Computational Linguistic Approach for Holistic User Behaviors Modeling Through Opinionated Data of Virtual Communities

Kashif Asrar, Syed Abbas Ali
Department of Computer & Information System Engineering, NED University of Engineering & Technology, Karachi, Pakistan

***Abstract*—This research is aimed at establishing a computational linguistic model for the detection of positive and negative statements, synthesized for the Pakistani microblogging site Twitter, particularly, in the Roman Urdu language. With increased freedom of speech people express their sentiments towards an event or a person in positive, negative, neutral, and sometimes sarcastic tones, especially on social media platforms. Pakistani social media users, like other multilingual countries, express their opinions through code switching and code mixing. Their language lacks correct grammar, informal and nonstandard writing, unrelated spelling, alternative analogies make it difficult for computational linguist to mine their data for computational research. To overcome this challenge, the study employed web scraping tools to retrieve a large number of Roman Urdu tweets. In order to establish a new positive and negative statements corpus, the text data is annotated through a sentiment analysis carried out by using TextBlob sentiment analysis and Bidirectional Encoder Representations from Transformers (BERT). Addressing this issue makes it possible to eliminate the gap that is evident in the models that do not identify Roman Urdu as a form of language. The findings are useful for the regulatory bodies and researchers since it offers a culturally and linguistically appropriate database and model targeting resource constraints and key performance metrics. It helps in content moderation and in making policies regarding the technological advancement within Pakistan.**

***Keywords*—Roman Urdu; positive and negative statements detection; sentiment analysis; BERT Model; Long Short-Term Memory (LSTM) networks; Pakistani social media**

I. Introduction

A. Background

The advancement in social media and virtual group discussions has created large databases of people's opinions in form of ratings, comments, discussions, and social media posts. This data comprises various user actions, tastes, and attitudes, which makes it highly valuable to computational modeling. Many of the large volumes of textual data originate from users and thus require a Computational Linguistic approach, which incorporates both natural language processing, sentiment analysis, and machine learning to build and model user behavior comprehensively [1]. Opinionated data is highly useful when it comes to capturing user behaviors because it is genuine. Unlike ordered or systematic data or simple questionnaires, this one is not elicited, which means it is voluntary and unbiased. Online reviews, which are feedbacks given by customers on e-commerce ventures, for example, not only show satisfaction levels of users, but also the features of product that bring out satisfaction or dissatisfaction. Just like that, engagements with forums offer information on cohesion within the community, the degrees of interest, and developing trends in a given society. Such data is usually sparse, relatively unorganized, noised and contextual in nature which makes it difficult to process. These concerns are well met by computational linguistics, allowing for obtaining insightful information reflecting the complexity of human activity [2].

There are some principles applied in computational linguistic strategies to apprehend users' behaviors. Sentiment Analysis, aimed at finding the positive, negative or neutral attitude of a text document, tops the list. Sentiment analysis may also provide understanding of the overall user satisfaction or frustration or enthusiasm. Techniques such as the deep learning algorithms are capable to capture features including sarcasm or viewing a video with mixed feelings, offering better perception on user attitudes. Another effective technique is topic modeling for investigation of the overall message in vast textual materials [3]. Discussions are sorted into topics using methods such as Latent Dirichlet Allocation (LDA) that create insights about users' interests, worries and emerging trends[16]. For instance, in analyzing a virtual health community, name of the game could be fitness, mental health, or nutrition among others. Furthermore, SNA is not a linguistic approach, but it provides a synchronous context analysis of users by mapping their interactions. Infusing linguistic information with network topologies make it easier to determine information flow, with the ability to detect opinion leaders, and unique clusters within a community [4].

However, modeling user behaviors from the opinionated data has several difficulties, such as; language is always oriented, and extracting informed semantic content demands cultural and contextual prisms alongside motivations of users. For example, a word like 'hot' has multiple definitions; it may mean heat, attractiveness, or popularity. Coancestry affects both parameters, and while there is a general agreement on its meaning and definition, certain questions, such as how it relates to genetic distance and which of the two parameters is affected more by coancestry, remain ambiguous at present and need further clarification to allow for a correct interpretation of the results. Moreover, getting data involved with virtual communities has scale and heterogeneity complications due to these. These platforms create tremendous amounts of data in various structures including qualitative textual form, multimedia and qualitative images thus demanding complicated procedures and large computing power to combine. For one, other hostile factors such as ethical issues such as privacy and data biases are also relevant. Bias is especially dangerous when it is carried in the datasets to generate models that are bias towards a particular ethnicity, gender, or age thus the importance of fairness in behavioral modeling [5]. The behavioural deterioration in online

discussion forums can be predicted by constructing behavioral sequences from temporal information and analysing n-gram features [10].

The holistic user behavior models can be applied across different fields and have numerous implications. In marketing, these models facilitate recommendation systems to provide more relevant recommendation to the viewers, thus improving their interactions. The concept helps the organizations to capture and know the users and the trends favored in a specific region or community. These models are used by social media platforms for moderating the content, for stopping spam, fake news, or for regulating the interaction. Looking at it strategically; even issues to do with urban planning can be informed by the contextual insights drawn from the online communities to make project well suited to the populace's sentiments [6].

B. Problem Statement

Specifically, the problem of positive and negative statements in social networks has a great impact on society, primarily in multicultural countries like Pakistan. However, social networking sites enable users to disseminate malicious contents since they are fake accounts and the population using the social media is large. This environment not only aggravates polarisation, but is also a threat to the social stability of society. To solve this problem there has to be measures that can be used to identify and prevent the act of positive and negative statements in real-time. However, there is still an insufficient number of culturally and linguistically appropriate detection measures that serve as a distinct disadvantage. Present day positive and negative statements detection algorithms are mostly developed from datasets derived from Western environments, which fail to capture non-western languages and expressions. Languages like Urdu and English are commonly used with code switching in social media in a country like Pakistan, but such systems fall short [9]. The use of code-mixed language brings into play factors like unconventional syntax, spelling irregularities, and semantics making it hard for machine learning models to handle positive and negative statements. Moreover, people's culture and their ways of interaction influence the definition of positive and negative statements, which aggravates the problem for automatic detection by using general approaches. Therefore, positive and negative statements remain prevalent in social media platforms which exposes the respective groups to more positive and negative statements. These challenges are addressed in this research by developing computational linguistic models of the sort needed for Pakistani Urdu comparing key performance metrics between neural and non-neural models. In this respect, the study is expected to develop culturally intelligent systems using sentiment analysis, text mining, and social network analysis to process the Urdu-English code-mixing data. The aim is to identify the impact of affective semantic resources in determining specific manifestations of positive and negative statements in Urdu-English code-mixed tweets in Pakistani context.

C. Objective

To assess the capability of computational linguistic models in recognizing specific forms of positive and negative statements in Roman Urdu data by developing a tailored corpus and leveraging advanced natural language processing techniques.

The rest of study is organized as follows: Section-II presents the related work, discussing previous studies and approaches relevant to our research. Section-III describes the proposed methodology, including the system architecture and the key algorithms used. Section-IV provides the results and discussion, highlighting the performance and implications of our findings. Section-V explains the significance and application. Section-VI concludes the study and outlines potential directions for future work.

II. Related Work

Identification of positive and negative statements has become a popular research domain in computational linguistics, especially in terms of multilingual and code-mixed data sets. Even though there are numerous studies trying to identify positive and negative statements, especially in languages such as English, the detection of positive and negative statements in Roman Urdu has not been explored enough. The identification of influential users in social network can be done with sentiment analysis relating sentiment with influence metrics [8]. This section presents a brief background on the different literatures available on positive and negative statements detection in terms of general approach, sentiment analysis, and the problems of working with code-mixed data especially in the Pakistani context. The generic sentiment analysis work flow is shown in Fig. 1:

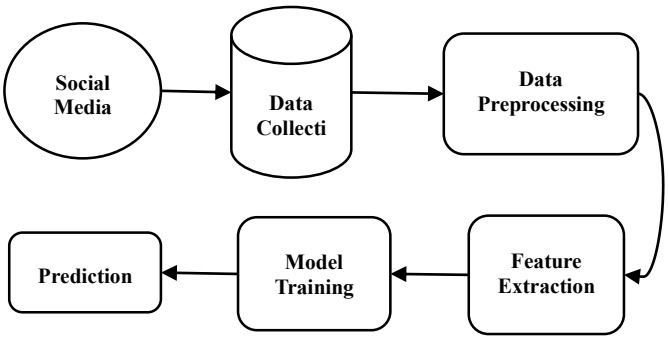

Fig. 1. Workflow of sentiment analysis.

A. Positive and Negative Statements Detection in Social Media

The spread of social media sites such as the Twitter, Facebook, or You-tube makes it easier for people to spread positive and negative statements, a challenge to computer aided recognition models. It is possible to use ML and DL to design models that can detect positive and negative statements in languages of interest [11] [20]. Also, in the early works, researchers deliberately used rule-based or a lexicon-based approach. However, with the recent studies they have incorporated various Natural Language Processing (NLP) algorithms like BERT and the Long Short-Term Memory (LSTM) networks [14]. But then, these are built on databases of words that are mostly English in most cases and this makes them unfit to operate on code-mixed languages like Roman Urdu.

B. Computational Approaches to Positive and Negative Statements Detection

There are a number of studies focusing on positive and negative statements classification that used various approaches: from classical machine learning methods, such as SVM, Random Forest (Davidson et al., 2017) to deep learning models like CNN and transformers including BERT [15]. Transformer models have emerged as highly effective models because of their high capacity for capturing contextual information in natural language. For example, classifiers based on BERT are one of the most accurate when detecting types of hatred speech at the moment [17]. However, these models are sensitive to domain specific information, therefore it becomes essential that there be the creation of corpora rich in linguistic as well as cultural characteristics of the language in question.

C. Challenges in Code-Mixed Data Processing

Interference has become common among the multilingual society, where people interchange two or several languages within a single conversation. Due to the variations in spelling, distinctive transliteration and absence of grammatical rules in the use of English words and phrases, the analysis of Urdu-English code-mixed (UECM) text is more complex in nature [2]. Also the code mixed data is not in any standard language hence does not exhibit the traditional linguistic characteristics that are expected in text data, therefore making it hard for the NLP Models to derive the features correctly. Furthermore, in code-mixed text processing, researchers have used word embedding and the character level features to work on text classification task [13]. However, due to the lack of annotated data for Roman Urdu, the respective positive and negative statements detection models are not yet significantly developed.

D. Sentiment Analysis and Annotation Techniques

Sentiment analysis is important in positive and negative statements identification because it categorizes messages as positive and negative statements, neutral or non-positive and negative statements. Traditional text mining techniques like lexicon-based techniques and polarity identification have also been adopted by researchers for text classification purposes [13]. Other advanced approaches employ deep learning approaches to learn the variation of sentiment in text information. Another important step is the procedure of annotation of the positive and negative statements data set as this allows to have good quality of the training sets. Other being TextBlob and machine learning equally for annotating sentiments and toxicity levels in the text [19]. To overcome the drawbacks of using only rule of thumb specific features for annotation, the proposed approach integrates pre-trained transformer models like BERT which helps in understanding the contextual semantics in case of code-mixed data.

E. Positive and Negative Statements Detection in the Pakistani Context

This situation has raised a question on the level of sensitization of Pakistan towards hate speech as there has been little conducted on the identification of abuse in the local language. Roman Urdu is the dominant type of text messaging in Pakistan; people switch between Urdu and English in social media. For the same reason, the availability of large-scale positive and negative statements datasets in Roman Urdu is still a problem for computational linguistics [12]. Some works that have been done in this regard are just limited and curate small datasets and they use basic ML-based classifiers to identify the objectionable language [13]. Nevertheless, these research studies are inconclusive in the formation of a framework for automated identification of positive and negative statements in Pakistani virtual communities.

F. Role of Transformer Models in Positive and Negative Statements Detection

The newer ways of developing NLP models are based on transformer structures like BERT and its brothers, and they stated to outperform almost all the text classification tasks including positive and negative statements detection. Scholars have proposed utilization of positive and negative statements datasets for adaptation of the transformer models optimization in both multilingual as well as code-mixed languages [18]. The above analysis of the BERT model in the Roman Urdu, specifically fine-tuning the BERT on customized datasets has registered great results in the detection of polite and abusive words [7]. However, these models should be subjected to further research in order to execute them and make them more efficient and effective in real-world scenarios given the dynamicity of positive and negative statements on social media.

III. Methodology

A. Research Design

This study utilized primary research method with the technique of web scraping as opposed to a review of literature. The purpose of the study is to prepare positive and negative statements corpus in Roman Urdu obtained from the tweets written by Pakistani tweeters. It is mainly confined to opinionated language in Roman Urdu which may occasionally have syntactical English and Urdu texts due to Pakistan's code-switching culture in social media. To achieve the purpose for the study, i.e. to recognize positive and negative statements in Pakistani language Roman Urdu, a custom dataset was used to handle Natural Language Processing and Computational Linguistic techniques appropriate to the socio-linguistic perspective of the virtual communities in Pakistan. In analyzing the sentiment of the text, the study employs TextBlob to check for polarity while BERT (Bidirectional Encoder Representations from Transformers) are used for deep contextual analysis and annotation of semantics. This makes sure that the produced model is closely context-adaptive and sensitive to discriminating positive and negative statements patterns prevailing in the Roman Urdu content of social media.

B. Data Collection

For the purpose of assembling the necessary data for this study, web scraping methods are adopted to scrape all the real-time content from Twitter and extract Roman Urdu text only. Since Roman Urdu writing is common by Pakistani users on the social media to report, comment, and share feelings and sentiments, opinionated and sentiments-expressing contents on Twitter are useful for constructing a domain-specific positive and negative statements corpus. The scraping process depends on the given keywords and hashtag specific to the topics that have positive and negative statements within the context of

Pakistani culture and language. The data gathered in the study is mainly in Roman Urdu with some switching between English and Urdu, a common practice in the Pakistani twitter world. Cleaning and normalization measures were taken into consideration to eliminate malignant contents such as advert content, URLs, emoji content, and duplicate content. Unlike secondary research, this study has constructed its data collection system from the ground up rather than drawing on previous databases. The positive and negative statements in the dataset are classified manually by analyzing and annotating the result of applying textBlob sentiment and BERT models that enhances the understanding of positive or negative sentiments as well as the context. This ensures development of robust, genuine and context-based dataset that accurately addresses the mechanism of positive and negative statements detection in Roman Urdu.

C. Data Analysis

After the creation of Roman Urdu dataset through web scraping from Twitter, the collected data was further preprocessed by removing noises from the text such as, URL addresses, emojis, and other non-text information. Subsequently, using TextBlob, the next step was to perform the first level of sentiment classification, separating the entries into positive, negative, or neutral. However, because positive and negative statements in the Roman Urdu language and the contexts in which it prevails is intricate, a second approach using BERT is made in the next research. BERT's neural organization means that it is capable of the idea of the context which is essential when interpreting sarcasm, and other forms of indirect hate messages that are sometimes written in Roman Urdu.

D. Performance Metrics

The parameters applied for examining the performance are as follows:

1) Accuracy (ACC): It is considered the most important parameter in evaluating the model's performance. This metric evaluates the number of samples for correct prediction over the number of all samples. The formula for calculating this metric is given in Eq. (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2) Recall: This parameter refers to the ability of the model to predict positive samples. This is calculated by dividing the number of samples that are categorized as true positive overall positive samples. The formula for calculating this metric is given in Eq. (2).

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

3) Precision: In this parameter, true positive identified the number of samples over several samples that are predicted as positive. Eq. (3) calculates the precision.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4) F1_score: In this parameter, the recall and precision are combined into a single metric. This is called the harmonic mean of recall and precision. Eq. (4) calculates the F1_score.

$$F1\_score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

E. Ethical Considerations

This study complies with high ethical standards, where vulnerable aspects such as data privacy, consent, and the appropriate use of social media content are concerned. Data is gathered using the web scraping technique from Twitter, although this is a social media platform; the identity of users and their confidentiality were kept anonymous. The app will not request or store names, faces, avatars, or locations of the application users. This study is concerned with the detection of positive and negative statements in Roman Urdu text only and no attempts are made to identify a person or a group of people for any unfair treatment. The objective is strictly academic and does not involve the promotion of any sort of prejudice or hatred whatsoever; it is solely for creating a corpus and the proposed model to better understand positive and negative statements in Pakistan. Collection of data is done in accordance to Twitter Developer Policy as well as the guidelines on usage of API and the terms of service. Further, the study also recognizes that positive and negative statements content is sensitive, and the dataset shall be used appropriately without re-emphasizing hatred in order to follow the best practice on ethicality of AI research.

IV. Results

A. Analysis of Model Performance: Neural versus Non-Neural Models

This analysis compares Neural Network-based models and Non-Neural (Traditional Machine Learning) models based on their performance across four key metrics: Test Accuracy, Test Recall, Test Precision, and Test F1 Score.

The graph structure utilizes color coding to categorize the models effectively. The neural models, represented in blue, include Multi-Layer Perceptron (MLP), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bi-GRU. On the other hand, non-neural models, marked in red, consist of traditional machine learning approaches such as Support Vector Machines (SVM) and Logistic Regression. This visual distinction helps in easily identifying and comparing the different model types. Fig. 2 represents four bar charts for test accuracy, test recall, test precision and test F1 scores respectively.

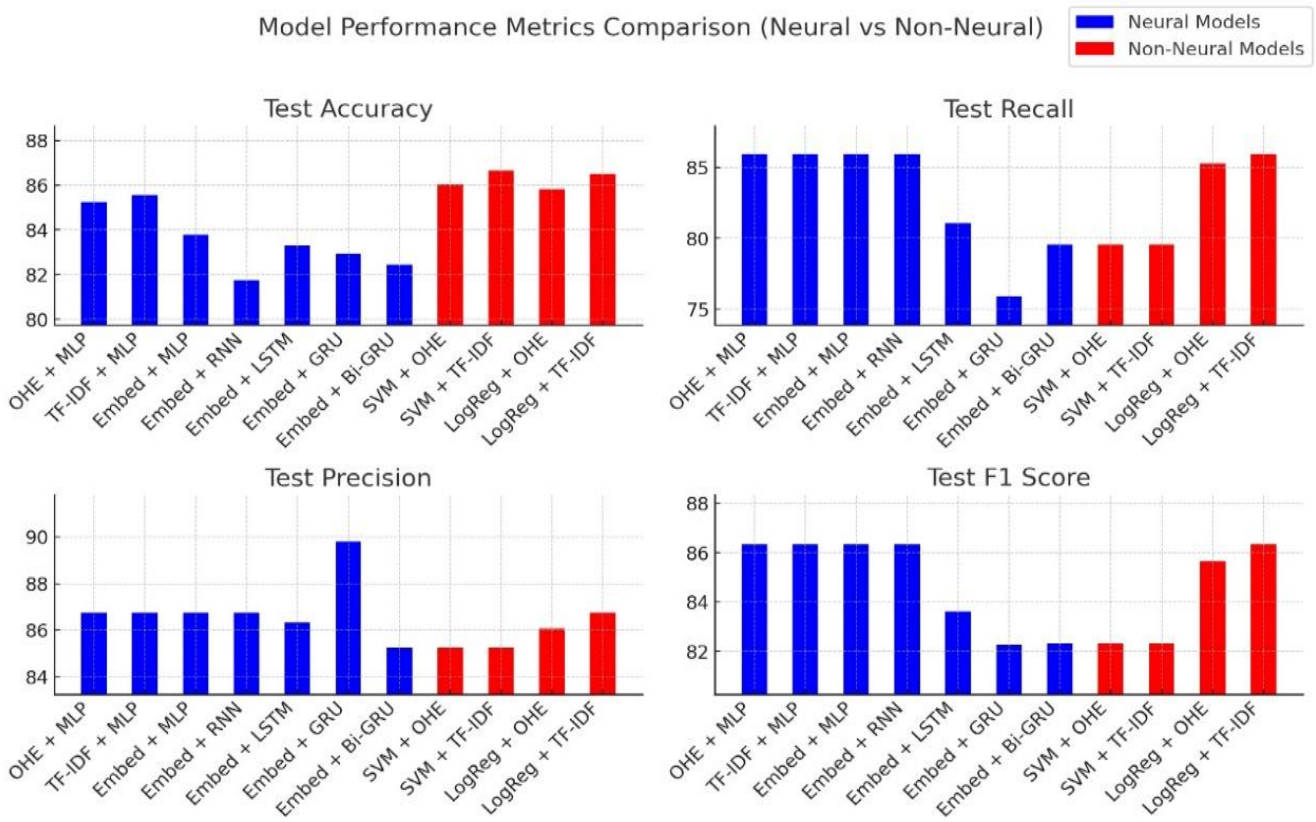

Fig. 2. Model performance metric comparison (Neural vs. Non-Neural).

The test accuracy analysis reveals that non-neural models outperform neural models, with SVM + TF-IDF achieving the highest accuracy at 86.65%, closely followed by Logistic Regression + TF-IDF at 86.49% and SVM + OHE at 86.04%. In contrast, neural models generally exhibit lower accuracy, Embedding + RNN performing the worst at just 81.74%, indicating that traditional machine learning approaches may be better suited for this particular task compared to more complex neural architectures.

The comparison between neural and non-neural models shows a clear dominance of non-neural approaches in terms of accuracy, with SVM and Logistic Regression performing particularly well. These traditional models likely benefit from better generalization when using structured text representations like TF-IDF and one-hot encoding (OHE). In contrast, neural models fail to surpass their non-neural counterparts in accuracy, which may be attributed to factors such as limited training data, suboptimal hyperparameter configurations, or an inherent mismatch between model complexity and task requirements. This suggests that, for this specific application, simpler machine learning methods with carefully engineered features may be more effective than complex neural architectures.

While non-neural models like SVM and Logistic Regression excel in accuracy, neural models demonstrate superior performance in recall, with MLP, RNN, and LSTM achieving the highest rates (up to 85.93%), except for LSTM (81.09%) and GRU (75.90%). In contrast, non-neural models lag behind with an average recall of ~79.57%, indicating they miss more true positive cases. This trade-off suggests that traditional methods prioritize precision and overall reliability at the expense of sensitivity, whereas neural networks are more effective at detecting positive instances—a critical advantage in scenarios, where missing positives is costly. If minimizing false negatives is the priority, neural models emerge as the preferred choice despite their lower accuracy.

In precision performance, the GRU-based neural model stands out with the highest score (89.80%), demonstrating its ability to minimize false positive predictions. Other neural architectures (MLP, RNN, LSTM) follow closely but trail slightly behind (~86.75%), while non-neural models like SVM and Logistic Regression exhibit marginally lower precision (85.26% to 86.08%). However, the GRU's high precision comes at a cost—its recall is notably low (75.90%), suggesting an overly conservative prediction approach that may miss true positives. In contrast, non-neural models strike a more balanced trade-off between precision and recall, making them a robust choice when both false positives and reliable detection are priorities. This highlights a key decision point: if minimizing false alarms is critical, the GRU excels, but if a balanced performance is needed, traditional models like SVM remain competitive.

The F1 score analysis reveals that MLP-based neural models and Logistic Regression (both at 86.34%) deliver the best balance between precision and recall, demonstrating that properly configured deep learning can match the performance of traditional methods. While GRU (82.27%) and Bi-GRU

(82.32%) suffer from lower F1 scores due to their poor recall—despite GRU's high precision—LSTM also underperforms (83.63%) because of its reduced sensitivity.

Notably, Logistic Regression emerges as a particularly strong baseline, achieving comparable results to neural networks while being far more computationally efficient. This suggests that for tasks requiring a harmonious trade-off between precision and recall, simpler models like Logistic Regression or well-tuned MLPs may be preferable over more complex architectures like GRU or LSTM, unless the use case. The results with findings are summarized in Table I.

TABLE I. SUMMARY OF BEST PERFORMING TYPE WITH RESPECT TO SPECIFIC METRIC

Metric	Best Performing Type	Key Finding
Accuracy	Non-Neural (SVM + TF-IDF)	Traditional ML models are superior in accuracy.
Recall	Neural (MLP, RNN)	Neural models are better at identifying positive cases.
Precision	Neural (GRU)	GRU is most precise but has lower recall.
F1 Score	Both (MLP & Logistic Regression)	Both perform equally well in balancing recall and precision.

B. *Neural versus Non-Neural Summary Findings*

The analysis reveals clear trade-offs between neural and non-neural approaches across different performance metrics. For applications requiring high accuracy, non-neural models like SVM with TF-IDF or Logistic Regression emerge as the top choices, delivering superior performance compared to their neural counterparts. These traditional methods not only achieve better accuracy but also offer significant advantages in training speed, with Logistic Regression completing training in just 1 to 2 milliseconds.

When recall is the primary concern, neural models demonstrate clear advantages. MLP and RNN architectures achieve the highest recall rates, making them ideal for scenarios where missing positive cases carries high costs. However, this strength comes with increased computational requirements, as neural networks generally require longer training times - particularly embedding-based models that need sequential processing. The MLP with OHE/TF-IDF configuration stands out as offering a particularly good balance between performance and computational efficiency.

Precision requirements present a different landscape, where the GRU model achieves the highest precision at 89.80%, significantly outperforming other approaches. However, this comes at the cost of substantially lower recall (75.90%), suggesting the model may be too conservative in its predictions.

For most practical applications needing a balance between precision and recall, either MLP-based neural models or Logistic Regression provide the best compromise, with the latter offering the additional benefit of faster training times and lower computational overhead.

The training time analysis as shown in Fig. 3 shows distinct patterns across model types. Non-neural models generally train much faster, with Logistic Regression being exceptionally quick (1 to 2ms), while SVM models show surprisingly long training times (886 to 982ms) for traditional methods. Among neural networks, training durations vary significantly - MLP and Bi-GRU models require substantial time (684 to 801ms), while embedding-based sequential models demand even more resources due to their architectural complexity. These timing differences create important practical considerations when selecting models for deployment, particularly in resource-constrained environments or applications requiring frequent retraining. The exact value of performance metrics along with training time is presented in Table II to classify the basis for opting particular model.

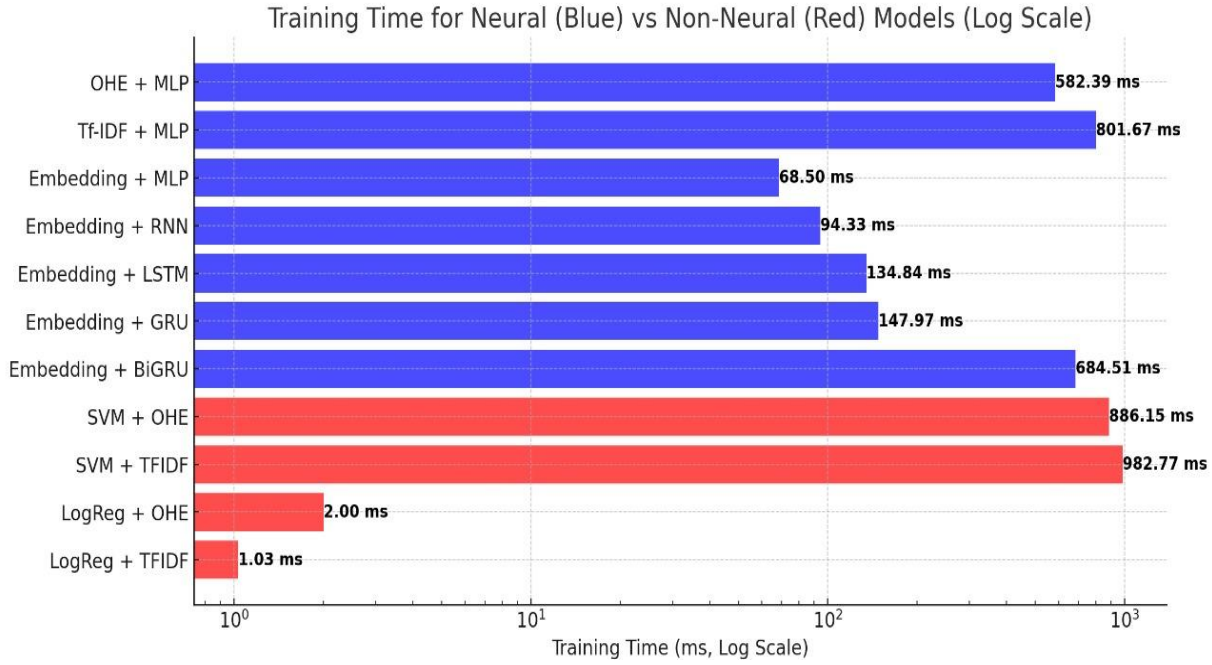

Fig. 3. Training time comparison (Neural vs. Non-Neural Model).

TABLE II. OVERALL PERFORMANCE OF NEURAL AND NON-NEURAL MODEL TYPES WITH RESPECT TO PERFORMANCE METRICS

Model Type	Name	Training Time	Test Accuracy	Test Recall	Test Precision	Test F1 Score
Neural	**OHE + MLP**	582.3880136	**85.25%**	85.93%	86.75%	86.34%
Neural	**Tf-IDF + MLP**	801.6679513	**85.55%**	85.93%	86.75%	86.34%
Neural	**Embedding Layer+ MLP**	68.50320315	**83.76%**	85.93%	86.75%	86.34%
Neural	**Embedding Layer+ RNN**	94.33099894	**81.74%**	85.93%	86.75%	86.34%
Neural	**Embedding Layer+ LSTM**	134.8372431	**83.28%**	81.09%	86.33%	83.63%
Neural	**Embedding+GRU**	147.9710679	**82.93%**	75.90%	89.80%	82.27%
Neural	**Embedding + Bidirectional GRU**	684.5123119	**82.45%**	79.57%	85.26%	82.32%
Non Neural	SVM + OHE	886.1468422	86.04%	79.57%	85.26%	82.32%
Non Neural	SVM + TFIDF	982.7742205	86.65%	79.57%	85.26%	85.32%
Non Neural	Logistic Regression + OHE	2.004743338	85.82%	85.24%	86.08%	86.66%
Non Neural	Logistic Regression + TF-IDF	1.02973169	86.49%	85.93%	86.75%	86.34%

V. DISCUSSION

The use of a computational linguistic approach is a strong foundation for modelling whole use activities via opinionated data in virtual communities. Unlike surveys and other quantitative methodologies, this method entails identification and analysis of text-based interactions, hence it reveals detailed sentiments, preferences and cognitive style [17]. Conversation analysis also becomes easier, thanks to sentiment analysis, you get topic modelling and natural language processing from social media data. These insights provide detailed information concerning user intentions, majority perspective and group behaviors thus providing a better chance of predicting their behaviors. Combining this data with machine learning models improves the choices, users, and interactions. This approach has potential for to use in marketing, content moderation, and community management and will thus foster advances in virtual community analysis [18].

A. *Generalizing Positive and Negative Statements Detection to Code-Mixed Data*

As a result of varying performances, computational linguistic models' ability to generalize knowledge from existing positive and negative statements datasets for the identification of certain categories of positive and negative statements in the code-mixed Urdu-English text is questionable. They show that the proposed models are capable of handling general positive and negative statements patterns, but determining their accuracy in handling code-mixed data is problematic due to linguistic and contextual peculiarities [14]. This was specifically found with emergent Bilinguals using mixed code in writing, where there is poor spelling and unconventional grammar and cultural references which are not characteristic of other datasets. Generally, the other available positive and negative statements datasets do not have the required linguistic variation in code-mixed texts which serve to reduce the overall effectiveness when used directly.

However, pre-trained language models such as BERT or multilingual BERT hold promising results in closing this gap by taking advantage of contextual embeddings; though, their results rely on further fine-tuning with domain data [17]. The inclusion of more labeled data related to Urdu-English code-mixed positive and negative statements improves the performance of the model in detecting context sensitive slurs, and indirectly downloaded bias and culture specific hate expressions. Furthermore, through the help of transfer learning and adding new augmented code-mixed data to the dataset increases generalization. However, there is no clear-cut benchmarks set for the evaluation of the models concerning code-mixed positive and negative statements recognition. Altogether, despite starting with computational linguistic models, certain modifications are required for effectively handling the complex dynamics of Urdu-English code-mixed positive and negative statements [18].

B. *Affective Semantics in Urdu-English Positive and Negative Statements*

Bibliosocial connotative support is helpful for recognizing the cultural and contextual portrayals of positive and negative statements in Urdu-English mixed data regarding others. In Pakistan's context, where Urdu and English are combined in many instances, in written forms and the context is multicultural and multilingual, the emotions, sentiments and socio-cultural aspects therefore must be considered for identifying positive and negative statements [18]. Resources such as sentiment lexicons, the models for emotion detection, and culturally sensitive databases assist basic computational mechanisms in distinguishing between profanity, and other words and phrases that might be harmless in other contexts. Another reason why positive and negative statements detection becomes a challenge is due to the use of Urdu-English code-mixing or a mixed language-Code mixing. For example, an Urdu sentence can be followed by an English phrase changing the message or tenor of the identical sentence according to general cultural disparities which are not seen from mere differences in the word structure [15]. Emotional semantic resources help to understand these mixed statements since they allocate emotional connotations to words or expressions. Urdu-English code-mixed data prove useful in the identification of cultural or emotional connotations that are typically omitted in the linguistic analysis. For instance, words such as 'kafir' (infidel) in Urdu or 'traitor' in English may have intense semantic prosodies hence used in specific political or religious discourses. An affective semantic resource that can detect such

emotional signals in order to differentiate between positive and negative statements and free speaking is used [19].

VI. Conclusion

Virtual communities have become new forms of socialization and an expression of coming up with opinions or even creating ideologies. These platforms capture immense volumes of user generated content in terms of behaviors, preferences and sentiments. To model such behaviors in an integrated fashion, complex computational linguistic techniques are needed for opinionated data. These data are generally in formats like posts, comments, reviews and discussions and helps in deriving the users' emotion, intention, social context etc. Nevertheless, this data is by nature, noisy, unstructured and context specific which presents many difficulties for further analysis. In order to effectively overcome these issues computational linguistics makes use of methodologies such as natural language processing (NLP), machine learning and semantic analysis. These methodologies assist in converting the plain, unformatted text to structured formats, where guidelines, behaviors, preferences and trends can be perceived.

Another research dimension that deserves special attention is the coupling of the topic modeling with the feature engineering stage. The techniques include Latent Dirichlet Allocation (LDA), and other methods, including BERTopic for neural-based topic modeling and for understanding major concepts or topics of discussion in a particular community. This is useful for giving a characteristic, or otherwise random nature of textual data, a semi-unified theme. Furthermore, syntactic analysis of the language in terms of Word Levenshtein distance, POS-tagging, and discourse analysis enhances the ability to identify the intent and structure of the users' communication profile. These features are useful in the process of sorting users by different parameters – formal or informal or persuasive language etc.

Besides textual features, as components of computational linguistic, contextual and social network analytical features are also used to model the user behaviors fully. Integrating text further with communication activities like likes, shares, and replies to activities enhances behavioral modeling through peer impacts and the network topology.

From the above analysis, the following should be recommended in order to model multichannel user behaviors in virtual communities. First, deep NLP methods, including transformers like BERT or GPT, should be employed to address sophisticated language patterns like sarcasm, idioms, cultural references, and so on in order to achieve more precise sentiment and topic analysis. This is particularly useful when used with images, videos, and metadata as it will be in combination with the textual analysis of the user behavior. Models should also address further temporal, situational, cultural views which enhance the accuracy of estimations and the relevance of discovered knowledge. Thus, it is crucial to implement ethical practices such as user privacy, security, and absence of bias within computational systems to gain people's trust and credibility. In addition, incorporating the social network analysis within the model might bring information on likes, comments, and shares to help identify the peer influence, and network effects while also identifying opinion leaders and future trends. Interdisciplinary combination of linguistic approaches, data analysis and computational methods on one hand and knowledge of psychological and sociological factors on the other are paramount in building rich models. Last of all, the continual training of these models against real-world situations provides for the accuracy and versatility of the models in the context of rapidly changing dynamic virtual communities. These approaches will result in sound, acceptable, and meaningful user behavior modeling.

References

[1] W. Chung and D. Zeng, "Dissecting emotion and user influence in social media communities: An interaction modeling approach," *Information & Management*, vol. 57, no. 1, pp. 103108, 2020.

[2] E. Purificato, L. Boratto, and E. W. De Luca, "User Modeling and User Profiling: A Comprehensive Survey," *arXiv preprint arXiv:2402.09660*, 2024.

[3] K. Kasianenko, S. Khanehzar, S. Wan, E. Dehghan, and A. Bruns, "Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20106–20135, Nov. 2024.

[4] K. F. Kahl, T. Buz, R. Biswas, and G. De Melo, "LLMs Cannot (Yet) Match the Specificity and Simplicity of Online Communities in Long Form Question Answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2028–2053, Nov. 2024.

[5] N. Borenstein, A. Arora, L. A. Kaffee, and I. Augenstein, "Investigating Human Values in Online Communities," *arXiv preprint arXiv:2402.14177*, 2024.

[6] H. Li and R. Zhang, "Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots," *Journal of Computer-Mediated Communication*, vol. 29, no. 5, p. zmae015, 2024.

[7] S. Biswas and G. Poornalatha, "Opinion Mining Using Multi-Dimensional Analysis," *IEEE Access*, vol. 11, pp. 25906–25916, 2023.

[8] S. Al-Otaibi, A. A. Al-Rasheed, B. AlHazza, H. A. Khan, G. AlShfloot, M. AlFaris, et al., "Finding influential users in social networking using sentiment analysis," *Informatica*, vol. 46, no. 5, 2022.

[9] M. Heidari and J. H. Jones, "Using BERT to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0542–0547, Oct. 2020.

[10] J. M. Tshimula, B. Chikhaoui, and S. Wang, "On predicting behavioral deterioration in online discussion forums," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 190–195, Dec. 2020.

[11] N. Borenstein, A. Arora, L. A. Kaffee, and I. Augenstein, "Investigating Human Values in Online Communities," *arXiv preprint arXiv:2402.14177*, 2024.

[12] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, "Trends in combating fake news on social media–a survey," Journal of Information and Telecommunication, vol. 5, no. 2, pp. 247–266, 2021.

[13] C. Messaoudi, Z. Guessoum, and L. Ben Romdhane, "Opinion mining in online social media: a survey," Social Network Analysis and Mining, vol. 12, no. 1, p. 25, 2022.

[14] S. Masud, M. Bedi, M. A. Khan, M. S. Akhtar, and T. Chakraborty, "Proactively reducing the hate intensity of online posts via hate speech normalization," in Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, 2022, pp. 3524–3534.

[15] L. Bharadwaj, "Sentiment analysis in online product reviews: mining customer opinions for sentiment classification," International Journal of Multidisciplinary Research, vol. 5, no. 5, 2023.

[16] S. Shrestha, I. Bittencourt, A. S. Varde, and P. Lal, "AI-based modeling for textual data on solar policies in smart energy applications," in Proc. 15th Int. Conf. Information, Intelligence, Systems & Applications (IISA), 2024, pp. 1–8.

[17] H. Chen et al., "Socialbench: Sociality evaluation of role-playing conversational agents," in Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 2108–2126.

[18] C. Saravanos and A. Kanavos, "Forecasting stock market volatility using social media sentiment analysis," Neural Computing and Applications, pp. 1–24, 2024.

[19] S. Li, F. Liu, Y. Zhang, B. Zhu, H. Zhu, and Z. Yu, "Text mining of user-generated content (UGC) for business applications in e-commerce: A systematic review," Mathematics, vol. 10, no. 19, p. 3554, 2022.

[20] M. Ranjan, S. Tiwari, A. M. Sattar, and N. S. Tatkar, "A new approach for carrying out sentiment analysis of social media comments using natural language processing," Engineering Proceedings, vol. 59, no. 1, p. 181, 2024.