1. **INTRODUCTION**

* POSOS is a French based startup related to health care and medication.
* Millions of questions are asked every year about drugs.
* There is a limited number of query types but the same question could be asked in many ways.
* Therefore, understanding what information people expect when asking a question is a great challenge.
* The goal of Posos challenge is about predicting for each question its associated intent.
* Questions must be classified according to a list of 52 different intent categories.
* Each question must fall under one of the intent.
* Given aquestion we should be able to predict the associated intent.
* Leveraging the latest natural language processing technology, POSOS is a French based startup founded in 2017, building a technology able to understand any drug related question and automatically gives a precise answer using multiple data sources whether official or customer based.
* The team combines both medical and scientific knowledge (Doctors of Pharmacy) and engineering skills (ENSMVA AND Epitech).
* By developing this solution, usable by many healthcare stakeholders, their team strives in reducing the drug misuse risks, potentially saving more than 144000 hospitalizations in France every year.

1. **PROBLEM STATEMENT**

To predict the expected intent number to each and every question given to which the question belongs to, which will help people to clarify their doubts regarding the intake of drugs.

1. **DURATION**

The duration of this challenge is from November 16, 2017 to December 12, 2018.

1. **OBJECTIVES**

* Making it possible to immediately answer any drug-related question, enabling for each and everyone a better healthcare
* Increase knowledge about drugs issues and needs related to the prescription, dispensing and use of medicines

1. **METHODOLOGIES**
   1. **PREPROCESSING:**

* Lower casing:

The first pre-processing step is to transform our questions into lower case. This avoids having multiple copies of the same words.

* Punctuation removal:

The next step is to remove punctuation, as it doesn’t add any extra information while treating text data. Therefore removing all instances of it will help us reduce the size of the training data.

* Stop words removal:

Stop words (or commonly occurring words) should be removed from the text data.

* Lemmatization:

Lemmatization is a more effective option than stemming because it converts the word into its root word, rather than just stripping the suffices. It makes use of the vocabulary and does a morphological analysis to obtain the root word. Therefore, **we preferred using lemmatization over stemming.**

1. **MODELS:**

The experiment was carried out on publically available database for prediction of expected answer. The dataset contains total of 10063 records that were divided into two sets, training set (8028) and testing set (2035).

The experiments were carried out by using different supervised machine learning algorithms.

1. DECISION TREE:

Decision tree is a popular classifier which is simple and easy to implement. There is a no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are easier to read and interpret.

1. STOCHASTIC GRADIENT DESCENT:

SGD is indeed a technique that is used to find the minima of a function. SGD Classifier is a linear classifier (by default in sklearn it is a linear SVM) that uses SGD for training (that is, looking for the minima of the loss using SGD).

1. LINEAR SVC:

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyper plane that divides, or categorizes, your data.

4. RANDOM FOREST CLASSIFIER:

**Random forests** or **random decision forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

**7. RESULTS**

|  |  |
| --- | --- |
| **MODELS** | **Accuracy** |
| Sophisticated gradient descent | 0.6083 |
| Linear SVC | 0.6058 |
| Decision tree | 0.5927 |
| Random forest classifier | 0.5728 |

1. **DISCUSSIONS AND CONCLUSION**

* The problem that we addressed is POSOS challenge in which we have to find the appropriate intent number of the questions given, to which they belong to.
  + - * Among all the models used, the accuracy of Sophisticated Gradient Descent is more i.e., 0.6083. So, we preferred to use this model.
* As the model’s accuracy is not more than 90%, we are again going with KDD process to get better result.

1. **REFERENCES**
   * + <https://challengedata.ens.fr/en/home>

* [https://www.posos.fr](https://www.posos.fr/)
* <https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/>