Data Mining Final Project

Spam Filtering- YouTube Comment

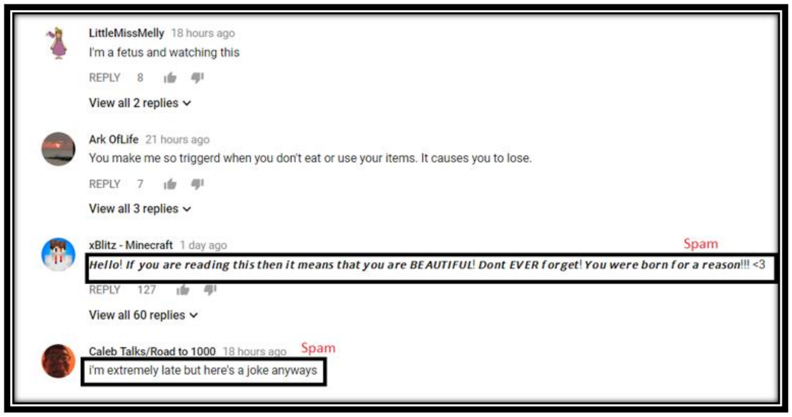
Team Members- Bharti Goel, Jaynab Khatun, Mounika Dudipala, Varsha Reddy Yasa.

Introduction

In this project we did spam filtering of YouTube comments using 5 different machine learning algorithms named Naïve Bayes, Naïve Bayes multinomial, J48, Random Forest and SVM. We have taken six data sets (five from UCI repository and one extracted and labelled by us). At the end we have used MCC based Freidman test to compare performance of different algorithms over different data sets.

Spam comment can be defined as undesired information with low quality content. It is difficult to distinguish between an informative and spam comment as spammers are quite smart nowadays and they are creating better spams (in order to get away from default spam detection). Due to monetization system of YouTube, spammers are more active and it is inconvenient, annoying and wasteful of computer resources.

Example of spam comments in YouTube videos.

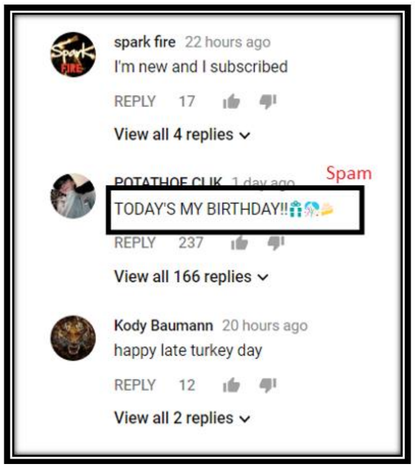


LittleMissMelly 18 hours ago
I'm a fetus and watching this
REPLY 8
View all 2 replies

Ark OfLife 21 hours ago
You make me so triggerd when you don't eat or use your items. It causes you to lose.
REPLY 7
View all 3 replies

xBlitz - Minecraft 1 day ago
Spam
Hello! If you are reading this then it means that you are BEAUTIFUL! Dont EVER forget! You were born for a reason!!! <3
REPLY 127
View all 60 replies

Caleb Talks/Road to 1000 18 hours ago
Spam
i'm extremely late but here's a joke anyways



spark fire 22 hours ago
I'm new and I subscribed
REPLY 17
View all 4 replies

POTATHOE CLIK 1 day ago
Spam
TODAY'S MY BIRTHDAY!!🎁🎊🍰
REPLY 237
View all 166 replies

Kody Baumann 20 hours ago
happy late turkey day
REPLY 12
View all 2 replies

Dataset Details

We have used six different data sets-

- First five datasets (Eminem, PSY, LMFAO, Katy Perry and Shakira) are taken from UCI repository. The link for the five data sets is https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection.

- The sixth dataset named Rotten Tomatoes is extracted from a YouTube video, and we labelled spam and ham manually.
- The collection is composed of 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). An example for an instance labelled spam in dataset is-:

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

- Details of all datasets used is represented in a following table.

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

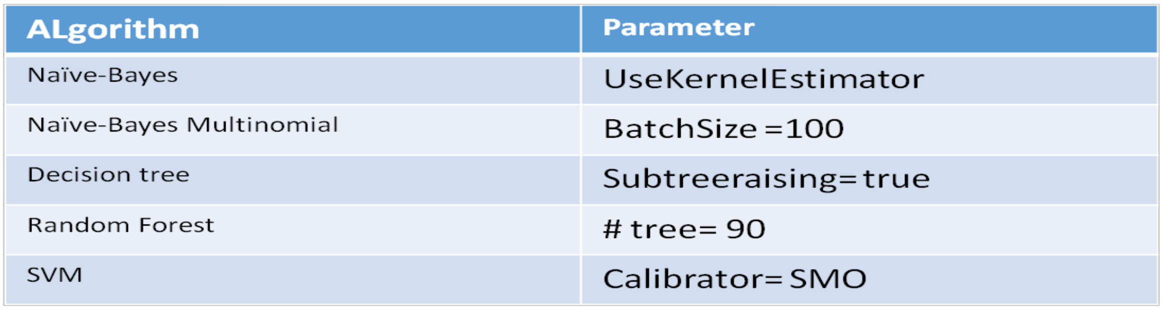
Data Cleaning/Preprocessing

- Data Cleaning includes removal of Extra attributes (COMMENT_ID, AUTHOR and DATE), Unicodes (/u,:P,-_-) and punctuations (, . ! ' ").
- Data Preprocessing includes stop-word removal (e.g. a, an, for, it, etc.), case conversion (each alphabet is converted into lowercase. E.g. cat and Cat are considered same) and stemming (We have used Weka stemmer IteratedLovinsStemmer). E.g. sleep and sleeping both are considered as same word after stemming.

Spam Detection Process

- **Attribute-** We used bag of words approach. Each word is considered as an independent attribute and number of times that word occur in a particular instance give value of the attribute. Content attribute of csv file is converted from string to word vector. Later, Supervised Machine learning is applied to the resultant attributes with tag as class attribute.
- **Split-** We did 1/3rd split, where first 66% is taken for training and 34% is taken as testing. In addition, comments were in order of their occurrence (chronological order) in real YouTube video.

- **Algorithms used-** There are so many algorithms but here we choose five best algorithms used for text classification for small sized datasets. We used Naïve-Bayes, Naïve-Bayes multinomial, decision tree, Random forest, supervised vector machine (SVM). Naïve Bayes is used because it is easy and fast to predict the class for text data set as it considers each word as independent feature. Naïve-Bayes Multinomial is better than normal Naïve-Bayes version. It gives a good result on large text data sets. This model based on word count adjusts the underlying calculation to deal with in data sets. Decision trees use the information gain for best feature selection and prune the tree to produce the result fastly. Here, we used J48 in WEKA on our data sets. Random forest is an ensemble approach that can also be thought of as a form of nearest neighbor predictor. SVM (linear kernel) used the kernel trick to transform the problem, able to apply linear classification techniques to non-linear data. We did the parameter tuning on each algorithm which gave is good result on our data sets. Following table gives details about parameter we have set for different model:

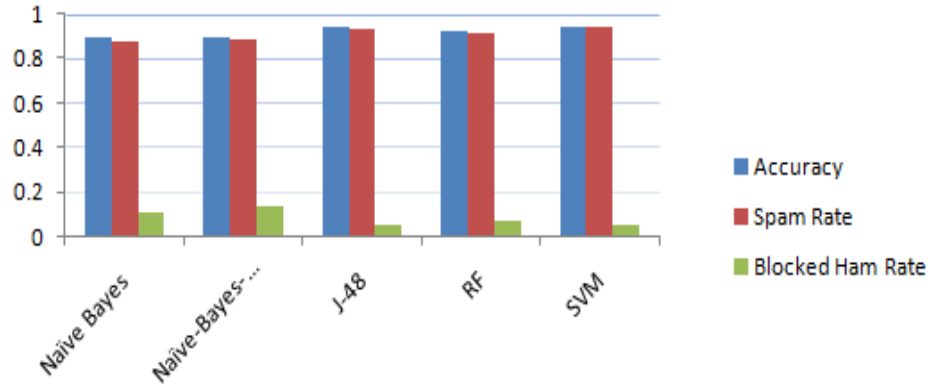


ALgorithm	Parameter
Naïve-Bayes	UseKernelEstimator
Naïve-Bayes Multinomial	BatchSize =100
Decision tree	Subtreeraising=true
Random Forest	# tree= 90
SVM	Calibrator=SMO

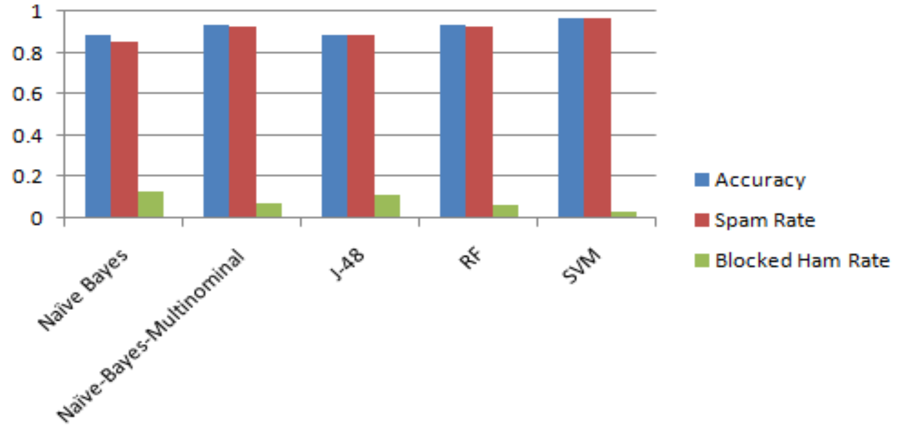
Results-

Metrics- We have used Spam blocked rate (true positive- when spam is classified as spam), blocked ham rate (false positive- when ham is classified as spam) and accuracy. Results are presented in following graphs as performance of each dataset at a time.

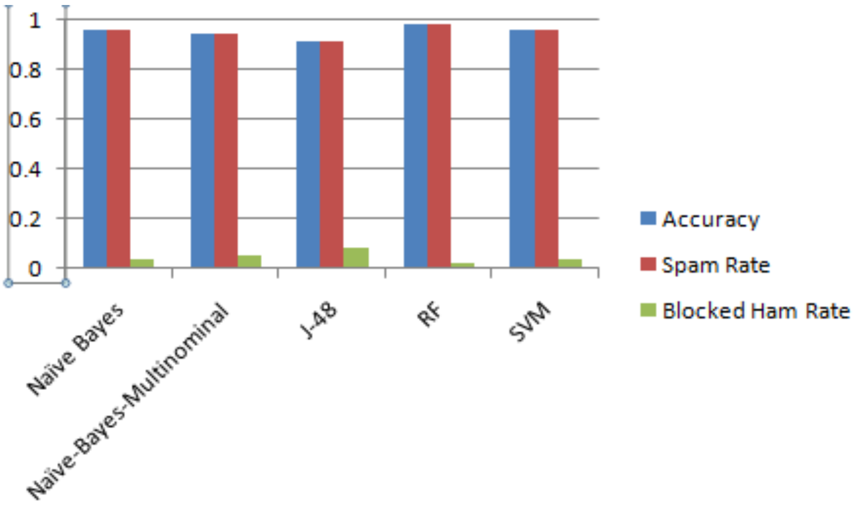
EMINEM Dataset: the size of data set is 454 instances. J48 and SVM gave a good result for Blocked Ham Rate which less than other model. From the table we noticed that precision is higher in SVM than J48.



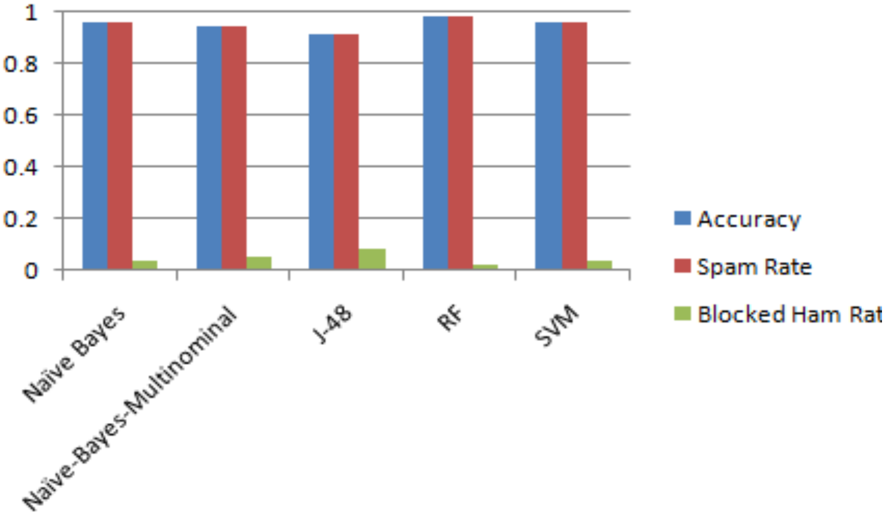
Shakira Dataset: the size of data set is 370 instances. Random Forest and SVM gave a good result for Blocked Ham Rate which less than other model. From the table we noticed that precision is higher in SVM than Random Forest.



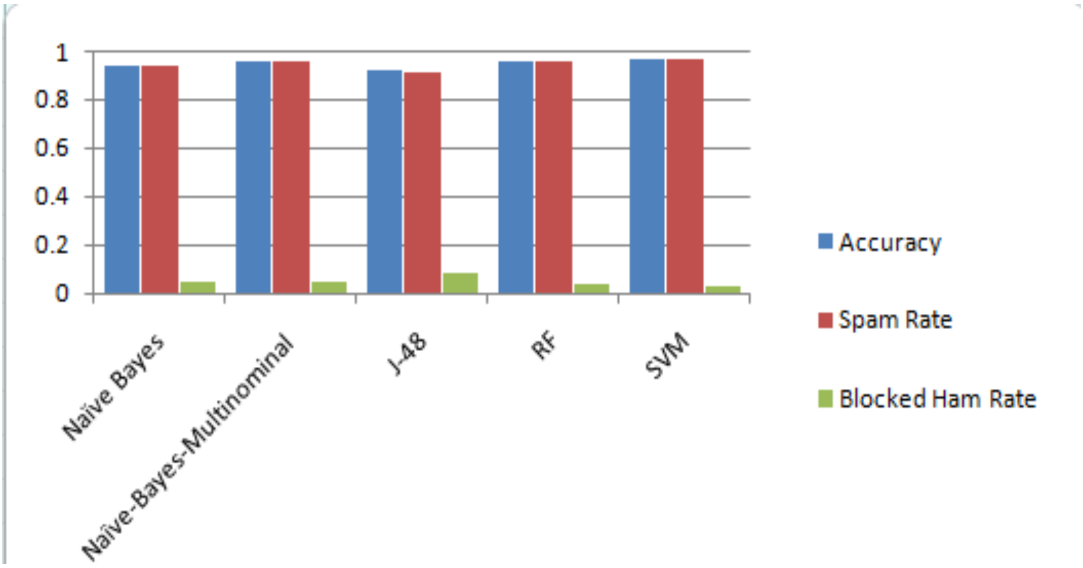
Psy Dataset: the size of data set is 350 instances. Random Forest gave a good result for Blocked Ham Rate which less than other model. From the table we noticed that precision is higher in Random Forest than other model.



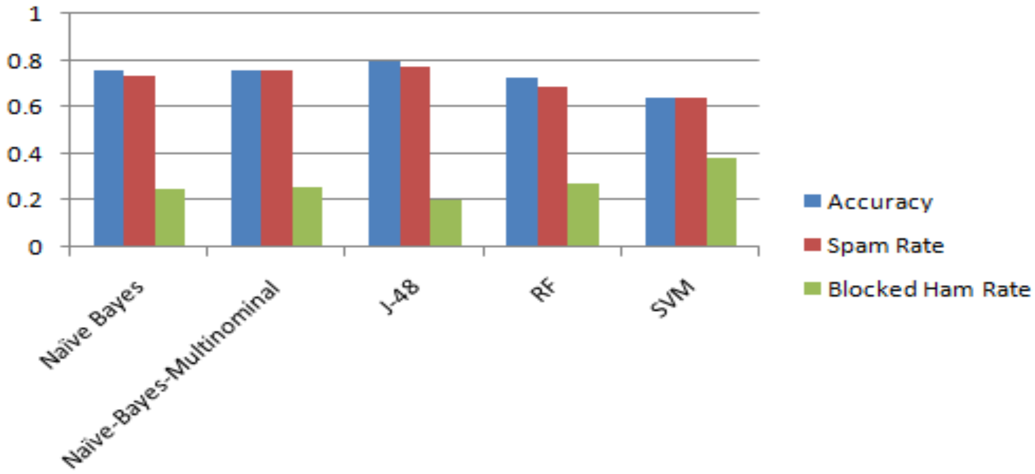
Katy Perry Dataset: the size of data set is 359 instances. Random Forest and Naïve Bayes gave a good result for Blocked Ham Rate which less than other model. From the table we noticed that precision is higher in Random Forest Naïve Bayes.



LMFAO Dataset: the size of data set is 438 instances. Random Forest and SVM gave a good result for Blocked Ham Rate which less than other model. From the table we noticed that precision is higher in SVM than Random Forest. Naïve Bayes gave a god FP rate.



Rotten Tomatoes: The size of data set is 352 instances. J48 model provided result for Blocked Ham Rate which better than model. From the table we noticed that precision is higher in J48. Still this result is not convincing to be good. Because the Blocked ham Rate is much higher than other dataset.

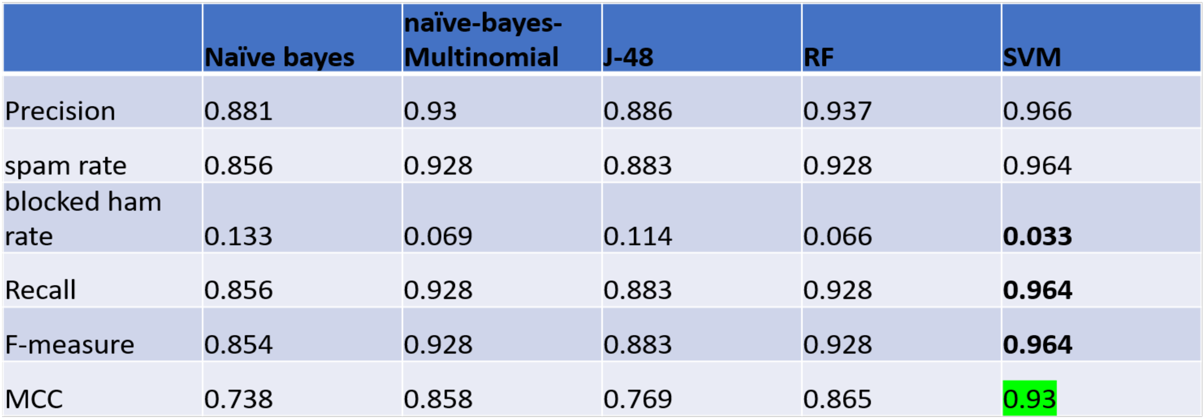


Classifier Comparison

To determine the accuracy of the classifiers and conclude the efficient classifier, we have to perform a statistical analysis of the classifiers. In this project, we performed non-parametric Friedman test to compare the classifiers that were applied on the datasets. In this test we are assigning ranks to classifiers, based on ranks Friedman test will check if null hypothesis can be ignored. These ranking were given based on the values of Matthews correlation coefficient (MCC). Higher the MCC value, higher the rank (smaller integer) of classifier. The higher values of MCC for each classifier are marked green in the tables below.

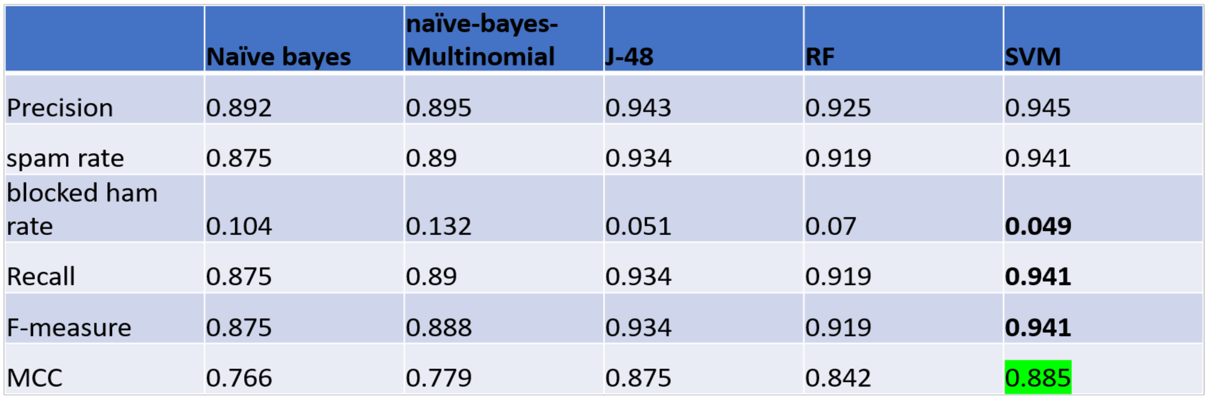
Tables below gives Precision, spam rate, blocked ham rate, recall, F-measure, and MCC.

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

Let us analyse the MCC values and ranks of five different classifiers namely Naïve Bayes, Naïve Bayes-Multinomial, J-48, Random Forest, Support Vector Machines on six different datasets namely Shakira, Eminem, Psy, Katy Perry, LMFAO, Rotten Tomatoes.

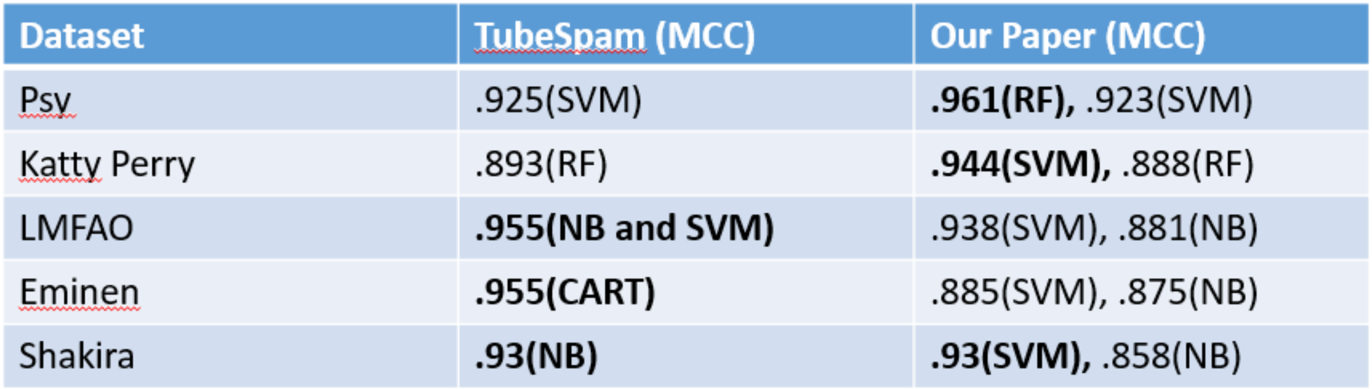
Rank table based on MCC values.

	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

Observations

From the above table, we notice that J-48, SVM has the largest range among all methods, which means they have achieved both very good and very bad results, not being consistently the best or the worst. NB presented the worst performance. RF has the best ranking position among all other classifiers from the table above. Null hypothesis can be rejected in this test with a confidence value of 99.9%. Thus, we can say that our analysis of Friedman test proves that Random Classifier is the accurate classifier for our six datasets.

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),** .858(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.**