Prodapt powering global telecom

Developing Machine Learning-based Approach for Optimizing Virtual Agent Training

Improves precision, recall & accuracy of NLU model

Credits

Sathya Ramana Varri C

Prashanth Suresh Babu

Sarvagya Nayak

Virtual agents not able to stand up to consumers' expectations

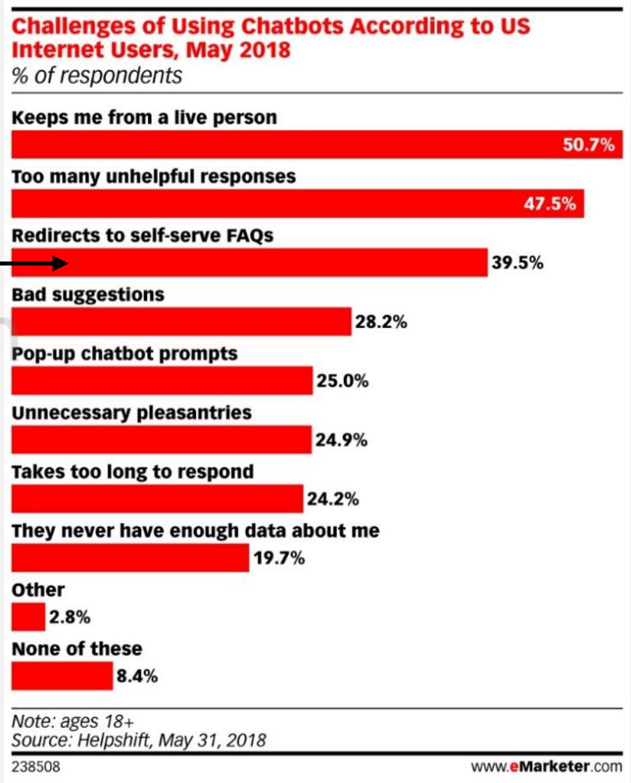
51% of US consumers don't have faith in VA's ability to respond correctly

Chatbots and virtual agents (VA) have high expectations in terms of customer engagements and overall customer experience. That's why Business Insider claims that **by 2020, 80% of the organizations will be using virtual agents.**

But the end consumers don't have faith in them. In fact, 51% of the US population thinks that VAs are a hindrance that keeps them from connecting to a live agent. 41% of respondents feel VAs don't provide enough detailed solutions and 37% feel they are generally not helpful.

The **major reason for the failure of these VAs** to satisfy consumers lies in their inability to identify the right intents. This, in turn, is the effect of wrong or inadequate training of VA's natural language understanding (NLU) engine.

Most often the identification of training data is done manually which is not enough. This insight talks about developing a **machine-learning (ML)-based Intent Analyzer tool**, which can identify the most effective data set for NLU training.



Machine learning-based Intent Analyzer tool identifies most relevant representative examples for NLU training

The conventional approach of identifying training data for VA NLU depends heavily on DSP's internal process experts. It involves choosing the most relevant few hundred examples of millions of historical chat. But, it is crippled with inefficiencies because it:

- Lacks coverage of all the examples needed for training
- Makes way for manual biases
- Highly time-consuming

Developing a ML-based intent analyzer tool is the most optimal approach for identifying representative training examples. This tool should perform:

Historical data of millions of chats

Data Import 1

Chatlogs are sourced as .csv files and imported as data frames using pandas library; chat column is sliced for further processing

Data Preprocessing 2

- Initial cleanup using regular expressions
- Cleanup and refinement of chat-specific components such as chat and time stamps
- Stop word and punctuation removal

Text Vector Processing 3

- TFIDF (Term Frequency Inverse Document Frequency) vectorization – identifies relevance of the word within the context of the document
- Stemming is performed to get root (stem) of word sans prefix/suffix

Clustering 4

- K-means clustering performed to identify useful logs
- Number of chats (e.g. N= 50) can be specified so that the top N chatlogs can be derived

Intent Analyzer Tool

ML-based Methodology

Most relevant chats per use case identified

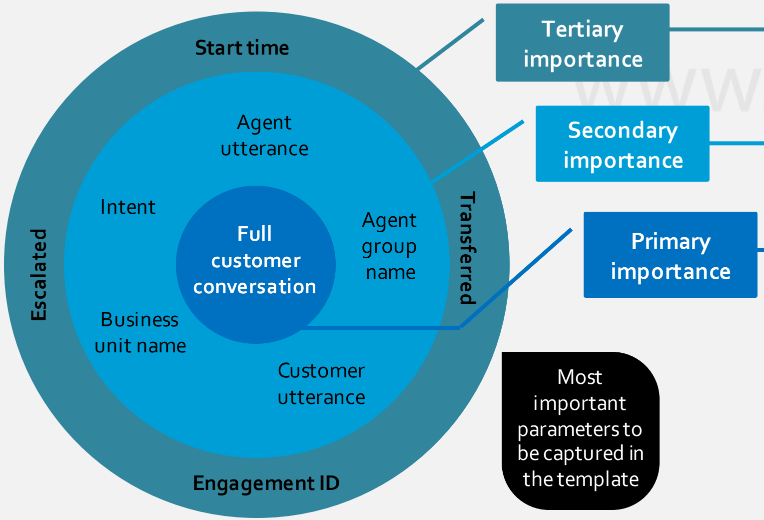
The subsequent slides give details on performing the above 4 steps in the best possible manner.

Data Import - Templatize the input to optimize the most time-consuming step

This step involves sourcing historical data from chatlogs for respective intents/use cases. Millions of chats are picked and imported to identify the most relevant handful of them.

Recommendations

Templatize | It is extremely important for the input data to be in a standard format. To ensure this standardization, it is recommended to create a **template** for it



Most important parameters to be captured in the template

Full Text	Agent Text	Customer Text	Business Unit Name	Agent Group Name	Start Time	Engagement Id	Transferred	Escalated
[a:] hello this is paul. i am happy to help[a:] may i have your firstand last name company name and wireless number please[c:] please tell me why there have been no account updates since feb 9[a:] also how may i help[c:] chris harris ch tarde[c:] xxx 7292929[c:] did you see my question[c:] hello.....[c:] ??????????[c:] is anyone there?[a:] thank you for validating your account information chris. thishelps us keep your account secure[a:] what type of updates are missing[c:] there are no voice usage numbers posted to any of our accounts since 2/9[a:] do you see view bill on the left as a blue link chris[c:] i need to look at unmilled usage[c:] i do see view bill[c:] i check the usage often[c:] but suddenly no more transactions/calls[a:] can you hover over view bill and click the first link please. let me know when that page has loaded[c:] loaded[c:]	[a:] hello this is paul. i am happy to help[a:] may i have your firstand last name company name and wireless number please[a:] also how may i help[a:] thank you for validating your account information chris. thishelps us keep your account secure[a:] what type of updates are missing[a:] do you see view bill on the left as a blue link chris[a:] can you hover over view bill and click the first link please. let me know when that page has loaded	[c:] please tell me why there have been no account updates since feb 9[c:] chris harris ch tarde[c:] xxx 7292929[c:] did you see my question[c:] hello.....[c:] ??????????[c:] is anyone there?[c:] there are no voice usage numbers posted to any of our accounts since 2/9[c:] i need to look at unmilled usage[c:] i do see view bill[c:] i check the usage often[c:] but suddenly no more transactions/calls[c:] loaded	BVA	BCS-Premier_e bill	2019-02-12T13:42:39Z	7.24618E+17	F	F

Sample from a template

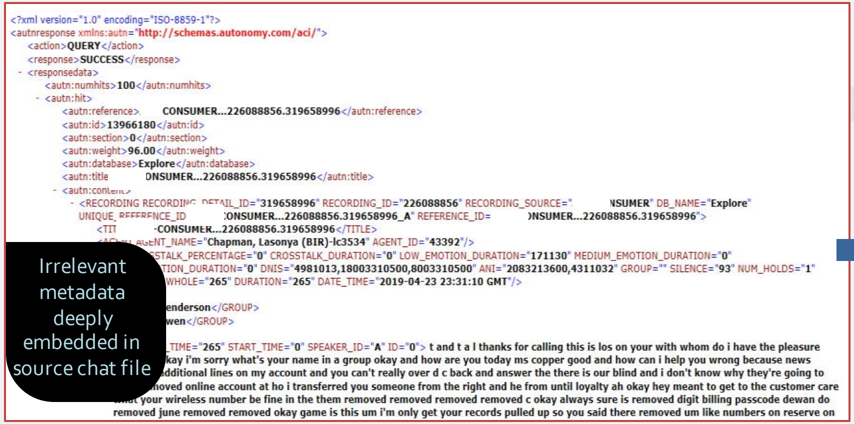
Data Import – Remove random noise and flatten the input file to remove metadata

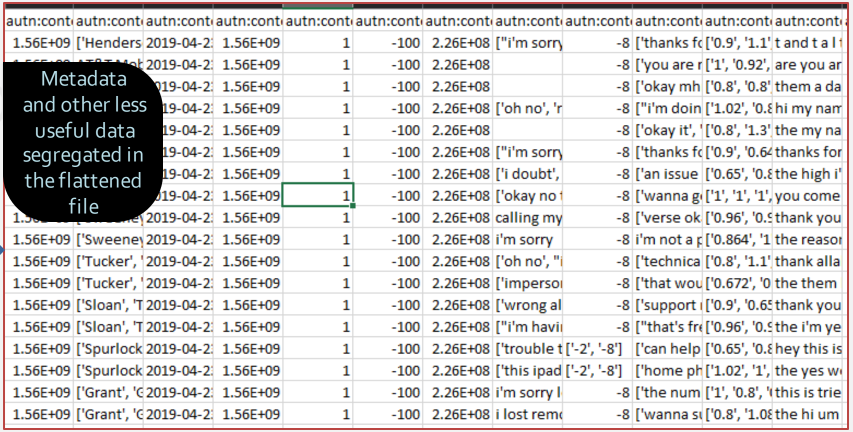
1
2
3
4

Recommendations

Remove random noise/white noise | Seasonality in data can lead to wrong inferences. For example, higher call drops or lower speeds during Thanksgiving or Christmas. To reduce that impact, choose the data set spread over a larger time period like 9-12 months.

Key-value pair | For efficient separation of metadata, flatten the file into excel file or other simpler formats





Reducing import time | By performing parallel processing and avoiding overloading memory

Data Preprocessing - Leverage raw text preprocessing, regular expression and lemmatization

1
2
3
4

The step involves initial clean up of the chat data by removing chat specific components such as timestamps, stop-words and punctuations.

Recommendations

Special character processing and text analysis

For removing special characters and analyzing text at high level perform 'raw text preprocessing' and 'regular expressions'

Raw text preprocessing

Removes chat specific notations and special characters which doesn't add any value to analysis. E.g – timestamps, special characters, etc.

[c:] i wish to make a payment arrangment for 13.26 on january 25th xxxx[c:] i have a question about a payment arrangement[c:] ##url#https://www.abcde.com/ esupport/article. html#!/iptv/km1025834

→

i wish to make a payment arrangment for 13.26 on january 25th xxxx i have a question about a payment arrangement url https://www.abcde.com/esupport/ article.html !/iptv/km1025834

Regular expression

Segregates numerals from alphabets and retains only special strings of alphanumeric values

my bill was 260$ and some change will my new bill be around 280$?[c:] i always thought it was lte since my phone has a symbol of 4g lte

→

my bill was xxx$ and some change will my new bill be around xxx$? i always thought it was lte since my phone has a symbol of 4glte

Lemmatization or stemming

Focuses on reducing the words to their root words

Lemmatization

the stemmed version of "I am paying the bills" is

Word	Root
Paying, pays, paid	Pay
Am, are, is	Be
Bill, bills, bill's	Bill

I am paying the bills → *I be pay the bill*

Rare word removal

They create noise by their association with other words. They might not be rare, but their usage in certain context can be misleading. E.g. – erroring, revert, captive, hill, angel.

Text Vector Pre-processing - Leverage term frequency inverse document frequency (TFIDF) for most effective vectorization

Text vector preprocessing helps in understanding the importance of words as per the relative context. It focusses on the difference in relevance in different circumstances.

Recommendations

Choice of vectorization method | Choose a method based on the type of text data. It is recommended to choose 'term frequency-inverse document frequency (TFIDF) vectorization' since it considers the relative importance of a word in each context.

Term Frequency-Inverse Document Frequency (TFIDF)

- TFIDF ensures that the chats are selected according to their relative importance
- More than their overall significance in everyday usage, it measures how critical they are in the context of the chat log corpus being analyzed
- This helps in identifying the most relevant chats as per the intent

N-grams and other multiword usage

Certain words have an entirely different meanings when used in combination with a few other words. Such words should be configured appropriately.

E.g. – the words "payment" and "arrangement" have different sense and relevance when used individually. But upon using collectively as "payment arrangement" it conveys some other meaning.

Hyper-parameter tuning

Tunes the parameters of the vectorization algorithm to optimize the output

E.g.

- *Max_df – sets the acceptable upper limit of frequency.*
  - *E.g. – 'IPTV max_df = .85' - Chat containing "IPTV" more than 85% of time will be ignored.*
- *Min_df – sets the acceptable lower limit of frequency*
  - *E.g. – 'IPTV min_df = .20' - Chat containing "IPTV" less than 20% of time will be ignored*
- *Max_features – defines the vocabulary size*
  - *E.g. – 'Max_features = 10,000' - This limits the number of words in vocabulary to 10,000 words*

TFIDF Input

E.g. – How to close my account,
Close my account,
Delete my account,
Close account

TFIDF Output

(0, 157)	0.6470729031869088
(0, 6)	0.5314083612599048
(0, 213)	0.4035924769447586
(0, 447)	0.3688020120578553

Clustering - Perform k-means clustering technique for effective classification

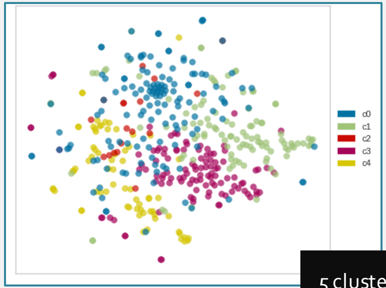
Clustering ensures that the top-N chats (where N is variable depending on business/NLU needs) are derived. These can be quickly analyzed to identify utterances, intents and entities. Additional ML processing such as entity or intent recognition can also be performed if required. All this results in significant time and effort saving.

Recommendations

Choice of clustering technique	Choose the technique that can work on huge volume of data like millions of customer chats. It is recommended to use k-means clustering for such a volume.
One-on-one mapping	Ensure one chat is mapped with only one intent i.e. avoid overlapping
Intent-specific scaling	Ability to scale the number of top-N use-cases based on intent call-volume (by varying the number of clusters): this enables the number of representative samples to be adjusted based on whether a given intent has more or less volume

```
(0, 157)      0.6470729031869088
(0, 6)        0.5314083612599048
(0, 213)      0.4035924769447586
(0, 447)      0.368802012057855З
```

Input vectors from TFIDF



5 clusters representing 5 intents

good morning! i am taking a trip to mexico next week and i have the international day plan or
hello erik brown and kelly brown are flying to cancun mexico tomorrow (phone numbers 773-
i need to know if my current plan coverage includes travel to mexico and if so what services n
hello i added international passport 3gb to my account for a long intl trip but i want to make s
im traveling on a cruise to mexico march 3 to the 10th are there plans available for using my p
i recurved international day use charges when i purchased the international passport with 1g
i'm traveling to london on saturday march 2 returning march 8th and i
hello i am traveling to hawaii in 2 week and wanted to check if my cur

Most relevant representative examples for the intent "international roaming"

Prodapt

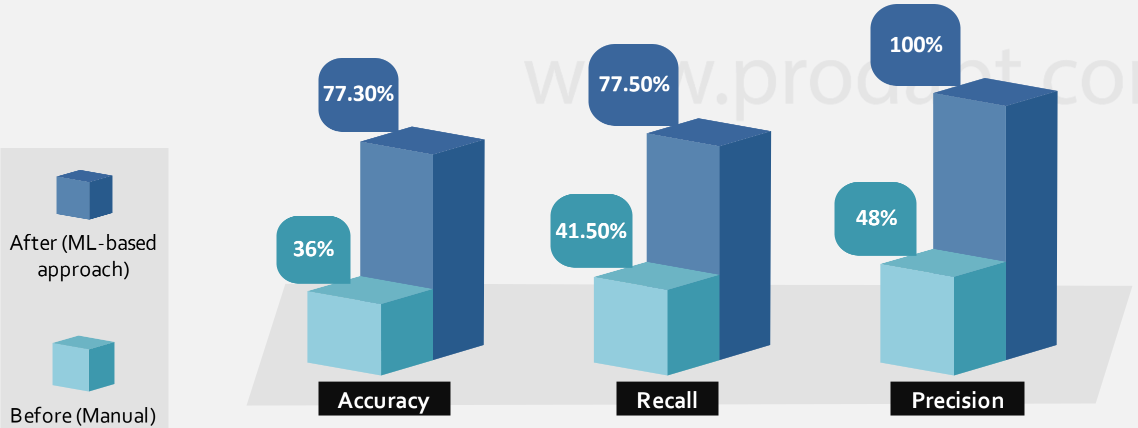
Key takeaways

Leveraging a machine learning-based approach for identification of training data for NLU can have following benefits

NLU confidence

The number of use cases crossing confidence threshold can increase by **160-180%.**

Confidence threshold – the minimum confidence level configured in VA below which it can't process the chat and transfers it to the live agent



Accuracy: Signifies how often the model is correct

Recall: Signifies the number of predicted positives out of total actual positives

Precision: Signifies actual positives out of all the positive predictions

Time efficiency

Can save up to **97%** of time in identifying the examples

Transfer to live agents

Can reduce by almost **80%**

Get in touch

USA

Prodapt North America
Tualatin: 7565 SW Mohawk St.,
Phone: +1 503 636 3737

Dallas: 1333, Corporate Dr., Suite 101, Irving
Phone: +1 972 201 9009

New York: 1 Bridge Street, Irvington
Phone: +1 646 403 8161

CANADA

Prodapt Canada Inc.
Vancouver: 777, Hornby Street,
Suite 600, BC V6Z 1S4
Phone: +1 503 210 0107

UK

Prodapt (UK) Limited
Reading: Davidson House,
The Forbury, RG1 3EU
Phone: +44 (0) 11 8900 1068

EUROPE

Prodapt Solutions Europe &
Prodapt Consulting BV
Rijswijk: De Bruyn Kopsstraat 14
Phone: +31 (0) 70 4140722

Prodapt Germany GmbH
Münich: Brienner Straße, 80333
Phone: +31 (0) 70 4140722

Prodapt Switzerland GmbH
Zürich: Mühlebachstrasse 54,
8008 Zürich

SOUTH AFRICA

Prodapt SA (Pty) Ltd.
Johannesburg: No. 3,
3rd Avenue, Rivonia
Phone: +27 (0) 11 259 4000

INDIA

Prodapt Solutions Pvt. Ltd.
Chennai: Prince Infocity II, OMR
Phone: +91 44 49033000

"Chennai One" SEZ, Thoraipakkam
Phone: +91 44 4230 2300

IIT Madras Research Park II,
3rd floor, Kanagam Road, Taramani
Phone: +91 44 49033020

Bangalore: "CareerNet Campus"
2nd floor, No. 53, Devarabisana Halli,
Phone: +91 80 4655 7008

THANK YOU!