Milestones in Bengali Sentiment Analysis leveraging Transformer-models: Fundamentals, Challenges and Future Directions

Saptarshi Sengupta, The Pennsylvania State University, USA

Shreya Ghosh, The Pennsylvania State University, USA

Tarikul Islam Tamiti, Rajshahi University of Engineering & Technology, Bangladesh

Prasenjit Mitra, L3S Research Cente, Hannover, Germany

Abstract

Sentiment Analysis (SA) refers to the task of associating a *view polarity* (usually, positive, negative, or neutral; or even fine-grained such as slightly angry, sad, etc.) to a given text, essentially breaking it down to a supervised (since we have the view labels apriori) classification task. Although heavily studied in resource-rich languages such as English thus pushing the SOTA by leaps and bounds, owing to the arrival of the Transformer architecture, the same cannot be said for resource-poor languages such as Bengali (BN). For a language spoken by roughly 300 million people, the technology enabling them to run trials on their favored tongue is severely lacking. In this paper, we analyze the SOTA for SA in Bengali, particularly, Transformer-based models. We discuss available datasets, their drawbacks, the nuances associated with Bengali i.e. what makes this a challenging language to apply SA on, and finally provide insights for future direction to mitigate the limitations in the field.

1 Introduction

Sentiment Analysis, an interdisciplinary field combining computational linguistics, text mining, and machine learning, has grown exponentially over the past decade. It has emerged as a powerful tool for understanding human emotions and opinions in numerous contexts, from product reviews to social media chatter (Yadollahi et al., 2017). However, much of this growth has focused on resource-rich languages, such as English. There remains, thus, a significant gap in research pertaining to low-resource languages, which are often overlooked due to the lack of readily available linguistic tools and datasets.

Among these low-resource languages, Bengali, is a language of considerable interest. Spoken by over 300 million people globally, it remains underrepresented in SA/NLP in general (Shammi et al., 2023). This underrepresentation is problematic as the rich cultural and socio-linguistic context of Bengali users is largely uncaptured, thus limiting our understanding and modeling of their sentiments.

SA in low-resource languages, particularly in Bengali, presents multiple challenges due to a variety of factors such as,

Lack of Annotated Corpora The most significant challenge is the scarcity of large, high-quality, and diverse annotated datasets in Bengali. These corpora are essential for training and validating SA models (Sen et al., 2022). Unlike languages such as English, where multiple large-scale annotated datasets exist, Bengali lacks such resources.

Linguistic Complexity Bengali is a morphologically rich language with complex verb forms, compound words, and a high degree of inflection. This richness adds an extra layer of complexity when developing SA models. Additionally, the language's orthographic variation and lack of standardized spelling further complicate text-processing tasks.

Sarcasm and Idiomatic Expressions Bengali, like many other languages, is rich in idiomatic expressions and sarcasm. Detecting sentiment in such cases requires a deep understanding of the language and cultural context, which is challenging for automated systems.

Limited NLP Tools There is a lack of comprehensive Natural Language Processing (NLP) tools for Bengali. Tools for tokenization, POS-tagging, stemming, and NER, which are readily available for languages like English, are still underdeveloped in Bengali.

Addressing these challenges necessitates a multifaceted approach; creation of large, diverse, and annotated datasets, the development of more sophisticated NLP tools for Bengali, and the application of culturally sensitive modeling techniques.

This paper presents a comprehensive survey of the current state of Bengali SA, aiming to highlight the challenges and opportunities associated with this low-resource language. It investigates the exist-

ing methods and resources in the field and discusses the challenges faced by researchers, such as the lack of un/labelled corpora, linguistic peculiarities of the language, and the cultural nuances that affect sentiment expression. Furthermore, this paper outlines potential strategies and future directions for improving SA in Bengali, thereby fostering a more inclusive representation of global languages in SA research. Through this survey, we aim to provide a solid foundation for researchers venturing into the field of Bengali SA, sparking new ideas, and encouraging the development of more robust, culturally sensitive SA models.

1.1 Types of SA

SA, also known as opinion mining, involves using text analysis, natural language processing, and computational linguistics to identify and extract subjective information from source materials. This process is employed in a variety of contexts, and the results can serve various purposes. Although different problems in their own right, we classify SA into the following categories to provide as much as coverage for Bengali as possible,

Hate Speech Detection: This SA variant is essential for identifying and flagging language that is deemed offensive, aggressive, or harmful (Romim et al., 2021). Its primary application is in moderating online platforms to promote healthy digital communication. Algorithms used in hate speech detection discern between harmless opinion expressions and damaging or toxic language.

Stance Detection: This form of SA is intended to establish the stance or viewpoint of a speaker or writer on a specific subject (Roy et al., 2021). The stance could be supportive, opposed, or neutral. Stance detection aids in assessing public sentiment concerning controversial subjects, brand opinion, or political inclination.

Emotion Mining: This SA category moves beyond simple positive or negative sentiment detection. It aims to pinpoint specific emotions such as joy, anger, sadness, fear, surprise, and so forth, as expressed in text. Emotion mining gives a more profound insight into user reactions to products, services, or events (Iqbal et al., 2022).

Aspect-Based SA (ABSA): ABSA offers a refined version of SA that doesn't just label an entire document as positive, negative, or neutral (Ahmed et al., 2021). Instead, it dissects the sentiment conveyed about different aspects or attributes within the text. For example, in a product review, ABSA can differentiate the sentiment towards the product's cost from the sentiment towards its features or usability.

Depressive Texts: This is a specialized branch of sentiment analysis (Hasib et al., 2023; Ghosh et al., 2023), that identifies and extracts subjective information from source text. This specialized task focuses on identifying signs of depression, based on the sentiments expressed in a user's written text.

1.2 Issues prevalent in SA

Each type of SA serves different needs and has its unique challenges. For instance, (a) **Distinguishing hate speech** from other forms of communication can be challenging. Satire, sarcasm, and cultural nuances can often be misinterpreted by algorithms. Furthermore, the definition of what constitutes hate speech can vary significantly across different cultures and legal systems, making it even more challenging to build universally applicable models. (b) One of the main challenges in stance detection is dealing with **implicit stances**, where the opinion isn't directly stated. Additionally, subjectivity and bias in labeling stance data can affect the accuracy of the model. Also, detecting the stance in less structured texts, such as social media posts, can be challenging due to the use of slang, emojis, and non-standard grammar. (c) Emotion mining is complicated by the fact that the **same text can evoke different emotions in different people**. Also, people express their emotions in varied ways, making the training data diverse and complex. The use of sarcasm, irony, and other forms of figurative language can lead to misinterpretation of the expressed emotion. (d) **Challenges in ABSA include accurately identifying aspects or features mentioned in the text**, especially when they are implicitly stated or referred to using different terminology. Also, determining sentiment toward a specific aspect can be complex when the overall sentiment of the text is different or when multiple sentiments are expressed in the same sentence. (e) **Handling the cultural nuances**, idioms, and language-specific expressions can be tricky. Translation can be used as a workaround, but this often leads to a loss of sentiment-bearing nuances.

Addressing these challenges requires continuous advancements in natural language processing, a deeper understanding of human emotions and language nuances, and the development of more sophisticated machine learning models. The choice

of which to use depends on the specific goals of the analysis, the nature of the data at hand, and the resources available for the task.

1.3 Related surveys

While there exist several surveys for SA in Bengali viz. Shammi et al. (2023); Sen et al. (2022); Hira et al. (2022); Banik et al. (2019); Alam et al. (2021), none of them truly provide broad coverage of Transformers for multilingual or Bengali applications. Our paper attempts to remedy this by discussing all (to the best of our knowledge) transformer variants applicable to Bengali text and showcasing their performance for the same. For an insight into pre-Transformer era techniques, we refer readers to any one of the above articles.

Datasets and Benchmarks

Table 1 provides a comprehensive summary of various research studies benchmarking distinct Natural Language Processing (NLP) tasks related to sentiment analysis. These tasks span a range of subdomains, including aspect-based sentiment analysis, emotion classification, hate speech detection, and depressive text detection. Each task has been examined within various contexts, such as food reviews, newspaper comments, e-commerce comments, and more.

The sentiment analysis category consists of a total number of samples ranging from approximately 1,000 to over 1,58,065. Model performance varies, with accuracy and F1 scores ranging from around 66% to above 94%. In the subdomain of aspect-based sentiment analysis, the tasks have been performed within contexts such as cricket and restaurant reviews. The total sample sizes for these studies sit at around 2,800 to 2,900. The F1 scores for these tasks are reported at 37% and 42%, respectively. Emotion classification tasks encompass a variety of contexts, including comments on government policies, socio-political issues, and YouTube comments. The reported accuracy and F1 scores for these studies are approximately 65% and 62% respectively. Hate speech detection tasks deal with detecting hate speech and cyberbullying, with F1 scores reported at 87% and 85%, respectively. Lastly, depressive text detection involves categorizing social texts into depressive and non-depressive classes, and models in this domain have achieved an accuracy of 94%.

The Transformer-based model, BanglaBERT, exhibited superior performance across various domains, notably outperforming other models by achieving a remarkable Weighted Average F1-score of 0.9331 (Kabir et al., 2023). Furthermore, BanglaBERT achieved state-of-the-art performance on the SentNoB dataset (Bhattacharjee et al., 2022). In the context of emotion classification, the Bangla-BERT model achieved a macro average F1-score of 24.61 across 22,698 Bangla public comments from social media platforms, covering 12 different domains (Islam et al., 2022), which is a quite difficult dataset. Lastly, Transformer-based models, BERT and ELECTRA, were deployed effectively for hate-speech detection, achieving accuracies of 85.00% and 84.92%, respectively, highlighting their potential in large-scale sentiment analysis tasks (Aurpa et al., 2022).

2 Proposed Architectures

As mentioned before, our aim in this survey is to discuss advances enabled by the Transformer architecture (Vaswani et al., 2017) and foundational variants originating from it. The Transformer model was introduced as a solution to bypass the recurrent language models of recent years such as ELMo (Peters et al., 2018). This is because, even though recurrent models were capable of capturing small-to-moderate length input dependencies, they were bottlenecked by their *scalability issue* i.e. there was no way to parallelize computation.

The Transformer then completely replaced recurrence with the *attention* mechanism and feed-forward layers which enabled the model to be scaled up to parameters not seen before in deep learning literature. Essentially, a Transformer is a sequence-to-sequence (seq2seq) model consisting of an encoder, which "encodes" the input text to an internal *contextualized* representation (using Multi-Headed Self Attention or MHA) and the decoder, conditioned on the encoder output, generates the target sequence (through masked MHA, as it is not allowed to look at tokens beyond its current timestep).

The original Transformer was intended mainly for sequence transduction tasks (given an input, covert it to the relevant output). However, architectures started emerging based on two-phase training of either the encoder, decoder, or the entire seq2seq setup. In the first phase (*pre-training*), the model is trained on a language modeling objective in a semi-supervised manner (the labels are obtained from the unlabelled corpus itself) followed by training

Table 1: Dataset Summary and Benchmark. Cl.: number of classes in the dataset, T.S.:total number of samples, A: Accuracy score in percentage, F1:F1-score in percentage, DT: Depressive text detection

Task	Paper	Context	Cl.	T.S.	Result
Sentiment Analysis	Junaid et al. (2022)	*(Food Review)* Customer Reviews about the quality of food from online food services (e.g., Foodpanda). Manually labeled as positive and negative	2	1040	A: 90.86
	Sharmin and Chakma (2021)	*(Social Media)* Reviews and comments on books, people, hotels, products, research, events etc. Manually labeled into three categories (positive, negative, neutral)	3	2979	A:66.06; F1:66.02
	Kabir et al. (2023)	*(BanglaBook)* Bangla book reviews classified into three broad categories: positive, negative, and neutral	3	158065	F1:93.31
	Haydar et al. (2018)	*(Facebook Comments)* Collected from e-commerce and restaurant FaceBook posts and categorized into positive, negative, and neutral class	3	34271	A: 80
	Islam et al. (2021)	*(SentNoB)* Comments on 13 different topics from Prothom Alo (Bangla newspaper) articles, YouTube videos, and social media	3	15728	F1:72.89 (Bhattacharjee et al., 2022)
	Jabin et al. (2022)	*(e-commerce Website)* Reviews collected from Rokomari.com. Labelled into positive and negative class	2	6652	A:94.5; F1:94.52
	Al-Amin et al. (2017)	*(Microblogging Website)* Comments collected. Labelled as positive, negative classes by 500 annotators.	2	16000	A:75.5
Aspect Based SA	Rahman and Kumar Dey (2018)	*(Cricket)* Comments from online sources on five aspect categories (batting, bowling, team, team management, other). Labelled into three sentiment polarities (positive, negative, neutral) and 5 aspect categories	3, 5	2900	F1:37
	Rahman and Kumar Dey (2018)	*(Restaurant)* Manually translated English restaurant dataset (Pontiki et al., 2016) into Bangla. Labelled into five aspect categories (food, price, service, ambiance, miscellaneous) and three sentiment polarities	3, 5	2800	F1:42
Emotion Classification	Rahman et al. (2019)	*(Facebook Comments)* Collected on socio-political issues of Bangladesh. Labelled into six fine-grained emotion classes: sadness, happiness, disgust, surprise, fear, and anger	6	5640	F1:62.39 (Parvin and Hoque, 2021)
	Tripto and Ali (2018)	*(YouTube Comments)* Bangla, English, and Bangla (Romanized) YouTube comments from 2013 to 2018. Labelled into 3 and 5-class sentiment and 6 emotions	3, 5, 6	15689	F1:65.97
	Iqbal et al. (2022)	*(BEmoC)* 7125 texts collected from Facebook, YouTube comments/posts, Bengali story books, etc. Labelled into six emotion categories: anger, fear, surprise, sadness, joy, and disgust by 5 annotators	6	7000	—
Hate Speech Detection	Karim et al. (2021)	*(Facebook, YouTube, and Newspaper Comments)* Extended Bangla Hate Speech dataset (Karim et al., 2020) for Personal, Geopolitical, Religious, etc.	6	8087	F1:87
	Aurpa et al. (2022)	*(Facebook Post)* Bangla texts containing cyberbullying (Ahmed et al., 2021) from comments on Facebook posts of celebrities, athletes, and government officials. Labelled into five harassment categories: sexual, non-bullying, trolling, religious, and threats.	5	44001	F1:85.00
DT	Ghosh et al. (2023)	*(Social Media)* 4784 Depressive and 10247 non-depression social texts collected and labeled as 0/1.	2	15031	A:94.32

(*fine-tuning*) on a labeled downstream task, such as named entity recognition, SA, etc. by applying a task-specific layer (or *head*) to the pre-trained checkpoint. Such two-phase training led to a slew of models which are popularly known today as *Foundation Models* (FMs).

In this work, we discuss only those architectures that are relevant to Bengali i.e. either have been pre-trained only on Bengali corpora or included as a part of the pre-training corpora (similar to the datasets, cf. section 2). With this in mind, we survey the models in figure 1. Note that not all of these models were tested on SA or its related tasks. However, we organize them here to direct interested users to use them for SA since these models have seen Bengali during pre-training.

2.1 Encoders

In this section, we describe those FMs that use only the Transformer encoder. These models usually follow a **Masked Language Modelling** (MLM) objective as their main task along with auxiliary objectives such as entity prediction, etc. During MLM, a random number of tokens in the input sequence are *corrupted* by means such as *replacement (masking)* (replacing individual/spans of tokens by special "mask"/random vocabulary tokens) or *exclusion* (deleting a span) and the idea is to either predict the replaced/masked spans or the original sentence [Table 3 (Raffel et al., 2020)]. We briefly discuss the relevant base architecture and models with Bengali as part of the training corpus, and/or additional training objectives.

1. **BERT** (Devlin et al., 2019) One of the first encoder-based models introducing the bidirectional attention mechanism, i.e. considering text from both sides of a token when computing its representation, along with the Next-Sentence Prediction (NSP) objective (to determine whether or not sentences A & B follow each other) to allow the model to learn properties at the sentence level.

   → M(multilingual) BERT[1] was simply pre-trained on 100 languages with the largest Wikipedia collections.

   → KooBERT[2] was pre-trained on texts from the social media platform *Koo India* [3]. The training corpus consists of 12 languages (including English) and we believe can be leveraged for useful *social media text analysis* in Indian languages.

   → TwHIN(Twitter heterogeneous information network) BERT (Zhang et al., 2022) was designed specifically for tasks involving social media data such as SA and hashtag prediction. TwHIN-BERT replaces the NSP objective with a task designed to determine if a pair of tweets are *socially similar* or not, by computing a contrastive loss between each pair of tweets in a batch.

2. **ALBERT** (Lan et al., 2020) is an efficient variation of BERT as it is smaller in size (parameters) and faster to train. ALBERT achieves these improvements due to three architecture choices, a) sharing parameters across all Transformer blocks b) decomposing the embedding lookup table into two sub-matrices to change hidden dimension size without increasing the number of parameters drastically, and c) Considering Sentence-Order Prediction (SOP) over NSP. It is a type of NSP, but, the idea here is to determine whether a pair of sentences are in the correct "order" rather than whether sentence A "follows" B.

   → IndicBERT Kakwani et al. (2020) pre-trained an ALBERT-based model (IndicBERT) on their IndicCorp corpus which contains text in 12 languages (11 Indian + English) on topics of social interest such as online news sources and magazines. Curiously, they remove the SOP objective during pre-training relying only on MLM (reason not provided in their paper).

   → SahajBERT Diskin et al. (2021) proposed a novel training paradigm in which an ALBERT model was trained on Bengali corpora using a *volunteer compute* approach i.e. pooling resources from several individuals to train a single model. This is a complete monolingual model consisting of the Bengali parts of the Wikipedia [4] and OSCAR (Ortiz Suárez et al., 2020) corpus.

[1] HuggingFace model: `bert-base-multilingual-cased`
[2] `https://huggingface.co/Koodsml/KooBERT`
[3] `https://www.kooapp.com/`
[4] `https://dumps.wikimedia.org`

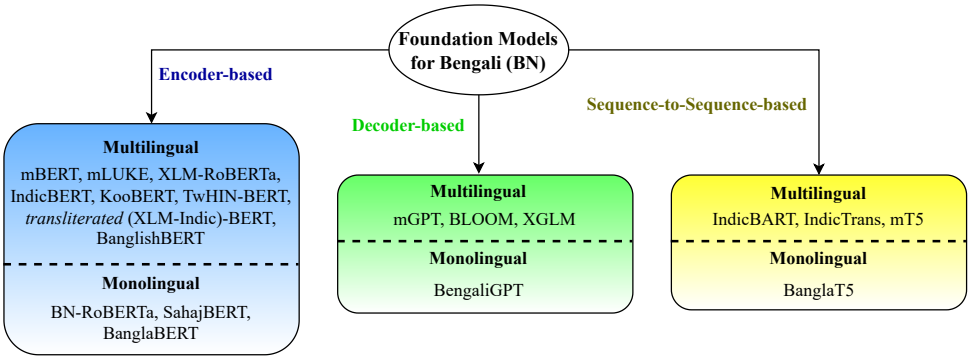

Figure 1: Models considered in this survey classified according to architecture type.

→ Transliterated-BERT Although not the official name, (Moosa et al., 2022) propose ALBERT and RemBERT (Chung et al., 2020) models trained on *transliterated* (writing a source language text in the target language script) versions of Indian languages. They conjectured that when several languages share similarities such as structure, common words, etc. it could be beneficial to transliterate them to a common script and train a FM on it. They chose a set of 20 languages (19 Indian + English) to transliterate to Latin script using the ISO 15919 standard and trained models on both the transliterated and multilingual corpus showing improvements with the former.

3. **RoBERTa** (Liu et al., 2019) is essentially BERT but with better optimization methods such as dropping the NSP objective (which they show degrades performance), using byte-pair encoding instead of wordpiece and training with much larger data, batches and steps. This results in a model, larger (parameters & vocabulary) & more efficient than BERT. bnRoBERTa (Jain et al., 2020) is a monolingual RoBERTa model trained only on the Bengali portion of OSCAR while XLM-(R)oBERTa (Conneau et al., 2020) is a multilingual model trained on a dataset of 100 languages from CommonCrawl with other pre-training objectives such as Translation Language Modelling (concatenating sentences in two languages + MLM across all randomly masked tokens) for cross-lingual tasks such as XNLI, etc. Yamada et al. (2020) proposed **LUKE**, a RoBERTa model trained with an masked entity prediction task (since it treats "words" and "entities" as different) in addition to MLM. These additional objectives aid in entity-forward tasks such as NER, Cloze-style QA, etc. mLUKE (Ri et al., 2022) then builds on LUKE by training a multilingual model (similar to XLM-R) on Wikipedia dumps of 24 languages.

4. **ELECTRA** (Clark et al., 2020) is a generative adversarial network (GAN) like *pre-training approach* in which two BERT models, termed generator and discriminator, are jointly trained. The generator is trained via standard MLM, producing outputs for the masked input sequence. After prediction, the masked tokens, in the original sequence, are replaced by sampling the generator output. The discriminator is then tasked to determine whether each token belongs to the original or replaced sequence. Such a setup has been shown to lead to a more efficient model in terms of training time and performance. BanglaBERT and BanglishBERT (Bhattacharjee et al., 2022) are two ELECTRA models trained on monolingual and EN (same as English BERT's pre-training corpora)-BN corpora. Owing to the small size of the Bengali Wikipedia dump, the authors collected content from the top sources (news, blogs, ebooks, etc.) as cited by Amazon Alexa rankings. **To the best of our knowledge, BanglaBERT seems to be the "best" model** to date on a wide range of Bengali tasks (BLUB (Bangla Language Understanding Benchmark))

2.2 Decoders

Unlike encoder models, Transformer decoders are more *classical* in that they are trained using the extant paradigm of **CLM** (Causal Language Mod-

elling) i.e. predicting the next word in a sequence being conditioned on the prior tokens, usually without auxiliary training objectives, unlike encoders. In decoders, the attention computation for a given token only has access to its preceding tokens (otherwise it would be cheating for the model to know what tokens come after it) unlike encoders which can use all of the tokens surrounding the masks.

While there are several decoder-based architectures, the most popular in this category seems to be the **GPT** (Generative Pre-Training) family of models (Zong and Krishnamachari, 2022). Each variant, GPT-1,2,3 etc. follows the training scheme for CLM but differs in ways such as using cleaner/diverse corpora and scaling up parameters resulting in better zero/few-shot capabilities.

Although GPT-3 (Brown et al., 2020) was shown to have reasonably good performance in languages apart from English (Armengol-Estapé et al., 2022) the training data would suggest that for languages, like Bengali, occupying a very small fraction of the dataset, downstream performance would not be good. In an effort to combat this, Lin et al. (2022) proposed XGLM, a multilingual generative model GPT-3 like model. Almost parallelly released was mGPT (Shliazhko et al., 2022) a GPT-2 (Radford et al., 2019) like architecture. While both are multilingual, there exists subtle differences between the two. XGLM used GPT-3 curie and wanted to examine the effect of model scale for low-resource languages but trained on a smaller/focused set of the same (30 total). mGPT on the other hand, wanted to replicate the GPT-3 architecture using open-source tools such as the Megatron-LM framework (Shoeybi et al., 2019), perform comparison with XGLM and cover more languages (60). The only monolingual GPT-based model seems to be Bengali-GPT (Ghosh, 2016) a GPT-2 model trained on the Bengali part of the mC4 corpus [5].

A surprise addition to this list of models is **BLOOM** (Scao et al., 2022), a BigScience [6] open-source multilingual model trained on scientific texts. Similar to mGPT, BLOOM uses Megatron's GPT-2 checkpoint to train on a diverse corpora of 46 different languages and 13 programming languages. Although their paper does not contain a breakdown of domains/topics considered per-language, we can see that with just 18 GB of Bengali corpora, they achieve impressive performance, at times comparable to high-resource languages like Spanish.

2.3 Sequence-to-Sequence

Finally, we discuss FMs consisting of encoders and decoders (seq2seq). In such setups, the encoder follows its MLM objective, formulates a representation for the input sequence, passes it over to the decoder which on being conditioned by it, performs CLM. However, during training, the decoder generates tokens in parallel irrespective of what the expected output is i.e. teacher forcing.

For Indic languages, the first major contribution has to be IndicTrans (Ramesh et al., 2022), a Transformer trained on their *Samanantar* dataset, a parallel corpus of 11 Indian languages to English. They mention that IndicTrans is a *uniscript* model i.e. all non-English text is converted to the Devnagari script which allows for broader lexical coverage across all the Indian languages.

T5 (Text-To-Text Transfer Transformer) (Raffel et al., 2020) makes the case that all standard NLP tasks, such as classification, question answering, etc. can be cast as seq2seq problems. They train T5 using a combination of supervised (as above) and unsupervised (MLM) tasks. For the former, the input must be formatted to their requirements such as prefixing the input with the task handle, etc. while for masking in general, T5 uses *span-masking* i.e. corrupting the input sequence by masking consecutive tokens, with the decoder tasked to predict the replaced spans. mT5 (Xue et al., 2021) is a multilingual-T5 model pre-trained on 101 languages from the mC4 corpus. It must be noted that since mT5 was trained using unsupervised CLM only, it will display random zero/few-shot performance, unless fine-tuned. BanglaT5 (Bhattacharjee et al., 2023) is, to the best of our knowledge, the only monolingual seq2seq model for Bengali. Pre-trained on the *Bangla2B+* corpus (Bhattacharjee et al., 2022), BanglaT5 achieves impressive performance in comparison to its other multilingual counterparts on their BanglaNLG benchmark.

Finally, we discuss **BART** (Lewis et al., 2020) a seq2seq model trained **very similarly to T5** (span-masking). However, it is our understanding that apart from a few architecture choices such as absolute v/s relative position embeddings, the key difference between BART and T5 is that the latter was trained in a multi-task setting whereas the former was trained simply with MLM + CLM. The

[5] https://huggingface.co/datasets/mc4
[6] https://bigscience.huggingface.co/

encoder's job is to reconstruct or *denoise* the corrupted input sequence (by a variety of techniques such as masking, sentence reordering, etc.) while the decoder learns standard CLM using the original input sequence and encoder output. mBART-50 (Tang et al., 2021) is then a multilingual extension of BART, covering bilingual pairs of 49 languages to English and pre-trained on publicly available parallel corpora from WMT, IWSLT, etc. IndicBART (Dabre et al., 2022) is a further extension of mBART covering 11 Indic languages not seen by it earlier.

3 Open Challenges

Seeing as SA in Bengali has come a long way, we highlight ongoing challenges in the field. As mentioned before, BN, like other low-resource languages, suffers from a **lack of resources**, datasets, and pre-trained models. Consider BERT for English. The corpora which was used, had a total of 3.3B tokens whereas the "best" model in Bengali, BanglaBERT, was trained with 2.1B tokens (**~1.5 times less**). Considering the linguistic complexity/diversity of BN, the value should have been 1.5 times more. This is a clear testament to the EN-BN performance gap. Efforts such as *AI4Bharat*[7] are thus a brilliant initiative in this direction.

Considering the socio-geographic distribution of BN, we recognize that there are two camps of writers viz. those who write in pure BN script & those who **code-mix** i.e write BN & another language (typically EN) in the same sentence. This makes it difficult to develop models for BN as users comfortable in both languages typically fall back on EN through transliteration leaving pure-script users with underpowered models. However, with tools such as *Google Keyboard* [8] support for Bengali has become more engaging and easy & with the appropriate permissions, we can curate a massive collection of user-generated pure-script text.

Mainly, there are two styles of BN, formal (*sAdhu*) & colloquial (*calit*) (Pal et al., 2021) and modern users typically favor the latter. This creates issues for models trained on more structured text, such as from Wikipedia, owing to most users relying on **relaxed grammar rules**. To alleviate this, models such as TwHIN-BERT should be developed for BN by training on a combination of both formal/colloquial and social-media texts. Additionally, as is a pervading issue in NLP nowadays, text mined from social media tends to be rife with biases. Thus, care needs to be taken to apply appropriate filtration before training our models.

Future Directions

While we motion for the need for larger and more diverse Bengali corpora, we recognize the difficulty of the task considering limitations in digitizing Bengali text such as scale, copyright issues and pipelines not being end-to-end (E2E) (Sankar et al., 2006). However, we recognize an opportunity here from the ongoing challenges. If the majority of BN users prefer to transliterate their texts, why not take advantage of that? Developing **E2E transliteration-translation datasets** will be useful for training models to synthesize pure-script monolingual data.

Inspired by BLOOM, we encourage researchers to explore **complex domains** in BN such as Medical (Sazzed, 2022), IT (Mumin et al., 2014), etc. Work along this direction can enable a wide range of applications in BN such as chat-agents capable of understanding both cultural nuances and medical jargon, aiding automated customer service etc.

Relying on text modality alone for detecting sentiment limits our models' ability to pickup on subtleties in communication such as prosody, body language, (Kundu et al., 2022). Human beings express sentiment via several *tells* such as facial expressions, voice modulation, etc. Looking beyond the extant text-based SA, a **multimodal** approach (Habimana et al., 2020; Zhu et al., 2023) to the problem should be investigated. **Memes** provide the perfect hunting ground for such analysis. Not only are they a source of entertainment, the paradoxical nature of memes makes it difficult for our models to gauge the conveyed message, enabling us to determine their robustness. With the release of datasets as *MemoSen* (Hossain et al., 2022), research in this direction will be truly propelled.

4 Conclusion

In this paper, we have surveyed a wide array of Transformer-based models which can be used to address SA in Bengali. Going over the SOTA, it is clear that while positive strides have been made to push Bengali to high-resource territory, much work remains to be done. Highlighting the existing architectures and challenges in the field, we believe that our paper can be treated as a means to beckon more research in this area.

[7] https://ai4bharat.org/

[8] https://en.wikipedia.org/wiki/Gboard

Limitations

In this paper, we focus our survey only on post-Transformer models, as other surveys (Shammi et al., 2023) (Sen et al., 2022) (Hira et al., 2022) (Banik et al., 2019) (Alam et al., 2021) on Bangla SA have covered pre-Transformer models extensively. We have also not independently verified claims made by the various papers meticulously since that goes beyond the scope of this paper. Our objective is to survey the state-of-the-art as it exists today in the literature assuming that the results claimed by the respective authors are reproducible.

Ethics Statement

This work primarily summarizes the work done on the Bengali language and no additional work using any human input was obtained. We provide full credit to all original works and have tried to carefully quote all parts of this work that have been obtained verbatim from these works. We believe our paper will enable the world to better understand the state-of-the-art in Bangla NLP, specifically sentiment mining. The impact will be of convenience, i.e., a quicker understanding of the existing tools, and thus the same benefits and costs/abuses of those, apply here. However, the ethical risks arising from this paper is smaller since these techniques already exist. We merely make finding them easier.

References

Md Faisal Ahmed, Zalish Mahmud, Zarin Tasnim Biash, Ahmed Ann Noor Ryen, Arman Hossain, and Faisal Bin Ashraf. 2021. Bangla text dataset and exploratory analysis for online harassment detection. *arXiv preprint arXiv:2102.02478*.

Md Al-Amin, Md Saiful Islam, and Shapan Das Uzzal. 2017. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In *2017 international conference on electrical, computer and communication engineering (ECCE)*, pages 186–190. IEEE.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Tanjim Taharat Aurpa, Rifat Sadik, and Md Shoaib Ahmed. 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24.

Nayan Banik, Md Hasan Hafizur Rahman, Shima Chakraborty, Hanif Seddiqui, and Muhammad Anwarul Azim. 2019. Survey on text-based sentiment analysis of bengali language. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 714–723.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*,

pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry V Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. 2021. Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34:7879–7897.

Ritobrata Ghosh. 2016. Bangla gpt-2.

Tapotosh Ghosh, Md Hasan Al Banna, Md Jaber Al Nahian, Mohammed Nasir Uddin, M Shamim Kaiser, and Mufti Mahmud. 2023. An attention-based hybrid architecture with explainability for depressive social media text detection in bangla. *Expert Systems with Applications*, 213:119007.

Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. 2020. Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences*, 63:1–36.

Khan Md Hasib, Md Rafiqul Islam, Shadman Sakib, Md Ali Akbar, Imran Razzak, and Mohammad Shafiul Alam. 2023. Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey. *IEEE Transactions on Computational Social Systems*.

Mohammad Salman Haydar, Mustakim Al Helal, and Syed Akhter Hossain. 2018. Sentiment extraction from bangla text: A character level supervised recurrent neural network approach. In *2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2)*, pages 1–4. IEEE.

Suma Hira, Atish Kumar Dipongkor, Saumik Chowdhury, Mostafijur Rahman Akhond, Syed Md Galib, et al. 2022. A systematic review of sentiment analysis from bengali text using nlp. *American Journal of Agricultural Science, Engineering, and Technology*, 6(3):150–159.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

MD Asif Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: a corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):135.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Khondoker Ittehadul Islam, Tanvir Yuvraz, Md Saiful Islam, and Enamul Hassan. 2022. Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 128–134.

Sanjida Jabin, Mahbuba Sumia Suhi, Md Fahim Arefin, and Khan Md Hasib. 2022. Comparison of different sentiment analysis techniques for bangla reviews. In *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 288–293. IEEE.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint arXiv:2011.02323*.

Mohd Istiaq Hossain Junaid, Faisal Hossain, Udyan Saha Upal, Anjana Tameem, Abul Kashim, and Ahmed Fahmin. 2022. Bangla food review sentimental analysis using machine learning. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0347–0353. IEEE.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain,

Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Rohit Kundu, Preethi Jyothi, and Pushpak Bhattacharyya. 2022. Survey: Exploring disfluencies for speech-to-speech machine translation.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2022. Does transliteration help multilingual language modeling? *arXiv preprint arXiv:2201.12501*.

Md Abdullah Al Mumin, Abu Awal Md Shoeb, Md Reza Selim, and M Zafar Iqbal. 2014. Sumono: A representative modern bengali corpus. *SUST Journal of Science and Technology*, 21(1):78–86.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Alok Ranjan Pal, Diganta Saha, Sudip Kumar Naskar, and Niladri Sekhar Dash. 2021. In search of a suitable method for disambiguation of word senses in bengali. *International Journal of Speech Technology*, 24:439–454.

Tanzia Parvin and Mohammed Moshiul Hoque. 2021. An ensemble technique to classify multi-class textual emotion. *Procedia Computer Science*, 193:72–81.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Md Rahman, Md Seddiqui, et al. 2019. Comparison of classical machine learning approaches on bangla textual emotion analysis. *arXiv preprint arXiv:1907.07826*.

Md Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Anurag Roy, Shalmoli Ghosh, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. An unsupervised normalization algorithm for noisy text: A case study for information retrieval and stance detection. *Journal of Data and Information Quality (JDIQ)*, 13(3):1–25.

K Pramod Sankar, Vamshi Ambati, Lakshmi Pratha, and CV Jawahar. 2006. Digitizing a million books: Challenges for document analysis. In *Document Analysis Systems*, volume 7, pages 425–436. Springer.

Salim Sazzed. 2022. BanglaBioMed: A biomedical named-entity annotated corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Ovishake Sen, Mohtasim Fuad, Md Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md Kamrul Hasan, Md Abdul Awal, Awal Ahmed Fime, Md Tahmid Hasan Fuad, Delowar Sikder, et al. 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning based methods. *IEEE Access*.

Shumaiya Akter Shammi, Sajal Das, Narayan Ranjan Chakraborty, Sumit Kumar Banshal, and Nishu Nath. 2023. A comprehensive roadmap on bangla text-based sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–29.

Sadia Sharmin and Danial Chakma. 2021. Attention-based convolutional neural network for bangla sentiment analysis. *Ai & Society*, 36:381–396.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.

Mingyu Zong and Bhaskar Krishnamachari. 2022. a survey on gpt-3. *arXiv preprint arXiv:2212.00857*.