UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea Triennale in Ingegneria Informatica e dell'Automazione

TESI DI LAUREA

Indagine sulle discussioni degli utenti di X sul conflitto tra Russia e Ucraina mediante l'analisi dei dati

Investigating X users' discussions on the Russia-Ukraine conflict using Data Analysis

Relatore

Prof. Domenico Ursino

Correlatore

Dott. Luca Virgili

Candidato

Riccardo Bastiani

ANNO ACCADEMICO 2023-2024

"Success is going from failure to failure without losing your enthusiasm."

attribuita a Winston Churchill

Sommario

L'analisi del social network è divenuta di fondamentale importanza negli ultimi anni con la sua capacità di analizzare eventi, opinioni ed idee. La libertà con cui gli utenti si affacciano all'attualità è senza precedenti, creando una quantità di dati di particolare valore. In particolare la guerra russo ucraina, che ha sconvolto il mondo riportando lo spettro della guerra nei paesi sviluppati è stato oggetto di grande interesse.

In questa tesi si analizza il rapporto simbiotico tra opinioni ed eventi, in particolare attraverso i tweet del social X. Inoltre ci si propone di valutare l'engagement come metrica e la predizione di trend emergenti attraverso il topic clustering.

Per fare questo abbiamo utilizzato un dataset curato su Kaggle contenenti i tweet di milioni di utenti. Lo abbiamo poi analizzato avendo cura di suddividere l'analisi in due macroblocchi. Il primo che riguarda la controffensiva ucraina e il secondo lo stallo successivo ad essa. Abbiamo innanzitutto eseguito delle analisi a largo spettro per avere meglio un'idea degli eventi e delle personalità prominenti. Successivamente abbiamo trovato degli eventi particolari e li abbiamo analizzati attraverso BERTopic.

Abbiamo scoperto in particolare dei pattern comuni tra gli eventi analizzati che non sarebbero stati trovati altrimenti.

Keyword: Social Network, X, Twitter, Conflitto Russia-Ucraina, Topic Modeling, Sentiment Analysis

Indice

Introduzione **1**

1 Panoramica sul conflitto tra Russia ed Ucraina **3**
1.1 Breve storia recente della Federazione Russa: 1991-2003 . . . 3
1.2 Gli antefatti alla rivoluzione ucraina del 2014: la rivoluzione arancione e le sue conseguenze . . . 4
1.3 I fatti di Euromaidan . . . 5
1.4 L'annessione della Crimea . . . 6
1.5 La guerra nel Donbass 2014-2022 . . . 6
1.6 L'invasione russa dell'Ucraina . . . 7
1.7 Il fronte sudorientale 8 aprile - 28 agosto . . . 10
1.8 La controffensiva ucraina: 29 Agosto - 11 novembre 2022 . . . 12
1.9 Lo stallo invernale . . . 13

2 Strumenti per l'analisi dei Social Network **16**
2.1 Social Network Analysis . . . 16
2.1.1 Perché le reti sociali sono così potenti? . . . 17
2.1.2 I network bimodali e le loro tipologie . . . 18
2.1.3 La diffusione delle informazioni . . . 19
2.1.4 Il concetto di omofilia . . . 20
2.2 Machine Learning . . . 21
2.2.1 Supervised learning . . . 21
2.2.2 Unsupervised learning . . . 22
2.2.3 Deep learning . . . 22
2.3 Natural Language Processing . . . 23
2.3.1 Text classification . . . 24
2.3.2 Information Retrivial . . . 25
2.3.3 Sentiment analysis e VADER . . . 25
2.3.4 Topic clustering . . . 26
2.4 Trasformers . . . 26
2.4.1 Introduzione all'architettura . . . 26
2.4.2 BERT . . . 28
2.4.3 BERTopic . . . 29
2.5 Python per l'analisi dei dati . . . 31
2.5.1 Introduzione al linguaggio . . . 31

2.5.2 Data Cleaning tramite pandas 32
2.5.3 Data Visualization tramite Seaborn, Plotly, Matplotlib 32

3 Analisi di X sul conflitto tra Russia e Ucraina **34**
3.1 Analisi esplorativa del dataset 34
3.1.1 Storia del dataset 34
3.1.2 Descrizione del dataset e statistiche 35
3.1.3 La problematica dei bot 37
3.2 Analisi della Controffensiva Ucraina 37
3.3 Analisi dello Stallo 45
3.4 Confronto tra le due Analisi 53
3.5 Analisi dei fattori predittivi dei trend su X 54
3.5.1 Analisi di un evento della Controffensiva Ucraina 54
3.5.2 Analisi di un Evento dello Stallo 59

4 Discussione **66**

5 Conclusioni **68**

Bibliografia **70**

Ringraziamenti **72**

Elenco delle figure

1.1 L'invasione iniziale, territori attaccati al 25 Febbraio 8
1.2 Situazione dell'invasione dell'Ucraina a fine marzo 2022 10
1.3 Situazione al 9 settembre, dopo la liberazione dell'Oblast di Kharkiv 13

2.1 La massa critica come punto di equilibrio 20
2.2 Breve storia dei più recenti ed importanti sviluppi dell'NLP 23
2.3 Architettura del Transformer 27
2.4 Pre-Training BERT model 28
2.5 Step del funzionamento di BERTopic 30

3.1 Utenti unici giornalieri durante il periodo della controffensiva ucraina 38
3.2 Utenti unici e tweet totali giornalieri durante il periodo della controffensiva . 39
3.3 Utenti che hanno postato maggiormente per giorno 40
3.4 Engagement giornaliero durante il periodo della controffensiva ucraina . . . 41
3.5 Engagement totale sovrapposto al grafico degli utenti unici e dei tweet totali 42
3.6 Utenti unici nel periodo di tempo considerato 47
3.7 Numero di utenti e numero di tweet giornalieri 48
3.8 Utenti unici per giorno 49
3.9 engagement totale sovrapposto ai tweet e agli utenti unici 50
3.10 Correlazione tra i topic identificati 55
3.11 Intertopic distance dei giorni tra il 4 ed il 10 settembre 56
3.12 Evoluzione dei topic durante il periodo dal 4 al 10 settembre 57
3.13 Intertopic distance del giorno 11 settembre. 58
3.14 Intertopic distance dei giorni dal 12 al 18 settembre 59
3.15 Evoluzione dei topic nel periodo tra il 12 e il 18 settembre 59
3.16 Interopic distance dei giorni precedenti al 25 febbraio 60
3.17 Heatmap dei giorni compresi tra il 22 e il 24 febbraio 62
3.18 Evoluzione dei topic nei giorni tra il 22 e il 24 febbraio 62
3.19 Interopic Distance del 25 Febbraio 64
3.20 Interopic Distance dei giorni 25, 26 e 27 febbraio 65
3.21 Evoluzione nel tempo dei topic dei giorni 25, 26 e 27 febbraio 65

Elenco delle tabelle

3.1 Username e numero di tweet degli utenti più attivi nel periodo tra il 29 Agosto e l'11 novembre . . . 37
3.2 Prime 10 date con il maggior numero di Utenti unici . . . 38
3.3 Utenti che hanno generato più engagement . . . 40
3.4 Engagement per gli utenti che hanno postato maggiormente . . . 41
3.5 Primi 10 picchi per Engagement derivati dalla distribuzione dell'engagement 42
3.6 Frequenza degli hashtag . . . 44
3.7 Sentiment score per i migliori Hashtag . . . 45
3.8 Username con il maggior numero di tweet . . . 46
3.9 Dieci giorni con maggior numero di utenti unici . . . 47
3.10 Engagement totale per Username . . . 49
3.11 Engagement totale degli utenti col maggior numero di tweet . . . 50
3.12 Hashtag frequencies and percentages . . . 51
3.13 Calcolo del sentiment dei tweet che contengono gli hashtag . . . 52
3.14 Comparazione del sentiment tra gli hashtag delle due parti del dataset . . . 53

Introduzione

Negli ultimi anni i social network si sono di fatto sostituiti agli spazi sociali Chin e Chignell [2007]; Kurka *et al.* [2015] e hanno trasformato in modo radicale il modo in cui comunichiamo sia a livello sociale sia professionale. Hanno abbattuto le barriere geografiche e permesso di comunicare e condividere emozioni, idee, opinioni in tempo reale. Hanno permesso di diffondere notizie e democratizzare l'informazione. Oggi chiunque di noi ha la possibilità di documentare fatti e avvenimenti.

Proprio per questa facilità di espressione, gli eventi di tutti i giorni e l'attualità sono sempre più discussi. Eventi di rilevanza locale e globale si fondono in un tutt'uno ogni giorno, tutti i giorni. Con la rapida evoluzione dei social, abbiamo anche assistito ad eventi tragici, che catalizzano l'attenzione e l'interesse delle persone. Eventi come il Covid, crisi economiche, terrorismo e guerre sono stati condivisi e discussi in ogni loro aspetto da una pletora di persone come mai prima d'ora.

Tra gli eventi che hanno sconvolto l'attualità la guerra in Ucraina è sicuramente tra le più importanti e catastrofiche dei giorni d'oggi Makhortykh e Sydorova [2017]; Racek *et al.* [2024]; Sufi [2023].

Il rinnovato timore di una guerra sul territorio europeo, la scoperta di essere sguarniti contro minacce esterne ed estremamente legati a regimi autoritari ha squarciato il velo degli spettri del passato, facendo ripiombare il continente e tutta la comunità occidentale in una situazione che non si vedeva dalla seconda guerra mondiale.

Essa è la prima guerra di tale rilevanza ad accadere nell'epoca in cui i social media sono ubiquitari Alieva *et al.* [2022]; Mir *et al.* [2023]. La diffusione dell'informazione, come mai prima d'ora in questo momento, ha permesso di avere video, immagini, reportage, informazioni di ogni tipo dal valore inestimabile per documentare la guerra in corso.

Uno dei social media più attivi e più interessanti per questa tipologia di informazione è X, che con la sua visione di microblogging si presta perfettamente ad interventi estemporanei e a rapida diffusione.

Grazie a questa sua capacità di essere sempre al centro dell'attenzione e alla velocità di diffusione delle stesse, lo studio dei social per comprendere cosa pensi la gente è di primaria importanza. La creazione di nuovi trend, nuove idee e l'orientamento di esse. La nascita di proteste, di movimenti popolari e di rivoluzioni vengono sempre veicolate nei social. La possibilità di una così rapida diffusione ha modellato la società al giorno d'oggi. Capire, analizzare e conoscere ciò che le persone esprimono è fondamentale per sapere come si muove il nostro oggi e come si muoverà domani.

In questa tesi ci proponiamo, di analizzare i social attraverso tecniche di data science per i social. Attraverso l'utilizzo di Python, in particolare di Pandas, di modelli di sentiment

analysis come VADER e di topic clustering come BERTopic. Attraverso questi strumenti studieremo i tweet per comprendere come gli eventi si sono dipanati e analizzeremo alcuni dei più interessanti, per portarli alla luce. Inoltre cercheremo di valutare se essi sono predicibili dall'analisi dei tweet precedenti.

In particolare il dataset, ottenuto da Kaggle, copre i tweet del primo anno e mezzo di guerra, dal febbraio 2022 al giugno 2023, catturando le fasi iniziali e i primi sviluppi del conflitto. Attraverso analisi sui tweet abbiamo trovato nell'engagement un predittore efficace nel mondo dei social. Esso, supera di gran lunga metriche come il numero di tweet e ci permette di andare ad identificare precocemente trend prima della loro massima popolarità. Aspetto che si è rivelato particolarmente utile nelle analisi degli eventi specifici, in particolare nell'analisi pre e post evento. Inoltre l'analisi degli hashtag ha fornito dettagli importanti sull'evoluzione delle idee e dei sentimenti relativi al conflitto. Ci ha permesso di tracciare la nascita e il declino di particolari temi, di notare la differente risposta dipesa dalle varie nazioni e di osservare il sentiment relativo al conflitto.

La tesi è organizzata come segue. Nel capitolo 1, andiamo a dare un breve contesto storico dei due paesi, cercando di capire alcune delle motivazioni moderne che hanno portato alla guerra. Nel capitolo 2, andiamo ad introdurre più nel dettaglio a cosa serve la Social Network Analysis e la sua importanza. Inoltre andiamo a descrivere alcune delle principali architetture nell'ambito del Machine Learning, della sentiment Analysis, gli strumenti correlati utilizzati come VADER e BERTopic ed infine una breve disamina sul linguaggio Python e sulle librerie più importanti utilizzate. Nel Capitolo 3, andiamo nel cuore dell'analisi dividendola in due blocchi. Per ognuno di questi andiamo a fare un'analisi generale dei tweet, con statistiche di sorta, valutazione dell'engagement, analisi degli utenti più importanti. Analisi degli hashtags e sentiment analysis dei tweet correlati ad essi. Infine andiamo, per ogni sezione ad analizzare un evento in particolare. Andando a trovare, predire ed analizzare trend e motivazioni che lo hanno reso importante agli occhi delle persone. Nel Capitolo 4, andiamo a discutere i risultati del capitolo precedente. Infine, nel Capitolo 5, andiamo a trarre le conclusioni della tesi e descrivere possibili lavori futuri.

CAPITOLO 1

Panoramica sul conflitto tra Russia ed Ucraina

In questo capito si analizzeranno, in primo luogo, gli anni precedenti l'invasione della Russia in Ucraina; successivamente, con un'analisi di ampio respiro, si suddividerà la guerra attuale in vari momenti temporali che riflettono l'andamento della stessa seguendo la più comune delle ripartizioni attualmente disponibili.

In particolare verranno presi in considerazione gli eventi del 2013-2014, la serie di manifestazioni chiamate "Euromaidan", la conseguente "rivoluzione di Maidan", l'invasione della Crimea e la guerra del Donbass che accompagna le vicende fino al 2022 con l'invasione terrestre su larga scala dell'Ucraina assieme alle fasi del primo anno della guerra.

1.1 Breve storia recente della Federazione Russa: 1991-2003

Con il discioglimento dell'Unione Sovietica il 26 dicembre del 1991, l'Ucraina, assieme a tutte le altre repubbliche ex-sovietiche, ha ottenuto l'indipendenza[1]. Questo evento ha segnato la fine di un'era e l'inizio di un nuovo capitolo per molte nazioni che, dopo decenni di controllo sovietico, si trovavano ora a navigare le acque turbolente dell'autonomia nazionale. Nonostante gli anni successivi abbiano visto un periodo di relativa pace per tutto il continente europeo, emergono nuove sfide e tensioni politiche e economiche. Nel 1999, in seguito al disastroso passaggio dall'economia pianificata all'economia di mercato, Vladimir Putin diventa il presidente della Federazione Russa, succedendo a Boris El'cin. Putin, prima di assumere la presidenza, aveva scalato le gerarchie come capo dei servizi segreti (FSB) , dimostrando una ferrea determinazione nel mantenere il controllo e stabilità all'interno del paese, anche utilizzando metodi estremi ed illegali.

Nello stesso anno, la Russia avvia un'offensiva contro la Cecenia, dichiarando illegittimo il governo ceceno. Questo conflitto culminerà con la presa della capitale Grozny nel febbraio 2000 e causerà oltre 1500 vittime. Considerata a posteriori come la prima mossa per assoggettare il potere permanentemente su di sè e ristabilire la figura dell'uomo forte. La guerra in Cecenia non solo ha rappresentato una violenta riaffermazione del controllo russo sulle regioni satelliti ma ha anche mostrato al mondo la risolutezza di Putin nel mantenere un totale controllo della Federazione Russa e di cambiare passo dopo le esperienze degli anni 90.

Nel 2003, le relazioni tra Russia e Georgia[2] cominciano a deteriorarsi ulteriormente. La Georgia, diventata sempre più filoccidentale, intensifica i suoi sforzi nella guerra interna contro le fazioni filorusse ed indipendentiste, alimentando le tensioni tra i due paesi. Queste

[1] https://en.wikipedia.org/wiki/History_of_the _Russian_Federation

[2] http://aei.pitt.edu/9382/2/9382.pdf

tensioni raggiungono un punto di crisi nell'aprile 2008, quando la Russia, a seguito della promessa della NATO di considerare l'ingresso della Georgia nell'alleanza atlantica, scatena una violenta risposta sul paese colchide, che vede l'espansione della NATO come una minaccia diretta alla sua sfera di influenza.

Nell'agosto dello stesso anno, la situazione degenera in un conflitto aperto con l'invasione russa della Georgia. La guerra, seppur breve, lascia segni profondi nella regione e nella geopolitica internazionale. Il conflitto termina pochi giorni dopo quando la Russia riconosce l'Abkhazia e l'Ossezia del Sud, due regioni separatiste filo-russe, come stati indipendenti. Questo riconoscimento non solo altera drasticamente la mappa politica della regione, ma stabilisce anche un precedente per l'intervento russo negli affari interni dei paesi dell'ex sfera sovietica, segnando un periodo di crescente instabilità.

1.2 Gli antefatti alla rivoluzione ucraina del 2014: la rivoluzione arancione e le sue conseguenze

Sin dal crollo dell'Unione sovietica l'Ucraina rimane particolarmente vicina alla Russia, attraverso diversi trattati e la cessione delle testate nucleari, viene considerata come uno stato cuscinetto tra Russia e il Blocco occidentale[3]. Con la salita al potere del presidente Leonid Kuchma, inizia un discreto periodo di crescita economica, a seguito della distensione post guerra fredda a cui però si aggiungono molti scandali politici e di corruzione. Il rapimento e l'omicidio del giornalista Georgij Gongadze, in particolare, solleva grandi accuse di corruzione e contribuiscono alla sfiducia verso il governo.

La tensione raggiunge un massimo relativo durante le elezioni del 2004. Il ballottaggio tra Viktor Yanukovyc, delfino di Kuchma e filorusso, contro il filoeuropeo Viktor Yushchenko, vengono criticate aspramente a causa dei brogli elettorali. Le accuse risultano così gravi da sfociare in un movimento di piazze, conosciuto come Rivoluzione Arancione, dal colore del partito del candidato filoeuropeo. A seguito delle quali le elezioni vengono rifatte sotto egida e controllo internazionale. Esse porteranno alla vittoria di Yushchenko.

Nel periodo della sua presidenza abbiamo un progressivo avvicinamento all'Unione europea e alla NATO, che negli anni avevano aperto al blocco ex sovietico. La presidenza risulta però molto travagliata, a causa della portata delle riforme e di alcune figure come il primo ministro ed oligarca Yulija Timoshenko che rendono la situazione molto difficile da gestire e portano così alle elezioni del 2006, che la vedono come come vincitrice, ma senza riuscire nell'intento prestabilito.

Nel 2010, Viktor Yanukovyc, eletto presidente dell'Ucraina, promuove varie modifiche della costituzione che accentrano i poteri nella figura del presidente. Questo cambiamento costituzionale gli consente di assumere un controllo quasi autocratico del governo, depauperando l'autorità del parlamento e delle istituzioni democratiche. Con il passare del tempo, Yanukovyc inizia a limitare le libertà civili, portando a una crescente repressione politica. Giornalisti, attivisti e oppositori politici diventano bersagli di persecuzioni sistematiche, spesso attraverso l'uso spregiudicato delle leggi, essi sono soggetti ad arresti arbitrari, processi manipolati e altre forme di intimidazione, creando un clima di paura e censura che soffoca il dissenso e la libertà di espressione e con una chiara visione a favore di un rafforzamento delle politiche verso la Russia che esercita una forte pressione per negoziare accordi controversi aumentando la sua influenza sul paese. Tra questi vi sono gli "accordi di Charkiv" del 2010, che estendono il diritto di permanenza della flotta russa del Mar Nero in Crimea in cambio di uno sconto sul gas naturale russo. Inoltre, viene introdotta la legge 'Sui principi della politica

[3] `https://www.europarl.europa.eu/cmsdata/267948/Matviichuk_Ukrainian%20NGO%27s.EuroMaidan.November%202013-February2014.Eng..pdf`

linguistica di Stato', che conferisce alla lingua russa lo status di lingua regionale nelle aree con significative popolazioni di lingua russa e principalmente la parte orientale del paese.

Nel 2012, a seguito di un accordo di libero scambio tra l'Unione Europea e l'Ucraina, la Russia intensifica le sue ingerenze negli affari interni ucraini. Preoccupata dalla crescente vicinanza dell'Ucraina all'UE, La Russia inizia una guerra commerciale con l'Ucraina, imponendo restrizioni sulle importazioni ucraine con l'obiettivo di indebolire l'economia ucraina e di bloccare l'accordo di libero scambio con l'UE, infliggendo un duro colpo agli scambi commerciali del paese.

Alcune delle condizioni previste nell'accordo di libero scambio miravano a ripristinare lo stato di diritto Tra le quali vi era la richiesta di liberazione degli attivisti e degli oppositori politici incarcerati, come Julija Tymošenko, nonché misure per ridurre i poteri accentratisi nella figura del presidente. Queste condizioni erano viste come passi necessari per allineare l'Ucraina agli standard democratici europei e per promuovere un governo più trasparente e responsabile.

Il 21 novembre 2013, sotto la crescente pressione economica e politica esogena e il tracollo degli scambi commerciali con la Russia, il governo ucraino decide di sospendere l'accordo con l'Unione Europea.

Questa decisione inaspettata, scatena una serie di proteste di piazza nella settimana successiva. Molti ucraini considerano la sospensione dell'accordo come un tradimento delle aspirazioni europeiste e di ammodernamento del paese e come un segno di un'ulteriore integrazione con la Russia. Le proteste, iniziate come dimostrazioni pacifiche, si trasformano rapidamente in un movimento di massa noto come Euromaidan, che chiede la ripresa delle trattative con l'UE e il rispetto dei diritti democratici. La situazione culmina in una crisi politica che porta a profonde divisioni interne e a ulteriori tensioni con la Russia

1.3 I fatti di Euromaidan

La corruzione dilagante nel governo ucraino e nelle amministrazioni locali, caratterizzata da un'impunità diffusa e una cronica mancanza di una visione economica a lungo termine, la crisi economica che ha colpito il paese, la tendenza al guardare con sempre maggiore interesse la Russia e la considerevole limitazione delle libertà personali portano ad un'esplosione del malcontento che si formalizza in proteste molto sentite da gran parte della popolazione più giovane

La sera del 21 novembre 2013, un considerevole numero di persone si riversa nelle piazze di Kiev per manifestare contro il governo. Le rivolte si estero rapidamente in tutto il paese, coinvolgendo sempre più cittadini nei giorni successivi. Le manifestazioni di piazza diventano permanenti, con accampamenti in molte città e un'occupazione continua delle piazze.

Il 30 novembre 2013, la situazione degenera. Un'azione di sgombero forzato da parte della polizia antisommossa contro un gruppo di giovani manifestanti provoca decine di feriti. La notizia dell'uso della forza si diffonde rapidamente in tutto il paese, generando un'ondata di sdegno e indignazione. Secondo sondaggi successivi, questo evento è stato indicato come una delle principali motivazioni che hanno spinto molti ucraini a partecipare alle proteste successive. Il giorno successivo, circa mezzo milione di persone si radunano per protestare contro la violenza della polizia e per chiedere nuove elezioni. La massiccia partecipazione intensifica gli scontri, rendendo la situazione sempre più tesa e violenta.

Nei mesi successivi, i manifestanti sono oggetto di una brutale repressione da parte del governo. Con segnalazioni di numerosi casi di torture, arresti arbitrari e incarcerazioni forzate. Le libertà civili pesantemente limitate, e oltre 130 giornalisti sono feriti durante le proteste. Il governo, a seguito della prosecuzione stabile delle proteste promulga leggi draconiane

anti-assembramento e anti-protesta, criminalizzando di fatto qualsiasi forma di dissenso pubblico. Tra la metà e la fine di gennaio, oltre 40 giornalisti e centinaia di manifestanti vengono feriti durante gli scontri con la polizia.

Il 20 febbraio 2014, le proteste raggiungono il culmine a seguito del fallimento nel ripristinare la costituzione precedente. Gli scontri sono sin da subito violenti come non mai con la situazione che peggiora quando i cecchini aprono il fuoco sulla folla. A fine giornata si contano oltre 100 morti e migliaia di feriti. Questo evento tragico segna un punto di svolta nella crisi.

A causa della gravità della situazione, il giorno successivo iniziano le trattative tra il governo e l'opposizione. Il presidente Viktor Yanukovyc rassegna le dimissioni e fugge dal paese, braccato dalla crescente pressione interna e internazionale. Il parlamento ucraino vota per il ripristino della costituzione del 2004 e approva una serie di riforme volte a ripristinare le condizioni democratiche precedenti. Questi cambiamenti hanno segnato un tentativo di riportare stabilità e legittimità al governo ucraino, dopo mesi di caos e conflitto.

1.4 L'annessione della Crimea

A seguito delle proteste di Euromaidan e dunque di una fortissima instabilità politica nelle regioni orientali e meridionali, significativamente più filorusse e maggiormente legate rispetto alla parte occidentale filoeuropea, si vennero a creare tensioni significative McDermott [2016].

La Crimea, penisola strategicamente importante sul Mar Nero, diventa rapidamente il fulcro di queste tensioni. La maggioranza della popolazione in Crimea di grande percentuale di etnia russa e con forti legami culturali e linguistici con la Russia. Le proteste filorusse in Crimea iniziarono quasi immediatamente dopo la caduta di Yanukovych. Queste manifestazioni furono rapidamente supportate dall'arrivo di forze militari non identificate, che la stampa internazionale soprannominò "omini verdi". Questi soldati, che la Russia inizialmente negò di aver inviato, erano in realtà forze speciali russe prive di insegne identificative. Gli "omini verdi", agendo assieme ai gruppi separatisti, presero rapidamente il controllo di edifici governativi e di infrastrutture strategiche.

Il 27 febbraio 2014, uomini armati occupano il parlamento della Crimea e issarono la bandiera russa. E nei giorni successivi ottengono il loro controllo su tutta la penisola, incontrando poca resistenza dalle forze ucraine locali.

Il 6 marzo, il parlamento della Crimea votò per unirsi alla Russia e indisse un referendum per il 16 marzo, considerato illegale dalla comunità occidentale. Il referendum si svolge in un clima di forte intimidazione e senza la presenza di osservatori internazionali. innumerevoli irregolarità, la presenza di militari armati ai seggi, esclusione di molti cittadini ucraine la mancanza di segretezza piagano il voto. Nonostante ciò, le autorità filorusse dichiarano che quasi l'intera popolazione avesse votato con percentuali bulgare riguardo l'annessione.

Il giorno successivo al referendum, il 17 marzo, il parlamento della Crimea dichiara l'indipendenza dall'Ucraina e chiede formalmente l'annessione alla Federazione Russa. La comunità internazionale, in particolare l'Unione Europea e gli Stati Uniti considerano assolutamente illecito il referendum e per tutta risposta impongono sanzioni economiche e diplomatiche particolarmente pesanti alla Russia

1.5 La guerra nel Donbass 2014-2022

Il Donbass, una regione dell'Ucraina orientale, è caratterizzata da una forte presenza di popolazione russofona e da stretti legami economici e culturali con la Russia. In particolare, gli *oblast* (divisioni amministrative) di Luhansk e Donetsk che hanno manifestato un significativo

malcontento rispetto all'orientamento pro-europeo del governo ucraino, in particolar modo dopo le proteste di Euromaidan. Il malcontento diventa rapidamente protesta separatista. Le manifestazioni, supportate dalla Russia portano alla proclamazione delle autoproclamate Repubbliche Popolari di Luhansk (LPR) e Donetsk (DPR) e, con l'aiuto di forze russe non ufficiali, del controllo di edifici governativi e infrastrutture.

A differenza della Crimea, il governo ucraino risponde militarmente. Nell'aprile 2014, lancia un'operazione antiterrorismo per riprendere il controllo della regione. Inizialmente, con successi significativi, riconquistando diverse città e riducendo l'area sotto il controllo separatista.

Nell'agosto 2014, la situazione cambia drasticamente quando la Russia intensifica il suo supporto ai separatisti, fornendo armi pesanti, personale militare e supporto logistico. Questi supporti militari, in particolare portano a due importanti sconfitte per l'Ucraina: la battaglia di Ilovaisk e la perdita di Novoazovsk. Esse costrinsero alla firma del primo Protocollo di Minsk il 5 settembre 2014, che prevedeva un cessate il fuoco immediato e il ritiro delle armi pesanti. Tuttavia, gli scontri continuarono, portando alla firma di un secondo accordo, Minsk II, nel febbraio 2015, grazie all'intercessione dell'UE e in particolare di Francia e Germania.

Minsk II prevedeva un cessate il fuoco più completo, il ritiro delle armi pesanti, lo scambio di prigionieri, riforme costituzionali in Ucraina per garantire maggiore autonomia alle regioni separatiste e il ripristino del controllo ucraino sul confine con la Russia. Nonostante l'accordo, le violazioni del cessate il fuoco continuarono da entrambe le parti.

Tra il 2015 e il 2019, la situazione nel Donbass rimane in uno stato di congelamento delle ostilità, con sporadici scontri e una situazione non troppo sentita dalla comunità internazionale per spingere importanti modifiche allo status quo. Nel 2019, l'elezione di Volodymyr Zelensky come presidente dell'Ucraina porta a nuovi tentativi di risoluzione pacifica. La "Formula Steinmeier", proposta dal presidente tedesco che prevedeva l'organizzazione di elezioni nelle regioni separatiste sotto supervisione internazionale.

La pandemia di COVID-19 nel 2020 complicò ulteriormente la situazione, con differenze nella gestione sanitaria tra le aree controllate dall'Ucraina e quelle sotto il controllo separatista.

Nel 2021, la tensione inizia nuovamente a salire con La Russia che ammassa circa 100.000 truppe vicino al confine ucraino, incluse 80.000 in Crimea e circa 15.000 al confine meridionale. Chiaramente considerato come un tentativo di pressione nei confronti della comunità internazionale. Nel dicembre 2021, il presidente statunitense Joe Biden minaccia severe sanzioni economiche contro la Russia in caso di invasione dell'Ucraina, che aleggiava da diversi mesi. In risposta, la Russia chiede garanzie che l'Ucraina e altri paesi ex-sovietici non sarebbero entrati nella NATO. Nonostante varie dichiarazioni e tentativi diplomatici di summit, la situazione degenera totalmente il 21 febbraio 2022 con l'ufficiale riconoscimento da parte della Duma russa e del presidente Vladimir Putin delle due repubbliche di Donetsk e Luhansk.

1.6 L'invasione russa dell'Ucraina

Il 24 febbraio 2022 la Russia invade l'Ucraina[4] facendo partire "l'operazione militare speciale" per la salvaguardia delle due autoproclamate repubbliche di Luhansk e Donetsk e delle minoranze russofone. Si riporta l'invasione iniziale al 25 Febbraio in Figura 1.1[5],

L'offensiva iniziale è caratterizzata da una rapida avanzata su 3 direttrici principali:

- Direttrice Nord: con truppe russe e bielorusse dalla Bielorussia verso Kiev, con l'obiettivo di accerchiare e conquistare rapidamente la capitale ucraina.

[4] https://en.wikipedia.org/wiki/Timeline_of_the_Russian_invasion_of_Ukraine

[5] https://www.understandingwar.org/backgrounder/russia-ukraine-warning-update-russian-offensive-campaign-assessment-february-25-2022

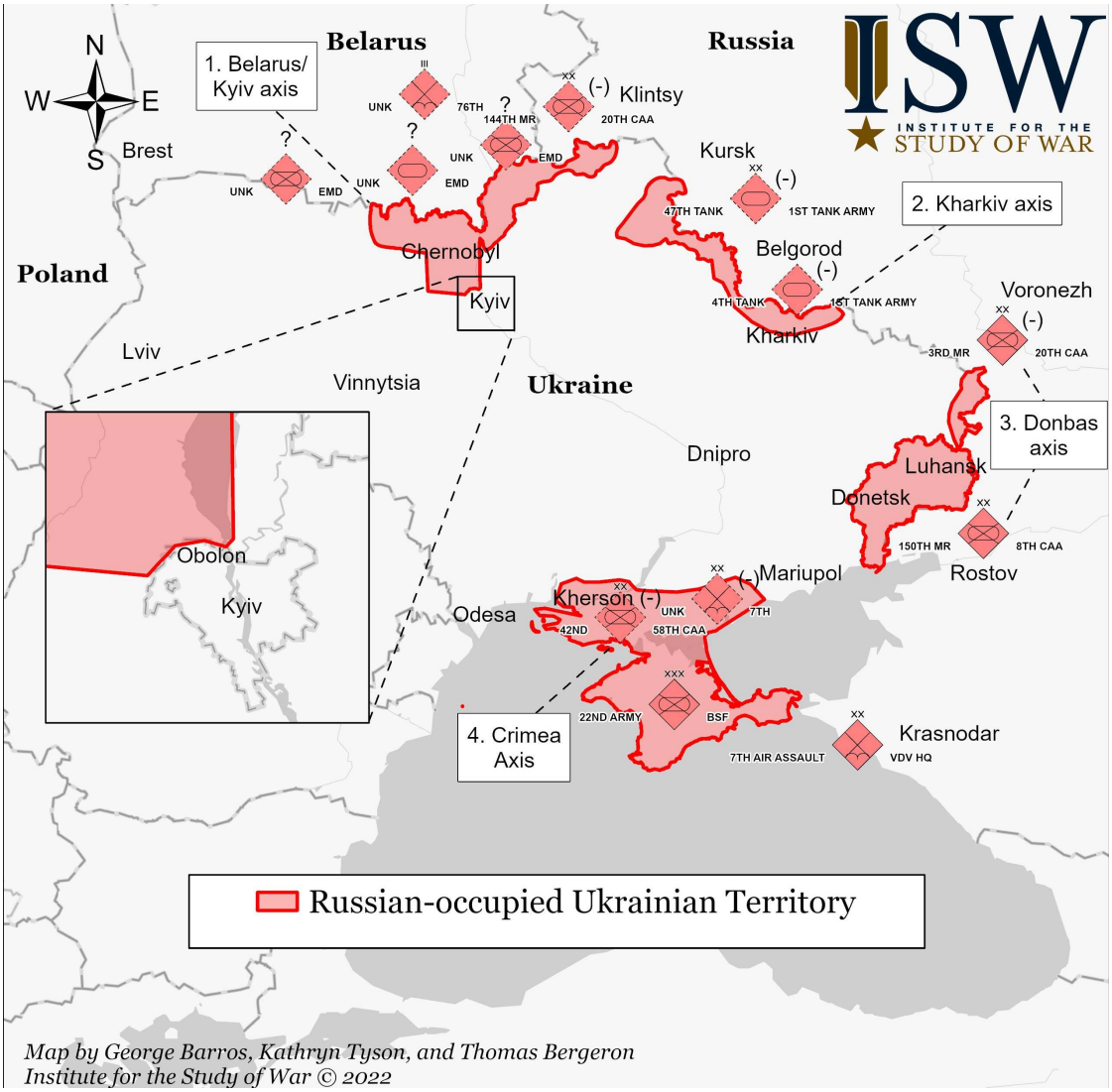

Figura 1.1: L'invasione iniziale, territori attaccati al 25 Febbraio

- Direttrice Nord-Est e Sud-Est: dirette verso Kharkiv, la seconda città più grande dell'Ucraina, e attraverso il Donbass.
- Direttrice Sud: Dalla Crimea occupata, puntando verso Kherson e Mariupol.

Nelle prime ore del 24 febbraio, la Russia lancia una serie di attacchi missilistici e aerei contro obiettivi militari in tutta l'Ucraina, colpendo aeroporti, basi militari, sistemi di difesa aerea e infrastrutture colpendo pesantemente tutte le maggiori città ucraine, tra cui Kiev, Kharkiv, Odessa e Lviv. Il governo ucraino reagisce immediatamente dichiarando la legge marziale e ordinando la mobilitazione generale e la chiamata alle armi per tutti gli uomini. Il presidente Volodymyr Zelensky rifiuta l'offerta di evacuazione degli Stati Uniti, pronunciando la frase diventata simbolo della resistenza ucraina: "Ho bisogno di munizioni, non di un passaggio". [6]

La comunità internazionale, per certi versi già sull'attenti dai giorni precedenti reagisce rapidamente all'invasione. I paesi della NATO, pur non intervenendo direttamente nel conflitto, fornirono armi, equipaggiamenti militari, assistenza finanziaria e umanitaria. L'Unione

[6] "The fight is here; I need ammunition, not a ride"

Europea e gli Stati Uniti, insieme ad altri paesi, accentuano le sanzioni del 2014 aggiungendo l'esclusione di alcune banche russe dal sistema SWIFT e colpendo i patrimoni all'estero di tutta l'intellighenzia russa[7].

Nei primi giorni dell'invasione, le forze russe avanzarono rapidamente su più fronti, incontrando una resistenza molto più forte del previsto. L'esercito ucraino, pur in inferiorità numerica e di equipaggiamento dimostra capacità di combattimento, reazione e di strenua volontà. aiutati inoltre dalla popolazione civile mobilitatasi in massa, con molti cittadini che si uniscono alle unità di difesa territoriale o parteciparono alla resistenza con metodi di guerriglia urbana

Nella direttrice sud, le forze russe ottengono inizialmente maggiori successi, conquistando rapidamente la città di Kherson. Questo permette l'avanzata verso la centrale nucleare di Zaporizhzhia, la più grande d'Europa, creando timori internazionali per la sicurezza dell'impianto.

Un aspetto cruciale del conflitto è che si delinea sin da subito come guerra ibrida condotta dalla Russia, che comprende massicce campagne di disinformazione e propaganda sui media tradizionali e sui social network. Questa strategia, nota come "maskirovka"[8], mira a confondere l'opinione pubblica internazionale, per diminuire il sostegno e a minare il morale ucraino.

Nelle settimane successive, la situazione sul campo si evolve rapidamente. Nonostante le forze russe continuino ad avanzare in alcune aree, l'esercito ucraino riesce a rallentare significativamente il loro progresso e in alcuni casi a contrattaccare. La battaglia per Kiev si trasforma in un lungo assedio, con le forze russe che non riescono a penetrare la città nonostante il notevole dispiego di forze.

Uno degli eventi più drammatici di questa fase è l'assedio di Mariupol, importante porto sul Mar d'Azov. La città venne bombardata per diverse settimane con la popolazione civile all'interno. Si include anche il famigerato attacco all'ospedale pediatrico, ampiamente condannato come crimine di guerra[9][10].

Verso la fine di marzo (Figura 1.2)[11] il piano russo di una rapida conquista della capitale e di un forzoso cambio di regime, risulta essere fallito. Le forze ucraine, sfruttando le linee di rifornimento sovraestese dei russi e la logistica carente per iniziare a riconquistare territorio intorno a Kiev. Il 29 marzo, la Russia annuncia il ritiro delle sue forze dalla regione di Kiev, effettivamente ponendo fine alla prima fase dell'invasione.

L'inizio di aprile vide la liberazione di diverse aree intorno a Kiev, inclusa la città di Bucha. La scoperta di centinaia di civili uccisi a Bucha fa il giro del mondo[12], portando ancora una volta ad accuse di crimini di guerra contro le forze russe, intensificando le sanzioni contro la Russia.

Il fallimento dell'offensiva iniziale russa segna un punto di svolta significativo nella guerra. La Russia è costretta a diminuire i fronti d'attacco e a rivedere gli obiettivi, concentrandosi successivamente sull'espansione del controllo nel Donbass e nel sud dell'Ucraina. Si da così inizio a una nuova fase del conflitto, caratterizzata da una guerra di attrito classica.

[7] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1484

[8] traslitterato dal russo: mascheramento, cioè la dottrina militare che si riferisce alla negazione e allo sviamento per confondere il nemico

[9] https://edition.cnn.com/interactive/2022/03/europe/mariupol-maternity-hospital-attack/index.html

[10] https://www.eeas.europa.eu/eeas/russiaukraine-statement-high-representative-borrell-and-commissioner-lenarcic-violations_en

[11] https://www.understandingwar.org/backgrounder/russian-offensive-campaign-assessment-march-28

[12] https://www.ohchr.org/en/press-releases/2022/12/un-report-details-summary-executions-civilians-russian-troops-northern

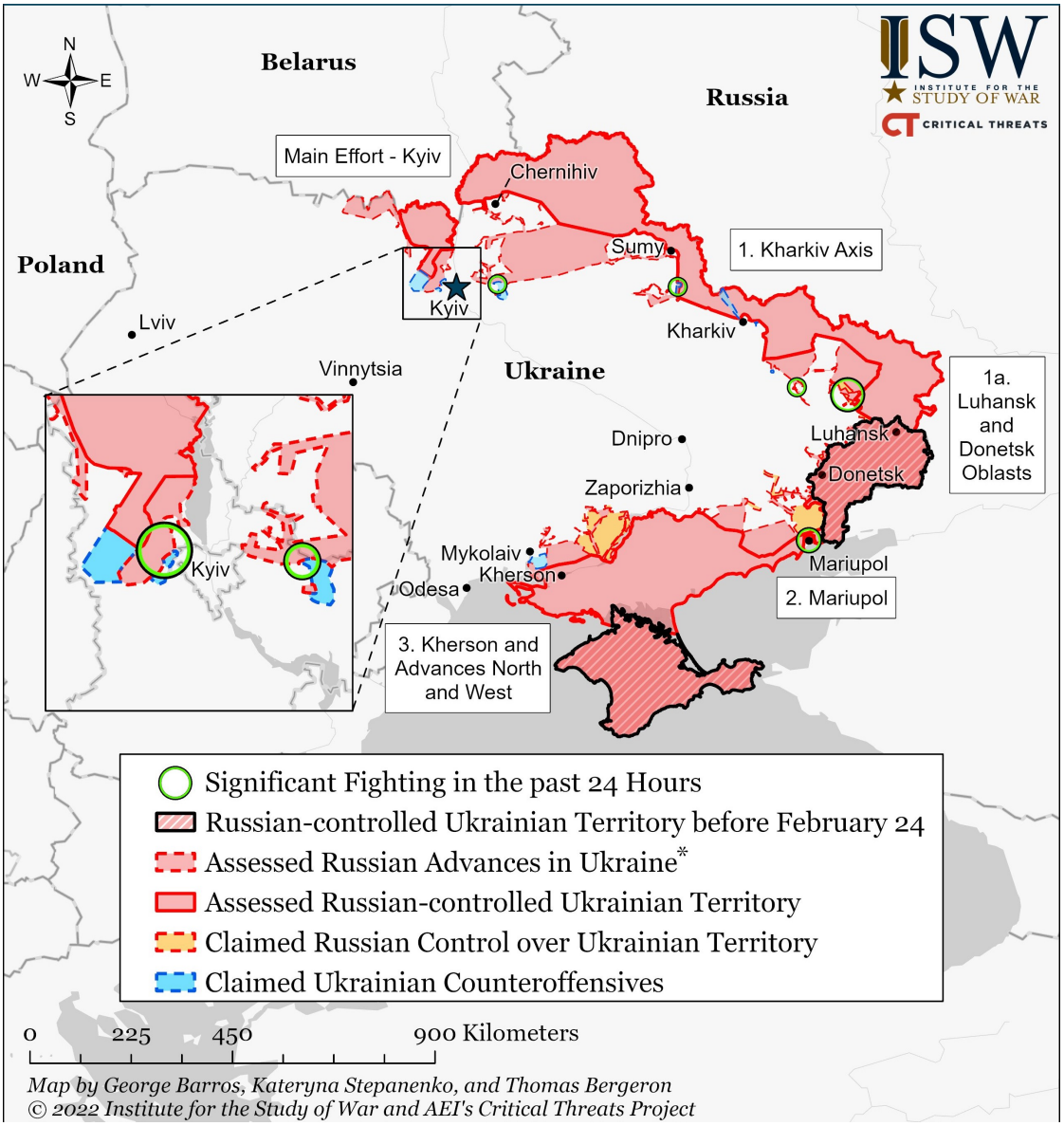

Figura 1.2: Situazione dell'invasione dell'Ucraina a fine marzo 2022

1.7 Il fronte sudorientale 8 aprile - 28 agosto

Ad aprile 2022, il focus del conflitto si spostò verso il sud e l'est dell'Ucraina. L'assedio di Mariupol divenne uno dei punti più critici e drammatici della guerra. La città, strategicamente importante per il suo porto sul Mar d'Azov, viene continuamente bombardata e La situazione umanitaria diventa rapidamente catastrofica, con civili intrappolati senza accesso a cibo, acqua e cure mediche.

Poco fuori dal perimetro della città, l'acciaieria Azovstal diventa simbolo della resistenza ucraina ospitando migliaia di soldati ucraini, principalmente del reggimento Azov, e centinaia di civili. La difesa dell'Azovstal durò per oltre due mesi in condizioni estreme, attirando l'attenzione mondiale e diventando un simbolo della determinazione ucraina.

Nell'est del paese, la Russia intensifica i suoi sforzi per conquistare completamente le regioni di Donetsk e Luhansk. Le città di Severodonetsk e Lysychansk divennero i principali obiettivi su cui vennero lanciati pesanti bombardamenti di artiglieria seguiti da avanzate di unità meccanizzate. In risposta, le forze ucraine iniziano ad adottare tattiche di guerriglia

urbana, rallentando l'avanzata russa e infliggendo pesanti perdite.

Parallelamente alle operazioni militari, si intensificano gli sforzi diplomatici. La Turchia, con il presidente Erdoğan riesce a mediare per la prima volta tra le due fazioni, ospitando negoziati tra le delegazioni russa e ucraina.

L'inizio di maggio vede la caduta di Mariupol, con la resa dell'Azovstal. Questo permette di stabilire un corridoio terrestre dalla Crimea occupata alle regioni separatiste del Donbass, realizzando uno degli obiettivi strategici dell'invasione e ottenendo il totale controllo del Mar d'Azov.

Nello stesso periodo, la Russia consolida il suo controllo su gran parte degli oblast di Luhansk e Donetsk con l'esercito ucraino che si deve ritirare da città di importanza strategica come Izium e Lyman.

A giugno, di fronte all'avanzata russa che aveva portato all'occupazione di circa il 20% del territorio ucraino, gli Stati Uniti decidono di inviare sistemi missilistici avanzati, in particolare gli M142 HIMARS (High Mobility Artillery Rocket System)[13]. Questo segna un punto di svolta significativo poiché, per la prima volta l'Ucraina diventa capace di colpire obiettivi russi a lungo raggio con alta precisione. La Russia reagisce duramente a questa escalation, considerandola una "linea rossa" e intensificando gli attacchi sia sul territorio sia, attraverso attacchi cibernetici, contro infrastrutture occidentali.

L'arrivo degli HIMARS cambia lentamente la dinamica sul campo di battaglia. Sebbene ancora in ritirata, come nel caso di Lysychansk a fine giugno, l'esercito ucraino, complice la guerra d'attrito e i numerosi colpi mandati a segno a depositi di munizioni, centri di comando e linee di rifornimento russe, rallentando significativamente l'avanzata nemica, costringendoli a combattere una guerra d'attrito estremamente sanguinaria.

A Luglio c'è la perdita totale dell'Oblast di Luhansk, confermato dal presidente Zelensky ma anche un aumento significativo degli aiuti occidentali. La Russia, di fronte alle crescenti difficoltà economiche causate dalle sanzioni, inizia a adottare misure di "economia di guerra".

Un importante sviluppo diplomatico fu il riconoscimento dell'Ucraina come membro associato nel programma di interoperabilità della NATO, un passo significativo verso l'integrazione delle forze armate ucraine negli standard dell'alleanza atlantica[14].

Le preoccupazioni per la sicurezza della centrale nucleare di Zaporizhzhia, la più grande d'Europa, diventano sempre più stringenti. L'Agenzia Internazionale per l'Energia Atomica (AIEA)[15] intensifica i suoi sforzi per garantire la sicurezza dell'impianto.

Ad agosto, l'efficacia crescente degli attacchi ucraini a lungo raggio, supportati dalle armi occidentali, cambia significativamente l'equilibrio sul campo. L'Ucraina riesce a colpire obiettivi strategici russi ben oltre la linea del fronte, compromettendo la logistica e il morale russo.

Un importante successo diplomatico viene raggiunto il 18 agosto con il summit trilaterale mediato dalla Turchia[16]. L'accordo più significativo è l'Iniziativa del Mar Nero per il Grano[17], che permette dopo quattro mesi la ripresa delle esportazioni di grano attraverso i porti del Mar Nero. Cruciale per prevenire una potenziale crisi alimentare globale, dato il ruolo chiave di Ucraina e Russia ed esportatori di grano[18]. Il summit inoltre porta anche a progressi nelle discussioni sulla sicurezza della centrale nucleare di Zaporizhzhia e allo scambio di alcuni prigionieri.

[13] https://www.defense.gov/News/News-Stories/Article/Article/3095394/us-provided-himars-effective-in-ukraine/

[14] https://www.nato.int/cps/en/natohq/topics_37750.htm

[15] in inglese IAEA:International Atomic Energy Agency

[16] https://www.un.org/en/black-sea-grain-initiative

[17] Black Sea Grain Initiative

[18] rispettivamente terzi e secondi al mondo

Questi sviluppi, uniti alle perdite sempre più ingenti russe portano a uno stallo relativo nel sud dell'Ucraina e segnarono l'inizio di un cambiamento nella dinamica della guerra. L'Ucraina, rafforzata dal supporto occidentale e dalle nuove capacità militari, riesce finalmente ad avere la possibilità di contrattaccare.

1.8 La controffensiva ucraina: 29 Agosto - 11 novembre 2022

La controffensiva ucraina, iniziata il 29 agosto 2022, segna un punto di svolta cruciale nella guerra. Preparata attraverso mesi di pianificazione e sostenuta da un massiccio afflusso di armi e addestramento occidentali, l'operazione si distingue per l'uso massiccio di droni che svolgono un ruolo chiave, sia per la ricognizione che per gli attacchi di saturazione a basso costo. Fondamentale è l'utilizzo di intelligence fornita dagli Stati Uniti e dal Regno Unito, che include posizioni satellitari dettagliate, informazioni su vulnerabilità e obiettivi strategici, nonché dati sui movimenti delle truppe nemiche. Inizialmente viene dichiarata una campagna nel sud del paese. Data la catena di comando ucraina più corta, la conoscenza maggiore del territorio e un'efficiente linea logistica, nel momento in cui le truppe russe si spostano al sud, parte una rapidissima controffensiva al nord, nell'oblast di Kharkiv (Figura 1.3)[19]. In pochi giorni, le forze ucraine liberano vaste aree, riprendendo il controllo di città strategiche come Izium e Kupiansk. A metà settembre, l'Ucraina ha liberato oltre 6.000 km² di territorio e tutto l'Oblast.

Questo successo porta alle prime critiche aperte all'interno della Russia, con figure come Ramzan Kadyrov, leader della Cecenia, che si espone pubblicamente nel criticare la condotta della guerra. In risposta ai successi ucraini, il 21 settembre la Russia annuncia una mobilitazione parziale, chiamando alle armi circa 300.000 riservisti, principalmente richiamati dalle parti più povere del paese.

Con l'avanzata ucraina, la Russia inizia a utilizzare l'energia come arma, minacciando e poi effettivamente interrompendo le forniture di gas attraverso il gasdotto Nord Stream 1 e 2. Questo aumento delle tensioni viene ulteriormente complicato da una serie di morti misteriose tra alti dirigenti di compagnie energetiche russe, alimentando speculazioni su possibili lotte interne di potere e sul dissenso crescente tra le alte sfere. La situazione si complica ulteriormente con il sabotaggio dei gasdotti Nord Stream, che li rende inutilizzabili per il trasporto del gas.

La scoperta di fosse comuni contenenti oltre 400 cadaveri dopo la liberazione di Izium, insieme a prove di torture, crea un grande turbamento in tutto l'occidente. Queste scoperte portano a un inasprimento delle sanzioni contro la Russia e rafforzano le accuse di crimini di guerra.

La guerra dell'informazione si intensifica, con l'eliminazione di oltre 1.600 account di propaganda russi da Facebook il 27 settembre. La tensione raggiunge nuovi picchi quando la Russia minaccia l'uso di armi nucleari nei territori occupati di Luhansk e Donetsk, a seguito di referendum ampiamente considerati illegittimi dalla comunità internazionale.

Nel frattempo, l'avanzata ucraina continua con la liberazione della città strategica di Lyman a ottobre. Nel sud, l'Ucraina riesce a riconquistare diversi villaggi, avvicinandosi al fiume Dnipro.

Il ponte di Kerch, collegamento vitale tra la Russia e la Crimea occupata, e simbolo della Crimea occupata, diventa un obiettivo strategico raggiungibile che viene colpito rendendolo inutilizzabile. Questo viene seguito da una delle maggiori rappresaglie missilistiche della Russia dall'inizio della guerra.

[19] https://www.understandingwar.org/backgrounder/russian-offensive-campaign-assessment-september-9

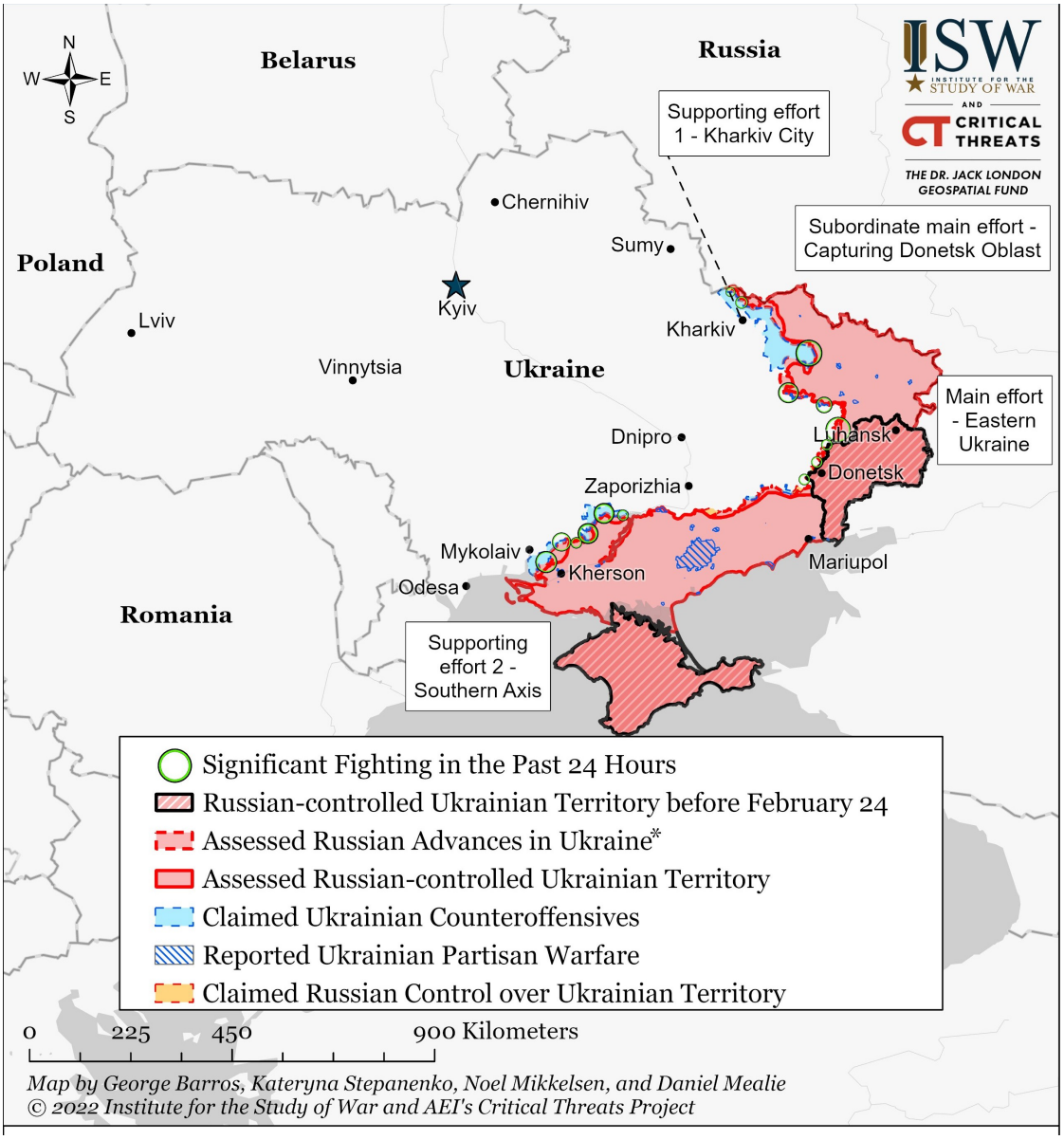

Figura 1.3: Situazione al 9 settembre, dopo la liberazione dell'Oblast di Kharkiv

L'11 novembre segna un momento cruciale: l'Ucraina riesce finalmente a sfondare le difese russe e a riprendere la città di Kherson, costringendo l'esercito russo a ritirarsi oltre il fiume Dnipro. Questo rappresenta una vittoria significativa per l'Ucraina ma anche la fine dello slancio. La controffensiva ucraina ha un impatto significativo sia sulla società russa, che inizia ad accusare pubblicamente le alte sfere dell'esercito e sulla diplomazia internazionale, che inizia a considerare la vittoria ucraina come possibile.

1.9 Lo stallo invernale

Dopo la riconquista di Kherson da parte dell'Ucraina, il fronte si stabilizza lungo il fiume Dnipro, mentre le Le forze russe si attestano sulla riva orientale, fortificando le loro posizioni.

Uno dei momenti di tensione è il 15 novembre quando un missile cade in territorio polacco, provocando due vittime e sollevando timori di un'escalation del conflitto che tuttavia finisce

in un nulla di fatto[20].

La Russia inizia a porre grande interesse negli attacchi missilistici contro le infrastrutture energetiche ucraine per provocare blackout in tutto il paese, aumentando i disagi ed esponendo la popolazione al freddo invernale.

Ad inizio dicembre, il freddo intenso unito alla pioggia e al fango rallenta notevolmente le operazioni su larga scala anche se intorno a Bakhmut nel Donbass i combattimenti non si fermano e la città diviene il fulcro di intensi scontri.

Gli attacchi russi alle infrastrutture ucraine proseguirono, causando gravi problemi di approvvigionamento energetico e idrico in molte città. La comunità internazionale incrementò gli aiuti umanitari per sostenere la popolazione ucraina durante l'inverno.

Il 21 dicembre, Zelensky per la prima volta esce dal paese e incontra a Washington, il presidente americano Joe Biden, creando aspettative per il continuo della guerra, assicurando risorse fondamentali per il paese e contratti di fornitura militari.

L'anno inizia subito con un'intensificazione dei combattimenti intorno a Bakhmut e Soledar. Il gruppo Wagner, una compagnia militare privata russa, inizia ad assumere un ruolo sempre più prominente in questi scontri, essendo meglio equipaggiati, piú pagati e piú esperti dell'esercito regolare.

Verso la fine del mese, diversi paesi occidentali annunciano l'invio di carri armati moderni all'Ucraina, inclusi i Leopard 2 tedeschi e i Challengers inglesi [21], segnando un'importante punto di svolta nel supporto militare.

Il 24 febbraio, primo anniversario dell'invasione, vede manifestazioni di solidarietà con l'Ucraina in tutto il mondo. I combattimenti a Bakhmut raggiungono livelli di intensità senza precedenti, in uno dei piú violenti assedi della guerra moderna, con entrambe le parti che subiscono pesanti perdite. La città nonostante non abbia particolare valore strategico e militare diviene simbolo di resistenza per l'Ucraina e un obiettivo di prestigio per la Russia.

Ad inizio marzo, con il fango che blocca le operazioni il fronte non si muove e la battaglia per Bakhmut continua ad essere il fulcro principale dei combattimenti. Le forze ucraine, pur subendo pesanti perdite, riescono a mantenere il controllo di parti della città.

Il 17 marzo, la Corte Penale Internazionale, in una sentenza particolarmente sentita emette un mandato di arresto per Vladimir Putin per crimini di guerra in Ucraina, in particolare per la deportazione forzata di bambini ucraini in Russia.

Verso la fine del mese, la Russia annuncia il dispiegamento di armi nucleari tattiche in Bielorussia, aumentando le tensioni con l'Occidente.

Ad inizio aprile i combattimenti intorno a Bakhmut proseguono con intensità, ma l'avanzata russa rallenta notevolmente. Nel mentre, le forze ucraine iniziano a effettuare contrattacchi localizzati ai fianchi della città, cercando di chiuderla a tenaglia. Intanto si intensificano le speculazioni su una possibile controffensiva ucraina di primavera, con entrambe le parti che si preparano per potenziali operazioni su larga scala.

Il mese successivo, con i primi caldi si assiste ad un aumento degli attacchi con droni sul territorio russo, inclusi alcuni che colpiscono Mosca.[22] Questi attacchi, sebbene non rivendicati ufficialmente dall'Ucraina, aumentano le tensioni e sollevano questioni sulla vulnerabilità del territorio russo, poiché ufficialmente l'attacco nel territorio russo risulta essere una linea rossa da non superare con tutte le parti che concordano.

Il 20 maggio le forze russe riescono a conquistare gran parte di Bakhmut, ma a un costo umano e materiale enorme.

[20] https://www.esteri.it/it/sala_stampa/archivionotizie/comunicati/2022/11/statement-by-the-high-representative-on-behalf-of-the-eu-on-the-explosion-in-poland/

[21] https://www.nytimes.com/2023/03/27/world/europe/ukraine-challenger-leopard-2-tanks.html

[22] https://www.theguardian.com/world/2023/may/30/moscow-drone-attack-mayor-reports-minor-damage-to-buildings

Il 6 giugno, l'esplosione della diga di Nova Kakhovka sul fiume Dnipro causa inondazioni devastanti e solleva accuse reciproche tra Russia e Ucraina a causa delle gravissime conseguenze umanitarie e strategiche con allagamenti in tutte le zone limitrofe.

Finalmente l'8 giugno l'Ucraina inizia a condurre operazioni offensive diverse aree del fronte, in particolare nelle regioni di Zaporizhzhia e Donetsk, segnalando l'inizio della seconda controffensiva russa.

CAPITOLO 2

Strumenti per l'analisi dei Social Network

In questo capitolo di analizzeranno gli strumenti utilizzato per l'analisi dei social network. Introducendo le basi della social network analysis e la sua importanza. Si introdurrà poi il machine lerning e le sue classificazioni, per spostarsi poi al natural language processing. Qui si farà la distinzione di alcune categorie fondamentali introducendo la sentiment analysis e VADER. Di seguito si introdurranno i trasformers, BERT e BERTopic. Infine una panoramica veloce su Python, Pandas e le librerie di visualizzazione.

2.1 Social Network Analysis

La Social Network Analysis può essere definita come "lo studio delle relazioni umane attraverso la teoria dei grafi". Essa condivide molte similitudini con i metodi statistici ma fa dell'analisi delle connessioni tra individui e gruppi il suo focus primario.

Fino ad una ventina di anni fa, il campo non godeva di particolare popolarità sia perché grandi moli di dati erano impossibili da ottenere sia per le difficoltà di visualizzazione degli stessi sia per la scarsa comprensione dell'utilità della materia. Con l'avvento dei social network tutto è cambiato rapidamente. Oggi, una qualsiasi piattaforma social in un solo giorno produce una quantità incredibile di dati, unito alla possibilità di accesso semplificato attraverso le API, ha dato un impulso senza precedenti al campo della social network analysis.

Il concetto principale della Social Network Analysis è dato dalle relazioni. Le relazioni umane definiscono sia chi siamo sia il modo in cui decidiamo di agire. Le interazioni e i modelli di comportamento giustificano il modo con cui lasciamo traccia nel mondo.

Bisogna fare una prima scrematura sul concetto di relazione. Le relazioni umane non sono discrete, ma sono più assimilabili ad un spettro continuo mentre nell'ambito dei social media abbiamo delle relazione binarie, che possono essere esemplificate dalla dicotomia amico/non amico.

Un utile quantificatore per iniziare a dare un peso ad esse è la frequenza di comunicazione. Esso risulta un eccellente framework con cui lavorare poichè possiede tutte le caratteristiche interessanti per un'analisi come misurabilità, accuratezza del risultato e influenza nelle relazioni. Nonostante essa non sia perfetta, poichè si possono portare molti controesempi, esso risulta un eccellente strumento iniziale per l'analisi.

Un'importante distinzione delle relazioni può essere data dalla simmetria o asimmetria delle stesse. Un chiaro esempio deriva nei casi in cui le relazioni di potere non permettono una perfetta simmetria, come nel caso di un titolare e di un dipendente. Esse sono particolarmente visibili e comuni nel mondo reale. D'altra parte, le relazioni simmetriche possono essere più facilmente trovate sulla rete, come ad esempio la richiesta di amicizia nei social, in cui dato

l'assenso di entrambi la relazione risulta perfettamente simmetrica, cosa rara nel mondo reale e soprattutto difficilmente studiabile.

Una volta identificate le relazioni, l'analisi si concentra su come queste formino le reti. I metodi per l'analisi variano in base alla disciplina e agli obiettivi. Uno degli approcci classici basati sull'econometria o sociologia è quello che corrisponde al concetto di omofilia, ovvero del fatto che persone con i medesimi interessi o stesse età hanno una probabilità maggiore di essere collegati da relazioni di amicizia o amore. Mentre nella social network analysis classica, assunta l'indipendenza degli eventi otteniamo un processo di Poisson, dove ogni evento può essere trattato come se fosse indipendente. Possiamo perciò semplificare di molto l'analisi andando ad ottenere un buon modello della realtà. Chiaramente essa è una semplificazione poiché assumere indipendenza non sempre è possibile e in quei casi l'analisi si complica incredibilmente.

2.1.1 Perché le reti sociali sono così potenti?

Per risponde alla domanda introduciamo un concetto fondamentale che riguarda la forza delle reti sociali e cioè la loro capacità di mantenere e potenziale i "legami deboli". Un legame debole viene definito come una connessione tra persone che richiede un minimo investimento emotivo tra le parti, talvolta addirittura nullo, caratterizzata da bassa frequenza di comunicazione. Essendo legami così economici rispetto a quelli classici, la possibilità di potenziarli e mantenerli risulta essere una caratteristica estremamente importante, poiché i legami deboli in una rete interpersonale reale possono trasmettere e mantenere informazioni a distanze enormi.

Il vantaggio di interagire attraverso legami a bassa intensità emotiva permette di dare opinioni su una vasta quantità di argomenti senza scadere nel conflitto con altri. Inoltre, la natura asincrona della comunicazione di questo tipo permette di mantenere un importante fattore legato alla novità e alla freschezza delle notizie che circolano.

Sebbene l'essere umano possiede un limite alla quantità di persone con le quali può mantenere un qualche tipo di legame, dato dal numero di Dunbar, la natura dei social network e la facilità con cui è possibile mantenere un legame debole, permette di aumentare considerevolmente il numero di persone alle quali si può comunicare.

La comunicazione asincrona, come nel caso di un post su Facebook o un tweet, permette di aumentare la portata dei legami deboli, sia che essi siano di durata momentanea, come nel caso di "meteore" sia nel caso che essi siano mediati attraverso un maggiore investimento emotivo trasformandosi in relazioni più profonde.

Questa tipologia di legami, è molto evidente in contesti sociali speciali come nel caso di rivoluzioni, guerre ed altri eventi ad altissimo impatto sociale.

Una volta ottenuti i dati necessari per l'analisi sulle reti, possiamo introdurre uno dei più importanti concetti: la centralità. Questa misura punta a misurare il potere e l'influenza, caratteristiche individuali, basandosi sulle loro relazioni all'interno della rete. Per ovviare al fatto che il concetto sia piuttosto vago, ci poniamo una domanda: "chi, in questa rete di relazioni, ha più peso?".

Normalmente in una rete esiste una persona che è ampiamente più popolare delle altre. Perciò la prima metrica, che possiamo introdurre, chiamata "grado di centralità"[1], cerca di identificare e di trovare le cosiddette "celebrità". Il "grado del nodo" sarà il numero di connessioni, come ad esempio il numero di follower su X.

La seconda metrica, la "centralità di vicinanza"[2] deriva dalla possibilità di influenzare un network da parte di una celebtrità o "ego" dipende dalla distanza tra lei e il resto del

[1] Degree of centrality in inglese
[2] Closeness centrality

network. La metrica misura la capacità di influenzare e raggiungere persone a grande distanza. Nonostante sia un concetto di primaria importanza, nel comprendere la centralità di un singolo utente è estremamente costosa da calcolare a livello computazionale.

Spesso il risultato di questa metrica differisce notevolmente dalla degree of centrality poichè la capacità di raggiungere molte persone non implica necessariamente la costruzione di relazioni significative. Questo concetto è fondamentale nel comprendere perché in grandi gruppi di persone, spesso coloro tra i più famosi non sono coloro che possono influenzare il network in profondità riguardo certi argomenti.

Una terza metrica è quella della centralità di intermediazione [3] che si basa sull'assunzione che un individuo possa ottenere potere se gestisce e regge un "bottleneck" o collo di bottiglia, permettendo e regolando l'afflusso di informazioni.

Inoltre essa identifica i "boundaries spanners"[4] ovvero persone che permettono di creare legami tra comunità pressoché indipendenti tra di loro. Creare legami tra queste bolle permette di assumere un alto livello di importanza perché particolarmente sensibili a operazioni di pubblicizzazione o collezione di informazioni.

Unendo le metriche insieme otteniamo una notevole capacià di analizzare il network. Tuttavia esistono figure che sfuggono a questo tipo di analisi come i cosiddetti "cardinali grigi" ovvero persone che hanno un immenso potere, sfruttando persone ben connesse e con legami importanti ma risultando loro stessi quasi invisibili. Metodi avanzati scovarli sono variegati, come l'analisi della centralità degli autovettori o algoritmi specifici.

Oltre alla questione dei cardinali grigi abbiamo una problematica alla quale non possiamo dare risposta solo utilizzando gli strumenti sopra citati e cioè: "Cosa spinge le persone ad aggregarsi attorno alle celebrità?"

Per rispondere a questa domanda andiamo ad analizzare le società, cioè gruppi estremamente piccoli di persone.

introduciamo gli "ego networks" come la sottoreti personali centrate su un singolo nodo. Sapere la dimensione di queste tipologie di network risulta essere fondamentale per capire qual è la portata possibile per un'informazione

Altra metrica, per il rispondere migliorare la comprensione delle dinamiche sociali è il coefficiente di clustering e cioè la misura la proporzione degli amici che sono a loro volta amici tra loro. Negli ego network un valore elevato significa un elevato livello di fiducia, mentre un basso valore suggerisce molti ascoltatori passivi e bassi livelli di fiducia su determinati argomenti.

2.1.2 I network bimodali e le loro tipologie

La maggior parte dei dati disponibili nelle reti sociali si presenta sotto forma di network bimodali, un concetto strettamente legato alla dualità delle persone e dei gruppi.

Le idee e le attitudini degli individui, così come le loro connessioni, sono modellate dalla loro presenza o meno in diversi gruppi sociali. Poiché le persone raramente appartengono a un solo gruppo, risulta interessante studiare le similitudini e le differenze tra gli individui attraverso le loro affiliazioni.

I network di affiliazione[5] sono un tipo semplice di network bimodali, dove due persone sono entrambe membri di un gruppo comune. In questi casi, é possibile ipotizzare debolmente che esse si conoscano a priori. Aumentando il numero di gruppi nei quali queste persone condividono la loro presenza, si può considerare una associazione piú forte tra di loro, fino a

[3] betweenness centrality
[4] una traduzione possibile è quella di superatori di confini
[5] attribute network

parlare di una vera identità di gruppo. Questo fenomeno riflette la tendenza delle persone a formare legami più stretti quando condividono piú contesti.

Gli Attribute Network, si basano sul concetto di omofilia, precedentemente citato e cioè l'idea che le persone che condividono interessi siano maggiormente propense ad avere legami tra di loro. Questo principio é ampiamente utilizzato per la costruzione di sistemi di suggerimento delle conoscenze online. Inoltre è anche interessante analizzare i network inversi, poiché nei casi riguardanti ambienti particolarmente divisivi, si possono identificare gruppi contrapposti. Un caso emblematico é quello dei discorsi politici e la possibilità di categorizzare affiliazioni.

2.1.3 La diffusione delle informazioni

Dati i network, come è possibile comprendere cosa si diffonderà? Nonostante una la risposta definitiva sia impossibile, possiamo trovare delle metriche e ragionare su di esse per comprendere i meccanismi di propagazione.

inizialmente ogni elemento, come ad esempio un video, passa attraverso la ego network cioè la rete personale dell'individuo, con un andamento in genere lineare nel numero di interazioni. E alla fine di questo processo, modellato come un processo di Poisson, il numero totale di visualizzazioni è strettamente correlato al numero di connessioni dirette nella rete sociale. la degree centrality.

Tuttavia è possibile che si raggiunga una soglia critica che permette la condivisione ad un livello molto superiore, seguendo un andamento esponenziale fino a raggiungere un punto di saturazione nella rete originale. A qeusto punto entrano in gioco i boundary spanner, individuo che fungono da ponte tra varie comunità, permettendo l'introduzione dei contenuti ad altre platee e continuando la diffusione. Questo processo è alla basa del fenomeno della viralità.

Il raggiungimento della soglia critica rappresenta uno dei più importanti fattori da considerare. Senza di essa il processo di condivisione non si alimenta nel modo corretto e non riesce ad espandersi in nuove reti sociali.

Per stimarla il numero trovato euristicamente è di circa il 7%. Se il 7% dell'audience che è destinataria interagisce allora c'è possibilità che si superi la soglia passando da una diffusione lineare ad una esponenziale.

La massa critica è intrinsecamente legata al costo della partecipazione. Il costo può essere di varia natura, dal denaro al tempo speso. Considerato che il costo, solitamente, rimane costante o addirittura diminuisce grazie alle economie di scala possiamo fare un'importante considerazione. La massa critica può essere considerata come il punto di equilibrio della retta costante del costo intersecata con la curva del vantaggio di partecipazione o di adozione.

Questo semplice modello, ci permette di comprendere le dinamiche della diffusione in vari contesti, dai movimenti sociali alle innovazioni.

Per raggiungere la soglia di massa critica, avere un ottimo contenuto rimane strategia più efficace. Essendo però un concetto vago, possiamo scomporlo in indicatori che ci possono aiutare a valutare la qualità.

- rilevanza, misura quanto il messaggio sia rilevante e pertinente per il target.
- risonanza, indica quanto il contenuto coincida o meno con i valori e le credenze di riferimento dell'utente.
- rigore, se il contenuto è positivo o negativo.
- immediatezza, permette di valutare la facilità assorbimento e se richiede di fare determinate azioni nell'immediato.

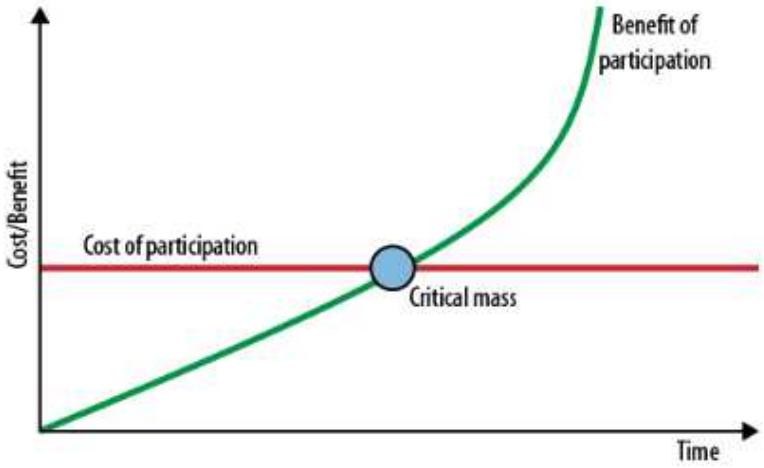

Figura 2.1: La massa critica come punto di equilibrio

- certezza, valuta se la probabilità dell'evento è positiva o negativa e in caso con quale probabilità accada.
- fonti, considera se le fonti sono affidabili o meno.
- valore aggiunto, valuta quanto il contenuto offra informazioni di qualità e che lo distinguono.

La valutazione di questi fattori è spesso molto complessa a causa delle preferenze personali ed eterogenee degli individui. Un modo per semplificare è l'utilizzo di una combinazione affine, in cui la somma dei pesi ha valore unitario e dove i pesi simboleggiano l'importanza di ciascun fattore. L'importanza relativa di questi fattori può anche variare notevolmente sia in base al contesto che al pubblico di riferimento.

Un'ulteriore considerazione da portare è la forma della rete. Nonostante sia chiaro che le informazioni e le notizie possano modificare la forma di una rete l'inverso è vero allo stesso modo. La forma di un network può controllare il modo con cui le informazioni di espandono.

Questo crea un sistema in controreazione dinamico che modifica continuamente la rete, aumentando la complessità della valutazione.

2.1.4 Il concetto di omofilia

L'omofilia, definita come il fatto che persone con i medesimi interessi o importanti similitudini hanno una probabilità maggiore di essere collegati da relazioni, gioca un ruolo fondamentale nell'analisi dei social network.

Possiamo dividerla in due grandi manifestazioni della stessa: Omofilia di status e omofilia di valore.

L'omofilia di status implica che gli individui con caratteristiche simili per estrazione sociale o background, tendono ad associarsi maggiormente tra di loro. Questa forma in particolare è radicata nel tessuto sociale e ammette pochi cambiamenti data la lentezza con cui la scala sociale funziona.

Mentre la seconda, l'omofilia di valore, indica che coloro che pensando similarmente, con credenze o interessi simili possano associarsi indipendentemente dalla classe sociale di riferimento, essendo questo tipo molto volatile e legato alla natura degli interessi e delle tendenze, tende a variare molto rapidamente.

Mentre l'omofilia è una forza sociale importante, non è l'unico fattore nella formazione delle relazioni sociali. Nel caso di persone non troppo simili nè dissimili sia per estrazione sociale che per interessi, un altro fattore entra in gioco. La curiosità o ricerca delle informazioni. Essa può aiutare a spiegare meglio la verosimiglianza delle connessioni anche se varia persona per persona ed è influenzata dalla tendenza personale di ricercare novità.

Possiamo notare come esistono dei limiti all'eccessiva similarità, così come all'eccessiva dissimilarità. la mancanza di nuove informazioni influisce negativamente sull'instaurazione di rapporti significativi. Allo stesso modo, estremizzando le dissimilarità può rendere difficile o impossibile la comunicazione limitando lo scambio di informazioni e la creazione di legami

Questo processo può essere anche inteso temporalmente come la rappresentazione del ciclo di vita. All'inizio la curiosità può soverchiare le tendenze omofile ma se non viene introdotta nuova informazione per troppo a lungo, tende a indebolirlo notevolmente a causa della troppa somiglianza.

L'analisi dei social network risulta di particolare interesse per una grande quantità di ragioni, data la sua complessità e difficoltà. Gli strumenti utilizzati, a partire dai grafi allo sviluppo di algoritmi e di modelli statistici sono di grande rilevanza e versatilità. Inoltre essa non solo permette di affrontare l'ampio spettro di sfide poste dai social ma anche di affrontare altri campi del sapere come l'economia, la psicologia e la politica.

2.2 Machine Learning

Il machine learning è un ramo dell'intelligenza artificiale che si concentra sullo sviluppo di algoritmi e modelli statistici, per lo sviluppo di processi automatizzati per l'estrazione di pattern dai dati. l'approccio è dato dall'idea che i sistemi possano imparare dai dati, identificare pattern e prendere decisioni senza essere stati effettivamente programmati per tutti i possibili scenari.

Abbiamo varie tipologie di approccio:

- Supervised Learning
- Unsupervised Learning
- Deep Learning

2.2.1 Supervised learning

Il supervised learning è uno dei principali approcci del machine learning, è caratterizzato dalll'utilizzo di dataset labeled, ovvero etichettati. l'etichettatura, in inglese labeling di un dataset consiste nell'identificazione di non strutturati e l'aggiunta di etichette per aiutare il modello nel fare previsioni migliori.

L'obiettivo è quello di trovare relazioni tra i dati di input (features) e quelli di output desiderati (labels). L'algoritmo analizza il dataset delle coppie di allenamento per estrapolare e generalizzare le relazioni andando oltre i dati di allenamento.

In genere si ha un divisione in due blocchi; Classificazione e Regressione con la prima che cerca di raggruppare i dati andando a prevedere categorie con applicazioni che spaziano dalla gestione dello spam all'identificazione di frodi.

Mentre nella regressione cerchiamo di trovare un valore numerico come nel caso della previsione dei prezzi o la stima delle vendite di un determinato prodotto.

Infine le tecniche più comuni che possono essere elencate nell'ambito del supervised learning sono:

- regressione lineare e logistica

- alberi decisionali e random forest
- reti neurali
- k-nearest neighbors
- Support Vector machines

2.2.2 Unsupervised learning

L'unsupervised learning, si approccia in modo differente e antitetico rispetto al supervised learning, infatti si usano solo ed esclusivamente dati unlabeled e si cerca di trovare somiglianze e pattern intrinseche nei dati, senza perciò affidarsi all'essere umano, risultando dunque come algoritmi che "imparano da soli". Particolarmente utile nei casi in cui ci siano molti dati, come ad esempio il meteo, in cui il labeling risulterebbe non conveniente per la mole oppure quando non si hanno noti i risultati da ricercare a priori.

Possiamo categorizzare le tecniche utilizzate in 3 grandi famiglie: clustering, association e Dimensionality reduction.

La tecnica più utilizzata è il clustering, che attua un raccoglimento dei dati in base alla loro somiglianza o, ad eventuali, caratteristiche in comune. Largamente utilizzato in un gran numero di applicazioni come l'analisi delle immagini o il riconoscimento delle frodi. A sua volta il clustering si suddivide in svariati sottogruppi, tra i quali possiamo trova lo Hierarchical clustering o Probabilistic Clustering.

Per quanto riguarda l'association, in italiano analisi delle associazioni in cui si utilizza un approccio basato sulle regole di associazione e cioè la ricerca della probabilità che alcuni eventi accadano insieme. L'utilizzo più classico è dato dall'analisi del carrello in cui oltre agli elementi acquistati si consigliano degli oggetti che potrebbero essere correlati.

Infine la Dimensionality reduction, chiamata riduzione della dimensionalità serve per ridurre il numero di variabili in un dataset mantenendo quelle rilevanti, partendo dal presupposto che non tutte abbiano lo stesso peso. Si cerca così di ottenere un miglioramento sia dal punto di vista della semplificazione dei dati sia sull'efficienza.

2.2.3 Deep learning

Il deep learning rappresenta uno dei campi più avanzati e di interesse attualmente. caratterizzato dall'utilizzo di Reti Neurali Artificiali (ANN) composte da molteplici strati, da cui deriva il termine 'deep' o profondo. Questo approccio, inizialmente ispirato al funzionamento del cervello, si è evoluto ed è diventato uno dei più potenti strumenti al giorno d'oggi per affrontare problemi tra i più disparati e con complessità estrema. L'idea di fondo è quella di "permettere ai modelli computazionali che sono composti da diversi strati di apprendere la rappresentazione dei dati con molti livelli di astrazione"[6] e dunque ogni strato permette di riconoscere e analizzare parti sempre più complesse del problema, che è anche ciò che lo differenzia dagli approcci più classici.

L'idea fondamentale dietro il deep learning è quella di permettere ai computer di apprendere rappresentazioni gerarchiche dei dati. Ogni strato della rete neurale impara a riconoscere caratteristiche sempre più astratte e complesse, partendo da elementi semplici nei primi livelli fino a concetti sofisticati negli strati più profondi. Questa capacità di apprendimento automatico distingue il deep learning dagli approcci tradizionali di machine learning, dove le caratteristiche dovevano essere spesso progettate manualmente dagli esperti.

[6] "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" da Deep Learning, Bengio, LeCun, Hinton

Il deep learning ha profondi legami con gli approcci sopra citati sia con quello supervisionato sia con quello non supervisionato. Nel primo caso le reti neurali profonde hanno migliorato ampiamente la classificazione e la regressione, andando ad avere risultati migliori rispetto ai classici algoritmi. Nel secondo caso gli autoencoder hanno migliorato ampiamente la facilità con cui è possibile attuare la riduzione della dimensionalità.

La grande versatile e l'estrema scalabilità sia in base alla quantità di dati sia alla potenza computazionale disponibile, ha permesso di ottenere eccellenti risultati, riuscendo a sfruttare dataset e hardware di dimensioni variabili. tra le architettura quelle più importanti troviamo

- CNN, convolutional neural networks, che hanno avuto un importantissimo ruolo nel riconoscimento delle immagini e nella computer vision in generale
- RNN e LSTM, rispettivamente Recurrent Neural Networks e Long Short Term Memory che hanno apportato ampi miglioramenti nell'analisi del linguaggio naturale, con l'LSTM che si configura come un miglioramento attraverso meccanismi di memoria a lungo termine.
- Attention-Based Architecture, come i Trasformers che hanno rivoluzionato lo scenario negli ultimi anni e che sono attualmente lo stato dell'arte per varie branche come l'elaborazione del linguaggio naturale.

2.3 Natural Language Processing

Il Natural Language Processing (NLP) consiste in una serie di tecniche per rendere accessibile il linguaggio umano ai computer. Negli ultimi anni, la sua efficacia e pervasività sono diventate onnipresenti.

La nascita del campo di studio inizia nel 1950 con i primi tentativi di traduzione automatica, con scarsi risultati. Nei decenni successivi gli approcci basati su regole permettono la nascita dei primi successi come ELIZA, uno dei primi chatbot che tentava di ricreare uno psicoterapeuta. Solo dagli anni 80 però assistiamo all'introduzione dei metodi statistici grazie alla maggior potenza di calcolo e diffusione dei computer. Nel decennio successivo si introducono le prime Reti neurali e le catene di markov.

Solo dopo il primo decennio del duemila abbiamo le prime vere e proprie rivoluzioni come l'introduzione da parte di google del modello Word2Vec e successivamente architetture come Long Short Term Memory. Per arrivare con l'introduzione dei transformer e dei modelli pre-trained come BERT nel 2018 e dell'architettura GPT nello stesso anno.

l'immagine va scalata

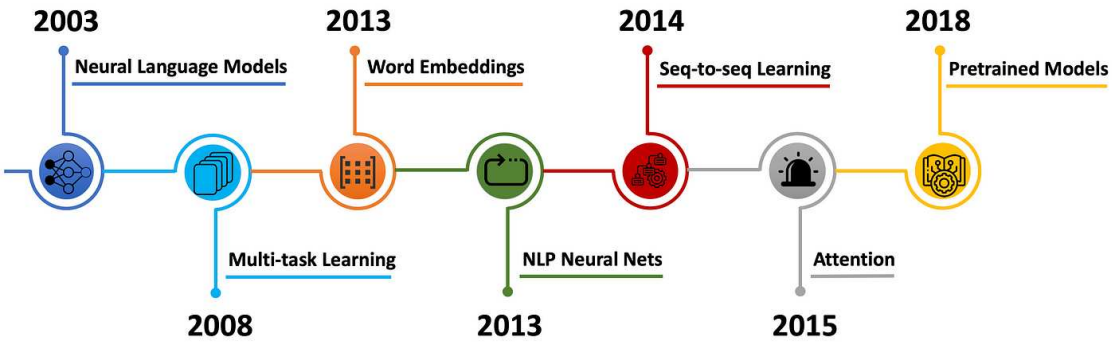

Figura 2.2: Breve storia dei più recenti ed importanti sviluppi dell'NLP

I più importanti campi di utilizzo dell'NLP sono variegati e diffusi come ad esempio le traduzioni in tempo reale nel web, su piattaforme streaming come youTube o nella grande maggioranza dei social media.

Altra importante campo è la classificazione dei testi utilizzata per la gestione dello spam, la categorizzazione delle email e dei contenuti testuali nei social media.

I motori di ricerca hanno beneficiato, anch'essi dei progressi della tecnica diventando sempre più sofisticati e capaci di provvedere risultati più coerenti con le ricerche.

Mentre l'NLP si concentra sullo sviluppo di algoritmi e metodi per rappresentare e processare il linguaggio, la linguistica computazionale si focalizza più sullo studio linguaggio utilizzando strumenti computazionali anche se le due disciplina si intersecano e si influenzano.

Gli approcci moderni all'NLP si basano principalmente su algoritmi di machine learning e, più recentemente, di deep learning. Poichè questi metodi hanno la capacità di gestire grandi moli di dati e affrontare problemi complessi come la traduzione automatica tra lingue diverse, la comprensione del contesto e la generazione di testo.

La peculiarità dell'NLP è data al fatto che affronta sfide particolari dovute alla natura del linguaggio. A differenza di immagini o audio, i testi sono discreti, il che crea difficoltà nell'elaborazione e nella valutazione dei risultati, poiché non è possibile approcciarlo come problemi di natura continua.

Altro problema unico è dato dalla composizionalità del linguaggio che richiede modelli in grado di catturare complessità e sfumature, poiché il significato non deriva solo dall'unione delle lettere in parole, ma anche dal modo in cui queste sono combinate e dalla loro posizione. L'ambiguità intrinseca del linguaggio richiede la capacità di gestire molteplici interpretazioni e basarsi su contesti mutabili.

Una sfida da tenere in considerazione, data la differenza che c'è con la linguistica computazionale è data dal bilanciamento tra l'utilizzo di conoscenze linguistiche pregresse e l'apprendimento dai dati. (forse paragrafo da eliminare, è un po' buttato là)

Molti dei problemi sono classicamente definiti e formulati come problemi di ottimizzazione, dove si cerca di massimizzare un modello rispetto agli input e output.

dove vale in generale la formula:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}(x)} \Psi(x, y; \theta)$$

con x,y rispettivamente input e output, Psi funzione obiettivo o di modello e theta vettore di parametri del modello, solitamente ricavato dai dati.

Tra le varie macroaree possiamo citarne alcune di maggiore interesse:

- Text Classification
- information retrivial
- machine translation
- text generation
- sentiment ed emotion analysis
- document clustering e topic modeling

2.3.1 Text classification

La text classification, è una delle branche classiche dell'NLP. Si contraddistingue per il fatto che si cerca di categorizzare del testo, di svariata natura, in un numero finito di categorie.

Esso viene largamente utilizzato nelle aziende sia per la gestione delle recensioni che per l'assistenza ai clienti, nel giornalismo per ricerca delle fake news e nella scoperta di nuove notizie e in molti altri campi.

Data la grande quantità di tipologia di testi e documenti che vengono generati e creati, l'approccio manuale è ampiamente fuori discussione sia per i costi sia per le tempistiche.

ci sono varie tecniche utilizzate, ovviamente anch'esse hanno subito una rapida evoluzione.

Gli approcci a dizionario e rule-based, nelle quali si creano delle regole basate su pattern e su categorie già specificate ha visto un rapido declino.

Successivamente l'introduzione di tecniche di machine learning ha contribuito a cambiare la visione generale introducendo algoritmi come Naive Bayes, che si basa sull'assumere indipendenza e sull'utilizzo del teorema di bayes, random forest, che si basa sull'utilizzo di molto alberi decisionali(da cui il termine forest)e Support Vector Machine, che ricerca iperpiani ottimali e si comporta estremamente bene quando abbiamo una dimensionalità molto alta.

Ovviamente l'avvento del deep learning il panorama è cambiato e si è adottato sempre di più LSTM, CNN, RNN e in particolare l'architettura basata su trasformer da cui abbiamo BERT che ha rivoluzionato l'ambito.

2.3.2 Information Retrivial

L'information retrivial, si occupa del recupero delle informazioni rilevanti in grandi moli di dati testuali. Il campo ha tra i maggiori utilizzatori i motori di ricerca e i sistemi di raccomandazione automatica.

L'obiettivo che si cerca di ottenere è di fornire a una vasta platea di utenti le informazioni più rilevanti nel minor tempo possibile e dunque non si limita alle parole chiave ma anche all'analisi del contesto e cercare di capire cosa l'utente desideri. In particolare il comprendere e gestire le sfumature del linguaggio è una delle parti più complesse e difficili.

Per quanto riguarda le tecniche inizialmente TF-IDF, Term-Frequency Inverse Document Frequency, nella quale si da più importanza a termini non usuali ma che compaiono maggiormente nell'analisi. Nonostante, assieme a tecniche più sofisticate, risultino ancora largamente utilizzati, negli ultimi anni anche in questa branca il deep learning e l'architettura trasformer ha rivoluzionato l'ambito.

Possiamo anche notare come l'assimilazione di nuove tecnologie abbia reso possibile e sempre più interessante l'approccio multimediale del recupero non solo di testo ma anche di altre tipologie di dati come immagini o video.

2.3.3 Sentiment analysis e VADER

La sentiment analysis, assieme alla emotion analysis, In italiano analisi dei sentimenti e delle emozioni, è un campo di particolare interesse dell'NLP. Esso si basa sull'estrazione di opinioni e sentimenti dai testi, cosa di particolare interesse se legata a grandi moli di dati come quelli che possono provenire dai social media o dalle recensioni di grandi siti come Amazon.

La gestione intelligente di queste analisi risulta essere particolarmente spendibile in ambito aziendale, permettendo di monitorare percezioni dei clienti ed identificare problemi. Inoltre ben si interseca con la politica e il giornalismo nell'analizzare l'opinione pubblica su questioni ad alto impatto e rilevanza, come elezioni e guerre.

Gli strumenti utilizzati sono variegati e se prima si basavano sulle semplici associazioni e regole linguistiche e in seguito con il miglioramento della tecnica gli algoritmi Support Vector Machins e naive bayes hanno dato un grande impulso.

Ovviamente con l'avvento dell'era del deep learning, delle reti neurali e in particolare dell'architettura Trasformer e dei modelli pretrained come BERT che sono diventati rapidamente lo stato dell'arte. Da un lato per l'ampia capacità di catturare sfumature prima non identificabili,Dall'altra per la possibilità del fine-tuning e della capacità di adattamento a vaste aree del sapere.

La grande novità è data anche dalla possibilità di superare la semplice dicotomia tra emozione positiva e negativa ma di andare a ricercare l'emozione in modo più dettagliato, creando uno strumento di fondamentale importanza

Possiamo adesso introdurre VADER, Valence Aware Dictionary and sEntiment Reasoner che si configura come uno strumento basato su regole sintattiche e lessicali per l'analisi del sentiment e delle emozioni, specificamente pensato per l'analisi dei social media.

Si utilizza un dizionario di parole e regole che aiutano a determinare se un dato testo risulta essere positivo, negativo o neutro. Ma oltre a questo si determina anche l'intensità del sentimento attraverso la punteggiatura, che dunque può risultare un discriminante notevole.

Operativamente dalla libreria python NLTK[7], dopo aver importato il 'lexicon' cioè la lista di caratteristiche lessicali, come ad esempio le singole parole, che vengono etichettate in base al loro essere positive o negative.

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
```

Il SIA, sentimentIntensityAnalyzer ci fornisce il polarity score che è un valore che identifica l'intensità dell'emozione analizzata. Vader si basa su 4 valori: pos, neg, neu e compound. Dove i primi 3 si riferiscono ad emozioni positive, negative e neutre. Il compound score invece è quello più interessante ed utilizzato e serve per valutare complessivamente il sentiment, normalizzato tra -1 e +1, con i valori prossimo allo zero implicano neutralità e logicamente valori negativi per sentimenti negativi e positivi per sentimenti positivi. Ricordiamo inoltre che i valori più vicini agli estremi sono particolarmente difficili da ottenere, soprattutto su vasti testi.

Possiamo perciò considerare valori positivi da un valore superiore a 0.05 e al contrario epr i valori negativi.

2.3.4 Topic clustering

La document clustering e topic modeling, serve per estrarre gli argomenti principali da testi, accorpandoli in concetti e cercando di trovare delle tematiche. Ciò permette di prendere documenti di grandezza variabile e sintetizzarli, cercando di non perdere informazioni nel processo. La possibilità di trovare tematiche inoltre risulta estremamente importante laddove abbiamo grandi moli di dati non strutturati e vogliamo cercare di comprendere tendenze e pattern, come ad esempio nei social media.

Gli strumenti sono anche qui molto variegati ma in particolare la libreria BERTopic è stata una tra le più interessanti e spedibili.

2.4 Trasformers

2.4.1 Introduzione all'architettura

Nel 2011, con il celebre paper "Attention is all you need", si assiste all'introduzione del concetto di "Self-Attention". L'attenzione è definita come la capacità di un modello di assegnare pesi differenti a parti diverse dell'input.

[7] Natural Language ToolKit

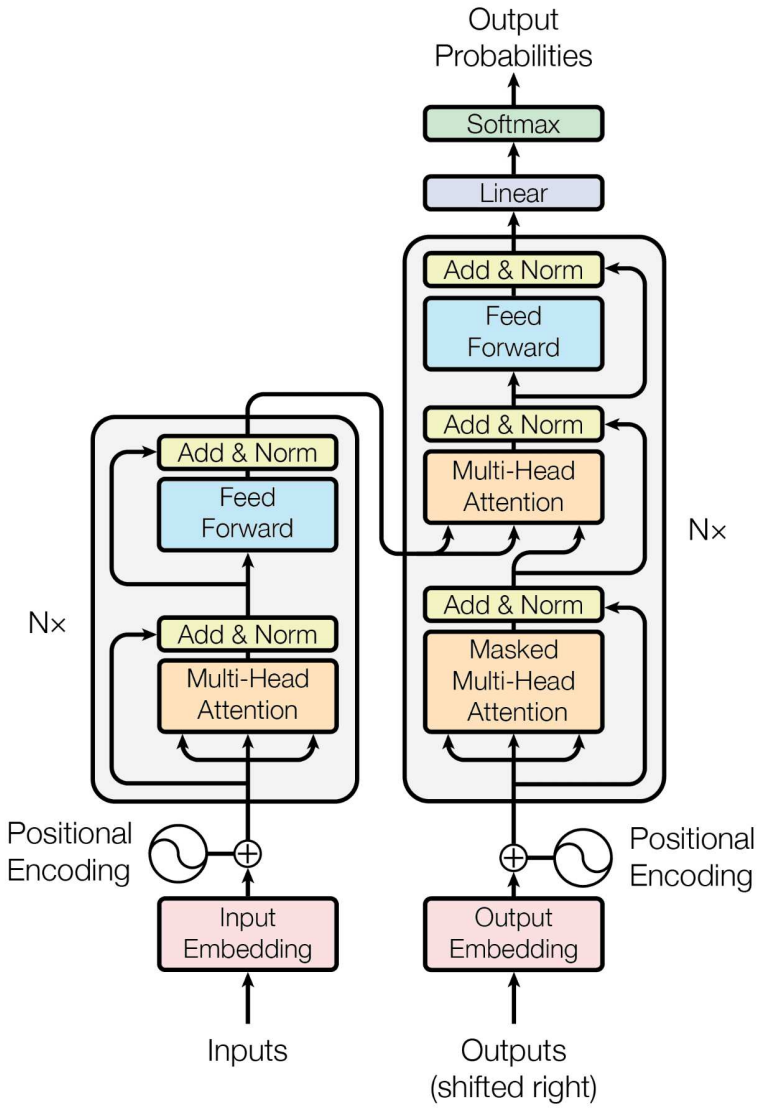

Figura 2.3: Architettura del Transformer

La self-attention è un meccanismo di attenzione che permette al modello di processare una sequenza di input tenendo conto delle relazioni tra tutti gli elementi. riuscendo così a lavorare efficacemente su un grande numero di problemi.

L'architettura è costituita principalmente da due componenti, l'encoder e il decoder.

- l'encoder, riceve una sequenza di input e lo mappa in una sequenza continua. Esso è costituito da più strati identici.
- Il decoder invece prende la sequenza dell'encoder e genera un output utilizzando gli output precedenti assieme all'attention del dell'encoder. Anch'esso è costituito da più strati identici

Matematicamente l'attention può esser definita come il mappare una query (o un'informazione in generale) e una coppia chiave-valore ad un output, con tutti e 4 costituiti da vettori. L'output risultante sarà la somma pesata dei valori. In particolare si calcola il prodotto scalare tra query e chiave e si ottiene un punteggio query-chiave, poi i punteggi

vengono divisi per la dimensione delle chiavi e si applica la funzione softmax che permette di normalizzare i punteggi e assicurare che la somma dei pesi sia unitaria. A questo punto i pesi risultanti vengono usati per calcolare la media pesata.

Ricordando che Q viene dal decoder mentre K e V dall'encoder che rappresenta l'informazione codificata.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Inoltre è importante notare che non viene utilizzata una attention singola ma una "multi-head" attention che è un'estensione e opera in parallelo su più attention indipendenti, permettendo di considerare più aspetti dell'input.

Sin da subito i risultati sono stati particolarmente eccellenti nell'ambito del linguaggio naturale grazie alla possibilità di parallelizzazione dei processi, al contrario delle altre architetture usate nell'elaborazione del linguaggio. Oltre alla possibilità di scalabilità e alla capacità di evidenziare correlazioni anche a grande distanza tra gli input.

2.4.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) è uno dei modelli di linguaggio più celebri basati sull'architettura Transformer. Sviluppato da Google nel 2018, BERT ha avuto profonde ripercussione nell'ambito dell'NLP grazie al fatto di avere encoder bidirezionali.

Per quanto riguarda il funzionamento si utilizza un vocabolario di 30.000 token, convertiti in una rappresentazione distribuita delle parole[8]. La rappresentazione vettoriale permette di memorizzare informazioni e significati. Questi vettori vengono poi fatti passare in diversi strati di Trasformers, che elaborano l'input in modo bidirezionale.

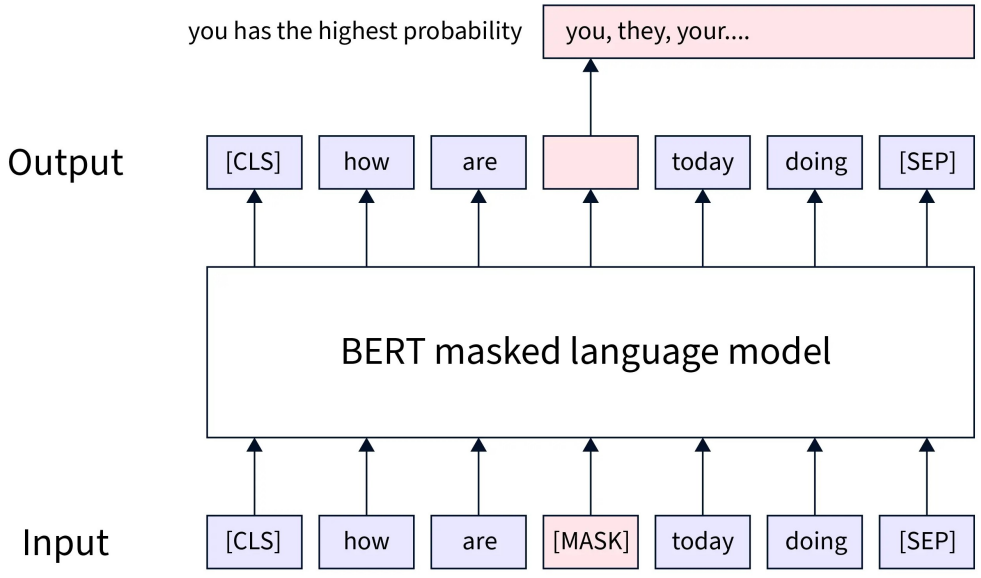

Figura 2.4: Pre-Training BERT model

Abbiamo 2 fasi principali, il pre-training e il fine-tuning.

[8]word embeddings

Durante la fase di pre-trainingl'obiettivo è quello di sfruttare la tecnica del transfer learning, in cui si utilizza un modello addestrato per fargli eseguire compiti in attività che posseggono un certo grado di similarità. Per far questo BERT utilizza la self-supervision, un approccio che permette al modello di generare automaticamente le proprie etichette consentendo l'utilizzo di una mole colossale id dati senza gli svantaggi dell'etichettatura manuale.

Nel pre-training, abbiamo due principali compiti

- Masked Language Model: Nella quale un certo numero di token nell'input vengono mascherati casualmente (mask token), e il modello deve predire questi token mascherati. Questo costringe BERT a sviluppare le associazioni tra parole e contesti e ottenere limitato grado di conoscenza della struttura delle frasi. I token speciali CLS e SEP servono rispettivamente per delimitare l'inizio e fine della sequenza
- Next Sentence Prediction: Il modello deve capire se due frasi sono adiacenti o meno. anche se aiuta limitatamente il miglioramento.

Fine-tuning: Dopo il pre-training, si può passare alla fase di fine-tuning su obiettivi specifici utilizzando dati etichettati. In genere si aggiunge tipicamente un layer output in base all'obiettivo richiesto.

Sin dalla pubblicazione BERT si è assestato rapidamente come il modello allo stato dell'arte, fino all'arrivo dell'architettura GPT, e in particolare ha permesso un incredibile miglioramento nell'ambito dell'elaborazione del linguaggio naturale tra cui in particolare nella sentiment e topic analysis.

In particolare nell'ambito della sentiment analysis, il vettore associato al token CLS viene mappato e utilizzato come rappresentazione per tutta la sequenza. Esso viene trasformato in uno scalare per esere poi processato da una funzione nello strato aggiunto in fase di fine-tuning ottenendo così l'output.

2.4.3 BERTopic

BERTopic rappresenta una tecnica di topic modeling sviluppata da Maarten Grootendorst. Essa si avvale della combinazione dell'architettura Trasformer con sia dell'algoritmo c-TD-IDF. Abbiamo tre fasi principali: l'embedding dei documenti, il clustering degli stessi e infine la generazione dei topic globali.

Nella fase di embedding dei documenti, BERTopic si avvale di Sentence-BERT per convertire i testi in rappresentazioni vettoriali. In questo modo, avvelendosi del modello pre-addestrato, si creano embedding che permettono di rappresentare i documenti in uno spazio multidimensionale. La vicinanza vettoriale implica

Per quanto riguarda la parte di document embedding possiamo dire che "In BERTopic, eseguiamo l'embedding dei documenti per creare rappresentazioni nello spazio vettoriale che possono essere confrontate. Assumiamo che i documenti contenenti lo stesso argomento siano semanticamente simili. Per eseguire la fase di embedding, BERTopic utilizza il framework Sentence-BERT (SBERT) permette agli utenti di convertire frasi e paragrafi in rappresentazioni vettoriali utilizzando modelli linguistici pre-addestrati."

Per il document clustering invece si utilizza UMAP[9] che è un famoso algoritmo di riduzione della dimensionalità. Esso è molto utile per l'efficacia nello scaling su grandi dataset e per preservare la struttura globale dei dati. successivamente si applica HDBSCAN per il raggruppamento di documenti in cluster tematici.

[9]Uniform Manifold Aproximation and Projection for Dimension Reduction

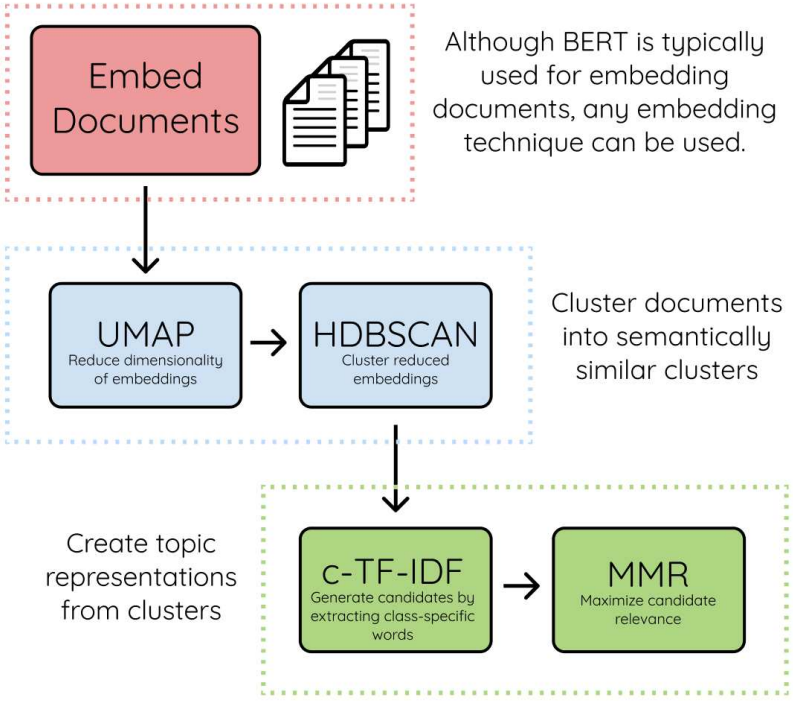

Figura 2.5: Step del funzionamento di BERTopic

Nella generazione dei topi ci si basa su c-TF-IDF, che si distingue dall'algoritmo tradizionale utilizzato nell'ambito dell'information retrival in quanto tratta ogni cluster come un singolo documento anzichè come un insieme di documenti, e successivamente calcola i valori rispetto ai cluster stessi.

introduciamo brevemente c-TD-IDF che è una variante basata sulle classi del Term Frequency-inverse document frequency. così strutturato.

$$\text{c-TF-IDF}_{x,c} = \text{tf}_{x,c} \cdot \log\left(\frac{A}{f_x}\right) \tag{2.4.1}$$

dove abbiamo

- x rappresenta un termine specifico o una parola
- $tf_{x,c}$ è la frequenza della parola x nella classe c
- A è il numero medio di parole per classe
- f_x è la frequenza della parola x tra tutte le classi

in questo modo possiamo ottenere i nostri cluster invece che singoli documenti, permettendoci di generare distribuzioni per ogni cluster.

In questo modo abbiamo l'aggregazione dei documenti per cluster, il calcolo della frequenza dei termini per ogni cluster e otteniamo i termini più rappresentativi per ogni topic.

Questa tipologia di approccio porta a diversi vantaggi, come l'identificazione di termini distintivi per ogni topic. Una cospicua riduzione del rumore, il miglioramento della coerenza dei topic e la facilitazione dell'interpretazione dei topic attraverso parole chiave più significative.

A livello operativo possiamo semplicemente utilizzarlo out-of-the-box semplicemente attraverso:

```
model = BERTopic()
topics, probabilities = model.fit_transform(docs)
```

Una volta eseguito e processate le stringhe, ci vengono restituiti due valori per ciascun elemento: topic e probability

Topic rappresenta l'associazione di ciascun documento del topic designato. Ognugno di essi è contrassegnato da un identificatore che permette di raggruppare i documenti simili. La probability invece, rappresenta la confidenza nell'assegnazione di un documento ad un determinato topic.

A questo punto possiamo effettivamente accedere ai topic. Il primo nella classificazione è il "-1". Esso, è speciale e funge da contenitore per tutti i documenti che non sono stati associati per mancanza di confidenza nei topic primari. Spesso è possibile andare a fare una analisi ulteriore su questi outlier per valutare se essi siano stati categorizzati bene oppure no, trovando contenuti anomali.

Una volta allenato il modello una grande potenzialità di BERTopic è data dalla possibilità di generare delle visualizzazioni coerenti. Una di esse La distanza tra i topic[10] permette di rappresentarli in uno piano cartesiano dove la vicinanza indica similarità tematica. Molto utile per visualizzare cluster correlati semanticamente. Allo stesso modo la possibilità di visualizzare l'evoluzione dei topicin un determinato lasso di tempo permette di trovare trend, picchi temporali o fluttuazioni nell'interesse di determinati argomenti.

2.5 Python per l'analisi dei dati

2.5.1 Introduzione al linguaggio

Python è uno dei linguaggi di programmazione più importanti disponibili attualmente. Classificato come linguaggio ad alto livello, Python si distingue per la sua sintassi chiara e leggibile. Questa caratteristica lo rende particolarmente adatto ad un enorme numero di persone, contribuendo significativamente alla sua estrema popolarità PYthon è un linguaggio interpretato e, sebbene possa comportare una velocità di esecuzione leggermente inferiore rispetto a linguaggi compilati come C++ o Java, offre in cambio una straordinaria versatilità. Tra vantaggi Abbiamo

- la possibilità di ridurre i tempi di sviluppo, evitando lunghe problematiche legate a una sintassi particolarmente restrittiva.
- integrazione con altri linguaggi e sistemi
- alta espressività, il codice risulta facilmente leggibile e servono meno righe di codice per
- linguaggio dinamico, in cui non abbiamo problemi di tipizzazione forte come nel C++

Grazie a queste caratteristiche python risulta essere la scelta preferenziale per un'ampia gamma di settori che spaziano dalla ricerca ad ogni tipo di industria. Attualmente, si stima che Python occupi il terzo posto assoluto tra i linguaggi di programmazione più utilizzati, e in senso stretto secondo tra i linguaggi di programmazione superato solo da JavaScript. Questo successo si estende anche al mondo delle startup: Python risulta essere il linguaggio più utilizzato tra le cosiddette "Unicorn companies"[11]

La bassa barriera d'ingresso è un altro fattore chiave del successo di Python. Questa accessibilità ha reso Python lo standard de facto in molti ambiti di ricerca, sostituendo

[10] intertopic distance

[11] le startup con una valutazione superiore al miliardo di dollari.

linguaggi più complessi che richiedevano competenze informatiche avanzate sin dall'inizio. Ciò ha permesso un ampio accesso alla programmazione per ricercatori, scienziati e analisti senza dover investire troppo tempo nell'apprendimento di concetti di programmazione avanzati.

Uno dei maggiori pregi oltre a quelli elencati in precedenza risulta essere la comunità di programmatori e il suo ecosistema in generale. La disponibilità di un numero elevatissimo di librerie e framework specializzati facilita enormemente lo sviluppo in vari campi:

- Data Science e Machine Learning: Librerie come NumPy, Pandas, Scikit-learn e TensorFlow hanno reso Python il linguaggio di riferimento per l'analisi dei dati e l'intelligenza artificiale.
- Sviluppo web: Framework come Django e Flask permettono lo sviluppo di applicazioni web scalabili e si facile manutenzione.
- Grafica e Visualizzazione: Librerie come Matplotlib e Seaborn offrono potenti strumenti per visualizzare dati.
- Internet of Things (IoT): ampiamente utilizzato per la programmazione di dispositivi embedded e IoT.

2.5.2 Data Cleaning tramite pandas

Pandas è una delle librerie più celebri ed utilizzate di Python che fornisce strumneti per la manipolazione e la gestione dei dati ed è considerata la libreria di base per tutto ciò che riguarda il mondo del Data Science.

In particolare introduce due oggetti, le Series e i Dataframe che possono essere considerati come la trasposizione di vettori e matrici.

Le series sono definite come array monodimensionali che possono contenere un qualsiasi tipo di dato, sono etichettati e vengono rappresentati come una colonna. I Dataframe, invece, sono assimilabili a matrici e dunque sono bidimensionali e pensate come una collezione di series.

Inoltre Pandas permette la gestione di valori nulli, classica problematica ardua da risolvere, le funzioni per la gestione dei dataframe, il groupby per il raggruppamento, operazioni di filtraggio e operazioni di ordinamento, strumenti di I/O e funzioni statistiche.

Per questi motivi risulta essere indispensabile anche data la sua velocità nella gestione di grandi moli di dati, con tutte le strutture dati ottimizzate e una grandissima integrazione con altri linguaggi e tipologie di file.

2.5.3 Data Visualization tramite Seaborn, Plotly, Matplotlib

Le librerie di visualizzazione sono un importante tassello nell'ecosistema di Python. Seaborn, Plotly e Matplotlib sono sicuramente le 3 più famose ed importanti.

Matplotlib, la più antica, permette la modifica di ogni aspetto del grafico, a discapito di una difficoltà maggiore. Inoltre è ampiamente utilizzata nell'ambito della ricerca e della pubblicazione scientifica data l'ampia libertà che offre.

Seaborn è una libreria costruita sopra matplotlib, essa si concentra soprattutto sull'estetica dei grafici e delle visualizzazioni. Inoltre si concentra nell'ambito statistico permettendo di creare grafici come heatmap e grafici di distribuzione congiunta molto facilmente. Inoltre un grandissimo vantaggio è dato dal profondo legame con Pandas, in particolare nell'ambito della visualizzazione dei dataframe.

Infine Plotly, essa è una libreria più recente, che a differenza delle altre due si concentra principalmente sulla visualizzazione interattiva e tridimensionale. Permettendo la lettura

semplificata e sopportando l'integrazione con il web e con le dashboard risulta uno strumento indispensabile per visualizzare grafici con il quale l'utente può interagire liberamente.

CAPITOLO 3

Analisi di X sul conflitto tra Russia e Ucraina

Iniziamo con la descrizione del dataset, a seguire una panoramica sulla discussione su X. A seguire l'analisi pre, durante e post di un evento importante sulla controffensiva e poi di un evento altrettanto importante sullo stallo, Per il secondo avremo delle modifiche sia a livello temporale che metodologico. Infine traiamo delle conclusioni dai dati raccolti.

3.1 Analisi esplorativa del dataset

3.1.1 Storia del dataset

Il dataset utilizzato é stato trovato su Kaggle, community sul Machine Learning e Data science sotto il nome di "Ukraine conflict twitter dataset"[1]. Esso consiste in una raccolta di tweet, eliminati i duplicati, che spaziano dal 22 febbraio 2022 al 13 giugno 2023.

Il dataset ha subito diverse variazioni a partire dalla sua costituzione. Nasce come un progetto limitato a pochi giorni dato l'incredibile interesse e risonanza mediatica dell'evento e non era considerato come un progetto permanente, chiaramente verificabile attraverso le modifiche repentine e rapide che ha sostenuto nei primi mesi di vita.

A causa del largo seguito, quasi 80 mila visualizzazioni e 15 mila download ad oggi, il progetto inizia ad espandersi e ad assettarsi su degli standard buoni per l'eventuale analisi. Elenchiamo ora alcune delle modifiche più importanti:

- Tra fine Luglio 2022 e inizio Agosto 2022 c'è stata una modifica del web crawler[2] per diminuire il rumore riguardante i tweet non correlati, poiché inizialmente il numero di tweet estratti risultava essere eccessivo, questo ha permesso di evitare di ottenere troppi dati non correlati.
- 9 Agosto, a causa di un considerevole aumento degli hashtag **#ukraine** con argomenti non correlati é stato modificato ulteriormente il metodo di estrazione interessandosi principalmente ai tweet originali e non retwittati. Questo ha migliorato considerevolmente la qualità dei post.
- 17-19 agosto 2022, a causa di un bug di ReactJS, un terzo del dataset é stato eliminato e solo successivamente ripristinato. Questo ha portato alla perdita di alcuni tweet.

[1] https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows
[2] Un bot che naviga ed ottiene documenti testuali in una pagina web.

Altre importanti date che sanciscono modifiche consistenti al dataset e alla possibilità di analisi sono date da questi due eventi. Essi vanno presi in considerazione per evitare di trarre conclusioni affrettate. Come si vedrà infatti nel periodo di tempo coerente con queste modifiche, la natura del dataset cambia a livello quantitativo, migliorando la possibilità di analisi ma inficiando in parte l'analisi dei dati. Per ovviare al problema bisognerà tenere in considerazione alcune motivazioni di carattere meteorologico e bellico. Di seguiti alcune problematiche da tenere in considerazione:

- il 2 febbraio a causa di un problema con il web crawler la quantità di tweet ottenuti diminuisce. Questo problema verrà risolto il 26 febbraio aggiungendo inoltre i retweet.
- Tra l'8 e il 19 aprile 2023 a causa di un'ondata di ban, 6 degli account da sviluppatore del curatore del dataset vengono sospesi permanentemente portando ad un minore afflusso di tweet.

Il periodo di interesse analizzato risulta essere dal 29 Agosto 2022 al 7 giugno 2023. Esso è ben documentato e privo di grandi problematiche. Inoltre è facilmente caratterizzabile attraverso due principali blocchi.

Il primo blocco, che va dal 29 agosto 2022 all'11 novembre 2022 è quello della controffensiva ucraina. Il secondo invece riguarda lo stallo generale della guerra che va dal 12 novembre 2022 al 7 giugno 2023.

Questa divisione, soprattutto riguardo la parte finale è, in parte, forzata dall'acquisto di Twitter, e del conseguente cambio di nome in X, da parte dell'imprenditore Elon Musk.

Il cambio di gestione ha provocato una modifica sulle politiche delle API e degli account sviluppatore. La possibilità di ottenere dati è passata ad essere a pagamento, andando ad inficiare progetti di natura individuale. Il dataset in questione cessa la sua attività di aggiornamento il 16 giugno 2023. Poco dopo quella che gli storici classificano come la fine dello stallo.

La decisione di non utilizzare la prima parte del dataset è data dalle problematiche citate precedentemente e alla difficoltà di analisi in primo luogo. Infatti, il processamento di moli di dati così importanti richiede infatti potenza computazionale molto più elevata di quella disponibile.

3.1.2 Descrizione del dataset e statistiche

Il dataset è diviso in file CSV ed è aggiornato giorno per giorno. Per agevolare l'analisi, è stata effettuata un'unione dei file tramite pandas e attraverso la funzione concat, disponibile dalla libreria (Algoritmo 3.1).

```
for filename in files_to_combine:
    dfs.append(pd.read_csv(filename))
combined_df = pd.concat(dfs, ignore_index=True)
```

Listing 3.1: codice per unire i file csv

La variabile file_to_combine serve per elencare tutti i file giornalieri da inserire. La suddivisione ha seguito quella spiegata nel paragrafo precedente. Inoltre, abbiamo avuto altre sottodivisioni, rispettivamente di 7, 3 e 1 giorno utilizzate, principalmente, per l'analisi con BERTopic.

Il dataset possiede 29 colonne ma non tutte sono state utili nell'analisi. Andiamo a descrivere brevemente le colonne interessanti per l'analisi:

- *userid*, indica l'identificativo numerico dato da twitter ad un determinato account.

- *username*, rappresenta il nome dell'utente.
- *acctdesc*, contiene descrizione della pagina home dell'utente.
- *location*, indica illuogo da dove gli utenti tweettano, composta dal nome di città e stato.
- *following*, rappresenta numero di persone seguite da un determinato utente.
- *followers*, rappresenta numero di persone che seguono un determinato utente.
- *totaltweets*, rappresenta numero di tweet totali fatti dall'utente.
- *usercreatedts*, indica quando è stato creato l'account.
- *tweetid*, codice numerico legato al tweet specifico in questione
- *tweetcreatedts*, permette di sapere quando è stato creato un determinato tweet, in formato DATE con data e timestamp
- *retweetcount*, indica quante volte uno specifico tweet è stato retweettato
- *text*, il testo del tweet, comprende anche gli hashtags e le menzioni.
- *hashtags*, contiene gli hashtag correlati allo specifico tweet.
- *language*, rappresenta la lingua del tweet.
- *coordinates*, contiene le coordinate sotto forma di dizionario.
- *favorite_count*, indica il numero di volte in cui un tweet è stato messo nei preferiti.
- *is_retweet*, contiene un valore booleano che indica se il tweet è stato retwettato o meno.

Tra le varie colonne, alcune di esse hanno delle problematiche legate ai valori nulli o non definiti, tra le quali le più colpite in particolare sono:

- *location* è una colonna particolarmente colpita dalla presenza di valori nulli e con aggiornamenti spesso parziali. In particolare essa non è ben formattata poiché a volta si indica la città e la nazione, altre volte solo lo stato, altre volte solo la città. La quantità di valori nulli si attesta sul 33.78% per la prima parte riguardante la controffensiva e per il 40% nella parte dello stallo.
- *coordinates* contiene molti valori nulli. Per quanto riguarda la parte del dataset sullo stallo siamo di fronte a 14739879 di valori nulli e cioè circa 99% e 4314749 nella parte della controffensiva per un totale del 99.6%.
- *is_retweet* è anch'essa una colonna problematica a causa della situazione iniziale del dataset e delle problematiche precedentemente menzionate.
- *languages* possiede un 11% di lingue non definite.

Il totale dei tweet risulta essere, per quanto riguarda la controffensiva ucraina, pari a 4329940 tweet. La seconda parte del dataset possiede le stesse colonne per un totale di 14776078 di tweet. Il che porta ad un totale combinato di 19106018 tweet.

Dopo l'unione dei file il passo successivo è stato quello del filtraggio per lingua inglese. Nonostante ci siano 64 lingue nel dataset, la lingua inglese è maggioritaria (49.53% del totale).

Al secondo posto, troviamo un 11% di lingua non definita e a seguire tedesco e il francesce con rispettivamente il 6.8 e il 5.8%. L'italiano è al quinto posto con il 4.55%.

Dobbiamo considerare che la lingua russa, al 2% è viziata dal fatto che twitter non viene utilizzato nel paese e principalmente solo le persone più istruite o che vivono all'estero lo utilizzano.

3.1.3 La problematica dei bot

Andando ad eseguire le prime analisi si é notata la problematica legata alla presenza di bot. Nonostante molti di essi siano difficili da notare, alcuni di essi sono facilmente osservabili.

Attraverso l'analisi degli utenti più attivi per giorni, l'account "FuckPutinBot" è stato oggetto di considerazioni. Da un lato l'incredibile mole di tweet automatizzati postati ogni giorno (in ogni lingua dell'Unione Europea) ha inquinato il corretto svolgimento delle analisi (come ad esempio la valutazione degli utenti più attivi), dall'altro l'ovvio bias che esso rappresenta. L'utente in questione da solo costituisce il 2% dell'intero dataset, cosa che ci ha costretto ad eliminarlo dall'analisi a monte.

Ci sono stati altri due casi che hanno richiesto la rimozioni di bot durante l'analisi, in particolare nell'analisi dell'attività degli utenti giorno per giorno.

Infine notiamo come quasi nessuno di questi bot sia stato sospesa al giorno attuale, nonostante l'elevata mole di tweet.

3.2 Analisi della Controffensiva Ucraina

Nel periodo che intercorre tra il 29 agosto e l'11 novembre fronte dei poco più di 4 milioni di tweet, abbiamo avuto una quantità di utenti che hanno twettato pari a 634367. Andiamo a ricercare i primi 20 utenti per numero di tweet in Tabella 3.1.

Username	Numero di Tweet
ArvadaRadio	13576
rogue_corq	12343
knittingknots	8834
Ludmila_Volkov	7934
IdeallyaNews	7692
wRLMyxE7ZEr5CKr	7306
_Thirunarayan1	7091
UkraineAlert	6474
Elpoliticonews	5940
otfyxXWzpwdfTra	5732

Tabella 3.1: Username e numero di tweet degli utenti più attivi nel periodo tra il 29 Agosto e l'11 novembre

Possiamo adesso dare un primo sguardo per valutare a quale tipologia di utenti ci troviamo di fronte.

- La prima tipologia è data da giornali, radio e media di informazione in generale. A questa categoria appartengono una maggior parte degli utenti che hanno twettato. Questo è chiaramente dettato dal fatto di portare informazione, in particolare ArvadaRadio, IdeallyaNews, UkraineAlert, Elpoliticonews sono mezzi di informazione che provengono da diverse parti del globo.
- Persone interessate, analisti militari e propagandisti di ambo le fazioni. La seconda categoria è sicuramente quella più interessante con utenti quali rogue_corq, analista militare pro-ucraina, kardinal691, ludmila_volkov e knittingknots utenti particolarmente attivi, principalmente pro-ucraina.
- tutti gli altri account sono principalmente account di aziende, siti web e aggregatori generici di notizie.

Andiamo adesso a valutare la distribuzione del numero di utenti unici nel periodo considerato (Figura 3.1, essa è una statistica interessante per avere un primo approccio ai momenti di maggiore importanza.

Notiamo che la distribuzione possiede molti punti di massimo locale e con una tendenza all'aumento degli utenti man mano che si va verso il massimo assoluto, per poi osservare avere una diminuzione del numero di utenti.

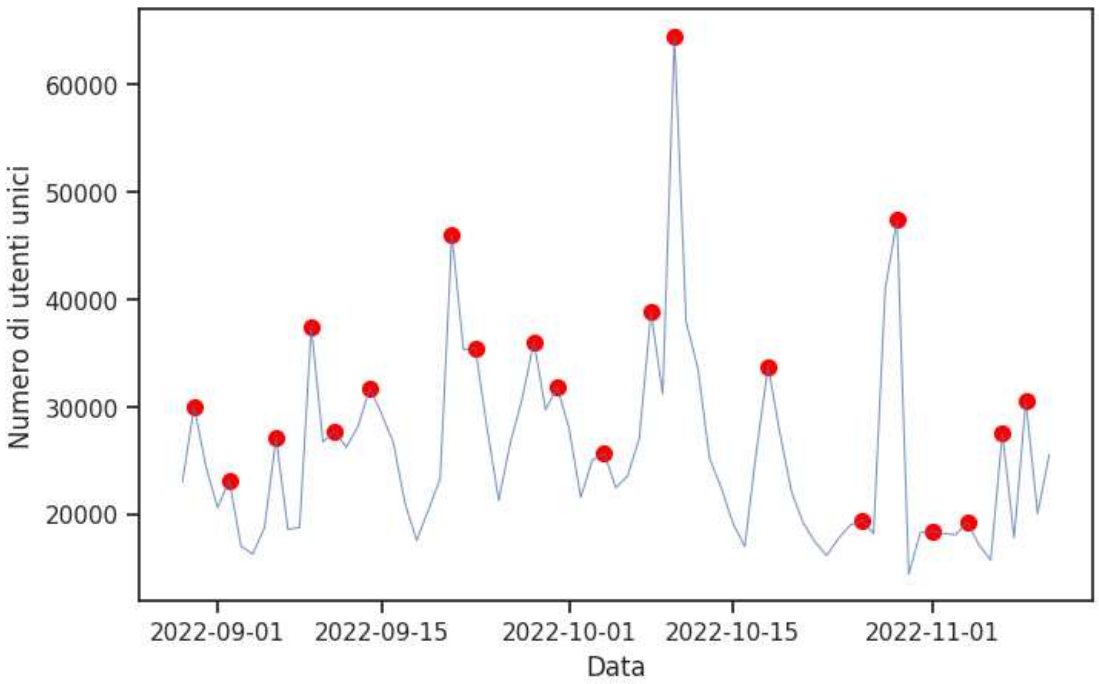

Figura 3.1: Utenti unici giornalieri durante il periodo della controffensiva ucraina

Dal grafico possiamo ottenere i picchi relativi con le date. Ordinandoli dal maggiore al minore e limitandoci ai primi 10, si ottiene Tabella 3.2.

Date	Numero di utenti
2022-10-10	64538
2022-10-29	47398
2022-09-21	45999
2022-10-08	38829
2022-09-09	37463
2022-09-28	36020
2022-09-23	35473
2022-10-18	33740
2022-09-14	31726
2022-09-30	31814

Tabella 3.2: Prime 10 date con il maggior numero di Utenti unici

Osservando la Tabella 3.2, possiamo andare a verificare se i punti di picco corrispondano o meno ad eventi importanti, in particolare:

- 10 ottobre: in risposta all'attacco al ponte dell'8 ottobre, la Russia lancia un pesantissimo attacco missilistico generalizzato su tutto il territorio ucraino, provocando decine di

morti e centinaia di feriti, oltre a gravi disservizi della rete elettrica.

- 29 ottobre: la base navale di Sebastopli viene attaccata in un'operazione di sabotaggio ucraina. Il risultato è il danneggiamento della nave ammiraglia russa e la distruzione di diversi UAV. In risposta a questo evento la Russia esce dal trattato sull'export di grano.
- 21 Settembre: in un discorso preregistrato, il presidente russo Vladimir Putin annuncia l'inizio della mobilitazione parziale.
- 8 ottobre: un'esplosione sul ponte di Crimea causa un collasso dello stesso e la morte di 3 persone.

Il numero di utenti unici può essere sovrapposto al numero di tweet totali. Questo ci aiuta nel trovare giorni che sono stati più interessanti. Presumendo che ad un maggior numero di tweet corrisponda un maggior numero di interazioni. E andare a valutare se sia possibile trovare altri eventi rilevanti. Anche in questo contesto andiamo a valutare quali eventi risaltino maggiormente. Da Figura 3.2, notiamo che dalla sovrapposizione tra i due grafici, si nota subito un andamento estremamente simile tra i due. In particolare per quanto concerne gli eventi associati ai massimi dei tweet tre di essi sono particolarmente notabili e sono coerenti con gli eventi sopracitati.

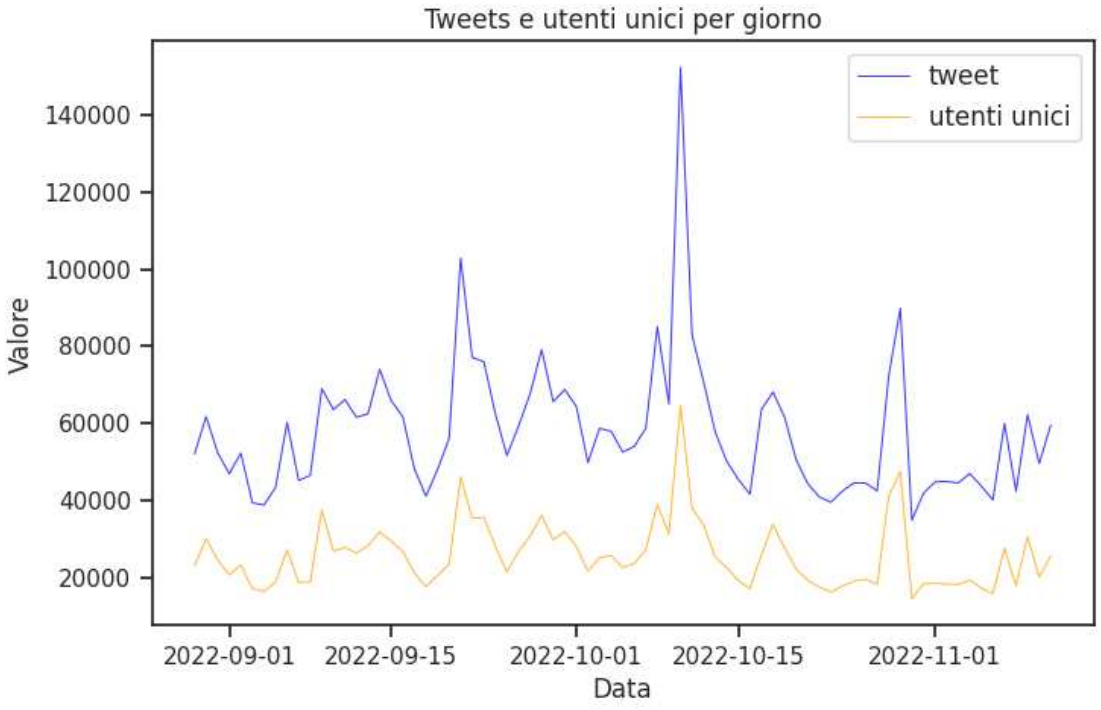

Figura 3.2: Utenti unici e tweet totali giornalieri durante il periodo della controffensiva

Altra metrica interessante può essere quella di verificare quali siano stati gli utenti più attivi giorno per giorno (Figura 3.3). Questa analisi ci serve sia per verificare se esistono degli utenti che sono dei bot, in caso di valori estremamente elevati, che per valutare la tipologia di utenti che è capace di postare così tanto.

Da Figura 3.3, notiamo un andamento irregolare dato da un particolare utente "HydroU-nofficial", con picchi 3 volte superiori a quelli del terzo per un totale di più di 2000 tweet in un singolo giorno. Andando ad ispezionare i suoi tweet si nota come esso sia un account di pubblica utilità nell'annuncio di disservizi della rete idroelettrica del territorio del Canada.

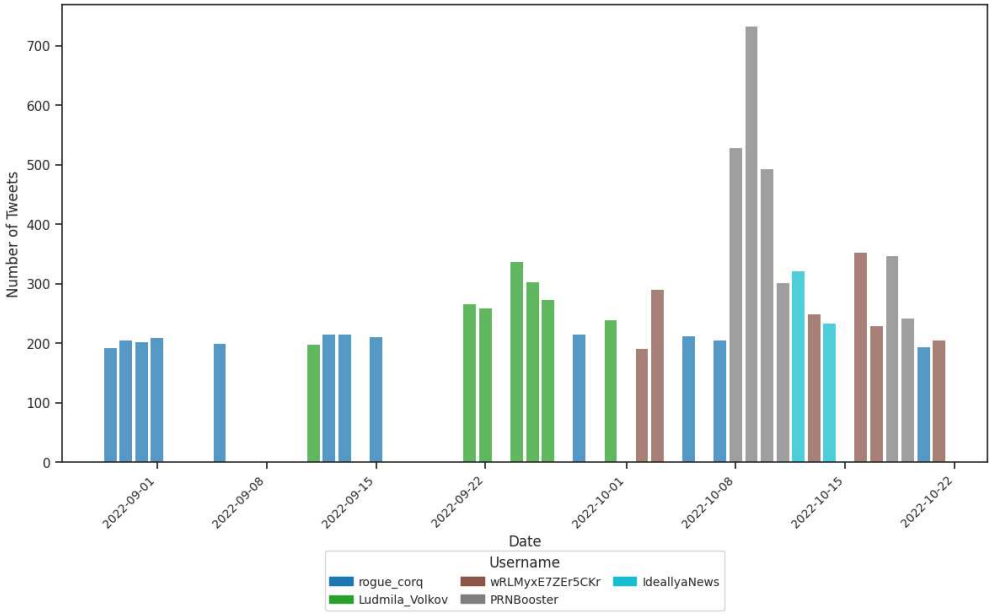

Figura 3.3: Utenti che hanno postato maggiormente per giorno

Esso è rientrato forzatamente nel dataset perché nel periodo considerato ha apposto hashtag a favore dell'ucraina, senza però andare a mettere contenuti interessanti e perciò si è deciso di eliminare dal dataset. Otteniamo una distribuzione coerente con le aspettative degli utenti che hanno maggiormente tweettato in assoluto. Alcuni degli utenti che si distinguono sono: PRNBooster, Ludmila_Volkov, rogue_corq e wRLMyxE7ZEr5CKr e solamente il primo non rientra negli utenti più assidui per numero di post. Andando a valutare i suoi tweet esso risulta essere un account aggregatore di notizie riguardanti la guerra.

Su X, in particolare, possiamo utilizzare un'euristica di ricerca che somma il numero di retweet e il numero di volte a cui il tweet è stato messo mi piace[3], per trovare l'engagement. Esso punta a valutare il grado di interesse e di interazioni tra gli utenti, andando a basarsi sull'interesse suscitato dal singolo tweet invece che sul numero totale. Possiamo in questo modo lavorare sulla ricerca di utenti capaci di influenzare gli altri. in particolare, andando ad agire sul nostro periodo di tempo interessato, troviamo i seguenti utenti,che risultano essere gli utenti con il maggior engagement, riportati in tabella 3.3

Username	Engagement totale
UAWeapons	2240212
nexta_tv	2132698
UAarmy_animals	885405
IAPonomarenko	874225
Blue_Sauron	712592
Tendar	704873
strategywoman	669274
SarahAshtonLV	644532
TheStudyofWar	538414
UKR_token	479653

Tabella 3.3: Utenti che hanno generato più engagement

[3] favourite in inglese

Possiamo adesso trovare la quantità di engagement creato dagli utenti che hanno postato maggiormente. Come possiamo notare in tabella 3.4 la differenza tra i 2 gruppi è estremamente ampia, ciò suggerisce che questa seconda metrica permetta di trovare utenti molto più interessanti da analizzare.

Username	Engagement
_Thirunarayan1	10640
UkraineAlert	8333
knittingknots	4714
Elpoliticonews	3080
rogue_corq	2245
IdeallyaNews	1041
ArvadaRadio	1023
wRLMyxE7ZEr5CKr	522
otfyxXWzpwdfTra	505
Ludmila_Volkov	436

Tabella 3.4: Engagement per gli utenti che hanno postato maggiormente

Prima di passare all'analisi degli utenti nello specifico andiamo a visualizzare la distribuzione dell'engagement in 3.4. Nello specifico, calcoliamo i massimi locali dell'engagement per valutare se ci sono degli eventi degni di nota che differiscono da quelli con in cui ci sono stati il maggior numero di post.

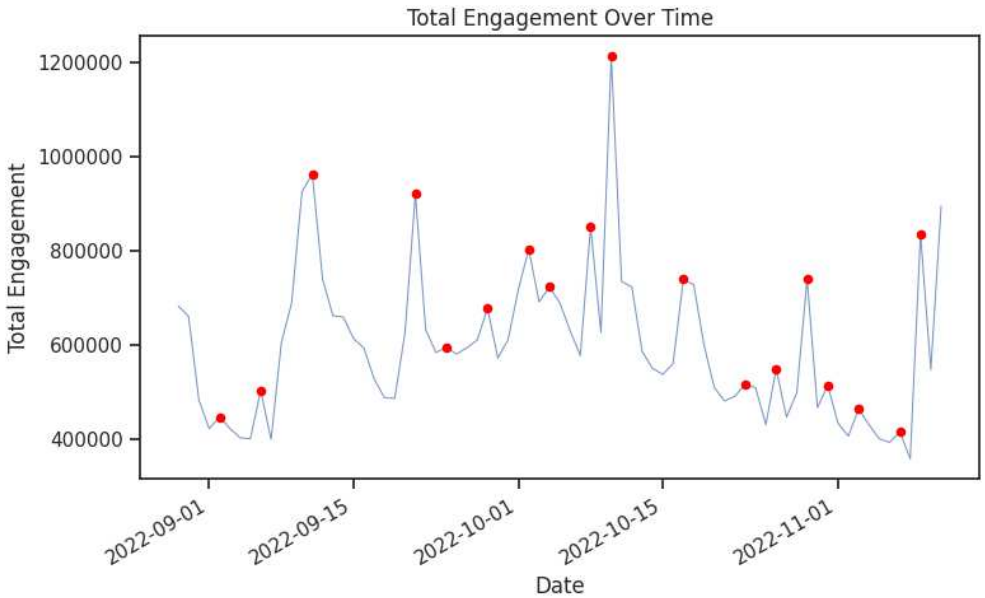

Figura 3.4: Engagement giornaliero durante il periodo della controffensiva ucraina

dalla quale possiamo otteniamo i 10 picchi, ordinati per quantità di engagement, come riportato in 3.5

Andiamo adesso a sovrapporlo al grafico dei tweet e degli utenti unici per verificare se ci sono degli eventi degni di nota, ovvero eventi di nicchia, in cui l'engagement è alto ma non sono altrettanto alti il numero di utenti e i tweet in assoluto. Per migliorare la visibilità scaliamo l'engagement di un fattore 10 (Figura 3.5).

Data	Engagement
2022-10-10	1213524
2022-09-11	963287
2022-09-21	920134
2022-10-08	851392
2022-11-09	834961
2022-10-02	802305
2022-10-29	739448
2022-10-17	738746
2022-10-04	722518
2022-09-28	677020

Tabella 3.5: Primi 10 picchi per Engagement derivati dalla distribuzione dell'engagement

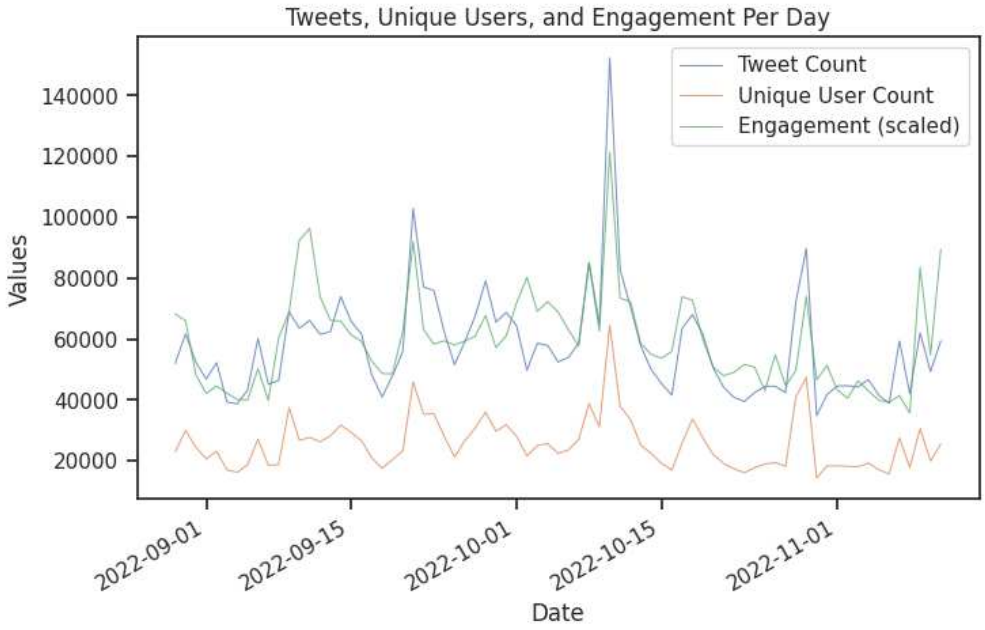

Figura 3.5: Engagement totale sovrapposto al grafico degli utenti unici e dei tweet totali

Possiamo andare adesso ad analizzare le date e vedere quali di esse sono più significative in assoluto. Dal grafico sovrapposto otteniamo 2 date particolari che risultano essere degli outlier: l'11 settembre e il 2 ottobre. Per quanto riguarda le altre date esse sono ancora coerenti, andiamo a descrivere brevemente gli eventi successi nelle due date in particolare:

- 11 settembre: nella notte tra il 10 e l'11 settembre la Russia si ritira dalla cittadina di Izyum, sancendo un'importantissima vittoria ucraina che libera così l'intero Oblast di Kharkiv e portando alla luce le prime crepe nell'esercito russo costretto a riassestarsi rapidamente.
- 2 ottobre: un'importante avanzata nel sud del paese permette di arrivare sulle sponde del fiume Dnipro, sancendo così l'inizio della ripresa della città di Kherson, pochi giorni dopo.

Ora che abbiamo una visione più chiara degli eventi e dell'evoluzione dell'engagement, andiamo ad analizzare nel dettaglio gli utenti che ne hanno creato maggiormente. Andia-

mo ad utilizzare BERTopic per valutare quali siano i loro topic di interesse e le keyword maggiormente interessanti:

- UAWeapons. Tra i primi topic possiamo trovarne alcuni riguardanti l' equipaggiamento militare perso, nell'ambito OSINT[4], con un particolare focus alle perdite vicino alle varie città.
- Nexta TV. i topic sono legati a perdite militari, relazioni internazionali, allerte riguardanti danni alla rete elettrica.
- UAarmy_animals, che è un account ufficiale dell'esercito ucraino legato alla diffusione di animali da compagnia incontrati o posseduti dai soldati. Essi sono particolarmente amati in generale.
- IAPonomarenko, è un corrispondente sul campo. I topic sono legati a equipaggiamento di guerra, analisi del conflitto, attacchi dei droni, e situazioni particolarmente famose nella stampa come l'assedio della Azovstal e delle città di Kherson.
- Blue_Sauron, analista che fornisce informazioni dal campo sulla guerra. I suoi topic sono legati, principalmente a combattimenti, video di droni, bombardamenti.
- Tendar, anch'esso analista militare, i suoi topic riguardano principalmente Operazioni di combattimento, situazioni militari specifiche, bombardamenti e OSINT.
- strategywoman, giornalista e reporter di Kyiv. I topic sono particolarmente legati agli update giornalieri sulla situazione a Kyiv.
- SarahAshtonLV, sottoufficiale dell'esercito ucraino, principalmente i topic riguardano update sulla guerra e sulla situazione LGBT.
- TheStudyofWar, account dell'Institute for the Study for War, think tank americano. i loro topic sono estremamente specifici e legati ad analisi e predizioni militari.
- UKR_token, giornale ispanico che tratta di argomenti legati alla guerra.

Possiamo concludere che molti degli utenti ad alto numero di engagement si suddividono in enti governativi, giornalisti e persone con un background militare come analisti o soldati. In particolare il terzo gruppo di persone ha presa su una determinata ed ampia fetta della popolazione per vari motivi, ma tutti ampiamente dentro la guerra. Questo è un ottimo indicatore della bontà dell'engagement rispetto al numero di tweet.

Possiamo adesso andare a chiederci quale sia il ruolo degli hashtag e quali sono i sentimenti legati al loro utilizzo nei tweet, identificando trend temporali per verificare la fazione sostenuta dagli utenti che li utilizzano.

Andiamo ad estrarre i 25 hashtag più comuni (Algoritmo 3.2). La colonna è nella forma di dizionario e dunque ha bisogno di alcune accortezze. Attraverso la funzione extract_hashtag che si utilizza *eval* per convertire la stringa in un dizionario Python. Il codice ritorna una lista con il valore del campo "text" di ciascun dizionario nella lista dei dizionari.

```
def extract_hashtags(hash_dict):
    try:
        hashtags_list = eval(hash_dict)
```

[4]Open Source INTelligence, cioè la branca dell'intelligence che utilizza e analizza fonti aperte e disponibili al pubblico

```
        return [item["text"].lower() for item in hashtags_list]
    except:
        return []

df['hashtags'] = df['hashtags'].apply(extract_hashtags)
```

Listing 3.2: funzione per estrarre gli hashtag dalla colonna preposta

Dopo aver estratto i 25 hashtag più comuni possiamo valutare la loro frequenza relativa sia sul totale dei tweet (Tabella 3.6.

Hashtag	Occorrenze	Frequenza relativa (%)	Frequenza sul totale (%)
ukraine	1215255	7.05	28.07
russia	730431	4.23	16.87
putin	455228	2.64	10.51
russiaisaterroriststate	250815	1.45	5.79
ukrainerussiawar	217737	1.26	5.03
standwithukraine	193274	1.12	4.46
ukrainewar	186163	1.08	4.30
nato	159349	0.92	3.68
russian	155888	0.90	3.60
kherson	154168	0.89	3.56
biden	140611	0.82	3.25
usa	131371	0.76	3.03
slavaukraini	129793	0.75	3.00
ukrainerussianwar	126996	0.74	2.93
china	111283	0.65	2.57
russiaukrainewar	95136	0.55	2.20
ukraine	94148	0.55	2.17
ucrania	91242	0.53	2.11
rusia	90260	0.52	2.08
ukrainian	85203	0.49	1.97
zelensky	83390	0.48	1.93
war	82069	0.48	1.90
canada	69236	0.40	1.60
russland	68492	0.40	1.58
ukrainewillwin	68424	0.40	1.58

Tabella 3.6: Frequenza degli hashtag

Come possiamo notare in Tabella 3.6, la maggior parte degli hashtag è chiaramente correlato ad eventi o fazioni della guerra. In particolare possiamo notare come diversi hashtag vengano riproposti a causa di differenze nella lingua o inversioni di parole. Inoltre sembrerebbe esserci un certo bias positivo nei confronti dell'Ucraina andando ad osservare la frequenza degli hashtag come "russiaisaterroriststate", e "standwithukraine", entrambi nella top 10.

Valutiamo se questo corrisponde a realtà. Andiamo adesso a considerare il sentiment dei top hashtag. Nonostante essi non siano spesso frasi compiute, essi ci permettono di dare un giudizio negativo, positivo o neutro rispetto al contenuto del tweet.

Con l'utilizzo di VADER, prendiamo gli hashtag dal testo, invece che dalla colonna. Dei migliori 25 hashtag andiamo ad analizzare il sentiment del tweet in cui esso è presente.

Andiamo poi a confrontarli per ottenere un giudizio. i risultati sono sintetizzati nella Tabella 3.7.

Hashtag	Sentiment Score
ukraine	-0.09
russia	-0.09
putin	-0.13
russiaisaterroriststate	-0.10
ukrainerussiawar	-0.08
standwithukraine	0.01
ukrainewar	-0.09
nato	-0.11
russian	-0.11
kherson	-0.11
biden	-0.07
usa	-0.09
slavaukraini	0.04
ukrainerussianwar	-0.07
china	-0.03
russiaukrainewar	-0.10
ukraine	-0.08
ucrania	-0.05
rusia	-0.05
ukrainian	-0.10
zelensky	-0.05
war	-0.08
canada	-0.07
russland	-0.37
ukrainewillwin	0.00

Tabella 3.7: Sentiment score per i migliori Hashtag

Il risultato che ne consegue, è diverso rispetto all'ipotesi iniziale. Se da un lato abbiamo della positività riguardo ad alcuni hashtag, la quasi totalità e cioè 22 su 25 risultano avere un sentiment score negativo e nessuno supera la soglia di positività di 0.05. Possiamo argomentare che nell'analisi del sentiment i tweet posseggono affermazioni negative sia per contrarietà rispetto alla fazione sia per gli orrori che un conflitto può portare. Ci sono però degli hashtag che portano alla luce un certo bias anti-atlantista a causa della negatività degli hashtag NATO, USA e Canada. Importante notare come il sentiment riguardante gli hashtag in lingua specifica differiscano particolarmente da quelli generici e in lingua inglese. Come la differenza tra russia e russland, la prima in lingua inglese mentre la seconda in lingua tedesca.

3.3 Analisi dello Stallo

La seconda parte del dataset, comprende tweet che vanno dall'11 novembre al 7 giugno 2023. Il maggior intervallo temporale è supportato dalla più lenta evoluzione degli eventi storici. Infatti nel periodo invernale abbiamo pochi movimenti complessi sul terreno e molta guerra di posizione.

Bisogna prendere in considerazione la questione della storia del dataset, che in questo periodo risulta essere rilevante. Nel mese di febbraio 2023 ci sono stati problemi legati alla quantità di tweet però come vedremo non inficia particolarmente l'analisi. In compenso nei mesi tra febbraio e fine maggio abbiamo un afflusso maggiore di tweet dato in parte dalla modifica al dataset.

Fatte queste considerazioni possiamo valutare come, nel mese di maggio e giugno abbiamo altri picchi, che mancavano nei mesi invernali, nonostante la sospensione di diversi account sviluppatore.

Una delle possibili cause è data dal fatto che gli eventi sviluppatisi nei periodi primaverili possono essere più interessanti al grande pubblico dovuta alla ripresa delle operazioni di guerra. Infatti data la natura pianeggiante del terreno di combattimento e data l'impossibilità di avanzamento sul ghiaccio e sul fango, quando le temperature tendono ad alzarsi c'è anche un disgelo delle operazioni.

Andando a verificare la grandezza del dataset abbiamo 14776079 di tweet che risulta essere corposo aumento rispetto alla prima parte.

Anche qui abbiamo deciso di andare a considerare il dataset, per quanto concerne l'analisi testuale e generale, della parte in lingua inglese.

In questo caso abbiamo 65 lingue totali, di cui la lingua inglese costituisce il 56% del totale, un 7.3% di lingue non classificate seguite da un 6.6% di francese, 6.4% di spagnolo e 5.8% di tedesco.

Le percentuali sono all'incirca coerenti con la prima parte del dataset anche se, tra le lingue europee notiamo un lieve aumento percentuale, legato soprattutto alla diminuzione delle lingue non classificate.

Il numero di utenti unici che hanno tweettato è di 1950635. Anche qui andiamo a valutare chi sono gli utenti che hanno il maggior numero di tweet per valutare in che tipologia di utenti ricadano (Tabella 3.8).

Username	**Tweet**
Hkjhgc2	61098
UlfaniaEda	53591
yusr35144430	33118
HiHikeep4	30814
rogue_corq	26739
Hike150Hike	26462
queen_ukraine	25647
RabiaSalem02	24974
Wejdan_cameron	22745
belal4abty	21250
NewsExplorerFr	18718
LendaThode	17299
HiGk2k	17013
LindaLopez67	16929
IdeallyaNews	16128

Tabella 3.8: Username con il maggior numero di tweet

Possiamo andare a dare una prima distinzione degli utenti che appaiono nella Tabella 3.8:

- Il primo gruppo è dato da coloro che sono anche nella prima parte del dataset: IdeallyaNews e rogue_corq.

- Il secondo gruppo è dato da giornali ufficiali come NewsExplorerFr.
- Il resto degli utenti può esser categorizzato come propagandisti o retweetter di notizie.

Adesso, visualizziamo la quantità di utenti unici nel periodo considerato, con i relativi massimi locali (Figura 3.6).

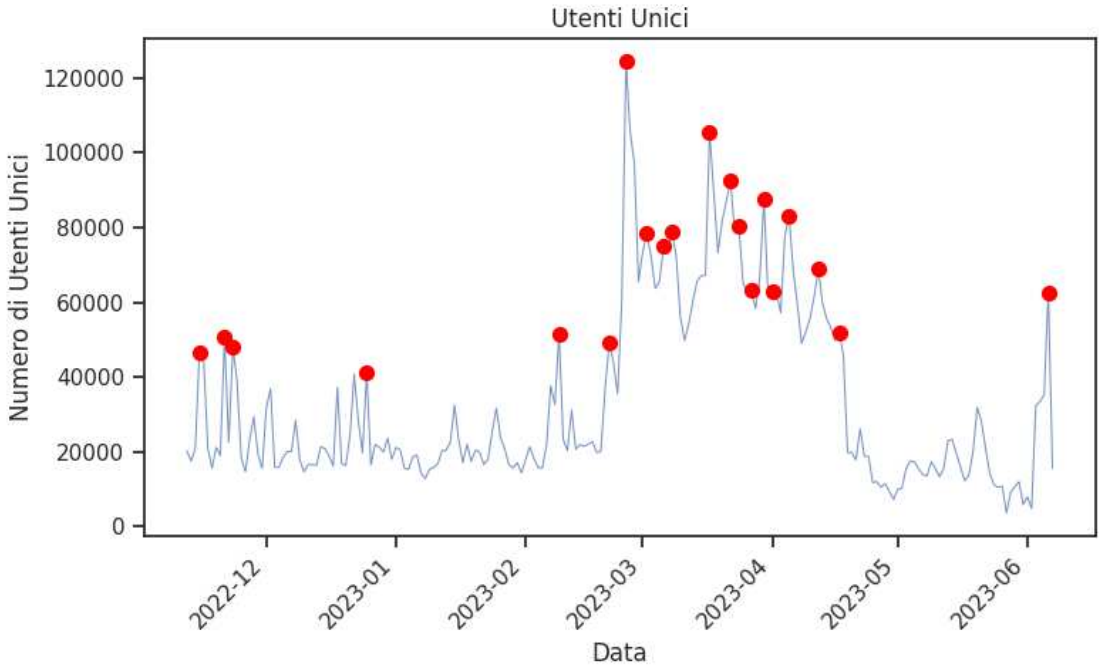

Figura 3.6: Utenti unici nel periodo di tempo considerato

Data la maggior quantità di dati e di massimi locali ci limitiamo nel numero di valori da visualizzare, sintetizzati in Tabella 3.9.

Date	valori di picco
2023-02-25	124348
2023-03-17	105374
2023-03-22	92322
2023-03-30	87624
2023-04-05	83158
2023-03-24	80350
2023-03-08	78891
2023-03-02	78336
2023-03-06	74869
2023-04-12	68853

Tabella 3.9: Dieci giorni con maggior numero di utenti unici

Notiamo come nei periodi di riferimento per la maggior quantità di eventi essi si concentrano nel periodo del dataset in cui sono aumentati i retweet. Da questi picchi possiamo andare a valutare se ci sono eventi in particolare che hanno attirato l'attenzione degli utenti per spingerli a postare maggiormente oppure se queste oscillazioni sono dovute ad una problematica del dataset.

Prima di iniziare a rispondere è bene sovrapporlo al numero di tweet per valutare se ci sono o meno delle discordanze (3.7).

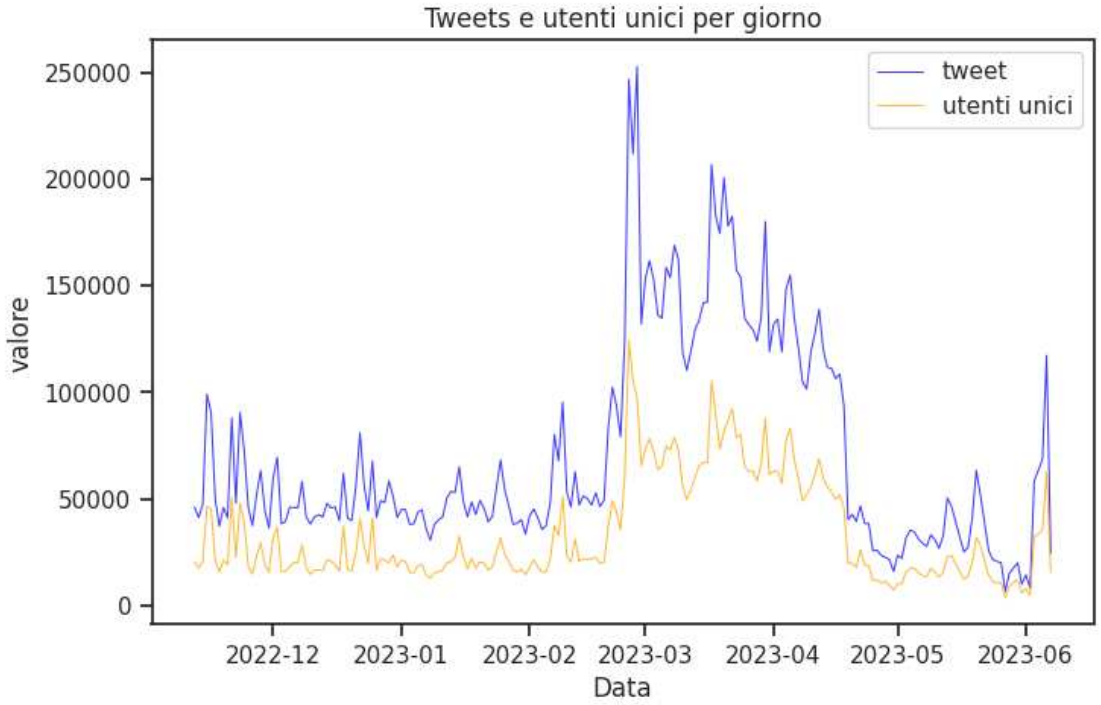

Figura 3.7: Numero di utenti e numero di tweet giornalieri

Andando a guardare la Figura 3.7, notiamo come ci sia una forte correlazione tra i due grafici ma è anche interessante notare come ad alcuni dei picchi degli utenti unici non corrispondano al numero di tweet. Analizziamo ora gli eventi trovati:

- 25 febbraio: i carri armati tedeschi Leopard arrivano in Ucraina, assieme alla visita del presidente polacco Morawiecki. Inoltre gli USA autorizzano un pacchetto di aiuti da 2 miliardi, come chiaro segnale di supporto in occasione del'anniversario della guerra.
- 27 febbraio: Putin decide di reclutare prigionieri di guerra ucraini per creare un battaglione speciale, contravvenendo a diverse leggi internazionali.
- 17 marzo: la corte internazionale dell'Aia condanna il presidente Vladimir Putin.
- 20 marzo: c'è l'invio di un sostanzioso pacchetto di aiuti da parte dell'unione europea.
- 6 giugno: c'è un attacco con i droni ai danni della diga di Nova Khakovka, nella zona sud del paese e il disastro che ne è conseguito.

Verifichiamo ora quali sono stati gli utenti che hanno postato maggiormente giorno per giorno durante il periodo considerato. Troviamo subito degli utenti con valori estremamente elevati, in particolare l'utente "Telehuntwatch". Analizzando i suoi tweet, risulta essere un bot che sponsorizza l'affiliazione ad un canale telegram riguardante la guerra. In Figura 3.8 possiamo vedere il grafico modificato eliminando l'utente in questione. il grafico risulta essere abbastanza uniforme con pochi valori estremi anche nei giorni di maggior interesse. In particolare due utenti risultano essere particolarmente arrivi nel periodo tra marzo e fine maggio: UlfaniaEda e Hkjhgc2 che sono presenti anche nella parte della controffensiva a cui si aggiunge l'utente Ezybad.

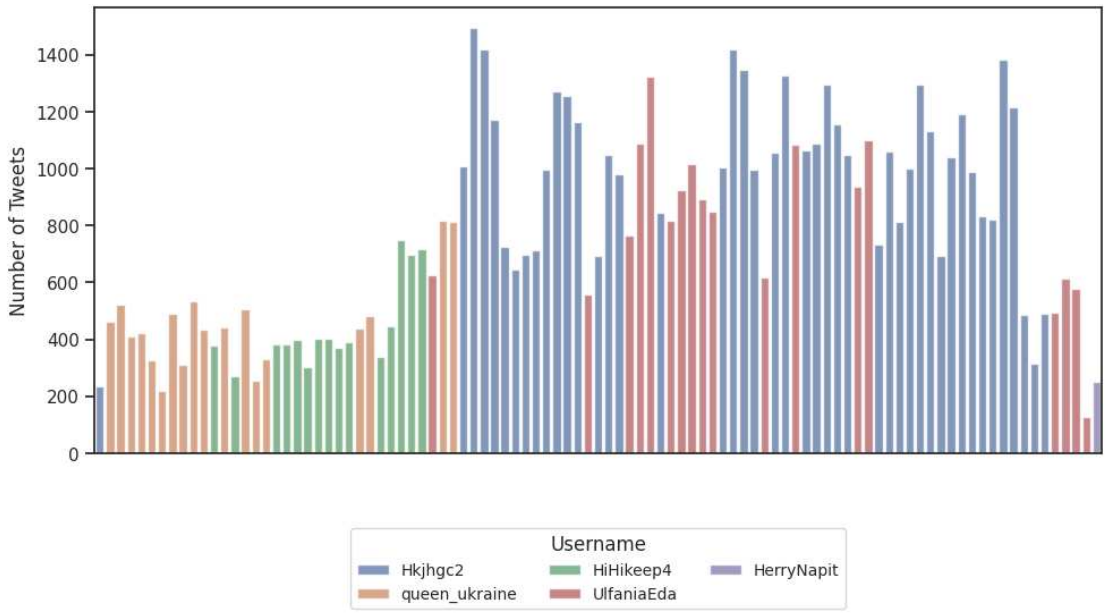

Figura 3.8: Utenti unici per giorno

Come già fatto in precedenza l'analisi dei tweet non è sufficiente. Andiamo a supportarla con il calcolo e la visualizzazione dell'engagement, definito come somma dei retweet e dei favorite.

L'ipotesi anche questa volta è che gli utenti che postano maggiormente non siano coloro tra i quali portano ad avere maggiori interazioni. Inoltre andiamo a cercare se esistono eventi di nicchia.

Anche in questo caso otteniamo degli username che sono diversi da quelli dei maggiori tweettatori, come possiamo vedere in Tabella 3.10. Un'eccezione notevole è quella di HerryNapit che è presente tra gli utenti che hanno postato maggiormente in un dato giorno e in particolare il giorno 5 giugno, suggerendo che esso può avere dei legami con l'evento del giorno successivo.

Username	Engagement
nexta_tv	2206930
Tendar	1926255
strategywoman	1537144
UAWeapons	1270394
UAarmy_animals	1027033
GlasnostGone	778600
TheStudyofWar	772103
IAPonomarenko	660130
jjaranaz94	637557
HerryNapit	438357

Tabella 3.10: Engagement totale per Username

Andiamo adesso a valutare le differenze di engagement con gli utenti che fanno più tweet, come si può notare in Tabella 3.11. La situazione ricalca quella della prima parte del dataset.

Possiamo notare come nella lista degli utenti che hanno il maggior livello di engagement 7 sono presenti nella prima parte del dataset legato alla controffensiva, suggerendo che il

Username	Engagement
rogue_corq	10904
Hkjhgc2	4835
Wejdan_cameron	3832
Hike150Hike	3226
HiHikeep4	2342
HiGk2k	2335
yusr35144430	2183
LendaThode	1936
UlfaniaEda	1660
queen_ukraine	1587

Tabella 3.11: Engagement totale degli utenti col maggior numero di tweet

loro status è costante nel tempo, mentre per quanto riguarda le altre 3 persone esse sono: GlasnostGone, jjaranaz94, HerryNapit. Prima di andare ad analizzarli passiamo alla visualizzazione dell'engagement, del numero di utenti e dei tweet in Figura 3.9 (l'engagement è scalato di un fattore 1000 per facilitare la visualizzazione).

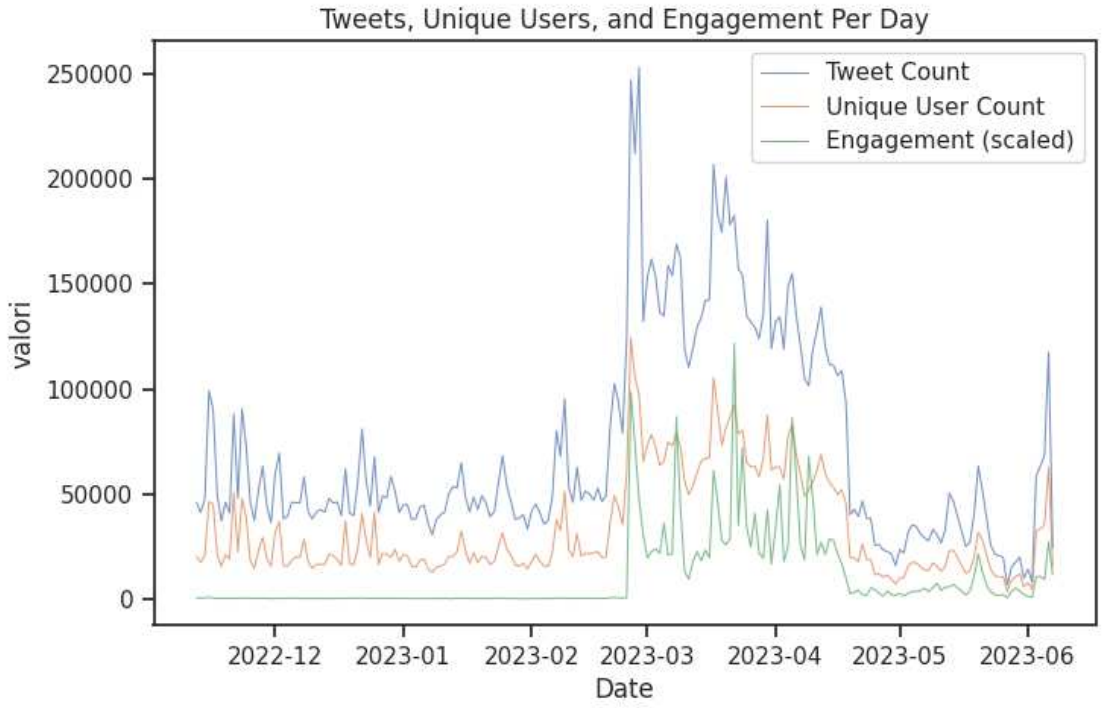

Figura 3.9: engagement totale sovrapposto ai tweet e agli utenti unici

Da questo secondo grafico, otteniamo delle date interessanti, con delle differenze tra le due metriche:

- 22 marzo: in questa giornata il presidente ucraino porta delle medaglie ai soldati nella città di Bakhmut, andando pericolosamente sulla linea del fronte.
- 9 aprile: per il primo giorno dopo più di un anno e mezzo di guerra, l'Ucraina ritorna ad esportare elettricità. Segnale che la strategia russa di attaccare le centrali non ha provocato danni a lungo termine.

In questo caso gli eventi sopracitati ricadono nella categoria degli eventi di nicchia o per gli addetti ai lavori, essendo non particolarmente significativi.

Andiamo adesso ad analizzare quali sono gli argomenti che trattano, con BERTopic, i 3 utenti precedentemente trovati:

- GlasnostGone, analista militare e youtuber. I suoi topic riguardano principalmente operazioni militari e notizie flash.
- jjaranaz94, account totalmente particolare, sospeso successivamente che parla di musica e di covid in tutti i suoi topic.
- HerryNapit, un giornalista yemenita famoso per aver fatto ampi reportage sulla situazione in Yemen e nel periodo considerato ha coperto la situazione ucraina.

Possiamo concludere che in questo caso quasi tutte le personalità con alto engagement si sono rivelate essere interessanti e legati principalmente a giornalisti di guerra, analisti militari, tv e giornali e organizzazioni governative.

Andiamo adesso ad estrarre gli hashtag dalla colonna hashtag, otteniamo la Tabella 3.12 con occorrenze, frequenza relativa e frequenza sul totale.

Hashtag	**Occorrenze**	**Frequenza relativa (%)**	**Frequenza sul totale (%)**
ukraine	4875926	8.24	33.00
russia	2246430	3.80	15.20
bakhmut	1257331	2.13	8.51
putin	953419	1.61	6.45
canada	891051	1.51	6.03
germany	857356	1.45	5.80
russiaisaterroriststate	752710	1.27	5.09
usa	712656	1.20	4.82
ukrainerussiawar	680712	1.15	4.61
standwithukraine	656019	1.11	4.44
zelensky	650249	1.10	4.40
nato	597475	1.01	4.04
ukrainewar	593518	1.00	4.02
rusia	548404	0.93	3.71
america	502168	0.85	3.40
china	421921	0.71	2.86
russian	409336	0.69	2.77
ucrania	375966	0.64	2.54
ukrainerussianwar	365769	0.62	2.48
slavaukraini	362764	0.61	2.46
ukrainewillwin	355596	0.60	2.41
kyiv	324420	0.55	2.20
ucraina	323370	0.55	2.19
ukrainian	312997	0.53	2.12
war	264713	0.45	1.79

Tabella 3.12: Hashtag frequencies and percentages

In questo caso abbiamo comprensibilmente alcuni hashtag che riguardano la situazione in generale ed alcuni come quelli di Bakhmut, centro nevralgico della maggior parte degli scontri.

Tra gli hashtag non abbiamo particolari differenze, anche se canada e germany hanno avuto un numero di occorrenze molto superiore, in questa parte del dataset. In particolare quello rigurdante la germania è a causa della questione dei carri armati Leopard, particolarmente sentita.

Andiamo adesso ad utilizzare VADER per valutare il sentiment dei tweet che presentano questi hashtag. Sempre considerando che un hashtag possa convogliare idee riguardanti i bias pro o contro le condizioni e situazioni della guerra.

L' hashtag con il maggior sentiment negativo che troviamo è Bakhmut, chiaramente legata alla lunghissima guerra di logoramento che si è dipanata durante tutto il periodo considerato (Tabella 3.13). Anche in questo caso troviamo delle differenze sostanziali tra i valori nelle lingue nazionali e in lingua inglese, con differenze che possono essere estremamente visibili come nel caso di USA e America o Ukraine e Ucraina.

Hashtag	**Sentiment**
ukraine	-0.12
russia	-0.11
bakhmut	-0.25
putin	-0.15
canada	-0.15
germany	-0.25
russiaisaterroriststate	-0.13
usa	-0.02
ukrainerussiawar	-0.09
standwithukraine	0.02
zelensky	-0.05
nato	-0.08
ukrainewar	-0.10
rusia	-0.05
america	-0.21
china	-0.05
russian	-0.14
ucrania	-0.05
ukrainerussianwar	-0.10
slavaukraini	0.04
ukrainewillwin	-0.00
kyiv	-0.25
ucraina	0.05
ukrainian	-0.09
war	-0.10

Tabella 3.13: Calcolo del sentiment dei tweet che contengono gli hashtag

Come il precedente caso, notiamo che gli hashtag sono tendenzialmente negativi, con qualche rara eccezione di neutralità data da quegli hashtag che sono più legati al bias positivi riguardo l'ucraina.

La situazione legata alla valutazione del sentiment degli hashtag anche in questo caso risulta essere tendenzialmente negativa per la maggior parte di essi. Da una parte abbiamo una visione che è andata peggiorando anche degli hahstag più neutrali, dall'altra alcuni sono molto variabili in base alla lingua. Come possiamo notare il sentimento legato alla situazione legata a bakhmut è tra i più negativi che sono stati analizzati con un -0.25 che non lascia

adito ad interpretazioni. La situazione, ridottasi ad un "tritacarne" ha pesantemente mosso il sentiment generale. Inoltre anche altri hashtag come Germany, Kyiv, Canada e America sono particolarmente negativi. Indicando una certa negatività riguardo al blocco atlantico. Di particolare interesse è la questione legata alle lingue nazionali rispetto a quella inglese, Questo si riflette sulla dicotomia tra usa e america, con il primo che possiede il quarto valore più elevato mentre l'altro ampiamente negativo. Allo stesso modo Ucraina e ukraine seguono lo stesso rapporto. Il sentimento di coloro che hanno a cuore la causa invece è generalmente positivo e costante nell'esserlo.

3.4 Confronto tra le due Analisi

Confrontando le due analisi generali possiamo valutare l'evoluzione del sentiment tra le due divisioni del dataset. Abbiamo preso i 25 hashtag più utilizzati tra i due dataset e messi a confronto, ed abbiamo calcolato il sentiment dei relativi tweet in Tabella 3.14.

Hashtag	Sentiment controffensiva	Sentiment stallo	Differenza
ukraine	-0.09	-0.12	0.03
russia	-0.09	-0.11	0.02
bakhmut	-0.13	-0.25	0.12
putin	-0.13	-0.15	0.02
canada	-0.07	-0.15	0.08
germany	-0.25	-0.25	0.00
russiaisaterroriststate	-0.10	-0.13	0.03
usa	-0.02	-0.02	0.00
ukrainerussiawar	-0.08	-0.09	0.01
standwithukraine	0.01	0.02	-0.01
zelensky	-0.05	-0.05	0.00
nato	-0.11	-0.08	-0.03
ukrainewar	-0.09	-0.10	0.01
rusia	-0.05	-0.05	0.00
america	-0.21	-0.21	0.00
china	-0.03	-0.05	0.02
russian	-0.11	-0.14	0.03
ucrania	-0.05	-0.05	0.00
ukrainerussianwar	-0.07	-0.10	0.03
slavaukraini	0.04	0.04	0.00
ukrainewillwin	0.00	-0.00	0.00
kyiv	-0.25	-0.25	0.00
ucraina	0.05	0.05	0.00
ukrainian	-0.10	-0.09	-0.01
war	-0.08	-0.10	0.02

Tabella 3.14: Comparazione del sentiment tra gli hashtag delle due parti del dataset

Come ci si poteva aspettare l'hashtag Bakhmut risulta essere quello con l'evoluzione negativa maggiore, data dal maggior peso con il prolungarsi degli scontri. Altri hashtag che sono genericamente legati alla guerra hanno visto il loro valore tendere sempre di più verso il sentiment negativo. Al contrario gli hashtag pro Ucraina hanno visto il loro valore rimanere costante.

Altro elemento interessante è dato dalla consistenza dei 7 utenti che hanno avuto il maggior grado di engagement in ambo le parti del dataset che ricordiamo essere: nexta_tv, Tendar, strategywoman, UAWeapons, UAarmy_animals, IAPonomarenko, TheStudyofWar

andiamo a dividere l'engagement per il numero totale di tweet, questo per avere un'idea di quanto engagement per giorno ci sia e se ci sono delle differenze di sorta tra i due dataset. Per quanto riguarda la controffensiva, non abbiamo particolari variazioni, i valori sono tendenzialmente compresi tra 14 e 16 per tutti i giorni i riferimento. Mentre per la parte della stallo abbiamo ampie differenze per le quali ci costringe a dividerlo in due parti. La prima parte con un rapporto che si attesta tra i 4 e i 9 per i giorni fino al 20 di febbraio, indicando un generale disinteressamento nei confronti degli eventi. Nel prosieguo invece i valori sono compresi tra i 100 e i 200 nella maggior parte dei giorni, con i picchi che arrivano a 600 in corrispondenza degli eventi importanti trovati precedentemente. In questo secondo caso notiamo come la differenza di engagement sia notevolmente più pronunciata, a causa della modifica del web crawer nella seconda parte del periodo.

3.5 Analisi dei fattori predittivi dei trend su X

Per rispondere alla domanda che ci siamo fatti sulla questione riguardante la possibilità di anticipare e predire trend ed eventi in base all'analisi dei tweet, dobbiamo andare ad analizzare alcuni dei massimi locali di interesse.

Per quanto riguarda la parte legata alla controffensiva abbiamo pensato di scegliere l'11 settembre.

Nonostante esso sia un giorno che non corrisponde ad un massimo a livello di tweet, la differenza tra l'alto livello di engagement e il relativo basso livello di tweet ed utenti unici indica un evento particolare e relativamente di nicchia.

3.5.1 Analisi di un evento della Controffensiva Ucraina

Per prepararci all'analisi, utilizziamo BERTopic per andare ad analizzare la settimana precedente, il giorno stesso e i giorni successivi. In questo modo abbiamo una panoramica degli eventi precedenti e successivi per andare a valutare l'evoluzione dei trend.

Iniziamo con i tweet che vanno dal 4 settembre fino al 10 settembre essi vedono un'analisi su 366195 Tweet. La clusterizzazione dei topic risultante può essere sintetizzata in questi eventi:

- Pericolo alla centrale nucleare: con le keywords plant,nuclear,iaea,power plant,nuclear power,zaporizhzhia,nuclear plant. Questo primo evento risulta esser legato alla situazione della centrale nucleare e al report dell'ente per l'energia nucleare riguardo ai possibili danni subiti e al fatto che si sia perso il contatto con parte della griglia di alimentazione.

- Avversione verso la NATO: con le keywords nato, fuck nato, nato countries, ukraine nato. Il secondo risultato si scaraventa duramente contro l'immobilità dell'alleanza atlantica, in particolare con Jens Stoltemberg, il segretario generale. Dall'analisi degli hashtag vista nella sezione generale abbiamo notato una tendenza alla negatività con un -0.11 di sentiment score. L'evento in particolare serve alla propaganda russa per circostanziare il conflitto, aggiungendo minacce di guerra nucleare.

- Donald Trump e Biden: con le keyword Biden. speech, maga, republicans, enemy state, biden. Il terzo cluster si focalizza su un particolare discorso di Donald Trump contro il presidente Biden. Il topic nell'evento è il primo topic spurio e cioè che ha

solo relativamente interesse nella questione ucraina. D'altro canto entrambi si sono particolarmente spesi riguardo alla guerra ed entrambi sono personaggi importanti nel futuro e nello svolgimento della guerra stessa.

- Supporto all'Ucraina: con le keyword slavaukraini, slavaukraini iaponomarenko. Questo topic si collega all'hashtag legato al supporto dell'ucraina. Interessante anche la menzione dell'account @iaponomarenko che risulta essere un corrispondente per l'Ucraina e che è stato tra le persone che hanno maggiormente creato engagement in entrambe le parti del dataset.
- Accordi con la Corea del Nord: con le keyword korea, north korea, northkorea ,artillery shells, shells, rockets, russia buying. In questo caso abbiamo la notizia dell'acquisto, da parte della Russia di missili Nord Coreani che da sempre crea clamore a causa del continuo armamento della nazione asiatica.

Questo ci permette di fare una prima cernita sugli argomenti di interesse. Un potente tipo di visualizzazione è dato dalla Heatmap (Figura 3.10), che serve per valutare quanta correlazione ci sia tra i primi 20 topic.

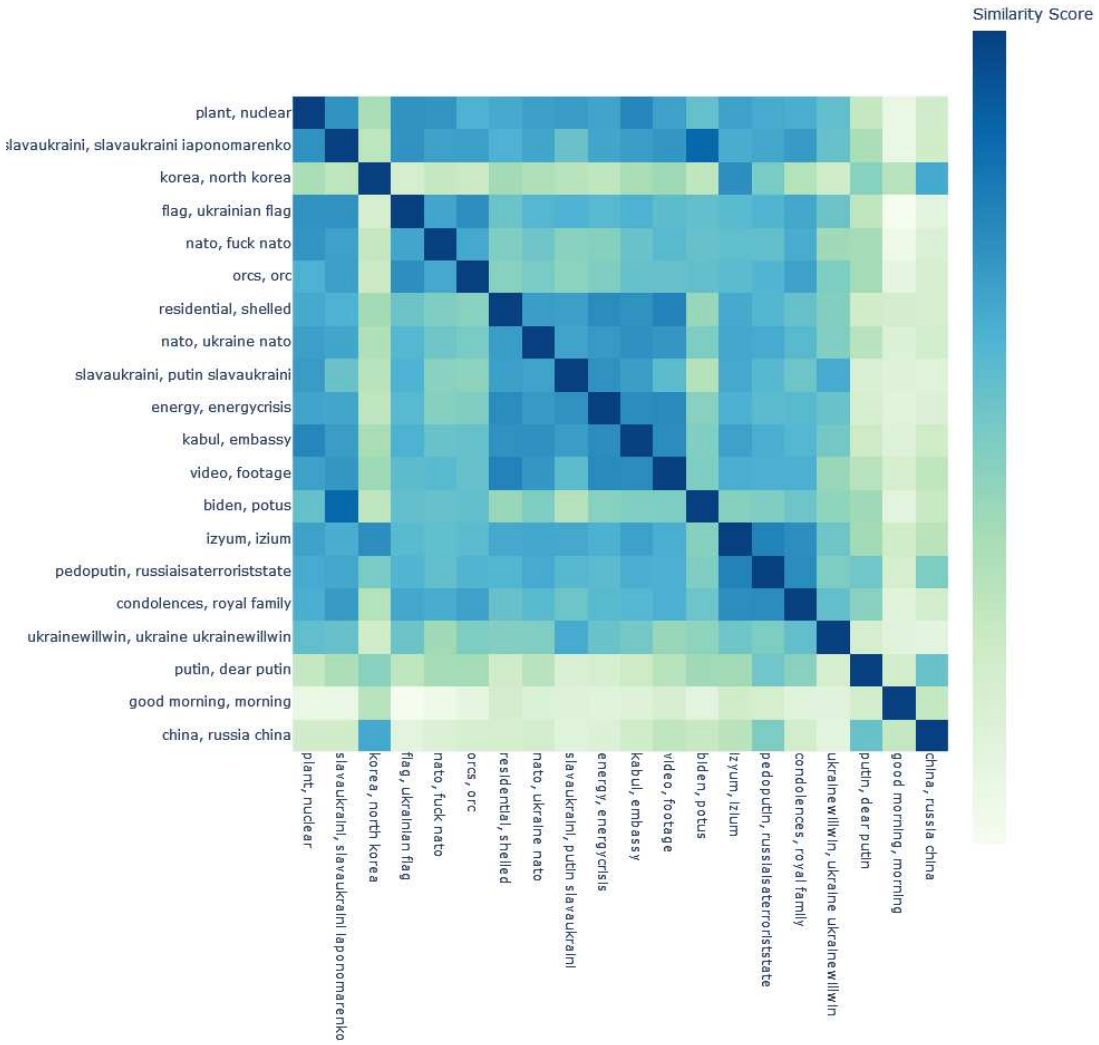

Figura 3.10: Correlazione tra i topic identificati

Da Figura 3.10, possiamo notare come la maggior parte dei topic effettivamente siano correlati tra di loro. Molti topic riguardanti La centrale Nucleare, l'arrivo degli HIMARS,

l'indipendenza di Kyev, il ponte Antonovsky di Kherson, e le notizie dalla Crimea e dalla Corea del Nord. Da qui possiamo anche notare come la maggior parte dei topic spuri non siano di fatto correlati, garantendoci la bontà della coerenza di BERTopic nel suo clustering.

Altra potente ed interessante visualizzazione è data dalla distanza tra i topic, anche chiamata Intertopic Distance. Essa si configura come la distanza tra i cluster dove quelli più semanticamente simili sono ravvicinati attraverso il coherence score, per valutare quanto essi siano comprensibili dall'essere umano. In Figura 3.11, si riporta l'intertopic distance dei topic presenti tra il 4 ed il 10 settembre.

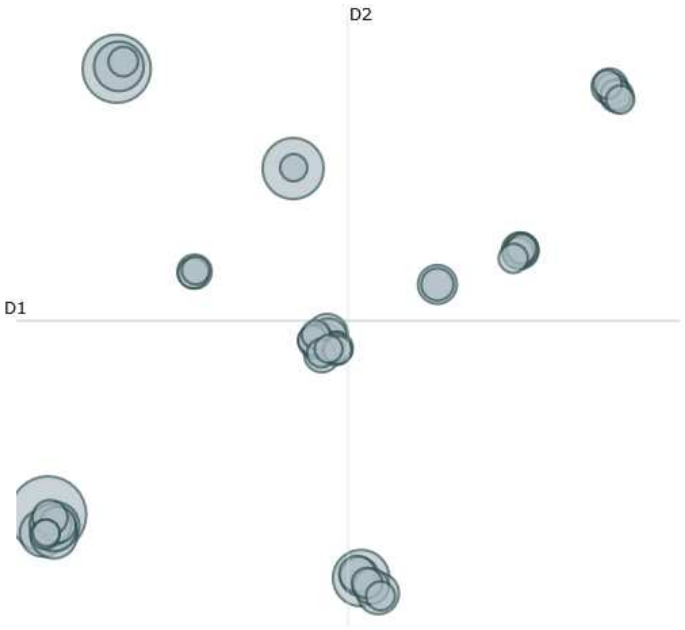

Figura 3.11: Intertopic distance dei giorni tra il 4 ed il 10 settembre

In Figura 3.11 notiamo diversi cluster ben spaziati. La parte bassa con i cluster più grandi è infatti quella legata ai primi topic riguardanti la guerra in Ucraina. Andando verso i quadranti 1 e 4, in particolare l'1, troviamo un'ampia distanza rivelando topic che non sono correlati e che esulano dalla guerra.

Inoltre, per rispondere più agevolmente alla domanda posta è di fondamentale importanza sapere quali siano state le variazioni di interesse dei topic specifici lungo il periodo di tempo considerato.

La questione del graficare lungo un periodo di tempo i topic ha comportato diverse difficoltà data l'estrema necessità di potenza alla grande quantità di tempo. Per la parte riguardante la controffensiva non è stato necessario andare a lavorare sulla forma del dataset né sull'eliminazione di parti di interpunzione, hashtag e chiocciole, vista la quantità minore di tweet considerata (Figura 3.12).

In Figura 3.12 si nota come alcuni topic degradino, come quello legato alla centrale nucleare e ai suoi problemi. La liberazione Izyum, risulta essere l'evento più interessante che inizia a farsi largo nei giorni tra il 9 e il 10 settembre, andando a sovrastare e a predire l'evento legato all'11 settembre e cioè la totale liberazione della città. Le prime avvisaglie sono chiare dai giorni precedenti e guadagnano di intensità nei giorni.

Per quanto riguarda il giorno stesso, l'11 settembre, andando ad analizzare i topic essi sono particolarmente di giubilo. Con tutti i primi topic, ad eccezione dei primi due. Il topic 1 riguarda le condizioni disastrose della situazione delle città e il topic 2 dei carri armati

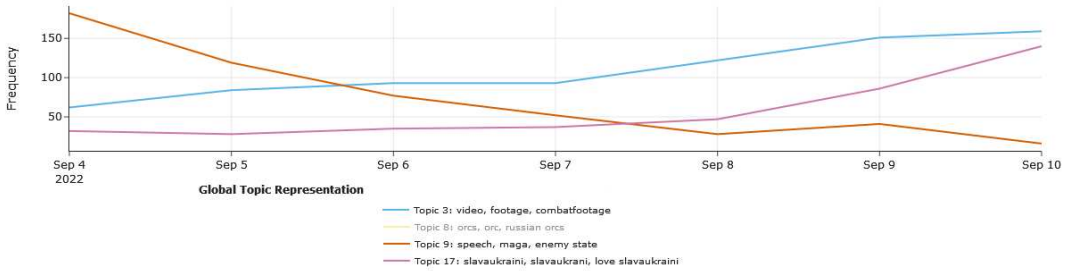

Figura 3.12: Evoluzione dei topic durante il periodo dal 4 al 10 settembre

abbandonati. Il resto dei primi topic sono tutti largamente in supporto alla vittoria ucraina. Diamo ora una visione più completa:

- Infrastrutture elettriche ed idriche: con le keyword electricity, water, regions. Come detto sopra le condizioni problematiche all'ingresso della città hanno creato particolare clamore.
- Veicoli militari catturati e abbandonati: con le keyword tanks, abandoned, captured. La rapida conquista della città e l'altrettanta rapida ritirata russa ha lasciato un grandissimo quantitativo di veicoli militari in buono stato. La loro cattura è stata un notevole vantaggio militare data la disparità dei mezzi in campo.
- Sostegno all'Ucraina: con keyword come ukrainewillwin, ukrainerussianwar, ukraine. Chiaramente legati al giubilo per la conquista della città e la rapidissima ritirata russa.
- Sostegno all'Ucraina: anche qui con keyword di supporto e sostegno come: slavaukraini, slavaukrani, iaponomarenko. Con la figura di iaponomarenko che possiamo ricordare essere una delle figure chiave, presente nella lista delle persone più influenti per engagement e capace di creare grandissima diffusione con le sue notizie dal fronte.
- Sostegno e ringraziamento per gli alleati: con le keyword ukrainewillwin, thank, europe, help, usa. In cui si ringrazia per il supporto dato da parte del blocco occidentale.
- Centrale nucleare di Zaporizhzhia: con le keyword nuclear, zaporizhzhia, plant, power. Che sottendono alle paure legate ai possibili gravi danni alla centrale.

Anche l'intertopic distance in Figura 3.13 riflette questo eccellente accorpamento.

Come possiamo vedere i cluster tendono ad essere tutti molto semanticamente vicini con poche eccezioni. Solamente il topic a destra dell'asse cartesiano ha una distanza maggiore ma principalmente per l'uso di parole ucraine. Infatti esso è legato alle parole heroyamslava e slavaukraini.

Andiamo infine a valutare la situazione nei giorni successivi. La situazione risulta molto più grave ed emotivamente difficile a causa dell'evento principale che è la scoperta di fosse comuni. Inoltre l'interesse varia particolarmente con i primi topic che divergono in più direzioni verso topic differenti.

- Fosse comuni a Izyum. Keywords: mass, bodies, graves, grave, mass grave, burial, mass burial, burial site. Entrati ad Izyum, subito si ha la scoperta di fosse comuni con centinaia di cadaveri solleva le accuse di crimini di guerra e sdegno generale in tutto il mondo.

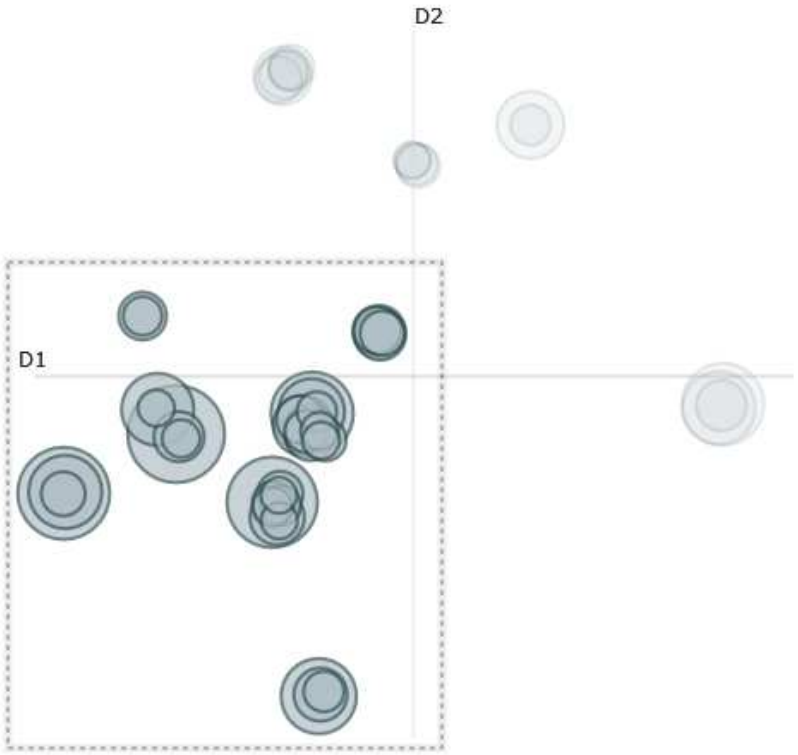

Figura 3.13: Intertopic distance del giorno 11 settembre.

- Escalation del conflitto tra Armenia e Azerbaijan. Keywords: armenia, azerbaijan, armenian, armenia azerbaijan, armenians. Nel Nagorno-Karabakh, regione contesa tra Armenia e Azerbaijan, subito dopo la conquista della città di Izyum, si sono riaccese le tensioni con scontri armati in questa area del Caucaso. Questo ha subito attirato e sviato l'attenzione internazionale su questo altro focolaio.
- Incontro tra Xi Jinping e Putin. Keywords: xi jinping, jinping, xijinping, china russia. L'incontro tra i due presidenti e alleati ha voluto segnalare un rafforzamento delle relazioni tra i due paesi con implicazioni significative per gli equilibri geopolitici globali, specialmente nel contesto del conflitto in Ucraina e il segnale che le attuali difficoltà russe non sono sufficienti per gli accordi di pace.
- Zelensky visita le città liberate. Keywords: zelenskyy, visits, president. Il presidente Zelensky visita le città recentemente liberate dall'occupazione russa. Con queste visite simboliche si cerca di rafforzare il morale della popolazione e dimostrare il controllo del governo ucraino sui territori riconquistati.

In questo caso la situazione è dicotomica. Da una parte troviamo la questione ucraina, dall'altra l'alto interesse legato al focolaio nel Nagorno-Karabakh. In particolare i due topic si trovano ad avere le stesse dimensioni per quanto riguarda la frequenza. Questa divisione a metà è sugellata dall'intertopic distance in Figura 3.14.

Abbiamo 2 chiari blocchi. Quello in alto che parla di ucraina e delle fosse comuni e quello in basso che parla della guerra del Nagorno Karabakh e della questione Cino-russa.

Volendo valutare come la situazione di dipani nel tempo notiamo alcune particolarità, in Figura 3.15. L'avvenuta conoscenza delle fosse comuni, evento macabro e di grande impatto

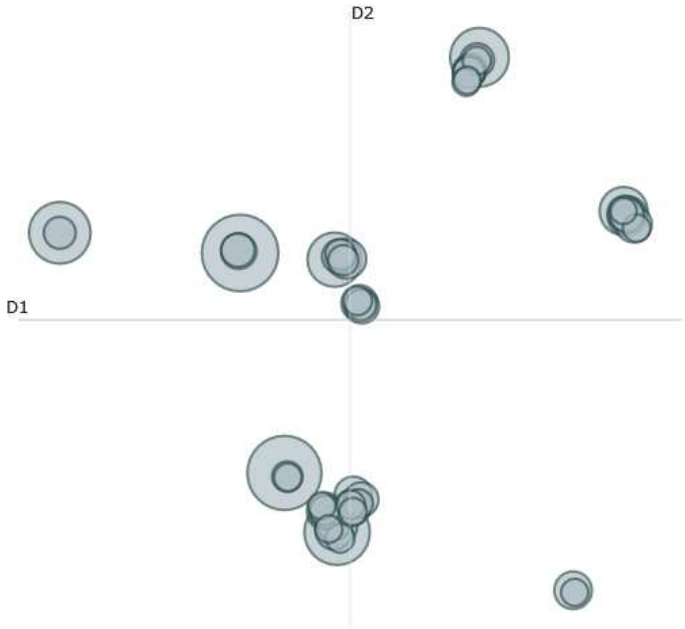

Figura 3.14: Intertopic distance dei giorni dal 12 al 18 settembre

emotivo, viene istantaneamente subissato dalla questione Azero-armena. Solo pochi giorni dopo, con la caduta di interesse ci si riappropria della questione ma con un interesse minore. Si può affermare, vista la posizione della guerra, le forze in campo e il loro passato che essa sia legata al sobillamento da parte della Russia. In parte per sviare l'attenzione sulle sconfitte recenti, in parte per evitare che tutta la comunità occidentale si focalizzi sull'evento di Izyum.

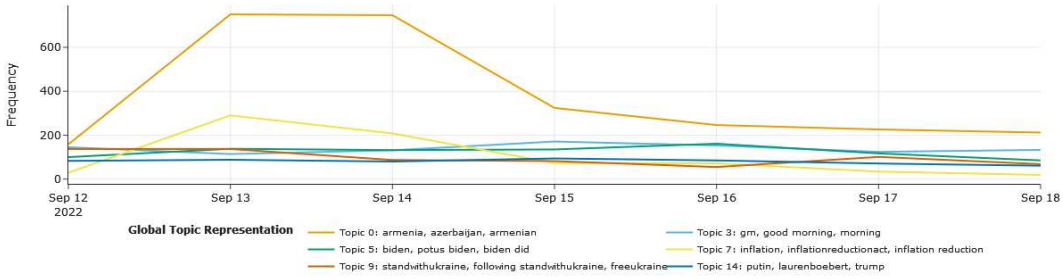

Figura 3.15: Evoluzione dei topic nel periodo tra il 12 e il 18 settembre

3.5.2 Analisi di un Evento dello Stallo

Data la maggiore ampiezza della seconda parte del dataset e l'elevata richiesta di potenza computazionale di BERTopic, si è resa necessaria una riduzione del periodo di analisi da 7 a 3 giorni pre e post evento. Questo permette un'analisi gestibile, mantenendo una prospettiva sufficientemente ampia per cogliere eventi e dinamiche

Per identificare gli eventi più significativi da analizzare, abbiamo esaminato attentamente i picchi di engagement nel dataset. Due picchi particolarmente evidenti si sono verificati il 25 e il 27 febbraio, segnando l'inizio di una nuova fase del conflitto. In particolare, l'anniversario della guerra, l'invio dei carri armati Leopard e un cospicuo pacchetto di aiuti. Questi picchi sono caratterizzati da un notevole aumento del flusso di utenti su X e dalle modifiche del

processo di collezione dei tweet. Questo oltre ad essere un momento topico della fase della guerra essendo l'evento con il maggior numero di tweet disponibile.

Un altro periodo significativo è quello relativo al periodo tra 17 e 22 marzo. Il 17 Marzo a seguito della sentenza della corte dell'Aia. Ora è interessante notare come sia uno dei pochi casi di discrepanza tra numero di utenti e Engagement in cui, il 17 Marzo, come evento non ha prodotto un corrispondente picco di engagement, suggerendo che, nonostante l'interesse iniziale, non ha stimolato una discussione sostenuta. Al contrario, il 22 marzo ha visto un picco di engagement, probabilmente in risposta alle gravi minacce di guerra nucleare. Questo periodo coincide anche con la visita del presidente cinese Xi Jinping in Russia. È interessante notare che poco dopo l'annuncio di questa visita, la Corte Penale Internazionale ha emesso un mandato di arresto per Vladimir Putin, sancendo ancora maggiormente uno strappo nell'ambito della politica internazionale.

Un ultimo picco significativo, che giunge alla fine del dataset, si è verificato il 6 giugno con il crollo della diga di Nova Kakhovka, un evento che ha avuto profonde implicazioni umanitarie e strategiche.

Abbiamo deciso di analizzare il picco di Febbraio anche per la sua importanza a livello storico il suo impatto sulla dinamica del conflitto, in contrapposizione con una decisione esterna come quella della corte internazionale dell'Aia. Questo permette di esaminare come eventi di natura diversa influenzino il discorso pubblico e l'engagement online.

La presenza di due picchi ravvicinati a febbraio ci ha portato a focalizzare l'analisi sul periodo dal 22 al 24 febbraio 2023. Inoltre per motivi di grandezza del dataset abbiamo dovuto lavorare solo sulla parte inglese attraverso il filtraggio per lingua dei tweet e su di una finestra temporale di 3 giorni. Il risultato di BERTopic è riportato in Figura 3.16.

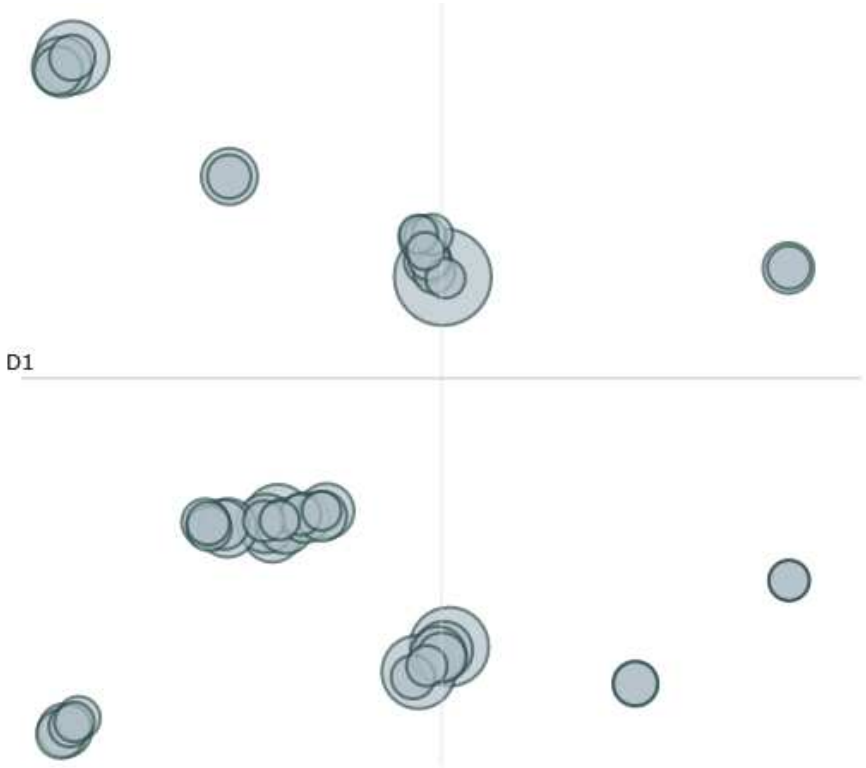

Figura 3.16: Interopic distance dei giorni precedenti al 25 febbraio

La notizia principale è data dall'anniversario del primo anno della guerra. In questo

caso BERTopic non è riuscito ad eseguire una clusterizzazione corretta, questo nonostante i tentativi di ottimizzazione, inclusa la riduzione del volume di tweet e una pulizia del testo. In particolare, gran parte del contenuto relativo all'anniversario è stato raggruppato nel topic indefinito con la grande maggioranza della parte concernente il primo anno di guerra.

Nonostante queste criticità, sono emersi dei topic interessanti:

- Critiche all'amministrazione Biden: Il topic riguarda principalmente questioni interne agli stati uniti, con critiche e richieste di impeachment per il presidente. Questo ci indica che il peso politico degli USA nella guerra è di primaria importanza, sia per il sostegno militare sia per l'ampio impatto politico.
- Tensioni in Trasnistria: Il topic riflette le preoccupazioni per le tensioni nella zona separatista e filorussa della Moldavia. Suggerendo i timori per un possibile nuovo fronte della guerra, essendo la regione confinante con l'Ucraina.
- Risoluzone ONU: Il topic parla della risoluzione dell'ONU che spinge per un cessate il fuoco e la fine della guerra. In particolare la Cina, si è astenuta e poco dopo si è incontrata con lo stesso Puntin, lanciando un forte segnale alla comunità internazionale.
- Critiche a Zelensky: Al contrario della narrativa legata ai primi mesi di guerra, il topic contiene le critiche verso il presidente ucraino, con addirittura accuse di tirannia. Questo svela un'evoluzione della percezione della popolazione ucraina e della fiducia accordata.

L'analisi di BERTopic mostra una situazione molto differente rispetto alla fiducia e all'unità rispetto alle fasi iniziali del conflitto. Tutto ciò può essere considerato normale a causa dell'incredibile pesantezza della guerra e delle sue conseguenze.

l'heatmap infatti risulta particolarmente scorrelata. Con pochi topic interessanti come possiamo vedere in figura 3.17, essa conferma l'analisi fatta, dal quale possiamo vedere come il blocco più simile è quello riguardante la politica estera e solo sucessivamente quello legato al conflitto. Indice sia dell'importanza su scala globale del conflitto sia della difficoltà di clusterizzazione.

Andando ad osservare la intertopic distance in Figura 3.18 ci accorgiamo di come non ci sia un particolare filo conduttore e della dispersione nell'identificare temi e trend coerenti tra loro. La dispersione dei temi, esistente anche negli stessi cluster di topic è segnalata dalla presenza di argomenti non correlati come Justin Trudeau e il world Pancake Day che equivale al martedì grasso. La frammentazione è anche sintomo del rallentamento della guerra avvenuto nei mesi precedenti e dello spostamento dei riflettori mediatici e dalla difficoltà di mantenerlo per un enorme periodo di tempo.

Valutiamo se l'analisi attraverso l'evoluzione della quantità di tweet per topic nel tempo può offrirci prospettive più chiare, andando ad identificare dei trend emergenti. Il problema più rilevante in questo caso è dato dal gestire l'elevato volume di tweet senza avere la potenza necessaria. Per ovviare al problema il modello è stato allenato nuovamente dopo esser andati a modificare il dataset, come precedentemente descritto.

Con questo approccio si possono distinguere 2 tendenze in salita. In particolare con lo sviluppo nel tempo notiamo che il topic legato al minuto di silenzio per l'anniversario dell'inizio della guerra, che nelle altre visualizzazioni e anche nell'originale divisione in topic non era riuscito ad essere presente, appare. Inoltre risulta chiaro come ci siano delle trattative di pace che iniziano ad apparire tra i topic.

Controlliamo se esse possano fungere da predittore per gli eventi dei giorni successivi.

Il giorno successivo l'analisi con BERTopic mostra invece un'eccellente livello di clusterizzazione, dove i topic sono ben definiti e privi di elementi estranei. In particolare possiamo notare:

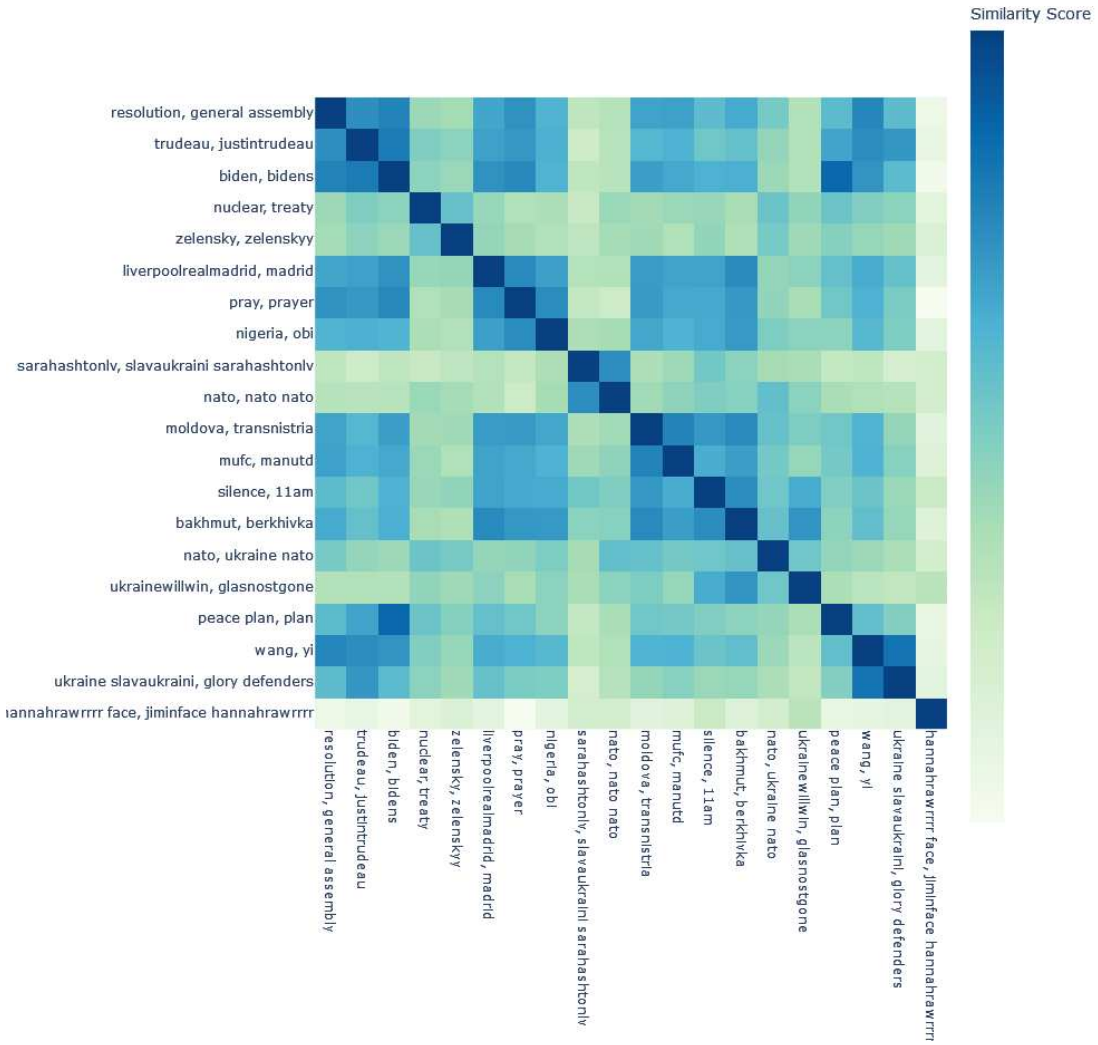

Figura 3.17: Heatmap dei giorni compresi tra il 22 e il 24 febbraio

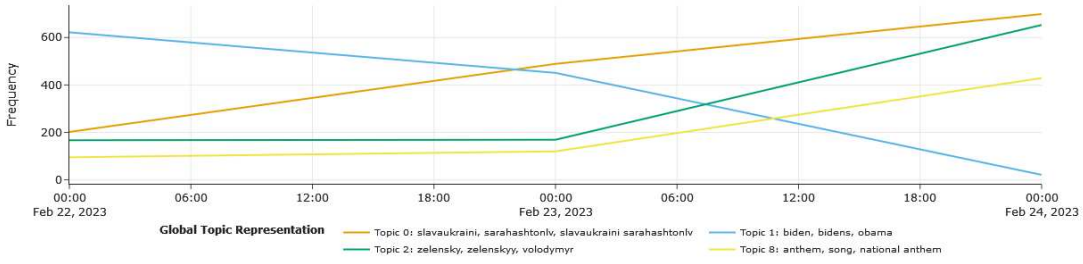

Figura 3.18: Evoluzione dei topic nei giorni tra il 22 e il 24 febbraio

- Isolamento della Russia. Keywords: osce pc, reading statement, statement mtg, emptied immediately. Questo topic parla dell'assemblea dell'OSCE[5] nella quale la Russia risulta essere isolata completamente, con un'uscita da parte di tutti i membri quando la Russia ha preso parola.

[5]Organizzazione per la sicurezza e la cooperazione in Europa (OSCE), la più importante organizzazione intergovernativa riguardante la sicurezza europea

- Gratitudine per i Leopard. Keyword: wheels thank,germany cat,ukraine weareallukrainians,freeukraine. Il topic si collega ai ringraziamenti per l'arrivo dei carri armati Leopard da parte della Germania che era stata a lungo criticata per la mancanza di interesse nell'invio. Tra le keyword è interessante notare come essi siano appellati ironicamente come gatti, correttamente notato dalla clusterizzazione.

- Polonia e problemi. Il topic si collega al discorso di Zelensky, nel quale i media polacchi hanno erroneamente rivelato la controfigura, creando un certo problema per la sua sicurezza.

- Anniversario della resistenza. Keywords: year functioning,today btw,btw year[6]. Il topic si collega all'anniversario della guerra, con i complimenti e il supporto al parlamento per esser riusciti a resistere in questi momenti difficili.

- Discorso del presidente Zelensky: il topic si collega chiaramente agli altri e riguarda il discorso del presidente Zelensky.

- Strategia difensiva a Bakhmut: questo topic è interessante perchè anticipa ed evidenzia la debolezza della strategia e della logistica russa sulla città di Bakhmut, punto focale della guerra. "The Ukrainian, primary strategy in Bakhmut seems to be mowing down Russian troops with machine cannons and machine guns. And they do this quite efficiently. The lack of Russian antitank gear is apparent. Russians continue being pure cannon fodder."[7]

Questa analisi rivela una narrazione dinamica del conflitto, capace di catturare sia gli eventi simbolici legati all'anniversario della guerra sia all'evoluzione sul campo che alle personalità chiave. L'eccellente clusterizzazione può essere notata ancora una volta dall'intertopic distance in Figura 3.19.

Possiamo facilmente notare come, tra i primi 50 topic abbiamo un blocco estremamente ben definito dove ricadono la maggior parte di essi. Solo pochi topic sono scorrelati e di poco conto.

Per quanto concerne l'analisi post evento, essa ha presentato sfide significative a causa della grande richiesta di potenza computazionale. Data la presenza di due picchi ravvicinati si è deciso di andare a valutare se essi sono visibili dall'analisi.

L'analisi dei topic riesce ad essere significativa ma dispersiva. In particolare possiamo trovare:

- Conferenza OSCE: il primo topic è uguale a quello del 25 febbraio che mantiene ovviamente la sua rilevanza.

- Esumazione ad Izyum: il secondo topic parla della conclusione dell'esumazione delle fosse comuni nella città di Izyum, che riporta l'attenzione sulle atrocità commesse nei periodi di occupazione della città.

- Errore mediatico polacco: in questo caso uniamo due topic riguardano sempre l'errore televisivo polacco ma che non sono stati accorpati insieme.

I restanti topic, sono concentrati principalmente sugli eventi legati al 25 febbraio, che indica una certa inerzia nella discussione. L'analisi della distanza tra i topic, come si può

[6] btw, acronimo di "By the Way", in italiano "comunque sia"

[7] Gli ucraini come strategia fondamentale sembrano falcidiare le truppe russe con cannoni e mitragliatrici, e lo fanno molto efficientemente. La mancanza di materiale per abbattere i carriarmati risulta essere apparente. La russia continua ad essere carne da cannone.

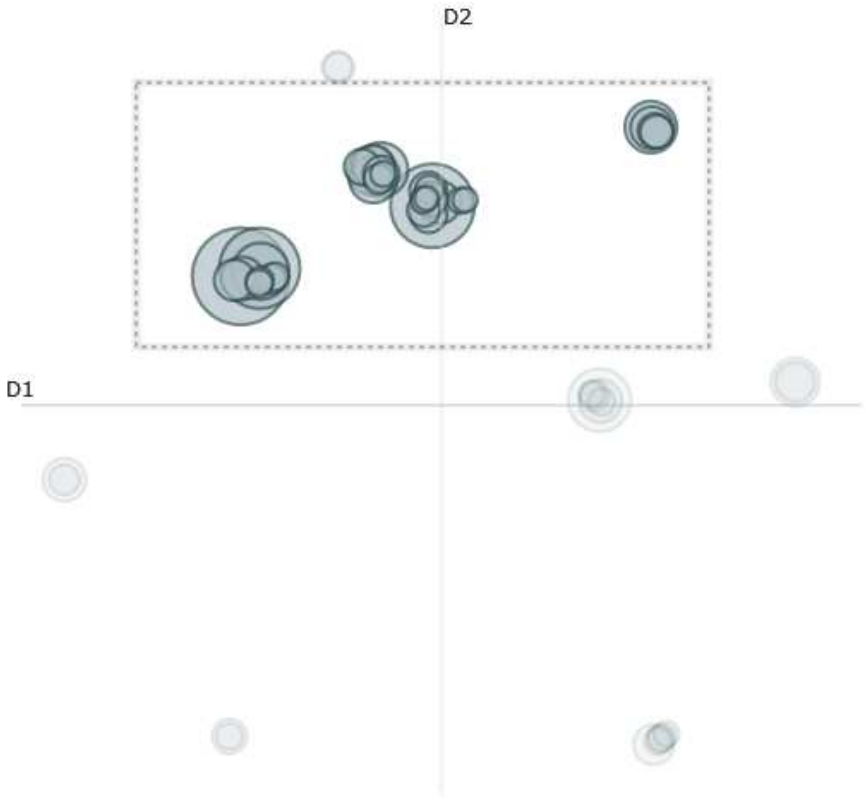

Figura 3.19: Interopic Distance del 25 Febbraio

vedere nella Figura 3.20. Nonostante l'analisi non abbia come scopo quella di distinguere i 2 topic, essendo essi semanticamente simili essa si è rivelata caotica e particolarmente dispersiva suggerendo una grande frammentazione dei trend nel periodo.

Volendo avere una visione più chiara delle motivazioni dei due picchi, andiamo ad eseguire sull'evoluzione dei topic durante il periodo di tempo considerato.

In questo caso è stata necessaria una pulizia più impegnativa rispetto a quelle precedenti:

- Rimozione dei link: La rimozione è dovuta alle difficoltà palesatesi nella mancata unificazione dei topic riguardanti l'errore mediatico polacco.
- Eliminazioni di hashtag e menzioni: Questo per diminuire il rumore di fondo data la grande mole di tweet.

La rimozione dei link, degli hashtags e delle menzioni ha effettivamente ridotto il carico di lavoro e ha permesso di trovare chiaramente la distinzione tra i due picchi, come possiamo vedere in Figura 3.21.

Grazie alle accortezze utilizzate è stato possibile portare a compimento l'analisi e portare alla luce chiaramente i due picchi:

- il primo picco, riguardante gli eventi del 25 febbraio, sono legati alla conferenza OSCE, sottolineando un importante evento diplomatico all'interno della comunità europea.
- il secondo picco è emerso come un rinnovato interesse per la situazione venutasi a creare ad Izyum

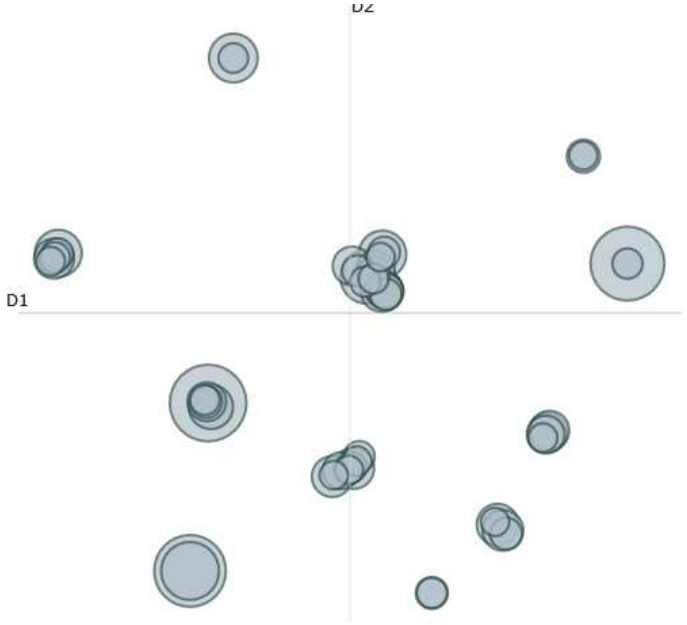

Figura 3.20: Interopic Distance dei giorni 25, 26 e 27 febbraio

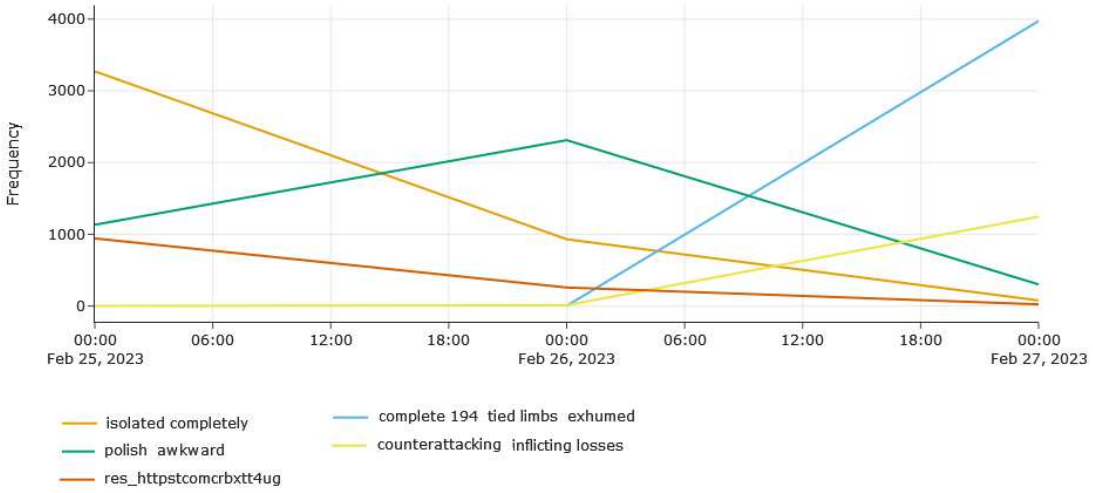

Figura 3.21: Evoluzione nel tempo dei topic dei giorni 25, 26 e 27 febbraio

Un'ultima analisi è stata effettuata, riguardante l'attività successiva a questi giorni ma senza trovare eventi di particolare interesse. Come abbiamo anche potuto notare non c'è stato un rinnovato interesse per le vicende dell'esumazione né altri eventi maggiori hanno preso piede.

A seguito delle analisi sull'intorno dell'evento possiamo asserire che è possibile predire in qualche misura gli eventi a partire dall'analisi dei tweet e dalla clusterizzazione. Infatti andando a sfruttare l'evoluzione nel tempo dei topic siamo riusciti sia in questo caso, sia nel precedente a scoprire eventi emergenti prima del loro massimo.

CAPITOLO 4

Discussione

L'analisi del dataset riguardante il conflitto ha fornito una panoramica generale sulle dinamiche degli utenti e degli eventi. In particolare su quelli più significativi e decisivi, sugli influencer chiave e sulle tendenze. Corredato assieme dall'analisi di eventi specifici al fine di prevedere trend emergenti.

Per quanto concerne l'analisi degli utenti con il maggior numero di tweet si è rivelato poco significativo e nella maggior parte dei casi inconcludente. La quasi totalità di essi si sono rivelati essere bot con scarsa utilità reale, con funzioni principalmente di aggregatori di notizie. Nonostante l'analisi e la ricerca dei bot sia un argomento di particolare importante, esso esula dalle nostre intenzioni iniziali.

Di conseguenza la quantità di tweet di un utente non è assolutamente un predittore valido. Per questo motivo la nostra attenzione si è spostata sull'engagement, come metrica più affidabile per l'individuazione di figure chiave tra cui nexta_tv, Tendar, strategywoman, UAWeapons, UAarmy_animals, IAPonomarenko e TheStudyofWar.

Essi sono effettivamente degli influencer in questo ambito, ognuno di essi ha largo seguito ed esprimono una buona commistione di personalità, troviamo infatti think thank strategici, giornalisti al fronte, televisioni, giornali e analisti militari.

Per la prima parte del dataset troviamo altri 3 utenti che hanno lasciato un segno cioè Blue_Sauron, SarahAshtonLV e UKR_token. Essi rimangono sempre utenti con alte interazioni ma non sono nei maggiori 10 per tutto il dataset. In particolare Blue_sauron è un analista militare mentre SarahAshtonLV è un sottoufficiale dell'esercito con valorose azioni di guerra mentre UKR_token è un giornale online.

Invece, per la seconda parte del dataset abbiamo altre 3 personalità che rientrano nel novero dei giornalisti e corrispondenti. In particolare essi sono GlasnostGone, youtuber e giornalista, jjaranaz94 ed HerryNapit. Un'eccezione interessante ed un unicum è rappresentata dall'ultimo dei tre, un giornalista yemenita con esperienza in zone di guerra, poiché è l'unico ad essere anche nel maggior numero di tweet per giorno, legato all'evento del 6 Giugno.

Per quanto riguarda l'identificazione di eventi significativi, sono emerse due metodologie efficaci. La prima si basa sulla sovrapposizione tra il numero di utenti unici e il numero di tweet totali, mentre la seconda invece riguarda la distribuzione dell'engagement giornaliero. Entrambe le metodologie sfruttano la ricerca dei massimi locali. Dalla prima si trovano in genere eventi ampiamente riconosciuti e ben delineati come il 25 febbraio, e dunque gli eventi legati all'anniversario della guerra e i nuovi pacchetti di aiuti. Dalla seconda metodologia

possiamo trovare degli eventi che rappresentano degli eventi più specifici e in genere legati a questioni tecniche, come quelli legati rispettivamente al 22 marzo e cioè la condanna della corte internazionale dell'Aia e al 9 aprile che riguarda la questione della vendita di energia elettrica.

Inoltre, l'analisi sugli hashtag ci ha dato una visione d'insieme delle tendenze generali. Il sentiment risultante è particolarmente negativo e questo è il risultato della natura particolarmente violenta e totale del conflitto, come esemplificato da "Bakhmut", che risulta essere il più negativo tra gli hashtag presenti, e dalla situazione venutasi a creare. Anche gli hashtag che sono di supporto alla causa ucraina non superano mai la soglia della neutralità di VADER posta a 0.05. È interessante notare la discrepanza tra hashtag in lingue diverse, come nel caso di "Russland", in tedesco che presenta tweet correlati con sentiment più negativo rispetto al suo equivalente inglese, così come successo per così come "USA" e "america".

Un'ipotesi potrebbe essere data dalla forte presenza di bot che hanno poco engagement ma permettono di dare una rappresentazione ancora più negativa. Questa idea è supportata dalla questione degli hashtag di lingue differenti.

Il prossimo passo è quello di rispondere alla domanda se l'analisi degli intorni di un macroevento possa predire lo stesso dal modo in cui gli utenti interagiscono. Ricordando che si è osservata una variazione significativa nel numero di tweet e nell'engagement nella seconda parte dell'analisi. Il cambiamento può essere attribuito in parte all'arrivo dell'inverno, che ha comportato una diminuzione degli eventi, in parte alle modifiche del web crawler. Inoltre l'analisi è dipesa, in particolare nella seconda parte del dataset dalle difficoltà relative alla grandezza dello stesso. Per ovviare ad esse sono state eseguite delle modifiche legate alla quantità di tweet e alla pulizia degli stessi.

Da una lato abbiamo potuto osservare come l'evoluzione dei topic possa fungere da predittore di eventi futuri nel tempo. Questo è legato al caso della liberazione di Izyum. Già due giorni prima dell'effettiva liberazione, essa ha guadagnato rilevanza nei topic, con un velocissimo aumento fino al massimo dell'11 settembre, giorno effettivo della liberazione. Tuttavia l'altro evento che ha catalizzato l'attenzione è stato il riaccendersi del conflitto nel Nagorno-Karabakh. Questa guerra, distinta per modi e motivazioni da quella ucraina è balzata agli oneri della cronaca e con una posizione di rilievo nei topic. La guerra che ha profonde radici culturali e storiche è stata teatro di scontri ad intermittenza da più di 30 anni. Anche se dal 2020 in poi non c'erano stati grandi movimenti. La concomitanza di questo evento, con gli eventi di Izyum, solleva ipotesi sulla possibile presenza di guerra ibrida e manipolazione mediatica.

L'analisi ottenuta nel secondo evento risulta altresì interessante. Innanzitutto abbiamo potuto osservare ancora una volta come alcuni eventi fondamentali possano essere anticipati. Nel nostro caso la principale notizia risulta essere legata alla conferenza OCSE, alle ripercussioni diplomatiche e all'anniversario della guerra. la capacità di predire eventi in base ai tweet è sempre presente, possiamo infatti notare come la quantità di tweet che parlano della conferenza stessa aumenti con il passare dei giorni, allo stesso modo nella proposta. Questo però risulta essere in misura molto minore che rispetto al caso precedentemente analizzato, suggerendo una dipendenza dal tipo di evento. La presenza di un secondo picco molto ravvicinato ha permesso di valutarli congiuntamente. Scoprendo così un collegamento con il primo evento e cioè la fine dell'esumazione dei corpi nella città di Izyum. A supporto dell'importanza sia dell'analisi dei topic che della loro visualizzazione nell'emergere di eventi non banali e non immediati.

CAPITOLO 5

Conclusioni

In questa tesi abbiamo esplorato il ruolo fondamentale dei social network, come luogo per la discussione di eventi e dello scambio di opinioni. Grazie al fatto che essi abbiano trasformato radicalmente il modo in cui le persone interagiscono. La capacità di X di veicolare messaggi rapidamente e ad una platea vasta è stato particolarmente utile. La rapida diffusione degli eventi legati alla guerra, il flusso continuo di informazioni in tempo reale ha permesso di analizzare la sempre mutevole percezione pubblica del conflitto. Tuttavia, la difficoltà intrinseca di analizzare un dataset di un evento di così ampio respiro e variabilità, ha reso necessaria la scelta di strumenti adeguati.

A partire dalla scelta dei periodi da analizzare, ai problemi risultanti dalle modifiche esterne legate alle scelte di X, alla scelta degli strumenti da utilizzare per la sentiment analysis e il topic modeling. Nelle analisi generali abbiamo portato alla luce la struttura del dataset, la questione dei bot e la ricerca legata all'engagement per distinguere tra attività artificiale e quella reale, andando a trovare persone capaci di influenzare le opinioni. Successivamente, nelle analisi relative agli hashtags abbiamo trovato una generalizzata negatività attraverso la sentiment analysis dei tweet collegati. Inoltre abbiamo potuto apprezzare come ci siano state evoluzioni, generalmente negative, durante il periodo considerato. Oltre alla presenza di differenti sensibilità ai diversi paesi. Nell'analisi dell'evento legato alla controffensiva siamo riusciti a trovare un modo efficace per predire gli eventi pochi giorni prima del loro accadimento ed abbiamo notato come sia visibile la manipolazione mediatica post evento. Questo è un risultato dello spostamento dell'interesse internazionale.

Nell'analisi del secondo evento, legato allo stallo, invece abbiamo analizzato allo stesso modo gli eventi ma interessandoci anche dell'analisi dei due picchi ravvicinati. Nell'analisi pre evento c'è dell'interesse verso l'importante conferenza che ha sancito lo strappo della Russia nel panorama diplomatico alla vigilia dell'anniversario della guerra. Successivamente a questo, gli eventi delle due analisi si sono riunificati nella fine delle esumazioni nella città di Izyum, evento cardine dell'analisi della controffensiva, che ha riportato con rinnovato interesse alla questione legata alle vittime civili.

L'analisi di un evento così attuale e dinamico non può certamente fermarsi qui. In primo luogo la naturale estensione di questa tesi è legata alla disamina ed analisi della parte più recente del conflitto, sino ad arrivare al momento attuale, permettendoci di tracciare l'evoluzione della narrativa e della percezione pubblica. In secondo luogo, dati i problemi riscontrati, sarebbe di grande utilità andare ad eseguire un'analisi specifica per identificare i bot e/o account che alimentano la disinformazione. Infine, andare ad utilizzare strumenti ancora più

sofisticati e moderni per eseguire un'analisi multimodale. Permettendo di includere oltre al testo anche video, immagini ed audio, largamente presenti nei social e di grande utilità ed importanza per riuscire a dare una rappresentazione ancora più dettagliata e sfaccettata degli eventi.

Bibliografia

ALIEVA, I., NG, L. H. X. e CARLEY, K. M. (2022), «Investigating the spread of Russian disinformation about biolabs in Ukraine on Twitter using social network analysis», in «2022 IEEE international conference on big data (big data)», p. 1770–1775, IEEE. (Cited at page 1)

CHIN, A. e CHIGNELL, M. (2007), «Identifying communities in blogs: roles for social network analysis and survey instruments», *International Journal of Web Based Communities*, vol. 3 (3), p. 345–363. (Cited at page 1)

DEVLIN, J., CHANG, M.-W., LEE, K. e TOUTANOVA, K. (2018), «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*.

GROOTENDORST, M. (2022), «BERTopic: Neural topic modeling with a class-based TF-IDF procedure», *arXiv preprint arXiv:2203.05794*.

HU, Y., FARNHAM, S. e TALAMADUPULA, K. (2015), «Predicting user engagement on twitter with real-world events», in «Proceedings of the International AAAI Conference on Web and Social Media», vol. 9, p. 168–177.

HUTTO, C. e VADER, E. G. (2014), «a parsimonious rule-based model for sentiment analysis of social media text. 2014», *DOI: https://doi. org/10.1609/icwsm. v8i1*, vol. 14550, p. 216–225.

KURKA, D. B., GODOY, A. e VON ZUBEN, F. J. (2015), «Online social network analysis: A survey of research applications in computer science», *arXiv preprint arXiv:1504.05655*. (Cited at page 1)

MAKHORTYKH, M. e SYDOROVA, M. (2017), «Social media and visual framing of the conflict in Eastern Ukraine», *Media, war & conflict*, vol. 10 (3), p. 359–381. (Cited at page 1)

MCDERMOTT, R. N. (2016), «Brothers Disunited: Russia's use of military power in Ukraine», in «The Return of the Cold War», p. 77–107, Routledge. (Cited at page 6)

MIR, A. A., RATHINAM, S., GUL, S. e BHAT, S. A. (2023), «Exploring the perceived opinion of social media users about the Ukraine–Russia conflict through the naturalistic observation of tweets», *Social Network Analysis and Mining*, vol. 13 (1), p. 44. (Cited at page 1)

PRINCE, S. J. (2023), *Understanding deep learning*, MIT press.

RACEK, D., DAVIDSON, B. I., THURNER, P. W., ZHU, X. X. e KAUERMANN, G. (2024), «The Russian war in Ukraine increased Ukrainian language use on social media», *Communications Psychology*, vol. 2 (1), p. 1. (Cited at page 1)

RUSH, A. M. (2018), «The annotated transformer», in «Proceedings of workshop for NLP open source software (NLP-OSS)», p. 52–60.

SUFI, F. (2023), «Social media analytics on Russia–Ukraine cyber war with natural language processing: Perspectives and challenges», *Information*, vol. 14 (9), p. 485. (Cited at page 1)

TSVETOVAT, M. e KOUZNETSOV, A. (2011), *Social Network Analysis for Startups: Finding connections on the social web*, " O'Reilly Media, Inc.".

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. e POLOSUKHIN, I. (2017), «Attention is all you need», *Advances in neural information processing systems*, vol. 30.

Ringraziamenti

Il primo ringraziamento va a Mamma che mi ha sostenuto e dato la forza per tutto questo tempo e a Papà, senza il quale non mi sarei mai interessato a questi argomenti.

Vorrei ringraziare il mio amico e correlatore Luca che ha reso possibile questa tesi e il professor Ursino.

Poi non posso non ringraziare ogni persona che mi è stata accanto durante questo lunghissimo percorso e che in un modo o nell'altro ha contribuito a farmi diventare la persona che sono:

Mirko, Gianluca, Jacopo, Lucio, Andrea, Roberto, Paolo, Camilla, Madda, Beatrice, Anna, Mery, Giulio, Vincenzo, Roman, Bruno, Lollo, Francesco, Filippo, Chiara, Bettina, Alessio, Ernesto, Ilaria, Alessandro, Melanie.