Infegy Linguistics
Testing Sentiment Performance

Introduction

93% accuracy. Jumping to the chase, Infegy Linguistics rated nearly two million consumer reviews with phenomenal accuracy. Read on to find out why that number alone doesn't tell the whole story, and learn how Infegy Linguistics performed overall.

Measurement of sentiment and tonality is an important part of modern language analysis. While many perform this type of analysis, few do it well. Within Infegy Linguistics, powering Infegy's social intelligence platform Infegy Atlas, sentiment analysis has been taken to the next level, with significant advances in accuracy and understanding of complex language.

What Makes Sentiment Analysis Difficult?

Human language is elaborate, with nearly infinite grammatical variations, misspellings, slang and other challenges making accurate automated analysis of natural language quite difficult. Traditional approaches to sentiment analysis are surprisingly simple in design, struggling with complicated language structures, and fail when contextual information is required to correctly interpret a phrase. Consider, for example, the following two sentences:

"I like that Corvette."

"That looks like a Corvette."

When analyzing sentiment, the first example would optimally be scored as positive, with the second marked neutral. However, the vast majority of systems will not mark these examples correctly, as the word expressing positivity in the first sentence, "like", is not expressing tone in the second. The best way for a system to correctly interpret this complexity is to understand the context around the word's usage. Consider this secondary example:

"I'm craving McDonald's so bad."

Again, most systems will mis-interpret this statement, seeing the word "bad", or even the phrase "so bad", and score the sentence as negative. Contextual understanding is critical for a system to be able to reach human-level accuracy.

infegy

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116

How Infegy Linguistics Operates

Infegy's approach to sentiment analysis works much more closely to typical human interpretation of text. While parsing, the system breaks down each sentence, identifying the subject, any modifiers and grammatical structure for each. Rather than just scoring words as positive or negative, this system is able to understand the context of each word's usage, enabling dramatic improvements in performance.

In the two Corvette examples above, the word "like" can be interpreted in two completely different ways with this system, understanding when it is or isn't expressing tonality. With our McDonald's example, the system understand's the phrase "so bad", in this context, is actually a modifier to the earlier verb "craving", and as such, is able to correctly score this sentence as strongly positive.

Testing Sentiment Accuracy

When validating a sentiment analysis system, the testing methodology is crucial. The data source, cleanliness of language, how it is scored, subject matter and volume of data tested are all significant variables that can dramatically affect results. For an optimal test, the data source should closely match the intended uses. For example, if your intended application is analysis of online dialog, the data used to test system accuracy should also be sourced from online dialog. Likewise, the data should closely match in its cleanliness, such as grammatical structure, frequency of words being mis-spelled and usage of slang. Volume of data tested is also important, and a general rule of thumb here is "the more the merrier".

The test that was constructed to substantiate Infegy Linguistics' accuracy figures was the largest test undertaken in the industry, and was specifically designed to ensure accuracy is validated across as many subject matter as is practical. For the benchmark, nearly two million (1,912,958) rated user reviews from a major online retailer were collected. Ensuring a wide spread of subject matter, the reviews cover 75 categories, from gourmet food to video games to jewelry, each having 11,294 to 38,961 reviews, with an average of 23,617. The reviews are well-distributed in content, some very short, some very long, with the average review containing 110 words. Spelling and grammar are also well-distributed, with some reviews having very good structure, and others being very poor, with no punctuation and many misspellings.

What to Measure

With the content for the testing taken care of, it is important to consider what measures of performance should be tracked. Unfortunately, many in the industry are focused on one single metric: precision, often referred to as accuracy. While certainly important, this measure alone does not tell us the whole story. Another metric, known as recall, is equally important to the understanding of how these systems perform. The key metrics are described below.

Precision/Accuracy: A measure of how often a sentiment rating was correct. For documents with tonality, accuracy tracks how many of those that were rated to have tonality were rated correctly.

Recall: A measure of how many documents with sentiment were rated as sentimental. This could be seen as how accurately the system determines neutrality. Generally, high recall scores are very difficult in tests of broad subject matter, as the system is required to understand ever-larger sets of words and language.

F1 Score: Also called F-Score or F-Measure, this is a combination of precision and recall. The score is in a range of 0.0 - 1.0, where 1.0 would be perfect. The F1 Score is very helpful, as it gives us a single metric that rates a system by both precision and recall. As such, it is commonly used amongst experts and researchers in the linguistics and natural language processing fields to simply describe the performance of such systems. The formula for calculating F1 Score is:

$$F1 = 2 * (precision * recall) / (precision + recall)$$

Why Recall Matters

For an example of the importance of recall, consider we have 100 documents discussing a bank. Of these documents, 10 are neutral, making statements such as, "I just went to the bank." 40 of them are positive comments about the bank, and the last 50 are all negative comments specifically mentioning fraud.

Now imagine we were to analyze this dataset with a system which does not understand fraud as being negative. It may correctly score all 40 positive comments, and mark the 50 fraud comments and 10 neutral comments as neutral. In this case, of the 40 comments the system rated, it got all 40 correct, so it would have a theoretical accuracy of 100%. However, it didn't rate any of the 50 comments on fraud. So of the 90 sentimental comments, only the 40 positive comments were rated, giving a recall score of 44% (40/90). Now consider the impact to a positivity result: The system would say the data is 40% positive, 0% negative, and 60% neutral. This would be very misleading data, as the true rating should be 40% positive, 50% negative and 10% neutral. Quite a difference!

In this example, the system may have a very high accuracy rating, but without knowing its recall, we cannot comfortably trust the results.

Interestingly, recall and accuracy are often at odds with each other, as attempts to boost recall often negatively impact accuracy and vice versa. Consider simply scoring the word "like". If a system assumes the word is always positive, it will boost recall, as that system will rate more documents as sentimental. However, it would likely hurt accuracy, as the word "like" is slightly more often used in a non-sentimental way ("this tastes like chicken"), causing the system to mis-score many comments as positive that should actually be marked neutral.

Test Results

To perform the test, the text from each of the 1,912,958 unique consumer reviews was sent through Infegy Linguistics, with the system rating sentiment for each. Every review was scored by the review's author as positive or negative, and the sentiment rating was compared to this for validation. So, how did Infegy Linguistics do? Here are the results:

Infegy Linguistics	
Documents Scored	1,912,958
Precision / Accuracy	93% (1733629/1863968)
Recall	97% (1863968/1912958)
F1 Score	0.952
Overall Correct	**91%** (1733629/1912958)

For sake of comparison, Infegy was given the opportunity to run this test on a leading competitor's system. The competitor is a third party provider, focusing entirely on natural language processing, with sentiment being a major component of their offering. The company is popular in several sectors, particularly social media monitoring and analysis, where they power the sentiment analysis behind several popular products.

The test which was run with the competitor had to be smaller in volume per their technical requirements, so their testing was done with a subset of the data. Random documents were sampled from the large test, keeping the category distribution and effective weighting the same, using as many reviews as the vendor would allow. The results they achieved are below:

Competitor	
Documents Scored	38,704
Precision / Accuracy	89% (22739/25584)
Recall	66% (25584/38704)
F1 Score	0.758
Overall Correct	**59%** (22739/38704)

The vendor claims precision in the "near 90%" range, and our test did confirm that claim. However, due to the system's lackluster recall result, overall accuracy is poor, coming in at just 59%, significantly lower than the 91% achieved by Infegy Linguistics.

Closing

After years of research and development, testing has officially confirmed Infegy Linguistics is well ahead of competition, with the system's deep understanding of context and grammatical structure enabling near-human levels of performance. Going forward, Infegy continues to invest significant resources into research and development behind these and other linguistics technologies, ensuring Infegy will remain the leader in the space, ever improving upon these capabilities.

About Infegy

Founded in 2006 in Kansas City, Missouri, USA, Infegy was born to create the future of market research and consumer insight. From day one, Infegy's goal has been to leverage the exponential growth of content shared online to better understand consumers. Early on Infegy began working with agencies to improve their in-house research. For these agencies, the incredible speed and flexibility of Infegy's technology was an immediate fit. Since then, Infegy has been expanding markets within enterprise, PR, research and analysis organizations and the financial industry.

Learn more or contact us at www.infegy.com

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116