ISSN 1870-4069

Machine Learning for Intent Classification in an Educational Chatbot

Carlos Natanael Lecona-Valdespino[1], José Moisés Aguilar-Ibarra[2], Guillermo Santamaría-Bonfil[3]

[1] Universidad Panamericana, Mexico

[2] Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

[3] Consejo Nacional de Ciencia y Tecnología, Mexico

0245614@up.edu.mx, jaguilari1800@alumno.ipn.mx, gsantamaria@conacyt.mx

Abstract. Nowadays, intelligent learning environments use advanced human-computer interfaces such as voice-commanded chatbots. A fundamental aspect of such chatbots is that the dialogue system is able to correctly identify the intention of the users. In this work, an user intention classifier for an educational chatbot is built using different NLP techniques as well as different ML algorithms. A corpus of phrases belonging to intentions from an educational chatbot is used to train and compare support vector machines, decision trees, random forest, and extreme gradient boosting. A key step in this analysis is the inclusion of the parts of speech filtering to reduce the number of features required for intent classification, hence, reducing the computational burden. Results showed that the random forest algorithm using the parts of speech filtering identifies user intentions accurately and with less computational effort, in comparison to the rest of the algorithms combinations.

Keywords: Educational conversational agent, machine learning, intent classification, natural language processing, recommender system.

1 Introduction

Chatbots are considered to be one of the key technologies powered by artificial intelligence to improve the communication and interaction of users with computers [4]. This technology reduces the workload required from users by providing them with friendly and human-like interfaces. For this, chatbots employ Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) tools to process and understand speech commands or written text. These systems range from simple frequent ask questions answering machines to more sophisticated systems that can sustain a conversation with a human about a specific topic.

Chatbots are used broadly in several industries such as social media platforms, customers services, medicine, and even as educational assistants [3, 8]. Regarding the application of chatbots to educational settings, studies revealed that chatbots can enhance the learning settings by allowing personalized real-time interactions, improving the communication among peers, and increasing the efficiency of the learning procedure [4].

One of the main abilities of any chatbot is the identification of users intentions. An intent refers to the action or goal a user looks forward to achieve. Hence, intent classification tries to identify what the user wants by applying Natural Language Processing (NLP) techniques along with Machine Learning (ML) algorithms [9]. By identifying the user intent from a predefined intention list, the chatbot dialogue manager can redirect the conversation to specific dialogue paths [2].

Several works have been developed for intent classification using a multitudes of NLP and ML techniques. In the case of the former, tokenization, stop-words removal, lemmatization, parts of speech (POS) tagging and filtering have been employed [9, 2, 8]. In the case of the latter, naïve bayes, logistic regression, Decision Trees (DT), artificial neural networks (multilayer perceptron, convolutional neural networks, long-short term memory, BERT), Support Vector Machines (SVM), naïve bayes using bag of words and word embeddings, majority voting, FastText, stand among the most recently employed [3, 9, 6, 2, 8].

While complex neural networks such as BERT have shown outstanding performance in NLP problems where there are large amounts of available data, intent classification datasets are, in average, around a few thousand samples [5]. Under such circumstances traditional or ensemble algorithms perform better than complex deep learning structures. A key NLP procedure is identifying the POS. Such procedure consists in tagging words in accordance to their grammatical function. Although, POS tags are different for each language, within the same language, words tagged with the same POS can be considered to play a similar grammatical function in the sentences. Such POS property can be employed to filter the less important features for each sentence. In fact, in [1] considers this preprocessing step key for improving the classification accuracy.

Therefore, in this work several NLP procedures are used to preprocess user utterances from a corpus of phrases of an educational chatbot. By cleaning and filtering important features using POS tagging, support vector machines, decision trees, random forest, and extreme gradient boosting are trained for intent classification. The results showed that the random forest algorithm using POS tagging filtering identifies user intentions with high accuracy and with less computational effort, by reducing the number of features employed for classification.

The rest of the paper is organized as follows: section 2 presents the materials and methods; section 3 presents the experimental setup and the results of the experimentation; section 4 presents conclusions and discusses future work.

2 Materials and Methods

In Fig. 1 the methodology to build a model for intent classification is presented.

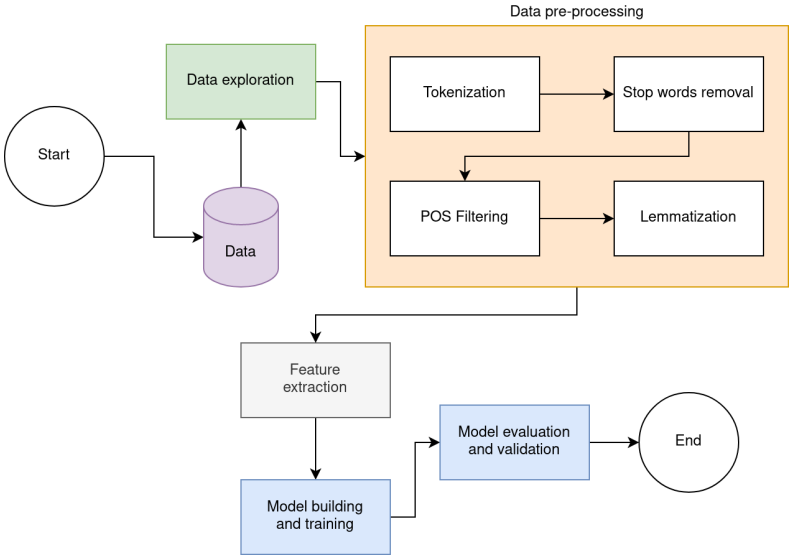


Fig. 1. Intent classification pipeline.

First, data corresponding to the intents of an educational chatbot used for training in energy topics [7] is collected and classified by intent.

Next, an exploratory data analysis is carried out to visualize i) words and ii) the POS that compose each intent. The dataset is then pre-processed using four NLP processes namely tokenization, stop words removal, POS (Part-of-Speech) filtering, and lemmatization.

From the pre-processed phrases features are extracted by building n-grams and transforming the text into TF-IDF vectors. Once the vectors are obtained, the dataset is separated into training and test sets. Using these, four ML models are trained and evaluated using three metrics: Balanced Accuracy (BA), Matthews Correlation Coefficient (MCC), and Cohen Kappa Score (KCS).

2.1 Data

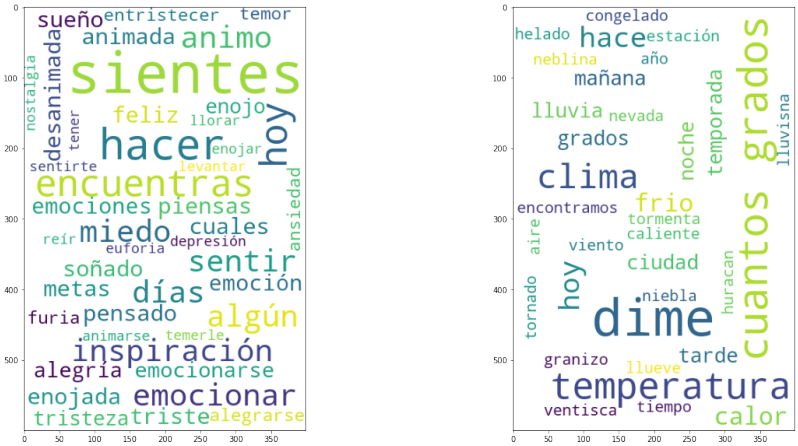
Initially, a corpus of 1850 phrases labeled their respective intention was developed for an educational chatbot used for virtual reality training environments [7]. The phrases are in Spanish and correspond to user requests to the chatbot to achieve one of the available actions.

The list of the eleven (11) available intents in the chatbot is shown in Table 1. In this table, first and second columns corresponds to the user intention and the name assigned in the dialogue manager of the chatbot, respectively. The third and four column presents the frequency of phrases and their corresponding proportion in the corpus.

Observe that, the corpus is highly imbalanced since the search intention corresponds to more than 60% of it. This is due to the chatbot main duty is to be capable of answering any knowledge question requested by the user as accurately as possible.

Table 1. Description of chatbot intentions.

User intention	Name	Frequency	Proportion
Lookup for a concept	Search	1200	0.65%
Teleportation within the virtual world	Navigation_microred_caseta	200	0.11%
Ask the chatbot to tell an interesting fact	Utility_check	50	0.03%
Ask for the date	Date_check	50	0.03%
Ask the chatbot to tell a joke	Joke_check	50	0.03%
Ask for the emotional state of the chatbot	Feel_check	50	0.03%
Ask for the time	Time_check	50	0.03%
Ask about the weather	Weather_check	50	0.03%
Thanking the chatbot	Gratitude_check	25	0.013%
End the conversation	End_conversation	25	0.013%



(a) (b)

Fig. 2. Word clouds for two different intents. On (a) the word cloud for Feel_check intention is shown, whereas on (b) the Weather_check is presented.

2.2 Data Exploration

To qualitatively understand differences between the available intentions a word cloud for each of these was generated. These word clouds allow to visualize the words that are most frequently used depending on each intent and the type of words that should be included for training.

Figs. 2a and 2b show the word clouds corresponding to feel_check and weather_check, respectively. Observe that, the feel_check intention employs words such as feel (sentir), emotions (emociones), joy (alegría), inspiration (inspiración), and so on. In contrast, the weather_check intention uses words such as temperature (temperatura), degrees (grados), weather (clima), and so on.

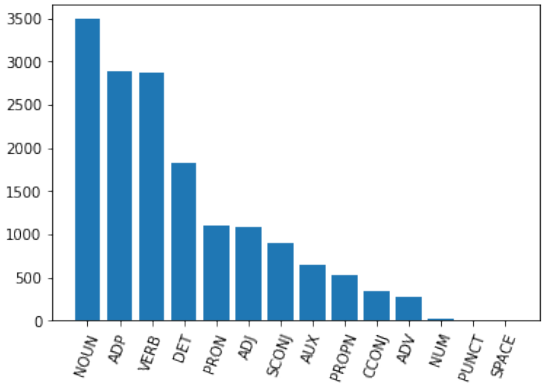


Fig. 3. POS Exploration.

2.3 Pre-Processing

Pre-processing consists in tokenization, stop words removal, POS (Part-of-Speech) filtering, and lemmatization. At the beginning, tokenization is performed to divide the sentences into words and apply the next pre-processing steps. Next, all stop-words are removed using a Spanish stop-words dictionary.

Afterwards, the POS filtering is applied to select only relevant words based on the data exploration and finally all words are normalized by lemmatization.

POS Filtering

Fig. 3 shows how many words belong to each part of speech present in the dataset. Based on this, it is possible to directly discard parts of speech that are insignificant and are frequent in the dataset, such as adpositions, which in Spanish are generally only prepositions, and determiners, which include articles, demonstratives, possessives, and quantifiers.

Different combinations of parts of speech may improve or worse the performance of the ML models. Therefore, in this work the POS considered are Adjective (ADJ), Adverb (ADV), Auxiliary (AUX), Noun (NOUN), Proper Noun (PN), Verb (VERB), Pronoun (PRON), and Subordinating conjunction (Sc). The reason for this is that by testing various combinations of POS, it was determined that the mentioned POS elements were the most relevant for the dataset used.

The combinations selected were i) {NOUN, VERB, PN}, ii) {NOUN, VERB, ADJ, PRON} , iii) {NOUN, VERB, ADJ, AUX, PN}, iv) {NOUN, VERB, ADJ, ADV, Sc}, and v) {NOUN, VERB, ADJ, PN}. In addition, the performance of the ML algorithms are compared with and without the proposed POS combinations.

2.4 Feature Extraction

To extract the features on which the models are trained on, uni-grams and bi-grams are built. Then, these are transformed into TF-IDF vectors.

A different feature space is generated for all combinations of POS filtering and No-POS filtering. In general, by including a POS filtering, the total number of features is decreased (ranges from 97 to 111) in comparison with No-POS filtering (114 features).

2.5 Machine Learning Classifiers

Once data is pre-processed, the dataset is splitted into training and testing subsets. The proportions used for these subsets were 80% and 20%, respectively. It is worth to note that, due to the classes imbalance, an stratification approach must be consider in the splitting of the data. In this work, sample proportions are maintained for both, the training and testing subsets.

On the other hand, classifiers can be grouped into single (DT or SVM) or ensemble (RF or XGBoost) algorithms. The main difference between these is that, the first looks forward to obtain a robust model with a good generalization, while the second achieves a good generalization by combining several classifiers by specific strategies such as bagging or boosting.

Therefore, in this work DT, SVM, RF, and XGBoost are employed for intent classification. From these, DT, RF, and XGBoost are tree-based classifiers. A DT partition the feature space into regions defined by thresholding each feature values. This algorithm is the most interpretable and the most prone to overfitting. RF and XGBoost combine several DTs to classify, improving the variance and the accuracy of the model using clever strategies. In the case of RF, DTs are trained in parallel, each receiving a random subset of data sampled with replacement.

In the case of XGBoost, a series of DT are stacked one after the other, each focusing upon the samples misclassified by the previous learner. Finally, SVM is a robust single ML model that is built upon the structural error minimization principle. SVM can handle noisy non-linear relationships by penalizing errors and mapping to larger feature spaces through the kernel functions.

Furthermore, each ML model was tuned by a grid search cross validation procedure for obtaining the best hyperparameter values. In this work, a 3-fold cross validation procedure was carried out. The parameters tested where: i) for SVM, penalization C (0.1-10), γ ($1e^{-3}$-1), and different kernels (liner, radial, and polynomial); ii) for DT and RF, parameters max_features (log_2 and the square root), max_depth (1-100), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), and in particular for RF, class_weight (stratified or sub-sampling) and number of estimators (10-200); iii) for XGBoost, learning rate (0-1), max_depth (10-100), and number of estimators (10-2000).

2.6 Model Evaluation

To compare the performance of the ML algorithms for classifying intentions from an educational chatbot, multi-classification measures are employed. In this work the Balanced Accuracy (BA), Cohen's Kappa Score (KCS), and Matthews Correlation Coefficient (MCC) are employed.

These measures are useful for assessing the overall performance of a classifier in a multi-class problem. Further, MCC and KCS provide robust measures in the presence of classes imbalance, such as the corpus used in this work.

Table 2. Performance results for intent classification of the educational chatbot corpus.

POS Filtering	SVM			DT			RF			XGBoost		
	MCC	BA	CKS	MCC	BA	CKS	MCC	BA	CKS	MCC	BA	CKS
NOUN, VERB, PN	76.1	53.7	75.2	75.2	**51.7**	74.3	76.9	53.6	75.9	75	**50.6**	74
NOUN, VERB, ADJ, PROP	78.5	54	**77.7**	74.5	48.5	73.4	79.4	53.4	78.5	66.6	39	61.5
NOUN, VERB, ADJ, AUX, PN	76.5	**54.5**	75.7	76.4	51.3	75.3	**80.4**	**55.2**	**79.5**	77	49.5	76
NOUN, VERB, ADJ, ADV, Sc	77.6	52.4	76.6	74.2	48.5	73.1	79.9	54.3	79.1	68.1	39.7	64.3
NOUN, VERB, ADJ, PN	**78.7**	54.3	77.6	**76.8**	51.3	**75.7**	**80.4**	**55.2**	**79.5**	**77.6**	50.4	**76.6**
No-POS filtering	76.5	**54.5**	75.7	**76.8**	51.3	**75.7**	**80.4**	**55.2**	**79.5**	77.5	50.4	76.5

2.7 Software

To implement the machine learning pipeline, python is used. SpaCy, a free open-source Natural Language Processing library, is used for tokenization, stop words removal, POS filtering, and lemmatization. On the other hand, Scikit-learn, a machine learning library, is used for feature extraction and all subsequent steps in the pipeline [4].

3 Experimentation

In this section the results for intents classification for the educational chatbot are presented. Table 2 shows the performance of training each ML classifier with and without POS filtering combinations.

The overall best performing model was RF as shown by BA, MCC, and CKS metrics, followed by SVM, XGBoost, and DT. Regarding, POS filtering we can observe that, the best performing POS combination is {NOUN, VERB, ADJ, PN}. Further, by reducing the number of features employed in the classification the performance of models is not only unaffected, it can even increase algorithms performance.

For instance, in the case of XGBoost and SVM, by reducing the number of features, the performance of the model is increased. In comparison, for DT and RF, reducing the number of features does not improves the accuracy of the models, however, it reduces the computational burden.

Furthermore, Figure 4 presents the confusion matrix for RF. On the x-axis the predicted values are shown, whereas on the y-axis the true labels are presented. Observe that intents corresponding to feel_check, navigation_menu, navigation_microred_caseta, search, and utility_check have in most cases, a perfect classification.

In contrast, intentions such as date_check, end_conversation, gratitude_check, joke_check, time_check, and weather_check have poor identification performance. An interesting fact is that, for weather_check, time_check, joke_check, gratitude_check, and end_conversation are misclassified as a feel_check intention. Similarly, date_check is misclassified with end_conversation, feel_check, joke_check. These points out that, phrases used for all these intentions have many common words worsening classes discrimination.

[4] https://colab.research.google.com/drive/1v3TPPkoyilhwTgXsmw-sm4BKdHIhIj7q?usp=sharing

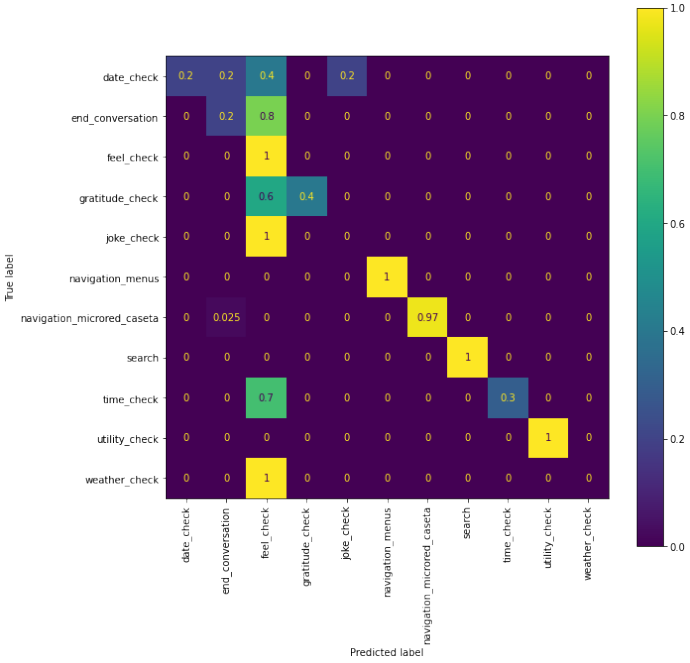


Fig. 4. Random forest confusion matrix.

4 Discussion and Conclusion

Based on a corpus consisting of 1850 phrases from an educational chatbot, a model for intent classification was developed using four different ML algorithms: SVM, DT, RF, and XGBoost. These are trained upon a common pipeline consisting in a pre-processing NLP stage (tokenization, stop–words removal, POS filtering, and lemmatization), feature creation using POS filtering and n-grams, and finally training and evaluating the models using three different metrics (BA, MCC, and CKS).

Results show that RF performs better than all other ML algorithms tested, achieving the highest score in every metric. Applying POS filtering in some algorithms improved their performance, and in others, the same performance is obtained. By applying the POS filtering an improvement in the computational burden of training ML models in larger dimensional spaces can be obtained, since less features are applied to obtain the best results. In particular, POS filtering has the greatest effect when applying it to XGBoost, where it produced the best result for all metrics, and SVM, where it improves the score in MCC and CKS.

Future work considers improving the dataset by increasing the number of phrases for each intent, in particular, those unrepresented in the dataset. Another venue of future work is the usage of word embeddings for feature extraction, and evaluating Deep Learning algorithms, such as DIET or BERT. Finally, the best resulting intent classifier will be implemented in the dialogue manager of an educational chatbot [7].

References

1. Chotirat, S., Meesad, P.: Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning. Heliyon, vol. 7, no. 10 (2021) doi: 10.1016/j.heliyon.2021.e08216
2. Hefny, A. H., Dafoulas, G. A., Ismail, M. A.: Intent classification for a management conversational assistant. In: 2020 15th International Conference on Computer Engineering and Systems (ICCES), pp. 1–6 (2020) doi: 10.1109/ICCES51560.2020.9334685
3. Helmi-Setyawan, M. Y., Maulana-Awangga, R., Rafi-Efendi, S.: Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot. In: 2018 International Conference on Applied Engineering (ICAE), pp. 1–5 (2018) doi: 10.1109/INCAE.2018.8579372
4. Hwang, G. J., Chang, C. Y.: A review of opportunities and challenges of chatbots in education. Interactive Learning Environments, pp. 1–14 (2021) doi: 10.1080/10494820.2021.1952615
5. Larson, S., Leach, K.: A survey of intent classification and slot-filling datasets for task-oriented dialog (2022) doi: 10.48550/arXiv.2207.13211
6. Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., Mars, J.: An evaluation dataset for intent classification and out-of-scope prediction. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1311–1316 (2019) doi: 10.18653/v1/D19-1131
7. Macias-Huerta, P. I., Santamarıa-Bonfil, G., Blanca-Ibañez, M.: CARLA: Conversational agent in virtual reality with analytics. Research in Computing Science, vol. 149, no. 12, pp. 15–23 (2020)
8. Santana, R., Ferreira, S., Rolim, V., de Miranda, P. B., Nascimento, A. C., Mello, R. F.: A chatbot to support basic students questions. In: IV Latin American Conference on Learning Analytics, pp. 58–67 (2021)
9. Schuurmans, J., Frasincar, F.: Intent classification for dialogue utterances. IEEE Intelligent Systems, vol. 35, no. 1, pp. 82–88 (2019) doi: 10.1109/MIS.2019.2954966