Spam Filtering- YouTube Comment

Team Members-
Bharti Goel,
Jaynab Khatun,
Mounika Dudipala,
Varsha Reddy Yasa.

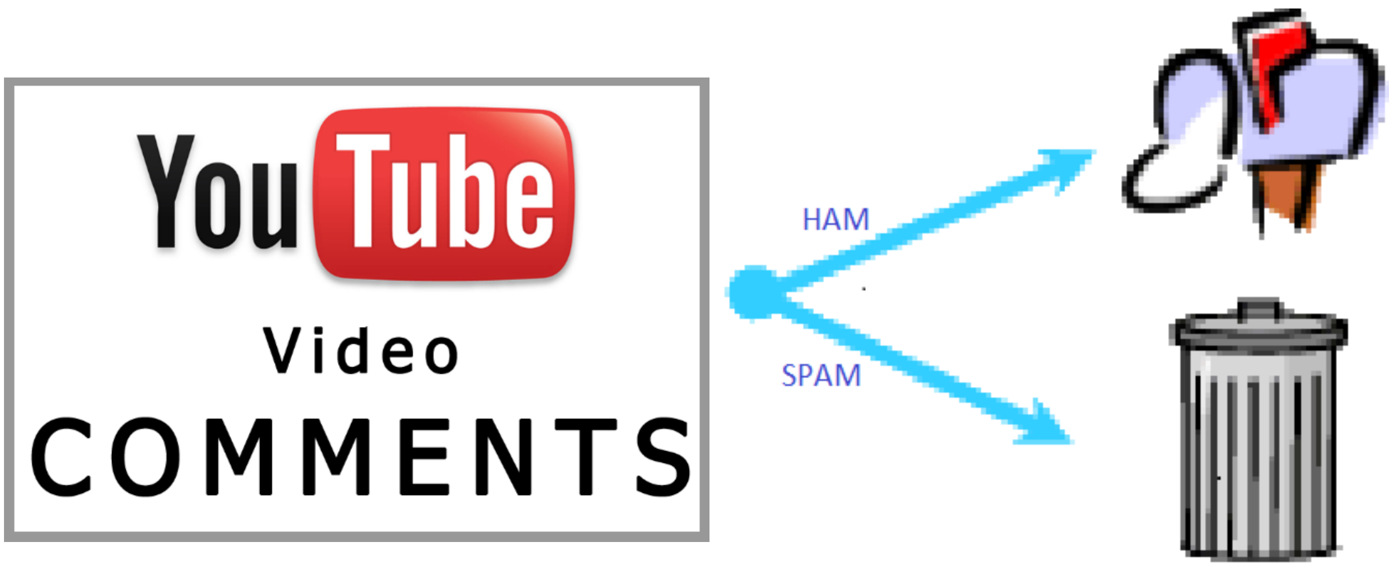
Outline

- Introduction
- Dataset Details
- Data Cleaning/Preprocessing
- Algorithms Used
- Results
- Analysis
- Conclusion

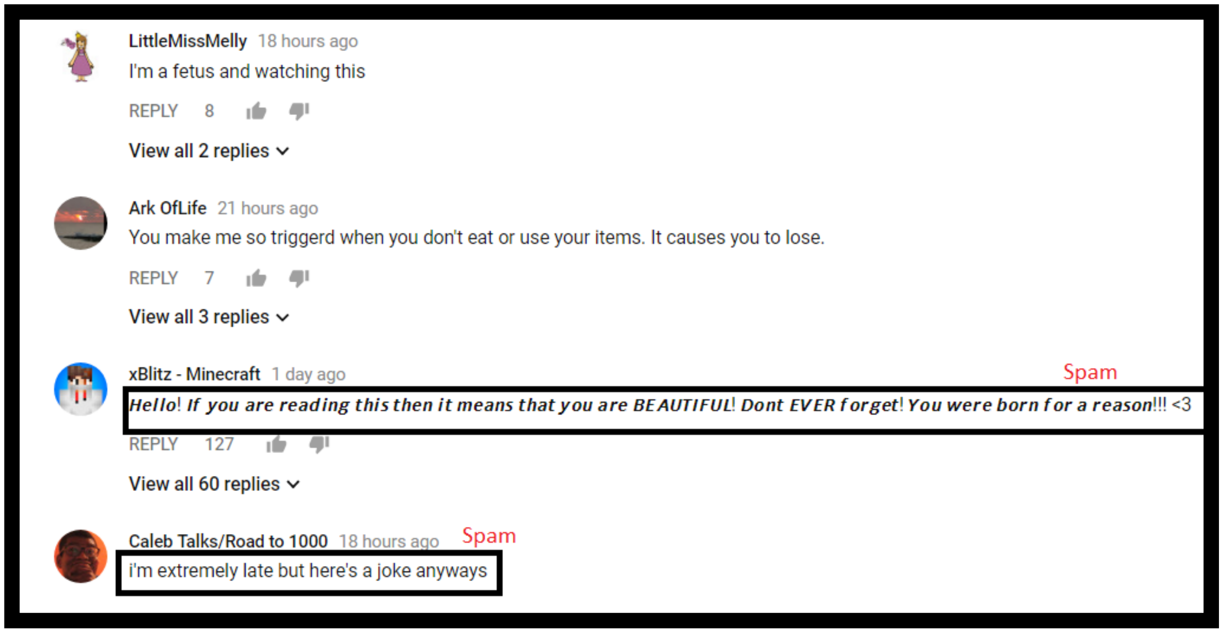
YouTube Spam Comment

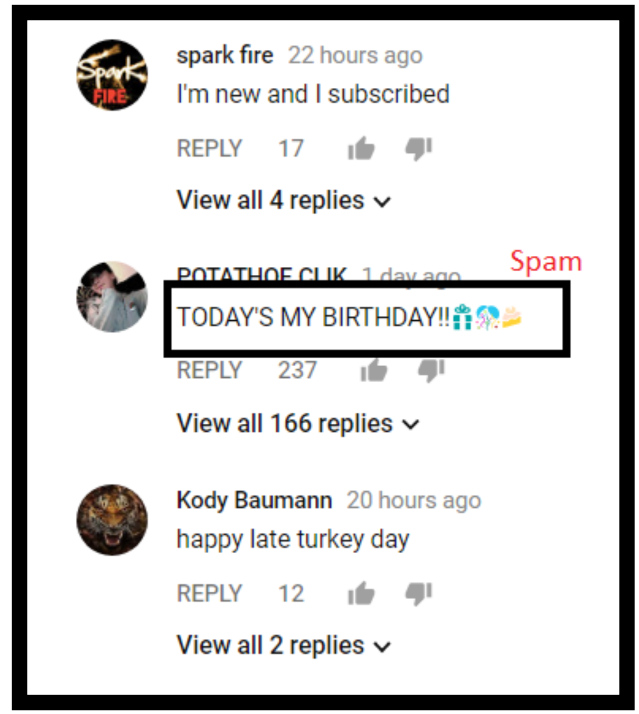
- Spam comment can be described as undesired information with low quality content.
- It is difficult to distinguish between an informative and spam comment.
- Due to monetization system of YouTube, spammers are more active.
- It is inconvenient, annoying and wasteful of computer resources.

Spam Detection



Examples of comment spam posted in YouTube





Dataset Details

- There are six datasets.
  - ➢ First five datasets were taken from UCI Repository. (https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection)
  - ➢ Sixth dataset (Rotten Tomatoes) is extracted from YouTube video and labelled manually (spam or ham).

Data set	YouTube ID	#spam	#ham	total
Eminem	uelHwf8o7_U	247	207	454
PSY	9bZkp7q19f0	175	175	350
LMFAO	KQ6zr6kCPj8	236	202	438
Katty Perry	CevxZvSJLk8	174	175	359
Shakira	pRpeEdMmmQ	174	196	370
Rotten Tomatoes	ZjKLItXpi1U	174	178	352

Cont..

- The collection is composed by one CSV file per dataset, where each line has the following attributes:

  COMMENT_ID, AUTHOR, DATE, CONTENT, TAG

- Each Instance is labeled as spam (represented as 1) or ham(represented as 0).

We offer one example below:

z12oglnpoq3gjh4om04cfdlbgp2uepyytpw0k, Francisco Nora,2013-11-28T19:52:35,please like :D https://premium.easypromosapp.com/voteme/19924/616375350,1

Data Cleaning / Preprocessing

Data cleaning process includes removal of

- Extra attributes- We have removed COMMENT_ID, AUTHOR and DATE.
  Considering only CONTENT and TAG.
- Unicode are removed (/u,:P,-_-)
- Punctuations (, ! . ‘ “)

Preprocessing includes-

- Stop-word removal – e.g. a, an, for, it.
- Case conversion- each alphabet is considered into lowercase. E.g. cat and Cat are considered same.
- Stemming- We have used Weka stemmer (IteratedLovinsStemmer). E.g. sleep and sleeping both are considered as same word after stemming.

Spam Detection Process-

- We have used bag of words approach.
  - ➢Where each word is considered as an independent attribute. And number of times that word occurs in a particular instance give value of the attribute.
- Content attribute is converted from string to word vector.
- Later, Supervised Machine learning is applied to the resultant attributes with tag as class attribute.

Algorithm Used

- Naïve-Bayes
  - Easy and fast to predict class of test data set

- Naïve-Bayes Multinomial
  - explicitly models the word counts and adjusts the underlying calculations to deal with in

- Decision Tree
  - Information gain and
  - Pruning the decision tree

Cont….

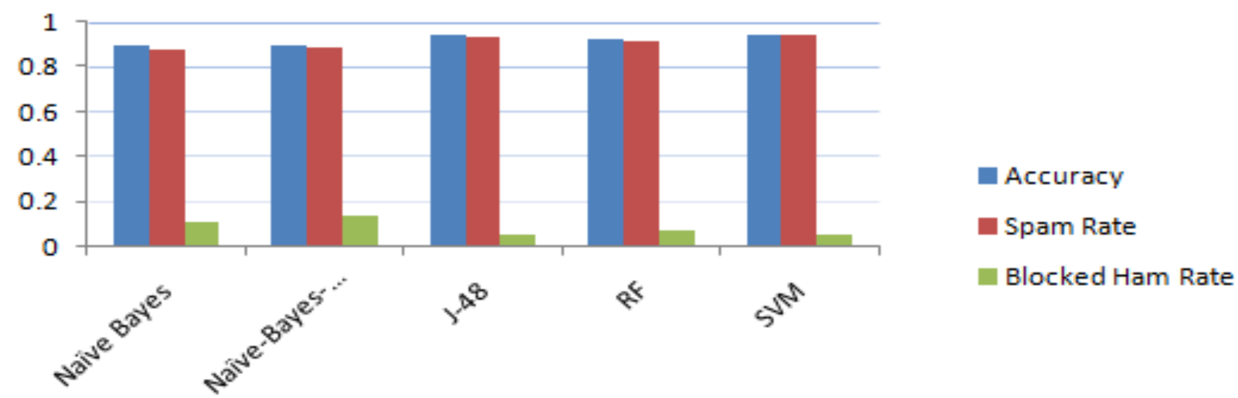
- Random Forest
  - ensemble approach that can also be thought of as a form of nearest neighbor predictor

- Supervised Vector Machine(SVM)
  - The kernel trick to transform the problem, able to apply linear classification techniques to non-linear data

ALgorithm	Parameter
Naïve-Bayes	UseKernelEstimator
Naïve-Bayes Multinomial	BatchSize =100
Decision tree	Subtreeraising= true
Random Forest	# tree= 90
SVM	Calibrator= SMO

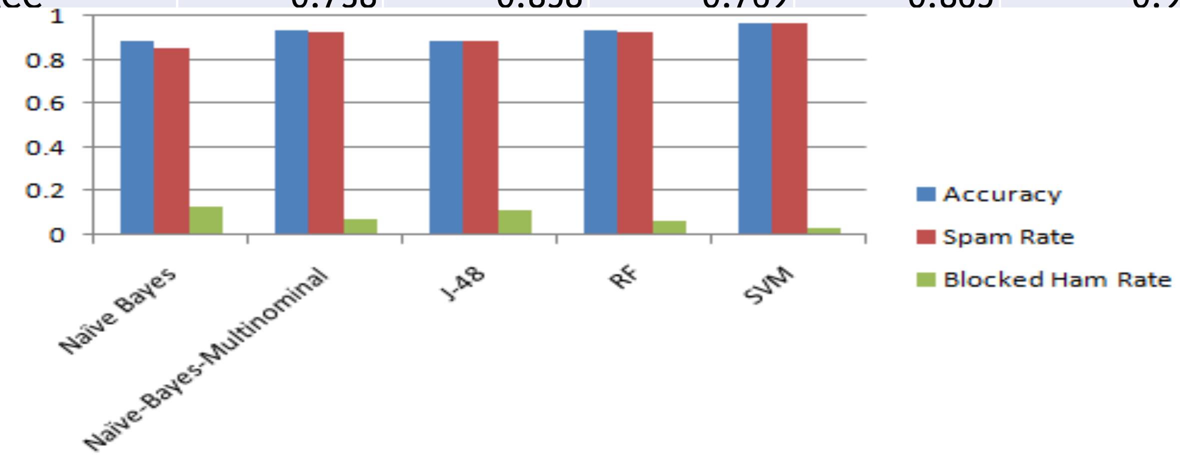
EMINEM Dataset

EMINEM	Naïve Bayes	Naïve-Bayes-Multinomial	J-48	RF	SVM
Precision	0.892	0.895	0.943	0.925	0.945
TP	0.875	0.89	0.934	0.919	0.941
FP	0.104	0.132	0.051	0.07	0.049
Recall	0.875	0.89	0.934	0.919	0.941
F-measure	0.875	0.888	0.934	0.919	0.941
MCC	0.766	0.779	0.875	0.842	0.885



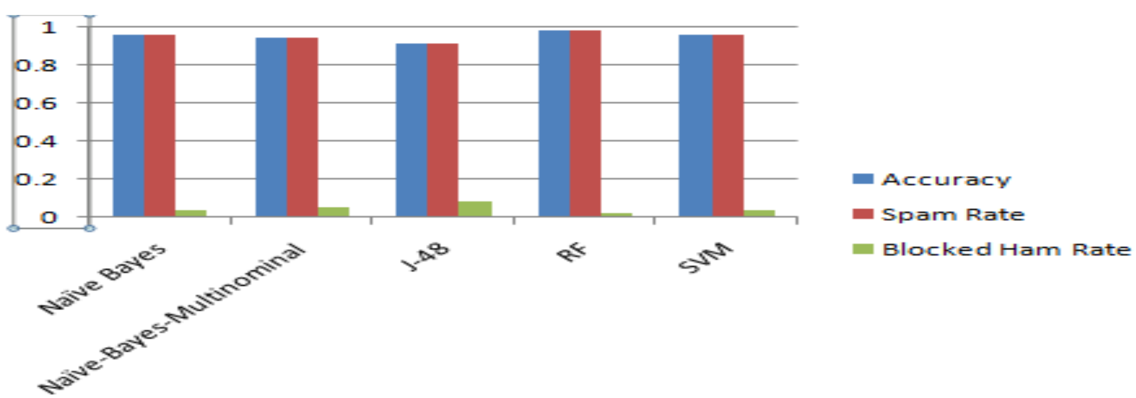
Shakira Dataset

Shakira	Naïve Bayes	Naïve-Bayes-Multinomial	J-48	RF	SVM
Precision	0.881	0.93	0.886	0.937	0.966
TP	0.856	0.928	0.883	0.928	0.964
FP	0.133	0.069	0.114	0.066	0.033
Recall	0.856	0.928	0.883	0.928	0.964
F-measure	0.854	0.928	0.883	0.928	0.964
MCC	0.738	0.858	0.769	0.865	0.93



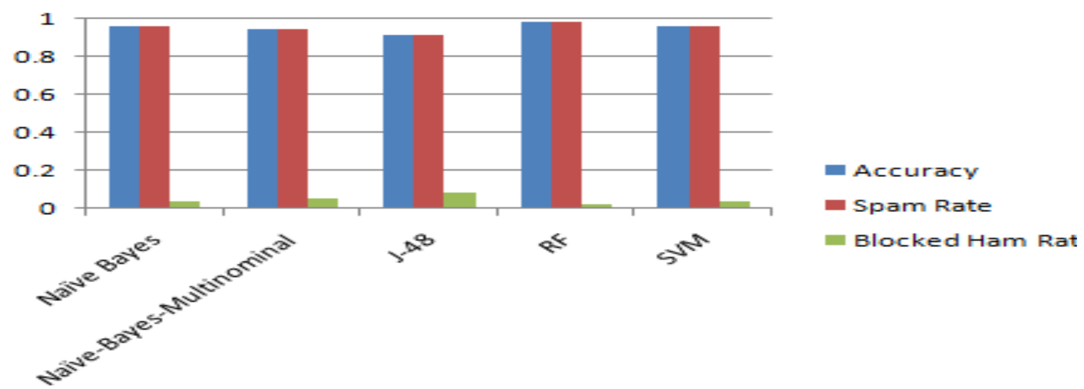
Psy Dataset

Psy	Naïve Bayes	Naïve-Bayes-Multinomial	J-48	RF	SVM
Precision	0.963	0.946	0.916	0.981	0.962
TP	0.962	0.943	0.914	0.981	0.962
FP	0.035	0.05	0.082	0.019	0.039
Recall	0.962	0.943	0.914	0.981	0.962
F-measure	0.962	0.943	0.914	0.981	0.962
MCC	0.924	0.888	0.829	0.961	0.923



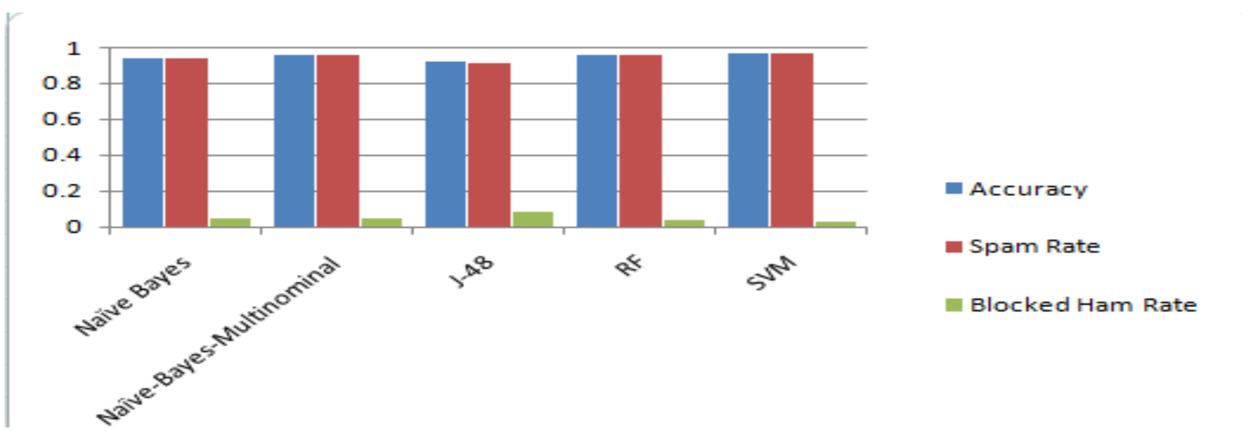
Katy Perry Dataset

Katy Perry	Naïve Bayes	Naïve-Bayes-Multinomial	J-48	RF	SVM
Precision	0.938	0.886	0.905	0.945	0.973
TP	0.933	0.886	0.905	0.943	0.971
FP	0.062	0.117	0.096	0.061	0.031
Recall	0.933	0.886	0.905	0.943	0.971
F-measure	0.933	0.886	0.905	0.943	0.971
MCC	0.871	0.771	0.809	0.888	0.944



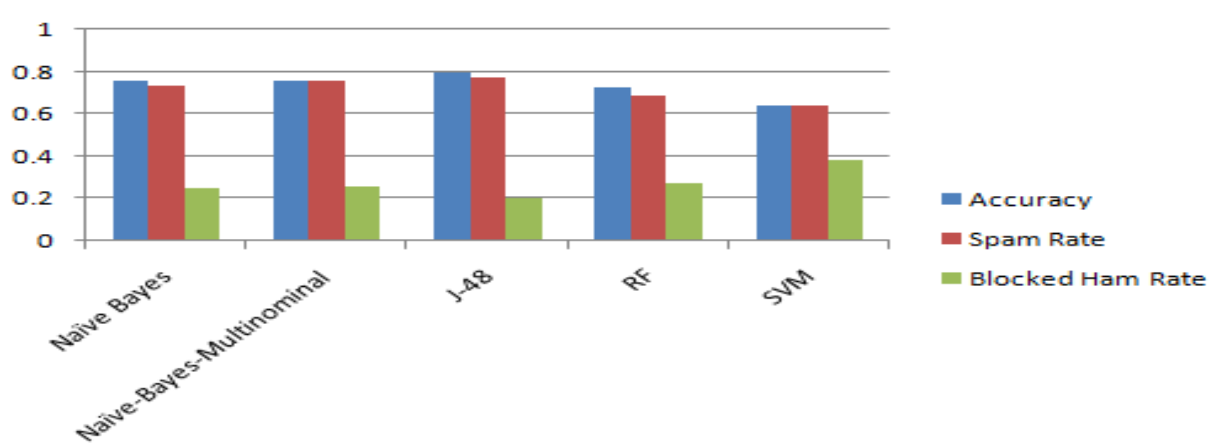
LMFAO Dataset

LMFAO	Naïve Bayes	Naïve-Bayes-Multinomial	J-48	RF	SVM
Precision	0.943	0.955	0.917	0.956	0.969
TP	0.939	0.954	0.916	0.954	0.969
FP	0.05	0.043	0.081	0.039	0.032
Recall	0.939	0.954	0.916	0.954	0.969
F-measure	0.939	0.954	0.916	0.954	0.969
MCC	0.881	0.907	0.831	0.909	0.938



Rotten Dataset

Rotten	Naïve Bayes	Naïve-Bayes-Multinominal	J-48	RF	SVM
Precision	0.755	0.757	0.792	0.729	0.639
TP	0.736	0.755	0.774	0.689	0.639
FP	0.247	0.259	0.206	0.27	0.385
Recall	0.736	0.755	0.774	0.689	0.639
F-measure	0.739	0.756	0.776	0.691	0.639
MCC	0.478	0.492	0.555	0.412	0.254



Comparing classifiers

- Why do we have to compare classifiers?
- How did we compare ?

Friedman Test

- It is a statistical analysis test which is non-parametric.
- The Friedman test checks if the null hypothesis(which states there is no difference between the results) can be rejected based on ranking position of each classifier over each dataset.
- The ranking was built using MCC rates, where the method with the highest MCC for a certain dataset is ranked as 1, and the method with the lowest MCC for the same dataset is ranked as n, where n is the number of classification methods

- Shakira

	Naïve bayes	**naïve-bayes-Multinomial**	**J-48**	**RF**	**SVM**
Precision	0.881	0.93	0.886	0.937	0.966
spam rate	0.856	0.928	0.883	0.928	0.964
blocked ham rate	0.133	0.069	0.114	0.066	**0.033**
Recall	0.856	0.928	0.883	0.928	**0.964**
F-measure	0.854	0.928	0.883	0.928	**0.964**
MCC	0.738	0.858	0.769	0.865	0.93

- Eminem

	Naïve bayes	**naïve-bayes-Multinomial**	**J-48**	**RF**	**SVM**
Precision	0.892	0.895	0.943	0.925	0.945
spam rate	0.875	0.89	0.934	0.919	0.941
blocked ham rate	0.104	0.132	0.051	0.07	**0.049**
Recall	0.875	0.89	0.934	0.919	**0.941**
F-measure	0.875	0.888	0.934	0.919	**0.941**
MCC	0.766	0.779	0.875	0.842	0.885

- Psy

	Naïve bayes	**naïve-bayes-Multinomial**	**J-48**	**RF**	**SVM**
Precision	0.963	0.946	0.916	0.981	0.962
spam rate	0.962	0.943	0.914	0.981	0.962
blocked ham rate	0.035	0.05	0.082	**0.019**	0.039
Recall	0.962	0.943	0.914	**0.981**	0.962
F-measure	0.962	0.943	0.914	**0.981**	0.962
MCC	0.924	0.888	0.829	0.961	0.923

- Katy perry

	Naïve bayes	naïve-bayes-Multinomial	J-48	RF	SVM
Precision	0.938	0.886	0.905	0.945	0.973
spam rate	0.933	0.886	0.905	0.943	0.971
blocked ham rate	0.062	0.117	0.096	0.061	**0.031**
Recall	0.933	0.886	0.905	0.943	**0.971**
F-measure	0.933	0.886	0.905	0.943	**0.971**
MCC	0.871	0.771	0.809	0.888	0.944

- LMFAO

	Naïve bayes	**naïve-bayes-Multinomial**	**J-48**	**RF**	**SVM**
Precision	0.943	0.955	0.917	0.956	0.969
spam rate	0.939	0.954	0.916	0.954	0.969
blocked ham rate	0.05	0.043	0.081	0.039	**0.032**
Recall	0.939	0.954	0.916	0.954	**0.969**
F-measure	0.939	0.954	0.916	0.954	**0.969**
MCC	0.881	0.907	0.831	0.909	0.938

- Rotten Tomatoes

	Naïve bayes	**naïve-bayes-Multinomial**	**J-48**	**RF**	**SVM**
Precision	0.755	0.757	0.792	0.729	0.639
spam rate	0.736	0.755	0.774	0.689	0.639
blocked ham rate	0.247	0.259	**0.206**	0.27	0.385
Recall	0.736	0.755	**0.774**	0.689	0.639
F-measure	0.739	0.756	**0.776**	0.691	0.639
MCC	0.478	0.492	0.555	0.412	0.254

Classifier Ranking:

	Shakira	Eminem	Psy	Katy perry	LMFAO	Rotten Tomatoes
NB	5	5	2	4	5	3
NB-M	3	4	4	5	3	2
J-48	4	2	5	3	4	1
RF	2	3	1	2	2	4
SVM	1	1	3	1	1	5

Observation

- J-48, SVM has the largest range among all methods, which means they have achieved very good and very bad results, not being consistently the best or the worst at all.
- However, NB consistently presented the worst performance.
- Null hypothesis can be rejected with 99.9% confidence rate.
- RF has the best ranking position.

Comparison with TubeSpam Paper.

Dataset	TubeSpam (MCC)	Our Paper (MCC)
Psy	.925(SVM)	**.961(RF),** .923(SVM)
Katty Perry	.893(RF)	**.944(SVM),** .888(RF)
LMFAO	**.955(NB and SVM)**	.938(SVM), .881(NB)
Eminen	**.955(CART)**	.885(SVM), .875(NB)
Shakira	**.93(NB)**	**.93(SVM),** .738(NB)

References

- Alberto, T.C., Lochter J.V., Almeida, T.A. **TubeSpam: Comment Spam Filtering on YouTube.** Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15), 1-6, Miami, FL, USA, December, 2015.
- Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
- [Online] www.youtube.com.

Thank you……