Uncovering TikTok's Top Trends: An NLP-based Analysis of User-generated Content

Aakar Dhakal
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
aakar.dhakal@g.bracu.ac.bd

S M Rakib Hasan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sm.rakib.hasan@g.bracu.ac.bd

Maisha Noor
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
maisha.noor@g.bracu.ac.bd

Md Mustakin Alam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—With over a billion active users, TikTok has emerged as one of the most widely used social media networks globally. However, substantial study on the platform's user-generated material is lacking. We show an analysis of the best patterns on TikTok using natural language processing (NLP) in this paper. Using state-of-the-art NLP methods, we collected and analyzed a considerable amount of user-generated material, including text, images, and videos. Our research reveals the subjects, sentiments, and activities that TikTok users communicate the most frequently. We also note the most typical keywords and grammatical constructions used in the top trends on the site. This study adds to the increasing amount of knowledge about social media platforms and offers insightful information about the topics and developments on one of the most widely used platforms worldwide.**

***Index Terms*—User-Generated Content, Sentiment Analysis**

I. Introduction

On social media, forums, blogs, e-commerce, and on other digital platforms, there is a vast amount of data accessible, with a sizable portion of it being user-generated content. As technology advances and the number of people who are comfortable using the Internet rises, there is an increasing quantity of user-generated content available online. User-generated content, or UGC, is produced by individuals based on their unique perspectives and experiences. This kind of content is developing into a valuable economic resource with growing commercial possibilities. On a variety of topics, including e-commerce, marketing, the healthcare system, the transportation system, brand research, customer service, and many other areas, the UGC offers the viewpoints of a large number of people.

Prospective customers increasingly depend on user-generated content (UGC) as e-word-of-mouth information to assess products and services. Managers are quickly realizing the necessity of using this data to gauge customer opinion and assess the success of their own and rival companies' goods. It improves knowledge of customer expectations and behavior while also spotting market opportunities. To identify whether a product or service is receiving favorable or negative feedback, it is essential to classify sentiment from customer reviews. It can be useful in locating significant problems with the goods or services. Meanwhile, NLP or Natural Language Processing is being used in social media to analyze user generated contents and derive insights in terms of marketing, consumer behavior, and sentiment analysis. Also, much research and analysis has been done in Facebook and Twitter but in the case of TikTok, there is still a huge scope for such research. With over one billion daily active users, TikTok is one of the most widely used social media apps thanks to its trending challenges, catchy music, and short-form videos. Therefore, TikTok's widespread use has made it an outstanding source of information for researchers to examine user behavior and trends through its user generated content. This paper seeks to outline the current state of UGC sentiment analysis research as well as trends in TikTok using NLP. To our understanding, UGC sentiment analysis has not been subjected to the content of TikTok to analyze trends. We will therefore provide an analysis of UGC sentiment analysis in this research by extracting information from TikTok contents.

II. Previous Works

User-generated content (UGC) is written by individuals and is based on their views and experiences [1]. Typically on the Internet, this content is produced by regular people who willingly contribute information, data, or media which is later presented to others in a useful or humorous way. Wikis, videos, and eatery reviews are a few UGS examples (Krumm et al., 2008). On a variety of topics, such as e-commerce, the healthcare system, marketing, customer service, transportation, brand study, and many others, the UGC requests community input. By sharing people's views about a particular entity or product, this information is developing into an increasingly significant economic resource that contains significant commercial potential (Berthon et al., 2015; Krumm et al., 2008; Putra Suryasari, 2021). However, the excess of reviews leads to becoming difficult for both management and potential customers to gather pertinent information. Consequently, there is a growing requirement for sentiment analysis studies using NLP models (Agarwal Mittal, 2013; D. Li et al., 2022). Sentiment analysis can be described by the computational study of evaluating people's attitudes toward a specific thing. It establishes the tone of an expression's feelings and can be used to understand a user's opinions, attitudes, and emotions. Automatic classification of the polarity of positive, negative, or neutral views is done using sentiment analysis techniques (Medhat et al., 2014; Priyadarshini Cotton, 2021; Serna et al., 2021). Sentiment analysis is used in many industries, including e-commerce, smart tourism, health care, company reputation management systems, and even to gauge public opinion regarding the COVID-19 pandemic. NLP (Natural Language Processing) is concerned with building systems that can understand and respond to text or speech as humans do, NLP makes use of computational linguistics combined with statisticals, machine learning, and deep learning models. NLP is being used in social media to analyze user-generated contents and derive insights in terms of marketing, consumer behaviour, and sentiment analysis.

[2]Consumer comment data from a large number of sources are used for latent semantic analysis in UGC information extraction. Customers' preferences can be derived from UGC information mining because customers communicate their desires. Based on feedback on the prior phone model and user perspectives, businesses can decide the contents and the direction of phone development. Through the subjective expression of emotions, Riaz et al. (2019) quantified emotional information and clearly understood users' tastes and behaviors. The objective of a win-win situation for customers and businesses is accomplished by improving the trade-off between warranty costs and customer happiness, paying close attention to high-level feedback, and lowering warranty costs. Sentiment analysis in particular areas is used to quantify consumer attitudes in the automotive industry, and supervised learning is used to find rivals with user-generated content. The primary task in determining a product's competitiveness, whether from the perspective of users or businesses, is to organize the text first before using structured vectors to use an optimization algorithm to find the best solution to the issue. Ramaswamy and DeClerck's (2018) recommendations for NLP technology to investigate how to comprehend customer views in the context of the text structuring issue demonstrate that further research will need to create a comprehensive corpus of pertinent text and employ deep-learning CNN models. The language model is one of the methods that convert language-based unstructured data into structured data. Later, it was revealed that Word2vec, Transformer, ELMO, and other pre-training models exist, but they can only execute one-way semantic analysis and have narrow application domains. The BERT model is used to generate the fundamental model of word vectors and sentence vectors, which is then used to conduct semantic extraction and analysis as well as analyze sentences having bidirectional semantics in user reviews. To discover the best solution, optimization algorithms have been extensively applied in many different fields. Aquila Optimizer technique was suggested by Abualigah et al. (2021b) based on natural prey-catching behaviors. To make it challenging to fall into local solutions during the optimization process, but challenging to understand the structure of the Bayesian network, Abualigah et al. (2022) proposed the Reptile Search Algorithm.

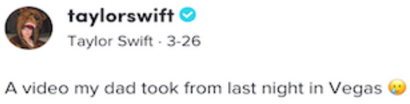
Another study [3] analyzed content from mentalhealth on TikTok. The authors analyzed 100 videos with more than 1 billion views in total. Additionally, the comments to each video were viewed and coded for content in the following themes: offering support or validation; mentioning experience with suicide or suicidal ideation; mentioning experience with self-harm; describing an experience with hospitalization for mental health issues; describing other mental health issues; and sharing coping strategies, experiences of healing, or ways to feel better. The only content category observed in most (51/100, 51 percent) of the videos included in the sample was "general mental health." The remaining content categories appeared in less than 50 percent of the sample. In total, 32 percent (32/100) of the videos sampled received more than the overall average number of likes (ie, more that 2.67 million likes). Among these 32 videos, 23 (72 percent) included comments offering support or validation and 20 (62 percent) included comments that described other mental health issues or struggles. The authors from another study [4] presented a method for extracting age-related stereotypes from English language Twitter data, generating a corpus of 300,000 over-generalizations about four contemporary generations (baby boomers, generation X, millennials, and generation Z), as well as "old" and "young" people more generally. They employed word-association metrics, semi-supervised topic modelling, and density-based clustering, and uncovered many common stereotypes as reported in the media and in the psychological literature, as well as some more novel findings. The author analyzed age-related data across six common topics expected to occur for all age groups: family and friends, finance, work, politics, technology, and health. For example, some of the most frequently mentioned over-generalizations for the four-generation groups in the friends and family category is, boomers are considered as terrible parents, Gen-X is also terrible parents but is better than boomers, millennials are not having more children and they give their children weird names, and Gen-Z is going to be best parents. However, this study was only limited to age groups and the authors did not consider economical, social, and ethnic categories. One study suggests [5] that persuasive messages on social media trigger users' persuasion knowledge if a brand is marked as their source. If, however, a user posts brand-related content, this can have persuasive effects without creating awareness of the persuasive potential. As it is only fair to the consumer to disclose covert advertising practices, this study argued for stricter regulation of campaigns aiming at a broad user participation. For instance, by also including ad disclosures on user-generated posts created in this context. As such, posts containing a hashtag or caption promoted by a brand could be marked as commercial content. This study highlights that user-generated content has a more persuasive effect on marketing than direct marketing by the companies.

The work [6] introduces TweetNLP, a platform for natural language processing (NLP) in social media that has its own Python module. It enables a range of NLP operations, such as sentiment analysis, named entity recognition, emoji prediction, and offensive language identification. The platform runs without the need for specialist hardware or cloud services and is driven by fairly large Transformer-based language models that are focused on social media text. The article describes the various language models used in TweetNLP, including the early pre-training on RoBERTa and XLM-R checkpoints and ongoing pre-training on Twitter-specific corpora. The paper[7] discusses the Twitter Vigilance platform is a tool developed by the DISIT Lab at the University of Florence for collecting and analyzing Twitter data for research purposes. It consists of three modules: TV, RTTV, and TVSolr, which allow for data collection, real-time analysis, and advanced search capabilities. The platform has been validated against a dataset of 270 million tweets and has been used in various case studies, such as monitoring city services, critical events, and user behaviour, among others. It has been adopted by over 30 users, mainly local public administrations and research groups, and contains over 100 channels for collecting information on different topics and trends.

III. Tik-Tok's Data

Since Tik-Tok does not shares it data publicly, it was challenging to get data from Tik-Tok. We have to use web scraping tools in Python to get data from Tik-Tok. We specifically targeted Tik-Tok videos with significant number of views, 1 million views and above. From those videos we collected video description i.e. caption and comments. We also included emoji and emoticon.

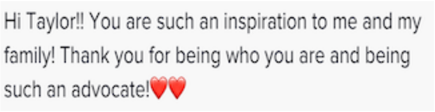
We aim to conduct Sentiment Analysis and Topic Detection and Classification on these data sets. We labelled the data for Sentiment Analysis into these the categories: Positive, Neutral, and Negative. Similarly, for Topic detection and classification we labelled the data as: Arts and Culture, Science and Technology, Politics, and Sports and Entertainment.



taylorswift
Taylor Swift · 3-26

A video my dad took from last night in Vegas 🥲

Fig. 1. Video Description.



Hi Taylor!! You are such an inspiration to me and my family! Thank you for being who you are and being such an advocate!❤️❤️

Fig. 2. Comment on a Tik-Tok Post.

IV. Language models

We will be using Transformer-based models because they well-suited for sentiment analysis and topic detection and classification tasks in natural language processing due to their ability to handle long-range dependencies and capture context effectively.

In sentiment analysis, transformer-based models can analyze the sentiment of a piece of text by processing the entire sequence of words as a whole, taking into account the relationship between words and their positions in the sequence. This is particularly useful in cases where the sentiment of a text is dependent on the context in which it appears.

Similarly, in topic detection and classification, transformer-based models can analyze the entire sequence of words and capture the semantic relationships between them. This allows the models to identify topics and classify them accurately, even when the text contains complex language structures and nuances.

V. Experimental Setup

Sentiment analysis and topic detection and classification using transformer-based models involve the following general steps:

Data Preprocessing: The text data is preprocessed by tokenizing it into individual words or subwords, converting the text into numerical representations, and splitting the data into training, validation, and testing sets.

Model Training: The transformer-based model is trained using the training set of data. During training, the model learns to identify patterns and relationships in the text data that correspond to sentiment or topic.

Fine-tuning: After the initial training, the model is fine-tuned using the validation set to optimize its performance. This involves adjusting hyperparameters and tweaking the model architecture.

Testing: Finally, the performance of the model is evaluated using the testing set of data. This is done by measuring metrics such as accuracy, precision, recall, and F1 score.

In the case of sentiment analysis, the transformer-based model can be trained to classify text as positive, negative, or neutral. The model processes the entire sequence of words, taking into account the context and relationship between words, to identify sentiment.

For topic detection and classification, the transformer-based model can be trained to identify and classify topics based on the content of the text. The model learns to identify common patterns and relationships in the text data that correspond to specific topics.

Overall, transformer-based models excel at these tasks due to their ability to capture complex relationships between words and their positions in the sequence, allowing them to identify sentiment and topics with high accuracy and efficiency.

VI. Results and Discussion

VII. Conclusion and Future Scope

In conclusion, our NLP-based analysis of user-generated material on TikTok has shed light on the predominant patterns and topics on the site. From our research, we were able to identify the most popular hashtags, challenges, and music genres, offering insight into the material that TikTok users respond to the most. Our findings contribute to the comprehension of user behaviour and preferences on TikTok and may be useful for content providers and marketers seeking to increase engagement on the site. In addition, given our analysis reveals the dearth of substantial research on TikTok data, it paves the way for more investigation of this prominent social media network. Because of the dynamic nature of TikTok's user-generated content, various difficulties were encountered during data collecting and preprocessing. In addition, because TikTok is a relatively new site, there is currently a dearth of studies on assessing its material, making it difficult to compare the results to the existing body of research. Future studies should investigate additional factors driving the popularity of TikTok trends, such as user demographics and social interactions, and expand the analysis to encompass other languages and countries. Investigating the influence of TikTok trends on society and culture, such as their impact on fashion, music, and social standards, is a promising alternative avenue. This study lays the groundwork for future research aimed at comprehending the birth and spread of viral material on TikTok.

References

[1] M. Li, G. Zhang, L.-T. Zhao, and T. Song, “Extracting product competitiveness through user-generated content: A hybrid probabilistic inference model,” Journal of King Saud University - Computer and Information Sciences, Mar. 2022, doi: 10.1016/j.jksuci.2022.03.018.

[2] N. Afriliana, N. S. Iswari and Suryasari, “Sentiment Analysis of User-Generated Content: A Bibliometric Analysis,” Journal of System and Management Sciences, Dec. 2022, doi:10.33168/JSMS.2022.0634.

[3] C. H. Basch, L. Donelle, J. Fera, and C. Jaime, “Deconstructing TikTok Videos on Mental Health: Cross-sectional, Descriptive Content Analysis,” JMIR Formative Research, vol. 6, no. 5, p. e38340, May 2022, doi: 10.2196/38340.

[4] K. C. Fraser, “Extracting Age-Related Stereotypes from Social Media Texts,” ACL Anthology, Jun. 01, 2022. https://aclanthology.org/2022.lrec-1.341/

[5] M. Mayrhofer, J. Matthes, S. Einwiller, and B. Naderer, “User-generated content presenting brands on social media increases young adults' purchase intention,” International Journal of Advertising, vol. 39, no. 1, pp. 166–186, Jan. 2020, doi: 10.1080/02650487.2019.1596447.

[6] “Automated Emerging Cyber Threat Identification and Profiling Based on Natural Language Processing,” IEEE Journals Magazine — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10077593

[7] J. Camacho-Collados, “TweetNLP: Cutting-Edge Natural Language Processing for Social Media,” arXiv.org, Jun. 29, 2022. https://arxiv.org/abs/2206.14774