HiringBranch

Accurately Measuring Soft Skills

Large Language Models Versus HiringBranch

Introduction

HiringBranch, an AI technology company dedicated to unlocking the power of soft skill measurement, has gone to great lengths to demonstrate its superior ability to do so. Why? Because AI is only as good as its outcomes.

AI is typically built on large data sets that often do not belong to the technology company developing the algorithms. Familiar large public datasets include the Chat GPT API or Google's FLAN-T5. These datasets are useful for creating a language model that can solve a general problem. What about when a large language model (LLM) is given a specific task? This paper explores LLMs and their ability to solve one important problem: accurate soft skill measurement.

Building AI Technology with Datasets

Large datasets such as the Chat GPT API, Google's FLAN-T5 LLM, the Mistral 7B LLM, and more are available to train new algorithms and create new language models.

Machine learning applications of technology rely on these types of datasets to build predictive models. Publicly available datasets are generally useful for creating a language model that can solve a general problem. After attempts with other LLMs, HiringBranch decided not to use these public datasets to create its language model, since the company set out to solve the particular problem of soft skill measurement.

Understanding Language Models

Language models are AI algorithms that rely on deep learning and datasets to be able to process information and make predictions or new content with it. They can be large language models (LLM) or small depending on the number of variables the model was trained on.

Language Model Accuracy Starts With Reliable Datasets

Where Does HiringBranch Data Come From?

HiringBranch owns more than a decade's worth of first-party training data tied to people performance. The data was collected and structured from the company's former product, LearningBranch. Today, the company has dedicated years to create the first proprietary language model (LM) dedicated to soft skills. The company has used its LM to develop algorithms trained to detect soft skills with precision. This collection of algorithms is the first-ever registered Soft Skills AI®.

To date, no other publicly-available datasets focus on soft skills in academia or industry.

As a result, HiringBranch turned inwards and created its own dataset using authentic customer interactions to equip its soft skill language model.

First, it turned to its team of linguists to understand how soft skills are linguistically expressed. Then, using structured and unstructured data, the company created its own proprietary dataset rich in soft skill information to train a new generation of algorithms.

The HiringBranch algorithms are designed to solve a specific problem: **accurately measure soft skills in real-time as a person is speaking or writing in a mock job experience**. What makes the HiringBranch hiring assessment so effective is the AI algorithms that power it, and the unique datasets they are trained on.

HiringBranch is committed to its technology's ongoing accuracy and validity testing to ensure it can deliver the best results to its clients.

Putting the HiringBranch Language Model to the Test

As such, HiringBranch ran a series of fine-tuning experiments, using the DistilBERT machine learning technology, to measure what the accuracy of its soft skill predictions would be after being trained on various datasets. While equipped with billions of variables, **large public datasets led to poor accuracy results** – meaning there wasn't enough data for an algorithm to learn to detect soft skills accurately (the data was either insignificant or hidden).

HiringBranch Language Model At Least 32% More Accurate at Measuring Soft Skills Than Next Best LLM

To achieve reliable hiring assessment results, HiringBranch created a language model using its proprietary data. The dataset includes hundreds of thousands of annotated soft skill data points that were used to train DistilBERT -- now collectively known as the HiringBranch Soft Skills Transformer v.1.77, or SST for short.

The SST predictions were compared to the Google Flan-T5-Base and Large datasets as well as the Mistral 7B. These comparisons were selected because they were considered the best open-source models available at that time of comparison (Apr 2024).

In the experiments, DistilBERT was asked to classify if the soft skill was present after being fed positive and negative samples from each. The experiments were more accurate on the Google Flan LLM than the Mistral 7B LLM, and most accurate using HiringBranch's SST.

The results of the dataset comparison are listed below for two soft skills:

Soft Skill Accuracy: HiringBranch vs LLMs

Model Name	Soft Skill 1	Soft Skill 2
HiringBranch SST	**95.68%**	**98.20%**
Google Flan-T5-Base	44.63%	41.56%
Google Flan-T5-Large	63.85%	63.72%
Mistral 7B	54.07%	56.96%

HiringBranch

The soft skills included in this study are proprietary and have therefore been anonymized in the table above.

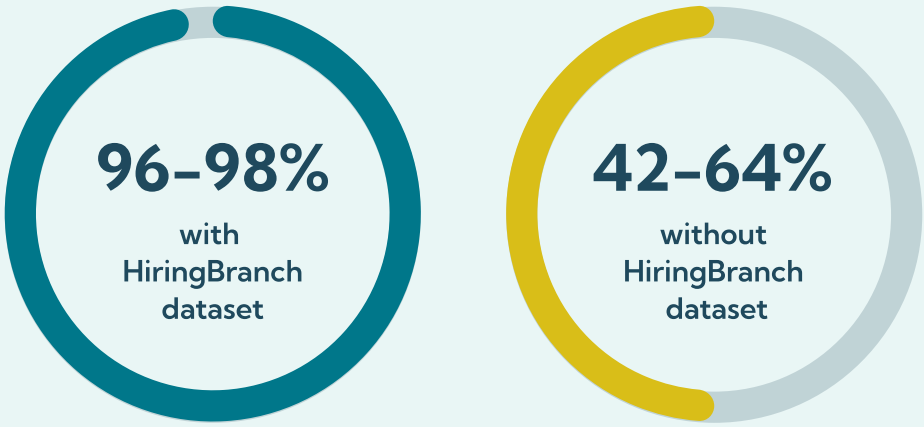
When the DistilBERT algorithms trained on Google's Flan T5-Large dataset they had the second best results with 64% accuracy across both soft skills measured.

When DistilBERT trained on the HiringBranch dataset, it measured soft skills at an average of 97% accuracy. That means when it comes to solving the problem of accurately detecting and measuring soft skills, the HiringBranch dataset is at least 32% more accurate than the nearest best LLM, and around 50% more accurate than the rest.

> "**When DistilBERT trained on the HiringBranch dataset, it measured soft skills at an average of 97% accuracy.....the HiringBranch dataset is at least 32% more accurate than the nearest best LLM.**"

From a hiring perspective, this means that nearly double the number of candidates have their soft skills represented accurately in the HiringBranch assessment than other skill assessments that have been developed using publicly-available datasets.

Language Model Soft Skill Accuracy



The Most Accurate Language Model for Soft Skill Measurement on the Market: Key Takeaways

Bigger isn't always better when it comes to datasets. There are large public datasets available that help technology companies develop AI products that can solve general problems. To solve a specific problem, like accurate soft skill measurement, a smaller, more specialized dataset has proven to yield more accurate results. This is because soft skills operate on a deeper linguistic layer of socio-pragmatic skills.

HiringBranch has used its proprietary dataset to create a language model that is much better at measuring soft skills than when trained on a large public dataset. The implications are significant and to date, millions of candidates have had a fairer and more accurate hiring process using the HiringBranch Soft Skills AI. The HiringBranch team is continuously testing and improving the accuracy of its AI algorithms using the most advanced data science techniques available.

Key Points Summary

- Public datasets that are available to train algorithms may not be appropriate for solving specific problems, like soft skill measurement.
- Smaller datasets may lead to more accurate results than larger ones if they have enough of the right data. A language model with greater accuracy will yield more effective results.
- The proprietary dataset, owned and created by HiringBranch, that its language model algorithm was trained on, has led to soft skill measurement results that are significantly more accurate than those trained on public datasets.
- The implications of accurate soft skill measurement are far-reaching, and today, HiringBranch has helped over 1 Million candidates access a fairer and more accurate hiring process using its Soft Skills AI.

Meet the HiringBranch Data Scientists

Assaf Bar Moshe, PhD,

Chief Research and Development Officer

As the head of the HiringBranch research team, Assaf is dedicated to bridging the gap between languages and computers by modeling natural languages as accurately as possible. As a scholar, his research focus is on Arabic dialectology, syntax, pragmatics, and historical linguistics, as well as Mandarin Chinese syntax and pragmatics. He has been lecturing about linguistics at Oxford University and the University of Heidelberg over the past decade. Assaf is proficient in English, Spanish, Chinese, German, Hebrew, and Arabic.

"With enough data and confidence in the data, management teams can use it to make decisions about the business, like which location candidates perform better, or which hiring partner brings more revenue etc. Using AI to process large amounts of data also means decision-making can happen faster. As a scientist, I'm excited by the potential of AI that we're already seeing within the recruiting and talent acquisition space. It's certainly increased the hiring confidence of our customers."

Vaibhav Kesarwani,

Senior NLP and ML Engineer

Vaibhav, aka "Vicky," has worked and studied machine learning for the past 7 years, completing his Master's at the University of Ottawa with a poetry analysis and classification thesis. Over the span of his career, he's worked on Tensorflow, Keras, DL4J, Scikit, Word2Vec, BERT, NLTK, WEKA and Gensim for applying NLP, ML and DL algorithms. Today he leads the AI efforts at HiringBranch.

"What's exciting about HiringBranch is the sheer scale. I've never seen machine learning on a scale like this before while providing immediate impact. Customers and applicants might see a single score, but the work to get to that score is immense. The complexity of the product is extremely large and I'm still learning every day."

HiringBranch

About This Study

This study comprised multiple phases including preparation, pre-processing, tuning, and more. The following sections describe the methodology behind this language model study in detail.

Dataset Preparation

Before HiringBranch started the computational aspect of the project, a group of linguists was gathered to identify and define the soft skills to be measured based on real CX interactions. This was foundational for the creation of a unique corpus for each soft skill, and required selecting thousands of representative CX interactions from proprietary HiringBranch data. Interactions were annotated based on string-matching rules and revised by linguists. Multiple rounds of annotation were possible based on the results of the first round. Rules and keywords were modified as necessary and a second or third round of annotation was completed where necessary, adding or removing data points as necessary. Once all annotators and linguists were satisfied with the quality of the annotations the dataset was finalized.

Model Pre-processing

Before model training began, the dataset pre-processing rules were defined. All variations of the data in the experiment stage were considered for objective comparison between the original dataset and the pre-processed dataset. Further, depending on the exact nature of the soft skill involved, and the text variations in the user responses, different pre-processing techniques of the original dataset were applied. Therefore, to evaluate the factual effectiveness of the pre-processing techniques, all the intermediate transformations were also included in the experiments. With the advent of deep learning and large language models, it is widely seen that the usability of pre-processing has declined over the years, and our experiments have also shown this to be a consistent pattern. HiringBranch still follows this to continue the tradition of data transformation. For some soft skills, we still see a slight increase in the accuracy scores on the pre-processed dataset compared to the original one.

Model Tuning

The dataset was divided into a development, training, and a test set. The data scientists kept the test set aside for the evaluation stage – the model never sees this data during the training stage. For model tuning, the model used the development set to tune its internal parameters and hyper-parameters. To start the first round or epoch of the training, HiringBranch had to provide the initial set of hyper-parameters, carefully selecting the learning rate, sequence length, warm-up, and weights (such as the development set accuracy being maximized). The experiments were set up with varying sets of parameters to deduce the ideal constant values used throughout the training stage. Once an ideal set was chosen, it was mostly unchanged during the various iterations to ensure a fair comparison during the evaluation stage.

Model Training

An optimum set of hyper-parameters was selected to begin training. Small tweaks were made to maximize accuracy metrics. Every experiment and confusion matrix was recorded in the process. All experiments were completed on a single set of user responses, and the series number for the experiments did not change. If a major addition or deletion to the original dataset was made, the series was incremented, denoting that the test set had changed and that the evaluation metrics would change too. For the first soft skill, a total of 26 experiments were completed, where changes to the original dataset were made, and experiments were conducted using various schemas like binary classification for discrete data and regression analysis for continuous data. A classification schema was selected as it provided the best results and applied to all subsequent soft skills. Therefore for the rest of the soft skills, a more focused experiment set was used that allowed the HiringBranch data scientists to go deeper, not wider.

Model Evaluation

For model evaluation, the confusion matrices for all the experiments were recorded in a single experiment series. The overall weighted F1-scores and the F1-scores of all the classes (majority and minority) were then compared. The overall metrics, though important, did not provide the complete metrics needed for evaluation, since the majority class had stellar accuracy but the minority did not.

For deep learning-based models, the input used was the CX interactions, either pre-processed or not pre-processed depending on the experiment series. For an LLM-based model, the input had a prompt prefix, CX interaction, and a prompt suffix. For the prefix and suffix, we had to experiment with a variety of prompts to see which gave the most relevant output. For different LLM models, we had to use different prompts as there are more subtle differences in the separators and vocabulary these models use to understand and respond. The prompts were very carefully selected, as each evoked certain variations in the output text that the parser was unable to parse. Therefore, prompts were inserted with restrictive language to control the response predictably. This is a classification task so a class label was needed along with a short description after a separator to make it easy for the parser to record the labels and do a consistent evaluation.

Further, this study experimented with zero-shot and few-shot scenarios. For zero-shot, the CX interaction was pushed to the model along with a task description as a prompt prefix and expected model prediction based on that. For few-shot, along with the task description, positive and negative samples were added so the model could see these examples and make better predictions. We chose these examples to be varied so that they offer a better representation of the dataset.

About HiringBranch AI

At HiringBranch we developed our own AI Governance Framework based on guidance from the Government of Canada's Algorithmic Impact Assessment, the Institute of Corporate Directors, and NuEnergy.ai. Today it is a proprietary SaaS solution used by some of the world's largest enterprises to improve their high-volume hiring outcomes. The AI is built from NLU, NLP, and supervised algorithms that work in conjunction with the HiringBranch proprietary People Skills Framework™. Using industry and role-based scenarios tailored to the employer that elicit open-ended responses, the technology is capable of text and speech analysis. HiringBranch has documented methodology to test datasets against biases and other unexpected outcomes using statistical analysis. Assessment scores are fully automated, applied across assessment methods, and aggregated. The technology's assessment data is analyzed and validated for test content, response methodology, assessment structure, scores, and outcomes.

HiringBranch

About HiringBranch

HiringBranch is an AI-powered communication and soft skills hiring assessment and training solution that guarantees a hiring performance improvement. 4X more reliable than traditional multiple choice assessments, HiringBranch tailors assessment content to specific roles and industries in an open-ended speech and chat candidate experience for outstanding results.

Founded in 2017 and is headquartered in Vancouver and Montreal, Canada, HiringBranch proudly serves large and medium-sized enterprises globally in retail, banking, insurance, telecommunications, health, and IT. The entire organization is committed to operating fairly, while fostering diversity and inclusion for all of our customers globally through unwavering and unbiased technology.

Start Measuring **Soft Skills** with HiringBranch

For more information about the HiringBranch hiring assessment and how it measures soft skills for high-volume recruiters, reach out to our team and ask how you can get started.

www.hiringbranch.com