YouTube spam detection

Artificial Intelligence for CyberSecurity

July 12, 2023

Alessandro Zanatta

University of Pisa

Dataset and goal

Total of 1956 YouTube comments from 5 different (famous) musical videos[1] with the following features:

- Comment ID
- Author
- Date
- Content
- Class

[1] https://archive.ics.uci.edu/dataset/380/youtube+spam+collection

Dataset and goal

Total of 1956 YouTube comments from 5 different (famous) musical videos[1] with the following features:

- Comment ID
- Author
- Date
- Content
- Class

Goal: recognize and differentiate between legitimate (ham) comments and spam comments!

[1] https://archive.ics.uci.edu/dataset/380/youtube+spam+collection

Example entries

	COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
0	LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: ...	1
1	LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A	adam riyati	2013-11-07T12:37:15	Hey guys check out my new channel and our firs...	1
2	LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	1
3	z13jhp0bxqncu512g22wvzkasxmvvzjaz04	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^_^	1
4	z13fwbwp1oujthgqj04chlngpvzmtt3r3dw	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtWQ Check this out .	1
...	...	...	...	...	...
1951	_2viQ_Qnc6-bMSjqyL1NKj57ROicCSJV5SwTrw-RFFA	Katie Mettam	2013-07-13T13:27:39.441000	I love this song because we sing it at Camp al...	0
1952	_2viQ_Qnc6-pY-1yR6K2FhmC5i48-WuNx5CumlHLDAI	Sabina Pearson-Smith	2013-07-13T13:14:30.021000	I love this song for two reasons: 1.it is abou...	0
1953	_2viQ_Qnc6_k_n_Bse9zVhJP8tJReZpo8uM2uZfnzDs	jeffrey jules	2013-07-13T12:09:31.188000	wow	0
1954	_2viQ_Qnc6_yBt8UGMWyg3vh0PulTqcqyQtdE7d4Fl0	Aishlin Maciel	2013-07-13T11:17:52.308000	Shakira u are so wiredo	0
1955	_2viQ_Qnc685RPw1aSa1tfrIuHXRvAQ2rPT9R06KTqA	Latin Bosch	2013-07-12T22:33:27.916000	Shakira is the best dancer	0

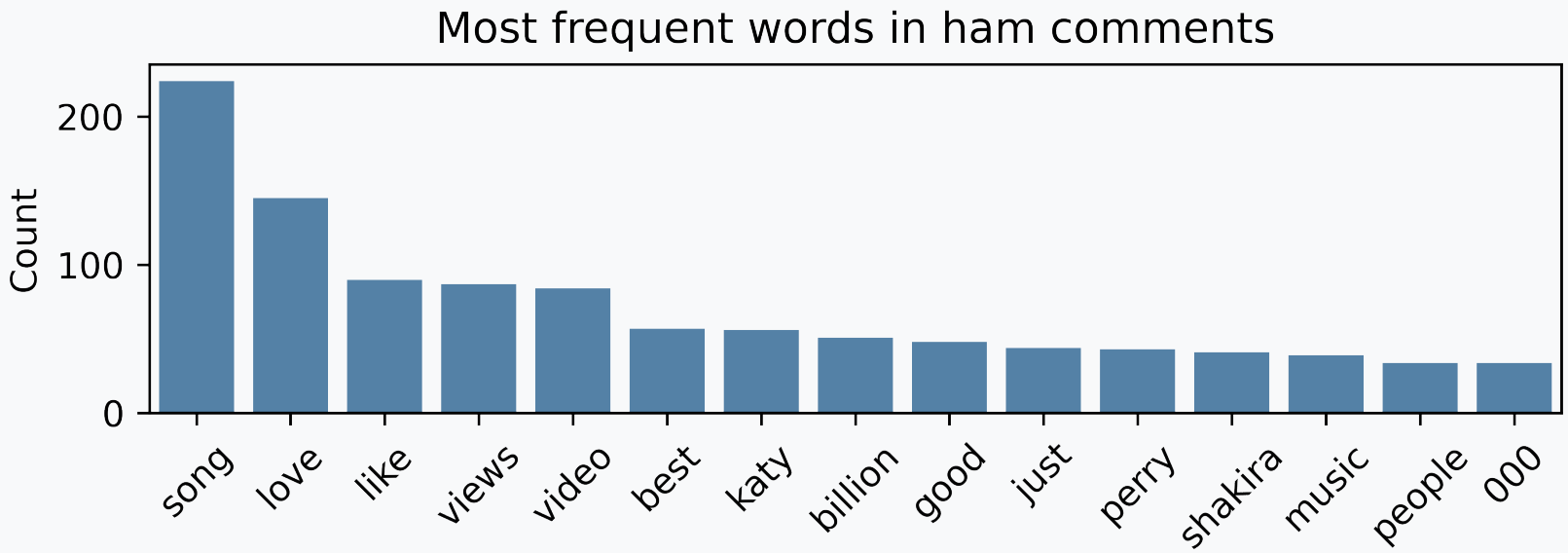
Class equal to one indicates spam!

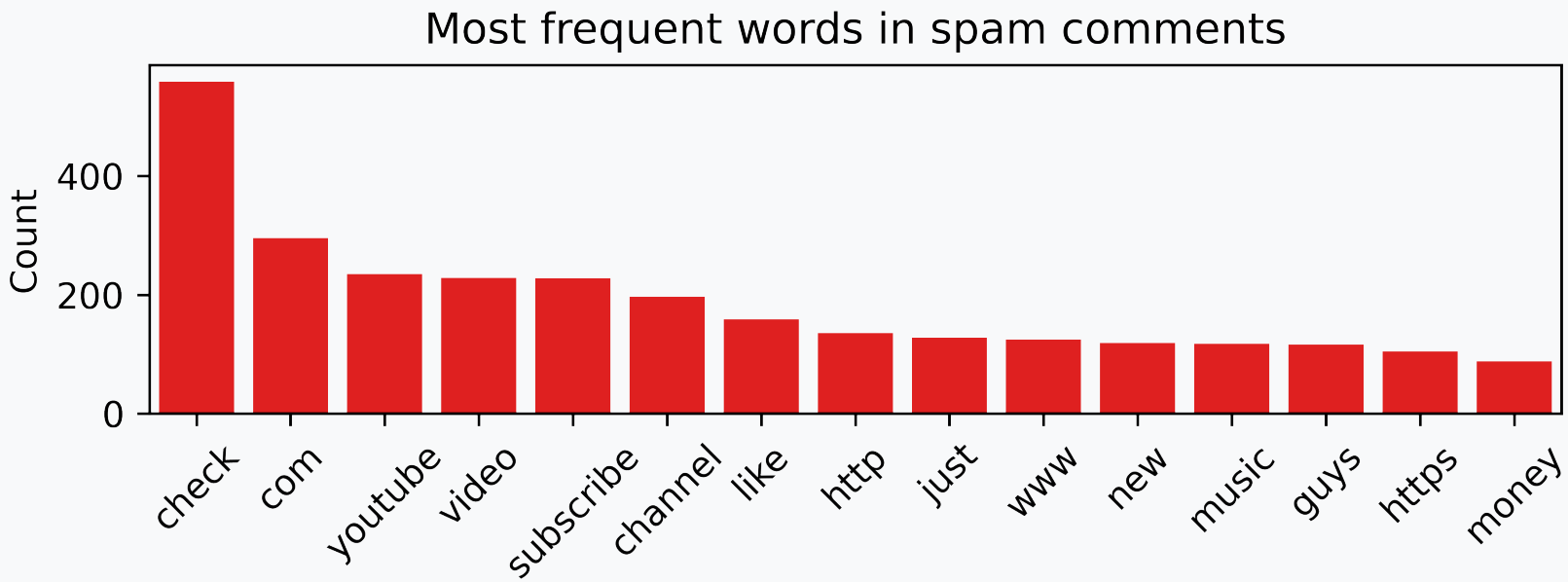
Word cloud

Ham word cloud

Spam word cloud

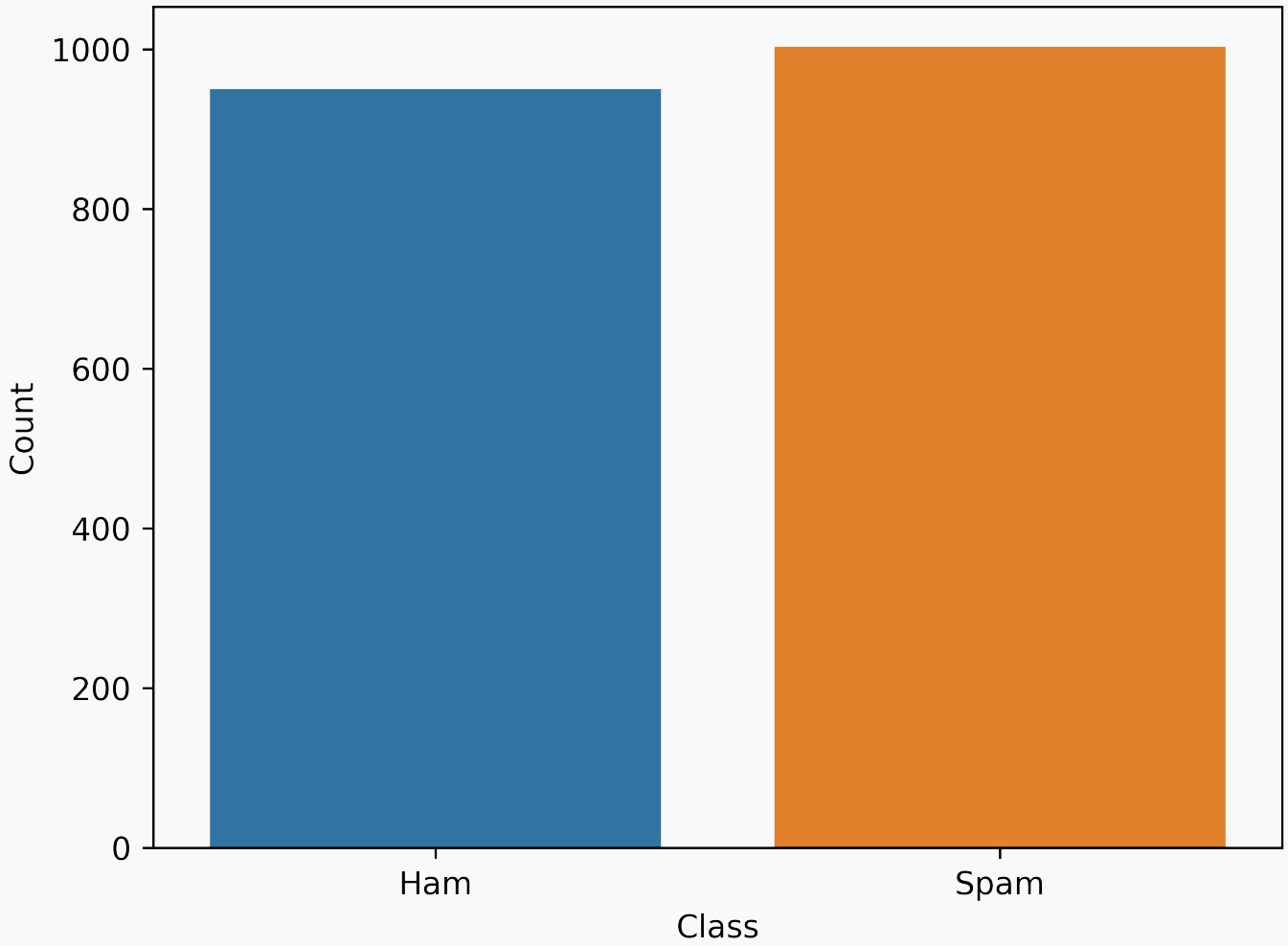
Most frequent words





Balanced dataset

Classes are (basically) balanced!



Cleaning

- Checked for null values (only a few in date)
- Removed useless features (comment ID, author and date)
- Removed duplicates (only 3 duplicates, which affected the balance positively)
- Replaced HTML tags and entities in comments (e.g. replaced
 with \n)

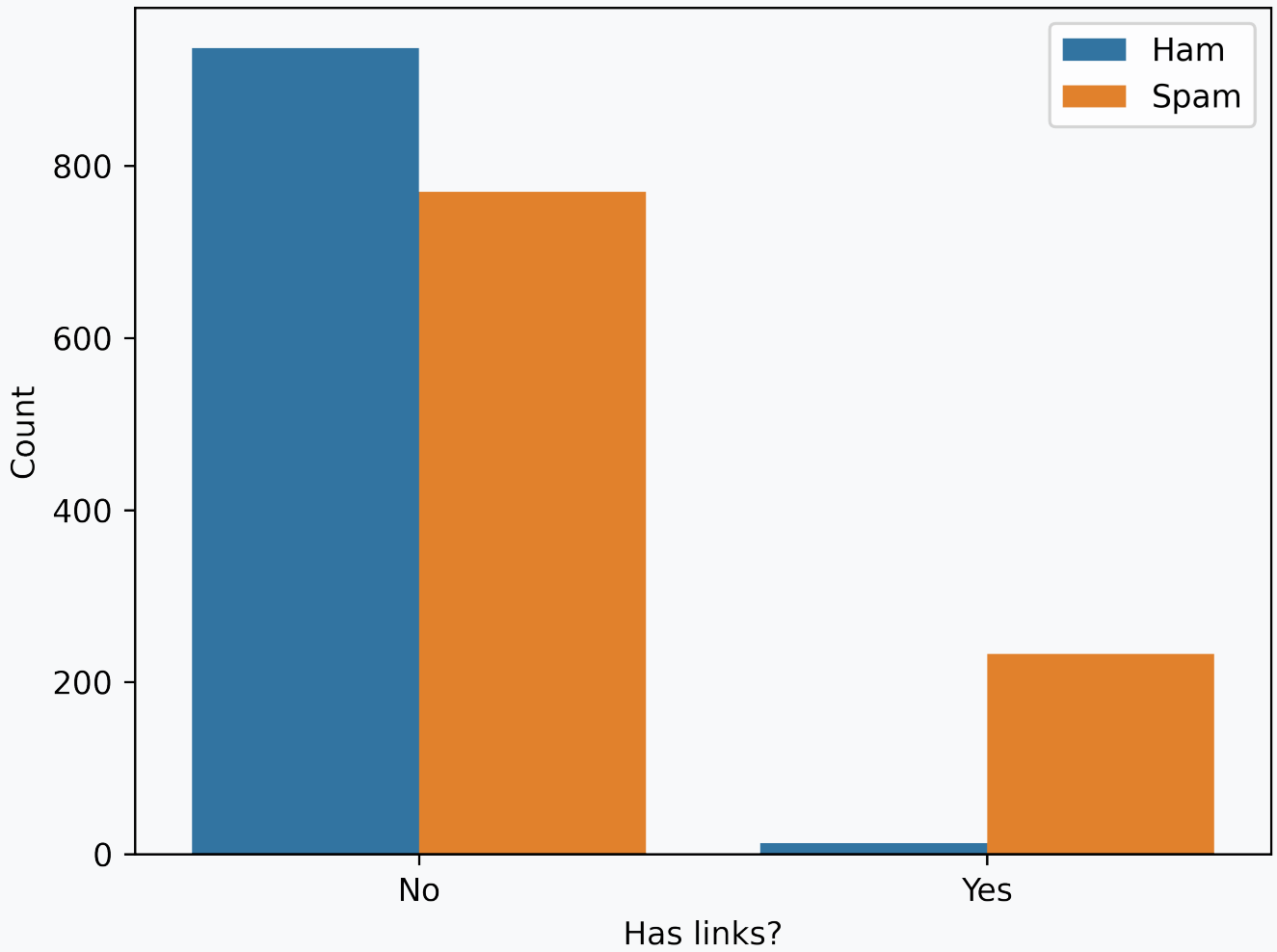
Adding useful features

Tried to extract possible features that may indicate a spamming behaviour:

- Links
- YouTube links (spamming one's channel)
- Use of non-ASCII characters (e.g. emojis)
- Number of characters words, and sentences
- Number of uppercase letters

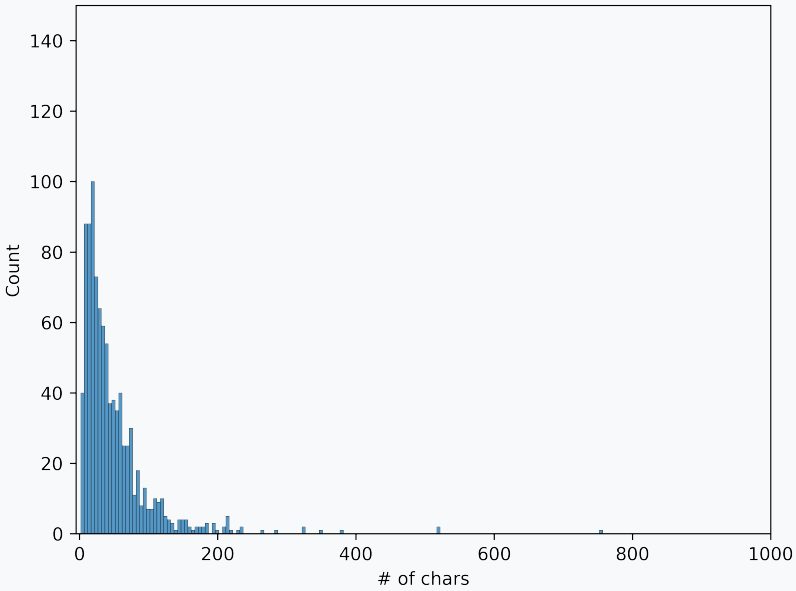
Adding useful features - Links

Presence of links may indicate spam:

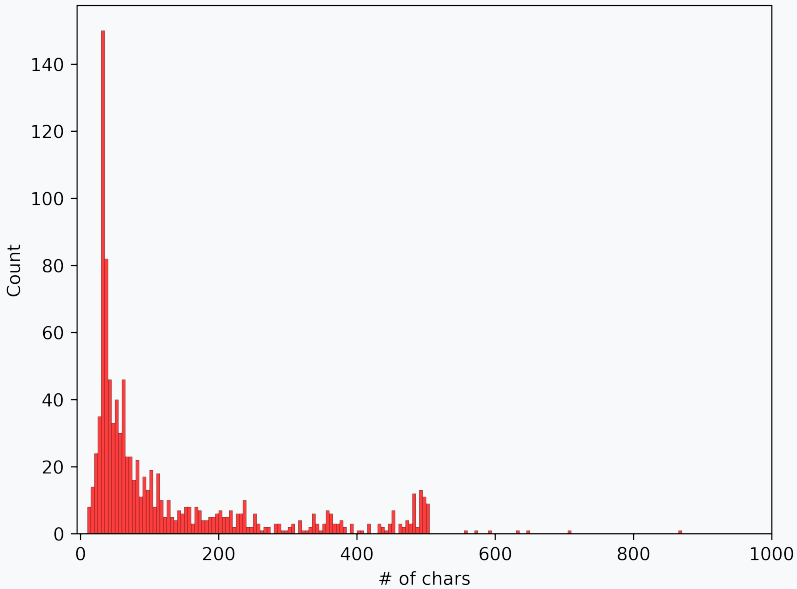


Adding useful features - Characters

Spam comments have a much higher peak, a longer tail, and a second smaller peak at about 500 characters.

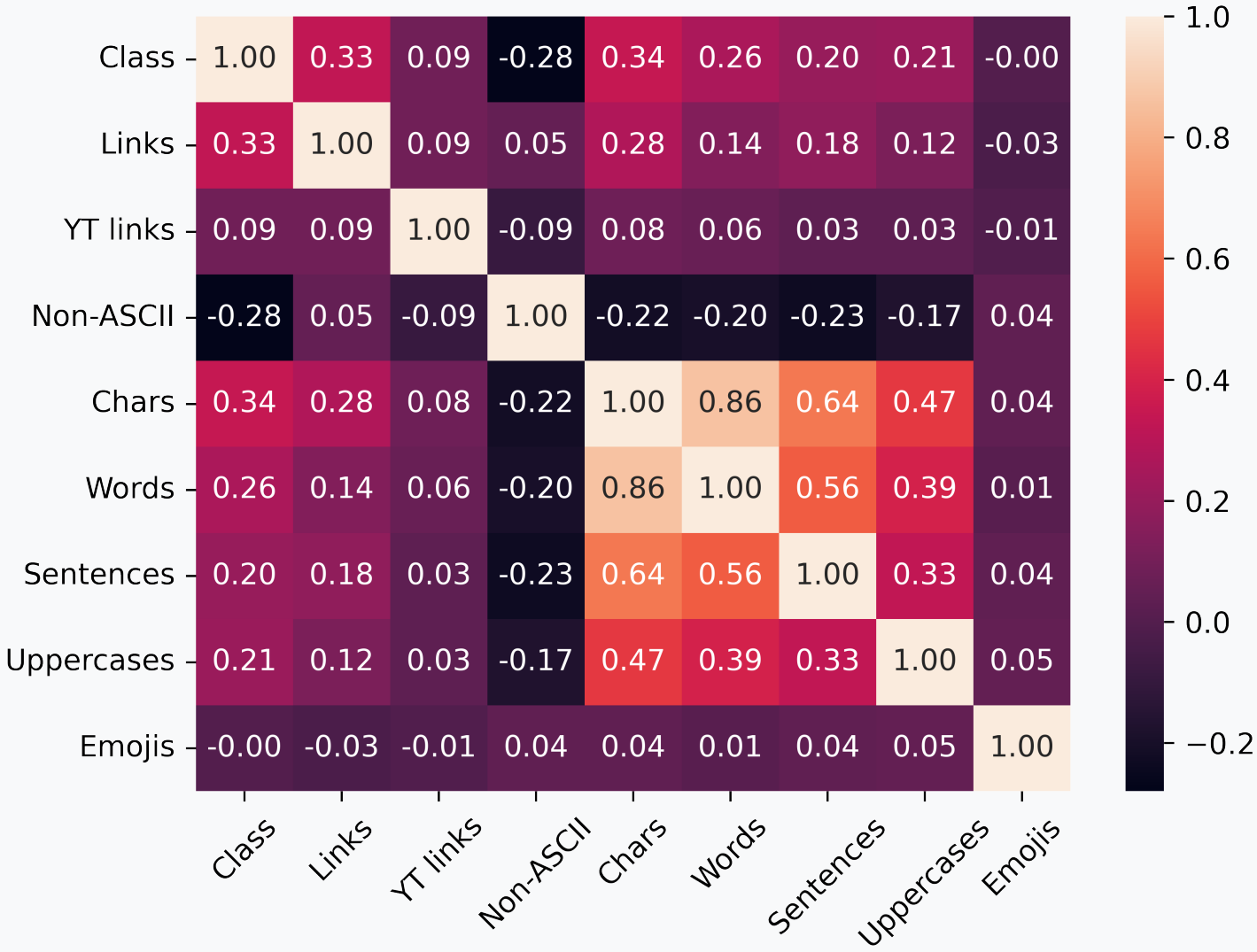


Ham



Spam

Adding useful features - Heatmap



Approach

Classification has been performed with 4 classifiers:

- Support Vector Machine
- Multinomial naïve Bayes
- Decision tree
- Random tree

and with 3 different preprocessings:

- Stemming with Porter stemmer
- Stemming with Snowball stemmer
- Lemmatization

Performance evaluation

Results obtain from K-fold (10 folds):

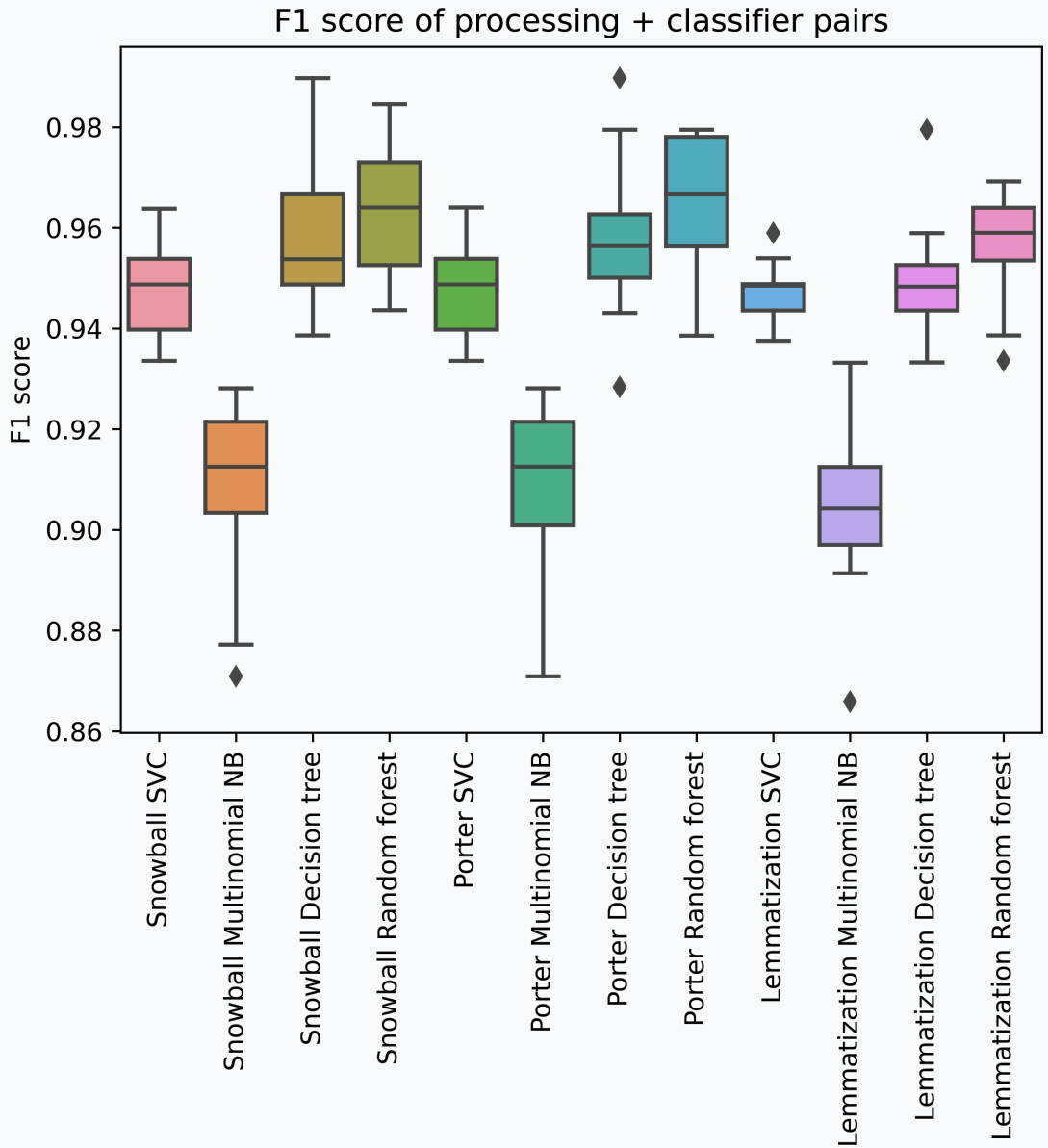
	SVM	Multinomial NB	Decision tree	Random forest
Snowball	0.948	0.908	0.957	0.964
Porter	0.949	0.907	0.957	0.962
Lemmatization	0.947	0.903	0.949	0.958

F1 score

	SVM	Multinomial NB	Decision tree	Random forest
Snowball	0.948	0.908	0.958	0.964
Porter	0.949	0.907	0.96	0.964
Lemmatization	0.947	0.904	0.949	0.961

Accuracy

Performance evaluation



Testing the null hypothesis

The random forest classifier has performed the best, whereas the preprocessing has a much smaller impact on final results. Wilcoxon test can be used to determine if there is a statistical difference between preprocessing methods (with a fixed classifier).

Preprocessing pair	**P-value**
Snowball - Porter	0.4962
Snowball - Lemmatization	0.1934
Porter - Lemmatization	0.1055

Random forest with different preprocessing

Using the conventional acceptance of statistical significance at 0.05 (5%), we refute the null hypothesis: *the difference between the three proprocessing methods is not statistically significant!*

Performance evaluation - Other results

In general, Wilcoxon test allows determining that in this dataset, for the tested classifiers and preprocessing algorithms:

- The use of a different preprocessing is usually not significant, but it is for the decision tree classifier
- The use of a different classifier is almost always significant

Conclusions and future work

Conclusions:

- The best classifier turned out to be the Random forest model. As the classifier has very good performances, the initial goal can be considered achieved
- The use of K-fold and of the Wilcoxon test ensure that results are statistically significant

Conclusions and future work

Conclusions:

- The best classifier turned out to be the Random forest model. As the classifier has very good performances, the initial goal can be considered achieved
- The use of K-fold and of the Wilcoxon test ensure that results are statistically significant

Improvements/future work:

- Trying other algorithms and/or preprocessing methods, which may lead to even higher performances
- The dataset is not large at all. To completely ensure that results can be trusted, it would be needed to use a much larger dataset (possibly with at least tens of thousand of comments). Unfortunately, no such dataset was found online