ASSIGNMENT III (CS 6474/CS 4803 Social Computing)

Due date	11:59pm, Nov 27, 2017
Topic	Build, replicate, and compare an abortion stance classifier based on Twitter data (see [1])
What to hand in?	i) 4-5 page report (double spaced, 11 point font, single column); ii) zipped/compressed folder of code
How/where to submit?	T-Square
Grade	10% [Total points: 100]

Dataset Source	www.munmund.net/courses/fall2017/notes/AbortionTwitterData.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 we covered how data arising in social computing systems can help us understand events, issues, and phenomena in the offline world. This assignment will test your understanding of applying these understandings to a hotly contested topic: the abortion debate in the United States.

Question

In this assignment you are supposed to build a classifier that will automatically infer the ideological stance of a tweet around the topic of abortion (for, against, and neural stances) and replicate and compare some of the findings given in Sharma et al. [1]. Refer to the table above to download the dataset. Your submission would include a 4-5 page report that discusses how you built the classifier, presents the performance of your classifier, and compares its performance with that given in [1]. You also need to submit your code/scripts in a compressed folder.

Note: as with Assignment II, 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 are most comfortable with.

Background and information about the dataset

As described in [1], the dataset contains 400 tweets (see raw_tweets.txt), posted between January 2015 and September 2016, coded by human annotators to indicate for, against, and neutral stances on the controversial topic of abortion.

Here's a brief summary of the dataset construction. Starting with the hashtag #abortion, frequently co-occurring or trending hashtags related to it were identified through a website called Hashtagify, a hashtag search engine. #prochoice and #prolife were two hashtags with the highest correlation with #abortion, and it inspired the choice to append #antilife and #antichoice to the seed list of hashtags to capture contrary ideologies on the abortion debate.

Thus, the final hashtags used to construct the dataset were: #abortion, #prochoice, #prolife, #antichoice, and #antilife.

For generating the labels of the set of 400 tweets, one human rater familiar with the abortion debate first examined a random sample of 200 tweets from the above dataset using an open coding approach, followed by employing an iterative process to categorize different tweets into "codes" relevant to the abortion debate. These hand-coded rules, referred to as "memos", were employed to create a codebook; this codebook contained the definitions of the codes, their correlations, and specific examples. See the files under the directory "memos" for a description – four memos are included here. These codes were then applied to annotate a second sample of 200 tweets into: *For Abortion* (tweets that voice support for abortion), *Against Abortion* (tweets that argue against the practice of abortion), and *Neutral to Abortion* (tweets that do not express an explicit stance on the issue). The enclosed file ground_truth.txt gives the assignment of these labels to each tweet present in raw_tweets.txt.

Contents of the report

(1) *Feature construction (10 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 tweets. Features can include (but not limited to) unigrams, bigrams, TF-IDF, Part-of-Speech tags, length of words etc. of the tweets. As used in [1] (see section 3.2.2), consider using the memos as features in your classification model. Summarily, you will need to discuss why you chose the particular set of features.

(2) *Description of the classifier (5 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 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 (5 points)*. Present how you evaluated how well your classifier of choice performed in distinguishing between the three abortion ideology classes. Particularly, discuss applicability of the concept of *k*-fold cross validation here. 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, etc.

(4) *Implementation (35 points).*

   a. Discuss how you preprocessed your data. If you used stopword removal, stemming, or tokenization over the content of the tweets, 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 raw tweets 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). Discuss how you implemented the memos as feaures. Also report if you did any kind of normalization or standardization of features, 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 your *k*-fold cross validation method, 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 (25 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 (like in [1]).

(6) *Replication and Comparison (20 points).* Discuss to what your classifier was able to replicate the performance of the abortion stance classifier given in [1] (see section 4.1). How does it compare? Is it as good? Is it worse? Is it better? Were the memos as features helpful? Justify your ideas behind your comparative observations.

References

[1] Sharma, E., Saha, K., Ernala, S. K., Ghoshal, S., & De Choudhury, M. (2017). *Analyzing Ideological Discourse on Social Media: A Case Study of the Abortion Debate.* In Proceedings of CSSA's Annual Conference on Computational Social Science (CSS) 2017. Link: http://www.munmund.net/pubs/CSS17_Abortion.pdf