Perceiving Health Posts on Instagram

Akram Shaik **Abdul Rahman Shaikh**

Abstract:

Image sharing platforms are increasingly being used by social media users to communicate several different topics through images. Instagram is one such social media platform which has become extremely popular and is the third most used social media platform after Facebook and YouTube. This surge of usage produces a huge corpus of image dataset which can be analyzed to mine important topics related to different fields of science. In our work, we extract images related to three main topics - #covid, #health and #healthylifestyle from Instagram, filter the images by location, cluster the filtered images and generate topics related to the images through two processes i) topics on captions generated through COCO algorithm (machine generated) ii) topics on hashtags provided by users (user generated). We analyze these two sets of topics generated and find that machine generated output fails to identify the appropriate content of the image and produces a very generic output, whereas user generated output produces topics related to the hashtag used to extract the image. We also find that most of the data on Instagram comes from the US. The perception of the three main topics across different regions of the world differ in some ways but the overall topics remain the same such as healthy lifestyle is usually perceived as eating healthy, working out or maintaining a good diet irrespective of the location.

Introduction:

Social media has revolutionized forms of communication in the past decade with the number of social media users increasing from being 0.9 billion in 2010 to reaching 3.08 billion in 2020 [1] with almost one third of the world being logged on various social media platforms and 89% young adults visiting at least one social media site every day [2] sharing opinions, messages, tweets, images and videos to events happening around them providing opportunity for researchers to study and analyze these interactions to reveal hidden and important patterns which could help in several major fields such as marketing, psychology and health. Events occurring through the daily life of social media users are often posted on social media platforms such as Facebook, Twitter, Instagram, Snapchat, and YouTube. Currently Facebook and YouTube dominate social media platforms since a decade with Instagram users growing significantly over the last six years making Instagram the third most used social media platform [3]. Instagram has provided the platform to share images as an effective way of communication to record major events through posts or share everyday experiences via Instagram stories evolving social networking sites to become more visual centric. Facebook and Instagram also provide good marketing platforms where influencers and businesses can market products or services to their users, comparing the two platforms, marketers are shifting their focus from Facebook to Instagram due to its huge potential in reaching wider and appropriate audiences [4]. With users uploading a plethora of images on Instagram every day provides tons of visually rich data related to major fields that can be analyzed and used to improve aspects of life.

Instagram has redefined image sharing on social media platforms, starting as a photo posting app in 2010 to allowing influencers, content creators, and businesses reach customers all over the platform, this platform turns 10 years in 2020 and has 1 billion monthly active users [5] with around 100 million photos being shared everyday [6] It is the most popular image sharing platform with more than twice the number of users when compared to Twitter [3]. Users of Instagram visit the platform daily to post experiences from their lifestyle or keep up to date with people or accounts they follow. The popularity of Instagram has prompted 71% of US businesses to use Instagram [6] to promote their businesses or communicate their message, and around 83% users on the platform discover a new product or service [7]. Instagram has a huge popularity among the young adults and is a great sales channel for businesses targeting such groups of customers. An Instagram post can be an image or video uploaded by the user which consists of four parts – image, caption, hashtags, and location. The caption of the image is usually the description of the image or text the user associates the image with and hashtags are topics or keywords written with the symbol '#' that could be related to the image or emotions or moments the user relates the images through. Hashtags also allows users to look for similar images of a certain topic or keyword that interests the user. Interaction on Instagram allows users to like or comment on a picture or video providing users to show their love or acceptance of a post which can also provide a metric to indicate popularity of the account or likeness of the post.

The global pandemic, **COVID-19**, has changed the way of communication in many areas ranging from classes to office meetings being moved from in person to visual interactions. With the lockdown being imposed all over the world, people have less things to do and virtually no place to visit. Communication through social media has become a necessity to stay in touch with the world, family, and friends as well as to keep oneself entertained during these times. Social media platforms have become a source of providing and receiving health information related to COVID and has played a major role in communicating information to their users [8]. The usage of these platforms have increased rapidly in major countries such as the UK where the government is paying influencers to communicate about the coronavirus [9] and in India where everyday usage of platforms such as Instagram increased by 59% [10]. Instagram has also been reported to be the most popular platform for UK users with 67% of users increasing their use of Instagram along with making decisions to buy products or services [11]. #COVID-19 and #coronavirus was used by two-thirds of Instagram users to communicate virus-related information [12].Researchers have published datasets related to coronavirus from several social media platforms [13][14] and have also analyzed the datasets to understand the global reaction to the epidemic. With such an increase in usage of the platform regarding coronavirus, data shared on Instagram has huge potential for revealing hidden patterns of different topics as the data is personal and informal.

This surge in the usage of Instagram generates lots of user data which has huge potential to reveal patterns helpful in marketing on the platform or provide a source of data showing user sentiments or opinion towards an important topic or event occurring in the world. In this paper, we analyze posts from Instagram related to health, fitness, and the novel coronavirus around different parts of the world to understand the diversity of these topics in different major countries where Instagram is most popularly used. This could help us further explore the role of location in perceptions of users toward these topics. We focus on Instagram as it is increasingly being used over the world and less analysis has been done by researchers on Instagram when compared to other social media platforms. The analysis on health-related topics on Instagram is due to the implication of Instagram being a good source to provide information by the WHO [15] as well as several other health organizations. Fitness has often been motivated through Instagram by influencers and the platform has a large amount of data related to fitness topics [16].

These images from the posts on Instagram have not been analyzed before and could reveal important information. We try to answer the following questions –

1) How Instagram users from different countries perceive topics related to health, fitness, and the coronavirus?

2) Which location on Instagram is generating the most content related to health, fitness, and the coronavirus?

3) How does Machine generated output compare with user generated output?

To answer these questions, we extract posts from Instagram through hashtags related to health, fitness, and the coronavirus. For each of these three topics we characterize posts based on the countries provided by the user and choose locations with posts higher than a threshold of 20. For each country we cluster images based on the content of the image to group together similar images and output the topics of the images. We also analyze the hashtags provided by the user for the clustered images to understand the common topics posted by Instagram users from the same location. This analysis can provide data important for health marketers or businesses on Instagram to focus on topics most posted by the user or provide location-based marketing to target the appropriate audience. Sentiment analysis of users on Instagram related to the three topics can also help healthcare officials to provide the suitable message on the social media platform to communicate about diseases.

The rest of the chapter is organized as follows - section 2 provides the literature review conducted, section 3 describes the data used in the analysis, section 4 explains the proposed methodology, section 5 contains the results of the experiment and section 6 presents the discussion and conclusion.

Related Works:

Since its launch in 2010, Instagram has offered a platform for users to create, share and tag a huge number of images or videos. However, within the context of Instagram, a handful of studies have been conducted by researchers in comparison to Twitter and Facebook [17]. Some researchers have utilized Instagram posts as predictive features to gain insights in various fields. Reece and Danforth [18] crowdsourced Instagram data from 166 users extracting 43,950 photos using Amazon Mechanical Turk and identified markers of depression using human assessment and machine learning predictive models. They found that depression markers are observable in Instagram user posts and can be detected from user photos before the first diagnosis. There was low correlation between human ratings and the predictive models with the human ratings being weaker predictors of depression. Ferwerda and Tkalcic [19] conducted an online survey to assemble 193 Instagram users whose accounts were extracted to collect 54,962 pictures for predicting personality of users through visual and content features. They found that the features can be used individually to predict personality, but no additional value is added when they are combined. Habibi and Cahyo [20] extracted 40 days of 99,237 text captions data from Instagram posts related to #kopi to cluster users characteristics based on hashtags. Davainyte et al. [21] investigated the influence of Instagram posts on user's choice of restaurant and meal preparations at their homes. They find that more time spent by users on Instagram influences the users more and state that consumer behavior can be influenced by social media platforms.

Major organizations in the world use Instagram as a medium to disseminate information to its many users around the world, health communication on Instagram has found to be educational and useful with the ability to propagate rich health information to its users [22]. Instagram has the potential to be used for health communication with young people as a platform to collect information or provide experiences to users around the world. Alkazemi et al. [23] studied 1000 Instagram posts by the Gulf Cooperation Council Ministries of Health to analyze how health information is discussed by the Ministries of Health and found that there could be improvement in the form of communication by the GCC Ministries of Health on Instagram although the platform has the ability to provide communication in the Gulf countries. Lee et al. [24] performed content analysis on 758 Instagram posts containing #mentalhealth. Most of the posts were related to general practices like exercise or healthy diets and many posts also had advertisements for therapies. The authors suggest that Instagram has the potential to be a communication platform between health professionals and people suffering from mental health problems. Kim and H. Kim [25] explored posts uploaded by centers for disease control and prevention on Instagram by using Microsoft Azure Cognitive Services. They found that text was incorporated in photos to deliver a message by the account which was found that it can be an inefficient way of communicating evident with low engagement. Photos with more human faces and expression of emotions on faces had less engagement than other types of photos. Yakar et al. [26] conducted analysis on images related to neurosurgery finding that social networks can provide better consultations, increase awareness and post new findings providing a new platform for communication between patients and neurosurgeons as well as offer educational support. Wong et al. [27] provide the pros and cons of using Instagram in medicine studies and found that it has great potential to be an educational tool due to its interactive framework as it also allows for collaboration.

Data posted by users on Instagram create a large proportion of cultural visual data which can provide insights to understand cultural and diverse differences between users. Hochman and Manovich [28] compared visual signatures of 2.3 million Instagram photos collected from 13 global cities, they also analyzed social, cultural and political perceptions of users in Tel Aviv, Israel over the course of three months. Leaver and Highfield [29] used #ultrasound and #funeral to extract three months of data from Instagram in 2014 and analyzed the emotions of users to birth and death finding that people express their emotions such as grief more on Instagram than on other social media platforms. Chang [30] analyzed Instagram data from the US to study the cultural differences, trends and hot topics among different states and cities. Singh et al. [31] measure p-diversity using photos from Instagram geotagged in New York City to estimate the race, age, and gender of Instagram users from NYC. They compare this measure with census-based metrics to develop cheaper and faster methods for studying diversity using social media photos. Lalancette and Raynauld [32] analyzed Canadian Prime Minister Justin Trudeau’s feed on Instagram in his first year since being elected to join office. This study takes on a hybrid quantitative and qualitative approach for in-depth examination in the political image of Trudeau focusing on several elements. The analysis contributes to the image based political communication in Canada by politicians and celebrities on social media.

Due to a vast number of health-related topics posted on Instagram, researchers have identified various topics related to different fields of health. Muralidhara and Paul [33] applied a topic modeling approach to 96,426 Instagram posts related to #health characterizing them into 47 health related topics. They stated that Instagram can be a source of public health information due to a large and diverse set of health topics being discussed on the platform. They also found that most image tags are not the best indicator for identifying the image. Malighetti et al. [34] extracted 500 images related to body images detecting emotions using Microsoft Azure Cognitive Services. They found that happiness and neutrality were expressed and recognized in all images. Murashka et al. [35] performed content analysis on 2000 fitspiration related images to identify the objectification elements in the images and discovered topics in 35,263 user comments. They found that one-third of images were objectified and could distract users from health goals whereas health topics were related to inspiration and health motivation. Santarossa et al. [36] analyzed the #fitspo trend, posts to motivate towards a healthy lifestyle, on Instagram using 10,000 posts. They conducted content analysis along with text and network analysis on the images and text captions of the posts, respectively. They found that most of the posts were related to positive feelings and personal accounts were more popular than non-personal accounts. Cohen et al. [37] conducted a content analysis on 640 Instagram posts posted by body positive accounts finding that most posts had diverse body attributes with broad conceptualization of beauty indicating that Instagram could provide another perspective for young women towards body positivity.

Hashtags related to images given by the users provide a hint of relevance to the image content, researchers have analyzed this relevance to identify relevant tags of the image. Argyrou [38] applied topic modeling with LDA to explore related topics of 1000 images to remove irrelevant hashtags and evaluated them through crowdsourcing platforms, they discovered that relevant hashtags coincided best with matching topics of the images. Giannoulakis et al. [39] examined 1000 to explore the HITS algorithm for identifying the right tags in a crowdsourcing environment. The authors kept the top four ranked tags by the HITS algorithm to compare with benchmark data. They found that HITS for selecting appropriate hashtags has good potential to find the right tag. Giannoulakis and Tsapatsoulis [40] studied 1000 Instagram images to explore the descriptive power of hashtags provided by the user. They also conducted an online questionnaire allowing participants to choose tags for 20 images. They found that 66% of hashtag choices by the participants correlated with the hashtags provided by the user. Fiallos et al. [41] selected 7382 photos from a dataset of 905,418 posts related to #allyouneedisecuador and ran through Microsoft Cognitive services to retrieve visual description of images. They detected relevant topics using topic modeling and clustering finding that there was a low similarity between user descriptions and topics generated.

The increase in the usage of e-liquids and vaping in teenagers has encouraged researchers to analyze Instagram, a social media platform favorite among teenagers, in vaping related topics. Ketonen and Malik [42] extracted 560,414 images mentioning #vaping from Instagram. They extracted image features using deep learning models then clustered the images into seven categories related to e-cigarettes using VGG16 and VGG19. Vassey et al. [43] classified 49,655 Instagram image posts related to vaping using deep learning models into categories men, women, e-liquids, and vaping devices. They found that there were 40% images describing e-liquids, 30% vaping devices and 13% depicting humans. Czaplicki et al. [44] extracted 14,838 posts from 5201 Instagram users related to Juul, a vaping device popular among teenagers and young adults. They analyzed the textual data of the posts finding that most of the posts were related to lifestyle and one-third were promotional. Zhang et al. [45] classified 840 images from Instagram related to waterpipes or hookah, they used a CNN to extract features from images and SVM to classify the images into hookah or non hookah images. They observed an increase in the average level of accuracy when SVM and CNN are combined in comparison to other standard methods.

With the rise of social media, traditional marketing methods have changed into online social media marketing where most companies market their products through photos, videos, advertisements, or live streams. Instagram is a popular visual centric social media platform which allows all businesses to promote their products to a target group of customers. De Veirman et al. [46] conducted two experiments to understand marketing through Instagram influencers finding that marketing products using influencers with high number of followers could lower brand uniqueness and attitudes. They also state that influencers with most followers are most popular and more likeable. Pilgrim and Bohnet-Joschko [47] studied 1000 posts from Instagram influencers finding that influencers usually communicate non-campaign-driven health and produce good visual content to gain friendship and trust of their followers to communicate with them. They also state that Instagram can be used to promote health measures and prevention campaigns.

Data:

The data source for this study comes from Instagram, an image sharing social media platform. Images related to health and medicine are extracted using a tool Instaloader. Instaloader is a tool which can be used to download images or videos from Instagram along with their descriptions and metadata of the images or videos. For our study, we only extract images from Instagram for the following hashtags using Instaloader - #health, #healthylifestyle and #covid. The hashtags #health and #healthylifestyle are selected manually since these are the hashtags mostly used by users on Instagram to depict the general topics related to health [48]. The #covid is selected to analyze the depiction of covid by the users on Instagram over different parts of the world. The images extracted for #covid are mostly from the year 2020 and the images for the rest of the hashtags range from a timeline of 2014 to 2020. The number of images collected for each hashtag is listed in the table 1 -

The images collected for each of the hashtags is then filtered by the location of the image as provided by the user. During the filtering process, the images that have no location associated with them are removed and images with locations are placed in folders containing the country code associated with the location. The number of images after filtering by location are listed in table 1 -

Hashtags	No. of total images collected	No. of images after filtering
#health	610,883	375,757
#healthylifestyle	428,796	239,575
#covid	652,201	122,403

Methodology:

For each hashtag, images were first filtered according to the location tagged by the user as described above. After filtering by location, images for each location underwent a preprocessing phase in which they were resized and denoised using Gaussian smoothing. After the preprocessing phase, a pre-trained model VGG16 on ImageNet was used for feature extraction of the images. Our approach for this study is described below:

1. Using VGG16 a pre-trained model in Keras to extract features of images for each hashtag and each country.

2. Using K Means to form a cluster on images of the same country (number of clusters depend on the number of images for each country)

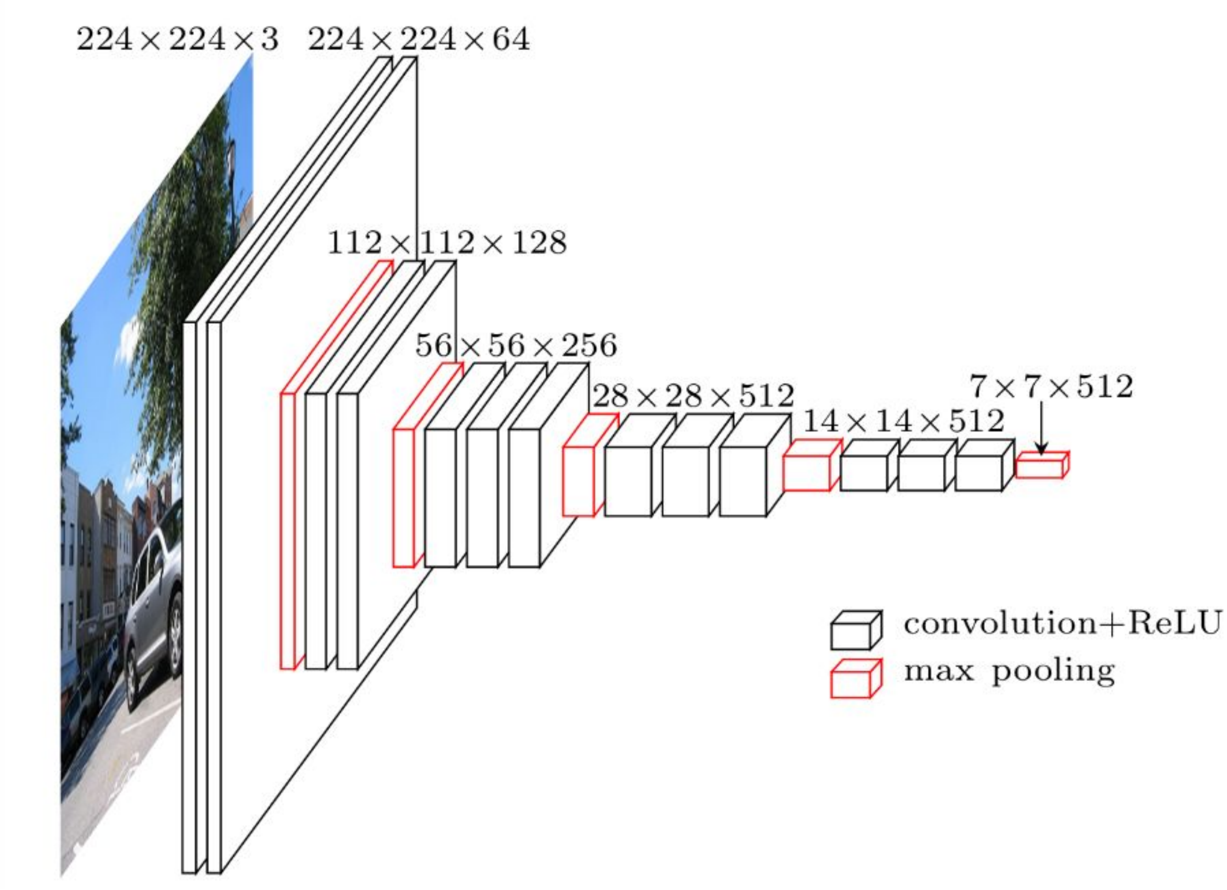
3. Using coco algorithm to predict captions for clustered images for each hashtag and each country

4. Performing Topic Modeling using LDA on the captions generated in Step 3 to find the topics related to clusters of images of each country.

5. Extracting hashtags of the clustered images.

6. Performing Topic Modeling using LDA on the hashtags extracted in Step 5 to find the topics related to hashtags of images in each country

Feature extraction of image is done through VGG16, a convolutional neural network, a Visual Geometry Group model with 16 weight layers. The input layer of VGG16 model takes an image size of 224 x 224 x 3, from the input layer to the last max pooling layer which is of size 7 x 7 x 512, feature extraction of image takes place. Each image after filtering by location is passed through this model to extract the image features. The model's architecture is displayed in Figure 1 as below.



After extraction of the features through VGG16, kMeans algorithm is used to cluster the images. The algorithm clusters data by separating images into disjoint clusters which are described by means of the images in the cluster. The means are cluster centroids and k means chooses centroids that minimize inertia, measure of how internally coherent clusters are. For each hashtag considered and each country extracted after filtering, clusters of the images are created through KMeans, the number of clusters for each country are decided based on the number of images present in the country.

After the clustering of images, we build a model to generate captions of the images using a convolutional neural network as an encoder and a recurrent neural network as a decoder. The CNN encoder has images as an input and the output is fed into the RNN decoder which outputs the caption for that image. The Microsoft Common Objects in Context (COCO) dataset is used to train the model on its large image and captions dataset. The trained features are passed through the encoder and the decoder outputs the captions of the images. The captions which are a machine generated output are stored for each country and each cluster created in the country.

After image captioning, the captions are passed through a topic model to discover the abstract topics occurring in the collection of captions of each cluster. Through this, the topics associated with the clusters are known for a machine generated output. We use Latent Dirichlet Allocation as our topic model to classify the text in the captions to a particular topic. Frequent occurring words are also generated per topic. During this step the captions are gone through a preprocessing phase in which the caption is split into words, the words are lowercase, punctuations are removed. Stopwords are removed and the words are lemmatized, verbs in past and future are changed to present, as well as words are stemmed keeping only the root word. We then use tf-idf on the captions for each cluster to generate topics. The number of topics is selected based on the captions in each cluster. For example - If there are captions less than 10 then only one or two topics are generated. The topics generated for each country per cluster is then visualized through a word cloud. These topics are through a machine generated algorithm.

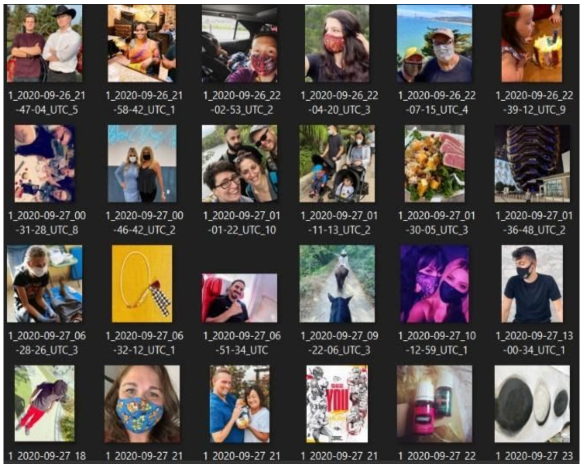
For the other part of our study, we extract hashtags of the images provided by the user after they have been filtered by location and clustered through KMeans. The hashtags depict the user perspective of the image posted, they also help categorize and organize similar content and make the image more discoverable by using the appropriate hashtag as well as market the content to the appropriate audience. For each of the three hashtags and each of the countries after filtering the hashtags related to the images clustered are extracted and stored to perform topic modeling. Topic modeling is performed on the hashtags collected for each of the clusters to classify the hashtags to a particular topic. The topics are generated for each country per cluster for the hashtags and then visualized through a word cloud. These topics are through a user generated algorithm.

The topics generated through the machine generated algorithm and through a user generated algorithm are compared to analyze the similarity and logical output of both the generated algorithms. We can then analyze the perception of Instagram users through the topics generated from the hashtags and the captions generated by the coco algorithm for clusters in each country. We can also compare the topics of different countries related to the three hashtags.

Results:

1. **#covid:** The clustered images after applying VGG16 for feature extraction and KMeans for clustering are:

The model performed well on some instances but performed poorly on some images. The images in figure 2 are clustered according to the similarity of people wearing masks, although some images are incorrectly added to the cluster.



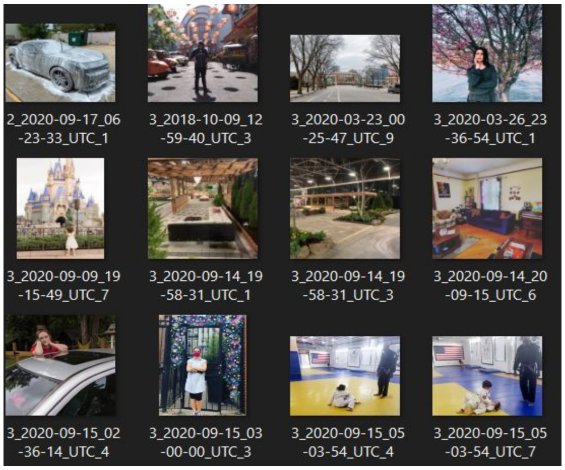
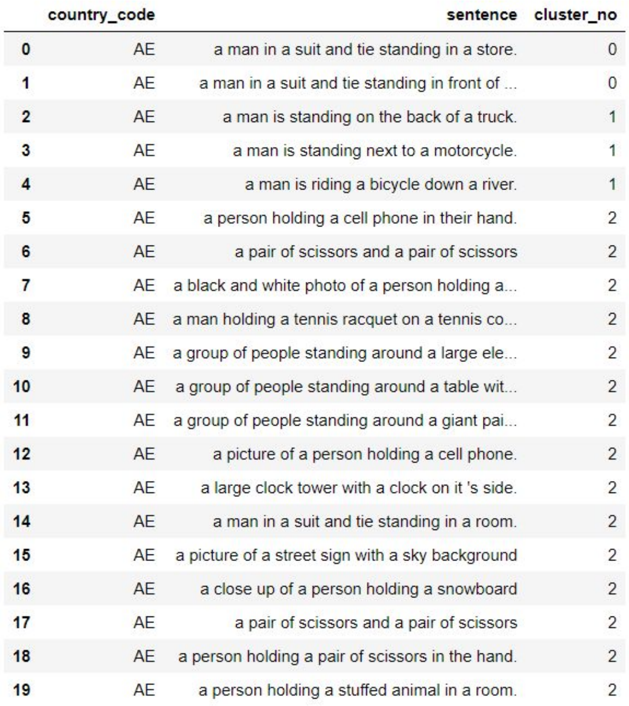


Figure 2 - Images clustered correctly

Figure 3- Images clustered incorrectly

The images in figure 3 are clustered incorrectly and is an example of how the model performed poorly. After the clustering process, captions are generated using the coco algorithm on each cluster for all the countries. Figure 4 shows the output generated by coco for the country AE (United Arab Emirates) for clusters 1,2 and 3. After the captions are generated by coco algorithm, topic modeling through LDA was performed, Figure 5 shows the word cloud of the topics generated for the country US for all the clusters.

On the other side, considering user generated output through hashtags, after the images have been clustered, we extracted hashtags for each country and each clustered image. Figure 6 portrays a sample of the dataset for the hashtags extracted by users. After hashtag extraction of the images, we ran the hashtags through LDA for topic modeling. Figure 7 shows the word cloud of the topics generated for US for user generated output of the hashtags for all the clusters.



	country_code	sentence	cluster_no
0	AE	a man in a suit and tie standing in a store.	0
1	AE	a man in a suit and tie standing in front of ...	0
2	AE	a man is standing on the back of a truck.	1
3	AE	a man is standing next to a motorcycle.	1
4	AE	a man is riding a bicycle down a river.	1
5	AE	a person holding a cell phone in their hand.	2
6	AE	a pair of scissors and a pair of scissors	2
7	AE	a black and white photo of a person holding a...	2
8	AE	a man holding a tennis racquet on a tennis co...	2
9	AE	a group of people standing around a large ele...	2
10	AE	a group of people standing around a table wit...	2
11	AE	a group of people standing around a giant pai...	2
12	AE	a picture of a person holding a cell phone.	2
13	AE	a large clock tower with a clock on it 's side.	2
14	AE	a man in a suit and tie standing in a room.	2
15	AE	a picture of a street sign with a sky background	2
16	AE	a close up of a person holding a snowboard	2
17	AE	a pair of scissors and a pair of scissors	2
18	AE	a person holding a pair of scissors in the hand.	2
19	AE	a person holding a stuffed animal in a room.	2

Figure 4 - Captions generated by coco algorithm (machine generated output)

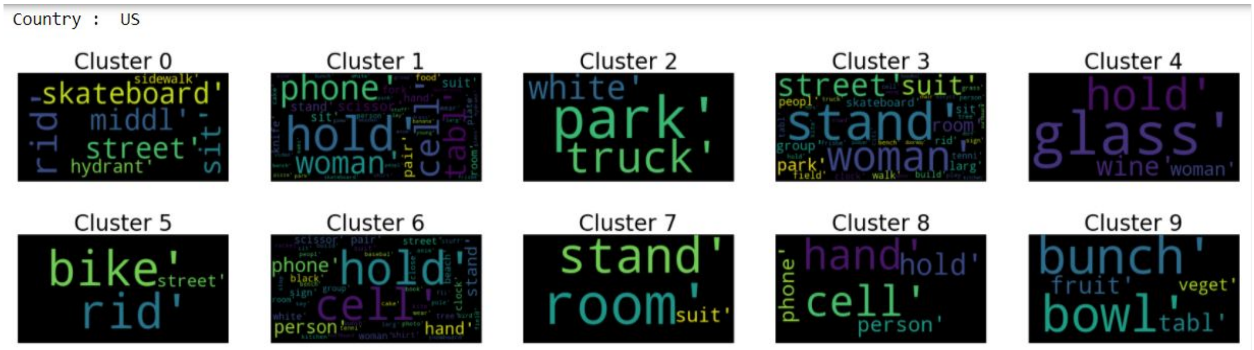
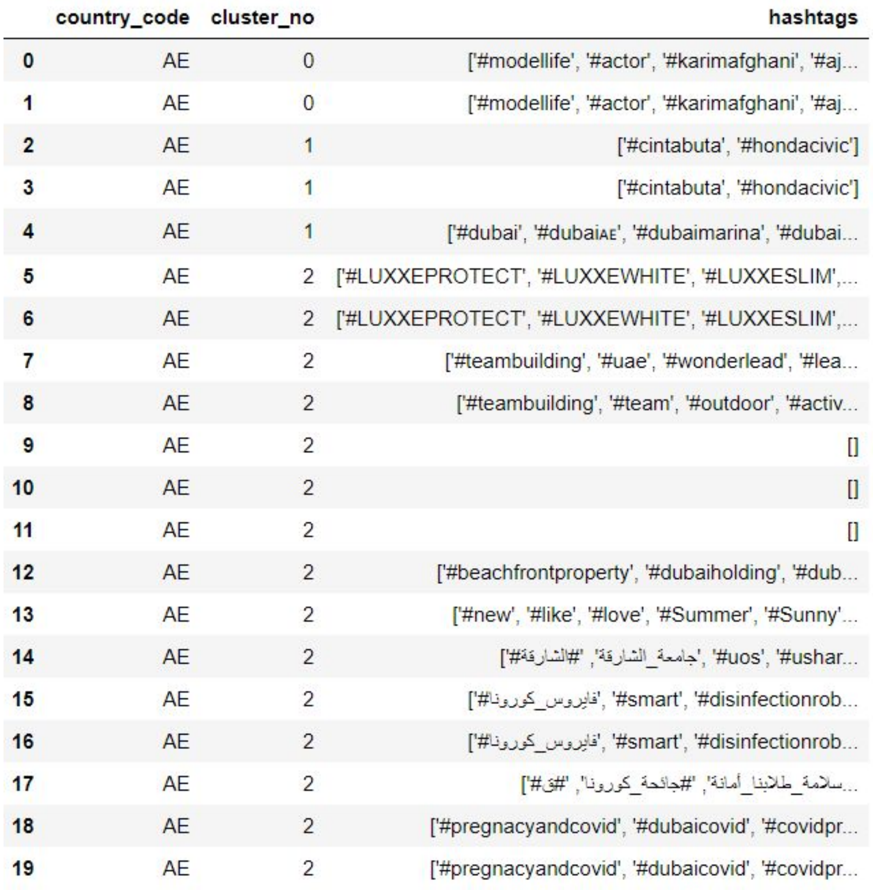


Figure 5 - Word cloud of Topic modeling through LDA on captions



	country_code	cluster_no	hashtags
0	AE	0	['#modellife', '#actor', '#karimafghani', '#aj...
1	AE	0	['#modellife', '#actor', '#karimafghani', '#aj...
2	AE	1	['#cintabuta', '#hondacivic']
3	AE	1	['#cintabuta', '#hondacivic']
4	AE	1	['#dubai', '#dubaiAE', '#dubaimarina', '#dubai...
5	AE	2	['#LUXXEPROTECT', '#LUXXEWHITE', '#LUXXESLIM',...
6	AE	2	['#LUXXEPROTECT', '#LUXXEWHITE', '#LUXXESLIM',...
7	AE	2	['#teambuilding', '#uae', '#wonderlead', '#lea...
8	AE	2	['#teambuilding', '#team', '#outdoor', '#activ...
9	AE	2	[]
10	AE	2	[]
11	AE	2	[]
12	AE	2	['#beachfrontproperty', '#dubaiholding', '#dub...
13	AE	2	['#new', '#like', '#love', '#Summer', '#Sunny'...
14	AE	2	['#الشارقة', '#جامعة_الشارقة', '#uos', '#ushar...
15	AE	2	['#فايروس_كورونا', '#smart', '#disinfectionrob...
16	AE	2	['#فايروس_كورونا', '#smart', '#disinfectionrob...
17	AE	2	['#ق', '#جائحة_كورونا', '#سلامة_طلابنا_أمانة...
18	AE	2	['#pregnacyandcovid', '#dubaicovid', '#covidpr...
19	AE	2	['#pregnacyandcovid', '#dubaicovid', '#covidpr...

Figure 6 - Hashtags extracted for the images after clustering (user generated output)

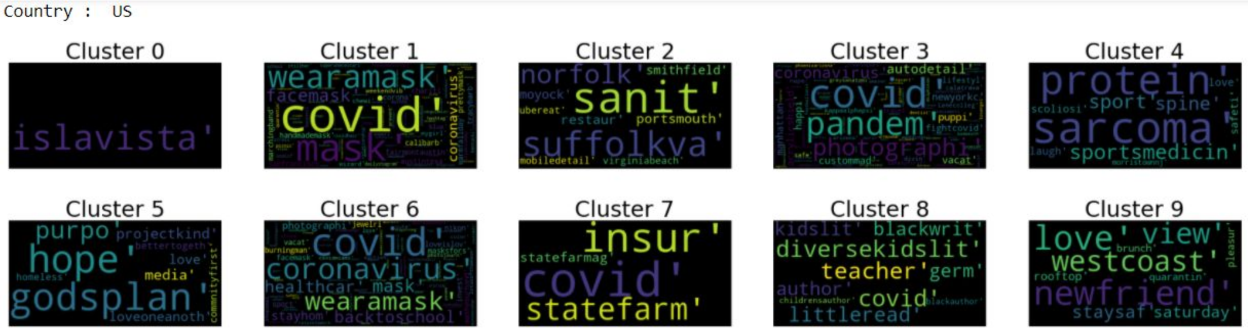


Figure 7 - WordCloud of topics generated by LDA for hashtags

1. **#health**

For #health the model performed well in most scenarios generating clusters of images similar. Figure 8 portrays one such cluster in which images of food have been clustered together.

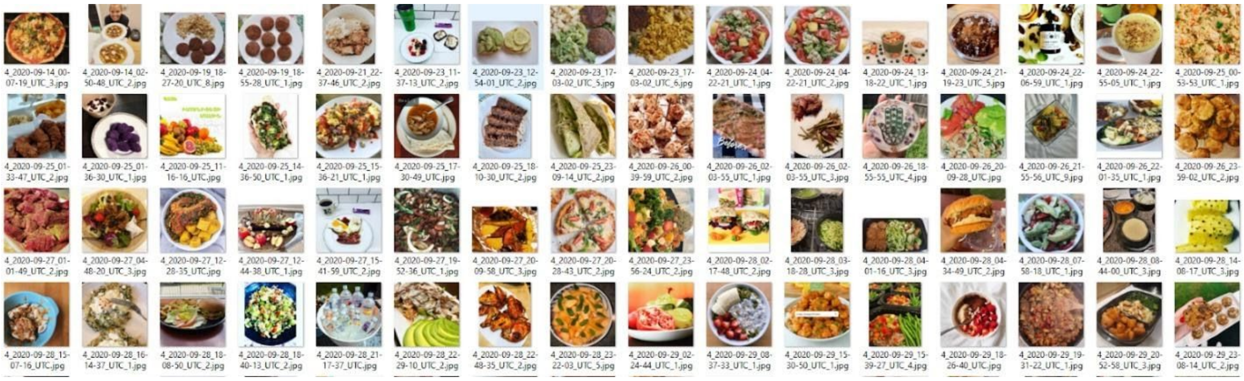


Figure 8 - VGG16 and KMeans performing well on the dataset

The model performed poorly in some instances creating a cluster of only one image and clustering images dissimilar to one another as shown in Figure 9

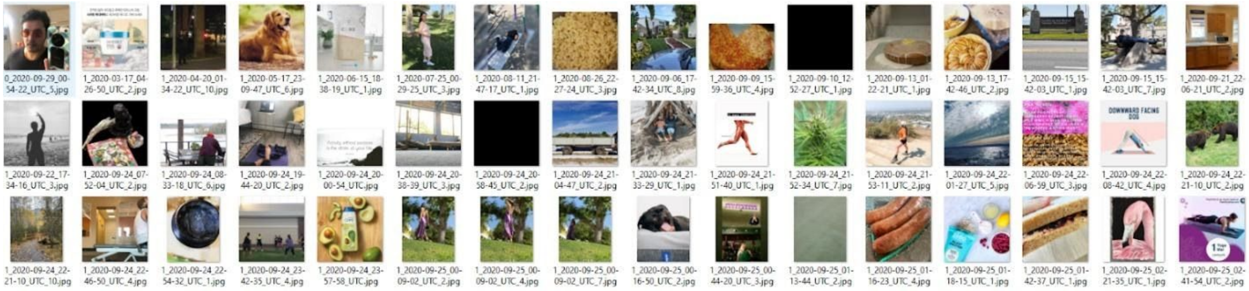
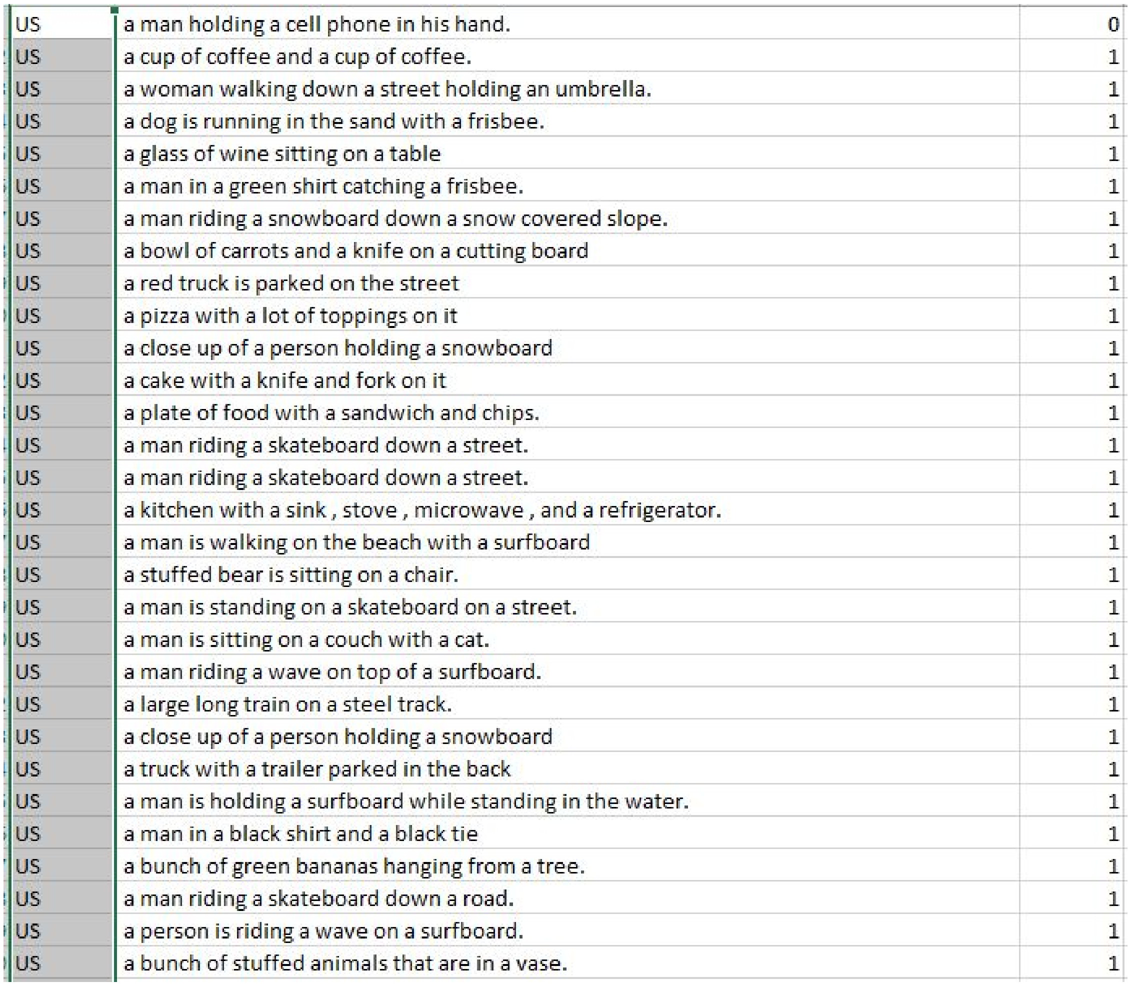


Figure 9 - VGG16 and KMeans performing poorly on the dataset

Similar to the process performed for #covid, we ran the clustered images through the COCO algorithm to generate machine output captions and extracted hashtags as user generated output then ran LDA models on the captions and hashtags to generate a WordCloud. Figure 10 shows the captions generated by coco, Figure 11 portrays the WordCloud of the topics generated on the captions which are machine generated. Similarly, Figure 12 shows the hashtags extracted for the clustered images and figure 13 portrays the topics generated on the hashtags for each cluster and country which is user generated.



US	a man holding a cell phone in his hand.	0
US	a cup of coffee and a cup of coffee.	1
US	a woman walking down a street holding an umbrella.	1
US	a dog is running in the sand with a frisbee.	1
US	a glass of wine sitting on a table	1
US	a man in a green shirt catching a frisbee.	1
US	a man riding a snowboard down a snow covered slope.	1
US	a bowl of carrots and a knife on a cutting board	1
US	a red truck is parked on the street	1
US	a pizza with a lot of toppings on it	1
US	a close up of a person holding a snowboard	1
US	a cake with a knife and fork on it	1
US	a plate of food with a sandwich and chips.	1
US	a man riding a skateboard down a street.	1
US	a man riding a skateboard down a street.	1
US	a kitchen with a sink , stove , microwave , and a refrigerator.	1
US	a man is walking on the beach with a surfboard	1
US	a stuffed bear is sitting on a chair.	1
US	a man is standing on a skateboard on a street.	1
US	a man is sitting on a couch with a cat.	1
US	a man riding a wave on top of a surfboard.	1
US	a large long train on a steel track.	1
US	a close up of a person holding a snowboard	1
US	a truck with a trailer parked in the back	1
US	a man is holding a surfboard while standing in the water.	1
US	a man in a black shirt and a black tie	1
US	a bunch of green bananas hanging from a tree.	1
US	a man riding a skateboard down a road.	1
US	a person is riding a wave on a surfboard.	1
US	a bunch of stuffed animals that are in a vase.	1

Figure 10 - Captions generated through the coco algorithm

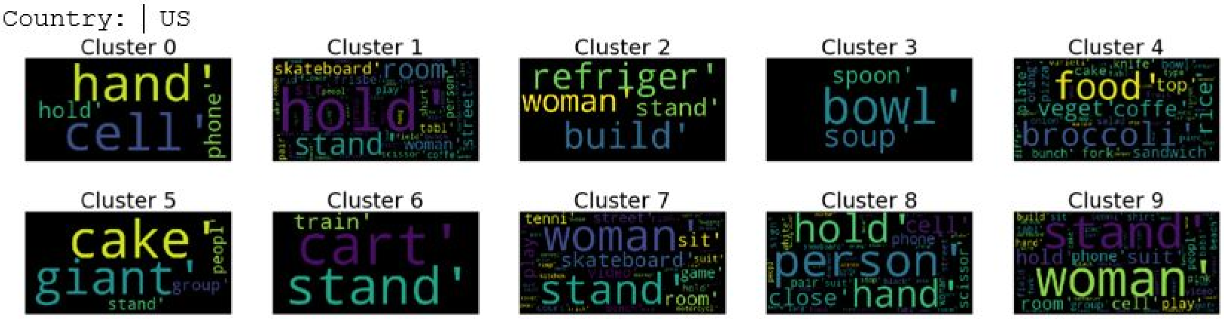
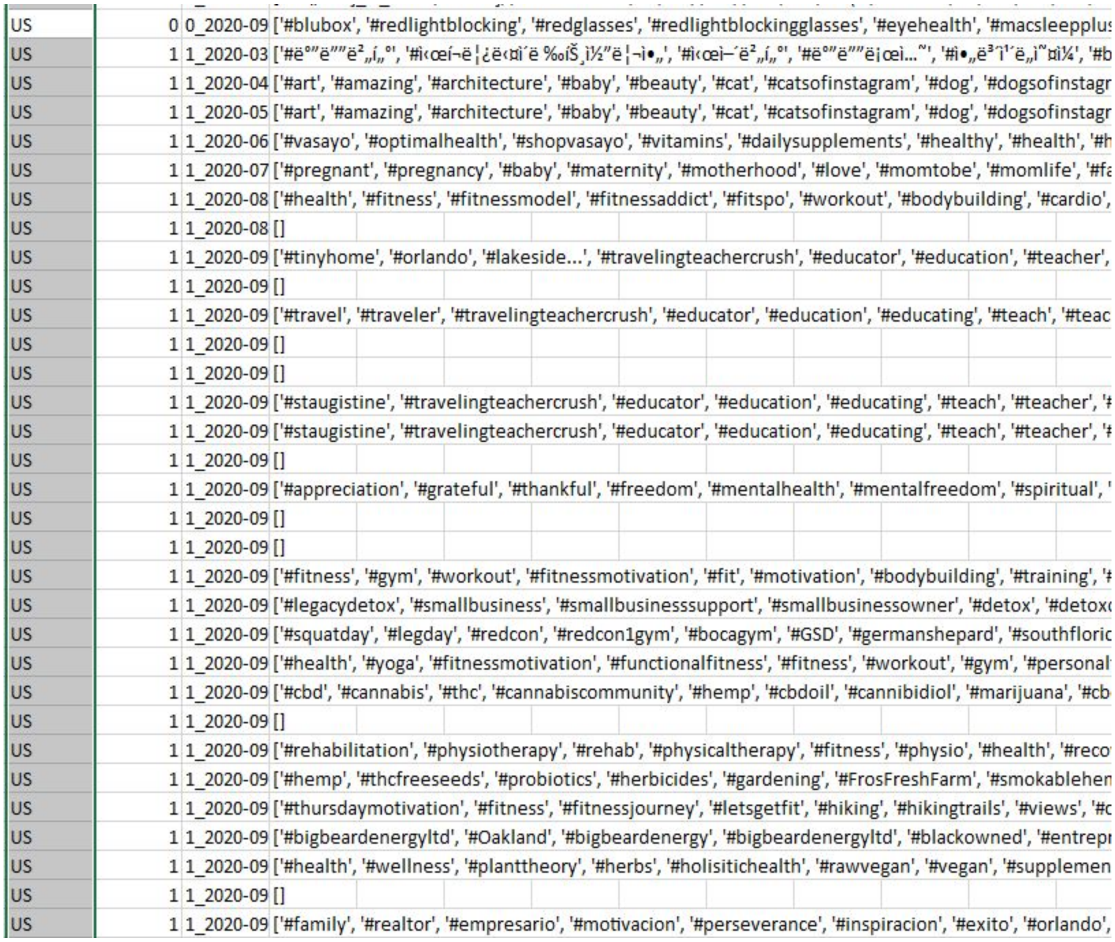


Figure 11 - WordCloud of topics generated on captions for country US for each cluster



US	0	0_2020-09	['#blubox', '#redlightblocking', '#redglasses', '#redlightblockingglasses', '#eyehealth', '#macsleepplus
US	1	1_2020-03	['#ë°"ë""ë²„í„°', '#ì‹œí¬ë¦¿ë‹¤ì´ë ‰íŠ¸ì½"ë¦¬ì•„', '#ì‹œì–´ë²„í„°', '#ë°"ë""ë¡œì…˜', '#ì•„ë³´ì¹´ë„ì˜¤ì¼', '#b
US	1	1_2020-04	['#art', '#amazing', '#architecture', '#baby', '#beauty', '#cat', '#catsofinstagram', '#dog', '#dogsofinstagr
US	1	1_2020-05	['#art', '#amazing', '#architecture', '#baby', '#beauty', '#cat', '#catsofinstagram', '#dog', '#dogsofinstagr
US	1	1_2020-06	['#vasayo', '#optimalhealth', '#shopvasayo', '#vitamins', '#dailysupplements', '#healthy', '#health', '#h
US	1	1_2020-07	['#pregnant', '#pregnancy', '#baby', '#maternity', '#motherhood', '#love', '#momtobe', '#momlife', '#fa
US	1	1_2020-08	['#health', '#fitness', '#fitnessmodel', '#fitnessaddict', '#fitspo', '#workout', '#bodybuilding', '#cardio',
US	1	1_2020-08	[]
US	1	1_2020-09	['#tinyhome', '#orlando', '#lakeside...', '#travelingteachercrush', '#educator', '#education', '#teacher',
US	1	1_2020-09	[]
US	1	1_2020-09	['#travel', '#traveler', '#travelingteachercrush', '#educator', '#education', '#educating', '#teach', '#teac
US	1	1_2020-09	[]
US	1	1_2020-09	[]
US	1	1_2020-09	['#staugistine', '#travelingteachercrush', '#educator', '#education', '#educating', '#teach', '#teacher', '#
US	1	1_2020-09	['#staugistine', '#travelingteachercrush', '#educator', '#education', '#educating', '#teach', '#teacher', '#
US	1	1_2020-09	[]
US	1	1_2020-09	['#appreciation', '#grateful', '#thankful', '#freedom', '#mentalhealth', '#mentalfreedom', '#spiritual', '
US	1	1_2020-09	[]
US	1	1_2020-09	[]
US	1	1_2020-09	['#fitness', '#gym', '#workout', '#fitnessmotivation', '#fit', '#motivation', '#bodybuilding', '#training', '#
US	1	1_2020-09	['#legacydetox', '#smallbusiness', '#smallbusinesssupport', '#smallbusinessowner', '#detox', '#detox
US	1	1_2020-09	['#squatday', '#legday', '#redcon', '#redcon1gym', '#bocagym', '#GSD', '#germanshepard', '#southflori
US	1	1_2020-09	['#health', '#yoga', '#fitnessmotivation', '#functionalfitness', '#fitness', '#workout', '#gym', '#personal
US	1	1_2020-09	['#cbd', '#cannabis', '#thc', '#cannabiscommunity', '#hemp', '#cbdoil', '#cannibidiol', '#marijuana', '#cb
US	1	1_2020-09	[]
US	1	1_2020-09	['#rehabilitation', '#physiotherapy', '#rehab', '#physicaltherapy', '#fitness', '#physio', '#health', '#reco
US	1	1_2020-09	['#hemp', '#thcfreeseeds', '#probiotics', '#herbicides', '#gardening', '#FrosFreshFarm', '#smokablehen
US	1	1_2020-09	['#thursdaymotivation', '#fitness', '#fitnessjourney', '#letsgetfit', '#hiking', '#hikingtrails', '#views', '#c
US	1	1_2020-09	['#bigbeardenergyltd', '#Oakland', '#bigbeardenergy', '#bigbeardenergyltd', '#blackowned', '#entrepr
US	1	1_2020-09	['#health', '#wellness', '#planttheory', '#herbs', '#holisitichealth', '#rawvegan', '#vegan', '#supplemen
US	1	1_2020-09	[]
US	1	1_2020-09	['#family', '#realtor', '#empresario', '#motivacion', '#perseverance', '#inspiracion', '#exito', '#orlando',

Figure 12 - Hashtags extracted for clustered images

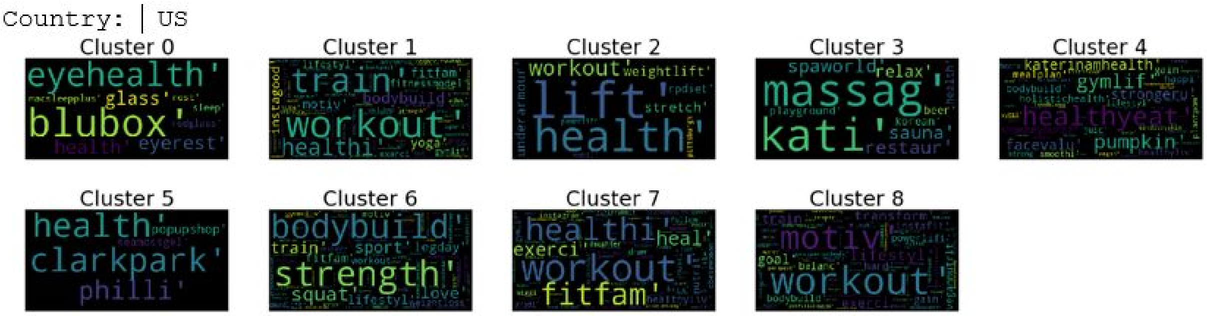


Figure 13 - WordCloud of topics generated on hashtags for country US for each cluster

1. **#healthylifestyle**

The process for data collected for #healthylifestlye is same as that of the other two hashtags, hence we will skip displaying the preprocessed output and display the wordcloud of the topics generated from the captions and wordcloud of topics generated through the hashtags. Figure 14 portrays the WordCloud of topics on machine generated output and Figure 15 shows the WordCloud of topics generated on user generated output.

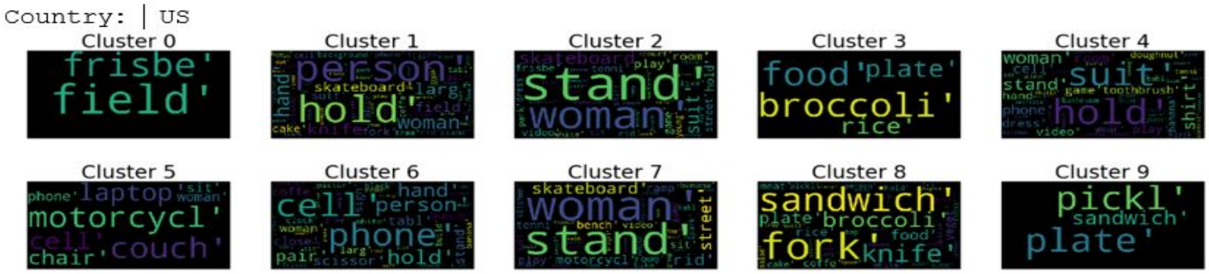


Figure 14 - WordCloud of Topics on captions generated through COCO

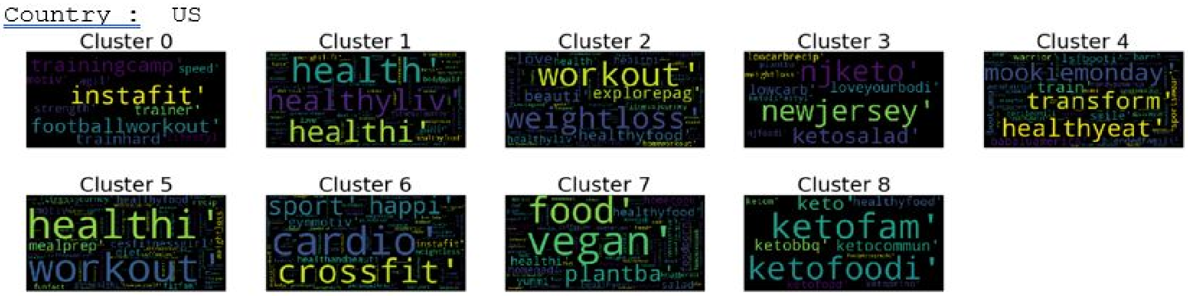


Figure 15 - WordCloud of Topics on hashtags provided by the user for US

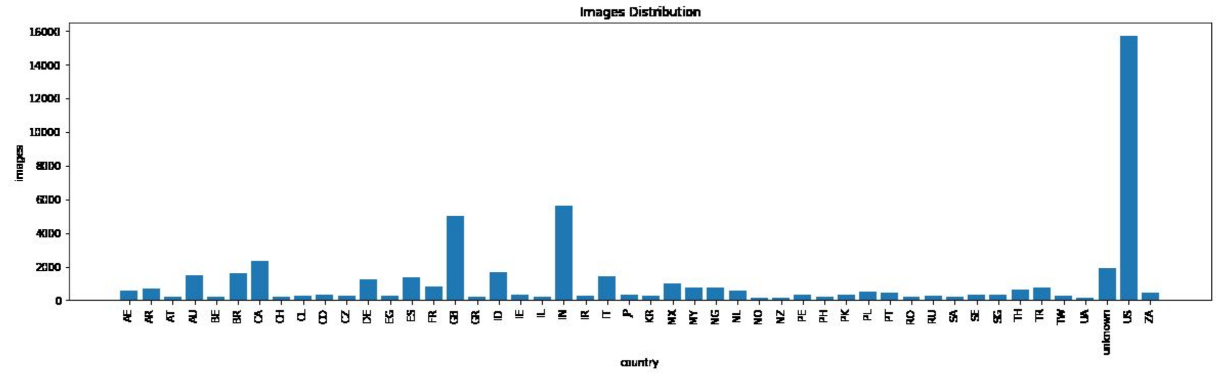


Figure 16 - Images generated by each country

Discussion:

From 6,52,201 images out of #covid data, it got reduced to 1,22,403 images after filtering with locations. Variance in the number of images from all countries was very high since out of 184 countries, 92 countries have images less than 50, 147 countries had images less than 500 and from 22 major countries the number of images ranges from 1000 to 9000. It was difficult to decide how many images should be considered for the analysis, So we chose random 1500 images from countries with more than 1500 images. To perform clustering initially we used the elbow method with the K-Means algorithm. But the number of clusters generated was more than 20, but considering the laptops we work on and the difficulty in managing those many clusters we decided to take different numbers of clusters for each country based on the number of images they have. From all clusters generated few were not good having only a 1 or 2 images in them out of thousands of images. In every country there were at least 2 clusters generated which were not correct or very small. The captions generated using coco algorithm were very raw and missed information from any complex images involving more than 1 object especially in images where objects are in same color but the captions are acceptable to understand what images are about. The topics generated using LDA on user hashtags involved words which are not always related to images like captions of dog be like #love, which is not easy to interpret through the picture. But topics generated on machine captions involved words which can be seen as the name of an object or actions.

It is important to notice that the difference between topics generated on captions from COCO and topics generated on hashtags provided by users is to a great extent. While coco outputs captions through its trained dataset, it fails to identify the bigger picture and generalizes the things happening in an image. Whereas, user input through hashtags may not be the ideal solution to generate topics related to the images it does provide a much better view of the image and the topics related to the image.However, in some instances hashtags need to be filtered appropriately to get the desired topics since hashtags provided by the user might also contain words which are not associated with the content of the image and could be misleading. From our analysis, we find that hashtags are the best way to provide topics related to images when compared to other machine generated outputs.

We also analyze the difference of interpretation by Instagram users around different countries for topics related to covid, health and healthylifestyle. We find that topics related to covid in different countries usually have the same motivation such as to stay home or wear masks, they also differ in some topics related to different incidents occurring in their countries. Health and healthylifestyle are perceived differently by users across different regions, in some countries there is more emphasis on food and in others there is more emphasis on working out, although the topics remain the same and the general overview of health or healthylifestyle is the same throughout the world ranging from eating good food to working out regularly.

For all the hashtags considered, images are mostly generated from the country USA. India and Britain are the second and third countries to generate the most content after the USA.

Future Work: In the current project we work on images from all countries and that increases the complexity of managing more clusters from all countries especially from countries having more images. In future we would like to work on a single major country since it has more images and we can create more clusters, this way there will be better cluster formation and the topics generated from them could present different results. We can implement different algorithms like VGG19 or ResNet for feature extraction and use algorithms other than COCO such as YOLO to improve captions for the images.

References:

[1] "How Many People Use Social Media in 2020." https://www.oberlo.com/statistics/how-many-people-use-social-media (accessed Sep. 21, 2020).

[2] “Demographics of Social Media Users and Adoption in the United States.” https://www.pewresearch.org/internet/fact-sheet/social-media/ (accessed Sep. 21, 2020).

[3] “The rise of social media.” https://ourworldindata.org/rise-of-social-media (accessed Sep. 21, 2020).

[4] “Facebook vs. Instagram in 2020: All You Need to Know.” https://www.socialbakers.com/blog/instagram-vs-facebook-advertising-differences-and-best-practices (accessed Sep. 21, 2020).

[5] “About the Instagram Company.” https://about.instagram.com/about-us (accessed Sep. 21, 2020).

[6] “• Instagram by the Numbers (2020): Stats, Demographics & Fun Facts,” Jan. 26, 2020. https://www.omnicoreagency.com/instagram-statistics/ (accessed Sep. 21, 2020).

[7] J. Chen, “Important Instagram stats you need to know for 2020,” *Sprout Social, https://sproutsocial. com/insights/instagram-stats*, 2020, [Online]. Available: https://sproutsocial.com/insights/instagram-stats/.

[8] Q. Tang, K. Zhang, and Y. Li, “The important role of social media during the COVID-19 epidemic,” *Disaster Med. Public Health Prep.*, pp. 1–5, Sep. 2020.

[9] E. Bolat, “Why the UK government is paying social media influencers to post about coronavirus,” *The Conversation*, Sep. 09, 2020.

[10] A. Bhattacharya, “For Indians under lockdown, social media is the go-to source for news and entertainment,” *Retrieved May*, vol. 15, p. 2020, 2020.

[11] FashionNetwork.com US, “Instagram was UK’s top social media platform in lockdown, key for shopping.” https://us.fashionnetwork.com/news/Instagram-was-uk-s-top-social-media-platform-in-lockdown-key-for-shopping,1237799.html (accessed Sep. 21, 2020).

[12] A. Rovetta and A. S. Bhagavathula, “Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags,” *J. Med. Internet Res.*, vol. 22, no. 8, p. e20673, Aug. 2020.

[13] E. Chen, K. Lerman, and E. Ferrara, “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set,” *JMIR Public Health Surveill*, vol. 6, no. 2, p. e19273, May 2020.

[14] K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, “A First Instagram Dataset on COVID-19,” *arXiv [cs.SI]*, Apr. 25, 2020.

[15] “World Health Organization (@who) • Instagram photos and videos.” https://www.instagram.com/who/ (accessed Sep. 21, 2020).

[16] M. Neal, “Instagram Influencers: The Effects of Sponsorship on Follower Engagement With Fitness Instagram Celebrities,” Rochester Institute of Technology, 2017.

[17] A. Singh, M. N. Halgamuge, and B. Moses, “An Analysis of Demographic and Behavior Trends Using Social Media: Facebook, Twitter, and Instagram,” *Social Network Analytics*, p. 87, 2019, Accessed: Sep. 27, 2020. [Online].

[18] A. G. Reece and C. M. Danforth, “Instagram photos reveal predictive markers of depression,” *EPJ Data Science*, vol. 6, no. 1, p. 15, Aug. 2017.

[19] B. Ferwerda and M. Tkalcic, “Predicting Users’ Personality from Instagram Pictures: Using Visual and/or Content Features?,” in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, Singapore, Singapore, Jul. 2018, pp. 157–161, Accessed: Sep. 27, 2020. [Online].

[20] M. Habibi and P. W. Cahyo, “Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 4, pp. 399–408, 2019.

[21] L. Davainyte, V. K. Kirketerp, H. Kraus, S. Urlass, and F. J. A. Perez-Cueto, “The influence of Instagram usage on cosumers ‘choice of restaurants and meal preparation at home.” 2019, [Online]. Available:

https://figshare.cardiffmet.ac.uk/articles/The_influence_of_Instagram_usage_on_cosumers_choice_of_restaurants_and_meal_preparation_at_home/9192497/files/16741907.pdf.

[22] M. N. Kamel Boulos, D. M. Giustini, and S. Wheeler, “Instagram and WhatsApp in Health and Healthcare: An Overview,” *Future Internet*, vol. 8, no. 3, p. 37, Jul. 2016, Accessed: Sep. 27, 2020. [Online].

[23] M. F. Alkazemi, J. P. D. Guidry, E. Almutairi, and M. Messner, “#Arabhealth on Instagram: Examining Public Health Messages to Arabian Gulf State Audiences,” *Health Communication*. pp. 1–9, 2020, doi: 10.1080/10410236.2020.1816283.

[24] N. Lee, K. Buchanan, and M. Yu, “Each post matters: a content analysis of# mentalhealth images on Instagram,” *J. Vis. Commun. Med.*, vol. 43, no. 3, pp. 128–138, 2020.

[25] Y. Kim and J. H. Kim, “Using photos for public health communication: A computational analysis of the Centers for Disease Control and Prevention Instagram photos and public responses,” *Health Informatics J.*, vol. 26, no. 3, pp. 2159–2180, Sep. 2020.

[26] F. Yakar, R. Jacobs, and N. Agarwal, “The current usage of Instagram in neurosurgery,” *Interdisciplinary Neurosurgery*, vol. 19, p. 100553, Mar. 2020.

[27] X. L. Wong, R. C. Liu, and D. F. Sebaratnam, “Evolving role of Instagram in# medicine,” *Intern. Med. J.*, vol. 49, no. 10, pp. 1329–1332, 2019.

[28] N. Hochman and L. Manovich, “Zooming into an Instagram City: Reading the local through social media,” *First Monday*, Jun. 2013, doi: 10.5210/fm.v18i7.4711.

[29] T. Leaver and T. Highfield, “Visualising the ends of identity: pre-birth and post-death on Instagram,” *null*, vol. 21, no. 1, pp. 30–45, Jan. 2018.

[30] S. Chang, “Instagram Post Data Analysis,” *arXiv [cs.HC]*, Oct. 07, 2016.

[31] V. K. Singh, S. Hegde, and A. Atrey, “Towards measuring fine-grained diversity using social media photographs,” 2017, [Online]. Available: https://wp.comminfo.rutgers.edu/vsingh/wp-content/uploads/sites/110/2017/10/ICWSM_Singh_Diversity.pdf.

[32] M. Lalancette and V. Raynauld, “The Power of Political Image: Justin Trudeau, Instagram, and Celebrity Politics,” *Am. Behav. Sci.*, vol. 63, no. 7, pp. 888–924, Jun. 2019.

[33] S. Muralidhara and M. J. Paul, “#Healthy Selfies: Exploration of Health Topics on Instagram,” *JMIR Public Health and Surveillance*, vol. 4, no. 2. p. e10150, 2018, doi: 10.2196/10150.

[34] C. Malighetti, S. Sciara, A. Chirico, and G. Riva, “Emotional Expression of# body on Instagram,” *Social Media+ Society*, vol. 6, no. 2, p. 2056305120924771, 2020.

[35] V. Murashka, J. Liu, and Y. Peng, “Fitspiration on Instagram: Identifying Topic Clusters in User Comments to Posts with Objectification Features,” *Health Commun.*, pp. 1–12, Jun. 2020.

[36] S. Santarossa, P. Coyne, C. Lisinski, and S. J. Woodruff, “#fitspo on Instagram: A mixed-methods approach using Netlytic and photo analysis, uncovering the online discussion and author/image characteristics,” *Journal of Health Psychology*, vol. 24, no. 3. pp. 376–385, 2019, doi: 10.1177/1359105316676334.

[37] R. Cohen, L. Irwin, T. Newton-John, and A. Slater, “#bodypositivity: A content analysis of body positive accounts on Instagram,” *Body Image*, vol. 29. pp. 47–57, 2019, doi: 10.1016/j.bodyim.2019.02.007.

[38] A. Argyrou, S. Giannoulakis, and N. Tsapatsoulis, “Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation?,” in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Sep. 2018, pp. 61–67.

[39] S. Giannoulakis, N. Tsapatsoulis, and K. Ntalianis, “Identifying Image Tags from Instagram Hashtags Using the HITS Algorithm,” in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 89–94.

[40] S. Giannoulakis and N. Tsapatsoulis, “Evaluating the descriptive power of Instagram hashtags,” *Journal of Innovation in Digital Ecosystems*, vol. 3, no. 2, pp. 114–129, Dec. 2016.

[41] A. Fiallos, K. Jimenes, C. Fiallos, and S. Figueroa, “Detecting Topics and Locations on Instagram Photos,” in *2018 International Conference on eDemocracy eGovernment (ICEDEG)*, Apr. 2018, pp. 246–250.

[42] V. Ketonen and A. Malik, “Characterizing vaping posts on Instagram by using unsupervised machine learning,” *Int. J. Med. Inform.*, vol. 141, p. 104223, Jun. 2020.

[43] J. Vassey, C. Metayer, C. J. Kennedy, and T. P. Whitehead, “#Vape: Measuring E-Cigarette Influence on Instagram With Deep Learning and Text Analysis,” *Frontiers in Communication*, vol. 4. 2020, doi: 10.3389/fcomm.2019.00075.

[44] L. Czaplicki *et al.*, “Characterising JUUL-related posts on Instagram,” *Tobacco Control*. p. tobaccocontrol–2018, 2019, doi: 10.1136/tobaccocontrol-2018-054824.

[45] Y. Zhang, J.-P. Allem, J. B. Unger, and T. Boley Cruz, “Automated Identification of Hookahs (Waterpipes) on Instagram: An Application in Feature Extraction Using Convolutional Neural Network and Support Vector Machine Classification,” *J. Med. Internet Res.*, vol. 20, no. 11, p. e10513, Nov. 2018.

[46] M. De Veirman, V. Cauberghe, and L. Hudders, “Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude,” *null*, vol. 36, no. 5, pp. 798–828, Sep. 2017.

[47] K. Pilgrim and S. Bohnet-Joschko, “Selling health and happiness how influencers communicate on Instagram about dieting and exercise: mixed methods research,” *BMC Public Health*, vol. 19, no. 1, p. 1054, Aug. 2019.

[48] “Hashtags for #health.” http://best-hashtags.com/hashtag/health/ (accessed Dec. 08, 2020).