ANALYZING COVID19 TWEETS USING HEALTH BEHAVIOURS THEORIES AND CLASSIFICATION MODELS

By

Boma Graham-Kalio

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2021

DEDICATION

I dedicate this thesis to my sister

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In order to explain people's health habits, Health Behaviour Theories has been used to analyze posts on social media during previous incidents. With regard to the pandemic, social media data can expose public attitudes and experience, which helps to reveal elements that impede or encourage attempts to reduce the spread of the disease.

This thesis aims to use Health Behaviour Theories (Health Belief Model, Social Norm and Trust) and Machine Learning Models to explain/examine people's behaviour and reactions towards COVID-19. Using text mining techniques, we analyzed COVID-19 comments from Twitter and used candidate key phrases which represents each construct to label the comments according to their Health Behaviour constructs. Next, we used three machine learning models (LinearSVC, Decision tree and logistic regression) to classify the comments into their construct. 10-fold cross validation was then used for evaluating the model to check for bias, while precision, recall and F1-Score were the metrics used for evaluating the classification results. In the multiclass (single label) classification result, decision tree and linearSVC performed best with an F1-score of 98%, while for the multiclass-multilabel classification result decision tree was the best with an F1Score of 1.00%. Finally, we performed thematic analysis based on each construct and further categorised them into themes which gave meaningful insight into each construct. The result from the thematic analysis revealed a total number of 32 themes from all the constructs.

LIST OF ABBREVIATIONS USED

HBM	Health Belief Model
CTA	Cue to Action
DecTree	Decision Tree
SVM	Support Vector Machine
LR	Logistic Regression
COVID - 19	Coronavirus Disease 2019
TF-IDF	Term Frequency-Inverse Document Frequency
ML	Machine Learning
NLP	Natural Language Processing
linearSVC	Linear Support Vector Classifier

ACKNOWLEDGEMENTS

I would like to express my special thanks to everyone who played a special part in the completion of this thesis. My parents who supported me with love, encouragement and prayers. Secondly my supervisors Professor Rita Orji and Professor Nur Zincir - Heywood who helped guide me on where to research exposing me to new knowledge as this project has made me gain extensive knowledge.

CHAPTER 1 INTRODUCTION

1.1 The Problem

Several viruses have attracted the attention of the medical and scientific communities in recent years for posing a significant risk to international public health. Coronaviruses are among these viruses, which have a wide international effect due to the severe respiratory syndromes they cause and the most well-known of which are Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). In December 2019, a recent outbreak of human infection caused by a novel coronavirus (2019-nCoV, now known as SARS-CoV-2) was discovered and registered in Wuhan City, Hubei Province, China[30].

The 2019 Coronavirus Disease known as COVID-19 is a fast transmission disease caused by Coronavirus 2 Severe Acute Respiratory Syndrome (SARS CoV2). COVID-19 has already been classified as a global pandemic since it has spread to countries all over the world. Since the first cases of COVID-19 were discovered, the number of fatalities has steadily increased due to the virus's lethality, prompting the WHO to declare COVID-19 a global pandemic. Due to the virus's high infectious and death rates, governments across the world have adopted a variety of policies aimed at mitigating the spread and effect of it. These actions began with the Chinese government placing an order to quarantine the people of Wuhan from January 23rd, 2020 after which, multiple countries (e.g., US, Italy, Argentina, Spain, Canada) followed suit by declaring state of emergency and implementing strict quarantine and social distancing [66].

To tackle the spread of COVID-19, most political officials/government leaders have adopted policies that promote and, in some cases, enforce, "social distancing." As a result of these policies, entertainment events have been canceled, schools and universities have closed, and companies have reduced their hours of service, implemented telecommuting, or closed entirely. Without a question, the pandemic and the steps put in place to combat it have had and may continue to have a significant effect on the lives of millions of people without effective vaccination. However, given

the unprecedented nature of the pandemic and the reactions to it, the public perception towards it is not clear and we are likely to be shocked by how people respond [66].

1.2 Motivation

To control the infection, Public health authorities have recommended a variety of behavioural precautions, such as social distancing and personal hygiene practices. These social and behavioural containment interventions are deemed to be successful in preventing COVID-19 cases from growing exponentially but this has also resulted to a lot of change [59]. Hence it is important to identify and understand people's perception towards the pandemic as this can support public health authorities to effectively resolve legitimate concerns and lead to the development of more user-centered interventions to help mitigate the spread of the virus.

The most important factor in promoting health is each individual's beliefs, values, tendencies, and habits which reinforces the right or wrong behaviors. To understand public perceptions and reactions, theories of human behaviors provide a conceptual framework for understanding the factors that influence specific behaviors and how they influence the behavior. The health belief model (HBM), Social Norm Theory and Trust are amongst the theories and models proposed by sociologists, psychologists, and anthropologists to explain the factors affecting health behaviors. Previous studies have verified the application of these three theories and models to understand and explain health behaviors that are effective in the area health education and promotion. The health belief model (HBM), Social Norm Theory and Trust are amongst the most popular behavior change theories that have been widely adopted and are used to explain health behaviors[39].

The HBM suggests that the following constructs play a major role in motivating people to engage in preventive health behaviors such as social distancing, hygiene, and other COVID-19 precautionary measures: *perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy*[51]. The general acceptance and popularity of the health belief model is due to its high predictive power which is designed to explain the reasons why people may or may not participate in the prevention health behavior. Specifically, the HBM proposes that an individuals' preventive behavior is affected by their beliefs in being at risk

(perceived susceptibility), the seriousness of risk (perceived severity), existence and effectiveness of a way to reduce the incidence or severity of disease (perceived benefits), and a person's feelings on the obstacles to performing a recommended health action(perceived barriers), and as a result, based on the assessment of these factors, they engage in screening and prevention activities [37][76][90].

The Social Norms Theory, also known as "mental representations of appropriate behaviour," can guide behaviour in specific circumstances or environments. Normative messages have been shown to encourage pro-social behavior, such as lowering alcohol intake, increasing voter turnout, and conserving energy. Social norms affect people's behavior: what they think others are doing, or what they perceive others approve or disapprove[26]. Previous research has identified a number of reasons for adhering to norms, including the ability to learn from others and obtain association or social acceptance [123][26]. Although people are conditioned by norms, their perceptions are often incorrect, for example People can underestimate health-promoting behaviors such as hand washing, while overestimating unhealthy behaviors[35][110].

During an outbreak, social networks can amplify the dissemination of both negative and beneficial behaviors, and these effects can spread to a network of connected friends (directly or indirectly)[24]. The virus spreads from person to person, since people who are centrally positioned in networks interact with more people, they are frequently the first to contract the virus[25]. However, by demonstrating constructive measures such as hand washing and physical distancing to a broader variety of people, these same central people could be instrumental in slowing the pandemic. According to some studies, a greater proportion of interventions may be attributed to have indirect effects on people who copy the actions of those who are involved in the intervention [10].

Trust is crucial during an infectious disease outbreaks, as shown by studies on occurrence such as SARS (2002–2003), swine flu (2009–2010), and Ebola (2014–2016)[101] [18][17][59]. Trust building could influence perceived severity and transparency, also willingness to adopt interventions such as physical distancing and personal hygiene [17]. Baruch Fischhoff mentioned

in a National Academy of Sciences workshop on infectious disease threats, "…trust building activities can enable organizations to get out in front and stay ahead of problems"[7].

Several characteristics of the COVID19 pandemic make building trust especially difficult; one of them is the uncertainty as COVID-19 is a new, invisible, and unfamiliar threat, which, as Paul Slovic comments[7], 'hits all the hot buttons that lead to heightened risk perception'. The rapid nature of the outbreak coupled with ongoing uncertainties create significant complications that hamper trust. Questions about the incubation time, infectivity before symptoms, seasonal dimensions, disease specificity for certain population groups, re-infection rates, and, perhaps most significantly, mortality rates are part of the uncertainties[7]. A more detailed explanation of these behaviour theories are contained in chapter2.

It is important to gain a deeper understanding of public perceptions of coronavirus in order to tailor educational effort. Understanding these perceptions over time will provide comprehensive data that can be used to develop and improve tailored public health interventions. The Confederacy of International Health Education Organizations (CNHEO) declared that in the design, implementation, and evaluation of health education initiatives, theory and technology for educational research should be used as these theories and models could be useful tools for health education experts [87].

Social media is ablaze with discussions around the pandemic and these conversations are often characterised by a number of traits suggesting that those utterances are worthy of closer examination. It is possible to determine public perceptions of strange physical distancing and personal hygiene by mining related content from social media sites. Traditional surveys have drawbacks, such as resource costs and the difficulties of tracking changes in real-time, but social media sites provide a unique opportunity to examine unfiltered opinions, messages and conversations of vast audiences while mitigating these limitations. In recent years, there has been a significant increase in the global adoption of personal communication technologies. This has been made possible in large part by the widespread availability of social media (SM), which has been aided by the rise in cell phone ownership. SM has proved to be an important communication tool in a number of fields, including education, marketing, and health communication. For example,

in the field of health communication, the US Centers for Disease Control and Prevention (CDC) and local health departments in the United States have used Twitter (a social media application) to communicate to people during epidemics. Another example is the United Kingdom and Norway, where health officials used Twitter to keep people informed about the West African Ebola epidemic in 2014 and 2015. The use of social media in health care will increase the quality of communication by speeding up interactions between health care agencies, providers, and patients [71].

People have expressed their opinions and shared knowledge, including misinformation, about the virus on social media sites such as Twitter since the early stages of COVID-19. As governments attempt to minimize its effect by implementing counter measures, people have used social media outlets to voice their opinions about the precautionary measures, the leaders enforcing them, and the ways their lives are transforming. Because of the "social distancing" initiatives put in place to mitigate it, the use of social media, such as Twitter, as forums to express opinions and exchange knowledge about COVID-19 will only gain more users as a result of people being at home more often than usual [66].

It is virtually impossible for public health officials to manually review social media posts on a daily basis. Machine learning approaches such as text classification or categorization may help with automated textual content analysis. Such tools can be used to automatically classify large amounts of social media content for real-time analysis, allowing public health officials to assess how well their health messages and safety measures are being received.

1.3 Solution

To understand the public perception, we employed health behavior theories (HBM, social norm and Trust) and machine learning models to explain/examine people's behavior and reactions towards COVID-19. To achieve this goal, we used a text-mining approach to collect public views (comments) on the COVID-19 pandemic from social media posts, where the data was mined from twitter. Using three (3) supervised ML models, we built the text classification models to classify these comments based on the eight adapted health behaviour constructs to understand and predict

the general behaviours of the public during the pandemic. Thematic analysis was then performed to analyze each construct comments to bring out meaningful themes.

1.4 Contributions

The thesis made three major contributions:

- Experimenting with traditional ML models on COVID-19 data (Evaluating with parametric and non – parametric models) to see the best performing model for classifying COVID-19 comments.
- Showing the possibility of using the six HBM constructs and two other constructs from other theories (Social Norm and Trust) making a total of eight constructs for classifying COVID-19 tweets.
- Conducting thematic analysis on the classified data to understand public opinions.

1.5 Research Questions

The following are the research questions that were investigated in this study:

RQ1: What is the best ML algorithm for classifying COVID-19 related tweets into their appropriate constructs?

RQ2: What insight can be derived from COVID-19 related tweets classified using the HBM, Social Norm and Trust.

1.6 Overview of Thesis

A detailed description of this study is explained from chapter 1 to 5.

CHAPTER 1 INTRODUCTION: This chapter introduces the thesis. It states the problem as well as the issues concerning the problem discussed in the thesis.

CHAPTER 2 RESEARCH BACKGROUND: This chapter contains a review of research related to this thesis. It presents a review of the HBM, Social Norm, Trust and text classification.

CHAPTER 3 RESEARCH METHODOLOGY: This chapter describes the steps taken to understand and predict public behaviour from social media comment using machine learning techniques based on the health behaviour theories and performed thematic analysis on the classified comments.

CHAPTER 4 RESULT AND DISCUSSION: This chapter contains details about machine learning result and discussion on the thematic analysis.

CHAPTER 5 CONCLUSION AND FUTURE WORK: This chapter summarizes the entire work and presents the limitations and future research directions.

CHAPTER 2 RESEARCH BACKGROUND

This chapter puts forward a detailed review of the related literature to elaborate the ideas presented in the introduction chapter. The literature review, in general, explains the health belief model, Social Norm and Trust in detail. We began by first discussing the Health Behaviour Theories adapted in our research. The Health Belief Model is first discussed followed by social norm and trust. Next, we discussed various studies that applied text mining approaches and lastly the ML models chosen for text classification in this study.

2.1 The Health Belief Model Theory

Though there are many theories about behaviour and behaviour change, the health belief model is one of the most well-researched and commonly applied theories about health-related behaviours[64]. The health belief model (HBM) is a theoretical model that focuses on the role of social and psychological factors in deciding health-related behaviours[52]. The model was established in the 1950s by a group of social psychologists named; Rosenstock, Hochbaum and Kegeles from the US Public Health Service, who wanted to know why certain people refused to participate in preventive health care services like immunization and tuberculosis screening, which could help with early disease detection and prevention [91][54]. In comparison to other theories exploring behavior modification or change, the HBM includes a belief component, an attitude component and a behavior component. The belief component is concerned with what the individual perceives to be the true condition, while the attitude component is concerned with how the individual feels about it. These two components function together to motivate the person to behave in a certain way[109].

Their original purpose was to concentrate on the efforts of individuals who tried to enhance public health by understanding why preventive health measures were not adapted. People's beliefs, such as perceived health benefits, barriers to practise, and self-efficacy, affect their dedication to health-promoting behaviours[65]. The HBM emphasises cognitive elements and the actions of people

also relies upon their rational expectations from a cognitive standpoint. During the SARS and H1N1 pandemics, empirical studies using HBMs investigated changes in health behaviour. According to these studies, people are more likely to adopt the recommended behaviour when they are convinced of the severity of the disease, believe they are highly vulnerable to it, are assured that a preventive behaviour is safe, and believe the costs of doing so are minimal [16][81][40]. Over the years, the model has been updated and extended to include a self-efficacy component based on Albert Bandura's research and a cues to action component, and it has been widely used by social science researchers to understand and predict health-related behaviours[97].

The model's six main components are cognitive-based, specifying particular factors that must be considered by a person who believes that he or she is healthy when determining whether or not to follow the recommended health behaviour. The HBM contains primary concepts that predict whether or not people can take measures to prevent, test for, or manage disease conditions. These concepts are: *Perceived susceptibility, Perceived Severity, Perceived Benefits and Perceived Barriers* to a behavior*, cues to action,* and *self-efficacy*[55]. The HBM Constructs are described below;

Perceived threat: This is the combination of perceived susceptibility and perceived severity of a health condition.

- **Perceived Susceptibility:** This term refers to beliefs about the likelihood of contracting a disease or condition. For example, A woman, for example, must believe she is at risk for breast cancer before she would consider having a mammogram.
- **Perceived Severity:** Feelings regarding the seriousness of the disease or of leaving it untreated require assessments of both medical and health effects (e.g., mortality, disability, and pain) and potential social consequences (e.g., impacts on working conditions, family life, and social relations).

The combination of susceptibility and severity was classified as perceived threat[55].

Perceived Benefits: If an individual perceives personal susceptibility to a serious health condition (perceived threat), the person's beliefs about the perceived benefits of the numerous available actions for reducing the disease threat will affect if this perception contributes to behaviour change. Other perceptions not related to health, such as a financial saving related to stopping smoking or impress a family member with a mammogram, can also influence behavioural decisions. Individuals with optimal beliefs in susceptibility and severity are unlikely to support any recommended health intervention unless they also believe the action has the ability to reduce the threat[55].

Perceived Barriers: The possible negative aspects of a particular health action that may function as a hindrance to the undertaking of recommended behaviour is referred to as perceived barriers. Individuals weigh the expected benefits of an intervention against perceived barriers in a kind of nonconscious, cost-benefit analysis. Example of perceived barrier; "It could help me, but it may be expensive, have negative side effects, be unpleasant, inconvenient, or time-consuming." Thus, "combined levels of susceptibility and severity provide the energy or force to act and the perception of benefits (minus barriers) provide a preferred path of action"[55].

Cues to Action: The concept of cues that can trigger actions was included in early formulations of the HBM. For example, Hochbaum [114] believed that readiness to act could only be amplified by other factors (perceived susceptibility and perceived benefits), especially cues to initiate action, such as bodily events, or environmental events, such as media coverage. Although the concept of cues as triggering mechanisms is appealing, cues to action are difficult to research in explanatory surveys; a cue can be as fleeting as a sneeze or the barely conscious perception of a poster [114].

Self-Efficacy: Self-efficacy is defined as "the conviction that one can successfully execute the behavior required to produce the outcomes"[9]. Bandura[63] distinguished between self-efficacy and outcome expectations, which are a person's estimate that a specific behaviour will result in specific outcomes. Expected results are similar to the HBM definition of perceived benefits, but they are not the same. Self-efficacy was introduced as a separate construct to the HBM by Rosenstock, Strecher, and Becker in 1988, along with the original concepts of susceptibility, severity, benefits, and barriers [55].

People must feel challenged by their current behavioral habits (perceived susceptibility and severity) and assume that improvement of a specific kind will result in a valued outcome at an acceptable cost (perceived benefit), according to the original HBM theory. They must also believe in their own competence (self-efficacy) in order to overcome perceived barriers to take action[50][51].

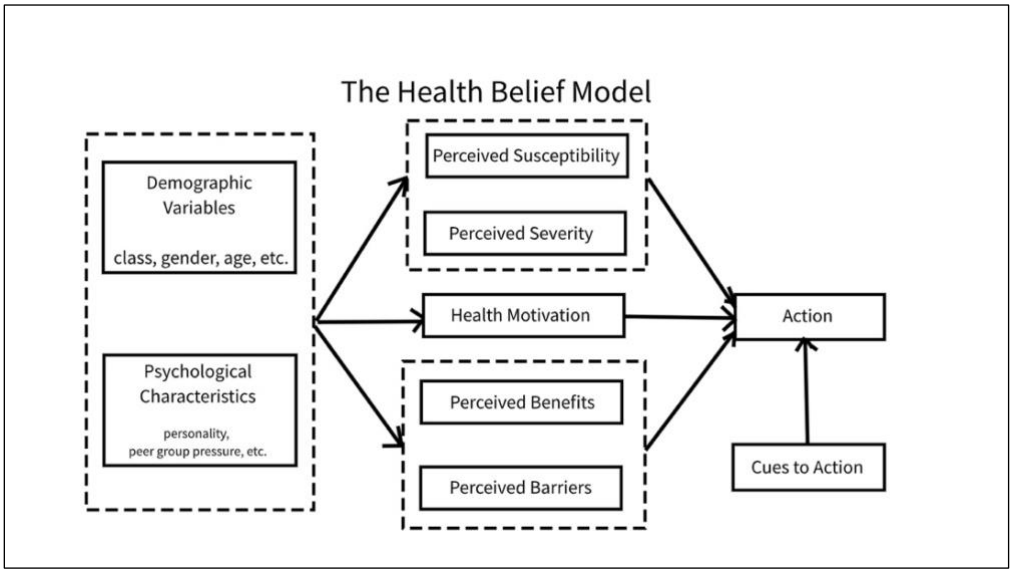

Figure 1 Health Belief Model Constructs [51]

Figure 1 shows the components of the HBM. Relationships between constructs are indicated by arrows. Knowledge and sociodemographic factors that may influence health perceptions are examples of modifying factors. The major constructs of the HBM are susceptibility, severity, benefits, barriers, and self-efficacy. Modifying factors, as well as cues to action, have an impact on these perceptions. Behaviour results from a combination of beliefs. Perceived susceptibility and severity are combined in the "health belief" box to identify threat. Although the HBM defines constructs that contribute to outcome behaviours, there are no established relationships between or among these constructs. The complexity has resulted in a wide range of HBM applications. For example, while many studies have tried to establish each of the major dimensions as being independent, others have tried multiplicative approaches and examining interactions between

constructs. To improve the utility of the HBM in predicting behaviour, analytical approaches to defining these relationships are needed. [51].

2.1.1 Applications of the HBM

The Health belief model has been applied in different aspects of health to understand or predict behaviour. This section review studies that adopted HBM using surveys, text mining and machine learning for analysis.

2.1.1.1 Applications of HBM using Survey

There are several approaches that has applied HBM using survey data to understand public opinion. In this section, I will briefly go over the papers that used these approaches.

The HBM has been used to investigate people's attitudes and behaviors in the aftermath of the 2009 H1N1 (Hemagglutinin type 1 and Neuraminidase type 1) influenza pandemic. Using the Health Belief Model, some studies, such as Fridman et al.[44], determined factors that influence a pregnant woman's acceptance of the H1N1 vaccine (HBM). During the 2009 H1N1 epidemic, a self-administered questionnaire based on the HBM was used in a cross-sectional study of postpartum women. Perceived vaccination barriers and perceived infection severity were independent predictors of vaccination.[44]. Based on the Health Belief Model, Mirkhalili et al.[89] sought to find predictors of influenza A (H1N1) preventive behaviors among Jiroft residents. In 2016, a descriptive cross-sectional analysis was performed on 400 Jiroft citizens. Cluster sampling was used to pick samples, and data was collected using an HBM-based questionnaire created by the researcher. The Pearson correlation coefficient and linear regression were used to analyze the data. Behavior and knowledge ($r=0.206$, $p=0.001$), benefits ($r=0.308$, $p=0.001$), susceptibility ($r=0.130$, $p=0.009$), and perceived severity ($r=0.248$, $p=0.001$) all had a significant positive correlation. The Health Belief Model constructs predicted 15% of the variance in H1N1 influenza preventive behaviors, with perceived benefits ($=0.233$) being the most powerful predictor [89].

The HBM has also been extensively used to study vaccination beliefs and behavioral patterns to identify patients' perception of disease and vaccination [69][57][6][79]. Donadiki et al. [38] used an investigative method to examine the key reasons for HPV (Human papillomavirus) vaccine rejection, as well as participants' expectations and attitudes toward HBM constructs (perceived susceptibility, perceived severity, perceived advantages, perceived barriers, cues to action, and self-efficacy) among a sample group of female university students in Athens using a self-allotted questionnaire. According to the study's findings, participants with negative attitudes toward vaccination, had a lack of knowledge about the HPV vaccine and its cost, and myths about HPV vaccination were less likely to be vaccinated [38]. Mehta et al. [72] have developed and tested a Health Belief Model (HBM)-based intervention to boost vaccination rates among college students. HBM-based intervention was compared to a traditional knowledge-based intervention in 90 college-aged men ages 18–25 years in a randomized controlled trial. The intervention group showed substantial positive improvements in information and HBM constructs, as measured by repeated measures of ANOVA (analysis of variance). The pretest/posttest regression analysis revealed that selfefficacy for receiving the vaccine ($p = 0.000$), perceived barriers ($p = 0.007$), and perceived severity ($p = 0.004$) were all significant positive predictors of vaccine acceptability [72].

As a result of the new viral disease COVID-19 pandemic, the HBM has been used in a number of studies to examine preventive behavior [56][41][77][122]. In March 2019, Shahnazi et al. [59] conducted a cross-sectional study on 750 people to determine COVID-19 preventive health behaviors based on the health belief model among people in Golestan Province. The results showed that female gender, perceived barriers, perceived self-efficacy, fatalistic beliefs, perceived interests, and city living had the highest COVID-19 preventive behaviors. Males and villagers required preventive interventions.[94]. Tong et al. also conducted a study to explore strategies for promoting adherence to COVID19 precautionary measures using the health belief model (HBM) and generalized social beliefs (i.e. social axioms). In April 2020, they conducted a telephone survey using a twostep stratified random sampling method, obtaining a probability sample of 616 adults in Macao, China (18–87 years old; 60.9 percent women). The findings revealed that participants adhered to some COVID19 precautionary measures (e.g., wearing a face mask; 96.4 percent) but not others (e.g. social distancing; 42.3 percent)[59].

Based on the health belief model, Walrave et al. [116] looked into the factors that influence app use intention. Furthermore, correlations between respondents' news consumption and their health status were investigated. In a survey of 1500 people aged 18 to 64 in Flanders, Belgium, the relationships between the model's constructs were investigated using structural equation modeling. In total, 48.70% (n=730) of respondents stated that they plan to use a COVID-19 tracing app. [116].

These studies showed that HBM can be used to understand adherence to COVID19 precautionary measures, as it has being used to understand beliefs and behavioral patterns to understand public perception towards diseases.

2.1.1.2 Applications of HBM to Social Media data using Text Mining or ML

During my research, I have found that not much research has applied HBM to social media data using text mining and ML approaches.

HBM using Text Mining

In relation to HBM with Text mining, To the best of our knowledge only one (1) study was found which applied text mining and health belief model to analyze social media data.

Using the Health Belief Model and quantitative content analysis, P.D et al analyzed Instagram posts about Zika. They found out that threat messages were abundant, but there was little engagement with the posts. Perceived severity (75.8%) and perceived susceptibility (59.9%) were also found to be the most frequently occurring HBM constructs, meaning that posts represent a high degree of perceived threat (perceived severity and susceptibility). Threat as the major factor which may influence people's response to the Zika virus [49]. This study shows how text mining was successfully applied with social media data using the health belief model.

HBM using Text Mining and Machine Learning

Only a few studies have applied machine learning and health belief model to analyze social media data. I will go over the works that relate to this task.

The Health Belief Model was applied to the human papillomavirus (HPV) vaccine on Tweets by Du et al. and Shapiro et al. [39][95]. Du et al. classified health beliefs by manually categorizing a subset of 6000 tweets based on their relevance to the HBM constructs using the key four (4) HBM constructs (perceived susceptibility, perceived severity, perceived benefits, and perceived barriers). The performance was measured in terms of sensitivity, precision, and accuracy. On testing sets, the models had satisfactory results in terms of sensitivity, specificity, and accuracy, with a mean accuracy of 80.50 % for identifying HBM-related tweets and between 80.33 % and 89.82 % for the four HBM constructs. Shapiro et al. [95] used data mining and machine learning techniques to conduct an international comparison of English language tweets about HPV vaccines and social connection amongst twitter users posting about the vaccines and the HBM was used as a basis for coding the types of concerns expressed on twitter with regards to the HPV vaccine. Concerns about HPV vaccines were expressed in 14.9 % of tweets in Canada, 19.4 % in Australia, and 22.6 % in the United Kingdom [39][95].

To quantify health beliefs, Raamkumar et al. and Wang et al. employed the health belief model (HBM) with four constructs namely: perceived susceptibility, perceived severity, perceived benefits, and perceived barriers to classify social media content using COVID-19 data with deep learning and ML classifiers and produced reasonable results. In Wang et al study, the HBM text classification models achieved mean accuracy rates of 0.92, 0.95, 0.91, and 0.94 for the constructs of perceived susceptibility, perceived severity, perceived benefits, and perceived barriers, respectively and In Raamkumar et al study the classification of all four HBM constructs performed over 0.86 [117][85].

2.2 Social Norm Theory

Social norms are principal determinants of health-related habits, and they appear in a number of prominent psychological health behavior theories. Social norms are commonly described as "rules and standards that are understood by members of a group and that guide or constrain social behaviors without the force of law," and they often relate to a perceived social pressure to participate or not engage in those behaviors [2]. Individuals' perceptions of normative behaviours

are used to guide behavioural patterns and intentions in social norms, which can also be based on direct and explicit interaction between group members [98]. The role of perceived normative peer behaviors and attitudes has emerged as a key predictor of health behaviors, despite the fact that social norms can be conceptualized in a number of ways [83]. The public actions of a small number of individuals (e.g., students visibly drunk), media coverage, discussion of such extreme minority behaviors, and the highlighting of such extremes in daily conversations are all factors that lead to such misperceptions. The scope and prevalence of negative health behaviors in the minority are exaggerated as a result of these factors, while the behaviors of the stable majority are overlooked [33]. For example, social norms theory has been widely applied to comprehend the reality of female genital cutting (FGC), a non-medically justified modification of women's genitalia that poses a global health threat to 140 million women and girls[115]. Existing programme implementations that centred on social norms around FGC offered valuable insights into the potential of addressing social norms for social change, implying that community-based initiatives can be effective in promoting behavioural change when they effectively incorporate a social environment-focused approach [36][28][75][107]. It was found to be successful in improving people's health-related behaviours because it included a social norms aspect within an intervention that also addressed people's individual attitudes and awareness, local institutional policies, political accountability, and community members' economic conditions [28][29].

In the context of the COVID-19 pandemic, recent research has attempted to identify the underlying drivers of social norm enforcement. In a study conducted in the Netherlands, Kuiper et al. [73]discovered that people who indicated higher levels of impulsivity were more likely to break physical distancing laws, and moral convictions about policy enforcement played a significant role in behavioral changes[73]. M¨uller and Rau[120] investigated the role of economic preferences on individuals' COVID-19 behavior. They report a negative association between risk preferences, complying to physical distancing (avoiding crowds) and panic buying in a sample study of German students [120].

2.3 Trust

The need for transparency from institutional and departmental leadership is intensified in times of crisis, such as during the coronavirus disease 2019 (COVID-19) pandemic. Due to the obvious confusion surrounding the virus's etiology and consequences, a cacophony of voices arose, with institutional correspondence often misaligned with media reports and an indistinguishable mix of misinformation, unverified rumors, and deliberately manipulated disinformation[104]. These rumors and hoaxes spread quickly, disrupting the communication ecosystems' authenticity balance. Furthermore, the most common argument concerned government initiatives or measures to combat the spread of Covid-19, suggesting that health agencies and governments had not sufficiently succeeded in providing accurate information in response to public[104].

According to cheng et al.[20], trust can take several forms and concentrate on various things depending on the situation. Barber argues[53] that trust is important in democratic societies because it helps to minimize social complexity. "The likelihood that someone will take an action that is beneficial or at least not harmful to us is high enough for us to consider participating in some kind of cooperation with them," says one description of trust. Simply put, trust is a reflection of uncertainty and expectation among various parties. There has been a substantial increase in interest in trust. Ongoing societal changes, described as late modernity and post-modernity, have prompted an increase in interest in trust. One of many social constructs, trust is a component of social truth. Relationships between social actors, both individuals and groups, are naturally associated with trust. Since trust is a social construct, it's fair to question if it can be trusted, i.e., if social trust works as anticipated[53].Trust can have a positive effect on a person 's behaviour and performances in addition to the social influence[70] [7]. In Devine et al.[34] study in relation to the COVID-19 pandemic, Trust was shown to be crucial for the recovery of the current crisis, and it is influenced in complicated ways by policy responses. Governments will also need to restore confidence in what will most likely be a somewhat different policy environment both domestically and globally after the crisis. They also stated that in the future, understanding the dynamics of trust, how it encourages and hinders policy responses, as well as the possible effects of these responses on trust, will be critical questions in policy and trust research[34].

2.4 Applications of Text mining to COVID19

Text mining technology is the use of computer technology to process vast amounts of text data in order to reveal useful information from further analysis [126]. In the biomedical field, text mining has been widely used to discover new knowledge from search problems [127] [58]. Several articles have been published that analyze COVID-19 social media data, and there are several research papers that address the use of social media data and the valuable contribution of information in the field of public health. Most of the research are on sentiment analysis, topic modeling or text mining. Some studies are summarized in this section.

Matteo et al. [27] conducted a comparative study of users' behavior on five separate social media platforms during the COVID-19 health emergency (Twitter, Instagram, YouTube, Reddit and Gab). They also evaluated user engagement using NLP and interest in the COVID-19 topic, and also the discourse's evolution over time. [27]. Elizabeth et al. used a text analytic method to classify topics and extract meanings from unstructured textual data [47]. From March, when the first case was announced, to June, when the normalization process began (18 March- 28 May), Gokhan et al. [99]looked at how society responded to the epidemic. In order to understand these reactions, a total of 567,018 texts using the hashtag #StayHome on the Twitter platform was fetched and analyzed. To make sense of what society has talked about, Sensitivity analysis has been used to see the weekly reactions of individuals that differentiate their positive/negative moods and hope levels. As a result, some differences have been observed in the emotions analyzed in two-week periods[99].

Using a Text Mining technique, Josimar et al.[23] studied the effect of coronavirus in the populations of Paris, Mexico, Brazil, and South America. From April 23 to June 18, data was analyzed and visualized using a cloud of words and a bar chart. The first is related to the health crisis and economic impact caused by the coronavirus, and the second is related to the health crisis and economic impact caused by the coronavirus[23]. In Brazil, the Research Foundation of the State of Sao Paulo (Fapesp) published a dataset in which a data mining technique was used to conduct an exploratory study of the datasets. There are some inconsistencies discovered, such as NaN values, null reference values for analytes, outliers in analyte results, and encoding problems.

The results were cleaned datasets for future studies, but at least 20% of it was discarded due to non-numerical, null values, and numbers outside of the reference range [92]. Josimar has used Complex Network Representation and Text Mining to examine the interaction of South American countries on Twitter and characterize the flow of data by said users. The experiments introduced the idea of existence of patterns, similar to Complex Systems, and the presence of users' groups publishing and communicating together during the time, with the possibility of identifying robots continuously sending posts [21]. Another research by the group used a text mining method to examine the publications of people in Mexico in order to better understand how a public health emergency of international significance plays out in social media and the impact of the spread of Covid-19 on society [22].

Stefanos et al. conducted another study that used the BERTmoticon model for multilingual emoji prediction [68] to better understand how Twitter users reacted emotionally to news about the coronavirus. The model can predict emojis for text written in 102 different languages. When the World Health Organization (WHO) declared coronavirus a global pandemic, there was a spike in sadness, followed by an uptick in outrage and disgust when the number of COVID-19-related deaths in the United States exceeded 100,000 [105].

Other contributions, such as the works of McAlaney et al.[68] and Boon-itt et al.[14] identified the topics and their overarching themes that emerged from the public COVID-19-related discussion. The most popular tweet topics, as well as clusters and patterns, were identified using a natural language processing approach and the latent Dirichlet allocation algorithm [68] analyzed discussion about the chagas disease and the data generated 16 topics that the public linked to the chagas disease when they talked about COVID-19 Keyword frequency, sentiment analysis, and topic modeling were used to identify and explore discussion topics over time. While Boon-itt et al.[14] analyses included frequency of keywords, sentiment analysis, and topic modeling to identify and explore discussion topics over time. The results of the sentiment analysis showed that people have negative outlook toward COVID-19 and based on topic modeling, the themes relating to COVID-19 and the outbreak were divided into three categories: the COVID-19 pandemic emergency, how to control COVID-19 and COVID-19 news.

2.5 Applications of Machine Learning Classifiers

Classification is the problem of determining which of a set of categories (classes) a new instance belongs to in machine learning. The problem is known as multiclass classification when the groups contain more than two different labels[108][46]. In the basic scenario of multiclass classification it is presumed that:

1. Each instance receives only one class label. In other words, this is a single-class classification as opposed to multi-label classification, which allows for multiple class labels per instance [108].
2. The class labels are independent, that is, there is no relationship between the class labels, in contrast to hierarchical classification, which aims to solve classification problems by organising classes into a hierarchical structure[46].

In this section we reviewed three ML classifiers (SVM, Logistic regression and decision tree) chosen from past work that has been applied in several studies.

2.5.1 Decision Tree

Decision trees (DT) [1] are effective algorithms that have been used in a wide range of studies. Some research are discussed in this section. Abdar et al [1] applied decision tree method to produce basic and understandable rules for a heart disease dataset. The obtained results showed that decision trees could generate simple rules among the dataset with high performance of 85.33%. The obtained results revealed that decision trees had a high output of 85.33 percent in generating simple rules from the dataset. Christopher et al. [15] created a basic decision aid for clinicians interpretingCOVID-19 suspect blood tests using. This study showed that using blood test analysis and machine learning (decision tree algorithm) as an alternative to transcription polymerase chain reaction (rRT-PCR) for identifying COVID-19 positive patients is feasible and clinically sound. The sensitivity was between 92 percent and 95 percent, and the accuracy was between 82 percent and 86 percent [15]. It has also been used to screen for COVID-19 using radiomics features derived from chest CT images. By segmenting whole lung volume and extracting 86 radiomics features, this study was able to differentiate pneumonia caused by COVID-19 from suspected cases.

Decision tree and logistic regression were part of the models applied in this study and the both had an accuracy of 0.976 [88].

2.5.2 Logistic regression

Logistic regression has also been applied in numerous medical data classification tasks. Some works are discussed below; Wen and Rose[119]used logistic regression to examine the actions and disease trajectories of breast cancer patients over time and successfully interpreted the results. Biyani et al. [11]used logistic regression to classify the sentiment data in a cancer survivor network, with a f1 score of 78%. The logistic regression model has also been used in COVID-19 studies, such as predicting epidemic trends[118]forecasting the spread of in Kuwait using compartmental and logistic regression models were they were both successful[3].

2.5.3 Support Vector Machine

The support vector machine has also grown in popularity as a tool for machine learning tasks such as classification, regression, and novelty detection. Several studies has successfully developed SVM models, Few studies are discussed in this section. Aramaki et al[5] proposed a method that uses Twitter API to extract influenza-related tweets, Afterwards the support vector machine (SVM) classifier extracting tweets that mention actual influenza patients. The experiment results show that the suggested solution is feasible, as it got a high score of 89 percent. SVM was used by Jung son et al. [103] to find predictors of drug adherence in HF patients. Leave-one-out cross-validation (LOOCV) was used on the data to assess the robustness of the predictions produced with the SVM models, with the highest detection accuracy from the assessment being 77.63 percent. To detect coronavirus infected patients using X-ray images, sethy[61] et al used a deep feature plus support vector machine (SVM) based methodology. The SVM distinguishes between corona-affected X-ray images and those that are not, such as COVID-19, pneumonia, and standard. SVM is tested for COVID-19 detection using the deep features of 13 different CNN (convolutional neural network) models. Using ResNet50's (Residual Networks) deep functionality, the SVM achieved the best performance. ResNet50 plus SVM has the highest accuracy of 98.66 percent.

2.6 Summary

In this chapter, we discussed health behavior theories and text classification, as well as related research that motivated the approach. First, we discussed the health behavior theories that were used in this study. Secondly, we went over the studies that applied the Health Behaviour Theories and results that was achieved using their approach. The Health Behaviour Theories reviews shows how it has been applied in a wide variety of health issue and also previous epidemics which has been successful in understanding public perceptions and creating solutions. Afterwards the classification algorithms that are used for text data classification were discussed in the next step. We also covered the research that provided insights about several works done on COVID19 data using text mining and machine learning approaches, which shows different approaches that have been taken for this purpose, such as using sentiment analysis to classify social media comments or topic modeling (Latent Dirichlet Allocation) to discover topics. Finally, I used them as an incentive to propose my approach as discussed in chapter 3.

From our literature review, most research that adopted behaviour theories used surveys to analyze public opinions which can be limiting in comparison to using social media data. Also, the few existing literatures that studied the impact of coronavirus using social media data did not use the full HBM theory and no further analysis was done with the data. In our study, we are adopting the full theory in addition to two other theories which were also mostly analyzed using surveys without considering the opinions of people on social media and according to research, it has been reported that people share their opinions, feelings or emotion on social media, which is also a larger platform in an unfiltered and unbiased manner.

CHAPTER 3 METHODOLOGY

The aim of this study is to use the health behaviour theories and machine learning models to explain/examine people's behaviour and reactions towards COVID-19. To achieve this aim, we employed a text-mining approach to collect public views (comments) on the COVID-19 pandemic from social media posts. Using supervised machine learning, we classified these comments based on eight adapted health behaviour constructs to understand and predict the general behaviours of the public during the pandemic. Figure 2 shows a general overview of the research method and below is a summarized list of steps we employed:

1. We created scripts to extract COVID-19-related user comments from twitter.
2. We then identified relevant health behaviour theories to classify our data. Based on research we choose the Health Belief Model (HBM), Social Norm and Trust. The HBM have six constructs. Social Norm and Trust theories can also be called constructs, making it a total number of eight constructs applied in this study.
3. We adapted the health behaviour constructs to COVID-19 context where each construct was defined in relation to COVID-19.
4. We randomly selected sample comments per construct based on the adapted definition of the health behaviour theories.
5. We then manually extracted meaningful key phrases from the data for each construct.
6. We preprocessed the data using natural language processing techniques.
7. We formed the ground truth dataset by automatically labeling the tweets using the generated candidate key phrases for both multiclass (single label) and multiclass-multilabel classification for machine learning classification.
8. We vectorized the ground truth data using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting technique before developing the machine learning classifiers.

9. We then performed multiclass (single label) and multiclass-multilabel classifications using three supervised machine learning models (trained on the vectorized data) for predicting the constructs.

10. We conducted thematic analysis on both the labelled and ML classified comments to extract meaningful themes that represents public perceptions towards COVID-19 pandemic using the eight constructs.

In the following sections, we present detailed descriptions of the steps we employed in achieving these classifications.

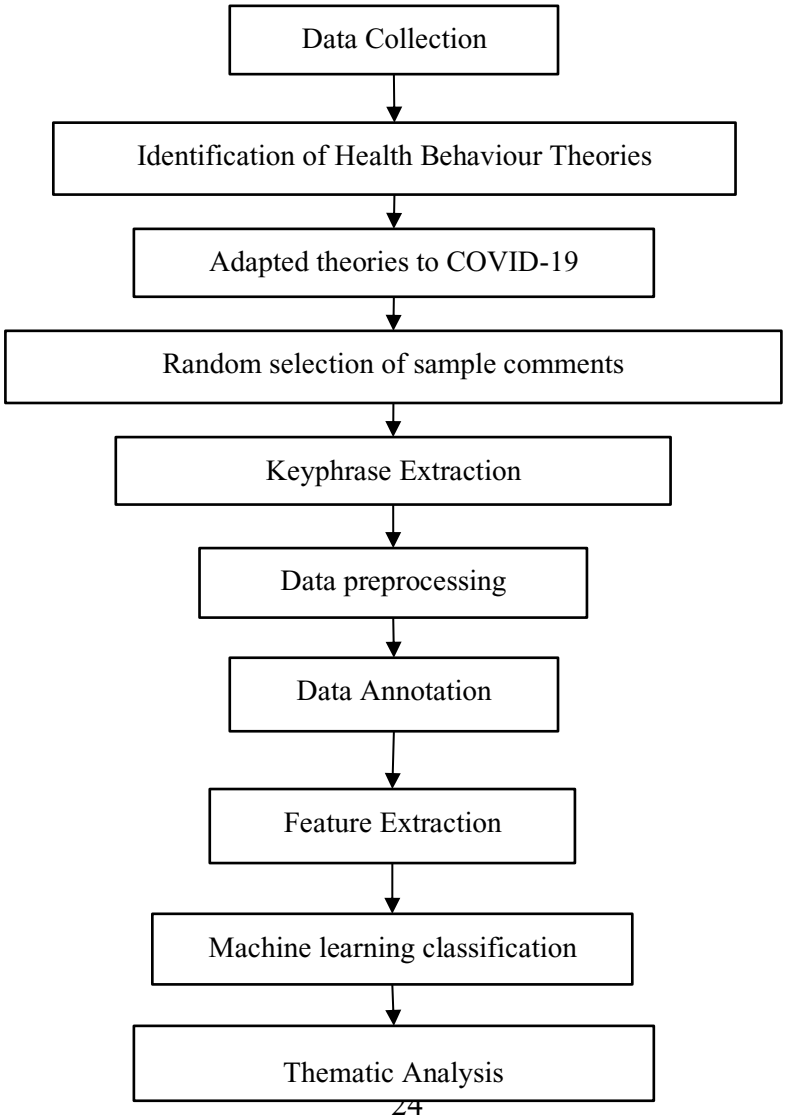

Figure 2 Overview of System Architecture

3.1 Data Collection

Data collection was the first step in the development process. Twitter was used to gather relevant data for this study as it is a widely used social media platform over a four(4) weeks period of time. Using the tweepy library to link to the Twitter API, the python scripting language was used to extract the data. Tweets with the following hashtags were used: #COVID19, #COVID, #ncov2019, #Covid_19. #CoronavirusPandemic, #Coronavid19, #Coronavirus. A total number of 5,562,550 was collected after the extraction process.

3.2 Identification of Health Behaviour Theories

We then identified the relevant health behaviour theories to be used in this study to label our data. Based on research we choose the Health Belief Model, Social Norm and Trust. The **HBM** was chosen because it is one of the most commonly used and well-researched models for explaining and predicting individual health behaviour changes[95]. It makes an attempt to predict health-related behaviour using specific belief patterns. Individual perceptions, modifying factors, and probability of action are the three categories that make up a person's motivation to engage in a healthy behaviour. Individual perceptions are factors that influence how people perceive disease and how important health is to them, as well as their perceived susceptibility and severity[82]. **Social Norm** was chosen because norms help to keep society in order, as they are "rules and standards that are understood by members of a group to guide or constrain social behaviors without the force of law,". They frequently relate to a perceived social pressure to participate or refrain from participating in those behaviours. Human beings require norms to direct and guide their behaviour, as well as to provide order and predictability in social relationships because they influence decisions[2]. Lastly previous research has shown **Trust** building could influence perceived severity and transparency, and also willingness to adopt interventions as it is one of the key factors for successful adoption of interventions. Which is important in this study as COVID-19 is a new, invisible, and unfamiliar threat[8]. The theories are explained more in the Literature Review.

Each theory has their own strength as the HBM focuses on attitudes and beliefs of individuals and shows the relationship between beliefs and behaviour, Social Norms guides and influences behaviour and Trust influence transparency which are different from other theories.

3.3 Adapted theories to COVID-19 Context (Health Behaviour Theories Definition)

In order to label our data to their appropriate health behaviour constructs. We first adapted the identified theories to COVID-19 context, where each construct was defined in relation to COVID-19. These definitions would help to determine what comments best fits each construct for labelling. The definitions for perceived susceptibility, perceived severity, perceived benefit and perceived barriers were adapted from a similar COVID-19 study [86], while the definitions of the other two(2) constructs from HBM, Social Norm and Trust were determined from their original definitions. Table 1 shows the definitions of the Health Behaviour constructs.

Table 1 Definition of the Health Behaviour constructs in relation to COVID-19

Construct	Definition
Perceived Susceptibility	Comments showing an assessment of the increased likelihood of infection with COVID-19, highlighting the increased local prevalence and high cases.
Perceived Severity	Comments indicating an assessment of the perceived seriousness and the consequences from getting infected by COVID-19. (e.g. Pneumonia, acute respiratory syndrome, kidney failure, hospitalization, highlighting mortality risk and death).
Perceived Benefits / Response efficacy	Comments supporting the precautionary measures (eg, school closure, working from home, of events cancellation, mass gatherings, wearing mask, sanitizing) to reduce the transmission of coronavirus disease.
Perceived Barriers	Comments portraying feelings of difficulties, discomfort, challenges, negative effects and perceived ineffectiveness of lockdown and other recommended precautionary measures. (e.g. loss of freedom, violation of individual rights, inconvenience, loss of income, boredom, anxiety etc.)

Perceived Self-Efficacy	Comments portraying the belief or confidence in one's ability to perform the preventive measures in order to mitigate the virus.
Cue to Action	Comments reminding users to engage in the preventive measures through awareness and reinforcement.
Social Norm	Comments indicating/demonstrating desirable(health-promoting) or non-desirable behaviours with regards to the recommended precautionary measures.
Trust	Comments exhibiting trust or distrust in the government, other organizations or public figures relating to COVID-19.

3.4 Random selection of sample comments

After defining the constructs in relation to COVID-19 context. An iterative process was carried out in selecting appropriate comments for each construct to enable us generate appropriate candidate keyphrases for labelling the comments to their corresponding Health behaviour constructs. We first randomly selected 30 comments per construct based on their definition and were reviewed together by five (5) reviewers. After careful review the best comments for each construct were selected, the process was carried out several times till all the reviewers reached an agreement of 98% for all the constructs. Table 2 shows sample comments of COVID-19 discussion on Twitter based on the health behaviour constructs.

Table 2 Reviewed Sample Comments

Construct	**Comments**
Perceived Susceptibility	*"new cases emerge across the country as us overtakes china and italy in number of coronavirus cases please stay inside this is real people."*
Perceived Severity	*"A 64-year-old man has died from complications resulting from #COVID19, the first death in #Brunei linked to the global #coronavirus pandemic."*
Perceived Benefits / Response efficacy	*"Everyone must now stay at home except in exceptional circumstances, to protect our NHS and save lives."*

Perceived Barriers	*"Lockdowns and school closures are affecting children's education, mental health and access to basic health services. For children on the move or living through conflicts, the consequences will be unlike any we have ever seen."*
Perceived Self-efficacy	*"I am an emergency physician. My increased exposure means that I have chosen to isolate from my family, to keep them safe. This is how I see my daughters (pictured with their cousins, with whom they're staying). If I can do this, you can stay home."*
Cue to Action	*"hey guys let us stay safe please practice social distancing or better still quarantine."*
Social Norm	*"this so called social distancing is a load of crap no one is observing it and while i am working people are hugging their friends and joking that we are all going to get covid anyway so what is the point shut the city down."*
Trust	*"never trust communists to keep their infection numbers low chinese hospitals are refusing to see potential coronavirus patients and instead forcibly removing them from the hospital premises they are lying to the world and only making things worse."*

3.5 Keyphrase Extraction

Now having a clearer understanding or structure of what type of comments best suits each construct, the data was manually analyzed to extract candidate keyphrases for each construct. This process was also an iterative process as it was reviewed several times by the same five reviewers in order to remove the weak keyphrases and pick out the best. The extracted candidate keyphrases were used to generate more keyphrases by using their synonyms from an online dictionary[19] based on the construct and syntax of the keyphrase.

3.6 Data Preprocessing

After extracting the candidate keyphrases, we carried out some pre-processing techniques to prepare the data for data annotation (data labeling) and ML Classification, using NLP techniques implemented in Python. Social media comments, in their raw form, are unstructured and cannot

be analyzed without proper processing to remove text elements that are irrelevant to the data. These irrelevant elements include punctuations, symbols, abbreviations etc. In text analysis, it is necessary to employ some preprocessing procedures (such as tokenization, removing punctuations etc.), to remove any textual elements that is irrelevant to the data context as this enables effective annotation of the comments and better performance in ML classification. Pre-processing data is also used to help the ML Classifiers understand the data samples better by making the data more consistent [60].

Figure 4 shows the preprocessing steps we employed. In the following section, we described each step as follows:

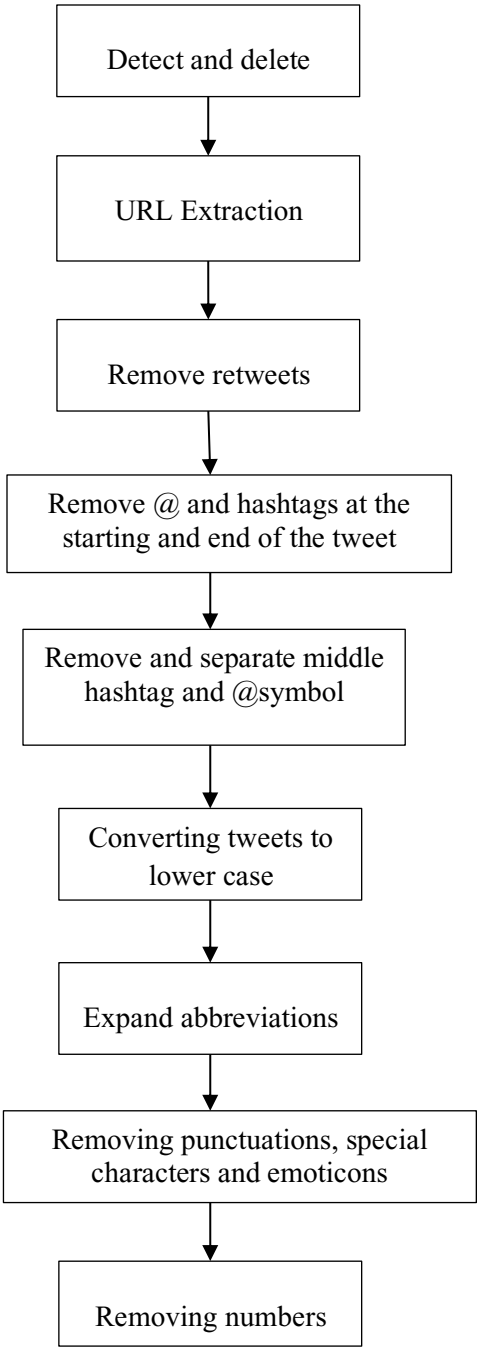

Figure 4 - Preprocessing Pipeline

3.6.1 Detect and delete language.

We used a Python library called Lang detect [62] to detect and delete non–English comments.

3.6.2 URL Extraction

We ran a python script to remove all URLs existing in the comments using regular expressions. This was carried out because URLs do not give any semantic knowledge in classifying data but can significantly reduce the feature size when performing machine learning. For example, text starting with HTTP://, https:// were removed [96].

3.6.3 Remove retweets

We also ran a script to remove all retweets from the data. "A Retweet is a re-posting of a Tweet" [48]. Most users include an 'RT' at the beginning of their tweets to indicate that they are reposting someone else's tweet. Our script removed all 'RT's from every comment [48]. For example, "*RT @sarah I've got pneumonia on both lungs and I'm fighting for me and my baby*".

3.6.4 Remove @ and hashtags at the starting and end of the tweet.

On Twitter, a post can point to other users using an @ token in front of the username and an # to tag tweets pertaining to a particular category that is trending. We specifically removed words starting with the @(mentions) and #(hashtags) symbols only at the beginning and end of each tweet in order to maintain the context of the tweet, because some of the mentions and hashtags in between are important key candidates that denotes the meaning of a constructs. For Example: "*I am feeling something worse ??is going to happen for India#COVID-19#StayHome\r\nPlease stay at your home ??????*".

3.6.5 Remove and Separate Middle hashtags and @ symbols

We ran a Python script to identify all words beginning with @(mentions) and #(hashtags) that occurred within the comments. For each of these words, whitespace characters were added before every uppercase character encountered in them, and the @(mention) or #(hashtag) was removed. For example, *"#StayHomeSafe" was converted to "Stay Home Safe*".

3.6.6 Converting tweets to lower case

In social media comments, it is very common to find words or groups of words with irregular casings, (*for example, "This is REAL").* Therefore, it is necessary to have all words in a consistent case. This would ensure that during text classification, all tokens are mapped to their corresponding feature irrespective of the casing[60]. This would prevent case sensitive issues with classifying data and ensure that the texts are consistent. We ran a script to convert all words to lower case characters.

3.6.7 Expand abbreviations.

In the process of text analysis, abbreviations can create noise[78]. To solve this, abbreviations are expanded. Abbreviations with apostrophes are expanded first, followed by those without apostrophes. For example, “can’t” is replaced by” can not”,” don’t” is replaced by” do not”, etc. These abbreviations were expanded to their full forms.

3.6.8 Removing punctuations, special characters and emoticons.

Since a text-based approach was used, any feature that are not in text format can lead to inconsistencies[60] and this can reduce the overall effectiveness of the system. In order to avoid redundant features, we removed irrelevant punctuation marks since they are not needed in the analysis.

3.6.9 Removing numbers

To refine a tweet content, numbers are removed to increase the effectiveness[48].

3.7 Data Annotation

After data preprocessing, similar to the approach used by existing research[86], we prepared the ground truth dataset by annotating the comments to their health behaviour constructs using the set of candidate keyphrases generated earlier to label the comments accordingly. The data was labelled according to two classification tasks in machine learning namely multiclass (single label) and multiclass-multilabel classifications, they are both described in chapter2.

Data annotation is simply the process of categorizing and labeling information so that machines can use it. It is especially useful for supervised machine learning (ML), where the system relies on labeled datasets to process, understand, and learn from input patterns to arrive at desired outputs. In machine learning, data labeling is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it. For example, labels might indicate whether a photo contains a bird or car, which words were uttered in an audio recording, or if an x-ray contains a tumor. Data labeling is required for a variety of use cases including computer vision, natural language processing, and speech recognition[124].

3.7.1 Data Annotation process

The proposed approach used an algorithm to identify the text to extract based on recognized text patterns known as pattern matching. Pattern matching checks if a specific sequence of data/ characters/tokens exists among a given data. It is used for Identification, Page Processing, Storage Processing or Pre-Classification[125].

The Pattern Matching process looks for a specified pattern within a user-defined value. It works by "reading" through text strings to match patterns that are defined using Pattern Matching Expressions, also known as Regular Expressions. The Pattern Matching Expressions are compared to text and token data in the Pattern Matching process as well as several other Quick Fields processes and features. When the text or token information matches the pattern, Quick Fields can perform some action, such as identifying the document as belonging to a document class, automatically annotating the text that meets the pattern or substituting different text[125].

Figure 5, shows the pattern matching approach, which classifies the data according to the keyphrase found in the text. We present an algorithm which accepts keyphrases and our corpus to label the tweets.

```
Perceived_Barriers_include_Keywords = [
      “frustration”,
      “suffering from hunger”,
      “cancelling flights”,
      “closures affecting”,
      “precarious position”,
      “experiencing difficulties”,
]

if any ([re.search(r’\b’ + kw + r’\b’, str(tweet)) for kw in
      Perceived_Barriers_include_Keywords]):
return “PerceivedBarriers”
```

Figure 5 Regular Expression Algorithm

We applied eight Regular Expressions algorithms for each classification task as we have eight constructs using the extracted keyphrases. Based on the two types of ML classification tasks applied in this study, the data was annotated differently for the multiclass (single label) and multiclass-multilabel classification task.

3.8 Feature Extraction

Twitter data are just sequences of string characters. To use automatic classification algorithms, special representation must be used to make it suitable for computation. Hence, we vectorized the ground truth data using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting technique. During the vectorization process, each word (term) in the corpus is assigned a unique number. Text data is converted into an N-dimensional vector, where N represents the number of words in the corpus. Each vector element's value represents the term frequency of the corresponding word[80].

3.8.1 Feature Weighting with Term Frequency-Inverse Document Frequency (TF-IDF)

The Term Frequency-Inverse Document Frequency (TF-IDF) [43] statistic weights terms by combining the frequency of a term in a document (TF) with how rare the term appears in comparison to the entire document collection (IDF). TF – IDF is computed as:

$$TF\ -IDF(d,t) = TF(d,t) \times IDF(t)$$

where d stands for document, t for term, TF for term frequency, and IDF for inverse document frequency. Term Frequency (TF) is the number of occurrences a term (feature) appears in a document and is computed as[43]:

$$\mathrm{TF}(\mathrm{d},\mathrm{t}) = \sum_{i \in \mathrm{d}}^{|d|} 1\{\mathrm{d_i} = \mathrm{t}\}$$

Document Frequency (DF) [100] is the number of documents that contain a specific word or term. Inverse Document Frequency (IDF), on the other hand, tackles the problem of DF not being a strong discriminator by taking into account the importance of words/terms in relation to the total number of documents and the number of documents in which the word/term is found.

$$IDF(t) = log\frac{1 + |d|}{d_t}$$

where d represents the total number of documents and dt represents the number of documents containing the word t. Each unique term in the document set is assigned a TF-IDF weight, and all terms are ranked from highest to lowest weight value, indicating their importance. The top k terms are chosen using a user-defined threshold k[43].

3.9 Text Classification with Machine Learning

After feature selection and transformation, the documents can be easily represented in a form that can be used by a ML algorithm. We then developed machine learning (ML) models to classify the comments into their corresponding health behaviour construct since our goal is to use the health behaviours theory and machine learning to explain/examine people's behaviour and reactions towards COVID-19 pandemic. We implemented the models/classifiers by performing multiclass (single label and multilabel) classifications using three supervised ML algorithms widely used for text classification problems for predicting the constructs.

Classification problems are classified based on the number of class labels that can be assigned to a particular input instance[43]. In this scenario of classifying health behaviour constructs, we employed two classification types which are multiclass (single label) and multiclass-multilabel methods to compare the result as most comments belongs to multiple labels as opposed to a single label. We employ three widely known supervised Machine learning models to perform the classification. We use one nonparametric (decision tree) and two parametric classifiers (linear support vector classifier and logistic regression) to classify twitter messages.

3.9.1 Multiclass Classification

Classification in machine learning is the problem of determining which set of categories (classes) a new instance belongs to. When there are more than two different labels in the categories, the problem is referred to as multiclass classification. The basic scenario of multiclass classification assumes that (1) and only one class label is assigned to each instance (this is a single-class classification). For example, consider a document classification problem in which each document must be classified based on the language in which it was written. If a document could only be written in one idiom, and the idioms available were Chinese, English, French, German, Portuguese, and Spanish, each document would be classified in one of these six classes. Each input instance is assigned to only one of the possible classes in this case and this is referred to as single-label classification. The majority of classification problems studied in ML are single-label classification problems[4].

3.9.2 Multilabel Classification

As previously discussed, most classification problems assign a single class to each example or instance. However, there are many classification tasks in which each instance can be assigned to one or more classes. This set of problems falls under the umbrella of multi-label classification. Document classification is a common example of multi-label classification problems, as each document can be assigned to more than one class. To perform this classification, the multi-label classification was transformed into multiple binary classification problems. This technique uses the existing single-label classifiers to perform classification after transforming the multi-label

problem into multiple single-label problems and combines the results of all the single-label classifiers to find results for multi-label classification[113]. In this case, each binary classifier is in charge of predicting only one label (one binary problem for each label). The model was trained and validated separately for each of the eight constructs, yielding eight binary classification models for each model.

In creating the Machine Learning Model, we have three (3) parts:

- Data Splitting (Training and Testing set)
- Machine Learning Algorithm implementation
- The Evaluation Process

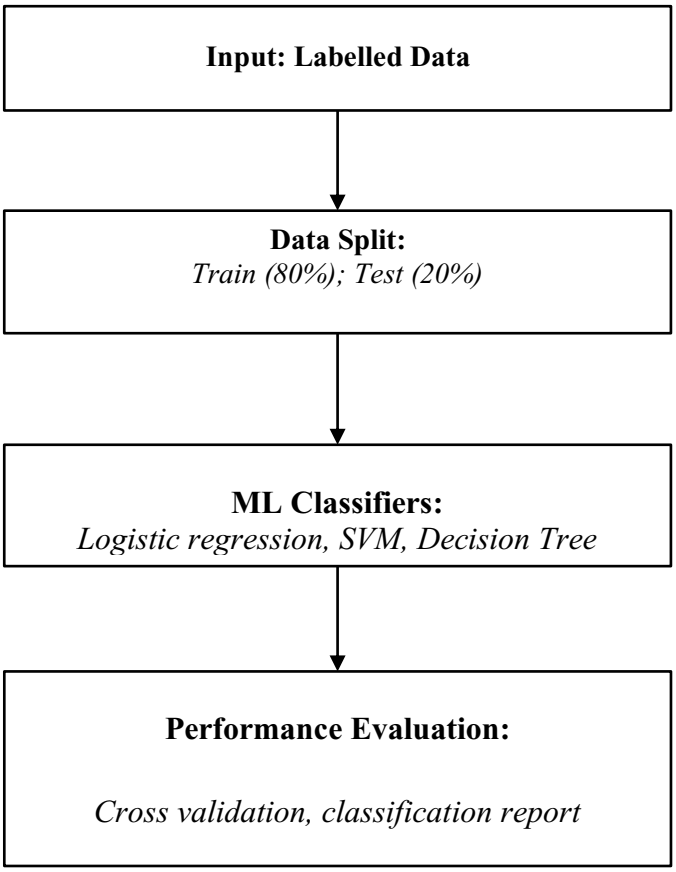

Figure 6 - Overview on Applying Machine Learning Model

3.9.3 Data Split

We split our data into different tasks, which are training datasets and testing datasets. Training dataset is used in training the models, While the testing dataset is used to evaluate the model's performance[128]. We built random subsets of the tweet dataset by dividing the dataset into two parts: 80 percent training and 20 percent testing.

3.9.4 Machine Learning Classifiers

The main objective of machine learning algorithms is to identify meaningful relationships in a body of training data that is presented as individual examples, as well as produce a generalization of those relationships that can be used to classify test data that is presented later. As a result, a variety of learning paradigms have emerged to deal with various situations[13]. Here we perform classification techniques to classify tweets during the pandemic to predict the general behaviours of the public.

Depending on the issue, there are a variety of machine learning algorithms to choose from. Three machine learning algorithms that are widely used for text classification are described in this chapter. We selected three state-of-the-art models in the classification tasks for our experiment as they are widely used for text classifications and performs well: Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT). We will experiment on each ML model's application in the health behaviour classification in our experiment since these models have been well-studied. We implemented the models/algorithms after splitting the data:

3.9.4.1 Decision Tree

One of the most common learning methods is the Decision Tree, which is a classification technique that focuses on an easily understandable representation structure. A Decision Tree is created by iteratively splitting a data set on an attribute that splits the data as much as possible into the various existing classes until a stop criterion is reached. Since Decision Trees can easily be visualized in a

tree-structured format, which is simple to understand for humans, the representation form allows users to get a quick overview of the data[84].

The root node, inner nodes, and end nodes, also called leafs, are the three types of nodes in Decision Trees. There are no incoming edges on the root node and marks the start of the decision-making process. There is precisely one incoming edge and at least two outgoing edges on the inner nodes. They have a test that is based on a data set attribute. For example, a test might ask, "Is the customer above 35 for the attribute age?". The Leaf nodes is made up of an answer to the decision problem, mostly represented by a class prediction. A decision issue may be, for example, determining whether or not a customer in an online store would make a purchase, with yes and no class predictions. No outgoing edges and only one incoming edge exist in leaf nodes. The decision made by the previous node is represented by edges[13].

All nodes that are separated by exactly one edge from n are referred to as children of n, given a node n, while n is referred to as the parent of all its child nodes. A Decision Tree is depicted in Figure 7. For example, a data record with the attributes cold, polar Bear will be passed down to the left subtree because his temperature attribute is cold, and then to the leaf “North Pole,” which would be classified with the corresponding label[67].

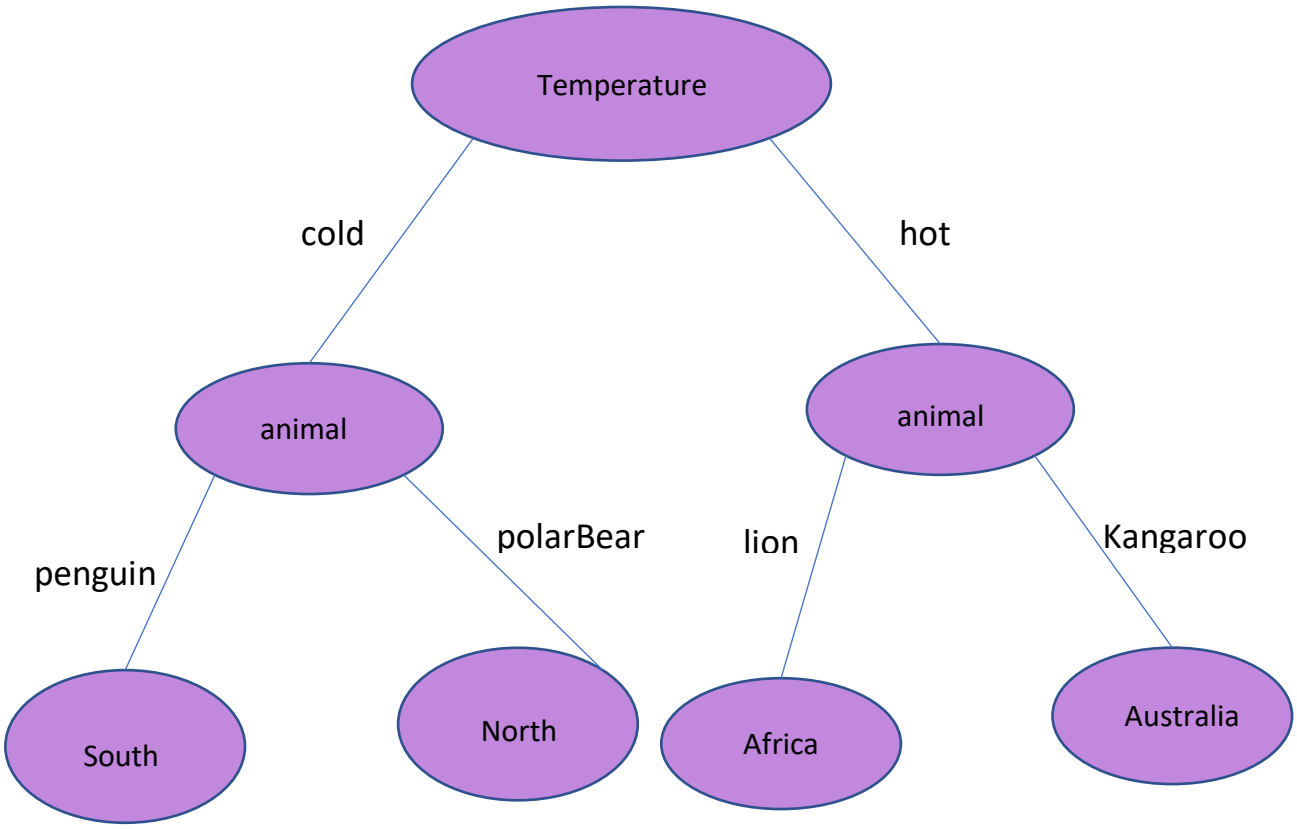

Figure 7 Decision Tree Example [84]

Training a Decision Tree is a popular data mining technique mainly used for the purpose of classification. Its aim is to predict the value of a target attribute using a set of input attributes. In a supervised example, a training set is used to identify patterns in the data and construct the Decision Tree. After that, they will use a collection of previously unseen examples to predict the value of their target attribute[13]. The training set contains data records in the form of:

$$(\vec{x}, Y) = (x_1, x_2, x_3, \ldots, x_n, Y)$$

with Y being the target attribute value, $\vec{x}$ being a vector containing n input values, where n is the number of attributes in the data set.

A training set containing a target attribute, input attributes, a split criterion, and a stop criterion is required to train a Decision Tree and thus construct a classifier. The split criterion assigns a value to all attributes at a given node. This value indicates how much information is obtained by breaking

the node using this attribute. Following that, the best value from all attributes is chosen, and the node is divided into the various outcomes of each attribute[67].

Common stop criteria are:

• The tree's maximum height has been reached.

• The node's number of records is less than the allowed minimum.

• In terms of gained knowledge, the best split criterion does not surpass a certain threshold[67].

3.9.4.2 Support Vector Machines

Support Vector Machines[74] are supervised learning methods that require classified and established data to classify new data. The most basic approach to classifying data is to try to construct a function that separates data points into corresponding labels with (a) the fewest possible errors or (b) the maximum possible margin. This is because larger empty areas next to the splitting function result in less errors because the labels can be better distinguished.

Figure 8 shows that a data set can be separated by multiple functions without causing any errors. As a result, the margin around a separating feature is used as an additional parameter to measure the separation's quality. The separation A is preferable in this case because it separates the two classes more precisely.

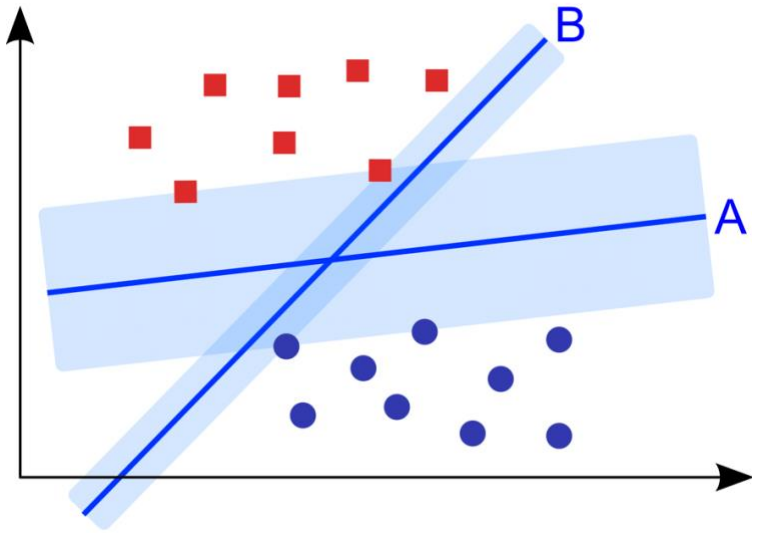

Figure 8 Support Vector Machine diagram [74]

Figure 8 shows Visual of a Support Vector Machine splitting a data set into two groups using two different linear separations, resulting in different sized margins across the splitting functions[67]. Formally, Support Vector Machines (SVMs) generate one or more hyperplanes in an n-dimensional space. The first phase in the data splitting process is to try to divide the data into corresponding labels. The example given uses a data set of n data points for predicting the likelihood of a customer making a purchase in an online store, with each data point consisting of a label y ∈ {purchase, nopurchase} and an attribute vector $\vec{x}$ containing the data values for that specific session. The resulting function can be used to classify future events if the data is entirely separable in a linear fashion[67][74].

3.9.4.3 Logistic Regression

In the same way, as linear regression calculates the target variable, logistic regression does the same. Rather than predicting values like linear regression, logistic regression estimates the odds of a specific occurrence happening. For example, when forecasting admissions to a school, logistic regression estimates the odds of students being admitted. The logistic regression algorithm transforms the response (target variable) into the probability of the event occurring. The effect of each variable on the odds ratio of the observed event of interest is the result. The main advantage is that it eliminates confounding effects by analysing the association of all variables at the same time[106]. The model builds a regression model to predict the probability that a given data entry belongs to the category labeled "1." Logistic regression models the data using the sigmoid function, just as linear regression assumes that the data follows a linear function[32][93].

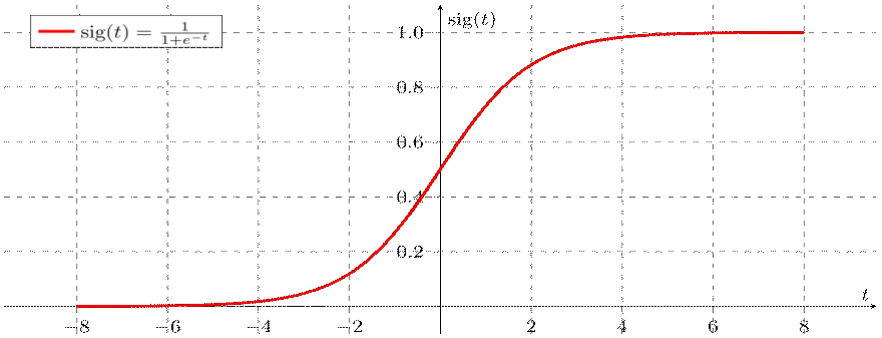

Figure 9 logistic regression [32]

3.9.5 Performance Measures

3.9.5.1 Cross validation

In machine learning, the Bias and Variance dilemma is a common problem to get trade-off between good generalization and avoiding over-fitting. As shown in Figure 10, a 10-fold cross-validation is a technique for randomly dividing a dataset into 10 folds. The data is divided into nine parts for training and one part for evaluating the model. The learning process is performed 10-times on the different sections of the training data by repeating the process, and each component is only used once for testing [128].

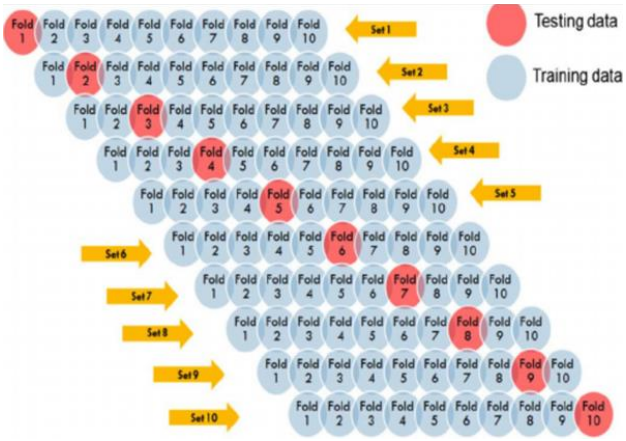

Figure 10 10-fold cross-validation procedure

We trained and evaluated each model on the vectorized documents by performing a 10-fold cross-validation experiment and conducted the 10-fold cross-validation using training data to compare the three models [128]. That procedure is illustrated below:

- The training data is divided into ten equal parts at random.
- Each part of the data serves as a test set in turns. Support Vector Machine, Logistic Regression and Decision tree are trained on the remaining data and validated on the test set.

- Then the ten tests results are averaged[128][32].

3.9.5.2 classification report

Furthermore, a classification report is created to confirm the result. The result of this process is presented in chapter4. We then compared their results using four evaluation metrics: accuracy, precision, recall, and F1 score. There are various performance measures in classification, we choose four common measures in our approach, which are *accuracy*, *precision*, *recall*, and *F1* score[31]. The F1 score (or F measure) is a preferred measure because it is the harmonic mean of precision and recall, thus accounting for the impact of each class on the overall score.

The classification report displays the model's precision, recall, and F1-score. Table 3 shows the meaning of all the symbols used in the equations below.

Table 3 Classification Report Symbol

TN	the prediction is negative, and the actual value is negative.
FN	the prediction is negative while the actual value is positive.
FP	the prediction is positive while the actual value is negative.
TP	the prediction is positive, and the actual value is positive".

1. Precision Score [102]

Precision is defined as the ratio of correctly classified true positives divided by the total number of true positives and false positives, as shown in Equation 1below.

Equation 1Precision equation

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

2. Recall Score [102]

The ratio of the total number of correctly labeled true positives divided by the total number of true positives and false negatives is known as recall, as shown in Equation 2 below.

Equation 2 Recall Equation

$$\text{Recall} = \frac{\text{T P}}{\text{T P} + \text{F N}}$$

3. F1 Score [31]

F1 is the weighted average of Precision and Recall. Because it takes into consideration both false positives and false negatives, it will be more useful in cases of uneven class distribution. This metric's equation is shown in Equation 3:

Equation 3 F1Score Equation

$$\text{F1 Score} = \frac{2\times(\text{Recall}\times\text{P recision})}{(\text{Recall}+\text{P recision})}$$

4. Accuracy Score [31]

Accuracy is defined as the classification ratio of all correctly classified data across the whole test set. Equation 4 below can be used to calculate it.

Equation 4 Accuracy Equation

$$\text{Accuracy} = \frac{\text{T P} + \text{T N}}{\text{T P} + \text{T N} + \text{F P} + \text{F N}}$$

We have two types of averaging evaluation methods based on these per-class measurements: macro average and micro average. “Micro-average Create a contingency table for all classes then compute the precision, recall and F1 measure of the whole dataset as one “big class””. “Macro-average Compute the precision, recall and F-measure for each class, then average the sum over number of classes”. We used macro-averaging measurements to evaluate the performance of our algorithms in our study[102][31].

3.10 Programming Language and Tools used

3.10.1 Jupyter notebook

Jupyter notebook is a free and opensource web application that allows users to write python code in an easily shared and monitored environment. To gain a deeper understanding of how the program works, it is simple to visualize and compile the code line by line[42].

3.10.2 Scikit-learn

The Scikit-learn library for Python programming language is used to develop the machine learning model. The library is based on SciPy (Scientific Python) and includes the following libraries, among others: • Numpy: multidimensional arrays package • Matplotlib: A comprehensive library for drawing 2D and 3D diagrams. • Pandas is a data structure and analysis tool[112].

Scikit-learn is a free machine learning software library designed to work with other frequently used Python libraries such as Numpy and SciPy. It focuses on data modeling and offers multiple supervised and unsupervised learning algorithms. Scikit-learn implementations of support vector machines, logistic regression, naïve Bayes and random forest were used for classification [42].

Scikit-learn implementations of support vector machines, logistic regression, naïve Bayes and random forest were used for classification[112].

3.10.3 Pandas and Numpy

Pandas is a Python package that makes it easier to analyze imported datasets. Panda is based on Numpy, a library that includes added support for multidimensional arrays[121].

3.10.4 Natural Language Tool Kit-NLTK

The Natural Language Toolkit (NLTK) is a Python package that consists of a set of natural language algorithms. This package is an open source and easy to use. It includes a number of useful pre-processing packages, such as tokenization, punctuation removal, stop words removal, and word count. NLTK assists the computer in comprehending, analyzing, and preprocessing text samples [112].

3.11 Thematic Analysis

Thematic Analysis (TA) is the process of identifying, analyzing, and reporting themes (patterns) within a given set of data[45]**.** As there is a dire need to delineate the on-going conversations on the infection with the intention of creating awareness on people's reaction, opinion, action and recommendation that are inimical to the wellbeing of the populace. Hence, we performed thematic analysis of the discussions on each construct by manually analyzing the comments to extract themes.

3.12 Summary

We successfully preprocessed and labelled the data. In the data preprocessing stage, lemmatization and stop words were not applied due to semantic meaning of the data . Then we applied three (3) ML models on multiclass and multilabel classification using supervised learning algorithm to classify the data which was successful as explained further in the result section. Lastly, we conducted thematic analysis of each construct to identify and investigate public reactions towards COVID-19 pandemic.

CHAPTER 4 RESULT AND DISCUSSION

In this chapter, we will present the results of our approach to see how each classifier performed on each construct using multiclass and multilabel classification.

4.1 Data Preprocessing

After Data preprocessing, we had a total of 696597 from the original data.

Table 4 Number of data after preprocessing

Original Data	**Clean Data**
5562550	644230

Due to the semantic meaning of the constructs, stop words and lemmatization were not part of the data preprocessing stage, as removing that process was necessary in bringing out the contextual meaning that denote the constructs otherwise it can bring out a lot of different meaning. The same thing applies for lemmatization.

4.2 Keywords Generated For Data Labelling

The keywords generated for each construct to perform automatic data annotation are shown below. In generating candidate keyphrases, perceived susceptibility and perceived severity were easier to generate as the rest were more contextual to get the right meaning of the construct. Also, most Cue to Action (CTA) and Perceived benefits comments were similar because most of the CTA comments reminded users to carry out the preventive health behaviors by adding the perceived benefits.

4.2.1 Cue to Action

stay at least six feet, practice social distancing, practice healthy habits, wash your hands ,stay home, stay healthy, remember to stay at home, respect social distancing, sanitize your hands,

please stay home, cover your mouth, remember to keep busy, stay at home, use your mask, drink a lot of water, eat a lot of fruit, sanitize yourself, frequently wash hand, adhere to social distance, avoid social gathering, avoid shaking hand, avoid handshakes, avoid hugs, avoid touching your nose and mouth, maintain social distancing, avoid mass gather, avoid public place, avoid touching your face, stop shaking hands, stop the spread, practice social distancing, stayaware, prevent the spread, slow the spread, prevent transmission.

4.2.2 Perceived Barriers

frustration, suffering from hunger, cancelling flights, closures affecting ,precarious position ,experiencing difficulties, facing difficulties, affecting childrens education, affecting mental health, affecting health services, challenged mental health, worry about losing money, causing discomfort, freedom of movement, unleashing abuse ,consequence of the emergency, consequence of the lockdown, consequence of social distancing, laid off work, loss of work, loss of job, restrictions on freedom, feels lonely, adversely impact, loss of income, loss of freedom, boredom, affect mental health, suffering from anxiety, unemployment claims, economic losses, cancelled flights, financial issues ,precarious position, shelters close, challenging for mental health, feel anxious ,feel distressed, bored, frustrated, ineffective things, defying lockdown, reduced hours.

4.2.3 Perceived SelfEfficacy

confident that I, confident in my, confident in our, confident that we , i can do this , we can do this, we can fight, I can fight , we can beat, i can beat, I can defeat, we can defeat ,we can overcome, i can overcome, we can conquer , i can conquer, we can prevail, i can prevail, i feel confident, we feel confident , i am motivated , we are motivated, i feel encouraged ,we feel encouraged , i am positive that, we are positive that, i feel positive that, self confident, we can stop, i have confidence, we have confidence, i can wear, we can wear , i can stayhome we can stayhome, i believe that i, i believe that we, i believe i can, we believe that, i shall overcome , we shall overcome , i am convinced that, i can survive, we can survive, i can pull through, we can pull through , i can get past , we can get past, i can get through , we can get through , i can control, we can control, i can

get control of, we can get control of, i can cope, we can cope, i do believe we ,i believe we can, i can manage , we can manage ,i can wash my hands, we can wash our hands, i can keep a safe distance, we can keep a safe distance, i can stay safe, we can stay safe, i can learn how to , we can learn how to, i can keep fighting, we can keep fighting, i can prevent, we can prevent, i will get through this, we will get through this, i can end this, we can end, we can put an end to, we will win, i am determined to, we can save the, we will play our part, i will play my part, we can control it, we can control this, we can deal with this, we can do it ,i can do it, we will be okay, i can chose to isolate, i can choose to stay, i can chose to wear, we will overcome this ,we shall overcome this, i can stop the, i can stop this.

4.2.4 Perceived Susceptibility

probability of getting infected, probability of getting sick, probability of catching, possibility of getting infected, possibility of getting sick, possibility of catching, might get sick, might get infected, might catch, may get sick, may get infected, may catch, most likely to get it, most likely to get this, most at risk, more likely to get it, more likely to get this, more likely to be infected, more at risk, likely to get sick, likely to get infected, likely to contract , likely to be infected, likely to be sick, likely to catch, likelihood of getting sick, likelihood of getting infected, likelihood of catching, risk of getting it, risk of getting this, risk of getting sick, risk of getting infected, risk of catching, risk of becoming severely ill, risk of becoming ill, risk for contracting, risk of contracting, risk of serious illness, risk of being infected, this is real, risk is real, huge probability, high probability, high risk of, high risk for, high risk from, susceptible, susceptibility, vulnerable, increased risk, at acute risk, multiple cases, new cases, more cases, cases rises, cases increases.

4.2.5 Perceived Benefits

save lives, save live, save lifes, saves life, flatten the curve, reduce the crowd, save a soul, flatten the curve, curb the spread, saving hundreds, slowing spread, help slow the spread, help protect yourself, reduce total number of cases, save a seniors life, reduce number of cases, slow spread.

4.2.6 Perceived Severity

death, die, dead, dying, hospitalized from the virus, suffering from the pandemic, suffering from the virus, sick from the virus, sick from the pandemic, severe illness, severly ill, serious illness, seriously ill, critical condition, infected patient, test positive, tested positive, casualties, painful symptoms, need critical care, mortality rate, this can kill, virus can kill, virus kills, the virus is deadly, breath struggle, positive cases, confirmed cases, critical care, hospitalized, respiratory failure, pneumonia.

4.2.7 Social Norm

hugging, wearing mask, wearing gloves, wearing gloves, wearing hand glove, wearing hand gloves, shaking, working, partying, goingout ,carrying mask ,carrying sanitizer, carrying handsanitizer, gathering , staying home, gathering, kissing, avoiding ,washing hands, covering mouth, keeping distance, eating out, travelling, sanitizing.

4.2.8 Trust

trust, falsely assuring, right information, irresponsible information, false information.

4.3 Data Annotation

Table 5 and Table 6 shows the result of the labelled data of each construct for the multiclass and multilabel classification task.

4.3.1 Multi – Class Classification

Table 5 shows the number of labelled data on each construct for the multi - class classification task.

Table 5 Number of comments in each construct

HBM Construct	**Sizes**
Perceived susceptibility	13454

Perceived severity	25087
Perceived benefits	6730
Perceived barriers	4537
Cue to action	29358
Self-efficacy	2314
Social Norm	1564
Trust	2073
Total	85,117

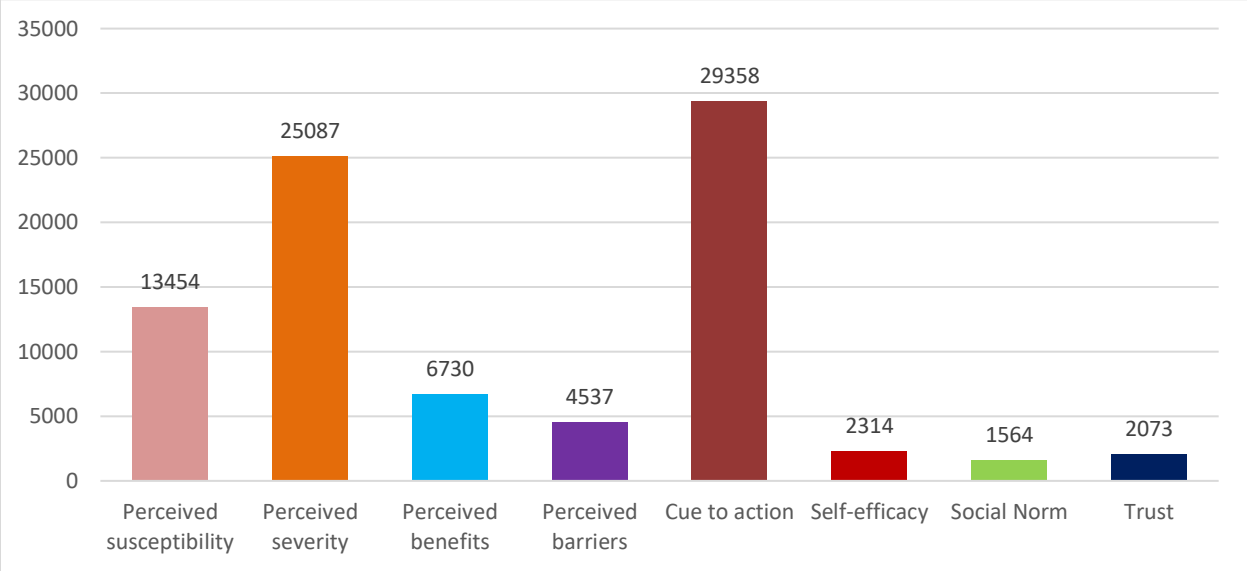

Figure 11 Visual representation of number comments in each construct

4.3.2 Multi Label Classification

Table 6 shows the number of labelled data on each construct for the multi – label classification task.

Table 6 Number of comments in each construct

HBM Construct	Sizes
Perceived susceptibility	13597
Perceived severity	27824
Perceived benefits	7102
Perceived barriers	4559
Cue to action	34007
Self-efficacy	2339
Social Norm	1564
Trust	2075
Total	93,067

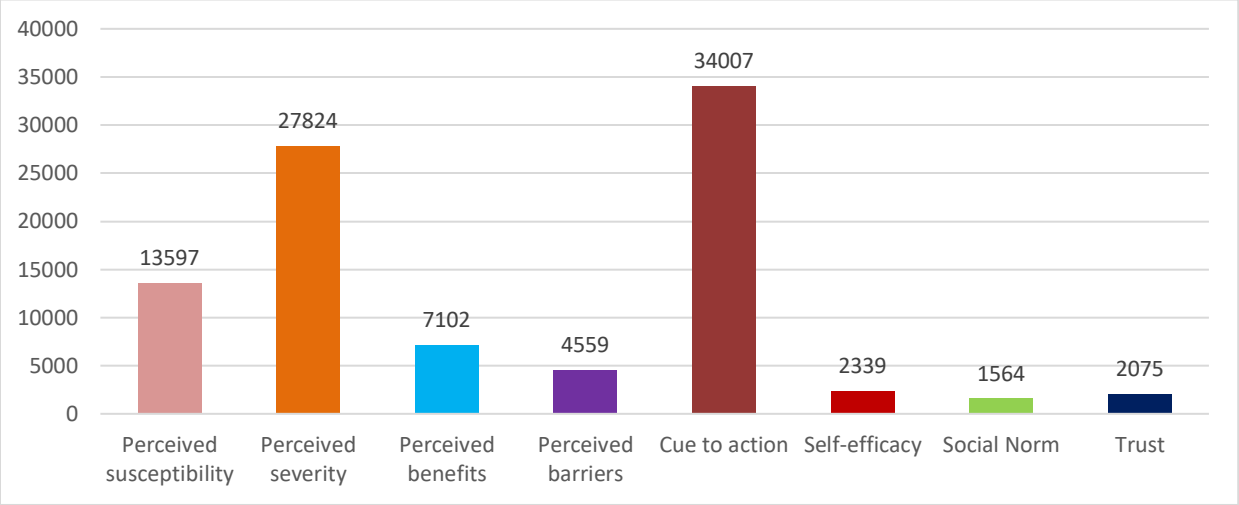

Figure 12 Visual representation of number comments in each construct

4.4 Machine Learning Classifiers

In this section, a summary of model performance on the training and testing dataset are shown using different comparison metrics. The performance metrics considered are accuracy, precision and recall. The results shown below are for the two types of classification applied in this study.

We first reviewed the result for the single label multi class dataset, which shows the result of the cross validation, training and testing classifications.

4.4.1 Multi-Class Classification

This section presents the result of the evaluation performance on the cross validation, training and testing set.

4.4.1.1 Cross Validation performance

The cross-validation technique was used to evaluate the 3 model's (Linear SVC, Decision Tree and LogisticRegression) performance, as shown in Table 7. The f1score showed that all the models performed above 0.73, which indicates that there was no over fitting in the model[12].

Table 7 Result of cross validation

Linear SVC								
Accuracy=0.95								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.94	0.99	0.91	0.86	0.96	0.95	0.92	0.99
Recall	0.97	0.95	0.90	0.67	0.98	0.94	0.68	0.97
f1-score	0.95	0.97	0.91	0.76	0.97	0.95	0.78	0.98
Logistic Regression								
Accuracy=0.94								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.93	0.99	0.92	0.85	0.94	0.95	0.95	1.00

Recall	0.97	0.91	0.88	0.67	0.98	0.93	0.62	0.92
f1-score	0.95	0.95	0.90	0.75	0.96	0.94	0.75	0.96
Decision Tree Classifier								
Accuracy=0.94								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.96	0.95	0.92	0.73	0.95	0.94	0.77	0.98
Recall	0.97	0.93	0.93	0.72	0.95	0.94	0.71	0.97
f1-score	0.97	0.94	0.92	0.73	0.95	0.94	0.74	0.97

4.4.1.2 Classification performance of the ML Classifiers on each construct

In this section, a summary of the model's performance on the training and testing dataset is shown using the classification report, which includes accuracy, precision and recall. The result is shown in Table 8, Table 9 and Table 10.

Table 8 Training and Testing Set Result on Linear SVC

Linear SVC								
Train								
Accuracy = 0.99								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.98	1.00	0.97	0.98	0.99	0.99	1.00	1.00
Recall	0.99	1.00	0.98	0.90	0.99	0.98	0.96	1.00
f1-score	0.99	1.00	0.97	0.94	0.99	0.98	0.98	1.00
Test								
Accuracy = 0.95								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.94	0.99	0.91	0.86	0.96	0.96	0.92	0.99

Recall	0.97	0.94	0.89	0.70	0.98	0.94	0.69	0.97
f1-score	0.96	0.97	0.90	0.77	0.97	0.95	0.79	0.98

Table 9 Training and Testing Set Result on Logistic Regression

Logistic Regression								
Train								
Accuracy = 0.96								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.95	1.00	0.94	0.91	0.97	0.97	0.97	1.00
Recall	0.98	0.95	0.92	0.76	0.99	0.95	0.79	0.96
f1-score	0.97	0.97	0.93	0.83	0.98	0.96	0.87	0.98
Test								
Accuracy = 0.94								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.93	0.99	0.92	0.85	0.94	0.96	0.94	1.00
Recall	0.97	0.90	0.87	0.69	0.98	0.93	0.64	0.92
f1-score	0.95	0.95	0.89	0.76	0.96	0.94	0.76	0.96

Table 10 Training and Testing Set Result on Decision Tree

Decision Tree Classifier								
Train								
Accuracy = 1.00								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f1-score	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test								
Accuracy = 0.95								

Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.97	0.96	0.91	0.79	0.95	0.94	0.77	0.98
Recall	0.97	0.95	0.92	0.76	0.96	0.94	0.70	0.98
f1-score	0.97	0.95	0.92	0.77	0.95	0.94	0.73	0.98

In Table 8 -Table 10, the training set shows all three (3) ML classifiers performed well with an accuracy above 96% and they are higher than the testing result which indicates they trained higher as they are training on the same set. In the testing set the result above shows that all the three (3) models had an accuracy above 94%. The classifiers with the highest accuracy were decision tree and linear SVC with 95% accuracy. Based on each construct, we used the f1 scores of each model to analyze their performance; Trust had the highest performance of 98% with linearSVC and decision tree, followed by cue to action with decision tree (97%), perceived barrier and perceived severity with linearSVC (97%), perceived susceptibilty with linearSVC(95%), social norm with linearSVC (79%) and perceived selfefficacy with linearSVC and decision tree(77%).

4.4.1.3 Best and Least performance on Machine Learning classifiers

Table 11 shows the best and least performing ML classifiers for each Health Behaviour Construct.

Table 11 Best and Least performing Machine Learning classifiers

HBM Construct	Best performing ML classifiers(F1Score)	Least performing ML classifiers (F1-Score)
Perceived susceptibility	linearSVC (0.95)	Decision Tree and Logistic Regression (0.94)
Perceived Severity	linearSVC (0.97)	Decision Tree (0.95)
Perceived Benefits	Decision Tree (0.92)	Logistic Regression (0.89)
Perceived Barrier	linearSVC (0.97)	Decision Tree and Logistic Regression (0.95)

Cue to Action	Decision Tree (0.97)	Logistic Regression (0.95)
Perceived Self Efficacy	Decision Tree and linearSVC (0.77)	Logistic Regression (0.76)
Social Norm	linearSVC (0.79)	Decision Tree (0.73)
Trust	Decision Tree and linearSVC (0.98)	Logistic Regression (0.96)

As seen from the analysis, the performance are all relatively high and did not have long gaps within the range of their performance, meaning the three classifiers all did relatively well, as the lowest was 73%. It is obvious from Table 11 that in almost all the datasets, Decision Tree and Linear SVC outperformed logistic regression, but not within a close range. In terms of the constructs; Trust had the highest performance (98%), followed by perceived severity, perceived barrier and cue to action (97%), then perceived susceptibility (95%) and Perceived Benefits (0.92), lastly social norm (0.79%).

4.4.2 Multi-label Classification

This section presents the result of the multi-label classification. For the multi label result, we focus mainly on the f1score and not the accuracy because it performs binary classification on each construct.

4.4.2.1 Classification performance of the ML Classifiers on each construct

A summary of the model's performance on the training and testing dataset in a multi label task is displayed using the classification report, which includes accuracy, precision and recall.

Table 12 Training and Testing Set Result on Linear SVC

Linear SVC
Train
Accuracy = 0.99

Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f1-score	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test								
Accuracy = 0.95								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	1.00	0.99	0.97	1.00	1.00	1.00	1.00
Recall	0.96	0.74	0.92	0.51	0.96	0.92	0.48	0.92
f1-score	0.98	0.85	0.96	0.67	0.98	0.96	0.65	0.96

Table 13 Training and Testing Set Result on Logistic Regression

Logistic Regression								
Train								
Accuracy = 0.96								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	1.00	0.99	0.96	1.00	0.99	0.97	1.00
Recall	0.78	0.49	0.86	0.66	0.91	0.85	0.78	0.74
f1-score	0.96	0.66	0.92	0.78	0.95	0.92	0.89	0.83
Test								
Accuracy = 0.94								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	0.99	1.00	0.98	0.86	1.00	0.99	0.83	0.98
Recall	0.68	0.13	0.46	0.19	0.52	0.50	0.03	0.16
f1-score	0.81	0.23	0.63	0.31	0.68	0.66	0.60	0.27

Table 14 Training and Testing Set Result on Decision Tree

Decision Tree Classifier
Train

Accuracy = 1.00								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	0.99	0.99	0.94	1.00	0.99	0.95	1.00
Recall	0.99	1.00	1.00	0.85	1.00	1.00	0.96	1.00
f1-score	1.00	0.99	1.00	0.90	1.00	0.99	0.95	1.00
Test								
Accuracy = 0.95								
Construct	Cue to Action	Perceived Barrier	Perceived Benefits	Perceived Self Efficacy	Perceived Severity	Perceived Susceptibility	Social Norm	Trust
Precision	1.00	0.98	1.00	0.95	1.00	0.99	0.94	1.00
Recall	0.99	0.99	1.00	0.85	1.00	0.99	0.96	1.00
f1-score	1.00	0.99	1.00	0.90	1.00	0.99	0.95	1.00

4.4.2.2 Best and Least performance on Machine Learning classifiers

Table 15 shows the best and least performing ML classifiers for each Health Behaviour Construct.

Table 15 Best and Least performing Machine Learning classifiers

HBM Construct	Best performing ML classifiers(F1Score)	Least performing ML classifiers (F1-Score)
Perceived susceptibility	Decision Tree (0.99)	Logistic Regression (0.66)
Perceived Severity	Decision Tree (1.00)	Logistic Regression (0.68)
Perceived Benefits	Decision Tree (1.00)	Logistic Regression (0.63)
Perceived Barrier	Decision Tree (0.99)	Logistic Regression (0.23)
Cue to Action	Decision Tree (1.00)	Logistic Regression (0.81)
Perceived Self Efficacy	Decision Tree (0.90)	Logistic Regression (0.31)
Social Norm	Decision Tree (0.95)	Logistic Regression (0.60)
Trust	Decision Tree (1.00)	Logistic Regression (0.27)

Table 15 shows that Decision tree performed better than SVM and linearSVC. For the multilabel classification Decision tree performed well in all the constructs from 95% and logistic regression did good in some constructs like cue to action with 81%, average in perceived susceptibility, perceived severity, perceived benefits and social norm (range of 60% - 66%), and low in perceived barrier (23%) and self-efficacy(31%).

4.4.3 Summary of ML Result Analysis

From the over analysis it is seen that cue to action and perceived severity mostly performed better within the three classifiers and the least were social norm and self efficacy. Furthermore, decision tree and liner SVC performed better than logistic regression. The difference in the result for multi class within the three (3) classifiers were not much compared to the difference in multilabel. In the multilabel decision tree performed far better than logistic regression and linearSVC was closer in range to decision tree performance. Between multiclass classification performance and multilabel classification performance linearSVC and Decision tree classifier were the best in performance.

4.5 Thematic Analysis

Next, we conducted thematic analysis on the health behaviors construct comments, where we randomly selected comments from both the ground truth dataset and classified dataset on the multi-class classification task. We had a total of 63,960 comments, where 45, 980 was the ground truth data and 17,980 the classified data. Table 16 Themes and sample comments shows the common themes from each construct.

Table 16 Themes and sample comments

Constructs	Themes	sample comments
Perceived Susceptibility	Senior Citizens and Tobacco products/drugs, Confirmed Cases, Role Model /Celebrity getting infected, Increase in the number of cases, People with underlying health issues/weak Immune system.	*"a thread **a weak immune system and drug use makes you more at risk of serious illness** due to covid19 there is always risk when taking **drugs** but now is an extra risky time its safer not to use for more information visit"*

Perceived Severity	Fear/Panic/Anxiety, Increase in Death rate.	*"hi everyone we are all going to die **covid19 is a threat** believe me we are behind usa by weeks new city has had a very **high death rate** because people are not taking it seriously our means of survival is to follow the rules social distancing and handwash"*
Perceived Benefits/ response efficacy	Protecting vulnerable people and loved ones, Saving Lives, Working from home, Reduction in confirmed cases, Going back to a normal life	*".... suck it up stay home as far as possible this is not about you its about everyone else its about your duty to the community and its about **protecting vulnerable people** like my dad so that families can reunite once again in the future."*
Perceived Barriers	Stress, Anxiety and Depression, Lack of Freedom/Restrictions, Increased risk of Domestic Abuse and Violence, Inadequate Medical Attention/ Affecting health services, Economic Isuues	*"it is **not easy to stay at home with anxiety depression migraine overthinking** mind and then facing all family members"*
Cues to Action	Increase in cases, Loss of loved ones/ infection of closed relation, Self – reporting, Posters, Symptoms, COVID-19 management Apps, Increase in death rate.	*"i think with the rate of increase in cases of corona virus in lagos state at this point should be activein the stay at home policy this will help lessen the spread increase of the virus"*
Perceived self-efficacy	Coming together to achieve the goal, leading by example, Reduction in cases (Precautionary measures working).	*"can help break the chain of the spread covid19 let us practice this simple and effective habit together we can fight."*
Social Norm	Lack of believe in the preventive measures (loss of hope), Kids not taking it seriously.	*"just drove past a school kids out for recess there was **no social distancing in fact children were hugging each other** and staff were standing side by side talking to each other that is why we need to"*
Trust	Lack of Trust in Authorities, Misinformation.	*"with the coverup in december and january we really **cannot trust the coronavirus numbers from the chinese government** without more credible and solid evidence to verify ..."*

The themes that emerged from the data are further explained in more detail below;

4.5.1 Perceived Susceptibility

As defined earlier, perceived susceptibility deals with the fear or vulnerability of a person, family, loved one's or community getting infected. This is the stage when one becomes aware or feel threatened about the disease. Most people felt threated or more at risk either because of their health conditions, age and the number of cases as explained in this section.

4.5.1.1 Senior Citizens and Tobacco products

According to our findings, research shows that **senior citizens /older population** and people that smoke **tobacco products or vaping** were considered to be more prone to the virus. Fatality rate was also noticed to appear more among the older population and amongst people that smoke. This corresponds with the actual result of the things that are happening in the world as reflected in the social media posts. see sample comments below:

> *"Even though, young and old have the same probability of getting CoronaVirus infection, the fatality rate is more among the **older population**. Developed countries with an older population are reporting more deaths."*
>
> *"**Seniors and people with disabilities** or compromised immune systems are at **higher risk of contracting covid19**..."*
>
> *"**Tobacco smokers are at greater risk of death if they get the virus,** they are also **catching COVID-19 at a higher rate** than nonsmokers, people whose lungs are damaged from smoking are very to wean from a vent and often die as a result"*

4.5.1.2 Confirmed Cases

In understanding the perceived susceptibility of COVID19, on a daily basis, country published the daily confirmed cases, and these results were also published on news and social media. Based on

our findings, it was shown that people believed they were susceptible when they were confirmed cases. Our findings showed that people believed they were more susceptible to getting the virus, when famous people, family or their relation started getting infected. sample comments are shown below:

> *"at work today last week at this time we had our first covid patient now* ***we have confirmed intubated and being ruled out this is real, the surge is only starting*** *,stay safe"*
>
> *"stay at home uk tells people, as global* ***confirmed cases*** *pass todays"*
>
> *"the united states just overtook china for the* ***most confirmed cases of covid19*** *in the world* ***it did not take long for that to happen stay home and save lives flatten the curve*** *before the healthcare system is overrun"*

4.5.1.3 Role Model /Celebrity getting infected

People were awoken and felt more vunerable to the virus, when celebrities or their idols tested positive for the virus. sample comments are shown below:

> *"****the news prince charles has tested positive for covid19*** *with mild symptoms will massively increase pressure for ramping up of testing capacity most people…*
>
> *"****this is real*** *donald trump talking about his coronavirus covid19 on fox news what do you think* ***spains deputy prime minister carmen calvo has tested positive for coronavirus sowho are you to be so carefree****"*

4.5.1.4 Increase in the number of Cases

As the number of cases increased, people feared they were more at risk to the disease, which was an eye opener. sample comments are shown below:

> *"i think with* ***the rate of increase in cases of corona virus in lagos lagos state at this point should active the stay at home policy*** *this will help lessen the spreadincrease of the virus"*

*"awful italy announces **new covid19 cases increase for a** total of italy reports new coronavirus deaths increase for a total of italys death toll is now exceeds china coronavirus pandemic"*

4.5.1.5 People with underlying health issues/weak Immune system

It was believed that people with weak immune system or underlying health issues were more prone to be infected. Studies show that people with weak immune system or underlying health issues were more prone to be infected. So, they believed in the early stages that they were more at risk and were more scared. In the tweet, a couple of people shared that they felt vulnerable to COVID19 and needed people to do their part, so they do not get infected. Sample Comments are below:

*"i am david i am years old and i have had a kidney transplant i have **a weakened immune systemthis puts me at a severe risk** i need your help to stay safe so please"*

*"given than most of those who die of covid19 **have multiple underlying health issues especially elderly** the question cdc needs to address is how much is life expectancy reduced by contraction of the virus need better data now"*

4.5.2 Perceived Severity

Perceived severity includes views regarding the disease itself (e.g., whether it is life-threatening or can trigger injury or pain) as well as the disease's wider effects on work and social roles. For example, a person may believe that COVID-19 is not a medically serious condition, but if he or she believes that missing work for many days would have serious financial implications, he or she may believe that COVID-19 is an especially serious condition[79].

4.5.2.1 Fear/Panic/anxious

Tweets extracted in this research revealed that people panicked and became anxious when things got severe. The more people believed the disease was severe and fatal with the daily confirmed death toll from the virus, the more it messed with their mental health. sample comments are shown below:

"many of us are losing sleep,bombarded with news and ***information about new coronavirus outbreak people are feeling anxious afraid and alone*** *raimund a psychologist working for msf in hong kong offers simple tips for how we can cope with the stress during covid2019 pandemic"*

"as a source of this pandemic china as country needs to be held reposibility and owes the entire world an explanation lakh of deaths across ***the world plus the mental health issues due to being in isolation and due to fear of death****"*

4.5.2.2 Increase in Death rate

The severe outcome of those infected by COVID19 is death, especially with the new strains of this virus increasing the infection rate and killing faster. This study demonstrated more evidence of the impact of the increase in death rate as it relates to COVID19 to severity perception. As a result of the severe impact of COVID19, which resulted in millions death, medical personnel were among those who died. Tweets from users revealed, this also increased the severity perception of people when many medical practitioners started dying and the news got out. Comments are displayed below:

*"****as the death toll rises the severity of this corona outbreak is becoming real*** *are you prepared it will be in your city soon this new book can help you get ready it is a comprehensive page guide to survive this novel coronavirus st reviews n28"*

*"****coronavirus death toll among doctors in italy rises*** *to covid19"*

"hereu2019s ***a mathematical fact assuming the total deaths continue to increase by a day*** *everyone on earth will be dead in just day's time thatu2019 saturday augustnnitu2019s time everyone pulls together and we will get through this"*

"personally, convinced that covid taken into account the ***huge losses of medical staff*** *amp trillions of economic losses requires bsl4 in testing treatment amp research biosafety rating is not a bureaucratic biological auditing only on inflation of mickeymouse studies but on harm"*

"medical staff infected *with coronavirus of the infected medical personnel are from teluk intan hospital perak please stay at home help our courageous overworked medical personnel save livesnmay all our patriotic frontliners recovern"*

4.5.3 Perceived Benefits/ response efficacy

To help curb the spread, most people took different preventive measure to assist in their own way through donation, giving different examples on the best practices of social distancing, and other precautionary measures such as printing out instructions and labels. Most people encouraged by reminding people of the benefits of carrying out the Public Health Measures. Public Health Measures such as (wearing masks, social distancing etc.) slows the spread of COVID19 while flattening the curve.

4.5.3.1 Protecting vulnerable people and loved ones

Most people felt the need to protect vulnerable people such as people with weak immune system and people they know, which led to carrying out the preventive measures. sample comments are shown below:

"one day when this is over we will spend time together and start hugging again but right now there is no time to waste do your part ***to protect vulnerable people do it for yourself do it for the others d****o it"*

"my mom is a doctor she stays away from me ***so I don't get infected,*** *I really miss her presence."*

4.5.3.2 Saving lives

PHMs has shown that staying home will help to save lives[111] and suppress the spread of COVID19. The impact of COVID19 has revealed that staying at home is the safest way to avoid being infected by the virus, as users in the sampled comment acknowledge. Comments are shown below:

"***stay home save lives*** *quarantine and chill"*

"*please share with other parents if your work is not critical in the response to coronavirus then please keep your child* ***at home stay home save lives*** *coronavirus"*

4.5.3.3 Working from home

A lot of companies have enforced that their staff work from home, this has allowed families to spend time with each other extensively. This seen as perceived benefits amongst users as there are so many benefits of working from home which includes less commute, saving costs and better work -life balance for some people and it was also perceived to help in slowing down the spread. See sample comments below:

"the new way to shop from the parking lot ***start work from home tomorrow yay i can wear slippers****"*

"we all want the economy to recover to go back to normal safely ***the goal of work from home quarantine is to slow the spread of coronavirus*** *we are not there yet numbers are still on the rise listen to the experts follow science follow the data "*

4.5.3.4 Reduction in confirmed case

The impact of preventing the spread is that the amount cases will reduce which users see as perceived benefit. As this will reduce the severity of COVID19. See comments below:

"that may well be believe the calls for instant answers by so many are contrary to the point of ***stay at home which is to reduce cases*** *and avoid overwhelmining health care level of community spread unknown and disease itself still not well understood restraint needed"*
"this initiative of ***lockdown is definitely going reduce new cases*** *of covid19 if everyone strictly follow social isolation"*

4.5.3.5 Going back to a normal life

The goal of exercising the PHMs is to flatten the curve and eventually get rid of the COVID19, allowing people to return to normalcy. Normalcy would imply that people would be able to travel and reunite with their families. See sample comments below;

"if you can stay home please do when this coronavirus gets wiped out ***then we can return to normal life*** *here is a thread from a dr in nyc on the front lines of this"*

"someone please save brazil from our stupid president while the world are in quarantine and taking care of themselves to stop spreading the virus our president wants to ***stop the quarantine and return to normal life****"*

4.5.4 Perceived Barriers

Perceived Barriers to taking action often influence health-related behaviours. The term "perceived barriers" refers to a person's perception of the barriers to behaviour change. Even if a person perceives a health condition as dangerous and believes that taking a specific action will effectively reduce the danger, barriers can prevent them from engaging in the health-promoting conduct. To put it another way, in order for behaviour change to occur, the perceived benefits must outweigh the perceived obstacles[79].

4.5.4.1 Stress, Anxiety and Depression (mental health)

Due to the uncertainty of what will happen, people were highly stressed and anxious about the virus as it rapidly changed the way we work socialize and live. Also, it had led to more people feeling depressed as they are restricted and have limited interactions with other people.

Example of sample comments:

"it is ***not easy to stay at home with anxiety depression migraine overthinking*** *mind and then facing all family members"*

"as a person with disabilities i was subject to social distancing and isolation before the coronavirus is not helping the fact that one ***of my anxiety triggers is germs been feeling stressed****"*

4.5.4.2 Lack of Freedom/Restrictions

The COVID-19 has brought about a lot of preventive measures to slow the spread of including lockdown, which has resulted in a lot of restrictions that made people lose their freedom in how they do things and move. See sample comments below:

> *"the emperors best weapons bread circus fear* ***more people died fighting for freedom*** *than will ever die from covid19 even once you have given your freedoms away"*
>
> *"****rather die from the coronavirus knowing that we preserved freedom*** *than to survive the coronavirus only to totally lose all of our liberty and freedom"*

4.5.4.3 Increased risk of Domestic Abuse and Violence

Research shows that people are scared of the risk of getting abused, hence the restriction is harder on them. Some people suffer abuse and neglect, and some kids are also not safe at home. Hence the women will be at increased risk of domestic abuse and violence when they are forced to stay at home, which might lead to breaking the preventive measure. See comments below:

> *"the very conditions needed to battle the* ***disease isolation, social distancing restrictions on freedom of movement are the very conditions that feed into the hands of abusers*** *who now find state sanctioned circumstances tailormade for unleashing abuse."*
>
> *"we are all being asked to hunker down to slow the spread of covid19 but what do you do* ***when staying home means running the risk of getting abused thats the reality many are facing*** *across the country right now and local advocates are worried hear from them tonight."*
>
> *"women will be at* ***increased risk of domestic abuse and violence when they are forced to stay at home*** *and police should be sensitive to their particular circumstances women who are victims of abuse should not end up being arrested."*

4.5.4.4 Inadequate Medical Attention / Affecting health services

Public Health Measures have resulted in limited access to medical attention, denying people the care they require. Only emergency surgery is performed in the medical field. As seen in the sample comment, women are denied constitutionally protected rights to abortion. Pregnant women in the pandemic go to their prenatal appointment without their partner and must wait until they are close to delivery before their partner can enter the delivery room. Due to covid hospitals are running out of a lot of things and there is decline in blood donation. Comments are shown below:

> *"geoffrey when did seasonal flu or pneumonia last* ***overwhelm health services*** *and when did the health services* ***last run out of icu beds and ventilators****"*

> "*help karachi due to panic situation because of coronavirus there is* ***decline in blood donation and therefore thalassemia patients are facing difficulties*** *people are requested to donate blood to save lives."*

4.5.4.5 Economic losses

Based on our findings the economy suffered great losses and investors were offering less to none, due to the impact of COVID19 affecting business, stock exchange market. See comments below:

> *"sam levey on who in the end will have to answer for the* ***economic losses incurred during the coronavirus pandemic****"*

> *"the golden leaf foundation is providing million in funding to launch a rapid recovery loan program in response to* ***economic losses for small businesses related to covid19*** *brooks pierce attorneys provide more information in this client alert"*

> *"****my debt portfolio is down by or so reasons business and economic losses expected due to covid19*** *leveraged investors desperately selling fixed income investments to pay for margin calls spreads widening"*

4.5.5 Cues to Action

As a cue prompts engagement in health-promoting behaviors. Cue to action is linked with perceived susceptibility, seriousness, benefits, and barriers as these factors are needed for an action to occur or that will prompt the action. As it links to the rest of the constructs, most cues were either alerted through the severity of the situation like death of an individual or loved one (Perceived Severity) or increase in cases (Perceived Susceptibility). Mostly Perceived Susceptibility, Perceived Severity and Perceived benefits influences Cues. Most people encouraged by reminding people of the Preventive Health Behaviors (PHB's) and benefits of their action.

4.5.5.1 Increase in cases

Reminding people to carry out the preventive measures due to high increase in cases. This action is driven by Perceived susceptibility. see sample comments below:

> *"in the wake of* ***increased cases of covid19 gov mr udom emmanuel has issued a stay-at-home order for week*** *with effect from monday march in a letter signed by ssg dr emmanuel ekuwem all borders amp roads leading in and out of the state are closed with immediate effect"*
>
> *"the united states just overtook china for the most confirmed cases of covid19 in the world, that did not take long for that to happen****, stayhome and save lives, flattenthecurve before the healthcare system is overrun****"*

4.5.5.2 Loss of loved ones/infection of closed relation

A family who has dealt with loss of a loved one due to the pandemic has seen the real impact of how the COVID19 virus has snatched a life. This enables them to create awareness and encourage behavioural change such as wearing mask, sanitizing of hands to combat the virus. Sample comments are shown below:

"a friend of a friend died today from covid19 he was years young ***this is real folks social distancing and really isolation*** *aint a snow day what you dont know youre carrying can kill do the right thing,* ***pls stay home people****."*

"i have a dear friend that said goodbye to his dad this week via facetime because he was like a lot of us and did not take covid19 seriously and took a trip he had booked he got sick got quarantined and died win a week ***just stay home you guys tell everyone you care about****"*

4.5.5.3 Self – reporting

Research showed that self – reporting cases will go a longer way in slowing down the spread of the disease as it will be detected earlier. sample comments are shown below:

*"****please share help researchers slow the spread of covid19*** *and identify at risk cases sooner by* self-reporting your symptoms *daily even if you feel well download the app"*

self-report daily help slow the *outbreak identify those at risk sooner we all can contribute to this and help scientists identify high risk areas who is most at risk and where it is spreading fastest*

4.5.5.4 Posters

Most public places have posters stating the public guidelines measures to stay safe and protect them from the virus, which serves as a reminder to the public. Comments are shown below:

"keep your distance save a life harry kelly aged from omagh wants to keep his family and friends safe and has ***designed these posters for the western trust to help get the message out there****"*

"hi friends i am going to be making ***a few posters this week to encourage people to stay home*** *if we want this virus to be over sooner* ***stay home as much as possible*** *dont gather i want us all to make it through this and i want it to be over quickly"*

4.5.5.5 Symptoms

Symptoms are an indication that alerts you that you might be sick. Research has shown that this is a fast cue to people or when another person is showing the symptoms, it alerts one to take safety measures. Sample comments are below:

> ***"if you are sick with mild to moderate covid19 symptoms stay home*** *contact your doctor after days if you are not feeling better dont try to get tested and dont go to the er by staying home you reduce the possibility of transmission to others"*
>
> *"everyone in my house is sick not go to the er sick but feel like total crap, have fevers but coughs congestion headaches last place i will just* ***be more careful and stayhome for a while****"*

4.5.5.6 COVID-19 management Apps

To help flatten the curve, different research has being ongoing, which has aided in developing different apps targeted at COVID19. sample comments are shown below:

> *"for those in this* ***app has just been launched with researchers*** *at guys and st thomas amp kings college* ***to help slow the spread of covid19 and identify at risk cases sooner*** *by selfreporting your symptoms daily even if you feel well download the app"*
>
> *"****handy healthy app that can help slow*** *the spread of covid19, even if we feel well download the app smart".*

4.5.5.7 Increase in death rate

Increase in death rate is a cue that people need to follow the health guideline in order to stay protected from the virus. This mostly prompts people to carry out the preventive measures. Sample comments are shown below:

> *"even if you feel you are invincible you can transmit covid to someone at high risk with medical systems worldwide breaking down the potential for* ***death increases*** *with each passing week as someone w two high risk people in my household* ***i beg you pls stay home****"*

"... coronavirus figures today deaths are up again daily toll up from to today but daily increase in infections falls again now down ***to a tragic increase in deaths, we really have to play our part by staying home****."*

4.5.6 Perceived self-efficacy

According to our findings, people's confidence was shown by believing they could fight this and by also encouraging people to carry out the preventive health measures with the believe that it can be fought together. This belief is driven by one's confidence. Because coronavirus is a pandemic and the PHMs put in a place is a "we thing", most Pself-efficacy was shown by generalizing their confidence as shown in the example below.

4.5.6.1 Coming together to achieve the goal

People showed believe in beating the virus by coming together to achieve this, as this is understandable as it's a pandemic and a contagious virus, which means it can be reduced by an individual effort, hence most people showed efficacy by including everyone. Sample comments are shown below:

*"****we all need to take steps to slow the spread*** *of the coronavirus tackling coronavirus is a national effort we will get through this but we need to take steps now to make sure that we limit the scope and impact of the virus here are some things we can do now to help ourselves"*

"can help break the chain of the spread covid19 let us practice this simple and effective habit ***together we can fight****."*

*"****we believe that together we can make a greater impact*** *during this hard time we are all facing here is a message from davidcoulthardf on behalf of the yianis christodoulou foundation we hope you are all staying safe and healthy in this challenging time"*

4.5.6.2 Reduction in cases (Precautionary measures working)

The fact that it was seen that social distancing was working in another country with reduced rate of infections increased the self-efficacy of people and gave them hope. Sample comments are shown below:

> *"this is what social distancing is doing we need to flatten the curve social distancing* ***physical distancing appears to be working*** *in Canada please continue to physical distance do not let up* ***we can beat****"*

> *"do you see* ***the dip in the curves its working*** *just keep going and we may not have to extend this thing after days* ***we can do this people stay*** *strong stay calm stay home"*

4.5.6.3 Leading by example

Self – Efficacy was also shown by leading by example. It was believed that by doing, it should encourage or motivate people. Sample comment is shown below:

> *"I am an emergency physician. My increased exposure means that I have chosen to isolate from my family, to keep them safe. This is how I see my daughters (pictured with their cousins, with whom they're staying).* ***If I can do this, you can stay home****."*

> *"i m down when pm said these words it is our duty now that we have to fight for our country and for our family dont panic only do one thing* ***stay home like I am doing****"*

4.5.7 Social Norm

In line with our research, due to the pandemic, our regular norms like handshake and hugging changed. Due to the preventive measures put in place new norms arose, examples of such norms are, wearing mask, keeping a distance of 6ft etc. They were mixed feelings about these norms which have their positive and negative side to it.

4.5.7.1 Lack of believe in the preventive measures (loss of hope)

Some people were not observing the precautionary measures, due to fear that it will affect everyone and another individual disbelieving as result of this act. Sample comment is shown below:

"this so-called social distancing is a load of crap no one is observing it and while i am working ***people are hugging their friends*** *and joking that we are all going to get covid anyway so what is the point shut the city down"*

"went to Costco cashiers ***are not wearing mask*** *at all asked few employees they said* ***mask won't prevent coronavirus*** *cant believe what i heard told them its airborne too so sad and puzzled why shoprite is the same too dont know with Walmart"*

4.5.7.2 Kids not taking it seriously

Kids are struggling to understand the importance of the health guidelines as they do not understand the severity of the virus and other kids are seeing their friends not performing this behaviour, which makes them to easily break the rule. Sample comment is shown below:

"*just drove past a school kids out for recess there was* ***no social distancing in fact children were hugging each other*** *and staff were standing side by side talking to each other that is why we need to"*

"so the ***grandchildren are going to*** *beaches bars swimming pools licking produce and products* ***not caring if they get the virus*** *or spread it are the inconsiderate…"*

4.5.8 Trust

Trust is crucial during an infectious disease outbreaks, as shown by studies and Trust building could influence perceived severity and transparency, also willingness to adopt interventions such as physical distancing and personal hygiene [17].

4.5.8.1 Lack of Trust in Authorities

Every country has handled COVID19 differently and given people a reason to know what comes from the Government. Most authorities have lied and withheld information which makes the public not to trust and question the government a lot as they are not getting valid responses. Sample comment is shown below:

> *"together we can have an impact on what happens in sask but only if **our government is honest about our readiness gives us as much information as possible** and acts quickly to slow the spread"*

> *"with the coverup in december and january we really **cannot trust the coronavirus numbers from the chinese government** without more credible and solid evidence to verify ..."*

4.5.8.2 Misinformation

Research shows that there has been a lot of hidden information and retraction this COVID period. Sample comments are shown below:

> "***rumors and misinformation about coronavirus** can cause harmful behaviors that increase personal amp public health risks that is why health allegheny has created a page dedicated to addressing rumors about covid19"*

> *"what is sad is all deaths will be blamed on covid19 and in reality they are not all dying from that but looks like its easier to say caronavirus killed everyone rather than knowing the real cause of death"*

> *"the chinese premier has warned local govts not to cover up new cases of covid19 as low daily rates of infection have led to relaxing travel restrictions in hubei province only one reason why the reported chinese data is dubious"*

4.5.9 Summary of the observations from thematic Analysis

During the analysis processed, we discovered that PSusceptibility and PSeverity mostly went together in discussions, they were used to interchangeably. To help curb the spread most people took different preventive measure to assist in their own way through donation, giving different examples on the best practices of social distancing, printing out instructions and labels. Cue to action and perceived benefits mostly fed off each other or balanced out each other as most individuals talked about the PBenefits by reminding individual of the preventive health measures. Therefore, they use each other to promote and influence health behaviours, it was also noticed that individuals showed PSelfefficacy by promoting PBenefits and Cue to Action as they are used for motivation. Pselfefficacy and Cue to Action also work together as in selfefficacy most people showed confidence by performing the action or ensuring that we can beat the virus while also reminding users of the preventive health behaviours. Perceived self-efficacy was mostly shown as a together, it was mostly referred to as a "we thing" or "together thing" because in order to beat the virus everyone has to play their part and in this case the spread can't be slowed by only few people performing the action.

CHAPTER 5 CONCLUSION AND FUTURE WORK

In this chapter, we summarized the thesis and highlight the limitations, contributions, and suggested potential directions for future work.

5.1 Conclusion

In line with our objective as mentioned earlier in chapter 1, this study evaluated and compared three state of the art ML algorithms for mapping COVID19–related social media discussions to the constructs of health behavior theories such as the HBM, Social Norm and Trust to understand people's reactions to COVID-19. We also evaluated and compared two types of ML classifications namely multiclass (single labeling) and multi labeling classification, which both had good performance with each classifier having an accuracy above 80 percent.

Our contribution to this study as stated in chapter 1, are: we showed that our ML text classifiers successfully yielded accurate classifications of COVID-19 tweeter comments using the HBM constructs, Social Norm and Trust in the context of people reactions to COVID-19. This further demonstrates the potential for developing Machine Learning prediction systems to classify big data from social media using behavioral models and frameworks. Afterwards, we conducted thematic analysis of each comments to analyze the difference public perceptions and the result showed different themes which buttress public concerns regarding COVID-19, which shows a clearer understanding as to why and what prompts certain behaviours based on the constructs from the behaviour theories. It was also interesting to see how the constructs drive each other. For example, Cue to Action and Perceived Benefits are mostly depicted in the same sentence to promote positive behaviour as they are both positive constructs.

The findings demonstrated the utility of the behavior theories in understanding public perception or opinions in regard to COVID-19. It also helps in providing a sound theoretical basis for future public health messaging and for rapidly measuring and assessing the effects of such messaging

and programs as there is an urgent need to improve health promotion and information campaigns to enhance the benefits and reduce the barriers to COVID-19 preventive measures.

In particular, this study suggests that methods for addressing COVID-19 concerns may benefit from targeting concerns about perceived barriers (including psychological barriers such as mental health and emotional problems), Social Norm and Trust related factors to carry out the preventive health measures.

5.2 Limitations

There are certain limitations in this study, which are as follows:

- Different geolocations were not taken into account as this can also have an impact on the result.
- Different age groups were not considered.
- Other social media platforms were not used in this study.

5.3 Future Works

The following summarizes the potential future work directions:

- This study can be extended to other health behaviour theories to analyze public opinions.
- More experiments can be carried out using neural network as is a more complex model.
- Further experiment can be done with other traditional ML classifiers.
- Further experiment can be done using different geolocations to see how it affects the result.
- It can also be applied to other social media platforms.
- Data balancing can be performed on the multilabel classification task to see if it will alter the result.

Bibliography

[1] Abdar, M. et al. 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*. 179, (Oct. 2019), 104992. DOI:https://doi.org/10.1016/j.cmpb.2019.104992.

[2] Ajzen, I. 2011. The theory of planned behaviour: Reactions and reflections. *Psychology and Health*.

[3] Almeshal, A.M. et al. 2020. Forecasting the Spread of COVID-19 in Kuwait Using Compartmental and Logistic Regression Models. *Applied Sciences*. 10, 10 (May 2020), 3402. DOI:https://doi.org/10.3390/app10103402.

[4] Aly, M. 2005. *Survey on Multiclass Classification Methods*.

[5] Aramaki, E. et al. 2011. *Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter*.

[6] Bakogianni, G.D. et al. 2010. HPV vaccine acceptance among female Greek students. *International Journal of Adolescent Medicine and Health*. 22, 2 (2010), 271–273. DOI:https://doi.org/10.1515/IJAMH.2010.22.2.271.

[7] Balog-Way, D.H.P. and McComas, K.A. 2020. COVID-19: Reflections on trust, tradeoffs, and preparedness. *Journal of Risk Research*. 23, (2020), 1–11. DOI:https://doi.org/10.1080/13669877.2020.1758192.

[8] Balog-Way, D.H.P. and McComas, K.A. 2020. COVID-19: Reflections on trust, tradeoffs, and preparedness. *Journal of Risk Research*. 23, (2020), 1–11. DOI:https://doi.org/10.1080/13669877.2020.1758192.

[9] Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*. 84, 2 (Mar. 1977), 191–215. DOI:https://doi.org/10.1037/0033-295X.84.2.191.

[10] Bavel, J.J.V. et al. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*. Nature Research.

[11] Biyani, P. et al. 2013. Co-training over Domain-independent and Domain-dependent features for sentiment analysis of an online cancer support community. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013* (New York, NY, USA, Aug. 2013), 413–417.

[12] Blockeel, H. et al. 2002. *Efficient Algorithms for Decision Tree Cross-validation*.

[13] Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993 | Semantic Scholar: *https://www.semanticscholar.org/paper/Book-Review%3A-C4.5%3A-Programs-for-Machine-Learning-by-Salzberg/3ecdaaa55313520b50ae17de9f4f6650403754a3*. Accessed: 2021-04-18.

[14] Boon-Itt, S. and Skunkan, Y. 2020. Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*. (2020). DOI:https://doi.org/10.2196/21978.

[15] Brinati, D. et al. 2020. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems*. 44, 8 (Aug. 2020), 1–12. DOI:https://doi.org/10.1007/s10916-020-01597-4.

[16] Brug, J. et al. 2009. Risk perceptions and behaviour: Towards pandemic control of emerging infectious diseases: Iional research on risk perception in the control of emerging infectious diseases. *International Journal of Behavioral Medicine*. Springer.

[17] Building Communication Capacity to Counter Infectious Disease Threats ... - National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Global Health, Forum on Microbial Threats - Google Books: *https://books.google.ca/books?hl=en&lr=&id=M1ksDwAAQBAJ&oi=fnd&pg=PR1&ots=Ia7eDiruzK&sig=MR38deqYkLwaPBOpdnwuCHCr9Xg&redir_esc=y#v=onepage&q&f*

=*false*. Accessed: 2021-03-15.

[18] Cairns, G. et al. 2013. Reputation, relationships, risk communication, and the role of trust in the prevention and control of communicable disease: A review. *Journal of Health Communication*. Taylor & Francis Group .

[19] Cambridge Dictionary | English Dictionary, Translations & Thesaurus: *https://dictionary.cambridge.org/*. Accessed: 2021-04-19.

[20] Cheng, X. et al. 2017. Understanding trust influencing factors in social media communication: A qualitative study. *International Journal of Information Management*. 37, 2 (Apr. 2017), 25–35. DOI:https://doi.org/10.1016/j.ijinfomgt.2016.11.009.

[21] Chire Saire, J.E. 2020. Characterizing Twitter Interaction during COVID-19 pandemic using Complex Networks and Text Mining. *arXiv*.

[22] Chire Saire, J.E. and Briseño, A.P. 2020. Text Mining Approach to Analyze Coronavirus Impact: Mexico City as Case of Study. *medRxiv*.

[23] Chire Saire, J.E. and Cruz, J.F.O. 2020. Study of coronavirus impact on Parisian population from April to June using Twitter and Text Mining approach. *medRxiv*.

[24] Christakis, N.A. and Fowler, J.H. 2013. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*. 32, 4 (Feb. 2013), 556–577. DOI:https://doi.org/10.1002/sim.5408.

[25] Christakis, N.A. and Fowler, J.H. 2010. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*. 5, 9 (Sep. 2010), e12948. DOI:https://doi.org/10.1371/journal.pone.0012948.

[26] Cialdini, R.B. and Goldstein, N.J. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology*. 55, 1 (Feb. 2004), 591–621. DOI:https://doi.org/10.1146/annurev.psych.55.090902.142015.

[27] Cinelli, M. et al. 2020. The COVID-19 social media infodemic. *Scientific Reports*. (2020). DOI:https://doi.org/10.1038/s41598-020-73510-5.

[28] Cislaghi, B. et al. 2016. *Values deliberation & collective action: Community empowerment in rural Senegal*. Springer International Publishing.

[29] Cislaghi, B. and Heise, L. 2018. Theory and practice of social norms interventions: Eight common pitfalls. *Globalization and Health*. 14, 1 (Aug. 2018), 83. DOI:https://doi.org/10.1186/s12992-018-0398-x.

[30] Costa, M.F. 2020. Health belief model for coronavirus infection risk determinants. *Revista de Saude Publica*. 54, (2020). DOI:https://doi.org/10.11606/S1518-8787.2020054002494.

[31] Crijns, T.D.L.G.V.D.F.G.D.S.T. 2016. *TEXT CLASSIFICATION CLASSIFYING EVENTS TO UGENDA CALENDAR GENRES*.

[32] Daniel, J. and Martin, J.H. 2020. *Speech and Language Processing*.

[33] Dempsey, R.C. et al. 2018. A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. *Frontiers in Psychology*. Frontiers Media S.A.

[34] Devine, D. et al. 2020. Trust and the Coronavirus Pandemic: What are the Consequences of and for Trust? An Early Review of the Literature. *Political Studies Review*. (Aug. 2020), 147892992094868. DOI:https://doi.org/10.1177/1478929920948684.

[35] Dickie, R. et al. 2018. The effects of perceived social norms on handwashing behaviour in students. *Psychology, Health & Medicine*. 23, 2 (Feb. 2018), 154–159. DOI:https://doi.org/10.1080/13548506.2017.1338736.

[36] Diop, N. et al. 2008. *Evaluation of the long-term impact of the TOSTAN programme on the abandonment of FGM/C and early marriage: Results from a qualitative study in Senegal*.

[37] Dodel, M. and Mesch, G. 2017. Cyber-victimization preventive behavior: A health belief

model approach. *Computers in Human Behavior*. 68, (Mar. 2017), 359–367. DOI:https://doi.org/10.1016/j.chb.2016.11.044.

[38] Donadiki, E.M. et al. 2014. Health Belief Model applied to non-compliance with HPV vaccine among female university students. *Public Health*. 128, 3 (Mar. 2014), 268–273. DOI:https://doi.org/10.1016/j.puhe.2013.12.004.

[39] Du, J. et al. 2020. Use of Deep Learning to Analyze Social Media Discussions About the Human Papillomavirus Vaccine. *JAMA network open*. 3, 11 (Nov. 2020), e2022025. DOI:https://doi.org/10.1001/jamanetworkopen.2020.22025.

[40] Durham, D.P. et al. 2012. Deriving Behavior Model Parameters from Survey Data: Self-Protective Behavior Adoption During the 2009-2010 Influenza A(H1N1) Pandemic. *Risk Analysis*. 32, 12 (Dec. 2012), 2020–2031. DOI:https://doi.org/10.1111/j.1539-6924.2012.01823.x.

[41] Elwalid Fadul Nasir 1 , Ahmed Khalid Elhag 2, H.M.A. 3 2020. COVID-19 Perceptional Disparity Among Dental Healthcare Personnel at King Faisal University: Applying Health Belief Model. *European Journal Of Dentistry*. (2020).

[42] Essential libraries for machine learning and Data science in python | by Siddharth Dash | Medium: *https://medium.com/@dash.siddharth01/essential-libraries-for-machine-learning-and-data-science-in-python-fafd06137988*. Accessed: 2021-04-18.

[43] Fernando Garcia-Constantino, M. 2013. *On The Use Of Text Classification Methods For Text Summarisation*.

[44] Fridman, D. et al. 2011. Predictors of H1N1 vaccination in pregnancy. *American Journal of Obstetrics and Gynecology*. Mosby.

[45] Gavin, H. 2012. *Understanding Research Methods and Statistics in Psychology*. SAGE Publications Ltd.

[46] Ghamrawi, N. and McCallum, A. 2005. Collective multi-label classification. *International*

Conference on Information and Knowledge Management, Proceedings (New York, New York, USA, 2005), 195–200.

[47] Glowacki, E.M. et al. 2021. Identifying #addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis. *Substance Abuse*. (2021). DOI:https://doi.org/10.1080/08897077.2020.1822489.

[48] Gohokar, I. et al. 2014. Multi-Class Tweet Categorization Using Map Reduce Paradigm Tweet categorization using Hadoop View project Gene Regulation View project Multi-Class Tweet Categorization Using Map Reduce Paradigm. *Article in International Journal of Computer Trends and Technology*. 9, 2 (2014). DOI:https://doi.org/10.14445/22312803/IJCTT-V9P117.

[49] Guidry, J.P.D. et al. 2019. Using the health belief model to analyze instagram posts about Zika for public health communications. *Emerging Infectious Diseases*. Centers for Disease Control and Prevention (CDC).

[50] HEALTH BEHAVIOR AND | Valentina Suárez - Academia.edu: *https://www.academia.edu/41912708/HEALTH_BEHAVIOR_AND*. Accessed: 2021-03-16.

[51] Health Behavior and Health Education: Theory, Research, and Practice - Google Books: *https://books.google.ca/books/about/Health_Behavior_and_Health_Education.html?id=1xuGErZCfbsC&redir_esc=y*. Accessed: 2021-03-15.

[52] Health Behavior and Health Education: Theory, Research, and Practice - Google Books: *https://books.google.ca/books?id=1xuGErZCfbsC&pg=PT116&lpg=PT116&dq=Hochbaum+(1958),+for+example,+thought+that+readiness+to+take+action+(perceived+susceptibility+and+perceived+benefits)+could+only+be+potentiated+by+other+factors,+particularly+by+cues+to+instigate+action,+such+as+bodily+events,+or+by+environmental+events,+such+as+media+publicity&source=bl&ots=-q297I3-3p&sig=ACfU3U3lcmc5PcIdUmkJfWh1i9CXeoVpfA&hl=en&sa=X&ved=2ahUKEwj02cz7irDvAhWVRTABHcIeDqEQ6AEwAnoECAQQAw#v=onepage&q=Hochbaum*

(1958)%2C for example%2C thought that readiness to take action (perceived susceptibility and perceived benefits) could only be potentiated by other factors%2C particularly by cues to instigate action%2C such as bodily events%2C or by environmental events%2C such as media publicity&f=false. Accessed: 2021-03-16.

[53] Huber, B. et al. 2019. Fostering public trust in science: The role of social media. *Public Understanding of Science*. 28, 7 (Oct. 2019), 759–777. DOI:https://doi.org/10.1177/0963662519869097.

[54] Janz, N.K. and Becker, M.H. 1984. The Health Belief Model: A Decade Later. *Health Education & Behavior*. (1984). DOI:https://doi.org/10.1177/109019818401100101.

[55] Jones, C.L. et al. 2015. The Health Belief Model as an Explanatory Framework in Communication Research: Exploring Parallel, Serial, and Moderated Mediation. *Health Communication*. 30, 6 (Jun. 2015), 566–576. DOI:https://doi.org/10.1080/10410236.2013.873363.

[56] Jose, R. et al. 2021. Public perception and preparedness for the pandemic COVID 19: A Health Belief Model approach. *Clinical Epidemiology and Global Health*. 9, (Jan. 2021), 41–46. DOI:https://doi.org/10.1016/j.cegh.2020.06.009.

[57] Juntasopeepun, P. et al. 2012. Factors influencing acceptance of human papillomavirus vaccine among young female college students in Thailand. *International Journal of Gynecology and Obstetrics*. 118, 3 (2012), 247–250. DOI:https://doi.org/10.1016/j.ijgo.2012.04.015.

[58] Kang, A. et al. 2021. Environmental management strategy in response to COVID-19 in China: Based on text mining of government open information. *Science of the Total Environment*. (2021). DOI:https://doi.org/10.1016/j.scitotenv.2021.145158.

[59] Kit Tong, K. et al. 2020. Adherence to COVID-19 Precautionary Measures: Applying the Health Belief Model and Generalised Social Beliefs to a Probability Community Sample. *APPLIED PSYCHOLOGY: HEALTH AND WELL-BEING*. 2020, 4 (2020), 1205–1223.

DOI:https://doi.org/10.1111/aphw.12230.

[60] Kowsari, K. et al. Text Classification Algorithms: A Survey. DOI:https://doi.org/10.3390/info10040150.

[61] Kumar Sethy, P. et al. 2020. *Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine*. Preprints.

[62] langdetect · PyPI: *https://pypi.org/project/langdetect/*. Accessed: 2021-03-30.

[63] Lin, Y. et al. 2020. Understanding COVID-19 vaccine demand and hesitancy: A nationwide online survey in China. *PLOS Neglected Tropical Diseases*. 14, 12 (Dec. 2020), e0008961. DOI:https://doi.org/10.1371/journal.pntd.0008961.

[64] Lindsey, N. and Regina, G. 2015. *THE HEALTH BELIEF MODEL AND WOMEN'S ADHERENCE TO A CARDIAC REHABILITATION PROGRAM*.

[65] Lo, S.W.S. et al. 2015. Factors associated with health-promoting behavior of people with or at high risk of metabolic syndrome: Based on the health belief model. *Applied Nursing Research*. (2015). DOI:https://doi.org/10.1016/j.apnr.2014.11.001.

[66] Lopez, C.E. et al. *UNDERSTANDING THE PERCEPTION OF COVID-19 POLICIES BY MINING A MULTILANGUAGE TWITTER DATASET*.

[67] LUCKERT, M. and SCHAEFER-KEHNERT, M. 2015. *Using Machine Learning Methods for Evaluating the Quality of Technical Documents*.

[68] Lyu, J.C. and Luli, G.K. 2021. Understanding the Public Discussion about the CDC in the COVID-19 Pandemic: A text-mining analysis of Twitter data. *Journal of medical Internet research*. (2021). DOI:https://doi.org/10.2196/25108.

[69] Marlow, L.A.V. et al. 2009. Predictors of interest in HPV vaccination: A study of British adolescents. *Vaccine*. 27, 18 (Apr. 2009), 2483–2488. DOI:https://doi.org/10.1016/j.vaccine.2009.02.057.

[70] Mayer, R.C. et al. 1995. AN INTEGRATIVE MODEL OF ORGANIZATIONAL TRUST. *Academy of Management Review*. 20, 3 (Jul. 1995), 709–734. DOI:https://doi.org/10.5465/amr.1995.9508080335.

[71] McInnes, C.J. and Hornmoen, H. 2018. 'Add twitter and stir': The use of twitter by public authorities in Norway and UK during the 2014-15 ebola outbreak. *Observatorio*. 12, 2 (2018), 23–46. DOI:https://doi.org/10.15847/obsobs12220181173.

[72] Mehta, P. et al. 2013. Designing and evaluating a health belief model-based intervention to increase intent of HPV vaccination among college males. *International Quarterly of Community Health Education*. 34, 1 (Jan. 2013), 101–117. DOI:https://doi.org/10.2190/IQ.34.1.h.

[73] Meier, K. et al. 2020. Public perspectives on protective measures during the COVID-19 pandemic in the Netherlands, Germany and Italy: A survey study. *PLoS ONE*. Public Library of Science.

[74] Meyer, D. 2014. Support vector machines: the interface to libsvm in package e1071. ... *Systems and their* ... 1, (2014), 1–8. DOI:https://doi.org/10.1007/978-0-387-77242-4.

[75] Miller, D.T. and Prentice, D.A. 2016. Changing Norms to Change Behavior. *Annual Review of Psychology*. 67, 1 (Jan. 2016), 339–361. DOI:https://doi.org/10.1146/annurev-psych-010814-015013.

[76] Mo, P.K.H. et al. 2019. Can the Health Belief Model and moral responsibility explain influenza vaccination uptake among nurses? *Journal of Advanced Nursing*. 75, 6 (Jun. 2019), 1188–1206. DOI:https://doi.org/10.1111/jan.13894.

[77] Mukhtar, S. 2020. Mental health and emotional impact of COVID-19: Applying Health Belief Model for medical staff to general public of Pakistan. *Brain, Behavior, and Immunity*. Academic Press Inc.

[78] Nadkarni, P.M. et al. 2011. Natural language processing: An introduction. *Journal of the*

American Medical Informatics Association. Oxford Academic.

[79] Nugrahani, R.R. et al. 2017. Health Belief Model on the Factors Associated with the Use of HPV Vaccine for the Prevention of Cervical Cancer among Women in Kediri, East Java. *Journal of Epidemiology and Public Health*. 02, 01 (2017), 70–81. DOI:https://doi.org/10.26911/jepublichealth.2017.02.01.07.

[80] Oliinyk, V.-A. et al. 2020. *Propaganda Detection in Text Data Based on NLP and Machine Learning*.

[81] Park, J.H. et al. 2010. Perceptions and behaviors related to hand hygiene for the prevention of H1N1 influenza transmission among Korean university students during the peak pandemic period. *BMC Infectious Diseases*. 10, 1 (Jul. 2010), 222. DOI:https://doi.org/10.1186/1471-2334-10-222.

[82] Patterson, N.M. et al. 2018. Using the health belief model to identify communication opportunities to prevent Chagas disease in Southern Ecuador. *PLoS Neglected Tropical Diseases*. 12, 9 (Sep. 2018). DOI:https://doi.org/10.1371/journal.pntd.0006841.

[83] Perkins, H.W. et al. 2010. Effectiveness of social norms media marketing in reducing drinking and driving: A statewide campaign. *Addictive Behaviors*. 35, 10 (Oct. 2010), 866–874. DOI:https://doi.org/10.1016/j.addbeh.2010.05.004.

[84] Quinlan, J.R. 1986. Induction of decision trees. *Machine Learning*. 1, 1 (Mar. 1986), 81–106. DOI:https://doi.org/10.1007/bf00116251.

[85] Raamkumar, A.S. et al. 2020. Use of health belief model–based deep learning classifiers for COVID-19 social media content to examine public perceptions of physical distancing: Model development and case study. *JMIR Public Health and Surveillance*. 6, 3 (Jul. 2020). DOI:https://doi.org/10.2196/20493.

[86] Raamkumar, A.S. et al. 2020. Use of health belief model–based deep learning classifiers for COVID-19 social media content to examine public perceptions of physical distancing:

Model development and case study. *JMIR Public Health and Surveillance*. 6, 3 (Jul. 2020). DOI:https://doi.org/10.2196/20493.

[87] Rahmati-Najarkolaei, F. et al. 2015. Factors predicting nutrition and physical activity behaviors due to cardiovascular disease in Tehran university students: Application of health belief model. *Iranian Red Crescent Medical Journal*. 17, 3 (2015), 1–6. DOI:https://doi.org/10.5812/ircmj.18879.

[88] Rezaeijo, S.M. et al. 2021. Screening of COVID-19 based on the extracted radiomics features from chest CT images. *Journal of X-Ray Science and Technology*. 29, 2 (Jan. 2021), 229–243. DOI:https://doi.org/10.3233/XST-200831.

[89] Rezaeipandari, H. et al. 2018. *Investigation of Predictors of Preventive Behaviors of Influenza A (H1N1) Based on Health Belief Model among People of Jiroft City, (Iran)*.

[90] Rosenstock, I.M. et al. *Social Learning Theory and the Health Belief Model reprint requests to*.

[91] Rosenstock, I.M. 1977. The Health Belief Model and Preventive Health Behavior. *Health Education & Behavior*. (1977). DOI:https://doi.org/10.1177/109019817400200405.

[92] Saire, J.E.C. 2020. Data mining approach to analyze Covid19 dataset of Brazilian patients. *arXiv*.

[93] Schein, A.I. et al. 2007. Active learning for logistic regression: an evaluation. *Mach Learn*. 68, (2007), 235–265. DOI:https://doi.org/10.1007/s10994-007-5019-5.

[94] Shahnazi, H. et al. 2020. Assessing preventive health behaviors from COVID-19: a cross sectional study with health belief model in Golestan Province, Northern of Iran. *Infectious Diseases of Poverty*. 9, 1 (Dec. 2020), 157. DOI:https://doi.org/10.1186/s40249-020-00776-2.

[95] Shapiro, G.K. et al. 2017. Comparing human papillomavirus vaccine concerns on Twitter: A cross-sectional study of users in Australia, Canada and the UK. *BMJ Open*. 7, 10 (Oct.

2017), e016869. DOI:https://doi.org/10.1136/bmjopen-2017-016869.

[96] Sharma, S. and Bhagat, A. 2017. Data preprocessing algorithm for Web Structure Mining. *Proceedings on 5th International Conference on Eco-Friendly Computing and Communication Systems, ICECCS 2016* (Apr. 2017), 94–98.

[97] Shillitoe, R.W. and Christie, M.J. 1989. Determinants of self-care: The health belief model. *Journal of Interprofessional Care*. 4, 1 (1989), 3–17. DOI:https://doi.org/10.3109/13561828909043602.

[98] Short, S.E. and Mollborn, S. 2015. Social determinants and health behaviors: Conceptual frames and empirical advances. *Current Opinion in Psychology*. Elsevier.

[99] Silahtaroğlu, G. et al. *İşletme ve Maliye Araştırmaları Dergisi, Cilt 2, Sayı 3, s*.

[100] Singh, A.K. and Shashi, M. 2019. Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*. 10, 7 (2019), 305–310. DOI:https://doi.org/10.14569/ijacsa.2019.0100742.

[101] Smith, R.D. 2006. Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Social Science and Medicine*. 63, 12 (Dec. 2006), 3113–3123. DOI:https://doi.org/10.1016/j.socscimed.2006.08.004.

[102] Sokolova, M. et al. 2006. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report* (2006), 24–29.

[103] Son, Y.J. et al. 2010. Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Informatics Research*. 16, 4 (2010), 253–259. DOI:https://doi.org/10.4258/hir.2010.16.4.253.

[104] Spreading (Dis)Trust: Covid-19 Misinformation and Government Intervention in Italy | Lovari | Media and Communication: *https://www.cogitatiopress.com/mediaandcommunication/article/view/3219/3219*.

Accessed: 2021-03-18.

[105] Stoikos, S. and Izbicki, M. 2020. Multilingual Emoticon Prediction of Tweets about {COVID}-19. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media* (2020).

[106] Tanja Crijns Dr. L. G. Vuurpijl Dr. F. Grootjen Dhr. S. Trooster 2016. *TEXT CLASSIFICATION CLASSIFYING EVENTS TO UGENDA CALENDAR GENRES*.

[107] Tankard, M.E. and Paluck, E.L. 2016. Norm Perception as a Vehicle for Social Change. *Social Issues and Policy Review*. 10, 1 (Jan. 2016), 181–211. DOI:https://doi.org/10.1111/sipr.12022.

[108] Tax, D.M.J. and Duin, R.P.W. 2002. Using two-class classifiers for multiclass classification. *Proceedings - International Conference on Pattern Recognition* (2002), 124–127.

[109] Taylor, D. et al. 2006. A Review of the use of the Health Belief Model (HBM), the Theory of Reasoned Action (TRA), the Theory of Planned Behaviour (TPB) and the Trans-Theoretical. *London, UK: National …*. (2006).

[110] The construction of social norms and standards. - PsycNET: *https://psycnet.apa.org/record/1996-98402-026*. Accessed: 2021-04-07.

[111] Tong, K.K. et al. 2020. Adherence to COVID-19 Precautionary Measures: Applying the Health Belief Model and Generalised Social Beliefs to a Probability Community Sample. *Applied Psychology: Health and Well-Being*. 12, 4 (Dec. 2020), 1205–1223. DOI:https://doi.org/10.1111/aphw.12230.

[112] Top 10 Python Libraries for Data Science | by Rashi Desai | Towards Data Science: *https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266*. Accessed: 2021-04-18.

[113] Tsoumakas, G. and Katakis, I. 2007. Multi-Label Classification. *International Journal of Data Warehousing and Mining*. 3, 3 (Jul. 2007), 1–13.

DOI:https://doi.org/10.4018/jdwm.2007070101.

[114] Urich, A. *The Health Belief Model*.

[115] Vissandjée, B. et al. 2014. Female genital cutting (FGC) and the ethics of care: Community engagement and cultural sensitivity at the interface of migration experiences. *BMC International Health and Human Rights*. BioMed Central Ltd.

[116] Walrave, M. et al. 2020. Adoption of a contact tracing app for containing COVID-19: A health belief model approach. *JMIR Public Health and Surveillance*. 6, 3 (Jul. 2020), e20572. DOI:https://doi.org/10.2196/20572.

[117] Wang, H. et al. Using Tweets to Understand How COVID-19-Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study. DOI:https://doi.org/10.2196/26302.

[118] Wang, P. et al. 2020. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons and Fractals*. 139, (Oct. 2020), 110058. DOI:https://doi.org/10.1016/j.chaos.2020.110058.

[119] Wen, M. and Rosé, C.P. 2012. *Understanding Participant Behavior Trajectories in Online Health Support Groups Using Automatic Extraction Methods*.

[120] What Determines the Enforcement of Newly Introduced Social Norms: Personality Traits or Economic Preferences? Evidence from the COVID-19 Crisis: *https://www.researchgate.net/publication/345991562_What_Determines_the_Enforcement_of_Newly_Introduced_Social_Norms_Personality_Traits_or_Economic_Preferences_Evidence_from_the_COVID-19_Crisis*. Accessed: 2021-03-18.

[121] What is Python Used For? 10+ Coding Uses for the Python Programming Language.: *https://www.freecodecamp.org/news/what-is-python-used-for-10-coding-uses-for-the-python-programming-language/*. Accessed: 2021-04-18.

[122] Wong, M.C.S. et al. 2021. Acceptance of the COVID-19 vaccine based on the health belief

model: A population-based survey in Hong Kong. *Vaccine*. 39, 7 (Feb. 2021), 1148–1156. DOI:https://doi.org/10.1016/j.vaccine.2020.12.083.

[123] Wood, W. 2000. Attitude Change: Persuasion and Social Influence. *Annual Review of Psychology*. 51, 1 (Feb. 2000), 539–570. DOI:https://doi.org/10.1146/annurev.psych.51.1.539.

[124] Woodward, K. et al. 2020. LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. *Personal and Ubiquitous Computing*. 24, 5 (Oct. 2020), 709–722. DOI:https://doi.org/10.1007/s00779-020-01427-x.

[125] Wu, S. and Manber, U. 1992. Fast text searching. *Communications of the ACM*. 35, 10 (Oct. 1992), 83–91. DOI:https://doi.org/10.1145/135239.135244.

[126] Zhong, N. et al. 2012. Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*. (2012). DOI:https://doi.org/10.1109/TKDE.2010.211.

[127] Zhu, F. et al. 2013. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*.

[128] *Toward Better Health Care Service: Statistical and Machine Learning Based Analysis of Swedish Patient Satisfaction Survey YU WANG KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ELECTRICAL ENGINEERING*.