SMS Spam Detection

Group Members :

1. Saurabh Jaiswal (13114055)
2. Saurabh Jain (13114056)
3. Saurabh Verma (13114057)

Introduction

- irrelevant or unsolicited messages , for the purposes of advertising.
- annual worldwide SMS traffic volume was 6.9 trillion at end-2010 to break 8 trillion by end-2011.
- in parts of Asia up to 30% of messages were represented by spam.
- 400 percent increase in unique SMS spam campaigns in the first half of the year 2012.
- 30 million smishing (SMS Phishing) messages are sent to cell phone.
- 140 bytes, which translates to 160 characters of the English alphabet.

DataSet

- a total of 5,574 short messages
- 4,827 legitimate messages
- 747 mobile spam messages
- a total of 81,175 tokens

Hams	63,632
Spams	17,543
Total	81,175
Avg per Msg	14.56
Avg in Hams	13.18
Avg in Spams	23.48

ham	What you doing?how are you?
ham	Ok lar... Joking wif u oni...
ham	dun say so early hor... U c already then say...
ham	MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham	Siva is in hostel aha:-.
ham	Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam	FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam	URGENT! Your Mobile No 07808726822 was awarded a £2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Features

For each of the message we have created these 7 features :

1. Number of characters.
2. Number of unique character
3. Weighted unique characters.
4. Word count
5. Is repeated
6. Repetition count top 3
7. Length of longest word

Methodology

➔ Machine Learning Methods used :

1. Linear SVM Classifier
2. Stochastic Gradient Descent (SGD) classifier.

➔ Both with 5 and 7 features each.

➔ Linear SVM Classifier using TF-IDF vector as feature vector.

Linear SVM using TF-IDF feature vector

- **Term Frequency :**

$$\mathrm{TF}_{i,j} = \frac{F_{ij}}{\max_z F_{zj}},$$

- **Inverse Document Frequency :**

$$\mathrm{IDF}_i = \log \frac{m}{n_i},$$

- **TF IDF weight :**

$$X_{ji} = TF_{ij} \text{ X } IDF_{ij}$$

Linear SVM using TF-IDF feature vector : Eg

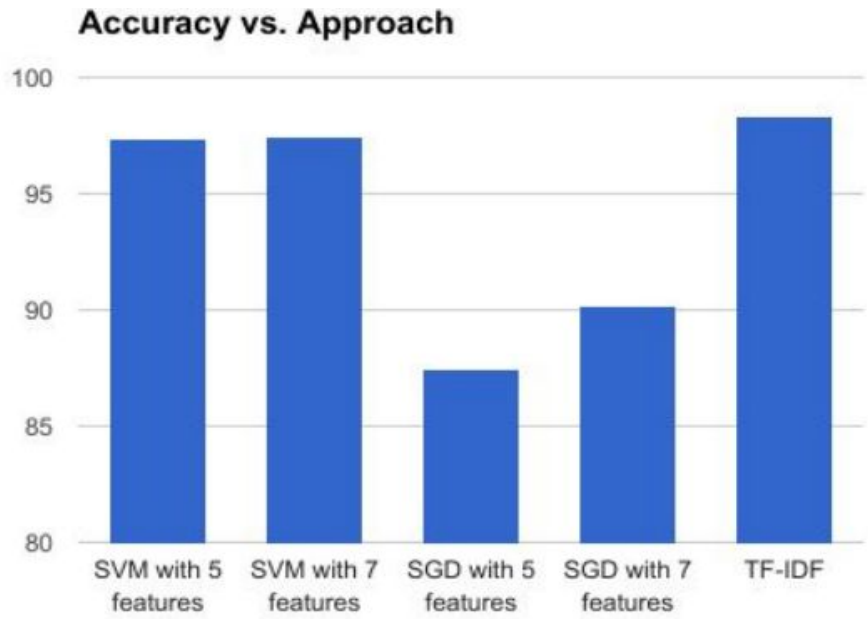
Message 1 : What you doing?how are you?

Message 2 : Sunshine Quiz! Win a super Sony DVD recorder.

	what	you	doing	how	are	Sunshine	Quiz	Win	a	super	Sony	DVD	Recorder
Freq 1	1	2	1	1	1	0	0	0	0	0	0	0	0
Freq 2	0	0	0	0	0	1	1	1	1	1	1	1	1
TF	0.5	1	0.5	0.5	0.5	1	1	1	1	1	1	1	1
IDF	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
X1	0.15	0.3	0.15	0.15	0.15	0.3	0	0	0	0	0	0	0
X2	0	0	0	0	0	0	0.3	0.3	0.3	0.3	0.3	0.3	0.3

Results - Average Accuracy

Approach	Accuracy
SVM with 5 features	97.4 %
SVM with 7 features	97.5 %
SGD with 5 features	87.5 %
SGD with 7 features	90.2 %
SVM with TF-IDF	98.3 %



Results - Accuracy

Conclusions

1. SVM with lesser no of features may give a good accuracy but in SGD we require more features to get better accuracy.
2. The best classifier turns out to be SVM with TF-IDF with 98.3% accuracy
3. SVM provides consistent results whereas SGD does not.

References

1. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
2. http://scikit-learn.org/stable/modules/sgd.html
3. https://opendatascience.com/blog/spam-detection-in-9-lines-of-code/