data

MDPI

Article

Multimodal Hinglish Tweet Dataset for Deep Pragmatic Analysis

Pratibha [1], Amandeep Kaur [1], Meenu Khurana [2] and Robertas Damaševičius [3,*]

[1] Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140601, Punjab, India; pratibha@chitkara.edu.in (P.); amandeep@chitkara.edu.in (A.K.)
[2] Chitkara University School of Engineering and Technology, Chitkara University, Baddi 173205, Himachal Pradesh, India; meenu.khurana@chitkara.edu.in
[3] Department of Applied Informatics, Vytautas Magnus University, LT-53361 Kaunas, Lithuania
* Correspondence: robertas.damasevicius@vdu.lt

Abstract: Wars, conflicts, and peace efforts have become inherent characteristics of regions, and understanding the prevailing sentiments related to these issues is crucial for finding long-lasting solutions. Twitter/'X', with its vast user base and real-time nature, provides a valuable source to assess the raw emotions and opinions of people regarding war, conflict, and peace. This paper focuses on collecting and curating hinglish tweets specifically related to wars, conflicts, and associated taxonomy. The creation of said dataset addresses the existing gap in contemporary literature, which lacks comprehensive datasets capturing the emotions and sentiments expressed by individuals regarding wars, conflicts, and peace efforts. This dataset holds significant value and application in deep pragmatic analysis as it enables future researchers to identify the flow of sentiments, analyze the information architecture surrounding war, conflict, and peace effects, and delve into the associated psychology in this context. To ensure the dataset's quality and relevance, a meticulous selection process was employed, resulting in the inclusion of explanable 500 carefully chosen search filters. The dataset currently has 10,040 tweets that have been validated with the help of human expert to make sure they are correct and accurate.

Keywords: hinglish; pragmatic analysis; sentiment analysis; tweet dataset

check for updates

Citation: Pratibha; Kaur, A.; Khurana, M.; Damaševičius, R. Multimodal Hinglish Tweet Dataset for Deep Pragmatic Analysis. *Data* **2024**, *9*, 38. https://doi.org/10.3390/data9020038

Academic Editor: Fabio Crestani

Received: 11 November 2023
Revised: 29 December 2023
Accepted: 1 February 2024
Published: 15 February 2024

1. Introduction

In the current digital era, where platforms of mass communication have risen to prominence, Twitter/'X' has emerged as a modern platform of public discourse [1]. The dialogue surrounding warfare [2], conflict, protest [3] and the pursuit of peace is of particular note, as it reflects the intricate web of human emotions both positive (empathy, support) and negative (anger, hate [4,5]) that are inextricably linked to these profound issues. The expression of sentiments [6,7] on this platform transcends the boundaries of ephemeral commentary and can include fakes or disinformation [8]. It serves as a barometer for the prevailing societal attitudes towards the most pressing issues of war and peace. The task of deciphering and understanding these digital exchanges is a scholarly pursuit of the highest order, bearing the potential to inform and refine the strategies that underpin conflict resolution [9] and the construction of peace. The nature of discourse on Twitter/'X', with its immediacy and the depth of reflection it permits, provides a unique lens through which the complex relationship between public sentiment and global events can be observed and understood.

Hinglish is a language that is spoken by large propotion of nation people. It is made up of Hindi and English. It's mostly used on social media, and conversation with it has become more popular in the last few years. The amount of Hinglish material on the Internet has grown a lot since the rise of social media. This lets researchers look into and study the language for many reasons, including text pragmatic analysis and mood analysis. Sentiment analysis is the process of figuring out how someone feels about something written down [10,11]. It can be used for many things, like analysing customer feedback, keeping an eye on a brand, or keeping an eye on social media. Still, Hinglish sentiment analysis might be hard because it uses two different languages and rules and frameworks that aren't usually used in language [12].Text pragmatic analysis, on the other hand, looks at how language is used to get something done in a certain setting. To do this, you have to look at a text's meaning beyond its formal meaning [13], taking into account the speaker's intentions, the audience, and the social and cultural setting in which the work is created [14].

The lack of public data sets, particularly for pragmatic studies, is one of the largest hurdles to the analysis of hybrid languages such as Hinglish [15,16]. It is difficult to construct accurate algorithms for sentiment and pragmatic analysis in the absence of appropriate data such as for low-resourced language [17]. Beside this, the complexity of Hinglish, which mixes two distinct grammatical systems and vocabularies, makes it challenging to develop suitable models for machine learning techniques [13]. Lack of standardisation is yet another big issue [18–20]. Hinglish is an informal language and spoken as colloquial, and distinct varieties of Hinglish may exist in different locations and ethnic groupings. This can lead to differences in syntax, spelling, and grammar. In order to construct models that effectively analyse Hinglish, it is necessary to have a comprehensive understanding of the language's variants and nuances [21,22].

To overcome these obstacles, it will be necessary to curate more comprehensive datasets, and so exact understanding becomes easy [23,24]. As social media continues to explore, scholars will have more possibilities to explore and analyse the language, ultimately leading to a greater knowledge of Hinglish and its application in various domain. In the next section, we try to paint the landscape of all aspects related to these subjects with the help of existing documentary evidence.

2. Analysis of Existing Datasets

Table 1 provide the list of some of the datasets that are publicly available in the public domain. It can be observed that the dataset that contains content related conflicts, war, crisis etc. are limited, especially in mix code languages.

Table 1. Dataset Description.

Dataset Name	Source	Time Period	Number of Tweets	Language
Russo Ukrainian War2 [26]	Twitter/'X'	22 February 2022	57,384,192	English Russia
Gaza Conflict [27]	Twitter/'X'	5 July to 26 August 2014	49,205,389	English, Spanish, Indonesian, French
War Between Ukraine and Russia [28]	Twitter/'X'	22 February 2022 to 8 January 2023	454,488,445	English, French, Italian, German
Tweets discussing the Russia/Ukraine War [29]	Twitter/'X'	23 February and 8 March 2022	5,203,764	English
End of US-Afghan War Tweet Data [30]	Twitter/'X'	11 August 2021 to 27 August 2021	359,904	English
Ukraine-Russia [31]	Twitter/'X'	22 February 2022 through 8 March 2022	63 millions	English, French, Italian, German, Ukrainian
Russia-Ukraine war-Tweets Dataset [32]	Twitter/'X'	31 December 2021 to 3 March 2022	1,316,605	English, Spanish, French, German
Kunduz Madrassa attack by Afghan [33]	Twitter/'X'	2 April 2018 to 8 April 2018	7500	English
Sarin Gas Massacre in Syria [34]	Twitter/'X'	August 2013	4 million	English

Table 1. *Cont.*

Dataset Name	Source	Time Period	Number of Tweets	Language
Hope speech detection in YouTube comments [35]	Youtube	November 2019 to June 2020	59,354	English, Tamil, and Malayalam
Hindu Nationalism Online: Twitter/'X' as Discourse and Interface [36]	Twitter/'X'	starting in December 2019 and concluding in May 2020	20,370,555	English
Kashmir Conflict [37]	Twitter/'X'	2 December 2019 to 12 January 2020	60k	English
HinGE A Dataset for Generation and Evaluation of Code-Mixed Hinglish Text [38]	Twitter/'X'	–	10,731	Hinglish
PHINC A Parallel Hinglish Social Media Code-Mixed Corpus for Machine Translation [39]	Twitter/'X' Facebook	–	13,738	Hinglish

It is obvious from the existing set of datasets that building repositories for pragmatic text analysis has been ignored by contemporary researchers. Foremost among these obstacles is the expensive nature of developing datasets [40,41]. Time-consuming and costly, the production of high-quality datasets for pragmatic analysis requires the collecting and annotation of massive volumes of data. Annotating text for pragmatic analysis necessitates a high level of consensus among annotators, which makes consistency in annotation a crucial issue that merits considerable consideration. Depending on their background and expertise, various annotators may interpret language usage differently. Privacy is a further concern that must be made, given that text datasets may contain sensitive information. Thus, precautions must be implemented to protect the privacy and secrecy of the data.The section gives the formulation of problem that can be derived from these current issues.

3. Problem Undertaken

The study of pragmatic behaviour analysis on Twitter/'X', such as the use of indirect speech acts or figurative language, holds significant potential to enhance the integrity of social media platforms by detecting and categorising tweets with uncommon communication patterns into a structured dataset [42]. The primary aim of this investigation is to enhance data availability and transparency, for Twitter/'X' researchers without employing advanced analytical techniques such as data mining, machine learning, and deep learning algorithms, by identifying diverse potential applications through the implementation of data analysis methodologies [41,43,44]. Hence, mathematically

Let T be the set of all tweets on Twitter/'X' . Let P be the set of peculiar tweets with communication patterns related to war and conflicts. The problem is to discover P from T and curate them into a structured dataset using of tweets, such as the use of indirect speech acts or figurative language. The aim of this research is to enhance data availability, transparency by identifying and categorising peculiar tweets in P with the help of five deep pragmatic analysis principles. The research aims to achieve this without using advanced analytical techniques like data mining, machine learning, and deep learning algorithms i.e through the process of human expert annotation. Formally, the problem can be stated as: Given T, identify and extract P using pragmatic behaviour analysis on Twitter/'X', and curate them into a structured dataset such that their nature can be categorised with a metric m. The research also aims to identify potential applications of P through the implementation of data analysis and curation methodologies. The potential applications P may include pragmatic and sentiment analysis s of world war related tweets.

Let T be the set of all tweets on Twitter/'X', and let $P \subset T$ be the subset of tweets that exhibit peculiar communication patterns, specifically those related to the concept of "world war III and conflicts".

3.1. Problem Definition

1. Identification Problem: Define T as a non-empty set, $T = \{t_1, t_2, \ldots, t_n\}$, where each t_i represents an individual tweet. Define P as a subset of T, $P = \{t_i \in T \mid t_i$ exhibits peculiar communication patterns as defined by a set of pragmatic criteria $C\}$. The problem is to determine a function $f : T \to \{0, 1\}$ such that for any tweet $t_i \in T$, $f(t_i) = 1$ if $t_i \in P$, and $f(t_i) = 0$ otherwise.
2. Curation Problem: Upon identification, curate P into a structured dataset D, where D is a finite ordered list of elements from P, $D = [p_1, p_2, \ldots, p_k]$, and $k \leq n$. This curation must be performed using pragmatic behaviour analysis, which includes the identification of indirect speech acts, figurative language, and other relevant pragmatic phenomena without the employment of advanced analytical techniques such as data mining, machine learning, or deep learning.
3. Categorial Problem: Define a category $m : P \to \mathbb{R}$ that quantifies based on the principles of deep pragmatic analysis of tweets within P. The category m should be able to measure the prevalence of the peculiar communication patterns within T in a transparent and reproducible manner.
4. Application Problem: Identify potential applications A of the curated dataset D, where A may include, but is not limited to, sentiment analysis s. The sentiment analysis s is a function $s : P \to S$, where S is the set of possible sentiment categories, and it aims to classify the sentiment of each tweet in P as it pertains to the topic of world war and conflicts.

3.2. Formal Problem Statement

Given the set T, the problem is to identify and extract a subset P using pragmatic behaviour analysis, and to curate this subset into a structured dataset D such that the nature of tweets in P can be categorised with a deep pragmatic analysis. Furthermore, the research seeks to explore the potential applications of the dataset D, including but not limited to the pragmatics of sentiment analysis s, with the aim of enhancing data availability and understanding of the communication patterns related to world war III on Twitter/'X'. This is to be achieved through human expert feedback and deep pragmatic analysis, eschewing the use of advanced computational analytical techniques.

4. Methodology

In this section, we write a systematic process of gathering and extracting pertinent data from the Twitter/'X' platform. We also expound on how the seed tags were meticulously chosen, and how these seed keywords were subsequently employed to gather a more extensive pool of data from the Twitter/'X' platform for better understanding. The ultimate goal is to create a gold standard dataset that would enable us to monitor the public sentiment/emotion concerning the world wars and conflicts.

The steps below will explain the methodology. It can be observed from the Figure 1 that the methodology consists of 8 steps. The first step is to define the scope for the collection of data. The second step is to construct seed keywords and search filters for the extraction of data. The third step is to run pre-processing operation on the tweets. In fourth step topic analysis with the help of algorithms such as LDA is done.From the outcome of LDA, the fifth step is convert pragmatic analysis based on 5 principles. The sixth step is assign labels based on topic inferring and deep pragmatic analysis.The seventh step is to validate the label with human expert evaluation. The eighth step is to publish data after validating through a case study.

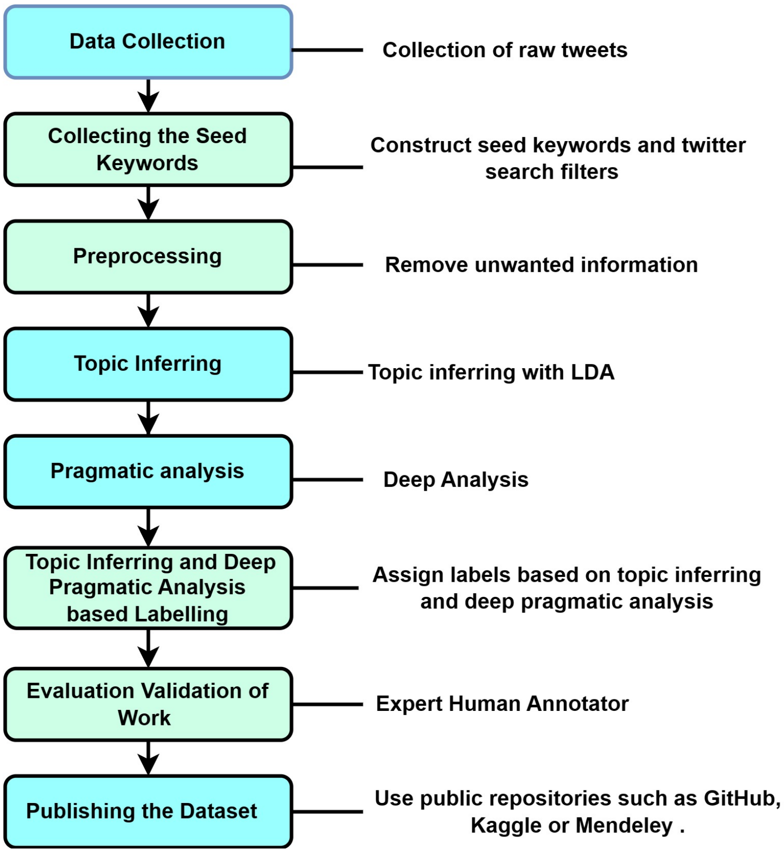
Data Collection
Collection of raw tweets
Collecting the Seed Keywords
Construct seed keywords and twitter search filters
Preprocessing
Remove unwanted information
Topic Inferring
Topic inferring with LDA
Pragmatic analysis
Deep Analysis
Topic Inferring and Deep Pragmatic Analysis based Labelling
Assign labels based on topic inferring and deep pragmatic analysis
Evaluation Validation of Work
Expert Human Annotator
Publishing the Dataset
Use public repositories such as GitHub, Kaggle or Mendeley .

Figure 1. Research Workflow.

4.1. Seed Words as Search Filters

Table 2 of seed keywords on world war sentiments, raw emotions, and opinions was developed with the purpose of determining the many forms of view that are present in people's perspectives on world wars. The seed keywords were chosen with great deliberation in order to accurately represent the main categories. For instance, the category of sentiments comprises the feelings of love, hatred, fear, hope, and fury, all of which are emotions that individuals may feel in response to war and conflict. Raw emotions are more specific feelings that people may have in response to a particular incident or circumstance related to war, such as sadness, excitement, disgust, and surprise.

Table 2. Seed keywords Category.

Category	Seed Keywords That Can Be Attached to Event of Wars and Conflicts
Sentiments	Love, Hate, Fear, Hope, Anger
Raw Emotions	Sadness, Joy, Disgust, Surprise
Opinions	Support, Opposition, Indifference, Ambivalence

Table 2 suggests that armed conflicts and hostilities can elicit a diverse array of raw emotion, sentiments and viewpoints among individuals. The events in question are associated with seed keywords such as love, hate, fear, hope, and anger, indicating that individuals may exhibit intricate emotional reactions to wars and conflicts. Apart from the fundamental emotions, individuals may encounter secondary emotions, including but not limited to, sadness, joy, disgust, and surprise. Diverse perspectives may exist among individuals regarding wars and conflicts, encompassing viewpoints that span from advocacy and dissent to apathy and uncertainty. In general, this implies that armed conflicts and hostilities can exert a substantial influence on individuals' affective and mental health, as well as their convictions and perspectives regarding these occurrences.

4.2. Expanding the Search Filters

To organise the data into multiple clusters of words related to sentiments and emotions attached to world wars, we followed a methodology mentioned in Table 4. Our objective was to categorise tweets pertaining to global conflicts, specifically world wars, into distinct sentiment classifications, encompassing emotions such as fear, wrath, grief, hope, and others. In order to accomplish this task, a combination of manual and automated methods was employed. Initially, a comprehensive compilation of keywords associated with the diverse range of emotions and attitudes linked to world wars, including but not limited to "fear", "anger", "hope", "despair", "patriotism", and "jingoism", was manually recognised. Subsequently, an automated programme, such as the Python Natural Language Toolkit (NLTK), was employed to ascertain additional lexemes and expressions that could potentially be linked to the aforementioned keywords. Subsequently, that helps in context or topic analysis were employed to categorise tweets into groups based on the similarity of terms and phrases associated with each sentiment category.

For example, here's a sample data table of how we organised tweets related to the sentiment of fear:

The tweets presented in Table 3 appear to convey a significant degree of apprehension regarding the potential occurrence of global conflict or other global events. The sentiment classification of all the tweets is "Fear". The linguistic expressions employed in the tweets indicate that the users are encountering feelings of anxiety, apprehension, and distress regarding the condition of the global affairs, leading to their experience of fear and discomfort. Hence, by organising the tweet data into multiple clusters of words in Table 4 related to sentiments and emotions, we can gain insights into how people feel about world wars, especially in the wake of conflicts such as ***Syrian Civil War (2011–present), Yemeni Civil War (2015–present), Nagorno-Karabakh Conflict (2020), Tigray Conflict (2020–present), Israel-Palestine Conflict (2021), Afghanistan Conflict (2001–2021) Ukraine and Russia, Indo-Pakistan Conflict, Indo-China Conflict, China-Taiwan, North Korea, US Cold War*** and so on. These insights can be used by researchers, policymakers, and other stakeholders to make informed decisions and take appropriate actions.

Table 3. Sentiment of fear.

Tweet ID	Tweet Text	Sentiment Category
1	"I'm so scared of what's going to happen in the world"	Fear
2	"The thought of war terrifies me"	Fear
3	"I can't sleep at night thinking about the possibility of a world war"	Fear
4	"This is really freaking me out"	Fear
5	"I'm filled with dread about the state of the world"	Fear

We have used open source social media intelligence tools such as Twint, Tweepy, and the Twitter/'X' API to get Twitter/'X' data.These tools provide access to vast quantities of data from the Twitter/'X' platform, allowing them to collect tweets on specific topics, such as those pertaining to global wars and conflicts [44,45]. With these tools, we were able to apply filters (seed words)and query parameters to obtain relevant data for their analysis and to refine their search results. These tools were also used to derive metadata from tweets, such as the creation date and time, emojis, user location, and used hashtags, which can provide additional insight into public sentiment and opinion on our topics of interest.

Table 4. Seed keywords.

Cluster	Keywords
Fear	Terrorism, Invasion, Attack, Nuclear weapons, Aatankwad, Aakraman, Hamla, Paramanu Hathiyaron, Bhayavah, Bhayankar, Khatarnak, Darawana, Aggression, Assault, Offensive, Incursion, Ambush, Raid, Onslaught, Nuclear Threat, Atomic Menace, Radiation Danger, Nuke Hazard, Violence, Hostility, Belligerence, Conflict, Encroachment, Trespassing, Intrusion, Extremism, Fanaticism, Radicalism, Militancy, Insurgency, Rebellion, Uprising, Sabotage, Subversion, Undermining, Treachery, Menace, Peril, Threat, Hazard, Provocation, Incitement, Stimulus, Aggravation, Irritation, Offensiveness, Displeasure, Dislike, Aversion, Menacing, Intimidating, Threatening, Scary, Attack on sovereignty, Offensive on territory, Breach of borders, Infringement on independence, Danger, Risk, Perilousness, Hazardousness, Destabilisation, Turmoil, Upheaval, Chaos, Agony, Distress, Torment, Affliction, Coercion, Duress, Pressure, Panic, Alarm, Hysteria, Fright, Brutality, Ferocity, Savagery, Barbarity, Military attack, Offensive campaign, Atomic Armament, Nuclear Missile, Radiological Weapon, Warhead, Insurrection, Revolt, Mutiny, Coup, Atrocity, Barbarism, Inhumanity, Cruelty, Confrontation, Hostilities, Clashes, Breach of defence, Infringement on protection, Threat to safety, Outrage, Fury, Indignation, Resentment, Intrusive, Obtrusive, Meddlesome, Nosy, Antagonism, Rivalry, Enmity, Assault with arms, Offensive with weapons, Armed attack, Strike with artillery, Radiation hazard, Radioactive peril, Rebellion against government, Uprising against authority, Insurgency against state, Revolt against administration, Human rights abuse, Oppression, Tyranny, Despotism, Battle, War, Combat, Skirmish.
Hope	Peace, Diplomacy, Resolution, Shanti, Shaanti, Aaram, Kootniti, Kootneeti, Rajneeti, Baatchee, Baatchit, Varta, Samadhan, Samaadhaan, Hal, Umed, Umeed, Asha, Shanti: Shantipurn, Niraashaant, Anand, Aaram: Vishraam, Thaharana, Rehat, Sakoon, Kootneeti/Rajneeti: Satta, Sarkaar, Prabandhan, Vyavastha, Sambhashan, Sandesh, Sanvaad, Prastutikaran, Samadhan/Samaadhaan: Hal, Halvayi, Nivaran, Nirdhaar, Aashirvaad, Aas, Aasha
Anger	Violence, Aggression, Conflict, Occupation, Intrusion, Struggle, Business, Anger, Opposition, War, Tension, Clash, Change, Retaliation, Fight, Resistance, Competition, Quarrel, Dilemma, Blow, Collision, Right, Expression, Freedom, Independence, Obstruction, Barrier, Battle, Rescue, Oppression, Reaction, Krodh, Hinsa, Pratidwand, Jang, Tanaav, Takraav, Badlaav, Pratikar, Dangal, Pratirodh, Mukabala, Jhagda, Uljhan, Thokar, Takkar, Pratipaksha, Adhikar, Abhivyakti, Mukti, Swatantrata, Roktham, Avrodh, Pratibandh, Samara, Paritran, Utpidan, Yuddh, Pratikriya, Anushasan, Aakramakata, Sangharsh, Vyavasaay.
Joy	Victory, Liberation, Freedom, Independence, Jeet, Azaadi, Swatantrata, Vijay, Jit, Fateh, Safalta, Baajti, Mukti, Chhoot, Nijaat, Nirmaan, Aatma, Aatma kush, Nirbharata, Swavalamban, Swabhimaan, Adhikar, Pratishodh, Kamyabi, Uddhar, Pragati, Utkarsh, Jayate, Siddhi, Triumph, Conquest, Subjugation, Overcoming, Accomplishment, Elation, Delight, Jubilation, Ecstasy, Exhilaration, Gratitude, Joyousness, Festivity, Rejoicing, Exultation, Thrill, Euphoria, Rapture, Bliss, Gleefulness, Merriment, Jollity, Glee, Cheerfulness, Happiness, Cheer, Enthusiasm, Excitement, Thriving, Flourishing, Advancement, Progression, Growth, Development, Prosperity, Success, Achievements, Triumphalism, Celebrations, Revelry, Exultancy, Fulsomeness, Contentment, Satiety, Gratification, Pleasure, Self-satisfaction, Delightfulness, Gladness, Exuberance, Zeal, Passion, Blissfulness.
Sad	Loss, Death, Trauma, Suffering, Sadness, Loss, Death, Trauma, Suffering, Haani, Maut, Trauma, Peeda, insecurity, vulnerability, despair, hopelessness, helplessness, sadness, sorrow, anguish, agony, misery, grief, mourning, Asuraksha, Bhay, Nirasha, Niraasha, Niraashrit, Dukh, Dard, Takleef, Bechaini, Udasi, Shok, Dukh, Vyatha, Dard, Duhkh, Matam, Rona, Bebasi, Tanhayi, Virah, Alvida, Tanhaai, Andhera, Ashantata, Vipada, Sangharsh, Kasht, Dardnak, Dardbhara, Duvidha, Asahayta, Samvedana, Sankat, Vilap, Nafrat, Anath, Nirjivta, Ulat Palat, Akshamta, Bebas, Bevakoof, Khafa, Betahasha, Nirasha, Vilamb, Vair, Bhagna Hriday, Dardnaak, Maafi, Mafinama, Nirnay, Nisantaan, Durghatna, Sthayitva, Asthayi, Virahita, Virodh, Ashru, Rulaana, Daridrata, Baychani, Tootna, Bhagna, Durbhagya, Durbalata, Nirlajja, Nirlajjata, Hani, Bayanak, Sunsaan, Vyakulta, Vipatti, Aansoon, Afsos, Afsosnaak, Asafalta, Asafal, Asafalata, Udhas, Udasi, Man-hi-man, Hichaki, Haath-diya, Dar, Udaasi, Lachar
Surprise	Breakthrough, Diplomatic relations, Unprecedented events, Aashcharyajanak Safalta, Rajneetik Rishte, Anoothi Ghatnaen, Astonishment, miracle, Revelation, Startling, Unexpected, Eye-opener, Stunner, Phenomenon, Thunderbolt, Wonderment, Mystery, Unexpectedness, Shocking, Stupefaction, Serendipity, Epiphany, Mind-blowing, Wondrous, Paradigm shift, Mind-boggling, Unpredictable, Unforeseen, rare, Out of the blue, Remarkable, Unanticipated, Puzzlement, Enigma, Surprise attack, Mind-bending, Revolutionary, Bewilderment, Unusual course of events, Newsworthy, Jolt, Awe-inspiring, Impressive, Catching off guard, Unexpected turn, Mysteriousness, Unexpected twist, Staggering

Table 4. *Cont.*

Cluster	Keywords
Disgust	Genocide, Atrocities, Human rights violations, War crimes, Nafrat, Narsanhaar, Apraadh, Manavaadhikaaron ka Ullekh, Yudh Apraadh, Brutality, Cruelty, Oppression, Injustice, Discrimination, Prejudice, Racism, Homophobia, Xenophobia, Bigotry, Intolerance, Hatred, Loathing, Abomination, Revulsion, Contempt, Disdain, Dislike, Disapproval, Disgust, Abhorrence, Repugnance, Aversion, Antipathy, Odium, Detestation, Despise, Scorn, Malice, Animosity, Hostility, Enmity, Agony, Torment, Aggravation, Malignity, Spitefulness, Vengefulness, Resentment, Bitterness, Displeasure. Discomfort, Discontent, Disquietude, Unease, Annoyance, Irritation, Frustration, Anguish, Misery, Wretchedness, Affliction, Tribulation, Hardship, Suffering, Opprobrium, Shame, Disgrace, Embarrassment, Humiliation, Degradation, Ignominy, Infamy, Scandal, Reproach, contumely, Insult, Defamation, Libel, Slander, Calumny, Falsehood, Deceit, Betrayal, Infidelity, Treason, Perfidy, Duplicity, Fraud, Corruption, Iniquity, Sin, Vice, Immorality, Decadence, Depravity, Aberration, Deviation, Perversion, Lewdness, Obscenity, Profligacy, Impurity, Indecency, Blasphemy, Sacrilege, Profanity, Heresy, Apostasy, Ignorance, Stupidity, Foolishness, Ineptitude, Incompetence, Ineffectiveness, Negligence, Sloth, Procrastination, Apathy, Indifference, Insensitivity, Callousness

The keyword cluster "Fear" connotes a perception of peril, menace, and aggression. The lexicon in question comprises terms such as terrorism, attack, invasion, aggression, extremism, rebellion, coercion, panic, and brutality. The cluster of keywords pertaining to "Hope" connotes a positive and optimistic outlook, encompassing concepts such as peace, diplomacy, negotiation, and resolution. The cluster of keywords pertaining to "Anger" connotes a state of discord, strain, and resistance. The lexicon comprises terms such as violence, aggression, struggle, oppression, retaliation, and clash. The cluster of keywords pertaining to "Joy" connotes optimistic emotions and favourable consequences, such as triumph, autonomy, self-sufficiency, accomplishment, advancement, and affluence. The analysis of the keyword clusters suggests that emotions such as fear, hope, anger, and joy are associated with war, crisis, and conflicts. The fear cluster is indicative of a perception of peril and aggression, whereas the hope cluster is suggestive of favourable consequences achieved via tranquilly, tact, and mediation. The cluster associated with anger connotes a state of discord and strain, while the cluster associated with joy connotes favourable emotions and consequences, such as triumph and affluence. For the extraction of the emojis from the each tweet text, we have used regular expressions that matches all ranges of Unicode characters that correspond to emoji categories, including the Miscellaneous Symbols and Arrows block, Dingbats block, Emoticons block, Miscellaneous Symbols block, Supplemental Symbols and Pictographs block, and the new Extended Emoji block introduced in Unicode 13.0 etc.

5. Topic Inferring and Content Analysis

For demonstrating the application 'A' of the dataset having 'T' tweets as a subset tweets (English + Hindi) were selected. The purpose of making sub set of tweets was to show case(s) of identifying hidden emotion and sentiments in the written utterances. The visualization of topic and content analysis done with the help of LDA can observed from Figures 2 and 3 . The application of deep learning on the subset of tweets can be observed from Table 5, which demonstrates deep pragmatics analysis for assigning appropriate labels on keywords extraction from the subset of tweets.

The topic analysis indicates a distinct and varied landscape of topics, as evidenced by the inter-topic distance observations in Figures 2 and 3. Notably, each cluster is unique in size, suggesting a diverse range of topic concentrations within the dataset. The variation in cluster sizes is further highlighted by the significant difference between the largest bubble (the first cluster) and the smallest (the tenth cluster), implying a disparity in the volume or complexity of topics they represent.

The presence of different silent terms in each cluster suggests a specialized vocabulary or focus unique to each topic, further emphasizing the distinctiveness of the clusters. This specialization is a common characteristic in topic modeling, where each cluster or topic contains words most representative of that topic.

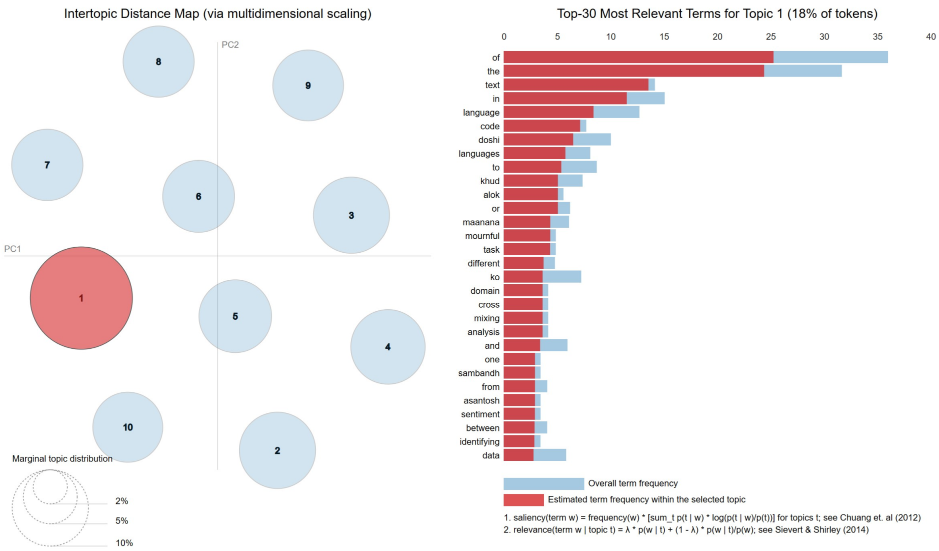

Figure 2. Topic Inferring [46,47].

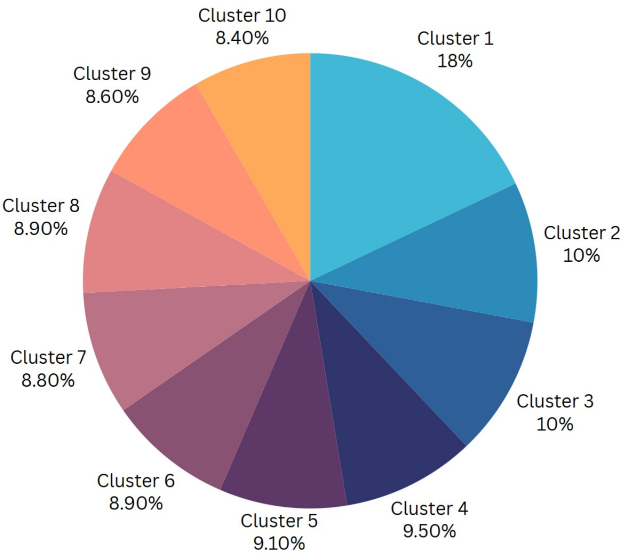

Figure 3. Percentage of Tokens by Cluster.

Table 5. Deep Pragmatic Analysis.

Cluster 1 Keywords	Data, code, doshi, Khud, alok, maanana. Task, different, domain, across. Mixing, Sambanndh, Asantosh, sentiment, text
Possible Annotations	Data Integration and Sentiment Analysis (L1), Self-Reflection and Opinion (L2)
Pragmatic Analysis	***This label (L1)*** captures the core concepts of combining data from different domains, potentially for sentiment analysis or opinion mining. The keywords "data", "code", "mixing", "across", "domain", "sentiment", and "text" strongly suggest this theme. This cluster could represent discussions about techniques for integrating data from various sources to understand sentiment or emotions expressed in text. It might involve discussions of tools, challenges, or practical applications in this field. ***This label (L2)*** highlights the presence of words like "khud" (self), "alok" (light, perspective), "maanana" (acceptance), "asantosh" (satisfaction), and "sentiment", which could indicate discussions about personal opinions, self-reflection, and satisfaction with outcomes. This cluster (L2) might contain tweets where people express opinions, reflect on their experiences, or share their satisfaction regarding various topics.

Table 5. *Cont.*

Cluster 2 Keywords	Shak, dil, virahit, say umeed, ko, suhkriya, doshi, disheartеd, guilt data, kal, discontent, shivering, infuriating, linguistics
Possible annotations	Emotional Linguistics (L3), Discontent (L4)
Pragmatic Analysis	The presence of words like “shak” (doubt), “dil” (heart), “virahit” (deprived), “umeed” (hope), “disheartеd”, “guilt”, “discontent”, and “infuriating” suggests a strong emotional context. These words collectively represent various sentiments and emotional states, indicating that the cluster 2 may involve texts dealing with feelings, emotional expressions, or discussions around emotional topics. ***The term “emotional linguistics (L3)*** ” suggests an analytical or structured discussion around language, possibly in the context of these emotional expressions. Several words like “disheartеd”, “guilt”, ″discontent”, and “infuriating” specifically point to negative emotions or states of dissatisfaction. This indicates that cluster 2 may represent discussions or texts that revolve around themes of regret, anger, frustration, or general dissatisfaction. Hence, the label ***Discontent (L4)***
Cluster 3 Keywords	ninda, ashanka, ashirwad, daya, ruffeled, achha, vibram, feathers, mad, heartendness, deep, tantrum mad, bitter, process.
Possible annotation	Emotional Turmoil (L5), Reflections (L6)
Pragmatic Analysis	Words like “ninda” (criticism or condemnation), “ashanka” (doubt or suspicion), “mad” (angry or intense), “ruffled”, “tantrum”, and “bitter” indicate strong negative emotions or states of disturbance. These terms suggest discussions or expressions of conflict, upset, or emotional unrest. Hence, **Emotional Turmoil** (L5) is best suited. On the other hand, words like “ashirwad” (blessings), “daya” (compassion or mercy), “achha” (good), and “heartendness” (perhaps a misspelling or variation of ′heartedness′ or ′heartening′) can imply a more reflective or positive aspect. This duality of negative and positive emotional language indicates a complex emotional landscape, perhaps reflecting on both the turmoil and the more compassionate or hopeful aspects of the human experience. Hence, the annotation **Reflections** (L6)
Cluster 4 Keywords	pashtaap, prayaschit, koti, sunehra, fumming, funereal, ghabrahat, naman, chill, agony, creeping, anguish, humbeled, words sullen,
Possible annotations	Contemplative Remorse (L7), Solemnity (L8)
Pragmatic Analysis	Keywords such as “pashtaap” (regret or remorse), “prayaschit” (atonement or penance), and “fumming” (a variant of ′fuming′, indicating anger or frustration) suggest a deep sense of reflection on past actions or emotions. This reflection is typically associated with feelings of guilt, regret, or a desire to make amends, indicative of a contemplative and remorseful state. Hence, the label ***Contemplative Remorse (L7)*** Words such as “funereal” (relating to a funeral or death), “ghabrahat” (anxiety or unease), “chill”, “agony”, “anguish”, and “sullen” (bad-tempered or gloomy) all contribute to a solemn or deeply serious tone. “Naman” (a gesture of respect or salutation) and “humbled” also suggest reverence or a subdued demeanor, often found in solemn or serious circumstances.
Cluster 5 Keywords	dukh, vismay, bhoot, krodh, mujbuti, sorrow, downcast, khed, deep, apology, good, creepy, gussa, prokop
Possible annotation	Sorrow (L9), Indignation (L10)
Pragmatic Analysis	***Sorrow (L9)***: Words like “dukh” (sorrow), “sorrow” itself, “downcast”, and “khed” (regret or sorrow) clearly point towards a theme of sadness and regret. These words suggest discussions or expressions that revolve around personal grief, disappointment, or general sadness. ***Indignation (L10)*** Terms such as “krodh” (anger), “gussa” (anger), and “prokop” (fury or rage) indicate strong feelings of anger or annoyance. “Vismay” (wonder or surprise) can sometimes be associated with shock or disbelief that could lead to indignation, depending on the context.
Cluster 6 Keywords	udaas, ahyankar, utsah, bharosha, downherted, bhavishya, pareshani, shakti, hona, umang, doshi, adhbut, honor, nayi, irate
Possible annotation	Sadness (L11), Resilience (L12)
Pragmatic Analysis	***Sadness (L11)*** The cluster includes words that cover a wide range of emotions. “Udaas” (sad), ″downhearted“, and ″irate“ suggest feelings of sadness and anger. In contrast, ″utsah“ (enthusiasm), “bharosha” (trust), “umang” (joy or enthusiasm), and “shakti” (strength) indicate positive emotions and qualities. ***Resilience and Hope (L12)*** Words like “bharosha”, “umang”, and “shakti” not only represent positive emotions but also suggest a sense of resilience and hope. “Bhavishya” (future) and “nayi” (new) reinforce this theme, indicating a forward-looking or hopeful perspective.

Table 5. *Cont.*

Cluster 7 Keywords	dukhi, maafi, thanks, bhavuk, upset, prerena, dhundla, vinamarta, bhayanak, chiddhana, garmi, uproar, dreary, angrily sambhavana
Possible annotation	Melancholic Stirrings (L13), Humility and Remorse (L14)
Pragmatic Analysis	Melancholic Stirrings: Words like "dukhi" (sad), "upset", "bhavuk" (emotional), "dhundla" (blurry or unclear, often metaphorically used to represent confusion or lack of clarity), and "dreary" suggest a theme of sadness, emotional depth, and a general melancholic or downcast mood. "Angrily" and "uproar" indicate a disturbance or intense emotional reaction, adding to the sense of emotional stirrings. Humility and Remorse: "Maafi" (forgiveness or apology), "vinamarta" (humility), and "thanks" imply a sense of remorse, gratitude, or humbleness. These terms suggest an acknowledgment of mistakes, appreciation for others, or a general attitude of humility and respect.
Cluster 8 Keywords	nirasha, khushi, chinta, anukool, visvash, rona, sankalap, ekta, pragati, ashirvad, wathful, stormy, heavy, reproach, vyaakul
Possible annotations	Mixed Emotions, Resilience and Unity in Adversity
Pragmatic Analysis	**Mixed Emotions** (L15) The cluster encompasses a range of positive and negative emotions, suggesting a focus on mixed feelings and experiences. Pragmatic analysis: The cluster touches on themes of disappointment ("nirasha"), happiness ("khushi"), worry ("chinta"), trust ("visvash"), crying ("rona"), and reproach ("reproach"). This suggests a discourse that grapples with both positive and negative aspects of life. ***Resilience and Unity in Adversity (L16)*** The keywords in the cluster emphasises unity and collective strength in the face of challenges. The presence of words like "ekta" (unity), "visvash" (trust), and "anukool" (supportive) suggests a discourse that focuses on finding strength in togetherness and overcoming obstacles together.
Cluster 9 Keywords	doah, dosh, apmaan, vivklit, khushi, laaz, sukoon, bhavana, bachaini, vidhva, blame, dekhbaal, frenyize, aas, sukoon
Possible annotation	Turmoil (L17), Serenity (L18)
Pragmatic Analysis	The label "Turmoil and Serenity in Self and Relationships" captures the essence of the cluster's topics, suggesting that the texts or tweets likely involve discussions or expressions navigating through emotional and moral complexities within oneself and in relation to others. It reflects the dynamic interplay between challenging and comforting emotions and situations, as well as the pursuit of understanding, peace, and resolution in personal and social contexts. ***Turmoil (L17)*** Words like "doah" (doubt or blame), "dosh" (fault or blame), "apmaan" (insult), "vivklit" (perplexed), "bachaini" (restlessness), and "frenzyize" (likely a misspelling or variation of "frenzied", indicating chaotic or wild behavior) suggest a state of emotional and moral turmoil. These terms indicate discussions or expressions of conflict, guilt, agitation, or blame. ***Serenity (L18)*** In contrast, words like "khushi" (happiness), "laaz" (shame but can also imply honor in certain contexts), "sukoon" (peace or tranquility), and "aas" (hope) reflect a more positive or peaceful emotional state. This indicates a movement or desire towards tranquility, contentment, and positive emotional and moral states.
Cluster 10 Keywords	utsaah, shama, ullas, santusti, jazbaat, ashvasan, samaghdari, gratitude, mukti, chidhaavat, human, influriation, inconsolable, Connection
Possible annotation	Positive Emotional Dynamics (L19), Gratitude and Liberation (L20)
Pragmatic Analysis	***Positive Emotional Dynamics (L19)***: Words like "utsaah" (enthusiasm), "shama" (forgiveness or patience), "ullas" (joy), "santusti" (satisfaction), "jazbaat" (emotions), and "ashvasan" (assurance) denote positive emotional states and qualities. These terms suggest expressions of joy, emotional richness, and a sense of fulfilment or contentment. ***Gratitude and Liberation (L20)***: The inclusion of "gratitude" and "mukti" (liberation or freedom) adds layers of thankfulness and the concept of emotional or spiritual freedom. This implies discussions or expressions that revolve around being grateful, finding peace, or achieving a sense of liberation.

The distribution of tokens across clusters reveals that the topics are not evenly distributed. The first cluster holds the highest percentage of tokens at 18%, indicating it covers a larger portion of the dataset or a more prevalent topic. In contrast, the tenth cluster contains the smallest percentage of tokens at 8.4%, suggesting it might represent a more niche or less discussed topic. The remaining clusters have relatively similar percentages of tokens, ranging from 8.6% to 10%, indicating a more balanced distribution of content among them.

5.1. Deep Demostration through Case Study

As mention earlier, To illustrate the potential value (P) of the dataset and its applicability/application, a comprehensive case study was meticulously designed. Initially, a subset of tweets was curated using a stratified random sampling method, where the seed keywords served as the stratification criteria to ensure representativeness across different emotional or thematic segments. Essentially, Table 5 serves as both a showcase of the dataset's rich potential for various applications, such as enhancing sentiment analysis models, informing socio-linguistic studies, or even aiding in monitoring public opinion trends on key issues.

The analysis shown in the Table 5 has been done on the basis of following principles:

1. Contextual Understanding Principle (CUP) This principle emphasises the importance of understanding each word and phrase within its linguistic and cultural context. It's crucial for analysing code-mixed languages like Hinglish, where cultural idioms and linguistic structures are deeply intertwined and complex.
2. Emotional and Sentiment Mapping Principle (ESMP) This principle involves identifying and mapping the emotional or sentiment value of words and phrases to understand the overall emotional tone of the text. It's key in recognizing and categorising emotional words and the broader emotional states they imply.
3. Lexical and Semantic Analysis Principle (LSAP) This principle focuses on analysing the meaning of words and phrases, both individually and collectively, to understand their semantic roles in the text. It's essential for suggesting particular themes or topics and understanding the implications of word combinations.
4. Cultural and Sociolinguistic Relevance Principle (CSRP) This principle recognizes the importance of cultural insights and sociolinguistic factors in interpreting language, particularly in contexts deeply embedded with cultural nuances like Hinglish. It emphasises understanding words beyond their direct translations to include cultural undertones.
5. Coherence and Cohesion in Text Principle (CCTP) This principle looks at how words and phrases contribute to creating coherent and cohesive messages, themes, or narratives within the text. It's important for noticing patterns or common themes across different texts to accurately reflect the predominant theme or sentiment of the cluster.

These principles (CUP, ESMP, LSAP, CSRP, and CCTP) serve as a foundational framework for performing deep pragmatic analysis, particularly in linguistically and culturally rich datasets like Hinglish text collections. They guide the process of analysing, understanding, and categorising text to reflect the underlying themes, sentiments, and cultural nuances accurately.

5.2. Inferences and Findings

From the cluster keywords and annotations analysis in Table 4, the emotions reflected across the clusters can be attributed to several emotion theories that include, Plutchik's Wheel of Emotions: Many of the emotions identified across clusters, such as sadness, anger, trust, and anticipation, can be associated with Plutchik's theory, which identifies eight primary bipolar emotions. The clusters reflect complex emotions that can be seen as combinations or intensities of these primary emotions, such as "frenzy" as an intense form of anger or "sorrow" as a form of sadness. Second theory that needs mention here is appraisal Theory: This theory suggests that emotions are the result of personal interpretation of an event, leading to emotional reactions. The presence of words related to personal experience, opinion, and reflection (e.g., "maanana" for considering or "ashirwad" for blessings) across clusters indicate emotions that might be interpreted through personal appraisal of situations, particularly in the personal and reflective labels. The third one is, cognitive Theories of Emotions, this theory focuses on the thought processes that precede the emotional reaction. The clusters, especially those with keywords indicating reflection, decision-making, or contemplation (like "prayaschit" or "sankalap"), reflect the cognitive aspects of emotions where emotions are the end result of complex thought processes. However, it is clear that there are vary array of emotions and no single theory can be attributed. In terms of Relevance and Specificity of the annotations, the labels are generally relevant and specific to the keywords presented in each cluster. They effectively capture the essence of the emotions and topics discussed within the clusters, indicating a thoughtful consideration of the words' emotional and pragmatic implications. Moreover, in terms of Cultural Sensitivity, these labels show sensitivity to the cultural context, particularly important for code-mixed language data like Hinglish. They incorporate the cultural nuances and emotional expressions specific to the linguistic context. Lastly, the application 'A' demonstrates, good level of depth of pragmatic analysis. The pragmatic analysis clearly shows that the methodology followed allowed us to consider the cultural and contextual implications of the words, going beyond the literal meanings to understand their deeper, more nuanced implications in social and emotional contexts.

Applicability and Application 'A' The analysis clearly demonstrates how to be directly applicable to the keywords and themes, indicating an understanding of how these words are used in real-life communication, especially in informal and personal communication like tweets. In nust nutshell, the analysis of these clusters reflect a rich tapestry of emotional expression that aligns with various emotional theories. The quality of the labels and pragmatic analysis appears to be high, considering the depth, cultural sensitivity, and relevance. The applications for such work are vast, ranging from enhancing technological solutions like sentiment analysis to contributing to fields like mental health and cultural studies.

In our process to elucidate the subjects at hand, we have identified the top 15 words that are not only relevant to the discourse but are distinctively chosen for their rich contribution and depth of meaning. These words have been selected due to their substantial informational content, ensuring that each term provides unique insights and adds significant value to our understanding of the themes of war and conflict. As we proceed, these words will be presented, serving as critical markers that encapsulate the essence of the topics with in the domain of war and conflict and enhance the comprehensiveness of the dataset in question.

6. Result and Discussion

The primary goal of the research was to develop a novel dataset focused on "raw emotions" elicited by war and conflicts. This dataset would differ from existing resources by providing more nuanced and possibly real-time emotional responses associated with such intense and complex situations. The term "raw emotions" refer to the immediate, unfiltered emotional reactions captured from various sources, possibly including social media, news reports, or direct narratives. The creation of this dataset involved several steps. The process of collecting, organizing, and validating data related to emotions in war and conflict scenarios. This involved extensive data sourcing, cleaning, topic inferencing and evualation and annotation by human experts to ensure the dataset's accuracy and relevance. The inter-rater agreement visualizations and analysis has been part of this process. This ensured that the emotional annotations are consistent and reliable across different annotators. Dataset Characteristics (Table 6): gives informaiton on data collected.

Table 6. Dataset Characteristics.

Serial Number	Variable	Description
1	Number of Search Filters	500
2	Total Tweets	10,040
3	Fields in each Tweet	Tweet ID, Tweet, Retweet

The Multimodal Hinglish Tweet Dataset for Deep Pragmatic Analysis is publicly available at [48]. Following points must be considered as new contributions to this domain.

Development of a Novel Dataset The research contributes to the field by developing a novel dataset focused on raw emotions elicited by war and conflicts. Unlike existing resources, this dataset provides nuanced and real-time emotional responses, filling a critical gap in current emotional analysis and conflict studies. This contribution is particularly valuable for its focus on immediate, unfiltered emotional reactions, which are essential for understanding the human aspect of conflicts.

Rigorous Data Curation Process The dataset's credibility is bolstered by a rigorous curation process involving collecting, organizing, cleaning, and validating data. The process includes topic inferencing, evaluation, and expert annotation to ensure the dataset's accuracy and relevance. The inclusion of inter-rater agreement visualization and analysis ensures the consistency and reliability of emotional annotations across different annotators, contributing to the dataset's quality.

Demonstration of Potential Value P Beyond dataset creation, the research demonstrates the dataset's utility in addressing real-world problems related to understanding emotions in conflict situations. By providing case studies on the application of the dataset in predictive modeling, sentiment analysis, and informing humanitarian and policy-making efforts, the research showcases the practical applicability and effectiveness of the dataset in real-world scenarios. Given that the dataset contains both textual and visual elements in the form of emojis, it can be accurately classified as a multi-modal dataset. The significance and importance of this meticulously curated collection of tweets pertaining to wars and conflicts are as follows.

1. The dataset contains tweets that are pertinent to the topic of world wars and conflicts, allowing researchers to concentrate their analysis on tweets that are most relevant to the issue at hand.Currently the world is suffering from multiple conflicts including Ukraine-Russia, Indo-Pakistan, Indo-China, China-Taiwan, North Korea, and the US Cold War, among others.
2. By compiling the odd tweets into a structured dataset, the research improves the data's transparency and makes it simpler for other researchers to reproduce the analysis and draw their own conclusions regarding the emotions, opinions and sentiments regarding world war and conflicts and their regional conflicts.
3. The clustering enables researchers to categorise odd tweets related to world war and conflicts, which can provide insight into how frequently this topic is discussed on Twitter/'X' and how it evolves over time.
4. The dataset is curated using open social media tools and with help of data mining methods the process of curation was completed [49]. This research will provide foundational insights into the communication patterns of Twitter/'X' users and help identify indirect acts or figurative language employed in the messages regarding war and conflicts.

   Filling Data Availability Gaps This research addresses a significant gap in publicly available datasets focusing on emotions in conflict and war scenarios. By providing this resource, the researchers have significantly contributed to the field, enabling other researchers, policymakers, and organizations to understand and respond to the emotional dimensions of conflicts more effectively.

   Additional Auxiliary Work The contribution extends to auxiliary work that enhances the main research. This includes developing tools for easier dataset access, conducting preliminary studies to validate the dataset's utility, and engaging with communities and experts for feedback. Such efforts ensure the dataset's continuous improvement and wider adoption, maximizing its impact.

   Integration of Deep Pragmatic Analysis with Topic Clustering A distinctive contribution of this research is the enhancement of topic analysis and clustering approaches, typically reliant on algorithms like Latent Dirichlet Allocation (LDA), with the incorporation of Deep Pragmatic Analysis. This integration addresses the limitations of traditional topic modeling techniques that often miss the subtleties of human communication such as sarcasm, metaphor, and context-dependent meanings. Deep Pragmatic Analysis delves deeper into the linguistic and contextual nuances, providing a more sophisticated and accurate understanding of text data, especially in the emotionally charged and complex narratives found in war and conflict contexts. This approach significantly improves the interpretability and usefulness of the topic models, making this a pivotal advancement in the field of emotional data analysis. Overall, the research work stands out for its comprehensive approach to creating a valuable emotional dataset, its rigorous validation process, and its emphasis on practical application in understanding and addressing the human impacts of war and conflicts.

Evaluation and Validation of Work

The process of creating new dataset and demonstration of its potential value '*P*' was illustrated using use case approach.The case study approach allows to showcase a real life problem of identifying multiple hidden feelings, emotions and sentiments in a subset of the dataset. For this topic analysis was done and we got 10 clusters of keywords, for identifying the theme of each keywords, deep pragmatic analysis was done. The outcome is showcased in Table 5. Hence to evaluate and validate this approach human experts roped in. The human had fair knowledge in area of current affairs,especially in domain of war and conflicts.They are also proficients in reading and writing both english and nuances of hinglish. Figure 4 depicts thier "inter-rater" agreement on the quality of annotation assigned by said methodology and more detailed information on the methodology, please refer to Appendix A.

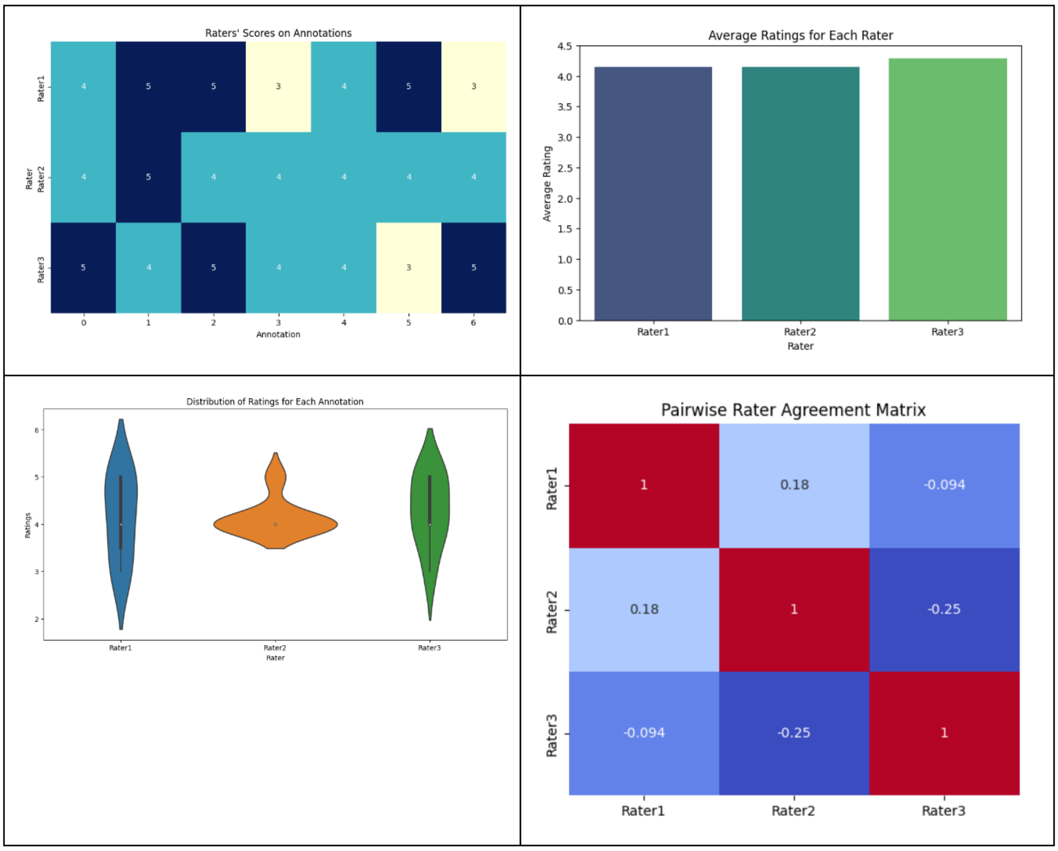

Figure 4. Analysis of Inter-rater agreement.

The dataset currently has 10,040 tweets that have been evaluated using a randomly selected subset of the tweets. Using a case study approach, the evaluation and validation of the deep pragmatic analysis technique have been done. From the outcome of the score given by the three human experts, the following inferences can be made:

1. The judges effectively consider the given annotations for each keyword cluster as appropriate emotional expressions/labels. The average intensity ratings across all raters were high (ranging from 4.14 to 4.29 on a scale of 1 to 5), indicating a strong emotional expression in the labels. The high agreement on primary emotions among raters, as confirmed by the pairwise rater agreement matrix values (positive values such as 0.18 to 1), supports a good level of consistent primary emotion identification.
2. The judges exhibit high confidence in the process of annotation, expressing a strong belief in the accuracy of the deep pragmatic analysis of five principles-based emotion annotations. The average confidence ratings range from 4.14 to 4.29 (on a scale of 1 to 5), indicating high confidence. The pairwise agreement matrix shows moderate to high consistency among raters (positive values, e.g., 0.18 to 1), as does the histogram of Kappa scores, which indicates a moderate level of agreement between rater pairs. In the end it can be said that there is a high level of agreement among raters regarding emotion intensity and primary emotion identification. This conclusion is supported by the consistency observed in the average ratings and pairwise agreement matrix values.

7. Conclusions

This research work gets into the construction and application of a novel Hinglish text dataset, a unique blend of Hindi and English used in various contexts of everyday life. Our initial approach involved crafting specific "seed words" to filter relevant posts from social media platforms like Twitter/'X' . These seed words were designed to be broad enough to encapsulate a variety of topics, ensuring a collection rich in cultural and linguistic nuances. We expanded the dataset beyond the initial scope of seed words to encompass a wider array of Hinglish texts. This expansion was carefully managed to maintain relevance and high quality, making the collection comprehensive and versatile for different kinds of studies and practical applications.One key application we explored with this dataset was in the domain of emotional and cultural studies, particularly focusing on how Hinglish is used to express emotions and sentiments across various themes. While part of our analysis included texts related to specific themes like 'war and conflict' to demonstrate the dataset's applicability in targeted domains, the methods and the dataset itself are designed for a much broader application. This ensures that our collection and analysis methods are not limited to any single domain but can be adapted for diverse research and practical needs. To understand the complex layers within the Hinglish texts, we employed deep pragmatic analysis. This approach involved advanced linguistic and cultural analysis techniques to uncover the nuanced meanings, emotions, and sentiments expressed in the text. The depth of our analysis comes from the combination of multiple linguistic methodologies, including sentiment analysis, emotion detection, and cultural reference identification, which together provide a comprehensive understanding of the text's meaning. Our research not only resulted in a valuable Hinglish text dataset but also established a detailed methodology for building and expanding such a dataset. The deep pragmatic analysis, coupled with cluster analysis, was instrumental in categorising raw emotions and expressions, thereby enhancing our understanding of the complex interplay of language, culture, and emotion.

Author Contributions: Conceptualization, P. and A.K.; methodology, software, P.; validation, M.K. and A.K.; formal analysis, A.K. and R.D.; investigation, P.; resources, A.K.; data curation, M.K.; writing—original draft preparation, P.; writing—review and editing, A.K. and R.D.; visualization, M.K.; supervision and project administration, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLTK	Natural Language Toolkit
DL	Deep Learning
LDA	Latent Dirichlet Allocation

Appendix A

Questions for Raters Evaluation

Read the sheet that randomly 7 emotions (out of 10) identified/annotations with respect to keywords clusters, context and corresponding tweets text. The identification was done using methodology explained in Methodology Section 4.

1. Emotion Intensity: On a scale of 1 (not at all) to 5 (extremely strong), how strongly are emotions expressed in this snippet?
2. Primary Emotion: What is the primary emotion expressed in this snippet? (Choose one from a provided list of emotions covered by the algorithm)
3. Secondary Emotions (Optional): Are there any additional emotions present, though less prominent? (Select from the list or indicate "None")
4. Confidence in Algorithm Annotation: How confident are you that the algorithm's emotion annotation for this snippet is accurate? (1-Not at all confident, 5-Extremely confident)
5. Clarity of Explanation (Optional, if the algorithm provides explanations): If the algorithm provides an explanation for its annotation, is it clear and understandable? (1-Not clear at all, 5-Very clear)
6. Additional Comments (Optional): Do you have any additional comments or observations about this annotation methodology used?

References

1. Zimbra, D.; Abbasi, A.; Zeng, D.; Chen, H. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *Acm Trans. Manag. Inf. Syst.* **2018**, *9*, 3185045. [CrossRef]
2. Tao, W.; Peng, Y. Differentiation and unity: A Cross-platform Comparison Analysis of Online Posts' Semantics of the Russian–Ukrainian War Based on Weibo and Twitter. *Commun. Public* **2023**, *8*, 105–124. [CrossRef]
3. Zadeh, M.H.; Cicekli, I. Protest Event Analysis: A New Method Based on Twitter's User Behaviors. *Inf. Technol. Control* **2023**, *52*, 457–470. [CrossRef]
4. Karayiğit, H.; Akdagli, A. BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Comments. *Inf. Technol. Control* **2022**, *51*, 409–428. [CrossRef]
5. Aldjanabi, W.; Dahou, A.; Al-Qaness, M.A.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics* **2021**, *8*, 69. [CrossRef]
6. Gunasekar, M.; Thilagamani, S. Improved Feature Representation Using Collaborative Network for Cross-Domain Sentiment Analysis. *Inf. Technol. Control* **2023**, *52*, 100–110. [CrossRef]
7. Liang, S.; Jin, J.; Du, W.; Qu, S. A Multi-Channel Text Sentiment Analysis Model Integrating Pre-training Mechanism. *Inf. Technol. Control* **2023**, *52*, 263–275. [CrossRef]
8. Tesfagergish, S.G.; Damaševičius, R.; Kapočiūtė-Dzikienė, J. Deep Fake Recognition in Tweets Using Text Augmentation, Word Embeddings and Deep Learning. In *Computational Science and Its Applications—ICCSA 2021: In Proceedings of the 21st International Conference, Cagliari, Italy, 13–16 September 2021*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2021; Volume 12954, pp. 523–538. [CrossRef]
9. Yinka-Banjo, C.; Ugot, O.A.; Misra, S.; Adewumi, A.; Damasevicius, R.; Maskeliunas, R. Conflict resolution via emerging technologies? *J. Phys. Conf. Ser.* **2019**, *1235*, 12022 [CrossRef]
10. Kaur, G.; Pratibha; Kaur, A.; Khurana, M. A Review of Opinion Mining Techniques. *ECS Trans.* **2022**, *107*, 10125. [CrossRef]
11. Tesfagergish, S.G.; Damaševičius, R.; Kapočiūtė-Dzikienė, J. Deep Learning-Based Sentiment Classification of Social Network Texts in Amharic Language. *Commun. Comput. Inf. Sci.* **2022**, *1740*, 63–75. [CrossRef]
12. Maity, K.; Saha, S.; Bhattacharyya, P. Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 2411–2420. [CrossRef]

13. Srivastava, A.; Hasan, M.; Yagnik, B.; Walambe, R.; Kotecha, K. Role of artificial intelligence in detection of hateful speech for Hinglish data on social media. In *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2020*; Springer: Singapore, 2021; pp. 83–95.
14. Kukkar, A.; Mohana, R.; Sharma, A.; Nayyar, A.; Shah, M.A. Improving Sentiment Analysis in Social Media by Handling Lengthened Words. *IEEE Access* **2023**, *11*, 9775–9788. [CrossRef]
15. Sasidhar, T.T.; Premjith, B.; Soman, K. Emotion detection in hinglish (hindi + english) code-mixed social media text. *Procedia Comput. Sci.* **2020**, *171*, 1346–1352. [CrossRef]
16. Gupta, R.; Srivastava, V.; Singh, M. MUTANT: A Multi-sentential Code-mixed Hinglish Dataset. *arXiv* **2023**, arXiv:2302.11766.
17. Tesfagergish, S.G.; Damaševičius, R.; Kapočiūtė-Dzikienė, J. Deep Learning-based Sentiment Classification in Amharic using Multi-lingual Datasets. *Comput. Sci. Inf. Syst.* **2023**, *20*, 1459–1481. [CrossRef]
18. Cui, J.; Wang, Z.; Ho, S.B.; Cambria, E. Survey on sentiment analysis: Evolution of research methods and topics. *Artif. Intell. Rev.* **2023**, *56*, 8469–8510. [CrossRef] [PubMed]
19. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
20. Chan, J.Y.L.; Bea, K.T.; Leow, S.M.H.; Phoong, S.W.; Cheng, W.K. State of the art: A review of sentiment analysis based on sequential transfer learning. *Artif. Intell. Rev.* **2023**, *56*, 749–780. [CrossRef]
21. Das, S.; Singh, T. Sentiment Recognition of Hinglish Code Mixed Data using Deep Learning Models based Approach. In Proceedings of the 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 19–20 January 2023; pp. 265–269.
22. Ledalla, S.; Rao, G.A.; Sesetti, A. Sentiment Analysis of Hinglish Reviews Using Hybrid Approaches. *Int. J. Health Sci.* **2022**, *6*, 5432–5445. [CrossRef]
23. Doğruöz, A.S.; Sitaram, S.; Bullock, B.E.; Toribio, A.J. A survey of code-switching: Linguistic and social perspectives for language technologies. *arXiv* **2023**, arXiv:2301.01967.
24. Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunsdon, T. Comparison of Topic Modelling Approaches in the Banking Context. *Appl. Sci.* **2023**, *13*, 797. [CrossRef]
25. Jain, L.; Sharma, M.; Abdulsada, Z.R. Offensive Tweets Detection in Hinglish Using HingBERT. *Int. Conf. Data Anal. Manag.* **2023**, *10*, 93–103.
26. Shevtsov, A.; Tzagkarakis, C.; Antonakaki, D.; Pratikakis, P.; Ioannidis, S. Twitter Dataset on the Russo-Ukrainian War. *arXiv* **2022**, arXiv:2204.08530.
27. Siapera, E.; Hunt, G.; Lynn, T. #GazaUnderAttack: Twitter, Palestine and diffused war. *Inf. Commun. Soc.* **2022**, *22*, 1297–1319.
28. Chen, E.; Ferrara, E. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia. *arXiv* **2022**, arXiv:2203.07488.
29. Smart, B.; Watt, J.; Benedetti, S.; Mitchell, L.; Roughan, M. #IStandWithPutin versus #IStandWithUkraine: The interaction of bots and humans in discussion of the Russia/Ukraine war. *Soc. Inform.* **2022**, *13618*, 34–53.
30. Askasnr, S. End of US-Afghan War Tweet Data. 2012. Available online: https://www.kaggle.com/datasets/aska88/end-of-usafghan-war-tweet-data (accessed on 11 August 2021).
31. Ashish, K.; Abhishek, M.; Ayush, A.; Rachna, J.; Monika, A. Sentiment Analysis on Multilingual Data: Hinglish. In *International Conference on Data Analytics & Management*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 607–620.
32. Agarwal, N.S.; Punn, N.S.; Sonbhadra, S.K. *Exploring Public Opinion Dynamics on the Verge of World War III Using Russia-Ukraine War-Tweets Dataset;* Knowledge Discovery and Data Mining-Undergraduate Consortium: Washington, DC, USA, 2022.
33. Naz, H.; Ahuja, S.; Kumar, D.R. DT-FNN Based Effective Hybrid Classification Scheme for Twitter Sentiment Analysis. *Multimed. Tools Appl.* **2021**, 80, 11443–11458. [CrossRef]
34. Staal, N. War of the Tweets: An Analysis of American and Russian Information Operations on Twitter following the August, 2013 Sarin Gas Massacre in Syria. Royal Millitary Collge of Canada, 2016. Available online: https://espace.rmc.ca/jspui/handle/11264/1041 (accessed on 1 February 2024).
35. Chakravarthi, B.R. Hope speech detection in YouTube comments. *Soc. Netw. Anal. Min.* **2022**, *12*, 75. [CrossRef] [PubMed]
36. Bhatia, K.V. Hindu nationalism online: Twitter as discourse and interface. *Religions* **2022**, *13*, 739. [CrossRef]
37. Rastogi, S.; Bansal, D. Visualization of Twitter sentiments on Kashmir territorial conflict. *Cybern. Syst.* **2021**, *52*, 642–669. [CrossRef]
38. Srivastava, V.; Singh, M. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. *arXiv* **2021**, arXiv:2107.03760.
39. Srivastava, V.; Singh, M. PHINC: A Parallel Hinglish Social Media Code-Mixed Corpus for Machine Translation. *arXiv* **2020**, arXiv:2004.09447.
40. Kaur, G.; Kaur, A.; Khurana, M. A stem to stern sentiment analysis emotion detection. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13–14 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
41. Alslaity, A.; Orji, R. Machine Learning Techniques for Emotion Detection and Sentiment Analysis: Current State, Challenges, and Future Directions. *Behav. Inf. Technol.* **2024**, *43*, 139–164. [CrossRef]
42. Ruytenbeek, N.; Decock, S.; Depraetere, I. Experiments into the influence of linguistic (in) directness on perceived face-threat in Twitter complaints. *J. Politeness Res.* **2023**, *19*, 59–86. [CrossRef]

43. Sharif, W.; Mumtaz, S.; Shafiq, Z.; Riaz, O.; Ali, T.; Husnain, M.; Choi, G.S. An empirical approach for extreme behavior identification through tweets using machine learning. *Appl. Sci.* **2019**, *9*, 3723. [CrossRef]
44. Ramesh, T.; Lilhore, U.K.; Poongodi, M.; Simaiya, S.; Kaur, A.; Hamdi, M. Predictive analysis of heart diseases with machine learning approaches.*Malays. J. Comput. Sci.* **2022**, 132–148.
45. ElKafrawy, P.; Mahgoub, A.; Atef, H.; Nasser, A.; Yasser, M.; Medhat, W.M.; Darweesh, M.S. Sentiment Analysis: Amazon Electronics Reviews Using BERT and Textblob. In Proceedings of the 20th International Conference on Language Engineering, Cairo, Egypt, 12–13 October 2022.
46. Sievert, C.; Shirley, K. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 29 June 2014.
47. Chuang, J.; Manning, C.D.; Heer, J. Termite: Visualization Techniques for Assessing Textual Topic Models. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 21–25 May 2012; pp. 74–77.
48. Pratibha.; Kaur, A.; Khurana, M. Multimodal Hinglish Tweet Dataset for Deep Pragmatic Analysis. 2023. Available online: https://data.mendeley.com/datasets/y63frd6pmf/3 (accessed on 29 December 2023).
49. Verma, K.; Bhardwaj, S.; Arya, R.; Islam, U.; Bhushan, M.; Kumar, A.; Samant, P. Latest tools for data mining and machine learning. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 18-23.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.