**The science behind the service**

Last Updated: 2017-09-19[Edit in GitHub](https://github.com/IBM-Bluemix-Docs/tone-analyzer/blob/master/science.md)

The IBM Watson™ Tone Analyzer service is based on the theory of psycholinguistics, a field of research that explores the relationship between linguistic behavior and psychological theories. The service uses linguistic analysis and the correlation between the linguistic features of written text and emotional and language tones to develop scores for each of these tone dimensions.

Psycholinguistics researchers have worked to understand whether the words that we use in our day-to-day lives reflect who we are, how we feel, and how we think. After several decades of research in this area, it is now accepted in psychology, marketing, and other fields that language reflects more than just what we want to say. The frequency with which we use certain types of words can provide clues to personality, thinking style, social connections, and emotional states.

For example, people exhibit various tones in their daily communications: joyful or sad, open or conservative, analytical or informal ([Gou et al., 2014](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-gou), and [Jian et al., 2014](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-jian)). These tones can impact the perception of a person's online identity and the effectiveness of their communications in different contexts.

Moreover, in business email communications, people are likely to perceive negative emotions with greater intensity than they do positive emotions ([Byron, 2008](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-byron)). And in social media, people present different online identities that impact the impression that others have of them ([DiMicco & Millen, 2007](https://console.bluemix.net/docs/services/tone-analyzer/references.html" \l "bib-dimicco)).

Many people naturally read a message and judge the tones conveyed by the sender. But can a computer detect the tones disclosed by a message accurately and automatically? This is one of the many challenging questions to which researchers in the artificial intelligence and cognitive sciences fields are seeking answers. First with the Personality Insights service and now with the Tone Analyzer service, IBM is beginning to answer this question.

**Overview of related research**

Research has shown a strong and statistically significant correlation between word choice and personality, emotions, attitudes, intrinsic needs, values, and thought processes. Several researchers have found that people vary in how often they use certain categories of words when writing for blogs, essays, and tweets, and that these communication mediums can help predict different aspects of personality:

* [Fast and Funder (2008)](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-fast)
* [Gill et al. (2009)](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-gill)
* [Golbeck et al. (2011)](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-golbeck)
* [Hirsh and Peterson (2009)](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-hirsh)
* [Yarkoni (2010)](https://console.bluemix.net/docs/services/tone-analyzer/references.html#bib-yarkoni)

Most of these prior works are based on finding psychologically meaningful word categories from word usage in writing. This research serves as the basis for IBM's work on the Tone Analyzer service. Relying on the scientific findings from psycholinguistics research, IBM is working to infer people's personality characteristics, their thinking and writing styles, their emotions, and their intrinsic needs and values from the words that they write. IBM uses its machine-learning models to evaluate these characteristics by assessing various features of a person's writing.

**The general-purpose model**

The general purpose endpoint analyzes written content for a set of tones that are applicable to a broad range of uses. The Tone Analyzer service computes a scorecard that includes the following tones:

* **Emotional tone** is derived from IBM's work on emotion analysis, which is an ensemble framework that infers emotions from a given text. To derive emotion scores from text, IBM use a stacked generalization-based ensemble framework; stacked generalization uses a high-level model to combine lower-level models to achieve greater predictive accuracy. Features such as n-grams (unigrams, bigrams, and trigrams), punctuation, emoticons, curse words, greetings (such as "hello," "hi," and "thanks"), and sentiment polarity are fed into machine-learning algorithms to classify emotion categories.
* **Language tone** is calculated from linguistic analysis based on learned features.

**Measuring the quality of the service**

IBM measured the quality of the Tone Analyzer service along the dimensions of tone mentioned in the previous section:

* *Emotional tone* categories were benchmarked against standard emotion data sets such as ISEAR and SEMEVAL. Results show that the average performance of the ensemble model (macro-average F1 score is around 41 percent and 68 percent, respectively, for the two data sets) is statistically better than the best reported accuracy of the state-of-the-art models (whose macro-average F1 scores are around 37 percent and 63 percent, respectively).
* *Language tone* was evaluated with an in-depth study of more than two hundred thousand sentences collected from sources such as debate forums, speeches, and social media. Of these sentences, IBM randomly selected 1330 sentences for analytical tone and 1000 sentences each for confident and tentative tones. IBM then submitted these sentence to the Tone Analyzer service and also asked humans to analyze them.

For the human analysis, IBM used a crowd-sourcing platform called [CrowdFlower](https://www.crowdflower.com/) to annotate the selected sentences with different labels. Only raters with approval ratings over 85 percent were allowed to participate in the annotation tasks. Five annotators labeled each sentence, and IBM used the most prevalent of the five annotated results to determine the final labels.

* + For *analytical tone*, humans labeled 915 of the 1330 sentences as analytical, 411 as non-analytical, and 4 as not understandable. By comparing the predicted label with these ground-truth labels, IBM found that its analytical tone detection received an F1 score of 0.7518.
  + For *tentative tone*, humans labeled 292 of the 1000 sentences as tentative, 706 as non-tentative, and 2 as not understandable. By comparing the predicted label with the ground-truth labels, IBM found that its tentative tone detection received an F1 score of 0.6369.
  + For *confident tone*, humans labeled 623 of the 1000 sentences as confident, 374 as non-confident, and 3 as not understandable. By comparing the predicted label with the ground-truth labels, IBM found that its confident tone detection received an F1 score of 0.7288.

Overall, the differences between the predicted and ground-trust labels are not statistically significant. This indicates that the service performs well.

**The customer-engagement model**

The customer engagement endpoint identifies tones from the customer-care domain. To select the tones to evaluate with the customer-engagement model, IBM first conducted a study to identify which dimensions of tone are perceived as important in the domain:

1. IBM selected a set of 53 tones from the tone dimensions used in marketing, the dimensions used to describe writing styles, and emotion and personality scales from psychology.
2. IBM asked workers on CrowdFlower to rate the extent to which the 53 tone attributes describe a specific utterance in 1000 customer-care conversations. To simplify the rating task in the context of crowd-sourcing, IBM divided the 53 tones into four subsets. Human annotators needed only to rate a subset of the tones; IBM then determined ratings for all of the tones from aggregations of these results.
3. IBM performed factor analysis on a 53-by-53 correlation matrix and found at least seven significant factors (dimensions). IBM determined the names of factors to represent the most important concepts reflected by each of the dimensions.

These steps identified seven important tone dimensions for the customer-care domain: frustration, satisfaction, excitement, politeness, impoliteness, sadness, and sympathy.

**Data collection**

IBM used Twitter customer-support forums as the source of conversational data for its analysis of tone in the customer-care domain. Many companies use Twitter as a channel for providing support to their customers. Agents are hired and trained to monitor tweets that mention the company, to direct help requests, and to provide fast support to address customers' needs.

To collect as many conversations as possible, IBM selected 62 brands with dedicated customer-service accounts on Twitter. IBM chose these brands to cover a large variety of industries, geographical locations, and organizational scales. Overall, IBM collected 2.6 million user requests between June 1 and August 1, 2016.

**Data annotation**

IBM sifted the collected data set to retain only those conversations that received at least one reply and that involved only one customer and one agent. IBM also removed from the data set all non-English conversations and conversations that contained requests or answers with images. To preserve users' and companies' privacy, IBM replaced all *@mentions* in each conversation with *@customer*, *@[industry]\_company* (for example, *@ecommerce\_company* or *@telecommunication\_company*), *@competitor\_company*, or *@other\_users*.

This selection process identified approximately 96 thousand conversational utterances to be annotated by CrowdFlower workers. To promote greater understanding of the annotation context, IBM asked the workers to annotate on a conversation level, labeling all utterances involved in a conversation. Because the subjective nature of some of the proposed tones could lead to inconsistent perceptions, IBM asked the annotators to indicate how strongly they felt that utterances demonstrated the tones on a four-point Likert scale ranging from "Very strongly" to "Not at all." A Likert scale is preferable to a simple binary yes/no scale because it accommodates a higher tolerance for diverse perceptions and generates less-sparse labels.

Five different annotators labeled each conversation. IBM used only US-based annotators with an acceptable level of performance from previous annotation tasks. IBM also required annotators to answer five training questions correctly before they could proceed with the real annotation tasks. The training questions ensured that annotators had a sense of what each tone means in the customer-service context. IBM also embedded validation questions in the real tasks to further validate the quality of annotators' labels.

IBM used the average of the five annotators' scores as the final label for each utterance. IBM then used a score of 1 as the threshold for converting the labels to binary values. IBM chose 1 as the threshold based on observations of the label distributions and on the performance of the models with different thresholds. The final results produced the following distribution of utterances: 9536 frustrated, 10,806 satisfied, 6107 excited, 63,853 polite, 1688 impolite, 24,954 sad, and 10,475 sympathetic.

**Learning tone models**

Based on these customer-engagement conversations, IBM trained a machine-learning model based on the Support Vector Machine (SVM) to predict tone for new customer-care utterances. For the machine-learning model, IBM leveraged several categories of features, including n-gram features, lexical features from various dictionaries, the existence of second-person references in the conversation, some dialogue-specific features such as saying thank you or apologizing, and several higher-level features such as the existence of consecutive question marks or exclamation marks.

IBM found that about 30 percent of the samples had more than one associated tone. IBM therefore elected to solve a multi-label classification task rather than a multi-class classification task. For each tone, IBM trained the model independently by using a One-vs-Rest (OVR) paradigm. The paradigm used the utterances for each class as positive samples and all other utterances as negative samples. IBM identified the tones predicted with at least 0.5 probability as the final tones. For several tones, the training data was heavily unbalanced; IBM therefore identified the optimal weight value of the cost function for each tone during training.

IBM used precision, recall, and F-score to evaluate the accuracy of its classification model. The model demonstrated high accuracy when compared to a benchmark data set.