Titolo Progetto

Christian Braga

This version: March 15, 2025 (First version: February 16, 2023)

QUI SCRIVI ABSTRACT PROGETTO IN BREVE PERÓ PUÓ ESSERE CONSIDERATO COME UN PICCOLO RIASSUNTO DI MEZZA PAGINA (NON DOVREBBE ENTRARE NEL COUNT)

Table of contents

1 Introduction

DA RIVEDERE / RISCRIVERE DEVO AGGIUNGERE TESTO MIGLIORE PER QUANTO RIGUARDA I PRECEDENTI LAVORI CHE SONO STATI FATTI. da strurrure come primo paper caseiro molto interessante. The Israeli-Palestinian conflict, an enduring and complex territorial dispute, has shaped the Middle East for over a century. Originating from conflicting national aspirations and colonial legacies, this conflict involves deep-seated historical, religious, and social issues. The disagreement centers around the claims to land considered sacred by both Israelis and Palestinians, compounded by the broader implications of international policies and interventions. Significant events such as the Balfour Declaration of 1917 and the ensuing British Mandate period catalyzed the conflict, setting the stage for decades of strife, including several wars and uprisings like the intifadas.

In today's digital age, the battlegrounds of this conflict extend into the virtual realms of social media, where public sentiment and information flow freely and swiftly across global networks. Platforms like YouTube have become critical in disseminating perspectives and shaping discourse regarding this conflict. The comments and discussions on such platforms offer a rich dataset for analyzing public sentiment and the prevailing themes of discussion concerning the Israeli-Palestinian conflict.

The importance of analyzing public sentiment through social media lies in its ability to provide real-time insights into the perspectives of diverse global audiences. This analysis not only helps in understanding how the conflict is viewed worldwide but also aids institutions in grasping the nuances of public opinion to better inform policy and communication strategies. By studying comments on prominent news channels like Al Jazeera English, CNN, and Sky News, this research aims to uncover the prevailing sentiments and topics of discussion that emerge from different viewer demographics, providing a contemporary snapshot of public opinion surrounding the Israeli-Palestinian conflict.

The motivation behind this study is to bridge the informational gap between public perception and the often complex realities of geopolitical conflicts. By leveraging advanced analytical tools such as BERT for sentiment analysis and BERTopic for topic modeling, this paper seeks to offer a systematic analysis of public discourse, contributing valuable insights into the dynamics of information dissemination and reception in the context of a highly mediated international conflict. These new NLP models provide a significant advancement over traditional machine learning methods like logistic regression and Latent Dirichlet Allocation (LDA). The enhanced capabilities of models like BERT allow for a more nuanced understanding of text data, enabling more accurate sentiment analysis and more effectively identified themes and topics from vast amounts of unstructured text data.

This study aims to answer several pivotal questions about the dynamics of public sentiment and topic discussion concerning the Israeli-Palestinian conflict as manifested on YouTube. These questions are designed to dissect various aspects of social media discourse, providing a multi-dimensional analysis of how digital platforms contribute to shaping public perceptions. The research questions are as follows:

Prevailing Sentiment and Primary Topics of Discussion: What is the prevailing sentiment among YouTube users regarding the Israeli-Palestinian conflict, and what are the primary topics of discussion for each side? This question seeks to identify the dominant emotions—whether positive, negative, or neutral—and the thematic concerns expressed by users in the comments sections of selected YouTube videos.

Evolution of Sentiment and Discussion: Has the sentiment and the topics of discussion evolved from the initial stages of the conflict to recent times? By examining comments from different periods, this study aims to trace any shifts in public opinion and the changing contours of the discourse over time.

Consistency Across News Channels: Is the sentiment expressed by YouTube users consistent across different news channels? Additionally, do users from different geographical or demographic backgrounds exhibit specific preferences for the news channels they use to gather information? This question addresses whether the portrayal of the conflict varies by news outlet and how that might influence viewer perceptions and preferences.

Perceptions of News Channel Bias: Is it possible to determine any relevant information about different news channels based on user sentiment and preferences, such as whether users perceive that an information network implicitly supports one faction over another? This involves analyzing comments for indications of perceived biases and evaluating how these perceptions might affect trust and credibility attributed to each channel.

By addressing these questions, the research aims to provide a nuanced understanding of how digital public spheres influence and reflect contemporary views on longstanding geopolitical issues. This inquiry not only enhances our understanding of the conflict's portrayal in media but also offers insights into the broader implications of digital media consumption on public opinion and international relations.

2 Methodology

2.1 Transformers models

The advent of transformer models in natural language processing (NLP) has marked a revolutionary shift in how text analysis is conducted, setting a new standard beyond the capabilities of traditional machine learning methods like Logistic Regression. Unlike earlier approaches that often relied on rigid, shallow learning mechanisms, transformers introduce a deep, context-aware architecture fundamentally enhancing the understanding and generation of text. Transformers, first introduced in the paper "Attention is All You Need" by Vaswani et al. (2017), utilize a mechanism known as self-attention. This method allows the model to weigh the significance of different words in a sentence, irrespective of their positional distance from each other. Thus, unlike previous sequence-based models that processed data linearly (e.g., RNNs and LSTMs), transformers can interpret sentences in a non-linear fashion, capturing nuanced meanings more effectively.

Revolutionizing Text Analysis: The transformer architecture has been particularly groundbreaking for its ability to handle large-scale language modeling tasks. This capability stems from two key aspects: Pre-training and Fine-tuning Paradigm: Transformers are typically pre-trained on vast amounts of text in an unsupervised manner, learning a general understanding of language syntax and semantics. This pre-training phase involves tasks like predicting the next word in a sentence or filling in missing words. Once pre-trained, these models can be fine-tuned with smaller, task-specific datasets, which is highly efficient for adapting the model to specific NLP tasks such as sentiment analysis, question answering, or text summarization. Text Classification Capabilities: For text classification, transformers represent a significant advancement. By understanding context across longer stretches of text and capturing subtler nuances in language, they provide superior performance in distinguishing sentiments, categorizing text, and identifying thematic elements. This deep contextual awareness allows for more accurate and nuanced classifications than was possible with traditional models, which often misinterpret the context or the polysemy of words.

In conclusion, transformer models have not only enhanced the efficiency and accuracy of text analysis tasks but have also broadened the scope of what can be achieved in NLP. Their deep learning approach, combined with the pre-training and fine-tuning strategy, allows them to excel particularly in text classification, making them the preferred choice for cutting-edge NLP applications.

2.2 Sentiment Analysis: Bert

To perform the sentiment analysis in this project, we will use the BERT model. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a groundbreaking model in NLP introduced by researchers at Google in 2018. It represents a significant advancement in the field due to its deep learning framework that fundamentally processes words in relation to all the other words in a sentence, rather than in isolation.

Characteristics and Strengths of BERT:

Bidirectional Context: BERT's major innovation is its bidirectional training of the transformer, an approach that allows the model to learn the context of a word based on all of its surroundings (left and right of the word). This differs from previous models that typically read the text input sequentially (left-to-right or right-to-left). The bidirectional nature of BERT allows it to capture a more nuanced understanding of language context and flow than single-direction models.

Fine-Tuning Adaptability: Once pre-trained on a large corpus of text, BERT can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, including sentiment analysis, without substantial modifications to the architecture. This makes BERT incredibly versatile and efficient for adapting to any specific NLP task.

Handling of "Polysemy": BERT excels in understanding polysemous words (words with multiple meanings depending on context), a common occurrence in human languages. For instance, the word "bank" can have different meanings in "river bank" versus "savings bank", and BERT's bidirectional context helps it discern these differences effectively.

Why BERT is Ideal for Sentiment Analysis:

Contextual Awareness: In sentiment analysis, the context in which a word appears is crucial. BERT's ability to analyze the context from both directions makes it particularly powerful for understanding the sentiment expressed in complex sentences. Subtlety in Language: BERT's deep learning framework is adept at picking up subtle cues in language that may indicate sentiment, such as sarcasm or politeness, which traditional models might overlook. Efficiency in Training: With BERT, you can achieve high accuracy in sentiment analysis with relatively little data compared to what is typically required for training deep learning models from scratch. This is because the pre-trained BERT model has already learned rich representations for a wide range of language patterns during its pre-training on large text corpora.

Nel caso specifico di questa analisi, BERT verrá utilizzato per svolgere sentiment analysis, il modello pre trainato che é possibile importare dalla libreria ?? sará fine tunato su un dataset in cui é stato effettuato il labelling manuale, cosí facendo il modello sará adattato a riconoscere i pattern del contesto a cui é sottoposto e quindi essere in grado di classificare in maniera efficace i commenti yotube.

Le tre classi di commenti che il modello tenterá di classificare sono:

2.3 Topic Modelling: Bert-topic

In the context of this project to perforform topic modelling i used another transformer model: BERT-topic. BERTopic is an advanced modeling tool used for identifying and visualizing topics from large collections of textual data. Unlike traditional clustering algorithms which often rely on bag-of-words approaches, BERTopic leverages state-of-the-art language models based on BERT (Bidirectional Encoder Representations from Transformers) to enhance the understanding of textual context and semantic relationships.

BERTopic utilizes dimensionality reduction and clustering techniques to organize text data into distinct topics, making it highly effective for extracting meaningful insights from unstructured text. This is particularly valuable in your analysis as it allows for the automated grouping of comments into thematic categories that represent the various viewpoints and topics of discussion related to the Israeli-Palestinian conflict.

3 Data

3.1 Data organization

Come anticipato lanalisi tenta di catturare l'opinione del pubblico attraverso le loro interazioni sulla piattaforma online: YouTube. YouTube è una piattaforma di condivisione video online dove gli utenti possono caricare, visualizzare, valutare, condividere e commentare video. Fondata nel febbraio 2005, offre una vasta gamma di contenuti generati dagli utenti, inclusi video musicali, clip di film, trasmissioni in diretta e reportage giornalistici, rendendola una delle principali fonti di intrattenimento e informazione nel mondo digitale. Specificatamente la nostra analisi andrá a targetizzare gli utenti di questa piattaforma social che hanno interagito tramite la funzione 'commento' con video riguardanti il conflitto Israelo Palestinese. La poplazione target é perció costituita da Youtube Users, the sample dagli utenti che hanno commentanto i video target, e la fondamentale unitá di analisi é il commento espresso.

Tale funzione non ci permette tuttavia di risalire all'orgine geografica dell'utente o di avere maggiori informazioni sulla sua persona. Di conseguenza, al fine di rispondere alle research questions, ho dovuto adottare qualche accorgimento che mi permettesse di organizzare i dati in maniera tale che fossero in grado di fornire una visione quanto piú globale possibile delle opinioni riguardo al conflitto.

Per fare ció sono partito dall'assunzione che gli utenti youtube siano piú propensi a consumare contenuti prodotti da network d'informazione che hanno la loro stessa origine o la loro stessa visione. Prendendo in considerazione perció video pubblicati da network da diverse parti del mondo avrei potuto costruire un sample che fosse omogeneo in termine degli argomenti di discussione ma eterogeneo in termini del tipo di utenti coinvolti.

I network di informazione dai quali ho selezionato i video per svolgere l'analisi (che nel contesto di questo studio identificheremo come 'Inference Datasets') sono stati:

- Al Jazeera English, un canale di notizie internazionale in lingua inglese, parte della rete mediatica di Al Jazeera con sede a Doha, Qatar. Lanciato nel 2006, il canale offre copertura giornalistica globale, concentrando l'attenzione su eventi del Medio Oriente, Asia, Africa e oltre. È noto per il suo approccio approfondito e le analisi dettagliate delle notizie internazionali, spesso mettendo in luce storie trascurate dai principali media occidentali. Ho scelto questo network nell'ipotesi che fosse visitato soprattuto da utenti provenienti dal medio oriente, viste le sue origni e argomenti di pubblicazione.
- CNN (Cable News Network) è un canale di notizie via cavo e satellite statunitense fondato nel 1980. È uno dei principali network di informazione a livello mondiale, noto per la sua copertura continua 24 ore su 24. CNN offre notizie, reportage e analisi su politica, economia, salute, scienza e intrattenimento, servendo un pubblico globale e fornendo perció un punto di vista sulle opinoni e le notizie diffuse da uno dei principali canali d'informazione statunitensi.
- Sky News è un canale di notizie britannico, lanciato nel 1989 da Sky Group. È noto per la sua copertura di notizie 24 ore su 24, che fornisce aggiornamenti continui sugli eventi internazionali, politici, economici e di cronaca. Sky News si distingue per la sua copertura approfondita delle notizie nel Regno Unito e nel mondo, spesso con una prospettiva britannica. Selezionando Sky News ho provato a fornire una prospettiva proveniente dal continente Europeo.

Inoltre al fine di approfondire l'analisi ed investigare anche l'evoulzione del sentiment e gli argomenti di discussione nel corso del tempo, ho selezionato per ogni canale di informazione 2 video, il primo datato ottobre 2023 in corrispondenza dell'attacco terroristico by Hamas-led gunmen who breached the border

fences between Gaza and Israel. This assault resulted in the tragic deaths of at least 1,400 people, many of whom were civilians, including children. This violent event is characterized by its scale and the severe human toll it took, leading to a widespread international reaction. Identificando quindi, nel contesto della nostra analisi questo momento come lo starting point delle opinioni degli utenti a riguardo del conflitto. Il secondo video di ogni canale invece è datato 2024, approssimativamente 1 anno dopo quell'evento. Così facendo ho potuto ottenere delle stime che tenessero in considerazione anche l'aspetto temporare e idenficare possibili informazioni interessanti sull'evoluzione del sentiment e degli argomenti di discussione.

An additional selection that had to be made at the level of individual comments concerned the language of expression. To ensure the effective functioning of models like BERT and BERT-topic, which are designed primarily for English language processing, I had to consider only comments expressed in English. This choice was driven by the need to maintain the accuracy and reliability of the topic modeling and sentiment analysis, as these models are optimized for English due to the extensive training data available in this language.

Per concludere nel contesto dell'analisi é stato necessario creare due tipi di dataset, un dataset per fare il fine-tuning di BERT, pre far si che il modello pre trainato si adatti e impari, sulla base delle osservazioni fornite ad eseguire la sentiemnt analysis e perció imparare a classificare i commenti sulla base del contenuto. E un secondo tipo, che ho denominato inference dataset, composti dai dati presi dai network precedentemente descritti, e sui quali sono andato effettivamente a svolgere l'analisi vera e propria.

3.2 Fine - Tuning Dataset

The first dataset that I had to build was the one used for fine-tuning the BERT model. Fine-tuning BERT is essential for tailoring the model to specific tasks or domains, such as sentiment analysis on YouTube comments. Originally, BERT is pretrained on a vast corpus of text to comprehend general language patterns and structures. However, it does not inherently grasp the nuances specific to particular datasets or topics. By fine-tuning BERT on a carefully curated dataset, the model can adapt to the specific linguistic and contextual cues of the data it will analyze, such as recognizing sentiment indicators unique to discussions around polarizing topics.

Per selezionare i video dai quali trarre i commenti per la creazione di questo dataset é stato necessario, ovviamente selezionare video riguardanti il conflitto israelo palestinese, cosí che il modello potesse acquisire knoledge riguardo agli argomenti di discussione e le tematiche trattate, ma in particolare é stato necessario selezionare video che contenessero commenti molto polarizzanti e che fossero ricchi di argomenti di discussione, ció é stato fondamentale per creare un modello che fosse in grado di generalizzare, capire gli argomenti di discussione e assegnare il corretto sentiment al commento.

A tal proposito ho manualmente cercato video che rispettassero questi requisiti, trovandone 6 da cui ho appunto estratto i commenti, mischiato randomicamente per garantire la creazione di un sample uniforme.

I video selezionati per creare il dataset di fine tuning (aggiusta titoli):

1. AL JAZEERA ENGLISH 7 OCTOBER DOCUMENTARY [1 metti i link in appendix]
2. CNN MOSTRA CAM DEL SOLDATO DI HAMAS NELL'ATTACCO DEL 7 OTTOBRE [2]
3. AL JAZEERA ENGLISH INVESTIGATING WAR CRIME IN GAZA [3]
4. CNN PEOPLE IN GAZA FEEL ABANDONED BY THE WORLD 40K PALESTIANS KILLED ETC [4]
5. Gaza towers collapse after explosion (CNN) [5]
6. Israel-Gaza: At least half of Gaza's buildings damaged or destroyed, new analysis shows [6]

Cosí facendo ho ottenuto un dataset composto da 19121 commenti, dai quali ne ho manualmente assegnato una label a 1486 commenti, costruento l'effettivo dataset per il finetuning

3.3 Inference datasets

Per la scelta dei dataset sui quali eseguire l'analisi, come anticipato ho selezionato video dai canali: Al Jazeera English, CNN, Sky News. Per ognuno dei quali ho scelto due video, uno pubblicato a ottobre 2023, per le ragioni precedentemente spiegate e uno pubblicato all'incirca 1 anno dopo. Un ulteriore aspetto che ho preso in considerazione in questo contesto é stato selezionare video che presentassero un ammontare di commenti (come??)

Inference Datasets:

Cosí facendo ho ottenuto due video per ogni canale d'informazione, avendo l'opportunitá di investigare eventuali correlazioni relative a questi, oppure aggregare i dati in funzione temporale. Descrivi velocemente le variabili di ogni singolo dataset.

3.4 Dataset Creation

Per creare tutti i dataset dei commenti é stato necessario utilizzare l'API messa a disposizione da Youtube, in breve il processo per la creazione del dataset di fine tuning é stato questo:

1. Library Utilization: The project began by incorporating the googleapiclient.discovery library, which facilitated access to YouTube's API, allowing for an automated way to fetch comments from specified videos.
2. API Key Creation: An essential step was generating an API key, a unique identifier that authorized and enabled access to YouTube's video data through the API. This key ensured secure and authorized data retrieval without exposing sensitive account details.
3. Fetching Comments: To systematically gather comments, first i extraced the video id from each youtube's video, then a Python function get_video_comments was defined. This function accepts a video_id and an optional parameter max_results to specify the number of comments to fetch per API call. The function iterates over all available comments for a video, extracting crucial information such as the comment text, number of likes, number of replies, and the publication date. This iterative process continues until all pages of comments are fetched, as managed by YouTube's API pagination.
4. Data Aggregation and Preprocessing: After collecting comments from various videos, the datasets from different video sources were concatenated to form a comprehensive dataset. This dataset was then shuffled to ensure a random distribution of data, which helps in unbiased model training.
5. Language Filtering: To maintain consistency in language processing and enhance the accuracy of the sentiment analysis model, a filtering step was included using the langdetect library. This library helped identify and retain only those comments written in English, ensuring the model trained on a uniform dataset.
6. Sentiment Column Addition: For the purpose of model training and subsequent predictions, a 'sentiment' column was added to the dataset. Initially, this column was populated with default values or placeholders, which were later replaced with actual sentiment labels either through manual labeling or automated predictions.

While for the inference datasets i followed the same process but without shuffuled the data and without aggregating the comments of the different videos in one dataset.

4 Sentiment Analysis

Una volta aver allenato il BERT model sul dataset di finetuning lo ho utilizzato per effettuare le predizioni sui dataset di inference. Il primo aspetto che é interessante investigare a questo punto riguarda la distibuzione complessiva del sentiment su tutti i video che ho selezionato, in modo da avere una visione globale delle opinioni degli utenti Youtube per poi investigare ulteriormente. Per fare ció é possibile realizzare un grafico a barre su tutti i dataset uniti.

(mettere numeri figure) Analyzing the sentiment across all sampled datasets, it is evident that the majority of comments express a neutral opinion. This observation suggests that comments under YouTube videos serve as a platform where individuals tend to express their thoughts in a manner that may encourage dialogue rather than merely passing judgment or condemning what they see. Naturally, the resultant sentiment is influenced by the sample of videos we have selected. Additionally, it is noteworthy that the number of comments supporting Israel exceeds those in favor of Palestine. In subsequent analyses, we will delve deeper into this phenomenon to understand these data thoroughly.

The prevalence of neutral comments might stem from various factors. Firstly, viewers who comment might prefer to discuss the content of the videos in a balanced or analytical way, avoiding polarizing opinions to foster a more constructive conversation. Secondly, neutral comments could be indicative of viewers trying to understand both sides of the conflict without showing explicit bias. This neutrality might reflect a broader desire among YouTube users to engage in discussions that are informed and thoughtful rather than confrontational.

Al fine di trarre maggiori dettagli interessanti circa il sentiment espresso dagli utenti ho poi investigato la distribuzione delle loro opinioni rispetto a i diversi canali d'informazione:

Analizzando questi grafici emergono dati interessanti. Possiamo in un certo senso confermare quanto visto in precedenza per quanto riguarda la prevalenza di commenti neutrali, che in proporzione sono quelli di maggior frequenza sia nei video pubblicati da Sky News che dalla CNN, mentre per Aljazeera i commenti neutrali sono superati da quelli Pro Israele. Tale risultato é indicazione del fatto che questo canale d'informazione genera una maggiore polarizzazione delle opinioni; gli utenti che hanno commentato i video di Aljazeera tendono a mostrare maggiormente la loro opinione e il loro sostegno a una causa, piuttosto che presentare argomentazioni neutrali. La maggior affluenza di commenti di una fazione piuttosto che l'altra puó essere dovuta principalmente a due fattori: O un sostegno di questa per gli argomenti presentati nei video, oppure una forte opposizione. Mediante le successive analisi di topic modelling sará interessante investigare e cercare di capire se possiamo identificare dei pattern di preferenza o opposizione rispetto a un canale piuttosto che un altro.

Concentrandosi sui commenti a sostegno di una fazione piuttosto che l'altra possiamo notare come anticipato che Aljazera é il canale con la proporzione maggiore di commenti pro israele : il 41.1%, seguito da Sky con il 35.1% e infine CNN con il 22.9%. Per quanto riguarda i commenti pro Palestina possiamo notare che sono piú frequenti nei video di CNN con il 26.1%, seguiti dal 24.2% su Sky e il 23.6% per Aljazeera.

Dai risultati ottenuti sembra che per quanto riguarda l'opinione espressa dagli utenti ci sia una sorta di contrarietá tra i canali Aljazeera e CNN. Il primo infatti ha la maggior presenza di commenti pro israele e la minor affluenza di commenti pro palestina tra i 3 canali. Mentre é esattamente il contrario per CNN che ha la maggior frequenza di commenti pro Palestina e la minore di pro israele. Questo aspetto sará imporante da tenere in considerazioni durante l'analisi di topic modelling per verificare se gli utenti di una certa fazione sono pro o contro questi canali e spiegare tali dati.

Un altro dato interessante da analizzare riguarda l'evoluzione del sentiment dopo un anno dall'inizio del conflitto:

Come possiamo notare dalla figura n? dopo un anno dall'attacco dell'ottobre 2023, delle truppe Palestinesi a Israele, che nel contesto della nostra analisi abbiamo classificato come l'inizio del conflitto, il sentiment esresso dagli utenti é cambiato. Il primo dato che é possibile notare é la diminuzione della proporzione dei commenti neutrali dal 2023 (43.5%) al 2024 (37.2%). Ció potrebbe essere indice del fatto che con il passare del tempo e il susseguirsi del conflitto gli utenti si siano fatti una propria idea a riguardo, e sia cresciuta di conseguenza la tensione. Tensione che si é manifestata con un incremento delle opinioni in supporto a una fazione piuttosto che l'altra. Come possiamo notare dal bar chart infatti, la proporzione del sentiement espresso si é uniformata. Osservando l'evoluzione della proporzione di commenti pro Palestina, ho riscontrato un incremento di quasi il 6%, incremento che invece non si é verificato nel caso di commenti pro israele, la cui proporzione é rimasta quasi invariata. A fronte della dimuzione di commenti neutrali riscontrata in precedenza questi dati ci mostrano come, dopo 1 anno di conflitto, la causa palestinese sembri essere quella piú sentita. Sulla base del nostro sample potremmo infatti dedurre la presenza di un trend che ha portato utenti che inizialmente avevano un'opinone neutrale a sostenere poi la causa palestinese. Tuttavia non possiamo con certezza confermare ció in quanto le suddette analisi sono comunque influenzate dal fatto che i video presi in considerazione per quanto coerenti nel trattare il conflitto israelo palestinese, singolarmente presentano diversi aspetti introducendo del bias in queste considerazioni.

Investigando piú nello specifico l'evoluzione del sentiment rispetto ai diversi canali di informazione possiamo notare che nel caso del canale Aljazeera riscontriamo una decisiva diminuzione dei commenti pro israele e una crescita delle opinioni pro palestina, allineando questo canale al trend precedentemente evidenziato. Anche l'evoluzione del sentiment nel canale di CNN si allinea a quanto abbiamo potuto constatare nell'analisi generale. Notiamo infatti una decisiva diminuzione dei commmenti neutrali su questo canale, sintomo di una maggiore polarizzazione che si manifesta con un ingente incremento dei commenti pro israele, diventando il canale con la proporzione maggiore sotto questo aspetto, ma con anche la crescita dei commenti pro palestina. Analizzando SKY invece possiamo notare come la distribuzione del sentimento espresso sia rimasta piú o meno invariata, mostrando nel tempo una maggiore incidenza di commenti neutrali, seguiti da i commenti pro israele e infine pro palestina.

Per concludere questa analisi temporale per specifico canale possiamo affermare che l'evolzione del conflitto ha portato una maggiore polarizzazione nelle opinioni espresse dagli utenti Youtube, CNN si configura come il canale in cui questo trend si é verificando maggiormente essendo l'unico in cui i commenti faziosi sono stati maggiori di quelli neutrali nel 2024 nonostante nel 2023 fosse quello con l'incidenza maggiore di quest'ultimi. Sky news é invece il canale rimasto piú equilibrato nel tempo e Aljazera é quello che ha ottenuto il maggior incremento percentuale di commenti pro palestina dal 2023 al 2024 e la maggior diminuzione di commenti pro israele tra i 3, rimanendo comunque in maggioranza. (valuta se inserire analisi lunghezza commenti rispetto al sentiment)

5 Topic Moddelling: BERT-Topic

La sentiment analysis ci ha permesso di ricavare interessanti informazioni sulla distribuzione del sentiment, ora mediante il modello BERT-topic ho investigato per ciascun sentiment i principali argomenti di discussione.

Per avere una prima overview iniziale degli argomenti di discussione ho creto un semplice word cloude che ci consente di notare quali sono state le parole piú frequenti:

Dal grafico si evince che le parole piú diffuse sono sicuramente “Israel”, “Hamas”, “people”, “gaza”, “War” facendoci capire che discussions are centered around the geopolitical dynamics and human impacts of the conflict. Anche parole come “killed,” “people”, “peace”, underline the significant concerns for human welfare and the active advocacy for action and support among viewers.

Investigando piú nel dettaglio gli argomenti di discussione sull’intero dataset ho svolto grazie al modello BERT-topic un’analisi di topic modelling.

Il topic 0, che rappresenta l’argomento piú comune di discussione possiamo notare che sia incentrato come potevamo aspettarci sulle due fazioni rivali, hamas e Israles.

Spiccano successivamente argomenti With terms such as “school,” “shout,” “middle,” “class,” this might indicate discussions about educational or societal issues, possibly reflecting on situations or events occurring in schools or involving children in conflict zones.

Topic 2: This topic, featuring words like “bid,” “joe,” “iran,” “billion,” and “en,” could be related to international relations or economic discussions, perhaps involving US policies or financial aspects concerning the Middle East.

Topic 3: With words like “October,” “oct,” “7th,” “9th,” and “Sunday,” this topic appears to capture conversations related to specific dates or events, probabilmente riferendosi all’evento del 7 ottobre che ha segnato l’escalation del conflito. Appaiono anche argomenti di discussione contenenti parole come “putin”, “nato”, “Russia”, “Ukraine”, probabilmente facendo riferimenti o analogie rispetto a un altro grande conflitto che é in atto al momento in Europa.

Per concludere possiamo notare come siano comuni espressioni emotive probabilmente relativamente ai contenuti dei video o agli eventi del conflitto, con parole come “horrible”, “awful”, “sorry” che dimostrano un coinvolgimento emotivo da parte degli utenti rispetto alle vicende riportate.

Proseguendo, al fine di determinare quali sono gli argomenti di discussione ci ciascuna fazione, ho eseguito la medesima analisi nello specifico per i commenti: Pro Israel, Pro Palestine, Neutral.

Analizando piú nello specifico gli argomenti di discussione nei commenti pro Israele possiamo notare come il Topic 0 sia centrato attorno a parole chiave come “Hamas,” “Israel,” “Gaza,” “people,” indicando che la discussione si concentra sugli attori principali del conflitto israelo-palestinese. La presenza di termini geografici e nominativi di gruppi coinvolti suggerisce un’analisi diretta delle dinamiche politiche e sociali del conflitto.

Topic 1: Si focalizza su temi educativi o di contesto scolastico, con parole come “shout,” “school,” “elementary,” “class,” e “grade.” Potrebbe riflettere discussioni su come il conflitto è percepito o insegnato nelle scuole, oppure metafore educative usate per discutere il conflitto.

Topic 2: Riguarda la narrativa di ostaggi e vittimizzazione, con termini come “hostages,” “dom,” “victim,” “hostage,” e “free.” Questo suggerisce che sia molto sentito dai sostenitori della causa Israeliana la questione

relativa agli ostaggi e alla loro liberazione, probabilmente riferendosi agli eventi del 7 ottobre 2023 dove i miliziani di Hamas hanno catturato numerosi ostaggi in seguito all'evento ? nome.

Topic 3: È legato a specifici eventi temporali, con termini come "October," "8th," "happened," "7th," e "Sunday." Questo implica discussioni su eventi che sono accaduti in particolari date.

Topic 5 e 6: Mostrano una miscela di termini più casuali o generici come "haven," "se," "guess," "saw," "won," e parole legate a giochi o competizioni come "stupid," "games," "prizes," "play," "win." Questi topics potrebbero riflettere come gli utenti utilizzano il gioco e la competizione come metafore per discutere il conflitto, oppure potrebbero semplicemente distogliere l'attenzione dai temi più pesanti del conflitto.

In sintesi, il grafico mostra come i commenti pro-Israele trattino una gamma di temi che vanno dal politico e storico, al personale e metaforico, con un forte focus su eventi specifici e reazioni emotive. Questo riflette la complessità del discorso intorno al conflitto dove le percezioni, le esperienze personali e la storia collettiva si intrecciano nei modi più vari.

Analizzando invece le parole piú frequenti nei commenti che sostengono la causa israeliana possiamo confermare pressapoco quanto affermato in precedenza. Le parole con maggiore peso, come "October", "8th", e "bear", indicano un'accentuata discussione su eventi specifici, forse riferendosi a date significative o eventi accaduti durante il mese di ottobre. L'uso di parole come "shout" e "fight" suggerisce un tono di confronto o di azione diretta, mentre termini come "hostages" e "eaten" potrebbero riflettere le narrative di conflitto e le percezioni di aggressione o vittimizzazione.

Le parole "school", "elementary", e "class" si evidenziano ancora una volta, suggerendo che le discussioni possono anche toccare gli impatti del conflitto sulla vita quotidiana, in particolare l'educazione dei giovani.

In sintesi, il grafico mostra una miscela di terminologia che riflette l'intensità delle discussioni, con una combinazione di riferimenti diretti al conflitto, eventi specifici e temi educativi. Questa varietà di parole chiave suggerisce che, pur essendo prevalentemente focalizzati sul sostegno a Israele, i commentatori pro-Israele integrano nei loro messaggi una gamma di considerazioni che spaziano dall'educativo al contestualmente specifico e emotivamente carico.

Passando ai commenti pro Palestina: Analizzando i commenti pro palestina possiamo notare che Mentre i commenti pro-Israele erano focalizzati su argomenti di conflitto, vittoria, e discussioni dirette sugli eventi, i commenti pro-Palestina sembrano toccare temi più vari e complessi che includono anche aspetti storici, educativi e umanitari.

Topic 0 e Topic 1: Simili ai commenti pro-Israele, questi topic includono parole legate a "Israel," "Gaza," "people," e "Palestine," indicando una discussione diretta sul conflitto.

Nel Topic 3 con parole come "liar," "unk," "real," "tine," e "les" suggeriscono una discussione sulla veridicità e l'autenticità delle informazioni, probabilmente i sostenitori della causa palestinese non si trovano d'accordo con i fatti narrati e mettono in dubbio la veridicitá delle informazioni diffuse o l'affidabilitá dei canali d'informazione.

Nei successivi argomenti di discussione é interessante notare come siano presenti tematiche religiose ("Jesus," "Christ," "God") e concetti di amore e pace ("love"), che potrebbero essere indicazione del fatto che alcuni utenti, indipendentemente dalla propria religione, esprimano sentimenti di vicinanza e amore nei confronti del popolo palestinese.

Per concludere nel topic numero 8 possiamo notare come le parole "nazis", "hitler", "holocaust" vengano usate dai sostenitori della causa palestinese probabilmente per descrivere le azioni della controparte paragonandole con fatti molto gravi al pari di quelli avvenuti durante il secondo conflitto mondiale.

Nel caso delle parole più frequenti nei commenti pro palestina, possiamo notare che non forniscono ulteriori informazioni rispetto a quanto precedentemente rilevato, ma confermando in un certo senso le tendenze sopra individuate.

Prendendo invece in considerazione i commenti neutrali:

gli argomenti di discussione dei commenti classificati come neutrali possiamo notare innanzi tutto una tendenza a discutere dei network di informazione, le abbreviazioni 'al', 'ja', 'zee' sembrano suggerire discussioni riguardanti aljazeera, la stessa cosa per 'cnn', 'news', 'propaganda', confermando la presenza da parte degli utenti di pareri contrastanti rispetto alle informazioni diffuse su Youtube. Molto probabilmente infatti, alcuni network vengono reputati come faziosi e di non presentare informazioni in maniera parziale, oppure presentando fatti ritenuti non veritieri.

Topic 4 e 5: Temi legati al tempo e alla famiglia o alle vittime del conflitto. Le date in Topic 4 possono riferirsi a eventi specifici, mentre le parole come "children," "kids," "heart," "killed," e "innocent" in Topic 5 sottolineano la discussione sulle conseguenze umane del conflitto.

Parole come "sad," "sorry," "load," "sham," e "mm." invece rivelano una risposta emotiva al conflitto o alle sue rappresentazioni, mostrando compassione o disappunto.

A differenza dei commenti pro Israele e pro Palestina, i commenti neutrali tendono a mostrare un'ampia varietà di prospettive che non si limitano a sostenere una causa specifica. Invece, esplorano le implicazioni più ampie del conflitto, il ruolo dei media, l'impacto culturale, e mostrano una sensibilità per le vittime e le conseguenze umane del conflitto. Questo indica un approccio più olistico e riflessivo nei confronti del conflitto, evidenziando un interesse per le cause profonde

Come precedentemente evidenziato, le parole piú frequenti confermano quanto detto e confermano il trend per il quale i commenti neutrali tendono a concentrarsi su una riflessione più ampia e meno polarizzata del conflitto, considerando la storia, l'educazione, i media e la cultura come elementi centrali per comprendere e discutere il contesto Israelo Palestinese.

Per concludere L'analisi complessiva dei topic modellati dai commenti su YouTube legati al conflitto Israelo-Palestinese mostra distinte narrazioni nei commenti categorizzati come pro-Israele, pro-Palestina, e neutrali, riflettendo la complessità e la varietà delle prospettive e delle reazioni degli utenti.

Commenti Pro-Israele: La discussione in questa categoria si focalizza sugli attori principali del conflitto, con una frequente menzione di Hamas, Gaza e Israele, indicando una discussione diretta sulle dinamiche politiche e sociali. Si notano anche temi di ostaggi e vittimizzazione, che mostrano una preoccupazione per la sicurezza e per le azioni di guerra, oltre a discussioni legate a specifici eventi temporali che hanno segnato momenti di tensione.

Commenti Pro-Palestina: Nei commenti pro-Palestina emergono temi che includono la storia e gli aspetti umanitari del conflitto, con una forte presenza di componenti religiose e di discussioni su giustizia e diritti umani. Le parole chiave come "Jesus," "Nazis," e "holocaust" indicano la tendenza a collegare la situazione palestinese a contesti storici di oppressione.

Commenti Neutrali: Mostrano un'ampia gamma di temi che vanno oltre il conflitto diretto, con discussioni su media, educazione e cultura. Parole come "school," "CNN," e "documentary" suggeriscono un dibattito su come il conflitto è rappresentato e percepito, mentre le date e le menzioni di eventi specifici evidenziano la sensibilità verso il contesto temporale degli eventi.

In conclusione, l'analisi dei topic mostra come i vari gruppi di commentatori utilizzino YouTube come piattaforma per esprimere non solo opinioni dirette sul conflitto, ma anche per riflettere su questioni più ampie

come la storia, i diritti umani, l'educazione e l'impatto dei media. Le differenze tra i gruppi di commenti sottolineano l'esistenza di una narrativa complessa e stratificata che va oltre la semplice dicotomia pro-Israele o pro-Palestina, includendo una varietà di prospettive che riflettono la profondità e la complessità del discorso pubblico su questi temi.

6 Conclusion

In this study, we explored the intricate dynamics of public sentiment and discussion topics regarding the Israeli-Palestinian conflict as expressed through comments on YouTube videos. Our investigation utilized advanced NLP techniques, leveraging BERT for sentiment analysis and BERTopic for topic modeling, to dissect the varied perspectives reflected in the digital public sphere. The research addressed several key questions, providing a multi-dimensional analysis of how sentiments and discussions have evolved and varied across different viewer demographics and news channels.

Research Questions and Findings: Prevailing Sentiment and Primary Topics of Discussion: The analysis revealed a predominance of neutral sentiments, indicating a preference among YouTube users for balanced discussion or a reflective approach rather than direct confrontation. Pro-Israel and pro-Palestine sentiments were also notable but to a lesser extent. The primary topics of discussion included geopolitical dynamics, historical contexts, and human impact, with frequent mentions of key figures and locations such as Hamas, Israel, and Gaza.

Evolution of Sentiment and Discussion: Over time, there was a noticeable shift from neutral to more polarized sentiments, reflecting an increase in user engagement and possibly growing tensions or heightened awareness of the conflict. This evolution suggests that digital platforms like YouTube not only serve as arenas for information dissemination but also for dynamic interaction that can shift public perceptions and engagement over time.

Consistency Across News Channels: Sentiments varied across different news channels, reflecting perhaps the editorial slants and regional focuses of the channels. For instance, Al Jazeera English tended to have a higher incidence of pro-Palestine sentiments, whereas CNN showed a more mixed sentiment distribution. This variation underscores the influence of media framing and audience targeting in shaping public discourse on international conflicts.

Perceptions of News Channel Bias: The analysis indicated that users are keenly aware of and often comment on perceived biases within news channels. Comments frequently pointed out either a favoritism towards Israel or Palestine, highlighting a critical consumption of news where viewers are not only passive recipients but active critics of the media.

Conclusions: The findings from this study illuminate the significant role that digital platforms play in the contemporary discourse surrounding international conflicts like the Israeli-Palestinian conflict. YouTube, as a global platform, offers a unique lens into the public's perceptions and sentiments, which are deeply influenced by historical narratives, media framing, and individual worldviews. The shift from neutral to more defined sentiments over the study period highlights the evolving nature of public engagement with such conflicts, suggesting a growing propensity among viewers to align themselves more distinctly with one side or the other as the conflict progresses.

Furthermore, the distinct topics and sentiments articulated across different viewer demographics and news channels emphasize the fragmented nature of public opinion on this issue, shaped by a complex interplay of cultural, social, and political factors. This fragmentation is also a reflection of the broader global discourse on peace, justice, and international relations.

Ultimately, this study not only advances our understanding of how the Israeli-Palestinian conflict is represented and perceived in the digital age but also sheds light on the broader implications of how modern conflicts are discussed and understood in public forums. These insights are crucial for policymakers, media professionals, and scholars interested in the intersections of media, public opinion, and international conflict.

By leveraging the capabilities of advanced NLP tools, this research contributes to a more nuanced and sophisticated understanding of the digital narratives that shape and reflect global perceptions of pivotal geopolitical issues.