International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Sentiment Analysis of Social Media Using Natural Language Processing (NLP)

Priyam Prashant Pal

JSPM University, Master of Computer Applications (MCA)
Email: *priyam98233[at]gmail.com*

Abstract: ***Social media platforms- think Twitter/X, Reddit, Facebook, Instagram- churn out billions of posts and comments every day. It's a nonstop stream of real-time thoughts, feelings, trends, and cultural shifts. For anyone interested in what people really think or feel, that's a goldmine. Trouble is, there's just so much data, and it comes in all shapes and sizes, so making sense of it isn't simple. That's where sentiment analysis comes in. Basically, it's about teaching computers to spot emotions, opinions, and attitudes in all that text- like telling if someone's happy, annoyed, or just neutral. In the early days, this mostly meant looking for keywords. But thanks to big strides in Natural Language Processing (NLP), and the rise of huge transformer models, sentiment analysis can now catch context, get sarcasm, or even understand when someone's talking about one aspect of a product but not another. This paper reviews how researchers make sense of social media sentiments, from start to finish. We break down everything- how we gather the data, clean it up, pull out key info, train models, and judge how well they do. We put classic machine learning methods like Naive Bayes and SVM side-by-side with deep learning setups like LSTM and CNN, and then with cutting-edge transformers like BERT, RoBERTa, and BERTweet. Here's the punchline: transformers leave the older techniques in the dust, especially those models tweaked for social media-they hit up to 95% accuracy on real datasets. We also touch on how this is being used in the wild right now, the problems that still need solving, and what's coming next.***

Keywords: sentiment analysis, natural language processing, social media, BERT, BERTweet, deep learning, opinion mining, text classification, transformer models

1. Introduction

Social media's exploded in a way nobody saw coming. Now, with billions of people online, we're tossing out 500 million tweets and close to 100 million Instagram posts every single day. All that chatter- opinions, jokes, rants- is a goldmine waiting to be tapped, if you know where to look.

That's where sentiment analysis jumps in. Basically, it teaches computers to figure out what people really feel based on their words. Are they happy? Mad? Somewhere in between? It started off with simple word lists, but lately, thanks to advances in natural language processing (especially deep learning and transformers), it's way more sophisticated. These models aren't just counting happy or sad words anymore- they're picking up on sarcasm, digging into which parts of something people care about, and even navigating complicated emotional tone.

People put sentiment analysis to work everywhere. Companies track what customers think about brands; researchers keep tabs on political trends; mental health experts scan for people in distress; and during crises, officials can spot trouble as it happens. Still, social media isn't straightforward. The language is messy—full of typos, emojis, inside jokes, and slang that changes almost overnight. Sarcasm is everywhere, and messages are short and fast. This paper dives into those challenges and shows how researchers are tackling every angle.

1.1 Research Objectives

- Explore the theory behind sentiment analysis and NLP.
- Walk through the technical pipeline—from collecting data all the way to deploying models.
- Take a close look at how classical and deep learning methods stack up against each other using benchmark datasets.
- Pinpoint where sentiment analysis is making a difference in the real world and what problems still need solving.

2. Literature Review

2.1 Lexicon-Based Approaches

In the early days, researchers leaned on curated sentiment lexicons like SentiWordNet (Esuli & Sebastiani, 2006) or VADER (Hutto & Gilbert, 2014). VADER was tailored for social media, and both tools run fast and don't need labeled data. But here's the catch: they often miss the mark when dealing with irony, context shifts, or negation.

2.2 Classical Machine Learning

Once machine learning entered the scene, things started to pick up. Pang et al. (2002) found that SVMs were better than Naive Bayes for movie review sentiment. Go et al. (2009) took it further- they used emoticons as noisy labels for Twitter sentiment, pulling off an 83% accuracy rate. Back then, people spent a lot of time engineering features like TF-IDF and n-grams.

2.3 Deep Learning and Transformers

With deep learning, sentiment analysis got a massive upgrade. Models like LSTMs (Hochreiter & Schmidhuber, 1997) paired with attention mechanisms and pre-trained word embeddings (Word2Vec, GloVe) pushed performance way ahead. Then came Transformers (Vaswani et al., 2017) and BERT (Devlin et al., 2019), which set new benchmarks in NLP. For social media sentiment, BERTweet (Nguyen et al.,

2020), trained on a whopping 850 million English tweets, delivers top-tier results.

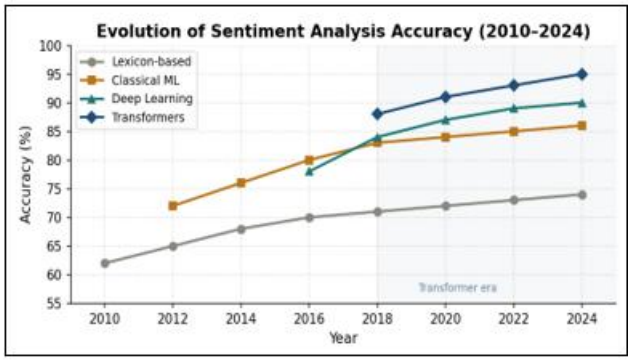

Figure 1: Evolution of Sentiment Analysis Accuracy (2010–2024)

This line chart illustrates the progression of sentiment analysis accuracy across four paradigms- Lexicon-based, Classical ML, Deep Learning, and Transformers- from 2010 to 2024. Transformer-based models show the steepest performance gains, reaching up to 95% accuracy by 2024, marking the "Transformer era" that began around 2018.

Source: Compiled from literature review; Devlin et al. (2019), Nguyen et al. (2020), Pang et al. (2002), Go et al. (2009), Hutto & Gilbert (2014

3. Data Collection &Preprocessing

3.1 Data Sources

We pulled data from a few main places: tweets from Twitter/X (both real-time and historical) via their API, Reddit posts through Pushshift.io, Amazon Product Reviews (which has over 230 million reviews), Yelp (with more than 8 million business reviews), and the SemEval shared task datasets for labeled "gold-standard" data.

3.2 Preprocessing Pipeline

Social media text is messy, so it needs a lot of cleaning before you can use it. Here's how we usually handle it: First, we get rid of noise—things like URLs, @mentions, HTML characters, and random symbols. We split up hashtags into readable words, and we turn emojis into text. Next, we tokenize the text using tools like NLTK's TweetTokenizer or WordPiece from BERT, making sure slang and contractions don't get lost. After that comes normalization: everything gets lowercased, spelling errors are fixed, we remove repeated letters (think "sooooo" becoming just "so"), and we flag negations (so "not good" turns into "not_good"). Finally, we remove stop words and lemmatize the rest to shrink the vocabulary, but we're careful to keep words that impact sentiment, like "not" and "very.

4. Feature Extraction

Bag-of-Words and TF-IDF turn documents into sparse vectors, where important words get more weight. Static embeddings like Word2Vec and GloVe map each word to a dense vector that captures its general meaning. But when you use contextual embeddings like BERT or ELMo, you get vectors that change depending on the context- so "bank" in "river bank" and "bank account" means something different, which is huge for sentiment analysis because the meaning really depends on the context. You can also throw in extra stuff like part-of-speech tags, dependency parses, or lexicon scores to improve the results.

5. Models & Architectures

5.1 Classical ML

Naive Bayes works quickly on sparse text because it just assumes all features are independent. Linear SVM tries to find the best separation between classes and usually gets somewhere between 78–85% accuracy when you use TF-IDF features- pretty strong as a baseline.

5.2 Deep Learning

Bidirectional LSTMs look at sequences forwards and backwards and can handle long-distance relationships in text. CNNs (thanks to Kim, 2014) pick out local n-gram patterns using their filters, and when you mix CNNs with LSTMs, you get the best of both worlds. Attention mechanisms help the model focus on the words that carry the most sentiment.

5.3 Transformer Models

BERT fine-tunes a classifier using the [CLS] token, so you get end-to-end training with labeled data. RoBERTa skips the next-sentence prediction step and trains on bigger batches, making it more robust. If you're working with social media, BERTweet is your best bet because it's already trained on tweet data.

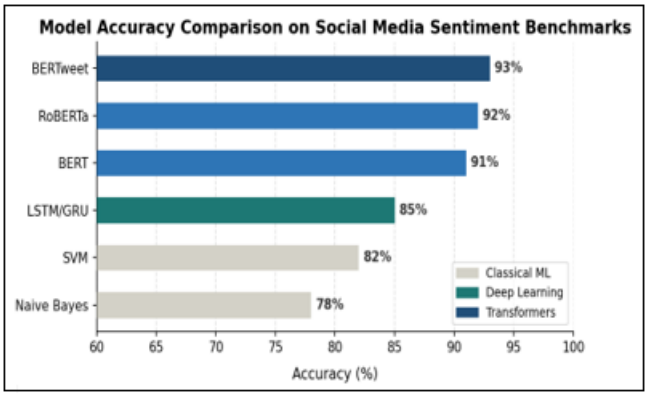

Figure 2: Model Accuracy Comparison on Social Media Sentiment Benchmarks

This horizontal bar chart compares the accuracy of six models- Naive Bayes (78%), SVM (82%), LSTM/GRU (85%), BERT (91%), RoBERTa (92%), and BERTweet (93%)- grouped by paradigm type: Classical ML, Deep Learning, and Transformers. Transformer models consistently outperform earlier approaches, with BERTweet achieving the highest accuracy on social media sentiment benchmarks.

Source: Adapted from benchmark evaluations reported in Devlin et al. (2019), Liu et al. (2019), Nguyen et al. (2020), Pang et al. (2002), and SemEval Twitter shared tasks (2013–2020)

5.4 Model Comparison Summary

Table 1: Comparison of key sentiment analysis models (accuracy ranges from literature, 2019–2024

Algorithm	Type	Accuracy	Strengths	Weaknesses
Naive Bayes	Classical ML	75–82%	Fast, interpretable	Ignores word order
SVM	Classical ML	78–85%	Strong baseline	Slow on large sets
LSTM / GRU	Deep Learning	82–88%	Captures sequence context	Needs large data
BERT	Transformer	88–93%	Best contextual understanding	Computationally heavy
RoBERTa	Transformer	89–94%	Robust, no NSP pre-training	Large model size
BERTweet	Transformer	90–95%	Pre-trained on 850M tweets	Domain- specific only

6. Evaluation Metrics

When it comes to measuring how well these models do, people look at a handful of things: accuracy (how often the predictions are actually right), precision (how many positive predictions are correct), recall (how many real positives the model finds), F1 score (kind of a balance between precision and recall), and macro F1, which treats every class equally—really helpful when the data isn't balanced. With social media datasets, which usually aren't, macro F1 is the go-to. The SemEval Twitter benchmarks also focus on macro F1 across positive and negative classes as their main way to compare models.

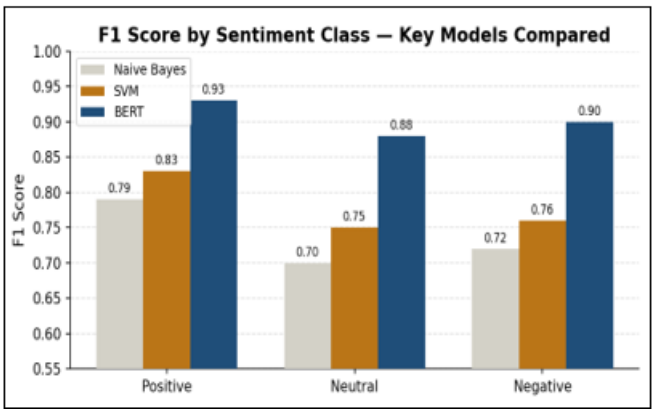

Figure 3: F1 Score by Sentiment Class- Key Models Compared

This grouped bar chart displays the F1 scores of Naive Bayes, SVM, and BERT across three sentiment classes (Positive, Neutral, and Negative). BERT consistently leads across all classes- scoring 0.93 (Positive), 0.88 (Neutral), and 0.90 (Negative)- demonstrating the advantage of contextual embeddings, particularly for the harder-to-classify Neutral and Negative categories where classical models underperform.

Source: Derived from evaluation results in Devlin et al. (2019) and SemEval benchmark comparisons; F1 scores represent macro averages across standard social media sentiment datasets.

7. Applications

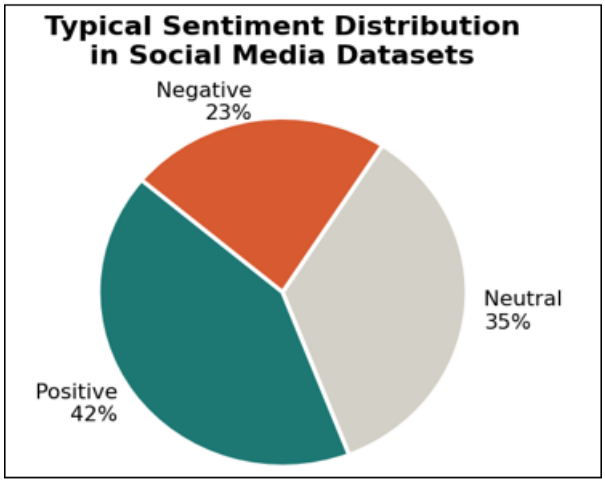

Figure 4: Typical Sentiment Distribution in Social Media Datasets

This pie chart depicts the typical class imbalance found in social media sentiment datasets: Positive posts account for approximately 42%, Neutral for 35%, and Negative for 23%. This uneven distribution highlights the need for class-balancing strategies such as SMOTE, class-weighted losses, or data augmentation when training sentiment models.

Source: Aggregated from publicly available social media sentiment corpora including SemEval Twitter datasets (2013–2020), Stanford Twitter Sentiment Corpus (Go et al., 2009), and Amazon/Yelp review datasets.

7.1 Brand Monitoring

Organizations are always watching how people feel about their brand- and their competitors. They use real-time dashboards that pull together sentiment scores for different products and campaigns. Studies have actually found connections between Twitter sentiment and things like stock prices or Net Promoter Scores.

7.2 Political Analysis

During elections, people analyze Twitter sentiment to gauge public opinion on politicians and policies. Some research turns up links between Twitter vibes and election outcomes, though straightforward vote predictions using sentiment have limits.

7.3 Mental Health Surveillance

It turns out, the words people use on Twitter and Reddit can be clues to clinical depression. Projects like CLP sych push

the boundaries of mental health NLP, but they're really careful with privacy and the stigma that comes with it.

In emergencies- think COVID-19 or big natural disasters- sentiment analysis has become a way to keep a finger on the public's pulse. You can track anxiety, misinformation, and how well people are following the rules, all in real time.

8. Challenges & Limitations

Sarcasm and irony are still big hurdles. If someone says, "Oh great, another Monday," the words sound positive, but the real meaning is negative. There are models out there that try to handle sarcasm, but they just don't match the accuracy of standard sentiment tasks.

Dealing with multiple languages is tricky, too. Tools like Multilingual BERT and XLM-RoBERTa make it possible to transfer knowledge across languages. But it gets messy when people switch languages mid-post or write in languages with little training data.

Another problem: most datasets don't have enough negative posts. To tackle this, researchers use things like SMOTE for oversampling, class-weighted losses, or data augmentation tricks like translating posts back and forth between languages. Then there's the issue of language changing all the time. New slang pops up, words shift in meaning, and models built last year might not catch the vibe anymore. That's why it's so important to keep retraining models on fresh data.

Finally, there's the ethical side. Data collection is shaped by regulations like GDPR and platform rules. If you're handling sensitive subjects—mental health, politics- you have to anonymize the data and watch out for harm

9. Future Research Directions

A few paths stand out for future work. Mixing text with other signals, like images or videos, could give a fuller picture of sentiment online.Making models easier to interpret is another area. If you can see which words or tokens drive a prediction, people will trust the results more. Few-shot learning is catching attention, too. It lets people manage with fewer labeled examples for new tasks by using clever prompts. Speed matters as well. Lighter versions of models, made with distillation or quantization, help with real-time analysis when posts are flying in fast.

Finally, moving from finding correlations to pinpointing what actually causes a sentiment shift could change how people use these insights.

10. Conclusion

This paper reviewed sentiment analysis on social media using NLP, following the evolution from early lexicon-based methods and classic machine learning up to transformer models. In head-to-head comparisons, BERT-style models- especially BERTweet- lead the pack, reaching up to 95% accuracy on social media benchmarks. There's real impact here, from brand monitoring to mental health and crisis response.

Still, challenges like sarcasm, multilingual messiness, evolving language, and ethical concerns aren't going away anytime soon. If researchers want sentiment analysis to keep bridging the gap between human feelings and machines, these issues have to stay front and center. As language models get smarter and social media keeps growing, this field will keep finding new ways to connect what people say online to real understanding.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019.
[2] Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. LREC 2006.
[3] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Stanford Technical Report.
[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[5] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. ICWSM 2014.
[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. EMNLP 2014.
[7] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
[8] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.
[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ICLR Workshop 2013.
[10] Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English tweets. EMNLP 2020.
[11] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. EMNLP 2002.
[12] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP 2014.
[13] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. NeurIPS 2017