Exploring Sentiment Trends on Twitter: A Machine Learning Approach for Analyzing Public Opinion

Smt.Rekha Vijay Patil, Darshan Ravindra Patil, Vaibhav Rameshchandra Chavan

[1] *Lecturer, Department of information technology, Maharashtra, India*
[2] *Student, Department of information technology, Maharashtra, India*
[3] *Student, Department of information technology, Maharashtra, India*

ABSTRACT

With the exponential growth of web technology, the volume of data on the internet has reached unprecedented levels. The internet has evolved into a platform for online learning, idea exchange, and opinion sharing. Social networking sites such as Twitter, Facebook, and Google+ have gained immense popularity, enabling users to share views, engage in discussions, and post messages globally. This survey focuses on sentiment analysis of Twitter data, which is crucial for analyzing opinions expressed in tweets, known for their unstructured and heterogeneous nature. We provide an overview and comparative analysis of existing sentiment analysis techniques, including machine learning and lexicon-based approaches, along with evaluation metrics. We explore the use of machine learning algorithms such as Naive Bayes, Maximum Entropy, and Support Vector Machine for sentiment analysis of Twitter data streams. Additionally, we discuss the general challenges and applications of sentiment analysis on Twitter.

Keyword: *sentiment analysis; text classification; natural language processing; Twitter*

1. INTRODUCTION

The advent of the Internet has revolutionized the way people express their views and opinions, shifting towards online platforms such as blog posts, forums, product review sites, and social media. Social networking sites like Facebook, Twitter, and Google Plus have become integral to daily life, enabling millions to share emotions, opinions, and experiences. Online communities serve as interactive media where consumers not only inform but also influence others through forums and discussions. This digital era has resulted in a massive influx of sentiment-rich data in various forms, including tweets, status updates, blog posts, comments, and reviews. For businesses, social media offers a platform to connect with customers and advertise their products/services. User-generated content plays a pivotal role in decision-making processes, as individuals rely heavily on online reviews and social media discussions when considering a purchase or service. The sheer volume of user-generated content necessitates automation, leading to the widespread use of sentiment analysis techniques. Sentiment analysis (SA)tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about Textual information retrieval traditionally emphasizes processing, searching, or analyzing factual data, which inherently possesses an objective component. However, there exists another dimension of textual content that expresses subjective characteristics, including opinions, sentiments, appraisals, attitudes, and emotions. These subjective elements are at the core of Sentiment Analysis (SA), which presents numerous challenging opportunities for developing new applications. The proliferation of online information sources such as blogs and social networks has contributed significantly to the growth of available information, further highlighting the relevance of SA. For instance, SA can be leveraged in recommendation systems to predict item recommendations by considering factors such as positive or negative opinions about those items. Several research questions are raised as follows.

Abbreviation	Description
TSA	Twitter-based Sentiment Analysis
SNS	Social Networking Service
SA	Sentiment Analysis
OM	Opinion Mining
NLP	Natural Language Processing
NB	Naïve Bayes
SVM	Support Vector Machine
POS	Part of Speech
BN	Bayesian Network
ME	Maximum Entropy
DAG	Directed Acyclic Graph

Table 1. The description of abbreviation.

2. TWITTER

Various microblogging platforms like Twitter, Facebook, and Instagram were born out of the emergence of SNS. Twitter is a widely used SNS that allows users to exchange 140-character messages (referred to as “tweet”). More than 300 million people have signed up to use Twitter, which generates over 500 million updates each day. Because of the ease with which it can be shared, Twitter has grown to be one of the most important sources of user-generated data. The following is a list of the most important features of Twitter.

Tweet: A tweet is a 140-character maximum data unit that can be transmitted using Twitter. Its content ranges from how people feel or what they think about certain events, to photos, videos, and links, etc., all of which can be easily shared with the users’ contacts.

Handle: This refers to the behavior of tweet updating or public messaging to other users. It is written as “@username,” and the @ symbol is used to refer to the person or organization with whom the tweets are connected.

Hashtag: Hashtag is a kind of metadata tag used in various SNS that allows users to adopt dynamic, user-generated tags to make it easier for others to find the tweets related to a specific topic.

Follow: This is an activity of registered users to pursue people, companies, or any organization that they are interested in and to receive updated tweets in real time. Twitter is more than just a tool for staying in touch with friends and sharing one’s own daily activities, its true strength lies in the dissemination of information and the following of others.

Retweet: It is one of the most useful tools for disseminating information on Twitter, in which users are allowed to re-post the tweets they are interested in. Here, the original tweets generally remain unchanged, followed by the abbreviation of the original username of the authors.

Search: This powerful feature allows users to search keywords and phrases on Twitter to find updated tweets about their interests in real time. People are more likely to join Twitter because of this search function, which facilitates the discovery and dissemination of relevant content. Related your research work Introduction related your research work Introduction related your research work.

Table 2. One example of a tweet including user opinions.

Source	Username	Post
Twitter	BaskFan	I like watching basketball@NBA

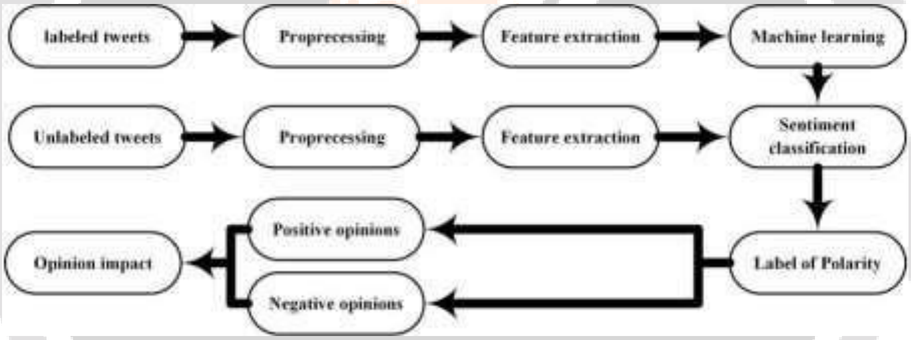
3. SENTIMENT ANALYSIS

Opinion mining, a subfield of linguistics and natural language processing, focuses on sentiment analysis, evaluating the polarity of words and phrases to extract views and feelings from textual data. Organizations and individuals often conduct studies to understand public sentiment on various issues.

Recent studies have expanded sentiment analysis into various applications. Liu et al. proposed a sentiment-based approach for sales pattern forecasting, while McGlohon et al.developed statistical and heuristic models to estimate product and merchant quality. Chen et al.used sentiment analysis to uncover hidden relationships in political contexts, developing novel opinion scoring models. Yano and Smith used statistical modeling to explore links between comment volume and political sentiment.

In the realm of social media, Twitter conversations have become a focal point for sentiment analysis. Deep learning approaches have been employed to analyze optimistic and pessimistic emotions expressed in Twitter conversations. Tamar Ginossar et al. studied the cross-platform spread of information through Twitter conversations, while Rabindra Lamsal et al.developed forecasting models for virus prevalence using Twitter conversation workload.

Sentiment analysis has found applications in business and social studies, with companies like Google and Microsoft developing their own sentiment analysis systems for industrial and commercial activities. However, sentiment analysis on Twitter poses challenges, primarily due to the constraint on message size. With tweets limited to 140 characters, extracting sentiment from such brief messages presents difficulties.



A sentiment analysis system typically receives data from various sources, including blogs, comments, reviews, etc., in different formats such as XML, HTML, and PDF. Techniques like tokenization, stemming, and stop-word removal are applied to standardize and transform the data from the corpus into text format for training datasets.

The selection of relevant features is a critical stage in sentiment analysis, as different combinations of features can significantly impact the final performance of sentiment analysis tasks. The polarity label of the test data is then determined using a text classifier trained and built using machine learning techniques.

4. REPRESENTATION OF FEATURES

Feature representation is a crucial preprocessing step in sentiment analysis, involving the transformation of text content into a feature vector. Common methods for expressing features in sentiment analysis include:

N-gram: Identifying a single feature in a text or speech corpus as a continuous sequence of n terms. Unigram refers to an n-gram of size one, and bigram refers to size two. The term frequency-based unigram is often used, where each word is considered a feature, and its occurrence frequency is the feature value.

Part of Speech (POS) tagging: Assigning a POS tag (verb, adverb, adjective, etc.) to each word in a text or corpus. The Penn Treebank POS tags are commonly used.

Table 3. Penn treebank POS tags.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Proposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund, or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNPS	Proper noun, plural	WDT	Wh-determiner

Negation: Negation is a crucial linguistic feature that significantly impacts the polarity of a sentence. The placement of negative words is critical in determining the scope of their influence. For example, in the statement "I like playing basketball but I am tired today," the negative term "tired" affects the overall sentiment because of its proximity to "basketball".

5. DIFFERENT LEVEL OFANALYSIS

Sentiment analysis classification can be categorized into three main types: document-level, sentence- level, and aspect-level.

5.1. Document-Level Sentiment Analysis

Document-level sentiment analysis involves classifying opinions as either negative or positive within the entire document. This approach treats the document's opinion as a single entity and employs supervised and unsupervised learning methods. Supervised learning divides documents into groups for specialized training datasets, while unsupervised learning measures semantic orientation to detect polarity.

5.2. Sentence-Level Sentiment Analysis

In sentence-level sentiment analysis, each sentence is assessed independently, and its overall tone is evaluated. The pre-judgment stage is crucial, as only subjective instances are further analyzed, while objective ones are typically discarded.

5.3. Aspect-Level Sentiment Analysis

Aspect-level sentiment analysis conducts a detailed analysis, examining the sentiments of individual components within the content. It involves three main steps: identification, categorization, and aggregation. Identification identifies relevant target pairs, categorization classifies their sentiments, and aggregation integrates sentiment values for a comprehensive perspective.

6. THE APPORACHES FOR TWITTER SENTIMENT ANYLISIS

The methodologies for sentiment analysis can be generally divided into three main taxonomy of sentiment analysis is shown in Figure 2.

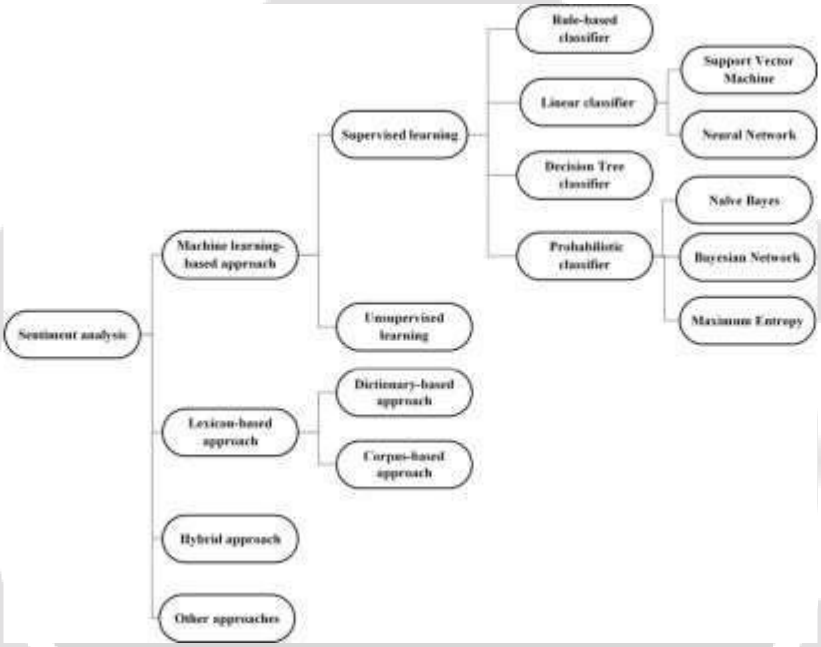


Figure 2 shows the taxonomy of sentiment analysis.

6.1. Machine Learning-Based Approach

The classification stage in sentiment analysis uses a classifier that is trained using machine-learning techniques. This approach can be broadly split into two types: supervised learning and unsupervised learning. The training dataset and linguistic characteristics are utilized for automatic text categorization in supervised learning, and primary supervised learning methodologies are outlined as follows.

6.1.1. Probabilistic Classifier

Mathematical models are utilized to predict categorization based on the input. Probabilistic classifiers such as the Naïve Bayes classifier (NB), Bayesian Network (BN), and Maximum Entropy classifier (ME) are commonly employed in data analysis. The NB classifier, based on Bayes' theorem, is widely used to determine the best class match. BN is another probabilistic model that uses Bayesian inference for probability calculation. Directed Acyclic Graphs (DAGs) are used to represent variables and their conditional interdependencies. ME computes the probability of a feature belonging to a specific category.

6.1.2. Linear Classifier

The linear classifier is generally used to determine the class to which a feature belongs. Classification decisions are based on linear predictor functions that combine feature values linearly. Support Vector Machine (SVM) and Neural Network (NN) are two widely used implementation methodologies.

6.1.3. Rule-Based Classifier

This classifier represents feature space information using a set of "IF-THEN" rules for classification. It classifies features into predefined classes based on these rules.

6.1.4. Decision Tree Classifier

This non-parametric approach to supervised learning continually partitions the feature space into sub- feature spaces for classification and regression. The goal is to use decision rules to predict the class label of the feature. While the supervised learning-based method is efficient for sentiment analysis, manually preparing labeled data for the classification system can be challenging. An unsupervised learning-based approach has been developed to solve this problem, identifying the degree of polarity using subjective indicators generated from the sentiment lexicon.

6.2. Lexicon-Based Approach

The lexicon-based method uses a sentiment lexicon to measure the strength of expressed feelings. A preset list of words is often used to create a sentiment lexicon. Dictionary-based and corpus-based methods are the two most common techniques for building a sentiment lexicon. The dictionary-based technique uses lexicographical information, such as a dictionary, to define sentiment words, while the corpus-based method typically uses co-occurrence scenarios with already established sentiment terms [69]. Table 5 lists publications that use the lexicon-based sentiment analysis approach.

Table 5. The list of lexicon-based approaches proposed for sentiment analysis.

Objective and Algorithm Used	Data Scope	Dataset
Classification of text using fine-grained attitude labels, semantic, lexicon created by own	User-generated personal story	Dataset from Experience Project website
Lexicon-based approach, document discourse structure, sentiment classifier, semantic, lexicon created by own	Movie review	IMDB
Lexicon-based comments-oriented news sentiment analyzer, NLP, PMI-IR, taxonomy lexicon	News information	N/A
Comparative analysis of emotion detection, supervised and lexical knowledge-based approach, SVM	Corpus of emotions	ISEAR, Emotinet

6.3. Hybrid Approach

The hybrid approach combines machine-learning and lexicon-based methods. Research has demonstrated that this approach can enhance classification performance. Table 6 summarizes publications that have employed the hybrid approach.

Table 6. The list of hybrid-based approaches proposed for sentiment analysis.

Objective and Algorithm Used	Data Scope	Dataset
Neural-network-based hybrid approach, sentiment classifier	Blogger comments and product reviews	Datasets collected from LiveJournal, Review Centre
Comparative study of ensemble technique for sentiment analysis, NB, SVM, maximum entropy	Movie review, product review	Cornell movie-review corpora
A system for subjectivity and sentiment analysis (SSA), manually created polarity lexicon	Chat messages, Arabic tweets	multi-domain sentiment dataset from Amazon
Rule-based multivariate feature selection, linear kernel SVM	Online review	DAR, TGRD, THR, MONT
Hybrid method combining rule-based classification and	Movie review, product review, and MySpace	Epinions, Edmunds,

6.4. Other Approaches

Some methods in TSA literature do not fit neatly into the previously mentioned categories and can be classified as "graph-based approaches." These approaches aim to construct a connected social graph for effective label propagation, assuming mutual influence among individuals. Speriosu et al. initially developed such approaches for TSA, using various objects like tweets, hashtags, and unigrams as nodes in the graph. Cui et al. introduced another label propagation method based on extracting and analyzing emotion tokens. More recently, Cambria et al.presented a graph-based technique that performs reasoning tasks by developing a morphology-aware concept parser. However, constructing the social graph is time-consuming, and its availability depends greatly on the diversity of the corpus. This area of study requires further investigation.

7. DISCUSSION

The machine-learning-based approach to TSA is the most popular, where conventional machine learning algorithms are trained using a subset of available features to predict the sentiment polarity of a given text. Combining multiple classifiers often yields better results than using individual ones. However, this approach has limitations. The size of the training dataset significantly affects classification performance, as most machine-learning algorithms require a large number of manually annotated tweets. Despite efforts like distant supervision to generate annotated tweets at scale, low-quality annotation can hinder TSA efficiency. Domain dependence is another limitation, as prediction accuracy is highly reliant on classifiers trained with domain-specific data.

Lexicon-based approaches, using sentiment lexicons, categorize TSA tasks without requiring annotated tweets. However, these approaches can struggle with words not in the lexicon and lack context independence, ignoring the relationship between sentiment and word context. Hybrid approaches have been proposed to overcome these limitations, offering superior performance in specific domains but at a higher computational cost.

8. CONCLUSION

In recent years, there has been a growing interest among researchers in analyzing tweets based on the sentiments they convey. This interest stems from the vast number of tweets posted on Twitter, which provides valuable insights into public sentiment on various subjects. This survey aims to introduce the fundamental concepts and techniques for sentiment analysis of tweets, and more than 60 publications were reviewed and categorized to showcase the latest developments in the field. Additionally, studying sentiment analysis through recent TSA applications can be beneficial. It is expected that TSA will continue to rapidly evolve as a research field in the coming years, with more studies on TSA anticipated in the future.

9. REFERENCES

[1]A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2]R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[3]Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision".
Stanford University, Technical Paper,2009

[4]L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013,