POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Discussion Topics on Instagram about COVID in Italy: techniques and applications

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Federico Ferri, 10530525

Advisor:
Prof. Marco Brambilla

Academic year:
2021-2022

Abstract: In the current world where social media are prevalent in our lives, they also became an important tool to express our opinions and emotions on current public debates.
Monitoring and understanding such opinions is a very complex task because the communication format is the natural language.
This thesis goes thought the steps of creating an Instagram dataset, comparing the traditional topic mining technique, that uses TF-IDF, with the more modern language model based technique. The results found the language model based technique to be much better overall and it can be considered as a drop-in replacement for more traditional techniques.
Moreover language model topic mining techniques are applied both on the posts and on the comments to analyze the Italian COVID debate that took place on Instagram.
To further improve the results and the demographic information, sentiment and emotion analysis is employed and also techniques to estimate the age and the gender from profile images.
Significant results include the estimation of trending topics in the COVID debate and the identification of relevant demographics such as the organized no-vaxxers.

Key-words: social media, topic mining, instagram, covid, italy

1. Introduction

In today's world dominated by the pervasiveness of technology in all aspects of our lives we have very potent means to communicate with each other such as social media. Such platforms born out of the necessity to interact and keep in touch with friends and relatives have become a tool to express ourselves on any matter even with people we have never met before thus creating a permanent digital record of conversations that were previously relegated to oral forms.
This opens up the possibility to gather and monitor such conversations, to obtain a reasonable picture of the discussion around a topic without using techniques such as surveys which are much more expensive and intrusive [5].
In particular in the social media spectrum there are various different platforms that attract different demographics, the most famous are Facebook, Twitter, TikTok, Instagram, Youtube, Snapchat, Linkedin, Pinterest, etc.. in particular with the focus on the Italian market the most used social media is Facebook followed by

Instagram and Linkedin [35].

To extract value from such social media information it is necessary to understand it, in particular to understand the meaning that is conveyed through messages and interactions.
Traditionally the analysis performed on social media contents have always been orientated to techniques such as sentiment analysis and similar techniques without understanding the meaning and the evolution of the conversation [33].
This thesis aims to develop a novel technique to be used on social media content, in a reliable way, to understand what are the topics of discussion, how they change in time and what opinion the people have about those topics. Such a technique must be able to collect data from social media platforms and it should leverage the state of the art technologies in the Natural Language Processing (NLP) field to extract meaning from such data and finally it must be able to convey to the observer such data through appropriate visualizations.
This tool can be used by researchers to measure the public sentiment with objective metrics thus uncovering the discussions that previously were held offline. Moreover results from a tool like this can be used effectively by policy-makers to better understand the political climate with a population and thus it can be used to adjust agendas to such a climate.

2. Background

The findings in this thesis are enabled by many advancements in different research fields spanning from computer science to sociology. Most importantly the advancements in the big data and NLP fields are the most notable.
The first step into this process is the scraping of the social media platform chosen for the task, but this can be a very short lived success since those platform evolve quite rapidly and requires a constant maintenance.
This can be done programmatically using Application Programming Interfaces (APIs) offered directly by the social media itself [11], such as the Facebook Graph API. This approach has now become quite unusual after the massive Cambridge Analytica scandal in 2018 [6] and nowadays only a few social media offer such an interface for research purposes.
Another less comfortable approach is scraping the content from the social media itself by impersonating a legitimate user [34]. This technique, that can be performed with tools like Selenium, is quite time consuming because every interaction with the website have to be scripted beforehand and it does not allow much flexibility in the exchange with the social media.
To store such a massive amount of data that can be acquired from social media a specialized database that can handle different varieties of data is necessary [31]. In particular it can be mentioned MongoDB that allows to work natively with JSON data which is typically used by APIs in social medias infrastructure.

Analyzing data from social media is making use of various research fields in particular from the filed of NLP since most of the content is in the form of text.
An simple technique to capture the important parts of a document is TF-IDF which computes a frequency score for each term and can be used to highlight the terms that have a spike in importance in a certain corpus [26].
A more complex technique to do this is to use language models, such as BERT, which leverage a lot of recent developments in the NLP field [10]. In particular a BERT model can be used as a tool to create sentence embeddings, a vectorial representation of a text sentence [28]. Such embeddings should be of great quality since language models encompass the nuances of a specific language and can make use of such knowledge to differentiate and better contextualize text corpuses compared to earlier techniques.
Such BERT sentence embedding models are very language specific since every language is different with different grammar and rules. Thus a language specific model is needed to obtain the best results. Fortunately such pre-trained language specific models exists and are available [29].
To cluster results several existing techniques have been considered, k-means clustering which is a very well known and available clustering algorithm is the more traditionally used.
Moreover clustering in high dimensional spaces is quite challenging thus it's suggested to reduce dimensions before attempting to cluster such data. To reduce the dimensionality of embeddings we can use algorithms such as UMAP that specialize in this kind of reduction by preserving both the local and global structure of embeddings [24].
To further improve the clustering on this reduced dimensionality data, a better than k-means clustering algorithm can be used. HDBSCAN allows to cluster vectors while being able to identify and exclude outliers [7].

For computing sentiment and emotions scores, which are classic metrics used when analyzing social medias, the state of the art approach is based, again, on language models because they are more capable to understand the context of words compared to previous dictionary based techniques [4]. Moreover to even better exploit

such advantage the language used to train the model should reflect the language where the model is used on which is the case with FEEL-IT which is based on UmBERTo a Roberta-based Language Model trained on large Italian Corpora [21, 22].

To work with images, that represent a crucial part in modern social media, the analysis relied on using OCR techniques to extract available text inside the image but it also relied on face detection and age/gender recognition. This is necessary due to the high amount of profiles a social media usually have and such profiles are often accompanied by a profile image.
To perform such analysis it is first necessary to detect a face inside an image using a detector, RetinaFace is an architecture to detect and localize a face inside an image while being not very resource intensive [9] and it can be implemented using MobileNets a very resource efficient convolutional neural network [16].
After the face is located inside an image, using a second convolutional neural network, it is classified by age and gender. This step is performed with a very efficient ShuffleNet [39] based network trained on the UTKFace dataset a labelled collection of over 20k faces that was used in a related research on face age progression and regression [38].

Finally for general purpose supervised text classification, FastText is a good fit to have a nice balance between speed and precision [18]. This tool runs on CPU thus making it more accessible to users. Moreover to improve performances to a level comparable to more modern, but complex, techniques such as language models there are pretrained word vectors available in a variety of languages including Italian [13].

3. Related work

The field of social media analysis (SMA) is very crowded with existing research since it's a good source of information for various industries and sectors, in particular businesses related to the service sector are the ones using SMA the most [20].
Moreover the most utilized social media platform for SMA is definitely Twitter since it offers, by far, the easiest access to information and the most common analysis technique is sentiment analysis.

About sentiment analysis commonly the most used techniques are Lexicon based methods and Machine learning methods or a combination of the two. Both of this methods are not state of the art, compared to language models based approaches [4], but work quite well with simple sentences without much sarcasm, slang or negations.
The type of analysis often performed using sentiment analysis is to understand what the people feel about a topic, for example: an election result or a brand preference. This kind of analysis does not have to ability to really understand concepts but only to measure known elements [12].

In the domain of image analysis there is the trend to replicate the analysis, which are common in the text domain, by using images. In particular, using Instagram images, posts can be clustered into topics which is an activity usually done using hashtags. Even the sentiment analysis can be performed using the images instead of the more commonly used hashtags.
This shows that is possible to gather from images a much greater amount of information that was previously not considered because thought to be too hard to access [30].
The topic of using images containing people to extract biometric information is not very well researched outside the world of defence and security.

Commonly for social media topic analysis TF-PDF and k-means clustering have been used to analyze the topics people are talking about. An example of such as technique has been applied on a small dataset of Egyptian political Tweets in Arabic [27].
Indeed it's clear that the techniques to use are very context specific and depend much on the dataset and use case presented [3]. Some common tools can be used to give a level of uniformity to the research and task at hand.

Recently there is more activity on the use of language models to identify topics and MultiModalTopicExplorer is a proposed tool, that has yet to be developed, which is able to identify the topics that people are talking about and how they change over time all into a single tool. There is a preliminary study on the feasibility on a 4chan dataset that is very promising [1].

4. Comparison of topic mining techniques and applications on an Italian Instagram COVID dataset

The main idea is divided into various steps with the objective of creating a pipeline that can be validated and then replicated for current and future uses.

4.1. Context

The focus of this paper is on Italy and the Italian conversations happening online because most of the research that already exists has an English-speaking point of view. Moreover my ability to speak both languages and the technical knowledge allow me to conduct such research.

The first step into performing an analysis on social media is to obtain a substantial dataset. This dataset has to be representative of where the conversation actually happens, in the Italian market that would be Facebook or Instagram [35]. Since Facebook is globally more widely used for researches on social media [33], Instagram seems a more interesting candidate.

Instagram is a 2010 social media platform mainly based around the sharing of pictures. This platform allow users to upload photos or videos that can be modified directly inside the application. Sharing of these contents depends on the user: if a user chooses a private profile his content can only be viewed by his friends, instead if the profile is public the content can be accessed by everyone.
Contents such as posts are often accompanied by a text description, also profiles can have a text biography and a profile picture. Text can also be embedded into images or videos through the app editing tools.
Contents can be liked and commented by the owner and other users. Comments can also recursively be commented as a reply and liked by the owner and other users thus creating discussion threads.
Instagram is owned by Meta, the parent company of Facebook, from whom it shares many features and policies. In particular from a policy point of view an API to programmatically navigate contents is not available. Moreover scraping is not permitted in any form in their user license [19].

The dataset should consist of good variety of posts, those posts can be images or videos together with a text description. Also posts metadata are useful such as the publication date and the amount of activity generated by the post including likes and comments.
The dataset should also include the comments associated with every post, in the form of the text of the comment, the date when it was left and the interactions in the form of likes.
The profiles of the people who comment are also crucial, in particular their nickname, possibly their name, biography and profile image.

This dataset for the topic mining techniques comparison should be as horizontal as possible by spanning a variety of accounts such as news, fashion, sport and in general every popular account where people tend to leave comments and opinions to gather many different topics and opinions. This makes the comparison of topic mining techniques more significant since more topics are present.
On the contrary for the topic of interest analysis a vertical dataset is needed. This dataset should include only posts that are strictly about the COVID topic and it should be very refined to not include any noise. To filter out the irrelevant data from this dataset a combination of techniques can be used to obtain the best results. In particular a series of supervised classifiers can be used to refine the results and only keep the relevant data. To train such classifier manual labelling on a sample dataset is required.

4.2. Topic mining comparison

With such an horizontal dataset in hand it is interesting to perform a comparison between knowledge discovery techniques. Such techniques are used to extract the topics people are talking about. In particular it would be interesting to compare traditional and well tested techniques such as TF-IDF to more advanced and cutting edge technologies such as language models. In the language models landscape since training a language model is a huge task, due to the need of a large corpus in the language of choice and the required specialized hardware, pretrained language models are the way to go. For this task it is required an Italian pretrained language model.

The goal is to determine if for social media datasets, in particular from Instagram, traditional techniques of topic mining are still relevant. Moreover the usage of language models in the topic mining sector is a very recent trend and it would be significant to confirm if a social media dataset is compatible with this kind of

techniques and yields good results. Finally a comparison between the two would be based on the amount of topics identified, their quality, the management of outliers and also the computational complexity measured by the need for specialized hardware and the time needed to perform the computation.

4.3. Topic of interest analysis

The final goal for this research is to identify from the previous phase a winning strategy for topic mining and apply it to a topic of interest to identify what are the people views on it and their evolution.

In particular an interesting topic to analyze would be COVID, since 2020 this pandemic has changed our lives and has been at the center of our attention.
The history of the pandemic and its progression is a very publicly well known argument and it would be interesting to find parallelism in the digital world.
Moreover analyzing what the people think about COVID and the consequences, that touched all of us, is a very interesting result since it would allow to quantify opinions.

This analysis should be performed on the vertical dataset which is a refined single topic collection of posts, comments and users on the COVID topic.
Such an analysis should include comments that would allow to view the opinion of users even if the number of comments is usually much greater than the number of posts.
Users that leave a comment are also a significant source of information, by using some analysis techniques it can be determined their age, gender and possibly some other information such as interests or passions. This should allow to have an even clearer picture on what demographics is associated to which opinion.
Finally the comments in the vertical dataset can be treated as if they were posts and thus clustered using the same techniques. This would allow to find common thoughts between comments and quantify the strength of such opinions. This would be like the equivalent of conducting a survey to gather people opinion on topics but on a massive scale.

5. Implementation

The implementation of the final solution was completed in steps, the following sections represent the macro areas of development.

5.1. Data acquisition

Obtaining the required data from Instagram is not easy. While some libraries to interact with Instagram already exist [14] they all present the limitation of accessing only login free contents such as the last published posts and the first few comments for each post.

Moreover Instagram employs anti-scraping techniques such as rate limiting both in the number of requests per time period and in general on the maximum number of contents that can be obtained.
IP address based rate limiting techniques are also applied to further limit the scraper ability.

Having a pool of Instagram accounts to scrape from is mandatory since the limits on a logged out user are very strict for the type of scraping required. To create new accounts on Instagram it is required a phone number and there is a limit on the number of accounts per phone number. However from the mobile application it appears to be possible to circumnavigate this limitation and create many different accounts from the same mobile number if the creation of the new account is done while already logged in with a verified account.

The scraper technology chosen is Selenium [17], a browser automation framework. This framework allows to control a web browser of choice by exposing the received data through an interface accessible from the controlling algorithm and by exposing another interface to direct the browser to perform actions.
Out of the box Selenium was not enough to scrape Instagram, some precautions had to be taken against Instagram discovering that an automation framework was used or they would immediately ban the account used to scrape.
Giveaways that an automation framework is being used are as easy as matching a specific User Agent, but can also be quite complex such as detecting if a webdriver is present. The decision was then to use undetected_chromedriver a specifically patched Selenium webdriver that already implements all of the required

mitigations [36].

Furthermore an algorithm had to be created to automate the login into an existing account, reach the Instagram homepage and then navigate to a selected page and scrolling it to the bottom cyclically to collect the existing posts in a reverse chronological order, with some delays to simulate a human behaviour.
The posts from the account are collected from the network logs of the browser to preserve the original JSON responses from the Instagram API which are more machine friendly.

A second algorithm to scrape comments is very similar to the previous one but it lands on a single post page and starts scrolling and expanding comments threads to load all of the comments asynchronously. The same network logs capture is used to retrieve the API responses containing the comments in JSON format.
Those comments also include the nickname, name and thumbnail of the profile picture of the person leaving the comment.

Lastly a similar algorithm to retrieve the user profiles of the people that leave comments opens their profile page and retrieves their name, full picture and biography if present. This works even for private profiles since this information is public anyway.

All of the scraped data is saved as-is into a local MongoDB instance that is able to accept and work with the JSON format [31]. Images are also saved into MongoDB using GridFS [25] a virtual file system running on top of the database.

5.2. Volumes analysis

The tools to analyze such data are written using the Python language, pymongo is used to interface with MongoDB, pandas to create refined and specific datasets from the ones available and matplotlib to visualize such datasets.

The quantitative analysis consists mostly on displaying the amount of posts and/or comments through time to give an idea of how many there are and when they are concentrated.

5.3. Traditional topic mining

To implement the traditional topic mining pipeline the sklearn Python library offers a TF-IDF vectorizer ready to use. Moreover to filter out useless tokens from text, nltk offer a very rich list of Italian stopwords to be directly integrated into the vectorizer for removal.

The text used for topic extraction is the one from posts descriptions and posts images obtained from the accessibility captions already provided by Instagram using OCR scans.

Before clustering using k-means an appropriate value for k has to be found. Using the elbow method [23] it's possible to determine such value to be used as an indication of the number of clusters.
Then k-means is applied and clusters are available for further analysis such as a quantitative analysis to compare clusters variations during time and word clouds representations of the most common tokens to identify the cluster topic.
Word clouds are available from the worcloud Python library that is capable of identifying and displaying the most common tokens.

5.4. Language models topic mining

To make use of language models for clustering there is an amazing library that makes use of pretrained BERT models in various languages, including Italian. Moreover it uses UMAP for dimensions reduction, HDBSCAN for clustering, a specialized version of TF-IDF to create topic representations and even MMR [8] to adjust topic representations. This library is BERTopic [15] and it's available for Python.

To make use of this library is necessary to use a machine fitted with a modern Nvidia GPU since it uses PyTorch. Free offerings such as Google Colab also work since they offer this kind of hardware.

By using the BERTopic library some analysis such as quantitative over time and the token importance per

topic are available inside the library itself.

The topics extracted using language models can then be compared to the traditional topics to evaluate the clustering algorithms performance in terms of the quality of such topics.

5.5. Demographic analysis

To analyze more in depth the people who leave a comment a few techniques can be used to obtain such result. In particular regular expressions (regex) can be employed on biographies to match known patters to extract the age or the birthday.
The regex to search for a birthday is combined with another filter that excludes ages below 13 which is the minimum legal age for Instagram.

```
\d{1,2}[/ -\.~]\d{1,2}[/ -\.~]\d{2,4}
```

Another regex is to search for age directly, also this one filtered with the minimum legal age filter.

```
(?:(?:[^\d]|^)(\d{2})[^\d]{0,4}(?:anni|age|y| 🎂 )|(?:anni|age| 🎂 )[^\d]{0,4}(\d{2})(?:[^\d]|$))
```

Finally also the profile image can be leveraged to extract the age and even the gender, using a combination of RetinaFace to detect the face inside the image and a neural network trained on the UTKFace dataset to guess the age and the gender. Luckily there is already a Python library capable of doing this in a single step, FaceLib [2].

5.6. COVID topic analysis

To conclude the analysis posts about COVID must be isolated. This is not a trivial task since there are posts where the COVID topic is present but it's not the core topic: e.g. "The tourism habits shift in Florence during the pandemic" is a post about tourism but with the context of COVID.
To perform this task a two steps system is necessary, first the posts that are not even COVID related are excluded by simply matching a list of COVID related tokens and by keeping only the posts that match at least one token. Such stemmed tokens are:
covid, *corona virus*, *coronavirus*, *pandemi*, *quaranten*, *mascherin*, *vaccin*, *restrizion*, *isolamento*, *pfizer*, *zeneca*, *johnson*, *zona rossa*, *zona arancio*, *zona gialla*, *zona bianca*, *green pass*, *pandemi*, *terapie intensive*, *prima dose*, *prime dosi*, *seconda dose*, *seconde dosi*, *terza dose*, *terze dosi*, *booster*, *no vax*, *lockdown*, *delta*, *omicron*, *variant*, *guarigion*, *iorestoacasa*, *stateacasa*

The second step is based on a binary supervised classifier trained with a dataset composed of around 500 manually labelled posts on FastText that yielded a 92.5% test accuracy in correctly classifying if a post is strictly COVID related.

Lastly posts can be clusterized and the extracted topics further analyzed by finding common elements and quantitative analysis over time to find how the conversation evolves.
The analysis on the comments is done for each of the topics to work only with comments related to a specific subject. The analysis uses sentiment and emotion detectors to measure how they change over time to gauge what people feel about specific topics.
Moreover also comments can be clusterized, this will be indeed more challenging from a speed and technical point of view given the amount of comments but those topics obtained from comments are representative of people's opinion on a topic. Sentiment and emotion classifiers can be again applied to further characterize an opinion on a topic.
Finally the age and gender information of the people who leave comments can be used to determine if certain opinions are more in line with a specific demographics.

6. Experiment and results

In this section the main challenges and results are presented with a focus on the most interesting findings.

6.1. Experimental settings

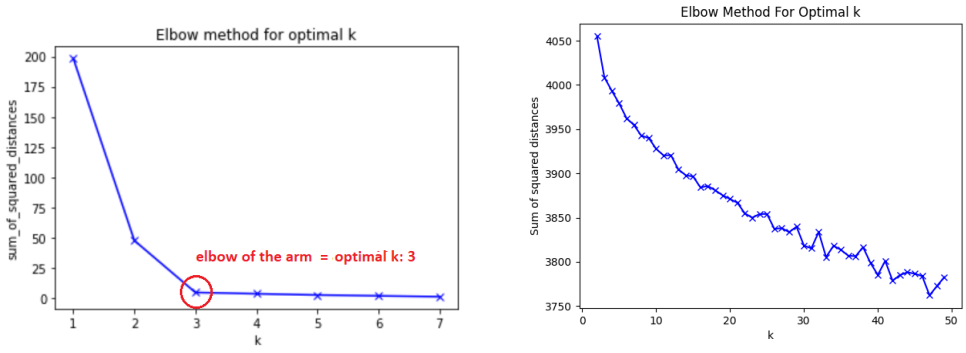
The first dataset is the one containing posts, comments and accounts from 9 main sources. Those sources were identified starting from Instagram suggestions tailored for Italy. They include *corriere*, *will_ita*, *pastorizian-everdiesreal*, *chiaraferragni*, *gliautogol*, *fedez*, *il_post*, *larepubblica* and *commenti_memorabili* which are all well known Italian Instagram pages with a considerable amount of followers.
This dataset contains around 4100 posts, 660k comments and spans from April 2021 to October 2021.
This dataset is more optimized for the comparison of the topic mining techniques and will be used for this purpose.

The other dataset is created from a single source, *corriere*, since it offers the most comments on the COVID topic and is composed only of posts whose main topic is COVID.
The relevant posts are 1400 and the relevant comments are 168k from a period spanning from May 2020 to February 2022.
This second dataset is more fitted for the single topic, in depth, analysis.

6.2. Issues

While scraping posts on Instagram it appears that there is an hard limit that arises after obtaining around 15k posts from the platform. After this any account gets blocked due to rate limiting. This makes it very hard to reach in the past since posts are in a descending chronological order.
Instead for comments scraping it averages 1200 comments per scraping account before getting blocked by Instagram. To simplify the scraping process the choice was not to expand the scraping to the threaded comments since it would be harder to classify them because they are not referred directly to the post but to another comment.
Moreover for the account scraping (referred to the people who left a comment) the limits are definitely more strict averaging only 120 accounts per scraping account before reaching an hard ban from Instagram. This makes scraping accounts unsustainable so the choice was to only rely on the thumbnail of the profile picture which is available inside the comment to find the age and the gender of the person who left the comment.
Of the 2789 accounts completely scraped for 904 (32%) of them at least one method of extraction worked. For 775 (28%) of them the FaceLib library gave a result, using regex only 231 (8%) had a positive result. So by using only the profile image, instead of the harder to get biography, the demographic data is still retrievable for most of the cases (85%).

While using the elbow method to find the optimal k value for the clustering process the resulting chart was quite bizarre compared to a standard expected result (1).



(a) Expected result. (b) Real result.

Figure 1: Elbow method differences.

In a different paper a similar result appeared and the authors settled on picking a k value before the line became wobbly [37]. In this case a correct k value to pick would be around 18.

6.3. Topic mining comparison

The clustering process on the first dataset was very successful for both the traditional and the language model approach. Both yielded high quality topics that are well defined and separated from each others.

Topics extracted using the traditional pipeline are denoted as *TT1...TT18*. The majority of the clusters found (*TT4, TT5, TT7, TT8, TT14, TT16, TT17, TT18*) coincides with the editorial plans of the pages scanned. Those are posts all similar to each other by having at least some elements in common such as the name of the journalistic column in them. This makes it probably quite easy for the clustering algorithm to group them together.
Moreover for the gender equality topic (*TT1*) most of the posts are from *corriere* that has a dedicated editorial plan called *27esimaora* on the topic, which probably helped the classifier too.
The fashion topic (TT2) only contains *chiaraferragni* posts which uses the same hashtags in the text multiple times (such as *#adv, #supplied*), again this most likely helped the clustering process.
For social issues (*TT3, TT10*), COVID (*TT11*) clusters they are a bit mixed up, they contain too many different conversations and were probably put together by keywords.
Finally the football cluster (*TT15*) mostly contains content from *gliautogoal* page which also made it easier for the clustering process.
All the other clusters were indeed quite correctly clusterized even if it does not appear that there is a deep understanding of the issues discussed but only a superficial understanding of the keywords.

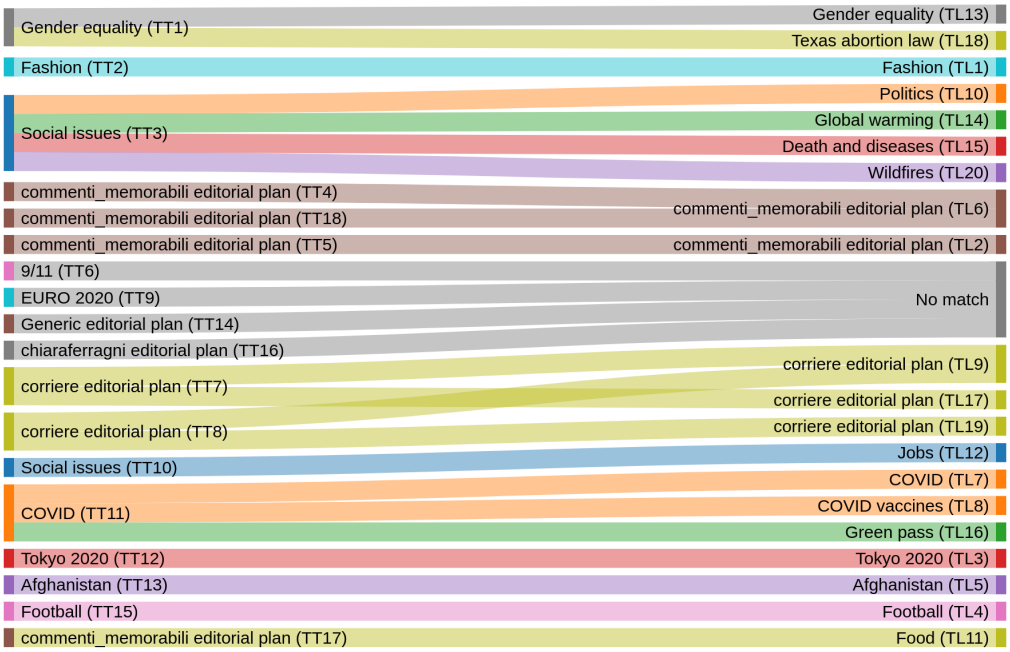


Figure 2: Topic mining and clustering techniques comparison. On the left using traditional techniques, on the right using language models.

Topics from the language model method are denoted as *TL1...TL20*. Similarly to the previous clustering method there are the editorial plan clusters (*TL2, TL6, TL9, TL17, TL19*). We can observe that there are fewer and smaller editorial plan clusters this time which means that some posts were better classified for their topic instead that for the journalistic column that they are part of.
A great improvement can be seen on topics such as gender equality (*TL13*) and the Texas abortion law (*TL18*) which were previously put together and are now divided. Moreover also a greater variety of sources can be seen which denotes a better understanding of the topic by the model. It is peculiar to see how the model was able to understand the what (abortion law) and the where (Texas) and put it together.
A similar improvement can be seen in the topics politics (*TL10*), global warming (*TL14*), death and diseases (*TL15*) and wildfires (*TL20*). Such topics were previously grouped together but the language model was able to differentiate them which is a demonstration of deeper topic understanding. Moreover in the wildfires topic (*TL20*) the ability to group together different locations of wildfires (Sicily, Sardinia, Calabria, Greece) denotes the ability to understand that the topic can be associated with geographical locations.
For the COVID (*TL7*), COVID vaccines (*TL8*) and green pass (*TL16*) clusters the same as in previous in-

stances can be noted, a better degree of specialization of the topics given by the ability of the language model to understand links between concepts.
Even for topics such as fashion (*TL1*), jobs (*TL12*) and food (*TL11*) a better variety of sources can be observed and a better and more clear understanding of the topics too.
Finally in the language model topic mining we cannot reference some of the topics observed in the traditional topic mining results such as 9/11 (*TT6*) or EURO 2020 (*TT9*). This is because of the volumes, they are a very small volumes compared to the other topics so they are not in the top spots of the language model extracted topics.

This comparison revealed the superiority (2) of language models topic mining and clustering techniques which is given by the ability of language models to better understand topics. This new language models technique could be considered as a drop-in replacement for the more traditional techniques since no major results degradation was observed.

6.4. COVID debate analysis

Using the COVID specific dataset and by applying the language model clustering we obtain 19 distinct topics. Of those clusters some of them contain similar topics to each other. By applying a visualization which is highly inspired by LDAvis [32], a great visualization technique typically reserved for LDA we obtain that there are 4 macro-clusters that include all of the extracted clusters (3).
The similarity of clusters within the macro-clusters is expected since it is already a very refined dataset on a single topic (COVID) and it's also a desired behaviour since this helps in creating a simplified view for the analysis.

Those macro-clusters are about vaccines, provisions, disease and politics.

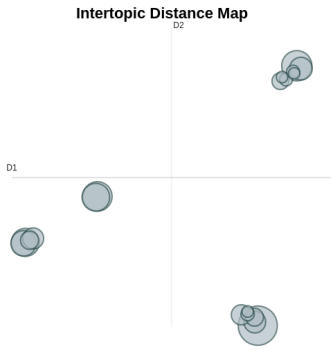


Figure 3: There are 4 macro-clusters in the COVID specific dataset.

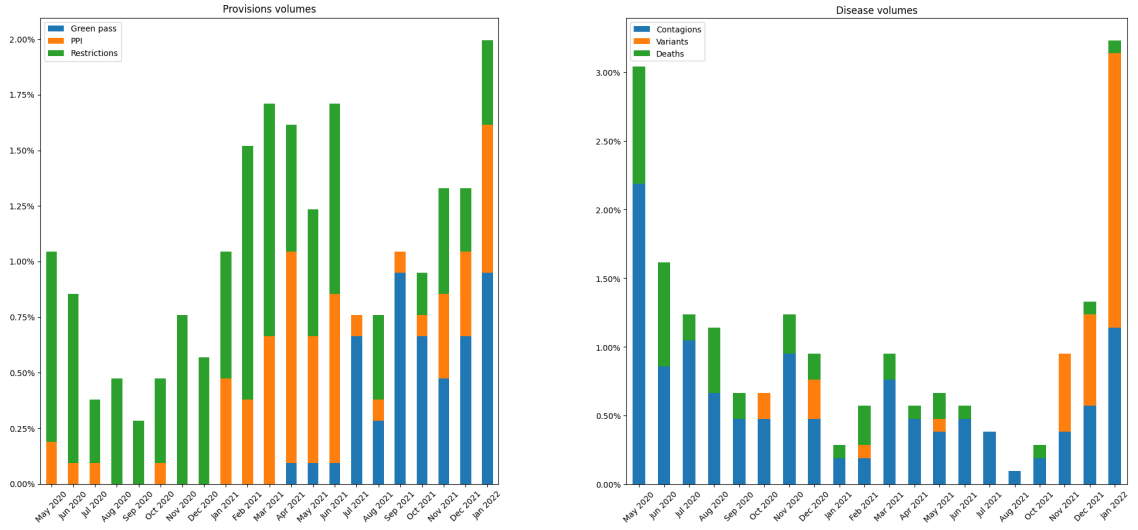
Starting from the **vaccines** mega-cluster, this group is composed by the following clusters:
- *Vaccination campaign*: includes all posts about the progresses in such campaign.
- *Restrictions with respect to vaccines*: all of the restrictions that will impact people who are not vaccinated.
- *Testimonies with respect to vaccines*: interviews with celebrities or repented no-vaxxers about the positives of vaccines.
- *Vaccine*: about efficacy, type of vaccines and new developments.
- *No-vax*: all of the posts regarding the organized opposition to vaccines.

From this mega-cluster it is possible to observe the March-April 2021 debate about vaccine priority or the opposition to the vaccination campaign always present but stronger on September, October and November 2021.

The second mega-cluster is about **provisions** and is composed of the following clusters:
- *Restrictions*: such as limitations to free movement, lockdowns, self-certification modules to leave home and red, orange, yellow and white areas rules.
- *PPE*: everything about personal protective equipments such as masks (surgical and FPP2) or gloves.
- *Green pass*: all the posts about the European digital COVID certificate.

An interesting observation that reflects real events can be made about PPE that were very marginal in the first COVID wave of 2020 but present in the second wave at the beginning of 2021 and at the end of 2021. This is reflected by politics considering PPE useful only for medical staff during 2020 and then switching to a more broad use of PPE (4a).

Moreover about the green pass we can observe it from mid 2021 but it really took off in the summer of 2021 where it was announced as mandatory for the cold season. We can also observe in the conversation how the conversation shifted from traditional lockdowns and different color areas to the green pass.



(a) The shift in the provisions conversation.

(b) The deaths discussion it's reducing over time.

Figure 4: Quantitative analysis on interesting mega-clusters.

The third mega-cluster is about the COVID **disease** and is composed of the following clusters:

- *Contagions*: mostly the posts about the daily contagions count.
- *Deaths*: the posts about the COVID deaths.
- *Variants*: the COVID variants such as the delta or omicron variants.

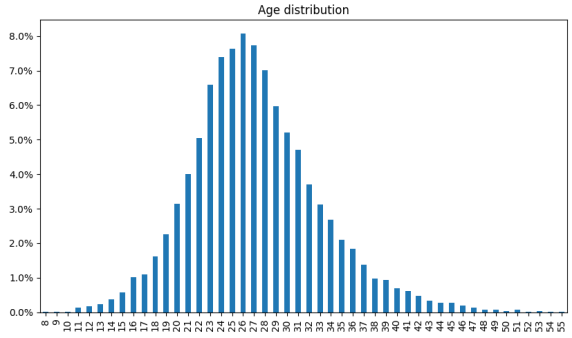
This mega-cluster presents an abnormal trend because it is decreasing for all of 2020 and 2021 while it has a huge peak at the end of 2021 while the omicron variant was very widespread. In the beginning we can observe that deaths posts are significant compared to contagions posts. This relationship trends downwards till mid 2021 where deaths posts are mostly gone. In the end of 2021 where contagion posts really skyrocket we don't see the same happening for deaths which reflects real world observations (4b).

Finally the last mega-cluster is about **politics** and is composed of the followings:

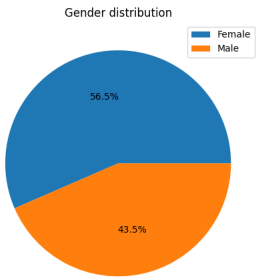
- *Foreign politics*: mostly includes speeches from UK or Germany leaders.
- *Italian politics*: speeches from Italian political personalities.
- *Festivities and events*: the discussion around the handling of festivities and big events.
- *Economy*: around the economical crisis created by COVID restrictions and the recovery fund allocation.
- *Education*: school closures and everything around the education environment.

For topics such as festivities we can observe that they only emerge during the summer period and the Christmas season.
About the school topic it emerges around September 2020 for the back to school period for just disappearing till March 2021 as it was in the real world because of the second COVID wave. The school discussion goes on until September 2021 when in the real world it was clear that the school year was safe. The school topic jumped back out in 2022 with the omicron variant that forced many classes in quarantine.
The economic discussion can be observed when there are no COVID contagions peaks so mostly during the summer of 2020 and the spring of 2021. There another peak in 2022 again due to omicron that impacted many workers due to quarantines.
In the political side of the discussion the Italian politics gets more attention except around April 2021 where the UK choices on COVID were quite controversial.

The analysis of the comments started with a primary assessment of the dataset to identify the general demographic (5) and the sentiments and emotions.



(a) Average age is 27.1 years old.

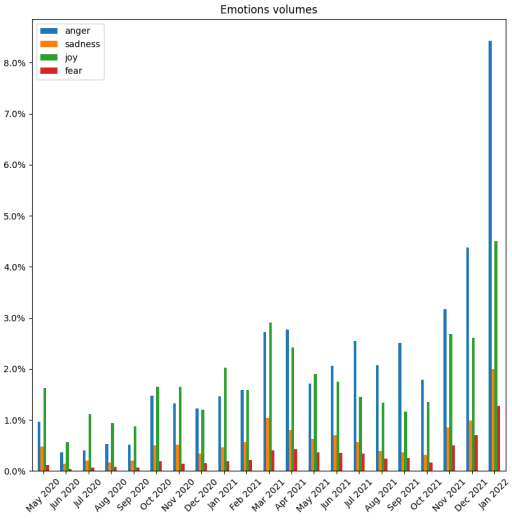


(b) The predominant gender is female.

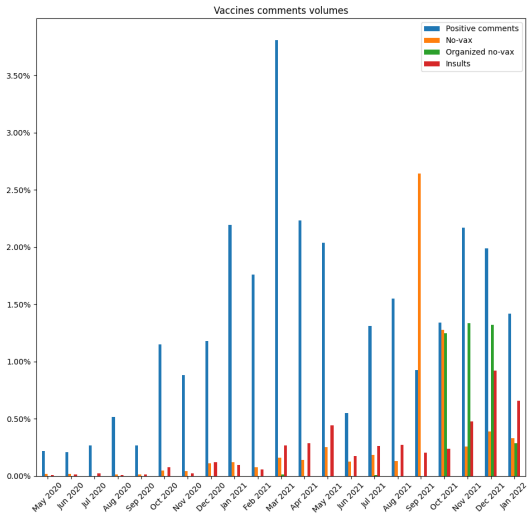
Figure 5: Demographic analysis on the dataset comments.

This preliminary analysis shows that there is a mounting sense of anger and negative sentiment (6a). This is mainly from the vaccines mega-cluster but can also be seen in the provisions mega-cluster. Completely opposite to the previous mega-clusters is the disease one, here a decrease in anger and fear can be observed thought the whole period except for the 2022 omicron peak.

For each of the mega-clusters, to analyze the comments, the choice was to use the language topic mining again but this time on the comments. The results helps to understand what is the topic of the conversation that people are having.



(a) The general trend is a growing sense of anger.



(b) The comments types are changing in phases.

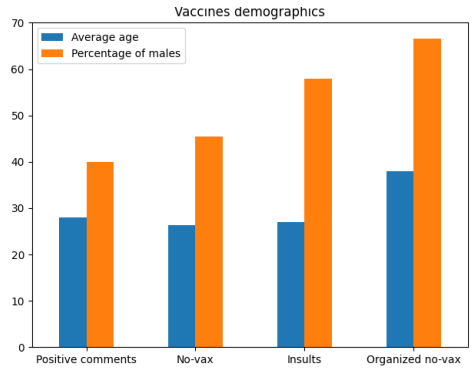
Figure 6: Quantitative and emotion analysis on the comments.

For the **vaccines** mega-cluster the comments are mostly positive around March 2021 which was around the start of the vaccination campaign then the insults and no-vax comments took over with a peak around September 2021. Afterwards no-vax comments decrease to give space to the organized no-vax comments which are mostly copy paste comments all similar to each others (6b).
About the demographics of the comments it shows that organized no-vaxxers have a significantly older demographic at around 40 years old and are predominantly males (7a).

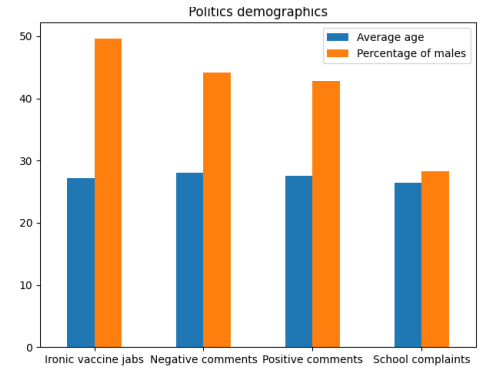
Instead in the **provisions** mega-cluster both the comments that supports the dispositions and the ones that dislike them are growing albeit the second category faster. We can observe peaks that are roughly around the time that people got their vaccine jab.
The demographics is more balanced but as previously observed males are more likely to be in the more controversial conversations.

In the **disease** mega-cluster we can observe that for the whole of 2020 and 2021 positive and supportive

comments are the majority but at the end of 2021 there is a reversal in the trend with the growth of negative comments.
The demographics is yet again more balanced but as always the more negative comments are mostly males.



(a) Organized no-vax are older and males compared to the average.



(b) The school complaints are gender imbalanced.

Figure 7: Demographic analysis on the comments of vaccines and politics mega-clusters.

Finally for the **politics** mega-cluster the comments are mostly positive and supportive with a small negative peak in the end of 2021. There are conversations about vaccine doses around the periods where most people got their vaccine shot and those conversations are mostly ironic about the necessity to always have a new dose. Finally about the school complaints there is a huge spike in March of 2021 against the closures of the schools. The demographics is balanced except for the school complaints topic where the gender is predominately female (7b).

7. Conclusions

The comparison of the topic mining methods highlighted the superiority of the language models methods compared to traditional techniques.
Moreover using a combination of topic mining on the posts and the comments and sentiment analysis it was possible to understand what the people are talking about and how the conversation evolves. Finally with the help of profile images it was also possible to define which people are associated with a certain opinion which gives a face to the discussion.

The creation of this pipeline was quite complex but it is a very scalable effort, once it's completed it can be used on multiple datasets and sources.
Instead the collection and creation of the dataset is the least scalable part of the entire pipeline, it is always a cat and mouse game where the social media try to defend their data with ever changing techniques which make this process very susceptible to external changes and not very scalable.

A great evolution of this pipeline would be a real time one which would make it even more useful to monitor and react to people sentiment over the main public discussion topics.

Unfortunately at the same time and given the current political circumstances a tool like this one would make it very easy to find and identify people based on their opinion on any topic.

References

[1] Seyed Hossein Alavi and Felipe Gonzalez-Pizarro. Multimodaltopicexplorer: Topic modeling for exploring multi-modal data from asynchronous online conversations. 2022.

[2] Sajjad Ayoubi. Facelib. `https://github.com/sajjjadayobi/FaceLib`, 2020.

[3] Loris Belcastro, Riccardo Cantini, and Fabrizio Marozzo. Knowledge discovery from large amounts of social media data. *Applied Sciences*, 12(3):1209, January 2022.

[4] Federico Bianchi, Debora Nozza, and Dirk Hovy. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021.

[5] Cody L. Buntain, Erin McGrath, Jennifer Golbeck, and Gary LaFree. Comparing social media and traditional surveys around the boston marathon bombing. In *#Microposts*, 2016.

[6] Marcus Burkhardt, Anne Helmond, Tatjana Seitz, and Fernando Van der Vlist. THE EVOLUTION OF FACEBOOK's GRAPH API. *AoIR Selected Papers of Internet Research*, October 2020.

[7] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg, 2013.

[8] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[11] Lusiana Citra Dewi, Meiliana, and Alvin Chandra. Social media web scraping using social media developers API and regex. *Procedia Computer Science*, 157:444–449, 2019.

[12] Zulfadzli Drus and Haliyana Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714, 2019.

[13] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[14] Chris Greening. Instascrape. `https://github.com/chris-greening/instascrape`, 2020.

[15] Maarten Grootendorst. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020.

[16] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[17] Jason Huggins. Selenium. `https://www.selenium.dev`.

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[19] Meta Platforms Ireland Limited. Terms of use. `https://help.instagram.com/581066165581870/?helpref=uf_share`.

[20] Shadrack Stephen Madila, Mussa Ally Dida, and Shubi Kaijage. A review of usage and applications of social media analytics. *Journal of Information Systems Engineering and Management*, 6(3):em0141, May 2021.

[21] Bernardo Magnini, Amedeo Cappelli, Emanuele Pianta, Manuela Speranza, V Bartalesi Lenzi, Rachele Sprugnoli, Lorenza Romano, Christian Girardi, and Matteo Negri. Annotazione di contenuti concettuali in un corpus italiano: I - cab. In *Proc.of SILFI 2006*, 2006.

[22] Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. I - cab: the italian content annotation bank. In *LREC*, pages 963–968. Citeseer, 2006.

[23] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018.

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[25] MongoDB. Gridfs. `https://docs.mongodb.com/manual/core/gridfs`.

[26] Shahzad Qaiser and Ramsha Ali. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 07 2018.

[27] Ahmed Rafea and Nada A. Mostafa. Topic extraction in social media. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 94–98, 2013.

[28] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[29] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[30] Richard Rogers. Visual media analysis for instagram and other online platforms. *Big Data & Society*, 8(1):205395172110223, January 2021.

[31] Nicolas Ruflin, Helmar Burkhart, and Sven Rizzotti. Social-data storage-systems. In *Databases and Social Networks on - DBSocial '11*. ACM Press, 2011.

[32] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. 06 2014.

[33] Sheela Singh, Priyanka Arya, Alpna Patel, and Arvind Kumar Tiwari. Social media analysis through big data analytics: A survey. *SSRN Electronic Journal*, 2019.

[34] Vidhi Singrodia, Anirban Mitra, and Subrata Paul. A review on web scrapping and its applications. In *2019 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6, 2019.

[35] Statista. Number of users of leading social networks in italy in march 2021. `https://www.statista.com/statistics/787390/main-social-networks-users-italy/`, 2021.

[36] ultrafunkamsterdam. undetected_chromedriver. `https://github.com/ultrafunkamsterdam/undetected-chromedriver`, 2020.

[37] Aleksandra Urman, Justin Chun ting Ho, and Stefan Katz. Analyzing protest mobilization on telegram: The case of 2019 anti-extradition bill movement in hong kong. *PLOS ONE*, 16(10):e0256675, October 2021.

[38] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[39] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices, 2017.

Abstract in lingua italiana

Nel mondo attuale in cui i social media sono prevalenti nelle nostre vite essi sono anche diventati uno strumento importante per esprimere le nostre opinioni ed emozioni sui temi di dibattito pubblico.
Il monitoraggio e la comprensione di tali opinioni è un compito molto complesso perché il formato di comunicazione è il linguaggio naturale.
Questa tesi ripercorre i passaggi della creazione di un dataset basato sui contenuti di Instagram, confrontando la tradizionale tecnica di topic mining che utilizza TF-IDF con la più moderna tecnica basata sui modelli linguistici. I risultati hanno riscontrato che la tecnica basata sui modelli linguistici è molto più performante e può essere considerata un sostituto delle tecniche più tradizionali.
Successivamente queste tecniche di topic mining basate sui modelli linguistici vengono applicate sia sui post che sui commenti per analizzare il dibattito, in italiano, sul COVID che si è svolto su Instagram.
Per migliorare ulteriormente i risultati e le informazioni demografiche, vengono utilizzate tecniche di analisi del sentimento e delle emozioni e anche tecniche per stimare l'età e il genere partendo dalle immagini di profilo.
Risultati più significativi includono la stima degli argomenti di tendenza nel dibattito sul COVID e l'identificazione di dati demografici rilevanti su gruppi quali quello dei no-vax organizzati.

Parole chiave: social media, topic mining, instagram, covid, italia