Leading Social Media Company Improves Platform with High Quality Data Collection

Situation

A leading social media company needed a large amount of data to improve its machine learning model. This would allow its tool to better understand user-generated messages by identifying user intent, sentiment, and entities (people, places, events) from natural language.

The training model required very large datasets—thousands of phrases representing different ways users might input requests. While the company was able to pull data from its own user-generated content, the amount of data available for each scenario wasn't enough to allow it to build the product as fast as it needed to. The model also required examples of phrases that were not clear or relevant to a user's request. Training the model with false positives and false negatives was an important requirement for this project.

Solution

The company had a tight internal deadline by which to complete this project, and needed to partner with a firm that could deliver a large amount of relevant, high quality data in a short amount of time. With a minimal turnaround and using an internal tool, Appen was able to recruit hundreds of participants within a few days and collect thousands of samples, which allowed the client to meet its internal deadlines. In less than two months, more than one million samples were collected across many different categories including transportation, events, movies, and sports. This data was then used to improve the platform's help center, ads, videos, and other features. These samples included enough variation in language, slang, and idioms for the data scientists to rely on one dataset for the whole end-to-end process.

About Appen

Appen is a global leader in the development of high-quality, human annotated datasets for machine learning and artificial intelligence. With over 20 years of experience, expertise in more than 180 languages, and access to a crowd of over 400,000 worldwide, Appen partners with global companies to enhance their machine learning-based products.

Benefits

As a result of this project, the client released its product on time with the data it required to meet its users' needs. The firm quickly and efficiently improved its machine learning model with access to a large amount of high quality data. The geographic and demographic diversity of the rater pool proved immensely valuable to the training model. The crowd model also allowed the firm to significantly control project costs compared to other methods of data collection.

Key Success Factors

Appen's ability to deliver training data across a diverse set of users in a tight frame—while maintaining a high level of quality—was a key success factor for this project. Appen's agility in responding to requests continues to add value to this client as it develops new features.

Appen at a Glance

Expertise in over 180 languages and dialects

Access to a curated crowd of over 400,000

20+ years of experience providing high quality, human annotated data to support machine learning for speech, search, eCommerce and more