IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Sentiment Analysis on Social Media and E-Commerce Reviews using Supervised Learning Techniques

Srilakshmi Sriram
Data Engineer, Comcast, Chennai, India

Abstract: *Sentiment analysis is a field of Natural Language Processing (NLP) that focuses on identifying and extracting subjective information from textual data. With the rapid growth of online platforms, users are continuously generating large amounts of data in the form of reviews, tweets, and posts. This paper presents an integrated approach to sentiment analysis using Twitter and Amazon product review datasets. We applied lexicon-based methods for social media content and supervised learning techniques, particularly the Random Forest classifier, for e-commerce review data. Preprocessing steps such as tokenization, stemming, and stop-word removal were performed. Results show that the supervised model on Amazon reviews achieves high accuracy while lexicon-based Twitter sentiment analysis provides insight into real-time public opinion. Our findings suggest that while lexicon-based approaches are suitable for short, real-time content like tweets, supervised models provide superior accuracy for structured review data, making a combined strategy optimal for business intelligence. The study concludes by suggesting that combining both methods can improve business intelligence strategies..*

Keywords: Sentiment Analysis, NLP, Supervised Learning, Twitter API, Amazon Reviews, NLTK, Opinion Mining

I. INTRODUCTION

Sentiment analysis—also called opinion mining or emotion AI—is the application of natural language processing, text analytics, computational linguistics, and biometric techniques to systematically detect, extract, measure, and interpret emotional or subjective information from text. Sentiment analysis is extensively applied to voice of the client accoutrements similar as reviews and check responses, online and social media, and healthcare accoutrements for operations that range from marketing to client service to clinical drug. Being approaches to sentiment analysis can be grouped into three main orders knowledge- grounded ways, statistical styles, and cold-blooded approaches. Knowledge-grounded ways classify textbook by affect orders grounded on the presence of unequivocal affect words similar as happy, sad, hysterical , and wearied.

Some knowledge bases not only list egregious affect words, but also assign arbitrary words a probable" affinity" to particular feelings. Statistical styles influence rudiments from machine learning similar as idle semantic analysis, support vector machines," bag of words"," Point wise collective Information" for Semantic Orientation, and deep literacy. More sophisticated styles try to descry the holder of a sentiment and the target. Grammatical reliance relations are attained by deep parsing of the textbook. mongrel approaches influence both machine literacy and rudiments from knowledge representation similar as ontology and semantic networks in order to descry semantics that are expressed in a subtle manner.

A. CHARACTERISTICS OF SENTIMENT ANALYSIS

It is estimated that 80 of the world's data is unshaped, in other words it is unorganized. Huge quantities of textbook data(emails, support tickets, exchanges, social media exchanges, checks, papers, documents, etc.), is created every day but it is hard to dissect, understand, and sort through, not to mention time- consuming and precious. Sentiment analysis, still, helps businesses make sense of all this unshaped textbook by automatically tagging it.

DOI: 10.48175/IJARSCT-27668

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Sentiment analysis allows businesses to harness tremendous quantities of free data to understand client requirements and station towards their brand. Organizations examiner online exchanges to ameliorate products and services and maintain their character. The analysis takes client care to the coming position. client support systems with incorporated SA classify incoming queries by urgency, allowing workers to help the most demanding guests first. Sentiment analysis is an important tool for pool analytics.

The introductory sentiment analysis of textbook documents follows a straightforward process:

1. Break each textbook document down into its element corridor(rulings, expressions, commemoratives, and corridor of speech)

. Identify each sentiment- bearing expression and element.

3. Assign a sentiment score to each expression and element(- 1 to 1)

4. Optional Combine scores for multi-layered sentiment analysis.

B. COMPONENTS OF SENTIMENT ANALYSIS

Sentiment analysis involves context-aware text mining that extracts subjective insights from source material, enabling businesses to gauge public sentiment toward their brand, products, or services by monitoring online interactions.

Tokenization

Given a character sequence and a defined document unit, tokenization is the task of mincing it up into pieces, called commemoratives, at the same time throwing away certain characters, similar as punctuation.

Stemming

Stemming is principally removing the suffix from a word and reduce it to its root word. For illustration " Flying " is a word and its suffix is " ing ," if we remove " ing " from " Flying " also we will get base word or root word which is " Fly ." We use these suffixes to produce a new word from original stem word.

Stop Word Filtering

A stop word is a used word(similar as " the ," " a ," " an ," " in ") that a hunt machine has been programmed to ignore, both when indexing entries for searching and when reacquiring them as the result of a hunt query. Stop words can be easily removed by maintaining a predefined list of commonly used non-informative words. The NLTK library in Python provides built-in stop word lists for 16 different languages to facilitate this process.

POS Tagging

Part- of- speech trailing(POS trailing or POS tagging or POST), also called grammatical trailing or word- order disambiguation, is the process of marking up a word in a textbook(corpus) as corresponding to a particular part of speech,(1) grounded on both its description and its environment — i.e., its relationship with conterminous and affiliated words in an expression, judgment , or paragraph. A simplified form of this is tutored to academy- age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Sentiment Classification

Sentiment classification is the task of looking at a piece of textbook and telling if someone likes or dislikes the thing about which they are talking. The input X is a piece of textbook, and the affair Y is the sentiment which we want to prognosticate, similar as the star standing of a movie review.

C. TYPES OF SENTIMENT ANALYSIS

Firstly we need to understand the methods that social media vendors use to determine sentiment. There are many types of sentiment analysis. However, we will concentrate on three.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

A. Manual processing

Sentiment analysis conducted manually by humans is considered the most developed and reliable method for evaluating emotional tone. Despite its accuracy, it is not flawless. Due to the rapid expansion of social media content, this method has become less practical at scale. As a result, most organizations now pair manual efforts with automated tools to handle large volumes of data efficiently.

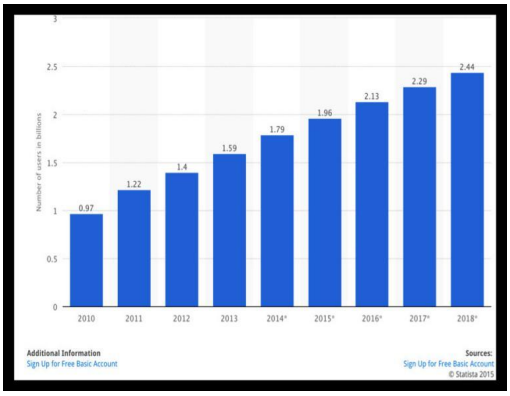

Fig 2.1 Number of social media users

B. Keyword processing

Keyword-based sentiment analysis involves evaluating individual words and assigning them sentiment values—either positive or negative. The cumulative sentiment of a text is then calculated based on these individual word scores. For instance, words like "great," "love," and "excellent" are typically rated positively, whereas words such as "dislike," "poor," and "terrible" receive negative scores.

This approach is known for its speed, simplicity, and low computational cost, making it easy to implement. However, it comes with several limitations. It struggles with context sensitivity, such as detecting sarcasm, understanding idiomatic expressions, or handling polysemous words like "sick," which can mean either "ill" or "amazing" depending on context. Additionally, inconsistencies can arise from varying sentiment scores assigned by different researchers. This method also lacks the sophistication to interpret phrases or non-adjective sentiment indicators accurately. Despite its limitations, many sentiment analysis systems—particularly in regions like Australia—still rely on keyword-based algorithms due to their efficiency.

C. Natural Language Processing

Natural Language Processing (NLP)—often referred to as text analytics, data mining, or computational linguistics—is the field of computer science that focuses on enabling machines to understand and interpret human language in a meaningful way. NLP recognizes that individual words form phrases, phrases form sentences, and sentences collectively express complex ideas. It works by analyzing the structure and semantics of language to extract meaning, with applications in areas such as speech-to-text conversion, language translation, and grammar correction.

NLP involves programming algorithms to understand human languages (like English) based on grammatical rules similar to those taught in school. While NLP offers a more advanced alternative to keyword-based processing, it is not without challenges. Detecting sarcasm, hyperbole, and informal language—including social media slang and acronyms (e.g., OMG, BFF, BTW)—remains a significant hurdle.

Some examples of modern social media slang include:

- Youturn: Following someone with the intent to unfollow them after they follow back (common on Twitter)
- Wallflower: A user who observes but rarely posts content
- Face Crawling: Requesting Facebook likes persistently

DOI: 10.48175/IJARSCT-27668

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

- Hash-Browning: Overusing hashtags in a single post
- Metapals: Online connections with whom one has never met in person.

Furthermore, users' express sentiment in nuanced ways—such as the difference in tone between "I'm fine!!!" and "I'm fine." Changing topics abruptly in a post also adds complexity, making accurate interpretation even more difficult for NLP systems.

II. LITERATURE REVIEW

Sentiment analysis has evolved significantly over the years, with applications across multiple domains such as marketing, politics, healthcare, and e-commerce. This section reviews foundational and contemporary research on sentiment analysis methods, datasets, tools, and applications in social media and e-commerce platforms.

A. Foundations of Sentiment Analysis

Sentiment analysis, also known as opinion mining, is rooted in natural language processing (NLP) and computational linguistics. The field initially focused on binary classification—positive or negative sentiment—but has since expanded to include neutral sentiment, emotion detection, and subjectivity classification [1]. Early sentiment analysis models relied heavily on lexicons and rule-based systems. Pang and Lee (2002) pioneered the use of machine learning in sentiment classification of movie reviews using Naïve Bayes and SVM models [2].

B. Opinion Mining and Subjectivity Detection

Opinion mining is broader than sentiment classification and includes identifying the opinion holder, target, polarity, and strength [3]. Subjectivity detection is a key preprocessing step in distinguishing between factual statements and opinionated text [4]. Turney's (2002) unsupervised semantic orientation approach using Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) set the stage for semantic-based analysis [5].

C. Supervised Learning in Sentiment Analysis

Supervised machine learning algorithms like Naïve Bayes, Support Vector Machines (SVM), Decision Trees, Logistic Regression, and Random Forests have been widely applied to sentiment classification tasks. More recently, ensemble techniques and deep learning methods (CNNs, LSTMs) have outperformed traditional classifiers in accuracy [6]. However, for domain-specific texts (like e-commerce reviews), Random Forest models often provide a balance of interpretability and performance [7].

D. Social Media Sentiment Analysis (Twitter)

Twitter has become one of the most studied platforms due to the availability of public data through APIs. Research on Twitter sentiment analysis has focused on real-time opinion tracking, political forecasting, crisis detection, and public health monitoring [8]. Challenges include tweet length limitations, slang, sarcasm, and misspellings. Studies have applied both lexicon-based methods (e.g., VADER, AFINN) and supervised classifiers (e.g., SVM, XGBoost) to Twitter datasets with varying success [9].

E. Sentiment Analysis in E-Commerce Reviews

E-commerce platforms like Amazon, Flipkart, and eBay collect massive amounts of user-generated content in the form of product reviews and ratings. These reviews serve as rich data sources for sentiment classification. Studies have shown that sentiment polarity in reviews can strongly influence purchase decisions and product rankings [10]. Supervised models trained on annotated product reviews (e.g., Amazon datasets) have achieved high accuracy, especially when combined with feature engineering and topic modeling techniques [11].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, June 2025 Impact Factor: 7.67

F. Lexicon-Based vs. Machine Learning Approaches

Lexicon-based approaches rely on predefined dictionaries of words associated with sentiment scores. While they perform well in domains with limited training data, they often struggle with domain-specific jargon, negation handling, and sarcasm [12]. Machine learning models, on the other hand, learn patterns from labeled data but require substantial preprocessing and training time. Several hybrid approaches have been proposed to leverage the interpretability of lexicons with the adaptability of machine learning models [13].

G. Tools and Datasets

Popular tools and libraries include NLTK, TextBlob, Scikit-learn, VADER (specifically for social media), and TensorFlow/Keras for deep learning implementations. Benchmark datasets include:

- Twitter Sentiment140 [14]
- Amazon Product Review Dataset [15]
- IMDB Movie Reviews
- Yelp Reviews

These datasets vary in structure, size, and labeling method (manual vs. star ratings), which impacts model performance and generalizability.

H. Recent Trends and Challenges

Recent research emphasizes the need for context-aware sentiment analysis, which accounts for user profiles, temporal dynamics, and cross-lingual sentiment [16]. Sarcasm detection, code-mixed language processing, and multimodal sentiment analysis (combining text, image, audio) are emerging areas of focus. Additionally, ethical considerations such as data privacy and algorithmic bias are increasingly being discussed in the literature [17].

III. METHODOLOGY

The primary objective of this study is to compare lexicon-based sentiment analysis (Twitter data using TextBlob) with supervised learning-based classification (Amazon product reviews using Random Forest). The methodology is divided into the following stages:

- **Data Acquisition**: Tweets are fetched using the Twitter API via Tweepy. Amazon reviews are obtained from publicly available datasets.
- **Preprocessing**: Both datasets are cleaned using standard NLP techniques—tokenization, stop-word removal, stemming, and POS tagging.

Sentiment Analysis:

- Twitter (Lexicon-Based): TextBlob is used to analyze sentiment polarity scores.
- Amazon Reviews (Supervised Learning): A Random Forest classifier is trained using pre-labeled data.
- **Evaluation**: Sentiment distribution is visualized, and classification accuracy, precision, recall, and F1-score are calculated for the Amazon dataset.

IV. DATA COLLECTION AND PREPROCESSING

A. Twitter Data Collection

Using Tweepy, tweets related to a specific keyword are fetched. Filters such as language (‘en’) and tweet count are applied.

Tweets were collected using the Twitter Developer API, filtered for English language and specific hashtags related to trending topics. Data was stored in JSON format and then parsed into CSV.

Sample Code Snippet:

```
import tweepy
from tweepy import OAuthHandler
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, June 2025 Impact Factor: 7.67

```
# API authentication (keys omitted for security)
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
tweets = api.search_tweets(q='mobile phone', count=1000, lang='en')
```

Sample Tweets:

```
sample_tweets = [
   "I love the new phone! It is amazing and works perfectly.",
   "This product is terrible. I am extremely disappointed.",
   "It's okay, not the best but not the worst either.",
   "Absolutely fantastic experience, will buy again!",
   "Worst customer service ever. Totally unacceptable.",
   "I'm feeling neutral about this event, nothing special.",
   "Great job by the team, very happy with the results!",
   "Not satisfied with the quality of the product.",
   "Meh, it's just average, nothing to write home about.",
   "Excellent features, exceeded my expectations."
]
```

B. Amazon Dataset

The dataset "Amazon Unlocked Mobile Reviews" was sourced from Kaggle. It contains features such as product name, brand, review text, star rating, and review votes.

C. Preprocessing Steps

Convert text to lowercase
Remove stop words
Perform stemming using NLTK's Porter Stemmer
Tokenize sentences and words
POS tagging for sentiment-bearing words

```
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

stop_words = set(stopwords.words('english'))
ps = PorterStemmer()

tokens = word_tokenize(review.lower())
filtered_tokens = [ps.stem(w) for w in tokens if w not in stop_words]
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

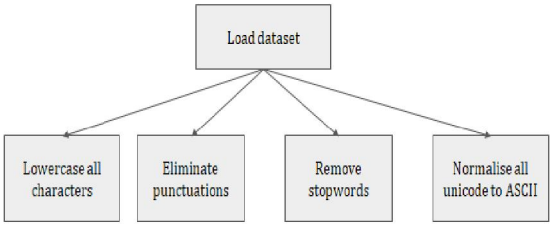

Figure 3.1 Data Cleansing

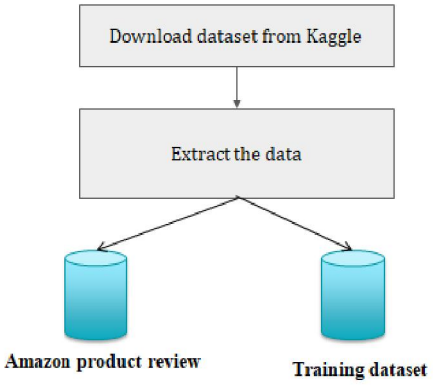

Figure 3.2 Data Extraction

	A	B	C	D	E	F	G	H	I	J	K	L
1	Product N	Brand	Price	Rating	Reviews	Review Votes						
2	"CLEAR CL	Samsung	199.99	5	I feel so LI	1						
3	"CLEAR CL	Samsung	199.99	4	nice phon	0						
4	"CLEAR CL	Samsung	199.99	5	Very plea	0						
5	"CLEAR CL	Samsung	199.99	4	It works g	0						
6	"CLEAR CL	Samsung	199.99	4	Great pho	0						
7	"CLEAR CL	Samsung	199.99	1	I already h	1						
8	"CLEAR CL	Samsung	199.99	2	The chargi	0						
9	"CLEAR CL	Samsung	199.99	2	Phone loc	0						
10	"CLEAR CL	Samsung	199.99	5	I originally	0						
11	"CLEAR CL	Samsung	199.99	3	It's batter	0						
12	"CLEAR CL	Samsung	199.99	3	My fiance	0						
13	"CLEAR CL	Samsung	199.99	5	This is a gr	0						
14	"CLEAR CL	Samsung	199.99	5	These guy	2						
15	"CLEAR CL	Samsung	199.99	1	I'm really	1						
16	"CLEAR CL	Samsung	199.99	5	Ordered t	1						
17	"CLEAR CL	Samsung	199.99	2	Had this p	0						
18	"CLEAR CL	Samsung	199.99	5	I was able	6						
19	"CLEAR CL	Samsung	199.99	5	I brought t	0						
20	"CLEAR CL	Samsung	199.99	4	I love the	1						
21	"CLEAR CL	Samsung	199.99	3	unfortuna	0						
22	"CLEAR CL	Samsung	199.99	4	The batter	0						
23	"CLEAR CL	Samsung	199.99	4	pros-beau	0						
24	"CLEAR CL	Samsung	199.99	1	I purchase	19						
25	"CLEAR CL	Samsung	199.99	4	Phone goe	0						
26	"CLEAR CL	Samsung	199.99	4	Phone's sp	0						
27	"CLEAR CL	Samsung	199.99	5	the phone	0						

Amazon_Unlocked_Mobile

Figure 3.3 Dataset

DOI: 10.48175/IJARSCT-27668

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

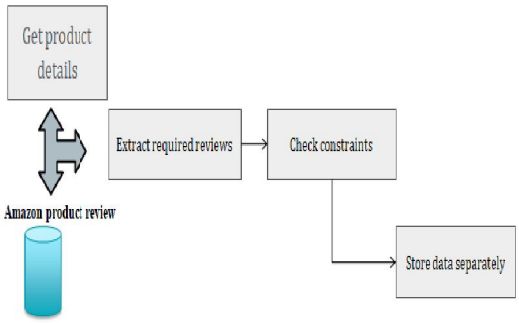

Figure 3.4 Input Processing for Amazon Reviews

V. RESULTS

A. Twitter Sentiment Distribution

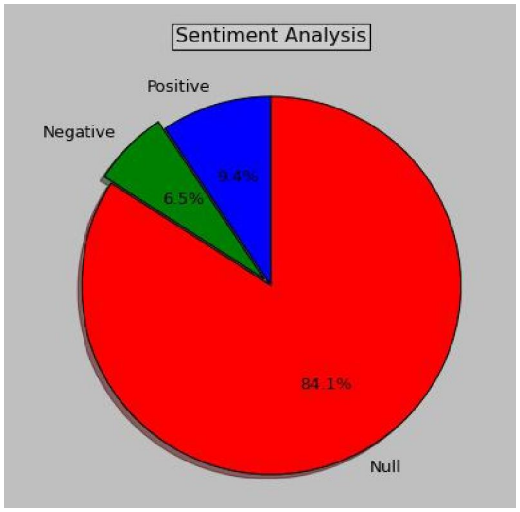

Fig. 5.1. Twitter Sentiment Pie Chart

Out of 1000 tweets:
Positive: 84%
Negative: 7%
Neutral: 9%

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

B. Amazon Sentiment Review for Different features of Mobiles:

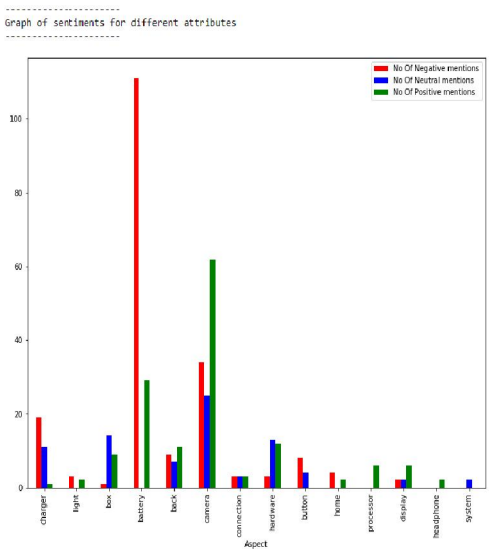

Fig. 5.2. Confusion Matrix for Amazon Sentiment Classification

VI. DISCUSSION

Lexicon-based methods are lightweight and suitable for real-time use, but they struggle with sarcasm and idioms. Supervised learning models, on the other hand, offer better performance when sufficient training data is available. Challenges faced include class imbalance, noisy data, and preprocessing of informal text. The Amazon dataset provided cleaner input, leading to better classifier performance. Twitter's short and informal format made it harder to analyze using rigid lexicons.

VII. CONCLUSION AND FUTURE WORK

This study combined lexicon-based and supervised learning approaches for sentiment analysis across two different platforms. The hybrid use case demonstrates how different methods can complement each other for stronger business intelligence. Future work includes expanding to multilingual sentiment detection, deploying models as REST APIs or Django web applications, and incorporating deep learning methods for better semantic understanding.

REFERENCES

[1]. M. Rambocas, and J. Gama, "Marketing Research: TheRoleof SentimentAnalysis". The 5th SNA-KDD Workshop'11. U nivers ityof Porto, 2013.

[2]. A K. Jose, N. Bhatia, and S. Krishna, "TwitterSentiment Analysis". National Institute of TechnologyCalicut,2010.

[3]. P. Lai, "ExtractingStrongSentimentTrendfromTwitter". Stanford University, 2012.

[4]. Y. Zhou, and Y. Fan, " A Sociolinguistic Study of American Slang," Theory and Practice in Language Studies, 3(12), 2209–2213, 2013. doi:10.4304/tpls.3.12.2209-2213

[5]. M. Comesaña, A. P.Soares, M.Perea, A.P. Piñeiro, I. Fraga, and A.Pinheiro, " Author ' s personal copy Computers in Human Behavior ERP correlates of masked affective priming with emoticons," Computers in Human Behavior, 29, 588–595, 2013.

[6]. A.H.Huang, D.C. Yen, & X. Zhang, "Exploring the effects of emoticons," Information & Management, 45(7), 466–473, 2008.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

[7]. D. Boyd, S. Golder, & G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," System Sciences (HICSS), 2010 …. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5428313

[8]. T. Carpenter, and T. Way, "Tracking Sentiment Analysis through Twitter,". ACM computer survey. Villanova:VillanovaUniversity, 2010.

[9]. D. Osimo, and F. Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis," Proceeding of the 12th conference of Fruct association, 2010, Un ited Kin gdom.

[10]. A Pak,and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer Application, France:Universitede Paris-Sud, 2010.

[11]. S.Lohmann, M. Burch, H. Schmauder and D. Weiskopf, "Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds," Annual conference of VISVISUS. Germany: University of Stuttgart, 2012.

[12]. H. Saif, Y.He, and H. Alani, "Semantic Sentiment Analysisof Twitter," Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute, 2011.

[13]. A Agarwal, B. Xie, I. Vovsha, O. Rambow, and R.Passonneau, "Sentiment Analysis of Twitter Data," Annual International Con fe renc es. New York:Columbia University, 2012.

[14]. J. Zhang, Y. Qu, J. Cody and Y. Wu, " A case study of Microblogging in the Enterprise: Use, Value, and Related Issues," Proceeding of the workshop on Web 2.0., 2010

AUTHOR'S PROFILE

Srilakshmi Sriram is a Data Engineer with a Bachelor's degree in Computer Science and Engineering, awarded in 2020. She began her professional journey at LTIMindtree, where she worked from October 2020 to April 2023. Since then, she has been serving as a Data Engineer II at Comcast. Currently, she is pursuing her Master of Science in Business Analytics (2024–2026), demonstrating a strong commitment to continuous learning and academic growth.

Her research interests include data mining, business intelligence, and criminology, reflecting a multidisciplinary approach to analytics and societal impact. She has published a paper on Generative AI in the International Journal for Research in Applied Science and Engineering Technology (IJRASET), Volume 12, Issue III, March 2024.

She is a recipient of the Award for Excellence(August 2022) and the Spotlight Award (H1 2024), recognizing her contributions to the field of data engineering. She is currently based in Chennai, India.