Technical Review of IBM Watson Natural Language Understanding API

Sonam Jain
Text Information Systems
University of Illinois Urbana-Champaign, USA
sonyj9@illinois.edu

Sai Aitha
Text Information Systems
University of Illinois Urbana-Champaign, USA
sa60@illinois.edu

ACM Reference Format:
Sonam Jain and Sai Aitha. 2024. Technical Review of IBM Watson Natural Language Understanding API. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 SUMMARY

IBM Watson Natural Language Understanding is an API that offers a variety of tools related to natural language processing (NLP). These tools assist users in analyzing unstructured text data to gain insight on different attributes of the text. This technical review evaluates the API in terms of its main functionalities: sentiment analysis, entity extraction, keyword tagging, speed and scalability and multilingual support. We will assess and compare its ease of integration and performance with other prominent NLP tools. By testing the API with real-world datasets and applications, we hope to make sense of its strengths and limitations.

2 INTRODUCTION

Natural Language Processing (NLP) transforms text-based information systems by enabling machines to understand, analyze, and respond to human language. It powers tools like chatbots, search engines, and content summarization systems, making them more intuitive and efficient. It also plays an important role in helping organizations analyze large amounts of unstructured text data. Tools like the IBM Watson API offer cloud-based NLP solutions.

This review seeks to answer key questions:

- How effective is the API in terms of common NLP tasks such as sentiment analysis and entity extraction?
- How fast and scalable is it?
- How well does it support other languages?
- How does IBM Watson compare to competing APIs like Google Cloud NLP and AWS Comprehend?

3 DESCRIPTION

The IBM Watson NLU API offers several core features:

- **Sentiment Analysis**: Classifies text as positive, negative, or neutral.
- **Entity Extraction**: Identifies named entities (e.g., people, places, and organizations) and categorizes them into semantic classes.
- **Keyword and Concept Tagging**: Highlights keywords and extracts abstract concepts from text.
- **Custom Models**: Provides users with options to train models tailored to specific domains.
- **Multilingual Support**: Handles over 13 languages, including English, Spanish, and Chinese, for a global user base.

The API is available through RESTful endpoints and offers official SDKs for Python, Node.js, and Java. Users authenticate via IBM Cloud API keys, ensuring secure and efficient access.

Our methodology involves hands-on testing with real-world datasets to evaluate NLP tasks such as sentiment analysis and entity extraction. We conducted the same tests using IBM Watson alongside Google Cloud NLP and AWS Comprehend to evaluate how IBM Watson's accuracy and speed compare with its competitors.

To evaluate sentiment analysis, a balanced dataset of 500 reviews was created using predefined positive, neutral, and negative phrases. Each review was assigned a ground truth label (positive, neutral, or negative) based on the sentiment conveyed by the phrase. Predictions from each tool were compared with ground truth labels. Accuracy was calculated as the percentage of correct predictions.

For entity extraction, we evaluated the accuracy of each tool in identifying entities such as organizations (ORG) and geopolitical entities (GPE) from a dataset of news articles. Ground truth entities were generated using the spaCy NLP library. We extracted organization (ORG) and geopolitical (GPE) entities from the same dataset of news articles. We compared the entities extracted by each tool against the ground truth. To ensure consistency, text normalization was applied to both extracted entities and ground truth data. The accuracy of each tool was calculated as the percentage of correctly identified entities over the total ground truth entities.

Our methodology for evaluating keyword and concept tagging involves leveraging real-world datasets of tweets to test the performance of IBM Watson NLU, Google Cloud NLP, and AWS Comprehend. The evaluation focuses on how accurately each tool can extract meaningful keywords or concepts from a set of 50 tweets. Using spaCy's en-core-web-sm model, ground truth entities were generated by extracting organizations (ORG) and geopolitical entities (GPE) from the tweets. These entities served as the baseline for evaluating keyword extraction accuracy. The extracted keywords were compared against the ground truth entities prepared using spaCy. Accuracy was calculated as the percentage of correctly extracted keywords relative to the total ground truth keywords across all tweets.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We evaluated the multilingual support by performing sentiment analysis across multiple languages (English, Spanish, and Chinese). The evaluation measures each tool's accuracy in classifying sentiments as positive, negative, or neutral. A multilingual dataset of sentences was prepared, including 34 sentences each in English (en), Spanish (es), and Chinese (zh). Each sentence was manually labeled with a ground truth sentiment (positive, negative, neutral). For each tool, the predicted sentiments were compared to the ground truth. Accuracy was computed as the percentage of correct predictions.

To measure speed and scalability, we measured Single Request Response Time by measuring the time taken by each tool to process a single NLP request. We measured Parallel Request Response Time by simulating high concurrency with multiple parallel requests to evaluate scalability. We calculated the average response time across all requests.

4 EVALUATION

4.1 Performance

To comprehensively evaluate the performance of the IBM Watson NLU API, we tested it alongside Google Cloud Natural Language API and AWS Comprehend. The tools were benchmarked using various datasets representing real-world scenarios such as customer reviews, news articles, and tweets. These tests covered sentiment analysis, entity extraction, keyword and concept tagging, speed and scalability, and multilingual support. The results are summarized below.

1. Sentiment Analysis Sentiment analysis was evaluated using a simulated dataset of 500 customer reviews from an e-commerce platform:

- **IBM Watson NLU:** Achieved an accuracy of 68.53%.
- **Google Cloud NLP:** Achieved an accuracy of 94.62%.
- **AWS Comprehend:** Achieved an accuracy of 71.71%.

Summary: Google Cloud NLP demonstrated superior sentiment detection, followed by AWS Comprehend. IBM Watson lagged in accuracy.

2. Entity Extraction Entity extraction was tested using 50 news articles

- **IBM Watson NLU:** Correctly identified 43.71% of entities.
- **Google Cloud NLP:** Correctly identified 66.34% of entities.
- **AWS Comprehend:** Correctly identified 70.65% of entities.

Summary: AWS Comprehend had the highest accuracy for entity extraction, followed by Google Cloud NLP, with IBM Watson NLU performing much less effectively.

3. Keyword and Concept Tagging This feature was evaluated using 50 tweets to test how well each tool could extract meaningful keywords or phrases:

- **IBM Watson NLU:** Extracted accurate and relevant keywords 50% of the time.
- **Google Cloud NLP:** Extracted relevant keywords with 59.52% accuracy.
- **AWS Comprehend:** Extracted keywords with 73.81% accuracy.

Summary: IBM Watson demonstrated the lowest accuracy in keyword tagging, followed by Google Cloud NLP, with AWS Comprehend leading.

4. Speed and Scalability Each tool was evaluated for response time and scalability under load:

- **IBM Watson NLU:** Single request response time was 1.09 seconds and parallel average response time was .04 seconds.
- **Google Cloud NLP:** Single request response time was 0.3 seconds and parallel average response time was .04 seconds.
- **AWS Comprehend:** Single request response time was 0.35 seconds and parallel average response time was .02 seconds.

Summary: Google Cloud NLP had the fastest single request response times, followed by AWS Comprehend, with IBM Watson being slower. IBM Watson and Google Cloud NLP had similar parallel average response times while AWS Comprehend had a parallel average response time that was halved.

5. Multilingual Support Multilingual support was tested with a dataset of 100 sentences in English, Spanish, and Chinese:

- **IBM Watson NLU:** Supported all tested languages and provided an average accuracy of 70% across them.
- **Google Cloud NLP:** Supported all tested languages and provided an average accuracy of 80% across them.
- **AWS Comprehend:** Supported all tested languages and provided an average accuracy of 74% across them.

Summary: IBM Watson demonstrated the weakest multilingual support. Google Cloud NLP was the strongest in multilingual support.

Overall Performance Assessment

- **Best Overall Performance:** Google Cloud NLP, due to its high accuracy in sentiment analysis, robust multilingual support, and speed in handling individual requests.
- **Best for Entity and Keyword Extraction:** AWS Comprehend, which excelled in extracting meaningful entities and keywords and performed well under heavy parallel loads.
- **Worst Overall Performance:** IBM Watson NLU, while a viable option, needs enhancements in accuracy across tasks and better handling of nuanced and complex data.

4.2 Ease of Integration

- **Setting up:** We began our work in a clean virtual environment, free of pre-installed libraries or tools. To use each tool, obtaining API keys was a necessary first step. Based on our experience, Google Cloud stood out for its exceptional documentation, making it straightforward to acquire and use the required credentials. IBM Watson followed closely, with API keys readily accessible from the settings page after account creation. However, AWS credentials proved the most challenging to set up, as they required extensive configuration of the account before gaining access to the necessary keys.
- **Cost Analysis:**
  - **IBM Watson Natural Language Understanding:** IBM Watson provides a tiered pricing model. The free tier allows for a certain number of API calls per month, with limitations on features and usage. Paid tiers offer increased API call limits and access to advanced features, with costs scaling based on usage.
  - **Google Cloud Natural Language API:** Google charges based on the number of text units processed, where one text unit equals 100 Unicode characters. The pricing varies

depending on the specific features. Google offers a free tier providing 5,000 text units per month.

- **AWS Comprehend:** AWS uses a pay-as-you-go model, charging per unit processed. Each operation has a specific cost per unit. AWS offers a free tier for Comprehend, which includes 50,000 units per month for the first 12 months.

5 DISCUSSION

The IBM Watson Natural Language Understanding (NLU) API shows a mixed performance across various natural language processing tasks. While it provides reliable multilingual support and reasonable developer usability, its overall accuracy in tasks like sentiment analysis, entity extraction, and keyword tagging lags behind competitors such as Google Cloud NLP and AWS Comprehend.

5.1 Strengths of IBM Watson NLU

1. Multilingual Support: IBM Watson NLU demonstrated consistent multilingual support, accurately processing English, Spanish, and Chinese with an average accuracy of 70While this is the weakest among the three tools evaluated, it still shows IBM Watson's ability to handle diverse languages, making it a good option for applications requiring global language coverage.

2. Developer Usability: The integration process for IBM Watson NLU was straightforward and well-documented. There is a large amount of documentation for popular programming languages such as Python, Java, and Node.js so developers can quickly integrate the API into their applications. The error-handling mechanisms were especially helpful because they provided detailed feedback on invalid inputs, so it was very easy to debug when testing. In comparison, AWS Comprehend offers less detailed guidance in its documentation, but Google provides similar levels of documentation.

5.2 Weaknesses

1. Sentiment Analysis: IBM Watson's sentiment analysis accuracy significantly trailed Google Cloud NLP and AWS Comprehend. It struggled with nuanced or mixed sentiments, often misclassifying examples like "It's fine, but I've seen better." as positive due to an overemphasis on certain keywords. This limitation impacts its reliability in analyzing customer feedback or reviews with complex sentiments.

2. Entity Extraction: IBM Watson performed the weakest in entity extraction, failing to identify entities with the precision and consistency demonstrated by AWS Comprehend and Google Cloud NLP. It failed to identify more than half the entities, making it a mostly unreliable tool for this task.

3. Keyword and Concept Tagging: IBM Watson's keyword extraction accuracy was the lowest among the tools evaluated. This weakness reflects its difficulty in extracting meaningful and relevant phrases from content, limiting its effectiveness in tasks like content summarization or search engine optimization.

5.3 Opportunities for Improvement

1. Enhanced Sentiment Analysis: IBM Watson could benefit from improvements in handling nuanced and mixed sentiments. This might involve refining its sentiment scoring algorithm to better account for context and balancing positive and negative tones in mixed statements.

2. Improved Entity Extraction: Use advanced context-based models to enhance the accuracy of entity recognition in ambiguous expressions.

3. Keyword and Concept Extraction: Focusing on refining keyword extraction accuracy would improve its relevance for applications like content tagging and search optimization.

5.4 Practical Implications

While the IBM NLP tool offers valuable features for certain use cases, its limitations in accuracy across various tasks affect its applicability in competitive, high-demand scenarios compared to its competitors.

1. Moderate Scalability IBM Watson demonstrates strong scalability in parallel processing, with an average response time of 0.04 seconds for concurrent requests. This makes it a viable option for applications needing real-time insights. On the other hand, IBM Watson's slower single-request response time (1.09 seconds) compared to Google Cloud NLP (0.3 seconds) and AWS Comprehend (0.35 seconds) may hinder its performance in time-sensitive applications.

2. Integration with IBM Ecosystem IBM Watson integrates seamlessly with other IBM tools like Watson Assistant and Watson Discovery, providing an ecosystem for enterprise applications. While beneficial for existing IBM clients, this ecosystem would not be ideal for organizations using other cloud platforms like AWS or Google Cloud, especially when considering IBM Watson's weaker performance in standalone evaluations

3. Cost Implications IBM Watson's cost structure is typically usage-based, making it accessible for small-scale projects. However, for applications requiring high precision or handling large datasets, its lower accuracy across key tasks may lead to inefficiencies and increased costs, as additional processing or manual validation might be needed.

6 CONCLUSION

In conclusion, our evaluation of the IBM Watson Natural Language Understanding (NLU) API underscores significant challenges in its performance when compared to competitors like Google Cloud NLP and AWS Comprehend. Across critical tasks such as sentiment analysis, entity extraction, and keyword tagging, IBM Watson NLU consistently ranked last, achieving the lowest accuracy in each category. This performance gap highlights the need for substantial improvements if Watson is to compete effectively in scenarios requiring high precision and nuanced language handling.

Despite these shortcomings, IBM Watson NLU offers some notable advantages. Its ease of integration stands out as a significant strength, offering a straightforward interface and seamless compatibility across various programming environments. This makes it an attractive choice for developers seeking quick and convenient implementation without extensive setup or configuration. Additionally, IBM Watson NLU's cost structure is relatively competitive, making it a viable option for organizations looking for a budget-friendly NLP solution.

However, the broader implications of this analysis suggest that IBM Watson NLU may not currently be the optimal choice for organizations that require top-tier performance in NLP tasks. The results also underscore the rapid pace of innovation in NLP technology, where leaders like Google and AWS continue to refine their offerings to maintain their status of being trailblazers in the field of technology. For IBM Watson to remain competitive, targeted improvements in accuracy and nuanced language handling are essential. If these enhancements are implemented, IBM Watson could provide a more balanced offering, combining ease of use with stronger performance metrics.

In its current form, IBM Watson NLU is best suited for projects that value ease of use and rapid deployment over absolute precision. Its practical integration features make it a viable option for developers with simpler use cases, but its competitive standing will depend on addressing the substantial performance gaps identified in this review.

—

References

- https://cloud.ibm.com/docs/assistant?topic=assistant-api-overview
- https://cloud.google.com/natural-language/docs
- https://docs.aws.amazon.com/comprehend/latest/APIReference/welcome.html
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval.
- https://www.datacamp.com/blog/what-is-named-entity-recognition-ner
- https://towardsdatascience.com/keyword-extraction-process-in-python-with-natural-language-processing-nlp-d769a9069d5c?gi=b0a878a48332
- https://serdarcelebi.medium.com/simple-sentiment-analysis-using-aws-comprehend-and-visualization-with-aws-quicksight-services-38861c9ca676