Twitch Emote Prediction: Extensions and Benchmarking using Transformers

Yasmine Hejazi
UC Berkeley ISchool
yhejazi@ischool.berkeley.edu

Abstract

The rise of live-streaming platforms like Twitch.tv has revolutionized how users communicate in the form of plain text and ideogram emojis. Despite extensive research on emoji use in a social media setting, emote research in the context of live-streaming platforms has received little attention from the standpoint of Natural Language Processing. In this paper, I investigate the relation between words and Twitch emotes by predicting which Twitch emotes are evoked by text-based Twitch chat messages. I train three models based on Long Short-Term Memory networks (LSTMs) and transformer models (fine-tuned BERT) to compare previous research to the latest up-and-coming technologies. The experimental results show that the fine-tuned BERT model outperforms a majority-class predictor baseline and the two LSTM models, suggesting that transformers are able to better capture emote prediction in a setting of noisy, user-generated text.

1 Introduction

Over the past decade, live-streaming platforms have been growing in popularity and revolutionizing the way people consume content online. One of the largest platforms in the space is Twitch.tv, a platform initially designed for streaming video games that has now expanded to cover a wide range of content, including "just chatting", music creators, and game shows. Twitch was acquired by Amazon in 2014 for almost one billion dollars, and has since grown to welcome over 31 million daily active users (Twitchadvertising.tv 2022).

Users on these platforms watch content creators and comment live on the stream to share their reactions or opinions with the community and streamer. Specifically, the Twitch chat has become home to a very unique community of viewers with a language of its own. The comments in Twitch chat tend to be short in length, rapid-fire, and lacking context as they are typically in reaction to what's happening on the live stream. The Twitch chat is heavily littered with Twitch "emotes," similar to text emojis but containing more meme-like significance.

Twitch emotes have become a language form of their own and have played a huge role in the development of the Twitch community, progressing to the point where people are saying the emote names as real words in daily conversation. There are millions of Twitch emotes that come from many different origins. First, there are platform-provided emotes that all Twitch communities have access to. Then, there are subscriber-level emotes that are rewarded to subscribers of channels that have grown to certain levels of success. Finally, many popular emotes are continuously introduced through third-party extensions like BTTV (betterttv.com). The viewer can render an emote by typing a pre-defined string i.e.; "Kappa" → .

In this paper, I investigate the relationship between words and Twitch emotes by developing a dataset of Twitch chat messages and modeling the underlying sentiment of Twitch emotes. Specifically, I aim to predict which Twitch emotes are evoked by text-based Twitch chat messages. This is anticipated to be a tough task relative to previous social media emoji prediction research because the meaning of Twitch emotes tend to evolve over time and over different Twitch communities. Furthermore, the messages in Twitch chat are noisy, fast, user-generated text that often lacks the context of what is happening on the stream.

Predicting Twitch emotes is valuable because it can provide insights into the behavior and emotions of Twitch users and communities. By analyzing the usage and meaning of different emotes, researchers and marketers can gain a deeper understanding of community opinion and the relevance of industry products, behavior prediction, and spam detection (Kobs et al., 2020).

In this task, I evaluated 4 models: (1) a baseline most-common-class predictor; (2) a Long Short-Term Memory network (LSTM); (3) a bidirectional LSTM network; and (4) a fine-tuned BERT model. In evaluation, I show that the use of transformers outperforms the previous benchmark of technologies (Barbieri et al., 2017). Given the involvement of 30 classes in this multi-class experiment, I used F1-score as the chosen metric of success.

2 Related Work

Twitch emotes are used with a similar communication style to emojis, which are visual cues used widely on social media and instant messaging to express ideas and feelings (Jibril and Abdullah, 2013). Due to their broad usage, the meaning and utilization of emojis has become a hot topic in the NLP research community. Several works studied emoji semantics and usage in the context of embeddings such emoji2vec and EmojiNet (Barbieri et al., 2016; Eisner et al., 2016; Reelfs et al., 2020; Wijeratne et. al., 2016).

Emoji prediction among the common tasks included in NLP research for emojis. LSTM-based classifiers such as Deepmoji are heavily referenced in emoji research and baselines used today (Barbieri et al. 2018; Felbo et al. 2017). Recently, past multi-class emoji prediction efforts have been surpassed by models based on transformer network architectures such as Bidirectional Encoder Representations from Transformers (BERT) (Ma et al., 2020) and RoBERTa (Barbieri et al. 2020).

There are few studies on Twitch and Twitch emotes; because of this, public or available Twitch emote data is nearly nonexistent. Notable Twitch emote studies include sentiment analysis using a convolutional neural network (Kobs et al., 2020) and extended research using traditional machine learning methods (Dolin et al., 2021).

This paper extends off of another noteworthy research study that aimed to achieve classification of the 30 most frequent Twitch emotes (Barbieri et al., 2017a). In this research, the authors utilized a Bidirectional Long Short-Term Memory (LSTM) network to classify a single emote that pertains to a Twitch chat message.

BERT, a model based on transformers utilizing attention, achieved groundbreaking results in numerous Natural Language Processing tasks including classification and sentiment analysis (Devlin et al., 2018). BERT is largely impactful and is the basis for many variations of BERT networks that pre-trained with certain contextual representations such as BioBERT (Lee et al., 2020). Considering the latest research on the effectiveness of transformers in classification tasks, the goal of this paper is to apply a transformer architecture previously used on emoji prediction - a fine-tuned BERT model - into the new context of Twitch chat and emotes (Ma et al. 2020).

3 Data Gathering and Preprocessing

A VOD (Video on Demand) is an archive of content previously streamed live on Twitch. Twitch saves these past broadcasts and their respective chats for a limited amount of time. The broadcast will be removed after some time between 7 and 60 days, depending on the level of the streamer.

With the lack of available Twitch emote data, I attempted to follow a similar data collection and preprocessing procedure as described in Barbieri et al. 2017a in order to provide a close data and model comparison. I created a scraper that collects Twitch chat text data for up to 18 of the latest VODs available for the top 100 streamers at the time of this paper. The top 100 streamers were identified according to the average view count on twitchtracker.com on March 12, 2023. VOD IDs and the respective chat messages were scraped based on availability within the date range of March 12, 2023 to March 20, 2023.

For preprocessing, I lowercased all text to reduce noise in our dataset and removed messages with less than three tokens to reduce data sparsity.

In this corpus, I only kept messages that included one of the thirty most frequent emotes. The possible emotes were limited to the 75 most-used Twitch emotes across the Twitch platform according to streamscheme.com (Goodman 2022). I limited the dataset to messages with one type of emote to avoid class overlap.

The final dataset yields a corpus of 858,770 chat messages in total across 93 streamers on Twitch.

Scraping Step	Count
Raw Data	49.7M
One Unique Emote	4.1M
3+ Tokens	906K
30 Emotes	859K

Table 1: Counts for each significant scraping step.

4 Models Description

In this section, I will describe the methodology followed to construct the four models experimented in this study. These four models include (1) a baseline most-common-class predictor; (2) a Long Short-Term Memory network (LSTM); (3) a bidirectional LSTM network; and (4) a fine-tuned BERT model.

4.1 Baselines

Three baselines were built to compare performance to that of the proposed BERT model. The first baseline is a simple most-common-class predictor that predicts the majority class for all instances. Occupying 21.44% of our Twitch chat corpus, KEKW is the emote that most frequently appeared.

Name	Emote	Count
KEKW		184.09K
LUL		177.8K
Pog		61.59K
Sadge		46.81K
FeelsBadMan		36.11K

Table 2: Top Occurring Emotes in Corpus.

The second baseline model utilizes a Long Short-Term Memory (LSTM) network. LSTMs are proven to be effective on many tasks including classification tasks, including modeling of tweets, which demonstrates an opportunity to apply to Twitch emote prediction (Barbieri et al., 2017b). The inputs of the LSTMs are word embeddings of 50 dimensions. The word embeddings were initialized using a lookup table of GloVe embeddings with a vocabulary of 400,000. Out-of-vocabulary words (OOVs) were represented as a fixed vector at the end of the vocab and handled as a separate word.

Regarding the latest benchmark set for Twitch emote prediction, the third baseline model uses a word-based Bi-directional LSTM network. The B-LSTM architecture has been a top performer in both the context of Twitter emoji prediction (Barbieri et al., 2017a) and Twitch emote prediction (Barbieri et al. 2017b). The B-LSTM architecture I implemented utilizes the same GloVe embeddings as the previous LSTM baseline to achieve the emote prediction task. In the B-LSTM, the forward network reads the message from left to right and the backward network reads in the opposite direction. The learned vector is then passed through the same non-linearity and affine transformation as our first LSTM model before a softmax classification layer.

4.2 Fine-Tuned BERT

Motivated by the overwhelming success of pre-trained transformer models, and also the latest benchmarking using BERT-based models on Twitter emoji prediction, I constructed a fine-tuned BERT model for Twitch emote prediction (Ma et al., 2020). I start with the pre-trained BERT base uncased (bert-base-uncased) model released by Google, stacked with a single linear layer and a softmax classification layer to shrink the space to our 30-class predictions.

5 Results

5.1 Model Comparison

This is a multi-class classification task where I predicted for the 30 different labels displayed in Table 4. I compared a fine-tuned BERT model to three baseline models: the most-common-class predictor, the LSTM model, and the B-LSTM model. I used F1-score to evaluate the models given the high number of classes in the experiment. I report the accuracy and macro-average precision, recall, and F1-score in Table 3.

Model	Accuracy	Precision	Recall	F1-Score
Majority	0.21	0.01	0.03	0.01
LSTM	0.42	0.71	0.32	0.38
B-LSTM	0.44	0.70	0.34	0.41
BERT	**0.51**	0.64	0.44	**0.48**

Table 3: Results of Twitch emote prediction experiments.

In both F1-score and accuracy evaluation metrics, the fine-tuned BERT model outperforms the other three models with an F1-score of 0.48 across all 30 Twitch emotes.

The largest signaling difference between the LSTM models and the BERT model is that the LSTM models produce many more false positives by overpredicting the two majority classes: KEKW and LUL . While this pattern can also be found in the fine-tuned BERT model, the BERT model does a better job of minimizing the false positives of these two classes as described in the next section.

As an example, I inputted the message ”wait for it” into each model. The LSTM models both predicted KEKW, which is the majority class emote that holds a notion of laughing. On the other hand, the BERT model produced a more meaningful prediction, PauseChamp , an emote that is commonly used when there is anticipation or build-up to an exciting ”PogChamp” moment.

5.2 Champion Model Analysis

In this section, I provide a deeper analysis of our champion model (BERT) performance.

Despite the overwhelming class imbalance throughout the 30 classes, the best emote F1-scores do not necessarily correspond to the most frequent emotes. I report in Table 4 that the best predicted emotes are MorphinTime (F1=0.98), FeelsGoodMan (F1=0.88), and FeelsBadMan (F1=0.85). On the other hand, the most frequent emotes have relatively mediocre F1-scores: KEKW (F1=0.50), LUL (F1=0.48), and Pog (F1=0.43).

I display a confusion matrix of the champion model's performance for each emote in Figure 1. The confusion matrix shows that the champion model produced many false positives for KEKW and LUL, showing evidence that the model has a bias towards the two most frequent emotes in our corpus. Despite these two classes having the most false positives, I also observed that the two emotes contest with each other; when the true label is KEKW, the model predicted LUL 32% of the time. I hypothesize this to be a result of semantic overlap between the two emotes as they both contain positive, laughing sentiments in visual appearance along with the shared contexts they are used in. Furthermore, when the actual label is Kappa , the model predicted LUL 57% of the time and KEKW 25% of the time, resulting in Kappa having the worst individual emote F1-score of 0.05. Kappa is used most in sarcasm and trolling, and sarcastic expressions can overlap into situations where a laughing sentiment is detected. Having over 6 times more representation than Kappa, KEKW and LUL appear to be the emotes of choice for a prediction that could contain a laughing connotation, sarcastic or not.

The top performing emote, MorphinTime , performs so well because it is almost always used in its own unique content. This emote is mostly used in variations of the chat spam message "spam this static to support fnatic." Fnatic is the world's leading Esports organization for games like League of Legends, Valorant, and CS:GO. This kind of message is one of many chat spam messages called "copypastas" (TwitchQuotes 2018).

In Table 5, I show some additional predictions made by the fine-tuned BERT model. In Sentence 1, the model predicted KEKW , inferring that the chatter was laughing at the streamer for losing to the game mechanics. However, the actual label for this text was Sadge , implying the chatter was sad for the streamer. These both seem to be valid emotes to the input text, just with different intentions. Considering there are many different people and different intents, it

Emote	Name	P	R	F1	Count
	KEKW	0.44	0.58	0.50	37.2K
	LUL	0.40	0.59	0.48	35.6K
	Pog	0.41	0.45	0.43	12.2K
	Sadge	0.44	0.34	0.39	9.3K
	FeelsBadMan	0.88	0.82	0.85	7.2K
	FeelsGoodMan	0.89	0.86	0.88	5.7K
	Kappa	0.75	0.03	0.05	5.7K
	PepeLaugh	0.73	0.46	0.38	5.3K
	YEP	0.47	0.08	0.13	4.5K
	PogU	0.62	0.10	0.16	4.1K
	monkaOMEGA	0.88	0.82	0.85	3.7K
	AYAYA	0.73	0.61	0.67	3.6K
	MorphinTime	0.99	0.97	0.98	3.6K
	NotLikeThis	0.27	0.16	0.20	3.3K
	gachiBASS	0.81	0.84	0.83	3.3K
	BibleThump	0.56	0.29	0.38	2.8K
	Pepega	0.67	0.36	0.46	2.8K
	PogChamp	0.88	0.36	0.51	2.7K
	TriHard	0.86	0.73	0.79	2.5K
	Kreygasm	0.38	0.34	0.36	2.2K
	WeirdChamp	0.65	0.37	0.47	2.1K
	4Head	0.68	0.44	0.54	1.6K
	PauseChamp	0.48	0.20	0.28	1.8K
	DansGame	0.57	0.33	0.42	1.7K
	FailFish	0.52	0.36	0.42	1.5K
	ResidentSleeper	0.49	0.19	0.28	1.4K
	widepeepoHappy	0.56	0.47	0.51	1.3K
	5Head	0.80	0.08	0.14	1K
	SwiftRage	0.71	0.56	0.63	1K
	4Weird	0.64	0.40	0.49	0.9K

Table 4: Detailed results for each class in Twitch emote prediction for the fine-tuned BERT model. I report the Precision, Recall, F1-Score, and test set frequencies for each emote.

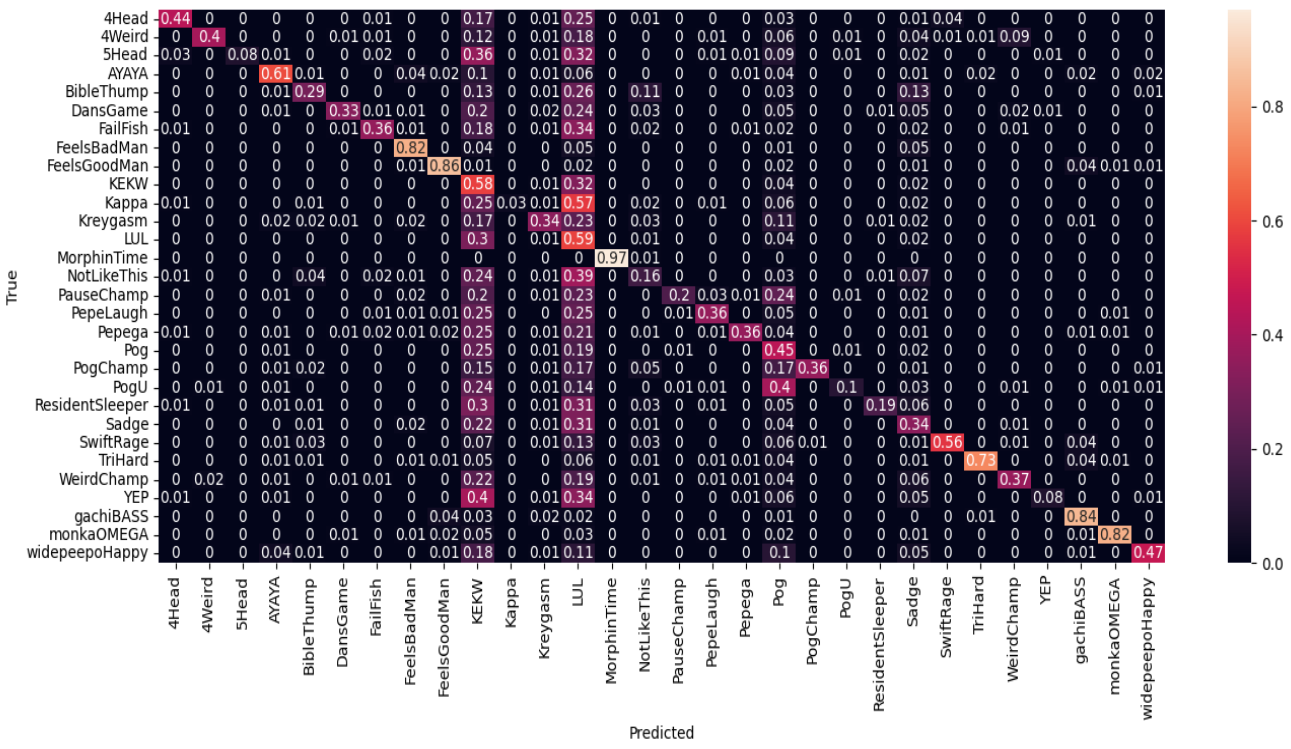

Figure 1: Confusion Matrix of the fine-tuned BERT model performance for the top 30 emotes

can be difficult to distinguish the ambiguity of this text without further context or previous user behavior. Another type of error is the consequence of the existence of multiple emotes with similar meanings. Sentence 2 supports my earlier observation that KEKW and LUL are frequently confused by the model as they are both used interchangeably in situations where something funny has happened.

Text	Prediction	Label
lost to mechanics		
lil bro is fuming		
i love uuu <3		
congrats on the new place. looks awesome		
you don't skip		

Table 5: Example predictions by fine-tuned BERT model.

Emotes can also have completely opposite meanings and still be a source of prediction errors due to possessing multiple semantic uses. AYAYA is typically used as a sign of excitement or happiness while BibleThump is typically used to express disappointment and is one of the most used crying emotes on Twitch. Both are cute and expressive emotes that can be used to quantify someone's love (happy love or crying from love), and therefore the model confuses the two in Sentence 3. Another aspect is that AYAYA originates from an anime meme and is associated with cuteness and anime. It is likely that the model predicted AYAYA due to the cute emphasis in Sentence 3 such as representing ”you” as ”uuu.”

Finally, I display example outputs where the BERT model predicted the emote correctly in some abstract scenarios. The model predicted Pog correctly in Sentence 4, considering that Pog has become its own adjective that is used interchangeably with ”cool” or ”excellent.” The model also correctly predicted DansGame in Sentence 5 where the chatter is likely outraged from the streamer skipping a game level or a song on their playlist.

6 Conclusion

The rise of live-streaming platforms like Twitch.tv has revolutionized how users communicate online. Research on emoji usage has been successful using transformers, but that research has yet to come into the context of live-streaming platforms where user text tends to be short in length, rapid-fire, and lacking in context. In this paper, I formulated the task of modeling the usage of Twitch emotes in order to achieve semantic understanding to aid in Twitch chat comprehension, user behavior, and community opinions.

To achieve this task, I created a corpus of Twitch chat messages and benchmark the dataset with two LSTM-based models and a BERT model based on a pre-trained BERT. The evaluation shows that the BERT model is more capable of performing Twitch emote prediction than its competing baselines, suggesting that transformers are able to better capture emote prediction in a setting of noisy, user-generated text.

As a next step, I propose to conduct a deeper data preprocessing procedure to include the removal of hyperlinks, non-ASCII characters, and repetitions from common spam messages. I would also create token grouping representations for mentions of other Twitch users or streamers in the community. Future work could also include more context into the model, introducing a representation of what was said or what happened on the stream before the emote message appeared. Context can be expanded with a representation of the previous chat messages before the message appeared. Finally, because of the large class imbalance between the 30 emotes, further modeling could be performed on a subset of emotes to attempt to achieve improved of tailored performance.

To support reproducibility and future research, the data and code are made available on a GitHub repository.

References

[Barbieri et al. 2016] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this emoji mean? a vector space skip-gram model for twitter emojis. *In Language Resources and Evaluation conference, LREC.* , Portoroz, Slovenia, pages 526–534.

[Barbieri et al. 2017a] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are Emojis Predictable? In *European Chapter of the Association for Computational Linguistics, EACL*, Association for Computational Linguistics. Valencia, Spain.

[Barbieri et al. 2017b] Francesco Barbieri, Luis Espinosa-Anke, Miguel Ballesteros, Juan Soler-Company, and Horacio Saggion. 2017. Towards the Understanding of Gaming Audiences by Modeling Twitch Emotes. *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 11–20, Copenhagen, Denmark. Association for Computational Linguistics.

[Barbieri et al. 2020] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *In Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

[Devlin et al. 2018] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXis, abs/1810.04805.

[Dolin et al. 2021] Pavel Dolin, Luc d'Hauthuille, and Andrea Vattani. 2021. FeelsGoodMan: Inferring Semantics of Twitch Neologisms. ArXis, abs/2108.08411.

[Eisner et al., 2016] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. *In Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

[Goodman 2022] Goodman Luci. 2022. 75 most popular Twitch emotes! - meaning origin StreamScheme. https://www.streamscheme.com/resources/twitch-emotes-meaning-complxete-list-monkas-pogchamp-omegalul-kappa

[Jibril & Abdullah 2013] Tanimu Ahmed Jibril and Mardziah Hayati Abdullah. 2013. Relevance of emoticons in computermediated communication contexts: An overview. *Asian Social Science*, 9(4):201.

[Kobs et al. 2020] Konstantin Kobs, Albin Zehe, Armin Bernstetter, Julian Chibane, Jan Pfister, Julian Tritscher, and Andreas Hotho. 2020. Emote-Controlled: Obtaining Implicit Viewer Feedback Through Emote-Based Sentiment Analysis on Comments of Popular Twitch.tv Channels. *Trans. Soc. Comput. 3*, 2(7):34 pages.

[Lee et al. 2020] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

[Ma et al. 2020] Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Emoji Prediction: Extensions and Benchmarking. ArXis, abs/2007.07389.

[Reelfs et al. 2020] Reelfs, J. H., Hohlfeld, O., Strohmaier, M., & Henckell, N. 2020. Word-emoji embeddings from large scale messaging data reflect real-world semantic associations of expressive icons. ArXis, abs/2006.01207.

[TwitchQuotes 2018] TwitchQuotes. 2018. Spam this static to help Fnatic. www.twitchquotes.com/copypastas/2757

[Wijeratne et al. 2016] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. EmojiNet: Building a machine readable sense inventory for emoji. *Computation and Language* ArXis, abs/1610.07710.