India.

All aforementioned features are explained in more detail in the following sections.

2 On AIr

In this section, we present our fed public healthcare intervention workflow, designed and centered around imparting healthcare information effectively. Our proposed machine learning methodology is complemented with active user feedback and works as follows.

2.1 Proposed Framework

In the early stages of learning, News articles and WHO guidelines are first preprocessed to ensure the data is of standard quality. The preprocessed text is further summarized using machine learning techniques for extractive text summarization. Further, both the summarized texts are converted to word embeddings. A similarity metric is leveraged to measure the similarity of the two text documents on the calculated embeddings. For each news article, the WHO guideline with the highest similarity score is provided to the user in the local language in the form of text and audio both. This serves the intent of augmenting the users daily news consumption with an appropriate WHO guideline to increase healthcare awareness. The user reviews the matching of an article and WHO guideline as either relevant or irrelevant and active user feedback is taken to improve the user experience and to provide increasingly relevant content. For each WHO guideline, we define two classes, Relevant Articles Cluster and Irrelevant Articles Cluster as shown in Fig 1. The news articles considered relevant form the first class and irrelevant articles form the second. These clusters get updated periodically for each guideline. For every update in the database we one of the following scenarios.

Scenario A - New News Article For every new article in our data, the sentence similarity model is deployed with the following input parameters. For each guideline in our data, the similarity score is now calculated between the news article and each news article in the relevant and irrelevant cluster of the specific guideline. These scores are averaged to give a final similarity score between the news article and the relevant and irrelevant cluster of each guideline. The news article along with the guideline with the highest similarity score for its relevant cluster is served to the user. The relevant and irrelevant class of each guideline is updated as per the reviews periodically.

Scenario B - New Guideline For every new guideline in our data, the sentence similarity model is deployed against all the news articles of a particular day. The top 10 matchings with the highest scores are provided to the user. With time, owing to the reviews from the users, the relevant and irrelevant clusters are updated for that particular guideline.

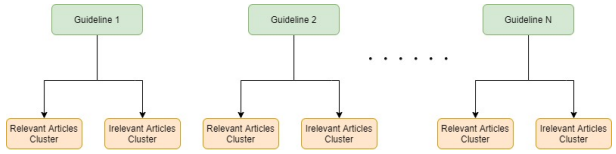

Figure 1: WHO Guidelines. Each WHO guideline has a relevant and irrelevant news article cluster felicitating improved matching.

Numerous techniques have been explored in each portion of the proposed framework and have been explained in detail.

2.2 Dataset

We have validated our approach using the following datasets.

WHO Guidelines This dataset comprises of WHO guidelines obtained from publically available WHO reports with special emphasis on Water, Sanitation , and Hygiene (WASH) from various WHO reports published. The dataset comprises more than 400 WHO articles manually scraped from individual reports owing to the varied format of each report. The dataset comprises the title of the guideline, the guideline, a category in which it belongs, and the URL of the WHO published report.

Some of the broad categories of these reports are

- Corona virus Prevention
- Guidelines for safe recreational water environments
- Water, sanitation and hygiene in health care facilities
- Progress on Sanitation and Drinking Water
- A practical guide to Auditing water safety plans
- Progress on Drinking Water, Sanitation and Hygiene

This dataset comprises of news articles scraped from publically available news articles. The dataset comprises the article headline, article text, URL of the article and the date of publishing. We have maintained the following news article datasets.

English The English News Article dataset consists of news articles extracted from ‘The Hindu’. The news articles

are filtered using the following keywords *'Handwash', 'Hygiene', 'sanitation', and 'health'*.

Hindi The Hindi News Article dataset consists of news articles extracted from 'Jagran'. The news articles are filtered using the following keywords *'svachta' , 'safai', 'haath dhona', 'saaf', 'haath ragad'* which are Hindi translations of 'Cleanliness', 'handwash', 'hygiene' and 'sanitation'.

2.3 Sentence Matching Model

In this section, we explain in detail the sentence matching model.

Preprocessing The dataset as described in Section 2.2 needs to be preprocessed to transfer text from human language to machine-readable format for further processing. The following preprocessing is done.

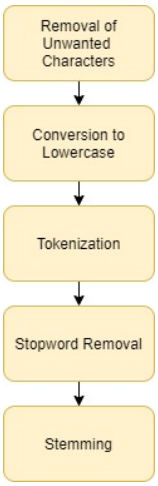

Figure 2: Stages in preprocessing. Various preprocessing methods are used to transform the text into a more digestible form so that machine learning algorithms can perform better.

News Article Dataset *Removal of Unwanted Characters* All characters except A-Z and ['.', ',' ,':'] are removed.

Conversion to lowercase The entire text is converted into lowercase. Lowercasing significantly helps with the consistency of expected output.

Tokenization This is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens.

Stop word removal Stop words are the most common words in a language like 'the', 'a', 'on', 'is', 'all'. These words do not carry important meaning and are removed from texts.

Stemming Stemming is a process of reducing words to their word stem, base or root form (for example, books book, looked look). We have deployed the Porter Stemming algorithm.

Summarization To incorporate the varied length of different articles, numerous summarization methods were explored to eliminate the effect of text length on the effectiveness of the proposed framework. Text summarization based on different methods has been tested which has allowed us to select the best framework for having an optimized summary of the news article. We have explored models related to extractive summarization, where most relevant sentences from the articles get extracted based on their importance.

Graph-Based Approaches The following graph-based approaches were explored, Text Rank with one-hot vectors, Text Rank with Word2Vec, Text Rank with Glove, Lex Rank, Reduction, PyTextRank, TextRank with Gensim. The majority of these approaches are based on Text-rank which is an extractive and unsupervised text summarization technique. TextRank algorithm plays a vital role in graph-based text summarization approaches. TextRank algorithms work simultaneously like PageRank algorithms. PageRank algorithms were primarily used for web search engine which has been incorporated by Google. The procedure followed in this algorithm is stated as follows, concatenation of whole text is done converting it to a single text, then these texts get split into sentences. The vector representation of these sentences is done using different word embeddings. This is done to calculate the similarity score between sentence vectors based on which similarity matrix gets created. Based on the similarity matrix graphs are created, leading to sentence rank calculation. The graphs are created based on the similarity matrix where each sentence is treated as vertices and similarity scores as edges weight.

Topic-Based Latent Semantic Analysis uses the Bag of Words model for generating a term-document matrix, in the term-document matrix each row represents the terms and columns represent the documents in which the terms are present. Latent Semantic Analysis based text summarization is an unsupervised robust Algebraic-Statistical method that extracts hidden semantic structures of words and sentences i.e. it extracts the features that cannot be directly mentioned. These features are essential to data but are not original features of the dataset. In Latent Semantic Analysis, the input document is first represented as a document term matrix used to represent the importance of words in sentences. There are different approaches to fill out the cell values such as Frequency of word, Binary representation, TF-IDF (Term Frequency-Inverse Document Frequency), Log entropy, Root type. After the matrix is generated, a singular value decomposition is performed on the generated document term matrix. SVD is an algebraic method that can model relationships among words/phrases and sentences. Using the results of SVD different algorithms are used to select important sentences. Here we have used the Topic method to extract concepts and sub-concepts from the SVD calculations and are called topics of the input document. These topics can be sub-topics, and then the sentences are collected from the main topics. A high number of common words among sentences indicates that the sentences are semantically related. The meaning of a sentence is decided using the word it contains, and the meaning of words is decided using the sentences that contain the word.

Feature-Based Luhns method is a simple technique to generate a summary from given words. It begins with transforming the content of sentences into a mathematical expression, or vector(represented below through binary representa-

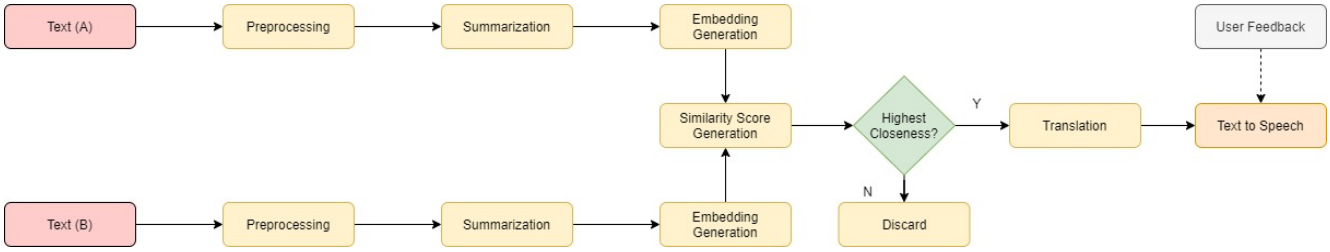

Figure 3: Methodology. The pipeline takes in news articles and WHO reports and constructs two-level sentence similarity between titles and the full-text to construct a similarity score. Finally the relevant texts are subject to text to speech translation for consumption in local language (Hindi).

tion). Here we use a bag of words, which ignores all the filler words. Filler words are usually the supporting words that do not have any impact on our document meaning. Then we count all the valuable words left to us. In this step, we evaluate sentences using sentence scoring techniques. Tf-IDF can also be used to prioritize the words in a sentence. Once the sentence scoring is complete, the last step is simply to select those sentences with the highest overall rankings.

Vocabulary Minimization The KL is a measure of difference (divergence) between two probability distributions P and Q. KL divergence is a method for scoring sentences in the summarization, it shows the fact that good summaries are intuitively similar to the input documents. It describes how the importance of words alters in the summary in comparison with the input, i.e. the KL divergence of a good summary and the input will be low.

Embeddings Embedding is a technique to transform text and convert them into a form, such that a machine can process it. It is one of the most popular representations of document vocabulary. The transformation is done in such a way that machine level analysis can be carried out on them. An embedding is a learned representation for text where words that have the same meaning have a similar representation. Our methodology employs the following embeddings.

Word2Vec [7] is a statistical method for efficiently learning a standalone word embedding from a text corpus. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. Two different learning models were introduced that can be used as part of the word2vec approach to learning the word embedding, Continous Bag of Words (CBOW) and skip-gram model. The CBOW model learns the embedding by predicting the current word based on its context. The continuous skip-gram model learns by predicting the surrounding words given a current word. The key benefit of the approach is that high-quality word embeddings can be learned efficiently (low space and time complexity), allowing larger embeddings to be learned (more dimensions) from much larger corpora of text (billions of words).

GloVe (Global Vectors for Word Representation) [8] Glove is an extension to the word2vec method for efficiently learning word vectors. Classical vector space model representations of words were developed using matrix factorization techniques such as Latent Semantic Analysis (LSA) that do a good job of using global text statistics but are not as good as the learned methods like word2vec at capturing meaning and demonstrating it on tasks like calculating analogies. GloVe is an approach to marry both the global statistics of matrix factorization techniques like LSA with the local context-based learning in word2vec. Rather than using a window to define local context, GloVe constructs an explicit word-context or word co-occurrence matrix using statistics across the whole text corpus. The result is a learning model that may result in generally better word embeddings.

Google Sentence Encoder [1] encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks. The model is trained and optimized for greater-than-word length text, such as sentences, phrases or short paragraphs. It is trained on a variety of data sources and a variety of tasks to dynamically accommodate a wide variety of natural language understanding tasks. The input is variable-length English text and the output is a 512-dimensional vector. The universal-sentence-encoder-large model is trained with a Transformer encoder.

TF-IDF (Term frequency inverse document frequency) TF-IDF stands for term frequency-inverse document frequency, and the TF-IDF weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

This is composed of 2 parts. TF, which measures how frequently a term occurs in a document, and IDF, which measures how important a term is by giving higher weight to words occurring only in a few documents.

$$TF_i = \frac{N_i}{N} \tag{1}$$

Where N_i is the number of times i appears in a document and N is the total number of terms in the document.

$$IDF_i = Log\frac{N}{DF_i} \tag{2}$$

Where N is the total number of documents and DF_i is the number of documents in which word i occurs.

Similarity Metric To generate a similarity score, between the news article and the WHO guideline, the following similarity metrics have been used.

Cosine Similarity is a metric used to measure the similarity between the two documents. It is independent of the size of the documents. Cosine similarity calculates similarity by measuring the cosine of the angle between two vectors. This is calculated as

$$cos(\theta) = \frac{\mathbf{A.B}}{\mid \mathbf{A} \mid . \mid \mathbf{B} \mid} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Mathematically speaking, Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any angle in the interval $(0, \pi]$ radians. It is thus a judgment of orientation and not magnitude, two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

Word Mover Distance (WMD) suggests that distances and between embedded word vectors are to some degree semantically meaningful. It utilizes the property of word vector embeddings and treats text documents as a weighted point cloud of embedded words. The WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to travel to reach the embedded words of another document. WMD shows that this distance metric can be cast as an instance of the Earth Movers Distance (a well-studied transportation problem for which several highly efficient solvers have been developed).

WMD enables us to assess the distance between two documents in a meaningful way, even when they have no words in common. The method also uses the bag-of-words representation of the documents (simply put, the words frequencies in the documents). The intuition behind the method is that we find the minimum traveling distance between documents, in other words the most efficient way to move the distribution of document 1 to the distribution of document 2.

2.4 Translation and Text2Speech

The news article and matched WHO guidelines with the highest similarity score are converted into the local language (Hindi) using Google Could Platform. Googles pre-trained neural machine translation delivers fast and dynamic translation results and Google Cloud's Text-to-Speech converts text into human-like speech.

2.5 Evaluation Metrics

As our entire methodology is aimed at efficiently providing healthcare information to the masses, our success needs to be a measure of the acceptability of the masses. Inter-rater reliability is the extent to which two or more raters (or observers, coders, examiners) agree. It addresses the issue of the consistency of the implementation of a rating system. Inter-rater reliability can be evaluated by using a number of different statistics. High inter-rater reliability values refer to a high degree of agreement between two examiners. Low inter-rater reliability values refer to a low degree of agreement between two examiners. We have evaluated our performance using the following commonly accepted statistics.

Percentage agreement Percentage Agreement amongst raters is a statistic calculated as the number of agreement scores divided by the total number of scores. In the case of multiple users, this technique gets complex. If total raters are n then it has to check for n(n-1)/2 combinations. Also, it does not take the chance of agreement into account and overestimate the level of agreement. Hence it is not reliable alone.

Cohen Kappa Cohens kappa coefficient [6] is a statistic that is used to measure inter-rater reliability for qualitative (categorical) items. It is generally thought to be a more robust measure than a simple percent agreement calculation since k takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters that classify N items into C mutually exclusive categories. To find the coefficient, we use the following formula

$$k = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

2.6 Experimentation

The first step of our experimentation involved the creation of the datasets mentioned in Section 2.2. Manual scraping of the WHO Guidelines dataset was done and a total of nearly 400 WHO articles were stored in a local CSV file. English News Articles and Hindi News Articles of the current day were also scraped and stored in their respective datasets.

In the first set of experiments, WHO Guidelines and English News article dataset was used. These datasets were preprocessed by the steps removing unwanted characters as mentioned in Section 2.3. After the preprocessing was done, pairs were generated where the current days' news articles are mapped against the entire WHO guidelines database. Pair generation is followed by evaluating the sentence similarity amongst the News Article Text vs WHO article Text. Numerous summarization techniques have been explored. The summarization techniques have been evaluated qualitatively since a quantitative measure was not appropriate for

Model	Embedding	Preprocessing	TF-IDF	Similarity Metric
1	Word2Vec	X	X	Cosine
2	Word2Vec	✓	X	Cosine
3	Word2Vec	X	✓	Cosine
4	Word2Vec	✓	✓	Cosine
5	Word2Vec	X	X	Word Mover Distance
6	Word2Vec	✓	X	Word Mover Distance
7	Glove	X	X	Cosine
8	Glove	✓	X	Cosine
9	Glove	X	✓	Cosine
10	Glove	✓	✓	Cosine
11	Glove	X	X	Word Mover Distance
12	Glove	✓	X	Word Mover Distance
13	Google Sentence Encoder	X	X	Cosine

Table 1: Combinations of approaches tested. These included pre-processing word-embedding models and similarity metrics for evaluation by eight human users.

our use and LexRank based summarization has been deployed. After the generation of pairs, the set of models mentioned in Table 1 are employed on each of these sets of pairs. Table 1 specifies the following technical details

- Embedding used for each model
- If preprocessing was done
- If TF-IDF weighting was done
- Similarity Metric used

After running these models on each of the four sets of pairs generated, the sim scores generated are obtained. The text similarity models work as shown in Fig 3. The scores obtained differ from one model to another. The obtained similarity scores are sorted and the top 5 pairs obtained from each similarity model on a particular set of pairs are filtered out. These top pairs were presented to reviewers to classify as relevant and irrelevant (binary classification). The reviewer classified the pair of a news article and WHO guideline as relevant if the pair had some relation while if they didnt have anything in common, a reviewer marked that pair as irrelevant.

Since the idea of relevance and irrelevance is subjective, therefore a total of 8 reviewers were used. The reviewers were given a set of total 65 stories(top 5 stories of each model) and were asked to press the 'tick' button if they though the news article and the WHO report were relevant and the 'cross' button otherwise as shown in Fig 4. To find the best model, percentage agreement and Kappa score were calculated for each of the Models and the particular model with the highest Kappa score among all these models was selected. The best model selected based on these scores was used for future tasks as well.

After the model is selected, we translate both the WHO dataset and the News articles into Hindi. The translated text is converted into speech for ease of access and understanding. The methodology incorporates a feedback system wherein for each News-article and WHO report presented to the user, they can classify the matching as relevant or irrelevant. After learning from the feedback, Relevant and Irrelevant clusters for each WHO Guideline are formed as shown in Fig 1. For any incoming news article, the Guideline for which the Relevant cluster has the highest similarity to the news article is provided to the user in audio-text format. With each feedback, the model ensures the delivery of increasingly relevant and effective content by updating the clusters frequently. The same workflow was followed by scraping Hindi news articles, which were converted into English for running sentence similarity models. As done above, the resultant pairs are provided to the users in the form of Hindi text and speech. The application feed displays various news articles related to sanitation and hygiene as shown in Fig 4. User can switch between the news article and the matched WHO guideline and review the corresponding matching as relevant or irrelevant using the appropriate buttons.

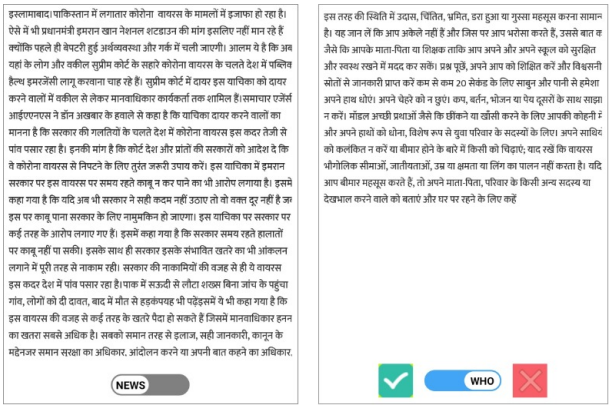

Figure 4: User Reviews. The user is provided the News Article and matched WHO Guideline. The matching is marked as Relevant/Irrelevant using the given buttons.

2.7 Observations and Results

To evaluate the performance of our healthcare intervention methodology. We relied on various inter-rater reliability

evaluation metrics as mentioned in Section 2.5. After calculating the similarity scores for a set of News articles and WHO guidelines using models defined in table 1, 8 different users were asked to classify the matched pairs as relevant or irrelevant. We calculated percentage agreement and Kappa Score for these 8 raters, on the pairs of News articles and WHO Guidelines and the following results are shown in Table3 and Table2 for the Hindi and the English News article dataset as mentioned in Section 2.2. In both cases, we can see that the model 2 (Preprocessing + Word2Vec Embedding + Cosine Similarity) gave the best results with 0.54912 Kappa Score and 79.2420 Percentage Agreement. As seen in Fig 6 and Fig 5, the Hindi News Article dataset provided better results in terms of both metrics. Hence for this app, model 2 was chosen.

Model No.	Kappa Score	% Agreement
2	0.54912	79.2420
5	0.47789	77.9810
4	0.389124	68.91235
3	0.41358	70.85732
8	0.38273	66.52417
10	0.332817	65.88290
11	0.261023	62.22067
7	0.278911	64.43729
9	0.258790	62.74201
13	0.23483	67.78201
1	0.167124	60.47632
12	0.10573	58.37529
6	0.0453067	55.78439

Table 2: Results of inter-rater agreement between eight users on Hindi News Articles. Cohen's Kappa and Percentage agreement were the two metrics used to evaluate the models.

Model No.	Kappa Score	% Agreement
2	0.51746	77.3809
5	0.47997	77.3809
4	0.400599	70.23809
3	0.391801	69.64285
8	0.352647	67.85714
10	0.296398	64.88095
11	0.262283	63.69047
7	0.240861	63.69047
9	0.238095	61.90476
13	0.206349	70.23809
1	0.184815	59.52380
12	0.085714	59.52380
6	0.020408	57.14285

Table 3: Results of inter-rater agreement between eight users on English News Articles. Cohen's Kappa and Percentage agreement were the two metrics used to evaluate the models.

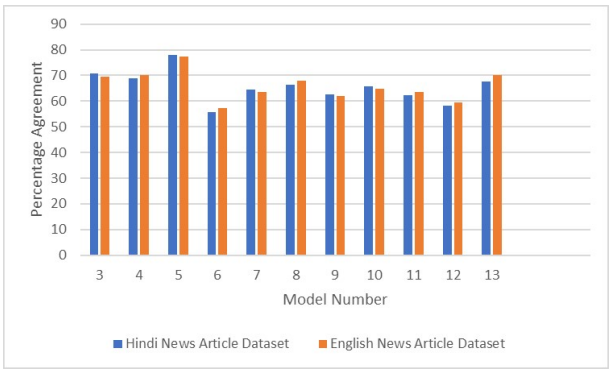

Figure 5: Percentage Agreement Score for models enumerated in Table 1. It is seen that Cohen's Kappa provided a better discrimination among models and was further used for model selection.

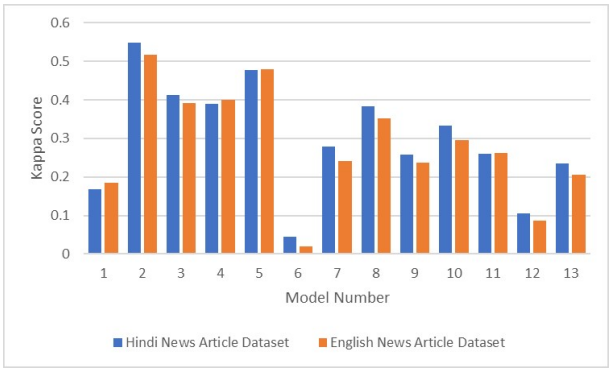

Figure 6: Cohen's Kappa Score for models enumerated in Table 1. The model yielding highest agreement among humans was selected for deployment.

3 Contact Tracing

During an outbreak of a contagious disease, it is essential to know the history of an infected person and trace down people they met over the last few days. In particular, for COVID 19, we need to investigate the last two weeks of interactions to be able to isolate and identify the potential patients. Currently, we are dependent on manual processes where we have dedicated experts who help trace the interactions with the infected patient. However, this process of contact tracing is prone to human errors which can cause missing out the intricate details of everyday interactions. To solve this and ease the process, we introduce a novel way of contact tracing, which uses the power of Bluetooth sensors of our smartphone and automatically records our past interactions with other users. We extended this approach to help maintain social distancing from infected/potential patients and notify users whenever they are at potential risk of infection if we have relevant data.

3.1 Proposed Architecture

In the proposed architecture, we are relying on Google Firebase[1] as our backend service for syncing unique user IDs for every smartphone. Once the user is authenticated via their phone number, the phone number is encrypted and stored in the backend and the user is assigned a randomized user ID. In case the user's mobile number is present in the list of quarantined/confirmed patients provided by the central data store, the state of the user is updated accordingly. We leverage Bluetooth low energy (BLE) Advertising to broadcast user information to all nearby devices. Whenever the user has a potential risk of infection, we notify the users accordingly. We also use notifications to enforce social distancing when a user has a potential risk of infection based on the state of nearby devices. We store all the interactions in the device itself. Once the user is identified as a potential patient, we can extract these interactions with user consent. The backend service then uses this interaction history to notify other users who have their user ID stored in the potential patient's interaction history.

3.2 Methodology

Authentication For authenticating users, we ask them to register using their mobile phone number and the same is verified using a one-time password (OTP) verification. Currently, we are using firebase authentication for this. The phone number is encrypted using a cryptographic function on the backend service to ensure the privacy of the user. Simultaneously it assigns a unique randomized user ID to the user for identification. Apart from that, the application fetches the current state of the user as stored on the backend service.

Recording Whenever someone installs the application and opens it, there is a background service that starts. It takes care of continuously BLE advertising and scanning. It can detect another device nearby with the same application installed using a unique identifier. It also broadcasts user information so the interaction can be recorded in the local database.

[1]https://firebase.google.com/

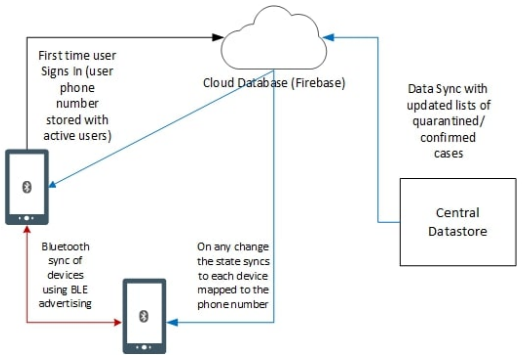

Figure 7: Proposed Contact Tracing architecture. This depicts the entire flow from user registration, peer to peer communication of user devices via Bluetooth and the data sync flow from the central database to the user device.

Reporting Whenever we find that a user is potentially infected, we can fetch the interaction history from the device on the users consent. We later use the interaction history to trace other users who have come in contact with the user and send an appropriate notification to alert them about it. Along with this, we can also use the user information from nearby devices to send real-time notifications if we detect that person is at risk because of the users nearby.

3.3 Current Implementation

Since we do not have a database with identified confirmed/quarantined users, we have integrated our current implementation with WashKaro without authentication to only record the interactions in the local device. Whenever the user is in a crowded place it urges the user to maintain social distancing to eliminate the potential risk of infection.

Figure 8: The current user interface of Contact Tracer. When the person is safe, the user is prompted with the photo on the left. Whenever the user comes in proximity to someone else using the app, the image is changed to the image on the right.

3.4 Observations and Limitations

Currently, this functionality is available only for Android smartphones with Bluetooth capability. Using the Global-

Stats statcounter[2], we can say that Android and iOS combined observed a mobile operating system market share of 99.29% as of March 2020. The Android app is published on Google Play Store while the iOS app is under implementation. Once it's complete, we can support inter-platform BLE scanning.

4 Symptom Self-Assessment

In the current scenario of the COVID-19 pandemic, with lockdown and social distancing measures in place, it might not be possible for the infected population to physically access hospitals. In such situations, app-based symptom questionnaires provide an accessible alternative. Thus, we devised a self-assessment tool to screen for the symptoms of COVID-19. This self-assessment questionnaire can potentially lower the burden on our already stretched-out healthcare systems by enabling quicker identification of symptomatic cases. The Suspect Cases can be guided via the Government helpline numbers and informed about proper self-quarantine protocols, nearby hospitals admitting COVID suspects, and testing centers, thereby enabling better control and limitation of the spread of the disease.

4.1 Methodology

We defined the Suspect Case using the WHO Interim Guidance on Global surveillance for COVID-19 caused by human infection with COVID-19 virus [10]. Using a flowchart as shown in Fig 9, to classify the cases as Suspect case (A), (B) and (C), the 7-point questionnaire was designed using the case definitions from the WHO Interim Guidance verbatim. Based on the application of the WHO criteria on the answers to the 7 questions, the user is notified- You are suspected/not suspected of having COVID 19.

4.2 Future Directions

We plan to translate this questionnaire to the other local Indian languages to widen the reach of the self-assessment tool. For the Suspect Cases, we can administer a second questionnaire to further stratify the risk of Acute respiratory distress syndrome(ARDS) and septic shock by assessing the severity of symptoms and looking for identified risk factors like age and pre-existing comorbidities that are not included in the WHO Interim Guidance. This can aid in making decisions regarding home quarantine against hospital admission. The app can also be used to identify other suspect cases in the same household. Further, to assist the government authorities to identify those requiring testing for COVID-19, we can ask for contact details of the Suspect Cases with informed consent and relay them to appropriate government authorities to enable targeted testing. A follow up of the suspect cases through push notifications, advising testing and recording test results, can help ensure that complacence does not set in.

4.3 Limitations

While this tool can aid in the identification of possibly infected individuals, we recognize the potential limitations. It

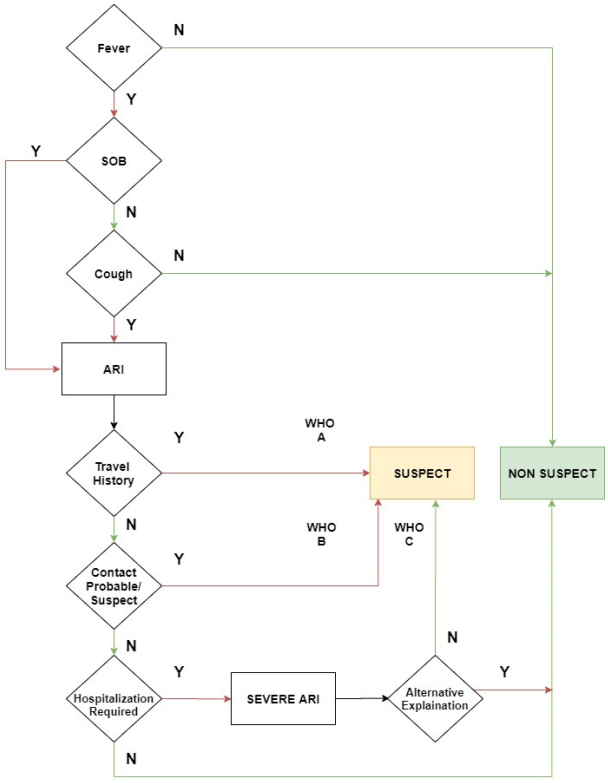

Figure 9: Self Assessment Tool Flowchart. Based on the WHO Interim Guidance verbatim.

[2]https://gs.statcounter.com/os-market-share/mobile/worldwide

is possible to have false-positive results in the population as the questionnaire is self-administered. It may also be difficult for a significant population to self-administer the questionnaire as they might not be comfortable with the English language. Strict use of the WHO guidelines might lead to misclassification of the infected individuals with atypical or mild symptoms as Non-suspects, which might lead to a false sense of security.

5 *WashKaro* Application

Our android application *WashKaro* is available for free download on Google Play Store - http://tiny.cc/WashKaro

5.1 User Interface Design

Clear and effective communication of preventive measures and updated information that includes statistics, guidelines, and the news is essential in a pandemic where some users are overwhelmed with information while some dont have access to any. To achieve this designing a trustworthy app that helps navigate the information clutter is important. It is necessary for preventive intervention applications like *WashKaro* to be recognised as sources of credible information by the users. This can be achieved by building an application that has compelling design, up-front disclosure, is comprehensive and has up-to-date content [2].

We use infographics for the clear and effective delivery of preventive measures [5]. Illustrations are used whenever possible to make the user aware of the data being collected in an attempt to make them comfortable, especially in the case of features like Contact Tracer that seem intrusive. Careful consideration is given to the content presented because the wrong information at the wrong time - negatively impacts the user. Keeping the same in mind, all notifications sent to the user are of a positive tone, prompting the user to carry out preventions like washing hands often and wearing masks when in crowded areas.

As this application is targeted towards the lesser-educated section of the society, and onboarding section is added to help the time users. There is an optional questionnaire comprising of a few basic questions on patient demographics to help us understand the prevalence and impact of interventions planned.

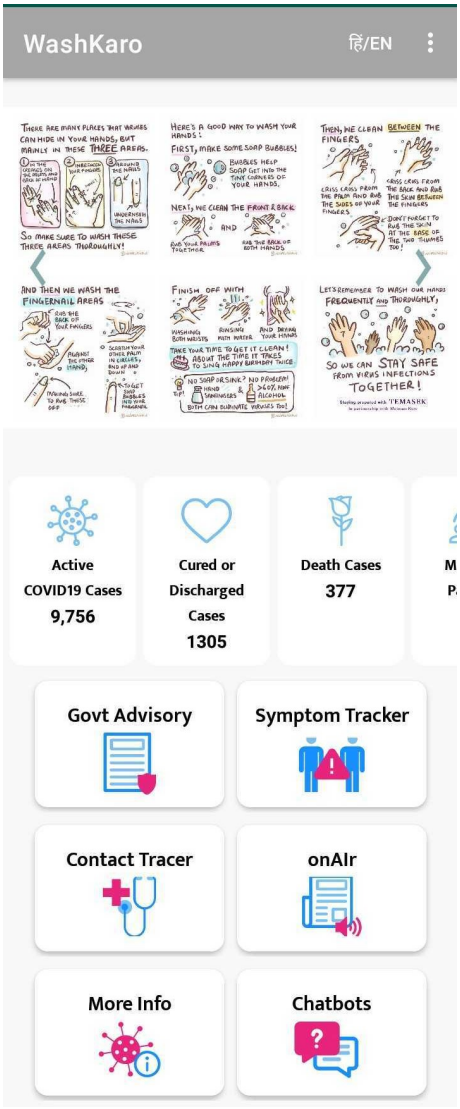

Figure 10: The current home screen user interface of the app displaying all the features.

6 Conclusion

To the best of our knowledge, this is the first ”Infodemic Management Suite” that uses state-of-the-art machine learning and is available in the hands of the users through m-Health technologies. Currently, it addresses the issue of ongoing WASH awareness in a local language in India, but we plan to expand this to include more regional and international languages. This is a daily-learning platform that allows user feedback on the relevance of content. The results of the technical approach taken in this work were evaluated by a panel of eight human users to choose the most appropriate model. However, this study has several limitations. The models have been trained on a relatively small corpus and we have only implemented the approach in Hindi, which is the most widely understood language in India. We do plan to incorporate more languages and local context to the application. All the humans evaluating the models were from a similar educational background. We hope to overcome this limitation through the feedback obtained from users of the *WashKaro* app. We also plan to devise a ranking score for the feedback providers based upon their reputation score for Public Health published via an accompanying website. Finally, the most important limitation is the lack of assessment of the interventional impact of this application. We plan to address this through a phased roll out with the primary health clinics in Delhi and appropriate partnerships delivering digital health interventions on-ground. Regardless, our current work highlights the potential of machine learning, m-Health and natural language processing in addressing primary health challenges and provides a framework for replicating such studies in a variety of public health challenges

including the COVID-19 pandemic.

7 Acknowledgements

This work was partly supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. Tavpritesh Sethi also acknowledges support from the Center for Artificial Intelligence at IIIT-Delhi, Prof. Rakesh Lodha from AIIMS New Delhi and Mr. Roshan Shankar, advisor to the Government of NCT of Delhi. Rohan Pandey and Vaibhav Gautam acknowledge the Department of Computer Science, Shiv Nadar University.

References

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[2] Aurora Harley. Trustworthiness in web design: 4 credibility factors. https://www.nngroup.com/articles/trustworthy-design/, 2016.

[3] Noshir Kaka, A Madgavkar, Alok Kshirsagar, Rajat Gupta, James Manyika, K Bahl, and Shishir Gupta. Digital india: Technology to transform a connected nation. *McKinsey Global Institute, March*, 2019.

[4] D Levy, C Wasden, D DiFilippo, and P Sur. Emerging mhealth: paths for growth. *PwC M-Health*, pages 1–44, 2012.

[5] Paul Benjamin Lowry, David W. Wilson, and William L. Haig. A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust. *International Journal of HumanComputer Interaction*, 30(1):63–93, 2014.

[6] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[9] C Jason Wang, Chun Y Ng, and Robert H Brook. Response to covid-19 in taiwan: Big data analytics, new technology, and proactive testing. *JAMA*, 2020.

[10] Global surveillance for covid-19 caused by human infection with covid-19 virus. https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf.

[11] India's water and sanitation crisis. https://water.org/our-impact/india/.

[12] Progress on household drinking water, sanitation and hygiene 2000-2017. special focus on inequalities. new york: United nations childrens fund (unicef) and world health organization, 2019. https://www.who.int/water_sanitation_health/publications/jmp-report-2019/en/.