ASSIGNMENT II (CS 8803 Social Computing)

Due date	3:30pm, October 15
Topic	Build and report the performance of a binary classifier to distinguish between legitimate and spam SMSes, based on their text features
What to hand in?	3 page report
How/where to submit?	T-Square
Grade	10% [Total points: 20]; 2% extra credit for novel features and for comparing multiple classifiers

Dataset Source	www.munmund.net/courses/fall2014/notes/smsspamcollection.zip
Useful libraries	[Python] nltk, scikit-learn (You are also free to use your favorite programming language, statistical tool, or library)

Our past several lectures covered basic statistical methods useful for analyzing data arising in social computing systems and problems. We also covered some papers on analysis of social text. This assignment will test your understanding of the concepts and the materials in general in these two spaces.

Question

In this assignment you are supposed to build a classifier that can distinguish between legitimate (ham) and spam SMS. Refer to the table above to download the dataset. Your submission would include a three page report that discusses how you built the classifier, and then presents the performance of your classifier. The classifier would be based on text features extracted from the ham and spam messages.

Note: you do not have to write a classifier from scratch. You are free to use one or more of the many open-source (or other) tools and packages that allow you to use a variety of different classifiers. Some example libraries you can use are listed in the table above. However you can pick any package or programming language you like or are most comfortable with.

Background and information about the dataset

The dataset is called the SMS Spam Collection v.1 dataset released and maintained by Tiago A. Almeida and José María Gómez Hidalgo. It is a public set of labeled SMS messages that have been collected for spam research. It has 5,574 English, real and non-encoded messages, tagged by humans to be legitimate (ham) or spam. The dataset contains just one text file (along with a ReadMe document). In the main file, each line has the correct class (ground truth) followed by the raw message. Some examples are given below:

```
ham   What you doing?how are you?
ham   MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
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  Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of
Australia? Text MQUIZ to 82277. B
spam  URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on
02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU
```

For the more curious minds, detailed information about how the dataset was constructed is given in the Appendix section later in the document.

Contents of the report

(1) *Feature construction (3 points).* Like we covered in our lectures, constructing a classifier involves extracting relevant and meaningful features from the data under consideration. In your report, you will need to first present the various features you derived from the textual content of the ham and spam messages. Features can include (but not limited to) unigrams, bigrams, TF-IDF, Part-of-Speech tags, length of words etc. of the messages. Also you will need to discuss why you chose the particular set of features. Feel free to refer to the papers in the References section for the purpose.

(2) *Description of the classifier (2 points).* Discuss what is the particular classifier you chose (e.g., *k* Nearest Neighbor, Support Vector Machine, Naïve Bayes or some other), and a justification or rationale behind its choice and applicability to the dataset in question. That is, if you picked classifier X, why is it a good fit for this problem? Why is it a better choice compared to another classifier Y?

(3) *Evaluation technique (2 points).* Present how you evaluated how well your classifier of choice performed in distinguishing between ham and spam messages. Particularly, discuss applicability of the concept of *k*-fold cross validation here, which we had covered in our Statistics/Data Mining review lectures. You will also need to present what metrics you used to evaluate performance of the classifier. For instance, typical metrics would include percentage accuracy, precision, recall.

(4) *Implementation (7 points).*

   a. Discuss how you preprocessed your data. If you used stopword removal, stemming, or tokenization over the content of the messages, you need to report it here. Point to the particular libraries or functions you needed for this.

   b. Discuss how you extracted the features you presented above from the SMS dataset. This needs to include what libraries and which particular functions you used for extracting each feature. If you did not use an existing library, you need to write about the method you used to compute the features from the data. Report if you did some filtering or feature selection to disregard not-so-common features (e.g., if you ignored all unigrams which occurred less than five times). Also report if you did any kind of normalization or standardization of each feature, and your justification behind doing or not doing so.

   c. Next, discuss how you implemented/used a library for your chosen classifier. Report what were the inputs and outputs to the particular library function you used and if/how you tuned parameters of the classifier (e.g., if you chose SVM, report the particular kernel you used).

   d. Discuss how you partitioned the dataset for *k*-fold cross validation, along with what was your chosen *k* here. Here you will also discuss based on your chosen *k*-fold cross validation setup, what were your training and test sets in each of the *k*-iterations.

   e. Discuss how you calculated the metrics of performance evaluation, e.g., accuracy, precision, recall etc. It is again okay to use an existing library that gives precision and recall values, in which case you need to present in your report which libraries/functions you used for the purpose, and what was your input and output to those functions.

(5) *Analysis of results (6 points).* Report the performance of your classifier based on the above discussion. You will need to use charts, graphs, or tables to report actual numbers—i.e., the values of the

performance metrics you chose above (accuracy, precision, recall etc.). These numbers should be reported for each of the *k* iterations of the *k*-fold cross validation setup. You should also report the average performance over all *k* cross validation folds, corresponding to each evaluation metric.

(6) *Extra credit (.4 points).*
   a. There will be 1% extra credit (.2 points) for extracting novel text based features from the ham/spam messages (don't be afraid to be creative here; clue: many words in the messages are non-standard English).
   b. There will be another 1% (.2 points) for comparing (per the above evaluation technique) two or more classifiers in their ability to distinguish the two sets, spam and ham, when applied to the same data. For instance, you can compare your chosen classifier SVM's performance over another classifier Naïve Bayes, and in the analysis section discuss which classifier performs better based on your chosen evaluation metrics like accuracy, precision, recall.

References

[1] Almeida, T.A., Gómez Hidalgo, J.M., Silva, T.P. Towards SMS Spam Filtering: Results under a New Dataset. *International Journal of Information Security Science* (IJISS), 2(1), 1-18.

[2] Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010, July). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, p. 12).

[3] Cormack, G. V., Gómez Hidalgo, J. M., and Puertas Sánz, E. Feature engineering for mobile (SMS) spam filtering. Proceedings of the *30th Annual international ACM Conference on Research and Development in information Retrieval* (ACM SIGIR'07), New York, NY, 871-872, 2007.

[4] Cormack, G. V., Gómez Hidalgo, J. M., and Puertas Sánz, E. Spam filtering for short messages. Proceedings of the *16th ACM Conference on Information and Knowledge Management* (ACM CIKM'07). Lisbon, Portugal, 313-320, 2007.

[5] Gómez Hidalgo, J.M., Cajigas Bringas, G., Puertas Sanz, E., Carrero García, F. Content Based SMS Spam Filtering. *Proceedings of the 2006 ACM Symposium on Document Engineering* (ACM DOCENG'06), Amsterdam, The Netherlands, 10-13, 2006.

[6] McCord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In *Autonomic and Trusted Computing* (pp. 175-186). Springer Berlin Heidelberg.

Appendix

This SMS corpus was collected from free or free for research sources on the web:

- A collection of 425 SMS spam messages was manually extracted from the Grumbletext Website. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages. The Grumbletext Website is: http://www.grumbletext.co.uk/.
- A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. The NUS SMS Corpus is available at: http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/.
- A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available at http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf.
- Finally, the SMS Spam Corpus v.0.1 Big dataset was also incorporated. It has 1,002 SMS ham messages and 322 spam messages and it is public available at: http://www.esp.uem.es/jmgomez/smsspamcorpus/.