# Task 2 Report

Julio Sánchez de las Heras Martín Consuegra[1,1*], Javier Santana Delgado[1,1*] and Alejandro Riquelme Castaño[1,1*]

*Corresponding author(s). E-mail(s): julio.sanchez6@alu.uclm.es; javier.santana1@alu.uclm.es; alejandro.riquelme1@alu.uclm.es;

**Abstract**

This report consist of supervised learning techniques applied to comments on purchases

## 1 Problem Description

The goal of this task is to create a model that predicts if a comment is related to a **camera** or an **auto**.

## 2 Methods and materials

So, at the beginning we have a dataset with two columns: first column classifies the comments in auto or camera and the second column is the comment itself. In addition, the dataset has 600 entries and a mean number of 530 words.

We will use that dataset for the classification model doing Support Vector Machine and Decision Tree Algorithm. Then, we evaluate the results using several evaluation metrics.

## 3 Experiments and results

### 3.1 Experiments

First of all, we have the data that we are going to work with in a csv file. The first step is preprocessing process and once the preprocessing is done, we will continue with vectorization, feature selection and classification algorithm to create a model that predicts if a comment is related to a camera or an auto.

Related to the preprocessing we begin **removing useless** data such as **parenthesis, at sign, etc**. Then, we **convert capital letters into lowercase letters** and we **lemmatize** all **terms** in order **to reduce the amount of text** and obtain the most meaningful part of each word. Continuing with the process we **remove contractions, repeated words and emoticons**. Eventually, we correct wrong words and save all into a pickle file. At this point, we have the data preprocessed.

Next, we vectorize every opinion by following different configurations such as: **TFIDF, TFIDF + N-grams, TFIDF + N-grams + POS tagging and TFIDF + N-grams + POS tagging + number of words**. TFIDF returns the one word combinations relevance in a document based on a formula. TFIDF + N-grams is similar to TFIDF but in this case we have several words combinations. POS tagging indicates type of words included in a document such as verbs, nouns, etc. Eventually, we visualize the number of words per document that will be relevant for the classification process.

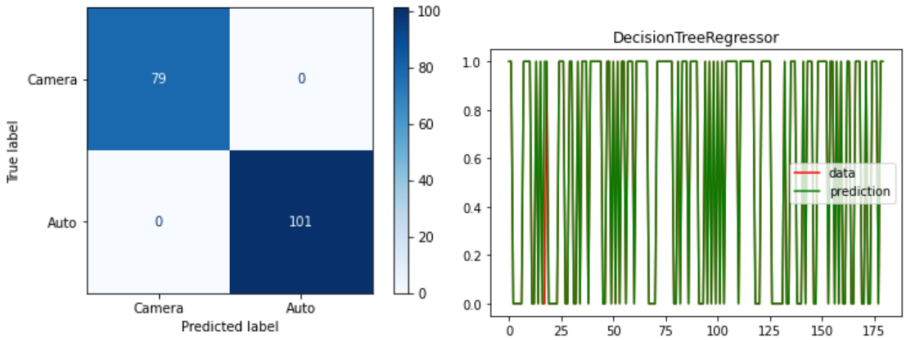
Then, we r**emove 70% of the lowest relevant features**. So, we consider the **30% of the most relevant features using SelectKBest** considering each setup used in vectorization process.

Lastly, we have to do the classification model doing Support Vector Machine and Decision Tree Algorithm for each previous case obtained through SelectKBest.

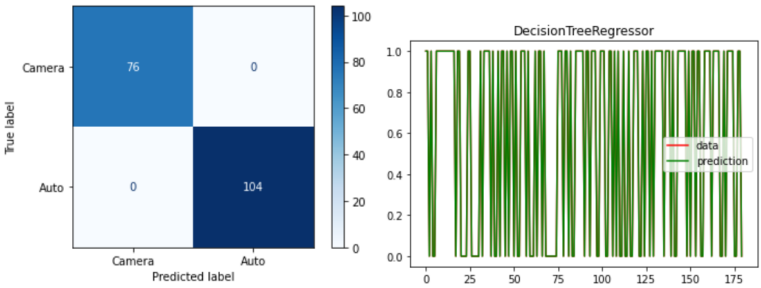
## 3.2 Results

Once the classification algorithm has been done, we analyzed the results.

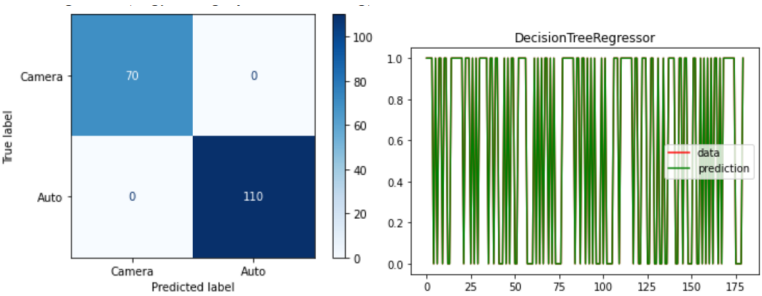
For TF-IDF, we have a **precision, recall, accuracy and f-measure value of one**. We can see the confusion matrix below (left image) and the decision tree obtained (right image).



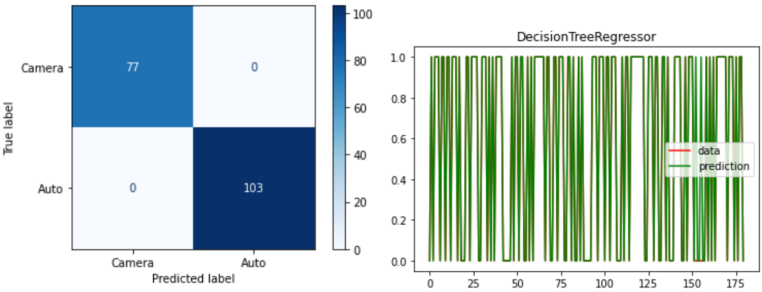
For TF-IDF with N-grams, we have a **precision, recall, accuracy and f-measure value of one**. We can see the confusion matrix below (left image) and the decision tree obtained (right image).




We repeat the same operation for TF-IDF with N-grams and POS Tagging, we have a **precision, recall, accuracy and f-measure value of one**. We can see the confusion matrix below (left image) and the decision tree obtained (right image).




For TF-IDF with N-grams, POS Tagging and number of words (right image), we have a **precision, recall, accuracy and f-measure value of one**. We can see the confusion matrix below (left image) and the decision tree obtained (right image).



# 4 Conclusion

As we have seen in this task, text is becoming more and more important for machine learning processes because people have tools like Twitter to share their opinions.

This information can be very relevant to make decisions in a company. However, these types of features are more complex to process than numerical features. Therefore, a preprocessing step is needed to deal with this information by correcting misspelled words, removing emoticons, etc. Due to the fact that the writing of different texts varies a lot depending on the studies of each person.

That is why in this task we try to eliminate all the above mentioned in order to obtain texts clean of unnecessary data and we also manage to bring back the different words that are used (verbs, nouns...). With this clean data we use it for the different classification models.

To conclude, we are a group that belongs to another intensification but by working so many hours on this task, we have finally found it interesting and we have a very good current thinking about people who process large amounts of data.