Cross-Event User Reaction Prediction on a Social Network Platform

Pramod Bide[1]
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India 400058

Sudhir Dhage[2]
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India 400058

***Abstract*—Social network surges with multiple tweets with mixture of multiple emotions by many users when events like rape, robbery, war and murder, we use this user data to analyze user emotions between cross-events and try to predict user reactions for the next possible such event. Cross-events are a series of events that belong under the same umbrella of topics and are related to the events occurring prior to it. The proposed system solve this problem using collaborative filtering using Topical and Social context. The Text Rank Algorithm is an unsupervised algorithm used for keyword extraction. Count Vectorizer is used on preprocessed text to get the frequency of words throughout the text which is used as training data to get a probability of emotion using a logistic regression model. We incorporated social context along with topical context to account for homophily and used the Low-rank matrix factorization method for user-topic prediction. The model as an output gives a total of 8 emotions which include Shame, Disgust, Anger, Fear, Sadness, Neutral, Surprise and Joy. Finally, the model is able to predict emotions with an accuracy of 95% considering cross events.**

***Keywords*—Twitter; cross events; collaborative filtering; logistic regression; social and topical context**

I. Introduction

With the advancement in technology, global news can quickly spread across the globe in a very small amount of time. Popular news spreads quickly through the internet like wildfire and all the people across the world try putting forth their opinion on social media. Social media has given rise to a platform where people can express their views on a particular topic. It has led to a formation of a community where like minded people group together. One of the most popular social media platforms to voice opinions is twitter where opinions about various global news are shared. There are over 330 million active users on twitter monthly and 145 million active users on twitter daily. There have been 1.3 billion twitter accounts created ever since twitter was launched in 2006. Thus, twitter's popularity and purpose can be a platform where a person's subjective feeling about a particular news is found in what a person tweets about it. It is a reflection of emotions a person feels towards the news and can be therefore equated. Twitter data has been used for multiple analysis because it acts as a dataset which tells us about the user. It has been used for applications like emotion detection [8], opinion tracking [9] and so on. However in our research work we have focused on a completely different domain of exploration that is to predict how a user will react to various events or sub events based on his historical tweets. Through our model we can identify the emotion a user will show if any futuristic or hypothetical event is provided to a model. This kind of user reaction prediction can have a wide application in various domains. The government can use the data to predict the overall response on policies or it can also be used by companies to predict what a user reaction can be when a new product is launched in the market [10] [16] [17]. Additionally it can be used to model recommendations of new products [11].To achieve this goal, we model the user-topic opinion prediction problem as a collaborative filtering task and present the topical context and social context incorporated matrix factorization method (TcScMF) framework, which includes social context and topical context as regularisation constraints. This paradigm is quite broad, and it may easily be applied to various social network setups or expanded to include other requirements. A real-world data set is gathered from Twitter, and labelled positive/negative user-topic opinions are obtained for evaluation by assessing sentiment in the observed tweets with a credible tool. We compare the proposed framework to state-of-the-art collaborative filtering methods in the trials. The experimental results show that using the TcScMF framework with social and topical context improves prediction accuracy.

A key point in our methodology is that we make use of ingenious ways to gather meaningful and clean data from twitter. One major problem that we faced while improving the accuracy of the model was sparseness of the user topic matrix. It was indeed very difficult to identify topics where the set of users are closely related to each other and have reacted to similar topics. We solve this problem by improving our data collection methodology where we first focus on identifying closely related users followed by filtering out common topics where they have reacted. This allowed us to improve the accuracy of our model by twice its original value.

Other than this, we were also keen on finalizing the ideal metric to compare users and topics in order to incorporate social and topical context. We experimented with different vector similarity measures like Cosine Similarity, Soft Cosine Similarity, Jaccard Similarity and Spearman's Rank Correlation Coefficient and we concluded that the choice of the similarity function is closely related to the type of dataset that we are working with. In our case we proceed with cosine similarity considering the size of our dataset as well as the results reflected by our model.

Last but not the least, we also expand the range of emotions that we are working with. Most data analysing techniques usually classify the given data in a binary fashion. But we incorporate a range of emotions and make sure that our model

works with this range at every stage including the loss function and the accuracy measure step. This allows our model to be robust in nature and accommodate a diverse dataset and reflect the same in the results produced.

The structure of the remaining paper Is as follows. Section II is the literature survey that we undertook before working on this. It identifies the current research work and gaps in them. In Section III we talk about our method of identifying the dataset and what the dataset looks like. In Section IV we have described our model in detail. In Section V we talk about results and in Section VI what we conclude from the research work. In Section VII, we talk about the references.

II. Literature Survey

Nicolas Esquivel et al. [1] presents a CLSTM-NN to forecast the existence of crime events across Baltimore. Matrixes of previous criminal occurrences, in particular, are utilised as input to forecast the existence of at least single event that happens. The dataset was acquired from the Open Baltimore portal's Public Safety domain. The model performed better, with an accuracy of 0.86 utilising sequences of matrices of events that occurred over the course of seven days. The results suggest that model can learn past geographical patterns and forecast the presence of crimes in the future. The spatiao-temporal resolution of forecasts is hampered by poor performance for a small percentage of crime episodes.

Yizhou Xu et al. [2] proposed to model which would predict alarm events that are due, using similar alarm patterns in flood alarm sequences. It begins by arranging the alert patterns in order of resemblance to the current alert in descending order. A Bayesian approximation is used to calculate the probabilities and confidence levels for all projected alert events. When nuisance alerts were received, they were ignored and produced no anomalies.

A method that takes into consideration the behaviours and characteristics of the user, to identify and accurately predict hot events is given by Xichan Nie, et al. [3] The similar topics on twitter are collected and segregated using semantic similarity, all this is done after applying natural language processing on the keywords extracted from these tweets. Then a relationship between the users is derived. The user information proved to be useful and gave better results on experimentation. When compared to previous models, the suggested method enhanced prediction precision by 27%, 23.5 percent accuracy, as well as 20 percent recall rate, indicating that the model efficiently anticipated hot events. Other similar methods also work on identifying "social hotspots" such as Krishna et al. [23] and Xiao et al. [19].

Alberto Rossi, et al. [4] develops an attention mechanism as well as a LSTM network - RNN method for modelling taxi driver performance and storing the semantics of famous attractions in order to anticipate a cab's next destination using spatial location from LBSNs. The datasets used were taxi paths datasets. The results show that LSTM lowers the EDS in Porto and Manhattan by 10.5 percent and 18 percent, respectively, compared to MMLP. The suggested model, like most deep-learning algorithms, lacks explainability. Because travel utility features such as journey distance, cost, and time are not accessible at the time of the prediction, the approach cannot use them. Owing to the length limit, it may be difficult to create a precise estimation that will identify catastrophic occurrences using tweets, which may lack appropriate context and be difficult to discriminate due to word ambiguity. Song, Guizhe and Huang, Degen [5] designed a model named SentiBERT-BiLSTM-CNN which detects diaster using Tweets. To generate sentimental contextual embeddings from a Tweet the proposed pipeline uses SentiBERT, for feature extraction they used a 1D convolutional layer. The suggested model outperforms the competition in the F1 score, making it a viable model for Tweets-based disaster prediction. The CNN model gave a precision of 0.8064, BiLSTM gave 0.8571 while SentiBERT-BiLSTM-CNN gave a precision of 0.9305 making it the best model. Because specific keywords may occur in both catastrophe and non-disaster Tweets, the model's recommended accuracy can be considerably improved. However, it is difficult to successfully employ the words as additional information to help enhance the detection accuracy.

Gan, Mingxin and Xiao, Kejun [6] concluded that prior studies had failed to account for the sequential characteristics of users' click behaviour, hence the focus was on overcoming such restrictions. R-RNN is a model for understanding a user's interest from his general click history, according to the study. The Amazon Dataset was used to conduct the study. Few of the previous models are all outperformed by the recommended model, thanks to the newly introduced click behaviour patterns and the R-RNN for CTR prediction design's adopted RNN, which gathers user stats from the most recent click sequences.

Song et al. [7] suggested a semi-online Computational Offloading Model. Reinforcement learning is used to investigate user behaviour in a sophisticated action space in order to catch unknown environment information. The research proposes a dynamic edge computing simulation environment to show that user behaviour has a significant impact on system utilisation. According to their research, the mean size of offloading activities accomplished is roughly ten thousand. Large-scale Computation offloading projects could not be resolved using these strategies. This paradigm offloads compute chores based on changing contexts, although it is made up of MEC systems loosely.

A variety of media is used by different authors; C.Fu et al. [13] use micro blog user features, while Z.Zhang et al. [15] apply a situational analysis method on data from multimedia social networks.Other approach's such as M.Nyugen et al.'s hybrid generative model [12] and Chen et al's ensemble methodology [14] seek to combine multiple techniques to give a better result.

Through their work in [18] N Zhang et al. proposed a novel user behavior prediction model which uses automatic annotation. The model uses a combination of the Discontinuous Solving Order Sequence Mining (DVSM) behavior recognition model coupled with the LSTM based behavior prediction model. Factorization machines are applied to predict user behavior in the work of Y Wang et al. [20].

In their paper[22], Hao. et al. perform analysis on huge multi-data source of comprehensive quality. They analyse the user behavior acquisition and simulation prediction framework construction method which relies on user perception.

Hui Zhang et al.'s research[21] proposes a solution for user

prediction that concerns the single user - multiple terminal use-case. They perform weight correction that is based on adapted feedback, which is used to create a Markov model. This model can predict the user's future service states. A corresponding heavy tailed model is used to predict the duration of the service as well. Their work further integrates preference of service with the attributes of the terminal. This establishes a matching metric of services and terminal. This enables the proposal of terminal service model for recommending the best service terminal to each user.

III. Dataset

The dataset was collected by scrapping twitter data. The data collected by us have many users who might have tweeted over multiple different topics. We used sns scrape a python package to scrape data because of it increased limits.There were multiple major events identified and under every major events many sub events were considered. For example Covid-19 was the major event and the sub events under it were lockdown, vaccination and so on. The methodology we decided to use was to identify the top most 100 influential users which solved our problem of overlapping users. After that we scrapped only their tweets from their profile about our list of subtopics. The dataset thus formed had over 10,00,000 tweets across all sub events.

IV. Proposed Methodology

A. Keywords Extraction

A keyword extraction models goal is to automatically find a group of phrases in a tweet that best characterise the content.TextRank algorithm has been used for Keywords extraction from tweets. This algorithm is fully unsupervised.

Non-printable characters (if any) are removed from the raw input tweet before it is converted to lower case. The tokenization of the processed input text is performed using Natural Language Toolkit (NLTK) library methods. To allow the words to be lemmatized based on their Part-of-speech (POS) tags, Natural Language toolkit is used to Part-of-speech (POS) tag the input tweet. Lemmatization is used to normalise the tokenized text (mostly nouns and adjectives). Different grammatical equivalents of a word are replaced by a single fundamental lemma in lemmatization. The lemmatized text is then Part-of-speech (POS) labelled. Later on, the tags are utilised for filtering.

Any word from the lemmatized text that is not a noun, adjective, or a foreign word is regarded a stop word in this context. This is predicated on the premise that keywords are often nouns, adjectives, or gerunds. Punctuations are also included to the list of stopwords. Even after we eliminate the stop words, there may still be some exceedingly frequent nouns, adjectives, or gerunds that are poor candidates for being keywords. An external file containing a list of stopwords is loaded, and each word is added to the preceding stopwords to form the final list stopwords-plus, which is then turned into a set. Stopwords-plus are all stop words and possible phrase-delimiters combined.The contents of this collection are used to divide the lemmatized text into n-gram phrases later on. However, we simply eliminate the stop words and operate with a bag-of-words method. The stop words are then removed from the lemmatized text. Only unique words from the processed text are stored in a set.

TextRank is a graph-based approach, which necessitates the creation of a graph. Each word in the dictionary will act as a graph vertex. The phrases in the vertices will be indicated by their index in the list. The edge connections between all vertices are stored in the weighted edge matrix. A graph is made with undirected, heavy edges. The weight of the connecting edge between the word vertex represented by index a and the word vertex represented by b is stored in weighted edge[a][b]. When weighted edge[a][b] is 0, it signifies that there is no edge or relationship between the words represented by index a and b. If the words co-occur inside a window of a defined window size in the processed text, there is a relationship between them and hence between a and b, which represents them. For each link detected between the same words in various regions of the text, we raise the value of the weighted edge[a][b] by (1/(distance between positions of words now represented by a and b)). The covered co-occurrences list which contains of a list of pairs of absolute positions in processed text of words whose co-occurrence at that location has already been checked, is managed in such a way that the same two words located in the same positions in processed text are not counted repeatedly while sliding the window one text unit at a time. All vertices start with a score of one. Because self-connections are ignored, weighted edge [a][b] are 0 initially. The total number of undirected edges associated with the vertex represented by p are stored in x[p].

The formula for scoring a vertex represented by a is as follows:

$$s[a] = (1 - r) + r[\sum(j)((we[a][b]/x[b]) * s[b])] \quad (1)$$

In the Eq. (1), r is the damping factor and b is one of the vertices that has a relationship to a. 'we' represents weighted edge. The score is updated repeatedly until convergence is reached.

Using stopwords as delimeters, lemmatized text is partitioned into phrases. These phrases are also potential possibilities for keyphrase extraction. Then a list of unique phrases is created and removing single-word key-candidates in favour of multi-word alternatives. Candidate-key phrases are scored and a list of key phrases is compiled by listing untokenized variants of tokenized words. Phrases are graded by summing the scores of their constituents i.e. words or text-units that were ranked by the graph algorithm. Keys are ranked based on their computed scores.

B. Emotion Detection

For setting up the emotion detection domain of our proposed solution, we leveraged neattext python library. A data set having more than 34 thousand tweets and its corresponding emotion was used for this. Firstly, all the username and tags were removed from each individual tweet and after that removal of stopwords took place. After this step all the stopwords like 'an', 'be', 'some', 'for', 'do', 'its', 'of', 'as', etc. are removed from the tweets for optimising the emotion prediction, reducing the computation time and resources.

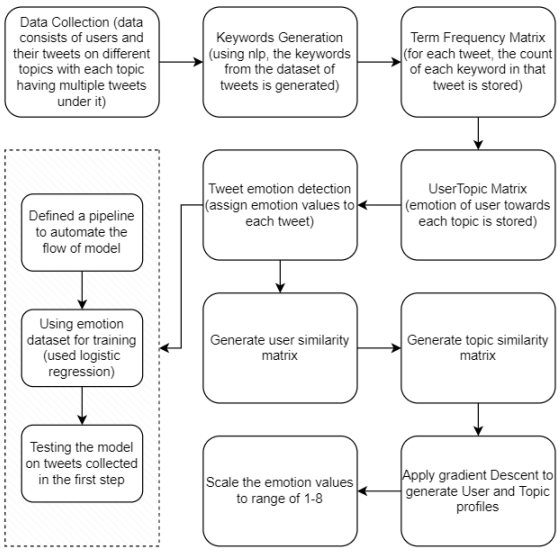

Fig. 1. Social and Topical Context System Framework.

The algorithm was trained using a dataset of 34762 tweets that included their moods. The data set is separated into training and test sets after usernames, tags, and stopwords are removed, and the model is trained on this data. Count Vectorizer is a tool that converts a text into a vector depending on the count of each word in the text. Logistic Regression is a machine learning classification algorithm which is applied to train the model precisely and predict the probability of an emotion. Finally, the web scraped twitter data set is passed to the model and emotion of each tweet is predicted.

C. Social and Topical Context System Framework

In the above Fig. 1, the basic flow of the execution of the model is given, starting from data collection to getting the results from the trained model. The initial step is the data collection, in which data about different twitter users and the tweets of specific topics they have tweeted on, is gathered. The next step is the processing of the tweets to generate the keywords file, which is done using nlp(natural language processing) which has been described above. Then the term frequency matrix is generated following which the UserTopic matrix is generated. Emotion detection model is applied on the collected tweets, the process of which is given in detail in the 'Emotion Detection' section. After completing these steps, we then generate the similarity matrices for users and tweets and after training the model, we then scale the emotion values which we get in the result in the range of 1-8.

D. Model Outline

The main purpose of the model is to analyse the users in a given set and predict their reactions on related events. It has been historically proven that the views of any user is influenced by the closest networks of users with similar interests as the user in question along with the similarity of the actual topic to other topics the user has already reacted upon. Although this system should be enough to deduce the reaction of a user in any two non-related events as well, we found that solely relying on user similarity or topic similarity yielded poor results as compared to using both of them. The most effective use of the above mentioned strategy is in the case of multi-phase events which are closely related to each other. At the same time, the similarity measures used to calculate the similarity of different users and different topics also plays a key role in the preparation of the model. Inline with our previous approach, we experimented with different similarity measures, some of which are Cosine Similarity, Manhattan Distance, Jaccard Similarity, Euclidian Distance and we found that the choice of the similarity function is closely related to the type of dataset we are working with. In our original scenario where we were dealing with sparse data, we found that Cosine Similarity was very effective however, implementing the same for dense matrices proved to be computationally expensive and in such scenarios, a simple Euclidian calculation was far easier.

E. Data Scarcity Problem and our Solution

In a quest to find similar users who reacted on similar topics, we found that such datasets are very difficult to acquire and in most cases, the data itself appeared to be biased as we would end up reading multiple accounts of the same user. Eliminating the above problem left us with a very scarce user-topic matrix where only a couple of users might have reacted on the same topic and vice versa. In order to make the user-topic matrix denser, we come up with a unique solution which is explained as follows; We first divide the user-topic matrix into different sections, essentially following a divide and conquer strategy. In the smaller subsections of the matrix, we make sure that multiple users have tweeted or reacted on the same topic and similarly, a single user has tweeted on multiple topics in the same sub-section. We found experimentally, that such local optimizations eventually lead to global patterns, and helps us develop a more robust model.

F. Social and Topical Context Mode Framework

We follow the below mentioned algorithm which incorporates our unique approach to solve the data scarcity problem and gives an outline of the model pipeline which is explained in detail later in the paper. Table I indicates all the parameters used in designing the Social Context and Topical Context Mode Framework (ScTcMF).

G. Example

Let say we have 4 users, User A, User B, User C and User D. We'll consider the main topic Covid and use the tweets by the users on the sub topics - 1st Dose, Omicron, Booster Dose and Lockdown.

User A (Newsweek) tweeted "Kate Middleton shares COVID vaccine photo, "hugely grateful" after 1st dose" on the sub topic 1st Dose and the emotion for the tweet was predicted to be Joy. User A tweeted "The highly mutated and contagious Omicron variant has driven the country to record the most cases in a day that any country has reported." on the sub topic Omicron and the emotion for the tweet was predicted to be Joy. User A tweeted "People struggling to get booster dose appointments could leave millions of Americans vulnerable as Omicron spreads." and the emotion for the tweet was

TABLE I. IMPORTANT PARAMETERS IN OUR MODEL

Sr. no.	Parameter	Description
1	α	Balances the error function between front terms and the social context normalisation terms
2	β	Controls the topical context's regularisation requirement
3	λ_0	Controls the reach of regularization
4	λ_1	Controls the reach of regularization
5	A	User-Topic Opinion Matrix
6	E	Latent Representation of User
7	G	Latent Representation of Topic
8	$\|.\|_F$	Frobenius norm of a matrix
9	P	Indicator Matrix
10	$\odot$	Hadamard product
11	Q(v,w)	Weighted number representing the similarity of two social friends' past opinions
12	F(v)	Set of social friends of u(v)
13	Tr(.)	Matrix Trace
14	$Diag_Q$	Diagonal Matrix
15	L_Q	Laplacian Matrix

Algorithm 1 Social and Topical Context Model Framework(ScTCMF)

1) Collect data for different kinds of events and identify sub-events of each of the events.
2) Identify the most active users from each of the sub-events and collect their twitter data.
3) Use the keywords extraction pipeline to extract important words from each and every tweet. This is further used for emotion detection.
4) Use the emotion detection pipeline to understand the text of the tweets and extract meaningful emotions in a range of 1-8.
5) Use TF-IDF to calculate the frequencies of each word of each tweet across multiple tweets of the same sub-event.
6) Initialize hyper-parameters like social and topical regularization factors and learning rate.
7) Calculate user and topic similarity using Laplace reductions and the original frequency matrix calculated previously. While doing so, choose an appropriate similarity function.
8) Perform stochastic gradient descent to estimate the user-topic matrix thereby predicting the user's reaction to unknown events.
9) Compare the result with the original matrix to retrieve the accuracy of the model. Different deviations can be used to calculate the accuracy in different scenarios.

predicted to be Disgust. User A tweeted "Xi'an residents can't leave homes, purchasing food difficult as COVID lockdown continues" on the sub topic Lockdown and the emotion for the tweet was predicted to be Sadness.

User B (bsindia) tweeted "TOP HEADLINES — PM @narendramodi receives his 1st #Covid19vaccine dose at AIIMS; @TheOfficialSBI reduces #homeloan rates to 6.7% ; European #stocks rebound as bond markets stabilise and more..." on the sub topic 1st Dose and the emotion for the tweet was predicted to be Joy. User B tweeted "Amid surging COVID-19 cases, the Delhi Disaster Management Authority has decided to impose a weekend curfew in the national capital." on the sub topic Omicron and the emotion for the tweet was predicted to be Anger. User B tweeted "A Delhi-based doctor said that a booster dose of Cov vaccine is a must as the protection cover of two doses declines over three to six months" on the sub topic Booster Dose and the emotion for the tweet was predicted to be Sadness. User B tweeted "#Mumbai Mayor Kishori Pednekar on Tuesday said if the daily COVID-19 cases here cross the 20,000-mark, a #lockdown will be imposed in the city as per the Union government's rules." on the sub topic Lockdown and the emotion for the tweet was predicted to be Anger.

User C (ChannelNewsAsia) tweeted "Duchess of Cambridge 'hugely grateful' for 1st vaccine dose" on the sub topic 1st Dose and the emotion for the tweet was predicted to be Joy. User C tweeted "Commentary: With Omicron threat, will returning to offices and schools bring new anxieties?" on the sub topic Omicron and the emotion for the tweet was predicted to be Joy. User C tweeted "Australia to shorten COVID-19 booster dose intervals from January" on the sub topic Booster Dose and the emotion for the tweet was predicted to be Joy. User C tweeted "Amid Omicron surge, UK PM Johnson resists another lockdown" on the sub topic Lockdown and the emotion for the tweet was predicted to be Neutral.

User D (TheQuint) tweeted "#Live — Around 90 percent adult population #vaccinated With #1stDose." on the sub topic 1st Dose and the emotion for the tweet was predicted to be Sadness. User D tweeted "#Podcast — Given the explosive growth of #COVID cases in India, are we underplaying the threat of #Omicron and its potential impact on our fragile health care system? We discuss with @MenonBioPhysics and @RajeenJayadevan. Tune in!" on the sub topic Omicron and the emotion for the tweet was predicted to be Fear. User D tweeted "#BharatBiotech on 20 Dec, sought approval from the #DCGI for the conduction of phase-3 trials for its intranasal vaccine (BBV154), which is to be used as a booster dose." on the sub topic Booster Dose and the emotion for the tweet was predicted to be Fear. User D tweeted "#LIVE — 'We will have to impose #lockdown in #Mumbai if daily #COVID cases cross the 20,000-mark,' said Mayor Kishori Pednekar." on the sub topic Lockdown and the emotion for the tweet was predicted to be Anger.

Now we'll take a different sub topic say "Social Distancing" and we'll see the tweets and its emotions predicted by our model for the above users. The model predicts that User A will tweet "Many health experts are now calling on the South Korean government to reimpose social distancing measures." on the given sub topic Social distancing and the emotion predicted by the model for the tweet was found

to be Sadness. The model predicts that User B will tweet "#Britain's economic growth slowed more than expected in July as concern about the spread of the delta variant of #Covid-19 overshadowed the government's decision to end most social distancing rules" on the given sub topic Social distancing and the emotion predicted by the model for the tweet was found to be Fear. The model predicts that User C will tweet "France plans tighter social distancing rules, booster ramp-up to fight COVID-19 wave" on the given sub topic Social distancing and the emotion predicted by the model for the tweet was found to be Sadness. The model predicts that User D will tweet "Amid videos of crowds flouting social distancing norms in #SarojiniNagar market surfaced, the Delhi High Court on 24 December rapped the New Delhi Municipal Council and Delhi Police for allowing illegal vendors to operate from there." on the given sub topic Social distancing and the emotion predicted by the model for the tweet was found to be Fear.

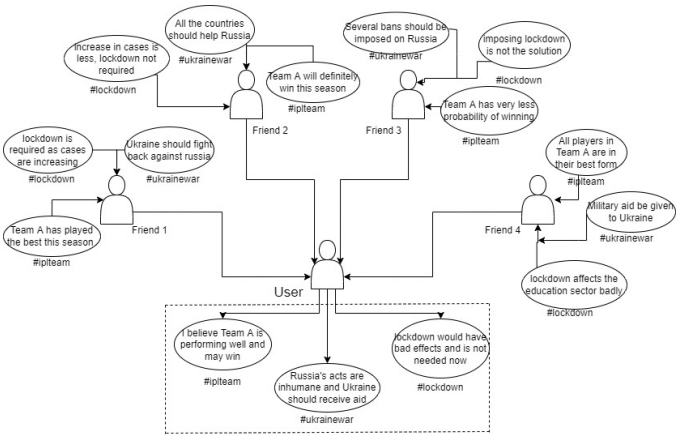

Fig. 2. Opinion Prediction.

H. User Topic Matrix

For user topic view prediction, a low rank matrix factorization algorithm was applied. The user-topic prediction model incorporates social and topical context mathematically. Matrix Factorization is commonly used in cutting-edge Collaborative filtering works.

In our scenario, because the user topic opinions in Twitter data is so less, the matrix A is highly scattered. Based on the assumptions that only a few factors impact the opinions, a more concise but accurate depiction is provided and goal to replicate the matrix by a multiplication of low rank factors.

$$A \approx EG^T \tag{2}$$

In Eq. (2), E $\in \mathbb{R}^p$ and G $\in \mathbb{R}^q$ where $p = a * c$ where $q = b * c$ with c << min(a,b) the row vector $E(k,:)$, $1 \leq$ k $\leq$ a and $G(o,:)$, $1 \leq$ o $\leq$ b are the existing depictions of user e_k and topic g_o. Here R denotes Real numbers, E is the Latent Representation of user and G represents the Latent representation of Topic. By minimising the following aim, the matrix factorization technique approximates the matrix A given in Eq. (3):

$$\min_{E,G} \|A - EG^R\|_F^2 \tag{3}$$

In order to simulate the labelled viewpoints i.e. opinions we make use of matrix P $\in \mathbb{R}^r$ where r = a*b. If e_k provided his view for g_o then we will consider the value at k,o index in the matrix P to be 1 and equal to 0 otherwise. Normalisation terms have been added to avoid overfitting. λ_0 and λ_1 are the control parameters. The basic low rank factorisation model is as follows illustrated by Eq. (4):

$$\min_{E,G} \|P \odot (A - EG^R)\|_F^2 + \lambda_0 ||E||_F^2 + \lambda_1 ||G||_F^2 \tag{4}$$

Fig. 2 illustrates the opinion prediction for user under influence. As shown, the user has four friends with different opinions on three different topics and based on the opinions the user's opinion on these topics is impacted. Considering the topic "#Ukrainewar" the user's friends have shown support to Ukraine in their tweets and also few tweets with support to Russia. The opinion of the user for this topic is also seen to be in favour of Ukraine. Thus the tweet of the user shows a similar emotion to that of the majority of his followers or tweets of people he/she sees.

I. Implementing Social Context

Homophily is a phenomenon that occurs in social networks. Forming a following connection on any social media platform typically suggests that the follower and the buddy have similar interests, and hence have more comparable perspectives on the same issue. Both users friends and followers are considered to be his social buddies. To establish the social context, we convert the directed network to undirected network. The Twitter users' social environment may thus be represented as an undirected weighted network with a symmetric adjacency matrix Q. Social friends are more likely to have similar viewpoints on issues than non social friends. Based on the above premise, we evaluate previous opinion similarity across social friends to represent user viewpoint variety and suggest a normalisation constraint as shown below,

$$\frac{1}{2} \sum_{v=1}^{a} \sum_{w \in F(v)} Q(v,w) \|E(v,:) - E(w,:)\|_F^2 \tag{5}$$

In Eq. (5), F(v) indicates a group of social acquaintances and Q(v,w) is the weighted number representing the similarity of two social friends' past opinions. Large value of Q(v,w) indicates that divergence will be less and vice versa.

Matrix form of the above equation,

$$\begin{aligned} X &= \frac{1}{2}\sum_{v=1}^{a}\sum_{w\in F(v)} Q(v,w)\|E(v,:)-E(w,:)\|_F^2 \\ &= \frac{1}{2}\sum_{v=1}^{a}\sum_{w\in F(v)}\sum_{x=1}^{c} Q(v,w)(E(v,x)-E(w,x))^2 \\ &= \sum_{v=1}^{a}\sum_{w\in F(v)}\sum_{x=1}^{c} Q(v,w)[E^2(i,k)-E(v,x)E(w,x)] \\ &= \sum_{x=1}^{c} E^T(:,x)(Diag_Q - Q)E(:,x) \\ &= Tr(E^R L_Q E) \end{aligned} \tag{6}$$

In Eq. (6), $Diag_Q$ is the diagonal matrix, L_Q is the laplacian matrix and Tr denotes the trace of the matrix. The resulting Matrix Factorisation model which involves social context is as follows illustrated by Eq. (7),

$$\begin{aligned} \min_{E,G} \|P\odot(A-EG^O)\|_F^2 + \lambda_0||E||_F^2 + \\ \lambda_1||G||_F^2 + \alpha Tr(E^O L_Q E) \end{aligned} \tag{7}$$

The regularisation parameter balances the reconstruction error between the social context regularisation term and the front terms, $\alpha \geq 0$. The User Opinion Similarity (UOS) is calculated by the given formula in Eq. (8),

$$UOS(e_v,e_w) = \frac{\sum_{x=1}^{b} A_{i_x}\cdot A_{j_x}}{\sqrt{\sum_{x=1}^{b} A_{v_x^2}}\sqrt{\sum_{x=1}^{b} A_{w_x^2}}} \tag{8}$$

A mapping is used to constrain the range of User Opinion Similarity between [0,1] where $UOS(e_v,e_w)$ = $(UOS(e_v,e_w)+1)/2$ is given by Eq. (9),

$$Q(v,w) = \begin{cases} UOS(e_v,e_w), & \text{if } e_w \in F(e_v) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

J. Implementing Topical Context

The topical context interpretation model is that consumers will supply more similar opinions to two topics that are more related in content.

$$\frac{1}{2}\sum_{v=1}^{b}\sum_{w=1}^{b} R(v,w)\|G(v,:)-G(w,:)\|_F^2 \tag{10}$$

In Eq. (10), G(v,w) represents the similarity index between the topics rv and rw. Greater value of R(v,w) infers that the topics rv and rw are very similar to each other having more similar opinions of users. On the other hand, low value of R(v,w) infers that the topic representations G(v,:) and G(w,:) have large distance between them. As per the topical context derivations mentioned earlier, we can obtain the matrix equivalent of the above equation as follows in Eq. (11):

$$\begin{aligned} Z &= \frac{1}{2}\sum_{v=1}^{b}\sum_{w=1}^{b} R(v,w)\|G(v,:)-G(w,:)\|_F^2 \\ &= \sum_{x=1}^{d} G^R(:,x)(Diag_R - R)G(:,x) \\ &= Tr(G^R L_R G) \end{aligned} \tag{11}$$

Likewise, $Diag_R$ is a diagonal matrix and L_R is the laplacian matrix. The model with topical context regularization is given below in Eq. (12),

$$\begin{aligned} \min_{E,G} \|P\odot(A-EG^R)\|_F^2 + \lambda_0||E||_F^2 + \\ \lambda_1||G||_F^2 + \beta Tr(G^R L_R G) \end{aligned} \tag{12}$$

Here, A regularisation parameter $\beta \geq 0$ is used to adjust the regularisation requirement of topical context. After analyzing many approaches for measuring similarity using topic distribution, we chose Cosine Similarity due to its efficacy and simplicity. By using the unique terms showing up in the tweets collection as features upon removing stop words and the term frequency as a feature value, a cosine similarity between term frequency vectors could have been used to quantify the content based resemblances between similar topics. A term frequency vector fvi could be formed for each topic ti by taking the unique terms making an appearance in the tweets gathering as features after removing stop words and the term frequency as a feature value. The similarity values in this definition vary from 0 to 1 since the word frequency cannot be negative. As a result, Topic Content Similarity was assigned (TCS).

$$TCS(r_v,r_w) = \frac{\sum_{x=1}^{B} fv_{v_x}\cdot fv_{w_x}}{\sqrt{\sum_{x=1}^{B} fv_{v_{x^2}}}\sqrt{\sum_{x=1}^{B} fv_{w_{x^2}}}} \tag{13}$$

In Eq. (13), fvv and fvw denote the term frequency vectors of rv and rw respectively, and B is the number of features in the vectors. Finally the element R(v,w) can be shown as follows in Eq. (14):

$$R(v,w) = \begin{cases} TCS(r_v,r_w), & \text{if } v \neq w \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

K. Framework with Topical and Social Context

We created and comprehended functions for social context and topical context, and then used them to describe regularisation restrictions. In this part, we develop a shared framework that incorporates both social and topical background.

This framework is used to minimize the objective function

given below in Eq. (16):

$$f(E,G) = \|P \odot (A - EG^R)\|_F^2 + \lambda_0||E||_F^2 + \lambda_1||G||_F^2 + \frac{\alpha}{2}\sum_{v=1}^{a}\sum_{w\in F(v)} Q(v,w)\|E(v,:) - E(w,:)\|_F^2 + \frac{\beta}{2}\sum_{v=1}^{a}\sum_{w=1}^{b} R(v,w)\|G(v,:) - G(w,:)\|_F^2 \quad (15)$$

$$f(E,G) = \|P \odot (A - EG^R)\|_F^2 + \lambda_0||E||_F^2 + \lambda_1||G||_F^2 + \alpha Tr(E^R L_Q E) + \beta Tr(G^R L_R G) \quad (16)$$

where $\alpha \geq 0$ and $\beta \geq 0$ are the social and topical normalisation parameter that can be changed to impact the findings in different ways. α=0 and $\beta > 0$ infers that the framework discusses just the topical context. On the other hand $\alpha > 0$ and β = 0 infers that the framework discusses just the social context. When α = 0 and β = 0 the method reverts to the simplest matrix factorization. In the proposed method it is suggested to keep $\alpha > 0$ and $\beta > 0$.

For proper training and simplicity purpose we implement gradient descent to find the local maxima/minima and solve the objective function, thus updating $E_{r_{+1}}$ and $G_{r_{+1}}$. The formula for which are given in Eq. (17) and Eq. (18),

$$E_{r_{+1}} = E_r - \gamma \nabla_E f(E_r, G_r) \quad (17)$$

$$G_{r_{+1}} = G_r - \gamma \nabla_G f(E_r, G_r) \quad (18)$$

Here the step size is γ, $\nabla_E f(E_r, G_r)$ and $\nabla_G f(E_r, G_r)$ are gradients in step r+1 which are termed as partial derivatives to E and G,

$$\nabla_E f(E_r, G_r) = -2(P \odot A_r)G_r + 2P \odot (E_r G_r^R)G_r + 2\lambda_0 E_r + 2\alpha L_q E_r \quad (19)$$

$$\nabla_G f(E_r, G_r) = -2(P \odot A_r)^R E_r + 2(P \odot (E_r G_r^R))^R E_r + 2\lambda_1 G_r + 2\beta L_R G_r \quad (20)$$

V. Experimental Results

Our proposed model compares non zero values of the original matrix with a value of the model predicted matrix. We have considered a total of 8 emotions which include Shame, Disgust, Anger, Fear, Sadness, Neutral, Surprise and Joy and they were numbered from 1-8.

When there was no deviation i.e. we tried for an exact match of emotions we get an accuracy of 65%. If we allow the prediction model to have a deviation of +1 or -1 from the original emotion, we get an accuracy of 84%. If we allow the prediction model to have a deviation of +2 or -2 from the original emotion, we get an accuracy of 95%.

We must note that +1 or -1 deviation should also be acceptable since many a times shame can be misjudged as disgust and vice versa. The model generated the best results when the gradient descent steps were fixed at 100. In both the original matrix factorization and the regularisation terms, the regularisation parameters are set to balance the reconstruction error. As a result, α and β are critical in deciding how much the framework approach can benefit from the social and topical environment's regularisation limits.

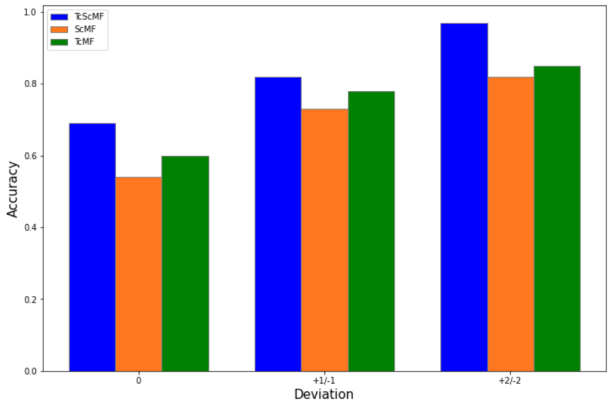

Fig. 3. Accuracy Comparison.

Fig. 3 gives a brief description of the accuracy of the model with different toleration levels. In order to calculate the accuracy of the model we first obtain a matrix that contains known values. This is because it is not necessary that every user in our original dataset or test set has reacted to every different tweet in the dataset. Hence to eliminate verifying against known data we first take a subset of the original dataset and maintain only known values in the matrix. The format of the matrix is constructed in a manner that each row represents unique users in the dataset and the columns represent unique topics. Hence each cell represents the opinions of a user on a given topic. Since our scale of emotions ranges from 1-8 we replace the value of cells with unknown values with 0. Our model is able to predict the reaction of every user on every topic, hence it returns a populated user and topic matrix whose product results in a matrix that is similar to the original user topic matrix. Once we obtain the predictions of the model, we iterate over each user and consecutively each topic and then compare the predicted value with the original value. Note that this is done only for cells whose value is not 0 in the original user topic matrix. Next we define a tolerance level which ranges from 0 to 2. This means that if a user had a reaction of 1 in the original matrix and if the model predicted 2, then we will consider this as a correct prediction only if the tolerance level is 1 or 2. This is because we assume that a tolerance level of 1 allows a deviation of 1 from the original value. Let's consider an example where we have users A and B and topics X and Y. Now user A has only reacted to topic X with an emotion of 2 and user B has reacted to topic Y with an emotion of 7. The original matrix in this case will look as follows: [{2, 0}, {0, 7}]. Now let's assume that after passing our data through the model we obtain a predicted matrix which

looks like this: [{3, 1}, {6, 7}]. Now while calculating the accuracy of the model we will iterate over user A and B and in a nested for loop iterate over topics X and Y. We will first encounter the first cell (0, 0). We will consider this value since the user A had reacted to topic X in the original scenario. Now if the tolerance level is 0, the predicted and original value will have a difference of 1 which does not fall within our tolerance level/deviation allowed. Hence we will not consider this as a correct prediction. Next we will ignore cells (0, 1) and (1, 0) since their values are not known (0) in the original dataset. Moving onto cell (1, 1), we will consider this value as a correct prediction since the difference between the predicted value and original value is 0 and they are an exact match. Hence in this scenario, out of 2 known values our model was able to predict 1 value properly as per our tolerance level and hence the accuracy turns out to be ½*100 which is 50%.

Fig. 3 displays the Accuracy comparison of the methods ScMF, TcMF and the ScTcMF incorporated by us. The results prove that ScTcMF: Framework with Topical and Social Context is the most efficient method. While testing our model, The values of λ_0 and λ_1 in the matrix factorization method was set to 1 and the values of α was set to 10 and that of β was set to 0.01. The Learning Rate was considered to be 0.01 and Number of Steps was set to 100. The Hyper parameters were tweaked so that we could get better accuracy.

VI. Conclusion

Our objective while writing this paper was to create a model and a framework that utilizes previous user reactions given a series of cross events to achieve user topic opinion prediction for future events. This was a novel idea because there are solutions that are used in predicting user emotion, but none of them explored this across cross events, which had its own set of challenges. We had to go through multiple iterations of data collection to get the right dataset which included multiple users that have tweeted on multiple topics. In this approach, we searched users on the basis of their activity across all the tweets and the top 100 among them were used as the dataset. This was crucial as a sparse matrix used for the low-rank factorization method would have resulted in poor model accuracy. Using this data, the keywords were extracted and the emotion was detected on a scale of 1 to 8 using the ScTcMF method. Finally, the results demonstrated that both social context and topical context can help improve the performance of the user-topic opinion prediction and by incorporating them we were able to get the model to the desired accuracy.

References

[1] Nicol áas Esquivel, Orietta Nicolis, Billy Peralta e Jorge Mateu. “Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks”. Em: IEEE Access 8 (2020), pp. 209101–209112. DOI: 10.1109/ACCESS.2020.3036715.

[2] Yizhou Xu, Jiandong Wang e Yan Yu. “Alarm Event Prediction From Historical Alarm Flood Sequences Ba- sed on Bayesian Estimators”. Em: IEEE Transactions on Automation Science and Engineering 17.2 (2020), pp. 1070–1075. DOI: 10.1109/TASE.2019.2935629.

[3] Xichan Nie, Wanshan Zhang, Yang Zhang e Dunhui Yu. “Method to Predict Bursty Hot Events on Twitter Based on User Relationship Network”. Em: IEEE Access 8 (2020), pp. 44031–44040. DOI: 10.1109/ACCESS.2020. 2977424.

[4] Alberto Rossi, Gianni Barlacchi, Monica Bianchini e Bruno Lepri. “Modelling Taxi Drivers’ Behaviour for the Next Destination Prediction”. Em: IEEE Transactions on Intelligent Transportation Systems 21.7 (2020), pp. 2980– 2989. DOI: 10.1109/TITS.2019.2922002.

[5] Guizhe Song e Degen Huang. “A Sentiment-Aware Con-textual Model for Real-Time Disaster Prediction Using Twitter Data”. Em: Future Internet 13.7 (2021). ISSN: 1999-5903. DOI: 10.3390/fi13070163. URL: https://www.mdpi.com/1999-5903/13/7/163.

[6] Mingxin Gan e Kejun Xiao. “R-RNN: Extracting User Recent Behavior Sequence for Click-Through Rate Pre-diction”. Em: IEEE Access 7 (2019), pp. 111767–111777. DOI: 10.1109/ACCESS.2019.2927717.

[7] Shinan Song, Zhiyi Fang, Zhanyang Zhang, Chin-Ling Chen e Hongyu Sun. “Semi-Online Computational Offlo-ading by Dueling Deep-Q Network for User Behavior Prediction”. Em: IEEE Access 8 (2020), pp. 118192–118204. DOI: 10.1109/ACCESS.2020.3004861.

[8] B. Pang and L. Lee, “Sentiment Analysis and Opinion Mining,” Foundations and Trends in Information Retrieval, vol. 1, pp. 1-135, 2008.

[9] X. An, R. A. Ganguly, Y. Fang, B. S. Scyphers, M. A.Hunter, and G. J. Dy. Tracking Climate Change Opinions from Twitter Data. In KDD’14: Workshopon Data Science for Social Good, 2014.

[10] Lynn R. Kahle (1997, June). The real-time response survey in new product research: it’s about time. Journal of Consumer Marketing, Vol 14, Issue 3, pp.234 – 248

[11] Z. Su, Z. Lin, J. Ai and H. Li, ”Rating Prediction in Recommender Systems Based on User Behavior Probability and Complex Network Modeling,” in IEEE Access, vol. 9, pp. 30739-30749, 2021, doi: 10.1109/ACCESS.2021.3060016.

[12] M. Nguyen and Y. Cho, ”A Hybrid Generative Model for Online User Behavior Prediction,” in IEEE Access, vol. 8, pp. 3761-3771, 2020, doi: 10.1109/ACCESS.2019.2962539.

[13] C. Fu et al., ”Forwarding Behavior Prediction Based on Microblog User Features,” in IEEE Access, vol. 8, pp. 95170-95187, 2020, doi: 10.1109/ACCESS.2020.2995411.

[14] L. Chen and H. Deng, ”Predicting User Retweeting Behavior in Social Networks With a Novel Ensemble Learning Approach,” in IEEE Access, vol. 8, pp. 148250-148263, 2020, doi: 10.1109/ACCESS.2020.3015397.

[15] Z. Zhang, R. Sun, X. Wang and C. Zhao, ”A Situational Analytic Method for User Behavior Pattern in Multimedia Social Networks,” in IEEE Transactions on Big Data, vol. 5, no. 4, pp. 520-528, 1 Dec. 2019, doi: 10.1109/TBDATA.2017.2657623.

[16] G. Zhao, X. Qian and X. Xie, ”User-Service Rating Prediction by Exploring Social Users’ Rating Behaviors,” in IEEE Transactions on Multimedia, vol. 18, no. 3, pp. 496-506, March 2016, doi: 10.1109/TMM.2016.2515362.

[17] Q. Yanfang and L. Chen, ”Research on E-commerce user churn prediction based on logistic regression,” 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017, pp. 87-91, doi: 10.1109/ITNEC.2017.8284914.

[18] N. Zhang, Y. Yan, X. Zhu and J. Wang, ”A novel user behavior prediction model based on automatic annotated behavior recognition in smart home systems,” in China Communications, doi: 10.23919/JCC.2022.00.005.

[19] Y. Xiao, J. Li, Y. Zhu and Q. Li, ”User Behavior Prediction of Social Hotspots Based on Multimessage Interaction and Neural Network,” in IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 536-545, April 2020, doi: 10.1109/TCSS.2020.2969484.

[20] Y. Wang, W. Shang and Z. Li, ”The application of factorization machines in user behavior prediction,” 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-4, doi: 10.1109/ICIS.2016.7550927.

[21] H. Zhang and J. Chen, ”A Novel User Behavior Prediction and Optimization Algorithm for Single-User Multi-terminal Scenario” 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 2016, pp. 144-147, doi: 10.1109/ISCID.2016.2042.

[22] C. Hao and Y. Zhou, ”Design and Implementation of User Behavior Acquisition and Simulation Prediction Framework for Mobile Intelligent Terminal Based on User Perception,” 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), 2020, pp. 125-128, doi: 10.1109/TOCS50858.2020.9339698.

[23] B. Ravi Krishna, P. Akhila, S. Sowjanya and B. Keerthana, "Prediction of Hot Topic in Social Media Based on User Participation Behavior in Social Hotspots," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, pp. 1545-1548, doi: 10.1109/ICECA52323.2021.9676120.