Homework 3 – Artificial Intelligence

Document Summaries

Alexandra DINIŞOR – 342C4

1. Naïve Bayes classifier for Text categorization – Performance analysis

- **Precision**

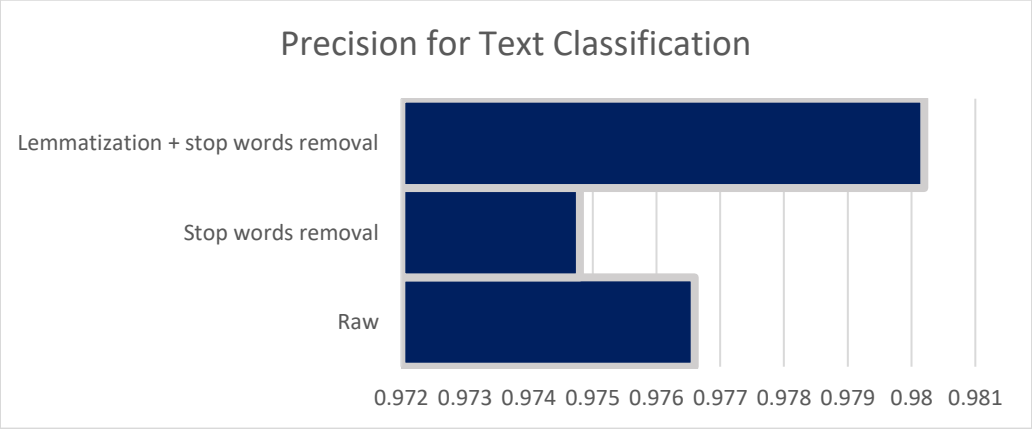

Fig. 1

Applying different text preprocessing steps (stop words removal, word lemmatization) helps reduce words (for instance: “is”, “the”) that are irrelevant for a news category.

Due to dimensionality reduction of the representation space, lemmatization and stop words removal make the most precise text classifier.

- **Recall**

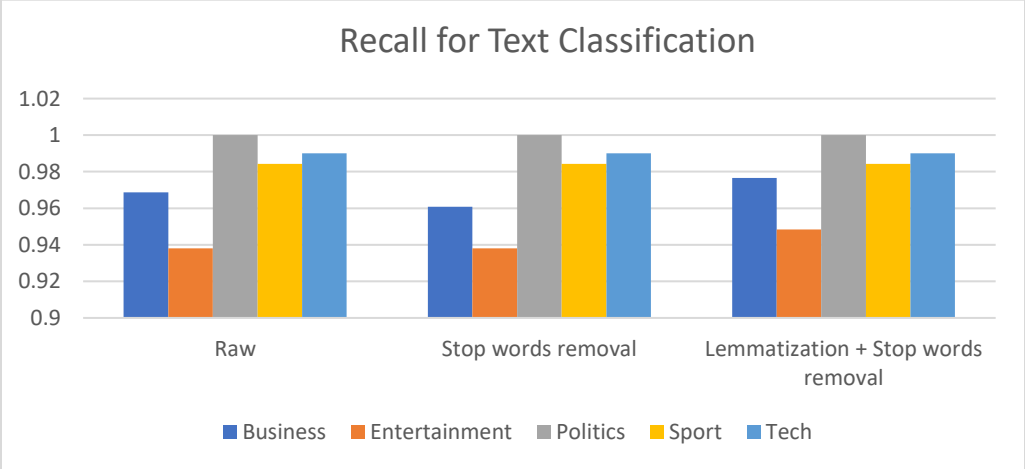

Fig. 2

The model tends to perform better on all categories (business, entertainment, sport, politics, tech) when using all the preprocessing steps - lemmatization and removing of the stop words. They are good ways of improving the accuracy of the classifier, when measuring recall it gets the biggest percentage (98,2%).

2. Extractive summarization on Naïve Baye classifier - Performance analysis

- **Unigrams**

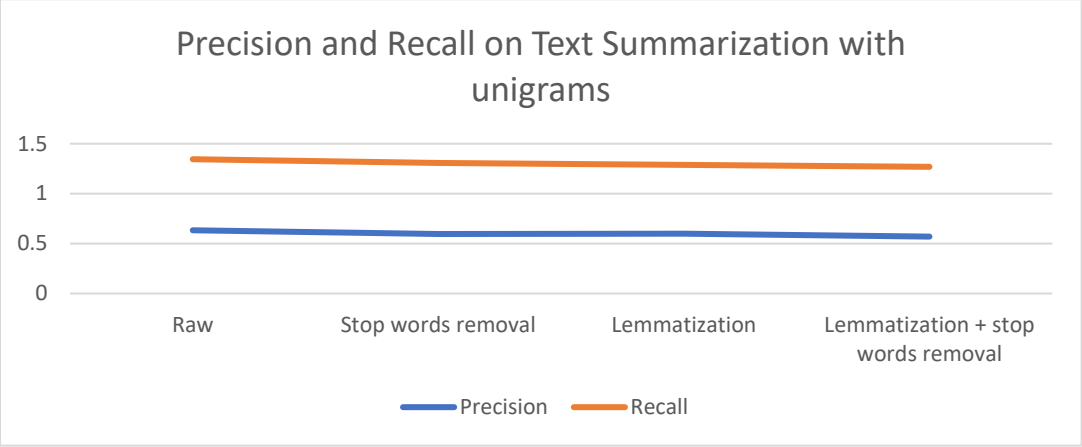

Fig. 3

Measuring Precision and Recall on unigrams in the context of ROUGE-1, the model performs the best on raw input (precision=63,3%, recall=71,1%)

- **Unigrams vs. Bigrams**

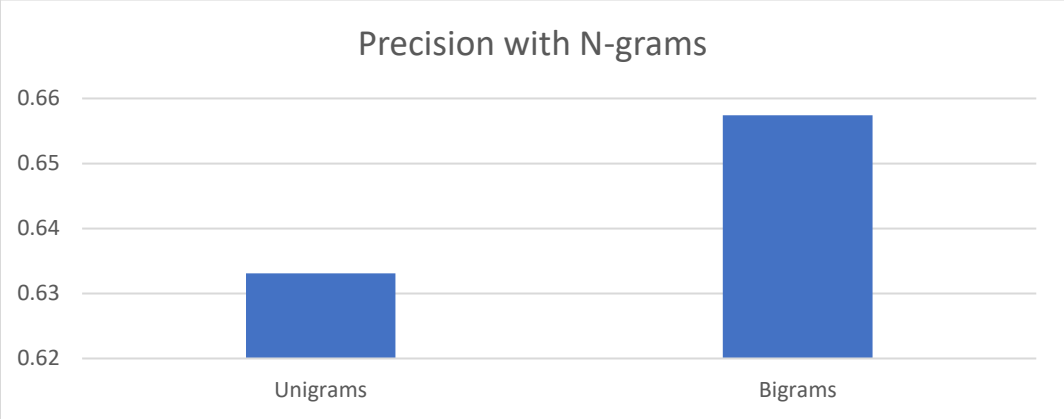

Fig. 4

The largest probability that a sentence selected by a summarizer matches with a sentence belonging to the extractive summary was obtained by the summarizer with Naive Bayes classifier on raw input with no preprocessing method. The value of precision is significantly the highest with unigrams (63,3%) as well as with bigrams (65,7%).

3. Confusion matrix without normalization for the Naïve Bayes Multiclass classifier

Predicted values	Class Labels	business	entertainment	politics	sport	tech
	business	**124**	**1**	**0**	**0**	**1**
	entertainment	**0**	**91**	**0**	**1**	**0**
	politics	**2**	**3**	**104**	**1**	**0**
	sport	**0**	**0**	**0**	**126**	**0**
	tech	**2**	**2**	**0**	**0**	**99**
Actual values						

Fig. 5 Raw dataset.

Predicted values	Class Labels	business	entertainment	politics	sport	tech
	business	**123**	**1**	**0**	**0**	**1**
	entertainment	**0**	**91**	**0**	**1**	**0**
	politics	**2**	**3**	**104**	**1**	**0**
	sport	**0**	**0**	**0**	**126**	**0**
	tech	**3**	**2**	**0**	**0**	**99**
Actual values						

Fig. 6 Dataset with stop words elimination.

Predicted values	Class Labels	business	entertainment	politics	sport	tech
	business	**125**	**1**	**0**	**0**	**1**
	entertainment	**0**	**92**	**0**	**1**	**0**
	politics	**1**	**3**	**104**	**1**	**0**
	sport	**0**	**0**	**0**	**126**	**0**
	tech	**2**	**1**	**0**	**0**	**99**
Actual values						

Fig. 7 Dataset with lemmatization and stop words elimination.

The database was composed of 2225 documents from BBC News, test set being selected from the original document base and representing 25% of data set.