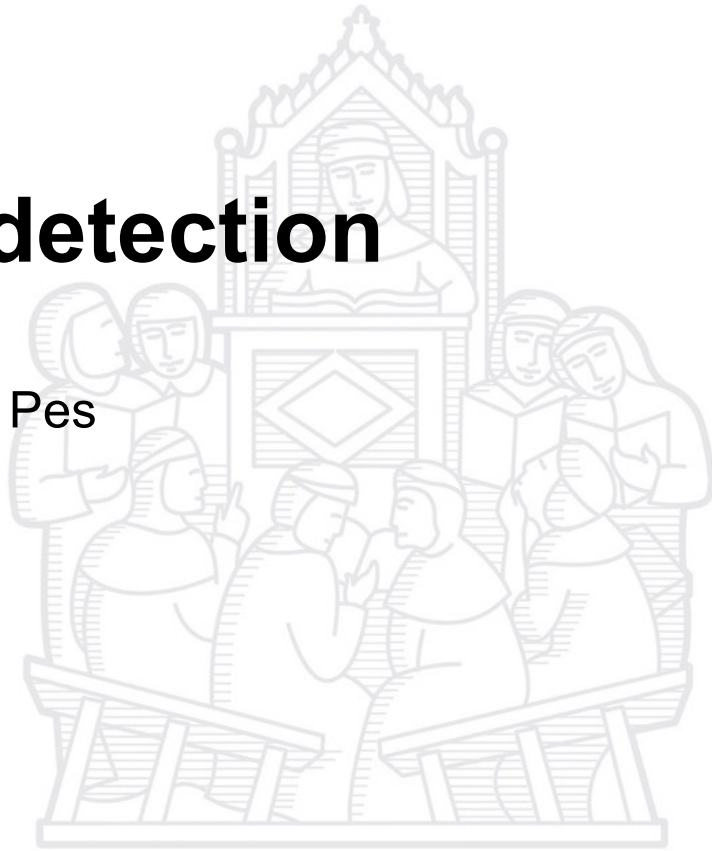
Classifiers for spam detection

Giulio Lanza, Giaime Paolo Pes

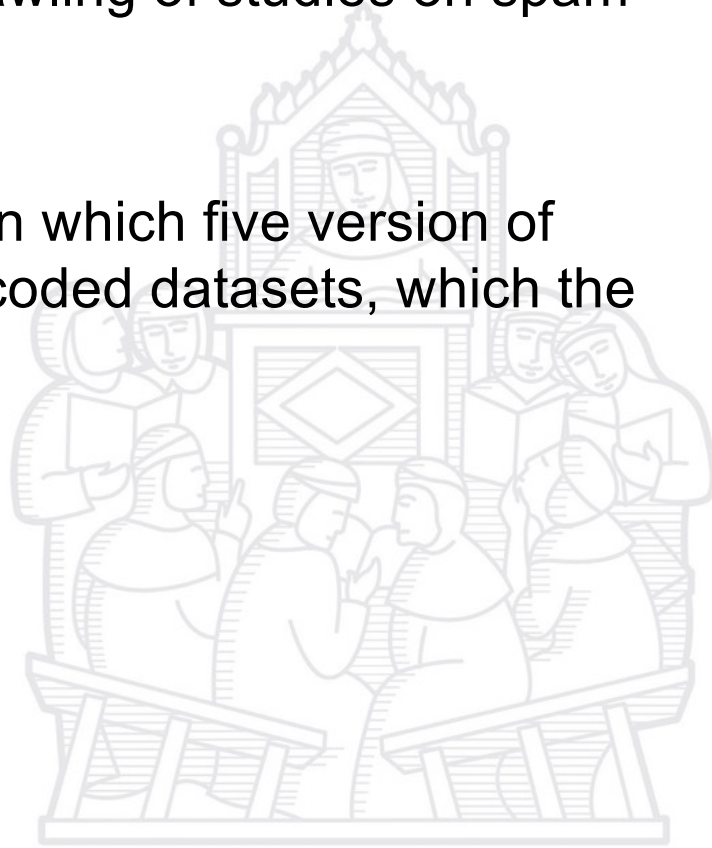


Sant'Anna
School of Advanced Studies – Pisa

Context

- The publication of the Enron Corpus, a database of 600,000 emails from the Enron Corporation, allowed a sprawling of studies on spam email detection

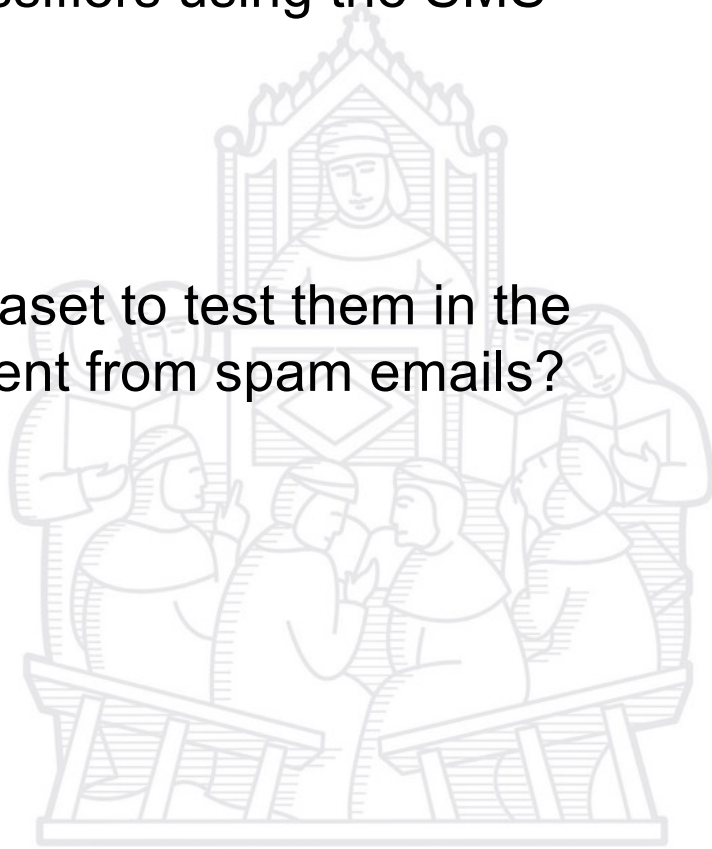
- One such example is Metsis et al. (2006), in which five version of Naive Bayes are compared on six, non-encoded datasets, which the authors provide



Goals

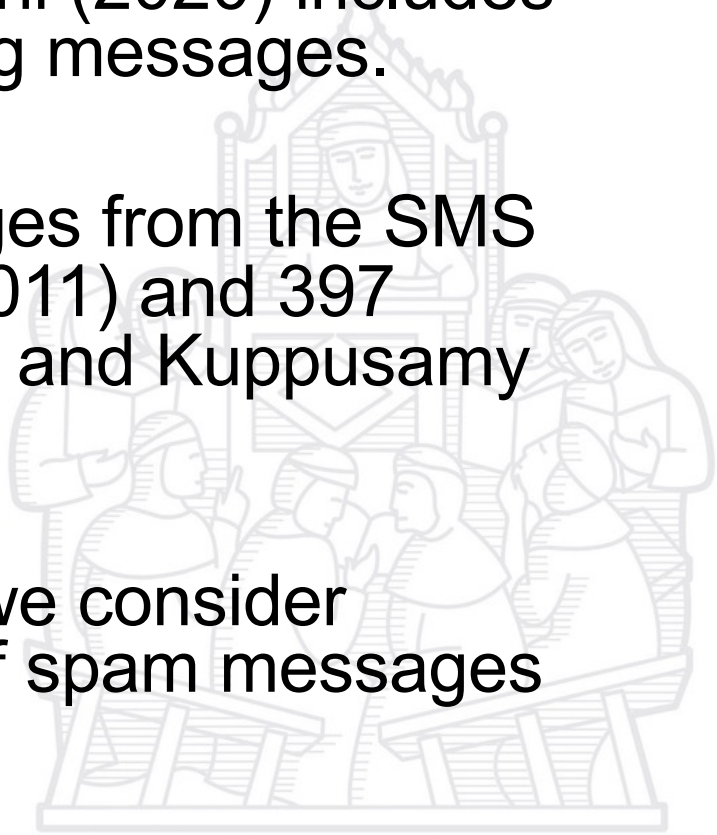
- Taking inspiration from Metsis et al. (2006), we're interested in comparing the performance of different classifiers using the SMS Dataset studied in Mishra et Soni (2020).

- We also train classifiers on the Enron 1 dataset to test them in the SMS dataset: are spam SMS notably different from spam emails?

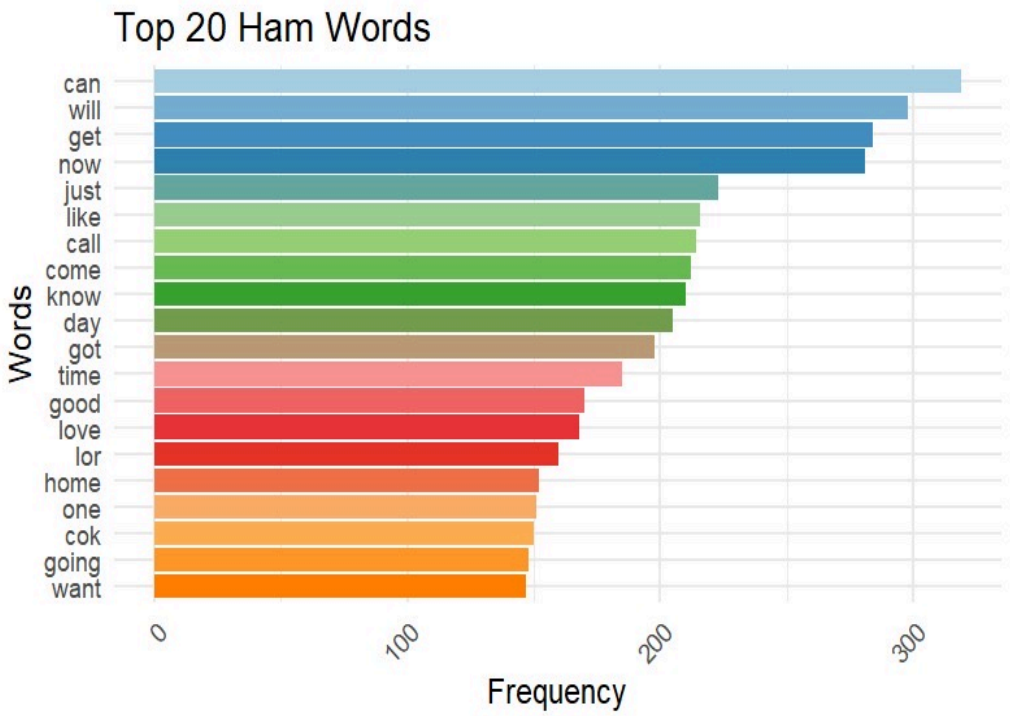


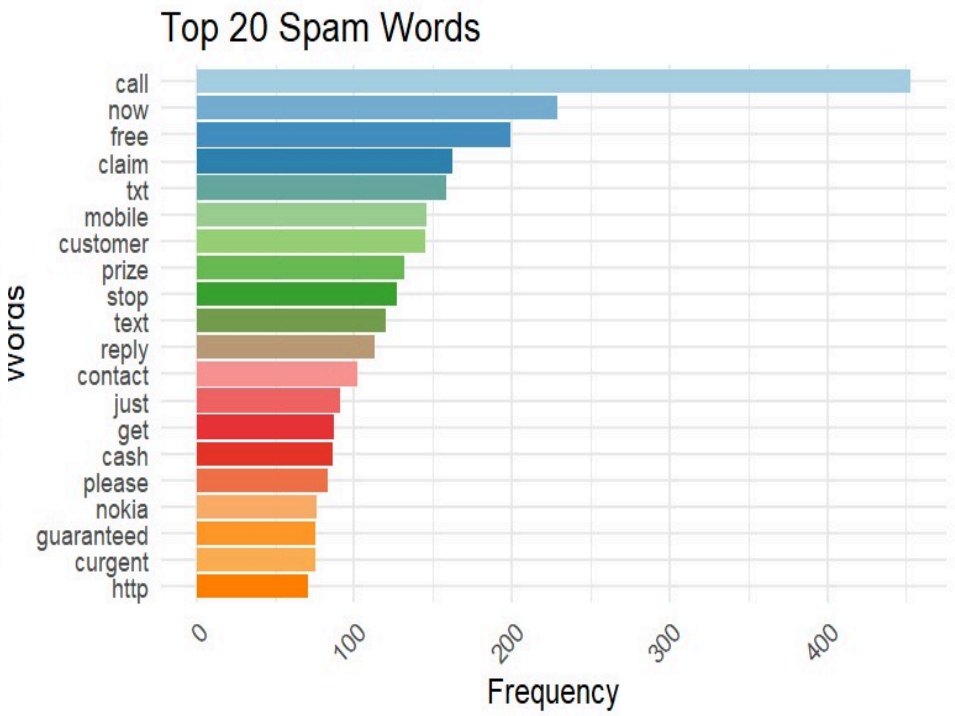
Dataset Description

- The SMS Dataset by Mishra et Soni (2020) includes recent forms of spam and smishing messages.
- It is made up of 5,574 text messages from the SMS Spam dataset by Almeida et al. (2011) and 397 smishing messages from Sonowal and Kuppusamy (2020)
- For the purpose of this research, we consider smishing messages as a subset of spam messages



Overview


Top 20 Ham Words
Words
can
will
get
now
just
like
call
come
know
day
got
time
good
love
lor
home
one
cok
going
want
0
100
200
300
Frequency


Top 20 Spam Words
Words
call
now
free
claim
txt
mobile
customer
prize
stop
text
reply
contact
just
get
cash
please
nokia
guaranteed
curgent
http
0
100
200
300
400
Frequency

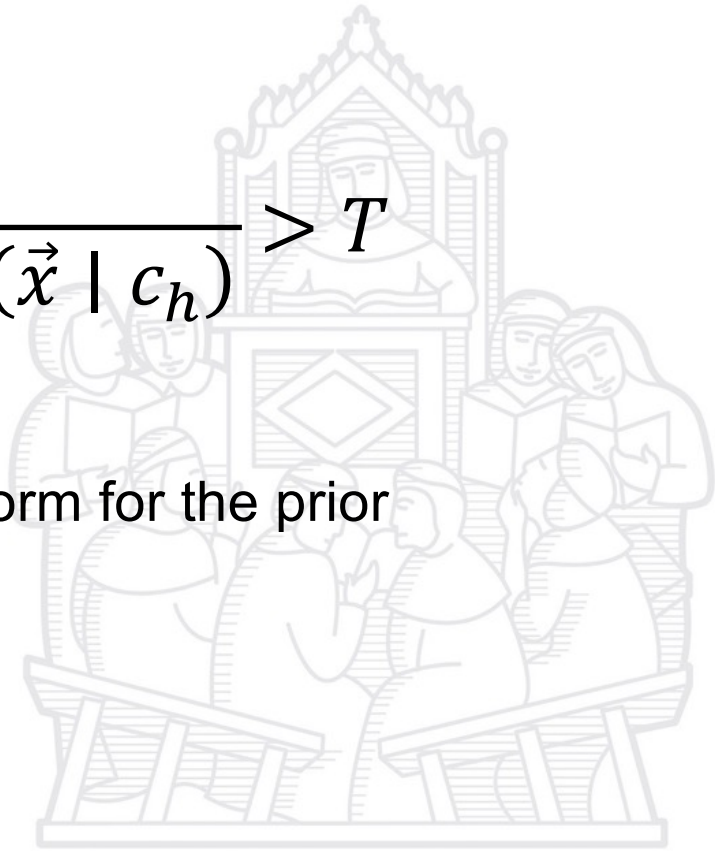
Heterogeneity across classes

Naive Bayes Classifiers

From Bayes' theorem, classifying a message as spam means computing:

$$\frac{p(c_s) \cdot p(\vec{x} \mid c_s)}{p(c_s) \cdot p(\vec{x} \mid c_s) + p(c_h) \cdot p(\vec{x} \mid c_h)} > T$$

The challenge is to find an appropriate form for the prior



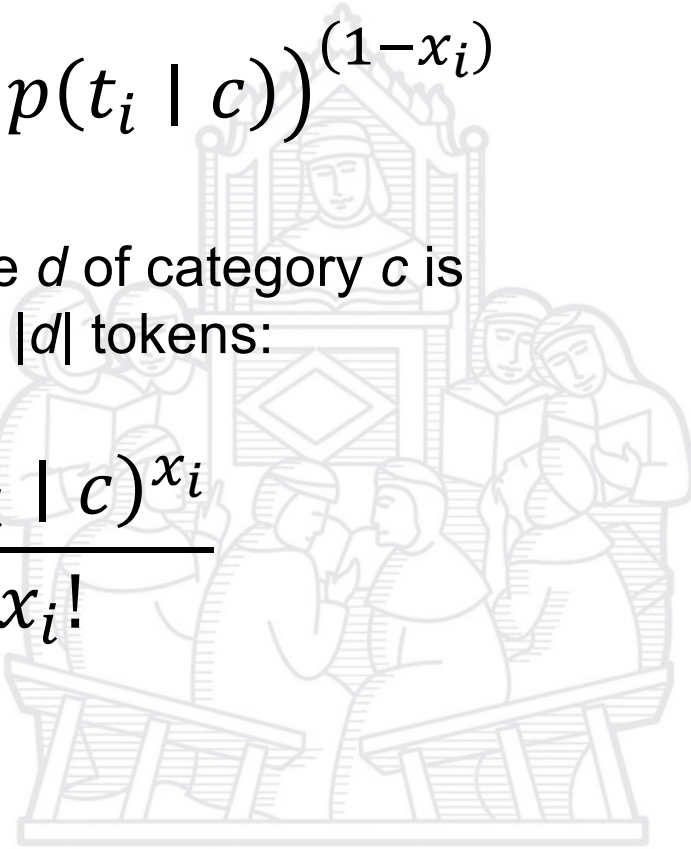
Multi-variate and multinomial Bernoulli

The multi-variate Bernoulli NB treats each message *d* as a set of tokens, containing each t_i that occurs in *d*

$$p(\vec{x} \mid c) = \prod_{i=1}^{m} p(t_i \mid c)^{x_i} \cdot \left(1 - p(t_i \mid c)\right)^{(1-x_i)}$$

In the multinomial NB model each message *d* of category *c* is seen as the result of picking independently |*d*| tokens:

$$p(\vec{x} \mid c) = p(|d|) \cdot |d|! \cdot \prod_{i=1}^{m} \frac{p(t_i \mid c)^{x_i}}{x_i!}$$



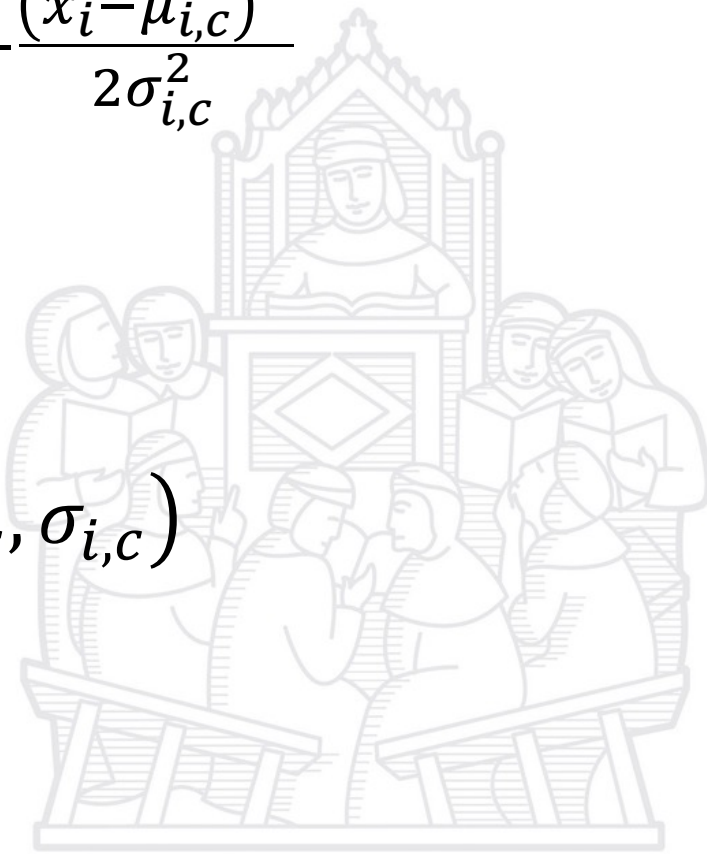
Multi-variate Gauss NB

Finally, the Gauss NB makes the assumption that tokens follow a normal distribution in each class

$$g(x_i; \mu_{i,c}, \sigma_{i,c}) = \frac{1}{\sigma_{i,c}\sqrt{2\pi}} e^{-\frac{(x_i-\mu_{i,c})^2}{2\sigma_{i,c}^2}}$$

Hence:

$$p(\vec{x} \mid c) = \prod_{i=1}^{m} g(x_i; \mu_{i,c}, \sigma_{i,c})$$

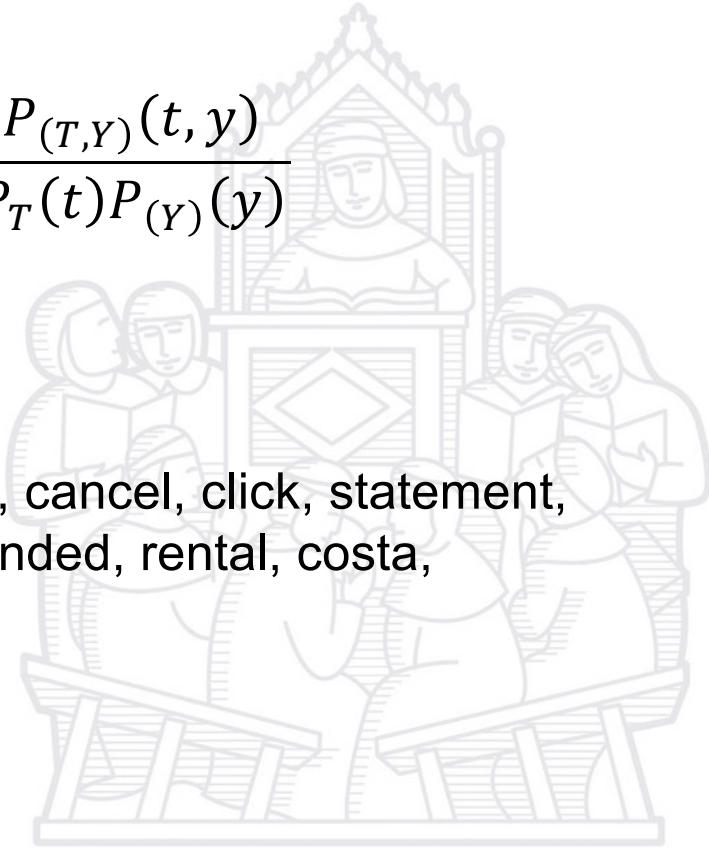


Attribute selection

Following common practice in text classification, we use mutual information to capture the information gain of each token:

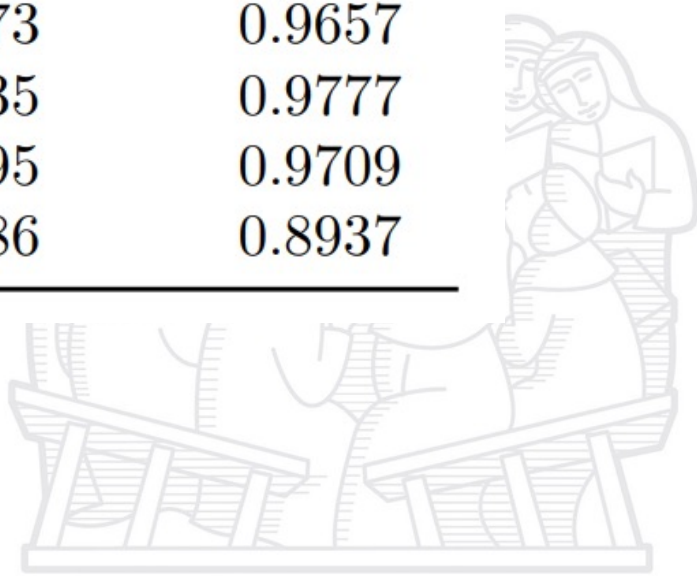
$$I(T;Y) = \sum_{t \in T} \sum_{y \in Y} P_{(T,Y)}(t,y) \log \frac{P_{(T,Y)}(t,y)}{P_T(t) P_{(Y)}(y)}$$

A few of the selected attributes:
anytime, valued, latest, address, bus, advise, tried, cancel, click, statement, gonna, city, black, points, bslvyl, weekends, suspended, rental, costa, asking…



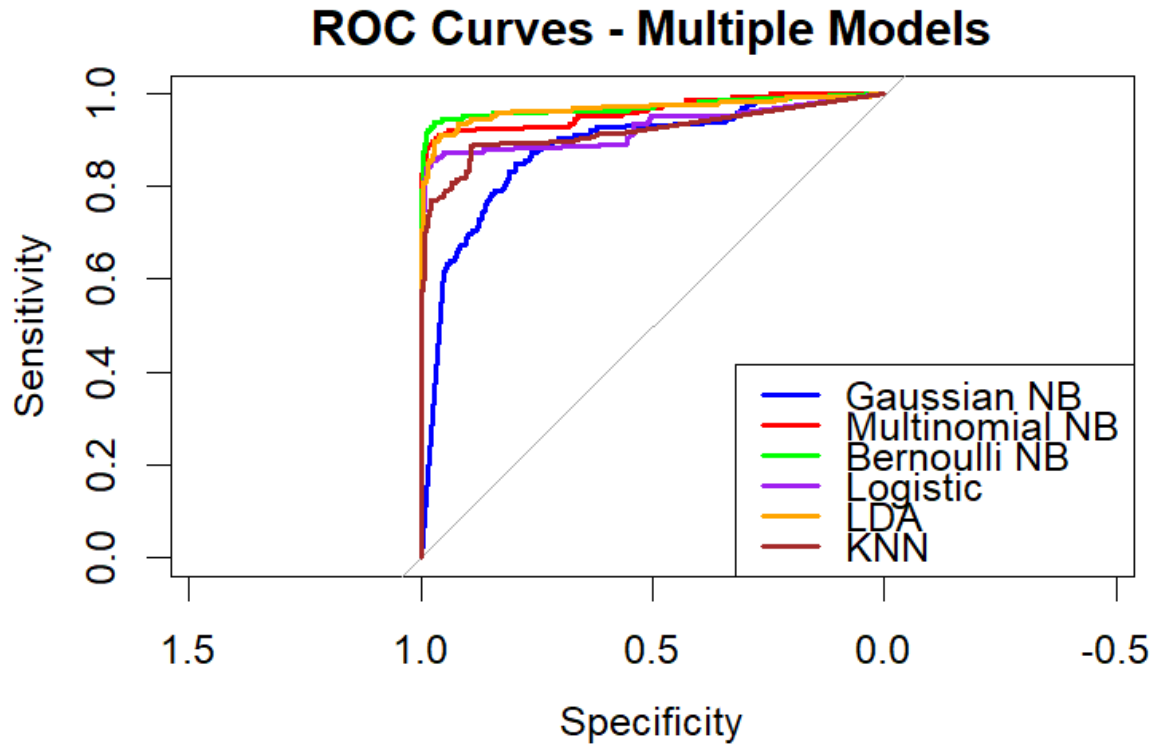
Summary of statistics

Model	Sensitivity	Specificity	Accuracy
KNN	1.0000	0.4578	0.9229
LDA	0.9940	0.8072	0.9674
Logistic	0.9870	0.8373	0.9657
Bernoulli	0.9950	0.8735	0.9777
Multinomial	0.9860	0.8795	0.9709
Gaussian	0.9361	0.6386	0.8937



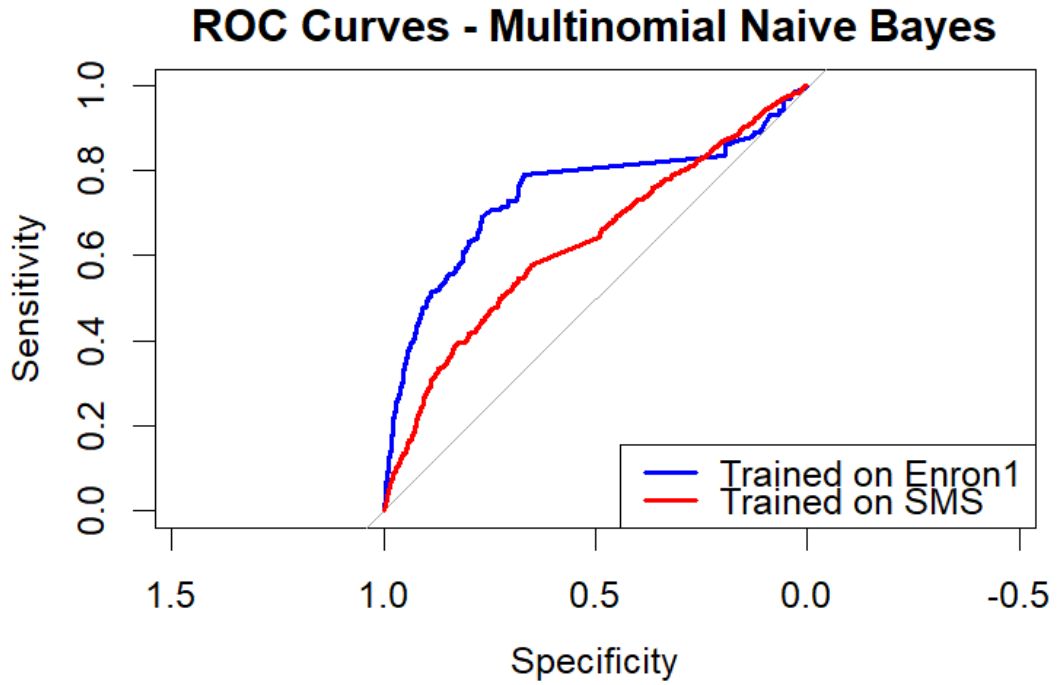
ROC curves

Cross-validating with 10 folds yields promising results, especially for the simple Bernoulli NB classifier



Combining Enron1 and SMS

We train the Bernoulli NB model on Enron1 and then test it on the SMS dataset, and vice versa



Bibliography

- Almeida, T. A., Gómez Hidalgo, J. M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: new collection and results. In ACM Symposium on Document Engineering.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd Edition). Springer, Berlin.
- Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam Filtering with Naive Bayes - Which Naive Bayes?. In CEAS.
- Mishra, S., & Soni, D. (2020). Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis. Future Generation Computer Systems.
- Sonowal, G., & Kuppusamy, K.S. (2020). PhiDMA – A phishing detection model with multi-filter approach. Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 1.