bitext we help AI understand humans

Bitext Sample Pre-built Customer Support Dataset for English

Overview

This dataset contains example utterances and their corresponding intents from the Customer Support domain. The data can be used to train intent recognition models Natural Language Understanding (NLU) platforms.

The dataset covers the "Customer Support" domain and includes 27 intents grouped in 11 categories. These intents have been selected from Bitext's collection of 20 domain-specific datasets (banking, retail, utilities...), keeping the intents that are common across domains. See below for a full list of categories and intents.

Utterances

The dataset contains over 20,000 utterances, with a varying number of utterances per intent. These utterances have been extracted from a larger dataset of 288,000 utterances (approx. 10,000 per intent), including language register variations such as politeness, colloquial, swearing, indirect style... To select the utterances, we use stratified sampling to generate a dataset with a general user language register profile.

The dataset also reflects commonly ocurring linguistic phenomena of real-life chatbots, such as:

- spelling mistakes
- run-on words
- missing punctuation

Contents

Each entry in the dataset contains an example utterance from the Customer Support domain, along with its corresponding intent, category and additional linguistic information. Each line contains the following four fields:

- flags: the applicable linguistic flags
- utterance: an example user utterance
- category: the high-level intent category
- intent: the intent corresponding to the user utterance

Linguistic flags

The dataset contains annotations for linguistic phenomena, which can be used to adapt bot training to different user language profiles. These flags are:

B - Basic syntactic structure
S - Syntactic structure
L - Lexical variation (synonyms)
M - Morphological variation (plurals, tenses...)
I - Interrogative structure
C - Complex/Coordinated syntactic structure
P - Politeness variation
Q - Colloquial variation
W - Offensive language
E - Expanded abbreviations (I'm -> I am, I'd -> I would...)
D - Indirect speech (ask an agent to...)
Z - Noise (spelling, punctuation...)

These phenomena make the training dataset more effective and make bots more accurate and robust.

Categories

The intent categories covered by the dataset are:

ACCOUNT
CANCELLATION_FEE
CONTACT
DELIVERY
FEEDBACK
INVOICES
NEWSLETTER
ORDER
PAYMENT
REFUNDS
SHIPPING

Intents

The intents covered by the dataset are:

cancel_order
complaint
contact_customer_service
contact_human_agent
create_account
change_order
change_shipping_address
check_cancellation_fee
check_invoices
check_payment_methods
check_refund_policy
delete_account
delivery_options
delivery_period
edit_account
get_invoice
get_refund
newsletter_subscription
payment_issue
place_order
recover_password
registration_problems
review
set_up_shipping_address
switch_account
track_order
track_refund

About Us

Bitext delivers the most precise and granular text analytics solution on the market, with an accuracy rate above 90%.

We are computational linguists first. Our technology really understands sentence structure and its different layers of meaning, so it always produces the richest results.

info@bitext.com

Our Locations

Madrid, Spain
José Echegaray 8, building 3, office 4
Parque Empresarial Las Rozas
28232 Las Rozas

San Francisco, USA
541 Jefferson Ave., Ste. 100
Redwood City
CA 94063