Applied Computer Science 2025, vol. 21, no. 2, pp. 128–138
https://doi.org/10.35784/acs_6936

Submitted: 2024-12-11 | Revised: 2025-03-01 | Accepted: 2025-05-11

Keywords: artificial intelligence, sentiment analysis, natural language processing

Agnieszka WOJDECKA[1], *Jakub GROMADZIŃSKI* [1], *Krzysztof WALCZAK* [1*]
[1] Poznań University of Economics and Business, Poland, agnieszka.wojdecka@wp.pl, jakub.gromadzinski@ue.poznan.pl, krzysztof.walczak@ue.poznan.pl
[*] Corresponding author: krzysztof.walczak@ue.poznan.pl

Effectiveness of large language models and software libraries in sentiment analysis

Abstract

This paper investigates the effectiveness of selected tools for sentiment analysis, focusing on both dedicated software libraries (NLTK, Pattern, TextBlob) and large language models (ChatGPT and Gemini). The evaluation was conducted in two stages: sentiment analysis of 30 synthetic opinions of varying linguistic complexity, and analysis of 5 sets of real user reviews collected from the web. The results show that large language models - although not explicitly designed for sentiment analysis - achieved the highest accuracy, with ChatGPT consistently producing the lowest deviation from human ratings. In contrast, software libraries showed greater variation, especially in the presence of complex linguistic structures. These findings highlight the potential of large language models in sentiment analysis tasks and underscore their robustness in interpreting nuanced language.

1. INTRODUCTION

Natural language is the primary means of human communication and is characterized by its complexity and constant evolution. While its rules are logical and intuitive to humans, they often pose significant challenges for computers to interpret. Natural language processing (NLP) systems use advanced algorithms to identify these rules and generate appropriate responses, facilitating communication between humans and machines. This capability enables a range of applications, including automated text classification, opinion mining, customer feedback analysis, and large-scale information extraction. Although the field of NLP is developing rapidly (Abro et al., 2023), its progress is accompanied by persistent challenges arising from the intricacies of human language (Khurana et al., 2023).

Sentiment analysis (Hussein, 2018), a key component of natural language processing, has become essential for understanding the emotions, opinions, and attitudes conveyed in online content. Among the wide range of web-based materials, discussion posts, product reviews, and social media comments are particularly valuable sources of insight into public sentiment on various topics. These texts-often short and informal, with domain-specific vocabulary and irregular syntax-pose significant challenges for sentiment analysis. While it is possible to analyze such content manually, it is time consuming and labor intensive. In this context, automatic sentiment analysis is not only desirable, but increasingly necessary.

The core task of sentiment analysis is to classify text into categories of positive, negative, or neutral sentiment (Xu et al., 2019; Nandwani & Verma, 2021). Recent advances in machine learning, particularly in natural language processing, have led to the development of tools and systems that can automatically assign sentiment categories to opinions. As a result, sentiment analysis can now be efficiently applied to large volumes of data from multiple sources.

This paper evaluates the effectiveness of selected sentiment analysis tools, including three software libraries (NLTK, Pattern, and TextBlob) and two large language models (LLMs): ChatGPT and Gemini. The tools were tested in a two-stage evaluation: first on synthetic opinions designed with different levels of linguistic complexity, and then on real user reviews collected from the Amazon platform. Expert human reviews served as a reference baseline. The results provide a comparative analysis across different linguistic contexts, showing that large language models - although not specifically designed for sentiment analysis - achieve higher accuracy and greater consistency than the libraries, particularly when handling complex or nuanced language.

The rest of this paper is organized as follows. Section 2 presents an overview of sentiment analysis, including common approaches, levels of analysis, and linguistic characteristics of user opinions. Section 3 introduces the selected sentiment analysis tools and describes the two-stage study design. Section 4 reports the evaluation results based on synthetic and real opinion data. Section 5 discusses the results, focusing on the comparative performance of the tools. Finally, Section 6 concludes the paper and outlines directions for future research.

2. SENTIMENT ANALYSIS OF USER OPINIONS

Sentiment analysis is a branch of natural language processing that focuses on identifying and interpreting the feelings, emotions, and attitudes expressed in textual opinions about products, services, or issues. These insights can support decision-making by both consumers and organizations. Sentiment analysis involves the classification of subjective information, typically into categories such as positive, negative, and neutral (Kaur & Bhatia, 2016). This classification is accomplished using lexicon-based (dictionary) approaches, machine learning methods, or hybrid techniques. Machine learning algorithms are often preferred for basic sentiment classification because of their ability to learn from training data and produce accurate results. Dictionary-based methods, on the other hand, are better suited for broader, domain-independent analyses. They are also valued for their simplicity and computational efficiency, as they do not require complex model structures (Medhat et al., 2014).

2.1. Approaches to sentiment analysis

The goal of sentiment analysis is to identify the tone, mood, and emotion expressed in opinions. There are three main approaches to this task. The lexicon-based approach relies on predefined sets of words or text corpora, using lexical and syntactic rules to assess sentiment (Bonta et al., 2019; Taboada et al., 2011). The machine learning approach applies supervised or unsupervised learning techniques to automatically classify sentiment based on patterns learned from the data. Finally, the hybrid approach combines elements of both lexicon-based and machine learning methods, often using lexicon-based features as input for training sentiment classifiers to improve classification performance.

Lexicon-based methods perform sentiment analysis at the sentence or statement level using predefined word lists and syntactic rules, without requiring training data. They provide a lightweight and interpretable approach to sentiment classification. These methods assign a polarity score to each token-the smallest linguistic unit into which text can be divided, such as a single word or punctuation mark-based on its emotional connotation. Scores typically range from -1 to 1, with -1 indicating negative sentiment, 1 indicating positive sentiment, and 0 indicating neutrality. The closer the score is to either extreme, the stronger the sentiment expressed. Polarity scores are typically assigned separately for positive, negative, and neutral sentiment, and an overall polarity score is then calculated by aggregating the individual token scores.

While this approach is efficient and easy to implement, it has several limitations. One major drawback is the inability to account for contextual ambiguity - words that can change meaning depending on usage. For example, the word "long" may have a negative connotation in the sentence "There was a long line at the checkout counter," but a positive connotation in the sentence "Cathy has beautiful, long hair. To overcome this limitation, lexicons can be adapted to specific domains, or entirely new domain-specific lexicons can be developed (Wankhade et al., 2022).

Among lexicon-based techniques, dictionary and corpus-based methods are the most common. The dictionary-based approach begins by compiling a list of words labeled with specific sentiment polarities (positive or negative). This list is then expanded using synonyms and antonyms from online dictionaries, in an iterative process that continues until no new relevant words are found. In the final step, each word is assigned a sentiment score, often as a numerical value. Manual verification and error correction at this stage can significantly improve model performance.

While dictionary-based methods allow for the rapid construction of sentiment lexicons, they often suffer from limited context sensitivity and domain specificity. For this reason, they are often complemented by more advanced text analysis techniques. Continued research in this area has led to a wide range of algorithms for identifying text polarity, contributing to the ongoing development of sentiment analysis tools (Liu, 2012).

Another widely used approach to sentiment analysis relies on machine learning algorithms that automatically classify text polarity based on patterns identified in the data. These algorithms do not require predefined word lists; instead, they learn from labeled examples to generalize sentiment classifications. Common machine learning classifiers used for sentiment analysis include decision trees, random forests, support vector machines, naive Bayes, and k-nearest neighbors. These models vary in complexity, interpretability, and suitability depending on the dataset and application context (Umarani et al., 2021; Rish, 2001; Peretz et al., 2024).

Within the machine learning approach, a distinction is made between supervised and unsupervised learning methods. In supervised learning, the model is trained on a dataset containing both input features and corresponding output labels. The algorithm learns to map inputs to outputs by minimizing the error between its predictions and the known target values. When discrepancies occur, the model adjusts its internal parameters to improve accuracy. This approach is commonly used in applications where historical labeled data is available to support the prediction of future outcomes (Chachal & Gulia, 2019).

2.2. Levels of sentiment analysis

Text sentiment analysis can be performed at different levels. Dividing the analysis into multiple levels provides a more comprehensive understanding of the text, from the overall context of the document to more detailed analysis at the sentence, phrase, and aspect levels. The four primary levels of sentiment analysis are *Document Level*, *Sentence* Level, *Phrase Level*, and *Aspect* (*Feature*) Level.

Document-level analysis involves assessing the sentiment of an entire document and assigning it a single sentiment classification. Both supervised and unsupervised machine learning techniques can be applied at this level. However, this approach is not widely used due to its relatively low accuracy, especially in texts containing mixed sentiments.

Sentence-level analysis examines individual sentences and assigns sentiment classifications to each. This approach is useful for texts that convey mixed emotions, allowing the aggregated data to reflect the overall sentiment of the document. Phrase-level analysis provides a more fine-grained approach by identifying sentiment in specific phrases within sentences, each of which may refer to different characteristics or aspects.

The most granular method is aspect-level analysis, which focuses on individual components or attributes within phrases (e.g., price, battery life, customer service). This level allows for highly accurate sentiment assessment by identifying opinions directed at specific product features (Wankhade et al., 2022).

2.3. Sentiment analysis of opinions

Opinions play a critical role in both business and everyday decision making. They enable companies to understand customer perceptions of their products and services, while also helping individual consumers make informed purchasing decisions based on the experiences of others. With the proliferation of social media and online review platforms, companies are increasingly relying on user-generated content as a source of insight, reducing the need for costly traditional surveys and market research (Liu, 2012). Sentiment analysis facilitates this process by automating the identification and interpretation of emotional tone in textual data.

However, the variety of ways in which opinions are expressed presents significant challenges to accurate sentiment classification. Opinions can be categorized as either regular or comparative. Regular opinions express feelings about a single object, whereas comparative opinions evaluate two or more objects based on shared attributes, typically indicating a preference for one of them (Liu, 2012). Examples of both types are shown in Table 1.

Tab. 1. Examples of regular and comparative opinions

Regular opinion	Comparative opinion
The new iPhone model has excellent photo quality.	The new iPhone model has better photo quality than the previous model.
This TV is cheap.	This TV is cheaper than others.

In practice, however, many opinions do not fall neatly into either category. Some convey sentiment through factual statements that carry emotional implications, while others use emotionally expressive language without clearly referring to a specific object. These ambiguous or context-dependent expressions present additional challenges for classification. For example, some statements may lack explicit positive or negative language, but still convey sentiment through the facts they present (Table 2, left). Others may contain emotionally charged language without expressing a clear evaluative stance toward an identifiable object (Table 2, right). In such cases, interpretation of sentiment often depends on the broader linguistic context or additional background information about the author.

Tab. 2. Examples of ambiguous or context-dependent opinions

Factual but emotionally suggestive	Emotionally expressive but semantically vague
This car uses a lot of fuel.	This product surprises me.
After a few days of use, the shoes fit well.	I feel affected by this book.

Another important characteristic of opinions is that they can express either rational or emotional evaluations. A rational review is based on logical reasoning, often focused on functional or usability aspects, and typically avoids emotional language. In contrast, an emotional evaluation is rooted in the author's personal feelings or state of mind, and may convey strong emotions even when not directed at a specific aspect of the object-or sometimes without a clear evaluative purpose at all (Liu, 2012). The distinction between rational and emotional evaluations is illustrated in Table 3.

Tab. 3. Examples of rational and emotional opinions

Rational opinion	Emotional opinion
This food processor has numerous functions.	I love my new food processor!
This washing machine uses a lot of water and is not energy efficient.	This washing machine is hopeless — it uses insane amounts of water and electricity!

In summary, the linguistic structure and emotional framing of an opinion play a critical role in how it is interpreted by sentiment analysis tools. The way an opinion is expressed-whether it is direct, comparative, emotional, or ambiguous-can have a significant impact on the accuracy of sentiment classification.

3. SENTIMENT ANALYSIS TOOLS AND STUDY DESIGN

3.1. Description of selected tools

A wide range of sentiment analysis tools has emerged in recent years, reflecting advances in natural language processing and the growing demand for automated opinion analysis. These tools differ in their underlying architecture, level of technical expertise required, and typical applications. For the purposes of this study, they are broadly categorized into three groups: *Business platforms*, *software libraries*, and *large language models* (Raiaan et al., 2024).

The first group of tools is primarily used for business purposes, especially by marketing companies that want to analyze customer sentiment towards the brand being marketed. These are easy-to-use platforms with pre-defined functionality and intuitive interfaces. They enable companies to monitor the emotional tone of brand mentions on social media and other websites in real time, often presenting the results in the form of clear, easy-to-interpret charts. An example of such a tool is Brand24, which continuously collects data and provides comprehensive sentiment analysis of online content.

The second set of tools consists of Python-based software libraries commonly used by programmers and researchers for text mining and sentiment analysis. These tools require more technical expertise than commercial platforms, but they provide flexibility and scalability for custom applications. In this study, we used three such libraries:

- *NLTK* (Natural Language Toolkit), which includes the VADER sentiment analyzer, designed for rule-based sentiment scoring of short text.
- *Pattern*, a lightweight library that provides built-in functions for polarity estimation based on lexical heuristics.
- *TextBlob*, a wrapper that combines features from NLTK and Pattern and outputs both polarity and subjectivity scores.

All three libraries use polarity scores ranging from -1 (negative) to 1 (positive), although their internal methods differ.

The third group are large language models. Although these models are not specifically trained for sentiment analysis, they are capable of performing it as part of broader natural language understanding tasks. Sentiment analysis using LLMs typically requires a carefully crafted instruction or prompt that includes all the relevant elements for the analysis. Models such as ChatGPT and Gemini have been adapted for general language tasks and can be effectively applied to sentiment classification. Their capabilities stem from being trained on large corpora of text, which allows them to understand sentence context and the meaning of individual words, even when words are polysemous. They also incorporate knowledge of grammar, syntax, and semantics, enabling them to accurately identify emotionally charged language and interpret complex linguistic phenomena such as sarcasm, irony, and subtle sentiment cues.

The goal of the study was to evaluate the effectiveness of selected tools for performing sentiment analysis on textual opinions. Tools from the second and third categories - software libraries and large language models - were included in the analysis. The following tools were chosen: NLTK (version 3.8.1), Pattern (version 3.6), TextBlob (version 0.18.0.post0), ChatGPT (GPT-3.5), and Gemini (version 1.0 Pro).

3.2. Study methodology

The study was conducted in two distinct phases to provide a more nuanced and structured evaluation of sentiment analysis methods. Stage I focused on a set of 30 *synthetic opinions-divided*into simple and complex linguistic forms-that were first rated by human respondents and then analyzed using selected software tools. Stage II involved the evaluation of five *real-world opinion sets* related to consumer products collected from an online platform. In both stages, the sentiment analysis results generated by the tools were compared with human judgments to assess consistency and accuracy. The complete evaluation workflow is shown in Figure 1.

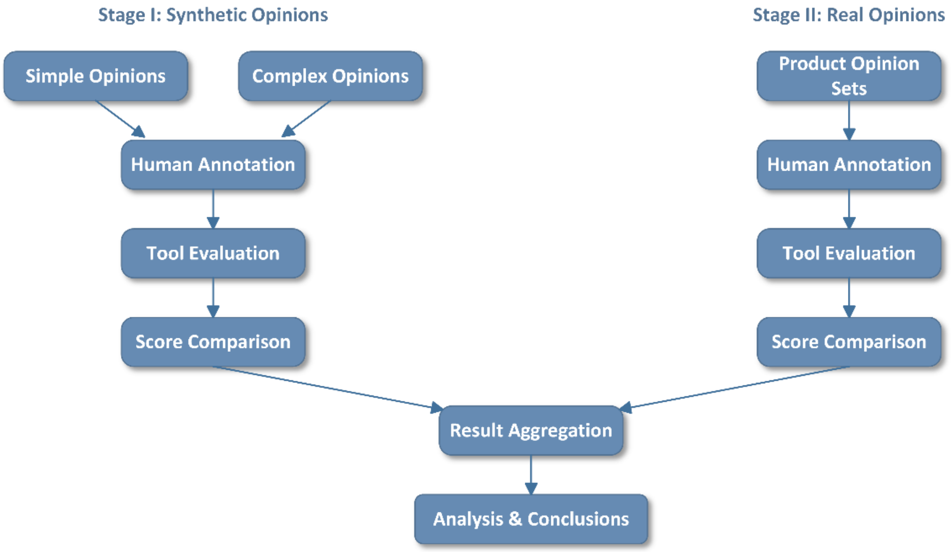

Fig. 1. Workflow of the sentiment analysis evaluation process

Each software tool assigned a sentiment score to individual opinions on a continuous scale from -1 to 1, with negative values indicating negative sentiment, positive values indicating positive sentiment, and values near zero representing neutrality. A group of respondents manually scored the same opinions by assigning each one to a predefined sentiment category, with each category assigned a fixed numerical score. These human-assigned scores served as benchmarks for evaluating the tool's performance. The sentiment categories and their corresponding scores are summarized in Table 4. The average human-assigned scores are shown in the Survey column of the results tables. The polarity ranges shown in the rightmost column are for color-coding purposes only within these tables.

Tab. 4. Sentiment categories and corresponding score ranges

Category	Human-Assigned Score	Polarity Range (for visualization)
Extremely negative	–1.0	[–1.0, –0.6)
Negative	–0.5	[–0.6, –0.2)
Neutral	0.0	[–0.2, 0.2]
Positive	0.5	(0.2, 0.6]
Extremely positive	1.0	(0.6, 1.0]

Effectiveness was assessed by calculating the mean difference between the scores of each tool and the corresponding mean scores of the survey, which served as expected values.

In the first phase of the study, the tools were used to assess sentiment on a synthetic dataset consisting of thirty opinions: ten simple opinions without complex linguistic structures and twenty containing various complex linguistic elements. The same set of opinions was also included in a survey and rated by a group of twenty randomly selected participants with a good command of English. To improve the readability of the results, each score was classified into a sentiment category based on predefined polarity ranges and then visually marked with color labels. The results of this evaluation are shown in Table 5.

Tab. 5. Sentiment scores assigned by selected tools and human evaluators for synthetic opinions (Stage I)

Opinion	Complexity	NLTK	Pattern	TextBlob	ChatGPT	Gemini	Survey
I love this product!	-	0.67	0.63	0.63	1.00	1.00	0.95
Didn't like the fact it was not protected well in the box.	-	-0.18	0.00	0.00	-0.60	-0.50	-0.48
It works how it should, everything is fine.	-	0.20	0.42	0.42	0.70	0.50	0.40
Nothing special, just OK.	-	-0.57	0.43	0.43	0.30	0.00	0.03
I'm actually very disappointed by this product due to its quality.	-	-0.53	-0.55	0.55	-0.90	-1.00	-0.83
I didn't have high expectations because of the low price, so I'm definitely not disappointed, but also not very satisfied.	-	-0.27	0.09	0.09	0.00	0.00	0.00
Do not buy it, it's garbage...	-	0.00	0.00	0.00	-1.00	-1.00	-0.95
For me it was the best purchase of this year.	-	0.64	1.00	1.00	1.00	1.00	0.95
I totally recommend this restaurant!	-	0.47	0.00	0.00	1.00	1.00	0.98
Works without problems.	-	0.31	0.00	0.00	0.70	0.70	0.35
Purchasing this 'gadget' is a great idea for those who have nothing to do with their money...	sarcasm	0.62	0.80	0.80	-0.70	-0.80	-0.80
The shoes are super comfortable, but they are of a lower quality compared to the previous model.	comparative	0.20	0.19	0.19	0.20	0.30	0.20
I wouldn't say that the product isn't worth its price, in fact, to me it's not bad.	multiple negation	0.29	0.10	0.32	0.40	0.00	0.40
The food was fine, but their water tasted like water from the well... also, it was rather cool so we had to sit with jackets on.	ambiguity	0.77	0.38	0.38	-0.20	-0.30	-0.40
Poor customer service.	tone	-0.48	-0.40	-0.40	-0.80	-1.00	-0.50
POOR CUSTOMER SERVICE!!!	tone	-0.61	-0.78	-0.78	-0.90	-1.00	-1.00
Imo this product is LIT.	slang	0.00	0.00	0.00	0.90	1.00	0.80
The packaging looked fantastic, according to the manufacturer it has many useful and practical features, but it turned out to be junk.	positive words in criticism	0.50	0.40	0.40	-0.80	-1.00	-0.65

Tab. 5. Sentiment scores assigned by selected tools and human evaluators for synthetic opinions (Stage I), continued

Opinion	Complexity	NLTK	Pattern	TextBlob	ChatGPT	Gemini	Survey
For this price I thought that the quality would be the bee's knees and it turned out to be a white elephant.	idioms	0.00	0.00	0.00	-0.60	-0.80	-0.45
I recommend this company for those who love spending hours on hold with customer service wasting their day!	sarcasm	0.65	0.63	0.63	-0.90	-1.00	-0.93
Firstly I was disappointed by the customer service, but the hotel room was surprisingly better than I expected and the breakfast was great.	syntactic ambiguity	0.91	0.14	0.14	0.20	0.50	0.45
As a lot of people say, it's not entirely true that it didn't impress, actually it did, and not in a negative way.	multiple negation	-0.61	-0.16	-0.16	0.60	0.70	0.45
New album is fire!	slang	-0.40	0.17	0.17	0.90	1.00	0.93
Who came up with something like this? 😍	emoticons	0.36	1.00	0.00	0.80	1.00	0.55
Who came up with something like this? 🤦	emoticons	0.36	0.00	0.00	-0.80	-1.00	-0.53
The food was OK, but the place itself needs some improvements.	subtle criticism	0.60	0.50	0.50	0.10	0.00	0.05
Previous model had more buttons than this one, which was better in my opinion.	comparative	0.44	0.28	0.28	-0.30	-0.30	-0.28
The scarf is pretty long. 👍	emoticons	0.49	0.23	0.10	0.50	0.70	0.48
The scarf is pretty long. 👎	emoticons	0.49	0.10	0.10	-0.50	-0.30	-0.50
Very knowledgeable vets with passion, unfortunately my cat was diagnosed with hyperthyroidism and needs to be treated with radioactive therapy.	domain terminology	0.15	-0.15	-0.15	0.10	-0.50	0.40

It can be observed that in many cases the results produced by different tools vary significantly and often differ from expert judgments. The most consistent results with human judgments were achieved by large language models - ChatGPT and Gemini. In contrast, the software libraries showed greater deviations from the survey results, especially when processing opinions with complex linguistic features.

The second phase of the study involved a general sentiment analysis of real-world opinion sets related to selected products. The opinions were obtained from the online shopping platform Amazon. Five different products from the electronics category were selected, and five user opinions were collected for each product. The selected products and the corresponding opinions are listed in Table 6.

Tab. 6. Opinions about selected products from the Amazon platform (original spelling retained)

Product	Opinion text
Apple AirPods	*The Apple AirPods are amazing and worth the price. The sound is crystal clear and the battery life is beyond believable. I have only charged them once since they arrived and I use them all the time.*
	Glad I broke down and bought a pair! Really love them and the fit of the ear is very comfortable. Like the quick connection and stable battery life. Worth the price!
	As always, my new pair of Apple AirPods do not disappoint. The only drawback is that I keep losing one of them.
	My old airpods broke so i saved up and bought a new pair definitely worth it i dont like the newer airpods cuz the pad things. These last long charge under 10 min and yeah def buy.
	My old airpods lost battery really fast, these lasted hours. They are much smaller and fit perfectly in my ears. The sound quality is amazing as well as the microphone quality.

Tab. 6. Opinions about selected products from the Amazon platform (original spelling retained), continued

Product	Opinion text
Datacolor Spyder Print printer	*I had high expectations of this device given the price, and because I also own the Spyder5 PRO from Datacolor, which despite having a glitchy room-light switching feature, has always worked perfectly for calibrating my screens. So, you could say I trusted the Datacolor brand. Unfortunately, this SpyderPRINT product is garbage.*
	The Printer is capable of producing quite good ICC printer profiles. However, it requires an enormous amount of patience, double-checking, re-checking, and do-overs. What's most frustrating is that it all comes down to the poor guide and housing materials.
	Used the device for about a year on an off with no issues. Worked well. Results were good. Then one day the device stopped working. Got in touch with customer support and their response was to basically try unplugging and re-plugging. If you do buy this, buy the insurance and demand a full replacement when the thing fails. Very disappointing. Will never buy a product from this company again.
	The cradle and the spectro of my Spyderprint 3 are made of a gummy, sticky plastic that is a real dirt magnet. After four years of use, mine is so disgusting that I wear cotton gloves when handling it.
	Of very poor quality and on top of that they do not allow you to return the product, or ask for a refund, terrible service from Amazon.
PlayStation 5	*The PlayStation 5 is a true masterpiece in the world of gaming, delivering an unparalleled experience that sets a new standard for next-gen consoles. From its sleek design to its powerhouse performance, the PS5 is a triumph in every aspect.*
	The PS5 just about perfectly nails it. Very, very impressed with this hardware! The bundled robot game is impressive, fun, and an excellent showcase for the system.
	The PS5 has quickly become a favorite of mine, I LOVE this thing, the DualSense controller feels very comfortable to hold with the one issue being it's atrocious battery life, everything works as intended and comes pre-installed with a fantastic game so there's something to play right away, would recommend for the whole family!
	Love it; was a little worried at first that I'd experience what others had and receive it somehow damaged or generally not working, but it arrived fairly quick and was in tip top shape when we took it out of the box! It has been a super useful console so far!
	I love my PS5. Glad I bought it. I don't really know what to write as a review. If you want one, you know what to expect and already have decided or else you wouldn't be looking at one. If you have no idea what this is and are buying it for someone else, just buy it. Whoever you're buying it for won't be disappointed.
Secret Mini Spy Camera	*Absolutely garbage, they can't hold a charge. They are made out of cheap plastic. I got three of them it was a waste of time and money. Just go with the ring cameras.*
	Over all if your looking for a easy to hide camera, these are alright. You can't hide them in stuffys or anything, but they are small and can fit in places nobody would ordinarily be looking. I gave it a 3 star rating because it's a good camera, but the battery life thing is what got it that 3 stars. Battery life is very important when it comes to having a wireless camera.
	Great way to monitor what's happening in your home or business. An affordable option that can be handy both day and night.
	For us, it was very difficult install, the picture was not clear, and basically, we did not end up using it at all. It was very disappointing.
	I got this for my daughter so she could use it as a nanny cam for my granddaughter, and she's told me it works perfectly. It's very simple to mount, and the adhesive on the mount is quite strong. The video quality is very clear, and it has night vision, as well as motion detection. I highly recommend this, especially for such a great price.
Smart Band	*The battery didn't stay charged and the connector broke. The watch kept disconnecting from phone.*
	It doesnt work at all!
	Good product for the money.
	This is cheap totally useless device. The bpm will differ as much as 20 to 30 bpm from reality.
	And do not get me started about the ridiculous claim that it can measure blood pressure.

These opinion sets were scored for emotional tone using the same software tools as in the first step, along with the scores of a group of thirty respondents. For each product, an average sentiment score was calculated based on the scores assigned by respondents to all related opinions (Survey column). To improve readability, both the tool results and the survey results are visually color-coded using the same polarity ranges as in the first stage. The results of the second phase of the study are presented in Table 7.

Tab. 7. Sentiment scores and classification of real product opinion sets (Stage II)

Product	NLTK	Pattern	TextBlob	ChatGPT	Gemini	Survey
Apple AirPods	0.6	0.2	0.2	0.9	0.6	0.8
Datacolor Spyder Print printer	-0.2	0.0	0.0	-0.9	-0.8	-0.8
PlayStation 5	0.9	0.3	0.3	1.0	0.9	1.0
Secret Mini Spy Camera	0.2	0.2	0.2	0.0	0.2	0.0
Smart Band	0.0	0.2	0.2	-0.5	-0.4	-0.3

As can be seen in Table 7, the large language models again produced results that were very close to the expert judgments, and in the vast majority of cases agreed with them. In contrast, the software libraries showed greater variability and more significant deviations from the expected values.

4. EVALUATION OF TOOL EFFECTIVENESS

Based on the results of both phases of the study, the effectiveness of each tool was evaluated by calculating the difference between its output and the corresponding survey-based reference value for each opinion or set of opinions. Specifically, the average absolute difference between the tool-generated scores and the survey scores was calculated for each tool. This metric provides a numerical measure of how closely a tool's output matches human judgment. Values range from 0 to 2, with 0 representing perfect agreement with the reference and 2 representing complete divergence. Thus, lower average differences correspond to higher efficacy, while higher values indicate lower agreement with human ratings.

The results were grouped into three categories to assess tool performance across different types of input:

- *Low linguistic complexity* - includes simple synthetic opinions without complex linguistic structures.
- *High linguistic complexity* - includes synthetic opinions with complex or nuanced language.
- *Real opinion sets* - includes real-world sets of user opinions with collective sentiment scores.

The results of the tool effectiveness evaluation are presented in Table 8.

Tab. 8. Comparison of tool effectiveness (lower values indicate higher accuracy)

Category	Average difference between result and the expected value				
	NLTK	Pattern	TextBlob	ChatGPT	Gemini
Low linguistic complexity	0.38	0.39	0.50	0.13	0.08
High linguistic complexity	0.68	0.60	0.60	0.12	0.24
Real opinion sets	0.28	0.56	0.56	0.08	0.12
Weighted overall score	**0.44**	**0.54**	**0.56**	**0.10**	**0.16**

Figure 2 provides a graphical representation of the results from Table 8.

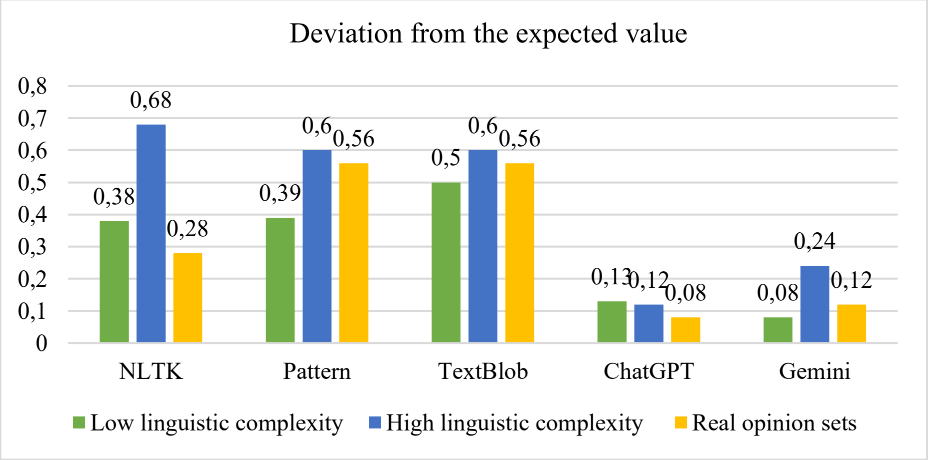

Fig. 2. Comparison of tool effectiveness depending on complexity (lower values indicate higher accuracy)

An overview of the aggregated results across all categories is shown in Figure 3.

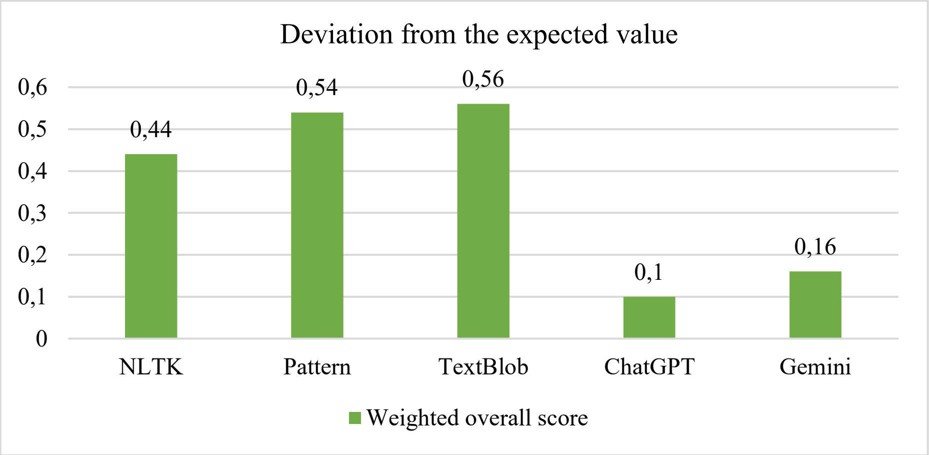

Fig. 3. Comparison of tool effectiveness overall (lower values indicate higher accuracy)

5. DISCUSSION

Analysis of the results shows clear differences in the performance of the tools in different linguistic contexts. In the low linguistic complexity category, the most accurate sentiment scores were produced by the large language models: Gemini achieved the lowest deviation (0.08), closely followed by ChatGPT (0.13). In contrast, the software libraries showed larger deviations, with NLTK at 0.38, Pattern at 0.39, and TextBlob slightly higher at 0.50.

For opinions involving complex linguistic structures, the LLMs again outperformed the traditional tools. ChatGPT maintained a low deviation of 0.12, and Gemini followed with 0.24, confirming their robustness in handling nuanced or ambiguous expressions. The software libraries, on the other hand, struggled in this scenario: both Pattern and TextBlob showed deviations of 0.60, while NLTK had the highest error at 0.68.

In the case of real-world opinion sets, the pattern persisted. ChatGPT again achieved the lowest deviation (0.08), and Gemini stayed close at 0.12. Among the libraries, NLTK performed relatively better in this category (0.28), though still significantly less accurate than the LLMs. Pattern and TextBlob again showed higher deviations of 0.56 each.

As summarized in Figure 3, the most effective tool overall was ChatGPT, with a weighted average deviation of 0.10 across all categories. Gemini followed with a deviation of 0.16, confirming the high reliability of LLMs for sentiment analysis tasks. In contrast, the software libraries were less consistent: NLTK scored 0.44, while Pattern and TextBlob had higher deviations of 0.54 and 0.56, respectively.

When analyzing individual categories, the largest discrepancies were observed in the context of complex linguistic structures, although this was mainly observed among the software libraries. This reinforces the conclusion that large language models are more effective at interpreting context-dependent and nuanced language. While LLMs already perform well in all categories tested, software libraries may still benefit from refinement or domain-specific adaptation to improve their performance in sentiment analysis tasks.

6. CONCLUSIONS

This study evaluated the effectiveness of selected sentiment analysis tools, including three software libraries (NLTK, Pattern, TextBlob) and two large language models (ChatGPT and Gemini). The evaluation was conducted in two phases: one with synthetic opinions of varying linguistic complexity, and another with real product reviews.

The key observation is that large language models consistently outperformed software libraries, especially in cases involving complex linguistic structures and subtle sentiment cues. Both ChatGPT and Gemini showed strong agreement with human judgments across all categories, with ChatGPT achieving the highest overall accuracy and Gemini performing best on opinions with low linguistic complexity. These results highlight the robustness and context sensitivity of LLMs in sentiment analysis.

However, the study has some limitations. The evaluation focused exclusively on English-language texts and included a relatively small number of synthetic and real opinions. In addition, the large language models were used as black-box systems, without specific sentiment analysis training or access to their underlying system instructions. The prompt wording may also have influenced the results.

Future work will include expanding the dataset, including non-English languages, and investigating fine-tuned or domain-adapted LLMs for sentiment analysis. Another direction is to explore hybrid approaches that combine lexicon-based methods with LLM results, and to investigate issues related to model explainability and bias in sentiment interpretation. Given the rapid evolution of large language models, it will also be important to continuously monitor and evaluate newly released versions to understand their behavior and assess their suitability for sentiment analysis tasks.

Funding

Supported by funds granted by the Minister of Science of the Republic of Poland under the „Regional Initiative for Excellence" Programme for the implementation of the project "The Poznań University of Economics and Business for Economy 5.0: Regional Initiative – Global Effects (RIGE)".

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

Abro, A. A., Talpur, M. S. H., & Jumani, A. K. (2023). Natural language processing challenges and issues: A literature review. *Gazi University Journal of Science*, *36*(4), 1522-1536. https://doi.org/10.35378/gujs.1032517

Bonta, V., Kumaresh, N., & Naulegari, J. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2),1-6. https://doi.org/10.51983/ajcst-2019.8.S2.2037

Chachal, A., & Gulia, P. (2019). Machine Learning and Deep Learning. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 2278-3075. http://dx.doi.org/10.35940/ijitee.L3550.1081219

Hussein, D. M. E. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, *30*(4), 330-338. https://doi.org/10.1016/j.jksues.2016.04.002

Kaur, F., & Bhatia, R. (2016). Sentiment analyzing by dictionary based approach. *International Journal of Computer Applications*, *152*(5), 32-34. https://doi.org/10.5120/ijca2016911814

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*, 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Liu, B. (2012) *Sentiment Analysis and Opinion Mining.* Morgan & Claypool.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*, 81. https://doi.org/10.1007/s13278-021-00776-6

Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier – An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*, *136*(Part B), 108972. https://doi.org/10.1016/j.engappai.2024.108972

Raiaan, M. A. K, Mukta, S. H., Fatema, K., Fahad, N. M., Sakib, S., & Mim, M. M. J. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839-26874. https://doi.org/10.1109/ACCESS.2024.3365742

Rish, I. (2001). An empirical study of the naïve bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 41-46.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Umarani, V., Juliana, A., & Deepa, J. (2021). Sentiment analysis using various machine learning and deep learning techniques. *Journal of the Nigerian Society of Phisical Sciences*, *3*(4), 385-394. https://doi.org/10.46481/jnsps.2021.308

Wankhade, M., Rao, A. C. S., & Kulkarni C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., & Wu, X. (2019). Chinese text sentiment analysis based on extended sentiment dictionary. *IEEE Access*, *7*, 43749-43762. https://doi.org/10.1109/ACCESS.2019.2907772