SIAM Text Mining Competition 2007

Matthew Eric Otey, Ph.D., Ashok N. Srivastava, Ph.D.,*
Santanu Das, Ph.D. and Pat Castle
{otey, ashok, sdas, pcastle}@email.arc.nasa.gov

1 Overview

The 2007 SIAM Text Mining Workshop (TM 07) is pleased to announce a text mining competition. The contest is open to all, including university students, postgraduates, researchers, academics, and practitioners (excluding members of the organizing committee) and offers an excellent opportunity to test one's text mining skills on realistic datasets. Selected competition participants will be invited to present their work at the TM 07 workshop.

2 Classification Task

The contest focuses on developing text mining algorithms for document classification. The documents in question are aviation safety reports documenting one or more problems that occurred on certain flights. The goal is to label the documents with respect to what types of problems they describe. Therefore the documents may belong to more than one class. The goal is to label the documents according to the classes to which they belong, while maximizing both precision and recall, as well as the classifier's confidence in its labeling. A labeled dataset will be provided containing the raw text of the reports for training and validation purposes, and an unlabeled test data set will be used to determine the winner of the competition. To learn more about the format of the supplied datasets, please see the description in section 4.

3 Contest Procedure

The participants will be given two data files. The first contains the raw text of all documents in the training set, while the second contains the labels of the documents in comma-separated value format (for more specific details, see section 4). This is data that participants can use for training and validation purposes. The organizers will also provide a small program implementing the cost function (see section 10) which participants can use as an aid when implementing their approaches. On a designated day (~~January 8, 2007~~ January 10, 2007), a

*NASA Ames Research Center, Itelligent Systems Division

test data set will be made available for a period of 48 hours. During this period participants will be able to download the test data set, apply their approaches, and make a submission to the organizers (see section 6 for more details on the submission). The organizers will then analyze the results and select finalists who will present at the TM 07 workshop during the SIAM International Conference on Data Mining in April 2007. The top three finalists will receive cash rewards. The exact rules of the contest can be found in Section 5, and the dates can be found in Section 9.

4 Data Format

The supplied datasets (training, validation and testing) will be in raw text format. All documents for each set will be contained in a single file. Each row in this file corresponds to a single document. The first characters on each line of the file will be the document number and a tilde separating the document number from the text itself. Each document will have been run through PLADS, which performs stemming, acronym expansion, and other basic pre-processing operations on the documents. PLADS will also remove non-informative terms (e.g., place names), and replace them with an underscore ("_"). The labels for the training and validation data set will be provided in full matrix format in a comma-separated value file, where the value in row i and column j is 1 if document i has label j, and -1 otherwise.

5 Rules

1. The contest is open to any party planning to attend SDM 2007. Group entries are permitted, but a person can participate in only one group.

2. Participants will be notified when the test dataset can be downloaded. Once the test data set has been posted, participants will have 48 hours to make their submission.

3. Only full submissions will be accepted (see Section 6 for details on what constitutes a full submission). Failure of participants to make a full submission will result in their disqualification.

4. Multiple submissions per group are not allowed. Only the first submission from an entrant before the deadline will be evaluated and all other submissions will be discarded.

5. The participants must make the source code of their implementation of their approach publicly available. Furthermore, any supporting libraries used by the participants must have publicly available source code.

6. The submitted source code must be able to compile and run on the evaluators' systems (see Section 7 for details on the available compilers and environments).

7. After submission, participants may be asked to validate their results by using their submitted source code to generate the same result files as they submitted earlier.

8. A Champion, First Runner-Up, and Second Runner-Up will be selected by the organizing committee. Details on the evaluation criteria are given in Section 10.

   The committee will also give adequate importance in the quality of the write-ups in determining the winners.

9. The prizes will be distributed as follows: $1500 for first place, $1000 for second place, and $500 for third place.

10. If a group is a winner of a prize, then the prize will be divided evenly among the group members.

11. If there exists a tie between two or more contestants, the committee has the right to redistribute the amount or number of awards.

12. In case of any dispute, the committee reserves the right to make a final decision in the matter.

6 Submission

Participants are required to email their submissions by the submission deadline in order to take part in the competition. A submission consists of:

- A write-up of the methods used not to exceed 10 8.5 by 11 pages using an 11pt font in a two-column format with 1-inch margins. The write-up should be in the style of a conference paper, and contain the full technical details of the algorithm used, its relation to previous work in text mining, and the details of the classification model that has been developed. labels of the test data set. The write-up should be written in English and be submitted in either Word, Postscript, or PDF format. paper explaining the details on their adopted method, in addition to the output files, according to the procedures written out in the "Submission" section. Any kind of failure to submit either the source code, output files, or technical paper in accordance with these procedures will be considered a violation of basic rules of this contest, and will result in immediate disqualification of the individual or group.

- The source code of the technique used. The source code that is submitted must be able to run on the evaluators' system. Details on the programming languages and environments that can be used are found in Section 7.

  Java, Matlab, etc., and should compile/run using standard compilers and interpreters, such as the GNU GCC tools, the current version of the Java Development Kit, Matlab or Octave, etc.

- The output file containing the participant's labeling of the test data set. The format of this file should be the same as that of the provided label file for the training and validation set. That is to say, it should be a file in comma-separated value format, where the jth value in line i of the file is 1 if the ith document in the test set should have label j, and -1 otherwise.

- The output file containing the participant's confidence in their labeling of the test data set. This file should also be in comma-separated value format, where the jth value in line i of the file is the participants confidence (a value between 0 and 1) that the corresponding value in the label file (1 or -1) is correct. That is to say, the confidence that document i should or should not have label j.

- The final submission should have a text file with the following information:

  1. Full Name
  2. Occupation
  3. Company/Institution
  4. Postal Address
  5. Telephone Number (including country/area codes)
  6. Email Address
  7. A statement that the participants agree that they are willing to make the source code of the implementation of their approach publicly available, and publish their submitted technical paper in the workshop proceedings.

- For group entries, contact information will be needed for all members of the group.

7 Evaluation Environment

The source code that is submitted must be able to run on the evaluators' systems. The first system is a cluster with the following hardware and software specifications:

- 64 Dual Intel Xeon EM64T 3.2GHz processors (each with 4GB RAM, 36GB Disk)
- 2Gbps Myrinet interconnect; MPI (mpich-mx) supported
- Linux Operating System: Debian 3.1, Kernel Version: 2.6.10
- GCC (C/C++) compiler suite: 3.3.5
- Java JDK/JRE: 1.5.0_05
- Perl: 5.8.4
- Python: 2.3.4

Other systems will be used to evaluate MATLAB submissions. The following versions of MATLAB and associated toolboxes are available:

- MATLAB Version 7.3.0.298
- MATLAB Compiler Version 4.5
- Neural Network Toolbox Version 5.0.1

- Optimization Toolbox Version 3.1
- Signal Processing Toolbox Version 6.6
- Statistics Toolbox Version 5.3
- Symbolic Math Toolbox Version 3.1.5

Please note that MATLAB submissions will not be evaluated on the cluster.

8 Competition Website

This document as well as the data sets and other information concerning this competition can be found on the web at: http://www.cs.utk.edu/tmw07. Also, at the URL http://www.cs.utk.edu/tmw06/Ashok-keyn.pdf one can find a presentation describing a previous approach to text mining this data set. Contact Murray Browne at mbrowne@cs.utk.edu., if you have additional questions.

9 Important Dates/Contest Schedule

- Contest begins now.
- Contest test data will be posted at Noon EST, ~~Monday, January 8, 2007~~ Wednesday, January 10, 2007.
- All submissions, including write-ups of results, are due by Noon EST, ~~Wednesday, January 10, 2007~~. Friday, January 12, 2007.
- Winners will be notified ~~Wednesday, January 17, 2007~~ Friday, January 19, 2007.
- Final papers from the winners are due February 16, 2007.

10 Cost Function

In describing the cost function, the following notation will be used:

1. D number of documents
2. C is the number of labels
3. A_j is the area under ROC curve for classifier predictions of label j. This quantity lies between 0.5 and 1.
4. t_{ij} is the target value for document i for label j. It is either $+1$ or -1, with the positive case indicating that the document has the corresponding label.
5. p_{ij} is the participant's prediction for document i of whether it has label j $(+1)$ or not (-1).

6. q_{ij} is the confidence in prediction for document i (must be between 0 and 1) of whether it should have label j.

7. Q is the cost function.

8. F_j is the frequency of documents having label j.

Participants must maximize the following cost function Q.
We define an intermediate cost function for label j as:

$$Q_j = (2A_j - 1) + \frac{1}{D}\sum_{i=1}^{D} q_{ij} t_{ij} p_{ij} \tag{1}$$

across the C categories. The index Λ defines one of three cases of weights defined below.
The final cost function is:

$$Q = \frac{1}{C}\sum_{j=1}^{C} Q_j \tag{2}$$

different aspects of the classification problem.
emphasizes the quality of the prediction while reducing the emphasis on the confidence of the predictions. emphasizes confidences in the predictions over the quality of the predictions.
The winner will be the team that achieves the largest value for Q. A first and second runner-up will be defined by those that achieve the second and third rank in terms of Q above.
A Figure of Merit will be assigned to each team using the following equation:

$$FOM = \frac{1}{C}\sum_{j=1}^{C} \frac{(F - F_j)}{F} Q_j \tag{3}$$

where:

$$F = \sum_{j=1}^{C} F_j. \tag{4}$$

This Figure of Merit will be used to break any potential ties, and measures the quality of predictions across all anomaly categories. Lower weighting is given to high frequency categories.

11 Disclaimers

The contestant takes the responsibility of obtaining any permission to use any algorithms/tools/data that are intellectual property of third party.