How Hashtags Are Used On Instagram Following A Globally Known Event And Their Relevance?

Sahej Soin

Received November 17, 2023
Accepted February 02, 2024
Electronic access February 15, 2024

As social media grows in popularity daily, so does the number of users. These large number of users form virtual communities as large as ones in the physical world. As these communities grow, they interact on social media and besides private messaging, all interaction is public. This makes it easy to extract these interactions in the form of metadata. Studying Instagram interaction posts related to a global event can provide valuable insights into user engagement and sentiments. Identifying most frequently used hashtags related to the global event can help analyze its reach and impact. Analyzing hashtags can help gain valuable insights into how users perceive and react to the global event on a large scale. It further helps in identifying individuals or accounts with a significant impact on the discussion. The findings can be leveraged to provide insights for planning and executing social media strategies for future global events. This data can be analyzed to provide information on the opinions and sentiments of the users. The paper focuses on this aim and tries to provide insights into the communities interacting during the time period of two popular events. The study uses the findings to gain insights into public discourse, the various types of reactions and attempts to categorize/segment the users' reaction to the global events. The study processed the data extracted from these two events and analyzed the quantitative information to provide qualitative insights into the communities' reactions to these two events and how they differed with one another using graphs. Graphs were created using the Gephi software to accomplish the same. The hashtags formed nodes. The edges between hashtags were directly proportional to the number of the times the hashtags were used with one other. This was done for each hashtag chosen for each of the two global events. The graphs offer information in the form of community cluster detection and degree (frequency). Cluster detection is accomplished using the Force Atlas algorithm for the graph layout. This causes hashtags that are frequently used together to group together – segregating them into separate communities. The size/radius of the hashtag nodes is determined by the degree of the hashtag. These features of the graph visualize the extent of the use of a hashtag, how often certain hashtags are used together, which sentiments, opinions and topics are related to the global event. Further statistical calculations also provided insights into the users' different reactions to the events by comparing datasets extracted from the graphs. The primary mode of comparison between the datasets is based on a calculated value – the 'Relevancy Ratio'. This value is a float between 0 and 1. The higher the value, the more 'relevant' the graph for a hashtag is considered to be. The datasets are created for each global event and include the relevancy ratio. For each dataset, the average (arithmetic mean), standard deviation and range are calculated and used for comparison. The dataset comparison resulted in finding significant differences in the average relevancy ratio and its spread. The Q dataset (hashtags related to Queen Elizabeth II's passing) was found to be more relevant on average but also inconsistent. The U dataset (hashtags related to the invasion of Ukraine by Russia) was found to be less relevant on average but was more consistent despite a larger number of graphs formed.

Introduction

Social Media is broadly defined as a technology or platform which allows for the sharing of content or multimedia among a virtual community [1–4]. The key aspects of a social media platform are the sharing of information and the community. Users of a social media platform can share information in a lot of ways, commonly through text, images, videos and links. The information is shared with other users and can be accessed by the community unless restricted by the user. This allows for interaction among the virtual community, enabling discussions and sharing of opinions.

Online communities exist on a variety of platforms, including social media, forums, and messaging apps. Understanding reactions across diverse platforms can be challenging due to the differences in user demographics, communication styles, and moderation policies. The rapid spread of information, both accurate and inaccurate, on social media platforms poses challenges for researchers. Distinguishing between reliable information and misinformation during global events is crucial. Algorithms used by social media platforms can amplify certain types of content, potentially influencing the direction of online discussions. Understanding the role of algorithms in shaping community reactions is an ongoing area of research. Ethical considerations

and privacy concerns may limit researchers' access to certain data, impacting their ability to study online community reactions comprehensively.

Through social media platforms, people – the users – share their opinions. These opinions show the thoughts of the person as well as their views on topics. As the community interacts, the information shared is an active representation of the thoughts, points of view and feelings of the users. Engagement within the community facilitates an efficient analysis of how a community reacts to information on a certain topic or theme.

Hashtags are keywords or phrases which are used to reference a topic or theme. A hashtag consists of a hash (#) symbol followed by the text. They are used to indicate what some information is related to. This allows users to find, share and react to information on a certain topic or theme easily by searching for the relevant hashtag. Hashtags are a popular part of social media platforms. As multiple hashtags can be used at once, we can track the relations between specific hashtags. These relations can then be graphed and analyzed to see how people connect different topics and themes.

Hashtags aggregate and organize conversations around a specific topic. When users include a hashtag in their posts, it creates a virtual space where discussions related to that topic can be easily tracked and analyzed. Trending hashtags often reflect what is currently capturing the community's attention. Monitoring popular and trending hashtags provides a real-time glimpse into the ongoing discussions and interests within a community. Hashtags reflect the users' opinions and sentiments. Features of hashtags such as degree/frequency allow for the popularity of a hashtag to be easily measured. This makes hashtags convenient to use and work with, while also providing a general overview/reflection of the user community.

Social Media is widespread with 59% of the global population using many different platforms[5]. This provides researchers with a large dataset – one of the largest in the world – to analyze people's thoughts and emotions towards global events[6] – events which are known by the public on a large scale such as political decisions in powerful countries, a major event in a celebrity's life or military action.

As the paper aims to understand reactions to an event, the more the data, the more information can be extracted and patterns analyzed. To accomplish this, a large dataset is needed. The two primary factors deciding the amount of data collected are the platform used (a social media site) and the event chosen to be analyzed.

The social media platform needs to be active and have a large database of users. Most importantly, it needs to have hashtags available which are commonly used. The largest social media sites with an active hashtag usage are Facebook, Instagram, Twitter and TikTok. This study is conducted in India, where TikTok has been banned since June 2020. Due to this, TikTok cannot be considered. Twitter has a character limit when posting tweets. This limits the number of hashtags posted. Due to this, Twitter is also ruled out to maximize the number of hashtags collected. Between Instagram and Facebook, Instagram was chosen due to its accessibility. Instagram is easier to collect data from and hence, the program to collect data can be developed quickly, allowing for more data to be collected within the same timeframe for the research. A pilot program was developed for both Instagram and Facebook in order to extract data through web scraping. The program made for Instagram was shorter, less complicated and overall, easier to develop and expand. Hence, Instagram was chosen over Facebook.

Instagram is one of the biggest social media platforms. The platform has over 500 million DAUs (daily active users) and 2.35 billion MAUs (monthly active users)[7]. Being one of the most popular social media platforms, Instagram's wide use of hashtags provides a vast amount of information to be analyzed and several different insights.

This large dataset can be analyzed to understand users' and communities' reactions to an event. The hashtags can be filtered based on frequency of usage to understand the most popular reactions to an event. Moreover, the study aims to understand the relations between different popular opinions on social media – in this case, Instagram. These reactions can be analyzed to understand and find different sentiments.

This information is extracted through Web Scraping[8]. The term "Web Scraping" refers to the process of extracting data or information from the World Wide Web. The data is "scraped" from the Internet. In this process, software or a bot/code is used to access the World Wide Web via the Internet. The web scraper (software which performs web scraping) uses either a web browser or HTTP (Hyper Text Transfer Protocol) to extract the data from a webpage. A program was developed to gather data through web scraping from Instagram. The code was written in Python and used a function to gather posts based on a given hashtag. These posts were used to extract captions and hashtags used. This method has been elaborated and explained in detail in Method Section.

This study investigates how various hashtags are used on Instagram with each other after a globally known event. The globally known events used are the invasion of Ukraine by Russia on 24th February, 2022 and the passing of Queen Elizabeth II on 8th September, 2022. The primary feature of a globally known event is that it is known by people all over the world. Events were chosen on the basis of how popular they were estimated to be. Besides popularity, another factor considered was the nature of the events. Events may be represented as facts – such as a celebrity's passing but they can also prove controversial such as geopolitical actions. The study also aims to explore how the nature of an event affects community reactions. The events were also chosen on the basis that they took place relatively close. This was to make sure that the number of posts and users on social media would not be affected significantly. Social media

sites grow at a rapid rate. Due to this, events which took place with a large time gap between them can have a large difference in the number of posts and hashtags used.

The invasion of Ukraine is considered a global event due to its significant impact on international relations, geopolitics, and the broader global community. The event involves multiple nations intergovernmental organizations such as North Atlantic Treaty Organization (NATO). This event involves direct conflict between Russia and Ukraine. Moreover, the geopolitics affects several countries. Due to this, this global event is considered controversial.

Queen Elizabeth II has been a symbol of continuity and stability for the United Kingdom and the Commonwealth. As the head of the Commonwealth, Queen Elizabeth II held a symbolic and unifying role for a diverse group of nations. Her passing is significant not only for the United Kingdom but also for the member countries of the Commonwealth, making it a global event. The death of Queen Elizabeth II does not surround controversy and can be considered to be largely interpreted as a fact.

Furthermore, this study is interested in investigating the relevance of posts to the hashtags examined. The relevance of a post refers to how connected it is to a given hashtag. Instagram users receive most of their content through their feed. This feed includes posts from accounts the user follows, suggested posts and ads from businesses that may be interesting and relevant to the user[9]. Instagram's algorithm for displaying a user's feed uses hashtags to promote different posts. More popular hashtags, such as those related to a globally known event, lead to a higher chance of a post being displayed on a user's feed. Hence, many content creators on Instagram will use a combination of hashtags to promote their own content which may not be related to the hashtag used. For example, advertisements may use popular hashtags to promote their post even if the post is not relevant towards the hashtag. This data is redundant and interferes with the analysis of the actual sentiments of the users.

To filter out this noise, a variable, relevance, will be used to determine how many posts are related to an event. The relevance of a single hashtag is a binary value – either 0 or 1. 0 indicates that the hashtag is not relevant while 1 indicates that the hashtag is relevant. This is determined by the identifier (name) of the hashtag itself. A hashtag is deemed relevant if the identifier has a direct connection to the event. The hashtag needs to have some clear features in the identifier which directly correlate to the event. General identifiers cannot be deemed relevant. For example, #peace does not provide clear relevance to the Invasion of Ukraine as it can be related to several other themes. However, #stopwar accomplishes this by directly referencing the ongoing conflict.

Related Work

There have been several studies involving data analysis of data extracted from social media using web scraping. These studies also follow a global event and collect data related to those events. For example, a paper by France Cheong and Christopher Cheong from RMIT University follows the 2010-11 Australian floods[10]. The paper uses hashtags, like this paper, as a parameter to collect posts. The paper also follows a similar analysis pattern by using graphs. A similar approach is followed in this paper but instead of users and user-resources, this paper focuses on hashtags and instead finds relations using hashtags alone. The study also shares a common limitation. The study was rate-limited (restricted) in the bandwidth they could use for extracting data from Twitter. This paper also faced the exact same issue but with Instagram. Limited bandwidth restricts the amount of data you can obtain through web scraping and acts as a limiting factor regardless of the bandwidth of the computer being used to web scrape.

Another paper, Opinion Mining on Social Media Data by Po-Wei Liang and Bi-Ru Dai from the National Taiwan University of Science and Technology[11], uses an algorithm to analyze the data found from web scraping. This paper follows a similar approach to microblogging and social media sites and emphasizes on the valuable data that can be found on the mentioned sites. The paper focuses more on sentiment analysis and opinion mining and analyses it using a proposed system architecture (the main theme of the paper). This study has a similar aim but instead of an algorithm or system, the data is analyzed through graphs. It offers for a more qualitative analysis as compared to the quantitative approach in the paper by Liang and Dai.

In a paper by Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan[12], web scraping is done to extract tweets from Twitter, another popular microblogging and social media site. These tweets were searched for specific terms or keywords but no hashtags. Later, however, a specific hashtag seen to be occurring frequently – #engineeringProblems – was analyzed. This paper follows a more direct approach and searches for the most popular hashtags directly. The difference can be credited to the time frame in which the web scraping was done. While the paper by Chen, Vorvoreanu and Madhavan searches for tweets in real time, this study collects data from a given timeframe, making it easier to find the most popular hashtags.

A paper by Arif Himawan, Adri Priadana and Aris Wahya Murdiyanto[13], focuses on account based data collection through web scraping. Although the method used falls under web scraping, it is very different from the method this paper uses. The method used in the referenced paper interacts directly with the framework of Instagram. It directly accesses the HTML of an Instagram page and then uses a library called BeautifulSoup to handle/parse the HTML document and access the elements in an organized and efficient way. The data is then exported/saved in

various formats – CSV, JSON and EXCEL. This study instead uses an Instagram based library which searches posts itself and then uses different other libraries and snippets of code to extract data. The study mentioned only focuses on the account data extraction whereas this paper also performs analysis on the data extracted.

Results

The global events searched were the Invasion of Ukraine by Russia and the death of Queen Elizabeth II. For both events, I skimmed Instagram for the most popular relevant hashtags at the time of the event.

Before the search was conducted, the following hypothesis was made regarding both global events:

The death of Queen Elizabeth II will provide more relevant hashtags and clear communities/clusters will be repeatedly seen in the graphs. The Invasion of Ukraine by Russia will provide less relevant hashtags and similar communities are not expected to be seen in different graphs.

The reasoning behind this is based on the nature of the respective events. The death of Queen Elizabeth II is a fact which is, although considered tragic, widely accepted. The majority of the world and especially, the United Kingdom was saddened by it. The unity and similarity in the reactions seen in the real world led to this hypothesis under the assumption that social media reactions would follow a similar pattern.

The direct opposite is predicted for the Invasion of Ukraine by Russia for the same reasons. War and invasions are extremely controversial events and the primary difference is seen between the two countries directly involved (Russia and Ukraine). Different communities and groups have different and various opinions on the same, resulting in various possible clusters. Most people, especially popular Instagram accounts, would refrain from commenting on the event so as to not hurt their social media following. This, combined with the popularity of the Invasion of Ukraine hashtags, means that the accounts could use those hashtags to stay popular or gain popularity while not also not offending any users – resulting in a lower relevancy ratio.

The web scraping program searched from 24th February, 2022 for the invasion of Ukraine and from 8th September, 2022 for the passing of Queen Elizabeth II. Each search continued for four weeks (28 days) after which it was stopped. The web scraping program ended the search on 23rd March, 2022 for the invasion of Ukraine and on 5th October, 2022 for the passing of Queen Elizabeth II.

The start dates were based on when the events started. This was done so all hashtags and posts were included directly from the start. For the passing of Queen Elizabeth II, 8th September, 2022 was decided as that was the date when Buckingham Palace announced the death of Queen Elizabeth II. Similarly for the Invasion for Ukraine, the date chosen was 24th February, 2022 as this was when Russia began invading inside the Ukraine border.

In order to minimize the effects of any possible external factors, the duration of the search was kept constant for both events. The duration was set at 28 days. This was so that a continued discussion of the events could be captured. The duration was limited to 28 days for primarily two reasons: limited computational power and depreciation of the interest in the events. The search was conducted on a work laptop which also being used for school and entertainment at the time. Due to the limited computational power and time, the search was limited to a 28 days. Also, Instagram trends are prone to sudden rise and sudden fall in popularity. This can cause a decrease in interest of the events and a depreciation in the discussion of the events overtime. After 28 days, the discussion is likely to decrease in its frequency as the community moves to newer trends and topics to discuss.

For the invasion of Ukraine, a total of nine hashtags were searched for.

For the death of Queen Elizabeth II, a total of six hashtags were searched for.

In total, 15 hashtags were searched for.

The hashtags were selected based on popularity during the search window. The most popular hashtags from Instagram during the specific search window were selected and then shortlisted based on if the hashtag was related to the event. This was done so a large amount of data could be collected. The study aims to analyze data and hence, more data is better. It gives broad insights across different perspectives and increases the overall validity of the findings. To collect as much relevant data as possible, the most popular hashtags were chosen.

Many hashtags are popular on Instagram but not all relate to the chosen event. Hashtags were only chosen if the hashtag had a direct mention to the event. For example, #putin refers directly to Vladimir Putin, the president of Russia. Hence, #putin was selected as Vladimir Putin has direct impact on the Invasion of Ukraine by Russia. This trend was followed for each hashtag found with two exceptions: #ukrainegirl and #peace.

#ukrainegirl has an ambiguous identifier. While it directly mentions Ukraine, it is followed by girl. This makes it unclear as to what the hashtag aims to mention. Ukraine is a direct mention but adding 'girl' seems peculiar in the sense that 'girl' does not correlate directly to the invasion. However, this hashtag was chosen as it attracted interest as to what other hashtags #ukrainegirl would attract. Moreover, #ukrainegirl proved to have been used in a lot of posts.

#peace also does not have any direct mention to the Invasion of Ukraine by Russia. However, #peace was selected due to the assumption that it would be used to promote stopping the war.

The hashtags chosen for both events are as follows:

1. Invasion of Ukraine: #kiev, #kyiv, #nato, #peace, #russia,

#russiaukraine, #ukraine, #ukrainegirl, #putin

2. Death of Queen Elizabeth II: #queenelizabeth, #queenelizabethfuneral, #queenelizabethii, #ripqueenelizabeth, #royalfamily, #royalfuneral

In order to keep a constant layout and spread of the graphs, all Force Atlas variables have been kept equal for each graph. The values are shown in the table below:

Repulsion Strength	500
Attraction Strength	1

Table 1 Force Atlas Variables

1The strength variables (Repulsion Strength and Attraction Strength) do not play a critical role in the layout of the graph. They have an effect only in the visualization of the graph. In order to visualize the degree (frequency) of each hashtag, the nodes were kept directly proportional to the degree. This caused some nodes with a large degree to overlap with each other, making it difficult to see smaller nodes and separate communities. In order to overcome this, the graph was designed to be spread out. To achieve this, the repulsion strength was set at 500 (a large value) and attraction strength at 1 (a small value). These values were set so communities would be easier to separate but internal nodes of a community far enough to distinguish. The values were derived after repeated testing for each graph in the Gephi software.

Invasion of Ukraine

Posts Collected	339
Nodes before filtering	965
Nodes after filtering	76 (7.88%)
Edges before filtering	13937
Edges after filtering	810 (5.81%)
Minimum Degree of filtered nodes	51
Maximum Degree of filtered nodes	611
Modularity Range	0-4, 5 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 2 Hashtag 3.1.1: #kiev

The graph is divided into five communities with different colors: light green, blue, dark green, orange and purple. The most central and the largest node is the “Ukraine” node showing that #ukraine was most widely used with not only #kiev but also with many other major hashtags. “Ukraine” is the most common hashtag in all the posts.

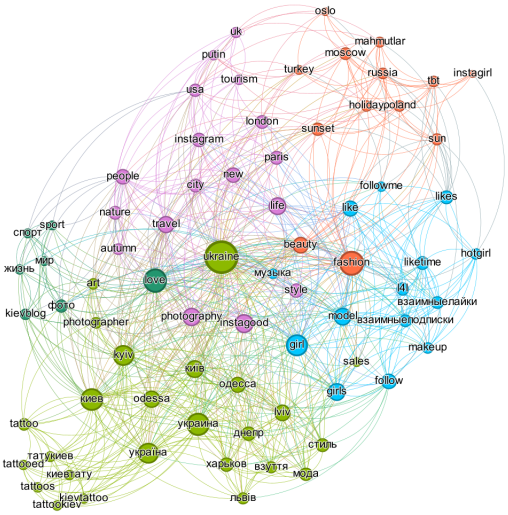

Fig. 1 Graph of #kiev

The blue community has hashtags #model, #girl, #girls, and #makeup which are related to female Instagram models. Other hashtags, including Russian ones, refer to likes and follows being requested by the posts. This is also related to posts which seek attention or user interaction, indicating that these posts were likely posted by accounts that ran businesses and made an income through Instagram. The businesses were likely makeup or other products for females as shown by the mention of #girl and #makeup.

The light green community is the most prominent in the graph. This community is spread out and has two sub-clusters inside itself, one present in the center bottom region of the graph and the other far away in the bottom left corner. This community has the most presence of Russian and Ukrainian characters and words. After translation, hashtags in the central cluster can be seen to be related to fashion as well as city/country names. The hashtags in the bottom left cluster are more focused towards tattoos. The hashtags are either in English, Ukrainian or Russian script and also mention specific locations – namely Kyiv and Odessa (major cities in Ukraine). As these hashtags are in local (Ukrainian) script/language and mention a service (tattoos), they are likely an indicator of local businesses in Ukraine which run promotions through Instagram.

Although both clusters belong to the same community and have the same modularity class value, they can be seen to be some distance apart in the graph. This indicates that the connections (edges) between the nodes are strong enough to be counted as a single community. However, the nodes form subsets (clusters) or groups which have a much stronger connection. This shows that the hashtags within the clusters are used together but not enough to form a separate community, indicating an

intermediate level of connection.

The purple community can be seen spreading out from the center of the graph towards the top. It mostly consists of tourism-related hashtags such as names of countries and cities, #tourism and #travel. This community transitions into the orange community in the top-right of the graph where the tourism theme merges with the vacation/holiday and nature theme. The orange community has mentions of places like Moscow, Russia and Turkey which are used along with sunset, sun and vacations. This could indicate that the posts mostly consisted of pictures of natural beauty in the mentioned places as supported by the use of #photography and #nature in the purple community.

The dark green community is much smaller than the other communities but contains one of the most used hashtags on Instagram - #love. This community has hashtags about sports, love and life indicating that most posts in this theme are related to enjoying life and recreational activities such as sports.

In summary, the five communities in this graph have the following themes:

1. Green: tattoos, photography, cities and countries in/near Ukraine
2. Purple: tourism, nature, photography
3. Blue: models, make-up, business, promotion
4. Orange: nature, sun, cities in Russia, vacation/scenic destinations
5. Dark Green: sports, enjoying life, love

It is interesting to note that despite having the #kiev hashtag, none of the major hashtags were related directly to the aspects of the invasion going on at the time. It can be concluded that the #kiev hashtag was more focused on the mentioned themes rather than the invasion of Ukraine, the country of which Kiev (formally called Kyiv) is the capital.

Posts Collected	345
Nodes before filtering	695
Nodes after filtering	50 (7.19%)
Edges before filtering	8688
Edges after filtering	348 (4.01%)
Minimum Degree of filtered nodes	36
Maximum Degree of filtered nodes	488
Modularity Range	0-4, 5 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 3 Hashtag 3.1.2: #kyiv

The graph is divided into five communities represented by the colors purple, light green, blue, orange and dark green. Note that

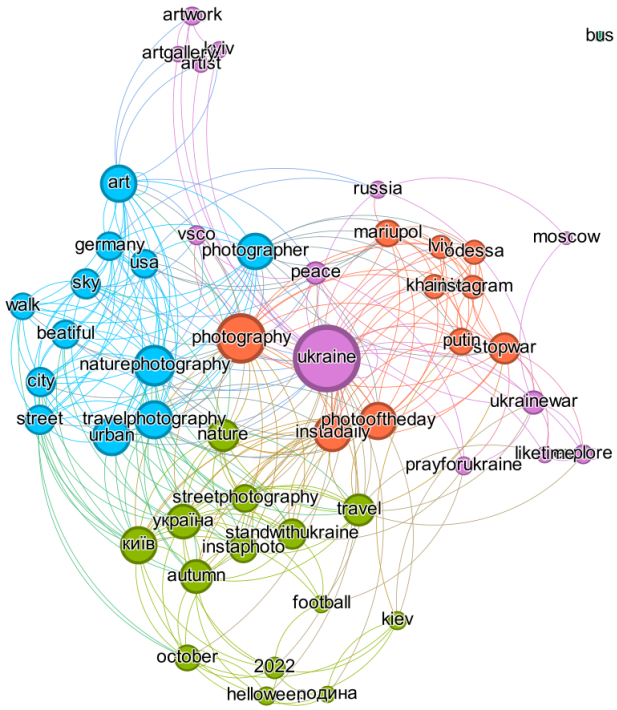

Fig. 2 Graph of #kiev

the dark green "community" is counted as a separate community because it is not connected to any of the other majorly used hashtags. It consists of only one single hashtag - #bus – which can be seen to be a stray node, disconnected from the rest of the graph. This means that this specific hashtag was not used with any of the other hashtags not even once. However, it did occur several times in the posts collected. As the hashtag was not used with any others, I will not consider this node as a part of the rest of the graph or as a separate community.

The main central node is #ukraine and is surrounded by the communities uniformly on its sides. The blue, orange and green communities are central and uniformly distributed while the purple community is divided into two sub-clusters, one at the top of the graph and one in the center and right side of the graph.

The blue community covers the tourism and photography theme. The mentions of places - #usa and #germany support the tourism theme. The photography theme is divided into two components. It includes hashtags of activities and experiences of urban life which are captured as photographs – namely #walk, #city, #street, #ubran and #travelphotography. The other component is #naturephotography which merges into the green community with mentions of #autumn and #nature.

The blue community can also be seen merging towards the top with the purple sub-cluster consisting of mentions of art, expanding on the creativity theme carried by photography. The art sub-cluster involves both the purple and blue community and follows an exclusive art theme only with hashtags such as #art, #artgallery, #artwork and #artist. It can be seen to be distant from the rest of the graph meaning that it is connected with only some central hashtags.

On the right side of the graph, the remaining purple community merges with the orange community to form a theme related to the invasion or war. Hashtags following the countries and war include #ukraine, #moscow, #russia, #putin, #prayforukraine, #ukrainewar and #stopwar. This combined cluster focuses on the countries involved and stopping the invasion/war. This community is the only section of the graph which is related to the main focus – the invasion.

The green community has very diverse hashtags and has no one distinct theme with hashtags that seem random such as #football or #halloween. The top left section of the green community is connected to the blue community but the bottom division of the green community can be seen as completely random and disconnected from the rest of the graph. In summary, the four communities in this graph have the following themes:

1. Blue: photography, urban life, art
2. Purple: stopping the war, art
3. Orange: stopping the war
4. Green: photography, supporting Ukraine

The two graphs mentioned above for #kiev and #kyiv refer to the same city, the capital of Ukraine. It is officially called Kyiv but the Russian derivation is Kiev. This can be used to understand and compare the two general viewpoints of using the same city as a hashtag but from different sides of the war. The Russian side using #kiev is less focused on the war and follows a more general trend shown throughout all of Instagram while the Ukrainian side using #kyiv is more focused on the war – primarily on stopping the invasion and supporting Ukraine.

The graph has four communities of colors green, purple, blue and orange. The green community is dominant on the right, the purple community at the bottom left and the orange community at the top. The blue community is largely present in the bottom right with some presence in the center.

The purple community is largely related to the aesthetic aspect of aviation and the military. With hashtags such as #avgeek, #boeing, #aviationdaily, #militarylove and #scenic, this community represents the aviation and military fanbase which is used with the #nato hashtag. It also spreads out into the central region where hashtags such as #airforce and #military indicate that these posts could have been made by people who are or were

Posts Collected	372
Nodes before filtering	1042
Nodes after filtering	79 (7.58%)
Edges before filtering	13899
Edges after filtering	705 (5.07%)
Minimum Degree of filtered nodes	43
Maximum Degree of filtered nodes	223
Modularity Range	0-3, 4 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 4 Hashtag 3.1.3: #nato

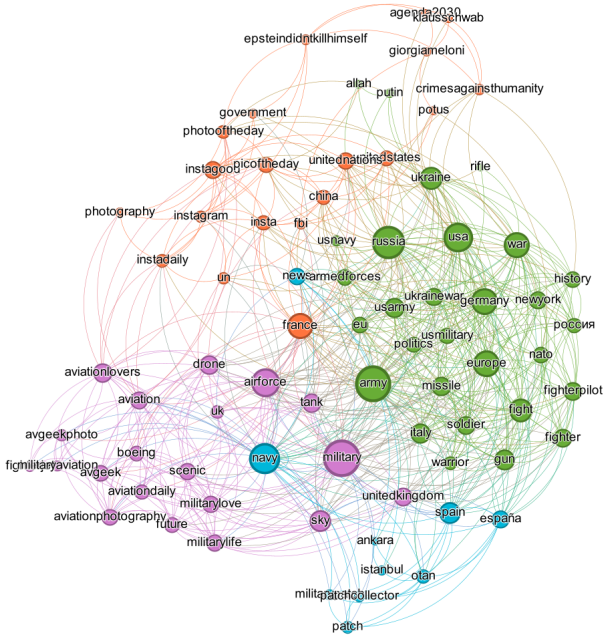

Fig. 3 Graph of #nato

in the military and used Instagram to show their love and/or admiration for the same.

The green and blue communities both primarily relate to European countries and the USA. These communities are more focused towards the political aspect of NATO, with hashtags related to the invasion of Ukraine such as #war, #ukraine, # (Russia). Although all hashtags in the right region are related to Europe, politics and the Ukraine war, it is interesting to note that there is no mention of stopping the war or helping Ukraine during the invasion despite Ukraine's clear intentions of joining NATO.

Moving towards the top of the graph, the nodes stray away from the military and aviation-related topics and transition into

the orange community which is related more to government organizations such as #unitednations and #fbi. At the very top of the graph, the hashtags are focused on political figures, #putin, #potus, #giorgiameloni, #klausschwab, as well as a mention of a conspiracy theory, #epsteindidntkillhimself.

In summary, the four communities of this graph have the following themes:

1. Green: military, Europe, Ukraine War/Invasion, politics
2. Purple: aviation (as a hobby or recreation), military
3. Blue: Europe, military
4. Orange: government bodies, political figures

Posts Collected	444
Nodes before filtering	1563
Nodes after filtering	116 (7.42%)
Edges before filtering	22157
Edges after filtering	1750 (7.9%)
Minimum Degree of filtered nodes	53
Maximum Degree of filtered nodes	780
Modularity Range	0-3, 4 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 5 Hashtag 3.1.4: #peace

The aim of retrieving posts with the #peace hashtag was to see if any users would post about the Ukraine invasion. The assumption was that users would be using #peace to support Ukraine and stop the invasion. This hypothesis was further supported by the #peace hashtag being one of the most popular hashtags.

However, the results show that Instagram users relate #peace to many different themes but not Ukraine. The four communities detected – green, blue, purple and orange – each represent different themes which can be seen as peaceful or relaxing. As none of the communities are related to the Ukraine Invasion, they do not provide any data for opinions or views for the event. Due to lack of relevant data, I shall not discuss this further in detail but only summarize the results as they do not play a major role related to Ukraine. The summary is as follows:

1. Purple: spirituality, religion, Christianity
2. Green: healing, health, meditation, spirituality, thankfulness, positivity
3. Blue: photography, happiness, aesthetic
4. Orange: nature, happiness, fashion

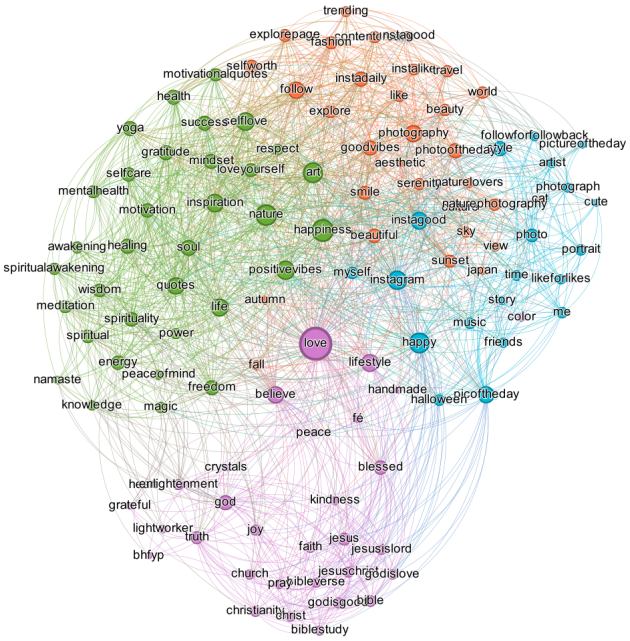

Fig. 4 Hashtag 3.1.4: #peace

The reason for the lack of relevant data for #peace does not have a definite cause. #peace was expected to be related to the Invasion of Ukraine as users would be able to promote #peace. Moreover, #peace is a very popular hashtag on Instagram with over 100 million posts (data from an Instagram search query itself). However, it is possible that #peace's popularity itself can be attributed as cause to the low relevance found in the graph. #peace was likely used in several posts not related to the invasion of Ukraine but to several other themes or topics. Due to its massive popularity, #peace was likely used for other, more popular, themes as compared to being used towards the Invasion of Ukraine. Instagram, however analyzed, is a social media platform which still presents several trends and fads as part of its userbase. It is possible these existing themes overshadowed the Invasion of Ukraine to a large extent, resulting in the lack of relevant hashtags.

The graph is divided uniformly into two major communities – orange and green- which dominate either side of the graph. A smaller minor purple community is scattered throughout the right section of the graph.

The green community is focused on mainly countries and contains mentions of the sport football in #worldcup, #messi, and #soccer. Upon inspection, it is observed that all countries mentioned play football and participated in the 2018 football world cup. This world cup was held in Russia, hence the relation.

The orange community has many themes, but the dominant ones are photography, aesthetic, fashion and modelling. There is

Posts Collected	354
Nodes before filtering	1031
Nodes after filtering	76 (7.36%)
Edges before filtering	15585
Edges after filtering	1040 (6.67%)
Minimum Degree of filtered nodes	53
Maximum Degree of filtered nodes	228
Modularity Range	0-2, 3 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 6 Hashtag 3.1.5: #russia

Posts Collected	411
Nodes before filtering	1180
Nodes after filtering	89 (7.54%)
Edges before filtering	16778
Edges after filtering	1270 (7.57%)
Minimum Degree of filtered nodes	53
Maximum Degree of filtered nodes	497
Modularity Range	0-3, 4 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 7 Hashtag 3.1.6: #russiaukraine

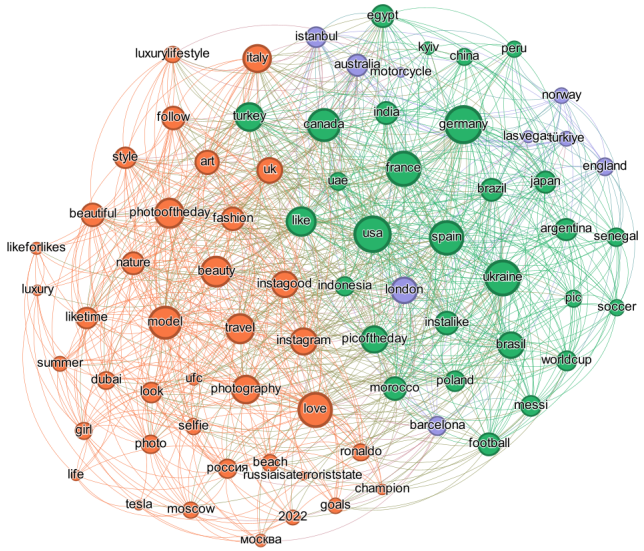

Fig. 5 Graph of #russia

a mention of #russiaisaterroriststate (russia_is_a_terrorist_state). However, this is the sole hashtag present in the graph which cannot be linked to any other theme. It is surprising to see that #russia was used primarily for football and modelling at the time of the Ukraine invasion with only one mention of the invasion going on at the time. As a direct mention of a country involved in the Invasion of Ukraine, the graph of #russia is rather underwhelming as barely any hashtags related to the event are present.

This graph, like the graph of #peace, does not focus on the Ukraine invasion at all and hence, I shall not discuss it further in detail. The summary of the results obtained from communities is as follows:

1. Green: football – countries and players

2. Orange: travel, photography, modelling, art

The graph for #russia is similar to #peace when comparing the lack of relevant data and its likely cause. #russia has a higher ratio of relevant hashtags. However, the cause can still be attributed to trends. Instead of popular Instagram themes like for #peace, #russia shows hashtags related to the 2018 FIFA World Cup™ hosted by Russia. Similar to #peace's low relevance caused by Instagram trends and fads, #russia's low relevance can be attributed to the popularity of the 2018 FIFA World Cup™ hosted by Russia.

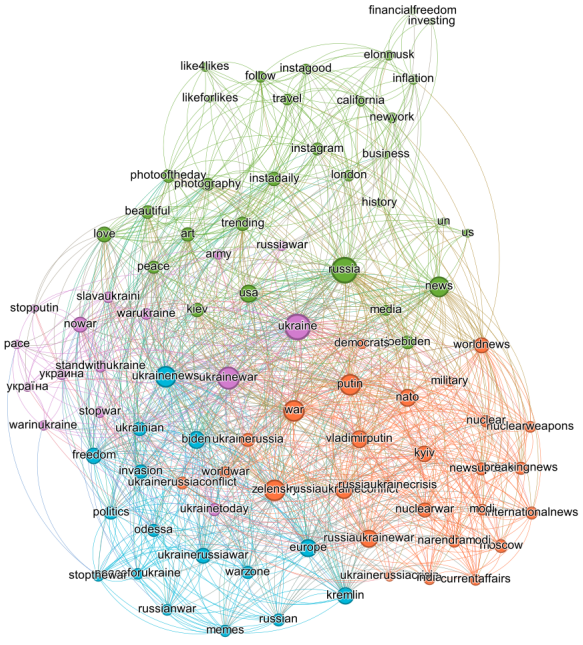

Fig. 6 Graph of #russiaukraine

This graph consists of four communities – blue, green, orange and purple. The orange community is present in the bottom right section of the graph. The blue community is adjacent to the orange community, spreading over the bottom left section. The green community spreads out at the top of the graph with some nodes being scattered towards the upward boundary of the graph. The purple community is scattered throughout the center and left regions of the graph.

The orange community focuses on the political aspect of the invasion/war. It has hashtags such as #nato, #worldnews, #currentaffairs, #nuclearweaponds, #putin and #zelensky. This community represents the posts that act as either informational (news) or opinionated posts. It also brings up the topic of nuclear weapons which is not present in any other graph. This community presents several relevant hashtags.

The blue community focuses on mostly neutral hashtags such as #invasion, #ukrainerussiawar, and #europe with the exception of #stopthewar. This community also has #memes, indicating that the #russiaukraine hashtag was used often with #memes to promote the posts which were likely memes. Moving further up in the blue community, the transition between information and support is made evident by the use of hashtags #freedom and #stopwar.

The purple community is solely supportive of Ukraine in the left region of the graph and opposes war as clearly shown by #nowar, #stopputin, #stopwar and #standwithukraine. This community also has hashtags such as #ukraine, #army and #russiawar which spread out in other communities' regions.

The green community is mostly unrelated to the invasion. Only a few hashtags are related to the invasion while the rest seem to have been used for promotional purposes. One particular unrelated topic where #russiaukraine is used is business and finance. The very peak of the green community has hashtags #business, #financialfreedom and #investing showing that #russiaukraine was used often to promote posts related to finance.

In summary, the four communities of this graph have the following themes:

1. Orange: politics, nuclear weapons, information/news
2. Blue: Russia, Ukraine, war
3. Purple: supporting Ukraine, stopping the war
4. Green: photography, finance, promotional hashtags

This graph has five communities – purple, blue, orange, light green, and dark green. The purple community is the largest and is present in the top and top left regions of the graph. The light green community is present in the right region of the graph with some nodes located further towards the bottom right of the graph. The blue community is present at the bottom of the graph. The orange community is small but has the most popular

Posts Collected	423
Nodes before filtering	1287
Nodes after filtering	98 (7.61%)
Edges before filtering	16037
Edges after filtering	1213 (7.56%)
Minimum Degree of filtered nodes	43
Maximum Degree of filtered nodes	240
Modularity Range	0-4, 5 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 8 Hashtag 3.1.7: #ukraine

Fig. 7 Graph of #ukraine

hashtag used alongside #ukraine – #instagram. The dark green community is scattered around the central nodes of the graph.

This hashtag was analyzed due to its popularity at the time of the global event – the invasion of Ukraine. As it directly mentions the name of the country being invaded, #ukraine was expected to have posts related to the invasion going on at the time. However, the graph produced has very few hashtags which are related to the event. The relevant hashtags are as follows:

1. #nowarinukraine (purple community) – supporting the anti-war stance
2. #standwithukraine (light green community) – supporting Ukraine

3. #vladimirputin, #putin (light green community) – referring to Vladimir Putin, president of Russia

4. #ukrainewar, #war (light green community) – referring to the invasion of Ukraine

As the graph does not provide any other useful insights towards views or opinions, it will not be discussed further.

Posts Collected	399
Nodes before filtering	1074
Nodes after filtering	79 (7.36%)
Edges before filtering	15427
Edges after filtering	1024 (6.64%)
Minimum Degree of filtered nodes	55
Maximum Degree of filtered nodes	684
Modularity Range	0-3, 4 class values
Search start date	24th February, 2022
Search end date	23rd March, 2022

Table 9 Hashtag 3.1.8: #ukrainegirl

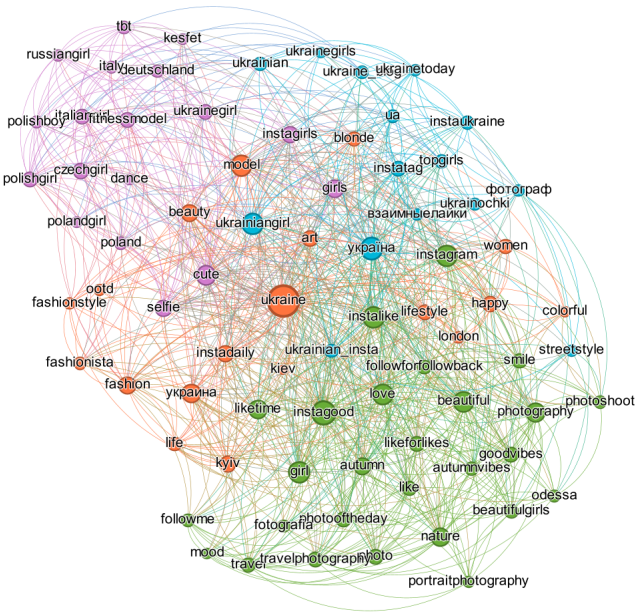

Fig. 8 Graph of #ukrainegirl

This hashtag was chosen due to its popularity at the time of the invasion. Although the hashtag is not related to the war and instead mentions a girl from Ukraine, it was chosen regardless due to its extensive usage. However, this hashtag has also proven to be completely unrelated to the invasion. The graph shows nodes that the hashtags are related to promoting posts and modelling rather than the invasion. This could be said to be expected due to the word 'girl' being present in the hashtag but it continues to show that users on Instagram prefer posts from other topics and themes rather than information about large-scale global events even when they are happening in real-time. Due to the lack of relevant information, this graph will not be discussed in detail.

The summary of the four communities in the graph of #ukrainegirl is as follows:

1. Green: photography, travelling, promotional hashtags

2. Purple: dance, countries

3. Orange: fashion, art

4. Blue: Ukraine, photography

Death of Queen Elizabeth II

Posts Collected	417
Nodes before filtering	1282
Nodes after filtering	97 (7.57%)
Edges before filtering	17890
Edges after filtering	1581 (8.84%)
Minimum Degree of filtered nodes	48
Maximum Degree of filtered nodes	282
Modularity Range	0-3, 4 class values
Search start date	8th September, 2022
Search end date	5th October, 2022

Table 10 Hashtag 3.2.1: #queenelizabeth

This graph consists of four communities: green, orange, purple and blue. The green community is dominant in the center and bottom regions of the graph. The blue community spreads from the bottom right into the green community towards the center. The orange and purple communities are scattered across the top region of the graph on the left and right respectively.

The green and blue communities have a much denser layout as compared to other communities showing that these hashtags were used together more frequently as compared to hashtags in other communities. These communities focus on the British Royal Family, namely its members and residences. This includes Catherine Middleton and Prince William, who are the heirs to the British throne after the current King Charles. These members of the royal family are mentioned in #katemiddleton, #princesofwales, #princewilliams, #dukeofcambirdge and #princeandprincessofwales (prince_and_princess_of_wales), showing that people put more emphasis on the secondary heir to Queen Elizabeth II rather than the primary heir, King Charles. This community also focuses on Lady Diana, a deceased member of the

royal family and former princess of Wales and wife of King Charles. Her popularity shows that even over 26 years after her death, people still remember her as a prominent icon.

The blue community also focuses on a couple from the British royal family – Prince Harry and Meghan Markle, the Duke and Duchess of Sussex. With hashtags such as #princeharry, #dukeofsussex, #duchess, #duchessofsussex and #harryandmeghan being used, they are also displayed as a popular (and controversial) couple in the British royal family. The blue community also mentions another (Amish) royal family with the hashtag #amishroyalfamily, of which there is no official information.

Users likely mention other deceased British royal family members as to remember them. Mentions of some British royal family members are seen several times in other graphs as well. This could indicate sentiments of sorrow as users remember members who had passed away. It is also possible that members who are widely liked by users are mentioned. Repeated member hashtags can be liked, remembered or even controversial.

Moving further away from the denser area of the graph, the nodes in the purple and orange communities can be seen to not be related to the British royal family. Both communities have hashtags related to photography, travel and modelling. Towards the top of the graph, the nodes represent hashtags which would be used with alternative interpretations or uses of the word 'queen' such as referring to drag queens or female celebrities. This is shown by the use of hashtags such as #adele, #arianagrande, #kimkardashian and #dragqueen. These nodes are also very far from the central royal family-related nodes showing that these hashtags were rarely used with mentions of the royal family and were rather used in relation to alternative uses of the word 'queen'. These alternatives uses likely use 'queen' as a term for empowering female celebrities as mentioned in hashtags such as #adele, #arianagrande etc.

In summary, the four communities in the graph of #queenelizabeth have the following themes:

1. Green: British royal family, Lady Diana, Prince William and Catherine Middleton
2. Blue: British royal family, Prince Harry and Meghan Markle, fashion
3. Orange: travel, photography
4. Purple: royal wedding, alternative uses of the word 'queen'

The graph of #queenelizabethfuneral is the smallest graph of both global events analyzed, consisting of only 44 nodes and two communities. The arrangement of the nodes using the Force Atlas algorithm can be seen to sync perfectly with the community distribution calculated by the modularity classes. Both communities primarily follow one topic.

The pink community focuses primarily on the members of the royal family as shown by the hashtags #princeharry, #meghanmarkle, #countessofwessex, #kingcharles and #Britishroyalfamily. The hashtags in this community also consist of #princessdiana and #princephilip, which refer to deceased members of the British royal family. This can show that people still actively remember them or have been reminded of them due to the passing of another member of the royal family – Queen Elizabeth II.

The green community refers primarily to Queen Elizabeth II and her life. Hashtags such as #queenelizabeth, #queenelizabeth2 and #hermajesty refer directly to Queen Elizabeth II. Other hashtags refer to Queen Elizabeth II's life, namely #windsorcastle, #windsor and #buckinghampalace, referring to Windsor

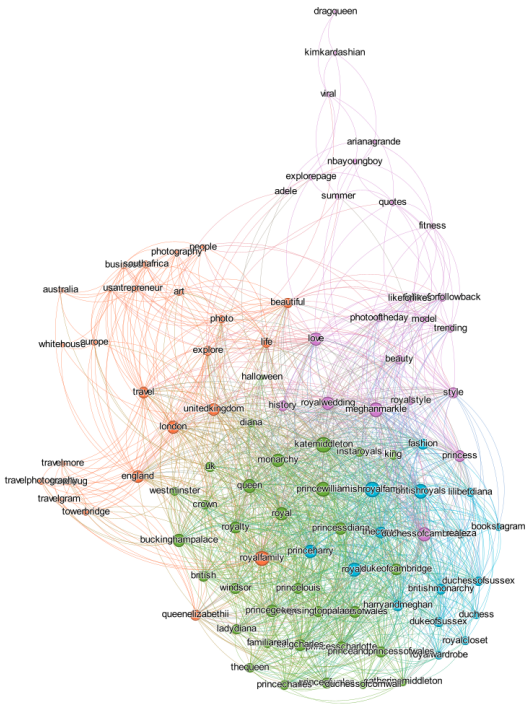

Fig. 9 Graph of #ukrainegirl

Posts Collected	441
Nodes before filtering	603
Nodes after filtering	44 (7.3%)
Edges before filtering	7810
Edges after filtering	567 (7.26%)
Minimum Degree of filtered nodes	49
Maximum Degree of filtered nodes	282
Modularity Range	0-1, 2 class values
Search start date	8th September, 2022
Search end date	5th October, 2022

Table 11 Hashtag 3.2.2: #queenelizabethfuneral

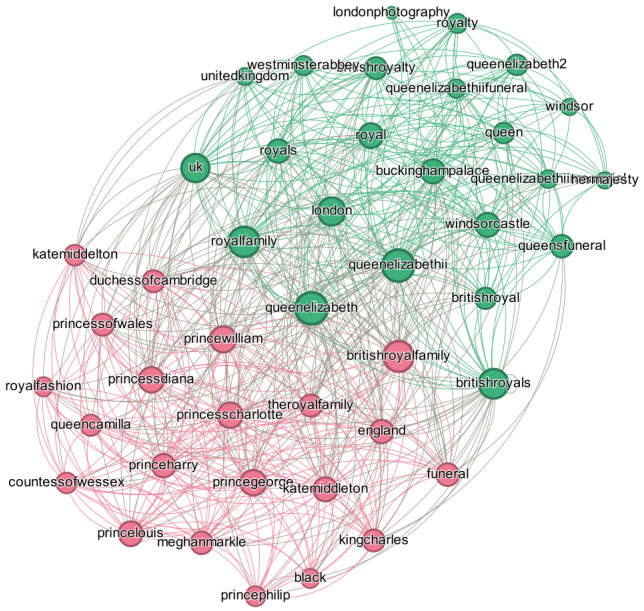

Fig. 10 graph of #queenlizabethfuneral

Castle and Buckingham Palace. These locations were where Elizabeth II lived and worked indicating important locations in her life.

There are still some irrelevant hashtags, namely #londonphotography and #black. Even in small graphs, it is evident that there will always be irrelevant hashtags which will act as noise that cannot be unfiltered by only degree filters. These hashtags, even if irrelevant, are used enough to appear on the filtered graph. The summary of the communities for the graph of #queenelizabethfuneral is as follows:

1. Pink: members of the royal family (current and deceased)

2. Green: Queen Elizabeth II and her life

This graph is also small, only one node larger than the graph of #queenelizabethfuneral but has three communities. All nodes are arranged together with adjacent nodes being approximately the same distance from one another. The purple community is dominant in the right region of the graph. The green and orange communities have scattered nodes in the central and left regions of the graph.

The purple community focuses on the members of the royal family. The graph shows mostly the same recurring members being mentioned such as #princeharry, #meghanmarkle, #princewilliam and #princessdiana, #princephilip showing that some members of the British royal family are more popular

Posts Collected	414
Nodes before filtering	602
Nodes after filtering	45 (7.48%)
Edges before filtering	7603
Edges after filtering	752 (9.89%)
Minimum Degree of filtered nodes	57
Maximum Degree of filtered nodes	230
Modularity Range	0-2, 3 class values
Search start date	8th September, 2022
Search end date	5th October, 2022

Table 12 Hashtag 3.2.3: #queenelizabethii

than others and even deceased members are still remembered actively.

The green community expands on the purple community's hashtags related to the royal family and also mentions some residences and locations related to the royal family such as #windsor, #Britishroyals, #kingcharles and #Britishmonarchy.

The orange community focuses directly on Elizabeth II excluding the node with #royalfamily. All other nodes in this community mention Elizabeth II directly either by referring to her as the queen or by referring to #thecrown, a series based on Queen Elizabeth II's reign.

#thecrown is directly related to Elizabeth II so although it is a series, it can be considered to mention Elizabeth II. The conclusion of #thecrown as relevant would make the graph of #queenelizabethii the only graph to not have any irrelevant hashtags. The summary of the three communities in this graph is as follows:

1. Purple: members of the British royal family

2. Green: members of the British royal family and related locations (residences and offices)

3. Orange: Elizabeth II

This graph has the most nodes of all graphs included in this paper after filtering – 116, the same as for #peace and consists of four communities – purple, green, orange and blue. All communities are primarily spread out over four sides of the graph and converge together in the center. The purple community lies in the right region of the graph, the green community on the left, the orange community at the top and the blue community at the bottom.

The purple community focuses on the members of the British royal family. It has hashtags referring to various members of the British royal family, mostly mentioning Prince Williams and Catherine Middleton with hashtags such as #princeofwales, #princessofwales, #dukeofcambirdge, #princewilliam, #princesscharlotte, #katemiddleton and #duchessofcambridge.

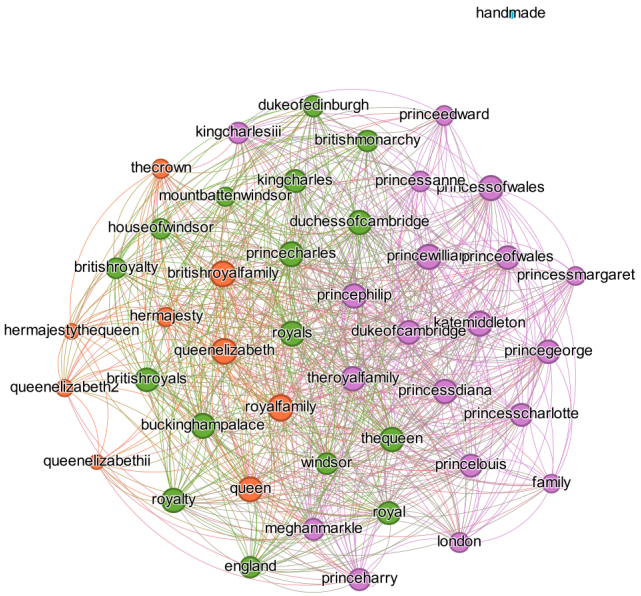

Fig. 11 Graph of #queenelizabethii

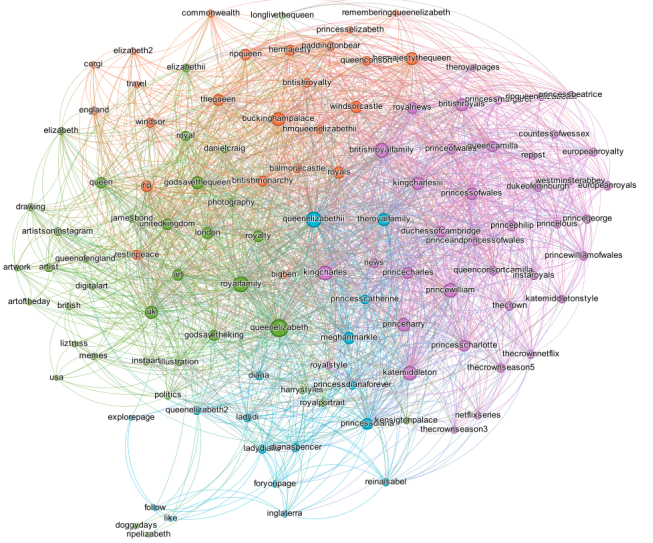

Fig. 12 Graph of #ripqueenelizabeth

Posts Collected	1299
Nodes before filtering	1538
Nodes after filtering	116 (7.54%)
Edges before filtering	21184 (10.15%)
Edges after filtering	2151
Minimum Degree of filtered nodes	45
Maximum Degree of filtered nodes	847
Modularity Range	0-3, 4 class values
Search start date	8th September, 2022
Search end date	5th October, 2022

Table 13 Hashtag 3.2.4: #ripqueenelizabeth

This community also refers to The Crown, a series by Netflix, based on Elizabeth II with hashtags such as #thecrownnetflix, #thecrownseason5 and #thecrownseason3.

The green community focuses on Elizabeth II but also has some irrelevant nodes. Irrelevant hashtags such as #usa, #memes and #art are popular on Instagram so it is expected to see some of these hashtags. However, the hashtags #jamesbond and #danielcraig are unexpected as they both refer to the fictional character James Bond, played by Daniel Craig recently. There are no other hashtags which might suggest a relation to James Bond so it is strange that these two hashtags are present in the top 7.54% of nodes.

The orange community focuses on remembering Elizabeth II and showing empathy towards her. Hashtags such as #rip, #restinpeace, #rememberingqueenelizabeth, #hmqueenelizabeth and #longlivethequeen represent posts which show sympathy towards Elizabeth II and show respect. Other hashtags simply refer to Elizabeth II such as #hermajesty and #elizabeth2. The community also mentions #corgi, a dog breed owned and preferred by Elizabeth II during her reign. This community also contains a hashtag which stands out because it has no relation to any other hashtags in the graph - #paddingtonbear.

However, a possible connection can be formed between James Bond and Paddington Bear. Posts related to Paddington Bear may mention its creator, Michael Bond as just Bond. The use of #bond could be used to link it to James Bond-related posts. Although there is a possible connection between James Bond and Paddington Bear, there is no clear connection of either character to Queen Elizabeth II.

The blue community is mostly related to Lady Diana, a deceased member of the British royal family. This community has few nodes compared to other communities and is focused on Lady Diana with mentions of #meghanmarkle and #princesscatherine. The summary of the four communities in this graph is as follows:

1. Purple: members of the British royal family, Prince Williams, Catherine Middleton

2. Green: Elizabeth II, James Bond

3. Orange: Elizabeth II (respect, sympathy), Paddington Bear
4. Blue: Lady Diana

Posts Collected	756
Nodes before filtering	1962
Nodes after filtering	147 (7.51%)
Edges before filtering	29566
Edges after filtering	2768 (9.89%)
Minimum Degree of filtered nodes	56
Maximum Degree of filtered nodes	669
Modularity Range	0-2, 3 class values
Search start date	8th September, 2022
Search end date	5th October, 2022

Table 14 Hashtag 3.2.5 #royalfamily

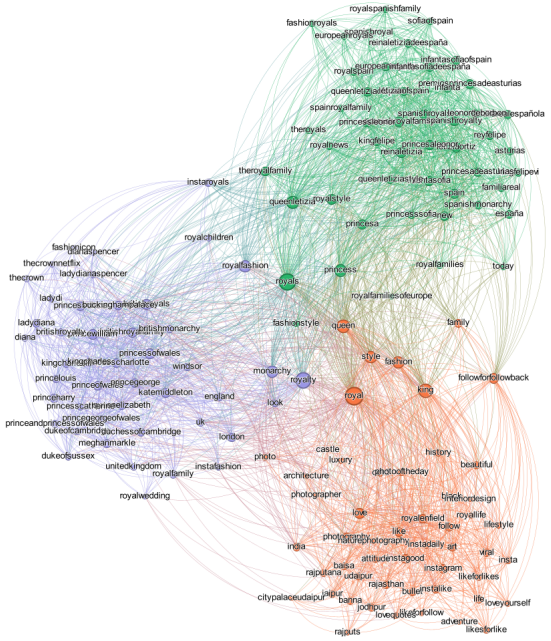

Fig. 13 Graph of #royalfamily

This graph can be seen to be divided into three major communities, both visually and using the modularity class. Each community is far away from the other and forms its own clusters. The three different communities represent three different royal families which can be explained by the hashtag used to generate the graph – #royalfamily. The three royal families shown are the British royal family (purple), the Spanish royal family (green) and the Rajasthani royal family (from Rajasthan, India). As I am only concerned with the British royal family and Elizabeth II, I will only see those nodes.

However, these are a very low number of nodes and concern only one topic – members of the royal family (a theme prevalent in all other graphs in section 4.2). As there is no large difference between the degree (number of occurrences in data) of the nodes, there is simply no data to analyze. The graph only provides the members of the British royal family as nodes – there is no proper data to be analyzed. Hence, I will not discuss this graph further.

Discussion

Statistics

The statistic discussed here will be a variable called 'relevancy ratio'. As the name suggests, this variable is defined as the ratio of the relevant to the total nodes in a graph. For each graph, the total number of 'relevant' nodes will be counted. These relevant nodes are defined as any nodes with the corresponding hashtag being related to the hashtag of the graph. The aim of this statistic is to find out what ratio or percentage of the total nodes in a graph are directly related to the topic rather than hashtags which may be related to irrelevant themes such as photography or tourism. For example, in the graph of #queenelizabethii, the node with #handmade is not relevant to the hashtag of the graph and will be deemed irrelevant. The same graph includes #thequeen, which is relevant to the topic and will be deemed relevant.

Relevancy ratio relies on the assumption that relevance can be determined directly by the hashtags used. Hashtags can play an important role in determining relevance. It is possible for Instagram users to use trending hashtags in order to promote their posts. However, relevant hashtags are still used. For example, several promotional hashtags were seen in the data collected. However, they were accompanied by relevant hashtags. Posts collected contain hashtags – both relevant and irrelevant. In order to quantify and attempt to measure relevance, both types of hashtags need to be considered. While promotional posts use relevant hashtags in order gain traffic, they also use irrelevant hashtags. Hence, irrelevant hashtags account for irrelevant posts. Posts could be analyzed individually with their hashtags and content both determining relevance. However, this violates ethics and Instagram's terms of use as automated collection of private information is not allowed. Furthermore, individually analyzing would require storing specific posts and accounts, which could potentially violate anonymity as it would include storing personal accounts and their information. In order to work with limited data, while still taking into account each aspect, relevancy ratio provides a satisfactory estimate of relevance. There are several qualitative variables which each play a role in determining relevance. Relevance ratio simplifies this complicated process by simplifying relevance to hashtags. There is a possibility that posts with multiple relevant hashtags actually

have irrelevant content. This can skew results and hence, lower validity of the new metric. However, as the metric accounts for irrelevant hashtags as well, it retains some validity. Although the metric is not 100% accurate or has 100% validity, the study is done on a large collection of hashtags – reducing the effect of outliers and hence, providing an estimate of relevance.

The formula for calculating the relevancy ratio is as follows:

Relevancy Ratio = Number of relevant nodes / total number of nodes

which may be shown simply as:

$R = \frac{(RelevantNodes)}{TotalNodes)}$

where R denotes relevancy ratio. For sake of simplicity, I will denote Relevancy Ratio as R from now on as it is shorter to read/write as well as easy to understand as a ratio. The following table includes the calculations done for this statistic for both events chosen. The values of R have been rounded off to four decimal places.

I divided the dataset into two parts based on the event the hashtag is based on. This produces two datasets for the passing of Queen Elizabeth and the Invasion of Ukraine. For convenience, these datasets will be denoted using Q and U respectively. The values in the datasets are as follows:

$Q = 0.4948, 0.8864, 0.9333, 0.6897, 0.3197$

$U = 0.1447, 0.3000, 0.1266, 0.0086, 0.0789, 0.4607, 0.0506$

In the U dataset, the two smallest values are 0.0086 and 0.0506. These values correspond to the graphs of #peace and #ukrainegirl respectively. These hashtags were selected arbitrarily even with no direct mention to the Invasion of Ukraine. Due to this reason, and their very low relevance, they have been removed from the U dataset in order to increase validity of the findings. The revised datasets are as follows:

$Q = 0.4948, 0.8864, 0.9333, 0.6897, 0.3197$

$U = 0.1447, 0.3000, 0.1266, 0.0789, 0.4607$

I will explore two different sets of data, each with its own distinct nature. By examining their characteristics, patterns, and relationships, I aim to gain insights and draw meaningful conclusions. It is important to note that these two datasets differ significantly in their nature and audience reception. The first dataset revolves around the controversial Invasion of Ukraine, which involves geopolitics, major governments, and conflicts between parties. In contrast, the second dataset pertains to the passing of Queen Elizabeth II, an event widely accepted as a fact with less controversy and a more unanimous audience. Considering these differences, the conclusions drawn from each dataset will be tailored to their respective contexts and will take into account the unique dynamics surrounding each event.

The following table shows the average (arithmetic mean), standard variation and range of the values in each dataset.

The arithmetic mean is used as an average calculation. The standard deviation is used as a measure of spread. Range is also used to measure spread. However, the range measure does not hold up for datasets of a small size. The range only takes two values (the maximum and minimum) into account. Hence, it cannot represent the dataset as a whole but only two extreme values which are possibly outliers.

Although the data points are limited in number (10 in total for both datasets), some conclusions can still be drawn from the respective datasets based on the arithmetic mean and standard deviation.

The standard deviation and arithmetic mean take into account every data point available. Hence, they represent the whole dataset. Standard deviation accurately measures the spread of the dataset with respect to the mean. Both measures take into account each datapoint with an equal weightage. This makes both measures high in validity as they represent, uniformly, the whole dataset.

There is a clear difference in the mean of the datasets. The Q dataset has a much higher mean than the U dataset. The Q dataset has, on average, almost four times the proportion of relevant posts as the U dataset. The Q dataset has much higher values than the U dataset as shown by this comparison. The mean of Q is 0.6648 while the mean of U is 0.2222. The mean of Q is almost three times that of U. Q has a mean approximately 199% greater than the mean of U. The means differ by 0.4426, almost 45% of the range of possible values from 0 to 1.

Although there is a massive difference in values, the Q dataset is much less consistent with a higher standard deviation and a higher range. The Q dataset has more spread out values as compared to the U dataset which has more consistent and less spread out data.

The Q dataset is found to be larger in terms of values than the U dataset, but the U dataset is found to be more consistent and less spread out. Note that the findings based on the comparison of the datasets is purely mathematical. They are as valid as the data used – obtained from the Relevancy Ratio. The findings from this comparison are limited in validity owing to the limitations of the Relevancy Ratio.

Difference in Datasets

From previous section, it can be concluded that on average, from all posts analyzed, hashtags from Queen Elizabeth's passing very more but are much more relevant. This shows that although there was a major difference in the relevant posts for each hashtag, there were more relevant posts as compared to hashtags from the Invasion of Ukraine. This means that a more percentage of people used the Ukraine hashtags for promotional and other purposes as compared to posts actually relating to their actual (intended) content.

The clear and significant difference in the datasets support the hypothesis. The relevancy ratio was higher on average for the Q dataset compared to the U dataset. While not explored in detail, the repeatedly occurring communities can be seen for the death of Queen Elizabeth II. Popular figures related to the British royal

Graph Hashtag	Total Nodes	Relevant Nodes	R
#kiev	76	11	0.1447
#kyiv	50	15	0.3000
#nato	79	10	0.1266
#peace	116	1	0.0086
#russia	76	6	0.0789
#russiaukraine	89	41	0.4607
#ukraine	98	18	0.1837
#ukrainegirl	79	4	0.0506
#queenelizabeth	97	48	0.4948
#queenelizabethfuneral	44	39	0.8864
#queenelizabethii	45	42	0.9333
#ripqueenelizabeth	116	80	0.6897
#royalfamily	147	47	0.3197

Table 15 Summary of Graph Hashtags

Parameter	Q	U
Arithmetic Mean	0.6648	0.2222
Standard Deviation	0.2323	0.1404
Range	0.6136	0.4521

Table 16 Summary of Parameters Q and U

family can be seen multiple times in different graphs. The same cannot be said the Invasion of Ukraine by Russia. The graphs showed many different communities and repeated topics were not as frequent compared to the graphs of the death of Queen Elizabeth II.

However, the findings cannot be claimed as fully valid. The statistics alone do not show a direct cause-and-effect relationship. The difference between datasets is clear, but only shows a difference, not possible causes. Causes can only be deduced and assumed based on likeliness (probability) dependent on several other variables not considered such as the actual content of the post or opinions mined through captions. As this was not done due to ethical concerns and Instagram's terms of use, a direct cause cannot be deduced.

The hypothesis stated that the Q dataset would have more relevant graphs (based on the relevancy ratio). This statistical/mathematical statement was found to be true. However, the cause does not necessarily match the reasoning behind the hypothesis.

The datasets Q and U have 5 and 7 data points respectively. The number of data points is low. This may cause lower validity as the dataset is small. A larger dataset would provide more validity. However, there is a significant difference between datasets. There is a large difference between the arithmetic mean. Due to the small size, there are a limited number of comparisons that can be made – which proves as a disadvantage. However, the comparisons that can be made show a significant difference and can lead to conclusions.

Methods

To find relations between hashtags, hashtags related to the events are chosen. Each hashtag searched provides some posts which would also have other hashtags mentioned.

Web Scraping

The programming language used for web scraping was Python – versions 3.10 and 3.11. Python is a versatile and easy language to work with due to minimal syntax and flexibility in terms of variables and data structures. Python also offers a wide variety of libraries which are necessary for web scraping. This makes developing a Python program faster and easier as compared to other languages such as C++ or Java. While other languages may be faster for execution, this specific application for web scraping is limited by network bandwidth. Thus, the program can only extract data as fast as the bandwidth permits. The other processing done for the data takes much less time to execute and hence, a program written in a faster language would only be marginally (and negligibly) faster.

The program uses the Instaloader library[7]. This library is used to find posts from a given hashtag and store the post as an object. This object is then used to retrieve other features of the post such as captions, date posted and most importantly, other hashtags. The program is split into three sections – initializing, finding and storing.

Initializing involves importing all external functions and libraries needed as well as declaring all global variables. The

program imports the Instaloader, Itertools and Datetime libraries. The Instaloader library is used to extract data from Instagram. The Itertools and Datetime libraries are used to iterate through all the posts in a given period. The global variables declared are a Hash Map (Python dictionary) to keep track of each hashtag's results and two counter variables to keep track of the number of unique hashtags and posts processed.

Finding involves searching for the posts with the given hashtag. I use the Instaloader library to log in to Instagram via a personal account and then make a scraper bot using the Instaloader.Instaloader class. The personal account is used to avoid creating a new account solely for web scraping. This is against Instagram's terms of use. To comply with the terms, the account used was an existing one. Existing accounts have preferences, followers and follow other accounts as well. These factors affect the account's recommended accounts and posts. However, the study solely focuses on searching posts by hashtags. This removes any biases that may be introduced by personal accounts. The posts collected are from a hashtag search/query which only presents posts with that specific hashtag present.

The bot is used to search for posts with a given hashtag in a specified timeframe. To search for posts, a function named 'collect_posts_from_hashtag' was defined which takes a hashtag, a start date, and an end date as parameters. The function uses the scraper bot to retrieve posts from Instagram and iterates through each post.

Specifically, the get_hashtag_post function from the Instaloader library is used. This built-in function collects all posts from the given hashtag and return an iterator object. This function is very useful as it provides all relevant posts without needing much code.

Each post is processed and its data, namely its caption and all the hashtags used, is stored. Storing involves saving the data collected in a text file. Two text files are generated for each hashtag which store the data extracted from each post processed and the number of times other hashtags were used. As the caption and hashtags are to be stored and they may contain unknown Unicode characters. To avoid any errors between data types or string processing, all reading and writing from the text files are done as bytes encoded using the standard utf-8 format. Each post follows a specific protocol to store the data by using delimiters in between different posts. The count of other hashtags' occurrences relative to another hashtag is stored as a local Hash Map variable. This variable is then used to store the count of other hashtags in a text file.

Graphing

To find the relations between the hashtags, I used basic Graph Theory to make an undirected and unweighted graph. I used nodes to represent all different hashtags. In each post, I found all hashtags used and used those hashtags to form edges. Repeating this process for each post, I formed multiple edges between hashtags. A second Python script was used for this. This script reads the stored data in the text file and uses the data to make a graph by forming edges between hashtags in the same post. This graph is then used to make a .gexf file. The .gexf file was then used in Gephi to visualize the data.

The software used to generate the graphs was Gephi version 0.10. Gephi offers a wide range of features including different graph algorithms, several statistics which can be calculated, a simple user interface and the ability to edit nodes based on their attributes. In addition, Gephi is also free to use. These features make Gephi a perfect choice for visualizing graphs.

The relation I sought to find in the graphs is the community formation. A community is defined as a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network [14]. Hashtags which are used more often with each other should form communities or groups and hashtags not used with each other should be further apart.

This layout requires communities to form based on their connections. Hashtags frequently used together will form communities. Similarly, distance between communities will also be determined by how frequently they're used together. To achieve the same, a force directed layout is needed. Force directed layouts will form layouts on the basis of connections (edges) and their density. The ideal layout should be fast enough to run efficiently but also provide high quality graphs. The aim of creating graphs is to create a visualization to clearly differentiate clusters, both through community formation and modularity. Different force directed algorithms were tested to find the optimal algorithm.

The Fruchterman-Reingold algorithm was tested first. This layout treats edges as springs and aims to minimize the (strain) energy stored in springs. This layout works well for a large number of nodes. However, after filtering the graph to remove noise, the number of nodes decreases significantly. After this, despite strong connections between certain nodes, the layout provides a uniform distribution. This is not what is wanted. The graph does not show any communities but instead a uniform, circular distribution.

A similar problem is faced by using the Yifan Hu algorithm. This algorithm formed some communities. However, they are not well defined. Repeated tweaking and tuning of layout variables can result in community formation in some graphs. However, this algorithm would require iterative methods for finding optimal parameter tuning. To retain validity, the difference between graphs, and especially between events, needs to be minimized. In order to follow the same, the Yifan Hu algorithm was rejected.

The optimal layout was found to be the Force Atlas algorithm. This algorithm works by simulating a physical system to arrange the nodes. Connected nodes attract each other while disconnected nodes repel each other – a similar mechanism to

gravity. For each node pair possible, the algorithm decides to either cause a force of attraction or repulsion. This causes connected nodes to group together and form clusters, separate from other clusters which may form.

However, there is a disadvantage of the Force Atlas algorithm. It has time complexity $O(n^2)$, meaning that the time taken for the graph to form is of the order n^2 where n is the number of nodes. This can cause problems for large graphs with a higher number of nodes. However, the community formation is of high quality, increasing the validity.

To distinguish between nodes, the visual appearance of a node is changed based on its modularity. The modularity of a node is used to differentiate between different communities. Modularity measures the strength of the structure of a graph by comparing the densities of different connections. Nodes in the same community will have stronger connections to each other and hence, have the same value for modularity. This results in each community having its unique modularity value. This numerical value is used to assign a color to each node so different communities are in different colors. This makes the graph easier to interpret and analyze.

In addition to changing the color, the size/radius of the nodes is also changed. Nodes represent hashtags so the graph should have some way of comparing the number of times a hashtag occurs. The number of times a hashtag occurs is the same as the number of edges it has. Hence, the degree attribute (number of edges) is used to calculate how big the node will be.

Noise Filtering

Some users might use popular hashtags to increase user traffic on their posts. To generate posts on a user's feed, Instagram uses an algorithm that takes into account the hashtags used. Using popular hashtags can increase the likelihood of the posts being shown to other users. This way of using hashtags can often result in irrelevant posts being collected. Posts that use trending hashtags will have other hashtags as well which will be related to their actual content. For example, a content creator may use a trending hashtag to promote their post which may not be relevant to the actual event but the post could also have content specific hashtags. To minimize this inaccuracy in data, hashtags that occur at a lower frequency will not be accounted for.

Each graph will have a different number of posts and nodes so I will only consider the ratio in this case. Setting an absolute value will result in different results for each graph. For example, a graph with 965 nodes (#kiev) and a graph with 695 nodes (#kyiv) have a large difference in the total number of nodes. Setting an absolute value such as 50 would result in a difference of the percentage of nodes represented. In this example, #kyiv would have 7.19% of its nodes in the final graph which is significantly greater than 5.18% for #kiev.

I want to have the most relevant nodes only so I will take the top 7.5% of nodes with the highest degree. For example, if a graph has 1282 nodes, I will filter the graph to have the 96 nodes with the highest degree. To achieve this, I will filter nodes to have a minimum degree which would make it so the top 7.5% of nodes are shown in the graph. It is likely that this exact value is not the number of nodes with the minimum degree so I will accept a difference of 0.5% nodes. This makes the range for nodes to be the top 7% to 8% in terms of degree.

Graphs in Gephi

After the .gexf file for a hashtag is created, it is loaded in Gephi and is modified accordingly. First, the nodes are filtered to be in the 7% to 8% range. Then, the statistics feature of Gephi is used to calculate modularity using the Louvain method. The colors of the nodes are assigned accordingly to the modularity class value. The nodes are then assigned a size/radius based on their degree so nodes with a higher frequency appear larger, indicating that the hashtag was used more.

The Force Atlas layout algorithm is then implemented to arrange the nodes of the graph. The repulsion strength (which affects the distance between two unconnected nodes) is set to a high value to make the graph more spread out. The attraction strength is set to a low value for the same reason. Making the graph more spread out reduces any chances of confusion between nodes and also makes the graph easy to interpret and analyze due to the easily differentiable clusters. The graph is then exported as a .png image file.

Ethics in Web Scraping

This sub section concerns the ethics regarding the web scraping performed. Ethics need to be followed and taken care of when performing web scraping. Automated scripts are designed to collect information from various sources on the internet. These scripts can be designed for malicious content such as collecting private info that is not meant to be collected, infringing copyright law or intellectual property law. This can lead to scams, fraud, blackmail, other crimes and can carry moral implications. In order to avoid this, ethics need to be take care of and hence, ethics a crucial part of the data collection process.

The program developed collects posts which are publicly available. Private or hidden posts cannot be accessed through the program. Instagram offers users private messaging and private accounts which cannot be accessed. Any data that is collected through the program is publicly available. Any Instagram user can view these posts. This data is publicly available to all Instagram users. As the program only accesses data available to any public users, it does not violate privacy.

Another ethical concern is that scraper bots can be used to artificially gain traffic. Bots often track specific accounts and

like and/or comment on posts. They can also follow accounts. The bots can be used to artificially gain traffic. This is not allowed by Instagram. The bot also does not interact with any posts. It only views and collects textual data from a post. The bot only collects the caption and then extracts hashtags from the caption. No other data is collected at all. This protects any private information which may be shared in captions. Moreover, the only useful data stored is the hashtags for each post.

A concern regarding Instagram is the network bandwidth. Instagram can be directly accessed at a very high bandwidth. However, an automated process can cause blockages. Automated access requests can clog up networks and are against Instagram's terms. To avoid this, the program itself included a network limiter. The requests to Instagram are sent at a fixed rate so that the program does not cause any problems towards the Instagram servers.

Conclusions

The paper explored the usage of hashtags on Instagram for two global events. The collected data was visualized as graphs and these graphs were then used to analyze various aspects of the users' reaction. This analysis not only gives insight into the several opinions of users but also how the opinions contrast across hashtags, as with #kiev and #kyiv. Deeper analysis can show that this contrast of hashtags goes beyond just reactions or opinions. It can show how different groups of people may have different views about the same topic/subject and how important different groups regard the topic/subject to be.

This analysis may even be used for researching people's views based on region and can have applications in geopolitics and human geography. A long term study spanning several years can be done to observe how different factors affect people's views and can have applications in psychology and business (consumer behavior). The format and methods of the study (using social media to track people's opinions and views) can prove to be useful and beneficial in many fields. Overall, the comparison of the datasets has shown that the hypothesis held up and was proven true.

Limitations

The data is retrieved using a program running on a computer (a laptop). The computer and network components have physical limitations which do not allow for unlimited data samples. For example, the bandwidth of the network used to retrieve data is finite and therefore, data can only be collected at a finite rate. Due to this, the data sampled in this study is limited.

Another limitation is the processor of the computer. The processor can process at a finite rate as well so the graphs are made at that speed itself. The algorithm used for generating graphs is the Force Atlas algorithm which can a time complexity of $O(n^2)$. This complexity, which may be considered fast for graph computations, cannot process a large number of nodes efficiently. Data was retrieved for #putin and #royalfuneral as mentioned in the Results and Discussion section. However, these two hashtags had a large number of nodes in the graph even after filtering – over 550. Due to the high order time complexity of the algorithm, this graph cannot be efficiently processed in the software Gephi. While running the algorithm, there are a lot errors and crashes due to the large amount of memory and processing required. For this reason, these figures could not be generated and hence, cannot be analyzed.

Graph analysis is a major part of the study. Although the graphs are generated through a fixed algorithm (Force Atlas), the analysis done is qualitative. The structure is generated through several computations. However, the basis of all interpretation and analysis carried out is the qualitative. The analysis involves commenting on the distance between communities, the different themes and their relations with the original hashtag for which the graph was made. The summarization of themes and possible causes behind occurrences of communities are generated through reasoning and general information widely available. It is possible that many different interpretations are possible due to this. The mentioned interpretations may vary and do not have a definite validity. Reliability of the analyses can only be concluded upon if multiple researchers carry out the analyses individually.

Instagram users may not represent a diverse and random sample of the overall population. Although the userbase is diverse, the posts collected are not guaranteed to have been posted by a diverse group of users. The platform may attract certain demographics more than others, leading to a skewed representation of users in research samples. Instagram users may have varying socioeconomic statuses, and certain groups may be overrepresented or underrepresented. Research findings may not generalize well to populations with different socioeconomic backgrounds. It is likely that users from specific regions are more likely to engage in specific discussions, which can cause representation of countries or regions to be skewed. This limits the generalizability of the study and can reduce validity.

Future Research

Currently, the research done comprises of minimal variables – ratio of relevant nodes to total nodes, their average, range, standard deviation and the graphs. For a more in-depth analysis of people's sentiments and opinions, research can be conducted with more variables collected from the same data. The captions, if lengthy enough, can be processed for opinion mining and can open up a whole new set of variables to analyze and compare. This analysis would require much more time and processing power for the same data.

Another aspect of the analysis that has not been explored in

detail is the 'repeated communities' part of the hypothesis. This can be accomplished by storing all the hashtags in the graph and then using a program to see which hashtags are used in various graphs. The degree and community of the recurring hashtags can also be analyzed to give insights into which hashtags are widely popular. This can also be extended to map individual user behavior which also can be recurring. While this is not studied in this paper, this can give a strong insight into user behavior and can be explored in future studies.

Another interesting variable to follow could be the accounts which upload the posts. With enough time, all accounts can be tracked and then see if any accounts follow each or have posted for both events. This can be used to see how an individual reacts to two (or more) different Global Events. An extension to this can also be produced to measure individual impact based on the number of likes and comments of the posts of a user. Furthermore, a large scale version of this study can be followed up with more Global Events such as the FIFA World Cup™ which took place in December 2022.

The geographical information of hashtags and posts can also be retrieved through the Instaloader library. Combined with sentiment tracking, a program can be constructed which can show sentiments combined with the location. This would provide a graph similar to the ones in this paper. The graphs could instead be laid out on a map and then sentiments placed as nodes on the map according to the geographical data obtained. For example, a map could be made restricted to the United Kingdom and then the posts collected from the passing of Queen Elizabeth II plotted along with sentiment analysis. This could give an overview of how different locations might have different reactions inside the United Kingdom.

This example can also extend to business applications. A company might run a similar program with their specific hashtags and analyze captions to track sentiments – primarily into positive, neutral and negative. This could then be used to track their company's reputation and success in specific geographical regions. This method can also prove efficient as it is not power consuming, requires limited network speeds and does not need powerful computers/hardware to run. A similar approach can be used for analyzing reactions to government policies.

While surveys are popular for gathering information on public opinions, they can be time consuming and prove expensive. The method explored in this study is cheap and, as done here, can be done using a work laptop. While social media can be explored by experts to gather detailed information, this method can be further improved to take into account comments as well and a well-trained NLP (Natural Language Processing) model can be used to perform opinion mining – potentially automating this process. Furthermore, a long term study can also be conducted. This study can be used to explore how opinions/sentiments change over time. This study would be limited to monitoring topics which are prone to change over long periods of time and are likely to be discussed in the future. An example can be analyzing views on homosexuality. This topic, which can be controversial, is likely to be discussed in the foreseeable future, especially with the current developments such as the growth of the LGBTQ community. A study could analyze the change in opinions and sentiments over time and monitor them in specific geographical areas as well. These sentiments can be converted to quantitative data and then be plotted across a time axis. Events such as changes in government policies can be marked to see how sentiments change around them. These ideas/studies can be extended for use in the fields of Political Science, Psychology as well as Sociology in order to explore human behavior as well as analyze efficiency of certain government policies.

These studies can be extended to be as detailed as possible with as many variables. However, an increase in variables and data needs to be matched with an increase in computational power and time. Tasks such as rendering graphs using Force Atlas require extensive RAM (Random Access Memory) and a strong processor. Storage of data does not pose a major issue as the text files needed to store the collected data are not of large size. For example, all data collected and programs made for this research require less than 5 megabytes of storage.

Another limitation future research may face is the validity of the research. The methods used to analyze the data collected are subjective. As the amount of data increases, the subjectivity can cause a large decrease in validity. This can be countered with multiple researchers. By increasing the number of different views and opinions, inter-rater validity can be increased.

References

1 *Merriam-Webster Online Dictionary, "Social Media noun Definition"*, https://www.merriam-webster.com/dictionary/social%20media.

2 C. Dictionary, *Social Media*, https://dictionary.cambridge.org/dictionary/english/social-media.

3 O. L. Dictionary, *Social Media"*, https://www.oxfordlearnersdictionaries.com/definition/english/social-media.

4 Wikipedia, *Social Media – Definition and Features"*, https://en.wikipedia.org/wiki/Social_media#Definition_and_features.

5 S. Kemp, *DIGITAL 2020: GLOBAL DIGITAL OVERVIEW"*, https://datareportal.com/reports/digital-2020-global-digital-overview.

6 L. Insider, *Global events definition"*, https://www.lawinsider.com/dictionary/global-events.

7 A. Graf and A. Koch-Kramer, https://pypi.org/project/instaloader/4.9.5/., Instaloader (Version 4.9.5) [library].

8 Wikipedia, *Web scraping"*, https://en.wikipedia.org/wiki/Web_scraping.

9 *Official Instagram Website, "How Instagram Feed Works"*.

10 F. Cheong and C. Cheong, PACIS 2011 Proceedings.

11 P. Liang and B. Dai, 2013 IEEE 14th International Conference on Mobile Data Management.

12 X. Chen, M. Vorvoreanu and K. Madhavan, *IEEE Transactions on Learning Technologies*, **7**, 246–259,.

13 A. Himawan, A. Priadana and A. Murdiyanto, *International Journal on Informatics for Development*, **9**, 59–65.

14 F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, *Proceedings of the National Academy of Sciences*, **101**, 2658–2663.