International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

SEMANTIC ANALYSIS OF ONLINE CUSTOMER QUERIES

Varun Patil[1], Prof. Maya Chaugule[2]

[1]*Varun Patil (K.L.S Gogte Institute of Technology, Department of Electronics and Communications, Belgaum, Karnataka, India)*
[2]*Prof. Maya Chaugule (K.L.S. Gogte Institute of Technology, Department of Electronics and Communications, Belgaum, Karnataka-India)*

Abstract - *Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. Often semantic analysis becomes difficult in Chatbot kind of environments due to lack of structure in user queries, lack of enough keywords, context maintenance as expected by end user etc. In this paper, we try to use Machine Learning approach with different algorithms for semantic analysis of customer queries received in Chatbot kind of environment of Banking Organization. Queries received will be directed to a Machine Learning classifier which analyses the query and maps to a particular Intent/Response. The main Objective will be to bring any improvement in classification algorithms with incorporation of technique via Context maintenance*

***Key Words*: Semantic Analysis, Classification, Machine Learning(ML), Support Vector Machines, Naïve Bayes**.

1.INTRODUCTION

Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of writing as a whole, to their language independent meanings. Semantic analysis is used in Question Answering Systems like chatbot to understand user intents and to respond with appropriate action response. Often user queries received in these environments are very short and lacks enough keywords, proper grammatical structure etc., due to which semantic analysis becomes difficult.

In this paper, we have developed a machine learning based classification model for analysis of customer queries. The proposed model also introduces context management via a tagging method to further improve classification accuracy. The classification thus performed helps to identify the user intent from his/her queries. An analysis study has been made to analyze the performance of different algorithms for classification in terms of accuracy, number of samples and the effect of context management.

2. ARCHITECTURE

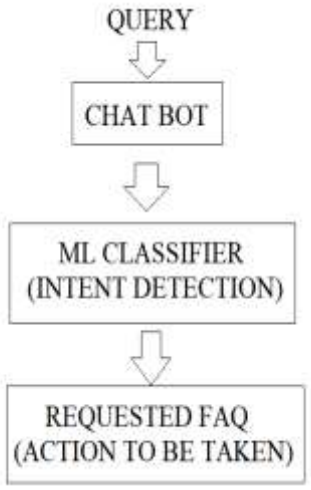

Fig 1: Architectural model

1] Chatbot: - A chatbot is an artificial intelligence (AI) software that conducts conversation with the customer or users in natural language through messaging applications and mobile applications or through the telephone.

2] Machine Learning Classifier: - A Machine Learning classifier is an algorithm that implements classification, especially a concrete implementation. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, which maps input data to a category.

3] Requested FAQ: - The requested FAQ is the result achieved by the classifier algorithm which provides a response for the query from the customer. The function of Machine Learning Classifier is to respond to the user's query and provide an appropriate FAQ response.

3. METHODOLOGY

Following is the step by step procedure for building the functional block model:

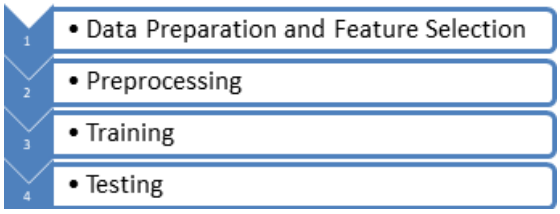

Fig 2: Functional Model

1] Data Preparation and Feature Identification

Selection of features

- **Questions**: These are frequently asked questions (FAQ) which are preprocessed that is the sentences are filtered with the help of tokenizing, removal of stopword and Lemmatization so that only keywords remain as features.
- **Answers**: These are the responses for the FAQ so that the missing keywords in the questions can be included in the training dataset.
- **Previous Tags**: Are the tags of the previously asked questions.

  Context Maintenance: For each FAQ there will be a previous tag attached which lets us know the previous question asked for context maintenance so that it learns from previous questions.

- **Target Class**: These are the categorized classes for the FAQ.

2] Preprocessing

In this step, data cleaning is performed to preprocess the data using Natural Language Processing steps so that unnecessary data is removed.

Tokenization: Tokenization describes splitting the sequence of string or paragraphs into sentences, or sentences into individual words.

Removing the Stopword: A majority of the words in a sentence are connecting parts of a sentence rather than showing subjects, objects.

Lemmatization: Lemmatization is a process where words a reduced to its base by removing inflection through dropping unnecessary characters, usually a suffix.

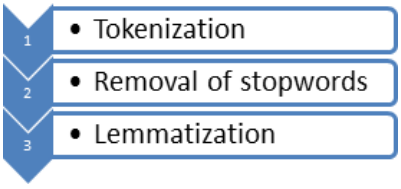

Fig 3: Preprocessing

3] Training

The training dataset was made up of 245 questions with respective answers, previous tags and target class. There were 35 target classes that is 7 variations of each questions (35*7= 245). Training was performed after the preprocessing step with 4 different algorithms namely Support Vector Machines, Naive Bayes, Logistic Regression, K-Nearest Neighbor. The analysis and performance for each algorithm is shown further in terms of accuracy. These algorithms were selected based on the literature survey performed to find the most used Machine Learning algorithms in industry for text classification.

Countvectorizer was made use of for preparing the word vector for textual features extraction. This helped in converting our textual data into vectorized format using Bag of Words concept. Bag of Word is a method for preparing textual content as input for our machine learning algorithms.

These classifiers were used from scikit-learn library. It is an open source machine learning library for the python programming language.

4] Testing

Testing was done using the train-test split library from scikit-learn with a distribution of 70% as training data and 30% as testing data.

4. RESULTS AND DISCUSSION

In the first round for selection of algorithms we have performed a train test split iteration whose results are given in figure 4.

Parameter	Support Vector Machine	Naive Bayes	Logistic Regression	K Nearest Neighbor
Accuracy	66%	61%	44%	32%

Fig 4: Comparison chart of accuracy for different algorithms.

Accuracy = Number of correctly predicted target class / Number of Target Class

From the first round of train-test analysis it was found that the accuracy rate for Logistic Regression and K-Nearest Neighbor (with K=3) were too low, therefore it was decided to go with only Support Vector Machines and Naive Bayes for our further testing process.

Since Naive Bayes and Support Vector Machine were found to give more accuracy compared to other algorithms. These two algorithms were further analyzed on two scenarios.

- **Analysis effect of number of samples**

As in figure 5 and 6 we found that when numbers of samples are increased (2, 4, 6 samples respectively) accuracy of both Support Vector Machine and Naive Bayes algorithms are increased. Also, SVM performs better than

Naive Bayes when samples are increased. These cases were analyzed along with context inclusion.

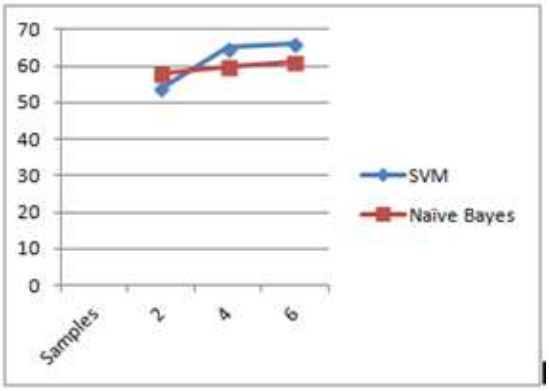

Fig 5: Accuracy vs. number of samples analysis without context maintenance

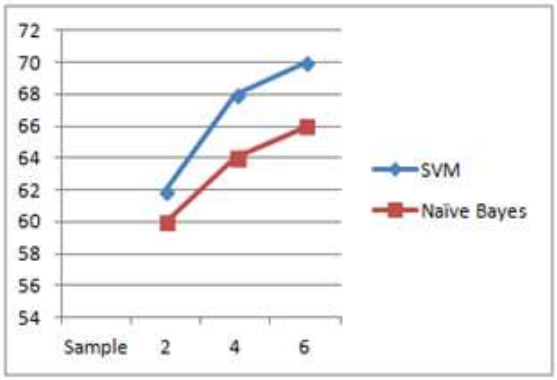

Fig 6: Accuracy vs. number of samples with context maintenance.

- **Analysis of effect of inclusion of context maintenance**

It is found that when context maintenance is included in terms of previous tags accuracy was improved for both of the algorithms as in figure 7 and 8.

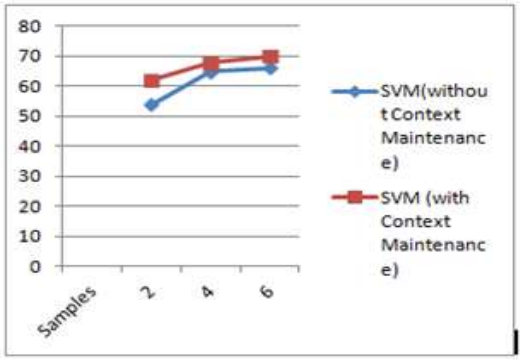

Fig 7: Accuracy with and without context maintenance (SVM vs. SVM)

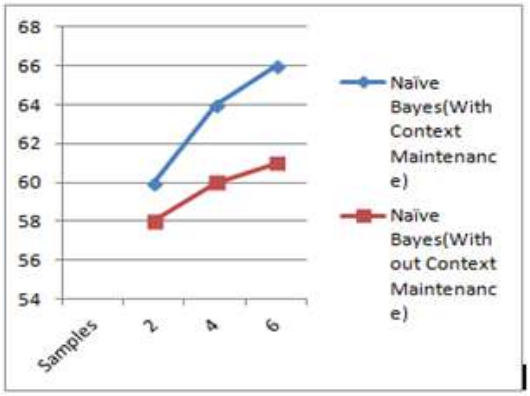

Fig 8: Accuracy with and without context maintenance (Naïve Bayes vs. Naïve Bayes)

5. CONCLUSION

The Machine learning based model which we have developed was found to be useful in semantic analysis of online customer queries. We found that SVM and Naive Bayes algorithms perform better compared to other algorithms analyzed in this paper. Accuracy can be further improved by increasing the number of samples. Also, we have developed a method of including previous tags in our machine learning model. Our analysis shows that this context management method improves the accuracy further in addition to the improvement achieved with increase in number of samples. Also, we have found that this method can help in prediction improvement in Chatbot environments where the query is normally very short but the context is readily available.

6. REFERENCES

[1] BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning, International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 05 | May -2017.

[2] Composite Naive Bayes Classification and Semantic Method to Enhance Sentiment Accuracy Score, Ardhian Ekawijana; Heri Heryono 2016 4th International Conference on Cyber and IT Service Management.

[3] Query Classification Using Convolutional Neural Networks, Hanxiao Zhang, Wei Song1, Lizhen Liu, Chao Du, Xinlei Zhao, 2017 10th International Symposium on Computational Intelligence and Design.

[4] Classification Model for Query Logs Based on Intent Mining, Zhang Shuang; Nianbin Wang, 2015 Seventh International Conference on Advanced Communication and Networking.

[5] Comparison of Question Answering Systems Based on Ontology and Semantic Web in Different Environment,

Journal of Computer Science 8 (9): 1407-1413, 2012 ISSN 1549-3636 © 2012 Science Publications.

[6] Intent Classification of Short-Text on Social Media, 2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015.