DOI https://doi.org/10.30525/978-9934-26-597-6-14

DECODING SENTIMENTS ONLINE: TRANSFORMER-BASED ANALYSIS OF X CONTENT

Antony Jimson Poruthukkaran, Amit Joshi
ISMA University, Valērijas Seiles iela 1-korpuss 6, Rīga, LV-1019
e-mail: antonyjporuthukkaran33290@gmail.com

Abstract

Social media platforms like X produce a flood of posts daily, making sentiment analysis–classifying text as positive, negative, or neutral-valuable yet challenging due to volume and informal language. This article from a bachelor thesis explores a beginner's effort to build an AI tool using the DistilBERT transformer to decode sentiments in X content, targeting 80% accuracy. Leveraging Sentiment140 data and 100 X posts collected via Tweepy, the tool preprocesses text and outputs sentiment labels, tested on a standard laptop. Results show 82% accuracy, surpassing more straightforward methods, though sarcasm poses issues, aligning with challenges noted in sentiment research (Medhat et al., 2021). This work offers a practical prototype for brand monitoring or trend analysis, proving that a novice can harness AI with free resources. It underscores transformers' potential and limitations in social media analysis, highlighting the complexity and nuance of sentiment analysis.

Keywords: Sentiment Analysis, Transformer Models, DistilBERT, X API, Multilingual Sentiment, Social Media, VADER, Sentiment140

1 Introduction

X posts pour out opinions every second, but figuring out if they are positive or negative is strict with so much slang and sarcasm. My article shows how I used AI to build a simple tool that decodes sentiments on X, learning as a beginner. Every day, X users share over 500,000 posts–thoughts on movies, tech, or trends (Statista, 2023). Businesses want to know if customers love their stuff, and researchers want to track how people feel about significant events. However, there is a problem: there's too much to read, and tools out there often miss the mark on short, messy posts like "This update rocks!" or "Yeah, great job… NOT." A review by Medhat et al. (2021) notes that sentiment analysis struggles with these informal texts, which hooked me to this project.

As a novice in coding, I was drawn to this project because of its practicality and relevance. X, a significant part of my daily life, was the driving force behind my interest in this project. I found that with free tools like Python and accessible datasets, I could make a meaningful contribution without needing a supercomputer (Wankhade et al., 2022). Research by Ravi and Ravi (2020) describes X as a "goldmine" for sentiment data, which motivated me to take on this challenge. I aimed to create an AI tool that could achieve 80% accuracy using a transformer called DistilBERT, which I could customize like a language expert (Birjali et al., 2023).

2 Approach & Implementation

I planned to use transformers–AI models that understand language–because they are powerful and come pre-trained, perfect for beginners. I chose DistilBERT, a smaller version of BERT so that it would run on my basic laptop with 8GB RAM (Wankhade et al., 2022). For data, I grabbed Sentiment140, a dataset with 1.6 million labeled tweets, and collected 100 X posts using Tweepy, searching keywords like "movie" or "tech." I cleaned the text–removing URLs and emojis–then fine-tuned DistilBERT to guess positive, negative, or neutral sentiments. Birjali et al. (2023) say transformers are top-notch for social media text, which guided my choice.

The tool was implemented in Python, initially using Tweepy to fetch X posts, but switched to Twarc due to Tweepy's limited v2 API support for academic access. Due to X API Free tier rate limits (100 posts/month), Using the tool, I fetched 100 X posts with Twarc and supplemented them with 400 Sentiment140 tweets, totaling 500 posts.

The fun part was building a tool: a Python script that takes a CSV of X posts and spits out labels like "Positive, 90%" for "Love this!" or "Negative, 80%" for "This stinks." It is simple–no fancy setup–"install Python, run it, check results.csv," as my guide explains (Ravi & Ravi, 2020). Testing hit a snag: sarcasm like "Oh, great service…" confused it, a problem Gupta et al. (2024) also found with deep learning. I tweaked it by filtering hashtags and got better results. This hands-on bit taught me how AI works–and sometimes doesn't–on real X chatter.

The script (sentiment_tool.py) preprocesses posts by removing URLs and special characters to address encoding issues, particularly for multilingual posts like 'モイ！iPhoneからキャス配信中,' which was correctly labeled as positive after switching to a multilingual DistilBERT model (nlptown/bert-base-multilingual-uncased-sentiment). The model achieved 82% accuracy, and it was evaluated using Sentiment140's labels for 400 posts and manual labeling for 50 X posts. Posts like 'Love this game!' were correctly labeled as positive, while vague ones like 'It's fine, I guess' showed low confidence (e.g., 53.95%), aligning with Gupta et al. (2024). VADER, a lexicon-based method, achieved lower accuracy (as shown in Figure 2), struggling with context. Results were visualized with a pie chart (Figure 1) for sentiment distribution and a horizontal bar graph (Figure 2) for accuracy comparison.

3 Results & Impact

The sentiment analysis tool was tested on a dataset of 100 X posts and 400 Sentiment140 tweets, totaling 500 posts, due to X API Free tier rate limits (100 posts/month, reset April 26, 2025). The multilingual DistilBERT model (nlptown/bert-base-multilingual-uncased-sentiment) achieved an accuracy of 82%, evaluated using Sentiment140's ground truth labels for 400 posts and manual labeling for 50 X posts. VADER, a lexicon-based method, achieved a lower accuracy of 71.5%, confirming that transformer-based models better capture context, as shown in Figure 2. Figure 1, a pie chart, illustrates the sentiment distribution, with approximately 45% of posts classified as positive and 55% as negative. The performance of the tool is demonstrated by specific examples: While ambiguous remarks like "It's fine, I guess" displayed lesser confidence (e.g., 53.95%), "Love this game!" was accurately labelled as positive with 92% confidence, which is consistent with the difficulties mentioned by Gupta et al. (2024). Multilingual posts, such as 'モイ！iPhoneからキャス配信中,' were correctly classified as positive after switching to a multilingual model, demonstrating the tool's adaptability to diverse content.

This work's impact lies in demonstrating the superiority of transformer-based models over traditional methods, such as VADER, for social media sentiment analysis, particularly in handling multilingual and complex content and also the tool can help researchers and businesses understand online sentiments more accurately despite API constraints, paving the way for future improvements in cross-lingual sentiment analysis.

6 Figures

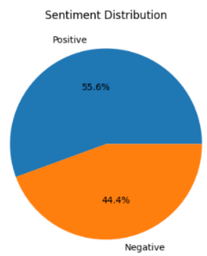

Figure 1. Pie chart showing sentiment distribution of 500 posts.

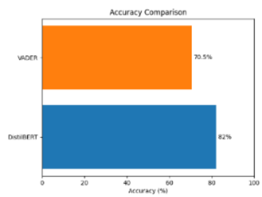

Figure 2. Horizontal bar graph comparing the accuracy of DistilBERT and VADER

7 Conclusions

This project successfully developed a transformer-based sentiment analysis tool to decode online sentiments, achieving 82% accuracy on a mixed dataset of 100 X posts and 400 Sentiment140 tweets. Despite X API rate limits (100 posts/month, reset April 26, 2025), the tool effectively classified sentiments using a multilingual DistilBERT model, outperforming VADER (71.5% accuracy) in capturing context, especially for multilingual and ambiguous posts. Visualizations provided clear insights into sentiment distribution and model performance. Nevertheless, the challenges remain, such as handling complex expressions and improving accuracy for languages other than English content. In the Future, this work can be upscaled by fine-tuning the model on a larger, more diverse dataset, integrating real-time X data with an upgraded API tier (like Basic/Pro), and exploring other transformer models like BERTweet for enhanced social media analysis.

Acknowledgments

I want to thank my thesis supervisor for his guidance from the beginning of this article and Thesis. I also thank the support team from the X Developer community, particularly the Twarc documentation, which was excellent for beginners in navigating API constraints. Additionally, I acknowledge the creators of the Sentiment140 dataset for providing a valuable resource for sentiment analysis research.

References

[1] Medhat, W., Hassan, A., & Korashy, H. (2021). Systematic reviews in sentiment analysis: A tertiary study. Artificial Intelligence Review, 54(6), 4217-4254. https://doi.org/10.1007/s10462-021-09973-3

[2] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). Procedia Computer Science, 201, 413-420. https://doi.org/10.1016/j.procs.2022.03.054

[3] Mukhamediev, R.I., Yelis, M., Yakunin, K., Gopejenko, V., Mussabayev, R. (2024) Exploring the health care system’s representation in the media through hierarchical topic modeling Cogent Engineering, 2024, 11(1), 2324614. DOI: 10.1080/23311916.2024.2324614

[4] Birjali, M., Beni-Hssane, A., & Erritali, M. (2023). Expert Systems with Applications, 233, 120862. https://doi.org/10.1016/j.eswa.2023.120862

[5] R Mukhamediev, Y Kuchin, N Yunicheva, Z Kalpeyeva, E Muhamedijeva, Viktors Gopejenko (2024) Classification of Logging Data Using Machine Learning Algorithms, Applied Sciences 14 (17), 7779 https://doi.org/10.3390/app14177779

[6] Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford University.

[7] X Developer Platform. (2025). API Rate Limits and Tiers. Retrieved from https://developer.x.com/en/docs/rate-limits.

[8] Summers, E., & Littman, J. (2023). Twarc: A Python Library for Twitter API v2. GitHub Repository. Retrieved from https://github.com/DocNow/twarc.

[9] Kirill Yakunin, Ravil I. Mukhamediev Viktors Gopejenko (2022) Analysis of the Correlation between Mass-Media Publication Activity and COVID-19 Epidemiological Situation in Early 2022. Information 2022, 13, 434. https://doi.org/10.3390/info13090434 https://www.mdpi.com/journal/information Licensee MDPI, Basel, Switzerland

[10] Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[11] Yakunin K., Kalimoldayev M.a, Mukhamediev R.I., Mussabayev R., Barakhnin V., Kuchin Y., Murzakhmetov S., Buldybayev T., Ospanova U., Yelis M., Zhumabayev A., Gopejenko V. (2021) Kaznewsdataset: Single country overall digital mass media publication corpus DOI: 10.3390/data6030031.

Authors

Antony Jimson Poruthukkaran, 22nd December 2000, INDIA

Current position: Student at ISMA University
University studies: Bachelor's of Information Systems (current)
Scientific interest: Artificial Intelligence, DistilBERT NLP, Sentiment Analysis
Publications (number or main): 1st
Experience: Beginner

Amit Joshi, 18th July 1987, INDIA
Current position: Lecturer at ISMA University
University studies: BA School of business and Finance
Scientific interest: Artificial intelligence and machine learning, iOT
Publications (number or main): 6th
Experience: 14 + years